

T.C.
İstanbul Üniversitesi
Sosyal Bilimler Enstitüsü
İşletme Anabilim Dalı
Sayısal Yöntemler Bilim Dalı
Doktora Tezi

**Veri Madenciliğinde Destek Vektör Makinaları ve
Karar Ağaçları Yöntemlerini Kullanarak Bilgi
Çalışanlarının Kurum Performansı Üzerine Etkisinin
Ölçülmesi ve Bir Uygulama**

Cemil KUZEY
2502080050

Tez Danışmanı
Prof. Dr. Öner ESEN

İstanbul, 2012



DOKTORA

TEZ ONAYI

ÖĞRENCİNİN

Adı ve Soyadı : CEMİL KUZİY Numarası :2502080050
Anabilim/Bilim Dalı : SAYISAL YÖNTEMLER Danışman : PROF.DR. ÖNER ESEN
Tez Savunma Tarihi : 05.03.2012 Saat :13:00
Tez Başlığı : Veri Madenciliğinde Destek Vektör Makinaları Ve Karar Ağaçları Yöntemlerini
Kullanarak Bilgi Çalışanlarının Kurum Performansı Üzerine Etkisinin Ölçülmesi Ve Bir Uygulama

TEZ SAVUNMA SINAVI, Lisansüstü Öğretim Yönetmeliği'nin 35. Maddesi uyarınca yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin **KABULÜ'NE** OYBİRLİĞİ / OYÇOKLUĞUYLA karar verilmiştir.

JÜRİ ÜYESİ	İMZA	KANAATİ (KABUL / RED / DÜZELTME)
1-PROF.DR. ÖNER ESEN		Kabul
2- PROF.DR İBRAHİM DOĞAN		Kabul
3-DOÇ.DR. MEHPARE TİMOR		Kabul
4-DOÇ.DR. EYÜP ÇETİN		Kabul
5-DOÇ.DR TUNÇHAN CURA		Kabul

**Veri Madenciliğinde Destek Vektör Makinaları ve Karar
Ağaçları Yöntemlerini Kullanarak Bilgi Çalışanlarının Kurum
Performansı Üzerine Etkisinin Ölçülmesi ve Bir Uygulama**
Cemil Kuzey

ÖZ

Bu tez çalışmasında, karar destek sistemleri ve veri madenciliği araçlarından olan karar ağaçları ve destek vektör makineleri (DVM) yöntemleri incelenmiştir. Karar ağaçları kolay yorumlanabilme ve sade görünüm özelliklerine sahip olduğundan popüler bir yöntemdir. Her iki yöntemde denetimli makine öğrenme (supervised machine learning) yöntemleridir. İstatistik öğrenme metotları istatistik ve makine öğrenimi alanlarının bir bileşimidir. DVM yöntemi istatistik öğrenme metotlarını içeren bir yöntemdir. Bilgi çalışanında olması gereken nitelikler keşfedici faktör analizi ile elde edildikten sonra, bu faktörlerin kurum performansı üzerindeki etkileri, karar ağaçları ve DVM modelleri ile detaylı olarak incelenmiştir. Çalışmada kullanılan veri kümesi, 2011 yılında bilgi çalışanları üzerinde yapılan anket yolu ile elde edilmiştir. Veri kümesinde 564 adet bilgi çalışanına ait kayıtlar mevcuttur. Ankette, bilgi çalışanı memnuniyeti 34 değişken ile ölçülürken, 4 değişken finansal ve 4 değişkende finansal olmayan kurum performansı için kullanılarak ölçülmüştür. Elde edilen ilgili veri kümesi kullanılarak DVM ve Karar ağaçları modelleri uygulanarak, modellerin performansları değişik araçlar ile kıyaslanmıştır. Analiz sonuçlarına göre, DVM ile C&R Tree karar ağacı modeli en yüksek doğruluk oranına sahip olduğu görülmüştür.

Measuring the Effect of Knowledge Workers on Organization Performance by Using Support Vector Machines and Decision Trees in Data Mining and an Application

Cemil Kuzey

ABSTRACT

In this thesis, decision trees and support vector machines methods, which are decision support systems and machine learning tools, are investigated. Decision trees are very popular methods because of its' simplicity to use and interpretability. Both of these tools are both supervised machine-learning methods. Statistical learning methods include statistics as well as machine learning. Support vector machines are methods that contain statistical learning tools. After obtaining the necessary attributes (factors) that knowledge workers should possess by using exploratory factor analysis (EFA), the effects of these attributes on organization's performance was examined in details by using decision tree as well as the support vector machine methods. Data set used in this study in 2011 was obtained through a survey carried out on knowledge workers. There are 564 records about knowledge workers in this data set. In the survey, the satisfaction of knowledge workers was measured by using 34 variables and the organization's performance was measured by using 4 variables for financial performance and by using 4 variables for non-financial performance. By applying SVM and decision tree methods on the acquired data set, the performances of the models were compared in terms of various performance measuring tools. According to the results of detailed analysis, SVM and the C&R Tree decision tree models had the highest performance results such as overall accuracy rates in many cases.

ÖNSÖZ

Bu tez çalışmasının amacı, popülaritesi gittikçe yükselen karar ağaçları ve destek vektör makineleri (DVM) modellerini kullanarak, bilgi çalışanlarında olması gereken niteliklerin belirlenmesi ve bu niteliklerin kurumların hem finansal hem de finansal olmayan performansları üzerindeki etkilerini detaylı olarak incelenmesini kapsamaktadır.

İnsan kaynakları sahasında yapılan uygulamaları incelendiğimiz zaman, DVM ve karar ağaçları modellerinin kullanımına pek az rastlanmaktadır. Çalışmanın birinci, ikinci ve üçüncü bölümlerinde destek vektör makineleri ile karar ağaçlarının temellerine ilişkin teorik bilgiler, önemli kavramların açıklanması ve modellerin algoritmik yapıları ele alınmıştır. Makine öğrenimi hem DVM hem de karar ağaçlarının temeli olduğundan, makine öğrenimi incelenmiştir. Daha sonra, bu tekniklerin uygulanması için insan kaynakları sahasında çalışanların memnuniyeti incelenmiştir. Uygulama için, bilgi çalışanları üzerinde bir anket yapılarak, sonuçların ilgili algoritmalar yardımı ile detaylı analizleri yapılmıştır. Kullanılan modeller sayesinde en yüksek oranda doğruluk düzeyini gösteren modeller tespit edilerek göreceli önem oranı yüksek faktörler elde edilmiştir.

Akademik hayatımda bana sabırla desteklerini eksik etmeyen, her zaman bana moral ve desteklerini eksik etmeyen değerli hocam, danışmanım sayın Prof. Dr. Öner ESEN'e teşekkür ederim. Akademik hayatıma uzun bir ara verdikten sonra benim tekrar bu alanda motive olmam için benden yardımlarını esirgemeyen Doç. Dr. Mehpere TİMOR'a ayrıca teşekkürlerimi sunmak isterim.

İstanbul, 2011

Cemil Kuzey

İÇİNDEKİLER

ÖZ	ii
ABSTRACT	iii
ÖNSÖZ	iv
İÇİNDEKİLER	v
TABLolar LİSTESİ	xiii
ŞEKİLLER LİSTESİ	xvi
KISALTMALAR LİSTESİ	xix
GİRİŞ	1
BİRİNCİ BÖLÜM	1
BİLGİ KEŞFİ VE MAKİNE ÖĞRENİMİ	1
1.1 Bilgi Keşfi Nedir (Knowledge Discovery)?	1
1.1.1 Veri Evrenin Yapısı	2
1.1.2 Tümevarımsal (Endüktif) Öğrenme	2
1.1.3 Tümevarımsal Öğrenme Hipotezi	3
1.2 Model Gösterimi	4
1.3 Makine Öğreniminde Temel Kavramlar	7
1.3.1 Veri Madenciliği	7
1.3.2 İşletme Veri Madenciliği	9
1.4 Makine Öğrenimi Tanımı	10
1.4.1 Makine Öğrenimi Çeşitleri ve kaynakları	13
1.4.2 Denetimsiz Öğrenme	14
1.4.3 Pekiştirmeli Öğrenme	15
1.4.4 Denetimli Öğrenme (Supervised Learning)	16
1.4.4.1 Eğitim Kümesi (Training Set)	17
1.4.4.2 Sınıflandırma Problemi	18
1.4.4.3 Performans Değerlendirmesi	20

1.4.4.3.1 Genelleme Hatası	21
1.4.4.3.2 Genelleme Hatanın Teorik Tahmini.....	22
1.4.4.3.2.1 VC- Çerçevesi (vapnik -Chervonenkis)	23
1.4.4.3.2.2 PAC- Çerçevesi (Probability Approximation Correct)	24
1.4.4.3.3 Genelleme Hatasının Ampirik Tahmini	25
1.4.4.3.4 Hesaplama Karmaşıklığı (Computational Complexity).....	28
1.4.4.3.5 Anlaşılabilirlik (Comprehensibility)	29
1.4.4.4 Büyük Veri Kümeleri İçin Ölçeklenebilirlik	30
1.4.4.5 “Boyutsallık Belası (Curse of Dimensionality)” Problemi	32
1.5 İstatistik Öğrenme Teorisi.....	34
1.5.1 Vapnik – Chervonenkis (VC) Boyut	34
1.5.2 VC- Boyut ve PAC Öğrenimi arasındaki ilişki	37
1.5.3 VC- Boyut ve Model Karmaşıklığı.....	38
1.5.4 Makine Öğrenimi İçin Teorik bir Çerçeve	42
1.5.5 Ampirik Risk Minimizasyonu (ARM)	43
1.5.6 VC- Güven.....	45
1.5.7 Yapısal Risk Minimizasyonu (YRM).....	48
İKİNCİ BÖLÜM	54
KARAR AĞAÇLARI	54
2.1 Karar Ağaçları (Decision Trees)	54
2.2 Karar Ağaçlarının Algoritmik Çerçevesi.....	57
2.3 Tek Değişkenli Bölme Kriterleri	58
2.3.1 Katıskı tabanlı (impurity-based) Ölçüt	58
2.3.2 Bilgi Kazancı (Information Gain).....	60
2.3.3 Gini İndeksi	60
2.3.4 Olabilirlik-Oran Ki-Kare İstatistikleri	61
2.3.5 DKM Ölçütü’	61
2.3.5 Normalleştirilmiş Katıskı Tabanlı Ölçüt	62
2.3.6 Kazanç Oranı (Gain Ratio)	62

2.3.6 Uzaklık Ölçümü (Distance measure).....	63
2.3.7 İkili Ölçütleri	63
2.3.8 Twoing Ölçütü.....	63
2.3.9 Ortogonal (ORT) Ölçütü.....	64
2.3.10 Kolmogorov – Smirnov Ölçütü’	65
2.3.11 AUC-Bölme Ölçütü	65
2.4 Çok Değişkenli Bölme Ölçütleri	66
2.5 Durdurma Ölçütleri	66
2.6 Budama Metotları	67
2.6.1 Maliyet-Karmaşıklık Budama (Cost – Complexity Pruning).....	68
2.6.2 İndirgenmiş Hata Budaması (Reduced Error Pruning).....	69
2.6.3 En düşük Hata Budaması (Minimum Error Pruning, MEP).....	69
2.6.4 Kötümser Budama (Pessimistic Pruning).....	70
2.6.5 Hata – Tabanlı Budama (Error-Based Pruning, EBP)	71
2.6.7 Optimal Budama.....	72
2.6.8 Minimum Tanım Uzunluğu Budama (Minimum Description Length, MDL)	72
2.7 Budama Metotlarının Kıyaslanması.....	73
2.8 Eksik Değerler (Missing Values).....	73
2.9 Yanlış Sınıflandırma Maliyetleri	76
2.10 Örnekleri Ağırlıklandırma	76
2.11 Karar Ağaçlarının Avantajları ve Dezavantajları	76
2.12 Karar Ağaçları İndükleyicileri.....	77
2.12.1 CHAID.....	77
2.12.2 ID3	78
2.12.3 C4.5.....	79
2.12.4 QUEST	79
2.12.4.1 Model Parametreleri.....	80
2.12.4.1.1 Alan seçimi.....	80

2.12.4.1.2 Bölünmüş Nokta Seçimi.....	81
2.12.5 CART(C&R Tree)	82
2.12.5.1 Algoritma Nasıl Çalışır?	84
2.12.5.2 CART Algoritmasında Ağaç Bölme Kuralları	86
2.12.5.3 Önsel Olasılıklar ve Sınıf Dengeleme.....	88
2.12.5.4 Eksik Değer İşlemleri	90
2.12.5.5 Öznitelik Önemi.....	92
2.12.5.6 Dinamik özellik İnşası	92
2.12.5.7 Maliyet – Duyarlı Öğrenim.....	93
2.12.5.8 Durdurma Kuralları, Budama, Ağaç dizileri ve Ağaç Seçimi	94
2.12.5.9 Olasılık Ağaçları	96
2.12.5.10 Teorik Temelleri	97
ÜÇÜNCÜ BÖLÜM	99
DESTEK VEKTÖR MAKİNELERİ (DVM)	99
3.1 Destek Vektör Sınıflandırıcı (DVS)	100
3.2 Yumuşak Marjlı DVS ve Optimizasyon	104
3.3 Çekirdek hilesi (Kernel trick)	107
3.4 Mercer Teoremi.....	110
3.5 Teorik Altyapısı	112
3.5.1 Teorem (Vapnik ve Chervonenkins)	114
3.5.2 Marj Tanımı	114
3.5.3 Fonksiyon Sınıfın Örtüsü	115
3.5.4 Marjlı VC Teoremi	116
3.5.5. DVS'nin Genelleme Sınırı	116
3.6 Destek Vektör Regressör (DVR).....	117
3.7 Yazılım Uygulamaları	120
3.8 Mevcut ve gelecekteki Araştırmalar	121
3.9 İşlemsel Etkinlik	121
3.10 Çekirdek Seçimi.....	122

3.11 Genelleme Analizi.....	123
3.12 Rademacher Karmaşıklığı.....	123
3.13 Karmaşıklık Analizi	124
3.14 Yapısal DVM Öğrenimi	124
3.15 Doğrusal Olmayan Destek Vektör Makineleri	126
DÖRDÜNCÜ BÖLÜM	128
ÇALIŞAN MEMNUNİYETİ	128
4.1 Çalışan Memnuniyetine Genel Bakış.....	128
4.1.1 Çalışan Memnuniyeti ve Performans	130
4.1.2 Çalışan Memnuniyeti ve Kurumsal Performans.....	131
4.1.3 Çalışan Bağlılığı, Kişisel Özellikler ve Çalışan Memnuniyeti.....	133
4.1.4 Çalışan Memnuniyeti ve İşten Ayrılma.....	135
4.1.5 Müşteri Memnuniyeti ve Hizmet Kalitesi	135
4.2 Çalışmalardaki memnuniyet faktörleri.....	136
4.2.1 Motivasyon	136
4.2.2 örgütsel vatandaşlık davranışları	138
4.2.3 İş Memnuniyetsizliği ve Tükenmişlik	139
4.2.4 Çalışma Koşulları	140
4.2.5 Demografik Faktörler	141
4.2.6 Kişilerarası Beceriler, Ücret	142
4.2.7 Çalışan Devri, Eğitim	144
4.3 Çalışan Memnuniyeti Anket Amacı.....	145
4.4 Bilgi İşçilerinin Memnuniyeti.....	145
4.4.1 Bilgi Birikimi ve Örgüt Performansı	145
4.4.2 Bilgi Çalışanı tanımı	146
4.4.2.1 UNESCO Tanımı	149
4.4.3 Bilgi işçilerinin gruplandırılması.....	149
4.4.4 Bilgi Çalışan Verimliliği ve kurum performansı	151
4.4.5 Bilgi işçilerinin Memnuniyetini Etkileyen Faktörler.....	152

4.4.6 Bilgi çalışanı verimliliği	156
4.4.7 Bilgi İşçilerinin Gruplandırılması.....	157
4.4.8 Bilgi işi ve işçilerinin özellikleri.....	158
4.4.9 Performans	161
BEŞİNCİ BÖLÜM	163
DVM VE KARAR AĞAÇLARI YÖNTEMLERİ KULLANILARAK BİLGİ	
ÇALIŞANININ KURUM PERFORMANSI ÜZERİNDEKİ ETKİSİNİN	
İNCELENMESİ	163
5.1 Analiz Aşamalarının şema ile gösterimi	163
5.2 Verinin Elde Edilmesi	165
5.3 Yöntem	165
5.4 Keşfedici Faktör Analizi (Exploratory Factor Analysis) Sonuçları	165
5.4.1 Keşfedici Faktör Analizi Modelinin Kurulması	167
5.4.2 Örneklem Büyüklüğünün İncelenmesi (KMO Ve Barlett Testi).....	170
5.4.3 Tanımlayıcı İstatistik Sonuçlarının değerlendirilmesi.....	171
5.4.4 R-Matrisi (Korelasyon Matrisi)	173
5.4.5 Faktörlerin Öz vektör Değerleri.....	174
5.4.6 Oransal Ortak Etken Varyans (communality)	175
5.4.7 Çoğaltılmış Korelasyon Matrisi (Reproduced Correlation Matrix).....	177
5.4.8 Faktörlerin Belirlenmesi (Rotated Component Matrix)	179
5.4.9 Güvenilirlik Analizi (Cronbach's Alfa).....	182
5.4.9.1 Tüm Çalışan Memnuniyetinin Güvenilirliğinin İncelenmesi	182
5.4.9.2 Faktörlerin Ayrı ayrı Güvenilirliklerinin Hesaplanması.....	184
5.5 Tanımlayıcı İstatistik Analiz Sonuçları.....	188
5.6 Önerilen bilgi çalışanı memnuniyetinin kurum performansı üzerindeki etkinliği modeli	192
5.7 Finansal ve Finansal Olmayan Kurumsal Performansların Başarılı ve Başarısız olarak Kümelenebilmesi	194
5.7.1 Aşamalı Kümeleme Yöntemi (Two Steps Clustering)	194

5.7.2 Kümelemede Analizde kullanılan Önem (Importance) Ölçütünün anlamı	195
5.7.3 Finansal kurum performans algısının “Başarılı” ve “Başarısız” olarak kümelenmesi	195
5.7.4 Finansal olmayan kurum performans algısının “Başarılı” ve “Başarısız” olarak kümelenmesi	196
5.8 Finansal performans algısının DVM ve karar ağaçları ile incelenmesi	197
5.8.1 Modellerin performans açısından kıyaslanması	198
5.8.1.1 Modellerin performanslarının kazançlar (Gains) grafiği ile gösterimi	202
5.8.1.2 Modellerin yanıt (response) grafiği ile gösterimi	203
5.8.1.3 Modellerin “Lift” grafiği ile gösterimi	204
5.8.1.4 Modellerin Özet Bilgileri	205
5.8.2 Finansal kurum performans algısının Destek Vektör Makineleri (Support Vector Machine) modeli yardımı ile detaylı incelenmesi	206
5.8.2.1 Destek Vektör Makineleri Modelinin Kernel Fonksiyonlarının performanslarının Kıyaslanması	208
5.8.2.2 Çok Terimli (Polynomial) Kernel Fonksiyonu Kullanılarak Destek Vektör Makineleri Yardımı İle Faktörlerin İncelenmesi	211
5.8.2.3 RBF Kernel Fonksiyonu Kullanılarak Destek Vektör Makineleri Yardımı İle Faktörlerin İncelenmesi	215
5.8.2.4 Lineer Kernel Fonksiyonu Kullanılarak Destek Vektör Makineleri Yardımı İle Faktörlerin İncelenmesi	218
5.8.2.5 Sigmoid Kernel fonksiyonu Kullanarak DVM Modelinin İncelenmesi	221
5.8.3 Finansal Kurum Performans Algısının Karar Ağaçları Modelleri Yardımı İle Detaylı İncelenmesi	223
5.8.3.1 C&R Tree Karar Ağacı Modelinin Detaylı İncelenmesi	223
5.8.3.2 C 5.0 Karar Ağacı Modelinin Detaylı İncelenmesi	236
5.8.3.3 QUEST Karar Ağacı Modelinin Detaylı İncelenmesi	240

5.8.3.4 CHAID Karar Ağacı Modelinin Detaylı İncelenmesi	245
5.9 Finansal Olmayan performans algısının DVM ve karar ağaçları ile incelenmesi	249
5.9.1 Modellerin performans açısından kıyaslanması	251
5.9.1.1 Modellerin Kazançlar (Gains) Grafiği	252
5.9.1.2 Modellerin yanıt (response) grafiği ile gösterimi	253
5.9.1.3 Modellerin “lift” grafiği ile gösterimi	254
5.9.1.4 Modellerin doğru Tahmin etme oranları	255
5.9.2 Finansal Olmayan Kurum Performans Algısının DVM Modeli İle İncelenmesi	256
5.9.2.1 Radyal Tabanlı (RBF) Kernel Fonksiyonlu DVM modeli	256
5.9.3 CHAID Karar Ağacı Modeli İle Finansal Olmayan Kurum Performansın İnceleme.....	260
5.9.4 C&R Tree Karar Ağacı Modeli İle Finansal Olmayan Kurum Performansın İnceleme.....	268
5.9.5 QUEST Karar Ağacı Modeli İle Finansal Olmayan Kurum Performansın İnceleme.....	275
5.9.6 C 5.0 Karar Ağacı Modeli İle Finansal Olmayan Kurum Performansın İnceleme.....	280
SONUÇ.....	285
KAYNAKÇA	291
EKLER.....	306
EK-1: SPSS-CLEMENTINE v.12 Analiz Özet Bilgileri.....	306
EK-2: ANKET SORULARI.....	310
ÖZGEÇMİŞ.....	318

TABLÖLER LİSTESİ

Tablo 1. Basitleştirilmiş Ağaç Büyütme Algoritma şeması	84
Tablo 2. Basitleştirilmiş budama algoritması	85
Tablo 3. Çalışan Memnuniyeti Anket Soruları ve ilgili değişkenler	166
Tablo 4. SPSS Clementine Faktör Analiz Ayarları	168
Tablo 5. KMO ve Barlett's Test Sonucu	171
Tablo 6. Tanımlayıcı İstatistik Sonuçları	172
Tablo 7. Korelasyon Matrisi (R-Matrisi)	174
Tablo 8. Faktörlerin Öz vektör Değerleri	175
Tablo 9. Oransal Ortak Etken Varyans (communality)	176
Tablo 10. Çoğaltılmış Korelasyon Matrisi (Reproduced Correlation Matrix)	178
Tablo 11. Bilgi Çalışan Memnuniyeti Faktörlerin Belirlenmesi	180
Tablo 12. Tüm Değişkenler için güvenilirlik analiz incelemesi	183
Tablo 13. Faktörlerin güvenilirliğinin gösterimi	184
Tablo 14. Yönetimin Tavrı faktörünün güvenilirlik analiz incelemesi	185
Tablo 15. Ücret/Ödüllendirme faktörünün güvenilirlik analiz incelemesi	186
Tablo 16. İş arkadaşları faktörünün güvenilirlik analiz incelemesi	186
Tablo 17. Çalışma şartları faktörü güvenilirlik analizi incelemesi	187
Tablo 18. İş güvencesi faktörü güvenilirlik analizi incelemesi	187
Tablo 19. Demografik Dağılım	189
Tablo 20. Bilgi çalışan memnuniyeti faktörleri	193
Tablo 21. Finansal kurum performans algısının kümelenmesi	196
Tablo 22. Finansal olmayan performans algısının kümelenmesi	197
Tablo 23. Test veri Kümesinde (Testing Set) modellerin performanslarının kıyaslanması	202
Tablo 24. Eğitim veri Kümesinde (Training Set) modellerin performanslarının kıyaslanması	202

Tablo 25. Radyal Tabanlı (Radial based kernel function) DVM modelinin performans değerleri.....	208
Tablo 26. Çok Terimli (Polynomial) Kernel DVM modelinin performans değerleri ...	209
Tablo 27. Sigmoid Kernel DVM modelinin performans değerleri	209
Tablo 28. Lineer Kernel DVM modelinin performans değerleri	210
Tablo 29. Çok terimli (polynomial) -DVM Faktörlerin göreceli önem oranları (Finansal Performans)	213
Tablo 30. Çok terimli -DVM modeli tahmin sonuçları.....	214
Tablo 31. RBF(Radyal Tabanlı)-DVM Faktörlerin göreceli önem oranları (Finansal Performans)	216
Tablo 32. RBF(Radyal Tabanlı)-DVM modelinin tahmin sonuçları	218
Tablo 33. Lineer-DVM Faktörlerin göreceli önem oranları (Finansal Performans)	219
Tablo 34. Lineer-DVM tahmin sonuçları.....	221
Tablo 35. Sigmoid DVM Faktörlerin göreceli önem oranları (Finansal Performans) ..	222
Tablo 36. C&R Tree modelinin performans ölçüm sonuçları.....	223
Tablo 37. C&R Tree modelinde faktörlerin göreceli önem oranları (finansal performans)	224
Tablo 38. C&R Tree karar ağacı modelinin tahmin sonuçları	230
Tablo 39. C&R Tree modelinde düğümlerin performans bilgileri.....	232
Tablo 40. C&R Tree modelinin risk ölçüm sonuçları	236
Tablo 41. C5 modelinde faktörlerin önem oranları	238
Tablo 42. C5 modelinin tahmin sonuçları (finansal performans).....	240
Tablo 43. QUEST modelinde faktörlerin göreceli önem oranları	242
Tablo 44. QUEST modelinin tahmin sonuçları.....	245
Tablo 45. CHAID modelinde elde edilen faktörlerin göreceli önem oranları.....	247
Tablo 46. Finansal olmayan kurum performans algısına ait kayıtlar	249
Tablo 47. Test Veri Kümesi için Modellerin ikili sınıflandırma kıyaslama sonuçları ..	251
Tablo 48. Finansal olmayan kurum performansında kullanılan DVM modellerin doğru tahmin etme oranları	255

Tablo 49. Finansal olmayan kurum performansında kullanılan karar ağaçları (CHAID, C&R Tree, QUEST, C5) modellerin doğru tahmin etme oranları	256
Tablo 50. Finansal olmayan kurum performans için RBF-DVM modeli	257
Tablo 51. RBF-DVM modeli tahmin sonuçları (Finansal olmayan kurum performans)	259
Tablo 52. CHAID karar ağacı modelinde faktörlerin göreceli önem oranları	261
Tablo 53. CHAID karar Ağacı Modelinin Kazançlar (Gains) açısından performansının İncelenmesi	266
Tablo 54. CHAID karar ağacı modelinin tahmin sonuçları	267
Tablo 55. C&R Tree modeli ile elde edilen faktörlerin göreceli önem oranları	270
Tablo 56. C&R Tree karar ağacı modeli tahmin sonuçları	275
Tablo 57. QUEST karar ağacı modelinde faktörlerin göreceli önem oranları	276
Tablo 58. QUEST karar ağacı modelinin tahmin sonuçları	280
Tablo 59. C 5.0 Karar ağacı modelinde faktörlerin göreceli önem oranları.....	281
Tablo 60. C 5.0 karar ağacı modelinin tahmin sonuçları	284

ŞEKİLLER LİSTESİ

Şekil 1. VC-Boyut.....	35
Şekil 2. Eksen hizalı dikdörtgen dört noktayı parçalara ayırabilir. Sadece iki noktayı içeren dikdörtgenler gösterildi.	37
Şekil 3. γ_1 marjlı sınıflandırıcıların bir sınıfının VC-boyutu	39
Şekil 4. γ_2 marjlı sınıflandırıcılar verilen veri kümesini parçalayamıyor	40
Şekil 5. γ_2 marjlı sınıflandırıcılar veri kümesinin bir altkümesini parçalayabilir. Burada açık gri renkli nokta silinen örneği temsil ediyor.....	40
Şekil 6. Marj ne kadar küçük olursa olsun, herhangi bir model sınıfın sınıflandırıcıları tarafından veri kümesi parçalanamaz.....	41
Şekil 7. Deneysel risk ve VC-güven arasındaki ilişki. Bunların toplamı geneleme sınırını verir	47
Şekil 8. VC-Güven h' de monotoniktir.....	48
Şekil 9. Yapısal risk minimizasyonunu kullanarak f_2^* optimal model bulmak	51
Şekil 10. VC boyut tarafından sıralandırılmış fonksiyonların iç içe geçmiş altkümeleri	52
Şekil 11. Doğrudan postalamaya cevap	55
Şekil 12. CART budama prosesinde bir safha: 17- terminal düğümlü alt ağaç. Vurgulanan düğümler sonraki aşamada budanacak olan düğümlerdir	96
Şekil 13. DVS de Doğrusal olarak ayrılabilen bir durum için Optimal hiperdüzlem gösterimi.....	100
Şekil 14. Analiz aşamaları	164
Şekil 15. SPSS Clementine ile faktör analizi ve kümeleme analizi gösterimi.....	168
Şekil 16. Bilgi çalışanları yaş dağılımı	192
Şekil 17. Önerilen bilgi çalışanının kurum performansı üzerindeki etkinliğini gösteren model.....	193
Şekil 18. Finansal performans algısı incelemesinin SPSS Clementine ortamında modellemesinin gösterimi	198
Şekil 19. Modellerin ikili sınıflandırıcı düğümü ayarlanması	199

Şekil 20. Modellerin performanslarının kazançlar (Gains) grafiği	203
Şekil 21. Modellerin yanıt (response) grafiği	204
Şekil 22. Modellerin “Lift” grafiği	205
Şekil 23. DVM modeli.....	207
Şekil 24. Çok terimli (polynomial) -DVM Faktörlerin göreceli önem oranları (Finansal Performans)	213
Şekil 25. RBF(Radyal Tabanlı)-DVM Faktörlerin göreceli önem oranları (Finansal Performans)	216
Şekil 26. Lineer-DVM Faktörlerin göreceli önem oranları (Finansal Performans).....	219
Şekil 27. C&R Tree modelinde faktörlerin göreceli önem oranları (finansal performans)	225
Şekil 28. Eğitim veri kümesi (Training set) ile elde edilen C&R Tree karar ağacı modelinin gösterimi	228
Şekil 29. Test veri kümesi (Testing set) ile elde edilen C&R Tree karar ağacı modelinin gösterimi.....	229
Şekil 30. C&R Tree modelinin kazançlar grafiği	233
Şekil 31. C&R Tree modelinin "lift" grafiği.....	234
Şekil 32. C&R Tree modelinin yanıt (response) grafiği	235
Şekil 33. C5 karar ağacı modelinin özellikli ayarlarının gösterimi	237
Şekil 34. C5 modelinde faktörlerin göreceli önem oranları ve sıraları.....	238
Şekil 35. C5 modelinin karar ağacı olarak gösterimi.....	239
Şekil 36. QUEST modelinin özellikli ayarlarının gösterimi.....	241
Şekil 37. QUEST modelinin karar ağacı gösterimi	244
Şekil 38. CHAID modelinin özellikli ayarlarının gösterimi.....	246
Şekil 39. CHAID modeli karar ağacının gösterimi.....	248
Şekil 40. Finansal olmayan kurumsal performans algısına ait kayıtların sayısı	250
Şekil 41. Finansal olmayan performans algısı incelemesinin SPSS Clementine ortamında modellemesinin gösterimi	250
Şekil 42. Modellerin finansal olmayan performanslarının kazançlar (gains) grafiği	253

Şekil 43. Modellerin finansal olmayan performanslarının yanıt (response) grafiği	254
Şekil 44. Modellerin finansal olmayan performanslarının "lift" grafiği ile gösterimi...	254
Şekil 45. Finansal olmayan kurum performans algısı için CHAID karar ağacı modelinin gösterimi.....	264
Şekil 46. Finansal olmayan kurum performans algısı için C&R Tree karar ağacı modeli	274
Şekil 47. Finansal olmayan kurum performans algısı için QUEST karar ağacı modelinin gösterimi.....	279
Şekil 48. Finansal olmayan kurum performans algısı için C 5.0 karar ağacı gösterimi	283

KISALTMALAR LİSTESİ

ARM: Ampirik Risk Minimizasyonu

Bİ: Bilgi İşçisi

BT: Bilgi Teknolojileri

BY: Bilgi Yönetimi

C&R Tree: Classification and Regression Tree

CHAID: Chi-Square Automatic Interaction Detection

DVM: Destek Vektör Makinesi

DVS: Destek Vektör Sınıflandırıcı

EFA: Keşfedici Faktör Analizi (Exploratory Factor Analysis)

İKY: İnsan Kaynakları Yönetimi

ÖVD: Örgütsel vatandaşlık davranışları

PAC: Probability Approximation Correct

PLS: Kısmi En küçük kareler

RBF: Radyal Tabanlı Fonksiyon

ROC Curve: Alıcı işletim karakteristik eğrisi (Receiver Operating Characteristic Curve)

VC: Vapnik – Chervonenkis

YRM: Yapısal Risk Minimizasyonu

GİRİŞ

Karar ağaçları, ağaç tabanlı sınıflandırma modelleri oluşturur. Kayıtları gruplara sınıflandırır veya bağımsız değişkenlerin değerlerine bağlı olan, hedef (bağımlı) değişken değerinin tahminini yaparlar. Karar ağacı en yaygın kullanılan mantık metodudur. Bir karar ağacının yapısı takip etmesi ve anlaşılması nispeten kolay olduğundan dolayı, popülerliği oldukça yaygındır. Karar ağaçları tahmin problemlerinin tüm türleri için kullanılmasına karşın, bunlar özellikle sınıflandırma problemleri için popüler metotlardır.

DVM (Destek Vektör Makineleri), çok boyutlu bir uzayda farklı sınıfların kayıtlarını ayıran hiperdüzlemler oluşturarak, sınıflandırma işlemlerini gerçekleştiren öncelikli bir sınıflandırma metodudur. DVM hem regresyon hem de sınıflandırma işlemlerini destekler ve sürekli ve kategorik değişkenleri işleyebilir. Doğrusal olmayan karar sınıfları oluşturmak suretiyle regresyon ve sınıflandırma işlemlerini yapan DVM metodu, bu sınırların mevcut olduğu öznitelik uzayının yapısından dolayı, sınıflandırma ve regresyon işlemlerini yaparken büyük esneklik sağlar.

Birinci bölümde makine öğrenimine ait temel kavramlara yer verilmiştir. En temel kavramlardan olan bilgi keşfi, kısaca incelendikten sonra makine öğreniminin alt yapısını oluşturan veri madenciliği ile ilgili bilgi verilmiştir. Daha sonra, işletmelerde veri madenciliğinin kullanım alanlarından bahsedilmiştir. Karar ağaçları ve DVM algoritmalarının temelini oluşturan makine öğrenimi hakkında temel bilgiler, tanımlar ve makine öğreniminin çeşitleri hakkında detaylı bilgiler aktarılmıştır. İstatistik öğrenme, makine öğrenimi ve istatistiğin bileşimi olduğundan, bu bölümün son kısmında istatistik öğrenme teorisine yer verilmiştir. İstatistik öğrenme alanında temel kavramlar olan, Vapnik Chervonenkis boyutu, güveni, model karmaşıklığı, ampirik (deneysel) risk minimizasyonu, yapısal risk minimizasyonu gibi önemli kavramlar açıklanmıştır.

İkinci bölümde karar ağaçları algoritmalarına yer verildi. Karar ağaçlarının algoritmik çerçevesi belirlendikten sonra, bölme kriterleri üzerinde duruldu. Daha sonra, karar ağacının nerede durdurulması ile ilişkili olarak, durdurma ölçütlerinden bahsedildi. Karar ağaçlarının daha basit bir görünüm kazanması ve yorumlanabilmesinin daha kolay olması için kullanılan, budama (pruning) metotları ile ilgili geniş bilgi sonraki kısımda yer almaktadır. Ayrıca, budama metotlarının kıyaslanması anlatıldıktan sonra, karar ağaçlarının avantajları ve dezavantajları üzerinde durulmuştur. Bu bölümün son kısmında, bilinen ve yaygın olarak kullanılan CHAID, QUEST, C&R Tree gibi karar ağaçları algoritmaları açıklanmıştır.

Üçüncü bölüm, DVM hakkında genel bilgiler içermektedir. Öncelikle DVS(destek vektör sınıflandırma) tanımı yapılarak, DVM'nin sınıflandırma metodu yönü üzerinde durulduktan sonra, doğrusal olarak sınıflandırılması mümkün olmayan veri kümeleri için kullanılan kernel (çekirdek) fonksiyonlarının temelini oluşturan önemli kavramlar, tanımlar ve teoremler ele alınmıştır. DVM teorik alt yapısının ana elemanı olan, VC-teoremi ve marj kavramları açıklanmıştır. DVM yönteminin sınıflandırma yönü ile birlikte regresyon yönü kısaca açıklanmıştır. Bu modellerin uygulanmasında kullanılan bilgisayar yazılımları hakkında kısaca bilgi aktarıldıktan sonra, DVM'de mevcut olan dört çekirdek fonksiyonu (kernel functions) seçimi hakkında kısaca bilgi verilmiştir.

Dördüncü bölümde, DVM ve karar ağaçları modellerinin uygulandığı bir konu olan “çalışan memnuniyeti” hakkında literatürde yapılan çalışmaların detayları verilmiştir. Çalışan memnuniyeti kavramından ne anlaşıldığı konusunda literatürde mevcut olan belli başlı tanımlara yer verilerek incelenmiştir. Ayrıca, çalışan memnuniyeti ile genel manada performans arasındaki ilişki literatürde mevcut olan kaynaklar yardımıyla aydınlatılmıştır. Çalışan memnuniyetini etkileyen faktörler üzerinde durulmuştur. Bu çalışmada, özellikle bilgi çalışanı (ya da bilgi işçisi) kavramı üzerinde durulduğundan, literatürde bilgi çalışanı kavramının tanımı ve özellikleri

hakkında genel bilgiler verilmiştir. Aynı bölüm içerisinde, bilgi çalışanının memnuniyetini etkileyen faktörler, çalışanların verimliliği ile devam edilmiştir. Bilgi çalışanların kurum performansları üzerindeki etkinlikleri hakkında yapılan çalışmalara ve araştırmalara yer verilmiştir.

Son bölümde, DVM ve karar ağaçları modelleri, bilgi çalışanları memnuniyeti üzerinde uygulanmıştır. Öncelikle, bilgi çalışanında olması gereken nitelikler, faktör analizi yardımıyla belirlendikten sonra, bu faktörlerin hem finansal hem de finansal olmayan kurum performansı üzerindeki etkileri DVM ve karar ağaçları modelleri yardımıyla detaylı olarak incelenmiştir. Analizler yapılırken, modellerin performanslarının kıyaslamaları detaylı olarak tablolar ve şekiller ile gösterilmiştir. Elde edilen sonuçlara göre, hangi modellerde hangi faktörlerin göreceli olarak etken faktörler olduğu elde edilmiştir.

BİRİNCİ BÖLÜM

BİLGİ KEŞFİ VE MAKİNE ÖĞRENİMİ

1.1 Bilgi Keşfi Nedir (Knowledge Discovery)?

Bilgi keşfi el ile incelenmesi çok zor olan çok büyük bir veri toplamasından faydalı bilgi çıkarımı yarı otomatik bir sürecidir. Yarı otomatiktan kasıt keşif sürecinde bilgisayar destekli araçlar kullanılır fakat bir analizcinin bu süreci yönlendirmesi kaçınılmazdır. Keşif süreci yolu ile alınan bilgi açıklayıcı kalıplar şeklini alırlar ve buna biz genellikle “modeller” adı verilir. Tüm modeller, gereksiz detayları göz ardı etme ve verideki ana trendleri özetleme eğiliminde olması gibi istenilen özelliklere sahiptirler. Bir model, terabaytlık bir veriyi özetleyebilir ya da temsil edebilir. Böylece, modeller büyük miktardaki veride saklı olan bilgiye ulaşma imkânı sağlar. Modellerden bir tanesi destek vektör makineleridir. Bilgi keşfi (Knowledge Discovery), çoğu zaman veri madenciliği (Data Mining) ile ilişkilidir. Veri madenciliği, veri tabanlarından bilgi çıkarımını hedefleyen bilgi keşfinin özel bir türü olarak ifade edilebilir. Ayrıca, veri madenciliği çoğu zaman, veri tabanlarında bilgi keşfi (Knowledge Discovery in Databases, KDD) olarak belirtilir¹.

Birçok yazar veri madenciliği ile veriden bilgi keşfi (Knowledge Discovery from Data, KDD) terimlerini benzer bir şekilde kullanmaktadırlar. Alternatif olarak, diğerleri, veri madenciliğini bilgi keşfi sürecinde esaslı bir adım olarak görmektedirler. Bilgi keşfi aşağıda belirtildiği gibi bir süreçten meydana gelir. Bilgi keşfi tekrarlamalı bir dizidir².

Bilgi keşfi süreci:

- veri temizleme

¹Lutz Hamel, **Knowledge Discovery With Support Vector Machines**, New Jersey, John Wiley & Sons Inc, 2009, p.3.

² Jiawei Han, Micheline Kamber, **Data Mining Concepts and Techniques**, San Fransisco, Elsevier, 2006, pp.5-7

- veri birleşimi
- veri seçimi
- veri dönüşümü
- veri madenciliği
- kalıp değerlendirilmesi
- bilgi sunumundan oluşmaktadır.

Veri tabanında bilgi keşfi birçok farklı adımlardan oluşan iyi tanımlanmış bir süreçtir. Veri madenciliği çok büyük veri tabanlarından gizli fakat faydalı bilginin keşfedilmesi ile sonuçlanan bir çekirdek adımdır. Bilgi keşfinin resmi bir tanımı şu şekildedir: “Bilgi madenciliği veri hakkındaki önceden bilinmeyen üstü kapalı fakat potansiyel olarak faydalı bilginin önemsiz olmayan çıkarımıdır” (Frawley, Piatetsky-Shapiro, 1996).

1.1.1 Veri Evrenin Yapısı

Bir veri evreninde, nesneler genellikle *özelliklerin* ya da *özniteliklerin* (değişkenler) bir derlemesi olarak ifade edilir. Nesnelerin bir kümesini tablo ile en yaygın şekilde ifade edilir. Tabloda her bir sütun bir özelliği ifade ederken, her bir sıra ise bir nesneyi ifade etmektedir.

1.1.2 Tümevarımsal (Endüktif) Öğrenme

Tümevarımsal öğrenme metodu, etki alanı ya da tanım bölgesi örneklerinden (domain examples) elde edilen özellikli bilgiden, genel etki alanlı veya tanım bölgesi bilgi elde eden yöntemler olarak tanımlanır³. Genel

³ “Object-oriented Induction” , (Çevrimiçi)<http://www.iiia.csic.es/Projects/FedLearn/OO-Induction.html>, 2011

olarak, endüktif öğrenme metotları, bir hipotez uzayı üzerinde arama metotları olarak karakterize edilebilir.

Makine öğrenimi tanımında endüktif bir süreç ifade edilir. Burada, sınırlı bir miktarda veri, eğitime kümesi olarak tanımlanır ve orijinal etiketlendirme sürecini tüm veri evreni üzerine yaklaştıran bir fonksiyon indüklenir. Yani, D eğitime kümesindeki özellikli örneklerden tüm X veri evrenine *genelleme* yapılmış olur. Bu süreç tümevarımsal öğrenme ya da endüktif öğrenme olarak adlandırılır. Tümevarımsal öğrenmenin kalbinde eğitime kümesi tüm evrenin doğru bir temsilcisi olduğu varsayımı yatmaktadır. Bu varsayım aşağıda ki gibi ifade edilir.

Endüktif bir algoritma ise “öğrenici” olarak bilinir. Endüktif algoritma, bir eğitime kümesi elde eder ve girdi öznitelikleri ve hedef öznitelikler (değişkenler) arasındaki ilişkiyi genelleştiren bir model oluşturur⁴.

1.1.3 Tümevarımsal Öğrenme Hipotezi

Hedef fonksiyonunu çok iyi bir şekilde, yeterince büyük eğitime örnekler kümesi üzerine yaklaştıran bir fonksiyon, hedef fonksiyonunu da çok iyi bir şekilde gözlemlenmeyen örnekler üzerine yaklaştıracaktır. Tümevarımsal öğrenmenin karışıklığı, “siyah kuğu problemi” ile açıklanabilir. Burada X kümesi, siyah ve beyaz mümkün olan tüm kuğuların evrenini ifade etsin ve D kümesi ise bir makine öğrenimi algoritması için eğitime kümesini göstere. Bu eğitime kümesinden, bir öğrenme algoritması, tüm kuğuların “beyaz” olduğunu gösteren bir model sonucuna varabilir, yani;

⁴ Maimon, O., Lior Rokach, “Introduction To Supervised Methods”, **Data Mining and Knowledge Discovery Handbook**, Ed. by Lior Rokach and Oded Maimon, 2005, Springer Science+Business Media Inc, p,135.

$$\hat{f}(x) = beyaz$$

Tüm $x \in X$ için. Bu açık şekilde orijinal sürece sadece bir yaklaşımdır,

$f: X \rightarrow \{beyaz, siyah\}$, bu birçok kuşu beyaz olarak etiketler, fakat aynı zamanda bazı kuşları ise siyah olarak etiketler. Modelimiz $\hat{f}$, kuşların rengine ait bilimsel soruları cevaplamak için zayıf bir seçim olur. Fakat eğer dünyada kuşların 99% beyaz ise, modelimiz tüm evren üzerinde hesaplandığı zaman 99%'luk bir doğruluğa sahip olur. Eğer kuşların renginin bir yaklaşımını istersek, bu da bize modelimiz çok iyi olduğunu gösterir.

Tümevarımsal öğrenmenin sorusu ve bir eğitime kümesinin veri evereninin iyi bir temsilcisi olup olmadığı konusu, takip eden modellerin uygulanmasına bağlıdır ve kesin bir önerme değildir. Veri evreninin bir temsilcisi olan, mümkün eğitime kümesi oluşturmak, istenilen bir durumdur. Yani “yeterince büyük” bir eğitime kümesi oluşturmak, çok istenilen bir durumdur. İstatistik örnekleme teorisinden faydalanılarak ve karmaşık teknikleri kullanılarak eğitime verisinin yeterince büyük olduğundan emin olabiliriz. Fakat eğitime kümesindeki nesneler hakkında her zaman bazı belirsizlikler eninde sonunda olacaktır. Bu belirsizlikleri, modelimizin genelleme yeteneğini ve beklenen doğruluğunu değerlendirmek için yardımcı teknikler kullanılır.⁵

1.2 Model Gösterimi

Hedef fonksiyon olan f 'nin, yaklaşım fonksiyonu olan $\hat{f}$ 'nin hesaplanabilir olması istendiğinden, $\hat{f}$ fonksiyonu, modellerinin uygun gösterimleri ile ilgilenilir. Tipik olarak, model gösterimi iki şekilde gösterilir:

- 1) Şeffaf gösterimler (Transparent)
 - a. “If-then-else” kuralları
 - b. Karar ağaçları

⁵ Hamel, a.g.e., p.9.

2) Şeffaf olmayan gösterimler (Nontransparent)

- a. Bir yapay sinir ağında elemanlar arasındaki bağlantılarda ağırlıklar
- b. Destek vektör makinelerinde (Support Vector Machines) vektörlerin lineer bileşimi

Şeffaf modellerin gösterimleri kolay olduğundan insanlar tarafından yardım edilmeden yorumlanabilir; şeffaf olmayan modeller yardımsız yorumlanamayabilir. Örneğin, “If – then –else” kurallarını kural metnine bakarak kolaylıkla yorumlayabiliriz. Diğer taraftan, tam olarak yapay sinir ağlarının öğrenilen bilgisini ağırlıklara nasıl depoladığını tam olarak anlamaksızın bir yapay sinir ağındaki ağırlıkları inceleyebiliriz.

Model gösterimi, belli hedef fonksiyonları nasıl modelleyebileceğimizi belirlediği için önemli bir konudur. Bu makine öğrenimi teorisinde, “dil meyilli (language bias)” olarak bilinir. Model gösterimi olarak bir veri tablosunu ele alalım. Bu durumda, model gösterimi basit olarak eğitime kümesini yansıtır ve dolayısıyla eğitime kümesindeki tüm nesneleri ezberler. Fakat tabloda olmayan nesneler için mantıklı sonuçlar bulmada başarısız kalacaktır. Böylece kullanılan model tablosundaki nesnelerin ötesinde bir genelleme yapamaz ve bundan dolayı veri evreni üzerinde orijinal etiketleme sürecinin bir yaklaşımı olarak zayıf bir seçimdir. Bu modelin kısıtlamaları, eğitime veri tablosunu, bizim model uygulaması olarak, seçmemizden kaynaklanmaktadır. Bir diğer model gösterimi olarak, modelin sabit içermesi durumudur. Modelin sunduğu nesne türü fark etmeksizin, her zaman aynı sabit cevabı üretecektir. Böylece model eğitime kümesinde hatalar yapacaktır. Ancak, eğer eğitime kümesi bir bütün olarak veri evreninin doğru bir gösterimi ise, modelin veri evreninde eğitime kümesindeki gibi aynı veya benzer doğruluğa sahip olacağını bekleyebiliriz. Böylece, model belli bir ölçüde genelleme yaptığını söyleyebiliriz; en azından birçok etiketi basit bir yapısına kodlar ve bu tek bilgi parçasını gözlemlenmeyen nesnelere etiketleri atamak için kullanır.

Karar ağaçları, sinir ağları ve destek vektör makineleri gibi model gösterimleri yukarıda bahsedilen iki uç durum arasında bulunmaktadır. Daha gelişmiş model gösterimlerine yol açan algoritmalar nesneleri ilgili etiketlere ilişkilendiren düzenlilikler keşfeder ve bu düzenlilikler daha sonra uygun model gösterimlerine kodlandır.

Genel olarak, şeffaf model gösterimleri şeffaf olmayan model gösterimlerine kıyasla performans açısından geri kalmaktadırlar. İnsanlar tarafından bir modelin yardımsız yorumlanabilmesi, kısıtların modelleme sürecine engel olmaktadır. Burada, bir şeffaf model belli olayları şeffaf olmayan modeller kadar etkin biçimde sınıflandıramamaktadır⁶.

Makine öğrenimi sadece bir veri tabanı problemi değildir; ayrıca yapay zekânın bir parçasıdır. Zeki olmak için, değişen bir çevre içinde bulunan bir sistem, öğrenme kapasitesine sahip olmak zorundadır. Eğer sistem öğrenebiliyorsa ve bu değişiklikleri uyarlayabiliyorsa, sistem tasarımcıları tahmin etmek ve mümkün tüm durumlar için çözümler sağlamak zorunda değillerdir. Makine öğrenimi, görme ile ilgili konuşma tanıma ve robotik alanlarında birçok problemler için çözümler bulmamıza yardımcı olmaktadır.

Makine öğrenimi örnek veri veya geçmiş tecrübeleri kullanarak bir performans kıstasını optimize etmek için bilgisayarları programlar. Bazı parametrelere kadar tanımlanmış bir model var olduğunu varsayalım. Öğrenme, eğitme verisini veya geçmiş tecrübenin kullanılması ile bilgisayar programının, modelin parametrelerini optimize etmesinden ibarettir.

Model gelecek ile ilgili tahmin yapmak için ön öngörücü olabilir ve ya veriden bilgi elde etmek için tanımlayıcıda olabilir veya her ikisi de olabilir. Bir örnekten

⁶ Hamel, a.g.e., p.11.

çıkarım yapmak temel görev olduğundan, makine öğrenimi istatistik teorisini matematik modeller oluşturmak için kullanır. Bilgisayar biliminin iki önemli yönü vardır. Birincisi, eğitim aşamasında optimizasyon problemini çözmek, büyük miktarda bir veriyi depolamak ve işlemek için verimli algoritmalara ihtiyaç vardır. İkinci olarak, model öğrenildikten sonra bunun temsil edilmesi ve algoritmik çözümlerin sonuçlarının çıkarılması etkin olmak zorundadır, yani bunun yer ve zaman karmaşıklığı tahmin doğruluğu kadar önemlidir⁷.

Öğrenme birliği (Learning Association)

Sınıflandırma (Classification)

Regresyon (Regression)

1.3 Makine Öğreniminde Temel Kavramlar

1.3.1 Veri Madenciliği

İlgilendiğimiz makine öğreniminde, yukarıda bahsedilen kavramsal problemler yer almamaktadır. Yani varsayımdan yola çıkan, gerçekte öğrenmenin ne olduğu ile ilgili belli bir felsefi duruşu olmayan bir öğrenmedir. Veri madenciliği, pratik bir konu olan ve teorik anlamda değil de pratik anlamda makine öğrenimini içeren bir konudur. Burada, veri madenciliği, veride yapısal kalıpları bulmak, açıklamak ve bu veriyi kullanarak tahmin yapmak için araç olarak teknikler ile ilgileniliyor. Veri madenciliğinde amaç, veride kalıpları bulmak ve açıklamaktır. Ayrıca, çıktı (output) kararı açıklamak için bilinmeyen örnekleri sınıflandırmak için kullanılan bir yapının gerçek tanımını da içerebilir. Performans açısından, elde edinilen bilginin açık bir gösteriminin sağlanması çok faydalıdır. Temelde bilginin edinilmesi ve bunu kullanma kabiliyeti “*öğrenme*” tanımını yansıtmaktadır. Birçok öğrenme teknikleri, öğrenilenin yapısal tariflerini arar. Bu tarifler oldukça karmaşık olabilirler ve kural setleri olarak ifade edilirler. Mesela, karar ağaçlarında olduğu gibi kural kümeleri olarak ifade edilir.

⁷ Ethem Alpaydın, **Introduction to Machine Learning**, MIT Press, 2004, pp.1-3.

Bu kurallar, insanlar tarafından anlaşılabilir. Ayrıca bu tarifler, ne öğrenildiğini açıklamak için bizlere hizmet ederler ve yeni tahminler için temel oluştururlar. Elde edilen tecrübeler, makine öğreniminin birçok uygulamalarında elde edilen açık bilgi yapıları ve yapısal tariflerin çok önemli olduğunu göstermiştir. İnsanlar, veri madenciliğini bilgi sahibi olmak için sıklıkla kullanırlar, sadece tahmin yapmak için değil⁸.

Veri madenciliği “veriden, potansiyel olarak faydalı, daha önce bilinmeyen ve üstü kapalı bilgiyi çıkarım” olarak ilk yıllarda (Frawley v.d., 1992) tarafından tanımlandı. Daha sonraki yıllarda veri madenciliği tanımı, bu tanım üzerinden birazcık genişletildi. Yeni tanımı, “bir veri kümesinde kalıpların veya ilişkinin bulunması için çeşitli algoritmaların uygulaması” olarak ifade edildi (Fayyad v.d., 1996). Veri madenciliği kavramı, bir işletme veri analizcisi için, sadece ilişkileri bulmak olmadığı gibi, veri madenciliği tanımı, verinin gerekli önışlenmesi, sonuçların yorumlanması ve karar verme aşamasında faydalı bir biçimde elde edilen bilginin sağlanması aşamalarını içermektedir. Diğer bir deyişle, bir işletme veri analizcisi klasik veri madenciliği tanımını ve bilgi keşfini bir süreçte toplamaktadır⁹.

Veri madenciliğinde ilk önce, verinin temizlenmesi, eksik değerlerin incelenmesi ve bazı türetilmiş değişkenlerin oluşturulması ile başlanılır. Daha sonra, “gerçek” modeldeki bazı önemli değişkenlerin belirlenmesi için basit algoritmalar kullanılarak ön modelleme ile devam edilir. Veri madencisi daha sonra, model parametrelerinde ya da değişkenlerde bazı ayarlamalar (yeniden kodlama gibi) yapar ve modeli “arıtır” (yeni bir versiyonunu oluşturur). Veri madencisi, modeli değiştirmeye ve

⁸ Ian H. Witten, Eibe Frank, **Data Mining: Practical Machine Learning Tools and Techniques**, San Francisco, CA, Elsevier Inc., 2005, p.9.

⁹ Robert Nisbet, John Elder, Gary Miner, **Handbook of Statistical Analysis and Data Mining Applications**, Elsevier, 2009, p.34.

uyumlu hale getirme işlemlerine, modelin tahmin edilebilirliğinde daha fazla bir ilerleme olamayana kadar devam edilir¹⁰.

Veri madenciliği, büyük miktardaki verilerin veri tabanlarında, veri ambarlarında ya da diğer bilgi depolarında toplanabildiği yerlerde, ilginç kalıpların keşfedilmesi işidir. Veri tabanı sistemleri, veri ambarı, istatistik, makine öğrenimi, veri görüntüleme, bilgi erişimi (alımı), ve yüksek-performanslı bilgi işlem gibi disiplinler arası yeni bir alandır. Katkıda bulunan diğer alanlar, sinir ağları (neural network), örüntü tanıma (pattern recognition), uzaysal veri analizi (spatial data analysis), resim veri tabanı, sinyal işleme (signal processing) ve işletme, ekonomi ve biyoinformatik gibi diğer uygulama alanlarıdır¹¹. Veri madenciliği, geleneksel veri analizinin aksine, buluş odaklıdır. Ayrıca, bu alan veri ambarı, makine öğrenimi, sinir ağları ve tümevarımsal mantık programlama interdisipliner alanlar ile yakından ilişkilidir. Veri madenciliği otomatik örüntü (patern) tanıma imkânı sağlar ve verideki geleneksel istatistik yöntemleri ile ortaya çıkarılması zor olan örüntüyü açığa çıkarmaya çalışır¹².

1.3.2 İşletme Veri Madenciliği

Veri madenciliği birçok işletme mekânlarında çok etkili olmuştur. İşleme dönüştürülebilir bilgilerin bulunması önemlidir. Kârlılığı geliştirmek için, somut bir şekilde kullanılabilen bilginin elde edilmesi işletmelerde çok önemlidir. Bazı ilk uygulamalar perakendecilik alanında yapılmıştır, özellikle pazar sepeti analizi (market basket analysis) şeklinde uygulanmıştır. Perakendecilik (retailing) uygulaması yanı sıra bankacılık, kredi kartı yönetimi, sigortacılık, telekomünikasyon, tele-pazarlama ve insan kaynakları yönetimi veri madenciliği uygulamalarındandır¹³.

¹⁰ Nisbet, Elder, Miner, **a.g.e.**, p.45.

¹¹ Han, Kamber, **a.g.e.**, p.40.

¹² Harleen Kaur, Siri Krishan Wasan, Empirical Study on Applicaitons of Data Mining Techniques in Healthcare, **Journal of Computer Science**, 2(2), 2006, pp.194-200.

¹³ David Olson, Dursun Delen, **Advanced Data Mining Techniques**, Springer-Verlag Berlin Heidelberg, 2008, p.7.

1.4 Makine Öğrenimi Tanımı

Makine öğrenimi olarak bilinen düşünce çizgisi, zeki makine arayışında, yapay zekâ toplumundan doğdu. Başlangıçta bu metotlar yapay sınır ağları ve karar ağaçları gibi iki gelişim yollarını takip etti. Birinci yol olan yapay sinir ağı, girdi değişkenlerine ağırlıklar atayarak, direkt olarak ifade etmek için doğrusal olmayan bir fonksiyon aradı (“neden”). Daha sonra, bunların etkisini biriktirerek ve bazı karar fonksiyonlarını takip ederek bir çıktı değeri (“sonuç”) elde etmek için “tepki” meydana getirir. Yapay sinir ağları sistemleri, insan beynini çalışma şekli ile benzerlik göstermektedir. Nöronların karmaşık ilişkileri ve beyin tarafından gerektiği gibi hızlı cevap vermek için bunların sinaptik ilişkili bağlantıları “öğretililebilir” ve “öğrenilebilirdir”. Bilgisayar bilimcileri, bir veri kümesinin girdi değişkenlerindeki karmaşık şablonları tanımak için kullanılan yapay sinir ağı biçimindeki sistemleri, ham terimler ile ifade etmeğe başladılar.

İkinci geliştirme yolu ise, “etkileri” direkt olarak ifade etmek ile ilişkili olan *karar ağaçlarıdır*. Bunun için, “kurallar” bulmak için metotların geliştirilmesi gerekmektedir. Bahsedilen kurallar, fonksiyonel ilişkiyi direkt olarak ifade etmeksizin girdi değerlerini birçok kutulardan bir tanesine ayırmak için hesaplanabilir. Bu metotlar, kuralları belirgin bir şekilde ifade etmeğe odaklanmaktadırlar ya da sonuçları ifade eden kurallar arasındaki (karar ağacı) ilişkiyi ifade etmeğe odaklanır. Bu metotlar, parametrik modellerin kısıtlamalarını önlemektedirler ve doğrusal olmayan olayların analizleri için çok uygun metotlardır¹⁴.

“Öğrenme” kelimesi sözlüklerde:

Çalışarak, tecrübe ile ya da öğretilerek bilgi elde etmek;

Bilgi ya da gözlem yolu ile haberdar olmak;

Bilgilendirilmek;

¹⁴ Nisbet, Elder, Miner, **a.g.e.**, pp.11-12.

Talimat almak.

Yukarıdaki tanımlar bilgisayar ile ilgili konularda yetersiz kalmaktadır. Çünkü makinenin bir şeyin bilgisini nasıl elde ettiğini bileceğiz ve makineye sorular sorulamayacaktır, sorulsa bile makinenin öğrenme kabiliyeti ve cevap verme kabiliyeti test edilemeyecektir. Bilgisayarların haberdar olabilmeleri veya bilinçli olabilmeleri soruları felsefi konulardır. Bu konular yerine, burada amaç, performansların geliştirilmesi ya da en azından performans için potansiyel ile ilgilenmektir. Yeni bilgiyi yeni durumlara uygulamadan, ezbere öğrenerek bir şeyden haberdar olunabilir. Hiç faydalanmadan talimat alınılabilir.

Veri madenciliği, çok büyük miktardaki verilerde, otomatik veya yarı otomatik olarak, işlevsel olarak kalıpları (patterns) keşfetme süreci olarak tanımlanır. Kalıplar faydalı olmalıdır. Öğrenim içinde aynı şekilde bir tanım yapılabilir: “Şeyler” davranışlarını gelecekte daha iyi performans gösterebilecekleri şekilde değiştirdikleri zaman öğrenirler.

Bu tanım, öğrenmeyi bilgidan ziyade performans ile bağlar. Öğrenme, davranış gözlemlenerek ve bunu geçmiş davranış ile kıyaslanması ile test edilebilir. Bu tanım daha objektif ve çok daha tatmin edici bir tanımdır. Fakat yine de problem vardır. Öğrenme konusu çok kaygan bir kavramdır. Birçok şey gelecekte daha iyi performans sergileyecek şekilde davranışlarını değiştirirler, fakat onların gerçekten öğrendiğini söylemek istemiyoruz. Örneğin, çok rahat terlikler ayak yapısını öğrenirler mi? Bir terlik olarak daha iyi bir performans göstermek için davranışını kesinlikle değiştirirler fakat buna öğrenme diyemeyiz. Bunun yanı sıra “*düşüncesiz*” öğrenme türünü belirten “*eğitme*” kelimesini sıklıkla kullanırız. Öğrenme, düşünme ve amaç anlamına gelir. Amaçsız öğrenme sadece eğitimidir. Bilgisayarlar ile ilgili öğrenme tanımında, operasyonel performans odaklı terimlerde kendine has bazı problemler vardır. Yani, bir şeyin gerçekten öğrendiğine karar vermek için bu şeyin belli bir amaca sahip olup

olmadığının bilinmesi gerekir. Makinelere bu tanım uygulandığında konu tartışmalı bir hal almaktadır¹⁵.

Makine öğreniminde temel soru “davranışlara dayalı olan kalıpları keşfetmek ve bunları açıklamak için bilgisayarlar kullanılabilir mi?”’dır. Bu durumu açıklamak için en kolay yol beklide sınıflandırma kullanılarak açıklanabilir. Belli bir objeye veya bir sınıfa aittir ya da değildir. Örneğin, bir kedi gördüğümüzde bunun memeliler sınıfına ait olduğunu biliyoruz. Objeler, eğer ilgili sınıfa ait ise, *doğru* olarak etiketleyen bazı durumlar ile ilişkili olan bazı süreçler mevcuttur. Eğer objeler sorudaki ilgili sınıfa ait değil ise *yanlış* olarak etiketlenir. Makine öğreniminin amacı, orijinal süreci mümkün olduğu kadar yaklaştıran bir etiketleme işlemi için, uygun bir modeli hesaplamaktır.

Tanım¹⁶: X , verilen bir veri evreni olsun ve S ise örnek kümesi olsun öyle ki , $S \subset X$. Bazı hedef fonksiyonları (etiketleme süreci) $f: X \rightarrow \{\text{doğru}, \text{yanlış}\}$ olsun. Ayrıca D etiketlenmiş bir eğitme kümesi olsun, öyle ki $D = \{(x, y) | x \in S \text{ ve } y = f(x)\}$. D kullanarak $\hat{f}: X \rightarrow \{\text{doğru}, \text{yanlış}\}$ gibi bir fonksiyon hesaplanır, öyle ki $\hat{f} \cong f(x)$, tüm $x \in X$ için. Burada, veri evreni olan X ilgili objelerinin bir kümesidir. Örnek kümesi olan S , veri evreninin bir alt kümesidir. Genel olarak, örnek kümesi gereklidir, çünkü ilgilendiğimiz birçok objelerin koleksiyonu, çok büyük veya sınırsız olma eğilimindedir. Çok büyük veri evrenleri için modeller oluşturmak, çok yavaş olabilir veya belki de sınırsız veri evreni için modeller oluşturmak imkânsız olabilir. Bundan dolayı, modeller oluşturma sürecini işlenebilir yapmak için, S örnek kümesi, veri evreninin bir temsilcisi olarak hareket eder. Hedef fonksiyonu olan f , gözlenebilir etiketler sağlayan bir süreçtir. Burada f , X deki herhangi bir eleman için, bu eleman gözlemlendiği zaman, $\{\text{doğru}, \text{yanlış}\}$ kümesinde uygun bir değer sağladığı varsayılır. Böylece, sürecin kendisine direkt olarak erişim olmamasına rağmen, bu veri evreninde, sürecin

¹⁵ Witten, Frank, **a.g.e.**, pp.7-9.

¹⁶ Hamel, **a.g.e.**, pp. 5-6.

elementlere atadığı etiketler her zaman gözlemlenebilir. Eğitim kümesi olan D' yi oluşturmak için, örnek küme S' deki objeler için etiketleri gözlemleyerek, hedef fonksiyonun bu özelliği kullanılır. Etiketlenmiş eğitim verisini kullanan makine öğrenimi, *denetimli (supervised learning)* olarak ifade edilir. Etiketlenmiş eğitim verisine ihtiyaç duymayan, makine öğrenme türüne ise *denetimsiz (unsupervised learning)* olarak adlandırılır. Makine öğrenimi tanımdaki denklem olan $\hat{f} \cong f(x)$, öğrenmenin, $\hat{f}$ fonksiyonunu bir yaklaşım olarak görülebileceğini ifade eder veya öğrenmenin, D deki eğitim örneklerine dayanan orijinal süreç olan f 'nin bir modeli olarak hesaplanması olarak görülebileceğini belirtir. Yani, makine öğreniminin sonucu, orijinal etiketleme fonksiyonunun bir modelidir. Fakat kolay olsun diye, çoğu zaman $\hat{f}$, eğitim verisi olan D 'nin bir modelidir denir. Bu ise, $\hat{f} \cong f(x)$ olarak belirtilen biçimsel görünüm ile uyumludur. Çünkü, eğitim verisindeki elemanlar orijinal etiketleme fonksiyonunun girdi-çıkı çiftleridir $(x, y) \in f, x \in S$ ve $y=f(x)$, ve bunun manası, f fonksiyonunu ve D eğitim verisini modellemek tek ve aynıdır. *{doğru, yanlış}* kümesindeki etiketlerin isimleri keyfidir, bunların yerine D ve Y veya 0 ve 1 'de yazılabilirdi. Burada önemli olan, bu küme iki farklı etiket içermektedir; bir tanesi sınıfa üyelik için, diğeri üyelik olmama içindir. Sınıflandırma problemi iki durumdan daha fazla olabilir. Yukarıda bahsedilen makine öğrenimi (machine learning) tanımından tek farkı, orijinal etiketlendirme süreci olan f ' in değer kümesi (codomain) ve bunun modeli olan $\hat{f}$ 'in uygun sayıda farklı etiketleri içeren bir küme olmasıdır. Orijinal etiketleme sürecinin bir modeline sahip olduktan sonra, bu modeli kullanarak X veri evrenindeki bir elemanın etiketi hesaplanabilir ya da tahmin edilebilir. Ayrıca, bu elde edilen model orijinal etiketleme süreci hakkında bazı bilgiler sağlar. Yani, bir model bazı açıklayıcı yeteneğe sahiptir.

1.4.1 Makine Öğrenimi Çeşitleri ve kaynakları

Makine öğreniminde, nelerin öğrenilmesi en önemli sorudur. Bir çeşit sayısal yapının öğrenilmesi genellikle göz önüne alınır. Farklı sayısal yapılar şunlardır:

- Fonksiyonlar
- Mantık programları ve kural kümeleri
- Gramerler
- Problem çözme sistemleri'dir.

Makine öğrenimi birçok kaynaktan oluşmaktadır. Bu farklı kaynaklar her biri farklı metotlar ve farklı kelimelerden oluşmaktalar. Bular tek bir disiplin altında toplanmaktadırlar. Aşağıda ayrı disiplinlerin makine öğrenimine katkıları bulunmaktadır¹⁷.

- Beyin modeller
- Uyarlamalı kontrol teorisi
- Psikolojik modeller
- Yapay zekâ
- Evrimsel modeller
- İstatistik

1.4.2 Denetimsiz Öğrenme

Denetimsiz öğrenmede amaç, girdiden doğru değerlerin, bir gözetmen tarafından sağlanan bir çıktıya olan haritalanmasıdır. Denetimsiz öğrenmede, gözetmen yoktur ve sadece girdi verileri mevcuttur. Burada amaç, girdide düzenliliğin veya kurala uygunluğun aranmasıdır. Belli kalıpların, diğerlerine göre çok daha sıkı olarak meydana gelecek şekilde girdi uzayında bir yapısı mevcuttur ve biz genellikle, nelerin meydana geldiğini ve ya gelmediğini görmek isteriz. Buna istatistikte, yoğunluk tahmini (density

¹⁷ Nils J. Nilsson, Introduction To Machine Learning: An Early Draft Of A Proposed Textbook, Stanford, CA, Stanford University, 2005, pp. 3-5.

estimation) adı verilir. Kümeleme (clustering), yoğunluk tahmini için kullanılan metotlardan bir tanesidir¹⁸.

Kümeleme uygulamalarından bir tanesi, görüntü sıkıştırma'dır. Bu durumda, girdi örnekleri RGB değerleri olarak görüntü piksellerinden oluşur. Kümeleme programı benzer renkli pikselleri aynı grupta gruplandırır ve bu gruplar görüntüde çok sık bir şekilde meydana gelen renklere karşılık gelir.

Denetimsiz öğrenmede, nesnelerin bir eğitime örnekleme verilir. Bu örneklemeden, “yapı” çıkarımı yapılması amaçlanır. Eğer eğitime nesnelerinde bazı yapılar mevcut ise, bu fazlalık içinden avantaj sağlamak mümkündür ve verinin kısa bir tanımı ve açıklamasını bulmak mümkündür. Veriyi göstermenin en genel yollardan bir tanesi, nesnelerin herhangi çiftleri arasında bir benzerlik belirlemektir. Eğer iki nesne çok fazla bir yapıyı paylaşıyorlarsa, aynı asıl örnekten veri çoğaltmak mümkündür¹⁹.

1.4.3 Pekiştirmeli Öğrenme

Bazı uygulamalarda, sistemin çıktısı eylemler dizisidir. Bu durumda, tek bir eylem önemli değildir. Hedefe ulaşmak için, doğru eylemler dizisi olan politika en önemli olan durumdur. Herhangi bir ara durumda, en iyi eylem diye bir şey yoktur, fakat eğer iyi bir politikanın parçası ise o eylem iyi bir eylemdir. Böyle bir durumda, makine öğrenim programı politikaların iyiliğinin belirleyebilmelidir ve ayrıca bir politika oluşturmak için geçmiş iyi eylem dizilerinden faydalanarak öğrenmelidir. Bu tür öğrenme metotlarına, pekiştirmeli öğrenme algoritmaları adı verilir. Buna örnek verecek olursak, bir oyunda tek bir hareket kendi başına o kadar önemli değildir; önemli olan doğru hareketler dizisidir. Bir hareket eğer iyi bir oyun politikasının parçası ise

¹⁸ Alpaydın, **a.g.e.**, p.11.

¹⁹ Ralf Herbrich, **Learning Kernel Classifiers: Theory and Algorithms**, Massachusetts, The MIT Press, 2002, p. 6.

iyidir. Oyun oynamak hem yapay zekâ ve hem de makine öğreniminde önemli bir araştırma sahasıdır²⁰.

Verilen bir ödülü maksimize etmek için, pekiştirmeli öğrenmede ki problemin ne yapılması gerektiğinin öğrenilmesidir; durumları eylemlere nasıl haritalanması meselesidir. Denetimli öğrenmenin (supervised learning) aksine, verilen bir durumda, öğrenme algoritmasına hangi eylemlerin yerine getirileceği söylenmez. Bunun yerine, öğrenci, eylem alındıktan hemen sonra gelmese bile, bazı ödüller tarafından alınan eylemler hakkında bilgi elde ettiği farz edilir²¹.

1.4.4 Denetimli Öğrenme (Supervised Learning)

Denetimli öğrenme probleminde, girdi-çıkı çiftlerinin bir örneği verilir, bu eğitme örneği olarak ta adlandırılır. Burada görev, herhangi bir girdiyi gelecek girdi-çıkı gözlemleri ile olan anlaşmazlığı minimize edecek şekilde bir çıkıya haritalayan belirleyici fonksiyon bulmaktır²².

Denetimli metotlar, girdi öznitelikleri (bağımsız değişkenler) ile bir hedef özniteliği (bağımlı değişken) arasındaki ilişkiyi bulmaya çalışan metotlardır. Bulunan ilişki bir yapı içerisinde temsil edilir ve bu yapı “model” olarak bilinir. Genellikle modeller veri kümelerinde saklı olan olayları açıklar. Ayrıca girdi öznitelikleri, bilinmek üzere, hedef özniteliğinin tahmini için kullanılabilir. Denetimli metotlar pazarlama, finans ve üretim gibi değişik alanlarda uygulanabilir²³. Sınıflandırıcı (classifier) ve regresyon modelleri olmak üzere iki temel denetimli modeller mevcuttur. Regresyon modeller, girdi uzayını gerçek değerli tanım kümesine haritalar. Örneğin, bir regresör,

²⁰ Alpaydın, **a.g.e.**, p.13.

²¹ Herbrich, **a.g.e.**, p.7.

²² **A.e.**, p.3.

²³ Oded Maimon, Lior Rokach, “Introduction To Supervised Methods”, **Data Mining and Knowledge Discovery Handbook**, Ed. by Lior Rokach and Oded Maimon, 2005, Springer Science+Business Media Inc, p.150.

karakteristik özellikleri verilen belli bir ürün için, talep tahmin edebilir. Bunun yanı sıra, sınıflandırıcılar, girdi uzayını önceden tanımlanmış sınıflara haritalar. Örneğin, sınıflandırma, banka ödemelerini zamanında yapanlar veya yapmayanlar şeklinde sınıflandırılarak kullanılabilir. Veri madenciliği sahasında ve diğer alanlardaki uygulamalarından dolayı, sınıflandırmanın faydaları çok muazzamdır. Sınıflandırma destek vektör makineleri, karar ağaçları, olasılıklı özetler ve cebirsel fonksiyonlar, vb. gibi birçok alternatif teknikler, algoritmalar temsil etmektedir.

1.4.4.1 Eğitim Kümesi (Training Set)

Veri kümesini üç kümeye bölmek için basit rastgele örneklendirme kullanılabilir²⁴:

- Eğitim kümesi (Training Set): modelin eğitilmesinde kullanmak için rastgele seçilen durumlardır
- Test Kümesi (Testing Set): modelin temizlenmesinden önce veya model geliştirmeleri eklenmeden önce modelin tahmin edilebilirliğinin belirlemek için kullanılan durumlardır.
- Doğrulama Kümesi (Validating Set): tüm modelleme yapıldıktan sonra modelin son performansını test etmek için kullanılan durumlardır

Mevcut veri kümesi iki alt kümeye ayrılır. Bunlardan birincisi eğitime kümesini (training set) oluşturmaktadır. Modellerin seçilen sınıfı içerisinde özellikli bir öğrenme modeli tanımlamak için kullanılır. Genellikle, eğitime kümesinin örneklem büyüklüğü nispeten küçük olacak şekilde seçilir. İstatistikî açıdan birkaç bin adet gözlem önemli kabul edilmektedir²⁵.

Tipik bir denetimli öğrenmede, bir eğitim seti verilir. Burada amaç, daha önce görünmeyen örneklerin tahminini yapmak için kullanılabilen açıklama şekli oluşturmaktır. Eğitim seti değişik dillerde açıklanabilir. Daha sıklıkla, belli bir *çanta*

²⁴ Nisbet, Elder, Miner, **a.g.e.**, p.70.

²⁵ Carlo Vercellis, **Business Intelligence: Data Mining and Optimization for Decision Making**, West Sussex, U.K, John Wiley & Sons Ltd., 2009, p.89.

şemanın, “*çanta örneği*” gibi ifade edilebilir. Burada çanta örnek ise, veri gruplarının bir derlemesidir. Bu gruplar kayıt, satır ya da örnekler olarak bilinir. Bunların kopyaları mevcut olabilir. Her bir veri grubu, öznitelik değerlerinin bir vektörü olarak tanımlanır. *Çanta şema* özniteliğin ve bunların tanım bölgelerinin açıklamalarına olanak sağlar. Bir “çanta şema”, $B(A \cup y)$ olarak tanımlanır, burada $A = \{a_1, \dots, a_n\}$ olacak şekilde n tane özniteliğe sahip girdi özniteliklerini belirtir ve y 'de sınıf değişkenini ya da hedef değişkeni temsil eder²⁶.

Öznitelik (bazen, değişken, özellik yada alan olarak adlandırılır) tipik olarak iki türdür: Nominal (değerler sıralı olmayan bir kümenin elemanlarıdır) ya da nümerik (değerler reel sayılardır). Örnek uzayı (tüm muhtemel örneklerin kümesi), tüm girdi özniteliklerin tanım kümelerinin bir Kartezyen çarpımı olarak tanımlanır: $X = \text{dom}(a_1) \times \text{dom}(a_2) \times \dots \times \text{dom}(a_n)$. Evrensel örnek uzayı U , tüm girdi öznitelik tanım kümelerinin bir Kartezyen çarpımı olarak tanımlanır, yani: $U = X \times \text{dom}(y)$. Eğitim kümesi, m adet değişkenler grubunun kümesinden oluşan bir *çanta örneğidir*. Eğitim kümesi, şeklen şu şekilde tanımlanır: $S(B) = (\langle x_1, y_1 \rangle, \dots, \langle x_m, y_m \rangle)$, $x_q \in X$ ve $y_q \in \text{dom}(y)$ 'dir²⁷. Eğitim demet kümesi, U üzerinde, sabit ve bilinmeyen ortak olasılık dağılımı U 'ya göre, tesadüfî ve bağımsız olarak üretildiği genellikle varsayılmaktadır. $y=f(x)$ fonksiyonu kullanılarak, bir değişkenler grubunun sınıflandırıldığı zaman, burada deterministik durum genellemesinin geçerli olduğu not edilmelidir.

1.4.4.2 Sınıflandırma Problemi

Bir sınıflandırma probleminde, gözlemler kümesi mevcuttur ve genellikle bir hedef sınıfı, bilinmeyen veri kümesinin kayıtları yardımıyla temsil edilir. Gözlemler, örneğin, cep telefonu kullanıcılarına karşılık gelebilir. İkili sınıf ise bir müşterinin halen

²⁶ Maimon, Rokach, a.g.e., p.134.

²⁷ A.e., p.134

aktif olup olmadığını gösterebilir. Her bir gözlem değeri, bilinen özniteliklerin verilen sayısı ile ifade edilir. Örneğin, bir önceki örnekte, öznitelik yaş, müşterinin kıdemi olarak ifade edilirken, hedef ise giden telefon trafiğini ifade etmektedir. Bir sınıflandırma algoritması, öznitelik değerleri bilinen, gelecekteki gözlemlerin hedef sınıfını tahmin edebilecek bir modeli belirlemek için, geçmişe ait mevcut gözlemleri kullanabilir. Değeri, tahmin edilecek hedef öznitelik sınıflandırma problemlerinde kategoriktir ve bundan dolayı sonludur. Birçok uygulamalarda hedef, bir ikili (binary) değişken tarafından temsil edilir²⁸.

Makine öğrenimi topluluğu, orijinal olarak kavram öğrenimi problemini ortaya attılar. Konseptler, ortak özellikler kümesine sahip objeler, olaylar ya da fikirler için mental kategorilerdir. Her bir konsept (kavram), daha geniş bir kümede tanımlı olan objelerin ya da olayların bazı alt kümelerini tanımlayan olarak görülebilir²⁹. Bir kavramı öğrenmek demek, bunun bir örnekler kümesinden genel tanımının sonucunu çıkarmaktır. Bu tanım, ya açıkça formüle edilir ya da örtük olarak bırakılır. Fakat her iki yolda, her bir olası örneği, kavrama atama yapar ya da yapmaz. Böylece, bir konsept (kavram), Örnek uzaydan Boolean kümesine olan bir fonksiyon olarak kabul edilir, yani: $c: X \rightarrow \{-1,1\}$. Alternatif olarak, c kavramı X 'in bir alt kümesi olarak kabul edilebilir, yani $\{x \in X: c(x) = 1\}$. C konsept sınıfı, bir konseptler (kavramlar) kümesidir. Diğer tanımlar ise, direkt olarak konsept öğreniminin bir uzantısı olan sınıflandırma problemi, ya da çoklu sınıflandırma problemi olarak bilinir. Bu durumda, tüm olası örnekleri önceden tanımlanan, sınıf etiketlerinin kümesine haritalayan bir fonksiyon aranır. Sınıflandırıcıların indüktörleri genellikle aşağıdaki gibi tanımlanır.

Tanım³⁰: S gibi bir eğitim kümesi verilmiş olsun, girdi özniteliklerin(değişkenler) kümesi $A = \{a_1, \dots, a_n\}$ olsun ve etiketlenmiş örnek uzayı

²⁸ Vercellis, **a.g.e.**, p.92.

²⁹ Tom Mitchell, **Machine Learning**, McGraw – Hill, 1997, p.20-25

³⁰ Maimon, Rokach, **a.g.e.**, p.135.

üzerindeki bilinmeyen sabit D dağılımından olan nominal hedef değişken y olsun, burada hedef minimum genelleme hatasına sahip bir optimal sınıflandırıcı indüklemektir.

Genelleme hatası, D dağılımı üzerinde yanlış sınıflandırma oranı olarak tanımlanır. Nominal değişken durumunda, şu şekilde ifade edilir:

$$\varepsilon(I(S), D) = \sum_{(x,y) \in U} D(x, y) \cdot L(y, I(S)(x)) \quad (1.1)$$

burada $L(y, I(S)(x))$ sıfır-bir kayıp fonksiyonudur ve aşağıdaki gibi tanımlanır:

$$L(y, I(S)(x)) = \begin{cases} 0 & \text{eğer } y = I(S)(x) \\ 1 & \text{eğer } y \neq I(S)(x) \end{cases} \quad (1.2)$$

Nümerik değişkenli durumda ise, toplam operatörü integral operatörü ile yer değiştirir.

1.4.4.3 Performans Değerlendirmesi

İndükleyicinin performans değerlendirme makine öğreniminin temel yönüdür. Bilindiği gibi, indükleyici eğitme kümesinin girdi olarak alır ve bilinmeyen bir örneği sınıflandırabilecek bir sınıflandırıcı meydana getirir. Hem sınıflandırıcı ve hem de indükleyiciler için bir değerlendirme kriteri mevcuttur. Performans değerlendirme modelin kalitesinin anlaşılması, KDD(Knowledge Discovery and Data Mining) yinelemeli sürecinde parametrelerin arıtılması ve verilen modeller (indükleyiciler) kümesinden en kabul edilebilir modelin seçilmesi açılarından önemlidir. Doğal olarak, yüksek doğruluk yüzdesine sahip modeller en iyi model olarak dikkate alınır. Bunun yanı sıra, hesaplama karmaşıklığı (computational complexity) ya da üretilen sınıflandırıcının anlaşılabilirliği (comprehensibility of the generated classifier) gibi başka değerlendirme kriterleri de mevcuttur.

1.4.4.3.1 Genelleme Hatası

Eğitme verisinin miktarı büyüdükçe, genelleme hatası azalır. Model sınıfın karmaşıklığı artarken, genelleme hatası önce azalır ve daha sonra artmaya başlar. Aşırı karmaşık bir sınıfa genelleme hatası eğitme verisinin miktarını artırmak suretiyle ancak belli bir noktaya kadar kontrol altında tutulabilir³¹.

Herhangi bir veri modeli için destek vektör makineleri genelleme hatası kriterine göre optimal olduğu asla ispatlanmamasına rağmen, bazı yazarlar destek vektör yaklaşımını “optimal” olarak kabul ederler³². Sınıflandırmada, yanlış sınıflandırma olasılığı genelleme hatası olarak bilinir ve optimalite tanımı için en popüler kıstaslardan biridir³³.

Kabul edelim ki $I(S)$, S 'de I gibi bir indükleyici tarafından üretilen bir sınıflandırıcıyı temsil etsin. $I(S)$ 'nin *genelleme hatası*, seçilen bir örneğin yanlış sınıflandırma olasılığıdır. Bir sınıflandırıcının sınıflandırma doğruluğu, bir eksi genelleme hatasıdır. *Eğitme hatası*, eğitme kümesindeki örneklerin sınıflandırıcı tarafından doğru bir şekilde sınıflandırılma yüzdesi olarak tanımlanır ve aşağıdaki gibi ifade edilir³⁴:

$$\hat{\epsilon}(I(S), S) = \sum_{(x,y) \in S} L(y, I(S)(x)) \quad (1.3)$$

Burada $L(y, I(S)(x))$ sıfır-bir kayıp fonksiyonudur. Genelleme hatası, doğal bir kriter olmasına rağmen, gerçek değeri sadece nadir vakalarda bilinmektedir. Bunun sebebi ise, örneğin etiketli uzayının D dağılımı bilinmediğinden kaynaklanmaktadır. Eğitim hatası, genelleme hatasının bir tahmini olarak alınabilir. Fakat eğitim hatasını olduğu gibi

³¹ Alpaydın, **a.g.e.**, p.39

³² Viladimir N. Vapnik, **The Nature of Statistical Learning Theory**, New York, Springer-Verlag, 1995, p.127.

³³ S. Raudys, “How good are support vector machines?”, **Neural Networks**, vol.13, 2000, pp.17-19.

³⁴ Maimon, Rokach, **a.g.e.**, p.136.

kullanmak iyimser yanlı tahmin sağlayacaktır, özellikle, eğer öğrenme algoritması eğitime verilerine aşırı uyuyor ise (over - fit). Genelleme hatasının tahmin etmek için teorik ve deneysel (ampirik) olmak üzere iki genel yaklaşım mevcuttur.

1.4.4.3.2 Genelleme Hatanın Teorik Tahmini

İndükleyicinin kapasitesi eğitime hatasındaki güven için belirleyici bir etkidir. Yani bu kapasite sınıflandırıcının indükleyebilme konusunda ki çeşitliliğinin göstermektedir. Düşük eğitime hatası, düşük genelleme hatası olduğunu göstermez. Çoğu zaman, eğitime hatası ve genelleme hatası için bir değişken olarak, eğitime hatasına atanan, güven arasında ters orantı (trade - off) vardır. Güven düzeyi, genelleme ve eğitime hataları arasındaki farkın ölçümünden elde edilir. VC- boyut, indükleyicinin kapasitesinin ölçülmesi için kullanılabilir.

Zayıf genelleme hatası nasıl ortaya çıkar? Büyük kapasiteye sahip indükleyiciler eğitime kümesinin büyüklüğüne kıyasla büyük sayıda serbest değişkenlere sahiptirler. Böylece, büyük kapasiteye sahip indükleyiciler düşük eğitime hatası elde ederler. Fakat sadece kalıpları ezberleyebilirler ya da kalıpları (pattern) aşırı uydurabilirler (over - fitting) ve bundan dolayı zayıf genelleme yeteneği sergiler. Düşük hata, yüksek genelleme hatası için, zayıf bir belirleyici olması muhtemeldir. Buna karşın, verilen belli sayıda örnekler için kapasite çok küçük olduğu zaman, indükleyiciler veriyi aşağı seviyede uydurabilir (under - fit), ve bu hem zayıf eğitime ve genelleme hatalarını sergiler. Yetersiz sayıda serbest değişkene sahip indükleyiciler için, eğitime hatası zayıf olabilir, fakat genelleme hatası için iyi bir belirleyicidir. Bu uç kapasiteler arasında, en iyi genelleme hatasını elde eden optimal kapasite mevcuttur.

Genelleme hatasının tahmin etmek için, dört tane teorik çerçeveler mevcuttur, bunlar PAC, VC ve Bayes tipi ve istatistiksel fizik. Bütün bu çerçeveler, eğitime hatalarının indükleyicilerin kapasitelerini ifade eden, ceza fonksiyonu ile birleştirirler

1.4.4.3.2.1 VC- Çerçevesi (vapnik -Chervonenkis)

Vapnik-Chervonenkis teorisi olarak bilinir (Vapnik, 1995). Örneklerden yola çıkarak, öğrenmek için tüm temel teorik yaklaşımlar içinde, en kapsamlı ve hem regresyon için hem de sınıflandırma için uygulanabilen yaklaşım VC-çerçevesidir. İndüklenen sınıflandırıcının, eğitime hatası ve genelleme hatası, eğitime kümesinin boyutu sonsuza giderken aynı minimal hata değerine yaklaşımları zorunlu olması, istatistikte tutarlılık olarak ifade edilir. Dolayısıyla VC-çerçevesi indükleme prosedürünün tutarlılığı için genel, yeterli ve gerekli koşulları sağlar. Çok yaygın olarak kullanılan ve kapasite ölçütü olarak bilinen, VC-boyut (VC-Dimension), Vapnik teorisi tarafından tanımlanmaktadır.

Teorem: Sınırlı VC-boyut, d olmak üzere, hipotez uzayı H 'nin genelleme hatasındaki sınır aşağıdaki gibi ifade edilir;

$$|\varepsilon(h, D) - \hat{\varepsilon}(h, S)| \leq \sqrt{\frac{d \cdot (\ln \frac{2m}{d} + 1) - \ln \frac{\delta}{4}}{m}}, \forall h \in H; \forall \delta > 0, \quad (1.4)$$

$1 - \delta$ olasılıklı, burada $\hat{\varepsilon}(h, S)$ kardinalite (nicelik) “ m ”nin eğitime kümesi S 'de hesaplanan “ h ” sınıflandırıcının eğitime hatasını temsil eder. Ayrıca $\varepsilon(h, D)$, D dağılımı üzerinde sınıflandırıcı “ h ” nin genelleme hatasını temsil eder.

VC- boyut tüm sınıflandırıcıların kümesinin bir özelliğidir ve H ile ifade edilir ve bu indükleyici tarafından incelenir. VC- boyut, geçerli sınıflandırıcıların bir kümesi tarafından parçalanabilen veri noktalarının maksimum sayısı olarak tanımlanır. Eğer 2^m olası yolda noktaları iki sınıfa ayıran sınıflandırıcılar mevcut ise S kümesi H tarafından parçalanır. Ayrıca, eğer H 'nin VC-boyutu d ise, H tarafından parçalanabilen d noktalarının en az bir kümesi mevcuttur. Fakat genel olarak, d noktalarının her kümesi H tarafından parçalanabileceği gerçekleşmeyecektir.

Bir indükleme prosedürünün güvenli olması için yeterli şart, indükleyicinin VC-boyutunun sınırlı olmasıdır. Bir doğrusal sınıflandırıcının VC-boyutu, girdi uzayının n boyutudur. Yani, sınıflandırıcının serbest parametre sayısıdır. Genel bir sınıflandırıcının VC-boyutu, serbest değişken sayısından çok daha farklıdır ve çoğu zaman bunun doğru hesaplanması çok zordur. Bu durumda VC – boyut için alt ve üst sınırların hesaplanması faydalıdır.

1.4.4.3.2 PAC- Çerçevesi (Probability Approximation Correct)

Örnek girdilerin bir kümesi ve bunların değerleri verildiği zaman, bir fonksiyonun tahmin edilmesi problemi bilinmektedir. Tahmin etmek istediğimiz fonksiyonun, fonksiyonların uygun bir biçimde sınırlanmış altkümesine ait olduğunu bilindiği zaman, en sonraki girdilerin değerlerini nerdeyse doğru bir şekilde tahmin edebileceğimiz mümkün girdiler ve bunların değerlerinin küçük bir kısmı mevcuttur. Yani, verilen örnek kalıpların (pattern) bir eğitime kümesi, etiketlenmiş örnekler ile tutarlı olan bir fonksiyon, hipotezlerin sınırlandırılmış bir kümesinden seçmemize izin vermek için yeterli olabilir. Öyle ki, etiketlenilmiş örneklerin seçildiği, aynı dağılıma göre seçtiğimiz bu fonksiyon, yüksek bir olasılıkla yaklaşık olarak rastgele seçilen sonraki örneklerde doğru olacaktır³⁵. Bahsedilen bu kavrama PAC(probably approximately correct) teorisinin gelişimine yol açtır. İlk önce Leslie Valiant³⁶ tarafından geliştirildi.

PAC-çerçevesi, “Olasılık Yaklaşım Düzeltme” olarak bilinir. “Probability Approximation Correct” terimlerinin kısaltmasıdır. PAC, “konsept (kavram) sınıf” nitelendirmek için kullanılabilir. “Konsept sınıf”, tesadüfî olarak makul sayıda çekilen

³⁵ Nilsson, **a.g.e.**, p. 107.

³⁶ L. Valiant, “A Theory of the Learnable”, **Communications of the ACM**, vol.27, 1984, pp.1134-1142.

eğitim örneklerinden, makul bir hesaplama düzeyinde güvenilir bir şekilde öğrenilebilir³⁷. Aşağıda PAC –öğrenilebilir tanımı aşağıdaki gibi verilmiştir.

Tanım³⁸ : Kabul edelim ki C , n tane öz niteliğe (değişkene) sahip örnek uzay olan X üzerinde tanımlanan bir konsept sınıfı olsun. Kabul edelim ki, I hipotez uzayı H 'yi dikkate alan indükleyici olsun. C , H 'yi kullanarak, I tarafından PAC-öğrenilebilir olduğu kabul edilir, eğer $\forall c \in C$, D dağılımı X üzerindedir, ε öyle ki $0 < \varepsilon < 1/2$ ve δ öyle ki $0 < \delta < 1/2$. En az $(1 - \delta)$ bir olasılıkla öğrenen I , $1/\varepsilon$, $1/\delta$, n , ve $\text{boyut}(c)$ 'de polinom olan $h \in H$ gibi bir hipotez çıkaracaktır. Öyle ki $\varepsilon(h, D) \leq \varepsilon$, $\text{boyut}(c)$, C 'deki c 'nin kodlama uzunluğunu temsil eder.

$c \in C \subseteq H$ hedef konseptin, ε hatası içinde $h \in H$ gibi bir hipotez elde etmek için, en az $(1 - \delta)$ olasılıkla, sınırlı hipotez uzayı olan H 'yi inceleyerek, “PAC öğrenme modeli” herhangi tutarlı “öğrenen I ” için, yeterli olan eğitim örnekleri sayısında, genel bir sınır sağlar. Daha özellikli olarak, eğitime kümesinin boyutu aşağıdaki gibi olmalıdır:

$$m \geq \frac{1}{\varepsilon} (\ln(1/\delta) + \ln|H|). \quad (1.5)$$

1.4.4.3.3 Genelleme Hatasının Ampirik Tahmini

Mevcut örnekleri, eğitime (training) ve test kümeleri olarak iki gruba ayırmak sureti ile, genelleme hatasının tahmin etmek, bir başka yaklaşım biçimidir. İlk önce, uygun bir sınıflandırıcının inşa edilmesi gerekmektedir. Dolayısıyla, indükleyici tarafından eğitime kümesi kullanılarak uygun sınıflandırıcı oluşturulur ve daha sonra test kümesinde sınıflandırıcının yanlış sınıflandırma derecesi ölçülür. Eğitime hatası çoğu zaman genelleme hatası için az değer biçtiği için, test küme hatası genelleme hatasının

³⁷ Mitchell, a.g.e., pp. 203-205

³⁸ A.e., 205-207

tahmini için, eğitime hatasına kıyasla daha iyi tahmin sağlamaktadır. Veri sınırlı olduğu zaman, veriyi yeniden örnekleme yaygın bir kullanımdır, yani, veriyi eğitime (training) ve test kümeleri olarak farklı şekillerde yeniden bölmektir. Bir indükleyici her bir bölme (partition) için eğitilir ve test edilir ve doğrulukların (accuracy) ortalaması alınmış olur. Bu verilen süreci takip etmek sayesinde, bir indükleyicinin gerçek genelleme hatasının daha güvenilir tahmini elde edilir³⁹.

Rastgele alt örnekleme (random sub-sampling) ve n-kat çapraz doğrulama (n-fold cross-validation) teknikleri, yeniden örneklendirme (re-sampling) için kullanılan yaygın tekniklerdir. Rastgele alt örneklemede, veri tesadüfî olarak ayrık eğitime ve test kümelerine birçok kez bölünür. Her bir bölünmede hatalar elde edilir ve ortalaması alınır. n-kat çapraz-doğrulama da ise, veri tesadüfî olarak n tane ayrışık yaklaşık olarak eşit büyüklükte altkümelere bölünür. İndükleyici birçok defa eğitilir ve test edilir. Her seferinde veri k-katlardan bir tanesinde test edilir ve geriye kalan n-1 tane katları kullanarak eğitilir.⁴⁰

İki veya daha fazla metodun tahmin doğruluğunun kıyaslanmasında, eğitime veri örneklerinin rastgele örneklenmesi ile ilişkili olan yanlılığı (bias) minimize etmek için, n-kat çapraz doğrulama adı verilen yöntem kullanılabilir. Bu metot, ayrıca rotasyon tahmini olarak da adlandırılır. Tüm veri kümesi, rastgele olarak $n - 1$ adet ayrışık, yaklaşık olarak eşit boyutlu alt kümelere bölünür. Sınıflandırma modeli n-kez eğitilir ve test edilir. Her keresinde, bir kat hariç diğer tüm katlar (fold) eğitilir ve sınıflandırma modeli geriye kalan tek kat üzerinde test edilir. Bir modelin genel doğruluğunun çapraz doğrulama tahmini n adet bireysel doğruluk ölçümlerinin ortalaması alınarak elde edilir⁴¹.

$$CVA = \frac{1}{n} \sum_{i=1}^n A_i, \quad (1.6)$$

³⁹ Maimon, Rokach, **a.g.e.**, p.156.

⁴⁰ Maimon, Rokach, **a.g.e.**, p.156.

⁴¹ Olson, Delen, **a.g.e.**, p.141.

Burada CVA apraz doęrulamanın doęruluęunu (Cross Validation Accuracy) belirtir. Ayrıca, n kullanılan katların sayısıdır ve A ise, her bir katın doęruluk ölçütüdür.

apraz-doęrulama genelleme hatası tahmini, genel yanlış sınıflandırmalar sayısının verideki örneklerin sayısına bölünmesi ile elde edilir. Yukarıda bahsedilen iki metodun avantajları ve dezavantajları mevcuttur. Rastgele alt örnekleme metodu, sınırsız sayıda tekrarlanabildięi için avantaja sahiptir. Fakat test kümeleri bağımsız bir şekilde örneklerin altında yatan dağılıma göre çekilmedięinden, dezavantaja sahiptir. Bundan dolayı, rastgele örnekleme ile eşleştirilmiş farklar (paired differences) için t-test kullanmak 1.tür hata (Type I error) olasılıęının artmasına yol açar. Yani gerçekte mevcut olmamasına rağmen anlamlı farkın oluşmasıdır. Her bir kat (fold) içinde oluşan genelleme hatasında, t-testi kullanma sonucunda düşük ihtimalli 1.tür hata meydana gelir, fakat bu, genelleme hatasının sağlam tahminini vermeyebilir. Sağlam bir tahmin elde etmek için, n -kat apraz-doęrulama teknięini n -kez tekrarlama çok yaygın bir uygulamadır. Fakat bu yöntem, bağımsız olmayan test kümeler vermektedir ve dolayısıyla 1.tür hata şansını yükseltir.

Tabakalama (stratification) teknięi, rastgele alt-örnekleme ve n -kat apraz doęrulama esnasında sıklıkla kullanılan bir yöntemdir. Tabakalama, tüm veri seti sınıf dağılımı, eğitme ve test kümelerinde korunduęunu garanti eder. Özellikle çok sayıda sınıflı veri kümeleri için, tabakalama teknięi tahmin edilmiş hata varyansının azaltılmasında yardımcı olmaktadır. Burada doęruluęu deęerlendirmek için, ikili t-testli tabakalı rastgele alt örnekleme kullanılır⁴². Bir n -kat apraz doęrulama (n -fold cross validation) için, mevcut verinin hepsi öğrenme örnekleme için kullanılabilir ve bunu yaklaşık aynı büyüklükteki n - paraya bölebilir. Bu, genellikle rastgele yapılır, fakat

⁴² Maimon, Rokach, **a.g.e.**, p.156.

tabakalama kullanılarak, n – çapraz doğrulama gruplarını birbirine benzer yapılabilir⁴³. Burada n sayısı, tipik olarak 5 ya da 10 olarak alınır. Birçok durumda n için çok büyük değer kullanılarak çok az kazanç elde edilir. n değerinin büyük olması, sınıflandırıcının oluşturulması için gerekli zaman miktarının büyük olması demektir. Bazı durumlarda, tahminin kalitesi, n değerinin çok büyük olması nedeni ile azalmış olur.

1.4.4.3.4 Hesaplama Karmaşıklığı (Computational Complexity)

Gerçek hayat problemlerinin pratik çözümleri ile ilgilendiğimizden, veri madenciliğinde kullanılan algoritmalar çok büyük boyutlu veri kümelerini işlemek zorundadırlar. Bir algoritma için, küçük bir örnek üzerinde çok iyi çalışması yeterli değildir; performansının büyük veri kümeleri ile ölçekli olması gerekmektedir. Hesaplama karmaşıklığı veya algoritmaların ölçeklenebilirliği algoritmaların etkinliğini belirtir⁴⁴.

Sınıflandırıcıları ve indükleyicileri kıyaslamak için bir diğer önemli kıstas ise hesaplama karmaşıklığıdır. Bu kriter, her bir indükleyicinin CPU kullanma miktarı olarak bilinmektedir.

- Yeni bir sınıflandırıcı oluşturmak için hesaplama karmaşıklığı: Özellikle veri madenciliği algoritmasını çok büyük veri kümelerine ölçeklendirmek için çok önemli bir metriktir.
- Bir sınıflandırıcının güncelleştirilmesi için hesap karmaşıklığı: Yeni bir veri verildiği zaman, hâlihazırdaki sınıflandırıcının yeni veriyi yansıtması için yeni bir sınıflandırıcı olacak şekilde güncelleştirilmesi için gereken hesap karmaşıklığıdır.

⁴³ C. D. Sutton, “Classification and Regression Trees, Bagging and Boosting”, **Handbook of Statistics 24: Data Mining and Data Visualization**, Ed. by Ed. by C.R. Rao, E.J.Wegman, J.L.Solka, Elsevier B.V., 2005, pp. 312.

⁴⁴ Nello Cristianni, John Shawe - Taylor, **Kernel Methods for Pattern Analysis**, New York, Cambridge University Press, 2004, p.12.

- Yeni bir olayın (örnek) sınıflandırılması için hesap karmaşıklığı: Oldukça ufak olduğundan dolayı, çoğu zaman ihmal edilir. Fakat bazı metotlar ya da belirli gerçek zamanlı uygulamalarında (k-nearest gibi) bu tür çok önemli olabilir.⁴⁵

Hesaplama karmaşıklığını (Wegman, 2005) aşağıdaki gibi dört başlık altında incelemektedir. Bunlar sırası ile şöyledir:

- a) Büyüklük sırasının dikkate alınması,
- b) CPU performansından kaynaklanan uygulanabilirlik sınırları,
- c) Dosya transfer performansında kaynaklanan uygulanabilirlik sınırları,
- d) Görsel çözünürlükten kaynaklanan uygulanabilirlik sınırları⁴⁶.

1.4.4.3.5 Anlaşılabilirlik (Comprehensibility)

Anlaşılabilirlik (Comprehensibility), performans değerlendirmesinde kullanılan kıstaslarındandır. Burada amaç, bir insanın indüklenen sınıflandırıcının ne kadar iyi kavradığıdır. Genelleme hatası, sınıflandırıcının veriye uyup uymadığını ölçerken, anlaşılabilirlik ise, o sınıflandırıcının “zihinsel uyumunu” ölçer. Sinir ağları, destek vektör makineleri gibi birçok teknikler özellikle doğruluğu elde etmek için düzenlenmişlerdir. Fakat bunların sınıflandırıcıları gerçek değerli parametrelerin büyük topluluklarının kullanılması ile temsil edilirler. Aynı zamanda, bunları anlamak oldukça zordur ve bunlar “kara-kutu (black -box)” modeller olarak bilinmektedirler. Çoğu zaman araştırmacının, indüklenen sınıflandırıcıyı denetlemesi çok önemlidir. Özellikle tıbbi teşhis gibi alanlarda, kullanıcının sistemin güvenli sonuçlar elde etmesi için, nasıl karar verdiğini anlamak zorundadır. Sistem, girdi verisinde daha önce önemi bilinmeyen, belirgin özellikler bulabilir. Dolayısıyla, bu buluşlar insanların incelemesi için ulaşılmasına imkân sağlanabilir⁴⁷. Anlaşılabilirlik, aynı indükleyicin farklı

⁴⁵ Maimon, Rokach, **a.g.e.**, p.157.

⁴⁶ E.J. Wegman, L. Solka Jeffrey, “Statistical Data Mining”, **Handbook of Statistics 24: Data Mining and Data Visualization**, Ed. by C.R. Rao, E.J.Wegman, J.L.Solka, Elsevier B.V, 2005, p. 2.

⁴⁷ L.K. Hunter, “Finding Relevant Biomolecular Features”, **ISMB**, 1993, pp.190-197.

sınıflandırıcıları arasında farklılık gösterebilir. Örneğin, karar ağaçlarında indüklenen ağacın boyutu (düğüm sayısı) önemlidir. Daha küçük ağaç, boyutları anlaşılması kolay olduğu için tercih edilmektedir. Bazı durumlarda, büyük ve dengesiz ağaçlarda kolayca yorumlanabilir. Bilindiği üzere, doğruluk ve karmaşıklık faktörleri sayısal olmasına rağmen, anlaşılabilirlik daha subjektiftir. Bir diğer önemli ayırım ise, karmaşıklık ve anlaşılabilirlik daha çok indükleyici metoduna bağlıdır ve çok daha az olarak ele alınan alan ile ilgilidir⁴⁸.

1.4.4.4 Büyük Veri Kümeleri İçin Ölçeklenebilirlik

Veri madenciliğinde en çok uygulanan yöntemlerden birisi sınıflandırmadır (classification). Literatürde belli sayıda sınıflandırma yaklaşımları önerilmiştir. Bu yaklaşımların kıyaslanmasında belli başlı kıstaslara dayanarak yapılır (Han ve Kamber, 2001). Bu kıstaslar, sınıflandırma model doğruluğu, hız, sağlamlık, yorumlanabilirlik ve ölçeklenebilirlik'tir. Ölçeklenebilirlik, çok büyük miktarda veri verildiği zaman, sınıflandırma modelinin oluşturulması için gerekli olan metot yeteneğidir⁴⁹.

Makine öğreniminde ve veri madenciliğinde ölçeklenebilirlik, bir algoritmanın büyüklük ile ayarlanma (ölçek) kabiliyeti ile ilgilidir. Problemin büyüklüğü artarken depolama ihtiyacı ve çalışma süresi olursuz (infeasible) olmaması, gerekli koşuldur. Ölçeklenebilir algoritmaların, karar ağaçları modellemesi, optimizasyon algoritmaları ve ilişki kurallarının madenciliği gibi, çok karmaşık problemler için tasarlanması, veri madenciliğinde en aktif araştırma alanı olagelmıştır. Açıkçası ölçeklenebilirlik, veri madenciliğinde temel bir problemdir. Çoğunlukla bunun algoritmik yönleri bakımından izlenmektedir. Kavramların da ölçeklenebilir olmasının önerilmesi ile problemin çift

⁴⁸ Maimon, Rokach, **a.g.e.**, p.157.

⁴⁹ Han, Kamber, **a.g.e.**, p.291.

eşliğı (duality) vurgulanması gerekmektedir⁵⁰. Bu konuda, genel inanış olan hipotez testinin veri madenciliğinde sunacağı hiçbir şeyinin olmadığı bir istatistik kavram olmasını ele alalım. Genel argüman, farklılıklar ya da ilişkiler çok az olmasına rağmen, test edilen her bir hipotezin istatistiksel olarak anlamlı olacak şekilde veri kümelerinin çok büyük olmasıdır.

İndüksiyon (tümevarım) makine öğrenimi, örüntü tanıma ve istatistik gibi birçok bilim dallarında merkezi problemlerden bir tanesidir. Fakat veri madenciliğı diğer geleneksel metotlardan farklı olmasının sebebi, bunun girdi verisinin çeşitli türlerinin çok büyük kümelerine olan ölçeklenebilirlik özelliğinden kaynaklanmaktadır. “ölçeklenebilirlik” nosyonu genellikle, çok sayıda kayıtların veya yüksek boyutsallık özelliklerinden en az bir tanesine sahip olan, veri kümelerini ifade eder. Klasik indüksiyon algoritmaları basit ve küçük boyutlu problemlere uygulanması ile başarılı sonuçlar elde edilmiştir. Fakat gerçek hayat ve büyük veri tabanlarında bilgi keşfetmeğe çalışmak, zaman ve hafıza problemlerini ortaya çıkarır. Birçok alanlarda, mesela astronomi, sağlık, moleküler biyoloji, finans ve pazarlama gibi birçok alanda büyük veri tabanları norm olmuştur. Dolayısıyla, veri madenciliğinin bu alanlarda kullanılması potansiyel verimli bir girişim oldu. Bunların uygulanması esnasında, bu alanlarda büyük pay sahibi olan şirketler araştırma topluluğundan çözümler beklemektedirler. Bilgi yoğunluğuna sahip şirketlerde, mesela bankalar ve telekomünikasyon gibi, her bir ya da iki yılda bir büyük oranda ham veri toplanmaktadır. Büyük veri ambarlarının yönetilmesi, verilen analiz edilmesi özellikli, çok pahalı yazılım ve donanım gerektirmektedir. Fayyad, v.d. (1996) göre, veri madenciliğı ile ilgilenen topluluklar için aşikâr zorluklar, veri madenciliğı algoritmalarının geliştirilmesi ve bunların gerçek hayat problemlerinde uygulanmasına olanak sağlamaktan geçtiğini belirtmektedirler. Gerçek

⁵⁰ Y.M. Benjamini, “Statistical Method for Data Mining”, **Data Mining and Knowledge Discovery Handbook**, 2nd Edition, Ed. by L. R. Oded Maimon, New York, 2010, Springer Science +Business Media, LLC., pp. 528.

dünya veri tabanlarının karakteristiklerinden biri, büyük hacme sahip olmalarıdır. Bu tür veri tabanları, çok sayıda sorunlar yaratmaktadır. Kısaca aşağıdaki gibi sıralanabilir:

- Hesaplamak karmaşıklığı,
- Doğru sınıflandırıcının bulunması zorluklarından kaynaklanan zayıf sınıflandırma doğruluğu
- Bellek problemi

Bahsedilen bu problemlerin, yani, büyük boyutlu veri tabanlarının çıkarmış olduğu sorunlar için çözüm olarak, aşağıdaki yaklaşımlar içermektedir⁵¹:

- Örneklem metotları (Sampling): istatistikçiler, popülasyondan farklı örneklem tekniklerini kullanarak kayıtları seçerler
- Birleştirme (aggregation): kayıtların bir grubunu tek olarak ele alarak ya da “gereksiz” kayıtların altkümelerini göz ardı ederek kayıt sayısını azaltmak.
- Eş zamanlı olarak problemin farklı taraflarını çözmek
- Çok sayıda kayıtlar ile başa çıkan algoritmalar olarak sağlayan verimli bellek metotlarının kullanımı
- Algoritmanın arama alanının azaltmak.

1.4.4.5 “Boyutsallık Belası (Curse of Dimensionality)” Problemi

Özniteliklerin (değişken) sayısı, girdi boyutunu belirler. Büyük boyutlu değişkenler sayısı arama alanının üssel olarak artırır ve bunun sonucu olarak, indükleyicinin istenmeyen ve gereksiz sınıflandırıcılar bulma ihtimali artacaktır. Denetimli sınıflandırma için etiketli örneklerin (labelled samples) gerekli sayısı boyutsallığın bir fonksiyonu olarak artar (Jimenez,L.O. ve Landgrebe, D.A. : , 1998). Fukunaga (1990) eğitime örneklerin gerekli sayısı doğrusal sınıflandırıcı için boyutsallıkla doğrusal olarak ilişkili olduğunu ve kuadratik sınıflandırıcı için

⁵¹ Maimon, Rokach, **a.g.e.**, p.159.

boyutsallıkla karesel olarak ilişkili olduğunu göstermiştir. Parametrik olmayan sınıflandırıcılar açısından, mesela Karar Ağaçları, durum daha da vahimdir. Çoklu yoğunlukların etkin tahminini elde etmek için, boyutların sayısı artıkça, örneğin sayısının üssel olarak artmak zorundadır (Hwang, J., Lay S., ve Lippman A., 1994). Bu olay genellikle “boyutsallık laneti (curse of dimensionality)” olarak adlandırılmaktadır. Bu terim ilk olarak (Bellman, 1961) tarafından icat edilmiştir. Küçük boyutlu örneklemede çok etkin olan karar ağaçları gibi indükleyiciler, boyutların sayısının artması ile anlamlı sonuçlar vermede başarısızdırlar. Ayrıca, daha az özelliklere (muhtemelen 10 tane den az değişken sayısı) sahip daha küçük sınıflandırıcılar insanlar tarafından çok daha kolay anlaşılmaktadır⁵².

Sözde boyutsallık laneti bir şebeke (grid) üzerinde hesaplama için ihtiyacı dramatik bir şekilde etkiler. Örneğin, k noktalı bir sisteme sahip bir-boyutlu bir problemi göz önüne alınsın. Bir kenarda, aynı yoğunluğu elde etmek için, d boyutta, k^d adet şebeke (grid) noktası gerekmektedir, öyle ki verinin boyutunda hesaplama karmaşıklığı üssel olarak arttığı ima edilmektedir. Böylece hesaplama karmaşıklığı, boyutla beraber üssel olarak artmaktadır ve ayrıca gerekli sayıda veri miktarı ihtiyacı da üssel olarak boyutla birlikte artmaktadır⁵³.

Daha önceki yazarlarında belirttiği gibi, kernel (çekirdek) metotlar, boyut arttıkça yanlılığın (bias) artmasından acı çekmektedirler. Kernel metotlar tarafından, tam yoğunluğun doğrudan tahmini, altı boyuta kadar mümkündür. Fakat kernel metotların altı boyuttan öte faydalı olmadığı manasına gelmez. Gerçekten, kernel metotlar istatistik ayırmacılık (statistical discrimination) gibi amaç için düzinelerce boyutta güçlü araçlardır⁵⁴.

⁵² Maimon, Rokach, **a.g.e.**, p.159-161.

⁵³ Wegman, Solka, **a.g.e.**, p.22.

⁵⁴ D.W. Scott, Stephan R.Sain, “Multidimensional Density Estimation”, **Handbook of Statistics 24: Data Mining and Data Visualization**, Ed. by C.R. Rao, E.J.Wegman, J.L.Solka, Elsevier B.V, 2005, p.257.

Sınıflandırma modeli oluşturulduktan sonra, bu model en iyi modeli seçmek için diğer modeller ile kıyaslanacaktır. Bu, kıyaslama kriterlerinden olan ölçeklenebilirlik (scalability), esas alınarak modelin giderek büyüyen veri hacmini işleme yeteneğine atıfta bulunur. İkinci olarak, ölçeklenebilirlik kriteri, verinin farklı alanlardan işleme yeteneğini de ifade edebilir⁵⁵.

1.5 İstatistik Öğrenme Teorisi

Bu kısımda maksimum –marjlı sınıflandırıcı kavramı istatistik öğrenme teorisi bağlamında formüle edilir ve mümkün olan en büyük marjlı doğrusal bir karar yüzeyi optimal sınıflandırıcıyı gerçekten temsil ettiği gösterilir.

Model karmaşıklığı VC-boyut açısından tanımlanır. Küçük marjlı karar yüzeyleri yüksek bir VC-boyuta sahiptirler ve bundan dolayı küçük bir VC-boyuta sahip olan büyük marjlara sahip karar yüzeylerinden daha çok karmaşık olarak kabul edilirler.

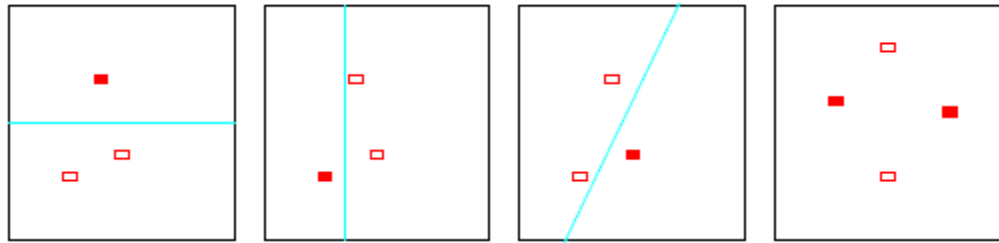
1.5.1 Vapnik – Chervonenkis (VC) Boyut

Vapnik – Chervonenkis (VC dimension) boyut bir fonksiyonlar sınıfının elemanlarının ne kadar kıvrak olduğunun belirlenerek, bu sınıfın karmaşıklığını ölçen bir yoldur⁵⁶. Örnekleme hatasının tahminlerinin kullanılmasındaki zorluk, uyumda kullanılan parametrelerin ya da karmaşıklığın (d) sayılarının belirtilmesidir. Vapnik-Chervonenkis (VC) teorisi karmaşıklığın böyle bir genel ölçümünü sağlamaktadır.

⁵⁵ (Filorin Gorunescu, **Data Mining: Concepts, Models and Techniques** (Intelligence System Library, Volume 12), Springer-Verlag Berlin Heidelberg, 2011, p.16.

⁵⁶ Trevor Hastie, R. Tibshirani, J. Friedman, **The Elements of Statistical Learning: Data Mining, Inference, and Prediction** (2nd.Edition), New York, Springer Science+Business Media LLC, 2009, p.238.

Varsayalım $\{f(x, \alpha)\}$ gibi bir fonksiyonlar sınıfı mevcut olsun ve α gibi bir parametre vektörü tarafından taransın (indekslensin), $x \in \mathbb{R}^p$ olsun. Şimdilik f , bir gösterge (indikatör) fonksiyon olsun, yani 0 veya 1 değerlerini alır. Eğer $\alpha = (\alpha_0, \alpha_1)$ ve f doğrusal gösterge fonksiyonlarsa $I(\alpha_0 + \alpha_1^T x > 0)$, o zaman sınıf f 'in karmaşıklığının $p+1$ demek söylemek makuldür. Aşağıdaki figürde, üçüncü düzlemdeki doğrusal gösterge fonksiyonlarının VC boyutunu göstermektedir. Fakat dördüncü düzlemde yoktur, çünkü bir doğrular kümesi tarafından dört adet nokta parçalara ayrılamadığından, son düzlemde VC boyut mevcut değildir. Genel olarak, p boyuttaki bir doğrusal gösterge fonksiyonu $p+1$ adet VC-boyut'a sahiptir. Bu ise serbest parametrelerin sayısını göstermektedir. İlk üç panelde düzlemdeki doğrular sınıfı üç noktayı parçalara ayırdığını göstermektedir. Son panel bu sınıfın dört noktayı parçalara ayırmadığını göstermektedir, çünkü boş noktaları bir tarafa ve dolu noktaları diğer tarafa koyacak şekilde hiçbir doğru bulanamaz. Böylece, düzlemdeki doğrular sınıfının VC-boyut'u üç'tür. Doğrusal olmayan bir sınıf dört noktayı parçalara ayırabilir ve böylece VC boyut üç'ten büyük olur.



Şekil 1. VC-Boyut⁵⁷

VC-boyut tahmin hatasının bir tahminini oluşturmak için kullanılabilir ve farklı sonuç türleri mevcuttur. VC-boyut kavramını kullanarak, bir fonksiyonlar sınıfı kullanıldığı zaman eğitim hatasının iyimserliği hakkında sonuçlar kanıtlanabilir.

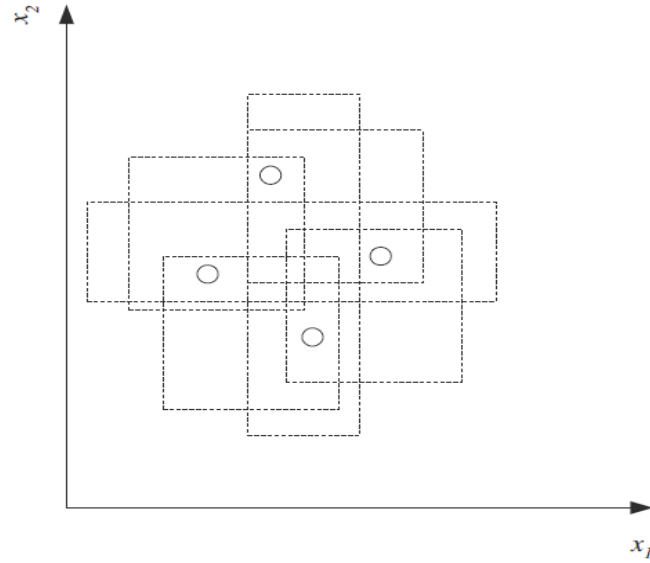
⁵⁷ A.e., p.238.

Kabul edelim ki, bizim elimizde N adet noktadan oluşan bir veri kümesi mevcut olsun. Bu N nokta, pozitif ve negatif olarak 2^N yolla etiketlenebilir. Bundan dolayı, 2^N farklı öğrenme problemi N veri noktaları tarafından tanımlanabilir. Bu problemlerin herhangi biri için, pozitif örnekleri negatif örneklerden ayıran $h \in \mathcal{H}$ olacak şekilde, bir hipotez bulabiliriz. O zaman $\mathcal{H}$, N adet noktaları *parçalara ayırır* deriz. Yani, N adet örnek yolu ile tanımlanabilen herhangi bir öğrenme problemi $\mathcal{H}$ den çekilen bir hipotez yolu ile hatasız öğrenilebilir. $\mathcal{H}$ tarafından parçalara ayrılabilen maksimum noktaların sayısı, $\mathcal{H}$ 'nin *Vapnik-Chervonenkis* (VC) boyutu adı verilir, ve $VC(\mathcal{H})$ olarak gösterilir, ve $\mathcal{H}$ 'nin kapasitesini ölçer⁵⁸.

Şekil 2'de eksen hizalı dikdörtgen, dört noktayı iki boyut olacak şekilde parçalara ayırabilir. O zaman, $VC(\mathcal{H})$ değeri dörttür, öyle ki burada $\mathcal{H}$, iki boyutta eksen hizalı dikdörtgenin hipotez sınıfıdır. VC boyutu hesaplarken, parçalara ayrılabilen dört noktanın bulunması yeterlidir; herhangi dört noktanın iki boyutta parçalara ayrılması çok gerekli değildir. VC boyut karamsar görünebilir. Hipotez sınıfı olarak bir dikdörtgen kullanarak sadece dört nokta içeren veri kümelerini öğrenebiliriz, daha büyük değil. Buda faydalı değildir. Fakat VC boyut örneklerin çekildiği olasılık dağılımından bağımsız olduğundan kaynaklanmaktadır. Gerçek hayatta, dünya pürüzsüzce değişiyor, birbirine yakın örnekler çoğu zaman aynı etikete sahipler ve olası tüm etiketleme konusunda endişelenmemize gerek yoktur. Hipotez sınıfımız tarafından öğrenilebilir dörtten daha fazla veri noktalarına sahip birçok veri kümeleri mevcuttur. Bundan dolayı, küçük VC-boyutlu hipotez sınıfları bile uygulanabilirler ve daha büyük VC boyutlulara tercih edilmektedirler⁵⁹.

⁵⁸ Alpaydın, **a.g.e.**, p.27

⁵⁹ Alpaydın, **a.g.e.**, p.28.



Şekil 2. Eksen hizalı dikdörtgen dört noktayı parçalara ayırabilir. Sadece iki noktayı içeren dikdörtgenler gösterildi.

1.5.2 VC- Boyut ve PAC Öğrenimi arasındaki ilişki

VC boyut ile PAC öğrenimini birbirine bağlayan iki teorem mevcuttur.

Teorem (Blumer, v.d., 1990): H hipotez uzayı, bir PAC öğrenilebilirdir, eğer bu uzay sonlu VC boyuta sahip ise.

Teorem: Bir hipotezler seti, H , PAC düzgün PAC öğrenilebilirdir eğer:

- $m \geq (1/\epsilon) \max \left[4 \log \left(\frac{2}{\delta} \right), 8VCdim \log \left(\frac{13}{\epsilon} \right) \right]$, ve
- çok terimli zamandaki (m 'de ve n 'de) eğitim kümesi ile uyumlu olan $h \in H$ gibi bir hipotez çıktısı veren bir algoritma mevcut ise. Burada m , örnek sayısını ve n ise boyut sayısını belirtmektedir. ϵ , doğruluk parametresini belirtir, ayrıca δ güven parametresini belirtir⁶⁰.

⁶⁰ Nilsson, a.g.e., p. 118.

1.5.3 VC- Boyut ve Model Karmaşıklığı

VC- boyut, bir sınıflandırıcının karmaşıklığının ölçüsüdür. Yani, bu ölçüm bir ikili sınıflandırıcının iki sınıf arasındaki sınırı ne kadar iyi ölçmesi olarak bilinir. Eğer VC – boyut çok büyük ise sınıflandırıcı daha karmaşık bir hal alır ve böylece ilgili sınıflara ait olan gözlemleri daha iyi bölebilir. Daha açık bir ifadeyle, karmaşık sınıflandırıcılar karmaşık sınıf sınırlarını daha az karmaşık sınıflandırıcılara oranla daha iyi modeller. Daha doğrusu, bir model ne kadar çok karmaşık ise, bir eğitme kümesindeki gözlemleri daha iyi böler⁶¹.

VC- boyutun model karmaşıklığını ölçmek için, bir yol olarak VC – boyutun formüleleştirilmesi gerekir. Aynı γ boyut’lu, marja sahip doğrusal sınıflandırıcıların bir sınıfını göz önüne alalım. $\hat{F}[\gamma]$, ilgili sınıfı ifade etsin. Model sınıf olan $\hat{F}[\gamma]$ ’nın rotasyon ve çevirisi altında kapalı olduğunu varsayalım; yani γ marjlı mümkün olan tüm modelleri içerir. Tüm $\hat{f} \in \hat{F}[\gamma]$ modeller için, tüm ρ rotasyonları için ve tüm τ çevirileri için, elimizde $\rho(\hat{f}) \in \hat{F}[\gamma]$ ve $\tau(\hat{f}) \in \hat{F}[\gamma]$ vardır. Herhangi iki model için, $\hat{f}, \hat{g} \in \hat{F}[\gamma]$, her zaman τ gibi bir çeviri ve ρ gibi bir rotasyon oluşturulur, öyle ki $\rho(\tau(\hat{f})) = \hat{g}$. Burada, destek hiperdüzlemlerle birlikte doğrusal bir karar yüzeyi ve γ boyutlu marjı, γ genişliğe sahip bir dikdörtgen olarak biçimlendirilir. Dikdörtgeni uzunlamasına ikiye ayıran bir doğru, gerçek bir karar yüzeyini temsil edecek şekilde düşünülebilir.

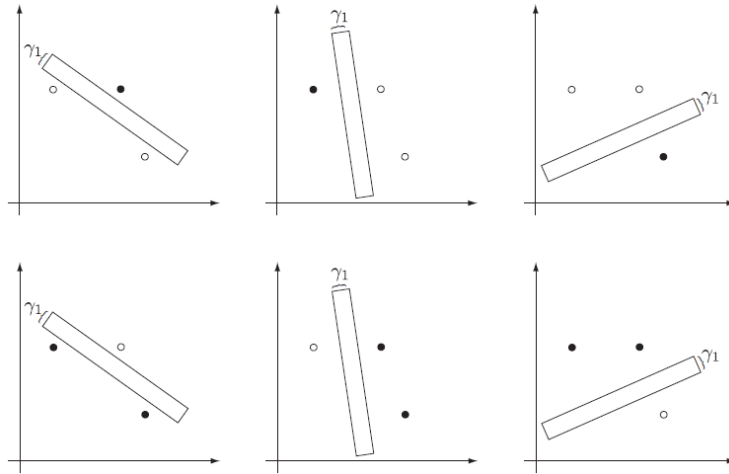
Sabit marj genişliğe sahip bir model sınıfın kavramı ile, VC-boyut daha sıkı bir şekilde tanımlanabilir. D gibi bir veri kümesinin göz önüne alınsın, o zaman bir model sınıftaki sınıflandırıcının VC-boyutu, D ’deki örneklerle olası tüm ikili etiket atamaları için model sınıftaki sınıflandırıcılar tarafından bölünebilen D ’deki örneklerin sayısıdır. Eğer D ’deki tüm örnekler olası tüm etiket atamaları için bölünebilirse, model sınıf D

⁶¹ Hamel, a.g.e, p.173.

veri kümesini *parçalar* (*shatter*) deriz. Bu sonuç bize VC-boyutun aşağıdaki tanımını verir:

Tanım: Bir model sınıfın olan ve D veri kümesi üzerinde tanımlanan $\hat{F}[\gamma]$ 'nin VC- boyutu, $\hat{F}[\gamma]$ tarafından parçalananan D 'nin en büyük sonlu alt kümesinin büyüklüğüdür⁶².

İçinde üç örnek olan bir veri kümesinin iki- boyutlu reel uzayda göz önüne alalım. Yani, $D \in \mathbb{R}^2$ ve $|D| = 3$. $\hat{F}[\gamma_1]$ D üzerinde tanımlansın. Marj γ_1 öyle seçilsin ki, olası tüm ikili etiket atanmaları için tüm üç örnek ayrılabilir. Bu durumda $\hat{F}[\gamma_1]$ 'nin VC-boyutu 3'tür ve $h_1 = 3$ yazılır. VC- boyut veri kümesinin boyutuna eşit olduğundan $h_1 = |D|$, bu durumda $\hat{F}[\gamma_1]$, D 'yi parçalar. Bu durum aşağıdaki şekil 3 gösterilmiştir.

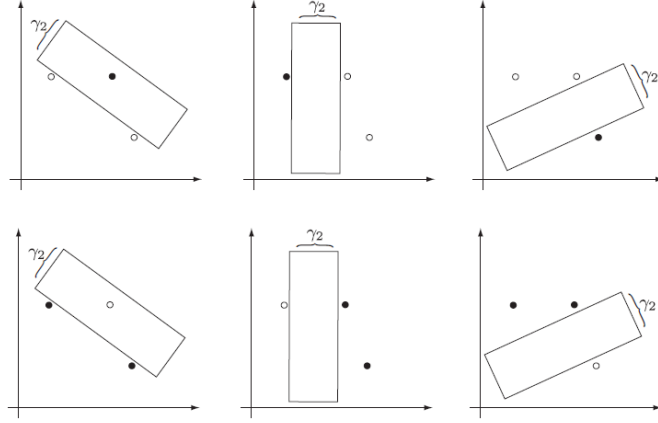


Şekil 3. γ_1 marjlı sınıflandırıcıların bir sınıfının VC-boyutu

Dolu ve boş toplar, D 'deki örneklerle farklı ikili etiket atamalarını göstermektedir. Tüm noktaların aynı etiket ile etiketlendiği durumlarda bölme imkânı olmadığı için göstermeğe gerek yoktur.

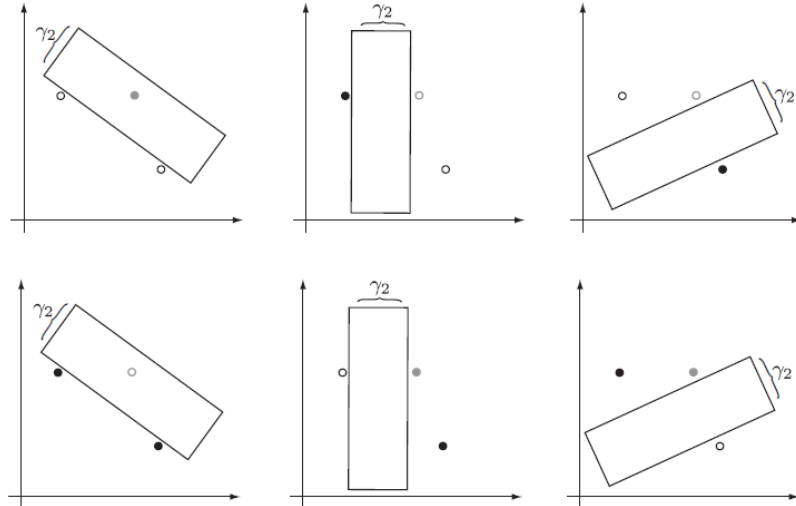
⁶² Hamel, a.g.e, p.173.

Aynı veri kümesi üzerinde, $\gamma_2 > \gamma_1$ olacak şekilde, $\hat{F}[\gamma_2]$ gibi sınıflandırıcıların ikinci bir sınıfını ele alalım. Özellikle, γ_2 'nin boyutu $\hat{F}[\gamma_2]$ 'deki sınıflandırıcılar tüm örnekleri mükemmel olarak bölemeyecek şeklindedir. Bu durum aşağıdaki şekilde gösterilmektedir.



Şekil 4. γ_2 marjlı sınıflandırıcılar verilen veri kümesini parçalamıyor

Eğer D 'deki bir örneklerden bir tanesini silersek, $\hat{F}[\gamma_2]$ 'deki sınıflandırıcılar D 'nin bu altkümesini parçalayabilir. Bu durumda aşağıda gösterilmiştir:

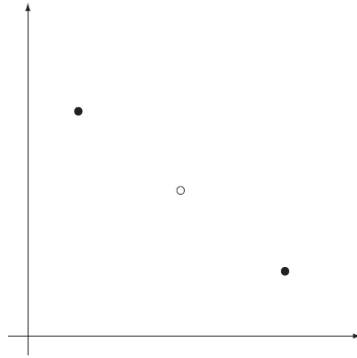


Şekil 5. γ_2 marjlı sınıflandırıcılar veri kümesinin bir altkümesini parçalayabilir.

Burada açık gri renkli nokta silinen örneği temsil ediyor

Burada açık gri renkli nokta, silinen örneği temsil ediyor. En soldaki sütundaki grafikler önemsiz ayrımları (bölmeleri) belirtir. Çünkü D'deki geri kalan örnekler aynı etiket ile etiketlenirilmişler. $\hat{F}[\gamma_2]$ iki elemanlı olan D'nin bir alt kümesini bölebildiği için, model sınıf olan $\hat{F}[\gamma_2]$ için VC-boyut 2'dir; $h_2 = 2$.

Büyük VC-boyut yüksek karmaşıklığa sahip modellerin sınıflarını belirtir. Yukarıdaki durumda, $h_1 > h_2$ olduğundan, $\hat{F}[\gamma_1]$ 'deki sınıflandırıcılar $\hat{F}[\gamma_2]$ 'daki sınıflandırıcılara göre daha karmaşıktır ve $\hat{F}[\gamma_1] \supset \hat{F}[\gamma_2]$ olarak yazılır. $\gamma_1 < \gamma_2$ olduğundan, bu bize küçük marjlı sınıflandırıcıların büyük sınıflandırıcılara göre çok daha karmaşık olduğunu ima eder. Ayrıca, karmaşık modeller daha az karmaşık modellere kıyasla veriye daha iyi uyabilirler. $\hat{F}[\gamma_1] \supset \hat{F}[\gamma_2]$ bize daha az karmaşık sınıflandırıcılar ile modellenen bir şey aynı zamanda çok daha karmaşık sınıflandırıcılar ile de modellenebileceğini belirtmektedir. Bir model sınıfın VC- boyutu, veriye bağımlıdır. Şekil 6'daki örneklerin bir düzenlemesi verilmektedir, burada model sınıfın marjını ne kadar küçük yapılırsa yapılsın herhangi bir model sınıfın sınıflandırıcıları tarafından parçalanmadığını gösterir.⁶³



Şekil 6. Marj ne kadar küçük olursa olsun, herhangi bir model sınıfın sınıflandırıcıları tarafından veri kümesi parçalanamaz

⁶³ Hamel, a.g.e, p.174.

Buradaki teorik gelişimde, tüm sınıflandırıcıların doğrusal çekirdek (linear kernel) fonksiyonunu kullandığını varsayıldı ve sabit (hard) marj dikkate alındı. Yukarıdaki oluşumlara, Polinom ve Gaussian çekirdek (kernel) fonksiyonlarına sahip sınıflandırıcılar dâhil ederek genişletilebilir. Bunun sonucu olarak, bazı veri kümelerindeki örnekler yukarıda gösterilen doğrusal karar yüzeyleri yerine doğrusal olmayan karar yüzeyleri yolu ile parçalanabilirler. Dolayısıyla, birey model sınıflarının karmaşıklığı marj boyutunun yanı sıra, maliyet sabiti ve birey model sınıflarının modellerinde kullanılan çekirdek (kernel) türüne de bağlıdır⁶⁴.

1.5.4 Makine Öğrenimi İçin Teorik bir Çerçeve

Makine öğreniminin teorik oluşumunda merkez olan matematiksel beklenti kavramı şu şekilde tanımlanır: $E[g] = \int g(x)dP(x)$, burada $g(x)$, $x \in X$ olacak şekilde tanım bölgesi üzerinde tanımlı bir fonksiyondur ve $P(x)$, X üzerinde bir olasılık dağılımıdır. $E[g]$, X tanım bölgesi üzerindeki fonksiyon değerlendirmelerinin toplamını ifade eder. Eğer k elemanlı $(x_1, \dots, x_k)$ tanım bölgesi olan X ayrık (discrete) ise , beklenen değer şu şekilde ifade edilir:

$$E[g] = \frac{1}{k} \sum_{i=1}^k g(x_i). \quad (1.7)$$

$E[g]$ beklenen değeri, tanım bölgesi X 'deki tüm fonksiyon ölçümlerinin ortalaması olarak adlandırıyoruz (Walpole, R.E. v.d., 2002). Şimdi, matematiksel beklentiği bizim sınıflandırma kavramın uygulayalım. Kabul edelim $p(\bar{x}, y)$ veri örneklerinin, $\bar{x} \in \mathbb{R}^n$, bileşik olasılığı olsun ve bunlara karşılık gelen etiketler $y \in \{+1, -1\}$; ayrıca L , 0-1 kayıp fonksiyonu olduğunu varsayalım, o zaman veri evreni üzerinde tanımlanan bazı $\hat{f} \in \hat{F}[\gamma]$ modeli için beklenen kayıp (expected loss) aşağıdaki gibi ifade edilir:

⁶⁴ A.e., pp. 172-175.

$$E[L(y, \hat{f}(\bar{x}))] = \int L(y, \hat{f}(\bar{x}))dP(\bar{x}, y). \quad (1.8)$$

Başka bir deyişle beklenen kayıp bir modelin temel veri evreninin üzerinde oluşturacağı beklenen hataların sayısıdır. Çoğu zaman aşağıdaki gibi ifade edilir:

$$R[\hat{f}] = E[L(y, \hat{f}(\bar{x}))], \quad (1.9)$$

burada $R[\hat{f}]$, beklenen risktir (expected risk). Bununla makine öğrenimi beklenen riskin minimizasyonudur,

$$\hat{f}^* = \operatorname{argmin}_{\hat{f} \in \hat{F}} R[\hat{f}], \quad (1.10)$$

burada $\hat{F}$, tüm model sınıfların sınıfını temsil eder, öyle ki tüm γ marjları için $\hat{F}[\gamma] \subset \hat{F}$. Optimizasyon sırasında beklenen riski minimize eden optimal model $\hat{f}^* \in \hat{F}[\gamma^*]$ bulmak için bu sınıftan modellerimiz olan $\hat{f}$ yazılır⁶⁵.

1.5.5 Ampirik Risk Minimizasyonu (ARM)

Deneyisel risk minimizasyonu (DRM) istatistik öğrenme teorisinde öğrenme algoritmalarının bir ailesini tanımlayan prensiptir ve öğrenme algoritmalarının performansında kuramsal sınırlar vermesi için kullanılır⁶⁶.

PAC (probably approximately correct) öğrenme teorisi, deneyisel (ampirik) risk minimizasyon (ERM) ile örnekler üzerinden ilgilenir. Modelsiz durumda, gerçek değerli bir fonksiyonun dağılımı olmaksızın öğrenme için bir çerçeve sağlar^{67,68}. Destek Vektör

⁶⁵ Hamel, **a.g.e**, p.176.

⁶⁶ Wikipedia, “Empirical risk minimization”, (Çevrimiçi) http://en.wikipedia.org/wiki/Empirical_risk_minimization#cite_ref-0, 2011.

⁶⁷ Viladimir V. Vapnik, **Estimation of Dependences Based on Empirical Data**, New York, NY, Springer –Verlag, 1982, pp.27-29.

metodu, genel formunda, kernel–tabanlı deneysel risk minimizasyon (ERM) algoritmasıdır. Güzel özelliklerinden dolayı, popülerliği artmaktadır ve deneysel (ampirik) performansı açısından ümit vericidir^{69,70}.

Yukarıda verilen $\hat{f}^* = \operatorname{argmin}_{\hat{f} \in \hat{F}} R[\hat{f}]$, formülasyonunda, $p(\bar{x}, y)$ bileşik olasılık dağılımını bilinmediğinden, makine öğrenimi mümkün değildir. Fakat D eğitme verisinde, gözlemler şeklinde bileşik olasılık dağılımı hakkında bazı bilgiye sahibiz,

$$D = \{(\bar{x}_1, y_1), \dots, (\bar{x}_l, y_l)\} \subset \mathbb{R}^n \times \{+1, -1\}. \quad (1.11)$$

Bu gözlemleri kullanarak risk tahmin edilebilir. Buna bazı model $\hat{f}$ 'lerin deneysel riski $R_{emp}[\hat{f}]$ adı verilir ve aşağıdaki şekilde tanımlanır:

$$R_{emp}[\hat{f}] = E[L(y, \hat{f}(\bar{x}))] = \frac{1}{l} \sum_{i=1}^l L(y_i, \hat{f}(\bar{x}_i)), \quad (1.12)$$

burada, $(\bar{x}_i, y_i) \in D$. Eğitme kümeleri genellikle sonlu olduğundan ve ayrık örnekler içerdiğinden dolayı ayrık matematiksel beklenti tanımını $E[g] = \frac{1}{k} \sum_{i=1}^k g(x_i)$ uygulanıyor. Beklenen riske benzer şekilde, deneysel riski minimize ederek en iyi model elde edilir:

$$\hat{f}^* = \operatorname{argmin}_{\hat{f} \in \hat{F}} R_{emp}[\hat{f}] = \operatorname{argmin}_{\hat{f} \in \hat{F}} \left(\frac{1}{l} \sum_{i=1}^l L(y_i, \hat{f}(\bar{x}_i)) \right). \quad (1.13)$$

Deneysel (ampirik) riski minimize etmek için olası tüm $\hat{F}$ modellerinin sınıfından bir model seçimine izin verildiği için, deneysel riski sıfıra ya da sıfıra yakın bir değere

⁶⁸ Vladimir V. Vapnik, Statistical Learning Theory, 1998) **Statistical Learning Theory**, New York, NY, Wiley, 1998, p.79.

⁶⁹ B. Schölkopf, C.J.C. Burges, A.J. Smola, **Advances in Kernel Methods: Support Vector Learning**, Cambridge, MA, MIT Press, 1999, pp.216-218.

⁷⁰ C. Cortes, Vladimir V. Vapnik, “Support Vector Networks”, **Machine Learning**, vol.20, 1995, pp. 273-297.

indiren bir model her zaman bulunur. Yani, ampirik (deneysel) riski minimize etmek aşırı iyimser olması muhtemeldir⁷¹.

1.5.6 VC- Güven

Beklenen risk, bir veri evreni için, optimal model bulmak için kullanılamayabilir. Deneysel riskin minimizasyonu çok iyimser olduğundan, optimal model bulmak için kullanılamayabilir. VC- boyuta dayalı bir modelin genelleme hatasının bir ölçümü, Vapnik tarafından ortaya atıldı. Bu ölçüm *VC-güven* olarak adlandırılır.

$$v(l, h, \eta) = \sqrt{\frac{h(\log(2l/h)+1)-\log(\eta/4)}{l}}, \quad (1.14)$$

burada l eğitme verisinin büyüklüğü, h ise el altındaki model sınıfın VC-boyutudur. η ise, $0 < \eta < 1$ olacak şekilde, bazı küçük sayıları belirtir. Burada görüleceği gibi, VC-güven olan v değeri, VC- boyut değeri olan h ile direkt olarak orantılıdır.

Sezgisel olarak büyük bir VC-boyut, karmaşık bir model anlamına gelir. Dolayısıyla bu, büyük bir genelleme hatası anlamına gelir. Çok karmaşık modeller eğitme verilerini aşırı uydurma (overfit) eğilimindedir. Bundan dolayı iyi genelleme yapamazlar. Bu, VC-güvenin tam olarak yakalamaya çalıştığı durumdur. Ayrıca, VC-boyut ters olarak veri kümesinin l büyüklüğü ile ters orantılı olduğu görülmektedir. Bunun anlamı eğitme verisi ne kadar büyük olursa, veri evreni hakkında o kadar çok şey biliyoruz demektir ve böylece genelleme hatası o kadar çok küçüktür.

Deneysel risk ile berber VC-güven bazı model $\hat{f}$ 'ler için bir üst sınır olarak kabul edilebilir.

$$R[\hat{f}] \leq R_{emp}[\hat{f}] + v(l, h_{\hat{f}}, \eta), \quad (1.15)$$

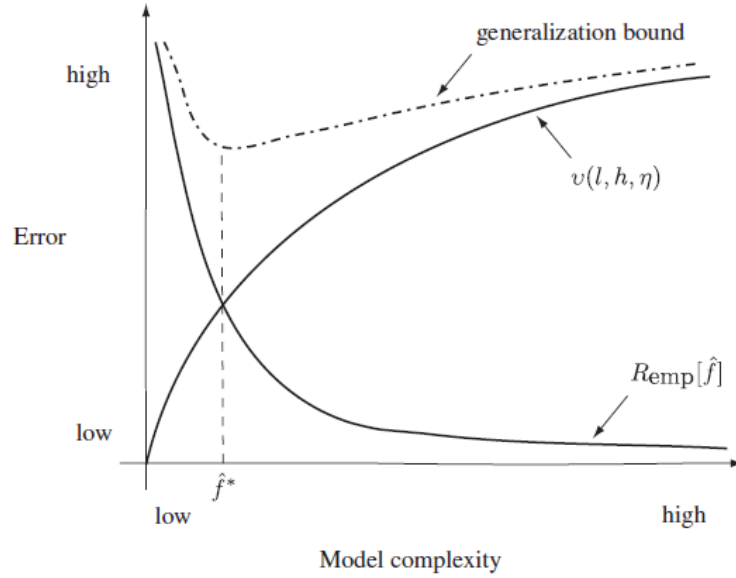
⁷¹ Hamel, a.g.e, p.177.

burada $h_{\hat{f}}$, $\hat{f}$ 'nin VC-boyutudur ve l eğitme verisinin boyutunu gösterir. Yani, deneysel (ampirik) risk ve VC-güven verildiğinden, bütün temel veri evreni üzerinde modelin beklenen kaybında bir üst sınır tahmin edilebilir. Vapnik bu üst sınırın $1 - \eta$ bir olasılıkla geçerli olduğunu göstermiştir.

Şekil7, deneysel risk $R_{emp}[\hat{f}]$ ile VC-güven $v(l, h_{\hat{f}}, \eta)$ arasındaki ilişkiyi göstermektedir. Modelin karmaşıklığı arttıkça, deneysel(ampirik) risk azalmaktadır. Buda bize karmaşık modellerin eğitme verilerinin çok iyi modellediğini göstermektedir. Burada, daha az karmaşık modellere kıyasla, karmaşık modeller eğitme verisinde olmayan veri üzerinde daha çok hatalar işleyecektir. Deneysel risk ve VC-güvenlerin toplamı, bu iki eğrinin bir zarfını temsil eder. Bu zarf çoğu zaman genelleştirme sınırı, (generalization bound) olarak bilinir. Test hatasının minimize edilmesi bize optimal modeli verdiğinden, genelleme sınırının minimize edilmesi model karmaşıklığı ve genelleme hatası arasında en doğru ödünleşim (trade-off) oluşturma ile eş değerdir. Bu da, optimal modeli verecektir. Bu bize optimal model $\hat{f}^*$ bulmak için aşağıdaki optimizasyon problemini verecektir,

$$\hat{f}^* = \underset{\hat{f} \in \hat{F}}{\operatorname{argmin}} (R_{emp}[\hat{f}] + v(l, h_{\hat{f}}, \eta)), \quad (1.16)$$

burada, $\hat{F}$ tüm model sınıflarının süper sınıfıdır. Şekilde görüldüğü gibi, bu optimizasyon probleminin çözümü genelleme sınırının minimum değerini veren modeldir.



Şekil 7. Deneysel risk ve VC-güven arasındaki ilişki. Bunların toplamı geneleme sınırını verir⁷²

$$R[\hat{f}] \leq R_{emp}[\hat{f}] + v(l, h_{\hat{f}}, \eta) \quad (1.17)$$

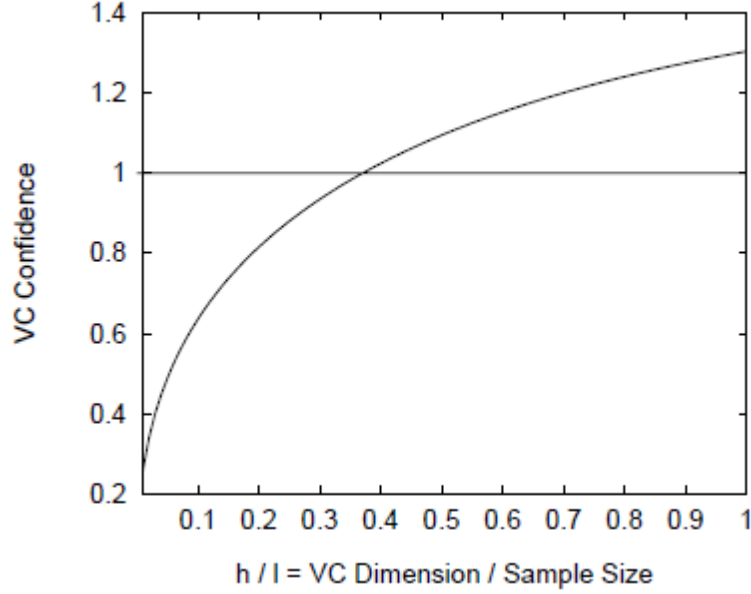
denklemindeki VC – güven seçilen fonksiyonlar sınıfına bağlıdır, buna karşın ampirik risk ve gerçek risk eğitme prosedürü yolu ile seçilen belirli bir fonksiyona bağlıdır. Seçilen fonksiyonların o alt kümesini bulunulması istenir, öyle ki bu altküme için risk sınırı minimize edilmiştir⁷³.

Aşağıdaki şekil 8’de $R[\hat{f}] \leq R_{emp}[\hat{f}] + v(l, h_{\hat{f}}, \eta)$ denkleminin sağ tarafı h ile değiştiğini göstermektedir, 95% güven düzeyi ($\eta=0.05$) ve eğitme örneği büyüklüğü 10000 verildiği farz edilmektedir. Burada görüldüğü gibi, VC güven h ’nin bir monotonik artan fonksiyonudur. Bu l ’nin herhangi bir değeri için geçerlidir⁷⁴.

⁷² Hamel, a.g.e, pp.177-179.

⁷³ J.B. Christopher, “A Tutorial on Support Vector Machines for Pattern Recognition”, **Data Mining and Knowledge Discovery**, 2, 1998, pp. 121-167.

⁷⁴ A.e. p.127.



Şekil 8. VC-Güven h 'de monotoniktir

1.5.7 Yapısal Risk Minimizasyonu (YRM)

Yapısal risk minimizasyonu (Structural Risk Minimization - SRM) (Vapnik ve Chervonekiz, 1974) tarafından geliştirildi. Öğrenme için sonlu eğitime veri kümelerinden model seçimi için bir tümevarımsal prensiptir. Kapasite kontrolün genel bir modelini açıklar ve ayrıca hipotez uzay karmaşıklığı (yaklaşım fonksiyonlarının VC boyutu) ile eğitime verisinin uyma kalitesi (ampirik hata) arasında ödünleşim (trade - off) eğrisi sağlar. Süreç aşağıdaki gibidir⁷⁵.

- 1- Tanım kümesinin önsel bilgisini kullanarak, n dereceli polinomlar, n adet gizli katman nöronlara sahip sinir ağlar, n düğümlü spline fonksiyonlar bir kümesi veya n kurallı bulanık mantık modeller gibi fonksiyonlar sınıfı seçilir.
- 2- Karmaşıklığı artırmak için fonksiyonlar sınıfını iç içe alt kümelerin bir hiyerarşisi olacak şekilde böler.

⁷⁵ M. Sewell, "Structural Risk Minimization" (Çevrimiçi) <http://www.svms.org/srm/>, 2006.

- 3- Her bir alt kümede ampirik (deneysel) risk minimizasyonu gerçekleştirilir (parametre seçimi)
- 4- Serideki ampirik riski ve VC boyutu toplamı en az olan model seçilir

Yapısal Risk minimizasyonu (SRM) (Vapnik, 1995) karmaşıklıkları açısından modellerin bir kümesini sıralı olarak kullanır. Örneğin artan dereceli polinomlar örnek olarak verilebilir. Karmaşıklık genellikle serbest parametrelerin sayısı tarafından verilir⁷⁶.

Uygun bir veri kümesi verildiğinde,

$$\hat{f}^* = \underset{\hat{f} \in \hat{F}}{\operatorname{argmin}} (R_{emp}[\hat{f}] + v(l, h_{\hat{f}}, \eta)) \quad (1.18)$$

denklemini, $\hat{F}$ 'de karmaşıklık ve genelleme hatası arasında en uygun ödünleşim (trade-off) yapan bir model bulmamıza izin verir. Verilen optimizasyon probleminde, bu belirtilen modeli tam olarak nasıl bulunur? Genellikle model sınıfı olan $\hat{F}$ sınırsızdır ve optimal model olan $\hat{f}^*$ bulmak için fazladan bir kılavuza gerek duymaksızın körü körüne gezinmesi muhtemelen verimli bir çaba değildir. Vapnik, model sınıfı $\hat{F}$ 'i körü körüne gezindirmek yerine, $\hat{F}$ 'in birey model alt sınıflarının VC-boyutları, optimal modeli bulmak için bir rehber olarak hizmet verebilir. Vapnik bu yöntemi yapısal risk minimizasyonu (structural risk minimization) olarak adlandırır. $\hat{F}[\gamma_1], \dots, \hat{F}[\gamma_k] \subset \hat{F}$ olacak şekilde lineer modellerin $\hat{F}$ bir sınıfa sahip olduğumuzu kabul edelim, burada

$$\hat{F}[\gamma_1] \subset \hat{F}[\gamma_2] \subset \dots \subset \hat{F}[\gamma_k] \text{ eğer } h_1 < h_2 < \dots < h_k, \quad (1.19)$$

⁷⁶ Alpaydın, **a.g.e.**, pp.80-81

burada h_i model sınıfı olan $\hat{F}[\gamma_i]$ 'nin VC-boyutudur. Lineer modellerin varsayıldığı farz edilirse, yukarıdaki denklem $(\hat{F}[\gamma_1] \subset \hat{F}[\gamma_2] \subset \dots \subset \hat{F}[\gamma_k])$ eğer $h_1 < h_2 < \dots < h_k$ çeşitli model sınıflarının marjları da kısmi olarak sıralıdır,

$$\gamma_k < \dots < \gamma_2 < \gamma_1.$$

Şimdi optimal modeli bulmak için elimizde etkili bir yöntem vardır. En az karmaşık model sınıfı olan $\hat{F}[\gamma_1]$ ile başlanılır ve

$$\hat{f}^* = \underset{\hat{f} \in \hat{F}}{\operatorname{argmin}} (R_{emp}[\hat{f}] + v(l, h_{\hat{f}}, \eta)) \quad (1.20)$$

denklemine göre genelleme sınırı minimize edilir. Keyfi bir model $\hat{f}_1 \in \hat{F}[\gamma_1]$ seçilir ve daha sonra uygun bir ρ rotasyonu ve τ çevirisi hesaplanır öyle ki

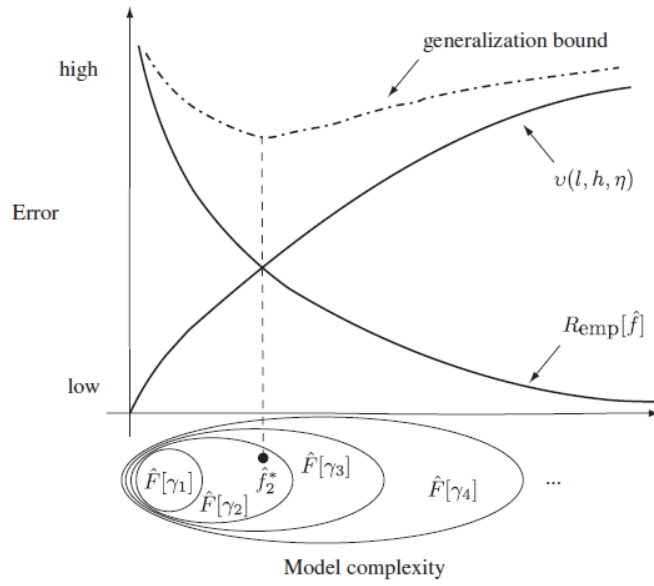
$$\hat{f}_1^* = \rho(\tau(\hat{f}_1)) = \underset{\hat{f} \in \hat{F}[\gamma_1]}{\operatorname{argmin}} (R_{emp}[\hat{f}] + v(l, h_{\hat{f}}, \eta)). \quad (1.21)$$

Daha sonra bir sonraki model sınıfı ele alınır, yani bu durumda $\hat{F}[\gamma_2]$, ve optimal model olan $\hat{f}_2^*$ aynı şekilde bir önceki optimal model gibi hesaplanır. Eğer bazı modellerin $\hat{f}_{i+1}^* \in \hat{F}[\gamma_{i+1}]$ genelleme sınırı $\hat{f}_i^* \in \hat{F}[\gamma_i]$ modelinin genelleme sınırından büyük durumu bulunursa o zaman arama sonlandırılır. Bu durumda,

$$\hat{f}^* = \underset{\hat{f} \in \hat{F}}{\operatorname{argmin}} (R_{emp}[\hat{f}] + v(l, h_{\hat{f}}, \eta)) \quad (1.22)$$

denklemine göre, optimal model $\hat{f}_i^*$ 'dir. Ayrıca, γ_i marjı optimal model performansını veri evreni üzerinde tamamen garanti eden maksimum marjdır. Bu yöntem yapısal risk minimizasyonu (structural risk minimization) olarak adlandırılır. Çünkü model sınıfların yapısı aramaya kılavuzluk etmek için kullanıldı. Aşağıdaki şekil yapısal risk

minimizasyonunu göstermektedir. Burada dört tane gittikçe artan karmaşık model sınıfları $\hat{F}[\gamma_1] \subset \hat{F}[\gamma_2] \subset \dots \subset \hat{F}[\gamma_4]$ gösterilmiştir ve $\hat{f}_2^* \in \hat{F}[\gamma_2]$ olan model genelleme sınırını minimize ettiği görülmektedir ve bundan dolayı optimal model olarak kabul edilir⁷⁷.



Şekil 9. Yapısal risk minimizasyonu kullanarak $\hat{f}_2^*$ optimal model bulmak

Vapnik'in yapısal risk minimizasyon yaklaşımı, büyüyen VC boyutların ($h_1 < h_2 < \dots$) modellerinin bir iç içe dizilerine uyar ve daha sonra üst sınırın en küçük değerine sahip olan modeli seçer. Üst sınırlar genellikle gevşek olur, fakat bu bunların model seçimi için iyi bir kıstas olmalarına reddetmez, öyle ki burada test hatasının göreceli (mutlak değil) büyüklüğü önemlidir. Bu yaklaşımın temel sakıncası, bir fonksiyonlar sınıfının VC boyutunun hesaplanmasındaki zorluktur. Çoğu zaman VC boyut için sadece ham üst sınır elde edilebilir ve bu yeterli olmayabilir. Örnek olarak

⁷⁷ Hamel, a.g.e, pp.179-180.

yapısal risk minimizasyon programının başarılı bir şekilde yürütüldüğü yer destek vektör sınıflandırıcılarıdır⁷⁸ (Support Vector Classifier).

Yapısal risk minimizasyon ilkesi açık şekilde matematiksel olarak iyi kurulmuştur. Fakat bunun anlaşılması zor olabilir. Destek vektör makinesi, yapısal risk minimizasyon yönteminin hedefini tamamen otomatik bir optimizasyon işlemi sayesinde, VC-boyutta bir sınırı ve eğitme hatasını eş zamanlı minimize ederek gerçekleştirmektedir⁷⁹.

Yapısal risk minimizasyonunu şu şekilde özetleyebiliriz:

$$R[\hat{f}] \leq R_{emp}[\hat{f}] + v(l, h_{\hat{f}}, \eta), \quad (1.23)$$

denklemindeki VC güven, seçilen fonksiyonlar sınıfına bağlıdır. Buna karşın ampirik (deneysel) risk ve gerçek riskler ise eğitme prosedürü tarafından seçilen belir bir fonksiyona bağlıdır. Seçilen fonksiyonlar kümesinin altkümesi risk sınırı bu altküme için minimize olacak şekilde bulunulması arzulanmaktadır. VC boyut h , tamsayı olduğundan, düzgünce değişecek şekilde olayları düzenlemeyebiliriz. Bunun yerine, tüm fonksiyonlar sınıfını iç içe geçmiş altkümelere bölünmesiyle elde edilen bir “yapı ” geliştirildi (Şekil 10)



Şekil 10. VC boyut tarafından sıralandırılmış fonksiyonların iç içe geçmiş altkümeleri

⁷⁸ Hastie, Tibshirani, Friedman, **a.g.e.**, p.212

⁷⁹ Matthias Rychetsky, **Algorithms and Architectures for Machine Learning Based on Regularized Neural Networks and Support Vector Approaches**, Shaker Verlag, 2001, pp 13-19.

Her bir alt küme için, ya h hesaplanabilmelidir ya da h 'nin kendisinde bir sınır elde edilmelidir. Yapısal Risk Minimizasyon (YRM), gerçek riskte sınırı minimize eden fonksiyonların altkümesini bulmaktan ibarettir. Bu basit olarak makinelerin bir dizisini eğiterek yapılabilir, her altküme için bir tane, öyle ki verilen bir altküme için eğitmenin amacı ampirik riski minimize etmektir. Daha sonra dizideki ampirik riski ve VC güveni en az olan o eğitilen makine seçilir⁸⁰. Böylece destek vektör makineleri için gerekli zemin oluşturulmuş oldu.

⁸⁰ Christopher, a.g.e., p. 128.

İKİNCİ BÖLÜM

KARAR AĞAÇLARI

2.1 Karar Ağaçları (Decision Trees)

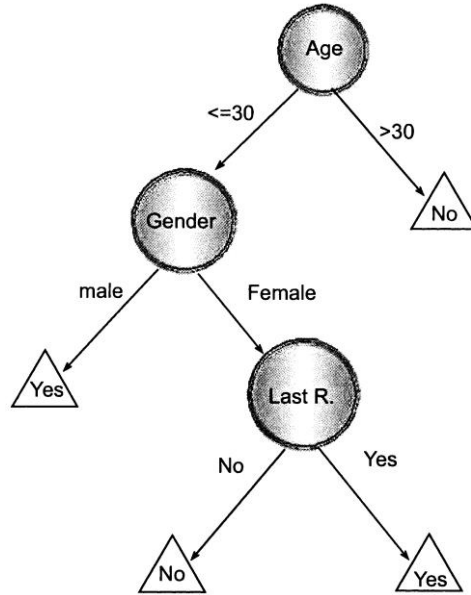
Sınıflandırma ve regresyonun kapsayıcı hedefi tahminde kullanılmak üzere bir modelin oluşturulmasıdır. Bir sınıflandırma veya regresyon probleminde, eğitime kayıtları (training records) bize verilmektedir. Her bir kayıt birçok öz nitelikten (attribute) oluşmaktadır. Tanım kümesi sayısal olan öz niteliklere sayısal nitelikler adı verilir, tanım kümesi sayısal olmayan niteliklere ise kategorik nitelikler (categorical attributes) olarak tanımlanır. Bağımlı (dependent) nitelik adı verilen bir tane seçkin nitelik mevcuttur. Kalan diğer nitelikler belirleyici (predictor) nitelik olarak adlandırılır; bunlar ya sayısal ya da kategorik olabilirler. Eğer bağımlı nitelik kategorik ise problem bir sınıflandırma problemi olarak belirtilir ve bağımlı nitelik *sınıf etiketi* (class label) olarak adlandırılır. Eğer bağımlı nitelik sayısal ise, probleme *regresyon problemi* denir¹.

Karar ağaçları sınıflandırıcıları temsil etmek için en popüler yaklaşımlardan bir tanesi olarak kabul edilir. İstatistik, makine öğrenimi, örüntü tanıma (pattern recognition), ve veri madenciliği bilim dallarında çalışan araştırmacılar mevcut veriden bir karar ağacı oluşturma konusunda çalışmaktadırlar. Bir karar ağacı, örnek uzayın yinelemeli bölümlendirmesi (recursive partitioning) olarak ifade edilen bir sınıflandırıcıdır. Karar ağacı köklü bir ağaç oluşturan düğümlerden oluşur, yani kendisine doğru gelen kenarlar olmayan ve “kök” olarak adlandırılan bir düğüme sahip yönlü bir ağaçtır. Diğer tüm düğümler tam olarak bir tane gelen kenara (incoming edge) sahiptirler. Giden kenarlara (outgoing edges) sahip düğümlere dâhili (internal) ya da test düğüm adı verilir. Diğer tüm düğümler yaprak (terminal ya da karar düğümleri olarak

¹ J. Gehrke, “Decision Trees”, **The Handbook of Data Mining**, Ed. by Nong Ye, New Jersey, 2003, Lawrence Erlbaum Associates Inc, pp. 3-4

bilinir) olarak adlandırılır. Bir karar ağacında, her bir dâhili düğüm, girdi özniteliklerin değerlerinin belli bir ayrık fonksiyonuna göre, örnek uzayını iki ya da daha fazla alt uzaylara böler. Basit ve en sık görülen durumda, her bir test tek bir özniteliği göz önünde bulundurur, öyle ki örnek uzayı özniteliğin değerine göre bölümlendirilir. Nümerik öznitelik durumlarında, durum bir aralık ifade eder².

Her bir yaprak (kanat) en uygun hedef değeri temsil eden bir sınıfa atanır. Alternatif olarak, yaprak, hedef özniteliğin olasılığı olacak şekilde belli bir değere sahip olasılık vektörüne sahip olabilir. Örnekler, ağacın kökünden yaprağa (dal, kanat) kadar bu örnekleri dolaşarak, yol (patika) boyunca testlerin sonucuna göre sınıflandırılır.



Şekil 11. Doğrudan postalamaya cevap³

Şekil 11’de potansiyel müşterilerin doğrudan postalamaya cevap verip vermeyeceğini çözmeğe çalışan bir karar ağacını göstermektedir. Dâhili (iç) düğümler halka ekinde gösterilir, bunun yanı sıra yapraklar (dal, kanat) üçgen şeklinde

² (Lior Rokach, Oded Maimon, “Decision Trees”, **Data Mining and Knowledge Discovery Handbook**, Ed. by Oded Maimon, Lior Rokach, 2005, Springer Science+Business Media Inc., p.165.

³ A.g.e, p.166.

gösterilmiştir. Bu ağaçta hem nominal hem de nümerik öznitelikler (değerler) mevcuttur. Herhangi bir analizci bu karar ağacını değerlendirerek, potansiyel bir müşterinin doğrudan postama ile ilgili cevabını tahmin edebilir ve tüm müşterilerin karakteristik davranışlarını anlayabilir. Her bir düğüm test ettiği özniteliği ile etiketlenir ve dalları buna karşılık gelen değerleri ile işaretlenir⁴.

Nümerik öznitelik durumunda, karar ağaçları hiperdüzlemlerin bir topluluğu gibi geometrik olarak yorumlanabilir; bu hiperdüzlemlerin her biri eksenlerden bir tanesine ortogonaldır. Karar vericiler, doğal olarak anlaşılması kolay olduğundan dolayı daha az karmaşık olan karar ağaçlarını tercih ederler. Breiman, L. v.d. (1984)'e göre karar ağacının karmaşıklığı bu ağacın doğruluğu üzerinde önemli bir etkisi vardır. Ağaç karmaşıklığı kullanılan durdurma kriterleri ve uygulanan budama metodu yolu ile dolaylı olarak kontrol edilir. Genellikle ağaç karmaşıklığı bazı metrikler tarafından ölçülür: toplam düğüm sayısı, toplam dal sayısı, ağaç derinliği ve kullanılan özniteliklerin (değerlerin) sayısı. Karar ağacı indüksiyonu ile kural indüksiyonu birbirleri ile çok yakından ilişkilidirler. Önceki parçayı oluşturmak için patika üzerindeki testleri birleştirerek ve dalın sınıf tahminini sınıf değeri olarak alarak, bir karar ağacının kökünden bunun bir dalına olan her bir patika bir kurala dönüştürülebilir. Örneğin, yukarıdaki şekil 11'e bakarak bir kural oluşturabiliriz: “eğer müşterinin yaşı otuzdan küçük veya eşit ve ya otuza eşit ise ve müşterinin cinsiyeti “erkek” ise, o zaman müşteri postaya cevap verecektir”.⁵ Elde edilen kural kümesi (rule set) daha sonra anlaşılabilirliğinin ve muhtemelen doğruluğunun geliştirilmesi için basitleştirilerek insanların kullanımına sunulabilir (Quinlan J., 1987).

Ağaç yapılı sınıflandırma ve regresyon yaklaşımları OLS regresyonu ve diskriminant analizi gibi normallik varsayımlarına ve kullanıcı tarafından belirtilen

⁴ Rokach, Maimon, **a.g.e.**, p.166.

⁵ Rokach, Maimon, **a.g.e.**, p.167.

model bildirimlerine dayanmayan sınıflandırma ve regresyon için alternatif yaklaşımlardır⁶.

2.2 Karar Ağaçlarının Algoritmik Çerçevesi

Regresyon ağaçları 1960 yılında Morgan ve Sonquist (1963) AID (Automatic Interaction Detection)⁷ geliştirmesiyle meydana gelmiştir. Daha sonra 1970’lerde sınıflandırma ağaçlarını oluşturmak için THAID (Morgan and Messenger, 1973) geliştirildi. Karmaşık bir program olan ve veriye ağaçları uydurmak için 1980’lerde istatistikçiler (Breiman, v.d., 1984) tarafından CART (Classification and Regression Trees) geliştirildi. Ağaç- tabanlı algoritma olan ve istatistikçiler tarafından geliştirilmiş QUEST (Quick Unbiased Efficient Statistical Tree) metodudur⁸.

Karar ağaçları indükleyicileri, otomatik olarak verilen bir veri kümesinden, bir karar ağacı oluşturan algoritmalarlardır. Amaç, genelleme hatasını minimize ederek optimal karar ağacını elde etmektir. Ancak, diğer hedef fonksiyonları tanımlanabilir, mesela, düğümlerin sayısı minimize edilebilir ya da ortalama derinlik minimize edilebilir. Verilen bir veriden elde edilen optimal karar ağacının indüksiyonu zor bir iş olarak nitelendirilir. Eğitim kümesi ile tutarlı olan bir minimal karar ağacı bulmanın NP-zor problem olduğu gösterilmiştir (Hancock, T.R., v.d., 1996). Ayrıca, bilinmeyen bir örneğin sınıflandırılması için gerekli olan, beklenen testlerin sayısına göre minimal ikili ağaç oluşturmak NP-Tam (NP- Complete) olduğu gösterildi(Hyafil L., 1976). Optimal karar ağacı algoritmasının kullanımı küçük problemlerde elverişlidir. Sonuç olarak, bu tür problemlerin çözümü için bulgusal (heuristics) metotlar gerekmektedir. Bu metotlar iki gruba ayrılabilir: yukarıdan – aşağı ve alttan-üste. Yukarıdan –aşağı karar ağaçları

⁶ C.D. Sutton, “Classification and Regression Trees, Bagging and Boosting”, **Handbook of Statistics 24: Data Mining and Data Visualization**, Ed. by Ed. by C.R. Rao, E.J.Wegman, J.L.Solka, Elsevier B.V., 2005, p.303.

⁷ J.N. Morgan, J. A. Sonquist, “Problems in the analysis of survey data, and a proposal”, **J.Amer. Statist. Assoc.**, vol.58, 1963, pp.415-434.

⁸ W.Y. Loh, X. Shih, “Split selection methods for classification trees”, **Statistica Sinica**, vol. 7, 1997, pp.815-840.

indükleyicilerinden bazıları ID3 (Quinlan,1986), C4.5 (Quinlan, 1993), CART (Breiman v.d., 1984). Bunlardan bazıları büyüme ve budama gibi iki kavramsal evrelerden oluşur (C4.5 ve CART gibi). Diğer indükleyiciler sadece büyüme aşamasına sahiptirler. Bu algoritmalar yukarıdan – aşağı olacak şekilde yinelemeli bir şekilde karar ağacını oluştururlar. Her bir yinelemede, algoritma girdi özniteliklerinin (değerlerinin) bir ayrık fonksiyonunun sonucunu kullanarak eğitim kümesinin bölmesini dikkate alır. En uygun fonksiyon seçimi, bazı bölme (splitting criteri) kistaslarına göre yapılır. Uygun bir bölme seçildikten sonra, her bir düğüm eğitim kümesini, bir durdurma kriteri sağlanana kadar, daha küçük altkümelere ayırır⁹.

2.3 Tek Değişkenli Bölme Kriterleri

Ayrık bölme fonksiyonları tek değişkenlidir. Tek değişkenlilikten kasıt, iç (dâhili) düğümün tek bir özneliliğin değerine göre bölünmesidir. Sonuç olarak, indükleyici, üzerinde bölme yapmak için en iyi niteliği arar. Birçok tek değişkenli kriterler mevcuttur. Bunlar farklı şekillerde ayırt edilebilirler:

- Ölçme kökenine göre: Bilgi kuramı, bağımlılık ve uzaklık.
- Ölçme yapısına göre: Katışkı (impurity) tabanlı ölçüt, normalleştirilmiş katışkı tabanlı ölçüt ve ikili ölçüt.

2.3.1 Katışkı tabanlı (impurity-based) Ölçüt¹⁰

X tesadüfî bir değişken k gibi ayrık değerler ile verilmiş olsun ve $P = (p_1, \dots, p_k)$ göre dağılsın, katışkı ölçütü $\phi: [0,1]^k \rightarrow R$ gibi aşağıdaki şartları sağlayan bir fonksiyondur,

- $\phi(P) \geq 0$
- $\phi(P)$ minimumdur, eğer $p_i = 1, \exists i$ için.

⁹ Rokach, Maimon, a.g.e., p.168.

¹⁰ Maimon, Rokach, a.g.e., p.153.

- $\emptyset(P)$ maksimumdur eğer $1 \leq i \leq k$, $p_i = 1/k$, $\forall i$ için.
- $\emptyset(P)$, P 'nin bileşenlerine göre simetriktir.
- $\emptyset(P)$, kendi aralığında düzgün bir fonksiyondur (her yerde türevlenebilir)

Eğer olasılık vektörü 1'in bir bileşenine sahip ise (x değişkeni sadece bir değer alır), değişken “saf (pure)” olarak tanımlanır. Eğer tüm bileşenler eşit ise, katışkı (impurity) seviyesi maksimuma ulaşır. S eğitime kümesi verilmiş olsun, hedef y niteliği şu şekilde tanımlanır:

$$P_y(S) = \left(\frac{|\sigma_{y=c_1} S|}{|S|}, \dots, \frac{|\sigma_{y=c_{|dom(y)|}} S|}{|S|} \right) \quad (2.1)$$

Ayrık nitelik a_i den dolayı bölme iyiliği (goodness of split), $v_{i,j} \in dom(a_i)$ değerlerine göre S 'nin bölümlenmesinden (partitioning) sonra hedef niteliğin katışıklığındaki azalması olarak tanımlanır:

$$\Delta\Phi(a_i, S) = \phi(P_y(S)) - \sum_{j=1}^{|dom(a_i)|} \frac{|\sigma_{a_i=v_{i,j}} S|}{|S|} \cdot \phi(P_y(\sigma_{a_i=v_{i,j}} S)) \quad (2.2)$$

.

Bazı yazarlar (Fayyad ve Irani, 1992) bu ölçütün kısıtlamalarını göstermişleridir. C-SEP (Class Separation) olarak adlandırılan ve bilgi kazancı (information gain) ve Gini indeksi gibi katışkı tabanlı ölçütlere göre daha iyi bir performans sergileyen nitelik seçim (attribute selection) ölçütlerini sunmuşlardır.¹¹

¹¹ U. Fayyad, K.B. Irani, “The attribute selection problem in decision tree generation”, **In Proceedings of Tenth National Conference on Artificial Intelligence**, Cambridge, MA, 1992, AAAI Press/MIT Press, pp. 104-110.

2.3.2 Bilgi Kazancı (Information Gain)

Bilgi kazancı (information gain), katışkı ölçüsü olarak entropi ölçüsünü kullanan bir katışkı tabanlı (impurity - based) ölçüttür¹²(Quinlan J., 1987).

$$\begin{aligned} \text{InformationGain}(a_i, S) = \\ \text{Entropy}(y, S) - \sum_{v_{i,j} \in \text{dom}(a_i)} \frac{|\sigma_{a_i=v_{i,j}} S|}{|S|} \cdot \text{Entropy}(y, \sigma_{a_i=v_{i,j}} S), \end{aligned} \quad (2.3)$$

öyle ki,

$$\text{Entropy}(y, S) = \sum_{c_j \in \text{dom}(y)} \frac{|\sigma_{y=c_j} S|}{|S|} \cdot \log_2 \frac{|\sigma_{y=c_j} S|}{|S|} \text{ dır.} \quad (2.4)$$

2.3.3 Gini İndeksi

Gini indeksi CART için¹³ (Breiman, v.d., 1984) tarafından geliştirilmiştir. Hedef niteliklerin değerlerinin olasılık dağılımları arasındaki ıraksamayı ölçen katışkı tabanlı (impurity - based) bir ölçüttür. Aşağıdaki gibi tanımlanır:

$$\text{Gini}(y, S) = 1 - \sum_{c_j \in \text{dom}(y)} \left(\frac{|\sigma_{y=c_j} S|}{|S|} \right)^2. \quad (2.5)$$

Sonuç olarak a_i niteliğini seçmek için değerlendirme ölçütü

$$\text{GiniGain}(a_i, S) = \text{Gini}(y, S) - \sum_{v_{i,j} \in \text{dom}(a_i)} \frac{|\sigma_{a_i=v_{i,j}} S|}{|S|} \cdot \text{Gini}(y, \sigma_{a_i=v_{i,j}} S), \quad (2.6)$$

¹² J.R. Quinlan, "Introduction of decison trees", **Machine Learning**, vol.1, 1986, pp.81-106.

¹³ L. Breiman, J. Friedman, R. Olshen, C. Stone, **Classification and Regression Trees**, Wadsworth Int. Group, 1984, pp.103-104

olarak tanımlanır.

2.3.4 Olabilirlik-Oran Ki-Kare İstatistikleri¹⁴

Olabilirlik – oran

$$G^2(a_i, S) = 2 \cdot \ln(2) \cdot |S| \cdot \text{InformationGain}(a_i, S), \quad (2.7)$$

olarak tanımlanır. Bilgi kazancı (information gain) ölçütünün istatistik önemini hesaplamak için bu oran çok faydalıdır. Sıfır hipotezi (H_0) girdi niteliği ve hedef niteliği şartlı olarak bağımsız olduğunu belirtir. Eğer H_0 hipotezi kabul edilir ise, test istatistiği serbestlik derecesi $(\text{dom}(a_i) - 1) \cdot (\text{dom}(y) - 1)$ eşit olacak şekilde χ^2 olarak dağılır.

2.3.5 DKM Ölçütü^{15,16}

DKM ölçütü ikili sınıf nitelikleri için katışkı tabanlı (impurity - based) bölme ölçütüdür. Katışkı tabanlı fonksiyon aşağıdaki şekilde tanımlanır:

$$DKM(y, S) = 2 \cdot \sqrt{\left(\frac{|\sigma_{y=c_1} S|}{|S|}\right) \cdot \left(\frac{|\sigma_{y=c_2} S|}{|S|}\right)} \quad (2.8)$$

Bu ölçüt, belli bir hata elde etmek için diğer katışkı tabanlı ölçütlere (bilgi kazancı ve gini indeksi) kıyasla daha küçük ağaçlar gerektirdiği teorik olarak ispatlandı (Kearns, M., Mansour Y., 1999)

¹⁴ F. Attneave, **Applications of Information Theory to Psychology**, Holt, Rinehart and Winston, 1959, pp 95-99.

¹⁵ T.G. Dietterich, M. Kearns, Y. Mansour, “Applying the weak learning framework to understand and improve C4.5”, **Proceedings of the Thirteenth International Conference on Machine Learning**, San Francisco, 1996, Morgan Kaufmann, pp. 96-104.

¹⁶ M. Kearns, Y. Mansour, “On the boosting ability of top-down decision tree learning algorithms”, **Journal of Computer and Systems Sciences**, vol. 58(1), 1999, pp. 109-128.

2.3.5 Normalleştirilmiş Katıřkı Tabanlı Ölçüt¹⁷

Yukarıda bahsedilen katıřkı-tabanlı kıstas çok büyük tanım kümesi değeri olan niteliklere karşı yanlıdır (biased). Bu ölçüt çok değeri girdi niteliklerini az değeri niteliklere tercih etmektedir (Quinlan J. , 1986). Örneğin, sosyal güvenlik numarasını gösteren bir girdi niteliğı muhtemelen en yüksek bilgi kazancını elde edecektir, fakat bu niteliğı (değişkeni) bir karar ağacına eklediğimiz zaman zayıf genelleştirilmiş doğruluk sonucu verecektir. Bu nedenler, aşağıda anlatılacak olan bölümlerde, katıřkı tabanlı ölçütleri “normalleştirmek” faydalıdır.

2.3.6 Kazanç Oranı (Gain Ratio)

Kazanç oranı bilgi kazancını aşağıdaki şekilde “normalleştirir”(Quinlan J. , 1993) .

$$GainRatio(a_i, S) = \frac{InformationGain(a_i, S)}{Entropy(a_i, S)}. \quad (2.9)$$

Bilindiğı üzere, bu oran payda sıfır olduğunda tanımsızdır. Ayrıca bu oran, payda çok küçük olduğu zaman nitelikler lehine bir eğilim gösterir. Sonuç olarak, iki aşama önerilir. Öncelikle bilgi kazancı (information gain) tüm nitelikler için hesaplanır. Sadece, en az ortalama bilgi kazancı kadar performans gösteren nitelikler dikkate alınır ve en iyi kazanç oranını elde eden nitelik seçilir. Kazanç oranı hem doğruluk açısından hem de sınıflandırıcı karmaşıklığı açısından basit bilgi kazancı ölçütüne göre daha iyi performans gösterdiği gösterilmiştir¹⁸.

¹⁷ J.R. Quinlan, “Introduction of decision trees”, **Machine Learning**, vol.1, 1986, pp.81-106.

¹⁸ J.R. Quinlan, “Decision Trees and Multivalued Attributes”. In J. R. (ed.), **Machine Intelligence**, v.11, Ed. by J. Richards, Oxford, England, 1988, Oxford Univ Press, pp.305-318.

2.3.6 Uzaklık Ölçümü (Distance measure)¹⁹

Uzaklık ölçümü kazanç oranında olduğu gibi katıksı ölçütünü normalleştirir. Fakat farklı şekilde normalleştirme işlemi yapar(Lopez de Mantras, 1991).

$$\frac{\Delta\Phi(a_i, S)}{-\sum_{v_{i,j} \in \text{dom}(a_i)} \sum_{c_k \in \text{dom}(y)} \frac{|\sigma_{a_i=v_{i,j} \text{ AND } y=c_k} S|}{|S|} \cdot \log_2 \frac{|\sigma_{a_i=v_{i,j} \text{ AND } y=c_k} S|}{|S|}} \quad (2.10)$$

2.3.7 İkili Ölçütleri

İkili kriterler ikili karar ağaçları oluşturmak için kullanılır. Bu ölçümler girdi nitelik tanım bölgesinin iki alt tanım bölgelerine bölünmesine dayalıdır. Kabul edelim ki S örneği üzerinde a_i niteliği için $\beta(a_i, \text{dom}_1(a_i), \text{dom}_2(a_i), S)$ ikili ölçüt değerinin göstereceği, $\text{dom}_1(a_i)$ ve $\text{dom}_2(a_i)$ a_i nin ilgili alt tanım kümesidir (alt domain). Nitelik tanım kümesinin iki ayrışık ve kapsamlı alt – tanım kümelerine optimal bölünmesi için elde edilen değer niteliklerin karşılaştırılması için kullanılır²⁰.

2.3.8 Twoing Ölçütü

Hedef niteliğin tanım kümesi oldukça geniş olduğu zaman gini indeksi problemler ile karşılaşabilir (Breiman v.d. 1984).bu durumda “twoing” ölçütü olarak adlandırılan ikili ölçütü uygulamak mümkündür. Bu ölçüt aşağıdaki şekilde tanımlanır:

¹⁹ R. Lopez de Mantras, “A distance -based attribute selection measure for decision tree induction”, **Machine Learning**, vol.6, 1991, pp.81-92.

²⁰ Maimon, Rokach, **a.g.e.**, p.155.

$$\begin{aligned}
& twoing(a_i, dom_1(a_i), dom_2(a_i), S) = \\
& 0.25 \cdot \frac{|\sigma_{a_i \in dom_1(a_i)} S|}{|S|} \cdot \frac{|\sigma_{a_i \in dom_2(a_i)} S|}{|S|} \cdot \left(\sum_{c_i \in dom(y)} \left| \frac{|\sigma_{a_i \in dom_1(a_i) AND y=c_i} S|}{|\sigma_{a_i \in dom_1(a_i)} S|} - \right. \right. \\
& \left. \left. \frac{|\sigma_{a_i \in dom_2(a_i) AND y=c_i} S|}{|\sigma_{a_i \in dom_2(a_i)} S|} \right| \right)^2 \tag{2.11}
\end{aligned}$$

Hedef nitelik (bağımlı değişken) ikili olduğu zaman, gini ve twoing kriterleri eşdeğerdirler. Çok-sınıflı problemler için, twoing ölçütü eşit şekilde ayrılmış bölünmeleri (splits) tercih eder²¹. “Twoing” ölçütü durumları en iyi iki alt sınıfa böler ve bölme noktasını (split point) tanımlayan her bir alt sınıfın varyansını hesaplar.

2.3.9 Ortogonal (ORT) Ölçütü²²

ORT kriteri Fayyad ve Irani (1992) tarafından sunuldu. Bu ikili ölçüt aşağıdaki şekilde tanımlanır:

$$ORT(a_i, dom_1(a_i), dom_2(a_i), S) = 1 - \cos\theta(P_{y,1}, P_{y,2}), \tag{2.12}$$

burada $\theta(P_{y,1}, P_{y,2})$ iki vektör olan $p_{y,1}$ ve $p_{y,2}$ arasındaki açıdır. Bu vektörler, sırası ile $\sigma_{a_i \in dom_1(a_i)} S$ ve $\sigma_{a_i \in dom_2(a_i)} S$ bölümlerindeki (partition) hedef niteliğinin olasılık dağılımını temsil ederler. Bu ölçüt, belirli problemler takımı için bilgi kazancı (information gain) ve Gini indeksine göre daha iyi performans sergilediği gösterilmiştir.

²¹ Nisbet, Elder, Miner, **a.g.e.**, p242.

²² U. Fayyad, K.B. Irani, “The attribute selection problem in decision tree generation”, **In Proceedings of Tenth National Conference on Artificial Intelligence**, Cambridge, MA, 1992, AAAI Press/MIT Press, pp. 104-110.

2.3.10 Kolmogorov – Smirnov Ölçütü^{23,24}

Kolmogorov- Smirnov uzaklığını kullanan bir ikili ölçüttür ve (Friedman, 1977) ve (Rounds, 1980) da önerilmiştir. İkili bir hedef niteliği $dom(y) = \{c_1, c_2\}$ olarak farz edersek, ölçüt aşağıdaki gibi tanımlanır:

$$KS(a_i, dom_1(a_i), dom_2(a_i), S) = \left| \frac{|\sigma_{a_i \in dom_1(a_i) \text{ AND } y=c_1} S|}{|\sigma_{y=c_1} S|} - \frac{|\sigma_{a_i \in dom_1(a_i) \text{ AND } y=c_2} S|}{|\sigma_{y=c_2} S|} \right| \quad (2.13)$$

Bu ölçüt (Utgoff, P.E., ve Cluse, J.A.: , 1996) da çoklu sınıflı ve eksik veri değerlerine sahip hedef niteliklerini halletmek için genişletilmiştir. Bunların sonuçları, önerilen metodun kazanç oranından (gain ration) daha iyi bir performans sergilediğini göstermiştir.

2.3.11 AUC-Bölme Ölçütü²⁵

AUC metrik kullanım fikri son zamanlarda (Ferri C. v.d., 2002) ortaya atıldı. ROC eğrisinin dışbükey zarfının altında kalan en büyük alanı elde eden nitelik (değişken) seçilir. AUC tabanlı bölme ölçütü hem sınıflandırma doğruluğu açısından hem de ROC eğrisinin altında kalan alan açısından diğer bölme kriterlerine göre daha iyi bir performans sergilediği gösterilmiştir ²⁶. Diğer katıksı ölçütünden farklı olarak, bu ölçüt üst (ebeveyn) düğümün katıksısı ile bölünmeden sonraki alt (çocuk) düğümün tartılı (weighted) katıksısı arasında bir kıyaslama gerçekleştirmez.

²³ J. Friedman, “A recursive partitioning decision rule for nonparametric classifiers”, **IEEE Trans. on Comp.**, C26, 1977, pp.404-408.

²⁴ E.Rounds, “A combined non-parametric approach to feature selection and binary decision tree design”, **Pattern Recognition**, vol.12, 1980, pp.313-317.

²⁵ C. Ferri, P. Flach, J. Hernandez – Orallo, “Learning Decision Trees Using the Area Under the ROC Curve”, **Proceedings of the 19th International Conference on Machine Learning**, 2002, Morgan Kaufman, pp. 139-146.

²⁶ Rokach, Maimon, **a.g.e.**, p.173.

2.4 Çok Değişkenli Bölme Ölçütleri

Birçok öznelikler (değişkenler) tek bir düğüm bölme testinde iştirak edebilirler. Açıkçası, en iyi çok değişkenli ölçüt bulmak en iyi tek değişkenli ölçütü bulmaya nazaran çok karmaşıktır. Ayrıca, bu tür ölçütler ağacın performansını önemli oranda artırabilmesine rağmen, bu ölçütler tek değişkenli ölçütlere kıyasla çok daha az popülerdirler. Birçok çok değişkenli bölme kısıtları girdi niteliklerin lineer bileşimlerine dayanmaktadır. “açgözlü arama (greedy search)” kullanarak en iyi lineer bileşen bulunabilir (Breiman, L. v.d., 1984, Murthy, 1998), bu metot sadece birçok metottan bir tanesidir. Bunun yanı sıra, lineer programlama, lineer diskriminant analizi ve diğerleri mevcuttur.

2.5 Durdurma Ölçütleri

Karar ağaçlarının büyüme safhası bir durdurma ölçütü başlatılana kadar devam eder. Aşağıdaki şartlar ortak durdurma kurallarıdır:

- Eğitim kümesindeki tüm örnekler y 'nin tek bir değerine aittir
- Maksimum ağaç derinliğine ulaşılır
- Terminal düğümdeki (uç düğüm) durumların sayısı üst düğüm (ebeveyn düğüm) için minimum durumların sayısından daha azdır
- Eğer düğüm bölünürse, bir veya daha fazla çocuk(alt) düğümdeki durumların sayısı, çocuk düğümler için minimum durumların sayısından daha az olacaktır
- En iyi bölme kısıtı (splitting criteria) belli bir eşik değerden çok büyük değildir. ²⁷

²⁷ Rokach, Maimon, a.g.e., p.174.

2.6 Budama Metotları

Sınıflandırma ağacının oluşumunda budama (pruning) aşaması aşırı uyumu (overfitting) problemini engellemek için ve yanlış sınıflandırma hatasını minimize etmek için ağacın gerçek büyüklüğüne karar verir. Aşağıdan – yukarı (bottom up) budamada, ağaç büyütme safhasında ağacın her bir yaprak düğümü, aile büyüklüğü olan n değerinin kullanıcı tarafından belirlenen c gibi bir eşik değerin altın düşene kadar büyümeye devam eder²⁸. Maliyet – karmaşıklığı (cost-complexity) budama, test kümesi adı verilen ek kayıtlar kümesi ile budama (Breiman v.d., 1984) ve minimum tanım uzunluğu²⁹ (MDL, Minimum Description Length) gibi yöntemlere dayalı budama çeşitleri aşağıdan- yukarıya budama için örnek gösterilebilir (Mehta, Raissanen, & Agrawal, 1995). Ağacın büyümesi esnasında, yukarıdan aşağı (top - down) budamada, her bir n düğümünde bir S_n istatistiği hesaplanır ve S_n değerine göre, n düğümünde ağaç büyümesi ya devam eder ya da durdurulur (Quinlan, 1993). Aşağıdan – yukarıya doğru budama, yüksek kaliteli ağaçlarda genellikle sonuç verir (Breiman, v.d., 1984; Lim, v.d., 2000; Quinlan, 1993), fakat aşağıdan – yukarıya olan budama, ağacın hiçbir parçası önce oluşturulup sonradan atılmadığından dolayı, işlemsel olarak daha verimlidir.

Sıkı bir durdurma kistası uygulayarak küçük ve yetersiz uyumlu (under-fitted) karar ağacı oluşturulur. Diğer taraftan, gevşek durdurma ölçütü kullanıldığı zaman ise, eğitme kümesine aşırı uyumlu (over-fitted) çok büyük bir karar ağacı oluşturulur. Budama metodu (Breiman v.d.,1984) tarafından önerilmiştir, ve yukarıda belirtilen ikilemi çözmek için geliştirilmiştir. Bu yöneme göre, gevşek bir durdurma ölçütü kullanılırsa karar ağacının eğitme kümesine aşırı uyumlu (overfit) olması sağlanır. Daha sonra aşırı uyumlu olan (overfit) bu ağaç, genelleme doğruluğuna katkısı olmayan alt

²⁸ J. Gehrke, **a.g.e.**, p.46.

²⁹ J.R. Quinlan, R.L. Rivest, “Inferring decision trees using minimum description length principle”, **Information and Computations**, vol.80, 1989, pp.227-248.

dallar çıkarılarak daha küçük bir ağaç olacak şekilde kesilir. Birçok çalışmada budama metotlarının uygulanması bir karar ağacının genelleme performansını iyileştirdiği gösterilmiştir³⁰. Budama metodunu diğer önemli tarafı “basitlik için işlem doğruluğudur” olarak da sunuldu³¹. Eğer hedef, yeteri kadar doğru ve özlü kavram açıklaması üretmek olduğu zaman, budama bu aşamada çok faydalıdır. Bu süreçte, başlangıç karar ağacı tamamen doğru bir ağaç olarak görülür. Böylece budanan bir karar ağacının doğruluğu başlangıç karar ağacına ne kadar yakın olduğunu göstermiş olur. Farklı budama teknikleri mevcuttur. Bunların birçoğu düğümlerin yukarıdan aşağıya ya da aşağıdan yukarıya düğümlerin gezinmesi olarak gerçekleştirir. Eğer bu operasyon belli bir kriteri iyileştirirse, bir düğüm budanır.

2.6.1 Maliyet-Karmaşıklık Budama (Cost – Complexity Pruning)

İki aşamada gerçekleşir, “en zayıf link budama” ya da “hata karmaşıklığı budama” olarak bilinir (Breiman v.d., 1984). Birinci safhada, $T_0, T_1, \dots, T_k$ gibi bir ağaç dizisi eğitime verisinde oluşturulur. Burada T_0 , budama öncesi orijinal ağaçtır ve T_k ise kök ağaçtır. İkinci aşamada, bu ağaçlardan bir tanesi genelleme hatasının tahminine dayanarak budanmış ağaç olarak seçilir. T_{i+1} ağacı, uygun yapraklı T_i öncel ağacındaki bir veya daha fazla alt-ağaçlar ile yer değiştirilmesi ile elde edilir. Budanmış alt ağaçlar, her budanmış yaprak için bariz hata oranındaki en küçük artışı elde eden ağaçlardır:

$$\alpha = \frac{\varepsilon(\text{pruned}(T,t),S) - \varepsilon(T,S)}{|\text{leaves}(T)| - |\text{leaves}(\text{pruned}(T,t))|}, \quad (2.14)$$

$\varepsilon(T,S)$, S örneği üzerinde T ağacının hata oranını gösterir ve $|\text{leaves}(T)|$ ise T 'deki yaprakların sayısını belirtir. Bunun yanı sıra $\text{pruned}(T,t)$ ise T 'deki t düğümünün uygun bir yaprak ile yer değiştirilmesi ile elde edilen ağacı belirtir. İkinci aşamada her bir

³⁰ Rokach, Maimon, **a.g.e.**, p.175

³¹ I. Bratko, M. Bohanec, “Trading accuracy for simplicity in decision trees”, **Machine Learning**, vol.15, 1994, pp. 223-250.

budanmış ağacın ($T_0, T_1, \dots, T_k$) genelleme hatası tahmin edilir. Daha sonra en iyi budanmış ağaç seçilmiş olur. Eğer verilen veri kümesi yeterince büyük ise, bunu eğitime kümesi ve budama kümesi olarak ikiye bölünmesi önerilmektedir. Ağaçlar eğitime kümesi kullanılarak oluşturulur ve budama kümesinde ise değerlendirilirler. Diğer yandan, eğer veri kümesi yeterince büyük değil ise, hesaplama karmaşıklığının etkilerine rağmen, çapraz-doğrulama (cross validation) yönteminin kullanılması önerilmektedir³².

2.6.2 İndirgenmiş Hata Budaması (Reduced Error Pruning)

Bu metot, karar ağaçlarını budamak için (Quinlan J. , 1987) tarafından geliştirilmiş basit bir yöntemdir. Altan üste doğru iç düğümlerde gezinirken, yöntem her bir iç düğümü en sık görülen ve ağacın doğruluğunu azaltmayan sınıf ile bu düğümlerin yer değiştirilip değiştirilmemesi için kontrol eder. Bu durumda, düğüm budanır. Bu yöntem herhangi bir ileri budamanın doğruluğu azaltması durumuna kadar devam eder. Doğruluğu tahmin etmek için, Quinlan (1987) bir budama kümesinin kullanılmasını önerir. Bu yöntem en küçük doğru verilen bir budama kümesi ile ilgili olan alt-ağaç ile sonlanır³³.

2.6.3 En düşük Hata Budaması (Minimum Error Pruning, MEP)

Bu yöntem (Olaru, C., Wehenkel L., 2003) tarafından önerildi. İç düğümlerin aşağıdan – yukarıya gezinmesi olarak çalışmaktadır. Her bir düğümde budamalı ve budamasız *1-olasılık hata oranı tahminin* kıyaslar. Burada *1-olasılık hata oranı tahminin* frekansları kullanarak basit olasılık tahmini için bir düzeltmedir. Kabul edelim ki S_t , bir t yaprağına ulaşmış olan örnekleri göstermiş olsun, daha sonra bu yaprakta beklenen hata oranı:

³² Rokach, Maimon, **a.g.e.**, pp.175-176

³³ Rokach, Maimon, **a.g.e.**, p.176

$$\varepsilon'(t) = 1 - \max_{c_i \in \text{dom}(y)} \frac{|\sigma_{y=c_i} S_t| + l \cdot p_{apr}(y=c_i)}{|S_t| + l}, \quad (2.15)$$

$p_{apr}(y = c_i)$, c_i değerini alan y 'nin a-priori (önsel) olasılığıdır ve l a-priori olasılığına verilen ağırlığı göstermektedir. Bir iç düğümün hata oranı, dallarının hata oranının ağırlıklı ortalaması şeklindedir. Ağırlık, her bir dal boyunca örneklerin oranına göre belirlenir. Hesaplama yinelemeli olarak yapraklara kadar gerçekleştirilir. Eğer bir iç düğüm budanmış ise, bu o zaman bir yaprak olur ve bunun hata oranı son denklemi kullanarak hesaplanır. Sonuç olarak, belli bir iç düğümün budanmasından önce ve budanmasından sonra hata oranını kıyaslayabiliriz. Eğer bu düğümün budanması hata oranının artırmıyor ise, budama kabul edilmelidir.

2.6.4 Kötümser Budama (Pessimistic Pruning)

Budama kümesi ya da çapraz doğrulama ihtiyacından kaçınır ve bunun yerine kötümser istatistik korelasyon testi (pessimistic statistical correlation test) (Quinlan J. , 1993). Temel fikir, eğitime kümesi kullanılarak tahmin edilen hata oranının yeterince güvenli olmadığıdır. Bunun yerine, daha gerçekçi bir ölçüt olan, binom dağılımı için süreklilik düzeltmesi (continuity correction for binomial distribution) kullanılmalıdır:

$$\varepsilon'(T, S) = \varepsilon(T, S) + \frac{|\text{leaves}(T)|}{2 \cdot |S|}. \quad (2.16)$$

Fakat bu düzeltme hala bir iyimser hata oranı üretmektedir. Sonuç olarak, iç düğüm olan t 'nin budanması dikkate alınabilir eğer bunun hata oranı referans ağacından bir standart hata içerisinde ise, yani (Quinlan J. , 1993) :

$$\varepsilon'(\text{pruned}(T, t), S) \leq \varepsilon'(T, S) + \sqrt{\frac{\varepsilon'(T, S) \cdot (1 - \varepsilon'(T, S))}{|S|}} \quad (2.17)$$

Son durum, oran için güven aralığına dayanır. Genellikle son durum kullanılır öyle ki T , kökü iç düğüm t olan bir alt-ağacı ifade eder ve S ise t düğümüne atıfta bulunan eğitme kümesinin kısmını belirtir. Kötümser budama yöntemi yukarıdan aşağıya iç düğümlerde gezinmeyi gerçekleştirir. Eğer bir iç düğüm budanır ise, o zaman bunun tüm torun düğümleri (astsalları) budama işleminden çıkarılır, bunun sonucunda oldukça hızlı bir budama işlemi yapılmış olur³⁴.

2.6.5 Hata – Tabanlı Budama (Error-Based Pruning, EBP)

Hata–tabanlı budama kötümser budamanın bir gelişimidir. Çok iyi bilinen C4.5 algoritmasında uygulanır. Kötümser budamada olduğu gibi, oranlar için istatistik güven aralığının üst sınırını kullanarak hata oranı tahmin edilir.

$$\varepsilon_{UB}(T, S) = \varepsilon(T, S) + Z_{\alpha} \cdot \sqrt{\frac{\varepsilon(T, S) \cdot (1 - \varepsilon(T, S))}{|S|}}, \quad (2.18)$$

$\varepsilon(T, S)$ eğitme kümesi olan S 'de, T ağacının yanlış sınıflandırma oranının belirtir. Z ise Standard normal kümülatif dağılımın tersidir, ve α istenilen anlam düzeyidir. $subtree(T, t)$, t düğüm tarafından kökleşmiş alt-ağacı ifade etsin. Ayrıca $maxchild(T, t)$, t 'nin en sık görülen çocuk (alt) düğümü olsun, yani S 'deki örneklerin çoğu bu belli çocuk(alt) düğüme ulaşır. Ayrıca S_t , t düğümüne ulaşan S 'deki tüm örnekler olsun. Yöntem alttan-üste tüm düğümleri gezerek, aşağıda verilen değerleri kıyaslar:

1. $\varepsilon_{UB}(subtree(T, t), S_t)$
2. $\varepsilon_{UB}(pruned(subtree(T, t), t), S_t)$
3. $\varepsilon_{UB}(subtree(T, maxchild(T, t)), S_{maxchild(T, t)})$

³⁴ Rokach, Maimon, **a.g.e.**, p.177.

En küçük değere göre, bu yöntem ya ağacı olduğu gibi bırakır ya da t düğümünü budar, ya da $maxchild(T,t)$ tarafından kökleşen alt-ağaç ile yer değiştirir.

2.6.7 Optimal Budama³⁵

Optimal budamayı bulma konusu (Bratko, I., Bohanec M., 1994)' de çalışıldı. İlk yapılan araştırma optimaliteyi garanti eder ve OPT olarak adlandırılır. Dinamik programlamaya dayalı olan algoritma, $\Theta(|leaves(T)|^2)$ 'nin karmaşıklığı ile optimal budamayı bulur, burada T başlangıç karar ağacıdır. OPT-2 olarak adlandırılan ve aynı şekilde dinamik programlama kullanarak optimal budama gerçekleştiren ikinci araştırma OPT'nin bir iyileştirilmiş halini sunmaktadır. Fakat OPT-2nin zaman ve uzay karmaşıklığı $\Theta(|leaves(T^*)|. |internal(T)|)$ 'dir. Burada T^* , hedef budanmış karar ağacıdır ve T ise başlangıç karar ağacıdır. Budanmış ağaç başlangıç ağaçtan çok daha küçük olduğundan ve iç düğümlerin sayısı yaprakların sayısından küçük olduğundan, hesaplama karmaşıklığı açısından OPT-2 genellikle OPT den daha verimlidir.

2.6.8 Minimum Tanım Uzunluğu Budama (Minimum Description Length, MDL)

Bir düğümün genelleştirilmiş doğruluğunu değerlendirmek için kullanılabilir. Bu metot bir karar ağacının boyutunu ağacı kodlamak için gerekli olan bitlerin sayısı yoluyla ölçer. MDL metodu çok daha az bitlerin yolu ile kodlanabilen ağaçları tercih eder. Bir t yaprağındaki bir bölünmenin maliyeti aşağıdaki gibi tahmin edilebilir:

$$Cost(t) = \sum_{c_i \in dom(y)} |\sigma_{y=c_i} S_t| \cdot \ln \frac{|S_t|}{|\sigma_{y=c_i} S_t|} + \frac{dom(y)-1}{2} \ln \frac{|S_t|}{2} + \ln \frac{\pi^{\frac{|dom(y)|}{2}}}{\Gamma(\frac{|dom(y)|}{2})}, \quad (2.19)$$

³⁵ Bratko, Bohanec, **a.g.e.**, pp. 223-250.

burada S_t , t düğümüne ulaşan örnekleri belirtir. Bir iç düğümün bölme maliyeti bunun alt (çocuk) maliyet toplamına dayalı olarak hesaplanır.

2.7 Budama Metotlarının Kıyaslanması

Birçok çalışmalar farklı budama tekniklerinin performanslarının kıyaslanmasını amaçlamışlardır³⁶. Sonuçlar bazı metotlar, örneğin maliyet-karmaşıklık budama, indirgenmiş hata budaması gibi, aşırı budama (over-pruning) eğiliminde olduğunu göstermiştir. Yani çok küçük fakat çok az doğru karar ağacı oluştururlar. Diğer metotlar, mesela hata-tabanlı budama, karamsar hata budaması ve minimum hata budaması gibi, küçük budamaya yönelik yanlıdır (biased). Birçok kıyaslamalar sonucunda başka budama metotlarına karşı, her hangi bir durumda, daha iyi performans gösteren bir budama metodunun olmadığı görülmüştür³⁷

2.8 Eksik Değerler (Missing Values)

Gerçek dünya veri kümelerinde, eksik değerler ortak bir deneyimdir. Bu durum eğitime kümesinde bazı değerler eksik olduğundan hem indüklemeyi hem de yeni bazı değerleri eksik değerlere sahip yeni bir örnekten dolayı sınıflandırmayı karmaşık halde getirebilir. Bu problem birçok araştırmacı tarafından ele alındı (Friedman, 1977; Breiman v.d., 1984; Quinlan, 1989). Eğitime kümesinde eksik değerler aşağıdaki gibi elen alınabilir: $\sigma_{a_i=?}S$, a_i değerleri eksik olan S deki örneklerin altkümesini ifade etsin. Nitelik a_i kullanarak bölme kıstasını hesaplarken, nitelik a_i deki değerleri bilinmeyen tüm örnekler basit bir şekilde görmezden gelinir, yani, bölme kıstası olarak $\Delta\Phi(a_i, S)$ kullanmak yerine, $\Delta\Phi(a_i, S - \sigma_{a_i=?}S)$ kullanır.

³⁶ J.R. Quinlan, "Simplifying decision trees", **International Journal of Machine Studies**, vol.27, 1987, pp.221-234.

³⁷ Rokach, Maimon, a.g.e., p.179.

Diğer taraftan, eksik değerlerin mevcut olduğu durumlarda, bölme kistası orantılı bir şekilde azaltılmalıdır (Quinlan, 1989). Yani, $\Delta\Phi(a_i, S)$ bölme ölçütünü kullanmak yerine, aşağıdaki düzeltmeyi kullanır:

$$\frac{|S - \sigma_{a_i=?}S|}{|S|} \Delta\Phi(a_i, S - \sigma_{a_i=?}S) \quad (2.20)$$

Kazanç oranında olduğu gibi, ölçüt değerin normalize olmuş durumunda, payda eksik değerlerin öznitelik alanında bir ek değer ifade eder gibi hesaplanmalıdır. Örneğin, eksik değerli Kazanç Oranı (Gain Ratio) aşağıdaki gibi hesaplanmalıdır:

$$GainRatio(a_i, S) = \frac{\frac{|S - \sigma_{a_i=?}S|}{|S|} InformationGain(a_i, S - \sigma_{a_i=?}S)}{-\frac{|S - \sigma_{a_i=?}S|}{|S|} \log\left(\frac{|S - \sigma_{a_i=?}S|}{|S|}\right) - \sum_{v_{i,j} \in dom(a_i)} \frac{|\sigma_{a_i=v_{i,j}}S|}{|S|} \log\left(\frac{|\sigma_{a_i=v_{i,j}}S|}{|S|}\right)}. \quad (2.21)$$

Bir düğüm bölündükten hemen sonra, her bir kenar çıkıntılardan(outgoing edges) bir tanesine, karşılık gelen $|\sigma_{a_i=v_{i,j}}S| / |S - \sigma_{a_i=?}S|$ ağırlıkla, $\sigma_{a_i=?}S$ eklenmesi gerekmektedir.

Eksik değerli yeni bir örneğin sınıflandırılması için aynı fikir kullanılır. Eğer bir örnek, bu düğümün bölme ölçütü bir eksik değerden dolayı değerlendirilebildiği bir düğüm ile karşılaşır, bu tüm giden uçlara tamamen geçirilir. Tahmin edilen sınıf, bu örneğin son bulduğu tüm yaprak düğümlerin ağırlıklı birleşiminde en yüksek olasılığa sahip sınıf olacaktır.

Diğer bir yaklaşım ise “vekil bölme (surrogate split)” olarak bilinen ve (Breiman, v.d., 1984) tarafından sunulan bir metottur ve bu CART algoritmasında uygulandı. Burada amaç ağaçtaki her bir bölme için farklı girdi özniteliği kullanan ve orijinal bölmeye en çok benzeyen bir vekil bölme (surrogate split) bulmaktır. Eğer

orijinal bölmede kullanılan girdi özniteliğin değeri eksik ise, o zaman vekil bölmeyi kullanma mümkündür. S örneği üzerindeki iki ikili bölmelerin arasındaki benzerlik aşağıdaki gibi tanımlanır:

$$res(a_i, dom_1(a_i), dom_2(a_i), a_j, dom_1(a_j), dom_2(a_j), S) = \frac{\left| \sigma_{a_i \in dom_1(a_i) \text{ AND } a_j \in dom_1(a_j)}^S \right|}{|S|} + \frac{\left| \sigma_{a_i \in dom_2(a_i) \text{ AND } a_j \in dom_2(a_j)}^S \right|}{|S|}, \quad (2.22)$$

burada ilk bölme öznitelik a_i 'yi ifade eder ve bu $dom(a_i)$ 'yi $dom_1(a_i)$ ve $dom_2(a_i)$ olarak böler. Alternatif bölme, öznitelik a_j 'yi ifade eder ve tanım bölgesini $dom_1(a_j)$ ve $dom_2(a_j)$ olarak böler.

Eksik değer diğer örneklerle dayalı olarak tahmin edilebilir³⁸. Öğrenme safhasında, eğer q değişkenler grubundaki (demet, veri grupları) nominal bir öznitelik olan a_i 'nin değeri eksik ise, o zaman bu aynı hedef öznitelik değerine sahip tüm örnekler üzerindeki bunun modu ile tahmin edilir. Yani,

$$estimate(a_i, y_q, S) = \operatorname{argmax}_{v_{i,j} \in dom(a_i)} \left| \sigma_{a_i = v_{i,j} \text{ AND } y = y_q}^S \right|, \quad (2.23)$$

Burada y_q , q değişkenler grubundaki (demet) hedef özniteliğin değerini belirtir. Eğer eksik öznitelik olan a_i sayısal ise, a_i 'nin modunu kullanmak yerine, bunun ortalamasının kullanılması daha uygundur³⁹.

³⁸ W.Y. Loh, X. Shih, "Split selection methods for classification trees", **Statistica Sinica**, vol. 7, 1997, pp.815-840.

³⁹ Rokach, Maimon, **a.g.e.**, pp.179-180.

2.9 Yanlış Sınıflandırma Maliyetleri

Birçok karar ağacı indükleyicileri başka bir sınıfa ait olan bir parçanın herhangi bir sınıfa sınıflandırmasından kaynaklanan sayısal cezalar sağlanabilir.

2.10 Örnekleri Ağırlıklandırma

Bazı karar ağaçları indükleyicileri (algoritmaları), farklı örneklerle farklı şekilde işlem yapabilir. Her bir örneğin katkılarını, analizde sağlanan bir ağırlığa göre (0 ile 1 arasında) ağırlıklandırarak gerçekleştirilir.

2.11 Karar Ağaçlarının Avantajları ve Dezavantajları

Bir sınıflandırma aracı olarak karar ağaçlarının literatürde birçok avantajlarından bahsedilmektedir.

- Karar ağaçları kendini açıklayıcı özelliğe sahiptir ve yoğun olduğu zaman bile takip etmesi kolaydır. Yani, eğer karar ağacı makul sayıda yapraklara sahip olduğu zaman, profesyonel olmayan kullanıcılar tarafından kolayca anlaşılabilir. Ayrıca, karar ağaçları kurallar kümesine dönüştürülebilir. Böylece, bu tarz temsil anlaşılabilir.
- Karar ağaçları hem nominal (kategorik) hem de sayısal (sürekli) girdi öznitelikleri ile işlem yapabilir.
- Karar ağacının gösterimi herhangi bir ayrık değerli sınıflandırıcıyı ifade etmek için çok zengindir.
- Karar ağaçları hatalar ihtiva eden veri kümelerini işleme yeteneğine sahiptir.
- Karar ağaçları eksik değerlere (missing values) sahip veri kümelerini işleme konusunda başarılıdır.
- Karar ağaçları parametrik olmayan bir metot olarak kabul edilir. Bundan dolayı, karar ağaçları uzay dağılımı ve sınıflandırıcının yapısı hakkında varsayımlara sahip değildir.

Bununla beraber, karar ağaçları dezavantajlara da sahiptir.

- Birçok karar ağacı algoritmaları, ID3 ve C4.5 gibi, hedef özniteliğin sadece kesikli ayırık değerler (discrete values) gerektirmektedirler.
- Karar ağaçlarının aç gözlü özelliği bir diğer dezavantaj sağlamaktadır. Bu dezavantaj ise karar ağaçlarının eğitime kümesine, alakasız özniteliklere ve ses'e karşı aşırı duyarlı olmasıdır (Quinlan, 1993) .
- Karar ağacı “böl ve yönet” metodu kullandığı gibi, eğer birkaç çok tane çok alakalı öznitelik mevcut ise çok iyi bir performans göstermektedirler, fakat eğer birçok karmaşık etkileşimler mevcut ise daha az performans gösterirler. Bunun sebeplerinden bir tanesi diğer sınıflandırıcılar bir karar ağacını kullanarak ifade etmesi çok zor olan bir sınıflandırıcıyı kompakt bir şekilde tarif edebilirler.⁴⁰

2.12 Karar Ağaçları İndükleyicileri

2.12.1 CHAID⁴¹

Yetmişli yılların başında başlayan, karar ağaçlarını oluşturmak için araştırmacılar uygulamalı istatistikte çeşitli yöntemler geliştirdiler, örneğin: AID (Sonquist v.d., 1971), MAID (Gillo, 1972) , ve CHAID (Kass, 1980). CHAID (Chisquare- Automatic – Interaction –Detection) orijinal olarak sadece nominal öznitelikleri (değişkenler) ile işlem yapacak şekilde geliştirildi. Her bir girdi özniteliği olan a_i için, CHAID hedef özniteliğine göre en az anlamlı derecede farklı olan V_i 'de değerler çifti bulur. Anlamlı farklılık, bir istatistik testinden elde edilen p değeri yolu ile ölçülür. Kullanılan istatistik testi, hedef özniteliğin türüne bağlıdır. Eğer hedef nitelik sürekli ise, bir F testi kullanılır.

⁴⁰ Rokach, Maimon, **a.g.e.**, pp. 183-184.

⁴¹ G. Kass“An exploratory technique for investigating large quantities of caregorical data”, **Applied Statistics**, vol. 29(2), 1980, pp.119-127.

Eğer değişken (öznitelik) nominal ise, bir Pearson chi-squared testi kullanılır. Eğer değişken (öznitelik) sıralı ise, o zamana olabilirlik-oran (likelihood-ratio test) testi kullanılır.

Seçilen her bir çift için, CHAID elde edilen p -değerinin belli bir birleştirme eşik değerinden büyük olup olmadığını kontrol eder. Eğer cevap pozitif ise, değerleri birleştirir ve ilave potansiyel çiftleri birleştirmek için arama yapar. Bu süreç anlamlı çiftler bulunmayana kadar tekrarlanır. Böylece hâlihazırdaki düğümü bölmek için kullanılan en iyi girdi özniteliği (değişken) seçilir, öyle ki her bir çocuk (alt) düğüm bir seçilen özniteliğin değerlerinin homojen bir grubundan oluşur. Unutmamak gerekir ki, eğer en iyi girdi özniteliğin düzeltilmiş p değeri belli bir bölme eşik değerinden küçük değil ise bölme gerçekleşmez. Aşağıdaki durumlardan bir tanesi karşılandığı zaman bu prosedürde durur:

- Maksimum ağaç derinliğine ulaşılır.
- Üst (ebeveyn) düğüm olmak için düğümdeki minimum sayıda duruma ulaşılır, böylece daha fazla bölünmez.
- Alt (çocuk) düğüm olmak için düğümdeki minimum durumlar sayısına ulaşılır.

CHAID eksik değerlerin hepsini tek bir geçerli kategori gibi muamele ederek işlem yapar. CHAID budama işlemi gerçekleştirmez.

2.12.2 ID3⁴²

Bu algoritma çok basit bir karar ağacı algoritması olarak kabul edilir (Quinlan J. , 1986). ID3 bilgi kazancını (information gain) bölme kistasları (splitting criteria) olarak kullanır. Ağacın büyümesi tüm örneklerin tek bir hedef özelliğın değerine ait olduğu zaman durur ya da en iyi bilgi kazancı sıfırdan büyük olmadığı zaman durur. ID3

⁴² J.R. Quinlan, "Introduction of decison trees", **Machine Learning**, vol.1, 1986, pp.81-106.

herhangi bir budama prosedürünü uygulamaz, ayrıca sayısal öznitelikler veya eksik değerler ile işlem yapmaz.

2.12.3 C4.5

ID3 nin bir evrimidir ve aynı yazar tarafından ((Quinlan J. , 1993)) tarafından sunulmuştur. Bölme kıstası olarak kazanç oranı (gain ratio) kullanır. Bölme işlemi bölünmesi gereken örneklerin sayısı belli bir eşik değerinden düşük olduğu zaman sonlanır. Ağacın büyüme işlemi gerçekleştikten sonra hata tabanlı budama (Error –based pruning) işlemi başlar. C4.5 sayısal öznitelikleri (değişkenleri) işleyebilir. Düzeltilmiş kazanç oranı ölçütü (corrected gain ratio) kullanarak eksik değerler içeren bir eğitime kümesinden indükleyebilir⁴³.

2.12.4 QUEST⁴⁴

QUEST (Quick, Unbiased, Efficient, Statistical Tree) algoritması tek değişkenli ve doğrusal bileşim bölmeleri destekler (Loh, W.Y., Shih X. , 1997). Her bir bölme için, ANOVA F- testi yada Levene's test (sıralı ve sürekli öznitelikler için) ya da Pearson's Chi-Square (nominal öznitelikler için) kullanılarak her bir girdi özniteliği (değişkeni) ile hedef özniteliği (bağımlı değişken) arasındaki bağlantılar hesaplanır. Eğer hedef değişken (öznitelik) çok terimli ise, iki süper sınıf oluşturmak için iki –ortalımalı kümeleme (two –means clustering) kullanılır. Hedef öznitelikli en yüksek bağlantıyı (association) elde eden öznitelik bölme için seçilir. Kuadratik diskriminant analizi (QDA) girdi özniteliği için optimal bölme noktasını bulmak için uygulanır. QUEST algoritmasının önemsiz yanlılığa sahiptir (bias) ve ikili ağaçlar sağlar. Ağaçları budamak için on –katlı çapraz - doğrulama (ten - fold cross - validation)⁴⁵ metodunu kullanır.

⁴³ Rokach, Maimon, **a.g.e.**, p.181.

⁴⁴ W.Y. Loh, X. Shih, **a.g.e.**, pp.815-840.

⁴⁵ Rokach, Maimon, **a.g.e.**, p.182.

QUEST bölünmüş alan seçimi (split - field selection), bölünmüş nokta seçimi (split - point selection) ayrı olarak ilgilenir. QUEST'te ki tek değişkenli bölme, yaklaşık olarak yansız alan seçimi gerçekleştirir. Yani, eğer tüm belirleyici (predictor) alanlar hedef alana göre aynı derecede bilgilendirici ise, QUEST belirleyici alanlardan herhangi bir tanesini eşit olasılıkla seçer. QUEST algoritması, CART algoritmasının birçok avantajlarını tanımaktadır, fakat CART'ta olduğu gibi ağaçlar hantal olabilir. Ağacın boyutunu küçültmek için otomatik maliyet – karmaşıklık budama (automatic cost –complexity pruning) tekniği, QUEST ağacına uygulanabilir. QUEST eksik değerleri halletmek için vekil (surrogate) bölme kullanır⁴⁶.

2.12.4.1 Model Parametreleri⁴⁷

2.12.4.1.1 Alan seçimi

- 1- Her bir belirleyici alan X için, eğer X kategorik bir alan ise, nominal ya da sıralı, bir X ve bağımlı alan arasındaki bağımsızlığın bir Pearson chi-square testinin p -değeri hesaplanır. Eğer X değeri sürekli ise (ölçek - seviyeli) F testi kullanılarak p değeri hesaplanır.
- 2- En küçük p değerinin önceden belirlenen (Benferroni-adjusted alpha level α_B) bir değer ile kıyaslanması
 - a. Eğer en küçük p değeri α_B 'den küçük ise, karşılık gelen belirleyici alan düğümü bölmek için seçilir. Sonraki adıma gidilir.
 - b. Eğer en küçük p değeri α_B 'den küçük değil ise, o zaman sürekli olan her bir X için bir p değeri hesaplamak için eşit olmayan varyanslar için Levene's test'i kullanılır. Bir başka deyişle, X 'in hedef alanın farklı seviyelerinde eşit olmayan varyanslara sahip olup olmadığının test edilmesi.

⁴⁶ SPSS, Clementine® 12.0 Algorithms Guide. Integral Solutions Limited, 2007, p.56.

⁴⁷ A.e., p.57.

- c. Leven's testinden olan küçük p değerini yeni bir Bonferroni – düzeltilmiş alfa düzeyi (Bonferroni –adjusted alpha level) olan α_L ile kıyaslanır.
- d. Eğer p değeri α_L 'den küçük ise düğümü bölmek için Leven testinden en küçük p değerli olan ilgili açıklayıcı alan seçilir.
- e. Eğer p değeri α_L 'den büyük ise düğüm bölünmez.

2.12.4.1.2 Bölünmüş Nokta Seçimi

- 1- Eğer Y sadece iki kategoriye sahip ise, bir sonraki adıma geçer. Yoksa Y 'nin kategorilerini iki üst sınıflara aşağıdaki gibi gruplara ayırır:
 - a. Y 'nin her bir kategorisi için X 'in ortalaması hesaplanır
 - b. Eğer tüm ortalamalar aynı ise, en büyük ağırlıklı frekans kategori bir üst sınıf olarak seçilir ve diğer tüm kategoriler diğer üst sınıf oluşturmak için birleştirilir.
 - c. Eğer ortalamalar tamamen aynı değil ise, başlangıç küme merkezleri iki en uç sınıf ortalaması olarak ayarlanması ile, iki – ortalamalar kümesi algoritması Y 'nin iki üst sınıfını elde etmek için o ortalamalara uygulanır (burada k -ortalamalar kümesinin bir özel durumunda, $k=2$ olarak alınır)
- 2- Kuadratik diskriminant analizi (QDA) bölme noktasının (split point) belirlemek için uygulanır. QDA genellikle iki kesme noktası (cut-off point) üretir, ilk üst sınıfın örnek ortalamasına en yakın olanı seçilir.

QEUST ilk önce, sembolik alanı bir sürekli ξ alanına, diskriminant koordinatlarını, belirleyicinin kategorilerine atama yaparak dönüştürür.

2.12.5 CART⁴⁸(C&R Tree)

CART (Classification and Regression Trees) açılımı Sınıflandırma ve Regresyon Ağaçları olarak tanımlanır. CART, 1984 tarihli bir monograf olup Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone (BFOS) tarafından beraber yazıldı. Bu algoritma yapay zekâ, makine öğrenimi, parametrik olmayan istatistik ve veri madenciliği için çok önemli bir dönüm noktasını temsil eder.

CART karar ağacı tekrarlamalı ve ikili bir bölme prosedürüdür. Bağımlı ve bağımsız değişkenler nominal ya da sürekli olabilir. Veri ham hali ile işlenir. Kök düğümünden başlanılarak veri, iki çocuk düğüme (alt düğüm) , ve daha sonra her bir çocuk düğümü tekrar iki torun çocuk düğümlerine ayrılır. Böylece ağaçlar durdurma kuralı olmaksızın en yüksek büyüklüğe ulaşır. Aslında, ağaç büyüme süreci veri eksikliğinden dolayı artık bölünme gerçekleşmez. Daha sonra maksimum büyüklükteki ağaç, yeni olan maliyet karmaşıklığı budama metodu yolu ile geriye doğru, yani kök’ e doğru budama yapılmaya başlanır. Budama işlemi yapılacak olan sonraki bölünme eğitme verisi üzerindeki ağacın toplam performansına faydası en az olan bölünmedir.

CART algoritmasının amacı, sadece bir ağaç ortaya çıkarmaktan ziyade, sıralı iç içe olan, her biri en uygun (optimal) ağaç adayı olan budanmış ağaçlar oluşturmaktır. “Uygun büyüklükteki” veya “dürüst” bir ağaç bağımsız test verisindeki budama serisinde her bir ağacın tahmin performansının hesaplanması ile belirlenir. C4.5 ‘e kıyasla, CART ağaç seçiminde dâhili (eğitme verilere dayalı) performans ölçütü kullanmıyor. Bunun yerine, ağaç performansı her zaman ya bağımsız test verileri üzerinde ya da çapraz- sağlama yolu ile ölçülür ve ayrıca ağaç seçimi sadece test veriye dayalı hesaplama sonrasında devam eder. Eğer test veya çapraz – sorgulama uygulanmamış ise, CART serideki hangi ağacın en iyi olduğu konusunda agnostik kalır.

⁴⁸ L. Breiman, J. Friedman, R. Olshen, C. Stone, **Classification and Regression Trees**, Wadsworth Int. Group, 1984, pp. 216-264

Bu yöntem eğitme verileri ölçütlerine dayanacak şekilde öncelikli model oluşturan klasik istatistik veya C4.5 gibi metotlara çok zıttır.

CART mekanizması opsiyonel otomatik sınıf dengelemesi ve otomatik eksik değer (missing value) işlemlerini içermektedir. Ayrıca CART, maliyet- duyarlı öğrenmeye, dinamik özellik inşasına ve olasılık ağaç tahminlerine fırsat tanımaktadır. Son raporlar yeni nitelikli önem sıralaması içermektedir. Farklı çapraz-sağlama klasörlerindeki ağaçların terminal düğümlerin sayısı ile hizaya gelemeyebileceğini farz edersek, CART yazarları budama dizisinde her bir ağaç için performans ölçümünde çapraz-sağlamanın nasıl yapıldığının göstererek çığır açtılar⁴⁹.

Ağaç diyagramının en tepesindeki olan kök düğüm tüm eğitme verilerini (training data) içermektedir. Bağımlı değişkene ait tüm verileri burada toplanmaktadır. Eksik veri olsun veya olmasın, bağımlı değişkene ait her bir örnek bağımsız değişken nitelikleri üzerinde veriler içermektedir. CART mevcut en iyi ayırıcıyı bulmak için veriyi arayarak başlar, her bir tahmin değişkeninin öznelik – değer çiftini bu tahmin değişkeninin bölme iyiliği (goodness of split) için test eder. CART ile elde edilen örneklerle baktığımızda, ağaçların dallanması (ayrışması) iki düğümde meydana gelmektedir, bundan dolayı CART iki bileşenli dallandırıcı kullanmaktadır.

Bütün bir ağaç oluşturmak için, CART her iki çocuk düğüm’de tasvir edilen ayrışma ya da dallanma süreçlerini tekrar ederek kök’ün torunlar oluşturur. Torunlar dallanarak torun çocuklar oluşturur ve bu süreç veri yetersizliğinden dolayı ayrışmanın imkânsız olduğu ana kadar devam eder. Ayrışmanın son bulduğu en uçtaki düğümlere terminal düğüm adı verilir.

⁴⁹ Dan, Steinberg, “CART: Classification and Regression Trees”, **The Top Ten Algorithms in Data Mining**, Ed. by, XindongWu, Vipin Kumar, Boca Raton FL, Chapman & Hall/CRC Taylor and Francis Group LLC, 2009, p.181.

2.12.5.1 Algoritma Nasıl Çalışır?

Hem regresyon hem de sınıflandırma için çok sayıda bölünme kurallarının mevcut olması, sürekli ve kategorik ayırıcıların ayrı işlenmesi, çok düzeyli kategorik ayırıcıların özel işlenmesi ve eksik verilerin sağlanması gibi nedenlerden dolayı CART algoritmasının tüm teknik detayları çok uzun ve karmaşıktır. Ağaç – büyüme sürecini takiben, ağaç budama süreci de karmaşık bir süreçtir ve son olarak ise ağaç seçimidir.

Aşağıda basitleştirilmiş ağaç büyütme algoritması verilmiştir. Aşağıda verilen basitleştirilmiş CART şeması, eksik değer (missing value) , sınıflandırma atamaları veya CART ile ilgili önemli detaylar hakkında herhangi bir dayanak sunmamaktadır. Verilen CART algoritması mümkün olan en büyük ağaç oluşturulmasının mekanizmasını göstermektedir.

Tablo 1. Basitleştirilmiş Ağaç Büyütme Algoritma şeması⁵⁰

```
BAŞLA:      Eğitme Verilerinin hepsini kök Düzümüne atama yap
            Kök düğümü terminal düğüm olarak tanımla
BÖLME:
Yeni Bölmeler=0
Ağaçtaki her terminal düğüm için:
Eğer terminal düğüm örnek büyüklüğü çok küçük ise
yaya düğümdeki tüm örnekler aynı hedef sınıfa ait SONRAKİNİ AL
Uygun bölme kuralını kullanarak düğümü en iyi şekilde iki çocuk
düğümlere bölen niteliği bul
Yeni Bölmeler+1
SONRAKİNİ AL
SONRAKİ
```

Ağaç büyütülmesinden sonra, CART sonraki aşamada iç içe dizi halinde budanmış alt ağaçlar üretir. Aşağıda verilen budama için basitleştirilmiş bir algoritma

⁵⁰ Steingberg, a.g.e., p.184.

şeması önselleri (priors) ve maliyetleri göz ardı eder. Bu, asıl CART budama algoritmasından farklıdır ve kısa, öz olması ve okuma kolaylığından dolayı burada yer verilmiştir. Maksimum büyütülmüş ağaç ($T_{\max}$) ele alınır ve tüm bölmeleri kaldırılır, eğitme verilerindeki ağacın doğruluğunu iyileştirmeyen iki terminal düğüm elde edilerek süreç başlar. Bu CART budama yönteminin başlangıç noktasıdır. Budama işlemi tekrarlı bir şekilde ağaç içerisindeki en zayıf halkaların (test verilerinde ağacın performansına en az katkı sağlayan dallanmalar) çıkarılması ile devam eder. Aşağıdaki figürde verilen algoritmada, budama işlemi sadece iki terminal düğümlerin ebeveynleri (üst düğümler) için sınırlandırılmıştır.

Tablo 2. Basitleştirilmiş budama algoritması⁵¹

```

TANIMLA:  $r(t)$  =  $t$  düğümündeki eğitme verileri yanlış sınıflama oranı
 $p(t)$  =  $t$  düğümündeki eğitme verileri yüzdesi
 $R(t) = r(t) * p(t)$ 
 $t_{\text{sol}} = t$  düğümü sol çocuk
 $t_{\text{sağ}} = t$  düğümü sağ çocuk
 $|T|$  =  $T$  ağacındaki terminal düğüm sayısı

BAŞLA:  $T_{\max}$  = Oluşan en büyük ağaç
Geçerli_Ağaç =  $T_{\max}$ 
İki terminal düğümlerin tüm  $t$  ebeveynleri
 $R(t) = R(t_{\text{sol}}) + R(t_{\text{sağ}})$  için tüm bölmeleri sil
Geçerli_Ağaç = Budama sonrası  $T_{\max}$ 

BUDAMA: Eğer  $|Geçerli\_Ağaç| = 1$  BİTTİ ye GİT
İki terminal düğümlerin tüm  $t$  ebeveynleri için
 $R(t) - R(t_{\text{sol}}) - R(t_{\text{sağ}})$  değeri minimum olması için  $t$  düğüm(leri)
Kaldır
Geçerli_Ağaç = Budama sonrası Geçerli_Ağaç

```

⁵¹ Steingberg, a.g.e., p.184.

CART budama algoritması, bütün bir alt ağacı tek bir budama hareketi ile ortadan kaldırabilen düğümler mekanizmasına ceza uygulaması açısından yukarıda bahsedilen algoritmadan farklıdır. Monograf uygulanan işlemi açık ve geniş özetini sunmaktadır.

2.12.5.2 CART Algoritmasında Ağaç Bölme Kuralları

CART bölme kuralları her zaman aşağıdaki şekilde çalışır:

“Özellik Sol’a ayrılır eğer ŞART sağlanırsa ve sağ’a ayrılır eğer ŞART sağlanmaz ise”. Burada ŞART şu şekilde ifade edilir “değişken $X_i \leq C$ ” sürekli özellikler için. Kategorik ve nominal özellikler için ŞART değerler listesinin bir üyesi olarak ifade edilir. Örneğin, ŞEHİR değişkeni üzerinde bölme şu şekilde ifade edilebilir. Özellik sol’a ayrılır eğer ŞEHİR {Chicago, Detroit, Nashville} kümesi içinde ise, Yoksa özellik sağ’a ayrılır. Daha sonraki kısımlarda belirtilecek olan seçilen bölme kurallarından bir tanesi kullanılarak, en uygun bölme ile ayırıcı ve ayırma (bölme) noktası CART tarafından otomatik olarak tespit edilir. CART gruplanmamış veriler ile çalıştığından dolayı, optimal bölmeler özelliklerin sıra – koruyan dönüşümlerine göre (örneğin logaritmik, karekök, üstel dönüşümler gibi) her zaman sabittir. CART yazarlarına göre (Breiman, v.d., 1984) ikili bölünmeler çoklu bölünmelere tercih edilmektedir, çünkü;

- 1) İkili bölünmeler verileri çoklu bölünmelere göre daha yavaş parçalara ayırır
- 2) Aynı özellik üzerinde tekrarlı bölünmelere izin verilir ve eğer seçilir ise, bir özellik için gerekli çok sayıda bölmeler üretecektir

CART yazarları ağaçların sınıflandırması için dört bölme kurallarını (Gini, twoing, ordered twoing, symetric gini) kullanarak örnekleri ele alırlar, fakat monograf birçok tartışmasını Gini üzerinde odaklar, buna ise çok iyi bilinen entropi kriterine benzemektedir. İkili (0/1) hedefi için, t düğümünün “Gini empürite ölçüsü”

$$G(t) = 1 - p(t)^2 - (1 - p(t))^2 \quad (2.24)$$

Burada $p(t)$, düğümdeki sınıf 1 'in görel frekansını ifade etmektedir.

$$G(t) = -p(t) \ln p(t) - (1 - p(t)) \ln(1 - p(t)) \quad (2.25)$$

olarak belirtilirse entropi kuralını verir.

P ebeveyn düğümün L ve R gibi sol ve sağ çocuklara bir kez bölünmesi ile elde edilen ilerleme (kazanç, gelişim)

$$I(P) = G(P) - qG(L) - (1 - q)G(R) \quad (2.26)$$

şeklindedir.

Burada, q (ihtimal ağırlıklı) sol' a giden örneklerin kısmıdır(fraksiyonu). CART yazarları Gini metodunu entropi yerine tercih ediyorlar çünkü Gini çok daha hızlı hesaplanabilir, simetrik yapılmış maliyetleri içermek için kolayca genişletilebilir ve daha az olası “ucu – kesik ” bölmeler oluşturur. “ ucu – açık” bölmeler bir bölmesi çok küçük ve diğeri ise çok büyük çocuk bölmelerinden meydana gelir. CART sonraki sürümlerinde entropi opsiyonel bölme kuralı olarak eklenmiştir. “twoing” kuralı hedef özelliğın dağılımının iki çocuk düğümlerde direkt kıyaslamasına dayanır.

$$I(split) = \{.25(q(1 - q))^u \sum_k |p_L(k) - p_R(k)|\}^2 \quad (2.27),$$

burada, k hedef sınıfları dizinler(endeksler), $p_L()$ ve $p_R()$ sıra ile sağ ve sol çocuk düğümlerinde ki hedefin olasılık dağılımlarını belirtir. Bu ayırıcı Messenger ve Mandell'in, 1972 düzeltilmiş sürümüdür. Twoing “iyileştirmesi” sol ve sağ çocuk olasılık vektörleri arasındaki farkı ölçer ve $q=0.5$ 'de maksimum değere sahip olan en önemli terim olan $[.25(q(1 - q))]$ kesin olarak eşit olmayan sol ve sağ düğüm boyutları

meydana getiren bölmeleri cezalandırır. Kuvvet terimi olan “u” ise kullanıcı tarafından kontrol edilebilen, eşit olmayan bölmeler için sürekli artan oranda ağır cezaya izin verir; örneğin, $u=10$ düzenlemesi tüm bölmelerin bölme özniteliğinin medyan değerinde zorla uygulanmasına benzerdir. Bizim pratik tecrübelerimizde, twoing kriteri çoklu sınıf hedeflerinde ve tabiatı gereği tahmini çok zor olan ikili hedefler de üstün oyuncu konumundadır. BFOS ayrıca hedefin sınıflarını sıralı olarak elen alan ve twoing bölme kriterinin başka bir çeşidini sunmaktadırlar. Önerilen bu kriterin “ordered twoing” bölme kriteri olarak bilinmektedir, regresyon kuralının karakteristik özelliğine sahip bir sınıflandırma kuralıdır, ve her bölmede (ayırma) en alt dereceli hedef sınıflarını en yüksek dereceli hedef sınıflarından ayırmaya çalışır.

Regresyon (Sürekli hedefler) için, CART, bir bölmenin iyileştirmesinin ölçülmesi için en küçük kareler kriteri (Least Square, LS, kareli tahmin hatalarının toplamı) ve en küçük mutlak sapma kriteri (Least Absolute Deviation, LAD, mutlak tahmini hataların toplamı) seçeneklerini temel olarak sunar. Sınıflandırma ağaçlarında olduğu gibi, en iyi bölme en büyük iyileştirmeyi elde eder.

2.12.5.3 Önsel Olasılıklar ve Sınıf Dengeleme

Sınıf dengeleme konusu bu alanda çalışan uygulayıcılar için makine öğrenimi alanında çok önemli bir konudur, çünkü eğitim verisi yüksek derecede dengesiz olduğu zaman birçok veri madenciliği metotları iyi sonuçlar vermemektedirler. Örneğin, birçok kredi kuruluşları için varsayılan kredi oranı genellikle tüm hesapların 5%’in altındadır, kredi kart hareketlerinde dolandırıcılık normal olarak 1%’in baya altındadır ve internet reklam “link tıklanma oranı” tüm gösterilen (izlenilen) reklamların 1%’den çok daha azdır. Birçok uygulayıcı, eğitim verilerindeki hedef sınıflar yaklaşık olarak eşit örneklem büyüklükleri verecek şekilde örneklendirildiğinden dolayı, her zaman kendilerini eğitim verileri ile sınırlandırır. Eğer ilgili sınıf çok küçük ise, yukarıdaki gibi örneklendirme dengelemesi analizciye çok küçük boyutlu eğitim örnekleme kalır.

Mesela, bir sigorta sahtekârlığı çalışmasında, şirket 70 adet dosyalanmış hilekârlık davası belirlemektedir. Analiz sadece dengelenmiş örneklem ile sınırlandırılırsa, analizecin sadece toplam olarak 140 (70 adet hilekârlık dosyası, 70 adet normal dosya) adet örneklem ile sınırlandırılmış olacaktır.

CART yazarları 1984’te bu konuyu açık bir şekilde ele almışlardır. Modelciyi herhangi bir örnekleme dengesi ile alakalı endişelerden uzak tutacak şekilde bir yol tasarlamışlardır. Eğitim verileri ne kadar çok dengesiz olursa olsun, CART, modelci tarafından herhangi bir hareket, hazırlık, örnekleme ve ağırlandırma gerekmesizin otomatik olarak dengesizliği giderecektir. Veriler bulunduğu şekli ile herhangi bir ön işleme yapılmaksızın modellenelir.

Yukarıda belirtilen esnekliği sağlamak için, CART “önseller” mekanizmasının kullanır. Önseller hedef sınıf ağırlıklarına çok yakındırlar fakat ağaçta CART tarafından rapor edilen herhangi sayılanları etkilemeyecek kadar görünmezdirler. Bunun yerine, önseller bölmelerin uyumunu belirlemek için üstlenilen hesaplamalarda gömülüdürler. Varsayılan sınıflandırma durumunda, CART her zaman kökteki sınıf frekansları ile alakalı her düğümdeki sınıf frekansını hesaplar. Bu işlem, sınıfları dengelemek için veriyi otomatik olarak yeniden ağırlandırmak ile eşdeğerdir ve optimal olarak seçilen ağacın dengelenmiş sınıf hatasının minimize etmesini garanti eder. Yeniden ağırlandırma işlemi tüm olasılıkların ve iyileştirmelerin (ilerlemelerin) hesaplanmasında gizlidir ve hiçbir kullanıcı müdahalesi gerektirmez; her bir düğümde rapor edilen örnek sayısı ağırlandırılmamış olan veriyi ifade eder. İkili (0/1) hedef için herhangi bir düğüm “sınıf 1” olarak sınıflandırılır sadece ve sadece

$$\frac{N_1(\text{Düğüm})}{N_1(\text{Kök})} > \frac{N_0(\text{Düğüm})}{N_0(\text{Kök})} \quad (2.28)$$

Bu ifade, K adet hedef sınıfı mevcut olduğunda, veride sınıfların gerçek dağılımına bakılmaksızın, her sınıf’a 1/K olasılık değeri kök düğümde atama

yapılmasını garanti eder. Monograf'ta bu varsayılan durum “önsel eşitlikler” olarak anılır. Bu, yeniden sınıf dengelenmesi için özel veri hazırlığına gerek kalmaksızın ya da manüel olarak elde edilmiş ağırlıkların sunulmasına gerek kalmadan, CART kullanıcılarına rahatlıkla dengelenmemiş veri ile çalışmalarına imkân sağladı. Dengelenmemiş veri ile çalışmak için, CART'ın varsayılan ayarlarını kullanarak çalıştırmak yeterlidir. “Önsel Veri (Priors Data)” opsiyonu seçilmesi ile örtük yeniden ağırlandırma kapatılabilir. Modelci, maliyetleri ya da eğitme verileri ve gelecek veri hedef sınıf dağılımları arasındaki potansiyel farkları yansıtmak için bir keyfi öncelikler seti belirtmek için seçebilir.

Not. Önsellerin ayarları, ağırlıklara rağmen, bir düğümde rapor edilen sayıları ya da her bir hedef sınıftaki rapor edilen kesirleri etkilemezler. Önseller herhangi bir düğüm atanan sınıfı ve ağaç - büyütme sürecinde bölme seçimini etkiler.

Önsellere olan güvenilirlik araştırmacının farklı hedef sınıflarından farklı oranlarda örnekleme konusunu görmezden gelmesi manasına gelmez; aksine araştırmacıya ne zaman ve nasıl örnekleme yapacağı konusunda geniş esneklik sağlamaktadır.

2.12.5.4 Eksik Değer İşlemleri

Eksik değerler genellikle gerçek dünya problemlerinde, özellikle işletme ile ilgili veritabanlarında görünür. Bunlarla uğraşma ihtiyacı tüm modelciler için can sıkıcı bir sorundur. CART'ın en önemli katkılarından bir tanesi eksik değerleri halletmek için tamamen otomatik ve etkili bir mekanizma içermesidir. Karar ağaçları üç aşamada eksik değerleri halletme mekanizmasına ihtiyaç duyar: a) ayırıcı (bölme) hesaplaması aşamasında b) eğitme verisinin bir düğüm üzerinden taşındığı zaman c) test verisinin son sınıf ataması için düğüm üzerinden taşındığı zaman (Quinlan, 1989). (a) ile ilgili olarak, CART'ın ilk sürümü, ayırıcı(dallandırıcı, bölücü) eksik olmayan veri altkütmesinde, her

bir ayırıcıyı performansına göre tam olarak hesapladı. CART'ın sonraki sürümleri eksikliğin derecesinin yansıtması için iyileştirme ölçütünü azaltan cezalar takımı sunar. Örneğin, eğer bir değişken bir düğümdeki kayıtların 20%'sinde eksik ise, bu değişkenin ilgili düğümdeki iyileştirme skoru 20% kadar azaltılabilir, ya da alternatif olarak 20%'nin yarısı ve vesaire. CART mekanizması, (b) (c)'ler için, eğitim verisinde eksik değer olsun ya da olmasın ağacın her bir düğümü için “vekil” ya da yedek ayırıcı meydana getirir. Böylece vekiller mevcuttur, tüm veride eğitilen bir ağaç eksik değer içeren yeni veriye uygulanmalıdır. Bu durum, eğitim verilerinde eksik değerlere müsamaha etmeyen ya da eksik değerlerin halledilmesi hakkında bilgiyi eksik değerler içeren eğitim verilerinden edinen makinelerle ile keskin bir çelişkidir. Friedman (1975) her iki sol ve sağ çocuk düğümlere doğru hareket eden eksik değere sahip ayırıcı örnekleri önerir, ve içinde örneğin bulunduğu tüm düğümlerin ağırlıklı ortalamasını alarak final sınıf ataması yapılmasını önerir. Quinlan, alternatif eksik değer halletme metotları çalışmasında Friedman yaklaşımının farklı bir yanını tercih eder. Quinlan test amaçlar için kendisinin uyguladığı yaklaşık yedekler (vekiller) temeli üzerinde agnostik kalıyor iken (Quinlan), eksik değer mevcut olduğunda, CART yedek performansının etkinliğinin kendi açımızdan değerlendirmelerimiz kesin olarak elverişlidir. Friedman, Kohavi, ve Yun (1996) çalışmalarında, Friedman CART kodunu 50% sinin sadece eksik değerleri halletmek için ayrıldığını not etmektedir, böylece Quinlan'ın deneysel versiyonu, CART yedek (vekil) mekanizmasını tekrarlaması olasılık dışı bir durumdur.

CART'da, eksik değerlerin halledilme mekanizması tamamen otomatik ve her düğümde lokal olarak uyarlanabilir. Ağacın her düğümünde, seçilen ayırıcı verinin ikili bölünmesine sebep olur (e.g. $X_1 \leq c_1$ ve $X_1 > c_1$). Yedek ayırıcı Z gibi tek özniteliktir, yedeğin kendisi de ikili ayırıcı şeklinde (e.g. $Z \leq d$ ve $Z > d$) olduğu yerdeki bölünmeyi tahmin eder. Başka bir ifadeyle, her ayırıcı tek bölme ikili ağaç ile tahmin edilmesi gereken yeni bir hedef haline gelir. Yedeklerin varsayılan kurala olan avantajını ölçen bir bağlantı puanı yolu ile yedekler sıralanır. Varsayılan kural, önseller için ayarlamalar yapıldıktan sonra tüm durumların daha büyük çocuk düğüme

gidilmesini tahmin eder. Yedek (surrogate) olarak nitelendirmek için, değişken bu varsayılan kuralı çok daha iyi yapmak zorundadır. Bundan dolayı yedekler bulmak her zaman mümkün olmayabilir. CART ağacında bir eksik değer ile karşılaşıldığında, örnek birinci sınıf yedeğe göre ya sağa ya da sola hareket ettirilir. Eğer bu yedek eksik ise bunun yerine ikinci sınıf yedek kullanılır ve bu şekilde devam eder. Eğer tüm yedekler eksik ise önseller için ayarlamalar yapıldıktan sonra varsayılan kural örneği en büyük çocuk düğüme atar. Bağlar bir örneği sola hareket ettirilmesi ile kırılır.

2.12.5.5 Öznitelik Önemi

Bir özniteliğin önemi, özniteliğin bir ayırıcı olarak bulunduğu tüm düğümlerdeki iyileştirmelerin toplamına dayanır. Yedeklerde iyileştirme hesaplarında dâhil edilir, bunun manası bir düğümü hiçbir zaman bölmeyen bir değişken bile çok büyük iyileştirme puanı alabilir. Bu değişken önem sıralamalarının, değişken maskelemesin ve öznitelikler arasındaki doğrusal olmayan korelasyonun ortaya çıkmasına imkân sağlar. Önem puanları seçmeli olarak ayırıcılara sınırlandırılabilir; sadece ayırıcıların ve hepsinin (ayırıcılarla ve yedeklerin) kıyaslanması çok faydalı bir teşhistir.

2.12.5.6 Dinamik özellik İnşası

Friedman (1975) her düğümden yeni özelliklerin otomatik olarak inşasından bahseder ve ikili hedef için ise tek özellik olan

$$x \times w$$

iki sınıf arasında, x sürekli prediktör öznitelikler vektörü ve w ise ortalamaların fark ölçekli vektörüdür (Fisher doğrusal diskriminant yönü). Bu düğümden tüm sürekli özniteliklerde lojistik regresyonun kullanılması ile ve tahmini logit'in prediktör olarak kullanılması ile benzerdir

CART monografında (yazılımında), yazarlar otomatik özellik seçimi içeren doğrusal birleşimlerin otomatik inşasını ele alıyorlar, bu imkân CART yazılımının ilk

sürümünde mevcut idi. BFOS ayrıca her bir düğüm içerisinde ayırıcıların Boolean kombinasyonlarının inşası için bir metot sunuyorlar, bu imkân çıkarılan yazılıma dâhil edilmedi. Veride yapıyı ortaya çıkarmanın en iyi yolun lineer kombinasyon ayırıcıların olduğu durumlar olmasına karşın (baknz Huang ver ark. daki. Olshen çalışma, 2004), genellikle böyle ayırıcılar her bir düğümde yüksek miktarda öğrenme gösterdiklerinden dolayı “overfitting” riskini artırdıkları bilinmektedir, böylece alt modellere yönelir

2.12.5.7 Maliyet – Duyarlı Öğrenim

Maliyetler istatistik karar teorisi için merkezidir, fakat maliyet-duyarlı öğrenme sadece Domingos(199) tan önce tutarlı bir ilgi topladı. O tarihten sonra, birçok konferans özellikle bu konu için adandı ve çok sayıda araştırma makalesi sonraki ilmi literatürde yerini aldı. Bundan dolayı CART monografı (yazılımı) maliyet-duyarlı öğrenim için iki strateji sunmasının ve CART’ı tarif eden tüm matematik düzenek (makine) yanlış sınıflandırma açısından çarpık olmasının not edilmesi çok önemlidir. i sınıfının bir örneğini j sınıfı olarak yanlış sınıflandırmanın maliyeti $C(i,j)$ dır ve aksi belirtilmediği müddetçe 1’e eşit olduğu varsayılır $C(i,i)$; $C(i,i)=0$ tüm i ’ler için. Maliyetlerin tüm seti her bir hedef sınıf için bir sıra ve sütundan oluşan C matrisi ile ifade edilir. Herhangi bir sınıflandırma ağacı terminal düğümü atamaları için hesaplanan toplam maliyete, tüm yanlış sınıflandırmalar üzerinde maliyetlerin toplanması ile sahip olabilir. Maliyet-duyarlılık öğrenimindeki konu ağacın oluşumunda ve budama aşamasında maliyetleri dikkate alan bir ağacın oluşumunu başlatmaktır.

Maliyetleri kullanım için ilk ve en basit metot ağırlıklandırmadan faydalanmadır: yanlış sınıflandırma için maliyetli sınıflara ait örnekler yukarı ağırlıklandırılır, ortak bir ağırlık ile verilen bir sınıfın tüm örneklerine uygulanır, Ting (2002) tarafından yeniden keşfedilen bir metottur. CART ta uygulandığı gibi, ağırlıklandırma net bir şekilde sonuçlandırılır böylece tüm düğüm sayıları ham ve ağırlandırılmamış şekillerinde rapor edilir. Çok sınıflı problemler için, yaklaşık olarak maliyetleri yansıtan izafî sınıf

ağırlıklarını elde etmek için, BFOS yanlış sınıflandırma maliyet matrisindeki girdilerin her sıra boyunca toplanacağını önerdi. Bu teknik matris içerisinde detayı önemsemez fakat şu anda sadeliğinden dolayı büyük ölçüde uyarlandı. Gini bölme kuralı için, CART yazarları tüm maliyet matrisini bölme kuralının içerisine yerleştirmenin mümkün olduğunu gösterirler, fakat sadece simetrik yapıldıktan sonra geçerlidir. “SymGini” bölme kuralı $C(i, j)$ ve $C(i, k)$ maliyetlerindeki farka duyarlı ağaçlar oluşturur ve bu kural simetrik olmuş maliyet matrisi karar vericinin probleminin kabul edilebilir bir temsili olduğu zaman en yararlıdır. Bun karşın, örnek ağırlıklandırma yaklaşımı i sınıfının nesnelerinin tüm yanlış sınıflandırılmalar için tek bir maliyet atar. BFOS başarılı maliyet-duyarlı öğrenim için tüm maliyet matrisini kullanarak ağaç budamanın esaslı olduğunu gözlemlediler.

2.12.5.8 Durdurma Kuralları, Budama, Ağaç dizileri ve Ağaç Seçimi

Karar ağaçları üzerine yapılan ilk çalışma budamaya izin vermiyordu, bunun yerine ağaçların büyümesi bazı durdurma koşulları ile karşılaşılana kadar devam ediyordu ve elde edilen ağaç final ağaç olarak nitelendirilirdi. Yazarlar ağaç büyümesini durdurmak için kullanılan herhangi bir durdurma kuralının CART monografında planlanmaması veri yapısında önemli eksiklik meydana getirmeyeceğini garanti edebileceklerini savunmaktadırlar. Böylece ağaçları durdurma kuralı olmaksızın büyütmeği seçtiler. Sonuçta elde edilen aşırı derecede büyük ağaç son optimal modelin çıkarıldığı ham materyal temin eder.

Budama mekanizması tam olarak eğitime verilerine dayalıdır ve $Ra(T) = R(T) + a|T|$ olarak tanımlanan maliyet-karmaşıklık ölçütü ile başlar. Burada “ $R(T)$ ” ağacın eğitime örnekleme maliyeti, “ $|T|$ ” ağaçtaki terminal düğüm sayısı ve “ a ” ise her bir düğüme uygulanan ceza değeridir. Eğer $a=0$ ise, böylece minimum maliyet-karmaşıklık ağacı açık bir şekilde mümkün olan en büyük ağaçtır. Eğer “ a ”nın artan bir biçimde olarak artmasına izin verilirse, minimum maliyet-karmaşıklık ağacı

daha küçük olacaktır çünkü son $R(T)$ değerini azaltan ağacın altındaki bölmeler kesilmiş olacaktır. “a” parametresi 0 dan başlayarak tüm bölmeleri budamak için yeterli olan değere kadar küçük adımlarla istikrarlı bir şekilde artırılır. BFOS bu şekilde elde edilen Q büyüklüğündeki herhangi bir ağaç Q adet terminal düğümleri olan tüm ağaçların içinde minimum olan $R(Q)$ gibi bir maliyet ortaya çıkarabileceğini ispatladı. Bu pratik olarak önemlidir çünkü optimal ağaç için test edilmesi gereken ağaçların sayısını köklü bir şekilde indirgemektedir. Kabul edelim ki maksimal ağaç $|T|$ adet terminal düğüme sahip olsun. Budama bölmeyi kaldırmayı, iki terminal düğüm üretmeği ve iki çocuk düğümü ebeveynlerinin içine almayı ihtiva eder böylece iki terminal düğümü bir tane ile yer değiştirir. Bu gibi budama ile elde edilen maksimal ağaçtan çıkartılan mümkün alt ağaçların sayısı sorudaki spesifik ağacın topolojisine bağlıdır fakat bazen “ $5|T|!$ ” ten büyük olabilir. Fakat incelememiz gereken verilen maliyet-karmaşıklık budama çok daha az ağaçtan oluşmaktadır.

Optimal ağaç test veride minimum maliyeti elde eden budanmış dizide tanımlanan ağaç olarak tanımlanır. Çünkü test yanlış sınıflandırma maliyet ölçümü örnekleme hatasına bağlıdır, budama dizisindeki hangi ağacın optimal olmasına ilişkin olarak belirsizlik her zaman mevcuttur. Hata eğrisinin (ağaç büyüklüğünün bir fonksiyonu olarak yanlış sınıflandırma hata oranı) ilginç bir karakteristiği büyük eğitim veri setleri için eğrinin minimumu etrafında eğri çoğu zaman düzdür. BFOS, tahmini maliyeti minimum maliyet ağacın (“0 SH”) 1 standart hata dâhilindeki en küçük ağaç olan “1SH” lık ağacın seçilmesini önerir. Onların “1SH” kuralı için olan iddiaları şundan kaynaklanmaktadır simülasyon çalışmalarında tekrarlamalar karşısında kararlı ağaç boyutu elde edilmesine karşın “0 SH” ağaç büyüklüğü tekrarlamalar karşısında oldukça değişebilir.

düğümün derinliği doğrultusunda yanlılık artmaktadır. Yanlılığı azaltmak için Laplace ayarı son ML literatüründe önerilmektedir (Provost ve Domingos, 2002). CART monografı (yazılımı) terminal düğüm tahminlerini ayarlamak için literatürde çok nadiren tartışılan biraz daha karmaşık metot sunuyor. “Breiman düzeltmesi” lakaplı, herhangi terminal düğümün tahmini yanlış sınıflandırma oranı olan $r \times (t)$ yukarı doğru

$$r \times (t) = r(t) + e/(q(t) + S) \quad (2.29)$$

ile ayarlar, burada $r(t)$ düğüm içinde eğitime örnekleme tahminidir, $q(t)$ ise düğümdeki eğite örnekleme fraksiyonudur, S ve e ise verilen bir ağaç için eğitime ve test hata oranları arasındaki farkın bir fonksiyonu olarak çözülen parametrelerdir. Laplace metodunun aksine, Breiman düzeltmesi düğümdeki ham tahmin olasılığına bağlı değildir ve eğer test verisi ağacın “aşırı uyum” olmadığını gösterirse düzeltme çok küçük olabilir. Block, Olshen, ve Walker (2002) bu konuyu detayları ile ele alırlar ve bir dizi deneysel çalışmalarda Breiman düzeltmesi için çok iyi performans değerlendirmesi yaparlar.

2.12.5.10 Teorik Temelleri

Karar ağaçlarının ilk çalışmaları tamamen teorik temellere dayanmayan ateorikti. Ağaçlar faydalı görünen metotlar olarak önerildi ve özellikleri ile ilgili sonuçlar deneysel (ampirik) örneklerdeki ağaç performansının gözlemlerine dayanırdı. Bu yaklaşım makine öğreniminde popüler kalırken, alandaki son eğilim çok daha kuvvetli teorik temellere ulaşmaktan taraftır. CART algoritması, teori problemini, kapsamlılık, önemli teknik anlayışlar ve anahtar sonuçlar için ispatlar sunarak çözmektedir. Mesela, araştırmacılar maksimal (mümkün en büyük) ağaç için beklenen yanlış sınıflandırma oranını elde ederler, Bayes oranın iki katı kadar yukarıdan sınırlı olduğunu gösterirler. Araştırmacılar ayrıca, ağaçlardaki “yanlılık/varyans” ödünleşimini (trade-off) ele alırlar ve yanlılığın özniteliklerin sayısından nasıl etkilendiğini gösterirler. Geniş olarak CART

yazarlarından Richard Olshen ve Charles Stone önceki çalışmalarına dayanılarak, monograf'ın son üç bölümü CART'ı “en yakın komşular” üzerinde teorik çalışma ile ilişkilendirir ve örneklem büyüklüğü sonsuza eğilim gösterirken aşağıdakilerin geçerli olduğunu gösterir:

- 1) Regresyon fonksiyonunun hesaplamaları gerçek fonksiyona yaklaşır
- 2) Terminal düğümlerin riskleri tekabül eden Bayes kurallarının risklerine yaklaşır

Bir başka deyişle, yeterince büyük örnekler ile CART ağacı hedef değişkenini yordayıcı (prediktör) değişkenleri ile ilişkilendiren gerçek fonksiyona yaklaşacaktır ve mümkün olan en küçük maliyet değerini elde edecektir (Bayes Oranı). Pratik olarak konuşacak olursak, bu tür sonuçlar günümüzde yaygın kullanılan örnek büyüklüğünden çok daha büyük örnek büyüklüğü ile gerçekleştirilir.

ÜÇÜNCÜ BÖLÜM

DESTEK VEKTÖR MAKİNELERİ (DVM)

Destek Vektör Makineleri (DVM), Destek Vektör Sınıflandırıcı (DVS) ve Destek Vektör Regresör (DVR) den oluşur, bilinen tüm veri madenciliği algoritmaları içinde en dayanıklı ve en doğru metotlardandır¹. DVM'ler Orijinal olarak Vapnik tarafından 1990larda geliştirildi (Vapnik, V.N, 1995; Vapnik, V.N., 1998), istatistik öğrenme teorisinde köklü sağlam bir teorik temele sahip, öğrenme için bir düzüne kadar az örnek gerektirir ve çoğu zaman boyutların sayılarına duyarsızdırlar. Geçen on yılda, hem pratik hem de teori alanlarında DVM ler hızlı bir şekilde geliştirildi.

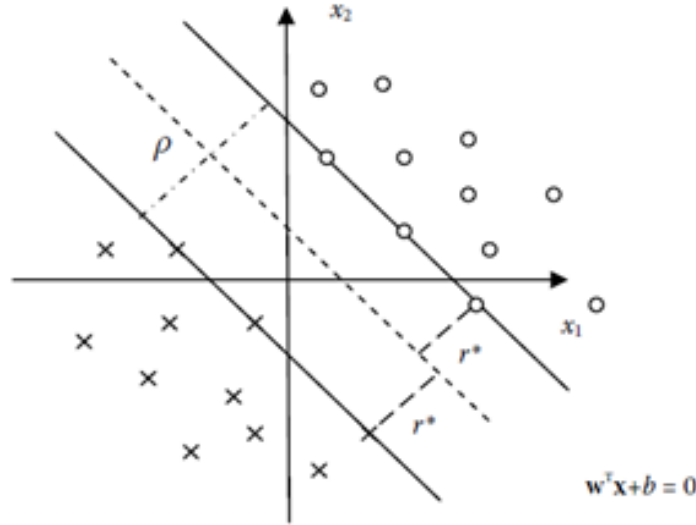
Destek vektör makinesi kökeni yapısal risk minimizasyon prensibine dayanmaktadır, ayrıca yapısal risk minimizasyon VC-boyut (Vapnik Chernonveki Dimension) içerisinde yer aldığından destek vektör makinesini daha da önemli yapmaktadır. Adından anlaşıldığı gibi, makine tasarımı destek vektörler olarak hizmet veren eğitme verisinin bir alt kümesinin çıkarımına dayanmaktadır ve bundan dolayı verinin bir istikrarlı özelliğini temsil eder. Destek vektör makinesi çok terimli (polinom) makine öğrenmesi, radyal temelli fonksiyon ağ ve iki katmanlı algılayıcı (two-layer perceptron) fonksiyonlarının içermektedirler. Böylece, bu metotlar eğitme verisinde bulunan yapısal istatistik düzenliliğin farklı temsillerini sağlasalar bile, bunların hepsi bir destek vektör makinesini çerçevesindeki ortak bir kökten gelmektedirler².

¹ C. Cortes, V.N. Vapnik, "Support Vector Networks", **Machine Learning**, vol.20, , 1995, pp.273-297.

² S. Haykin, **Neural Networks: A Comprehensive Foundation**, Tsinghua University Press, 2001, p.344.

3.1 Destek Vektör Sınıflandırıcı (DVS)

Doğrusal olarak ayrılabilen iki - sınıf öğrenme görevi için, DVC nin amacı verilen örneklerin iki sınıfın maksimal bir marj ile ayırabilen bir hiperdüzlem bulmaktır, bu tür hiperdüzlemin en iyi genelleştirme yeteneğini sunabildiği ispatlanmıştır. Genelleştirme yeteneği bir sınıflandırıcının eğitme verisinde çok iyi sınıflandırma performansı olduğu gibi, aynı zamanda eğitme verisi gibi aynı dağılımdan gelecek verisi için yüksek tahmin doğruluğunu garanti eder. Genelleştirme yeteneği bir sınıflayıcının eğitme verisinde çok iyi sınıflayıcı performansına sahip olduğunu(örneğin doğruluk) ifade etmesinin yanı sıra, aynı zamanda eğitme verisi gibi aynı dağılımdan gelen gelecek verisi için yüksek tahmin doğruluğunu garanti edeceğini ifade eder.



Şekil 13. DVS de Doğrusal olarak ayrılabilen bir durum için Optimal hiperdüzlem gösterimi

Bir marjın (ayırım, boşluk, ayırım, mesafe) bir hiperdüzlem tarafından tanımlanan iki sınıf arasındaki mesafe miktarı ya da olarak tanımlanır. Geometrik olarak marjın en yakın veri noktaları ile hiperdüzlem üzerindeki herhangi bir nokta arasındaki

en kısa mesafeye karşılık gelir. Şekil 14 bir iki-boyutlu girdi uzayı için yukarıdaki şartlar altında ilgili optimal hiperdüzlemin bir geometrik yapısını gösteriyor.

Kabul edelim ki w ve b sırası ile hiperdüzlemde ağırlık vektörü ve yanlı(bias) olsun. İlgili hiperdüzlem şu şekilde tanımlanabilir:

$$w^T x + b = 0 \quad (3.1)$$

örnek X ten optimal hiperdüzleme olan istenilen yönde geometrik uzaklık^{3,4}

$$r = \frac{g(x)}{\|w\|}, \quad (3.2)$$

burada $g(x) = w^T x + b$ hiperdüzlem tarafından tanımlanan diskriminant fonksiyonudur (Christianini, N, Shawe-Taylor, J., 2000) ve ayrıca, w ve x verilmek üzere, x 'in fonksiyonel marjını ile adlandırılır,

Sonuç olarak, DVS optimal hiperdüzlem için ayırma marjı (ρ) maksimize etmek için w ve b parametrelerinin bulunmasını hedefler, bu değer sırası ile iki sınıftan en kısa geometrik uzaklıklar olan r^* yolu ile bulunur, böylece DVS “maksimal marj sınıflandırıcı” olarak ta adlandırılır. Genellik kaybı olmaksızın, fonksiyonel marjı 1’ e eşit olacak şekilde düzeltilir⁵; yani, verilen bir eğitime kümesi $\{x_i, y_i\}_{i=1}^n \in \mathbf{R}^m \times \{\pm 1\}$ olmak üzere, aşağıdaki denklemler elde edilir

$$\begin{aligned} w^T x_i + b &\geq 1, \quad y_i = +1 \\ w^T x_i + b &\leq -1, \quad y_i = -1 \end{aligned} \quad (3.3)$$

³ R.O. Duda, P.E. Hart, D.G. Stork, **Pattern Classification**, Wiley, 2001, p.49 .

⁴ S. Haykin, **Neural Networks: A Comprehensive Foundation**, Tsinghua University Press, 2001, p.321.

⁵ Nello Christianini, John Shawe-Taylor, **An Introduction to Support Vector Machines and Other Kernel -Based Learning Methods**, Cambridge University Press, 2000, p.94.

Denklem (3) teki birinci veya ikinci parçalardaki eşitlikleri sağlayan özel veri noktaları $(\mathbf{x}_i, y_i)$ destek vektörler olarak adlandırılır, bu noktalar tam olarak optimal hiperdüzleme en yakın veri noktalarıdır⁶. Sonra, destek vektör $\mathbf{x}^*$ dan optimal hiperdüzleme karşılık gelen geometrik uzaklık

$$\mathbf{r}^* = \frac{\mathbf{g}(\mathbf{x}^*)}{\|\mathbf{w}\|} = \begin{cases} \frac{1}{\|\mathbf{w}\|} & \text{eğer } y^* = +1 \\ -\frac{1}{\|\mathbf{w}\|} & \text{eğer } y^* = -1 \end{cases} \quad (3.4)$$

Figürden anlaşılacağı üzere, ayırma marjı ρ

$$\rho = 2r^* = \frac{2}{\|\mathbf{w}\|} \quad (3.5)$$

Maksimum marjlı hiperdüzlemi bulunabileceğinden emin olmak için, DVS $\mathbf{w}$ ve b ye göre ρ 'yu maksimize etmeye çalışır:

$$\begin{aligned} & \max_{\mathbf{w}, b} \frac{2}{\|\mathbf{w}\|} \\ & s. t. \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, \dots, n \end{aligned} \quad (3.6)$$

Eşit şekilde,

$$\begin{aligned} & \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \\ & s. t. \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, \dots, n. \end{aligned} \quad (3.7)$$

Burada, sonraki optimizasyon adımlarının kolayca gerçekleştirmek için çoğu zaman $\|\mathbf{w}\|$ yerine $\|\mathbf{w}\|^2$ kullanılır. Genellikle, primal problem olarak bilinen (7) no'lu

⁶ Haykin, **a.g.e.**, p.321.

denklemden kısıtlara sahip optimizasyon problemini Lagrange Çarpanları metodunu kullanarak çözeriz,

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^n \alpha_i [y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1] \quad (3.8)$$

α_i i. eşitsizliğe göre Lagrange çarpanıdır.

$L(\mathbf{w}, b, \alpha)$ nın $\mathbf{w}$ ve b ye göre diferansiyelini alırsak ve sonuçları sıfıra eşitlersek, aşağıdaki iki optimallık koşulunu elde ederiz:

$$\begin{cases} \frac{\partial L(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = 0 \\ \frac{\partial L(\mathbf{w}, b, \alpha)}{\partial b} = 0 \end{cases} \quad (3.9)$$

Daha sonra aşağıdakileri elde ederiz,

$$\begin{cases} \mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{cases} \quad (3.10)$$

Elde edilen (10) no'lu denklemi (8) deki Lagrange fonksiyonunda yerleştirirsek, buna karşılık gelen dual problemi elde ederiz

$$\begin{aligned} \max_{\alpha} W(\alpha) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ s. t. \quad &\sum_{i=1}^n \alpha_i y_i = 0 \\ &\alpha_i \geq 0, i = 1, \dots, n. \end{aligned} \quad (3.11)$$

Ve aynı zamanda, Karush-Kuhn-Tucker tamamlayıcı koşulu

$$\alpha_i[y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1] = 0, \quad i = 1, \dots, n \quad (3.12)$$

Sonuç olarak, sadece optimal hiperdüzleme en yakın veri noktaları ve maksimal marjı belirleyen destek vektörleri olan $(\mathbf{x}_i, y_i)$ sıfır olmayan α_i değerlerine karşılık gelir. Diğer tüm α_i değerleri sıfıra eşit olur.

(3.11) nolu denklemdeki dual problem tipik konveks kuadratik programlama optimizasyonu problemidir. Birçok durumda, sıralı minimal optimizasyon (SMO, Sequential Minimal Optimization Algorithm) gibi bazı uygun optimizasyon tekniklerinin benimsenmesi ile verimli bir şekilde global optimuma yakınsar⁷.

Optimal Lagrange çarpanları olan α_i^* , optimal ağırlık vektörü olan $\mathbf{w}^*$ yı (10) no'lu denklemi yolu ile hesaplayabiliriz

$$\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i \quad (3.13)$$

Daha sonra, pozitif destek vektör $\mathbf{x}_s$ 'ten yararlanarak, karşı gelen optimal yanlı (bias) b^* aşağıdaki şekilde yazılabilir⁸

$$b^* = 1 - \mathbf{w}^{*T} \mathbf{x}_s, \quad y_s = +1 \text{ için} \quad (3.14)$$

3.2 Yumuşak Marjlı DVS ve Optimizasyon

Maksimal marjlı DVS, sonraki konu olan DVR dâhil edilmesiyle beraber, DVM algoritmalarının orijinal başlama noktasını temsil eder. Fakat birçok gerçek hayat problemlerinde, tüm noktaların doğrusal olarak ayrılabilir olma gerekliliği çok katı olabilir, özellikle birçok karmaşık doğrusal olmayan sınıflandırma durumlarında.

⁷ Christianini, Shawe-Taylor, **a.g.e.**, p.96.

⁸ Haykin, **a.g.e.**, p.324.

Örneklerin tamamen doğrusal olarak ayrılabilir olamadığı zaman, marjlar negatif olabilir. Bu durumlarda, primal problemin uygun (çözüm) bölgesi boştur ve buna uygun düşen dual problem ise sınırsız bir amaç fonksiyonudur. Buda en uygun şekle sokma probleminin çözmeği imkânsız kılıyor⁹.

Bu ayrılmaz problemleri çözmek için, genellikle iki yaklaşım benimsenir. Birincisi (3.7) no’lu denklemdeki katı eşitsizlikleri gevşetmek ve böylece sözde soft marj optimizasyon olarak adlandırılan metot elde edilir. İkinci metot ise doğrusal olmayan o problemlere “kernel trick (çekirdek Hilesi)” uygulayarak bunları doğrusal hale getirmek. Bu kısımda ilk olarak yumuşak marj optimizasyon sunulacak. Daha sonra, yumuşak marjı DVS’ ye kıyasla, genellikle DVC (3.7) no’lu denklemdeki optimizasyon probleminden elde edilen sabit (hard) marj DVC olarak adlandırılır

Veride karşı sınıfların karışmış birkaç noktasının olduğu durumları hayal edelim. Bu noktalar maksimum marjlı hiperdüzlem için bile var olan eğitme hatasını temsil eder. “soft marj” fikri DVC algoritmasını genişletmeği hedefler böylece hiperdüzlem birkaç böyle sesli (noisy) verinin var olmasına izin verir. Özellikle, artıran yapay değişken (slack) olan ξ_i sınıflandırıcı tarafından sınıflandırma ihlalinin miktarını açıklamak için sunuldu.

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

$$\text{s.t. } y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, i = 1, \dots, n, \quad (3.15)$$

burada C parametresi makinenin karmaşıklığı ve ayrılmaz noktalarının sayısı arasındaki “ödünleşim” kontrol ediyor. Bu “düzenleştirme” parametresi olarak görülebilir ve kullanıcı tarafından ya deney yolu ile ya da analitik olarak seçilir.

⁹ Christianini, Shawe-Taylor, **a.g.e.**, p.103.

Artıran yapay değişken olan ξ_i nin yanlış sınıflandırılan bir veri örneğinin hiperdüzleme olan uzaklık yolu direkt olarak bir geometrik açıklaması ile vardır. Bu uzaklık, desen ayrılabilirin ideal koşulundan gelen bir örneğin sapmasını hesaplar. Yukarıdaki bölümde sunulan Lagrange çarpanlarının aynı metodunu kullanarak, soft marj'ın dual problemini şu şekilde formüleştirebiliriz:

$$\begin{aligned} \max_{\alpha} \mathbf{W}(\alpha) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{s. t.} \quad &\sum_{i=1}^n \alpha_i y_i = 0 \\ &0 \leq \alpha_i \leq C, \quad i = 1, \dots, n. \end{aligned} \quad (3.16)$$

(11) no'lu denklemi (16) no'lu denklem ile kıyasladığımızda, ξ_i artan yapay değişkeninin dual problemde görünmediği önemlidir. doğrusal olarak ayrılamayan ve ayrılabilen durumlar arasındaki önemli fark $\alpha_i \geq 0$ kısıt'ı daha sıkı bir kısıt olan $0 \leq \alpha_i \leq C$ kısıt'ı ile yer değiştirilmesinden kaynaklanmaktadır. Aksi halde, bu iki durum birbirinin benzeridirler, ağırlık vektörü $\mathbf{w}$ ve yanlı (bias) b 'nin optimal değerlerinin hesaplamaları dahildir, özellikle destek vektörlerin tanımı^{10, 11}.

Ayrılmaz(inseparable) durumunda ki Karush-Kuhn-Tucker tamamlayıcı (tüm)ler) koşulu şu şekilde ifade edilir,

$$\alpha_i [y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i] = 0, \quad i = 1, \dots, n \quad (3.17)$$

$$\gamma_i \xi_i = 0, i = 1, \dots, n, \quad (3.18)$$

¹⁰ Christianini, Shawe-Taylor, **a.g.e.**, p.107.

¹¹ Haykin, **a.g.e.**, p.328.

burada γ_i' ler ξ_i' 'nin pozitifliğini uygulamak için sunulan ξ_i' 'e karşılık gelen Lagrange çarpanlarıdır¹². Eğer noktasında, primal problem için Lagrange fonksiyonunun ξ_i' 'e göre türevi sıfırdır, türevlerin hesaplamalarından aşağıdaki sonuç elde edilir

$$\alpha_i + \gamma_i = C \quad (3.19)$$

(18) ve (19) denklemler birleştirilirse ,

$$\xi_i = 0 \text{ eğer } \alpha_i < C . \quad (3.20)$$

Sonuç olarak, optimal ağırlık $\mathbf{w}^*$ aşağıdaki gibi olur

$$\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i \quad (3.21)$$

Optimal sapma(bias) değeri b^* eğitime kümesinde $0 \leq \alpha_i^* \leq C$ ve $\xi_i = 0$ olacak şekilde herhangi bir veri noktası olan $(\mathbf{x}_i, y_i)$ alarak ve (3.17) no'lu denklemde veri noktasını kullanarak elde edilebilir¹³.

3.3 Çekirdek hilesi (Kernel trick)

Çekirdek hilesi doğrusal olarak ayırlamayan problemleri çözmek için çoğunlukla kullanılan diğer bir tekniktir. Burada sorun, verilen veriler arasındaki iç çarpıma dayalı uygun bir çekirdek fonksiyonu tanımlamak, problemleri doğrusal olarak ayrılabilir yapmak için giriş uzayından (input space) bir yüksek boyutlu (sonsuz da olabilir) öznitelik uzayına(feature space) verinin doğrusal olmayan dönüşümüdür. Bunun altında yatan gerekçe kalıpların (patern) ayrılabilirliği üzerinde Cover Teoreminde bulunabilir; yani, karmaşık kalıba (patern) sahip bir sınıflandırma problemi yüksek boyutlu bir

¹² Haykin, **a.g.e.**, p.329.

¹³ Haykin, **a.g.e.**, p.329.

uzayda doğrusal olmayacak bir şekilde oluşmuş ise düşük boyutlu uzaydakine kıyasla büyük olasılıkla doğrusal olarak ayrılabilir¹⁴

Kabul edelim ki $\Phi : \mathbf{X} \rightarrow \mathbf{H}$ giriş uzayından $\mathbf{X} \subset \mathbf{R}^m$ öznitelik uzayına $\mathbf{H}$ doğrusal olmayan dönüşümü gösterebilir, problem $\mathbf{H}$ 'de doğrusal olarak ayrılabilir durumda olabilir. Buna karşılık gelen optimal hiperdüzlemi aşağıdaki gibi tanımlayabiliriz:

$$\mathbf{w}^{\Phi_T} \Phi(\mathbf{x}) + b = 0 \quad (3.22)$$

Genelliği bozmaksızın, sapma (bias) değerini $b=0$ eşitlersek ve (22) no'lu denklemi düzenlersek

$$\mathbf{w}^{\Phi_T} \Phi(\mathbf{x}) = 0 \quad (3.23)$$

Doğrusal ayrılabilir durumlara benzer olarak, benzer Lagrange çarpanları metodunu yardımı ile optimal ağırlık vektörü olan $\mathbf{w}^{\Phi*}$ öznitelik uzayında arıyoruz, ve aşağıdaki denklemi elde ederiz:

$$\mathbf{w}^{\Phi*} = \sum_{i=1}^n \alpha_i * y_i \Phi(\mathbf{x}_i) \quad (3.24)$$

Böylece, öznitelik uzayında hesaplanan optimal hiperdüzlem:

$$\sum_{i=1}^n \alpha_i * y_i \mathbf{w}^{\Phi_T}(\mathbf{x}_i) \Phi(\mathbf{x}) = \mathbf{0} \quad (3.25)$$

$\mathbf{w}^{\Phi_T}(\mathbf{x}_i) \Phi(\mathbf{x})$ terimi $\Phi(\mathbf{x}_i)$ ve $\Phi(\mathbf{x})$ vektörlerinin iç çarpımlarını gösterir. Bundan dolayı, burada iç çarpım çekirdek fonksiyonunu çıkarırız.

¹⁴ Haykin, a.g.e., p.330.

3.3.1 İç çarpım Çekirdek¹⁵

Çekirdek $\mathbf{K} = (\mathbf{x}, \mathbf{x})$ gibi bir fonksiyondur, tüm $\mathbf{x}, \mathbf{x} \in \mathbb{R}^m$ için

$$\mathbf{K} = (\mathbf{x}, \mathbf{x}) = \Phi^T(\mathbf{x})\Phi(\mathbf{x}) \text{ sağlanır.} \quad (3.26)$$

Burada Φ , giriş uzayı (input space) olan $\mathbf{X}$ 'dan öznitelik uzayı (feature space) olan $\mathbf{H}$ 'ye olan dönüşümdür.

Optimal öznitelik uzayında hiperdüzlem oluşturmak için Φ dönüşümünün somutlaştırılmış formunu dikkate almaksızın çekirdeği kullanabiliriz, bundan dolayı önemlidir. Ayrıca, Φ yüksek (sonsuz dâhil) boyutlu öznitelik uzayında açık bir şekilde formüle edilmemelidir. Sonuç olarak, çekirdeğin uygulanması algoritmayı boyuta duyarsız yapabilir, doğrusal olarak ayrılmaz problemleri etkin olarak çözmek ve yüksek boyutlu bir uzayda doğrusal bir sınıflandırıcıyı eğitmek amacıyla. Buda (25) no'lu denklemde $\Phi^T(\mathbf{x}_i)\Phi(\mathbf{x})$ yerine $\mathbf{K}(\mathbf{x}_i, \mathbf{x})$ kullanarak yapılabilir, sonra optimal hiperdüzlem şu şekilde elde olur:

$$\sum_{i=1}^n \alpha_i^* y_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}) = 0 \quad (3.27)$$

Belirtildiği gibi, çekirdek hilesi hesaplamayı basitleştirmek için cazip bir yöntemdir, bu sayede karmaşık öznitelik uzayını direkt olarak hesaplamaktan sadece iç çarpımların hesaplamasında olmadığı gibi aynı zamanda sınıflandırıcının tasarlanmasından da kaçınılır.

Fakat çekirdek hilesini uygulamadan önce, çekirdek fonksiyonunu nasıl oluşturacağımızı dikkate almalıyız, yani, bir çekirdek fonksiyonu hangi karakteristikleri sağlamalıdır. Bu soruyu cevaplamak için, öncelikle gerçek bir çekirdek fonksiyonu

¹⁵ Christianini, Shawe-Taylor, **a.g.e.**, pp. 30-33.

olarak kabul edildiği zaman $\mathbf{K} = (\mathbf{x}, \mathbf{x}')$ fonksiyonunun özelliğini karakterize eden Mercer teoremini bilmemiz gerekir

3.4 Mercer Teoremi¹⁶

Kabul edelim ki $\mathbf{K} = (\mathbf{x}, \mathbf{x}')$, $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$ (aynı biçimde $\mathbf{x}'$ için) kapalı aralığında tanımlı bir sürekli simetrik çekirdek olsun. Çekirdek $\mathbf{K} = (\mathbf{x}, \mathbf{x}')$ aşağıdaki seride genişletilebilir

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \varphi_i(\mathbf{x}) \varphi_i(\mathbf{x}'), \quad (3.28)$$

tüm i 'ler için $\lambda_i > 0$. bu genişlemenin geçerli olması ve yakınsaması için, aşağıdaki şart gerekli ve yeterlidir:

$$\int_b^a \int_b^a \mathbf{K}(\mathbf{x}, \mathbf{x}') \psi(\mathbf{x}) \psi(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0 \quad (3.29)$$

$$\int_b^a \psi^2(\mathbf{x}) d\mathbf{x} < \infty, \quad (3.30)$$

tüm $\psi(\cdot)$ için geçerlidir.

Teoremin ışığında, çekirdeğin oluşumunda, en faydalı karakteristiği özetleyebiliriz, bu Mercer çekirdeği olarak adlandırılır. Yani, giriş uzayı $\mathbf{X}$ ait olan herhangi rastgele sınırlı sayıda altkümeleri için, çekirdek fonksiyonu $\mathbf{K} = (\mathbf{x}, \mathbf{x}')$ tarafından oluşturulan ilgili matris olan

$$\mathbf{K} = (\mathbf{K}(\mathbf{x}_i, \mathbf{x}'_j))_{i,j=1}^n \quad (3.31)$$

Simetrik ve yarı kesin (semi-definite) matristir, buna Gram Matrisi adı verilir¹⁷

¹⁶ Haykin, a.g.e, p.331.

Bu koşul altında, pratikte bir çekirdek fonksiyonun nasıl seçilmesi konusunda hala biraz özgürlük vardır. Mesele, doğrusal çekirdek fonksiyonlarının yansırı, polinom ya da radyal temelli çekirdek fonksiyonları tanımlayabiliriz. Son zamanlarda DVC için ve birçok diğer istatistik testler için birçok çalışma farklı çekirdek araştırmaları alanlarında yapıldı.

Daha önce doğrusal ayrılmaz problemlerin çözümünde yumuşak marj DVS belirtilmişti. Çekirdek hilesi ile kıyaslandığında, iki yaklaşımda gerçekte problemleri farklı şekillerde çözmektedirler. Yumuşak marj orjinal giriş uzayında kısıtları gevşetir ve bazı hataların çıkmasına izin verir. Fakat, problem ağır derecede doğrusal olarak ayrılmaz olduğu zaman ve sınıflandırılmayan hatalar çok yüksek olduğu zaman, yumuşak marj kullanışsızdır. Çekirdek hilesi ayrılmaz problemleri ayrılabilir yapmak için veriyi çekirdek fonksiyon yardımı ile tam olarak yüksek boyutlu öznitelik uzayına eşleştirir. Fakat aslında problemin karmaşıklığından dolayı çekirdek hilesi her zaman problemin kesinlikle doğrusal olarak ayrılabilir olabileceğini garanti edemez. Bundan dolayı, pratikte çoğu zaman iki tekniğin farklı avantajlarını ortaya çıkarmak için bunları birleştirilir ve doğrusal olarak ayrılmaz problemleri daha etkin şekilde çözülür. Sonuç olarak, çekirdek yumuşak marj DVC için kısıtlara sahip optimizasyon problemine karşılık gelen dual form aşağıdaki gibidir:

$$\begin{aligned} \max_{\alpha} W(\alpha) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) \\ s. t. \quad &\sum_{i=1}^n \alpha_i y_i = 0 \\ &0 \leq \alpha_i \leq C, \quad i = 1, \dots, n. \end{aligned} \quad (3.32)$$

Benzer Lagrange çarpanları metodunu uygulanırsa, optimal sınıflandırıcı elde edilir:

¹⁷ Christianini,N., Shawe-Taylor, **a.g.e.**, pp. 118-120.

$$f(x) = \sum_{i=1}^n \alpha_i^* y_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}) + b^* \quad (3.33)$$

burada, pozitif destek vektör $y_s = +1$ için,

$$b^* = 1 - \sum_{i=1}^n \alpha_i^* y_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}_s) \text{ 'dır.}$$

3.5 Teorik Altyapısı

Yukarıdaki bölümlerde, DVC algoritmasını hem doğrusal olarak ayrılabilir durum ve hem de ayrılmaz durumlar için açıklamaya çalıştım. Çekirdek hilesi sınıflandırıcının ifade performansını daha çok geliştirir, öyle ki asıl doğrusal özelliği yüksel boyutlu öznitelik uzayında tutabilir ve boyut belasından kaçınabilir. Bu bölümde, DVC nin teorik temelleri üzerinde duracağım. Vapnic Chervonenkis (VC) teorisinin ışığı altında^{18, 19}, ilk önce sınıflandırıcının karmaşıklığının global şekilde nasıl kontrol edileceğinin konusunda kılavuzluk eden bir doğrusal sınıflandırıcının genel hata sınırı anlatılacaktır. Daha sonra algoritmanın iyi genelleme kapasitesini garanti etmek için DVC' de maksimum marjın önemini açıklamak için DVS'nın bir somut genelleme sınırı sonucu çıkarılacaktır.

VC teorisi, istatistik öğrenmede Olası Yaklaşık olarak Düzeltme (Probable Approximately Correct, PAC) öğrenme modelini geliştirir ve direkt olarak Destek Vektör Makinelerinin önerilmesine önderlik eder. VC teorisi VC-boyut^{20, 21} olarak bilinen yeni bir karmaşıklık ölçütü tanımlayarak genelleştirme hatasının tahmininde kullanılabilmesi için analitik genelleştirme sınırı sağlar.

¹⁸ O. Chapelle, P. Haffner, Vladimir N. Vapnik, "Support Vector machines for histogram-based image classification", **IEEE Trans. on Neural Networks**, Vol.10(3.5), 1999, pp.1055-1064.

¹⁹ C. Cortes, V.N. Vapnik, "Support Vector Networks", **Machine Learning**, vol.20, , 1995, pp.273-297.

²⁰ V.Cherkassky, F. Mulier, Learning From Data: Concepts, Theory and Methods, Wiley, 1998, p.107.

²¹ V. Cherkassky, X. Shao, F. Mulier, V. Vapnik, , "Model Complexity control for regression using VC generalization bounds", IEEE Transactions on Neural Networks, vol.10, 1999, pp. 1075-1089.

Aşık bir şekilde, farz edelim ki eğitme ve test verileri sabit ve fakat $\mathbf{D}$ gibi bilinmeyen olasılık dağılımına göre üretilmiş olsun, bir sınıflandırma fonksiyonu olan h nin $\mathbf{D}$ 'deki hatası $err_{\mathbf{D}}(h)$ aşağıdaki gibi tanımlansın

$$err_{\mathbf{D}}^{(h)} = \mathbf{D}\{(x, y): h(x) \neq y\} \quad (3.34)$$

beklenen hatayı hesaplar²²

PAC modelleri genelleme rastgele değişken $err_{\mathbf{D}}(h_s)$ nin dağılımını sınırlandırır ve ilgili PAC sınırı $\varepsilon = \varepsilon(n, H, \delta)$ formuna sahiptir; yani, PAC modeli, h_s hipotezinde, S eğitme verisinde hatanın olasılığı aşağıdaki ilişkiyi sağlar²³

$$\mathbf{D}^n\{S: err_{\mathbf{D}}(h_s) > \varepsilon(n, H, \delta)\} < \delta \quad (3.35)$$

Eğer $|H|$ hipotezleri S kümelerinde büyük hatalara sahip iseler, o zaman PAC sınırı aşağıdaki gibi olur

$$\varepsilon = \varepsilon(n, H, \delta) = \frac{1}{n} \ln \frac{|H|}{\delta} \quad (3.36)$$

PAC sınır, fonksiyon sınıfı H direkt olarak hata sınırını etkileyebileceğini sunar. VC teorisi PAC sınırını daha fazla sınırsız fonksiyon sınıfına genelleştirir ve VC boyut olan d gibi bir kavram tanıtır. VC boyut d , fonksiyon sınıfın mükemmel şekilde öğrenmesi için kullanılabildiği eğitme verisinin maksimum sayısını ölçer, eğitme verisinde sıfır hata oranlarını elde ederek, herhangi bir sınıf etiketlerinin bu noktalara atanması için. Daha sonra doğrusal sınıflandırıcının geliştirilmiş PAC sınırı aşağıdaki gibi tanımlanabilir:

²² Christianini, Shawe-Taylor, a.g.e., p.52.

²³ A.e, pp. 52-53.

3.5.1 Teorem (Vapnik ve Chervonenkins)²⁴

Kabul edelim ki H , VC boyutu d olan bir hipotez uzayını belirtsin. $\mathbf{X} \times \{-1,1\}$ de $1-\delta$ olasılıklı rastgele $\mathbf{D}$ gibi bir olasılık dağılımı için, eğitime kümesi S 'de rastgele hipotez $h \in H$ nin genelleştirme hatası

$$d \leq n, n > \frac{2}{\varepsilon} \text{ koşulu altında}$$

$$err_{\mathbf{D}}(h) \leq \varepsilon(n, H, \delta) = \frac{2}{n} \left(\log \frac{2}{\delta} + d \log \frac{2en}{d} \right) \quad (3.37)$$

Teoremin ışığı altında, (37) nolu denklemdeki ilk terim eğitime hatasıdır ve ikinci terim VC boyut d ye orantılıdır. Böylece, teorem gösteriyor ki, hipotez “ h ” ampirik risk hatasını küçük bir derecede kontrol ettiği müddetçe, eğer d 'yi minimize edebilirsek, gelecekteki hatayı minimize edebiliriz.

Yukarıdaki teorem bir doğrusal sınıflandırıcının genel hata sınırını sağlar ve sınıflandırıcının karmaşıklığını nasıl kontrol etmesi hususunda genele rehberlik eder. Aşağıda, DVS için sınır genelleştireceğiz ve DVS nin karşılık gelen genelleştirilmiş hata sınırı sonucunu elde edeceğiz.

3.5.2 Marj Tanımı²⁵

Girdi uzayı olan $\mathbf{X}$ i sınıflandırmak için, reel değerli fonksiyon sınıfı olan F 'i kullanmayı dikkate alalım, ve eşik(sınır) değeri 0 dır. $(x_i, y_i) \in \mathbf{X} \times \{-1,1\}$ olan örneğin, fonksiyon ya da hiperdüzlem olan $f \in F$ e olan marjı aşağıdaki gibi tanımlanır:

$$\gamma_i = y_i f(x_i) \quad (3.38)$$

²⁴ Christianini, Shawe-Taylor, **a.g.e.**, pp. 54-56.

²⁵ Christianini, Shawe-Taylor, **a.g.e.**, p. 59.

$\gamma_i > 0$ olması örnek (x_i, y_i) in doğru bir şekilde sınıflandırıldığını belirtir. S eğitme kümesine karşılık gelen f 'in marjinal dağılımı S deki örneklerin marjinal dağılımıdır. Marjinal dağılımın minimumu S eğitme kümesine karşılık gelen f 'in $m_s(f)$ olarak adlandırılır.

VC boyut d teorik olarak anlamlı olmasına karşın, pratikte d bazen sınırsız ve böylece genelleme sınırı birçok gerçek probleme uygulanamaz. Sonuç olarak, geleneksel VC boyut yerine DVS de marj ile alakalı benzer bir ölçüt sunulur:

3.5.3 Fonksiyon Sınıfın Örtüsü²⁶

Kabul edelim ki $F, \mathbf{X}'$ de reel değerli bir fonksiyon olsun. Girdi verilerinin bir serisi için

$$S = \{x_1, x_2, \dots, x_{1n}\}$$

F 'in γ -örtüsü sınırlı fonksiyon olan B kümesidir, öyle ki tüm $f \in F$ için, $g \in B$ mevcuttur,

$$\max_{1 \leq i \leq n} (|f(x_i) - g(x_i)|) < \gamma. N(F, S, \gamma)$$

Örtünün minimal büyüklüğünü belirtir. F 'in kapladığı(örtü) veri sayısı

$$N(F, S, \gamma) = \max_{S \in X^n} N(F, S, \gamma) \quad (3.39)$$

Daha sonra $N(F, S, \gamma)$ kullanarak, eğitme kümesi S de $m_s(f) = \gamma$ olan f hipotezi için Teorem (Vapnik ve Chervonenkins)'ni yeniden formüle edilir.

²⁶ Christianini, Shawe-Taylor, **a.g.e.**, p.59.

3.5.4 Marjlı VC Teoremi²⁷

Sınırlı reel değerli fonksiyon uzayı F olsun ve sabit $\gamma \in \mathbf{R}^+$. $\mathbf{X} \times \{-1,1\}$ de herhangi bir olasılık dağılımı $\mathbf{D}$ için, $1-\delta$ olasılıklı , $m_s(f) \geq \gamma$ marjına sahip eğitme kümesi S 'de bir hipotez $f \in F$ nin genelleme hatası aşağıda verilen denklemi $n > 2/\varepsilon$ şartı altında sağlar

$$err_D(f) \leq \varepsilon(n, F, \delta, \gamma) = \frac{2}{n} \left(\log \frac{2}{\delta} + \log N(F, 2n, \gamma/2) \right) \quad (3.40)$$

Yukarıdaki teorem eğitme verilerinden elde edilebilen genelleme hatasını sınırlandırmak için, $m_s(f)$ nin nasıl kullanılacağını gösterir. $N(F, 2n, \gamma/2)$ VC boyutun başka bir biçimi olarak görülebilir, burada çok büyük bir γ değeri çok küçük $N(F, 2n, \gamma/2)$ değerine karşılık gelir. Sonuç olarak, küçük boyutlu örnekler için büyük marj çok iyi sınıflandırıcının genelleme performansını ifade edeceği sonucu çıkarılabilir.

Marjlı VC teoremi VC teoreminin genelleştirilmiş bir biçimi olmasına rağmen, $N(F, 2n, \gamma/2)$ değeri gerçek dünya problemlerinde etkin bir şekilde sayısallaştırılamaz. Sonuç olarak, spesifik DVS algoritması için daha fazla somut hata sınırı sonucu elde edilir.

3.5.5. DVS'nin Genelleme Sınırı²⁸

Teorem: Kabul edelim ki girdi uzayı $\mathbf{X}$ yarıçapı R olan iç çarpım H uzayında bir hipertop (hyperball) olsun, $\mathbf{X} = \{\mathbf{x} \in H: \|\mathbf{x}\|_H \leq R\}$.

Fonksiyon sınıfı Ψ göz önüne alalım:

$$\Psi = \{\mathbf{x} \mapsto \mathbf{w}^T \mathbf{x}: \|\mathbf{x}\|_H \leq 1, \mathbf{x} \in \mathbf{X}\}$$

²⁷ Christianini, Shawe-Taylor, a.g.e., p. 60.

²⁸ A.e., p.63.

sabit $\gamma \in \mathbf{R}^+$. $\mathbf{X} \times \{-1,1\}$ de olasılık dağılımı $\mathbf{D}$ için, $1-\delta$ olasılıklı, $m_s(f) \geq \gamma$ marjına sahip eğitime kümesi S 'de $f \in \Psi$ hipotezinin genelleme hatası aşağıda verilen denklemi $n > 2/\varepsilon$, $64R^2/\gamma^2 < n$ şartı altında sağlar

$$err_D(f) \leq \varepsilon(n, \psi, \delta, \gamma) = \frac{2}{n} \left(\log \frac{4}{\delta} + \frac{64R^2}{\gamma^2} \log \frac{en\gamma}{4R} \log \frac{128nR^2}{\gamma^2} \right) \quad (3.41)$$

Girdi uzayının boyutu sınırdan görünmediği dikkat edilmesi gereken bir husustur. Bundan dolayı sınır sonsuz boyut uzayda kullanılabilir, bu da sınırın boyut probleminin üstesinden gelebileceğini belirtir. Ayrıca, örnekler iyi dağıldığı zaman, sınır rassal test örnekleri için küçük hata olacağını yüksek bir olasılıkla garanti edebilir. Bu durumda, marj γ örnek dağılımının kalitesi hakkında bir ölçüm olarak görülebilir ve böylece DVS algoritmasının genelleme performansını ölçebilir²⁹

3.6 Destek Vektör Regressör (DVR)

Şu noktaya kadar, DVS metodunun sınıflandırma kısmına odaklandık. Bu kısımda, Destek Vektör Makinelerini kullanarak doğrusal olmayan regresyon problemlerini çözümü dikkate alınacak, ve bu DVR olarak adlandırılır. Sınıflandırma algoritmasına benzer olarak, doğrusal öğrenme metotlarını ve çekirdek hilesi (kernel trick) kullanılarak elde edilen, doğrusal olmayan fonksiyonlardan yararlanarak maksimum marj metodunun temel özelliklerinin araştırılması amaçlanmaktadır. Ayrıca, uygun algoritmalar büyük boyutlar altında etkin olmalı³⁰.

Fakat regresyon problemleri için, toplanır gürültünün (additive noise) ardındaki dağılım uzun bir kuyruğa sahip olduğu zaman regresörde meydana gelen kötü performansa sahip olan aykırı değerlerin mevcudiyetinde, geleneksel en küçük kareler

²⁹ Christianini, Shawe-Taylor, **a.g.e.**, p.63.

³⁰ Christianini, Shawe-Taylor, **a.g.e.**, p.112.

tahmin edicisi (least-square estimator) uygulanabilir olmayabilir³¹. Bundan dolayı, modelde küçük değişikliklere duyarsız olan sağlam tahmin edici (robust estimator) geliştirilmesi gerekir; yani , sözde ε – duyarsız kayıp fonksiyonu arıyoruz.

Tanım (ε – duyarsız kayıp fonksiyonu)³²

Kabul edelim ki f 'i $\mathbf{X}$ 'de reel değerli bir fonksiyon olsun. ε – duyarsız kayıp fonksiyonu $L^\varepsilon(\mathbf{x}, y, f)$ aşağıdaki gibi tanımlanır.

$$L^\varepsilon(\mathbf{x}, y, f) = |y - f(\mathbf{x})|_\varepsilon = \max(0, |y - f(\mathbf{x})| - \varepsilon) \quad (3.42)$$

Eğer arzu edilen cevap y den olan tahmin edicinin çıktısı $f(\mathbf{x})$ hakkındaki sapmanın mutlak değeri ε 'dan küçük veya sıfıra eşit ise $L^\varepsilon(\mathbf{x}, y, f)=0$. Aksi takdirde, sapmanın mutlak değerinden ε değerinin çıkarılmasına eşittir.

Doğrusal olmayan regresyon modelini düşünelim

$$y = g(x) + v \quad (3.43)$$

Burada toplanır gürültü (additive noise) terimi olan v istatistiksel olarak çıktı vektör olan $\mathbf{x}$ dan bağımsızdır. Fonksiyon $g(.)$ ve gürültü v 'nin istatistiği bilinmiyor. Elimizde mevcut olan sadece eğitme verilerinin bir kümesi olan

$$S = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

ve bir fonksiyon sınıfı olan

$$F = \{f(x) = \mathbf{w}^T \mathbf{x} + b, \mathbf{w} \in \mathbf{R}^m, b \in \mathbf{R}\} \text{ vardır.}$$

³¹ Haykin, **a.g.e.**, p.340.

³² Christianini, Shawe-Taylor, **a.g.e.**, p.114.

Amaç uygun $\mathbf{w}$ ve b parametrelerini seçerek $f(x)$ bilinmeyen hedef fonksiyon $g(x)$ 'e yaklaştırmış olur. Primal problem aşağıdaki gibi temsil edilir:

$$\begin{aligned}
& \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \hat{\xi}_i) \\
& s. t. (\mathbf{w}^T \mathbf{x}_i + b) - y_i \leq \varepsilon + \xi_i, \quad i = 1, \dots, n \\
& y_i - (\mathbf{w}^T \mathbf{x}_i + b) \leq \varepsilon + \hat{\xi}_i, \quad i = 1, \dots, n \\
& \xi_i, \hat{\xi}_i \geq 0 \quad i = 1, \dots, n
\end{aligned} \tag{3.44}$$

Lagrange çarpanlarının benzer metodunu kullanarak, dual problem aşağıdaki gibi elde edilir:

$$\begin{aligned}
\max_{\alpha, \hat{\alpha}} W(\alpha, \hat{\alpha}) &= \sum_{i=1}^n y_i (\hat{\alpha}_i - \alpha_i) - \varepsilon \sum_{i=1}^n (\hat{\alpha}_i + \alpha_i) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\hat{\alpha}_i - \alpha_i) (\hat{\alpha}_j - \alpha_j) \mathbf{x}_i^T \mathbf{x}_j \\
& s. t. \sum_{i=1}^n (\hat{\alpha}_i - \alpha_i) = 0 \\
& 0 \leq \alpha_i, \hat{\alpha}_i \leq C, \quad i = 1, \dots, n
\end{aligned} \tag{3.45}$$

Optimizasyon problem olan Denklem (45) de İç çarpım çekirdeği gösterilebilir ve doğrusal olmayan fonksiyonlar çekirdek uzayında doğrusal öğrenme makineleri yolu ile elde edilebilmesi amacıyla regresyon algoritması öznelik uzayına genişletilebilir.

DVS'ye göre kıyaslandığında, DVR ε gibi ilave serbest parametreye sahiptir. İki parametre olan ε ve C yaklaştırma fonksiyonun (approximation function) VC boyutunu yanlı(bias) $b=0$ olduğu zaman kontrol ederler

$$f(x) = \mathbf{w}^T \mathbf{x} = \sum_{i=1}^n (\hat{\alpha}_i - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) \tag{3.46}$$

ε ve C kullanıcı tarafından seçilmelidir ve regresyon için karmaşıklık kontrolünü direkt olarak etkilerler. Daha iyi yaklaşım fonksiyonu elde etmek için ε ve C değerlerinin aynı anda nasıl seçilmesi gerektiği konusu açık araştırma problemidir.

3.7 Yazılım Uygulamaları

Bu alanda LibSVM³³, SVMlight³⁴ ve IBM Clementine v12 DVM uygulamaları alanında en ünlü yazılımlardır.

LibSVM Windows işletme sisteminde kullanılmak üzere derleyici dili (compiler Language) imkânı sağlar, bunun yanı sıra diğer işletim sistemlerinde kolayca geliştirmek, yenilenme yapmak ve uygulamak için C++ ve Java kodlarını da geliştirmiştir. Özellikle, LibSVM, diğer yazılımlara kıyasla, DVM algoritması ile ilgili kısmen daha az ayarlanabilir parametrelere sahiptir ve gerçek uygulama problemlerinin etkili çözümü için çok sayıda varsayılan parametreler sağlar.

SVMlight, C dilinde yazılmış olan diğer bir uygulamadır. Bir optimizasyon tekniği olan “Hızlı iniş yöntemine (steepest feasible descent)” dayalı etkili küme seçme tekniğini ve çekirdek değerlendirmelerinin iki etkili “daralma_shrinking” ve “önbellekleme_caching” hesaba dayalı ilkeleri benimser. SVMlight esasen iki C programlama içerir: “SVM_Learn”, eğitime örneklerinin öğrenimi ve ilgili sınıflandırıcının eğitimi için kullanılır ve “SVM_Classify”, test örneklerinin sınıflandırılması için kullanılır. Yazılım ayrıca genelleme performansının değerlendirilmesi için iki verimli tahmin yöntemleri sağlar: XiAlpha – tahminler, esas olarak hiçbir hesaplama pahasına olmadan işlem yapar fakat katı bir şekilde yanlıdır

³³ C.-C. Chang, C.-J Lin, “LibSVM: A library for support vector machines software”, (Çevirimiçi) <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>, 2001.

³⁴ T. Joachims, “SVM^{light} Support Vector Machine”, (Çevirimiçi) <http://svmlight.joachims.org/>, 14.08.2008.

(biased), ve diğer tahmin yöntemi ise Leave –one-out Test, hemen hemen tarafsızdır (unbiased).

Ayrıca, bir çok sayıda DVM algoritmasını içeren makine öğrenimi için araç kutuları (toolboxes) mevcuttur, bunlardan bazıları, Torch (C++), Spider(MATLAB), ve Weka (JAVA) dan oluşmaktadır.

3.8 Mevcut ve gelecekteki Araştırmalar

Son on yılda DVM hem teori ve hem de pratikte çok hızlı bir şekilde gelişti. Birçok gelecek araştırmaları devam etmektedir. Bu bölümde, temel ilerlemelerin yapıldığı ve bir çok araştırma problemlerinin açık olduğu başlıca araştırma yönlerinin bir kaç tanesine burada değinilecek.

3.9 İşlemsel Etkinlik

DVM'nin ilk sorunlardan biri büyük veri kümelerinde uygulanamaz algoritmalara sebebiyet veren, eğitime aşamasında maliyetli hesaplama karmaşıklığıdır. Fakat bu problem büyük başarı ile çözülüyor. Yaklaşımlardan bir tanesi, çok büyük bir optimizasyon problemini küçük problemler serilerine bölmektir, burada her problem sadece çok dikkatli seçilen birkaç değişkenden oluşmaktadır böylece optimizasyon problemi verimli bir şekilde yapılır. Bu süreç tüm ayrılmış olan optimizasyon problemleri başarılı şekilde çözülene kadar devam eder.

Daha yeni bir yaklaşım olan, bir durumlar kümesinin yaklaşık minimum çevreleyen topun bulunması gibi, DVM öğrenimleri problemini dikkate almaktadır³⁵. Bu çekirdek kümelerde DVM lerin öğrenme problemlerini çözmek çok hızlı ve iyi

³⁵ Ivor W. Tsang, Andras Kocsor, James T. Kwok. **Simpler core vector machines with enclosing balls**, Proceedings of the Twenty-Fourth International Conference on Machine Learning (ICML), Corvallis, Oregon, USA, June 2007, pp. 911-918.

yaklaşım çözümleri üretebilir. Örneğin, çekirdek vektör makinesi³⁶ ve daha ileri top vektör makinesi³⁷ saniyeler için milyonlarca veri için DVM'leri öğrenebilir.

3.10 Çekirdek Seçimi

Destek Vektör Makinelerinde çekirdek fonksiyonu seçimi genellikle Mercer Teoremini sağlamak içindir. Bundan dolayı, yaygın çekirdek fonksiyonları üç türden oluşur: sigmoid, polinom, ve radyal tabanlı (Radial Based Functions) fonksiyonlarıdır, öyle ki bunlar bazen çekirdek hilesinin (kernel trick) uygulanmasını sınırlandırabilirler. Son zamanlarda, Pekalska ve ark. , genel yakınlık ilişkisi haritalamaya dayanan çekirdek fonksiyonu tasarımı için yeni görünüm sağladılar³⁸ . Yeni çekirdek fonksiyonu Mercer şartlarını sağlamak zorunda olmadığı gibi sadece bir öznelik uzayı ile de sınırlandırılmak zorunda da değildir ve yaygın Mercer çekirdeklerinden daha iyi bir sınıflandırma performansı deneysel olarak göstermektedir. Fakat yeni genelleştirilmiş çekirdeğin teorik temeli daha ileri seviyede araştırmaya ihtiyaç vardır.

Ayrıca, birden fazla çekirdeği dikkate alan çoklu çekirdek öğrenimi diğer bir popüler yaklaşımdır; kombinasyonlar yolu ile çok iyi sonuçlar elde edilebilir.³⁹ Bu bir çekirdekler grubunun kullanımına benzerdir. Uygun bir amaç denkleminin ayarlanması ile çekirdeklerin karışımına izin vermek için çekirdek parametrelerinin iyi seçimi yapılabilir.

³⁶ I.W. Tsang, v.d., “Core vector machines: Fast SVM training on very large data sets”, **Journal of Machine Learning Research**, Vol.6, 2005, pp.363-392.

³⁷ I.W. Tsang, A. Kocsor, J.T. Kwok, **a.g.e.**, pp.911-918.

³⁸ E. Pekalska, P. Paclik, R.P.W. Duin, “A generalized kernel approach to dissimilarity - based classification”, **Journal of Machine Learning Research**, vol. 2, 2001, pp.175-211.

³⁹ J. Bi, T. Zhang, K. Bennett, “Column-generation boosting methods for mixture of kernels” **KDD**, 2004, pp. 521-526.

3.11 Genelleme Analizi

Genel olarak, çekirdek makinelerin genelleme hata sınırının tahmini VC boyut kullanılarak yapılıyor. Fakat sınır eğitme verilerine bağlı olmayan sabit bir karmaşıklık cezası içerir, bunun sonucu olarak, evrensel olarak etkili yapılamaz⁴⁰. Bu problemi çözmek için, klasik VC boyut yerine bir sınıflandırıcının karmaşıklığını değerlendirmek için alternatif olarak Rademacher in karmaşıklığı (Rademacher's complexity) sunuldu⁴¹, bu tesadüfî veriye uyan yeteneği sayesinde bir sınıflandırıcının kapasitesini (karmaşıklığını) ölçebilmek için sezgiye dayalıdır. Aşağıdaki şekilde tanımlanır:

3.12 Rademacher Karmaşıklığı⁴²

Bir $\mathbf{X}$ kümesinde D dağılımı yolu ile oluşturulan $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ örneği için ve $\mathbf{X}$ tanım kümesine sahip bir reel değerli fonksiyon sınıfı F için, F 'in ampirik Rademacher karmaşıklığı tesadüfî değişkendir

$$\hat{R}_n(F) = E_{\sigma} \left[\sup_{f \in F} \left| \frac{2}{n} \sum_{i=1}^n \sigma_i f(\mathbf{x}_i) \right| \middle| \mathbf{x}_1, \dots, \mathbf{x}_n \right] \quad (3.47)$$

burada $\sigma = \{\sigma_1, \dots, \sigma_n\}$ bağımsız düzgüm $\{\pm 1\}$ - değerli (Rademacher) tesadüfî değişkenleridir. F 'nin Rademacher karmaşıklığı

$$R_n(F) = E_s [\hat{R}_n(F)] = E_{s\sigma} \left[\sup_{f \in F} \left| \frac{2}{n} \sum_{i=1}^n \sigma_i f(\mathbf{x}_i) \right| \right] \text{ 'dır} \quad (3.48)$$

⁴⁰ P.L. Bartlett, S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results", **Journal of Machine Learning Research**, Vol.3, 2002, pp. 463-482.

⁴¹ P.L. Bartlett, "The sample complexity of pattern classification with neural networks: The size of the weight is more important than the size of the network", **IEEE Transaction on Information Theory**, vol.44(3.2), 1998, pp.525-536.

⁴² J. Shawe-Taylor, Nello Cristianini, **Kernel Methods for Pattern Analysis**, New York, Cambridge University Press, 2004, p.95 .

Formülün içerisindeki “sup” parçası sınıfın bir fonksiyonu ile tesadüfî etiketler arasındaki bulunabilen en iyi korelasyonu hesaplar. Ayrıca, çekirdek makinelerinde, Rademacher karmaşıklığı için üst sınır elde edilebilir:

3.13 Karmaşıklık Analizi⁴³

Eğer $k: \mathbf{X} \times \mathbf{X} \rightarrow \mathbf{R}$ bir çekirdek ise , ve $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ de $\mathbf{X}$ teki noktaların bir örneği olsun, daha sonra sınıflandırıcı F_B ’nin ampirik Rademacher karmaşıklığı aşağıdaki eşitsizliği sağlar

$$\hat{R}_n(\mathbf{F}_B) \leq \frac{2B}{n} \sqrt{\sum_{i=1}^n k(\mathbf{x}_i, \mathbf{x}_i)} = \frac{2B}{n} \sqrt{\text{tr}(\mathbf{K})}, \quad (3.49)$$

burada B, sınıflandırıcıdaki $\mathbf{w}$ ağırlıklarının sınırıdır.

Rademacher karmaşıklığı sadece ilgili çekirdek matrisinin iz’ini içerir, bu da somut eğitim verileri tarafından belirlenir. Bir sınıflandırıcının karmaşıklığını kontrol etmek ve ayrıca genelleme performansını tahmin etmek için geleneksel VC boyutuna Rademacher karmaşıklığının kullanılması daha uygundur

3.14 Yapısal DVM Öğrenimi

Marj maksimizasyonu, DVM algoritmasının başlangıç motivasyonudur⁴⁴. Sonuç olarak, DVM genellikle örneklerin sınıfların arasındaki ayrılabilirliği üzerinde odaklanır fakat sınıflar arasındaki yeterince eski veri dağılımı bilgisini kullanmaz. Meşhur “No free Lunch teoremi” başka algoritmalara doğal olarak üstün olan veya ek bilgi kullanmaksızın rastgele tahmin etmeye üstün olan bir kalıp (patern) sınıflandırma

⁴³ Shawe-Taylor, Cristianini, **a.g.e.**, pp. 99-100.

⁴⁴ Kaizhu Huang, Haiqin Yang, Irwin King, Michael R. Lyu, “Maxi-Min Margin Machine: Learning Large Margin Classifiers Locally and Globally”, **IEEE Transactions on Neural Networks**, 19(2), 2008, pp. 260-272.

metodu mevcut olmadığını belirtir⁴⁵. Uygulanacak olan sınıflandırma biçimini problemin türü, önceki bilgi ve eğitme örneklerinin miktarı belirtir. Farklı gerçek hayat problemlerine karşılık gelen farklı sınıflar farklı veri yapılarına sahip olabilirler. Bir sınıflandırıcı diskriminant sınırlarını sınıflandırma için önemli olan, özellikle sınıflandırıcının genelleme kapasitesi için, yapıları uydurmak için ayarlamalıdır. Fakat, geleneksel DVM yapıları ayırt etmez, ve

Türetilen karar hiperdüzlem tarafsız bir şekilde (unbiasedly) destek vektörlerin tam ortasında durur, bu da gerçek hayat problemlerinde optimal olmayan sınıflandırıcılara sebep olur^{46,47}.

Son zamanlarda, geleneksel DVM ye kıyasla yapısal bilgiye (structural information) daha çok ilgilenilmesi için bazı algoritmalar geliştirildi. Bu algoritmalar bir sınıflandırıcı tasarlanması için yeni bir görünüm sağlar, burada sınıflandırıcı veri dağılımının yapısına duyarlı olabilir. Bu algoritmalar genellikle iki yaklaşıma ile bilinirler. Birinci yaklaşım, manifold öğrenme yoludur. Bu yaklaşım verinin aslında girdi uzayının alt manifoldunda yaşadığını farz eder ve en tipik algoritma Laplace Destek vektör makineleri (LapSVM) içerir[38,39]. LapSVM ilk önce Laplace grafiği yolu ile oluşturulur. Daha sonra ilgili Laplace matrislerinin içerisinde verinin bir manifold yapısı DVM nin geleneksel yapısının içerisine ek terim olarak sunulur.

İkinci yaklaşım ise verinin önceki dağılım bilgisine sahip olan birçok kümeler içerdiğini varsayarak kümeleme algoritmalarını (clustering algorithms) yararlanmaktır⁴⁸. Bu varsayım manifold varsayımına göre daha geneldir, aslında, bu varsayım birçok popüler geniş marj makinelerine yol açtı. Son yaklaşım ise yapılandırılmış geniş marj

⁴⁵ Duda, Hart, Stork, **a.g.e.**, pp.454-457.

⁴⁶ Huang, Yang, King, Lyu, **a.g.e.**, pp. 260-272.

⁴⁷ D.S. Yeung, D. Wang, v.d., "Structured large margin machines: Sensitive to data distributions", **Machine Learning**, vol.68, 2007, pp.171-200.

⁴⁸ P. Rigollet, "Generalization error bounds in semi - supervised classification under the cluster assumption", **Journal of Machine Learning Research**, vol.8, 2007, pp.1369-1392.

makinelere (structured large margin machine ,SLMM)⁴⁹. SLMM ilk olarak yapısal bilgileri farklı sınıflarda yakalamak için kümeleme tekniklerini uygular. Daha sonra, SLMM ilgili yapı bilgisini kısıtların içerisine sokmak için, geleneksel Öklid uzaklığı yerine örneklerden karar hiperdüzlemlere olan bir mesafe ölçümü olarak Mahalanobis uzaklığını kullanır. Bazı popüler geniş marj makineleri, örneğin destek vektör makineleri, mini-max olasılık makinesi (MPM), maxi-min marj makinesi (M₄), hepsi SLMM nin bir özel durumu olarak görülebilir. SLMM daha iyi sınıflandırma performansı deneme yapılarak göstermiştir. Fakat, SLMM'nin optimizasyon problemi DVM deki karesel programlama (quadratic programming) yerine sıralı ikinci derece koni programlama (sequential second order cone programming, SOCP) olarak formüle edildiği için, geleneksel DVM ye kıyasla SLMM eğitime süresinde daha yüksek hesaplama maliyetine sahiptir. Bunun yanı sıra, büyük ölçekli veya çok sınıflı problemlerin genelleştirilmesi çok kolay değildir. Sonuç olarak, yeni bir yapısal destek vektör makinesi (structural support vector machine, SSVM) , SLMM deki kısıtlardan ziyade, DVM'nin klasik çerçevesinden faydalanmak için geliştirildi⁵⁰. Sonuç olarak, ilgili optimizasyon problemi yinede karesel programlama yolu ile DVM' deki gibi çözülebilir, ve bu çözümü seyrek tuttuğu gibi aynı zamanda ölçeklenebilir olarak da tutar. Ayrıca, , DVM ve yapılandırılmış geniş marj makinelere (SLMM) kıyasla, yapısal DVM' nin (SSVM) genelleme konusunda daha iyi olduğu teorik ve deneysel olarak gösterilmiştir.

3.15 Doğrusal Olmayan Destek Vektör Makineleri

Gerçek hayatta çok az sayıda veri kümeleri doğrusal olarak ayrılabilir. Basit doğrusal çerçevenin kolay bir şekilde veri kümesinin doğrusal ayrılabilir olmayan bir duruma genişletilebilir olması destek vektör makinelerini kayda değer kılmaktadır. Bu genişletmedeki temel amaç veri kümesinin doğrusal olarak ayrılabilir olmadığı bir girdi

⁴⁹ Yeung, Wang, v.d., **a.g.e.** pp.171-200.

⁵⁰ H. Xue, S. Chen, Q. Yang, "Structural support vector machine", **The Fifth International Symposium on Neural Networks**, Part I, LNCS5263, 2008, pp.501-511.

uzayını (input space) verinin doğrusal olarak ayrılabilirdiği öznitelik uzayı (feature space) olarak adlandırılan daha büyük boyutlu bir uzaya dönüştürmektir. Eğer bu dönüşümler dikkatli bir şekilde seçilirse, öznitelik uzayı ile ilişkili tüm hesaplamaları girdi uzayında gerçekleştirilebilir⁵¹. Bu dönüşümler ile ilgili fonksiyonlara çekirdek (kernel) fonksiyonları adı verilir ve doğrusal bir destek vektör makinesinden doğrusal olmayan bir destek vektör makinesine geçmek için kullanılan bu fonksiyonların sürecine çekirdek hilesi (kernel trick) adı verilir. Bu konuya yukarıda detaylı olarak yer verilmiştir.

⁵¹ Hamel, a.g.e, p.103.

DÖRDÜNCÜ BÖLÜM

ÇALIŞAN MEMNUNİYETİ

4.1 Çalışan Memnuniyetine Genel Bakış

Son yıllarda çalışan memnuniyeti büyük ilgi odağı haline gelmiştir. Birçok şirket çalışan memnuniyetini gözlemlemek ve artırmak için programlara önemli kaynaklar yatırımı yapıldı (Heskett v.d.,1997). İş tatmini, diğer yaşam alanları ile bireylerin memnuniyeti (örneğin, medeni tatmin) gibi duygusal bir durum olarak tanımlanmış olmasına karşın, iş memnuniyeti kişinin kendi işinin değerlendirmesini yansıtan bir davranışsal yapıdır. İş memnuniyetini ölçmek için, duygusal yönü genel olarak bilişsel yönüne nazaran daha az derecede vurgulanır¹. İş memnuniyeti gibi tutumlar iş ve organizasyon hakkındaki duygusal tutumlar ve düşünceler yolu ile oluşur. Çalışan memnuniyeti çalışanların işlerine ya da işçilerin işlerini sevmeye derecesindeki memnuniyeti olarak da tanımlanır. İş memnuniyeti insanların işleri hakkında ve işlerinin farklı yönleri hakkında hissettikleri olarak tanımlanır (Spector, 1997).

İş tatmini, kişinin işinden ne istediği ile bu işin sunduğu şeylerden ne algıladığı arasındaki algılanan ilişkinin bir fonksiyonudur (Locke, 1969). Memnuniyet, insanların işleri veya iş ortamları hakkında pozitif veya negatif değerlendirmeli yargıları gibi kabul edilir (Weiss, 2002).

“Genel memnuniyet” literatürde belirtilmesine rağmen, iş memnuniyeti çok boyutlu olarak görülmektedir. Memnuniyet kavramı çeşitli iş boyutları ile bağlantılıdır². Anderson (1984) çalışmasında beceri çeşitliliği, görev kimliği, görevin önemi, özerklik ve iş geribildirim gibi beş iş boyutunun önemli bir derecede iş tatmini ile ilişkili

¹ Remus Ilies and Timothy A. Judge, “An experience-sampling measure of job satisfaction and its relationships with affectivity, mood at work, job beliefs, and general job satisfaction”, **European Journal of Work and Organizational Psychology**, vol. 13(3), 2004, p.368.

² Carol H. Anderson, “Job Design: Employee Satisfaction and Performance In Retail Stores”, **Journal of Small Business Management**, 1984, pp.9-10.

olduğunu göstermiştir. Fakat beceri çeşitliliği ile iş memnuniyeti arasındaki ilişki tam olarak açık olmadığı çalışmada gösterilmiştir. Ayrıca, Teas³ perakende satış elemanının iş memnuniyeti ile performans geribildirimi ve diğer iş boyutları arasında pozitif bir ilişki bulmuştur.

Çalışanın iş memnuniyeti, birinin işi veya iş tecrübesinin değerlendirilmesinden kaynaklanan zevkli veya pozitif duygusal durum olarak da ifade edilir (Locke, 1976, p.1300). Çalışanın iş tatmini a) genel iş koşullarına küresel bir tepki olarak ve b) farklı iş faktörlerine (örneğin, iş, üst denetçi, ücret, iş arkadaşları vb.) karşı duygusal reaksiyonlar olarak işlevsel hale getirilmiştir⁴. İş doyumu, hem çalışanların bilişsel değerlendirmelerini ve hem de kişilerin işteki genel durumları hakkındaki duygularını içeren genel bir tutumu temsil eder. (Brown, Lam, 2008) çalışmalarında basit işle ilgili etkinin, iş yönlerinin bilişsel değerlendirmelerine nazaran, müşteri tarafından algılanan hizmet kalitesi ve iş memnuniyeti ile daha kuvvetli ilişkili olduğunu göstermişlerdir. Yapılan meta-analiz⁵ sonucunda bu ilişkinin istatistiksel olarak anlamlı ve çalışan memnuniyetinin müşteri memnuniyeti ve algılanan hizmet kalitesi ile önemli ölçüde bir bağlantının olduğu gösterilmiştir.

Çalışan memnuniyeti, müşteri memnuniyeti kadar örgütsel performansı etkiler. Çalışanlar işletmenin iç müşterileridir; onlar mevcut çalışma ortamını oluştururlar ve iş hedeflerine ulaşmak için iş dünyası ile işbirliği yapmaya isteklidirler.

İş tatmini çalışma ortamındaki iş performansını, devamsızlığı, fire (turnover) ve psikolojik gerginlikleri önemli derecede etkilediği tespit edilmiştir (Spector, 1997).

³ Kenneth R Teas, "A Test of a Model of Department Store Salespeoples' Job Satisfaction", **Journal of Retailing**, vol. 67, 1981, p.3.

⁴ Steven P Brown, A. Robert Peterson, "Antecedents and Consequences of Salesperson Job Satisfaction: Meta-Analysis and Assessment of Causal Effects," **Journal of Marketing Research**, vol.30 (1), 1993, pp.63-66.

⁵ Steven P. Brown*, Son K. Lam, "A Meta-Analysis of Relationships Linking Employee Satisfaction to Customer Responses", **Journal of Retailing**, vol.84(3), 2008, pp.243-245.

Memnun olmayan çalışanlar fazla fire vermeye ve devamsızlığa eğilimlidirler. İş memnuniyetinin önemi böylece performans, örgütsel verimlilik ve diğer konular ile bağlantılı olabilir (Dickter, v.d., 1996). Bundan dolayı çalışan memnuniyeti verimliliği artırabilir, personel devri azaltabilir ayrıca yaratıcılığı ve bağlılığı artırabilir. Bundan dolayı, işçi memnuniyeti konusu göz ardı edilmemelidir ve henüz çok az işletme ciddi bir şekilde işçi memnuniyetini dikkate almaktadır (Ulmer, v.d., 1999). Önceki çalışmalardan elde edilen bulgular çalışanların müşterilerin memnuniyetinde önemli rol oynadıklarını göstermektedir.

4.1.1 Çalışan Memnuniyeti ve Performans

Çalışmalar, çalışan memnuniyetinin şirketlerin finansal hedeflerini gerçekleştirmek için birincil bir rol oynadığını göstermektedir⁶. Eğer bir şirket çalışanlarına önem veriyor ve bakıyor ise, çalışanlarda müşterilerine önem verip onlar özen göstereceklerdir. Müşterilerin, çalışanlarının tatmin ve sorumluluk yüksek olan organizasyonlarda daha iyi tecrübelerle eğilimli oldukları doğrudur. Çalışanlara özen göstermek ve onlara bakmak, daha iyi bir ücret vermek, devamlı eğitim sağlamak ve çalışanların güvende hissedilmesi olarak tanımlanabilir (Koys, 2003).

Birçok çalışmada çalışan memnuniyeti ve finansal performans arasında anlamsız direkt ilişki olduğu gösterilmiştir. Örneğin, Wiley (1991) çalışmasında çalışan memnuniyeti ile finansal sonuçlar arasında istikrarlı bir negatif ilişki olduğunu belirtmiştir. Böylece, çalışan memnuniyet ile finansal performans arasında önemli oranda direkt bir ilişki mevcut değildir, dolayısıyla çalışan memnuniyeti ile finansal performans arasındaki ilişki doğrudan olmayan bir ilişki olarak açıklanabilir⁷.

⁶ Koys, D., "How the achievement of human-resources goals drives restaurant Performance", **Cornell Hotel and Restaurant Administration Quarterly**, vol. 44 (1), 2003. pp. 17–24.

⁷ Christina G. Chi, Dogan Gursoy, "Employee satisfaction, customer satisfaction, and financial performance: An empirical examination", **International Journal of Hospitality Management**, vol. 28, 2009, pp.245–247

Bazı yazarlar iş tatmini ile iş performansı arasındaki ilişkiyi ele almalarına karşın, elde edilen sonuçlar ilişkinin büyüklüyü veya yönü hakkında bir anlaşmaya varılmadığını göstermektedir. İlk araştırma memnuniyet ve performans arasında ilişkinin belirgin varlığını inkâr etmektedir (Brayfield ve Crockett, 1955). Çalışmalarında, aradaki ilişkinin açıkça belirgin olmadığını ortaya koydular, yüksek memnuniyetin ya yüksek ya da düşük verimliliğe (performans) eşlik ettiğini, diğer taraftan, düşük memnuniyetin ise ya yüksek ya da düşük verimliliğe (performans) eşlik ettiğini gösterdiler. Bazı yazarlar, çalışan memnuniyetinin performansın ilerlemesine yol açtığının göstermek suretiyle performans ve memnuniyet arasındaki ilişki kavramını desteklemektedirler (Herzberg, v.d., 1957). Bu konuda en kuvvetli görüşü Herzberg, v.d.(1957) belirtmiştir ve çalışmalarında defalarca iş memnuniyetinin iş performansına yol açtığını savundular. Yani daha iyi performans, iyileştirilmiş iş tutumları tarafından tetiklenme eğilimindedir, tersine, performanstaki bir düşüş ise iş tutumlarındaki negatif değişimler tarafından tetiklenme eğilimindedir⁸. Bir başka görüş, Herzberg, v.d. çalışmalarının tam aksine, Lawler ve Porter (1967) performansın iş memnuniyetine yol açtığı görüşündeler, yani performans ve iş memnuniyeti arasında tam ters ilişkinin olduğunu göstermeye çalışmışlardır.

4.1.2 Çalışan Memnuniyeti ve Kurumsal Performans

Çalışanların yanı sıra, kurumsal performans üzerinde ciddi etkileri olduğundan, çalışan memnuniyeti organizasyonlarda önemli bir konudur. Daha geniş anlamda, çalışan nüfus arasında genel yaşam tatmininin önemli bir parçasıdır. Öyle ki, düşük iş memnuniyeti devamsızlık, geç kalma, işçi devri ve hatta hastalık ve kazalar şeklinde oluşan bırakmanın önemli bir nedeni olarak öne sürülmüştür. Aksine, çok memnun çalışanlar daha az devamsızlık ve daha çok kararlılık eğilimindedirler. Bundan dolayı, daha az olasılıkla işlerini kaçırlar ve beklenen davranış ve hareketler konusunda daha

⁸ Iaffaldano, M.T., Muchinsky, P.M., "Job satisfaction and job performance: a meta-analysis", *Psychological Bulletin*, Vol. 97, 1985, pp. 251-52.

dikkatlidirler. (Zeffane, vd., 2007) çalışmasında kadınlar işlerinin içsel yönlerine önem verme eğiliminde olduklarını, buna karşın erkekler ise işlerinin çalışma koşulları ile ilişkili olan dışsal yönlerine değer verme eğiliminde olduklarını tespit etmişlerdir⁹. Son yıllarda çalışan memnuniyeti kavramı giderek artan bir şekilde araştırmacılar ve uygulamacıların ilgilendiği konulardan biri olmuştur.

Çalışan memnuniyeti çok disiplinli bir alan olduğundan, bu konuda yönetim organizasyon, örgütsel davranış, çalışma psikolojisi, insan kaynakları yönetimi gibi çok değişik alanlardan araştırmacılar çeşitli çalışmalar yapmışlardır. Bu çalışmalarda işçilerin memnuniyetinin farklı yönleri incelenerek, bu yönlerin değişik değişkenler üzerindeki etkileri incelenmiştir (Petrescu and Simmons, 2008). Ayrıca, Petrescu (2008) çalışmasında insan kaynakları yönetiminin çalışanlarının genel iş memnuniyeti üzerindeki etkilerini ve memnuniyetlerinin ücret ile olan etkilerini araştırdı. İKY uygulamaları istatistiksel olarak genel çalışan memnuniyeti ile anlamlı ve bazı durumlarda önemli etkisinin olduğu elde edilmiştir. Ayrıca, özellikle çalışanların sürekli eğitim ve iş özerkliğinden hoşlandıkları görülmüştür. Çalışanlar görsel performanslarının değerlendirilmesinden memnun kaldıkları izlenmiş ve dolayısıyla bazı gözetimler istenilir bir durum olduğu kabul edilmiştir yapılan araştırmada. Bunun yansısı, çalışanların katılımı yolu ile işçilere “söz hakkı” verilmesi çalışanların genel memnuniyetinde pozitif bir etki yaratmaktadır¹⁰.

Çalışan memnuniyeti konusu son yıllarda hem uygulamacılar hem de araştırmacıların üzerinde önemle durduğu konulardandır. Pozitif ve negatif etki ile iş tatmini arasındaki ilişkilerin memnuniyet ölçüsünün bir fonksiyonu olarak farklı olup

⁹ Zeffane, R., M. Ibrahim, R. ElMehairi, “Exploring the differential impact of job satisfaction on employee attendance and conduct”, **Employee Relations**, 2007, p. 237.

¹⁰ Petrescu, A.I., R. Simmons, “Human Resource Management Practices and Workers' Job Satisfaction”, **International Journal of Manpower**, vol. 29 no. 7, 2008, pp.651-667.

olmadığı Kaplan (2009) tarafından araştırılmıştır. Bu alanın uzmanları işçi memnuniyeti ölçümelerini ya duygusal ya da bilişsel olarak sınıflandırır¹¹.

İş tatmininin çeşitli teorileri psikologlar ve yönetim bilim adamları tarafından geliştirildi. Memnuniyet kaynaklarına farklı önem dereceleri atama eğilimindeler, bunlar içsel ya da dışsal olarak sınıflandırılır. İçsel kaynaklar kişinin tutum ve davranış gibi bireysel karakterine bağlıdır. Dışsal kaynaklar ise durumsaldır ve çalışma ortamı gibi çevreye bağlıdır. Dışsal kaynaklara dayanan teoriler genel olarak daha çok ekonomistler tarafından benimsenir, buna karşın içsel kaynaklar daha genel olarak sosyal bilimleri ile alakalıdır¹².

Çalışan memnuniyeti daha yüksek müşteri memnuniyetine sebep olur, bu memnuniyeti daha sonra firmanın performansını direkt olarak etkiler. Özellikle hizmet endüstrisinde, bazı çalışmalar çalışan memnuniyeti ile müşteri memnuniyeti ve şirket performansı aralarında pozitif bir ilişki buldular¹³

4.1.3 Çalışan Bağlılığı, Kişisel Özellikler ve Çalışan Memnuniyeti

Maltzer ve Renzl (2007) kişilik özelliklerine çalışan memnuniyetinin belirleyicileri olarak ve çalışan memnuniyetinin etkilerinin bir kuruluşa karşı duygusal bağlılığına odaklanmışlardır. Bu çalışmada, nevroitiklik çalışan memnuniyeti ile negatif bir şekilde ilişkili olduğu görülmüştür.

¹¹ Kaplan, S., C. R. Warren, A.P. Barsky, C. J. Thoresen, “A note on the relationship between affectivity and differing conceptualizations of job satisfaction: Some unexpected meta-analytic findings”, **European Journal of Work and Organizational Psychology**, vol 18, number 1, 2009, pp. 29-33.

¹² Luchak, A.A., “What kind of voice do loyal employees use?”, **British Journal of Industrial Relations**, Vol. 41 No. 1, 2003, pp. 115-20.

¹³ Anderson, E. W., v.d., “Customer satisfaction and shareholder value”, **Journal of Marketing**, vol. 68(4), 2004, pp. 172–173.

Bu çalışmada, çalışan memnuniyetinin belirleyicisi olarak kişilik özellikleri ve çalışan memnuniyetinin etkilerinin bir organizasyona karşı olan duygusal bağlılığa odaklanıldı. Ampirik çalışma nevrotiliğin çalışan memnuniyeti ile negatif bir şekilde ilişkili olduğunu gösterdi. İnsafli olmanın (vicdanlı olma, dürüstlük) bir etkisinin olmadığı ve uygunluk ise memnuniyeti pozitif şekilde etkilediği gösterilmiştir. Çalışan memnuniyeti duygusal bağlılık üzerinde güçlü bir etkisi vardır. Çoğu çalışmada, özellikle yönetimsel etkilerin çalışan memnuniyetine olan etkileri gibi çevre faktörlerine büyük ilgi verilmesine karşın, Meltzer ve Renzl (2007) bulguları bireylerin kalıcı özellikleri ile çalışma ortamındaki bireysel memnuniyet arasında önemli bağlantıları aydınlatmıştır. Bu sonuçlar hem teori hem de pratik öneme sahiptir, bu çalışma çalışan memnuniyetinin varyansının %20'den fazlasının kişisel farklılıklarından, yani kişisel özelliklerden kaynaklandığı göstermiştir. Çalışan memnuniyeti sonucunda yüksek motivasyona sahip, kendini işine adanmış çalışanlar meydana getirir, ayrıca daha etkili ve verimli işlerin meydana gelmesin yol açar. Bunun sonucu olarak da yüksek seviyede süreç ve ürün kalitesine vesile olur¹⁴.

Örgütsel kimlik, iş memnuniyeti ve örgütsel bağlılık arasındaki etkileşim vardır. Ayrıca, örgütsel bağlılığın ve kimliğin iş memnuniyeti üzerindeki etkisi incelenmiştir. Sonuçta, örgütsel bağlılık ve örgütsel kimlik iş memnuniyeti ile ters ilişkili olduğunu çalışmada göstermiştir. Ayrıca, iş tatmini üzerindeki bağlılığın etkisi kimliğin etkisinden daha kuvvetli bulunmuştur¹⁵. Çalışanlar ile kurumlar arasındaki ilişki çalışanların tutum ve davranışlarını etkiler (Cole and Bruch, 2006). İşçilerin tutum ve davranışlarına nelerin sebep olduğunun anlaşılması gerekir, bunun için işçilerin çalıştığı kurumlar hakkındaki algılarının belirlenmesi önemlidir. Kurum kimliği, iş tatmini ve kurum bağlılığı kurum tutum ve davranışlarının yorumlanmasında dayanak oluşturur. Hangi elemanların iş tatminini beslediğini anlamak önemlidir, çünkü bu elemanların etkileri

¹⁴ Eskildsen, J. K., Dahlgaard, J. J., "A causal model for employee satisfaction", **Total Quality Management**, 11(8), 2000, pp. 1081-1094.

¹⁵ Tuzun, I. K., "The impact of identification and commitment on job satisfaction", *Management Research News*, vol. 32 no. 8, 2009, pp. 728-730.

hem bireysel hem de kurumsal düzeylerde önemli sonuçlar doğurur. Yüksek iş memnuniyetine sahip çalışanların işlerine karşı yüksek pozitif tutum geliştirirler. (Tuzun,2009)

4.1.4 Çalışan Memnuniyeti ve İşten Ayrılma

Blau (2009) çalışmasında mesleki bağlılığın dört boyutunun (duygusal, normatif, birikmiş maliyetleri ve sınırlı alternatifler) kişinin işinden ayrılma niyetini açıklayıp açıklayamadığını test etmiştir. Çalışan memnuniyeti, işi terk etme niyeti ile kuvvetli bir korelasyona sahip olduğu gösterilmiştir. Elde edilen bu sonuç, literatürdeki modeller ile uygunluk göstermiştir. Mesleki bağlılığın bu dört boyutlu görünümü bir işçinin "iş kimliğini" nasıl oluşturacağını araştırmak için alternatif bir çerçeve olarak gösterilebilir. Bu dört boyut arasında, düşük olan işçilere kıyasla, daha yüksek olan işçilerin daha güçlü bir mesleki kimliğe sahip olduğu öne sürülebilir¹⁶.

Lance (1988) çalışan memnuniyeti, iş performansı ve işten ayrılma niyeti arasındaki ilişkiye çalışmada katkıda bulunmuştur. Ayrıca performans değişkeninin memnuniyet ve işten ayrılma niyeti arasında bir düzenleyici (moderatör) değişken olmasındaki rolünü test etmiştir. Performans ve işten ayrılma niyeti arasında istatistik olarak anlamlı negatif bir korelasyon elde edilmiştir¹⁷.

4.1.5 Müşteri Memnuniyeti ve Hizmet Kalitesi

Burke v.d. (2005) çalışmalarında çalışan memnuniyeti ölçümleri ile müşteri memnuniyeti arasındaki ilişkiyi, süreçlerin yeniden yapılanma girişimi öncesi ve sonrasını incelemişlerdir. Örgütsel performansı yeniden yapılandırma ve toplam kalite

¹⁶ Blau, G, "Can a four-dimensional model of occupational commitment help to explain intent to leave one's occupation", **Career Development International**, vol. 14, no. 2, 2009, pp.116-118.

¹⁷ Lance, C, "Job performance as a moderator of the satisfaction", **Journal of Organizational Behavior**, vol 9, 1988, pp.271-280.

yönetimi yolu ile geliştirilmesi konusu 1990'lı yıllarda önemle üzerinde durulmuştur. Bu çalışmada, araştırma ve pratik organlarının ortak kesitlerini incelemişlerdir. Birinci çalışma toplam kalite yönetimi ve yeniden yapılandırma ile işleri ve iş süreçlerini yeniden tasarlayarak kurum performansını artırmak. Diğer organ ise çalışan memnuniyeti ve müşteri memnuniyeti arasındaki bağlantı üzerine çalışılmıştır. Toplam kalite yönetimi ve yeniden yapılandırma zahmetlidir, fakat yukarıda bahsedilen verimlilik geliştirmeleri uygulanması sırasında bu talepleri ile başa çıkmak için iş taleplerinin ve çalışanların kapasitelerinin seviyelerinin gözlemlenmesi kritik öneme sahiptir. Burke v.d. (2005) ayrıca çalışanların müşterilerine kaliteli hizmet sunmaları için, yeni iş taleplerini içeren çalışan memnuniyetinin önemli bir rol oynadığını göstermiştir.

4.2 Çalışmalardaki memnuniyet faktörleri

Çalışan memnuniyetini düşünmeden, müşteri deneyimine odaklanarak aşırı miktarda zaman ve enerji harcamaktayız. Fakat çalışan memnuniyeti müşteri deneyimlerinin başarı ya da başarısızlığını belirleyeceği için çok önemlidir. Bazı çalışan memnuniyeti araştırmaları genel olarak çalışanların terfi için istikrarlı bir istihdam ve tatmin edici bir tazminat istediklerini göstermektedir. Diğer çalışmalar ise sosyal tatmin ve üstlerin özellikleri çalışanların memnuniyet seviyelerinde bir etkiye sahip olduğunu göstermektedir. Sonuç olarak bu çalışmalar iş tatmini değişik faktörlerden etkilendiğini desteklemektedirler.

4.2.1 Motivasyon

Motivasyon iş memnuniyeti için önemli faktörlerden biri olduğu tespit edilmiştir. Herzberg'in Motivasyon (Hijyen) teorisi, motivasyonu içerik teorilerinden biri olarak sınıflandırılır. Bu teoriler bazı yazarlar tarafından eleştirilmiştir. Bu eleştirilere bakılmaksızın, Herzberg gibi içerik teorisyenleri iş memnuniyeti konusuna önemli

katkılar sağlamışlardır. Herzberg (1959) çalışan memnuniyeti konusunda değişik alanlarda çalışmalar yapmıştır. Herzberg'in iki faktörlü yaklaşımı olan motivasyon – hijyen teorisi iş ve iş ortamı içerisinde memnuniyete veya memnuniyetsizliğe yol açan unsurların olduğunu iddia etmektedir. Herzberg (1959), İçsel bileşenlerin ya da iş içeriğinin çalışanların memnuniyetine yol açtığına ve dışsal bileşenler ya da iş ortamı olduğuna inanmaktadır. Motivasyon faktörleri “memnun ediciler” olarak listelenir, çünkü bizi en yüksek performans seviyelerine yükseltir. Medikal bir terim olarak bilinen hijyen faktörleri ise bizi optimum performans seviyesine ulaşmamızı engelleyen, “tatminsizler” olarak adlandırılır. Herzberg (1959), başarı, tanınma, kendi kendine çalışma, sorumluluk ve ilerleme için fırsat gibi beş faktörü memnun ediciler olarak adlandırmaktadır. Ayrıca, şirket politikası, denetim, ücret, kişiler arası ilişkiler ve çalışma koşulları gibi faktörleri ise “tatminsizler” olarak sınıflandırmaktadır.

İslam ve İsmail (2008) çalışmalarında çalışanların motivasyon ile ilgili faktörleri belirlemişlerdir. Elde edilen bulgular sayesinde çalışan motivasyon programları geliştirilirken, yöneticiler için bu bulgular faydalı öneriler sağlamaktadır. İşyerinde insanların nasıl motive olacağı ve çalışanların kendi kişisel hedeflerini ve kurumların hedeflerini gerçekleştirmek için uygun işyeri ortamlarının nasıl oluşturulacağı konuları yöneticilerinde devamlı sorduğu sorulardandır. Kendilerini motive eden faktörlerin başkalarını da motive eder varsayımı konusunda hatalar yapılmaktadır. Motivasyon kavramı insanoğlu için içsel kavramdır. İnsan yapısı karmaşık olduğundan, insanları motive eden faktörlerde karmaşık ve değişkendir. Verimliliği artırma, müşteri hizmetleri, teknoloji ve süreç yeniden yapılandırmadan daha fazlasını gerektirmektedir. Bu yaklaşımların başarılı olup olmaması daha çok bu programların uygulanması istenilen çalışanların motivasyonlarına bağlıdır. Başarılı bir motivasyon programının tasarlanması için çalışan dahil edilmesi çok önemlidir¹⁸. İslam ve İsmail (2008) bulgularında çalışanların motivasyonlarını etkileyen faktörler olarak parasal teşvikler,

¹⁸ İslam, R., A. Z. İsmail, “Employee Motivation: A Malaysian Perspective”, **International Journal of Commerce and Management** , vol. 18 no. 4, 2008, pp. 344-345.

ödöl ve tasdik etme sistemi, promosyon, iyi çalışma ortamı, çalışanların önerilerini dikkate alan bir sistem gibi sonuçları elde etmişlerdir.

4.2.2 örgütsel vatandaşlık davranışları

Foote ve Tang (2008) çalışmasında kendi kendini yöneten takımlar içindeki takım bağlılığı çalışan memnuniyeti ve örgütsel vatandaşlık davranışları (ÖVD) arasında düzenleyici veya aracı (moderatör) olduğunu gösteren bir model önermektedir. Çalışmadan elde edilen sonuçlarda çalışan memnuniyeti ve örgütsel vatandaşlık davranışları arasındaki ilişkinin önemli olduğu görüşmüştür. Ayrıca, çalışan memnuniyeti ve ÖVD arasındaki ilişki aracı değişken olan takım bağlılığı tarafından etkilenmektedir, öyle ki bu ilişki takım bağlılığı yüksek olduğunda yüksek olmaktadır. Bu çalışma daha önce bilinen fakat çok iyi tanımlanamayan çalışan memnuniyeti ve ÖVD arasındaki ilişkiyi anlamada önemli bir katkı sağlar. Ayrıca, takım bağlılığı ölçümü çapraz doğrulama, doğrulayıcı faktör analizi ve güvenilirlik testlerini kullanarak elde edilen uyum iyiliğine dayanan geçerli ve güvenlidir¹⁹.

Jain, v.d. (2008) çalışmalarında çalışma kontrol odağının (Work Locus of Control, WLOC) çalışan refahı ve örgütsel bağlılık arasında aracı bir değişken olarak rolünü incelemişlerdir. Elde edilen sonuçlarda çalışan refahı koşullu devam bağlılığı ile negatif bir şekilde ilişkili olduğu tespit edilmiştir, bu sayede çalışanlar devamlı katılım ile ilgili avantajları dikkate alırlar. Harici çalışma kontrol odağının ilişkide pozitif bir etkisi mevcuttur. Problemsiz varoluş olarak temsil edilen çalışma refahı belli bir organizasyon ile pozitif duygusal bağlılığı tahmin eder, ve çaba gibi temsil edilen dâhili çalışma kontrol odağı ise ilişkiyi negatif bir şekilde etkiler. Bu çalışmanın odağı ise çalışan refahının örgütsel bağlılık üzerindeki etkisini belirlemektir.

¹⁹ Foote, D.A., T. Tang, "Job satisfaction and organizational citizenship behavior (OCB) Does team commitment make a difference in self-directed teams?", **Management Decision** , vol. 46 no. 6, 2008, pp.933-935.

Hizmet sektöründe çalışan memnuniyetinin işletme performans üzerindeki etkileri araştırılmıştır. Çalışmada, çalışan memnuniyeti, hizmet kalitesi, müşteri memnuniyeti ve firma kârlılığı arasında ilişki yapısal eşitlik modeli kullanılarak incelenmiştir²⁰. Yee, v.d. (2008) çalışmalarında çalışan memnuniyetinin yüksek hizmet kalitesine yol açtığı ve bunun ise müşteri memnuniyetini direkt olarak etkilediğini göstermişlerdir. Memnun çalışanlar ve hizmet kalitesi daha sonra finansal başarıyı getirmektedir. Ayrıca bu ampirik çalışma işletme yöneticilerinin hizmet kalitesini ve müşteri memnuniyetini artırması için çalışan memnuniyetinin önemli bir husus olduğunu önermektedir ve hizmet sektöründeki organizasyonların operasyonel performansını artırmada işçilerin memnuniyetinin önemli bir rol oynadığı ampirik kanıtlar göstermiştir. İşlerinde tatmin olan işçiler istihdam edilen kuruluşlarında daha çok bağlanma eğilimindedir ve yüksek seviyede bir hizmet sağlamada daha çok kendilerini işlerine adanma eğilimindedirler.

4.2.3 İş Memnuniyetsizliği ve Tükenmişlik

İş memnuniyetinin, iş özellikleri ve demografik etkilerin orta düzey yöneticiler arasındaki tükenmişlik düzeyleri üzerinde incelemeler yapılmıştır. Bu çalışmanın bulguları orta düzey yöneticilerin tükenmişlik düzeylerinin orta düzeyde bulunmuştur. Ayrıca, iş memnuniyeti ve iş özellikleri tükenmişlik için önemli açıklayıcı değişkenler olduğu tespit edilmiştir. Aşırı iş yükü, üst düzey yönetim desteği eksikliği, görev karmaşıklığı ve rol belirsizliği gibi bazı iş özellikleri orta düzey yöneticilerin duygusal tükenmişliğini artırmakta ve performans düzeylerini ise azaltmaktadır. Yiyecek ve içecek yöneticilerin ve ön büro işlemleri yöneticileri arasındaki tükenmişlik düzeyleri orta düzey yöneticilerin tükenmişlik düzeylerine göre daha yüksek bulunmuştur. Üst düzey yönetim iş karmaşıklığını, görev belirsizliğini ve aşırı yük miktarını azaltarak orta

²⁰ Yee, R. W.Y., A. C. Yeung, Cheng, “The impact of employee satisfaction on quality and profitability in high-contact service industries”, **Journal of Operations Management**, vol 26, 2008, pp.651-668.

düzyer yöneticilere destek vermesi gerekmektedir. Tükenmişlik için belirlenen önemli değışkenler tanımlandığından, tükenmişliği ve bunun olumsuz sonuçlarını azaltmak için stratejilerin geliştirilmesi kolaylaştırılabilir. İş memnuniyetsizliği, tükenmişlik için işle ilgili önemli bir öncülü olduğu ampirik olarak tespit edilmiştir²¹.

4.2.4 Çalışma Koşulları

Hsu ve Wang (2008) iş memnuniyetinin çalışma koşulları ve bireysel sonuçlar arasında aracı olduğunu önermektedirler. Çalışmada ele alınan çalışma koşulları ile ilgili faktörler: Dışsal ödöl, denetim desteğı, adalet, özerklik, kurumsal imaj, yakınlık, çalışanların gelişimi. Bununla beraber, bireyler ile ilgili çalışan memnuniyet düzeyini etkileyen faktörler ise şunlardır: çalışan bağlılığı, görev performansı ve bağlamsal performanstır. Bu çalışmada, yapılan yapısal eşitlik modeli (Patial Least Squares Modelling) sonucunda çalışan memnuniyeti sırası ile çalışan bağlılığı, görev performansı ve bağlamsal performans ile pozitif olarak ilişkili olduğu tespit edilmiştir. Çalışan memnuniyeti, çalışan bağlılığı için en iyi belirleyicilerden biri olduğu görölmüştür. Memnun çalışanların motivasyon seviyelerinin daha yüksek olduğu, çok daha fazla çalıştıkları ve iş performans düzeylerinin daha yüksek olduğu gözlemlenmiştir. Çalışan memnuniyetini etkileyen çalışma koşulları ile ilgili faktörler etki kuvvetinin göre sırası ile çalışanların gelişimi, dışsal ödöl, denetim desteğı, kurumsal imaj, adalet, yakınlık ve özerkliktir. Çalışan memnuniyetini hangi faktörlerin etkilediğini bilmesi yöneticiler açısından önemlidir. Fakat, insan doğası çok karmaşık olduğundan bu çok kolay bir iş değildir. Eğer, bunun yapılamaması durumunda çalışanların hüsranı artacak ve bunun sonucunda da kötü bir performans elde edilecektir²².

²¹ Karauzum, A., N. Anafarta, S. Irmak, “Predictors of burnout among middle managers in the Turkish hospitality industry”, **International Journal of Contemporary Hospitality Management**, vol.20 no.2, 2008, pp.186-188.

²² Hsu, C.H., Y. C. Wang, “The development and empirical validation of the Employee Satisfaction Index model”, **Total Quality Management & Business Excellence**, vol.19 no. 4, 2008, pp.353-366.

4.2.5 Demografik Faktörler

Bowen ve Cattell (2008) çalışan memnuniyeti ile demografik faktörler, iş ortamı özellikleri, kariyer seçimi, cinsiyet, ırk ile ilgili konular ve işyerinde ayrımcılık ve taciz arasındaki ilişkinin önemini araştırmışlardır. Buna göre, cinsiyet ve ırk temelinde, algılanan iş memnuniyeti seviyelerinde anlamlı farklılıkların var olduğu tespit edildi. İşyeri özelliklerinin çoğunluğu anlamlı şekilde iş memnuniyeti ile ilişkili olduğu görülmüştür. Bahsedilen birçok işyeri özellikleri “kendini gerçekleştirme ihtiyacı” olarak sınıflandırılabilir, örneğin, iş yaparken kişisel memnuniyet duyguları, zorlayıcı ve yaratıcı iş üstlenmek, çeşitli ve tekrar etmeyen iş yapmak ve denetim seviyesi az olacak şekilde girişimcilik özelliklerinin ortaya konulmasına izin verilmesi. Anlamlı olan diğer faktörler “saygıya olan ihtiyaç” ve “ait olma ihtiyacı” kategorileri içerisinde yer almaktadır. Bu çalışmada ücret ve iş memnuniyeti anlamlı şekilde ilişkili çıkmamıştır. Kariyer seçimi konusunda doğrulayıcı duygular iş memnuniyeti duygusu ile pozitif şekilde ilişkili olduğu tespit edilmiştir. Cinsiyet ve ırk açısından, işyerinde ayrımcılık bir pek çok yönden iş memnuniyeti ile anlamlı şekilde ilişkilidir. Cinsiyet ve ırk gözetmeksizin tanınan ve esnek çalışma saatlerine sahip olduğunu hisseden çalışanlar yüksek oranda iş memnuniyetine sahip olduğu tespit edilmiştir. İşyerinde taciz ve ayrımcılık örnekleri nadir değildir. Anlamlı ayrımcılık şekilleri ırk ve dini inançlara dayanmaktadır. Bu şekilde ayrımcılık tecrübesi ile karşılaşan çalışanlar diğer çalışan arkadaşlarına göre daha düşük seviyede iş tatminine sahip olduklarını bildirmişlerdir²³.

Wickramasinghe (2009) bilgi teknolojileri (BT) firmalarında çalışanların iş memnuniyeti düzeyini incelemiştir. Genel iş memnuniyetini ve iş memnuniyetinin yönlerini (iş içeriği, ücret, terfi, denetim, iş arkadaşları) etkileyen faktörler olarak yaşı, çalışma süresini ve çalışma arkadaşlarını ele alarak inceleme yapmıştır. Yapılan ampirik çalışmanın sonuçları cinsiyet ve görev süresi (tenure) anlamlı iş memnuniyeti ölçüleri

²³ Bowen., P., K. Cattell, "Job satisfaction of South African quantity surveyors", **Engineering, Construction and Architectural Management**, vol. 15 no. 3, 2008, pp.260-269.

olduğunu belirtmiştir. Bayanlar işlerinde daha az memnun oldukları ve BT işlerine karşı ilgi kaybı hissetmelerine rağmen hâlihazırdaki işlerinde kalmak istedikleri ortaya çıkmıştır. Çok uzun süre BT işinde çalışanların işlerinden daha az memnun oldukları görülmüştür, ayrıca BT işlerine olan ilgi kaybı hissetmektedirler, dolayısıyla hâlihazırdaki işlerinden ayrılmak istemektedirler. BT çalışanları üzerinde yapılan çalışmalar genellikle batı kaynaklıdır. Birçok çalışma cinsiyet ile iş tatmini arasında korelasyon olduğunu belirtmiştir, fakat erkekler ve bayanlar arasında göreceli iş memnuniyeti ile ilgili olarak kesin kanıt elde edilmiş değildir. Yaş, iş memnuniyetinin yönleri ya da genel iş memnuniyeti ile anlamlı şekilde ilişkili olmadığı tespit edilmiştir. Çalışma süresi ile iş içeriği, terfi, denetim, BT alanına ilgi kaybı ve genel iş memnuniyeti arasında anlamlı negatif bir ilişkiye sahip olduğu bu çalışmada tespit edilmiştir. Bu çalışmada demografik değişkenlerin iş memnuniyetini nasıl şekillendirdiği görülmektedir²⁴.

4.2.6 Kişilerarası Beceriler, Ücret

Gu ve SenSiu (2009) oyun sektöründe iş memnuniyeti ve iş performansı arasındaki ilişkiyi ve memnuniyetin iş performansını nasıl geliştirdiğini incelemişlerdir. Kişilerarası beceriler iş gücünün önemli zayıflığıdır ve iş memnuniyeti iş performansı ile anlamlı şekilde korelasyona sahiptir. Ayrıca, eğitim imkânları, ücret ve ek ödemeler, iş arkadaşlarından ve üstlerden destek gibi faktörler iş memnuniyetinin önemli faktörlerindendir. Hizmet sektöründe, iyi teknik ve kişisel beceriler önemli olmasına karşın, ikisi tek başına müşterilerin memnuniyetinde yeterli değildir. Müşteri hizmetleri bu sektörde çalışan işçiler tarafından sağlanır ve çalışanların ruh hali sağlanan hizmetleri direkt olarak etkilemektedir. Bu nedenle, müşteri memnuniyeti çalışan memnuniyeti tarafından desteklenmektedir. Elde edilen sonuçlar doğrultusunda, oyun sektöründe özellikle bu çalışmanın yapıldığı saha çalışmasındaki kişiler arası beceriler diğer teknik

²⁴ Wickramasinghe, V., “Predictors of job satisfaction among IT graduates in offshore outsourced IT firms”, **Personnel Review** , vol.38 no.4, 2009, pp. 413-417.

veya oyun becerilerine oranla daha zayıf çıktığından, ilgili kuruluş çalışanların kişisel becerilerinin eğitimi konusunda daha dikkat etmelidir. Çalışanlar müşterileri selamlama, yardım etme ve tebessüm etme konularını öğrenmek zorundalar. Dil yeteneği diğer önemli bir faktördür, çünkü bu hem kişilerarası hem de teknik becerileri etkiler. Son olarak, oyun sektöründe çalışanların tam memnuniyeti iş becerileri ve çalışan bağlılığı ile önemli ölçüde ilişkili olduğu bulunmuştur. Bu çalışmanın sonuçlarına bakarak, tecrübeli çalışanların ve yönetimin yeni işe alınan çalışanları ve tecrübesiz işçileri iş ortamında eğitmesi uygun bir yaklaşımdır. Dolayısıyla, yeni çalışanların iş becerileri sürekli olarak gelişebilir ve tüm çalışanlar ile yönetim arasında hoş bir ilişki oluşturulabilir. İş başında eğitim modeldeki “eğitim” ve “destek” faktörlerini desteklemiş olur²⁵.

Graham ve Messner (1998) eğitim sektöründe çalışan memnuniyeti faktörlerini incelemiştir. Buna göre, müdürlerin işlerinden genel olarak memnun olduğu fakat ücret, ilerlemeleri için fırsatlar, ek ücret gibi faktörlerden daha az tatmin oldukları tespit edilmiştir. Okulların büyüklüklerine göre (küçük, orta, büyük) müdürlerin ücret, çalışma arkadaşları ile olan ilişkiler ve diğer faktörler açısından farklılık göstermektedir. Örneğin, orda büyüklükte çalışan müdürler işlerinden daha çok memnun olduğu analizler sonucunda gösterilmiştir. Bununla beraber, cinsiyet faktörü de incelenmiştir. Örneğin erkek müdürler bayan müdürlere göre ücret ve ek ücret faktörleri açısından daha çok memnun olduklarını belirtmişlerdir. Graham ve Messener (1998) müdürlerin çalışma sürelerinin dört veya sekiz yıldan fazla olanların ilerleme fırsatları faktörü açısından çok tatmin olmadıklarını tespit etmişlerdir. Ayrıca, az çalışma tecrübesine sahip müdürlerin ücretlerinden memnun olmadıklarını göstermişlerdir.

²⁵ Gu, Z., R. C. Sensiu, “Drivers of job satisfaction as related to work performance in Macao casino hotels”, **International Journal of Contemporary Hospitality Management** , vol 21 no. 5, 2009, pp.561-570.

4.2.7 Çalışan Devri, Eğitim

Bazı çalışanların bazılarının örgün eğitimleri olmayabilir. Eğitim yolu ve iş becerilerinin kazandırılması yolu ile bu çalışanların entelektüel yetenekleri geliştirilebilir ve dolayısıyla bunların çalışanların memnuniyetlerinde önemli bir etkisi vardır. Yönetim, işbirliği, iş güvencesi, çalışma koşulları ve eğitimi faktörleri çalışan memnuniyeti açısından incelenmiştir. Müşteriler genelde aynı çalışanları görmek istemektedir, istikrar müşteriler için önemlidir, dolayısıyla yüksek oranda işçi devri müşteri memnuniyetini etkilediğini ve bunun işletme için yüksek maliyete neden olduğu belirtilmiştir. Çalışan devri ile çalışan memnuniyeti arasında yakın bir ilişki olduğuna inanılmaktadır. Ayrıca, doğru eğitim iş memnuniyetine yol açar ve çalışanların işten ayrılma meyilleri daha az olacaktır. Bunun sonucunda çalışanlar işlerinde gurur duyacaklar ve kurum ile çalışanlar arasında kuvvetli bir bağ oluşacaktır. Çalışan memnuniyeti, insanlara iş yerinde önem verildiğinin gösterilmesi ile ortaya çıkar²⁶.

Chen, v.d, (2006) çalışmalarında yükseköğrenim kurumlarında çalışanların memnuniyetine odaklanmaktadır. Birçok çalışmalarda öğrenciler müşteri olarak ele alınır ve öğrencilerin bulundukları programlarda memnuniyet/memnuniyetsizlik seviyeleri incelenir. Çoğu zaman öğretmenlerin iş memnuniyetleri göz ardı edilir. Chen, v.d., (2006) çalışmasında çalışan mutsuzluğunun ilerleme önceliğini nasıl belirlediğini değerlendirmektedir. Bu çalışmanın analitik bulguları, yükseköğretim çalışanların yüksek ücretlere ve adaletli terfi sistemlerine odaklandığını göstermektedir. Ayrıca, Avrupa ve Amerika'daki kolej hocaların iş memnuniyet incelemeleri aynı sonuçları vermektedir. Bu çalışmadaki çalışan memnuniyeti modeli yükseköğretim sektöründe hem memnuniyet seviyelerini dikkate almaktadır hem de geliştirme stratejilerine karar vermede önem derecelerini de dikkate almaktadır. Çalışan öğretmenlerin memnuniyetini

²⁶ Loveday, M., "Employee satisfaction and HR at ISS", **Management Development Review**, vol.9 no. 4, 1996, pp. 9-12.

incelenmesinde organizasyon vizyonu, saygı, sonuç geribildirimi ve motivasyon, yönetim sistemi, ücret ve ek ödemeler ve çalışma ortamı gibi faktörler ele alınmıştır.

4.3 Çalışan Memnuniyeti Anket Amacı

Çalışan memnuniyeti anketlerinin amacı, analiz sonuçlarından iyileştirme kalite niteliklerini belirlemektir. Düşük kalite düzeylerine sahip durumlar genellikle iyileştirilmesi gerekir. Analiz sonuçlarından sonra çalışanların memnuniyetini artırmak için hangi kalite niteliklerinin geliştirilmesi konusu en önemli işlerdendir. Çalışanlar memnuniyet derecelerinin kalite niteliklerinin önemine dayanarak ölçtüklerinden, düşük kalite seviyesindeki bir niteliği seçerek iyileştirmek uygun bir yöntem değildir. Böylece, okulların iyileştirmesi gereken en önemli nitelikler çalışan öğretmenler tarafından önemli olarak derecelendirilen ve aynı zamanda memnuniyet skoru düşük olan nitelikler olmalıdır. Chen, v.d., (2006) bunun için en önemli uygulama modeli olan önem-memnuniyet (Importance-Satisfaction) modelini kullanmışlardır. I-S modelinde tüm kalite nitelikleri modele yerleştirilir ve iyileştirme stratejileri her bir niteliğin konumuna göre belirlenir²⁷.

4.4 Bilgi İşçilerinin Memnuniyeti

4.4.1 Bilgi Birikimi ve Örgüt Performansı

Çalışmalarda Bilgi Yönetimi uygulamaları (BY) ve işletme performansı arasındaki bağlantı incelenmektedir ve elde edilen sonuca göre BY uygulamaları ile firma performansı arasında kuvvetli ve pozitif bir ilişki var olduğu sonucu elde edilmiştir. Sonuçlar ayrıca, etkin bir BY stratejisi geliştirmek için insan boyutunun önemini vurgulamaktadır. Ayrıca, BY uygulamalarını benimseyen işletmeler rakiplerine

²⁷ Chen, H.S. , C. C. Yang, J. Y. Siau, H. H. Wang, “The development of an employee satisfaction model for higher education”, **The TQM Magazine**, vol. 18 no. 5, 2006, pp.484-500

göre daha iyi sonuçlar elde ettiği belirtilmiştir. BY alanındaki en büyük engellerden birisi önemli kavramların ölçülmesindeki yetersizliktir²⁸.

4.4.2 Bilgi Çalışanı tanımı

Bilgi işçisinin tanımı ve kapsamı konusunda çok farklı yaklaşımlar mevcuttur. Bu yaklaşımlar, bedeni çalışma ve mal üretiminin aksine zihni çalışma ve hizmet üretimi noktalarında kesişmektedir. Dura (1990) çalışmasında bilgi üretim ve dağıtımı, eğitim öğretim, araştırma ve geliştirme, iletişim ve enformasyon donanım ve hizmetleri kavramlarını bilgi işinin sınırlarına dâhil etmektedir.

Kelloway (2000) ve Davenport (2002) bilgi işini bilginin üretilmesi, uygulanması, ulaştırılması ve kazanılması olarak ve bilgi işçisinin de bilgiyi üreten, paylaşan ve günlük çalışma hayatında bu bilgileri kullanan işçiler olarak tanımlamaktadırlar.

Profesyonel hayat ve organizasyon için köklü değişiklikler ile birçok “bilgi işçisi” artık geleneksel disiplinler mesleklerin hiçbirine ait değildir. Disiplinler arası yaklaşımlar genellikle artık kabul görmektedir. Yüksek teknoloji çağında gerekli görülen yeni yetkinlikler kapsamlı iletişim, problem çözme ve koordinasyon becerilerinden oluşmaktadır. 21.yüzyılında iş ve istihdam tahminlerinde toplu istihdam fırsatlarının “bilgi işi” alanında olacağı belirtilmiştir. “Bilgi işi” geniş hizmet sektörünün aktiviteleri olan geçici çalışma ve dış kaynak kullanımı hizmeti, programlama, işletme ve yasal hizmetler, sağlık, mühendislik, mimarlık ve eğitim hizmetleri gibi alanları kapsamaktadır²⁹.

²⁸ Marques, Daniel Palacios and Fernando J. G. Simon, “The effect of knowledge management practices on firm performance”, **Journal of Knowledge Management**, Vol. 10, No. 3, 2006, pp. 143-145.

²⁹ Garrick, J., S., Clegg, “Knowledge Work and the New Demands of Learning”, **Journal of Knowledge Management**, vol. 4, number 4, 2000, pp. 279-286

“Beyaz-yakalı iş” olarak bilinen “Bilgi işi” terimi nispeten yeni bir kavramdır. İlk olarak Drucker (1959) tarafından kullanılmıştır ve “bilgi işçilerini (Bİ)” fiziki varlığı olmayan ya da maddi olmayan kaynaklar ile çalışan işçiler olarak tanımlamaktadır. Bir başka açıdan, “hizmet işçileri” terimi Bİ terimi ile ilişkilidir, herhangi bir zaman diliminde üretilen ve başka bir zamanda diliminde tüketilen mamul ürünlerin aksine, burada ürün üretilir ve aynı zamanda tüketilir (örneğin çağrı merkezleri temsilcileri) (Drucker, 1993). Sonraki yıllarda, Drucker (1994) bilgi işçisini, yeni bir ürün veya hizmet geliştirmek için örgün eğitim yolu ile kazanılan teorik ve analitik bilgiyi uygulayan yüksel seviyeli çalışanlar olarak tanımlamaktadır. Bilgi işçileri bazı iş özelliklerini paylaşırlar bile, potansiyel olarak önemli farklılıklar mevcuttur. Örneğin, beden işçileri (mavi yakalılar) ile bilgi işçileri arasında süreklilik mevcuttur, birçok iş her iki işçi grubunun bileşenlerini içerir. Bilgi işçileri, hem el işini hem de bilgi işini yapan “teknologlar” olarak tanımlanır ve bunlar bilgi işçileri grubu arasında en büyük ve en hızlı büyüyen tek grup olabilir³⁰.

Bilgi işçilerinin bir diğer alternatif sözlük tanımı ise yaşam mücadelesi için, bilginin kullanılması veya geliştirilmesi görevlerinde çalışan olarak tanımlanır. Bu tanım doğrultusunda, bir bilgi işçisinin görevleri planlama, elde edinme, araştırma, analiz etme, organize etme, depolama, programlama, dağıtma, pazarlama, karar verme ve son ürünü elde etmek için bilgi dönüşümünü gerektiren birçok diğer görevlerdir. Bu tanıma göre bilgi işçileri yöneticiler, mühendisler, muhasebeciler, avukatlar, finans uzmanları, sistem analistleri ve programcılar gibi meslekleri içerebilir³¹.

Ersel (2003) bilgi işçisini bilgiyi üreten, toplayan, depolayan, analiz ve koordine eden ve bilgiyle uğraşan çalışanlar olarak tanımlamaktadır.

³⁰ Drucker, P., “Knowledge-worker productivity: the biggest challenge”, **California Management Review**, Vol. 41 No. 2, 1999, pp. 79-85.

³¹ Horvath, D., “Knowledge worker definition”, (çevrimiçi) <http://http://searchcrm.techtarget.com/definition/knowledge-worker.html>, 2006

Bilgi işi tanımı halen anlaşılmasız olarak kalmaktadır. “Bilgi işi” kelimesi bir yazılımın yazılmasından reklam yazılımını kapsayacak şekilde kapsamlı bir şekilde kullanılmaktadır. Tanım, görevin kendisinden ziyade göreve neyin getirileceğine dönüşmektedir. Bu bilgiyi, perspektif, tecrübe, eğitim, beceri veya görevin sonucunun değerini artan şekilde belirten bir form haline getirir. Birçoğumuz kendimizi bilgi işçisi olarak adlandırmaktan kaçınmayacağız. Eğer küresel rekabetin değişen temelleri hakkındaki tahminler ciddiye alınırsa, sanayileşmiş bir ülkede bilgi işi tek sürdürülebilir ve en iyi kazanç yoludur. Bireylerin bilgiye nasıl yaklaştığının anlaşılması eksikliklerden bir tanesidir. Eğer bilgi yöneticileri değer katmak niyetindeler ise, bireylerin anlaşılması kritik öneme sahiptir. Dolayısıyla, bilgi iki açıdan kişiye has bir şeydir: kişilik insanların bilgiyi elde etmesinde, algılamasında, değer vermesinde ve kullanmasında kritik bir rol oynamaktadır, ikinci olarak bilginin oluşturulması bireylerin dünya görüşlerinden etkilenmektedir³².

Bilgi işi ve bilgi çağı akademik analize bir fren olduğunu iddia etmektedir. Bilgi işi ile ilgili bazı fikirler ve görüşler makul olmalarına karşın, bu terimleri çevreleyen belirsizlik ve karmaşıklık incelendiğinde bu argümanların etkisi ve kuvveti gittikçe zayıflama eğilimindedir. Aynı şekilde, bilgi işinin kritik öneme sahip olduğu ve yapıldığı örgütler bağlamında bilgi işini değerlendirirsek, bu argüman daha fazla zayıflama eğiliminde olacaktır. Eğer yeterince çaba harcanırsa, tamamen farklı sonuçlar için kavramlar geliştirilebilir ve kullanılabiliriz. “bilgi işi” gibi terimler etrafında tartışmak yerine , “bilgi işi” sadece belli bir azınlık gruba ait olduğunu iddia etmek yerine, tüm işçilerin bilgi işçisi olduğu, tümünün becerilere ve yeterli bilgiye sahip olduğu anlaşılmasından başlanılabilir. Örgütlerde çalışanlar ellerinde sınırlı kaynaklara sahip önemli kaynaklar olarak farz edildiği zaman, örgüt içerisindeki problemler ve yenilikler hakkında yeni ve farklı düşünme yollarını meydana getirecek kavramlar

³² William D. Guns, Liisa Välikangas, “Rethinking Knowledge Work: Creating Value through Idiosyncratic Knowledge”, **Journal of Knowledge Management**, Vol.1 No:4, 1998, pp. 287-293.

geliştirmek ve bunları uygulamaya girişmek mevcut popüler modellere ve yönetim ve örgütler hakkındaki düşünce tarzlarına tutulan sözcükler kullanmaktan daha iyidir³³.

4.4.2.1 UNESCO Tanımı

İnsan, kültür ve toplumun bilgi stokunu artırmak amacıyla herhangi bir yaratıcı sistematik faaliyetin yürütülmesi ve bu bilginin kullanılarak yeni uygulamaların tasarlanmasıdır. Temel araştırma, uygulamalı araştırma ve yeni araçlar ve süreç ürünleri meydana getiren deneysel geliştirme işleri bilgi işini içermektedir. Bilgi işçileri semboller ve kavramları manipüle ve organize ederler ve kendi akranları ile çalıştığı kurumlara kıyasla daha kuvvetli bir şekilde özdeşleşmektedirler, ayrıca kurumlarının uzun süreli başarısı için çok önemlidirler³⁴.

4.4.3 Bilgi işçilerinin gruplandırılması

Drucker (1994) bilgi işçilerini, yeni ürünler ve hizmetler geliştirmek amacıyla örgün eğitim yolu ile elde edilen teorik ve analitik bilgilerini uygulayan yüksek seviyeli çalışanlar olarak tanımlamaktadır. Yen işlerin çoğu endüstri işçilerinin sahip olmadığı veya bunları elde etmek için zayıf donanımlı olduklarından birçok yeterlilikler gerektirmektedir. Bundan dolayı, bunları elde etmek için sürekli olarak iyi bir eğitim alma alışkanlıklarına sahip olmaları gerekmektedir³⁵. Bilgi işçileri, çalışmak için ayrıca farklı yaklaşımlara ve farklı bir zihniyete sahip olmalıdırlar. Drucker bilgi işçisi grubunu çok geniş tutmuştur, öyle ki yöneticiler, öğretmenler, hekimler, ileri teknoloji uzmanları gibi gruplar dâhil olmaktadır. Drucker (1994) bilgi işçisi grubundaki çalışanların temel yaklaşımlarının, değerlerinin ve inançlarının önemli olduğunu vurgulamaktadır.

³³ Collins, D., “Knowledge Work or Working Knowledge ? Ambiguity and Confusion in the Analysis of the Knowledge Age”, **Employee Relations**, vol. 19, number 1, 1997, p. 41.

³⁴ UNESCO, Statistical Yearbook, Paris UNESCO, 1993, pp.12-13.

³⁵ Drucker, P., “The age of social transformation, **Atlantic Monthly**, 1994, pp.53-60.

Dove (1998) çalışmasında bilgi işçilerini üç grupta toplamıştır.

1-) Yeniliğe dayalı bilgi işinin oluşturulması. Mühendisler, yöneticiler ve yatırımcılar gibi işçiler, işlerini yapabilmeleri için yeniliğe bağımlıdır. Bu işçileri daha önceden belirlenmiş işleri yapmıyorlar, fakat işlerini ilk defa tanımlıyorlar ve yapıyorlar. Diğer bilgi işçilerinin işlerini yapabilmeleri için araçlar meydana getiriler.

2-) Geniş ve acil faydaya dayalı taşınabilir bilgi işi. Bu tür bilgi işçileri genel bir biçimde uygulamak üzere bilgiye sahiptirler. Bilgilerini çeşitli senaryolarda veya organizasyonlarda kullanabilirler. MBA mezunları ve yazılım programcıları bu bilgi işçileri sınıfına girerler. Bu işçiler işlemleri çalıştırmak için genel bilgilerini kullanırlar veya önceden tasarlanmış araçları kullanarak işlerini yaparlar. Birinci grupta tanımlana işçilere nazaran, bu gruptaki bilgi işçileri önceden oluşturulmuş işleri yapmak için bilgilerini kullanırlar.

3-) Dar fakat yüksek faydaya dayalı uzmanlık bilgi işi. Bu işçiler bir işi yapmak için ihtiyaç duyulan özellikli bilgiye sahiptirler. Yaptıkları işlerde uzman olarak kabul edilirler ve yaptıkları iş için özellikli uygulamalarda bilgiye sahiptirler, ayrıca bu bilgi işçilerinin bilgileri kolay bir şekilde diğer alanlara transfer edilebilir değildir. Özel yazılım dillerinde kodlar yazan programcılar bu tür bilgi işçisi grubunda örnek olarak gösterilebilir³⁶.

Yazarlar arasında bilgi işçilerinin gruplandırılması konusunda ortak bir nokta görülmemektedir. Örneğin, Coates (1986) bilgi işçilerini üç alternatif gruba ayırmaktadır: sekreterlik, mesleki ve idari. Bir kez ölçme sistemleri uygulandıktan sonra, birçok araştırmacı tüm farklı bilgi işçileri gruplarına ölçme sistemlerinin ayarlamasının zor olacağına inanmaktalar (Davenport, v.d., 2002 ; Dove, 1998).

³⁶ Dove, R., "The knowledge worker", **Automotive Manufacturing & Production**, Vol.110 No.6, 1998, pp.26-8

Sonuç olarak, bilgi işçisi tanımları özellikli değildir. Yazarların ortak bir tanımda bulunduğu görülmemektedir. Literatürün çoğu bilgi işini el işine görelili olarak tanımlamaktadır ve bilgi işçileri gruplarının içindeki herhangi farklılıkları ihmal eder. Ancak, bilgi işinin özellikleri hakkında bazı fikir birliği vardır. Örneğin birçok yazar bilgi işinin el işine nazaran daha az somut olduğuna inanmaktadırlar ve işçilerin beyinlerinin üretim araçlarını oluşturduğuna inanmaktalar (Drucker, 1999; Davenport, 2002).

4.4.4 Bilgi Çalışan Verimliliği ve kurum performansı

Ekonominin yapısı devamlı bir şekilde değişmektedir, öyle ki firmalar beden işçilerinin verimliliğine bağımlı iken, şirketler her geçen gün artan bir şekilde bilgi işçilerinin verimliliklerinde bel bağlamaktadırlar. Bilgi işçilerinin sayıları gittikçe arttığından, bunların verimliliklerinin stratejik planlamanın odağı olmalıdır. Çalışmalarda bilgi işçilerinin verimliliklerini ölçen evrensel olarak kabul edilen herhangi bir ölçüm yoktur. Ramirez ve Nembhard (2004) çalışmalarında, bilgi işçilerinin verimliliklerinin ölçümlerinin bir sınıflandırmasını incelemişlerdir.

Örtük bilgi ve kurumsal performans arasında ilişki araştırmacılar tarafından belli saha ve alanlarda incelenmiştir. Pathirage ,v.d., (2007) örtük bilginin inşaat alanında önemini vurgulamaktadırlar ve ayrıca örtük bilginin örgütsel performansa olan anlamlı katkılarını incelemişlerdir. Kuruluşlar insan ve bilgi kaynakları gibi ana kaynakları iyi kullanamadığı zaman, bu kaynakları israf edecektirler. Özellikle örtük bilgi (tacit knowledge) nispeten henüz tam olarak keşfedilmemiş olduğu kabul edilir ve daha iyi bir örgütsel performans için bu kaynakların yönetimi ve doğru anlaşılması büyük bir öneme sahiptir. Pathirage, v.d., (2007) çalışması inşaat sektörünün işgücü ve bilgiyle yoğunlaştırılmış yapısını ortaya çıkarmaktadır ve ayrıca inşaat alanında insan faktörü ve örtük bilginin anlam ve önemini vurgulamaktadır.

Bilgi işi (Knowledge Work) çok iyi araştırılmış konulardan olmasına rağmen bunun ölçülmesi oldukça zordur. Ayrıca ekip düzeyinde bilgi işi verimliliği etkileyen faktörler konusunda çok az fikir birliği sağlanmıştır. Bugünkü teoriler dağınık coğrafik bölgelerde ekiplerin çalışmasını ihmal etmektedir. Sijtsema, v.d., (2009) çalışmasında bilgi işi verimliliği konusundaki yönetim teorisinin farklı alanlarının (bilgi yönetimi, entelektüel sermaye ve öğrenme) teorilerini ve işi ve örgütsel psikolojiyi bütünleştirmişlerdir. Ayrıca bilgi işi verimliliğini etkileyen öğeler hakkında ve bilgi işi verimliliği alanında yapılan araştırmalar hakkında genel bir bakış sunmaktadır. Çalışmada üç soruya cevap aranmaktadır: Bilgi işi nedir? Bilgi işi verimliliği nedir? Küresel teknoloji şirketlerin dağınık ekiplerinde hangi bileşenler bilgi işi verimliliğini etkinleştirir ve hangileri engeller? Bilgi işi verimliliğini etkinleştiren veya engelleyen faktörlerin ekip görevleri, ekip yapısı ve süreçler, fiziksel, sanal ve sosyal çalışma alanları, örgütsel bağlam olduğu tespit edilmiştir. Bilgi işi verimliliğini etkileyen öğelerin bilinmesi sayesinde, şirketler dağınık bilgi işçilerinin verimliliğini geliştirmek için özellikli öğeleri canlandırabilir veya azaltabilir³⁷

4.4.5 Bilgi işçilerinin Memnuniyetini Etkileyen Faktörler

Literatürde çalışanların memnuniyetini etkileyen faktörlerden genel olarak incelenmiştir. Bilgi işçilerinin memnuniyetini etkileyen faktörler üzerinde çeşitli çalışmalar mevcuttur. Bu faktörler diğerlerinden farklı olduğu araştırmacılar tarafından incelenmiştir.

Narang, (2010) bilgi işçilerinin iş memnuniyetini ölçmek için güvenilir ve geçerli bir ölçek geliştirmiştir. Bilgi işçilerinin memnuniyetini etkileyen faktörler beş boyutta incelenmiştir: örgütsel destek, rekabetçi mükemmellik, baskıcı yönetim uygulamaları, adil ve şeffaf bir yönetim, denetim ve rehberlik. Bu çalışma bilgi işçilerinin iş

³⁷ Sijtsema, P. M., V. Ruohomaki, M. Vartiainen, “Knowledge work productivity in distributed teams”, *Journal of Knowledge Management* , vol. 13 no. 6, 2009, pp.533-546.

memnuniyetlerini etkileyen anahtar faktörlerin daha iyi anlaşılmasına yardımcı olmaktadır. Yapılan regresyon analizi sonucunda, bilgi işçilerinin memnuniyetini etkileyen faktörlerden olan rekabetçi mükemmellik ile adil ve şeffaf yönetim faktörleri diğer faktörlere göre çok daha ağır basmaktadır. Baskıcı yönetim ile denetim ve rehberlik önem sırasın göre en az öneme sahip faktörlerdendir ve aynı zamanda her ikisi de istatistiksel olarak anlamsızdırlar, ayrıca memnuniyet ile negatif olarak ilişkilidir. Örgütsel destek faktörü garip bir şekilde anlamsız bir boyuttur ve bilgi işçilerinin memnuniyeti ile negatif ilişkili olduğu regresyon analizi sonucunda görülmektedir. Bu beş faktörün göreceli önemi sıraları şu şekildedir, rekabetçi mükemmeliyetçilik, adil ve şeffaf yönetim, örgütsel destek, baskıcı yönetim uygulamaları, denetim ve rehberliktir³⁸.

Bilgi işçisi Peter Drucker tarafından ortaya atılmıştır, genellikle iş faaliyetleri içinde bilginin artırılması, manipüle edilmesi ve takas edilmesini içeren bir çalışanı tanımlamaktadır. İmalat sektörünün rekabetçi olabilmesi için, yüksek vasıflı ve iyi eğitilmiş çalışanlar yetiştirmek ve eğitmek zorundadır. Tam teşekküllü bilgi işçileri kadar olmasa da bunlara yakın seviyede olmalıdır. Çalışanların bilgi işçileri seviyesini yakalamaları için, DiGiovanni (2004) çalışmasında çeşitli düzeylerde çalışanlar için eşit şekilde etkili olacak olan eğitim ve gelişim aracı olarak “iş simülasyonu” uygulamasını fabrikada uygulanmak üzere keşfedildiğini örnek bir olay üzerinde göstermiştir. Bu yöntem geleneksel sınıf ve öğretmen tabanlı eğitim yöntemine alternatif bir yöntem olarak uygulanmıştır. Bu yöntemde yöneticiler ve çalışanlar beraber eğitime tabi tutuldular. Bu düzenleme sayesinde yönetim-çalışan ilişkilerini geliştirdi, takım çalışmasını teşvik etti ve heyecan duygusu yarattı. Bu şekilde yapılan eğitim³⁹ sayesinde uzun vadede bir anlayışın oluşturulması sağlanmış oldu. Sonuç olarak, çalışanlar bireysel kararlar ile ticari iş sonuçları arasında bir bağlantı olduğunu görmüş oldular. Çalışmada adı geçen ilgili fabrikada, yönetim değişimi zorlamadılar, bunun yerine uzun

³⁸ Narang, R., Alka Dwivedi, “Managing The Job Satisfaction of Knowledge Workers: An Empirical Investigation”, **Asia Pacific Journal of Business and Management**, Volume 1 (1), 2010, pp.1-5.

³⁹ DiGiovanni, M., “Developing Knowledge Workers on the Factory Floor”, **Development and Learning in Organizations**, vol. 18, number 4, 2004, pp. 9-12

vadede kabul ve anlayışa sebep olacak olana ortamı yönetmek için arayış içinde oldular. Sonuç olarak bu çalışmada çalışanların bilgi işçileri düzeyinde olmaları için gerekli olan faktörün devamlı değişimin direnilmesi yerine kucaklanmasıdır. Burada ayrıca, farklı bir eğitim sisteminin bir diğer önemli faktör olduğu ampirik olarak fabrika içinde yapılan örnek çalışmada gösterilmiştir.

DePablo ve Lyritas, (2008) çalışmalarına, örgütsel performansın artışına yönelik insan kaynakları yönetiminin faydalı olması için organizasyonel gereksinimlerin neler olduğunu detaylı bir şekilde derinlemesine literatür incelemesi yaparak analiz etmiştir. Ayrıca, öğrenen bir kuruluştaki insan kaynakları yönetiminin analizi için kavramsal bir çerçeve sağlama konusunu incelemiştir. Gelişmiş öğrenme ve semantik web teknoloji için gereksinimler çerçevesi en önemli katkılardandır, bu da çağdaş kuruluşlarda etkin yetkinlik yönetiminin stratejilerini yönlendirebilir. Gereksinimler çerçevesi bilgi yönetimi stratejilerinin kapsamlı bir spesifikasyonu için dayanak oluşturur⁴⁰. Bu çalışma faydalı bilgi kaynağı olmasının yanı sıra, stratejistler, insan kaynakları yöneticileri, bilgi işi çalışanları için tarafsız tavsiye niteliğindedir.

Çalışanların bilgi transferinin belirleyicileri olarak içe yönelik ve dışa yönelik motivasyonun etkileri Cruz, v.d., (2009) tarafından çalışılmıştır. Kâr amacı gütmeyen bir işletmenin analizini yapmak için vaka çalışması metodunu kullanılmıştır. Araştırma hipotezlerinin test etmek için PLS (partially least squares) yaklaşımı kullanılarak hem nitelleyici hem de sayısal analizler uygulanmıştır. Araştırma sonuçlarına göre, bilgi transferi içe dönük motivasyon sayesinde geliştiği görülmüştür, fakat dışa yönelik olan motivasyon faktörü bilgi transferi üzerinde anlamlı derecede etkili olmadığı saptanmıştır. İnsanlar, finansal ödüllerin aksine içe dönük motivasyondan dolayı kâr amacı gütmeyen kuruluşlar ile ilgilenmektedirler. Bilginin organizasyon içerisinde

⁴⁰ DePablos, P. O., M. Lytras, “Competencies and human resource management: implications for organizational competitive advantage”, **Journal of Knowledge Management** , vol. 12 no. 6, 2008, pp.48-56.

kalmasını garantiye almak için, örtük (tacit) bilgiyi açık (explicit) bilgiye dönüştürecek olan bir mekanizmanın tasarlanması yöneticilere önerilmektedir. İçe dönük motivasyon vasıtasıyla oluşturulan kurum bağlılığından dolayı, yazarlar yöneticilerin bu önerileri kendi bağlamlarında faydalı bulabileceklerini düşünmektedirler.

Bilgi yönetimi içinde anahtar bir unsur olan bilgi transferi son zamanlarda sürdürülebilir rekabet avantajı oluşturmada ve örgütsel verimlilikte önemli rol oynadığından dolayı takdir görmektedir. Çalışanlar, hem profesyonel hem de kişisel olarak ilerlemelerine imkân tanıyan uygun iş şartlarının sunulması yolu ile kendilerine çalıştıkları kurumların değer verdiğini algılamaları gerekmektedir. Çalışanlar ayrıca, yapmış oldukları işten dolayı bir esenlik duygusuna sahip olmalılar ayrıca çalışanların memnun olduklarını hissetmelidirler.

Dışa dönük motivasyon sonuçları yüksek kaliteli bireysel ve profesyonel hayat örnek olarak gösterilebilir. İçe dönük motivasyona, güzel bir çalışma ortamında yüksek seviyede özerklik ile çalışanların profesyonel gelişimlerini canlandırmalarına izin veren örgüte olan kuvvetli bağlılık, çalışanların etik ve ahlaki değerleri örnek olarak gösterilebilir. Bir organizasyon içerisindeki bilgi transferi çalışanların etkin bir şekilde beraber çalışmalarına yardımcı olur, bundan dolayı etkin yönetim için gereklidir. Yöneticiler, hem dışa yönelik hem de içe yönelik olan motivasyon sonuçlarını elde etmek için çalışanların arasında bilgi transferini maksimuma çıkarmak için motive edilmektedirler. Yöneticiler, örgüt verimliliğini artırmak için bir araç olarak bilgi transferini yapmaları konusunda çalışanlarını teşvik etmelidirler⁴¹.

Bilgi işçilerinin etkinliğini etkileyen faktörler üzerine çalışmalar yapılmıştır. Özerklik, dayanışma, ekip geliştirme, süreç ve bağlamsal (kavramsal) destek gibi faktörler ile bilgi işçi takımlarının etkinlikleri arasındaki ilişkinin nasıl olduğunu Janz

⁴¹ Cruz, N. M., V. M. Perez, C. T. Cantero, “The influence of employee motivation on knowledge transfer”, **Journal of Knowledge Management**, vol 13 no. 6, 2009, pp.478-480.

and Colquitt (1997) çalışmalarında incelemişlerdir. Elde edilen bulgulara göre taslak, süreç ve bağlamsal destek faktörleri arasında etkileşimin olduğu ve bu faktörlerin takım etkinliği üzerinde önemli etkileri olduğu görülmüştür. Özellikle, takım özerkliği ve takım iş motivasyonu arasındaki pozitif ilişki birbirine çok bağımlı koşullar altında çalışan takımlar olarak dönüşmüştür. Takıma verilen özerkliğin türen göre (planlama, ürün ve insanla ilişkili), etkileşimin etkisi değişmektedir. Gelişme açısından daha olgun takımlarda iş motivasyonu ve ekip süreci davranışları (yardımlaşma, paylaşma ve yenilik) arasındaki ilişkinin daha fazla pozitif olduğu görülmüştür. Bunların yanı sıra, süreç davranışları iler etkinlik arasında pozitif bir ilişki mevcuttur. Bu çalışma takım-tabanlı bilgi işçilerinin tasarlanması için yöneticiler tarafından pratik hayatta uygulanması önemli katkılar sağlayabilir, özellikle yöneticiler doğru şekilde ilgili süreç ve bağlamsal faktörleri belirleyebilirler⁴².

4.4.6 Bilgi çalışanı verimliliği

Yönetim biliminin 21. y.y. da yapacağı en önemli katkı bilgi işinin ve bilgi işçilerinin verimliliklerini artırılmasıdır. 21. Yüzyıl kurumun, ister ticari olsun ister ticari olmasın, en değerli varlığı ilgili kurumun bilgi işçileri ve bu çalışanların verimlilikleridir. Bilgi işçilerinin verimliliği üzerinde yapılan çalışmalar yeni sayılır. Geçmişte el işçilerinin verimlilikleri hakkındaki bilgiden çok daha fazla bilgi işçilerinin verimliliği hakkında bilgiye sahibiz, ayrıca birçok cevabı da biliyoruz. Henüz bilmediğimiz cevaplar ile ilgili zorlukları da biliyoruz ve bu konuda çalışılması lazım. Bilgi işçilerinin verimliliğini altı önemli faktör belirlemektedir⁴³.

a-) Bilgi işçisinin verimliliği “görev nedir? Ne olmalıdır? Katkıda bulunmak için ne beklemelisin? Görevini yaparken neler engel olmaktadır?” sorularının sorulmasını

⁴² Janz, B., Colquitt J., “Knowledge Worker Team Effectiveness”, **Personnel Psychology**, vol. 50, issue 4, 1997, pp. 55-62.

⁴³ Drucker, P., “Knowledge Worker Productivity : The Biggest Challenge”, **California Management Review**, vol. 41, issue 2, 1999, pp. 79-94.

gerektirmektedir. El işinin aksine, bilgi işi çalışanı programlamaz. Bilgi işinde mücadelede birinci gereklilik görevin ya da işin ne olduğunu tespit etmektir, böylece bilgi işçisi görevine odaklanması sağlanır ve elimine edilmesi gereken ne varsa mümkün olduğunca elimine edilir. Bu bilgi işçisinin görevin ne olduğunu kendisi tanımlar ve sadece bilgi işçisi kendisi bunu yapabilir.

b-) Bilgi işçilerinin bireysel verimlilikler için kendilerine sorumluluklar yüklenilir. Bilgi işçileri kendi kendilerini yönetmelidirler. Bilgi işçileri özerkliğe sahip olmalılar.

c-) Devamlı yenilik bilgi işçilerinin işinin, görevinin ve sorumluluğunun bir parçası olmalıdır.

d-) Bilgi işi, bilgi işçisinin tarafında devamlı öğrenmeği gerektirir. Fakat bilgi işi eşit şekilde devamlı eğitmeği gerektirir.

e-) Bilgi işçisinin üretkenliği, bilgi işçisinin “maliyet” yerine “varlık” olarak görülüp ve öylede muamele edilmesini gerektirir. Bilgi işçileri diğer tüm imkânlarla tercihen kurum için çalışmayı isterler.

4.4.7 Bilgi İşçilerinin Gruplandırılması

Bilgi işçileri, bireysel işlerde harcanan süreye dayalı olarak ya da sahip olunan bilgi türüne veya becerilere dayalı olarak çeşitli kategorilere halinde gruplandırılabilir. Bilgi işçilerinin farklı gruplara ayrılabilir olması gerçeği bilgi çalışanlarının çeşitli işlere sahip olduklarının bir göstergesidir. Bilgi işçileri, yenilikçi davranışlara karşı rutine davranışlarda faaliyette bulunulan zaman miktarına göre sınıflandırılabilir. Bazı çalışanlar, doğasında tekrarlayan ve rutine görevleri yerine getirmektedirler, fakat bazen bağımsız kararlar vermek için, çoğu zaman müşteri hizmetleri sorunları ile ilgili olarak, karmaşık bilgi kullanırlar. Diğer bir grup çalışanlar ise zamanlarının çoğunu bilgiye ulaşmada ve bilgi ile bağlantılı bağımsız kararlar vermede harcarlar. Bilgi işi ile uğraşan çalışanların ikinci şekilde sınıflandırılması şu şekildedir: özel bilgi işçileri, taşınabilir bilgi işçileri, bilgi oluşturan işçilerdir. Özel bilgi işçileri kurumun özellikli ürünleri veya

hizmetleri ile alakalı olarak önemli miktarda bilgiye sahiptirler. Bu bireyler kafalarında önemli şirket varlıklarını muhafaza eden birileri olarak düşünülebilir. Taşınabilir bilgi işçileri geniş ve anlık yardımcı bilgisine sahiptirler. Çeşitli kuruluşlar tarafından talep edilen bilgiye aşina çalışanlardır. Bilgisayar programcıları, kütüphaneciler ve işletme derecelerine sahip çalışanlar bu gruba örnek olarak gösterilebilir. Bilgi oluşturan çalışanlar çabalarının çoğunu ürün tasarımı ve geliştirilmesi gibi yenilikçi davranışlar üzerine odaklamaktadırlar. Bilim adamları ve bilgi sistemleri tasarımcıları bu gruba örnek olarak gösterilebilir⁴⁴.

4.4.8 Bilgi işi ve işçilerinin özellikleri

Bilgi işi karmaşık ve bu işle uğraşan çalışanların belli becerileri ve yetenekleri ile gerçek ve teorik bilgi ile aşinalıklara sahip olmaları gerekmektedir. Bu çalışanlar bilgiyi bulabilmeli, ulaşabilmeli, geri çağırabilmeli ve uygulayabilmesi gerekmektedir, ayrıca çalışanlar diğerleri ile iyi etkileşim içinde olmalı ve bu becerileri elde etmek ve geliştirmek için yetenek ve motivasyona sahip olmalıdırlar. Bu özelliklerin önemi işten işe göre değişmesine rağmen, bilgi işçileri bu temel niteliklere sahip olmak zorundadır. Birçok iş her zamankinden daha çok üniversite diploması gerektirmektedir ve bilgi işçilerinin eksikliği mevcuttur. Bilgi işi ile uğraşan araştırmacı bilim adamları ve tecrübeli fabrika müdürleri gibi bilgi işçilerinin emekliliğe ayrılıyor olacak olmaları bir diğer gelecek endişelerindendir⁴⁵.

- Gerçek ve teorik bilgiye sahip olma
- Bilgiyi bulma ve ona erişmek
- Bilgiyi uygulayabilme yeteneği
- İletişim becerileri
- Motivasyon

⁴⁴ William W. Prince, "Knowledge Workers"

(Çevrimiçi)<http://www.referenceforbusiness.com/management/Int-Loc/Knowledge-Workers.html>, 2011

⁴⁵ A.g.e., "Knowledge Workers" (Çevrimiçi)<http://www.referenceforbusiness.com/management/Int-Loc/Knowledge-Workers.html>, 2011

- Entelektüel yetenekler

Bilgi işçisinin belli başlı vasıfları olması gerektiği herkes tarafından bilinmektedir. Bu konuda üzerinde önemle durulmaktadır, örneğin Kelley (1990) bilgi işçisini “altın yakalı” işçiler olarak tanımlar ve bu bahsedilen altın yakalı bilgi işçileri problem çözme, yaratıcılık, yetenek ve zekâ gibi özelliklere sahip olduğundan dolayı işe alındığını vurgulamaktadır.

Kelley (1985) bilgi işçileri, bilgiyi kazanca dönüştürme yeteneklerine sahip olduklarından, bilgi çağında kurumun varlığını devam ettirmesi için çok önemli olduğunu belirtmektedir. Bilgi işçileri “altın yakalı” olarak adlandırılmasının sebebini örgütün ayakta kalması için bilgi işçilerinin bilgiyi önemli bir değere dönüştürmeleri olarak belirtilmiştir (Saravanamuthu, 2004). Altın yakalı işçiler olmalarının sebeplerinden, bilgi işçileri bilgiyi toplamaları, işlemeleri, analiz etmeleri ve yaymaları olarak belirtmiştir.

Stewart (1997) bilginin günümüzün rekabetçi bir silahı olduğunu ve doğal kaynaklardan, fabrikalardan, paradan daha değerli olduğunu belirtmektedir. Yeni yüzyılımızda, zenginlik bilginin ürünü olarak olduğunu vurgulamaktadır. Bilgi günümüzde ekonominin başlıca hammadde ve temel ürünleridir. Stewart (1997) “bilgi işini” hammadde ve ürünü bilgi olan işler olarak tanımlamaktadır.

Otomasyona uygun problemler ile ilgilenen “yöntemsel (prosedür ile ilgili)” işlerin aksine, bilgi işinin aşağıdaki gibi belirtilen özelliklere sahip bir yapısı vardır: çok kolayca tanımlanamayan; sınırları veya durdurma kuralları açık olmayan; herhangi bir özel durumda problem tanımı için soyutlama düzeyi hakkında karmaşık hüküm gerektiren; kuvvetli etik, politik, profesyonel boyutlara sahiptir; doğru ya da yanlış çözümleri yoktur, sadece iyi ya da kötü çözümlere sahiptirler; verilmiş çözümlere sahip değildirler, fakat keşfedilmeleri gerekmektedir, her bir tekrarlanma ya da deneme göz

önünde tutulur; onların nesnel başarı ölçüleri yoktur. Bilgi işi daha fazla disiplinler arası olurken, zor problemler ancak müzakere, tartışma, farklı perspektiflerin uzlaşılmasına meydan verecek takım elemanları arasında düşünüp taşınma yolu ile tatmin edici şekilde çözülebilir⁴⁶. Bilgi işi ve işçisinin tanımı ile alakalı tartışmaları daha ileri aşamaya götürmek yerine, bu kavramların anlamlarının, özelliklerinin anlaşılması daha önemli olmaktadır.

Bilgi toplumunda bilgi işçilerinin etkin işleyişlerine uygun özellikler mevcuttur. Temel seviyede amaç, bilgi teknolojilerinin veri ve bilgi işleme kapasitesi ile bu alanda çalışan insanların yaratıcılık ve yenilikçi kapasiteleri arasındaki birlikteliğin oluşturulması temel amaçtır. Bundan dolayı, bilgi işçileri yeni teknolojileri işlerinde uygulama konusunda becerikli olmalıdırlar. Böyle bir anlayış gereklidir böylece yaratıcılık ve yenilik gerektiren zamanlarını ve çabalarını katma değer faaliyetlerde yoğunlaştırmak için, programlanabilir işleri teknolojilere havale edebilirler. Daha da önemlisi bilgi işçileri örgütün “en iyi uygulamalarının” iş ortamının dinamiği ile uyumlu olup olmadığı konusunda karar verme kapasitesine sahip olmalıdırlar. Bahsedilen özelliklere sahip bilgi işçileri, örgütsel iş süreçleri içinde tasarlanması gerekli olan çift döngülü öğrenme ve öğrendiğini unutma devri için kritik elemanlardır. Eğer bilgi işçileri ticari girişim için çalışma şartlarındaki değişikliklerin etkilerini kavrarlarsa, örgütsel “en iyi uygulamalar” ile iş çevresinin dışsal gerçekliği arasında eş uyumlu (senkronize) olmaları için araç olabilirler. Öğrenme ve karar vermede özerklik için gereksinim göz önüne alındığında, bilgi işçileri ayrıca öz denetim ve kendi kendine öğrenme konularında rahat olma ihtiyaçları vardır. Bilgi işçileri yüksek derecede sorumluluk ve yetki gerektiren kurum içi girişimci suretinde hareket etmeleri gerekir ve ayrıca bunları

⁴⁶ Lang, J. C., “Managerial Concerns in Knowledge Management”, **Journal of Knowledge Management**, vol. 5, number 1, 2001, pp. 46-47

yerine getirmeleri için bilgi işçilerinin belli bir kapasite ve yeteneğe sahip olmaları gerekmektedir⁴⁷.

4.4.9 Performans

Zaim, v.d., (2007) bilgi yönetimi altyapısının ve bilgi yönetimi (BY) süreçlerinin etkilerinin BY uygulamalarının performansı üzerindeki etkilerini araştırmışlardır. Türkiye'deki GSM şirketlerinde çalışan bilgi işçileri ile yapılan birebir görüşmeye dayalı bir çalışmadır. Yapısal eşitlik modeli ile hipotezler test edilmiştir. Hipotezlerin testlerinin sonuçlarına göre hem BY süreçleri ve BY altyapısı pozitif ve anlamlı bir şekilde BY uygulamalarının performansını etkilemektedir. BY altyapısı BY performansını BY süreçlerine kıyasla daha fazla anlamlı şekilde etkilediği sonucu elde edilmiştir. Yani BY'nin arka planı ve içeriği BY'nin uygulama yönlerinden daha önemlidir. Çalışma vaka çalışması olduğundan genelleştirilemez, BY performansını etkileyen birçok faktör olabilir. Fakat BY performansının değerlendirilmesi, BY çabalarının etkinliğini, verimliliğini ve adaptasyonunu artırması beklenmektedir, böylece örgütün genel performansına daha fazla değer katacaktır. BY altyapısı ve süreçlerinin BY performansı üzerine yapılan ampirik çalışmaların kıtlığı nedeni ile, bu çalışma ilgili değişkenler arasındaki bağlantıyı açıklamak suretiyle önemli bir katkı sağlamaktadır⁴⁸.

Sürdürülebilir rekabetçi avantaj ve üstün performans sağlayan örtük bilgi, kurumlar için stratejik değerli bir kaynak olarak görülmesine rağmen, örtük bilgi ve kurumsal performans arasında nedensel bir ilişki kuran çok az ampirik kanıtlar ve araştırma çalışmaları mevcuttur. Bilgi yönetimi uygulamasında örtük bilgi stratejik bir faktör olarak kabul edilmektedir. Bundan dolayı, hem araştırmacılar hem de akademisyenler tarafından büyük ilgi gördü. Örtük bilginin etkin ve verimli şekilde

⁴⁷ Malhotra, Y., "Knowledge Management, Knowledge Organizations, Knowledge Workers", (Çevrimiçi) www.brint.com/interview/maeil.htm, 1998

⁴⁸ Zaim, H., E. Tatoglu and S. Zaim, "Performance of Knowledge Management Practices : a Casual Analysis", **Journal of Knowledge Management**, Vol. 11, No. 6, 2007, pp. 54-56.

yönetilmesi örgütler için bir anahtar başarı faktörü olmaya başlamıştır. Fakat, örtük bilginin sonuçlarını değerlendirmek ve bunun kurum performansı üzerindeki etkisini belirlemek bilgi yönetiminde en zorlayıcı konulardan bir tanesidir. Zaim (2007) diğer çalışmalara benzer olarak bu çalışmasında örtük bilginin kurum performansı üzerindeki etkilerini Türkiye’de ki bir bankadan elde edilen veriler yardımıyla araştırmıştır. Bu çalışmanın bulgularına göre örtük bilgi seviyesi ile finansal olmayan kurum performansı arasında pozitif doğrusal bir korelasyonun mevcut olduğunu onaylamaktadır. Örtük bilgi seviyesi görev bilgisi, uzmanlık bilgisi ve eğitim düzeyi olacak şekilde üç parçaya bölünmüştür. Kurum performansı üzerinde, görev bilgisi en etkili ve uzmanlık bilgisi ise ikinci en etkili faktörler olduğu belirlenmiştir. Ayrıca eğitim düzeyi nispeten finansal olmayan performans üzerinde daha az etkiye sahip olduğu görülmüştür⁴⁹.

⁴⁹ Zaim H., S. Zaim, "The Impacts of Tacit Knowledge on Organizations' Performance", 6'th International Congress on Knowledge, Management and Economy, İstanbul, Turkey, 2007, p.94.

BEŞİNCİ BÖLÜM

DVM VE KARAR AĞAÇLARI YÖNTEMLERİ KULLANILARAK BİLGİ ÇALIŞANININ KURUM PERFORMANSI ÜZERİNDEKİ ETKİSİNİN İNCELENMESİ

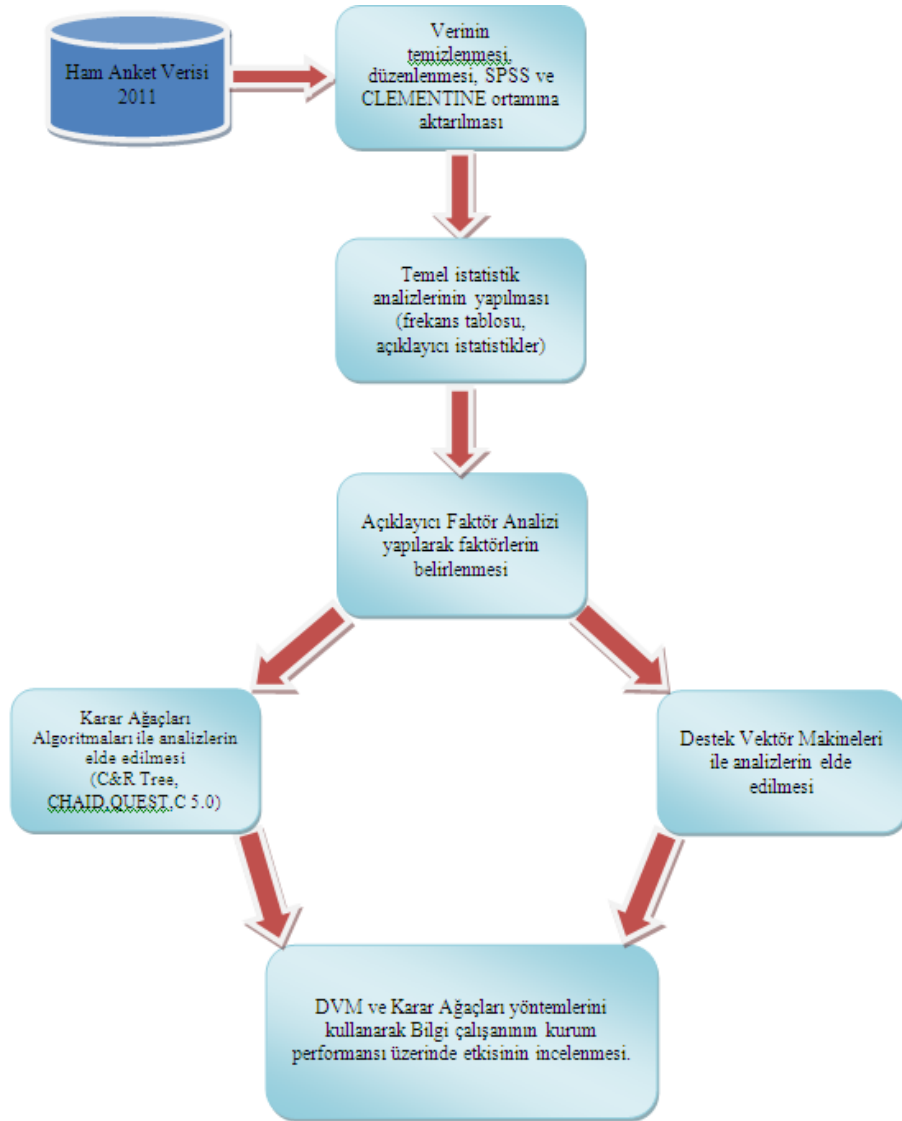
Çalışmada bilgi çalışanları olarak literatürde tanımlanan lisans, yüksek lisans, doktora mezunu çalışanlar ele alınmıştır. Kurum performansı başarılı ve başarısız olacak şekilde ikili sınıflandırma yapılmıştır. Uygulama için SPSS-Clementine¹ (v.12) modelleme programı kullanılmıştır. Verinin aktarılması ve modellemenin başarılı şekilde çalışması için IBM-SPSS (v.19)'dan faydalanılmıştır. Uygulama esnasında, faktör analizi, kümeleme analizi, destek vektör makineleri, karar ağaçları algoritmaları olan CHAID, C&R Tree, C 5 ve QUEST kullanılarak sonuçlar açık bir şekilde incelenmiştir.

5.1 Analiz Aşamalarının şema ile gösterimi

Elde edilen anket verisi kullanılarak şekil 14'de belirtildiği gibi aşamalar takip edilerek analiz başarılı bir şekilde gerçekleştirilecektir. Ham anket verisi ilk önce SPSS (v.19) istatistik programı ortamına aktarılacaktır. Clementine (v.12) SPSS ile ortak çalıştığından aynı veriler Clementine ile kolayca analiz yapılarak uygun modellerin belirlenmesi için Clementine programının kullanımına uygun hale getirilecektir. Sonraki aşamalarda ankette kullanılan değişkenler yardımı ile elde edilen kayıtların açıklayıcı temel istatistikî bilgileri elde edilerek incelenecektir. Ankette kullanılan bilgi çalışanı memnuniyeti değişkenleri (otuz dört adet) açıklayıcı faktör analizi uygulanarak bilgi çalışan memnuniyetinin faktörleri elde edilecektir. Faktör analizi sonrası, kurumsal performans değişkenleri finansal ve finansal olmayan performans olarak iki gruba ayrılmıştır. Bunun için Clementine v.12 yardımı ile finansal ve finansal olmayan

¹ SPSS Clementine®, Clementine® 12.0 Algorithms Guide. Integral Solutions Limited, 2007.

performansların ikili (binary) olacak şekilde başarısız (1) ve başarılı (2) olacak şekilde homojen bir sınıflandırma (cluster) analizi yapılacaktır. Son aşamalarda elde edilen bilgi çalışanın faktörlerinin kurumların finansal ve finansal olmayan performansları üzerindeki etkinlikleri DVM ve karar ağaçları modelleri kullanılarak detaylı incelemeler yapılacaktır.



Şekil 14. Analiz aşamaları

5.2 Verinin Elde Edilmesi

Bu tez çalışmasında uygulamada kullanılmak üzere değişik sektörlerde çalışan bilgi çalışanları üzerinde yapılan anket yolu ile veri kümesi elde edilmiştir. Lisans, Yüksek Lisans ve Doktora programlarından mezun olup halen değişik sektörlerde çalışan bilgi işçileri üzerinde çalışma yapılmıştır. 2300 bilgi çalışanı için önceden hazırlanan anket soruları gönderilmiştir, bunlardan 564 adet cevap geri dönmüştür. Bunlardan 56 doktora mezunu, 165 lisans mezunu ve 343 katılan yüksek lisans mezuniyet derecelerine sahip olan bilgi çalışanlarıdır.

5.3 Yöntem

Anket sorularının cevapları 1 (Kesinlikle Katılıyorum), 2 (Katılıyorum), 3 (Kararsızım), 4 (Katılmıyorum), 5 (Kesinlikle Katılmıyorum) olacak şekilde beşli likert ölçeği kullanılmıştır. Sorular en olumludan en olumsuz doğru derecelendirilmiştir. Tez çalışmasında kullanılan anket soruları EK-2’de verilmiştir. Ankette kullanılan ölçek Maltzer, (2007) tarafından geliştirilmiştir ve Zaim, H ve Zaim, S. (2007) çalışması baz alınarak bu çalışmada kullanılmıştır.

5.4 Keşfedici Faktör Analizi (Exploratory Factor Analysis) Sonuçları

Bilgi çalışanı memnuniyetini ölçmek için 34 değişken (soru) kullanılmıştır. İlgili sorular ve değişken isimleri tablo 3’te gösterilmektedir. Clementine v12. Ortamında yapılan faktör analizi sonuçlarına göre 34 değişken 5 faktör altında toplandığı görülmüştür. SPSS Clementine (v.12) ile elde edilen sonuçlar IBM SPSS (v.19) Programı ile de yapılarak sonuçlar karşılaştırılmıştır. Elde edilen bulgular birbirleri ile örtüştüğü görülmüştür.

Tablo 3. Çalışan Memnuniyeti Anket Soruları ve ilgili değişkenler

Değişken	Soru
M1	Çalıştığım kurumun bana değer verdiğini düşünüyorum.
M2	Çalıştığım kurumdaki görevimi ve sağladığım katma değeri biliyorum.
M3	Çalıştığım kurumda fikirlerim ve görüşlerim dikkate alınıyor ve onlara değer veriliyor.
M4	İş güvencesine sahip olduğumu düşünüyorum.
M5	Haklı bir gerekçe olmadıkça işime son verilmeyeceğini düşünüyorum.
M6	Kurumda yükselme ve terfilerin adil biçimde gerçekleştiğini düşünüyorum.
M7	Çalıştığım kurumda yükselme fırsatının olduğuna inanıyorum.
M8	Yöneticilerim, işimi daha iyi yapabilmem için bana yardımcı oluyorlar.
M9	İşimi düzgün yapabilmem için gerekli yetkiye sahibim.
M10	Kuruluşum, işimi düzgün yapabilmem için gerekli eğitim imkanlarını bana sağlamıştır.
M11	İşimi düzgün yapabilmem için gerekli geri beslemeyi yöneticilerimden alabiliyorum (olumlu/olumsuz uyarılar ve teşvikler)
M12	Çalıştığım kurumdan ayrılmayı düşünmüyorum.
M13	İş arkadaşlarımla kolektif çalışma yapabiliyoruz.
M14	İş arkadaşlarımla aramızdaki işbirliği ve yardımlaşma iyi düzeydedir.
M15	İş arkadaşlarım takım çalışmasına yatkındır.
M16	İş arkadaşlarımla müşterek kararlar alabiliyoruz.
M17	Yöneticilere fikirlerimizi korkmadan söyleyebiliyoruz.
M18	Yöneticiler hata yaptığımızda bizi tehdit etmeden ikaz ediyorlar.
M19	Genel olarak yöneticilerimize saygı duyuyoruz.
M20	Çalışanlar arasında güçlü bir iletişim sağlanmıştır.
M21	İhtiyaç duyduğumuz kişiye (yöneticiler dahil) istediğimiz zaman ulaşabiliyoruz.
M22	Ücretlerden memnunum.
M23	Ücret politikasının adil olduğunu düşünüyorum.
M24	Ücret dışı ödüllendirmelerden memnunum.
M25	Genel olarak işimden memnunum.
M26	Çalıştığım kurumdan memnunum.
M27	Çalıştığım işyeri ortamının fiziki olarak çalışmaya uygun olduğunu düşünüyorum.
M28	Çalıştığım işyeri ortamının sosyal açıdan (ilişkiler, kurum kültürü vs.) çalışmaya uygun olduğunu düşünüyorum.
M29	Daha iyi çalışma konusunda yöneticilerim tarafından teşvik edildiğimi düşünüyorum.
M30	Daha iyi çalışmam konusunda iş arkadaşlarım tarafından teşvik edildiğimi düşünüyorum.
M31	Genel olarak ücret dışındaki prim ve ikramiyelerden memnunum.
M32	Kurumun bana sağladığı sosyal imkânlardan memnunum.
M33	Kurumun bana sağladığı sosyal yardımlardan (eğitim, sağlık, kira vs.) memnunum.
M34	Çalıştığım Kurumu diğer arkadaşlarıma da tavsiye ediyorum.

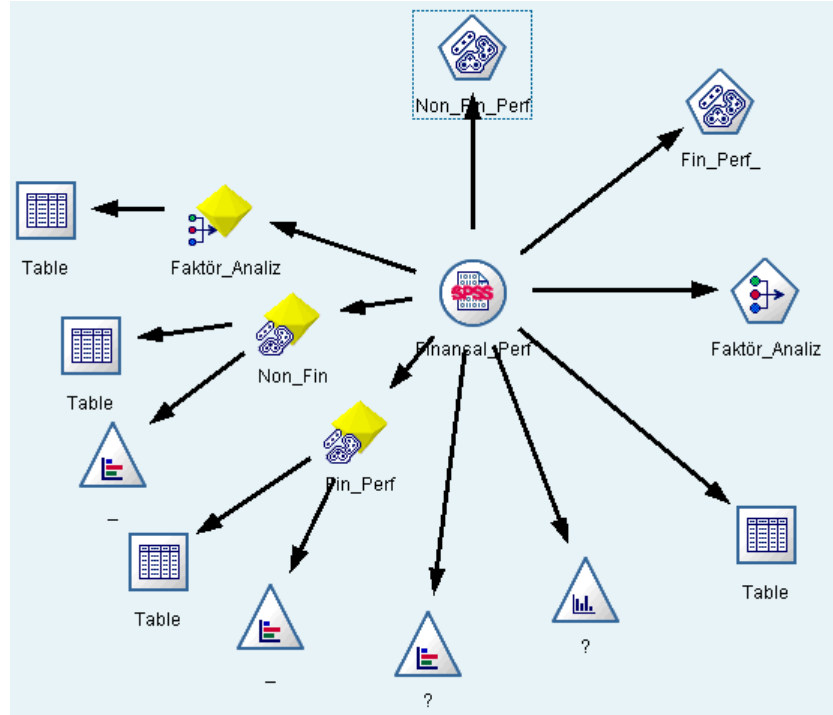
Faktör analizinde kullanılacak olan örneklem sayısı 564 adet olup analiz için her bir değişkenin 10 katı örneklem önerilmektedir (Nunnally, 1978). Dolayısıyla elimizdeki örneklem sayısı önerilen örneklem sayısından oldukça fazladır. Clementine V.12 ve IBM SPSS v.19 aynı anda kullanılarak faktör analizi yapılarak ilgili sonuçlar elde edilmiştir. Çalışmamda analiz boyunca Clementine V.12 kullandığımdan dolayı, buradan elde edilen sonuçlar analiz edilecektir. Kıyaslama yaptığımda, her iki program yolu ile elde edilen sonuçlar birbirleri ile benzeşmektedir.

Aşağıda SPSS Clementine Faktör analizi modeli kullanılarak, ilgili ayarlamalar yapılmıştır. Faktörleri elde etmek için ekstraksiyon yöntemi olarak “Temel Bileşenler (Principal Components)” yöntemi kullanıldı. Bir veri setinin temel boyutunun belirlenmesinde kullanılan yaklaşımlardan bir tanesidir. Temel bileşenler analizi orijinal veriyi doğrusal değişkenlerin bir kümesi olacak şekilde ayrıştırır. Veri içerisinde hangi doğrusal bileşenlerin mevcut olduğunu ve belli bir değişkenin o bileşene nasıl katkıda bulunacağını oluşturur. Temel bileşenler (Principal Component) analizi psikometrik olan sapasağlam bir yöntemdir ve diğerlerine nazaran kavramsal olarak daha az karmaşık bir teknik ve ayrıca diskriminant analizine birçok yönden benzediğinden dolayı tercih edilmektedir².

5.4.1 Keşfedici Faktör Analizi Modelinin Kurulması

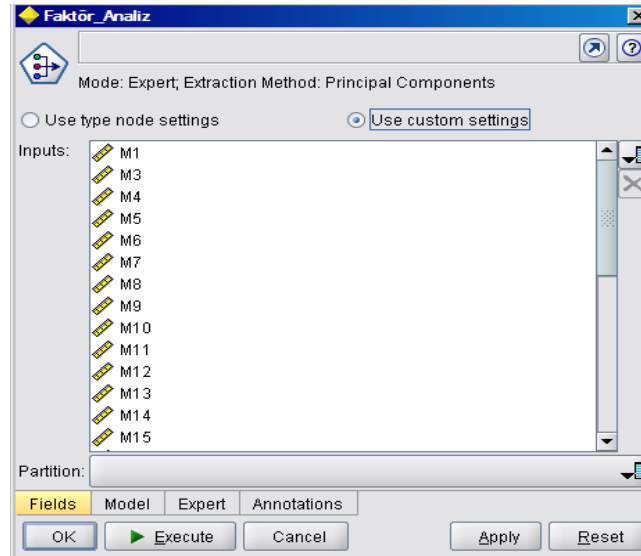
Faktör analizini başarılı bir şekilde değerlendirmek için, tablo 4’de görüldüğü gibi Clementine faktör analizi modelinin “expert” seçeneğini kullanarak, korelasyon matrisi, yaklaşım için maksimum iterasyon sayısı 25, ekstraksiyon faktörü için özdeğer (eigenvalue) seçeneği 1.0 ve üzeri, 0.4’ten küçük değerlerin faktör yüklemesinde gösterilmemesi seçeneği seçilmiştir ve ayrıca faktör rotasyonu için ise “Varimax” seçeneği işaretlenmiştir.

² Field, A., **Discovering Statistics Using SPSS**, London, Sage, 2005, pp. 630-632



Şekil 15. SPSS Clementine ile faktör analizi ve kümeleme analizi gösterimi

Tablo 4. SPSS Clementine Faktör Analiz Ayarları



Faktör_Analiz

Mode: Expert; Extraction Method: Principal Components

Model name: ☐ Auto ☒ Custom Faktör_Analiz

☐ Use partitioned data

Extraction Method:

Fields Model Expert Annotations

OK Execute Cancel Apply Reset

Faktör_Analiz

Mode: Expert; Extraction Method: Principal Components

Mode: ☐ Simple ☒ Expert

Missing values: ☒ Use only complete records

Fields: ☒ Correlation matrix ☐ Covariance matrix

Maximum iterations for convergence:

Extract factors:

☒ Eigenvalues over

☐ Maximum number

Component/factor matrix format:

☒ Sort values ☒ Hide values below

Rotation...

Fields Model Expert Annotations

OK Execute Cancel Apply Reset

Factor/PCA: Rotation

☐ No rotation ☐ Quartimax

☒ Varimax ☐ Equamax

☐ Direct oblimin ☐ Promax

Delta Kappa

OK Cancel Help

Faktör seçiminde “scree plot” gibi değişik görsel teknikler kullanılmasına rağmen, bunun yeterli olmadığı bilinmektedir. Kaiser (1960) faktör seçiminde öz değeri (eigenvalue) 1’den büyük olan faktörlerin seçilmesini önermektedir. Öz değer (eigenvalue) bir faktör tarafından açıklanan varyasyon miktarını belirtmektedir ve 1 öz değeri önemli bir varyasyon miktarını göstermektedir.

Faktörlerin birbirlerinden ayrışması ve anlaşılabilir olması için döndürme (rotasyon) uygulanması gerekmektedir. İki çeşit faktör döndürmesi mevcuttur: Dik döndürme (orthogonal rotation) ve eğik döndürme (oblique rotation). Dik (orthogonal rotation) da faktörler döndürülürken bağımsız tutulur. Döndürme öncesinde tüm faktörler bağımsızdırlar, yani birbirleri ile ilişkili değiller, dik döndürme tüm bu faktörlerin ilintisiz kalmasını sağlar. Eğik döndürmede ise faktörler birbirleri ile ilişkili olmaları sağlanır.

Teorik olarak faktörler bağımsızdırlar, yani birbirleri ile ilişkili değillerdir. Dolayısıyla, Field (2005)’e göre dik döndürme tekniklerinden bir tanesi seçilebilir. Çalışmada, faktörle birbirleri ile bağımsız oldukları göz önünde tutularak dik döndürme (orthogonal rotation) uygulandı ve dik döndürme tekniklerinden olan “varimax” tercih edildi. “varimax” faktörler içerisindeki dağılımı maksimize eder, böylelikle her bir faktör yüklemesine az sayıda değişken yükler ve daha kolay yorumlanabilir faktör kümeleri meydana getirir.

5.4.2 Örneklem Büyüklüğünün İncelenmesi (KMO Ve Barlett Testi)

Kullanılan örneklemin yeterliliği konusunda kullanılan bir başka teknik ise KMO (Kaiser- Meyer- Olkin) örneklem yeterlilik ölçütüdür. Analiz sonucuna göre, KMO örneklem yeterlilik ölçütü 0,96 olarak elde edilmiştir. Kaiser (1974)’e göre KMO test değeri 0,5 ve yukarısı kabul edilebilir, 0,5 – 0,7 arasının orta derece, 0,7 – 0,8 arasının iyi, 0,8 – 0,9 arasının çok iyi ve 0,9 ve yukarı KMO test değerinin örneklem yeterliliği

için süperdir. Buna göre çalışmada kullanılan örneklem yeterliliği 0,96 olduğundan oldukça yeterli olduğu KMO testi ile onaylanmıştır. Böylece faktör analizinin elimizdeki veri kümesine uygun olduğu konusunda emin olabiliriz.

Tablo 5. KMO ve Barlett's Test Sonucu

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.957
Bartlett's Test of Sphericity	Approx. Chi-Square	15379.236
	df	528
	Sig.	.000

Barlett ölçütü orijinal korelasyon matrisinin bir birim matris olup olmadığını test eder. Faktör analizinin uygulanması için değişkenler arasında bazı ilişkilerin olmasına ihtiyacımız var ve eğer R-matrisi birim matris ise, o zaman tüm korelasyon katsayıları sıfır olacaktır. Bundan dolayı, bu testin anlamlı olmasını istemekteyiz, yani anlamlılık düzeyinin %5 altında olması gerekmektedir. Anlamlı Barlett testi bize R-matrisinin birim matris olmadığını belirtir; böylece değişkenler arasında bazı ilişkilerin olduğunu gösterir. Bu veri kümesi için elde edilen Barlett testinin anlamlılık düzeyi oldukça anlamlılık olduğu görülmektedir ($p < 0,0001$) ve dolayısıyla faktör analizi veri kümemsi için uygun olduğunu belirtir.

5.4.3 Tanımlayıcı İstatistik Sonuçlarının değerlendirilmesi

Elde edilen açıklayıcı istatistik bilgilerine baktığımızda, faktör analizine dahil ettiğimi sorularda eksik değer (missing value) olmadığı görülmektedir. Her bir soru için M1'den M34'e kadar değişken ismi verildi ve bu değişkenlere karşılık gelen sorular tablo 3'de daha önce belirtilmiştir. Her bir sorunun ortalaması ve standart sapması tabloda verilmektedir. M5 (Haklı bir gerekçe olmadıkça işime son verilmeyeceğini düşünüyorum) sorusu en yüksek ortalama değere sahip olurken, M24 (Ücret dışı

ödüllendirmelerden memnunum) değişkeni ise en düşük ortalamaya sahiptir. Analize giren her bir değişken için veri sayısı 564 olduğundan, eksik değer bulunmamaktadır.

Tablo 6. Tanımlayıcı İstatistik Sonuçları

Descriptive Statistics			
	Mean	Std. Deviation	Analysis N
M1	3,576	1,174	564,000
M3	3,590	1,122	564,000
M4	3,798	1,206	564,000
M5	4,112	1,143	564,000
M6	2,888	1,329	564,000
M7	3,339	1,307	564,000
M8	3,496	1,216	564,000
M9	3,652	1,146	564,000
M10	3,433	1,223	564,000
M11	3,394	1,197	564,000
M12	3,512	1,341	564,000
M13	3,734	1,061	564,000
M14	3,835	1,025	564,000
M15	3,658	1,106	564,000
M16	3,759	1,064	564,000
M17	3,697	1,158	564,000
M18	3,583	1,196	564,000
M19	3,986	1,043	564,000
M20	3,358	1,110	564,000
M21	3,830	1,084	564,000
M22	3,144	1,289	564,000
M23	2,819	1,351	564,000
M24	2,544	1,384	564,000
M25	3,784	1,084	564,000
M26	3,693	1,198	564,000
M27	3,787	1,209	564,000
M28	3,598	1,202	564,000
M29	3,215	1,261	564,000
M30	3,287	1,181	564,000

M31	2,553	1,405	564,000
M32	2,920	1,414	564,000
M33	2,715	1,411	564,000
M34	3,372	1,300	564,000

5.4.4 R-Matrisi (Korelasyon Matrisi)

R-matrisinin birinci kısmı sorular arasındaki “Pearson” korelasyon katsayılarını vermektedir ve ikinci bölümde ise bunların tek-kuyruklu anlamlılık düzeylerini göstermektedir. Faktör analizini yapabilmemiz için sorular kendi aralarında ilişkili olacak fakat çok yüksek derecede ilişkili olmaması gerekmektedir. Aşağıdaki matrise baktığımızda tüm anlamlılık düzeyleri %5’ten küçük ($p < 0.05$) ve korelasyon katsayıları arasında %90’dan büyük sorular görülmemektedir. Matrisin determinantı sıfır olmadığından çoklu bağlantı (Multicollinearity) problemi verimizde mevcut değildir. Sorulardan hiçbirisi problem çıkarmadığından analizden soru çıkarmamıza gerek yoktur.

Tablo 7. Korelasyon Matrisi (R-Matrisi)

Correlation Matrix(a)																		
	M1	M3	M4	M5	M6	M7	M8	M9	M10	M17	M18	M19	M20	M30	M31	M32	M33	M34
M1	1	0,731	0,361	0,336	0,599	0,553	0,599	0,568	0,532	0,488	0,536	0,473	0,494	0,489	0,413	0,471	0,415	0,625
M3	0,731	1	0,382	0,364	0,598	0,553	0,596	0,584	0,47	0,532	0,515	0,437	0,489	0,486	0,381	0,411	0,354	0,547
M4	0,361	0,382	1	0,637	0,306	0,366	0,261	0,287	0,244	0,282	0,268	0,195	0,255	0,222	0,245	0,264	0,238	0,283
M5	0,336	0,364	0,637	1	0,379	0,372	0,336	0,351	0,199	0,372	0,36	0,299	0,26	0,241	0,239	0,182	0,182	0,314
M6	0,599	0,598	0,306	0,379	1	0,704	0,65	0,527	0,485	0,435	0,525	0,462	0,498	0,509	0,472	0,486	0,41	0,588
M7	0,553	0,553	0,366	0,372	0,704	1	0,596	0,482	0,437	0,361	0,403	0,383	0,446	0,396	0,405	0,446	0,369	0,546
M8	0,599	0,596	0,261	0,336	0,65	0,596	1	0,579	0,554	0,513	0,621	0,598	0,517	0,546	0,424	0,448	0,42	0,63
M9	0,568	0,584	0,287	0,351	0,527	0,482	0,579	1	0,492	0,536	0,533	0,434	0,487	0,47	0,347	0,375	0,302	0,49
M10	0,532	0,47	0,244	0,199	0,485	0,437	0,554	0,492	1	0,365	0,455	0,466	0,51	0,49	0,407	0,496	0,467	0,56
M17	0,488	0,532	0,282	0,372	0,435	0,361	0,513	0,536	0,365	1	0,623	0,477	0,478	0,4	0,343	0,327	0,333	0,504
M18	0,536	0,515	0,268	0,36	0,525	0,403	0,621	0,533	0,455	0,623	1	0,592	0,493	0,418	0,397	0,382	0,37	0,561
M19	0,473	0,437	0,195	0,299	0,462	0,383	0,598	0,434	0,466	0,477	0,592	1	0,542	0,448	0,342	0,345	0,324	0,544
M20	0,494	0,489	0,255	0,26	0,498	0,446	0,517	0,487	0,51	0,478	0,493	0,542	1	0,603	0,382	0,425	0,383	0,557
M30	0,489	0,486	0,222	0,241	0,509	0,396	0,546	0,47	0,49	0,4	0,418	0,448	0,603	1	0,442	0,49	0,389	0,585
M31	0,413	0,381	0,245	0,239	0,472	0,405	0,424	0,347	0,407	0,343	0,397	0,342	0,382	0,442	1	0,696	0,696	0,573
M32	0,471	0,411	0,264	0,182	0,486	0,446	0,448	0,375	0,496	0,327	0,382	0,345	0,425	0,49	0,696	1	0,795	0,65
M33	0,415	0,354	0,238	0,182	0,41	0,369	0,42	0,302	0,467	0,333	0,37	0,324	0,383	0,389	0,696	0,795	1	0,619
M34	0,625	0,547	0,283	0,314	0,588	0,546	0,63	0,49	0,56	0,504	0,561	0,544	0,557	0,585	0,573	0,65	0,619	1
Sig. (1-tailed)																		
M1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M3	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M4	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M5	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
M6	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0
M7	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
M8	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0
M9	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0
M10	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0
M17	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0
M18	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0
M19	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0
M20	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0
M31	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
M32	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0
M33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0
M34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0
a. Determinant = 7,62E-013																		

5.4.5 Faktörlerin Öz vektör Değerleri

R-matrisinin öz değerleri kullanılarak veri içerisindeki bileşenleri (öz vektörleri) hesaplanır. Clementine ve SPSS varsayılan olarak Kaiser kriterini kullanarak öz değer 1'den büyük olan faktörleri çıkarır. Tablo 8'de görüldüğü gibi her bir bileşenin çıkarım (extraction) öncesi ve sonrası ile döndürme (rotation) sonrası öz değerini göstermektedir.

Elde edilen sonuca göre, döndürme öncesinde, birinci faktör toplam varyansın %48'ini, ikinci faktör %7.9, üçüncü faktör %5.6, dördüncü faktör %4.1 ve beşinci faktör ise %3.2'sini açıklamaktadır. Döndürme (rotation) faktör yapılarını optimize etme etkisinden dolayı, döndürme sonrası sırasıyla faktörlerin toplam varyans açıklama oranları aşağıdaki gibi değişmektedir: %18.9, %16.5,%15.0, % 11.4 ve %7.1.

Tablo 8. Faktörlerin Öz vektör Değerleri

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	15.874	48.102	48.102	15.874	48.102	48.102	6.237	18.9	18.9
2	2.608	7.902	56.004	2.608	7.902	56.004	5.431	16.456	35.356
3	1.846	5.594	61.598	1.846	5.594	61.598	4.953	15.008	50.364
4	1.349	4.088	65.686	1.349	4.088	65.686	3.771	11.426	61.79
5	1.064	3.225	68.911	1.064	3.225	68.911	2.35	7.121	68.911
6	0.863	2.616	71.527						
7	0.829	2.513	74.04						
8	0.698	2.116	76.157						
9	0.664	2.012	78.169						
10	0.562	1.703	79.872						
11	0.533	1.615	81.487						
12	0.504	1.528	83.015						
13	0.492	1.492	84.507						
14	0.456	1.38	85.887						
15	0.45	1.364	87.251						
16	0.397	1.202	88.453						
17	0.349	1.056	89.51						
18	0.346	1.048	90.558						
19	0.324	0.983	91.541						
20	0.297	0.9	92.44						
21	0.282	0.856	93.296						
22	0.263	0.798	94.094						
23	0.238	0.721	94.816						
24	0.231	0.7	95.516						
25	0.222	0.671	96.187						
26	0.199	0.603	96.79						
27	0.186	0.564	97.354						
28	0.172	0.521	97.875						
29	0.16	0.485	98.36						
30	0.156	0.474	98.834						
31	0.146	0.442	99.276						
32	0.131	0.397	99.673						
33	0.108	0.327	100						
Extraction Method: Principal Component Analysis.									

5.4.6 Oransal Ortak Etken Varyans (communality)

Oransal ortak etken varyans tablosu faktör çıkarma (extraction) öncesi ve sonrası varyansı göstermektedir ve bir değişken içindeki ortak varyans oranı olarak tanımlanır.

Çıkarım öncesi tüm varyans ortaktır, dolayısıyla çıkarım öncesi ortak etken varyans değerlerinin (communality) 1'dir. Faktörlerin çıkarımları tamamlandıktan sonra, gerçekte ne kadar varyansın ortak olduğunu görmekteyiz. Örneğin birinci soru ile ilgili varyansın %62,2'si ortaktır. Aynı mantıkla diğer değişkenlerin oransal ortak etken varyansları yorumlanır.

Oransal Etken Varyans (Communality) tablosu faktörler tarafından açıklanan varyans oranı açısından da yorumlanabilir. Yani, faktör çıkarım (extraction) öncesi değişkenler kadar faktör mevcut olduğundan tüm varyans faktörler tarafından açıklanmakta ve dolayısıyla tüm oransal etken varyans değerleri 1'dir. Bunun yanı sıra, faktör çıkarım sonrası bazı faktörler iptal edildiği için ya da bilgiler kayıp olduğu için, mevcut faktörler tüm varyansı açıklamamaktadır fakat bazı oranlarda açıklamaktadır. Faktör çıkarım sonrası her bir değişkendeki elde tutulan faktörler tarafından açıklanan varyans miktarı tabloda verilmektedir.

Tablo 9. Oransal Ortak Etken Varyans (communality)

Communalities		
	Initial	Extraction
M1	1.000	0.622
M3	1.000	0.603
M4	1.000	0.755
M5	1.000	0.729
M6	1.000	0.628
M7	1.000	0.592
M8	1.000	0.721
M9	1.000	0.549
M10	1.000	0.508
M11	1.000	0.705
M12	1.000	0.527
M13	1.000	0.819
M14	1.000	0.863
M15	1.000	0.868
M16	1.000	0.857
M17	1.000	0.626
M18	1.000	0.698
M19	1.000	0.577
M20	1.000	0.644
M21	1.000	0.626

M22	1.000	0.684
M23	1.000	0.649
M24	1.000	0.767
M25	1.000	0.646
M26	1.000	0.745
M27	1.000	0.639
M28	1.000	0.725
M29	1.000	0.749
M30	1.000	0.649
M31	1.000	0.796
M32	1.000	0.739
M33	1.000	0.697
M34	1.000	0.740
Extraction Method: Principal Component Analysis.		

5.4.7 oğaltılmış Korelasyon Matrisi (Reproduced Correlation Matrix)

Değişken sayısı çok olduğundan ve elde edilen matrisin büyüklüğünden dolayı, aşağıdaki matris uygun bir şekilde düzenlenmiştir. Modelimizin uygunluğunu belirlemek için gözlenen korelasyonlar ile modele dayanan korelasyonlar arasındaki farka bakabiliriz. Bu farklar Tablo.10'daki matrisin rezidüel (residual) kısmında verilmiştir. İyi bir model için bu kısımdaki sayıların küçük olması gerekmektedir. Yani bu sayıların 0,05 'ten küçük olmasını istiyoruz. Büyük bir matrisi taramak yerine, matrisin altındaki kısımda verilen, mutlak değerce 0,05'ten büyük olan farkların sayısını vermektedir. Field (2005)' e göre, bu oranın %50'den küçük olması gerekir. Dolayısıyla bu farkların sayısı ne kadar sıfıra yakınsa (%) modelimiz o kadar iyi demektir. Bizim modelimizin çoğaltılmış (reproduced) korelasyon matrisine baktığımızda, 81 adet yani %15 gibi küçük bir oranda farkları mutlak değerce 0.05'ten büyük olanların sayısını ve oranını vermektedir. %50'den oldukça küçük olduğundan modelimiz uygun bir modeldir.

Tablo 10. Çoğaltılmış Korelasyon Matrisi (Reproduced Correlation Matrix)

	Reproduced Correlations												
		M1	M3	M4	M5	M17	M18	M19	M30	M31	M32	M33	M34
Reproduced Correlation	M1	.622a	.603	.394	.425	.509	.573	.530	.508	.425	.463	.408	.628
	M3	.603	.603a	.424	.463	.527	.567	.521	.497	.355	.384	.333	.576
	M4	.394	.424	.755a	.716	.290	.217	.146	.209	.237	.216	.180	.300
	M5	.425	.463	.716	.729a	.394	.333	.233	.198	.222	.170	.153	.295
	M17	.509	.527	.290	.394	.626a	.630	.549	.432	.368	.300	.307	.466
	M18	.573	.567	.217	.333	.630	.698a	.617	.447	.387	.354	.348	.545
	M19	.530	.521	.146	.233	.549	.617	.577a	.486	.349	.356	.331	.536
	M20	.520	.523	.225	.245	.503	.509	.525	.634	.404	.438	.389	.581
	M30	.508	.497	.209	.198	.432	.447	.486	.649a	.433	.494	.431	.606
	M31	.425	.355	.237	.222	.368	.387	.349	.433	.796a	.728	.736	.607
	M32	.463	.384	.216	.170	.300	.354	.356	.494	.728	.739a	.706	.658
	M33	.408	.333	.180	.153	.307	.348	.331	.431	.736	.706	.697a	.597
	M34	.628	.576	.300	.295	.466	.545	.536	.606	.607	.658	.597	.740a
Residualb	M1		.128	-.033	-.089	-.021	-.037	-.058	-.019	-.012	.008	.007	-.003
	M3	.128		-.041	-.099	.005	-.052	-.084	-.011	.025	.026	.021	-.029
	M4	-.033	-.041		-.079	-.008	.051	.049	.013	.008	.048	.058	-.017
	M5	-.089	-.099	-.079		-.023	.027	.067	.043	.017	.012	.028	.018
	M17	-.021	.005	-.008	-.023		-.007	-.071	-.032	-.024	.027	.026	.038
	M18	-.037	-.052	.051	.027	-.007		-.024	-.029	.010	.027	.023	.017
	M19	-.058	-.084	.049	.067	-.071	-.024		-.038	-.007	-.011	-.007	.008
	M20	-.026	-.034	.030	.015	-.025	-.016	.016	-.031	-.023	-.014	-.006	-.024
	M30	-.019	-.011	.013	.043	-.032	-.029	-.038		.010	-.004	-.042	-.021
	M31	-.012	.025	.008	.017	-.024	.010	-.007	.010		-.032	-.040	-.033
	M32	.008	.026	.048	.012	.027	.027	-.011	-.004	-.032		.090	-.008
	M33	.007	.021	.058	.028	.026	.023	-.007	-.042	-.040	.090		.021
	M34	-.003	-.029	-.017	.018	.038	.017	.008	-.021	-.033	-.008	.021	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are **81 (15.0%)** nonredundant residuals with absolute values greater than 0.05.

5.4.8 Faktörlerin Belirlenmesi (Rotated Component Matrix)

Faktörler arasında teorik olarak bir ilişki olmadığı ve birbirlerinden bağımsız olduklarını varsayımından hareketle dik döndürme (orthogonal rotation) tekniklerinden olan “varimax” kullanılarak Tablo 11 döndürülmüş bileşen matrisi (Rotated Component Matrisi) elde edildi. Stevens (1992)’e göre faktörlerin kolay yorumlanması için, faktör yüklemesi (factor loading) 0,4’den küçük olan değerler tabloda yer almamalıdır. Değişkenler faktör yüklemelerinin büyüklüklerine göre listelenmiştir. Bir soru tek kaldığından ve belirgin bir faktör oluşturmadığından analizden çıkartıldıktan sonra, bilgi çalışanı memnuniyeti anketindeki 33 adet değişken ile yapılan çalışmada, faktör analizi sonucunda beş faktörden oluştuğu görülmektedir. İlgili değişkenler ve ilgili faktör tabloda görülmektedir. Yönetimin tavrı, Ödüllendirme/Ücret, İş Arkadaşları, Çalışma Şartları ve İş güvencesi olmak üzere beş faktör mevcuttur. Elde edilen faktörler literatürde yapılan çalışmalar ile örtüşmektedir. “Yönetimin Tavrı” faktörü on iki değişkenden, “Ödüllendirme/Ücret” faktörü yedi değişkenden, “İş Arkadaşları” faktörü altı değişkenden, “Çalışma Şartları” faktörü altı değişkenden ve “İş Güvencesi” faktörü ise iki değişkenden oluşmaktadır. Analize başlamadan önce otuz dört adet değişken mevcut idi, ilk analiz sonrasında iki numaralı değişken olan M2 (Çalıştığım kurumdaki görevimi ve sağladığım katma değeri biliyorum) analizden çıkartıldı. Daha sonra otuz üç adet değişken ile faktör analizi adımları tekrar yapıldıktan sonra ilgili sonuçlar elde edilmiştir.

Tablo 11. Bilgi Çalışan Memnuniyeti Faktörlerin Belirlenmesi

Rotated Component Matrix(a)

Component					
Değişkenler	Faktör 1	Faktör 2	Faktör 3	Faktör 4	Faktör 5
	Yönetimin Tavru	Ödüllendirme/ Ücret	İş Arkadaşları	Çalışma Şartları	İş Güvencesi
Yöneticiler hata yaptığımızda bizi tehdit etmeden ikaz ediyorlar.	0,773				
İhtiyaç duyduğumuz kişiye (yöneticiler dâhil) istediğimiz zaman ulaşabiliyoruz.	0,712				
İşimi düzgün yapabilmem için gerekli geri beslemeyi yöneticilerimden alabiliyorum (olumlu/olumsuz uyarılar ve teşvikler)	0,702				
Yöneticilerim, işimi daha iyi yapabilmem için bana yardımcı oluyorlar.	0,696				
Yöneticilere fikirlerimizi korkmadan söyleyebiliyoruz.	0,689				
Genel olarak yöneticilerimize saygı duyuyoruz.	0,655				
Daha iyi çalışma konusunda yöneticilerim tarafından teşvik edildiğimi düşünüyorum.	0,601				
İşimi düzgün yapabilmem için gerekli yetkiye sahibim.	0,560				
Çalıştığım kurumda fikirlerim ve görüşlerim dikkate alınıyor ve onlara değer veriliyor.	0,528				
Çalıştığım kurumun bana değer verdiğini düşünüyorum.	0,513				
Kurumda yükselme ve terfilerin adil biçimde gerçekleştiğini düşünüyorum.	0,470				
Kuruluşum, işimi düzgün yapabilmem için gerekli eğitim imkanlarını bana sağlamıştır.	0,443				
Genel olarak ücret dışındaki prim ve ikramiyelerden memnunum.		0,848			
Ücret dışı ödüllendirmelerden memnunum.		0,812			

Kurumun bana sağladığı sosyal yardımlardan (eğitim, sağlık, kira vs.) memnunum.		0,781			
Ücretlerden memnunum.		0,764			
Kurumun bana sağladığı sosyal imkânlardan memnunum.		0,750			
Ücret politikasının adil olduğunu düşünüyorum.		0,723			
Çalıştığım kurumu diğer arkadaşlarıma da tavsiye ediyorum.		0,503			
İş arkadaşlarım takım çalışmasına yatkındır.			0,887		
İş arkadaşlarımla aramızdaki işbirliği ve yardımlaşma iyi düzeydedir.			0,878		
İş arkadaşlarımla müşterek kararlar alabiliyoruz.			0,865		
İş arkadaşlarımla kolektif çalışma yapabiliyoruz.			0,815		
Daha iyi çalışmam konusunda iş arkadaşlarım tarafından teşvik edildiğimi düşünüyorum.			0,602		
Çalışanlar arasında güçlü bir iletişim sağlanmıştır.			0,591		
Çalıştığım işyeri ortamının fiziki olarak çalışmaya uygun olduğunu düşünüyorum.				0,749	
Çalıştığım işyeri ortamının sosyal açıdan (ilişkiler, kurum kültürü vs.) çalışmaya uygun olduğunu düşünüyorum.				0,675	
Çalıştığım kurumdan memnunum.				0,542	
Genel olarak işimden memnunum.				0,501	
Çalıştığım kurumda yükselme fırsatımın olduğuna inanıyorum.				0,471	
Çalıştığım kurumdan ayrılmayı düşünmüyorum.				0,463	
İş güvencesine sahip olduğumu düşünüyorum.					0,840
Haklı bir gerekçe olmadıkça işime son verilmeyeceğini düşünüyorum.					0,796

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

5.4.9 Güvenilirlik Analizi (Cronbah's Alfa)

Güvenilirlik bir ölçümün tutarlılığını gösterir. Dolayısıyla çalışmada güvenilirlik analizi anket sorularının tutarlılığını ölçülmesinde kullanılır. Çalışmada kullanılan anketin geçerliliğini denetlemek için, faktör analizinde kullanılan ölçeğin güvenilirliğinin kontrol edilmesi faydalıdır³. Bunun için IBM SPSS v.19 ortamında, güvenilirlik analizi için “Cronbach's alpha” metodu kullanıldı. Öncelikle tüm değişkenleri (otuz üç adet) analize sokularak Cronbach's Alpha değeri bulundu, daha sonra faktör analizi sonucunda elde edilen beş faktör için ayrı ayrı Cronbach's alpha değerleri analize sokularak hesaplandı. Eğer birden fazla faktör mevcut ise, Cronbach (1951) güvenilirlik analizinin her bir faktör için ayrı olarak yapılmasını önermektedir. Bir anketin güvenilirliği değerlendirirken, Field (2005) eşik değer olarak Cronbach's Alpha değerinin 0,7 değerinin iyi ve 0,8 ve yukarı değerinin çok iyi olduğunu belirtmektedir.

5.4.9.1 Tüm Çalışan Memnuniyetinin Güvenilirliğinin İncelenmesi

Tüm değişkenler için yapılan güvenilirlik analizi sonucuna göre, Cronbach's Alpha değeri 0,97 olarak bulunmuştur. Field (2005) önerdiği eşik değerin çok yukarısında olduğu için, analizde kullanılan anket sorularının ölçeğinin güvenli olduğunu söyleyebiliriz. “Item-Total Statistics” tablosundaki son sütun olan “Cronbach's Alpha if Item Deleted” kısmı güvenilirlik analizi için önemli bir sütundur. Buranın iyi değerlendirilmesi gerekir. Burada amaç genel Cronbach's Alpha değerinin bu sütundaki değerlere yakın olması gerekir. Eğer bu sütundaki değerlerden bir veya birkaçı genel alfa değerinden çok farklı ise o ilgili soru veya sorular analizden çıkarılması gerekmektedir ve daha sonra faktör analiz yeniden yapılması gerekmektedir. Bu bilgilere göre, tüm soruların “Cronbach's Alpha if Item Deleted” değerleri ile genel

³ Field, A., **Discovering Statistics Using SPSS**, London, Sage, 2005, p.666.

Cronbach's alpha değerleri birbirlerine yakın olduğu için hali hazırdaki tüm sorular analizde kalmalıdır.

Tablo 12. Tüm Değişkenler için güvenilirlik analiz incelemesi

Reliability Statistics				
Cronbach's Alpha	N of Items			
0,965	33			
Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Çalıştığım kurumun bana değer verdiğini düşünüyorum.	110,0851	708,035	,735	,963
Çalıştığım kurumda fikirlerim ve görüşlerim dikkate alınıyor ve onlara değer veriliyor.	110,0709	712,137	,700	,963
İş güvencesine sahip olduğumu düşünüyorum.	109,8635	727,930	,398	,965
Haklı bir gerekçe olmadıkça işime son verilmeyeceğini düşünüyorum.	109,5496	728,159	,419	,965
Kurumda yükselme ve terfilerin adil biçimde gerçekleştiğini düşünüyorum.	110,7730	702,257	,728	,963
Çalıştığım kurumda yükselme fırsatımın olduğuna inanıyorum.	110,3227	707,665	,660	,964
Yöneticilerim, işimi daha iyi yapabilmem için bana yardımcı oluyorlar.	110,1649	704,774	,759	,963
İşimi düzgün yapabilmem için gerekli yetkiye sahibim.	110,0089	713,345	,664	,964
Kuruluşum, işimi düzgün yapabilmem için gerekli eğitim imkanlarını bana sağlamıştır.	110,2287	710,415	,665	,964
İşimi düzgün yapabilmem için gerekli geri beslemeyi yöneticilerimden alabiliyorum (olumlu/olumsuz uyarılar ve teşvikler)	110,2677	705,969	,753	,963
Çalıştığım kurumdan ayrılmayı düşünmüyorum.	110,1489	706,749	,656	,964
İş arkadaşlarımla kolektif çalışma yapabiliyoruz.	109,9273	716,639	,661	,964
İş arkadaşlarımla aramızdaki işbirliği ve yardımlaşma iyi düzeydedir.	109,8262	720,123	,620	,964
İş arkadaşlarım takım çalışmasına yatkındır.	110,0035	718,036	,608	,964
İş arkadaşlarımla müşterek kararlar alabiliyoruz.	109,9025	718,149	,632	,964
Yöneticilere fikirlerimizi korkmadan söyleyebiliyoruz.	109,9645	716,031	,612	,964
Yöneticiler hata yaptığımızda bizi tehdit etmeden ikaz ediyorlar.	110,0780	711,021	,672	,964

Genel olarak yöneticilerimize saygı duyuyoruz.	109,6755	718,255	,644	,964
Çalışanlar arasında güçlü bir iletişim sağlanmıştır.	110,3032	712,272	,706	,963
İhtiyaç duyduğumuz kişiye (yöneticiler dahil) istediğimiz zaman ulaşabiliyoruz.	109,8316	716,524	,648	,964
Ücretlerden memnunum.	110,5177	711,458	,614	,964
Ücret politikasının adil olduğunu düşünüyorum.	110,8422	707,995	,633	,964
Ücret dışı ödüllendirmelerden memnunum.	111,1170	705,450	,652	,964
Genel olarak işimden memnunum.	109,8777	710,832	,749	,963
Çalıştığım kurumdan memnunum.	109,9681	702,102	,815	,963
Çalıştığım işyeri ortamının fiziki olarak çalışmaya uygun olduğunu düşünüyorum.	109,8741	720,497	,514	,965
Çalıştığım işyeri ortamının sosyal açıdan (ilişkiler, kurum kültürü vs.) çalışmaya uygun olduğunu düşünüyorum.	110,0638	709,275	,696	,963
Daha iyi çalışma konusunda yöneticilerim tarafından teşvik edildiğimi düşünüyorum.	110,4468	699,040	,820	,963
Daha iyi çalışmam konusunda iş arkadaşlarım tarafından teşvik edildiğimi düşünüyorum.	110,3741	709,553	,705	,963
Genel olarak ücret dışındaki prim ve ikramiyelerden memnunum.	111,1082	705,400	,642	,964
Kurumun bana sağladığı sosyal imkanlardan memnunum.	110,7411	702,594	,677	,964
Kurumun bana sağladığı sosyal yardımlardan (eğitim, sağlık, kira vs.) memnunum.	110,9468	707,347	,612	,964
Çalıştığım kurumu diğer arkadaşlarıma da tavsiye ediyorum.	110,2890	697,158	,822	,963

5.4.9.2 Faktörlerin Ayrı ayrı Güvenilirliklerinin Hesaplanması

Cronbach (1951) tüm faktörlerin güvenilirliklerinin tek tek analiz edilmesini önermektedir. Bundan dolayı, elimizdeki beş faktörün güvenilirlik analizleri yapılarak aşağıdaki gibi her birinin “Cronbach’s alpha” değerleri elde edilmiş oldu:

Tablo 13. Faktörlerin güvenilirliğinin gösterimi

Faktör	Cronbach’s Alpa
Yönetimin Tavrı	0,94
Ücret/Ödüllendirme	0,90
İş Arkadaşları	0,93
Çalışma Şartları	0,87
İş Güvencesi	0,78

Beş faktörün “Cronbach’s alpa” değerleri eşik değer olan 0,7’den oldukça yüksek çıkmıştır. Çalışmada kullanılan anketin ölçümünün tutarlı olduğunu söyleyebiliriz. Her bir faktörün “Item-Total Statistics” tablolarını incelememiz gerekmektedir. Sorulardan bir veya birden fazlasının analizden çıkarılıp çıkarılmaması gerektiğini “Cronbach's Alpha if Item Deleted” sütununa bakarak tespit edebiliriz. Field (2005) önerdiği gibi, her bir değişkenin ilgili faktör deki “Cronbach's Alpha if Item Deleted” değeri ile ilgili faktörün genel “Cronbach’s Alpha” değeri arasında büyük bir fark olmadığı görülmektedir. Sorulardan hiçbirisi silindiği zaman, genel alfa değerini yükseltmeyeceğinden, analizimizdeki soruların çıkarılmaması gerekir.

Faktör 1: Yönetimin Tavrı:

Tablo 14. Yönetimin Tavrı faktörünün güvenilirlik analiz incelemesi

Reliability Statistics

Cronbach's Alpha	N of Items
0,938	12

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Çalıştığım kurumun bana değer verdiğini düşünüyorum.	38,764	100,614	0,739	0,622	0,932
Çalıştığım kurumda fikirlerim ve görüşlerim dikkate alınıyor ve onlara değer veriliyor.	38,750	101,683	0,728	0,624	0,932
Kurumda yükselme ve terfilerin adil biçimde gerçekleştiğini düşünüyorum.	39,452	99,026	0,704	0,534	0,934
Yöneticilerim, işimi daha iyi yapabilmem için bana yardımcı oluyorlar.	38,844	98,381	0,810	0,697	0,929
İşimi düzgün yapabilmem için gerekli yetkiye sahibim.	38,688	102,144	0,688	0,495	0,934
Kuruluşum, işimi düzgün yapabilmem için gerekli eğitim imkânlarını bana sağlamıştır.	38,908	102,045	0,642	0,471	0,936
İşimi düzgün yapabilmem için gerekli geri beslemeyi yöneticilerimden alabiliyorum (olumlu/olumsuz uyarılar ve teşvikler)	38,947	98,789	0,806	0,686	0,929
Yöneticilere fikirlerimizi korkmadan söyleyebiliyoruz.	38,644	102,876	0,646	0,492	0,935
Yöneticiler hata yaptığımızda bizi tehdit etmeden ikaz ediyorlar.	38,757	100,401	0,733	0,587	0,932
Genel olarak yöneticilerimize saygı duyuyoruz.	38,355	104,226	0,662	0,486	0,935
İhtiyaç duyduğumuz kişiye (yöneticiler dahil)	38,511	102,925	0,696	0,508	0,934

istediğimiz zaman ulaşabiliyoruz.					
Daha iyi çalışma konusunda yöneticilerim tarafından teşvik edildiğimi düşünüyorum.	39,126	97,645	0,809	0,678	0,929

Faktör 2: Ücret/Ödüllendirme:

Tablo 15. Ücret/Ödüllendirme faktörünün güvenilirlik analiz incelemesi

Reliability Statistics

Cronbach's Alpha	N of Items
0,896	7

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
İhtiyaç duyduğumuz kişiye (yöneticiler dahil) istediğimiz zaman ulaşabiliyoruz.	17,514	45,380	0,427	0,907
Ücretlerden memnunum.	18,200	39,255	0,735	0,876
Ücret politikasının adil olduğunu düşünüyorum.	18,525	38,850	0,720	0,878
Ücret dışı ödüllendirmelerden memnunum.	18,800	38,114	0,747	0,874
Kurumun bana sağladığı sosyal imkanlardan memnunum.	18,424	37,580	0,762	0,872
Kurumun bana sağladığı sosyal yardımlardan (eğitim, sağlık, kira vs.) memnunum.	18,629	37,992	0,736	0,876
Çalıştığım kurumu diğer arkadaşlarıma da tavsiye ediyorum.	17,972	39,058	0,741	0,875

Faktör 3: İş Arkadaşları:

Tablo 16. İş arkadaşları faktörünün güvenilirlik analiz incelemesi

Reliability Statistics

Cronbach's Alpha	N of Items
0,934	6

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
İş arkadaşlarımla kolektif çalışma yapabiliyoruz.	17,897	22,650	0,841	0,917
İş arkadaşlarımla aramızdaki işbirliği ve yardımlaşma iyi düzeydedir.	17,796	22,788	0,861	0,915
İş arkadaşlarımla takım çalışmasına yatkındır.	17,973	22,058	0,865	0,914

İş arkadaşlarımla müşterek kararlar alabiliyoruz.	17,872	22,396	0,868	0,914
Çalışanlar arasında güçlü bir iletişim sağlanmıştır.	18,273	23,446	0,706	0,934
Daha iyi çalışmam konusunda iş arkadaşlarım tarafından teşvik edildiğimi düşünüyorum.	18,344	22,897	0,706	0,935

Faktör 4: Çalışma Şartları

Tablo 17. Çalışma şartları faktörü güvenilirlik analizi incelemesi

Reliability Statistics				
Cronbach's Alpha	N of Items			
0,871	6			
Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Çalıştığım kurumda yükselme fırsatımın olduğuna inanıyorum.	18,374	23,574	0,600	0,862
Çalıştığım kurumdan ayrılmayı düşünmüyorum.	18,200	22,903	0,639	0,856
Genel olarak işimden memnunum.	17,929	23,874	0,741	0,839
Çalıştığım kurumdan memnunum.	18,020	22,286	0,812	0,824
Çalıştığım işyeri ortamının fiziki olarak çalışmaya uygun olduğunu düşünüyorum.	17,926	24,609	0,570	0,866
Çalıştığım işyeri ortamının sosyal açıdan (ilişkiler, kurum kültürü vs.) çalışmaya uygun olduğunu düşünüyorum.	18,115	23,385	0,694	0,845

Faktör 5: İş Güvencesi

Tablo 18. İş güvencesi faktörü güvenilirlik analizi incelemesi

Reliability Statistics				
Cronbach's Alpha	N of Items			
0,778	2			
Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
İş güvencesine sahip olduğumu düşünüyorum.	4,1117	1,307	0,637	.
Haklı bir gerekçe olmadıkça işime son verilmeyeceğini düşünüyorum.	3,7979	1,455	0,637	.

5.5 Tanımlayıcı İstatistik Analiz Sonuçları

Çalışmada kullanılan anket sonucu 564 bilgi çalışanı mevcuttur. Katılımcılardan erkeklerin oranı kadınlara göre oldukça yüksek gözükmemektedir. Erkekler yaklaşık %68 iken kadın katılımcı oranı %32'de kalmaktadır. Medeni durumda fark belirgin bir şekilde evli olanların oranı %66 ile bekâr olan bilgi çalışanlarına oranla yüksek çıkmıştır.

Yüksek lisans mezunu bilgi çalışanı sayısı 343 (%61) en yüksek orana sahip iken, lisans mezunu bilgi çalışanı %29 ve en az orana sahip olan bilgi çalışanı yaklaşık %10 ile doktora mezunu olan bilgi çalışanı grubudur.

Bilgi çalışanlarının meslek tecrübelerine bakıldığı zaman, 6-10 yıl arası çalışanların oranı %28 gibi diğerlerinden yüksek olmasına rağmen, 0-5 yıl ve 11-15 yıl meslek tecrübe gruplarının oranları arasında önemli derecede bir farkın görülmediği açıktır. Fakat 20 yıl ve üzeri çalışanların oranı diğerlerine göre az görülmektedir (%10).

Halen görev yapılan kurumdaki çalışma 0-5 yıl grubu %57 oranla en yüksek grubu oluşturmaktadır. Bu oran yarıya yarı düşerek 6-10 yıl arasında çalışan oranı %25'e düşmektedir 16 yıl ve üzeri gibi uzun süre bir kurumda çalışanların oranı oldukça azalmaktadır. Bunun sebebi, katılımcıların yani bilgi çalışanların oldukça genç bir grubu oluşturmamasından kaynaklanabilir.

Çalıştığı göreve isteyerek gelenlerin oranı %69'dur. Kısmi olarak evet diyenlerin oranı %23 iken, istemeyerek hali hazırda görevini yapanların oranı %8 gibi küçük bir oran elde edilmiştir. Bilgi çalışanlarının yapmış oldukları meslekler ile ilgili dışarıdan meslekleri ile ilgili eğitim alanların oranı %59 görülmektedir. Yani üniversite de alınan eğitim çağın gereklerini karşılamayabilir, özellikle bilgi çalışanları mesleklerini devam ettirebilmeleri için devamlı ekstra eğitim almaları gerekebilir. Bilindiği gibi bilgi

devamlı deęişken olduğundan, yıllar önce alınan eğitimler yeterli deęildir. Aslında bu oranın yüksek olması beklenir.

Katılımcıların pozisyonlarına baktığımızda orta düzey yöneticiler %32 ile en yüksek orana sahipler. Uzman ve memur pozisyonundaki bilgi çalışanları %26 olarak görölmektedir. Katılımcılardan üst düzey yönetici pozisyonundaki bilgi işçileri ve operasyon yöneticileri en az orana sahip gruplardır.

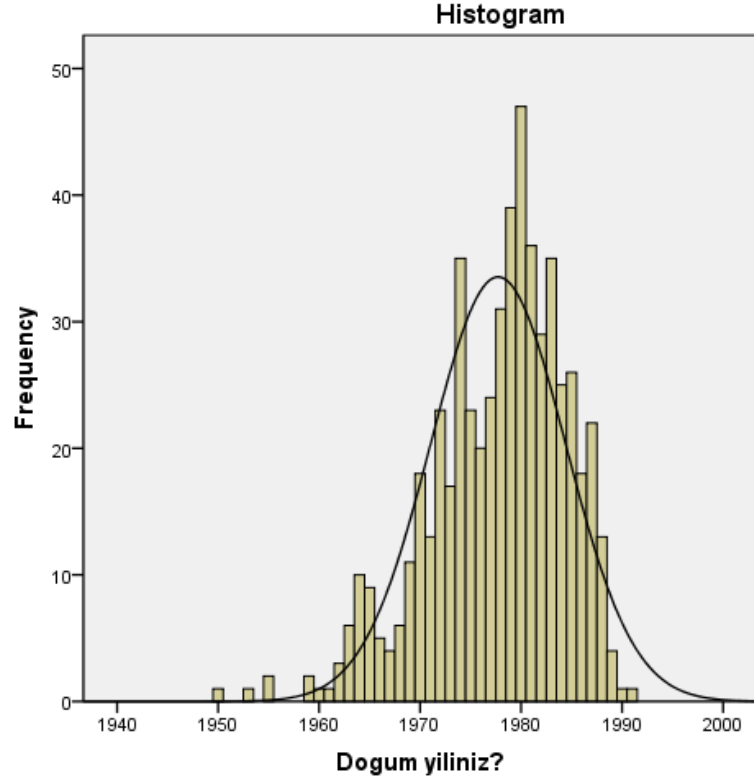
Tablo 19. Demografik Dağılım

Demografik Deęişken	Frekans	Yüzde
Cinsiyetiniz?		
Erkek	382	67.7
Kadın	182	32.3
Total	564	100
Medeni Durumunuz?		
Bekâr	193	34.2
Evli	371	65.8
Total	564	100
Eğitim Durumunuz?		
Doktora	56	9.9
Lisans	165	29.3
Yüksek Lisans	343	60.8
Total	564	100
Toplam Çalışma yılınız?		
20 yıl ve üstü	52	9.2
16-20 yıl	72	12.8
11-15 yıl	125	22.2
0-5 yıl	153	27.1
6-10 yıl	162	28.7
Total	564	100.0
Halen görev yaptığınız kurumdaki toplam çalışma yılınız?		
20 yıl ve üstü	15	2.7
16-20 yıl	29	5.1
11-15 yıl	61	10.8
6-10 yıl	138	24.5
0- 5 yıl	321	56.9
Total	564	100
Kaç yıldır mesleğinizi yapıyorsunuz:		

20 yıl ve üs	42	7.4
16-20 yıl	60	10.6
11-15 yıl	104	18.4
6-10 yıl	163	28.9
0- 5 yıl	195	34.6
Total	564	100
Çalıştığınız göreve isteyerek mi geldiniz?		
Hayır	47	8.3
Kısmen	130	23
Evet	387	68.6
Total	564	100
Mesleğinizle ilgili okul dışında mesleki eğitim aldınız mı?		
Evet	333	59
Hayır	231	41
Total	564	100
Pozisyonunuz?		
Üst düzey yönetici	42	7.4
Operasyon yöneticisi	43	7.6
Memur	147	26.1
Uzman	149	26.4
Orta düzey yönetici	183	32.4
Total	564	100
Doğum Tarihiniz?		
1980	47	8.3
1979	39	6.9
1981	36	6.4
1974	35	6.2
1983	35	6.2
1978	31	5.5
1982	29	5.1
1985	26	4.6
1984	25	4.4
1977	24	4.3
1972	23	4.1
1975	23	4.1
1987	22	3.9
1976	20	3.5
1970	18	3.2
1986	18	3.2
1973	17	3
1971	13	2.3
1988	13	2.3
1969	11	2
1964	10	1.8
1965	9	1.6
1963	6	1.1
1968	6	1.1

1966	5	0,9
1967	4	0,7
1989	4	0,7
1962	3	0,5
1955	2	0,4
1959	2	0,4
1950	1	0,2
1953	1	0,2
1960	1	0,2
1961	1	0,2
1990	1	0,2
1991	1	0,2
Eksik Değer	2	0,4
Toplam	564	100

Bilgi çalışanlarının yaşları ile ilgili bilgiler şekil 15 ve tablo 19’da görülmektedir. İki katılımcı yaşı ile ilgili bilgi vermemesine rağmen diğer bütün katılımcılar yaşlarını ankette belirtmişlerdir. Yaş değişkeni 1980 doğumlu bilgi çalışanlarının etrafında dağılmaktadır. Çok genç bilgi işçilerinin oranı ile 1955-1965 arası yaş grubu oranları diğerlerine kıyasla oldukça düşür görülmektedir. Yaş verilerine bakıldığı zaman mesleki tecrübe ile ilgili bilgiler ile örtüştüğü görülmektedir. Yani genç bilgi çalışanları aynı meslekte uzun süre kalması mümkün değildir. Bir diğer değişle, katılımcılarda yaşlı oran az olduğundan çalıştığı meslekteki tecrübe oranı değişkeni düşük oran vermesi doğaldır.



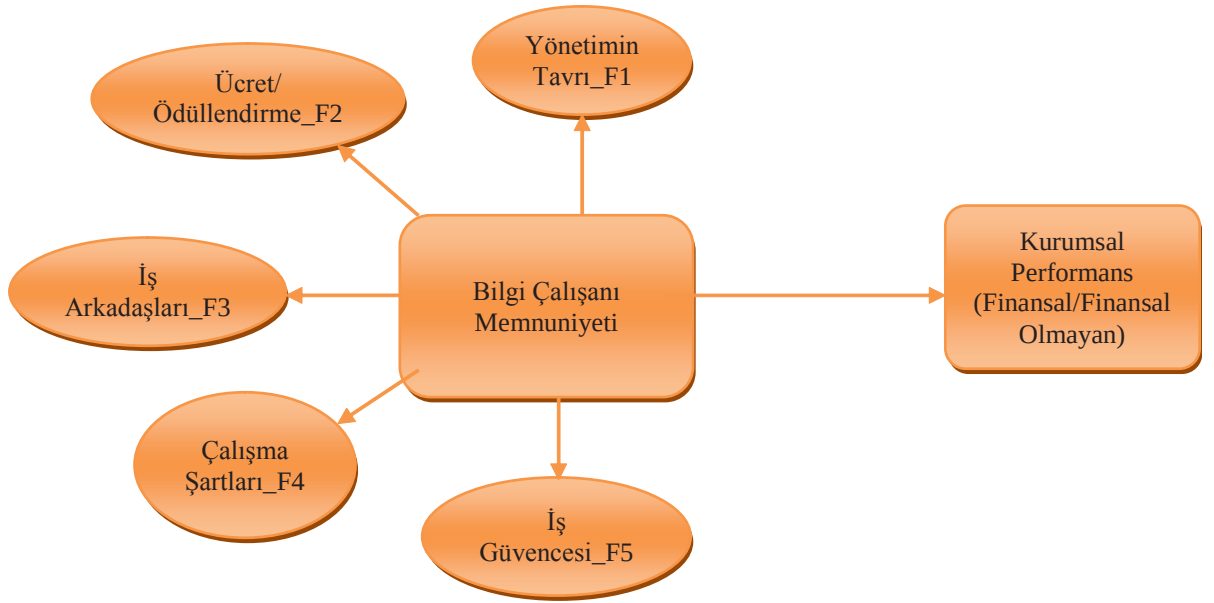
Şekil 16. Bilgi çalışanları yaş dağılımı

5.6 Önerilen bilgi çalışanı memnuiyetinin kurum performansı üzerindeki etkinliği modeli

Veri kümesine keşfedici faktör analizi (EFA) uygulandıktan sonra elde edilen faktörler tablo 19’da belirtilmiştir. Elde edilen faktörler kullanılarak önerilen model aşağıdaki gibi kurulabilir. Bilgi çalışan kurum performansı üzerindeki etkisini incelemek ve elde edilen bilgi çalışanında olması gereken niteliklerinin hangilerinin kurumun hem finansal hem de finansal olmayan performansı üzerinde en yüksek etkiye sahip olduğunu incelemek için aşağıdaki gibi bir model önerilmektedir.

Tablo 20. Bilgi çalışan memnuniyeti faktörleri

Faktör 1	Yönetimin Tavrı
Faktör 2	Ücret/Ödüllendirme
Faktör 3	İş Arkadaşları
Faktör 4	Çalışma Şartları
Faktör 5	İş Güvencesi



Şekil 17. Önerilen bilgi çalışanının kurum performansı üzerindeki etkinliğini gösteren model

5.7 Finansal ve Finansal Olmayan Kurumsal Performansların Başarılı ve Başarısız olarak Kümelenmesi

Kümeleme (cluster) modelleri benzer kayıtlar grubunu belirlemek üzerine odaklanırlar ve kayıtları bağlı oldukları gruplara göre isimlendirir. Bu işlem gruplar ve bunların özellikleri hakkında ön bilgi olmaksızın yapılır. Tahmin edilecek model için önceden tanımlanmış çıktı veya hedef alanı yoktur. Bu modellerin sınıflandırma performanslarını yargılayacak dışsal standartlar mevcut olmadığından, denetimsiz öğrenme modeller (unsupervised learning model) olarak adlandırılır. Bu modeller için doğru ya da yanlış cevaplar yoktur. Değerleri, veri içerisindeki ilginç grupları yakalama kabiliyetleri yardımı ile belirlenir. Kümeleme metotları kayıtlar arasındaki ve kümeler arasındaki mesafelerin ölçülmesine bağlıdır. Aynı kümeye ait kayıtların arasındaki mesafe minimize olacak şekilde, kayıtlar (elemanlar) kümelere atanır.

5.7.1 Aşamalı Kümeleme Yöntemi (Two Steps Clustering)

SPSS Clementine ortamında yapılan aşamalı kümeleme yönteminde kullanılan “The TwoStep Clustering Node” modeli çalışma şekli iki aşamada yapılır. İlk adımda, yönetilebilir alt kümeler olacak şekilde ham girdi verisini sıkıştırmak için veri içerisinde tek bir geçiş yapılır. İkinci adımda hiyerarşik kümeleme yöntemi kullanılarak aşamalı olarak alt kümeler daha büyük kümeler olacak şekilde birleştirilir. Kümeleme analizi homojen kümeler meydana getirir. Bu yöntem girdi verisindeki kalıpları ortaya çıkarır. Kayıtlar ait olduğu kümede birbirleri ile benzer fakat diğer kümelerdeki kayıtlar ile farklı olacak şekilde gruplandırılır.

5.7.2 Kümelemede Analizinde kullanılan Önem (Importance)

Ölçütünün anlamı

Önem (importance) ölçüsü ne kadar büyük ise, bir değişken için kümeler arası varyasyon daha az olasılıkla şansa bağlıdır. Önem ölçütü, bir istatistik testinin anlamlılık değerinin 1 sayısından çıkartılması ile elde edilir. Kategorik değişkenlerde Chi-Kare testi ve sürekli değişkenlerde ise t-testi kullanılır. Çalışmada kullanılan finansal ve finansal olmayan kurumsal performans algı değişkeni sürekli değişken olduğundan, t-testi kullanılarak önem ölçütü elde edilir. Sıfır hipotezi, küme içi ortalamalar küme boyunca aynı olduğunu gösterir. Eğer ilgili sürekli değişken gerçekten kümeyi belirlemede gerçekten etkili ise, sıfır hipotezi reddedilecektir ve anlamlılık düzeyi sıfıra yakın olacaktır. Böylece önem indeksi (importance measure) 1'e yakın olacaktır.

5.7.3 Finansal kurum performans algısının “Başarılı” ve “Başarısız” olarak kümeleneşmesi

Ankette elde edilen finansal performans algısının ikili hale (binary) getirilmesi için değişik teknikler mevcuttur. Ankette çalışmasında özellikle ortalaması alınan değişkenin ortalama üzeri değerlere başarılı (1), ortalamanın altında kalan bireylerin finansal performansı için ise “başarısız” (0) olarak atama yapılır. Buna göre, elimizdeki 564 bireyin 280 tanesi başarısız ve 284 tanesi ise başarılı olarak bulunmuştur. Bu tekniğin yerine, Clementine v.12 ortamında “aşamalı (Two Steps Cluster)” modeli kullanılarak aynı sonuçlar elde edilmiştir. Başarısız (1) olanların oranı %49, başarılı (2) olanlar ise % 51 olarak bulunmuştur. Yapılan kümeleme analizinin önem düzeyi 1 olarak elde edilmiştir. Yani, kümeler arası varyasyonun daha az olasılıkla şansa bağlı olduğunu belirtmektedir. Ayrıca, önem ölçütünün (importance measure) 1 olması finansal performans değişkeninin kümeleri belirlemede gerçekten etkili olduğunu belirtmektedir.

Tablo 21. Finansal kurum performans algısının kümelenmesi

Clementine “Two Steps Cluster” kümeleme Sonuçları					
1(Başarısız) 280 Records * Fin_Perf * Mean = 2,481 * Standard Deviation = 0,692			2 (Başarılı) 284 Records * Fin_Perf * Mean = 4,188 * Standard Deviation = 0,508		
CLSTR_FIN					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Başarısız (1)	280	49.6	49.6	49.6
	Başarılı (2)	284	50.4	50.4	100.0
	Total	564	100.0	100.0	
Sınıflandırma _Fin_Perf TwoStep					
	1	2	Overall	Importance	
	Count: 280 (49%)	Count: 284 (50%)	Overall: 564	>=0.95 >=0.90 <0.90 Unknown	
Finansal Kurumsal Performans	2.48 (0.69)	4.19 (0.51)	3.34 (1.05)	Important 1.00	

5.7.4 Finansal olmayan kurum performans algısının “Başarılı” ve “Başarısız” olarak kümelenmesi

Aynı işlemler finansal olmayan kurumsal performans algısı içinde yapıldıktan sonra aşağıdaki sonuçlar elde edilmiştir. Clementine V.12 yardımı ile yapılan aşamalı kümeleme (Two Steps cluster) yöntemi sonuçları ile geleneksel yöntemler olan medyan ve ortalama yöntemi ile elde edilen sonuçlar kıyaslanmıştır. Aşamalı kümeleme çok daha iyi homojen sonuçlar verdiği için ilgili yöntemin çıktıları aşağıda tablo olarak belirtilmiştir. Finansal olmayan kurumsal performans algısı başarısız (1) oranı 53.7% ile başarılı (2) oranı ise 46.3%'dir. Başarılı olarak kümelenen kayıtlar 261 iken başarısız olarak kümelenen kayıtların sayısı 303 olarak bulunmuştur. Önem düzeyi aşağıdaki gibi 1 olarak elde edilmiştir. Yani, kümeler arası varyasyonun daha az olasılıkla şansa bağlı

olduğunu belirtmektedir, finansal olmayan performans algı değişkeninin kümeleri belirlemede gerçekten etkili olduğunu belirtmektedir.

Tablo 22. Finansal olmayan performans algısının kümeleneşmesi

Clementine Aşamalı (Two Steps Cluster) Sınıflandırma Sonuçları					
1(Başarısız) 303 Records * Non_Fin_Perf * Mean = 2.584 * Standard Deviation = 0.771			2 (Başarılı) 261 Records * Non_Fin_Perf * Mean = 4.32 * Standard Deviation = 0.453		
CLSTR_NON_FIN_					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Başarısız (1)	303	53.7	53.7	53.7
	Başarılı (2)	261	46.3	46.3	100.0
	Total	564	100.0	100.0	
Sınıflandırma Non_Fin_Perf TwoStep					
		1	2	Overall	Importance
		Count: 303 (53%)	Count: 261 (46%)	Overall: 564	 >=0.95  >=0.90  <0.90  Unknown
Finansal Olmayan Kurumsal Performans		2.58 (0.77)	4.32 (0.45)	3.39 (1.08)	Important  1.00

5.8 Finansal performans algısının DVM ve karar ağaçları ile incelenmesi

SPSS- Clementine (v.12) ortamında kurulan modelin gösterimi aşağıdaki şekilde gibidir. SPSS ortamına atılan ham anket verisi üzerinde tanımlayıcı istatistik bilgileri (grafik, frekans tablosu, genel tablolar) elde edilmiştir. Verileri analize tabi tutmadan önce, veri kümesi içerisindeki değişkenlerin yapısı, türü ve analize dahil edilen/edilmeyen değişkenler ile ilgili özellikli ayarlamalar “Type” düğümü yardımı ile sağlanmıştır. SPSS verisi eğitime ve test verisi olarak bölmek için “partition” düğümü

sınıflandırıcı düğümü içerisinde bu algoritmaların varsayılan model parametreleri ayarları sistemde dâhil edilmiştir.

Kıyaslama

Estimated number of models to be executed: 5

☒ Use partitioned data

Rank models by: Overall accuracy

Rank models using: ☐ Training partition ☒ Test partition

Maximum models listed in summary report: 20

☒ Calculate variable importance

Profit Criteria

Costs: ☒ Fixed 5.0 ☐ Variable

Revenue: ☒ Fixed 10.0 ☐ Variable

Weight: ☒ Fixed 1.0 ☐ Variable

Lift Criteria

Percentile to use for lift calculation: 30

Fields **Model** Expert Discard Output Annotations

OK Execute Cancel Apply Reset

Kıyaslama

Estimated number of models to be executed: 5

Models used:

Use?	Model type	Model parameters	No of models
☐	Neural Net	Default	1
☒	C5.0	Default	1
☒	C&R Tree	Default	1
☒	QUEST	Default	1
☒	CHAID	Default	1
☐	Logistic	Default	1
☐	Decision List	Default	1
☐	Bayes Net	Default	1
☐	Discriminant	Default	1

☐ Restrict maximum time spent building a single model to 15 minutes

Stopping rules... Misclassification costs...

Fields **Model** **Expert** Discard Output Annotations

OK Execute Cancel Apply Reset

Şekil 19. Modellerin ikili sınıflandırıcı düğümü ayarlanması

Elde edilen sonuçlar hem eğitim hem de test kümeleri için iki ayrı tabloda verilerek algoritmalara karşılaştırılmıştır. Genel doğruluk (Overall accuracy) toplam sayıda ki kayıtlara kıyasla model tarafından doğru bir şekilde tahmin edilme oranı olarak tanımlanır. Test veri kümesini kullanılması ile kıyaslanan modellerin genel doğruluk oranlarına bakıldığında DVM (Support Vector Machine) %77,19 ile en yüksek genel doğruluk oranına sahiptir. Sonraki algoritmalar karar ağaçları algoritmaları olan C5 (%74,6), C&R Tree (%74,6), QUEST (%74,6) ve CHAID (%72,8) dir. Karar ağaçları algoritmaları genel doğruluk oranı açısından birbirlerine yakın olmakla beraber, DVM algoritması kadar iyi sonuçlar vermemektedir.

Alıcı işletim karakteristik eğrisi altında kalan alan (area under ROC (Receiver Operating Characteristic) Curve): ROC eğrisi modelin performansı için bir indeks sağlar. Eğri referans doğrusu üzerinde ne kadar yüksekte olursa test o kadar çok doğru demektir. Alanın 1 olması testin mükemmel ve 0,5 olması ise testin önemsiz olduğunu belirtir.

0.90-1 = Mükemmel (A)

0.80-0.90 = İyi (B)

0.70-0.80 = Orta (C)

0.60-0.70 = Zayıf (D)

0.50-0.60 = Başarısız (F) ⁴

Yukarıda belirtilen performans kriterlerine göre, ROC eğrisi altında kalan alan değeri en yüksek olan algoritmanın yine DVM (Support Vector Machine) %83,2 (iyi) ile en yüksek performans indeksine sahip iken, karar ağaçları algoritmalarından olan C&R Tree algoritması %80 ile ikinci en iyi performansa sahip algoritma olarak görülmektedir. CHAID algoritması %79,4 oranla iyi bir performans indeksine sahip iken C5 (%75,2) ve QUESR (% 72,4) orta derecede performans indeks değerlerine sahipler.

⁴ Thomas G. Tape , “The Area Under an ROC Curve”, (Çevrimiçi)<http://gim.unmc.edu/dxtests/roc3.htm>, 2011.

Kâr (Profit):

Clementine v.12 ikili sınıflandırıcı düğümünde (binary classifier node) belirtilen maliyet (cost), revenue (gelir) ve ağırlık (weight) kriterlerine dayanarak hesaplanır. Toplam kâr kümülatif yüzdellikler olarak belirtilir. İyi bir model için, kâr çok iyi tanımlanmış ve meydana geldiği yüzdellik dilimi ile beraber rapor edilen bir tepe noktası (pik) sağlaması gerekir. Kâr bilgi sağlamayan bir modelde, kâr eğrisi göreceli olarak düz olacaktır ve maliyet/gelir yapısına bağlı olarak artabilir, azalabilir ya da yatay olabilir. Bu çalışmada kurumların finansal performansları incelenirken kullandığımız algoritmaların ikili sınıflandırıcı yolu ile kıyaslanması sonucu elde edilen sonuçlarına göre, DVM maksimum kâr değeri 165 ile en iyi performansı sergilerken, karar ağaçlarından olan C&R Tree algoritmasının maksimum kâr değeri olan 165 ile ikinci en iyi performans gösteren algoritma olduğu görülmüştür. Bu sonuç diğer performans ölçümleri ile elde edilen değerler ile paralellik göstermektedir.

Lift: Genel örnekleme göre kümülatif dağılım dilimlerindeki isabet etme oranıdır. Örneğin üst dağılım (kantil) için lift değeri 3, isabet etme (hit) oranı genel örneklem için üç kez kadar yüksek olduğunu belirtir. İyi bir model için lift değeri üst dağılım için 1'den çok fazla yüksek olması gerekir ve düşük dağılımlarda lift oranı keskin bir şekilde 1'e iner. Lift bilgisine sahip olmayan modellerin lift değerleri 1 etrafında gezinir. Modelimizde DVM'nin lift değeri 1,68 ile en yüksek değere sahip algoritma iken, C&R Tree karar ağacı algoritmasının lift değeri 1,58 ile ikinci en iyi algoritma olduğu görülmektedir. Bu sonuçlar diğer performans ölçümleri ile paralellik göstermektedir.

İkili sınıflandırma düğümü yardımı ile kullandığımız modellerin algoritmaları kıyaslanmıştır. Performans ölçülerine bakıldığı zaman, DVM (Support Vector Machine) en yüksek performansa sahip olduğu görülürken, karar ağaçlarından C&R Tree en iyi performans sergilediği analizler sonucunda belirtilmiştir. Elde edilen sonuçlar test örneklem kümesine aittir. Eğitime veri kümesinde modeller eğitildiği için ve test veri

kümesinde aynı modeller test edildiği için performans ölçümleri test veri kümesine dayanmaktadır.

Tablo 23. Test veri Kümesinde (Testing Set) modellerin performanslarının kıyaslanması

Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
SVM	< 1	165.0	42	1.676	77.193	5	0.832
C5	< 1	145.0	61	1.396	74.561	2	0.752
C&R Tree	< 1	160.0	55	1.583	74.561	5	0.800
QUEST	< 1	145.0	49	1.451	74.561	5	0.724
CHAID	< 1	145.0	68	1.562	72.807	2	0.794

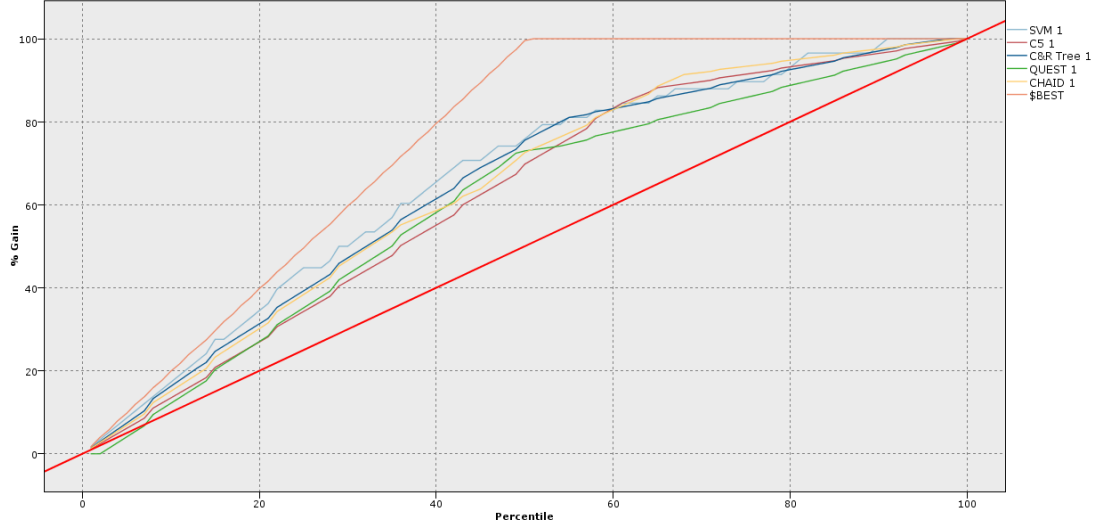
Tablo 24. Eğitim veri Kümesinde (Training Set) modellerin performanslarının kıyaslanması

Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
C&R Tree	< 1	559.706	63	1.618	74.667	5	0.811
QUEST	< 1	523.182	47	1.507	73.111	5	0.741
CHAID	< 1	513.35	47	1.592	72.667	2	0.777
SVM	< 1	515.0	55	1.578	71.556	5	0.782
C5	< 1	488.25	62	1.344	71.556	2	0.715

5.8.1.1 Modellerin performanslarının kazançlar (Gains) grafiği ile gösterimi

Modellerin kıyaslanmasında kullanılan ikili sınıflandırma (binary classification) düğümü yardımı ile hem tablo hem de grafik yardımı ile ilgili modellerin performansları kıyaslanabilir. Kazançlar grafiğine baktığımızda, %100' e doğru dike bir şekilde yükselen ve daha sonra düzelen model iyi bir model olarak kabul edilir. Buna göre,

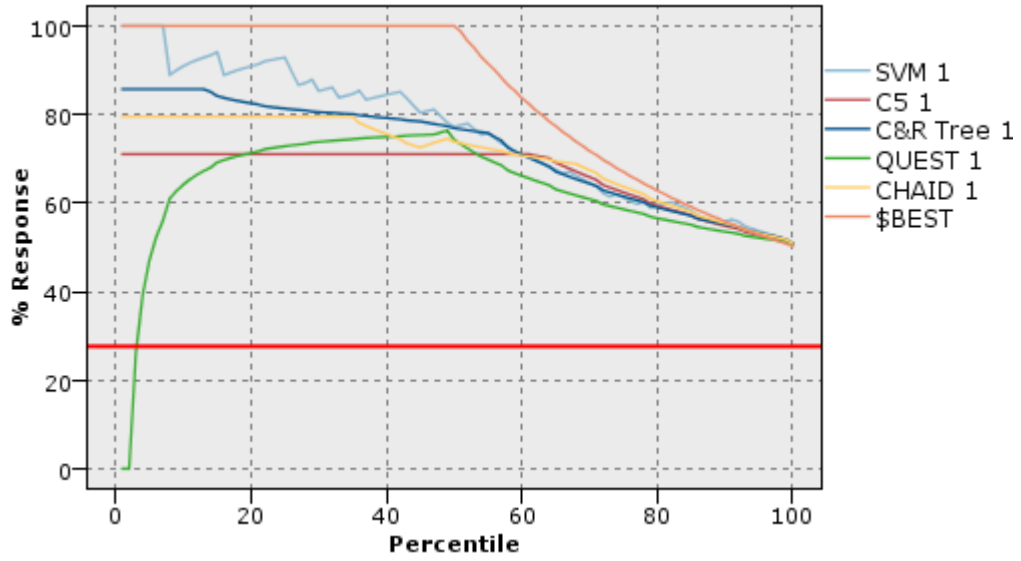
DVM (Support Vector Machine) ve C&R Tree Karar ağacı modeli hızlı bir şekilde %100 oranına yükselirken daha sonra diğer modeller gibi düzey konuma gelmektedirler.



Şekil 20. Modellerin performanslarının kazançlar (Gains) grafiği

5.8.1.2 Modellerin yanıt (response) grafiği ile gösterimi

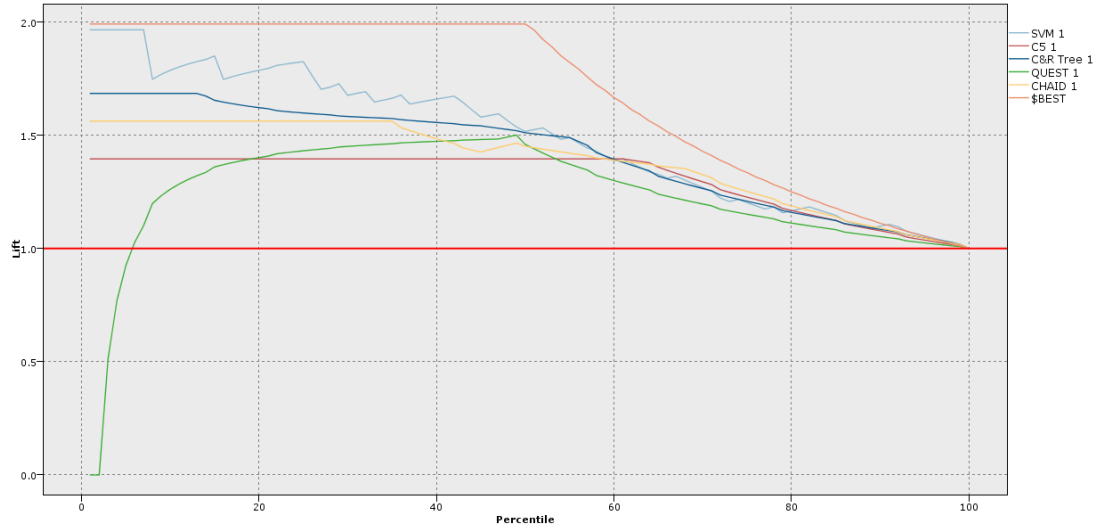
İyi bir model için, doğru %100 oranına yakın bir yerden başlar ve sağa doğru ilerlerken yüksekte düzlükte belli bir süre kalır daha sonra genel yanıt(tepki oranına) oranına kadar hızlı bir şekilde düşüşe geçer. Aynı şekilde, DVM modeli, C&R Tree Karar ağacı ve CHAID karar ağacı modeli %100 oranından başlar belli bir süre yüksek düzlükte devam ederken grafikte sağa doğru ilerledikçe hızlı bir şekilde genel yanı oranına doğru hızlı düşüşe geçmektedir. Diğer modeller, QUEST en düşük kazançlar (gains) açısından performansı göstermektedir. C5 modeli %70 gibi bir kazanç oranından başlamaktadır bu da DVM ve C&R Tree modeline göre düşük performans göstermektedir.



Şekil 21. Modellerin yanıt (response) grafiği

5.8.1.3 Modellerin “Lift” grafiği ile gösterimi

İyi bir model için “lift” grafiği 1,0 değerinin çok yukarısından başlaması gerekir, sağa doğru hareket ederken düz bir şekilde devam etmeli ve daha sonra hızlı bir şekilde 1,0 değerine doğru hızlı bir şekilde azalarak yaklaşmalıdır. “Lift” grafiğinde bakarak modellerimizin performansları incelendiğin de DVM (Support Vector Machine) en iyi performansı gösterirken, saha sonra C&R Tree karar ağacı modeli ile CHAID sırasıyla sonraki iyi performans gösteren modellerdir. C5 modeli de 1,0’den oldukça yüksekte olmasına rağmen DVM kadar iyi bir performans göstermemektedir. Sadece QUEST karar ağacı modeli 1,0 değerinin çok altında başlamaktadır. İyi bir model olmadığı grafikte görülmektedir.



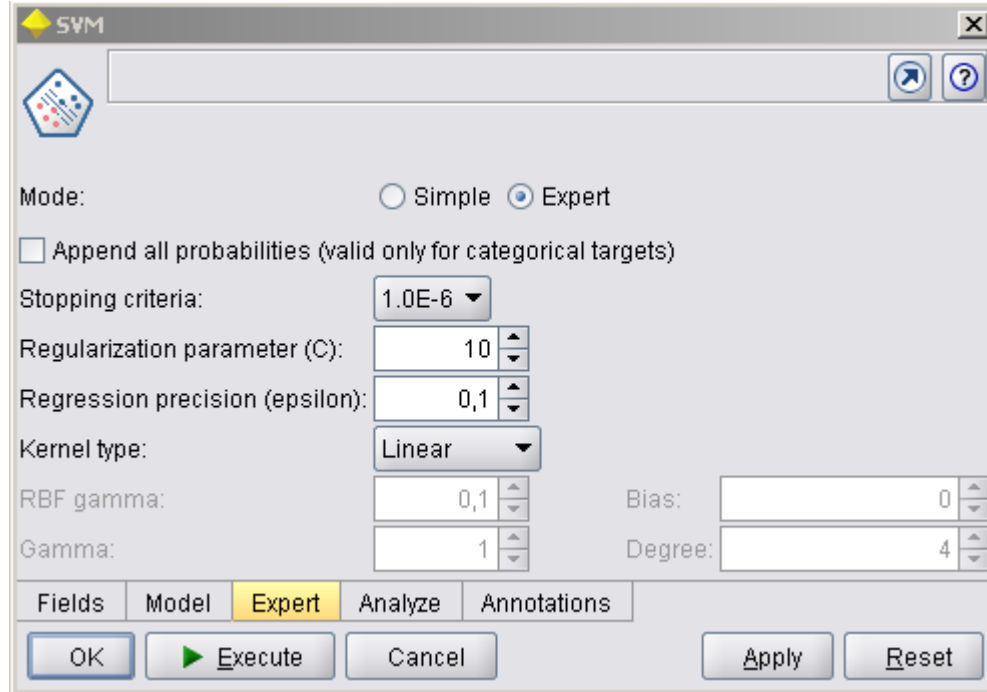
Şekil 22. Modellerin ‘Lift’ grafiği

5.8.1.4 Modellerin Özet Bilgileri

İkili sınıflandırıcı düğümü yardımı ile modelde kullanılan tüm (DVM, C&R Tree, CHAID, QUEST, C5) algoritmaların özellikli ayarlamalarına ilişkin bilgilerinin özeti EK-1’de belirtilmiştir. Algoritmalar varsayılan ayarları kullanılmıştır. Örneğin SVM’de hangi kernel fonksiyonunun kullanıldığı ve diğer tüm bilgiler EK-1’de detaylı bir şekilde gösterilmiştir. Hedef değişken yani bağımlı değişkenimiz finansal kurum performans algısı (CLSTR_FIN). Bağımsız değişkenler ise bilgi çalışanı memnuniyeti üzerinde yapılan faktör analizi sonucu elde edilen beş faktör olarak belirtilmiştir.

5.8.2 Finansal kurum performans algısının Destek Vektör Makineleri (Support Vector Machine) modeli yardımı ile detaylı incelenmesi

Destek Vektör Makineleri (DVM) algoritması kullanılarak finansal performans algısı incelenmiştir. Dönüşüm için kullanılan kernel fonksiyonları RBF (Radial Based Function), polinom, sigmoid ve lineer fonksiyonlar olduğu daha önce bahsedilmiştir. Clementine v.12 ortamındaki DVM (Support Vector Machine) ilgili dört kernel fonksiyonlarını desteklemektedir. Bağımlı değişken olarak finansal performans algısı (başarısız (1), başarılı (2)) olarak ve bağımsız değişkenleri de faktör analizi sonucu elde edilen bilgi işçisinin çalışma memnuniyeti faktörler analize sokulmuştur. Analizde verilerin 80% (450) eğitime (training) için ve 20%'si ise test (testing) için bölünerek analize dâhil edilmiştir. Bölünen veri kümesinin ilk kısmında modelin eğitilmesi için kullanılırken, bölünen ikinci veri kümesinde model test edilir ve düzeltilerek modelin geçerliliği (validate) yapılmıştır. Bir model doğru tahmin yapma yeteneğini hesaplamak için analiz (analysis) düğümü kullanılmaktadır. Analiz düğümü bir veya daha fazla modeller için tahmin edilen değerler (predicted) ile gerçek değerler (actual) arasında çeşitle kıyaslamalar yapmaktadır. İlgili modelin doğruluğu, modelin test sonuçlarına bakılarak ve veri kümesindeki elemanların modelimizi ne kadarı doğru ne kadarı yanlış tahmin ettiğine bakılarak belirtilir. İhtiyacımız olan sonuçların tamamı Clementine v.12 yardımı ile elde edilmiştir. Aşağıdaki sonuçlara göre, çok terimli (polynomial) kernel fonksiyonu en yüksek doğruluk oranını vermiştir, %78,95. İkinci en iyi DVM modeli %77,19 ile radyal tabanlı kernel fonksiyonu (RBF) olmuştur. Bunun yanı sıra, lineer kernel fonksiyonu %76,32 ve en düşük doğruluk oranını veren sigmoid fonksiyonu ise %71,93 oranlarında doğruluk (accuracy) değerlerini elde edilmiştir. Bilgi çalışan memnuniyetinin kurum finansal performans algısı üzerindeki etkisini incelemede en iyi doğruluk oranını veren DVM modeli çok terimli (polynomial) kernel fonksiyonudur.



Şekil 23. DVM modeli

Aşağıdaki analiz sonuçlarında, çakışım matrisleri (coincidence matrix) verilmiştir. Tahmin edilen alan ile gerçek alan arasındaki eşleşmelerin şablonlarını göstermektedir. Matris tablosundaki satırlar gerçek değerlerin sayılarını belirtirken sütunlar ise tahmin edilen değerlerin sayılarını belirtmektedir. Bu matris tahminde meydana gelen sistematik hataları belirlemede faydalı olmaktadır.

Analiz sonuçlarında ayrıca performans hesaplamaları mevcuttur. Kategorik bağımlı değişkene sahip modeller için performans hesaplama istatistiklerini göstermektedir. Her bir kategorik bağımlı değişkene ait istatistik bilgileri verilmiştir. İlgili kategoriye ait kayıtları tahmin etmek için modelin ortalama bilgi içeriğinin (bits) bir ölçümüdür. Sınıflandırma probleminin zorluğunu dikkate alır, böylece yaygın kategoriler için doğru tahmine nazaran, nadir kategoriler için doğru tahminler yüksek bir performans değerlendirme indeksine sahip olacaktır. Eğer model bir kategori için tahmin etmenin ötesinden fazla bir şey yapamıyor ise, o kategori için performans değerlendirme

indeksi sıfır olacaktır. DVM modelimizin performans değerlendirme indeksleri aşağıdaki tabloda verilmiştir. Kategorik değişkenlerimiz kurumsal finansal performans algısı olan başarısız (1) ve başarılı (2)'dir.

Bu çalışmada kullanılan kategorik bağımlı değişkenler finansal performans algısını değerlendirmede başarısız (1) ve başarılı (2) olarak belirtilerek ilgili istatistik performans sonuçları her bir model için aşağıdaki tablolarda belirtilmiştir. En yüksek doğruluk oranına sahip olan çok terimli (polynomial) kernel fonksiyonu DVM modelimizin performans değerleri başarısız (1) için 42,3% ve başarılı için 50,1% olarak bulunmuştur. İkinci en yüksek doğruluk oranına sahip radyal tabanlı (RBF) kernel fonksiyonunun performans indeks değerleri başarısız (1) için 43,5% ve başarılı için 43,5%. En düşük doğruluk orana sahip olan sigmoid kernel fonksiyonu başarısız (1) için 42,3% ve başarılı için (2) 32,5% olarak performans değerleri elde etmiştir.

5.8.2.1 Destek Vektör Makineleri Modelinin Kernel Fonksiyonlarının performanslarının Kıyaslanması

Tablo 25. Radyal Tabanlı (Radial based kernel function) DVM modelinin performans değerleri

Accuracy					
'Partition'	1_Training		2_Testing		
Correct	322	71.56%	88	77.19%	
Wrong	128	28.44%	26	22.81%	
Total	450		114		
Coincidence Matrix (rows show actuals)					
'Partition' = 1_Training			'Partition' = 2_Testing		
	1	2		1	2
Başarısız(1)	154	70	Başarısız (1)	44	12
Başarılı (2)	58	168	Başarılı (2)	14	44
Performans Evaluation					

'Partition' = 1_Training 1) 0.378 2) 0.340	'Partition' = 2_Testing 1) 0.435 2) 0.435

Tablo 26. Çok Terimli (Polynomial) Kernel DVM modelinin performans değerleri

Accuracy					
'Partition'	1_Training			2_Testing	
	Correct 327 72.67%			90 78.95%	
	Wrong 123 27.33%			24 21.05%	
	Total 450			114	
Coincidence Matrix (rows show actuals)					
'Partition' = 1_Training			'Partition' = 2_Testing		
	1	2		1	2
Başarısız(1)	174	48	Başarısız (1)	48	8
Başarılı (2)	75	151	Başarılı (2)	16	42
Performans Evaluation					
'Partition' = 1_Training			'Partition' = 2_Testing		
1) 0.343			1) 0.423		
2) 0.413			2) 0.501		

Tablo 27. Sigmoid Kernel DVM modelinin performans değerleri

Accuracy					
'Partition'	1_Training			2_Testing	
	Correct	313	69.56%	82	71.93%
	Wrong	137	30.44%	32	28.07%
	Total	450		114	
Coincidence Matrix (rows show actuals)					
'Partition' = 1_Training			'Partition' = 2_Testing		
	1	2		1	2
Başarısız(1)	141	83	Başarısız (1)	36	20
Başarılı (2)	54	172	Başarılı (2)	12	46
Performans Evaluation					

'Partition' = 1_Training	'Partition' = 2_Testing
1) 0.373	1) 0.423
2) 0.295	2) 0.315

Tablo 28. Lineer Kernel DVM modelinin performans değerleri

Accuracy					
'Partition'	1_Training			2_Testing	
Correct	320	71,11%		87	76.32%
Wrong	130	28,89%		27	23.68%
Total	450			114	
Coincidence Matrix (rows show actuals)					
'Partition' = 1_Training			'Partition' = 2_Testing		
	1	2		1	2
Başarısız(1)	153	71	Başarısız (1)	44	12
Başarılı (2)	59	167	Başarılı (2)	15	43
Performans Evaluation					
'Partition' = 1_Training			'Partition' = 2_Testing		
1) 0.371			1) 0.417		
2) 0.334			2) 0.430		

DVM için en iyi modeli elde etmek için dört kernel fonksiyonuna ait DVM modellerinin arasındaki uyum oranları tabloda verilmiştir. Aynı bağımlı değişkeni tahmin eden iki ve daha fazla oluşturulan modeller mevcut olduğu zaman, modeller tarafından oluşturulan tahminler arasındaki uyum verilmektedir. Kurumların finansal performans algısını tahmin etmeğe çalışan aynı bağımlı değişkenini tahmin eden dört kernel fonksiyonuna ait DVM modelleri arasındaki uyum 84,21% gibi iyi bir oran elde edilmiştir.

5.8.2.2 Çok Terimli (Polynomial) Kernel Fonksiyonu Kullanılarak Destek Vektör Makineleri Yardımı İle Faktörlerin İncelenmesi

En yüksek doğruluk oranına sahip olan kernel fonksiyonu çok terimli (polynomial) DVM olduğu yukarıda incelenmiştir. Bu modele ait özellikli parametre ayarlamalar Clementine V.12 sürümü kullanılarak özet bilgiler verilmiştir.

Çok terimli (polynomial)- DVM model bilgileri:

Fields

Target

CLSTR_FIN

Inputs

Faktör_1

Faktör_2

Faktör_3

Faktör_4

Faktör_5

Build Settings

Use partitioned data: true

Partition: Partition

variable_importance.LABEL: true

Calculate raw propensity scores: true

Calculate adjusted propensity scores: true

Based on: Testing partition

Mode: Expert

Append all probabilities (valid only for categorical targets): false

Stopping criteria: 1.0E-6

Kernel type: Polynomial

Regularization parameter (C): 10

Regression precision (epsilon): 0.1

RBF gamma: 0.1

Gamma: 1.0

Bias: 0.0

Degree: 4

Training Summary

Algorithm: SVM

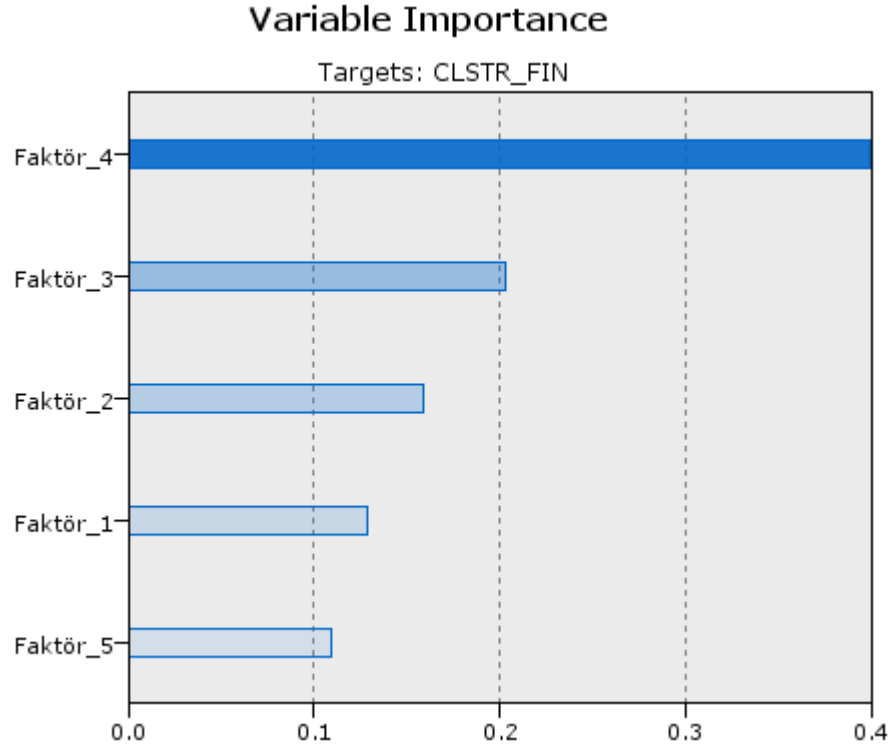
Model type: Classification

Değişken Önemi (Variable Importance):

Clementine v.12 DVM (SVM) algoritması ayarlarında değişken önem (variable importance) işaretlendiğinden, aşağıdaki gibi “değişken önem” grafiği verilmektedir. Modelde üzerinde odaklanılacak en önemli değişkenlerin bilinmesi gerekir. Değişken

önem grafiği model tahmininde her bir değişkenin göreceli öneminin belirtilmesi bu açıdan faydalı olmaktadır. Verilen değerler göreceli olduğundan, modele dâhil edilen tüm değişkenlerin değerleri toplamı 1,0'dır. Değişken önemi (variable importance) modelin doğruluğu ile ilişkili değildir. Sadece bir tahmin yapmada her bir değişkenin önemi ile ilgilidir, tahminin doğru olup olmaması ile ilgili değildir. Uygun istatistik önem ölçümü üreten, yapay sinir ağları, karar ağaçları (C&R Tree, C5, CHAID ve QUEST), Bayes ağları, diskriminant, DVM (Support Vector Machine) gibi modeller için mevcuttur. Değişken önem test veri kümesi üzerinden hesaplanır. Eğer veri kümesi test için ayrı bir veri kümesi ayrılmamış ise eğitime veri kümesi üzerinden hesaplamalar yapar.

Çok terimli kernel DVM modelinde elde edilen faktörlerin önem dereceleri hem grafikte hem de tablo halinde gösterilmiştir. Elde edilen çıktılarına göre çalışma şartları %39,95 göreceli önem oranının göre en etkili faktör olarak belirtilmiştir. İş arkadaşları faktörü %20,33 oranı ile ikinci en önemli değişken iken ücret/ödüllendirme faktörü göreceli önemi %15,87 ile üçüncü en iyi faktördür. %12,89 ve %10,96 göreceli önem oranları ile sırası ile yönetimin tavrı ve iş güvencesi faktörleri göreceli olarak diğer faktörlere göre az öneme sahip faktörler olarak belirtilmiştir. Elde edilen faktörlerin göreceli önem oranlarının toplamı 1 olduğu görülmektedir.



Şekil 24. Çok terimli (polynomial) -DVM Faktörlerin göreceli önem oranları
(Finansal Performans)

Tablo 29. Çok terimli (polynomial) -DVM Faktörlerin göreceli önem oranları
(Finansal Performans)

Faktörler	Değişken Önemi
Faktör_4 (Çalışma Şartları)	0.3995
Faktör_3 (İş Arkadaşları)	0.2033
Faktör_2 (Ücret/Ödüllendirme)	0.1587
Faktör_1(Yönetimin Tavrı)	0.1289
Faktör_5 (İş Güvencesi)	0.1096

Çok terimli DVM modelinde veri kümesinin %80 (450 adet) eğitildikten sonra veri kümesinin %20 (114adet) test edilerek modelin doğru tahmin etme yüzdesi % 78,95 olarak elde edilerek diğer kernel fonksiyonlarına göre en yüksek doğru tahmin etme oranına sahiptir.

'Partition'	1_Training		2_Testing	
Correct	327	72.67%	90	78.95%
Wrong	123	27.33%	24	21.05%
Total	450		114	

Çok terimli (Polynomial) kernel fonksiyonu kullanılarak DVM modelimizin tahmin sonuçları Clementine V.12 tablo düğümü vasıtasıyla aşağıda sadece ilk 20 verinin sonuçları verilmiştir. DVM modeli analiz sonrası bir dizi yeni değişkenler (alanlar) oluşturur. Bunlardan en önemlisi olan “\$S-alan adı” model tarafından tahmin edilen bağımlı (hedef) değişkenin değerini gösterir. Bu çalışmada DVM modeli tarafından tahmin edilen bağımlı değişken alanı “\$S-CLSTR_FIN” olarak belirtilmiştir. Finansal performans ağılsının DVM modeli tarafından tahmin edilen değeridir. Çok terimli DVM modeli tarafından tahmin edilen finansal performans algısı %78,95 oranında doğru tahmin edildiği tabloda görülmektedir. Ayrıca, “SP-CLSTR_FIN” finansal performans algısının model tarafından tahmin edilen değerinin olasılığını göstermektedir. Modelde SP-olasılık değerleri 0,50 ile 0,98 arasında değişmektedir. Yaklaşık veri kümesinin %60’ı SP-olasılık değerleri %70 üzerindedir. “\$SRP-CLSTR_FIN” ve “\$AP-CLSTR_FIN” yeni değişkenleri ham (raw propensity score) eğilim skoru ve düzeltilmiş eğilim skorları (adjusted propensity score) bağımlı (hedef alan) değişken için “gerçek” sonucun olma olasılığını belirtmektedir.

Tablo 30. Çok terimli -DVM modeli tahmin sonuçları

Table (80 fields, 564 records)															
	Faktor_2	Faktor_3	Faktor_4	Faktor_5	Fin_Perf	Non_Fin_...	Bin_Fin_...	CLSTR_...	Bin_Non_Fin_...	CLSTR_NON_...	Partition	\$S-CLSTR_FIN	\$SP-CLSTR_FIN	\$SRP-CLSTR_FIN	\$SAP-CLSTR_FIN
1	4.14	4.00	4.83	4.50	3.75	4.00	1.00	2	1.00	2	1_Training	2	0.857	0.857	0.841
2	2.43	4.33	4.50	3.50	3.00	3.25	0.00	1	0.00	1	12_Testing	2	0.612	0.612	0.651
3	4.00	4.33	4.00	4.50	4.00	3.75	1.00	2	1.00	2	1_Training	2	0.639	0.639	0.683
4	1.86	3.00	2.50	3.50	1.50	2.25	0.00	1	0.00	1	1_Training	1	0.762	0.762	0.196
5	1.86	3.67	3.83	4.50	3.50	3.75	1.00	2	1.00	2	1_Training	1	0.623	0.377	0.329
6	1.71	3.83	4.17	5.00	3.50	3.25	1.00	2	0.00	1	1_Training	1	0.555	0.445	0.420
7	3.14	4.00	3.67	2.00	3.00	2.50	0.00	1	0.00	1	1_Training	1	0.574	0.426	0.394
8	4.71	4.50	4.83	1.00	3.00	3.50	0.00	1	1.00	1	1_Training	1	0.743	0.257	0.210
9	4.71	4.33	4.50	4.00	3.25	3.00	0.00	1	0.00	1	1_Training	2	0.831	0.831	0.829
10	3.43	2.83	4.17	3.00	3.00	3.00	0.00	1	0.00	1	1_Training	1	0.667	0.333	0.279
11	2.57	3.83	2.50	1.00	2.75	2.25	0.00	1	0.00	1	1_Training	1	0.675	0.325	0.271
12	4.14	4.00	4.67	4.00	3.00	3.00	0.00	1	0.00	1	1_Training	2	0.817	0.817	0.822
13	4.00	4.67	4.33	2.50	3.50	3.25	1.00	2	0.00	1	1_Training	2	0.734	0.734	0.771
14	2.43	3.67	3.83	4.50	1.75	1.50	0.00	1	0.00	1	1_Training	1	0.595	0.405	0.365
15	4.29	4.83	4.83	5.00	4.25	4.25	1.00	2	1.00	2	1_Training	2	0.937	0.937	0.868
16	1.86	4.50	3.50	3.00	2.75	2.75	0.00	1	0.00	1	1_Training	1	0.743	0.257	0.210
17	3.00	3.00	2.83	3.50	3.25	4.00	0.00	1	1.00	1	1_Training	1	0.722	0.278	0.227
18	2.86	2.83	3.33	4.50	3.00	2.25	0.00	1	0.00	1	1_Training	1	0.694	0.306	0.252
19	2.29	1.50	2.83	5.00	4.00	4.00	1.00	2	1.00	2	1_Training	1	0.728	0.272	0.222
20	1.29	4.17	1.83	5.00	3.00	2.25	0.00	1	0.00	1	1_Training	1	0.924	0.076	0.118

5.8.2.3 RBF Kernel Fonksiyonu Kullanılarak Destek Vektör Makineleri Yardımı İle Faktörlerin İncelenmesi

Bilgi çalışanı memnuniyetinin kurumların finansal performans algısı üzerindeki etkisini incelemede DVM modelinin sonuçları detayları ile aşağıda verilmiştir. Bağımlı değişkenimiz finansal performans kategorik (başarılı/başarısız) ve bağımsız değişkenler ise faktör analizi sonucunda elde edilen beş adet faktörlerdir. İlk önce DVM modelimizin radyal tabanlı (RBF) kernel fonksiyonu kullanılarak ilgili sonuçlar elde edilmiştir. Clementine V.12 üzerinde DVM (Support Vector Machine) algoritmasında RBF fonksiyonu seçilerek ve bu fonksiyona ait özellikli ayarlamalar yapılarak sonuçlar aşağıdaki gibi belirtilmiştir

RBF- DVM Model bilgileri:

Fields

Target

CLSTR_FIN

Inputs

Faktör_1

Faktör_2

Faktör_3

Faktör_4

Faktör_5

Build Settings

Use partitioned data: true

Partition: Partition

_variable_importance.LABEL: true

Calculate raw propensity scores: true

Calculate adjusted propensity scores: true

Based on: Testing partition

Mode: Expert

Append all probabilities (valid only for categorical targets): false

Stopping criteria: 1.0E-6

Kernel type: RBF

Regularization parameter (C): 10

Regression precision (epsilon): 0,1

RBF gamma: 0,1

Gamma: 1,0

Bias: 0,0

Degree: 4

Training Summary

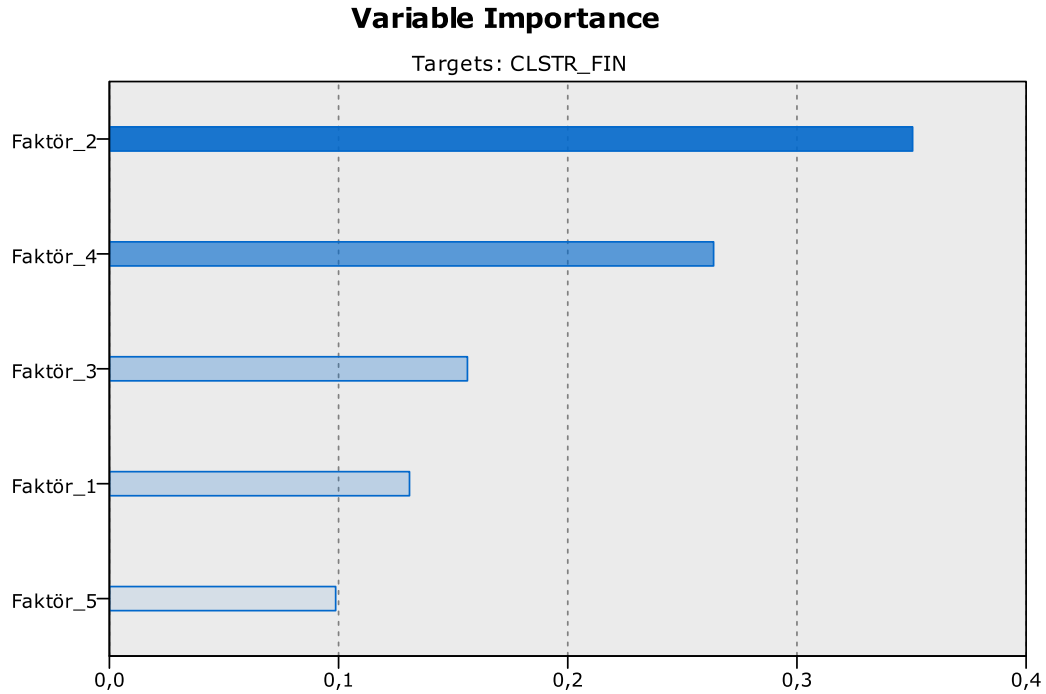
Algorithm: SVM

Model type: Classification

Önem grafiği çeşitli alanların tahminde göreceli etkisini gösterir. Kurumların finansal performans algısı üzerindeki elde edilen faktörlerin göreceli etkileri kolaylıkla görülmektedir. Faktör 2 (ücret/ödüllendirme) %35,05 ile en yüksek etkiye sahip iken, faktör 4 (çalışma şartları) %26,36 ile oldukça anlamlı bir etkiye sahip olduğu görülmektedir. Faktör 3 (İş Arkadaşları) ve Faktör 1(Yönetimin Tavrı) sırası ile %15,62 ve %13,09 ile sonraki en önemli göreceli etkiye sahip faktörlerdendir. Faktör 5 (iş güvencesi) ise en az göreceli (%9,87) etkiye sahiptir.

Tablo 31. RBF(Radyal Tabanlı)-DVM Faktörlerin göreceli önem oranları
(Finansal Performans)

Faktörler	Değişken Önemi
Faktör_2(Ücret/Ödüllendirme)	0,3505
Faktör_4 (Çalışma Şartları)	0,2636
Faktör_3 (İş Arkadaşları)	0,1562
Faktör_1 (Yönetimin Tavrı)	0,1309
Faktör_5 (İş Güvencesi)	0,0987



Şekil 25. RBF(Radyal Tabanlı)-DVM Faktörlerin göreceli önem oranları
(Finansal Performans)

RBF Doğruluk Oranı: DVM modeli RBF kernel fonksiyonu kullanılarak eğitim veri kümesinde eğitildikten sonra test veri kümesinde test edilerek % 77,19 doğruluk oranına sahiptir.

'Partition'	1_Training		2_Testing	
Correct	322	71,56%	88	77,19%
Wrong	128	28,44%	26	22,81%
Total	450		114	

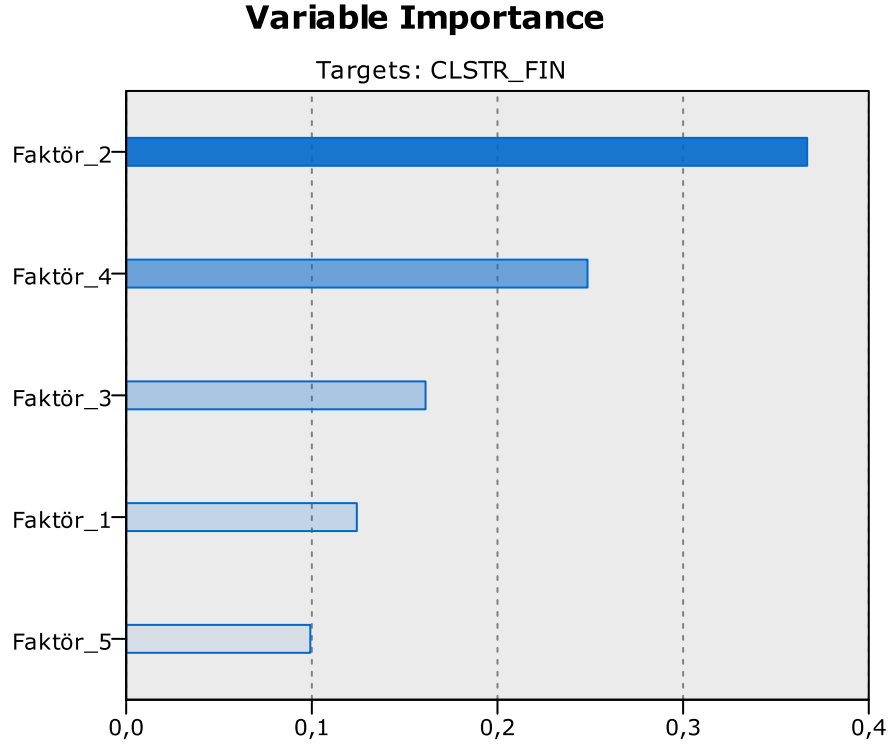
RBF kernel fonksiyonu kullanılarak DVM modelimizin tahmin sonuçları Clementine V.12 tablo düğümü vasıtasıyla aşağıda sadece ilk 20 verinin sonuçları verilmiştir. DVM modeli analiz sonrası bir dizi yeni değişkenler (alanlar) oluşturur. Bunlardan en önemlisi olan “\$S-alan adı” model tarafından tahmin edilen bağımlı (hedef) değişkenin değerini gösterir. Bu çalışmada DVM modeli tarafından tahmin edilen bağımlı değişken alanı “\$S-CLSRE_Fin” olarak belirtilmiştir. Finansal performans ağılsının DVM modeli tarafından tahmin edilen değeridir. Diğer testlerde elde edilen doğruluk oranı olan %77,19 oranında DVM modeli doğru tahminde bulunmuştur. Ayrıca, “SP-CLSTR_FIN” finansal performans algısının model tarafından tahmin edilen değerinin olasılığını göstermektedir. “\$SRP-CLSTR_FIN” ve “\$AP-CLSTRY_FIN” yeni değişkenleri ham (raw propensity score) eğilim skoru ve düzeltilmiş eğilim skorları (adjusted propensity score) bağımlı (hedef alan) değişken için “gerçek” sonucun olma olasılığını belirtmektedir.

Tablo 32. RBF(Radyal Tabanlı)-DVM modelinin tahmin sonuçları

Table (80 fields, 564 records)														
File	Edit	Generate												
	for_3	Faktor_4	Faktor_5	Fin_Perf	Non_Fin_...	Bin_Fin_...	CLSTR_FIN	Bin_Non_Fin_Perf	CLSTR_NON_FIN_...	Partition	\$\$-CLSTR_FIN	\$\$P-CLSTR_FIN	\$\$SRP-CLSTR_FIN	\$\$SAP-CLSTR_FIN
1	4.00	4.83	4.50	3.75	4.00	1.00	2	1.00	2_1_Training	2	0.799	0.799	0.799	0.840
2	4.33	4.50	3.50	3.00	3.25	0.00	1	0.00	1_2_Testing	2	0.553	0.553	0.553	0.525
3	4.33	4.00	4.50	4.00	3.75	1.00	2	1.00	2_1_Training	2	0.721	0.721	0.721	0.766
4	3.00	2.50	3.50	1.50	2.25	0.00	1	0.00	1_1_Training	1	0.800	0.800	0.800	0.178
5	3.67	3.83	4.50	3.50	3.75	1.00	2	1.00	2_1_Training	1	0.609	0.609	0.609	0.391
6	3.83	4.17	5.00	3.50	3.25	1.00	2	0.00	1_1_Training	1	0.579	0.579	0.579	0.340
7	4.00	3.67	2.00	3.00	2.50	0.00	1	0.00	1_1_Training	1	0.533	0.533	0.533	0.398
8	4.50	4.83	1.00	3.00	3.50	0.00	1	1.00	1_1_Training	2	0.783	0.783	0.783	0.827
9	4.33	4.50	4.00	3.25	3.00	0.00	1	0.00	1_1_Training	2	0.818	0.818	0.818	0.854
10	2.83	4.17	3.00	3.00	3.00	0.00	1	0.00	1_1_Training	2	0.549	0.549	0.549	0.518
11	3.83	2.50	1.00	2.75	2.25	0.00	1	0.00	1_1_Training	1	0.764	0.764	0.764	0.194
12	4.00	4.67	4.00	3.00	3.00	0.00	1	0.00	1_1_Training	2	0.797	0.797	0.797	0.839
13	4.67	4.33	2.50	3.50	3.25	1.00	2	0.00	1_1_Training	2	0.722	0.722	0.722	0.767
14	3.67	3.83	4.50	1.75	1.50	0.00	1	0.00	1_1_Training	1	0.535	0.535	0.535	0.395
15	4.83	4.83	5.00	4.25	4.25	1.00	2	1.00	2_1_Training	2	0.840	0.840	0.840	0.888
16	4.50	3.50	3.00	2.75	2.75	0.00	1	0.00	1_1_Training	1	0.661	0.661	0.661	0.260
17	3.00	2.83	3.50	3.25	4.00	0.00	1	1.00	2_1_Training	1	0.655	0.655	0.655	0.265
18	2.83	3.33	4.50	3.00	2.25	0.00	1	0.00	1_1_Training	1	0.619	0.619	0.619	0.298
19	1.50	2.83	5.00	4.00	4.00	1.00	2	1.00	2_1_Training	1	0.766	0.766	0.766	0.193
20	4.17	1.83	5.00	3.00	2.25	0.00	1	0.00	1_1_Training	1	0.829	0.829	0.829	0.167

5.8.2.4 Lineer Kernel Fonksiyonu Kullanılarak Destek Vektör Makineleri Yardımı İle Faktörlerin İncelenmesi

Lineer kernel fonksiyonu kullanılarak elde edilen DVM modelinde göreceli önem oranı en yüksek olan ve en etkili faktör %36,68 oranı ile ücret/ ödüllendirme faktörüdür. RBF DVM modelindeki gibi ilk iki en etkin faktörler Ücret/ödüllendirme ile Çalışma şartlarıdır (%24,85). Yine aynı şekilde, sırasıyla iş arkadaşları (%16,13), yönetimin tavrı (%12,42) ve iş güvencesi (%9,92) göreceli önem oranları ile üçüncü, dördüncü ve beşinci en önemli faktörler olduğu lineer DVM modeli yardımı ile elde edilmiştir. Genel doğruluk düzeylerine bakıldığında hemen hemen bir birlerin yakın olan RBF ve lineer DVM modelleri, çok terimli DVM modeline kıyasla daha az oranda doğruluk yüzdelerine sahiptirler.



Şekil 26. Lineer-DVM Faktörlerin göreceli önem oranları (Finansal Performans)

Tablo 33. Lineer-DVM Faktörlerin göreceli önem oranları (Finansal Performans)

Faktörler	Değişken Önemi
Faktör_2 (Ücret/Ödüllendirme)	0,3668
Faktör_4 (Çalışma Şartları)	0,2485
Faktör_3 (İş Arkadaşları)	0,1613
Faktör_1(Yönetimin Tavrı)	0,1242
Faktör_5 (İş Güvencesi)	0,0992

Lineer DVM modelimizin genel doğruluk oranı %76,32 olarak bulunmuştur. Yani model 450 adet kaydı eğittikten sonra 114 adet kayıt üzerinde test ederek 87 adet kaydı doğru tahmin etmiş ve diğer 27 adet kayıt yanlış tahmin edilmiştir.

'Partition'	1_Training	2_Testing
Correct	320 71,11%	87 76,32%
Wrong	130 28,89%	27 23,68%
Total	450	114

Lineer DVM model özellikli ayarlamalar özet bilgileri aşağıdaki gibi verilmiştir. Bağımlı değişken finansal performans algısı ve bağımsız değişken alanları ise faktör analizi sonucu elde edilen faktörlerdir. Kernel fonksiyon “Lineer” olarak seçilmiştir ve bu fonksiyona ait SPSS Clementine V.12 üzerinden özellikli ayarlamalar yapılarak ilgili sonuçlar elde edilmiştir.

Lineer-DVM model bilgileri:

Fields

Target
CLSTR_FIN

Inputs
Faktör_1
Faktör_2
Faktör_3
Faktör_4
Faktör_5

Build Settings

Use partitioned data: true
Partition: Partition
_variable_importance.LABEL: true
Calculate raw propensity scores: true
Calculate adjusted propensity scores: true
Based on: Testing partition
Mode: Expert
Append all probabilities (valid only for categorical targets): false
Stopping criteria: 1.0E-6
Kernel type: **Linear**
Regularization parameter (C): 10
Regression precision (epsilon): 0,1
RBF gamma: 0,1
Gamma: 1,0
Bias: 0,0
Degree: 4

Training Summary

Algorithm: SVM
Model type: Classification

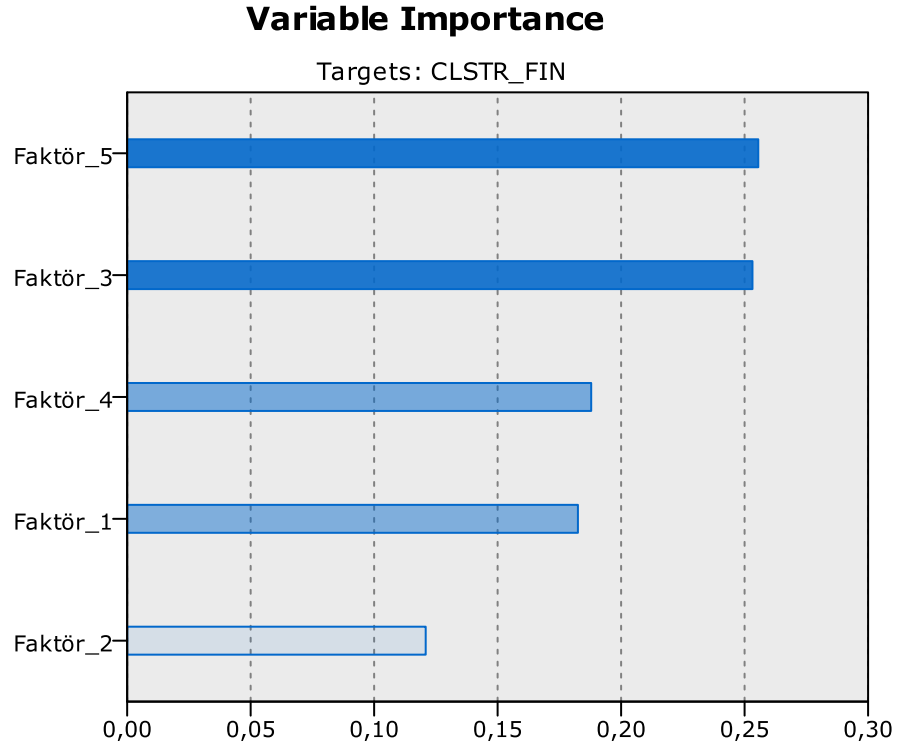
Lineer DVM modeli ile elde edilen tahmin değerleri (\$-CLSTR_FIN) , olasılık (\$SP-CLSTR_FIN)ve bunlara ait ham ve düzeltilmiş eğilim skorları (\$SPR, \$SAP-CLSTR_Fin) aşağıda ilk 20 adet kayıt için verilmiştir.

Tablo 34. Lineer-DVM tahmin sonuçları

	Faktör_2	Faktör_3	Faktör_4	Faktör_5	Fin_Perf	Non_Fin_...	Bin_Fin_...	CLSTR_...	Bin_Non_Fin_...	CLSTR_NON_...	Partition	\$S-CLSTR_FIN	\$SP-CLSTR_FIN	\$SRP-CLSTR_FIN	\$SAP-CLSTR_FIN
1	4.14	4.00	4.83	4.50	3.75	4.00	1.00	2	1.00	2	1_1_Training	0.791	0.791	0.791	0.823
2	2.43	4.33	4.50	3.50	3.00	3.25	0.00	1	0.00	2	1_2_Testing	0.537	0.537	0.537	0.472
3	4.00	4.33	4.00	4.50	4.00	3.75	1.00	2	1.00	2	1_1_Training	0.720	0.720	0.720	0.749
4	1.86	3.00	2.50	3.50	1.50	2.25	0.00	1	0.00	1	1_1_Training	0.796	0.796	0.796	0.169
5	1.86	3.67	3.83	4.50	3.50	3.75	1.00	2	1.00	1	2_1_Training	0.606	0.606	0.606	0.288
6	1.71	3.83	4.17	5.00	3.50	3.25	1.00	2	0.00	1	1_1_Training	0.569	0.569	0.569	0.329
7	3.14	4.00	3.67	2.00	3.00	2.50	0.00	1	0.00	1	1_1_Training	0.536	0.536	0.536	0.368
8	4.71	4.50	4.83	1.00	3.00	3.50	0.00	1	1.00	1	1_1_Training	0.787	0.787	0.787	0.820
9	4.71	4.33	4.50	4.00	3.25	3.00	0.00	1	0.00	1	1_1_Training	0.816	0.816	0.816	0.843
10	3.43	2.83	4.17	3.00	3.00	3.00	0.00	1	0.00	1	1_1_Training	0.541	0.541	0.541	0.478
11	2.57	3.83	2.50	1.00	2.75	2.25	0.00	1	0.00	1	1_1_Training	0.748	0.748	0.748	0.189
12	4.14	4.00	4.67	4.00	3.00	3.00	0.00	1	0.00	1	1_1_Training	0.785	0.785	0.785	0.818
13	4.00	4.67	4.33	2.50	3.50	3.25	1.00	2	0.00	1	1_1_Training	0.705	0.705	0.705	0.731
14	2.43	3.67	3.83	4.50	1.75	1.50	0.00	1	0.00	1	1_1_Training	0.537	0.537	0.537	0.366
15	4.29	4.83	4.83	5.00	4.25	4.25	1.00	2	1.00	2	2_1_Training	0.843	0.843	0.843	0.861
16	1.86	4.50	3.50	3.00	2.75	2.75	0.00	1	0.00	1	1_1_Training	0.652	0.652	0.652	0.248
17	3.00	3.00	2.83	3.50	3.25	4.00	0.00	1	1.00	1	2_1_Training	0.649	0.649	0.649	0.251
18	2.86	2.83	3.33	4.50	3.00	2.25	0.00	1	0.00	1	1_1_Training	0.616	0.616	0.616	0.279
19	2.29	1.50	2.83	5.00	4.00	4.00	1.00	2	1.00	1	2_1_Training	0.777	0.777	0.777	0.176
20	1.29	4.17	1.83	5.00	3.00	2.25	0.00	1	0.00	1	1_1_Training	0.803	0.803	0.803	0.166

5.8.2.5 Sigmoid Kernel fonksiyonu Kullanarak DVM Modelinin İncelenmesi

Doğruluk oranı en düşük olan DVM algoritmasıdır. Diğer algoritmalarla göre farklı değişken önem sıraları vermiştir. Göreceli önem değeri %25,55 ile Faktör_5 (İş Güvencesi) ve %25,32 ile Faktör_3 (İş Arkadaşları) en etken iki faktör sigmoid DVM modeli tarafından belirlenmiştir. Diğer taraftan, %18,79 önem oranı ile Faktör_4 (Çalışma Şartları), %18,79 önem oranı ile Faktör_1(Yönetimin Tavrı) ve son olarak ta %12,09 ile Faktör_2 (Ücret/Ödüllendirme) göreceli olarak en az etken faktör olarak model tarafından belirtilmiştir. Faktör_5 (İş Güvencesi) çok terimli DVM modelinde en az göreceli öneme sahip iken doğruluk düzeyi az olan sigmoid DVM modelinde en önemli etken faktör olarak gözükmemektedir. Dolayısıyla, model seçimimizde performans ölçümleri göz önüne alınarak doğruluk düzeyi yüksek olan çok terimli (polynomial) DVM en iyi DVM modeli olarak kabul edilerek analiz sonuçları bilgi çalışanı memnuniyeti faktörleri değerlendirilecektir.



Tablo 35. Sigmoid DVM Faktörlerin göreceli önem oranları (Finansal Performans)

Faktörler	Değişken Önemi
Faktör_5 (İş Güvencesi)	0,2555
Faktör_3 (İş Arkadaşları)	0,2532
Faktör_4 (Çalışma Şartları)	0,1879
Faktör_1(Yönetimin Tavrı)	0,1825
Faktör_2 (Ücret/Ödüllendirme)	0,1209

5.8.3 Finansal Kurum Performans Algısının Karar Ağaçları Modelleri Yardımı İle Detaylı İncelenmesi

5.8.3.1 C&R Tree Karar Ağacı Modelinin Detaylı İncelenmesi

CR&T Tree karar ağacı modelinin modelimizi doğru tahmin etme oranı %74,56'dır. 550 adet veri kümesi eğitildikten sonra 114 adet veri kümesi test edilerek modelimizin doğruluk oranı belirlenmiştir. Hem eğitime ve test veri kümelerine ait çakışım matrisleri (coincidence matrix) verilmiştir. Tahmin edilen alan ile gerçek alan arasındaki eşleşmelerin şablonlarını göstermektedir. Matris tablosundaki satırlar gerçek değerlerin sayılarını belirtirken sütunlar ise tahmin edilen değerlerin sayılarını belirtmektedir. Bu matris tahminde meydana gelen sistematik hataları belirlemede faydalı olmaktadır. Modelin doğruluğunun hesaplanmasında çakışım matrisleri kullanılarak elde edilir.

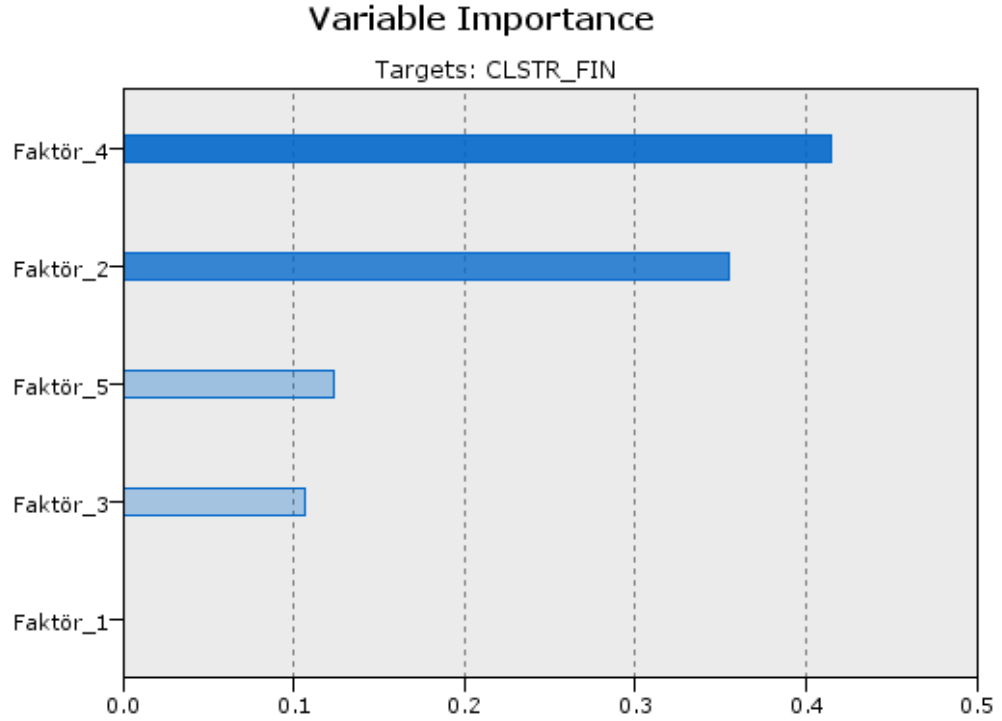
Tablo 36. C&R Tree modelinin performans ölçüm sonuçları

Accuracy					
'Partition'	1_Training			2_Testing	
	Correct	336	74.67%	85	74.56%
	Wrong	114	25.33%	29	25.44%
	Total	450		114	
Coincidence Matrix (rows show actuals)					
'Partition' = 1_Training			'Partition' = 2_Testing		
	1	2		1	2
Başarısız(1)	138	86	Başarısız (1)	37	19
Başarılı (2)	28	198	Başarılı (2)	10	48
Performans Evaluation					
'Partition' = 1_Training			'Partition' = 2_Testing		
1) 0.513			1) 0.472		
2) 0.328			2) 0.342		

Göreceli Değişken Önem Oranları: C&R Tree karar ağacı algoritması kullanılarak elde edilen modelde faktörlerin göreceli önem durumları hem grafik hem de tablo halinde aşağıda verilmiştir. Çalışma şartları %41,45 göreceli önem oranı ile en etken faktör iken, %35,46 oran ile ücret/ödüllendirme faktörü ikinci en etken faktör olarak elde edilmiştir. Çok terimli DVM (Destek Vektör Makineleri) modeli ile RBF DVM modelinde de çalışma şartları en önemli faktör olarak belirlenmişti. Ayrıca ücret/ödüllendirme faktörü de DVM modellerinde olduğu gibi ikinci önemli faktör olarak C&R Tree karar ağacı modeli ile paralellik göstermektedir. İş güvencesi ve iş arkadaşları sırasıyla %12,4 ve %10,69 ile birbirlerine yakın göreceli önem oranları ile sıralanmaktadır. Fakat yönetim tavrı %0 göreceli önem oranı ile en az öneme sahip (etkisiz) bir faktör model tarafından belirtilmiştir.

Tablo 37. C&R Tree modelinde faktörlerin göreceli önem oranları (finansal performans)

Faktörler	Değişken Önemi
Faktör_4 (Çalışma Şartları)	0.4145
Faktör_2 (Ücret/Ödüllendirme)	0.3546
Faktör_5 (İş Güvencesi)	0.1240
Faktör_3 (İş Arkadaşları)	0.1069
Faktör_1(Yönetimin Tavrı)	0.0000



Şekil 27. C&R Tree modelinde faktörlerin göreceli önem oranları (finansal performans)

Eğitme veri kümesi ve test veri kümesi ile elde edilen karar ağaçları 6 seviye ve 11 uç (terminal) düğümlerden oluşmaktadır. Birinci düğümde bazı istatistikler verilmektedir, finansal başarısız (1) kayıt sayısı 224 adet (%49,8) ve finansal başarılı (2) kayıt sayısı 226 adet (%50,2) olmak üzere 450 adet eğitim veri kümesi üzerinden karar ağacı elde edilmektedir. Başlangıç düğümde finansal başarılı olanları “isabet oranı (hit rate)” olan %50 değerini karar ağacı modeli kullanılarak alt grupları hedef alarak bu oranın geliştirip geliştiremeyeceğimize bakabiliriz. Elde edilen C&R Tree karar ağacında 6 seviye 11 adet uç (terminal) düğümden oluşmaktadır. Budama seçeneği seçilmedi, sadece varsayılan olarak karar ağacı oluşturulmuştur. Budama (pruning) seçeneği seçilerek elde edilen karar ağacının çok karmaşık olmayan sadece 3 seviye ve 3 terminal düğümden oluştuğu için yorum yapmak için uygun bir ağaç ortaya çıkmadığından budama yapılmadan elde edilen karar ağacı incelenmektedir.

Finansal başarılı (2) değişkeni baz alınarak incelendiğinde, karar ağacında ilk bölünme ücret/ödüllendirme faktörü üzerinden yapılmıştır. Sırasıyla diğer bölünmeler, çalışma şartları, iş güvencesi ve iş arkadaşları faktörleri üzerinden yapılmıştır. Ağacın sağ tarafı ağırlıklı olarak finansal başarılı (2) olanları ve ilgili faktörleri belirtmektedir. Ücret/ödüllendirme değeri 3.07'den büyük olanlar için finansal başarılı olanların isabet etme (hit) oranı %50 den %72 gibi yüksek bir orana yükselmekte iken, ücret/ödüllendirme değeri 3.07'den küçük olanlar için finansal başarılı (2) isabet oranı %50den %30'a kadar düşmektedir. Fakat ağacın bu kısmında yani ücret/ödüllendirme değeri 3.07'den küçük olan sol kısım alt gruplara bölünerek yönetimin tavrı ve iş arkadaşları faktörlerini içeren faktörlere bölünerek 8 ve 17 no'lu düğümlerde %53 gibi finansal başarılı (2) isabet oranı ilk oran olan %50'den yüksek çıkmaktadır. Kurallar kümesinde ve karar ağacında bu oranlar açık şekilde görülmektedir. Örneğin, ücret/ödüllendirme değeri 3.07'den küçük, yönetimin tavrı 3,21'den büyük ve tekrar ücret/ödüllendirme değeri 2,36'dan büyük olduğu zaman finansal başarılı algı (2) isabet oranı %53 olduğu görülmektedir.

Finansal başarı isabet oranı 100% gibi en yüksek olan düğüm 12 numaralı düğümdür. Karar ağacından ve kurallar kümesinden görüldüğü gibi, ücret/ödüllendirme faktörü 3,07 büyük, çalışma şartları 3,25 büyük ve tekrar çalışma şartları 4,92'den büyük olduğu alt gruplarda finansal başarı isabet oranı %100 gibi en yüksek isabet oranı elde edilmektedir. Aynı şekilde karar ağacı ve kurallar kümesine bakıldığı zaman ikinci en yüksek isabet etme oranı (hit) olan %77,9 ile 22 no'lu düğümdür. Bu alt grup karar ağacında ilk isabet oran olan %50,2 değerinden oldukça yüksektir. Bu değeri veren alt karar ağacı grubunda eğer Ücret/Ödüllendirme 3,07'den büyük, Çalışma Şartları 3,25 ile 4,92 arasında ve iş güvencesi faktör değeri 4,25'ten büyük olduğu zaman finansal başarılı (2) bağımlı değişkeni isabet etme oranı yaklaşık %78 olmaktadır. Bir sonraki en yüksek finansal başarılı (2) bağımlı değişkeni için isabet oranı yüksek olan düğüm 26 numaralı düğümdür. Burada, Ücret/Ödüllendirme 3,07'den büyük, çalışma şartları 3,25 ve 4,92 değerleri arasında, iş güvencesi 4,25 değerinden küçük ve iş arkadaşları faktör

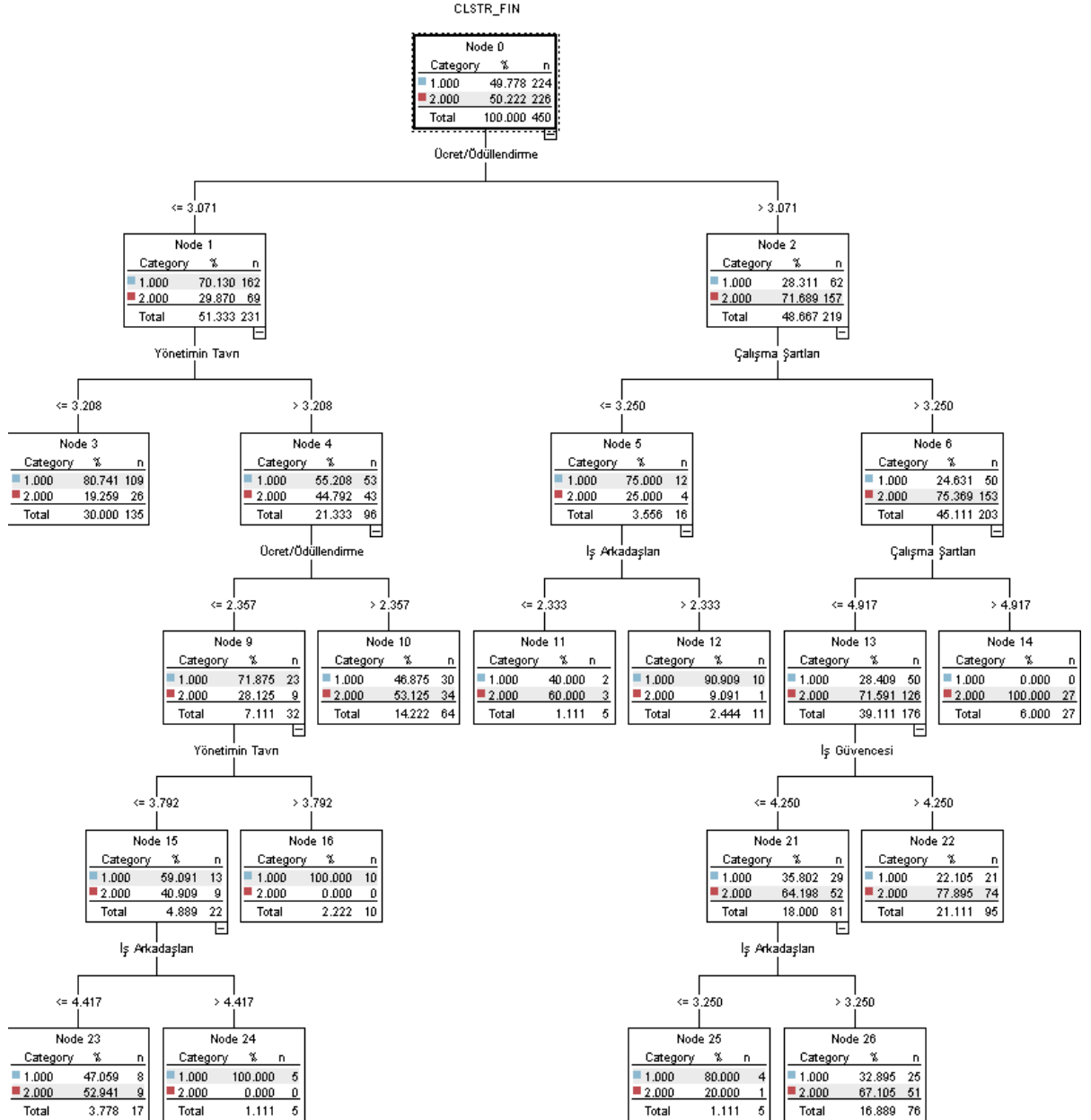
değeri 3,25'ten büyük olduğu zaman, finansal başarı (2) isabet etme oranı ilk %67 olmaktadır. Genel başarılı bağımlı değişkeni için isabet oranı olan %50'den büyük alt karar ağacı gruplarından bir diğeri %60 gibi bir isabet oranı olan grup Ücret/Ödüllendirme > 3,07, çalışma Şartları $\leq$ 3,25 ve iş Arkadaşları $\leq$ 2,33.

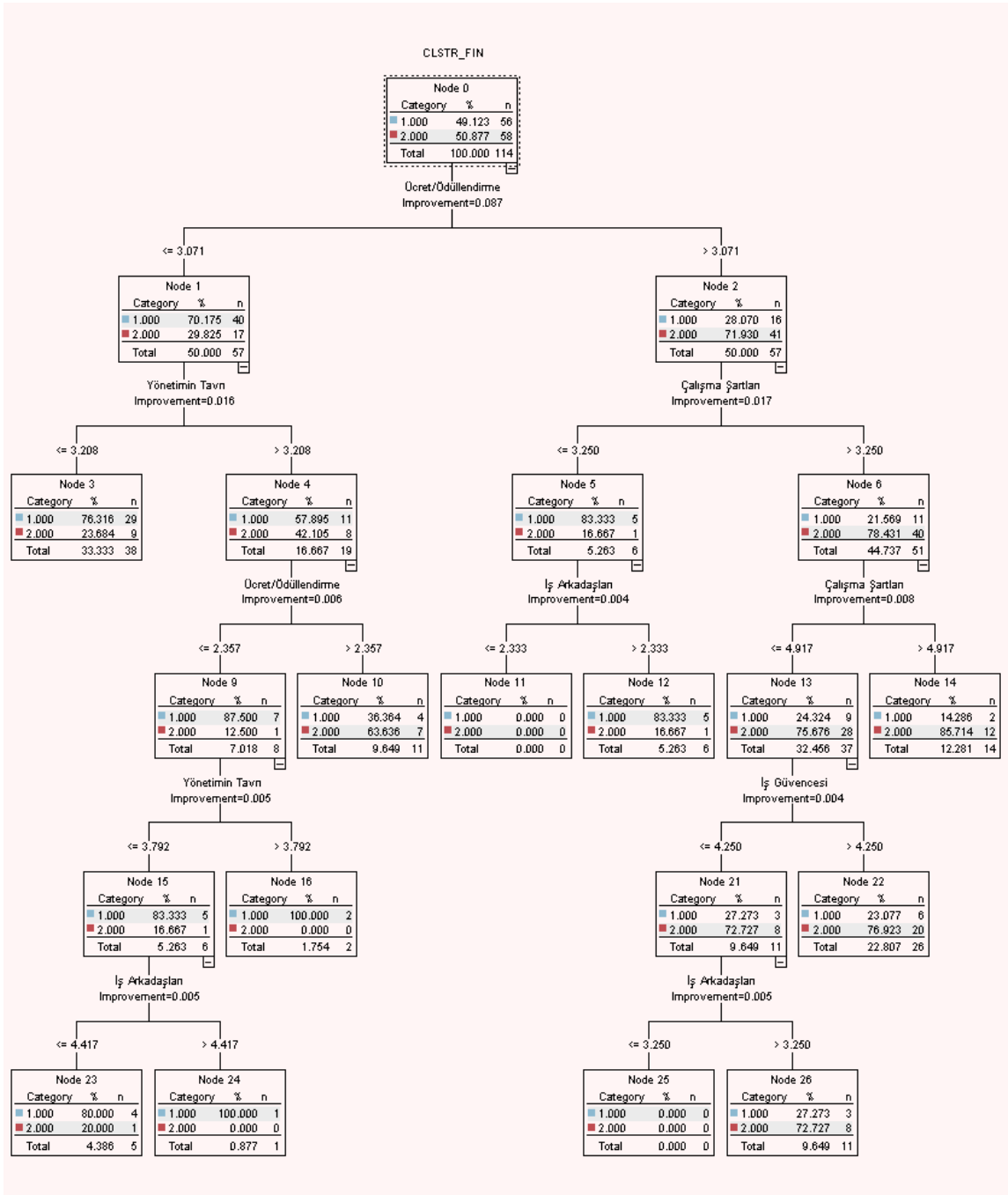
Aynı şekilde, finansal performans başarısızlık (1) algısı karar ağacı ve kurallar kümesi kullanılarak isabet etme (hit rate) oranları ve faktörler arasındaki ağ ilişkisi incelenebilir.

Karar ağacı yorumu kolay olduğu için popüler bir karar verme modelidir. Bizim modelimizde kullandığımız C&R Tree karar ağacı bilgi çalışan memnuniyeti faktörlerinin finansal başarı(2) algısı üzerindeki etkilerini anlaşılır bir şekilde göstermektedir.

Kurallar Kümesi:

Ücret/Ödüllendirme $\leq$ 3.07 [Mode: 1] (231)
Yönetimin Tavrı $\leq$ 3.21 [Mode: 1] $\Rightarrow$ 1 (135; 0.807)
Yönetimin Tavrı > 3.21 [Mode: 1] (96)
 Ücret/Ödüllendirme $\leq$ 2.36 [Mode: 1] (32)
 Yönetimin Tavrı $\leq$ 3.79 [Mode: 1] (22)
 İş Arkadaşları $\leq$ 4.42 [Mode: 2] $\Rightarrow$ 2 (17; 0.529)
 İş Arkadaşları > 4.42 [Mode: 1] $\Rightarrow$ 1 (5; 1.0)
 Yönetimin Tavrı > 3.79 [Mode: 1] $\Rightarrow$ 1 (10; 1.0)
 Ücret/Ödüllendirme > 2.36 [Mode: 2] $\Rightarrow$ 2 (64; 0.531)
Ücret/Ödüllendirme > 3.07 [Mode: 2] (219)
 Çalışma Şartları $\leq$ 3.25 [Mode: 1] (16)
 İş Arkadaşları $\leq$ 2.33 [Mode: 2] $\Rightarrow$ 2 (5; 0.6)
 İş Arkadaşları > 2.33 [Mode: 1] $\Rightarrow$ 1 (11; 0.909)
 Çalışma Şartları > 3.25 [Mode: 2] (203)
 Çalışma Şartları $\leq$ 4.92 [Mode: 2] (176)
 İş Güvencesi $\leq$ 4.25 [Mode: 2] (81)
 İş Arkadaşları $\leq$ 3.25 [Mode: 1] $\Rightarrow$ 1 (5; 0.8)
 İş Arkadaşları > 3.25 [Mode: 2] $\Rightarrow$ 2 (76; 0.671)
 İş Güvencesi > 4.25 [Mode: 2] $\Rightarrow$ 2 (95; 0.779)
 Çalışma Şartları > 4.92 [Mode: 2] $\Rightarrow$ 2 (27; 1.0)





Şekil 29. Test veri kümesi (Testing set) ile elde edilen C&R Tree karar ağacı modelinin gösterimi

Tahmin Sonuçları: C&R Tree Karar ağacı modeli tarafından elde edilen skorları ya da tahminler aşağıdaki gibi ilk 20 kayıt için elde edilmiştir. Bağımlı kategorik değişken CLSTR_FIN (Finansal Performans) gösterirken, \$R-CLSTR_FIN karar ağacı modeli tarafından üretilen tahmin skorlarını belirten değişkendir. Üretilen tahmin değerleri ile gerçek değerler tablo üzerinde kıyaslanır. Model çoğu zaman doğru tahminde bulunurken bazı kayıtlarda doğru tahmin yapılmamıştır. Zaten, yukarıda modelin doğruluk oranı %75 oranında olduğu belirtilmiştir. \$RC-CLSTR_FIN yeni değişkeni güven değerlerini (confidence value) belirtir.

Tablo 38. C&R Tree karar ağacı modelinin tahmin sonuçları

Table (81 fields, 564 records)															
	File	Edit	Generate												
	ir_3	Faktor_4	Faktor_5	Fin_Perf	Non_Fin_	Bin_Fin_	CLSTR_	Bin_Non_Fin_	CLSTR_NON_	Partition	\$R-CLSTR_FIN	\$RC-CLSTR_FIN	\$RRP-CLSTR_FIN	\$RAP-CLSTR_FIN	\$RI-CLSTR_FIN
1	4.00	4.83	4.50	3.75	4.00	1.00	2	1.00	2	1_Training	2	0.773	0.773	0.750	22
2	4.33	4.50	3.50	3.00	3.25	0.00	1	0.00	0	1_2_Testing	2	0.530	0.530	0.615	10
3	4.33	4.00	4.50	4.00	3.75	1.00	2	1.00	2	1_Training	2	0.773	0.773	0.750	22
4	3.00	2.50	3.50	1.50	2.25	0.00	1	0.00	0	1_1_Training	1	0.803	0.197	0.250	3
5	3.67	3.83	4.50	3.50	3.75	1.00	2	1.00	2	1_1_Training	2	0.526	0.526	0.286	23
6	3.83	4.17	5.00	3.50	3.25	1.00	2	0.00	0	1_1_Training	2	0.526	0.526	0.286	23
7	4.00	3.67	2.00	3.00	2.50	0.00	1	0.00	0	1_1_Training	2	0.667	0.667	0.692	26
8	4.50	4.83	1.00	3.00	3.50	0.00	1	1.00	0	1_1_Training	2	0.667	0.667	0.692	26
9	4.33	4.50	4.00	3.25	3.00	0.00	1	0.00	0	1_1_Training	2	0.667	0.667	0.692	26
10	2.83	4.17	3.00	3.00	3.00	0.00	1	0.00	0	1_1_Training	1	0.714	0.286	0.692	25
11	3.83	2.50	1.00	2.75	2.25	0.00	1	0.00	0	1_1_Training	1	0.803	0.197	0.250	3
12	4.00	4.67	4.00	3.00	3.00	0.00	1	0.00	0	1_1_Training	2	0.667	0.667	0.692	26
13	4.67	4.33	2.50	3.50	3.25	1.00	2	0.00	0	1_1_Training	2	0.667	0.667	0.692	26
14	3.67	3.83	4.50	1.75	1.50	0.00	1	0.00	0	1_1_Training	2	0.530	0.530	0.615	10
15	3.83	4.83	5.00	4.25	4.25	1.00	2	1.00	2	1_1_Training	2	0.773	0.773	0.750	22
16	4.50	3.50	3.00	2.75	2.75	0.00	1	0.00	0	1_1_Training	1	0.857	0.143	0.333	24
17	3.00	2.83	3.50	3.25	4.00	0.00	1	1.00	0	2_1_Training	1	0.803	0.197	0.250	3
18	2.83	3.33	4.50	3.00	2.25	0.00	1	0.00	0	1_1_Training	1	0.803	0.197	0.250	3
19	4.50	2.83	5.00	4.00	4.00	1.00	2	1.00	0	2_1_Training	1	0.803	0.197	0.250	3
20	4.17	1.83	5.00	3.00	2.25	0.00	1	0.00	0	1_1_Training	2	0.526	0.526	0.286	23

C&R Tree modelinin özet bilgileri:

Fields

Target

CLSTR_FIN

Inputs

Faktör_1

Faktör_2

Faktör_3

Faktör_4

Faktör_5

Build Settings

Use partitioned data: true

Partition: Partition

Calculate variable importance: true

Calculate raw propensity scores: true

Calculate adjusted propensity scores: true

Based on: Testing partition

Use frequency: false
Use weight: false
Levels below root: 5
Mode: Simple
Use misclassification costs: true
Training Summary
Algorithm: C&R Tree
Model type: Classification

Kazanç (Gains): Karar ağaçları algoritmalarında kazançlar (gains) ölçümü ağaçtaki tüm uç (terminal) düğümler için istatistikleri görüntüler. Kazançlar (gains) verilen bir düğümden (node) oran ya da ortalamanın tüm ortalama ne kadar farklı olduğunu ölçer. Fark ne kadar büyük olursa karar verme aracı olarak ilgili karar ağacı o kadar çok faydalı olduğunu gösterir. Sınıflandırma ağaçları için, kazanç endeksi yüzdesi (gain index percentage) her bir düğümden verilen bir hedef kategorisinin oranının genel orandan ne kadar farklı olduğunu belirtir.

Aşağıda elde edilen kazançlar endeksi değerleri incelendiği zaman, karar ağaçları incelenirken elde edilen sonuçlar göz önüne alındığı zaman 14, 22, 26 ve 11 numaralı düğümler en yüksek isabet oranlarını vermiştir. Beklenildiği gibi, 14, 22 ve 26 düğümleri sırasıyla en yüksek endeksi değerlerini vermektedir. Endeksi değeri %100 büyük olması rastgele seçim yerine bu düğümlerdeki kayıtların seçimi ile bilgi çalışanların finansal başarılı olanların elde edilmesi olasılığının daha yüksek olduğunu belirtir. 14 numaralı düğüm tüm örneklem için mümkün olan en yüksek orana sahiptir. Bu düğümün endeksi değeri 168 olduğundan, rastgele seçim yolu ile bu düğümden kayıtlar ile isabet etme oranı elde etmek yaklaşık 1,68 kat olduğunu belirtir.

Örneğin, test veri kümesinde finansal algı başarılı (2) değişkeni seçildiğinde, 14 numaralı düğümün endeksi (%) değeri 168,47 bir bütün olarak veri kümesi için bu düğümden kayıtların 1,68 kez olasılıkla hedef kategori altına olmasını belirtir. Yani, 14 numaralı düğümden finansal performans algısı başarılı (2) kategorisindeki bilgi

çalışanları toplam nüfusa oranla 1,68 kat daha finansal performans açısından başarılıdılar.

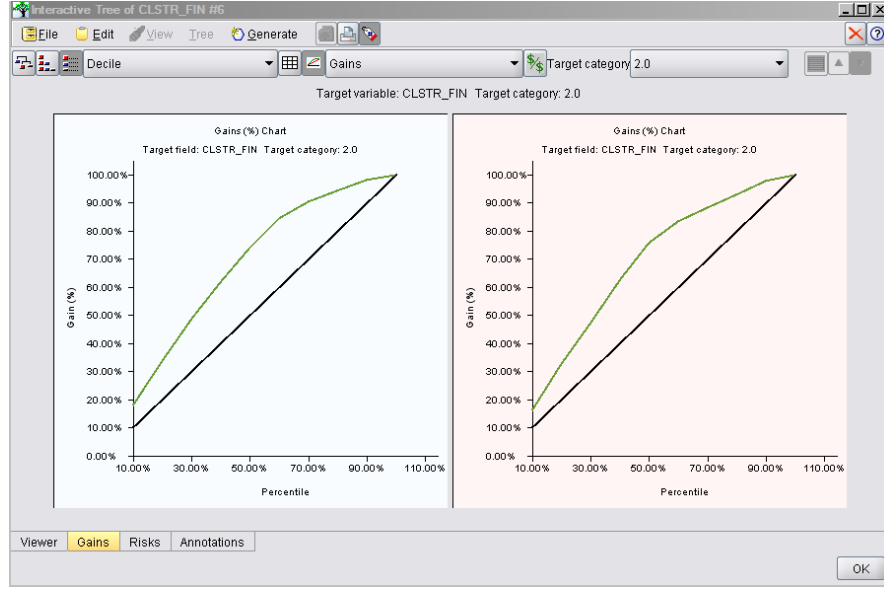
CR&R Tree karar ağacı algoritmasına ait olan düğüm sayıları desillere bölünerek incelenmiştir. İlk desilde (%10) sadece bir adet uç düğüm (14 numaralı uç düğüm mevcut). Test veri kümesi kısmında incelemelere devam edilirse, bu desilde 11 adet kayıt mevcut (9%), başarılı(2) finansal performans kategorisine ait kazanç sayısı 9 adet kayıt yani kazanç (gain %) oranı %16. Kazanılan %16'lık kayıttın %86'ı (response %) kurumların finansal olarak başarılı (2) hedef kategorisi altında bulunmaktadır.

Tablo 39. C&R Tree modelinde düğümlerin performans bilgileri

Training Sample							Testing Sample						
Nodes	Percentile	Percentile: n	Gain: n	Gain (%)	Response (%)	Index (%)	Nodes	Percentile	Percentile: n	Gain: n	Gain (%)	Response (%)	Index (%)
14,22	10.00	45.00	41.00	18.15	91.16	181.51	14	10.00	11.00	9.00	16.26	85.71	168.47
22	20.00	90.00	76.00	33.66	84.53	168.30	14,22	20.00	23.00	19.00	32.63	82.27	161.71
22,26	30.00	135.00	110.00	48.55	81.28	161.83	22	30.00	34.00	27.00	47.21	80.54	158.31
26	40.00	180.00	140.00	61.91	77.73	154.78	22,26	40.00	46.00	36.00	62.70	79.05	155.38
26,11,10	50.00	225.00	187.00	73.76	74.08	147.51	26,11,10	50.00	57.00	44.00	75.55	76.97	151.10
10,23	60.00	270.00	191.00	84.33	70.59	140.55	10,23,25,3	60.00	68.00	48.00	83.17	70.94	139.43
23,25,3	70.00	315.00	204.00	90.27	64.76	128.96	3	70.00	80.00	51.00	88.07	63.85	125.50
3	80.00	360.00	213.00	94.10	59.08	117.63	3	80.00	91.00	54.00	92.56	58.99	115.95
3	90.00	405.00	221.00	97.94	54.65	108.82	3	90.00	103.00	57.00	97.46	54.88	107.87
3,12,16,24	100.00	450.00	226.00	100.00	50.22	100.00	3,12,16,24	100.00	114.00	58.00	100.00	50.88	100.00

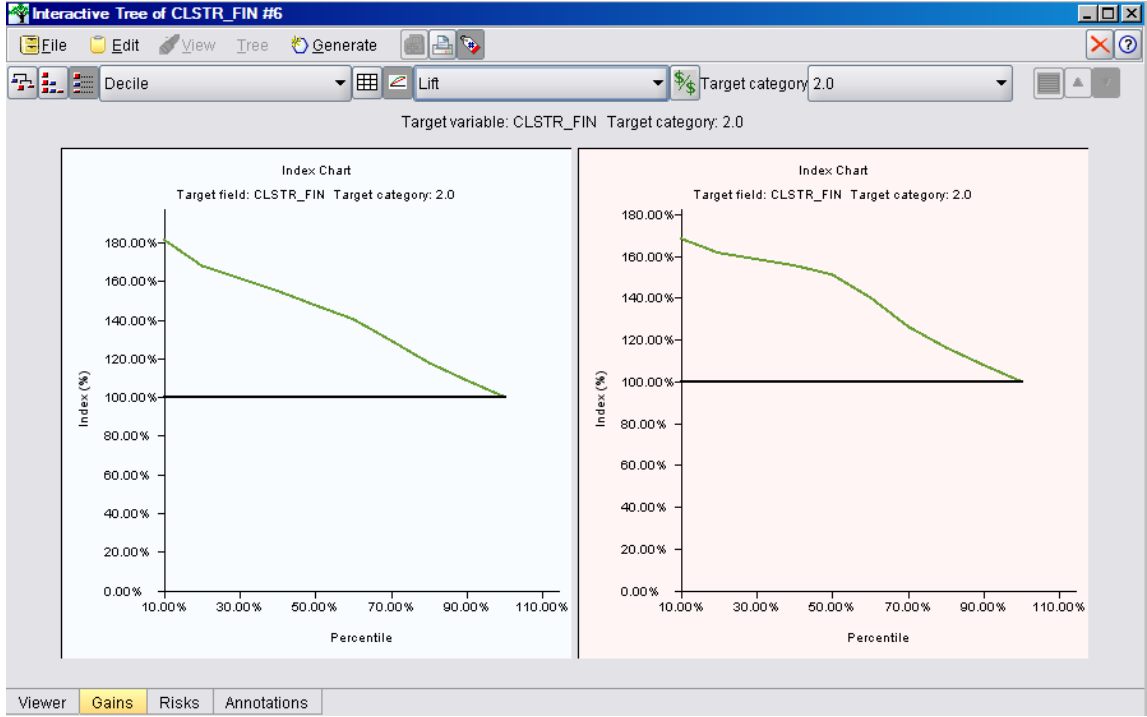
Kazançlar grafiği (gains chart): Kazançlar (gains) ağaçtaki kayıtların toplam isabet etme sayısına oranla her bir yüzdelik dilime (desil) isabet eden kayıtların oranı olarak tanımlanır. Diyagonal doğru eğer model kullanılmaz ise tüm örneklem için beklenen cevabın (expected response) grafiğini çizer. Eğri, kazanca dayalı yüksek yüzdelik dilimlerde yer alan kayıtları dâhil ederek cevap oranının (response) ne kadar geliştirilebileceğini belirtir. İyi bir model için, kazançlar grafiği (gains chart) %100'a kadar dik bir şekilde yükselir ve daha sonra düz bir şekil oluşturur.

Örneğin, test veri kümesinde en üst %55 dâhil edildiği zaman %80 oranında pozitif cevap (finansal başarı performans) kazancı elde edileceğini belirtir. Eğrinin eğimi ne kadar dik ise, kazanç (gain) o kadar yüksek olduğunu gösterir.



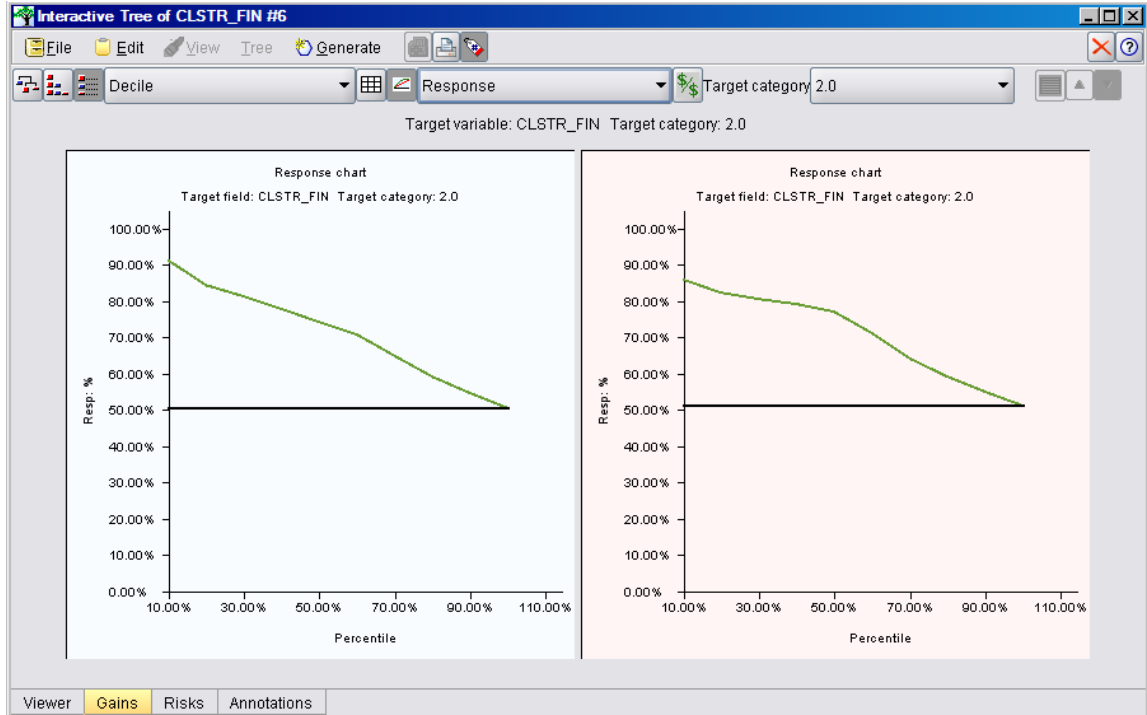
Şekil 30. C&R Tree modelinin kazançlar grafiği

“Lift” Grafiği: Lift(kaldırma) grafiği her bir yüzdelik dilimdeki isabet edilen (hits) kayıtların yüzdeleri ile eğitime veri kümesindeki tüm isabet edilen kayıtların (hits) oranları ile kıyaslama yapmayı sağlar. İyi bir model için, “lift” grafiği sol tarafta %100 (1,0) çok üzerinde başlar, sağa doğru hareket ederken yüksek bir düz plato şeklinde devam eder ve sağ tarafa doğru yani %100 (1,0) 'a kadar hızlı bir şekilde düşer. CR&Tree modelimizi kullanılarak elde edilen finansal başarı (2) kategorik değişkeni için elde edilen “lift” (kaldırma) grafiğimize bakarak modelimizin iyi bir model olduğunu belirtir.



Şekil 31. C&R Tree modelinin "lift" grafiği

Yanıt (Response chart) grafiği: Lift grafiğine çok benzerdir. Her bir yüzdelik dilimdeki isabet edilen (hits) kayıtların bir oranı olarak tanımlanır. 100% oranına yakın bir yerden başlar ve dereceli olarak genel yanıt (response) oranına kadar azalır. İyi bir model için, doğru %100 yakınında başlar, sağa doğru hareket ederken düz bir plato şeklinde devam eder ve grafiğin sağ tarafında genel cevap (overall response rate) oranına kadar keskin bir şekilde düşüş yapar. Grafiğimize baktığımızda modelimizin iyi bir model olduğu söylenebilir.



Şekil 32. C&R Tree modelinin yanıt (response) grafiği

Risk: Her bir seviyede yanlış sınıflandırma şansını gösterir. Kategorik tahminler için, risk yanlış şekilde sınıflandırılan kayıtların oranıdır. Yanlış sınıflandırma maliyetleri ya da öncelikli olasılık durumları mevcut olduğu zaman düzeltmeler yapılır. Bu model ile yapılan yanlış sınıflandırma oranı eğitim kümesinde 0,253 ve test veri kümesinde ise 0,254 olarak elde edilmiştir.

Tablo 40. C&R Tree modelinin risk ölçüm sonuçları

The screenshot shows a software window titled 'Interactive Tree of CLSTR_FIN #3'. It has a menu bar with 'File', 'Edit', 'View', 'Tree', and 'Generate'. Below the menu bar, there are two main sections: 'Training' and 'Testing'. Each section contains a 'Misclassification Matrix' table. The 'Training' section also shows 'Risk Estimate' as 0.253 and 'Standard Error' as 0.021. The 'Testing' section shows 'Risk Estimate' as 0.254 and 'Standard Error' as 0.041. At the bottom, there are tabs for 'Viewer', 'Gains', 'Risks', and 'Annotations', with 'Risks' currently selected. An 'OK' button is located at the bottom right.

		Misclassification Matrix		
Risk Estimate	Predicted	1.0	2.0	Total
0.253	Actual	138	86	224
Standard Error		28	198	226
0.021	Total	166	284	450

		Misclassification Matrix		
Risk Estimate	Predicted	1.0	2.0	Total
0.254	Actual	37	19	56
Standard Error		10	48	58
0.041	Total	47	67	114

5.8.3.2 C 5.0 Karar Ağacı Modelinin Detaylı İncelenmesi

Bu modelin elde edilmesinde kullanılan ayarlamalar aşağıdaki gibi yapılmıştır. Burada modelin doğruluğunun hesaplamasında kullanılan yöntemlerden çapraz doğrulama (cross validation) seçilmiştir. Örneklem sayısı küçük veri kümeleri için önerilmektedir.

C 5 modelinin özet bilgileri:

Fields

Target

CLSTR_FIN

Inputs

Faktör_4

Faktör_3

Build Settings

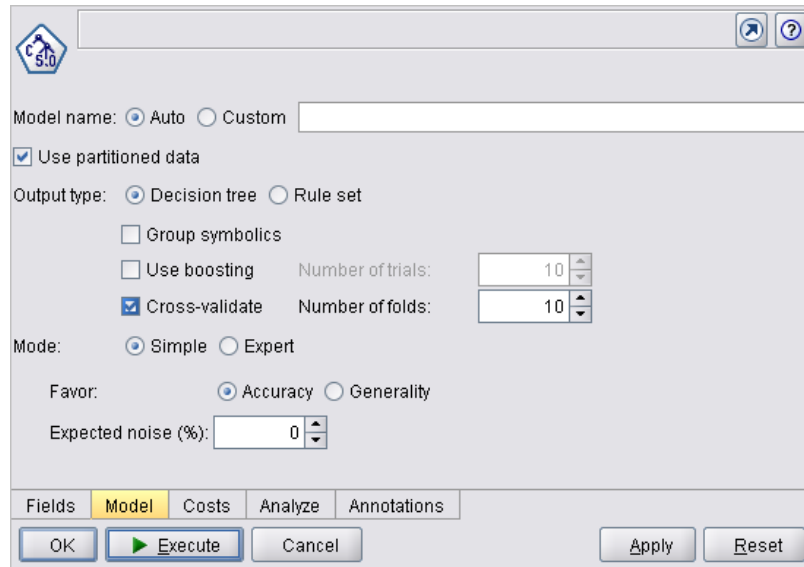
Use partitioned data: true

Partition: Partition

Calculate variable importance: true

Calculate raw propensity scores: true

Calculate adjusted propensity scores: true
Based on: Testing partition
Use weight: false
Output type: Decision tree
Group symbolics: false
Use boosting: false
Cross-validate: true
Number of folds: 10
Mode: Simple
Favor: Accuracy
Expected noise (%): 0
Use misclassification costs: true
Training Summary
Algorithm: C5
Model type: Classification



Şekil 33. C5 karar ağacı modelinin özellikli ayarlarının gösterimi

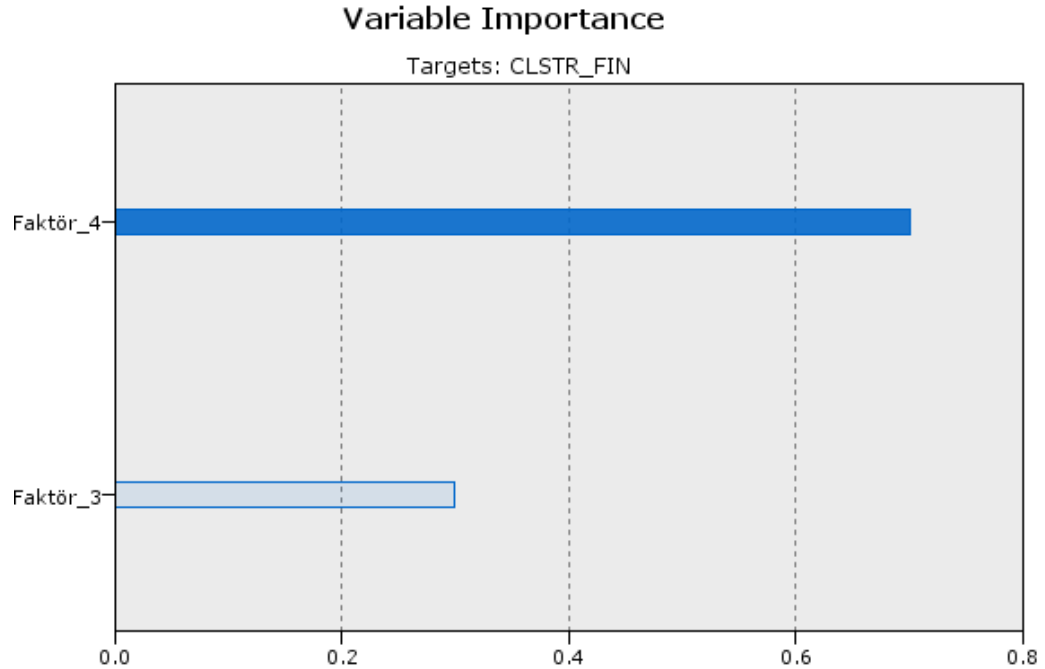
C5.0 karar ağacı modelimiz ile yapacağımız tahminin doğruluk oranı %74,56'dır. C&R Tree ve QUEST karar ağacı modelleri gibi doğruluk oranı aynı olmasına karşılık, C&R Tree modelimizin eğri altında kalan alanı (%80) diğer modellerimize göre en yüksek performans ölçümüne sahiptir.

Accuracy		
'Partition'	1_Training	2_Testing
Correct	322 71.56%	85 74.56%
Wrong	128 28.44%	29 25.44%
Total	450	114

Tablo 41. C5 modelinde faktörlerin önem oranları

Faktörler	Önem oranı
Faktör_4 (Çalışma Şartları)	0.7011
Faktör_3 (İş Arkadaşları)	0.2989

C.5 karar ağacı modelinde bilgi çalışanı memnuniyeti faktörlerinden finansal performans üzerinde en etken faktörler çalışma şartları ve iş arkadaşları faktörleri elde edilmiştir. göreceli önem oranı en yüksek olan faktör %70,11 ile çalışma şartları iken iş arkadaşlarının göreceli önem oranı %29,89 elde edilmiştir.



Şekil 34. C5 modelinde faktörlerin göreceli önem oranları ve sıraları

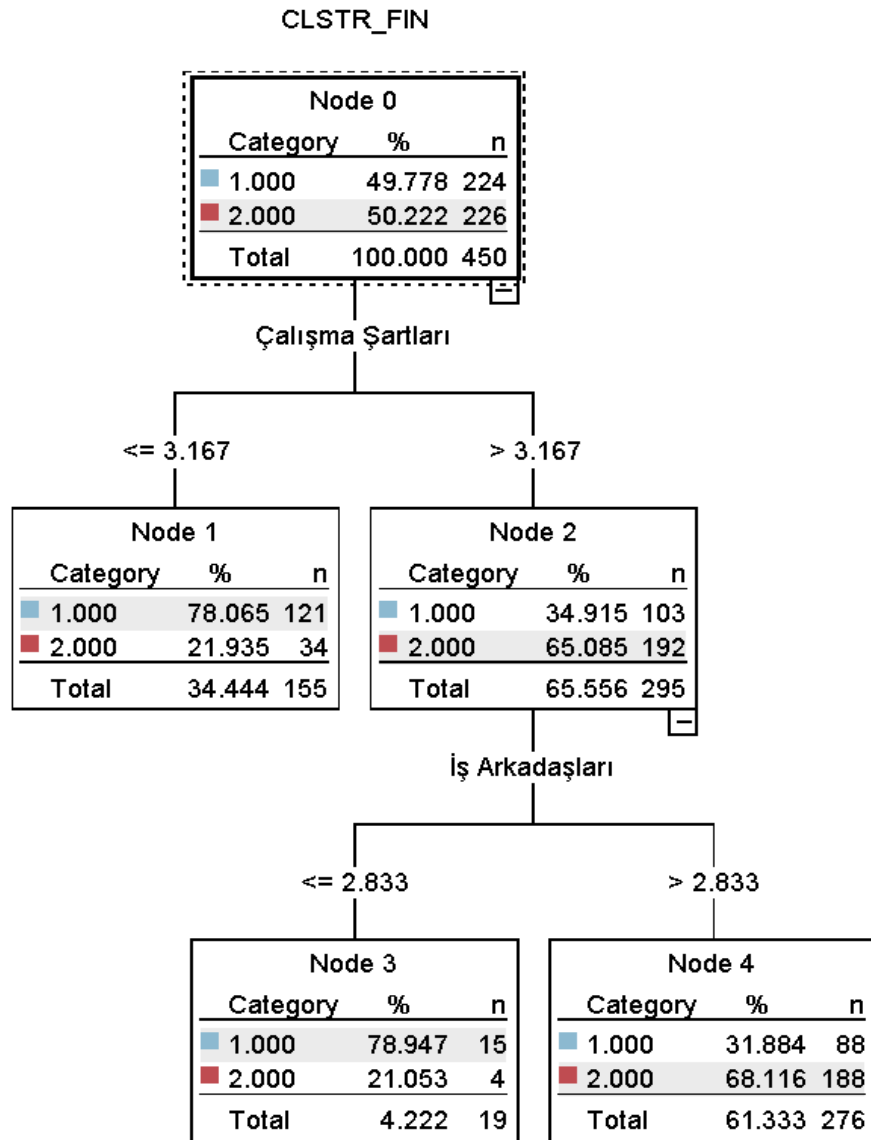
Kurallar Kümesi:

Çalışma Şartları ≤ 3.17 [Mode: 1] $\Rightarrow$ 1 (155; 0.781)

Çalışma Şartları > 3.17 [Mode: 2] (295)

İş Arkadaşları ≤ 2.83 [Mode: 1] $\Rightarrow$ 1 (19; 0.789)

İş Arkadaşları > 2.83 [Mode: 2] $\Rightarrow$ 2 (276; 0.681)



Şekil 35. C5 modelinin karar ağacı olarak gösterimi

Kurallar kümesi ve karar ağacına bakıldığı zaman, başlangıçta finansal başarı (2) performans algı isabet oranı %50.22 iken bu oran alt grupta artmaktadır. 4 numaralı düğümde finansal performans başarı (2) isabet oranı %68'e kadar yükselmektedir. Çalışma şartları 3,17 değerinden yüksek ve iş arkadaşları faktör değeri 2,83'ten büyük olduğu zaman, 4 numaralı düğüm isabet oranı artmaktadır.

C5.0 modeli ile elde edilen tahmini ve gözlenen finansal performans skorları elde edilerek tablo şeklinde sunulmuştur. Doğruluk oranı %75 olduğu bilinmektedir, yani bu model yardımı ile doğru tahmin oranı %75 olarak elde edilmiştir.

Tablo 42. C5 modelinin tahmin sonuçları (finansal performans)

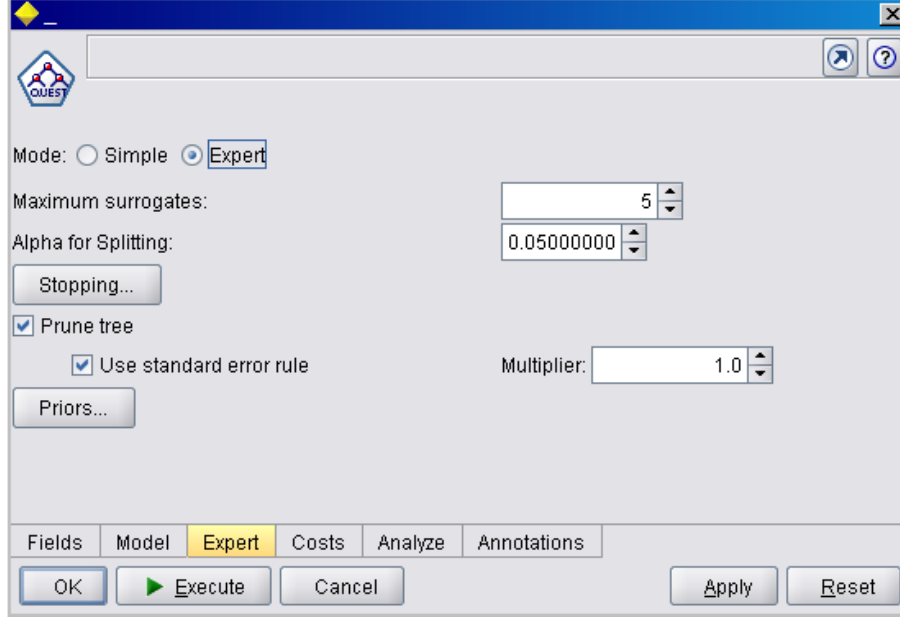
	Faktör_2	Faktör_3	Faktör_4	Faktör_5	Fin_Perf	Non_Fin_...	Bin_Fin_...	CLSTR_...	Bin_Non_Fin_...	CLSTR_NON_...	Partition	\$C-CLSTR_FIN	\$CC-CLSTR_FIN	\$CRP-CLSTR_FIN	\$CAP-CLSTR_FIN
1	4.14	4.00	4.83	4.50	3.75	4.00	1.00	2	1.00	2	1_Training	0.680	0.680	0.680	0.704
2	2.43	4.33	4.50	3.50	3.00	3.25	0.00	1	0.00	1	2_Testing	0.680	0.680	0.680	0.704
3	4.00	4.33	4.00	4.50	4.00	3.75	1.00	2	1.00	2	2_1_Training	0.680	0.680	0.680	0.704
4	1.86	3.00	2.50	3.50	1.50	2.25	0.00	1	0.00	1	1_1_Training	0.777	0.223	0.186	0.186
5	1.86	3.67	3.83	4.50	3.50	3.75	1.00	2	1.00	2	2_1_Training	0.680	0.680	0.680	0.704
6	1.71	3.83	4.17	5.00	3.50	3.25	1.00	2	0.00	1	1_1_Training	0.680	0.680	0.680	0.704
7	3.14	4.00	3.67	2.00	3.00	2.50	0.00	1	0.00	1	1_1_Training	0.680	0.680	0.680	0.704
8	4.71	4.50	4.83	1.00	3.00	3.50	0.00	1	1.00	1	1_1_Training	0.680	0.680	0.680	0.704
9	4.71	4.33	4.50	4.00	3.25	3.00	0.00	1	0.00	1	1_1_Training	0.680	0.680	0.680	0.704
10	3.43	2.83	4.17	3.00	3.00	3.00	0.00	1	0.00	1	1_1_Training	0.680	0.680	0.680	0.704
11	2.57	3.83	2.50	1.00	2.75	2.25	0.00	1	0.00	1	1_1_Training	0.777	0.223	0.186	0.186
12	4.14	4.00	4.67	4.00	3.00	3.00	0.00	1	0.00	1	1_1_Training	0.680	0.680	0.680	0.704
13	4.00	4.67	4.33	2.50	3.50	3.25	1.00	2	0.00	1	1_1_Training	0.680	0.680	0.680	0.704
14	2.43	3.67	3.83	4.50	1.75	1.50	0.00	1	0.00	1	1_1_Training	0.680	0.680	0.680	0.704
15	4.29	4.83	4.83	5.00	4.25	4.25	1.00	2	1.00	2	2_1_Training	0.680	0.680	0.680	0.704
16	1.86	4.50	3.50	3.00	2.75	2.75	0.00	1	0.00	1	1_1_Training	0.680	0.680	0.680	0.704
17	3.00	3.00	2.83	3.50	3.25	4.00	0.00	1	1.00	2	2_1_Training	0.777	0.223	0.186	0.186
18	2.86	2.83	3.33	4.50	3.00	2.25	0.00	1	0.00	1	1_1_Training	0.680	0.680	0.680	0.704
19	2.29	1.50	2.83	5.00	4.00	4.00	1.00	2	1.00	2	2_1_Training	0.777	0.223	0.186	0.186
20	1.29	4.17	1.83	5.00	3.00	2.25	0.00	1	0.00	1	1_1_Training	0.777	0.223	0.186	0.186

5.8.3.3 QUEST Karar Ağacı Modelinin Detaylı İncelenmesi

Diğer Karar ağaçları algoritmaları gibi QUEST karar ağacı modelinin doğru sınıflandırma oranı %74,56 olarak elde edilmiştir. Aynı şekilde, %80 büyüklüğünde ki eğitime (training) veri kümesi ile model eğitilir ve %20 oranındaki veri kümesi ile aynı model test edilerek geçerliği sağlanır.

Accuracy			
'Partition'	1_Training		2_Testing
Correct	329	73.11%	85 74.56%
Wrong	121	26.89%	29 25.44%
Total	450		114

QEUST karar ağacı algoritması ayarlamaları hem şekil hem de çıktı olarak verilmiştir. Kullanılan parametreler ve ilgili değerleri belirtilmiştir.



Şekil 36. QEUST modelinin özellikli ayarlarının gösterimi

QEUST modelinin özet bilgileri:

Fields

Target

CLSTR_FIN

Inputs

Faktör_1

Faktör_2

Faktör_3

Faktör_4

Faktör_5

Build Settings

Use partitioned data: true

Partition: Partition

Calculate variable importance: true

Calculate raw propensity scores: true

Calculate adjusted propensity scores: true

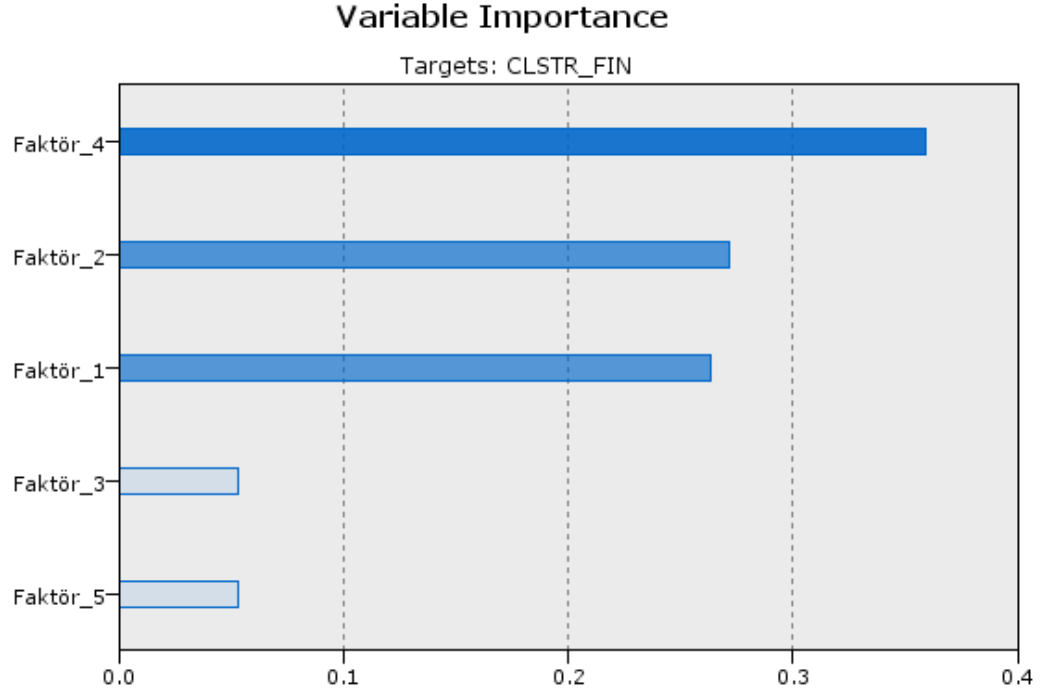
Based on: Testing partition

Use frequency: false

Levels below root: 5
Mode: Simple
Use misclassification costs: true
Training Summary
Algorithm: QUEST
Model type: Classification

Tablo 43. QUEST modelinde faktörlerin göreceli önem oranları

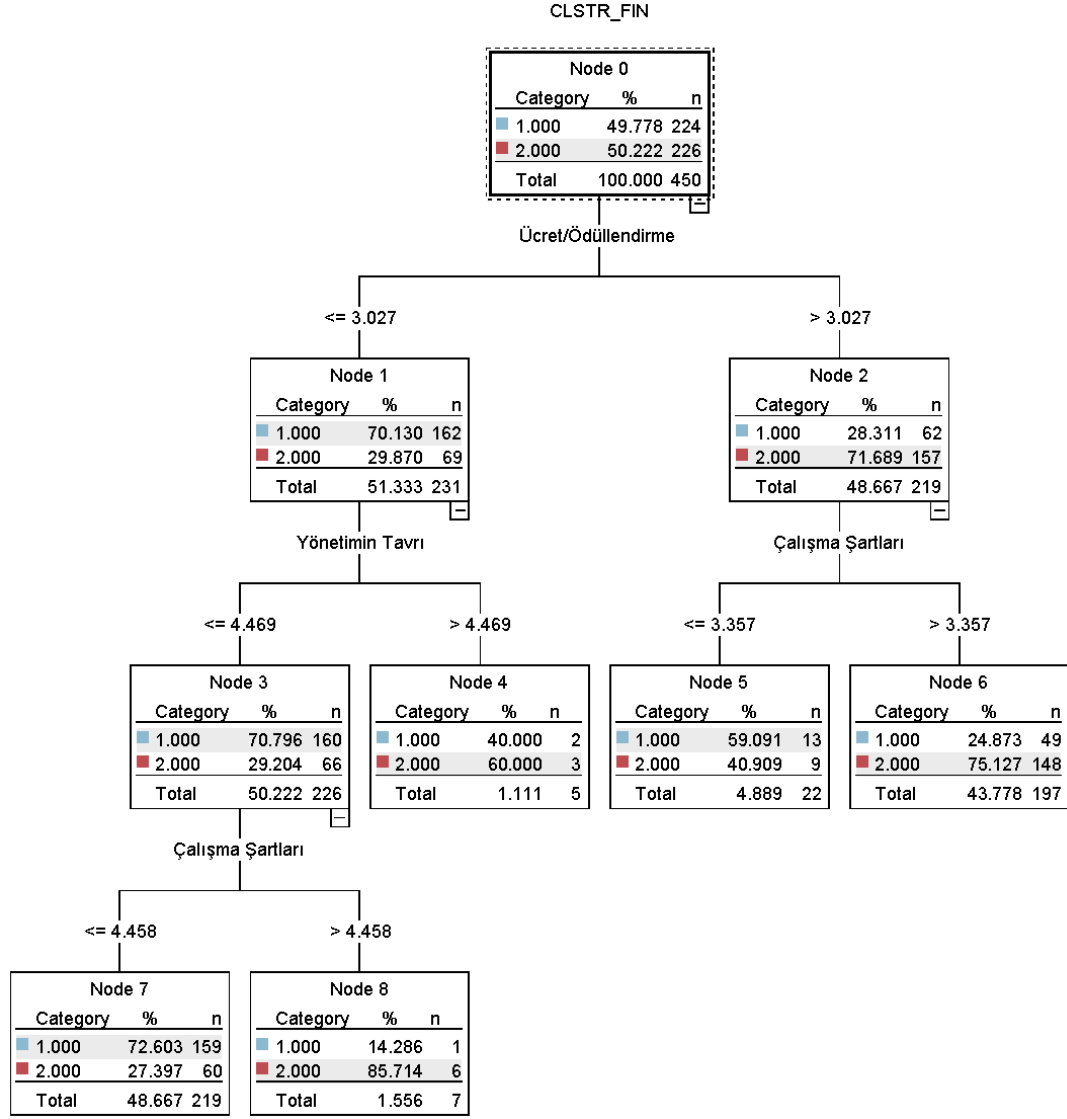
Faktörler	Önem oranı
Faktör_4 (Çalışma Şartları)	0.3587
Faktör_2 (Ücret/Ödüllendirme)	0.2720
Faktör_1(Yönetimin Tavrı)	0.2631
Faktör_5 (İş Güvencesi)	0.0531
Faktör_3 (İş Arkadaşları)	0.0531



Bu model yardımı ile elde edilen faktörlerin göreceli önem oranları ve sıraları mevcuttur. Çalışma şartları bu modele göre %35,87 ile en önemli faktör ilen ücret/ödüllendirme faktörü ise göreceli önem oranı %27,2 ile ikinci en önemli faktör olarak model tarafından belirtilmiştir. Yönetim tavrı faktörü %26,31 önem oranına

sahiptir. İş güvencesi ve iş arkadaşları faktörleri ilk faktörlere göre oldukça düşük (%5,31) göreceli önem oranlarına sahipler. C&R Tree karar ağacı modeli ile QUEST modelini kıyasladığımızda göreceli önem oranı açısından ilk iki faktörler olan çalışma şartları ve ücret/ödüllendirme faktörleri aynıdır.

Model yardımı ile elde edilen karar ağacı ve kurallar kümesi incelendiği zaman, 8,6 ve 4 numaralı düğümlerde finansal başarı performans isabet oranı en yüksek olduğu görülmektedir. En yüksek isabet oranı %85,7 ile 8 numaralı düğümdür. Başlangıçta %50 iken bu oran alt gruplarda bahsedilen düğümlerde oldukça artmaktadır. Yani bu alt guruplara dâhil olan kayıtlarda finansal performans başarı isabet etme oranı oldukça yükselmektedir. İlk bölünme ücret/ödüllendirme faktörü üzerinden yapılmaktadır. Ücret/ödüllendirme 3,03 değerinden küçük ve eşit, yönetimin tavrı 4,47'den küçük ve eşit ve çalışma şartları 4,46 değerinden büyük olduğu faktör değerleri durumunda finansal performans başarı isabet etme oranı yaklaşık %86 ya yükselmektedir. Ayrıca, diğer bir alt grup olan ücret/ödüllendirme 3,03 değerinden büyük ve çalışma şartları 3,36'den büyük olduğu zaman 6 numaralı düğümden finansal performans başarı değişkeni isabet oranı ikinci en yüksek değeri olan %75'e yükselmektedir. Son olarak ücret/ödüllendirme 3,03'ten küçük ve eşit, yönetim tavrı 4,47 faktör değerinden büyük olan alt grupta finansal performans başarı isabet oranı başlangıçtaki %50 değerden %60'a kadar yükselmektedir. C&R Tree modelinde olduğu gibi QUEST karar ağacında da ilk bölünme ücret/ödüllendirme faktörü üzerinden yapılmıştır.



Şekil 37. QUEST modelinin karar ağacı gösterimi

QUEST modelinin kurallar kümesi:

Ücret/Ödüllendirme ≤ 3.03 [Mode: 1] (231)
 Yönetimin Tavrı ≤ 4.47 [Mode: 1] (226)
 Çalışma Şartları ≤ 4.46 [Mode: 1] => 1 (219; 0.726)
 Çalışma Şartları > 4.46 [Mode: 2] => 2 (7; 0.857)
 Yönetimin Tavrı > 4.47 [Mode: 2] => 2 (5; 0.6)
 Ücret/Ödüllendirme > 3.03 [Mode: 2] (219)
 Çalışma Şartları ≤ 3.36 [Mode: 1] => 1 (22; 0.591)
 Çalışma Şartları > 3.36 [Mode: 2] => 2 (197; 0.751)

QUEST modeli test veri kümesini yaklaşık %75 olarak tahmin ettiği belirtilmiştir. Aşağıda tablo şeklinde ilk 20 kayıt için gerçek ve modelin tahmini değerleri belirtilmiştir.

Tablo 44. QUEST modelinin tahmin sonuçları

	Faktör_2	Faktör_3	Faktör_4	Faktör_5	Fin_Perf	Non_Fin	Bin_Fin	CLSTR	Bin_Non_Fin	CLSTR_NON	Partition	\$R-CLSTR_FIN	\$RC-CLSTR_FIN	\$RRP-CLSTR_FIN	\$RAP-CLSTR_FIN
1	4.14	4.00	4.83	4.50	3.75	4.00	1.00	2	1.00	2	1_Training	2	0.749	0.749	0.774
2	2.43	4.33	4.50	3.50	3.00	3.25	0.00	1	0.00	1	2_Testing	2	0.778	0.778	0.000
3	4.00	4.33	4.00	4.50	4.00	3.75	1.00	2	1.00	2	1_Training	2	0.749	0.749	0.774
4	1.86	3.00	2.50	3.50	1.50	2.25	0.00	1	0.00	1	1_Training	1	0.724	0.276	0.291
5	1.86	3.67	3.83	4.50	3.50	3.75	1.00	2	1.00	2	1_Training	1	0.724	0.276	0.291
6	1.71	3.83	4.17	5.00	3.50	3.25	1.00	2	0.00	2	1_Training	1	0.724	0.276	0.291
7	3.14	4.00	3.67	2.00	3.00	2.50	0.00	1	0.00	1	1_Training	2	0.749	0.749	0.774
8	4.71	4.50	4.83	1.00	3.00	3.50	0.00	1	1.00	1	1_Training	2	0.749	0.749	0.774
9	4.71	4.33	4.50	4.00	3.25	3.00	0.00	1	0.00	1	1_Training	2	0.749	0.749	0.774
10	3.43	2.83	4.17	3.00	3.00	3.00	0.00	1	0.00	1	1_Training	2	0.749	0.749	0.774
11	2.57	3.83	2.50	1.00	2.75	2.25	0.00	1	0.00	1	1_Training	1	0.724	0.276	0.291
12	4.14	4.00	4.67	4.00	3.00	3.00	0.00	1	0.00	1	1_Training	2	0.749	0.749	0.774
13	4.00	4.67	4.33	2.50	3.50	3.25	1.00	2	0.00	1	1_Training	2	0.749	0.749	0.774
14	2.43	3.67	3.83	4.50	1.75	1.50	0.00	1	0.00	1	1_Training	1	0.724	0.276	0.291
15	4.29	4.83	4.83	5.00	4.25	4.25	1.00	2	1.00	2	1_Training	2	0.749	0.749	0.774
16	1.86	4.50	3.50	3.00	2.75	2.75	0.00	1	0.00	1	1_Training	1	0.724	0.276	0.291
17	3.00	3.00	2.83	3.50	3.25	4.00	0.00	1	1.00	2	1_Training	1	0.724	0.276	0.291
18	2.86	2.83	3.33	4.50	3.00	2.25	0.00	1	0.00	1	1_Training	1	0.724	0.276	0.291
19	2.29	1.50	2.83	5.00	4.00	4.00	1.00	2	1.00	2	1_Training	1	0.724	0.276	0.291
20	1.29	4.17	1.83	5.00	3.00	2.25	0.00	1	0.00	1	1_Training	1	0.724	0.276	0.291

5.8.3.4 CHAID Karar Ağacı Modelinin Detaylı İncelenmesi

Elimizdeki veri kümesi için doğruluk oranı en düşük karar ağacı modellerindendir. Aynı veri kümesinin %80 (450) eğitildikten sonra, %20 (114) miktarı test edilmiştir. Buna göre test veri kümesinde modelimizin doğru tahmin etme oranı yaklaşık olarak %73 olarak elde edilmiştir.

Accuracy			
'Partition'	1_Training		2_Testing
Correct	327	72.67%	83 72.81%
Wrong	123	27.33%	31 27.19%
Total	450		114

CHAID modelinin özet bilgileri:

Fields

Target

CLSTR_FIN

Inputs

Faktör_2

Faktör_5

Build Settings

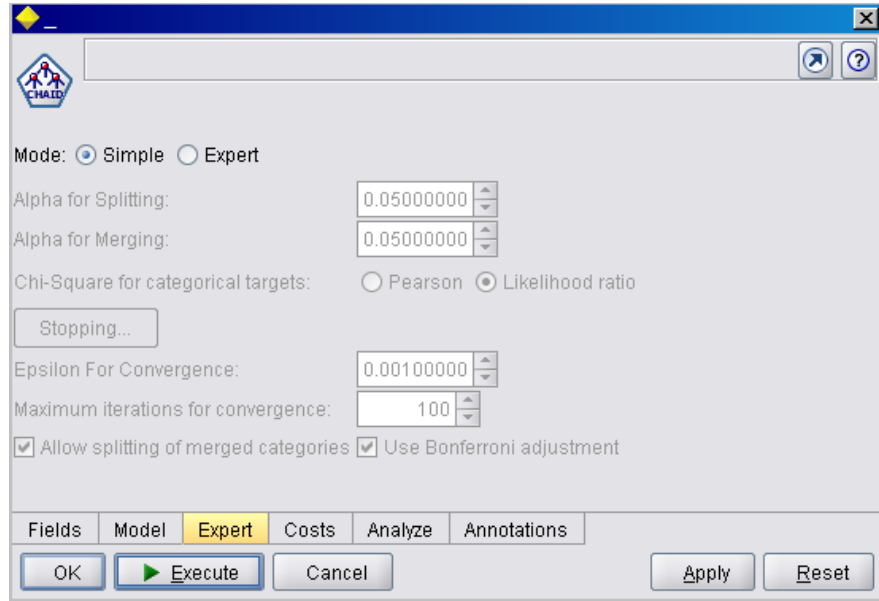
Use partitioned data: true

Partition: Partition

Calculate variable importance: true

Calculate raw propensity scores: false

Calculate adjusted propensity scores: false
Use frequency: false
Use weight: false
Levels below root: 5
Mode: Simple
Use misclassification costs: true
Training Summary
Algorithm: CHAID
Model type: Classification

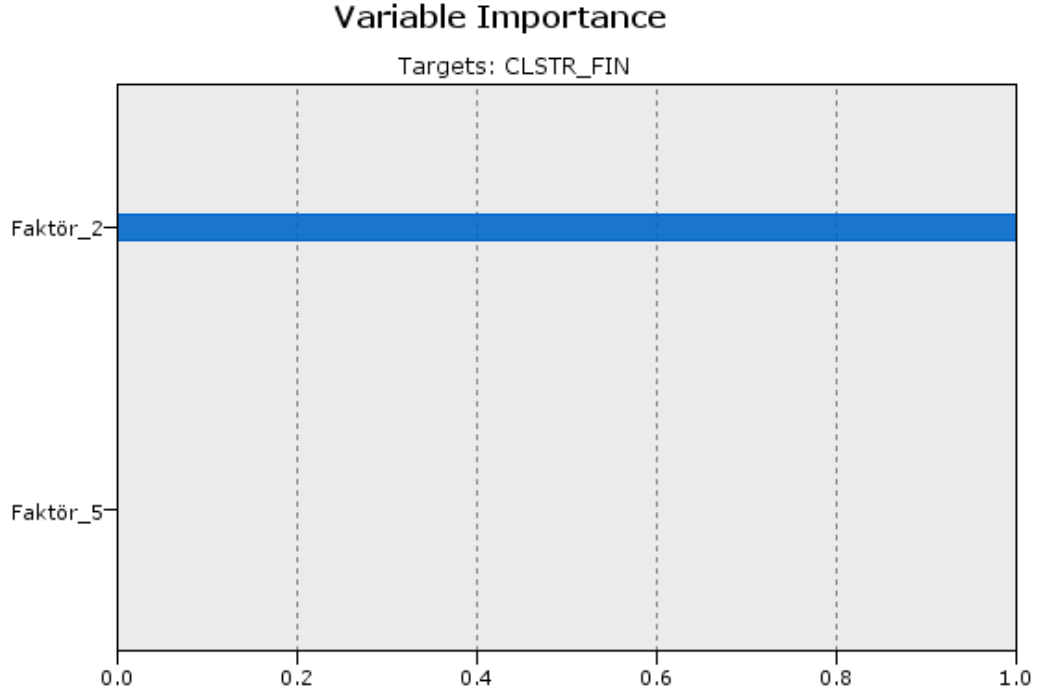


Şekil 38. CHAID modelinin özellikli ayarlarının gösterimi

CHAID karar ağacı modeli SPSS Clementine V.12 varsayılan ayarlamaları kabul edilerek modeli ilgili sonuçlar elde edilmiştir. Bu modelde elde edilen faktörlerin göreceli önem değerleri ve sıraları tablo ve grafik halinde elde edilmiştir. Sadece ücret/ödüllendirme faktörü göreceli önem oranı %100 ile en etken ve tek faktör olarak görülmektedir. Diğer faktörler model tarafından göreceli önemli olarak modele dahil edilmediği görülmektedir. Diğer karar ağaçları algoritmalarına kıyasla bu modelin doğruluk oranı düşük olduğundan, C&R Tree karar ağacı bizim model seçimimizde etkili bir karar verme modeli olarak seçilebilir.

Tablo 45. CHAID modelinde elde edilen faktörlerin göreceli önem oranları

Faktörler	Önem oranı
Faktör_2 (Ücret/Ödüllendirme)	1
Faktör_5 (İş Güvencesi)	0



CHAID modeli kurallar kümesi:

Ücret/Ödüllendirme ≤ 2.29 [Mode: 1] $\Rightarrow 1$ (128; 0.797)

Ücret/Ödüllendirme > 2.29 and Ücret/Ödüllendirme ≤ 3 [Mode: 1] $\Rightarrow 1$ (103; 0.583)

Ücret/Ödüllendirme > 3 [Mode: 2] (219)

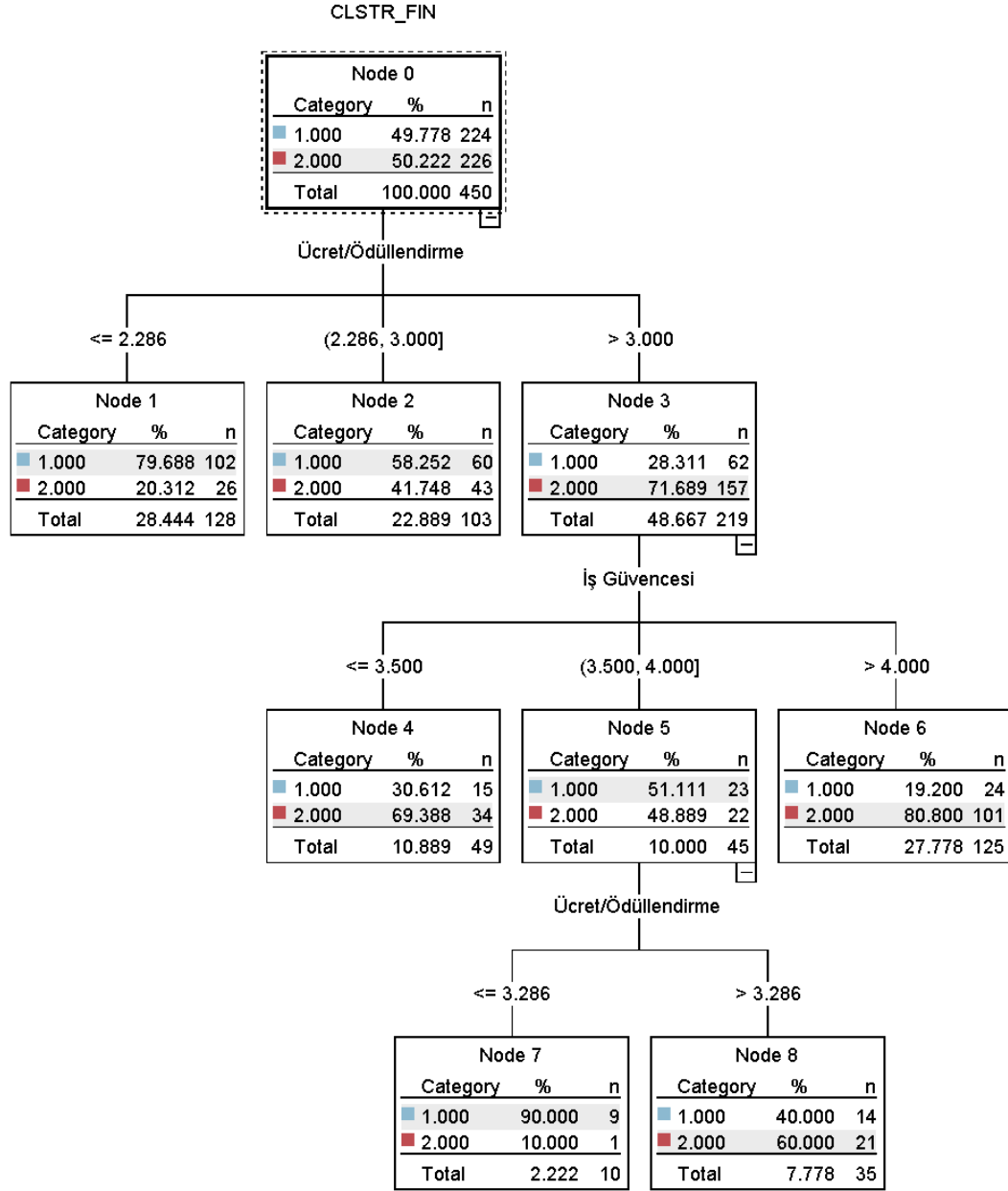
İş Güvencesi ≤ 3.50 [Mode: 2] $\Rightarrow 2$ (49; 0.694)

İş Güvencesi > 3.50 and İş Güvencesi ≤ 4 [Mode: 1] (45)

Ücret/Ödüllendirme ≤ 3.29 [Mode: 1] $\Rightarrow 1$ (10; 0.9)

Ücret/Ödüllendirme > 3.29 [Mode: 2] $\Rightarrow 2$ (35; 0.6)

İş Güvencesi > 4 [Mode: 2] $\Rightarrow 2$ (125; 0.808)



Şekil 39. CHAID modeli karar ağacının gösterimi

Kurallar kümesi ve CHAID karar ağacı bize diğer faktörler ile finansal performans arasındaki ilişkileri ve alt grupları belirtmektedir. Başlangıçta finansal performans başarı isabet oranı %50 iken bazı alt gruplar incelendiği zaman, bu oran %80 oranına kadar yükselmektedir. 4, 6 ve 8 numaralı düğümlerde pozitif isabet oranları

başlangıç isabet oranına kıyasla yüksek görünmektedir. En yüksek isabet oranı, 6 numaralı düğümdür, ücret/ödüllendirme 3'ten büyük ve iş güvencesi faktör değeri 4'ten büyük olduğu zaman finansal performans başarı isabet oranı %80 olmaktadır. Karar ağacında da görüldüğü gibi sadece ücret/ödüllendirme ve iş güvencesi faktörleri ile finansal performans arasında ilişki belirtilmiştir. CHAID algoritmasının “expert” kısmında değişik ayarlamalar yapılarak daha farklı sonuçlar elde edilebilirdi. Örneğin, Benforoni düzeltmesi (use Benforoni Adjustment) kullanılması dikkatli kullanılması gerekmektedir. Yanlı sonular verebileceği için seçilmesinin uygun olmayabileceği savunulmaktadır. Özellikle küçük örneklem büyüklüğüne sahip veri kümeleri için bu seçeneğin kullanılması önerilmektedir. Artan yanlış-pozitif (false-positive) maliyeti pahasına, bu seçeneğin kullanılmaması ile gerçek farklılıkların bulunması için analizimizin gücü artmaktadır.

5.9 Finansal Olmayan performans algısının DVM ve karar ağaçları ile incelenmesi

Finansal olmayan kurum performans kayıtlarına ait kümeleme sonuçları aşağıda belirtilmiştir.

Tablo 46. Finansal olmayan kurum performans algısına ait kayıtlar

Finansal Olmayan Kurum Performans	Sayı
Başarısız (1)	303 (%53,7)
Başarılı (2)	261 (% 46,3)

Finansal olmayan kurum performans başarısız (1) ve başarılı (2) algı sayı ve oranları hem tablo hem de grafik halinde SPSS Clementine V.12 yardımı ile elde edilmiştir. Başarılı algı oranı finansal başarı algısına göre düşük çıkmıştır.

5.9.1 Modellerin performans açısından kıyaslanması

Veri kümesi %80 (450 adet) ile eğitildikten sonra %20 (114 adet) ile test edilerek bilgi çalışanlarının finansal olmayan kurum performans algısı incelenerek uygun DVM (Destek Vektör Makineleri) modeli ile karar ağaçları modelleri belirlenmektedir. İkili sınıflandırma düğümü sayesinde ikili modeller aynı tablo üzerinde belirli performans ölçümleri açısından kıyaslanmaktadır. Veri kümesi eğitildikten sonra performans ölçümleri test veri kümesi sayesinde geçerliliği sağlanmaktadır.

Tablo 47. Test Veri Kümesi için Modellerin ikili sınıflandırma kıyaslama sonuçları

Model	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
SVM RBF	180.0	37	1.916	78.947	5	0.897
SVM Linear	175.0	38	1.916	78.947	5	0.895
SVM Polynomial	165.0	56	1.856	78.07	5	0.874
SVM Sigmoid	150.0	34	1.796	74.561	5	0.833
CHAID	165.0	52	1.838	79.825	4	0.831
C&R Tree	145.0	45	1.736	75.439	5	0.797
QUEST	155.0	43	1.669	78.07	5	0.79
C5	155.0	49	1.597	78.07	2	0.788

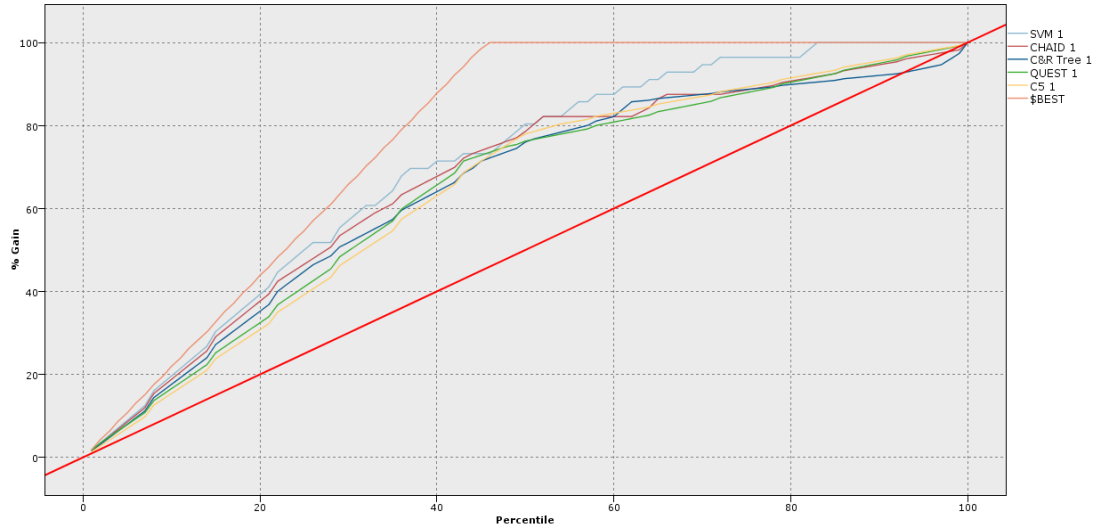
Test veri kümesi ile elde edilen performans ölçümlerine göre, genel doğruluk oranı %79,83 ile en yüksek performans değerine sahip olan karar ağacı modellerinden CHAID'dır. DVM(Support Vectör Machines) modellerinden RBF kernel fonksiyonlu DVM ile lineer kernel fonksiyonu kullanılarak elde edilen DVM her ikisinde %78,95 ile ikinci ve üçüncü en yüksek genel doğruluk oranlarına sahip modellerdir. C5, QUEST karar ağacı modelleri % 78,07 genel doğruluk oranına sahip iken, yine aynı %78,07 doğruluk oranı ile çok terimli kernel fonksiyonlu DVM (Polynomial SVM). En az doğruluk oranına sahip modellere %75,44 oranı ile C&R Tree karar ağacı modeli ile %74,56 sigmoid kernel fonksiyonlu DVM modelidir.

Eğri Altında Kalan Alan (Area Under Curve) kriterine göre modellerin performansları incelendiği zaman, tüme DVM (Destek Vektör Makineleri) modelleri en yüksek alana sahiptirler. Bu kriter modellerin performansları için bir endeks sağlar. Referans doğrusundan bir modelin eğrisi ne kadar uzak ise o model o derece çok daha iyi doğru tahminde bulunur. RBF kernel fonksiyonlu DVM en yüksek alana (%89,7) sahip iken karar ağaçlarından CHAID modeli % 83,1 olmaktadır. C&R Tree ve QUEST modelleri alan performansı açısından bir birlerine sırası ile %79,7 ve %79 oranları ile yakındırlar. Modellerin “Lift” performans kıstasları açısından değerlendirildiği zaman, tüm modelleri üst %30 alanda hepsinin 1,0 değerinin çok yukarısında görülmektedir. En iyi “lift” performansına sahip modelimiz lineer DVM ve karar ağaçlarından ise CHAID modeli görülmektedir.

En yüksek Kar (max. Profit) performans kıstasına göre yine RBF kernel fonksiyonlu DVM en yüksek kâra (180) sahip. Karar ağaçları içerisindeki modellerden ise CHAID modeli (Max Profit: 165) en yüksek kâra sahip model olarak belirtilmiştir.

5.9.1.1 Modellerin Kazançlar (Gains) Grafiği

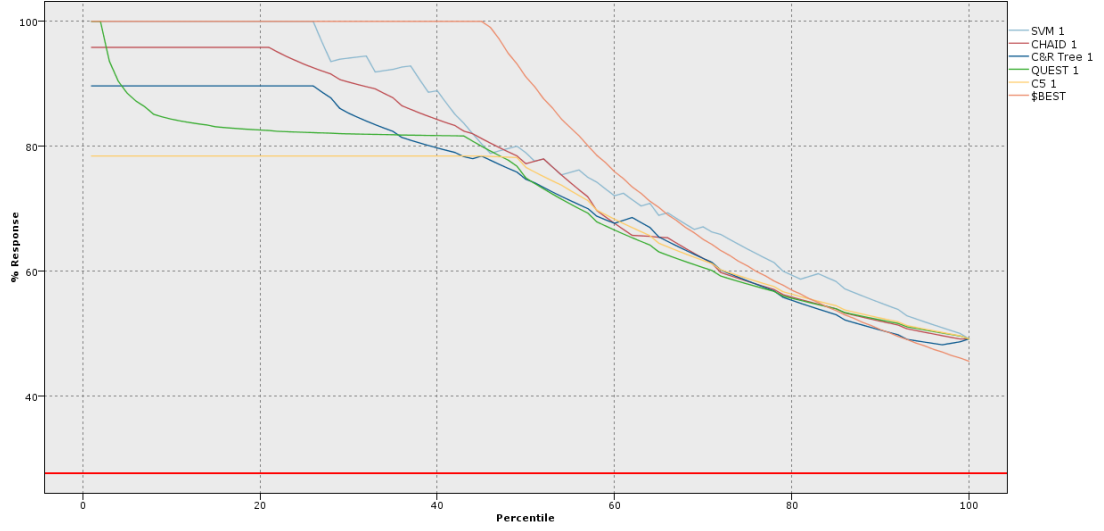
Kazançlar grafiğinde hızlı bir şekilde %100 yükselen DVM modeli ve daha sonra CHAID karar ağacı görülmektedir. Diğer modeller birbirlerine oldukça yakın olarak görülmektedir.



Şekil 42. Modellerin finansal olmayan performanslarının kazançlar (gains) grafiği

5.9.1.2 Modellerin yanıt (response) grafiği ile gösterimi

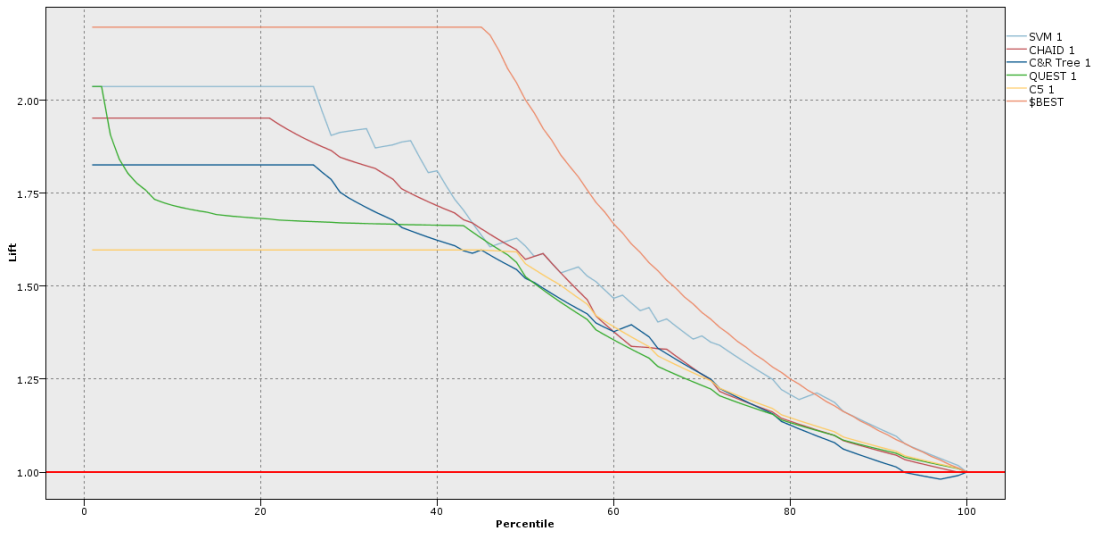
Yanıt (tepki) grafiği aynı şekilde performans modellerin performans açısından kıyaslanmasını yapmaktadır. İyi bir model %100 oranından başlar ve belli süre yatay şekilde devam ederken, grafik sağa ilerlerken hızlı bir şekilde azalarak düşmektedir. Buna göre DVM modelimiz tam %100den başlayarak yatay devam eder %25 yüzdeler diliminde azalarak hızlı bir şekilde düşmeye başlamaktadır. Yine aynı şekilde karar ağaçlarından CHAID modeli de %100 oranından başlayarak %20 yüzdeler diliminden sonra azalarak devam etmektedir.



Şekil 43. Modellerin finansal olmayan performanslarının yanıt (response) grafiği

5.9.1.3 Modellerin “lift” grafiği ile gösterimi

1,0 değerinin çok üzerinden başlayan model iyi performansa sahip model olarak belirtilmiştir. Buna göre sırasıyla DVM ile karar ağaçlarından CHAID ve C&R Tree modelleri 1,0 çok yukarı başlayarak belli bir süre yatay devam ederek hızlı bir şekilde azalarak 1,0 referans değerinde kadar azaldığı görülmektedir.



Şekil 44. Modellerin finansal olmayan performanslarının "lift" grafiği ile gösterimi

5.9.1.4 Modellerin doğru Tahmin etme oranları

Daha öncede ikili sınıflandırma kıyaslama tablosunda elde edilen doğruluk oranları ile aynı sonuçları verecek şekilde aşağıdaki bilgiler elde edilmiştir. Her bir modelin Hem eğitme (training) hem de test (testing) veri alt kümelerinde doğru ve yanlış tahmin etme oranları elde edilmiştir. Bir modelin doğruluğu o modelin veri kümesinin bir kısmını eğittikten sonra diğer veri kümesi ile test edilerek test kümesinde ki doğru tahmin etme oranı ile ölçülür. Aynı bilgiler ışığında RBF kernel fonksiyonlu DVM modeli ile CHAID karar ağacı modelleri en yüksek doğruluk oranlarına sahip modellerdir. Bu bilgiler performans grafikleri yorumlanmaları ile bir paralellik göstermektedir.

Tablo 48.Finansal olmayan kurum performansında kullanılan DVM modellerin doğru tahmin etme oranları

RBF Kernel Fonksiyonlu DVM Modeli				
'Partition'	1_Training		2_Testing	
Correct	334	74.22%	90	78.95%
Wrong	116	25.78%	24	21.05%
Total	450		114	
Lineer Kernel Fonksiyonlu DVM Modeli				
'Partition'	1_Training		2_Testing	
Correct	337	74.89%	90	78.95%
Wrong	113	25.11%	24	21.05%
Total	450		114	
Çok Terimli (Polynomial) Kernel Fonksiyonlu DVM Modeli				
'Partition'	1_Training		2_Testing	
Correct	337	74.89%	89	78.07%
Wrong	113	25.11%	25	21.93%
Total	450		114	
Sigmoid Kernel Fonksiyonlu DVM Modeli				
'Partition'	1_Training		2_Testing	
Correct	314	69.78%	85	74.56%
Wrong	136	30.22%	29	25.44%
Total	450		114	

Tablo 49. Finansal olmayan kurum performansında kullanılan karar ağaçları (CHAID, C&R Tree, QUEST, C5) modellerin doğru tahmin etme oranları

CHAID Karar Ağacı Modeli				
'Partition'	1_Training		2_Testing	
Correct	343	76.22%	91	79.82%
Wrong	107	23.78%	23	20.18%
Total	450		114	
C5.0 Karar Ağacı Modeli				
'Partition'	1_Training		2_Testing	
Correct	336	74.67%	89	78.07%
Wrong	114	25.33%	25	21.93%
Total	450		114	
C&R Tree Karar Ağacı Modeli				
'Partition'	1_Training		2_Testing	
Correct	351	78%	86	75.44%
Wrong	99	22%	28	24.56%
Total	450		114	
QUEST Karar Ağacı Modeli				
'Partition'	1_Training		2_Testing	
Correct	337	74.89%	89	78.07%
Wrong	113	25.11%	25	21.93%
Total	450		114	

5.9.2 Finansal Olmayan Kurum Performans Algısının DVM Modeli İle İncelenmesi

5.9.2.1 Radyal Tabanlı (RBF) Kernel Fonksiyonlu DVM modeli

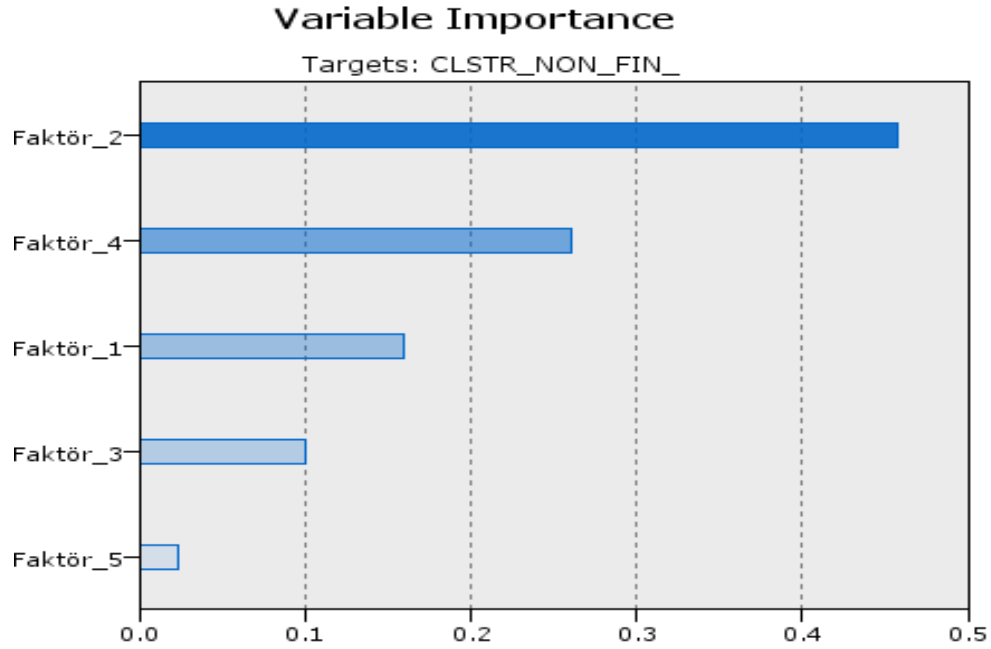
Yapılan ikili sınıflandırma düğümü kullanılarak modeller kıyaslanarak bu modellere ilişkin performans bilgileri daha önce verilmiştir. RBF-DVM modelimiz en yüksek doğruluk düzeyine sahip ve eğri altında kalan alan performans değeri diğer modellere oranla daha yüksek olduğundan RBF-DVM modeli en iyi doğru tahminde bulunacaktır.

Genel doğruluk oranı (Acuracy):	%78,95
Eğri Altında Kalan Alan (Area Under Curve):	0,897
En yüksek Kazanç (Max. Profit):	180

RBF- DVM modeline göre finansal olmayan kurum performansı açısından bilgi çalışanı memnuniyetinin faktörlerinin göreceli önem oranları ve sıraları tablo ve grafik yardımı ile verilmiştir. En yüksek göreceli önem oranı % 45,79 ile ücret/ödüllendirme faktörü iken, %26 oranı ile çalışma şartları ikinci en etken faktördür. Bunun yanı sıra, yönetimin tavrı %16 ile üçüncü en etken faktör iken, iş arkadaşları (%10) ve en az önem oranına sahip faktör işe iş güvencesi (%2,3) sondan son iki faktör olarak model tarafından elde edilmiştir.

Tablo 50. Finansal olmayan kurum performans için RBF-DVM modeli

Faktörler	Göreceli Önem Oranı
Faktör_2 (Ücret/Ödüllendirme)	0.4579
Faktör_4 (Çalışma Şartları)	0.2600
Faktör_1(Yönetimin Tavrı)	0.1591
Faktör_3 (İş Arkadaşları)	0.0999
Faktör_5 (İş Güvencesi)	0.0231



RBF-DVM model özet ayarlamaları:

Fields

Target

CLSTR_NON_FIN_

Inputs

Faktör_1

Faktör_2

Faktör_3

Faktör_4

Faktör_5

Build Settings

Use partitioned data: true

Partition: Partition

_variable_importance.LABEL: true

Calculate raw propensity scores: false

Calculate adjusted propensity scores: false

Mode: Expert

Append all probabilities (valid only for categorical targets): false

Stopping criteria: 1.0E-3

Kernel type: RBF

Regularization parameter (C): 10

Regression precision (epsilon): 0.1

RBF gamma: 0.1
 Gamma: 1.0
 Bias: 0.0
 Degree: 3
 Training Summary
 Algorithm: SVM
 Model type: Classification

RBF-DVM Modeli Tahmin Sonuçları: DVM modeli ile elde edilen tahmin skorları tablo halinde verilmiştir. Elimizde 564 adet kayıt olduğundan sadece ilk 20 kayıt için bilgiler verilmiştir. \$S-CLSTR_NON_FIN bağımlı değişken olan finansal olmayan kurum performans algısının tahmini değerlerini göstermektedir. Görüldüğü gibi birçok kayıt doğru tahmin edilmiştir. Bilindiği gibi, bu modelin doğru tahmin oranı yaklaşık %79 olduğu bilinmektedir.

Tablo 51. RBF-DVM modeli tahmin sonuçları (Finansal olmayan kurum performans)

	Faktör_3	Faktör_4	Faktör_5	Fin_Perf	Non_Fin_Perf	Bin_Fin_Perf	CLSTR_FIN	Bin_Non_Fin_Perf	CLSTR_NON_FIN	Partition	\$S-CLSTR_NON_FIN	\$SP-CLSTR_NON_FIN
1	3.143	4.000	4.833	4.500	3.750	4.000	1.000	2	1.000	2_1_Training	2	0.788
2	2.429	4.333	4.500	3.500	3.000	3.250	0.000	1	0.000	1_2_Testing	1	0.536
3	3.000	4.333	4.000	4.500	4.000	3.750	1.000	2	1.000	2_1_Training	2	0.666
4	0.857	3.000	2.500	3.500	1.500	2.250	0.000	1	0.000	1_1_Training	1	0.846
5	0.857	3.667	3.833	4.500	3.500	3.750	1.000	2	1.000	2_1_Training	1	0.729
6	0.714	3.833	4.167	5.000	3.500	3.250	1.000	2	0.000	1_1_Training	1	0.712
7	0.143	4.000	3.667	2.000	3.000	2.500	0.000	1	0.000	1_1_Training	1	0.590
8	0.714	4.500	4.833	1.000	3.000	3.500	0.000	1	1.000	1_1_Training	2	0.856
9	0.714	4.333	4.500	4.000	3.250	3.000	0.000	1	0.000	1_1_Training	2	0.824
10	2.429	2.833	4.167	3.000	3.000	3.000	0.000	1	0.000	1_1_Training	1	0.520
11	0.571	3.833	2.500	1.000	2.750	2.250	0.000	1	0.000	1_1_Training	1	0.759
12	0.143	4.000	4.667	4.000	3.000	3.000	0.000	1	0.000	1_1_Training	2	0.816
13	3.000	4.667	4.333	2.500	3.500	3.250	1.000	2	0.000	1_1_Training	2	0.714
14	2.429	3.667	3.833	4.500	1.750	1.500	0.000	1	0.000	1_1_Training	1	0.666
15	0.286	4.833	4.833	5.000	4.250	4.250	1.000	2	1.000	2_1_Training	2	0.852
16	0.857	4.500	3.500	3.000	2.750	2.750	0.000	1	0.000	1_1_Training	1	0.733
17	3.000	3.000	2.833	3.500	3.250	4.000	0.000	1	1.000	2_1_Training	1	0.742
18	0.857	2.833	3.333	4.500	3.000	2.250	0.000	1	0.000	1_1_Training	1	0.737
19	2.286	1.500	2.833	5.000	4.000	4.000	1.000	2	1.000	2_1_Training	1	0.811
20	0.286	4.167	1.833	5.000	3.000	2.250	0.000	1	0.000	1_1_Training	1	0.874

5.9.3 CHAID Karar Ağacı Modeli İle Finansal Olmayan Kurum Performansın İnceleme

Bilgi çalışanların finansal olmayan kurum performans algısını inçlerken kullanılan modeller arasında en yüksek genel doğruluk oranını veren model CHAID karar ağacı modelidir. CHAID karar ağacı modelinin doğru tahmin etme oranı yaklaşık olarak %80dir. İkili sınıflandırma düğümü yolu ile yapılan kıyaslama tablosunda aşağıdaki değerler elde edilmiştir. Genel doğru tahmin etme oranı diğer modellere göre en yüksek ilen, eğri altında kalan alan ne kadar büyük ise o kadar iyi bir model olduğu söylendiğinden, CHAID modeli DVM modelinden sonra eğri altında kalan alan performans ölçümü açısından iyi bir model olduğu söylenebilir.

Genel Doğruluk Oranı (Accuracy):	% 79,83
Eğri Altında Kalan Alan(Area under Curve) :	0,831
En yüksek kar (maksimum Profit):	165

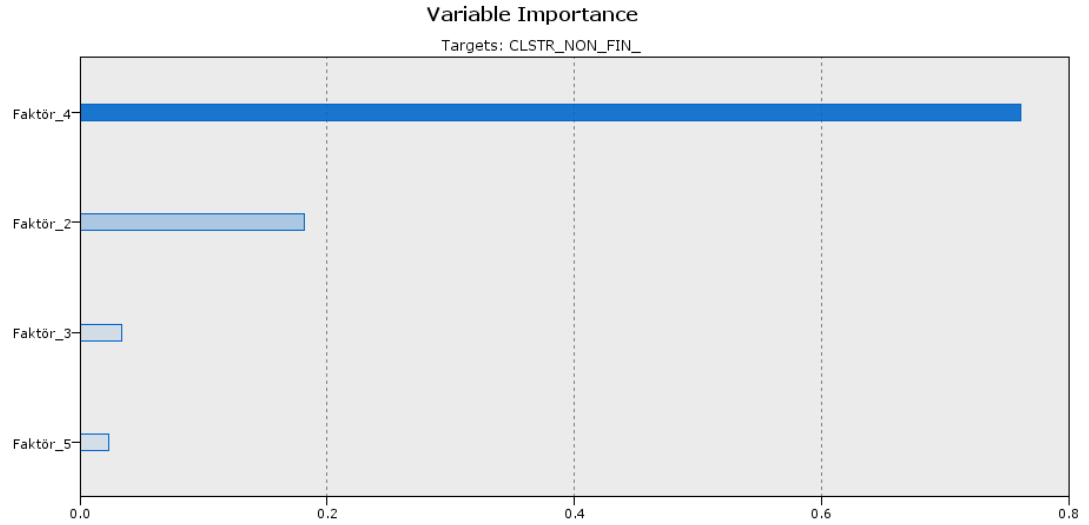
CHAID Karar Ağacı Modeli			
'Partition'	1_Training		2_Testing
Correct	343	76.22%	91 79.82%
Wrong	107	23.78%	23 20.18%
Total	450		114

Faktörlerin Göreceli Önem Oran ve Sıraları: Finansal olmayan kurum performans algısını etki eden faktörlerin CHAID karar ağacı modeli yardımı ile göreceli önem oranları ve önem sıraları hem tablo hem de grafik halinde SPSS Clementine V.12 yardımı ile elde edilmiştir. Çalışma şartları çok yüksek bir göreceli önem oranı olan yaklaşık %76 ilen en etken faktör olduğu model tarafından elde edilmiştir. Daha sonra, ücret/ödüllendirme faktörü %18 ile ikinci en etken göreceli faktör iken, iş arkadaşları ve iş güvencesi faktörleri diğer faktörlere oranla oldukça düşük göreceli önem oranları olan %3 ve %2 ile son iki en etken faktörler olarak model tarafından belirtilmiştir. DVM modelinde üçüncü en etken göreceli faktör olan yönetimin tavrı CHAID karar ağacı

modelinde göreceli etken faktörler arasında bulunmamaktadır. Özellikle çalışma şartları faktörü ve ücret/ödüllendirme faktörlerine yüksek göreceli oran yüklenmesi yapılmıştır.

Tablo 52. CHAID karar ağacı modelinde faktörlerin göreceli önem oranları

Faktörler	Göreceli Önem Oranı
Faktör_4 (Çalışma Şartları)	0.7614
Faktör_2 (Ücret/Ödüllendirme)	0.1815
Faktör_3 (İş Arkadaşları)	0.0340
Faktör_5 (İş Güvencesi)	0.0231



Karar Ağacı ve Kurallar Kümesi:

Çalışma Şartları ≤ 3.333 [Mode: 1] (174)
 İş Güvencesi ≤ 4 [Mode: 1] $\Rightarrow 1$ (118; 0.864)
 İş Güvencesi > 4 [Mode: 1] $\Rightarrow 1$ (56; 0.696)
Çalışma Şartları > 3.333 and Çalışma Şartları ≤ 4.333 [Mode: 2] (176)
 Ücret/Ödüllendirme ≤ 3.286 [Mode: 1] (91)
 İş Güvencesi ≤ 3.500 [Mode: 2] (36)
 İş Arkadaşları ≤ 3.333 [Mode: 1] $\Rightarrow 1$ (7; 1.0)
 İş Arkadaşları > 3.333 [Mode: 2] $\Rightarrow 2$ (29; 0.655)
 İş Güvencesi > 3.500 and İş Güvencesi ≤ 4 [Mode: 1] $\Rightarrow 1$ (19; 0.789)
 İş Güvencesi > 4 and İş Güvencesi ≤ 4.500 [Mode: 2] $\Rightarrow 2$ (13; 0.615)
 İş Güvencesi > 4.500 [Mode: 1] $\Rightarrow 1$ (23; 0.87)
 Ücret/Ödüllendirme > 3.286 [Mode: 2] $\Rightarrow 2$ (85; 0.682)
 Çalışma Şartları > 4.333 and Çalışma Şartları ≤ 4.667 [Mode: 2] $\Rightarrow 2$ (51; 0.706)
 Çalışma Şartları > 4.667 [Mode: 2] $\Rightarrow 2$ (49; 0.898)

Rules for 1 - contains 5 rule(s)

Rule 1 for 1 (118; 0.864)

if Çalışma Şartları $\leq$ 3.333
and İş Güvencesi $\leq$ 4
then 1

Rule 2 for 1 (56; 0.696)

if Çalışma Şartları $\leq$ 3.333
and İş Güvencesi $>$ 4
then 1

Rule 3 for 1 (7; 1.0)

if Çalışma Şartları $>$ 3.333
and Çalışma Şartları $\leq$ 4.333
and Ücret/Ödüllendirme $\leq$ 3.286
and İş Güvencesi $\leq$ 3.500
and İş Arkadaşları $\leq$ 3.333
then 1

Rule 4 for 1 (19; 0.789)

if Çalışma Şartları $>$ 3.333
and Çalışma Şartları $\leq$ 4.333
and Ücret/Ödüllendirme $\leq$ 3.286
and İş Güvencesi $>$ 3.500
and İş Güvencesi $\leq$ 4
then 1

Rule 5 for 1 (23; 0.87)

if Çalışma Şartları $>$ 3.333
and Çalışma Şartları $\leq$ 4.333
and Ücret/Ödüllendirme $\leq$ 3.286
and İş Güvencesi $>$ 4.500
then 1

Rules for 2 - contains 5 rule(s)

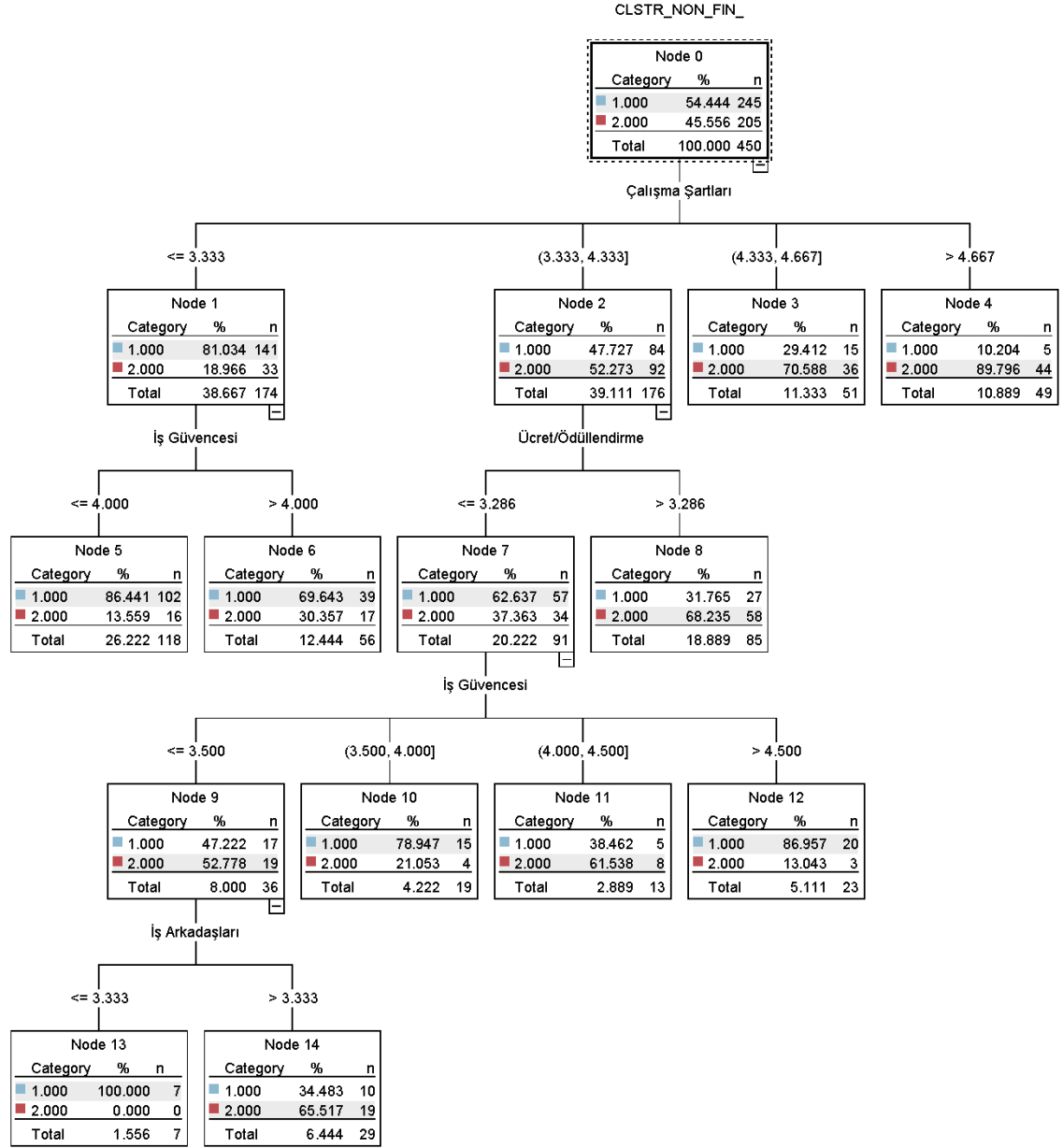
Rule 1 for 2 (29; 0.655)

if Çalışma Şartları $>$ 3.333
and Çalışma Şartları $\leq$ 4.333
and Ücret/Ödüllendirme $\leq$ 3.286
and İş Güvencesi $\leq$ 3.500
and İş Arkadaşları $>$ 3.333
then 2

Rule 2 for 2 (13; 0.615)

if Çalışma Şartları $>$ 3.333
and Çalışma Şartları $\leq$ 4.333
and Ücret/Ödüllendirme $\leq$ 3.286
and İş Güvencesi $>$ 4
and İş Güvencesi $\leq$ 4.500

then 2
Rule 3 for 2 (85; 0.682)
if Çalışma Şartları > 3.333
and Çalışma Şartları <= 4.333
and Ücret/Ödüllendirme > 3.286
then 2
Rule 4 for 2 (51; 0.706)
if Çalışma Şartları > 4.333
and Çalışma Şartları <= 4.667
then 2
Rule 5 for 2 (49; 0.898)
if Çalışma Şartları > 4.667
then 2



ayrılmıştır. Ağacın son bölünmesi iş arkadaşları faktörü üzerinden gerçekleşmiştir. Başlangıçta finansal olmayan başarılı performans oranı yaklaşık % 46 iken bu oran alt gruplarda daha da yükselmektedir. Çalışma şartları 4,67'den büyük olduğu zaman isabet etme oranı %90 gibi en yüksek bir oran elde edilmektedir. Çalışma şartları faktörünün diğer alt aralıkların da, örneğin (3,33;4,33] finansal olmayan kurum performans başarı isabet etme oranı %52 oranına yükselmektedir. Benzer şekilde çalışma şartları (4,33;4,67] aralığında isabet oranı başlangıç başarı isabet oranı %46'dan %71 oranına yükselmektedir. Bu değerler kurallar kümesinde de görülmektedir. Çalışma şartları faktör değerinin 3,33'den küçük ve eşit olduğu alt grupta başarı isabet oranı oldukça düşmektedir (%19). Sonraki en yüksek başarı isabet oranı yüksek olan 8 numaralı düğüm takip edildiği zaman, çalışma şartları (3,33;4,33] aralığında iken ve ücret/ödüllendirme 3,29 faktör değerinden büyük iken isabet etme oranı %68 olmaktadır. Sonraki en yüksek finansal olmayan başarı isabet oranları %66 ile 14 numaralı düğümde ve %62 oran ile 11 numaralı düğümlerde görülmektedir.

CHAID Karar Ağacı ayarlamaları Özet bilgileri:

Fields

Target

CLSTR_NON_FIN_

Inputs

Faktör_2

Faktör_3

Faktör_4

Faktör_5

Build Settings

Use partitioned data: true

Partition: Partition

Calculate variable importance: true

Calculate raw propensity scores: false

Calculate adjusted propensity scores: false

Use frequency: false

Use weight: false

Levels below root: 5

Mode: Simple

Use misclassification costs: true

Training Summary

Algorithm: CHAID

Model type: Classification

CHAID karar Ağacı Modelinin Kazançlar (Gains) açısından performansının İncelenmesi: Kazançlar (gains) bilgileri CHAID karar ağacında elde edilen düğümlere ilişkin istatistiksel bilgiler vermektedir. İlgili düğümdeki oran ya da ortalama genel oran ya da ortalamadan ne kadar farklı olduğunu gösteren bir ölçüm sağlar. Karar ağacında ve kurallar kümesinden bilindiği ve beklendiği gibi 4, 11, 3, 8 ve 14 numaralı düğümlerin endeks değerleri %100 oranından büyük çıkmıştır. Endeks değerinin %100'den büyük olması demek bu düğümlerden seçilen kişilerin (kayıtların) finansal olmayan kurum performansı başarısı açısından genel örneklem veri kümesinden rastgele seçilen kişilere nazaran daha başarılı olacağını belirtir. Örneğin, en yüksek endeks değerine sahip 11 numaralı düğümün endeks değeri %203. 11 numaralı düğümden seçilen kişilerin genel örneklemden seçilen kişilere kıyasla yaklaşık iki kat olasılıkla daha finansal olmayan kurum performans algısı açısından başarılı olduğunu belirtir. Aynı şekilde 4 numaralı düğümün endeks değeri 195 olmasından dolayı, bu düğümden seçilen kayıtlar (kişiler) yaklaşık olarak 1,95 kat olasılıkla genel örneklemden seçilen bilgi işçilerine (kayıt) kıyasla daha fazla başarılı olacağını belirtmektedir. Aynı mantık ile 3, 8, ve 14 numaralı düğümler bu şekilde yorumlanabilir. Zaten, karar ağacı ve kurallar kümesi incelenirken bu düğümlerin kazançlar (gains) bilgileri ile paralellik gösterdiği açıkça görülmektedir.

Tablo 53. CHAID karar Ağacı Modelinin Kazançlar (Gains) açısından performansının İncelenmesi

Interactive Tree of CLSTR_NON_FIN_													
File Edit View Tree Generate													
Decile													
Gains													
Target variable: CLSTR_NON_FIN_ Target category: 2.0													
Training Sample							Testing Sample						
Nodes	Node: n	Node (%)	Gain: n	Gain (%)	Response (%)	Index (%)	Nodes	Node: n	Node (%)	Gain: n	Gain (%)	Response (%)	Index (%)
4	49.00	10.89	44.00	21.46	89.80	197.11	4	24.00	21.05	23.00	41.07	95.83	195.09
3	51.00	11.33	36.00	17.56	70.59	154.95	3	9.00	7.89	7.00	12.50	77.78	158.33
8	85.00	18.89	58.00	28.29	68.24	149.78	8	16.00	14.04	11.00	19.64	68.75	139.96
14	29.00	6.44	19.00	9.27	65.52	143.82	14	3.00	2.63	2.00	3.57	66.67	135.71
11	13.00	2.89	8.00	3.90	61.54	135.08	11	1.00	0.88	1.00	1.79	100.00	203.57
6	56.00	12.44	17.00	8.29	30.36	66.64	6	15.00	13.16	1.00	1.79	6.67	13.57
10	19.00	4.22	4.00	1.95	21.05	46.21	10	5.00	4.39	3.00	5.36	60.00	122.14
5	118.00	26.22	16.00	7.80	13.56	29.76	5	33.00	28.95	6.00	10.71	18.18	37.01
12	23.00	5.11	3.00	1.46	13.04	28.63	12	5.00	4.39	0.00	0.00	0.00	0.00
13	7.00	1.56	0.00	0.00	0.00	0.00	13	3.00	2.63	2.00	3.57	66.67	135.71

Finansal olmayan kurum performans bağımlı değişkeni “CLSTR_NON_FIN” olarak değişken ismi ile adlandırılmaktadır. Model tarafından tahmin edilen finansal olmayan kurum performans değerleri “\$R-CLSTR_NON_FIN” değişkeni altında belirtilmiştir. Daha öncede bahsedildiği gibi bu değerler gözlenen değerler ile kıyaslandığı zaman %80 oranı doğru olarak tahmin edildiği görülmektedir. Kullanılan DVM ve Karar ağaçları modelleri arasında en yüksek doğruluk oranı sağlayan model CHAID karar ağacı modelidir.

Tablo 54. CHAID karar ağacı modelinin tahmin sonuçları

Table (78 fields, 564 records)																
	13	BOSLUK	Faktör_1	Faktör_2	Faktör_3	Faktör_4	Faktör_5	Fin_Perf	Non_Fin...	Bin_Fin...	CLSTR...	Bin_Non_Fin...	CLSTR_NON_FIN...	Partition	\$R-CLSTR_NON_FIN	\$RC-CLSTR_NON_FIN
1			4.17	4.14	4.00	4.83	4.50	3.75	4.00	1.00	2	1.00		2 1_Training	2	0.882
2			3.75	2.43	4.33	4.50	3.50	3.00	3.25	0.00	1	0.00		1 2_Testing	2	0.898
3			3.75	4.00	4.33	4.00	4.50	4.00	3.75	1.00	2	1.00		2 1_Training	2	0.878
4			2.92	1.86	3.00	2.50	3.50	1.50	2.25	0.00	1	0.00		1 1_Training	1	0.958
5			3.75	1.86	3.67	3.83	4.50	3.50	3.75	1.00	2	1.00		2 1_Training	2	0.800
6			3.75	1.71	3.83	4.17	5.00	3.50	3.25	1.00	2	0.00		1 1_Training	1	0.840
7			3.25	3.14	4.00	3.67	2.00	3.00	2.50	0.00	1	0.00		1 1_Training	2	0.845
8			4.42	4.71	4.50	4.83	1.00	3.00	3.50	0.00	1	1.00		1 1_Training	2	0.882
9			4.00	4.71	4.33	4.50	4.00	3.25	3.00	0.00	1	0.00		1 1_Training	2	0.898
10			3.83	3.43	2.83	4.17	3.00	3.00	3.00	0.00	1	0.00		1 1_Training	2	0.878
11			3.17	2.57	3.83	2.50	1.00	2.75	2.25	0.00	1	0.00		1 1_Training	1	0.858
12			4.83	4.14	4.00	4.67	4.00	3.00	3.00	0.00	1	0.00		1 1_Training	2	0.898
13			3.75	4.00	4.67	4.33	2.50	3.50	3.25	1.00	2	0.00		1 1_Training	2	0.898
14			3.58	2.43	3.67	3.83	4.50	1.75	1.50	0.00	1	0.00		1 1_Training	2	0.800
15			4.25	4.29	4.83	4.83	5.00	4.25	4.25	1.00	2	1.00		2 1_Training	2	0.882
16			3.42	1.86	4.50	3.50	3.00	2.75	2.75	0.00	1	0.00		1 1_Training	2	0.845
17			2.92	3.00	3.00	2.83	3.50	3.25	4.00	0.00	1	1.00		2 1_Training	1	0.958
18			2.42	2.86	2.83	3.33	4.50	3.00	2.25	0.00	1	0.00		1 1_Training	2	0.800
19			2.17	2.29	1.50	2.83	5.00	4.00	4.00	1.00	2	1.00		2 1_Training	1	0.890
20			3.25	1.29	4.17	1.83	5.00	3.00	2.25	0.00	1	0.00		1 1_Training	1	0.890

5.9.4 C&R Tree Karar Ağacı Modeli İle Finansal Olmayan Kurum Performansın İnceleme

Genel doğruluk oranı C5 (%78) ve QUEST (%78) karar ağaçlarına göre biraz düşük olmasına karşın performans ölçüm grafiklerinde DVM modeli ve CHAID karar ağacı modelinden sonra en iyi performans ölçümlerine sahip bir modeldir. Performans ölçümlerinde olan eğri altında kalan alan CHAID karar ağacından sonra en yüksek C&R Tree karar ağacına sahiptir. Dolayısıyla CHAID modelinden sonra bu modelin incelenmesi uygundur.

Genel Doğruluk oranı (Accuracy) :	% 75,44
Eğri altında kalan alan (Area Under Curve) :	0,797
Lift:	1,736

C&R Tree Karar Ağacı Modeli			
'Partition'	1_Training		2_Testing
Correct	351	78%	86 75.44%
Wrong	99	22%	28 24.56%
Total	450		114

C&R Tree Karar Ağacı Modelinin Özellikli Ayarlamalar Özet Bilgiler: Bu modele ait SPSS Clementine üzerinde yapılan ayarlamalar hakkında özet bilgiler aşağıdaki gibidir.

Özet Bilgiler:

Fields

Target

CLSTR_NON_FIN_

Inputs

Faktör_1

Faktör_2
Faktör_3
Faktör_4
Faktör_5

Build Settings

Use partitioned data: true
Partition: Partition
Calculate variable importance: true
Calculate raw propensity scores: false
Calculate adjusted propensity scores: false
Use frequency: false
Use weight: false
Levels below root: 5
Mode: Simple
Use misclassification costs: true

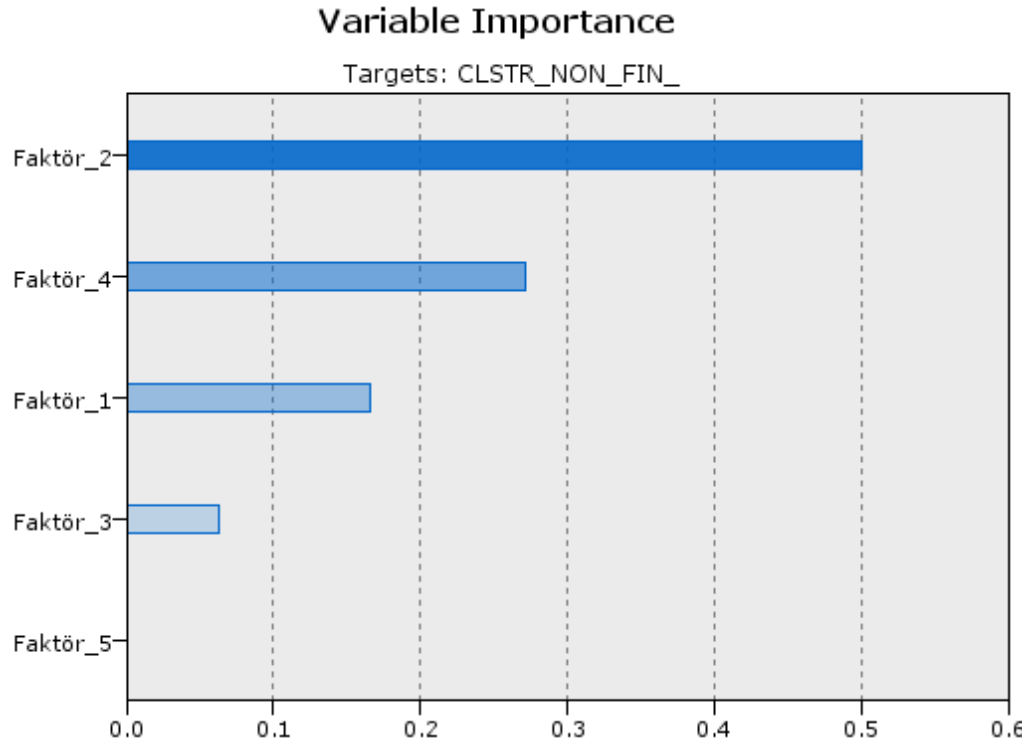
Training Summary

Algorithm: C&R Tree
Model type: Classification

Faktörleri Göreceli Önem Oranları ve Sıraları: Faktörlerin göreceli önem oranları aşağıdaki gibi elde edilmiştir. En yüksek göreceli önem oranı %50 ile ücret/ödüllendirme faktörü en etken faktör olarak model tarafından elde edilmiştir. DVM modelinde elde edilen sonuçlarda da bu faktör en etken faktör olarak belirlenmişti. Yaklaşık %27 göreceli önem oranına sahip çalışma şartları faktörü ikinci en etken faktör olduğu görülmektedir. İlk iki faktör olan ücret/ödüllendirme ve çalışma şartları DVM, CHAID ve C&R Tree modellerinde ilki iki en etken faktörler olduğu görülmektedir. CHAID modelinde etken faktörler sıralamasına göremeyen yönetimin tavrı faktörü, bu modelde yaklaşık %17 göreceli önem oranı ile üçüncü en etken faktör iken %6 oran ile iş arkadaşları en az etken faktör olarak model tarafından belirtilmiştir. İş güvencesi ne kadar göreceli önem faktörler arasında görülmesine rağmen %0 bir göreceli önem değerine sahiptir, yani diğer faktörlere göre C&R Tree modeline göre en etkisiz faktördür.

Tablo 55. C&R Tree modeli ile elde edilen faktörlerin göreceli önem oranları

Faktörler	Göreceli Önem Oranı
Faktör_2 (Ücret/Ödüllendirme)	0.5003
Faktör_4 (Çalışma Şartları)	0.2717
Faktör_1(Yönetimin Tavrı)	0.1657
Faktör_3 (İş Arkadaşları)	0.0623
Faktör_5 (İş Güvencesi)	0



Karar ağacının ve Kurallar Kümesinin Yorumlanması:

Ücret/Ödüllendirme ≤ 3.214 [Mode: 1]
 Çalışma Şartları ≤ 4.250 [Mode: 1]
 Çalışma Şartları ≤ 3.250 [Mode: 1] $\Rightarrow 1$
 Çalışma Şartları > 3.250 [Mode: 1]
 İş Arkadaşları ≤ 3.083 [Mode: 1] $\Rightarrow 1$
 İş Arkadaşları > 3.083 [Mode: 1]
 İş Güvencesi ≤ 3.750 [Mode: 2] $\Rightarrow 2$
 İş Güvencesi > 3.750 [Mode: 1] $\Rightarrow 1$
 Çalışma Şartları > 4.250 [Mode: 2]
 Ücret/Ödüllendirme ≤ 2.357 [Mode: 1] $\Rightarrow 1$
 Ücret/Ödüllendirme > 2.357 [Mode: 2] $\Rightarrow 2$
Ücret/Ödüllendirme > 3.214 [Mode: 2]
 Yönetimin Tavrı ≤ 4.208 [Mode: 2]
 Çalışma Şartları ≤ 3.250 [Mode: 1]
 Yönetimin Tavrı ≤ 3.542 [Mode: 1] $\Rightarrow 1$
 Yönetimin Tavrı > 3.542 [Mode: 2] $\Rightarrow 2$
 Çalışma Şartları > 3.250 [Mode: 2]
 Yönetimin Tavrı ≤ 4.125 [Mode: 2] $\Rightarrow 2$
 Yönetimin Tavrı > 4.125 [Mode: 1]
 Ücret/Ödüllendirme ≤ 4.071 [Mode: 1] $\Rightarrow 1$
 Ücret/Ödüllendirme > 4.071 [Mode: 2] $\Rightarrow 2$
 Yönetimin Tavrı > 4.208 [Mode: 2] $\Rightarrow 2$

Rules for 1 - contains 6 rule(s)

Rule 1 for 1 (143; 0.839)
 if Ücret/Ödüllendirme ≤ 3.214
 and Çalışma Şartları ≤ 4.250
 and Çalışma Şartları ≤ 3.250
 then 1

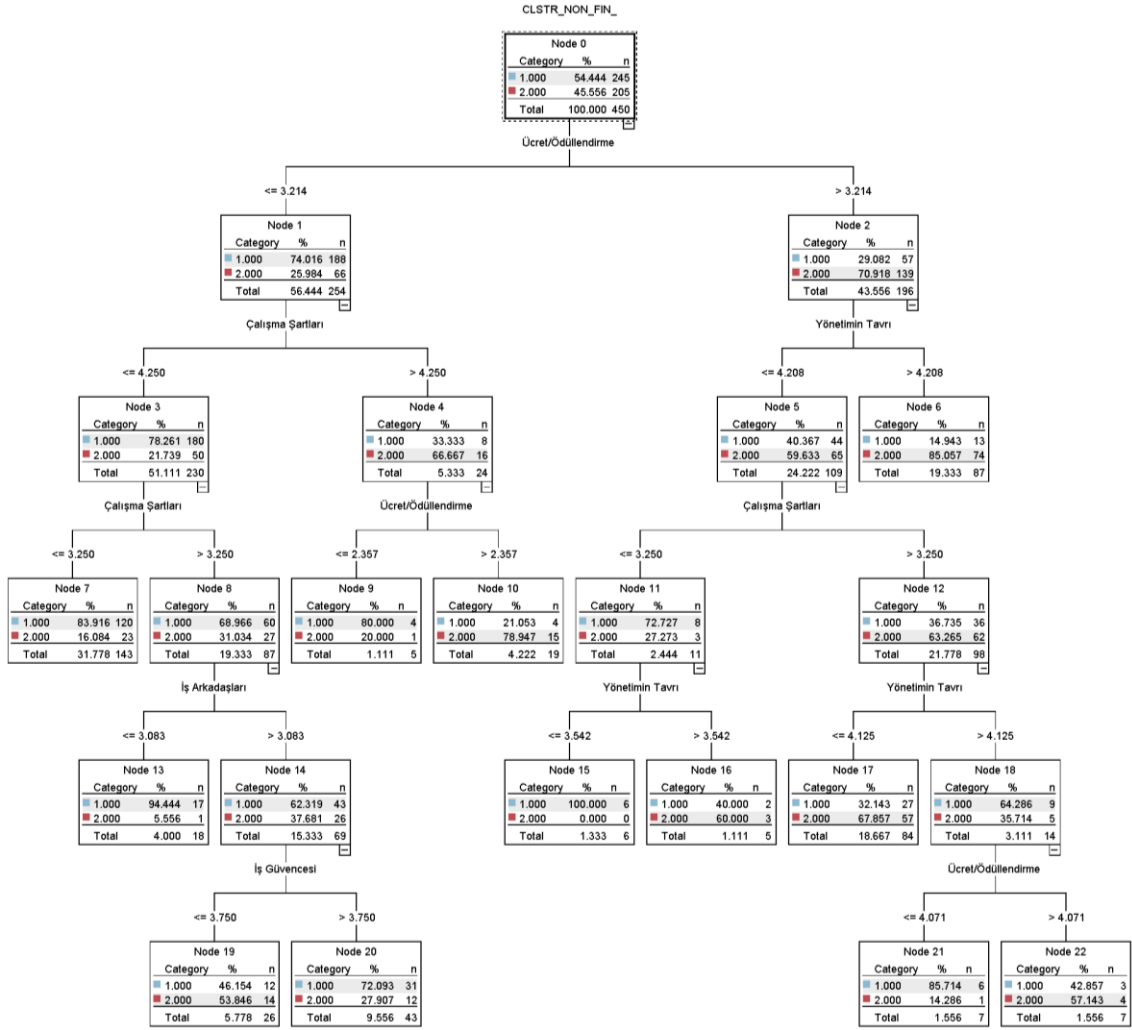
Rule 2 for 1 (18; 0.944)
 if Ücret/Ödüllendirme ≤ 3.214
 and Çalışma Şartları ≤ 4.250
 and Çalışma Şartları > 3.250
 and İş Arkadaşları ≤ 3.083
 then 1

Rule 3 for 1 (43; 0.721)
 if Ücret/Ödüllendirme ≤ 3.214
 and Çalışma Şartları ≤ 4.250
 and Çalışma Şartları > 3.250

and İş Arkadaşları > 3.083
 and İş Güvencesi > 3.750
 then 1
 Rule 4 for 1 (5; 0.8)
 if Ücret/Ödüllendirme <= 3.214
 and Çalışma Şartları > 4.250
 and Ücret/Ödüllendirme <= 2.357
 then 1
 Rule 5 for 1 (6; 1.0)
 if Ücret/Ödüllendirme > 3.214
 and Yönetimin Tavrı <= 4.208
 and Çalışma Şartları <= 3.250
 and Yönetimin Tavrı <= 3.542
 then 1
 Rule 6 for 1 (7; 0.857)
 if Ücret/Ödüllendirme > 3.214
 and Yönetimin Tavrı <= 4.208
 and Çalışma Şartları > 3.250
 and Yönetimin Tavrı > 4.125
 and Ücret/Ödüllendirme <= 4.071
 then 1

 Rules for 2 - contains 6 rule(s)
 Rule 1 for 2 (26; 0.538)
 if Ücret/Ödüllendirme <= 3.214
 and Çalışma Şartları <= 4.250
 and Çalışma Şartları > 3.250
 and İş Arkadaşları > 3.083
 and İş Güvencesi <= 3.750
 then 2
 Rule 2 for 2 (19; 0.789)
 if Ücret/Ödüllendirme <= 3.214
 and Çalışma Şartları > 4.250
 and Ücret/Ödüllendirme > 2.357
 then 2
 Rule 3 for 2 (5; 0.6)
 if Ücret/Ödüllendirme > 3.214
 and Yönetimin Tavrı <= 4.208
 and Çalışma Şartları <= 3.250
 and Yönetimin Tavrı > 3.542
 then 2
 Rule 4 for 2 (84; 0.679)
 if Ücret/Ödüllendirme > 3.214

and Yönetimin Tavrı <= 4.208
and Çalışma Şartları > 3.250
and Yönetimin Tavrı <= 4.125
then 2
Rule 5 for 2 (7; 0.571)
if Ücret/Ödüllendirme > 3.214
and Yönetimin Tavrı <= 4.208
and Çalışma Şartları > 3.250
and Yönetimin Tavrı > 4.125
and Ücret/Ödüllendirme > 4.071
then 2
Rule 6 for 2 (87; 0.851)
if Ücret/Ödüllendirme > 3.214
and Yönetimin Tavrı > 4.208
then 2



Şekil 46. Finansal olmayan kurum performans algısı için C&R Tree karar ağacı modeli

Karar ağacı ve kurallar kümesi incelendiği zaman bireysel düğümlere ait isabet oranları belirtilmiştir. Finansal olmayan kurum performans başarı bağımlı değişkeni başlangıçta isabet oranı yaklaşık %46 olmasına rağmen, 6 adet alt gruplar da bur oran %54 ile %85 oranları arasında değişmektedir. C&R Tree karar ağacı 6 seviye ve 12 adet uç (terminal) düğümlerden oluşmaktadır. 6,10,16,17,19 ve 22 numaralı düğümler başlangıç isabet oranı olan %46'dan yüksek isabet oranlarına sahiptirler. En yüksek isabet oranı 6 numaralı düğüme ait iken en az finansal olmayan kurum performans başarı

algısı en düşük isabet oranı 19 numaralı düğüme aittir. İlk bölünme ücret/ödüllendirme faktörü üzerinden yapılmaktadır. Ücret/ödüllendirme 3,21’den büyük ve yönetimin tavrı 4,21’den büyük olduğu zaman başarı isabet oranı %85 yükselmektedir. Sonraki en yüksek isabet oranı olan alt grup, ücret/ödüllendirme 3,21’den küçük ve eşit, çalışma şartları 4,25’den büyük ve ücret/ödüllendirme 2,36’dan büyük olduğu zaman isabet oranı %79’a yükselmektedir.

Finansal olmayan kurum performansı bağımlı değişken “CLSTR_NON_FIN” olarak belirlenmiştir ayrıca model tarafından tahmin edilen değişken ise \$R-CLSTR_NON_FIn olarak belirtilmiştir. Genel doğruluk oranı % 75 olduğu bilindiğinden, model birçok kaydı doğru şekilde tahmin ederken bazı kayıtların doğru olarak tahmin edilmediği görülmektedir.

Tablo 56. C&R Tree karar ağacı modeli tahmin sonuçları

Table (77 fields, 564 records)												
	Faktör_2	Faktör_3	Faktör_4	Faktör_5	Fin_Perf	Non_Fin_Perf	Bin_Fin_Perf	CLSTR_Fin	Bin_Non_Fin_Perf	CLSTR_NON_Fin	Partition	\$R-CLSTR_NON_Fin
1	4.167	4.143	4.000	4.833	4.500	3.750	4.000	1.000	2	1.000	2_1_Training	2
2	3.750	2.429	4.333	4.500	3.500	3.000	3.250	0.000	1	0.000	1_2_Testing	2
3	3.750	4.000	4.333	4.000	4.500	4.000	3.750	1.000	2	1.000	2_1_Training	2
4	2.917	1.857	3.000	2.500	3.500	1.500	2.250	0.000	1	0.000	1_1_Training	1
5	3.750	1.857	3.667	3.833	4.500	3.500	3.750	1.000	2	1.000	2_1_Training	1
6	3.750	1.714	3.833	4.167	5.000	3.500	3.250	1.000	2	0.000	1_1_Training	1
7	3.250	3.143	4.000	3.667	2.000	3.000	2.500	0.000	1	0.000	1_1_Training	2
8	4.417	4.714	4.500	4.833	1.000	3.000	3.500	0.000	1	1.000	1_1_Training	2
9	4.000	4.714	4.333	4.500	4.000	3.250	3.000	0.000	1	0.000	1_1_Training	2
10	3.833	3.429	2.833	4.167	3.000	3.000	3.000	0.000	1	0.000	1_1_Training	2
11	3.167	2.571	3.833	2.500	1.000	2.750	2.250	0.000	1	0.000	1_1_Training	1
12	4.833	4.143	4.000	4.667	4.000	3.000	3.000	0.000	1	0.000	1_1_Training	2
13	3.750	4.000	4.667	4.333	2.500	3.500	3.250	1.000	2	0.000	1_1_Training	2
14	3.583	2.429	3.667	3.833	4.500	1.750	1.500	0.000	1	0.000	1_1_Training	1
15	4.250	4.286	4.833	4.833	5.000	4.250	4.250	1.000	2	1.000	2_1_Training	2
16	3.417	1.857	4.500	3.500	3.000	2.750	2.750	0.000	1	0.000	1_1_Training	2
17	2.917	3.000	3.000	2.833	3.500	3.250	4.000	0.000	1	1.000	2_1_Training	1
18	2.417	2.857	2.833	3.333	4.500	3.000	2.250	0.000	1	0.000	1_1_Training	1
19	2.167	2.286	1.500	2.833	5.000	4.000	4.000	1.000	2	1.000	2_1_Training	1
20	3.250	1.286	4.167	1.833	5.000	3.000	2.250	0.000	1	0.000	1_1_Training	1

5.9.5 QUEST Karar Ağacı Modeli İle Finansal Olmayan Kurum Performansın İnceleme

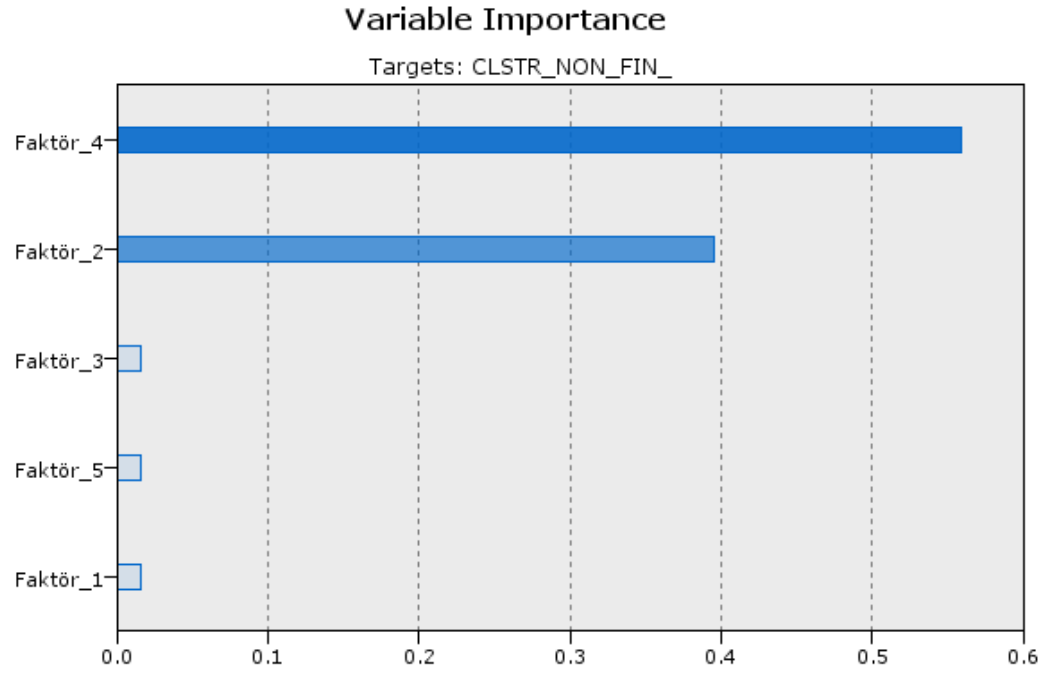
QUEST ve C5 modellerinin genele doğruluk oranları yaklaşık olarak %78 olarak elde edilmiştir. CHAID karar ağacı modelinden sonraki en yüksek doğruluk düzeyine sahip modellerdir.

QUEST Karar Ağacı Modeli				
'Partition'	1_Training		2_Testing	
Correct	337	74.89%	89	78.07%
Wrong	113	25.11%	25	21.93%
Total	450		114	

Faktörlerin göreceli önem oranları ve sıraları hem grafik hem de tablo halinde SPSS Clementine V.12 ile elde edilmiştir. DVM, C&R Tree ve CHAID karar ağaçları modellerinde olduğu gibi, ilk iki en etken faktörler çalışma şartları ve ücret/ödüllendirme faktörleri elde edilmiştir. En etken faktör göreceli önem oranı yaklaşık olarak %56 ile çalışma şartları faktörü iken, yaklaşık %40 göreceli oran değeri ile ücret/ödüllendirme faktörü ikinci en etken faktördür. Son üç faktör olan yönetimin tavrı, iş güvencesi ve iş arkadaşları yaklaşık %2 göreceli önem oranlarında sahiptirler.

Tablo 57. QUEST karar ağacı modelinde faktörlerin göreceli önem oranları

aktörler	Göreceli Önem Oranı
Faktör_4 (Çalışma Şartları)	0.5586
Faktör_2 (Ücret/Ödüllendirme)	0.3950
Faktör_1(Yönetimin Tavrı)	0.0155
Faktör_5 (İş Güvencesi)	0.0155
Faktör_3 (İş Arkadaşları)	0.0155



QUEST karar ağacı modeli özellikli ayarlamaları:

Fields

Target

CLSTR_NON_FIN_

Inputs

Faktör_1

Faktör_2

Faktör_3

Faktör_4

Faktör_5

Build Settings

Use partitioned data: true

Partition: Partition

Calculate variable importance: true

Calculate raw propensity scores: false

Calculate adjusted propensity scores: false

Use frequency: false

Levels below root: 5

Mode: Simple

Use misclassification costs: true

Training Summary

Algorithm: QUEST

Model type: Classification

QUEST modeli ile elde edilen karar ağacı ve kurallar kümesi aşağıda verilmiştir. İlk bölünme ücret/ödüllendirme üzerinden yapılmıştır. Ücret/ödüllendirme faktör skoru 3,24'den büyük iken bağımlı değişken olan finansal olmayan kurum performans başarısı başlangıçtaki isabet oranı %46'dan yaklaşık %71'e yükselmektedir. İsabet oranı alt gruplarda arttığı görülmektedir. 4 numaralı düğümde isabet oranı %100 oranına yükselirken, 6 numaralı düğümde ise yine %74 gibi bir orana yükselmiştir. Ücret/ödüllendirme 3,24 faktör değerinden küçük ve eşit ve çalışma şartları 4,58'den büyük olduğu zaman finansal olmayan kurum performans başarı oranı %46 olan başlangıç isabet oranından %100'e ulaşmaktadır. Yine bir başka alt grup olan, ücret/ödüllendirme 3,24'ten büyük ve çalışma şartları 3,44 faktör değerinden büyük olduğu zaman isabet oranı %74'e yükseldiği hem karar ağacında hem de kurallar kümesinden açıkça görülmektedir.

Rules for 1 - contains 2 rule(s)

Rule 1 for 1 (248; 0.758)

if Ücret/Ödüllendirme $\leq$ 3.237
and Çalışma Şartları $\leq$ 4.581
then 1

Rule 2 for 1 (16; 0.625)

if Ücret/Ödüllendirme $>$ 3.237
and Çalışma Şartları $\leq$ 3.438
then 1

Rules for 2 - contains 2 rule(s)

Rule 1 for 2 (6; 1.0)

if Ücret/Ödüllendirme $\leq$ 3.237
and Çalışma Şartları $>$ 4.581
then 2

Rule 2 for 2 (180; 0.739)

if Ücret/Ödüllendirme $>$ 3.237
and Çalışma Şartları $>$ 3.438
then 2

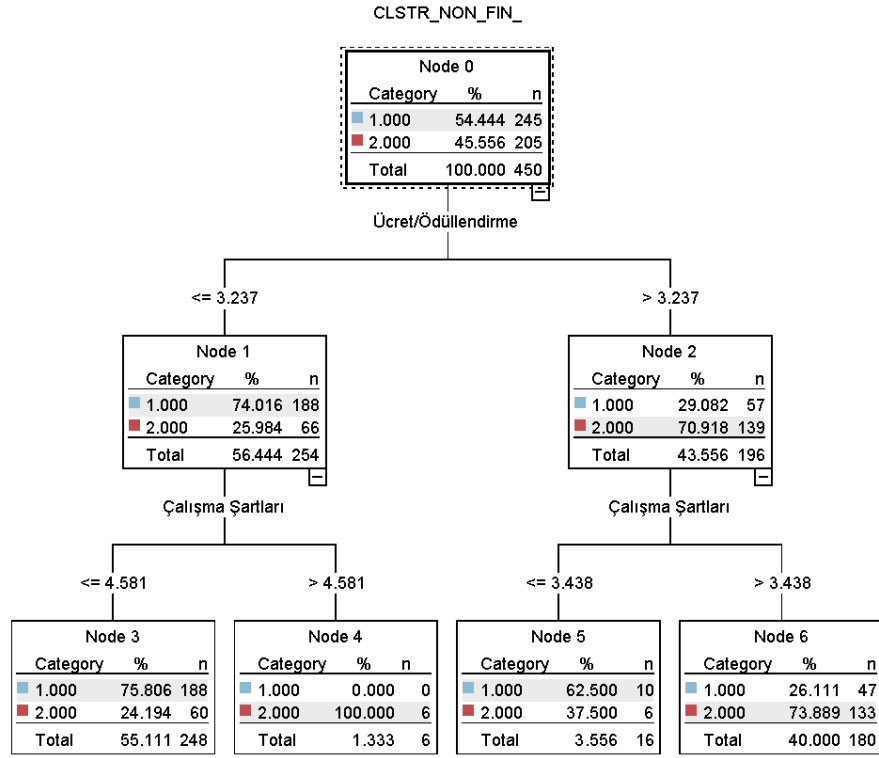
Ücret/Ödüllendirme $\leq$ 3.237 [Mode: 1] (254)

Çalışma Şartları $\leq$ 4.581 [Mode: 1] $\Rightarrow$ 1 (248; 0.758)

Çalışma Şartları $>$ 4.581 [Mode: 2] $\Rightarrow$ 2 (6; 1.0)

Ücret/Ödüllendirme $>$ 3.237 [Mode: 2] (196)

Çalışma Şartları ≤ 3.438 [Mode: 1] $\Rightarrow$ 1 (16; 0.625)
Çalışma Şartları > 3.438 [Mode: 2] $\Rightarrow$ 2 (180; 0.739)



Şekil 47. Finansal olmayan kurum performans algısı için QUEST karar ağacı modelinin gösterimi

Model yardımı ile elde edilen bağımlı değişkene ilişkin tahmin sonuçları \$R-CLSTR_NON_FIN\$ değişkeni olarak belirtilmiştir. Gerçek değerler ile kıyaslandığı zaman CHAID kadar iyi sonuç vermese de bu modele yakın bir doğru tahmin sonuçları model tarafından elde edilmiştir.

Tablo 58. QUEST karar ağacı modelinin tahmin sonuçları

Table (78 fields, 564 records)													
		Faktör_3	Faktör_4	Faktör_5	Fin_Perf	Non_Fin_Perf	Bin_Fin_Perf	CLSTR_FIN	Bin_Non_Fin_Perf	CLSTR_NON_FIN	Partition	\$R-CLSTR_NON_FIN	\$RC-CLSTR_NON_FIN
1	143	4.000	4.833	4.500	3.750	4.000	1.000	2	1.000	2	1.000	2	0.736
2	429	4.333	4.500	3.500	3.000	3.250	0.000	1	0.000	1	0.000	1	0.756
3	000	4.333	4.000	4.500	4.000	3.750	1.000	2	1.000	2	1.000	2	0.736
4	857	3.000	2.500	3.500	1.500	2.250	0.000	1	0.000	1	0.000	1	0.756
5	857	3.667	3.833	4.500	3.500	3.750	1.000	2	1.000	2	1.000	2	0.756
6	714	3.833	4.167	5.000	3.500	3.250	1.000	2	0.000	1	0.000	1	0.756
7	143	4.000	3.667	2.000	3.000	2.500	0.000	1	0.000	1	0.000	1	0.756
8	714	4.500	4.833	1.000	3.000	3.500	0.000	1	1.000	1	1.000	1	0.736
9	714	4.333	4.500	4.000	3.250	3.000	0.000	1	0.000	1	0.000	1	0.736
10	429	2.833	4.167	3.000	3.000	3.000	0.000	1	0.000	1	0.000	1	0.736
11	571	3.833	2.500	1.000	2.750	2.250	0.000	1	0.000	1	0.000	1	0.756
12	143	4.000	4.667	4.000	3.000	3.000	0.000	1	0.000	1	0.000	1	0.736
13	000	4.667	4.333	2.500	3.500	3.250	1.000	2	0.000	2	0.000	2	0.736
14	429	3.667	3.833	4.500	1.750	1.500	0.000	1	0.000	1	0.000	1	0.756
15	286	4.833	4.833	5.000	4.250	4.250	1.000	2	1.000	2	1.000	2	0.736
16	857	4.500	3.500	3.000	2.750	2.750	0.000	1	0.000	1	0.000	1	0.756
17	000	3.000	2.833	3.500	3.250	4.000	0.000	1	1.000	1	1.000	1	0.756
18	857	2.833	3.333	4.500	3.000	2.250	0.000	1	0.000	1	0.000	1	0.756
19	286	1.500	2.833	5.000	4.000	4.000	1.000	2	1.000	2	1.000	2	0.756
20	286	4.167	1.833	5.000	3.000	2.250	0.000	1	0.000	1	0.000	1	0.756

5.9.6 C 5.0 Karar Ağacı Modeli İle Finansal Olmayan Kurum Performansın İnceleme

C 5.0 kara ağacının genel doğruluk oranı CHAID karar ağacı modelinden sonra QUEST gibi en sonraki en iyi modeldir. Genel doğruluk oranı yaklaşık olarak %78 olarak bulunmuştur. Fakat performans ölçümleri daha önce incelendiği zaman C 5 modeli diğer modellere kıyasla en az performans ölçümlerine sahip bir modeldir. Kazançlar (Gains), genel yanıtlar (Responses) ve lift grafikleri incelendiği zaman en kötü performans ölçümleri göstermiştir.

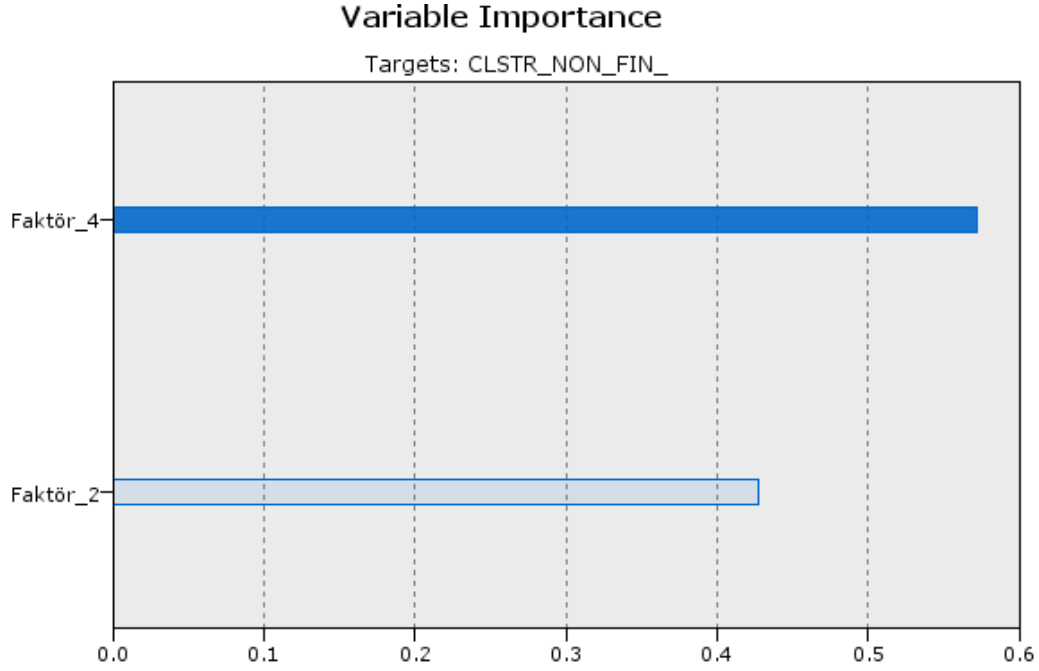
C5.0 Karar Ağacı Modeli				
'Partition'	1_Training		2_Testing	
Correct	336	74.67%	89	78.07%
Wrong	114	25.33%	25	21.93%
Total	450		114	

Faktörlerin göreceli önem oranları tablo ve grafik olarak SPSS Clementine vasıtasıyla elde edilmiştir. Çalışma şartları ve ücret/ödüllendirme faktörleri en etken faktörler olarak elde edilmiştir. DVM, CHAID, C&R Tree ve QUEST modellerinde de görüldüğü gibi en etken ilk iki faktörler çalışma şartları ve ücret/ödüllendirme faktörleri

olarak bulunmuştur. Bu modelde, %57 göreceli önem oranı ile en etken faktör çalışma şartlarıdır ve yaklaşık %43 göreceli önem oranı ile ücret/ödüllendirme faktörü ikinci en etken faktördür.

Tablo 59. C 5.0 Karar ağacı modelinde faktörlerin göreceli önem oranları

Faktörler	Göreceli Önem Oranı
Faktör_4 (Çalışma Şartları)	0.5725
Faktör_2 (Ücret/Ödüllendirme)	0.4275



Rules for 1 - contains 2 rule(s)

Rule 1 for 1 (230; 0.783)

if Ücret/Ödüllendirme $\leq$ 3.143

and Çalışma Şartları $\leq$ 4.167

then 1

Rule 2 for 1 (12; 0.75)

if Ücret/Ödüllendirme $>$ 3.143

and Çalışma Şartları $\leq$ 3.167

then 1

Rules for 2 - contains 2 rule(s)

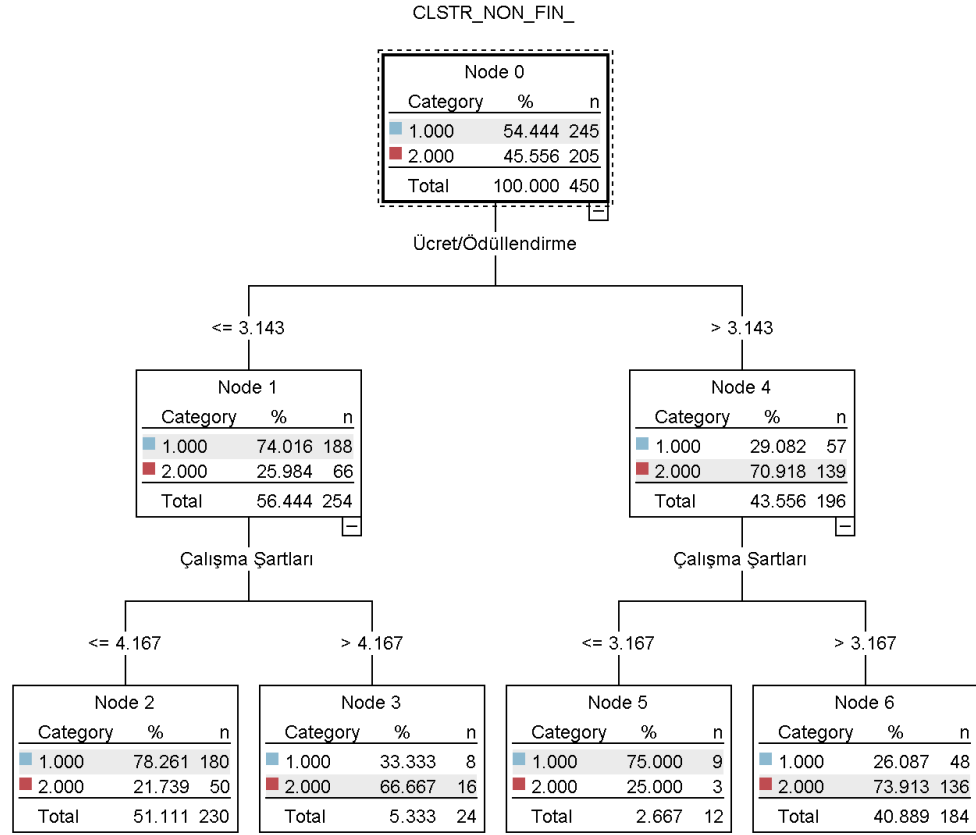
Rule 1 for 2 (24; 0.667)


```
if Ücret/Ödüllendirme <= 3.143
and Çalışma Şartları > 4.167
then 2
Rule 2 for 2 (184; 0.739)
if Ücret/Ödüllendirme > 3.143
and Çalışma Şartları > 3.167
then 2
```

```
Ücret/Ödüllendirme <= 3.143 [ Mode: 1 ] (254)
    Çalışma Şartları <= 4.167 [ Mode: 1 ] => 1 (230; 0.783)
    Çalışma Şartları > 4.167 [ Mode: 2 ] => 2 (24; 0.667)
Ücret/Ödüllendirme > 3.143 [ Mode: 2 ] (196)
    Çalışma Şartları <= 3.167 [ Mode: 1 ] => 1 (12; 0.75)
    Çalışma Şartları > 3.167 [ Mode: 2 ] => 2 (184; 0.739)
```

Karar ağacı ve kurallar kümesi bize C5.0 modelinin yorumlanmasında kolaylık sağlamaktadır. İlk bölünme ücret/ödüllendirme faktörü üzerinden yapılmaktadır. Başlangıç düğümde finansal olmayan kurum performans başarı isabet oranı yaklaşık olarak %46'dır. Karar ağacı üç seviye dört adet uç (terminal) düğümden oluşmuştur. En yüksek başarı isabet oranı 3 ve 6 numaralı düğümlerde görülmektedir. İsabet oranı 3 numaralı düğümden %67'ye yükselirken, 6 numaralı düğümden finansal olmayan kurum performans başarı oranı başlangıç isabet oranı olan %46'dan %74 gibi daha yüksek bir orana yükselmiştir.

Kurallar kümesi ve karar ağacında alt gruplar incelendiği zaman, ücret/ödüllendirme 3,14'den büyük ve çalışma şartları 3,17 faktör değerinden büyük olduğu zaman finansal olmayan kurum performans başarı oranı %74 olduğu görülmektedir.



Şekil 48. Finansal olmayan kurum performans algısı için C 5.0 karar ağacı gösterimi

C 5.0 modeli yardımı ile elde edilen tahmin sonuçları ve gerçek sonuçlar tablo halinde elde edilmiştir. Gerçek finansal olmayan kurum performans değişkeni CLSTR_NON_FIN olarak belirtilirken, model tarafından elde edilen yeni tahmini değişken \$C\$-CLSTR_NON_FIN olarak SPSS Clementine tarafından elde edilmiştir. Diğer karar ağaçları modelleri ve DVM modeli gibi C 5 modeli gözlenen kayıtlarının hepsini doğru tahmin edememiştir, bazı kayıtlar yanlış olarak tahmin edilmiştir. Doğruluk oranı zaten bize bunu göstererek, %78 gibi bir doğru tahmin etme oranı olduğu biliniyor.

Tablo 60. C 5.0 karar ağacı modelinin tahmin sonuçları

Table (78 fields, 564 records) #2													
		Faktör_3	Faktör_4	Faktör_5	Fin_Perf	Non_Fin_Perf	Bin_Fin_Perf	CLSTR_FIN	Bin_Non_Fin_Perf	CLSTR_NON_FIN	Partition	\$C-CLSTR_NON_FIN	\$CC-CLSTR_NON_FIN
1	143	4.000	4.833	4.500	3.750	4.000	1.000	2	1.000		2_1_Training	2	0.737
2	429	4.333	4.500	3.500	3.000	3.250	0.000	1	0.000		1_2_Testing	2	0.654
3	000	4.333	4.000	4.500	4.000	3.750	1.000	2	1.000		2_1_Training	2	0.737
4	857	3.000	2.500	3.500	1.500	2.250	0.000	1	0.000		1_1_Training	1	0.780
5	857	3.667	3.833	4.500	3.500	3.750	1.000	2	1.000		2_1_Training	1	0.780
6	714	3.833	4.167	5.000	3.500	3.250	1.000	2	0.000		1_1_Training	1	0.780
7	143	4.000	3.667	2.000	3.000	2.500	0.000	1	0.000		1_1_Training	2	0.737
8	714	4.500	4.833	1.000	3.000	3.500	0.000	1	1.000		1_1_Training	2	0.737
9	714	4.333	4.500	4.000	3.250	3.000	0.000	1	0.000		1_1_Training	2	0.737
10	429	2.833	4.167	3.000	3.000	3.000	0.000	1	0.000		1_1_Training	2	0.737
11	571	3.833	2.500	1.000	2.750	2.250	0.000	1	0.000		1_1_Training	1	0.780
12	143	4.000	4.667	4.000	3.000	3.000	0.000	1	0.000		1_1_Training	2	0.737
13	000	4.667	4.333	2.500	3.500	3.250	1.000	2	0.000		1_1_Training	2	0.737
14	429	3.667	3.833	4.500	1.750	1.500	0.000	1	0.000		1_1_Training	1	0.780
15	286	4.833	4.833	5.000	4.250	4.250	1.000	2	1.000		2_1_Training	2	0.737
16	857	4.500	3.500	3.000	2.750	2.750	0.000	1	0.000		1_1_Training	1	0.780
17	000	3.000	2.833	3.500	3.250	4.000	0.000	1	1.000		2_1_Training	1	0.780
18	857	2.833	3.333	4.500	3.000	2.250	0.000	1	0.000		1_1_Training	1	0.780
19	286	1.500	2.833	5.000	4.000	4.000	1.000	2	1.000		2_1_Training	1	0.780
20	286	4.167	1.833	5.000	3.000	2.250	0.000	1	0.000		1_1_Training	1	0.780

SONUÇ

Makine öğrenimi yapay zekâ alanının bir koludur. Deneysel verilere dayanan davranışların gelişmesi için bilgisayarlar yardımı ile algoritmaların geliştirilmesi ve tasarlanması ile ilgilenen bir bilimsel disiplindir. Makine öğreniminin ana odak noktası karmaşık örüntülerin tanınmasını otomatik bir şekilde öğrenmek ve veriye dayalı olacak şekilde doğru kararlar vermektir. Destek vektör makineleri (DVM) hem sınıflandırma hem de regresyona uygulanabilen makine öğrenimi alanından olan bir denetimli öğrenme yöntemidir. Bağımlı değişken kategorik olduğu zaman sınıflandırma yapabilen ve bağımlık değişken sürekli olduğu zaman regresyon analizi yapabilen bir yöntemdir. Destek vektör makineleri istatistik öğrenme teorisinde yer alan ve yapısal risk minimizasyonu ilkesine dayalı bir yöntemdir. Aynı şekilde, karar ağaçları veri madenciliği, istatistik ve makine öğrenimi alanlarında kullanılan en popüler sınıflandırma algoritmalarındandır.

Bu tez çalışmasında son yıllarda yaygın bir şekilde kullanılan destek vektör makineleri ve karar ağaçları yöntemleri kullanılmıştır. Bu yöntemler kullanılarak insan kaynakları yönetimi alanında bir uygulama yapılmıştır. Veri madenciliğinin temel kavramları, makine öğrenimi, istatistik öğrenme, destek vektör makineleri ve karar ağaçları yöntemleri ele alınmıştır. DVM ve karar ağaçları algoritmaları son yıllarda yaygın olarak kullanılmaktadır. Bilgi çalışanlarının kurum performansı üzerindeki etkilerinin incelenmesinde DVM ve karar ağaçları yöntemleri kullanılmıştır.

Bilgi işçisi ya da bilgi çalışanı kavramları literatüre bakılarak detaylı bir şekilde incelenmiştir. Belirli bir tanımlı olmamakla beraber, yaygın olarak kullanılan ve bu alanda hem teorik hem de deneysel çalışmalar yapan araştırmacıların tanımlarına ve açıklamalarına yer verilmiştir. Ayrıca, çalışanların memnuniyeti kavramına genel olarak

bakıldıktan sonra memnuniyeti etkileyen faktörler incelenmiştir. Aynı şekilde bilgi çalışanlarının memnuniyetini etkileyen faktörler de ele alınmıştır.

Çalışmada kullanılan veri kümesi 2011 yılında bilgi çalışanları üzerinde yapılan anket yolu ile elde edilmiştir. İstanbul ili içerisinde hizmet sektöründe çalışan bilgi işçileri üzerinde yapılan bir çalışmadır. 2300 çalışana gönderilen anketlerden 564'üne cevap gelmiştir. Dolayısıyla, elde edilen veri kümesinde 564 adet bilgi çalışanına ait kayıtlar mevcuttur. Ankette, bilgi çalışanının kurum performansı üzerindeki etkisi 42 değişken ile ölçülmüştür. Bu değişkenlerden 34'ü bilgi çalışan memnuniyeti, 4'ü finansal ve 4'ü ise finansal olmayan kurum performansı için kullanılmıştır. Anket sorularının cevapları, 1 (Kesinlikle Katılıyorum), 2 (Katılıyorum), 3 (Kararsızım), 4 (Katılmıyorum), 5 (Kesinlikle Katılmıyorum) şeklinde beşli likert ölçeği kullanılarak değerlendirilmiştir.

Sonuçların incelenmesi için elde edilen veriler düzenlenerek IBM-SPSS (v.19) istatistik programına uygun olacak şekilde düzenlenmiştir. Daha sonra, analizlerin yapılması için SPSS ile beraber çalışan SPSS-Clementine (v.12) modelleme programı kullanılarak modeller kurulmuştur. Bu modellere ilişkin analiz sonuçları, tablolar ve grafikler halinde elde edilmiştir.

Elde edilen anket verilerini kullanarak bilgi çalışanlarının memnuniyetlerine ilişkin faktörlerin belirlenmesinde, faktör analizi yönteminin uygun istatistiksel teknik olup olmadığını test etmek için, KMO örneklem uygunluk ölçümü, Barlett's testi ve ilgili matris sonuçları incelenmiştir. Bu sonuçlara göre keşfedici faktör analizinin (Exploratory Factor Analysis) uygun istatistiksel bir yöntem olduğu anlaşılmıştır. Bilgi çalışanlarının memnuniyetini ölçen ilgili değişkenler için uygulanan ölçeğin geçerliliği ve güvenilirliği test edilmiştir. Kullanılan ölçeğin geçerliliğini test etmek için keşfedici faktör analizi ve güvenilirliğini test etmek için ise Cronbach's Alpha katsayısı kullanılmıştır. Faktör analizinin uygun bir istatistiksel yöntem olduğu test sonuçları ile

belirlendikten sonra, analiz elimizdeki anket verisine uygulanarak bilgi çalışanlarında olması gereken 5 adet kritik faktör elde edilmiştir. Bu faktörler; yönetimin tavrı, ücret/ödüllendirme, iş arkadaşları, çalışma şartları, iş güvencesi'dir.

Anket verisinin kurumsal performans algısı bölümü, finansal ve finansal olmayan performans algılarından oluşmaktadır. Finansal ve finansal olmayan kurum performans algıları kümeleme analizi yolu ile başarılı ve başarısız olarak şekilde yapılmıştır. Herhangi bir değişkeni kategorik bir değişken yapmak için ortalama ve medyan yöntemleri yaygın olarak kullanılan yöntemler olmasına karşın, aşamalı kümeleme analizi (Two-Step Cluster Analysis) daha homojen kümeleme yaptığından bu yöntem seçilmiştir.

Bilgi çalışanlarında olması gereken kritik faktörler keşfedici faktör analizi yardımı ile belirlendi. Kurumlara yol gösterici olması için, hangi faktörlerin üzerinde durulması önemli bir noktadır, dolayısıyla karar destek sistemleri ve veri madenciliğinde yaygın olarak kullanılan DVM ve karar ağacı yöntemleri kullanıldı.

Bilgi çalışanlarının kurumların finansal performansları üzerindeki etkileri incelenmesi sürecinde DVM ve karar ağaçlarından C&R Tree (Classification and Regression Tree), C5, CHAID (Chi Squared Automatic Interaction Detection) ve QUEST (Quick, Unbiased, Efficient Statistical Trees) algoritmaları kullanılmıştır. Bu modellerin performansları, çeşitli performans ölçütleri açısından kıyaslanmıştır. Performans ölçümlerinin sonuçlarına göre DVM modeli, genel doğruluk oranı (overall accuracy rate), eğri altında kalan alan (area under curve) ve en yüksek kâr (maximum profit) performans ölçüm teknikleri açısından en iyi model olduğu görülmektedir. Karar ağaçlarından C 5, QUEST ve C&R Tree modellerinin genel doğru tahmin etme oranları aynı olmasına rağmen, C&R Tree karar ağacı modeli en yüksek kâr ve eğri altında kalan alan ölçümleri açısından DVM modelinden sonra en iyi ikinci model olduğu sonucuna varılmıştır. Destek vektör makineleri içerisinde daha detaylı bir inceleme yapılmıştır.

DVM Kernel fonksiyonları olan RBF (Radial Based Function), lineer, çok terimli, Sigmoid modelleri tek tek incelendikten sonra, çok terimli Kernel fonksiyonlu DVM modeli en yüksek doğruluk oranına sahip model olduğu görülmüştür. Çok terimli (polynomial) DVM modeline bakıldığı zaman, bilgi çalışanlarında olması gereken kritik faktörlerin göreceli önem sırasına göre, Çalışma Şartları (%40), İş Arkadaşları (%20), Ücret/Ödüllendirme (%16), Yönetimin Tavrı (%13) ve İş Güvencesi (%11) olarak bulunmuştur. Radyal tabanlı Kernel fonksiyonlu (RBF) DVM modelinde Ücret/Ödüllendirme (%35), Çalışma Şartları (%26), İş Arkadaşları (%16), Yönetimin Tavrı (%13) ve İş Güvencesi (%10) olarak belirlenmiştir. Lineer Kernel fonksiyonlu DVM modelinde ise göreceli önem oranı en yüksek ilk iki faktör Ücret/Ödüllendirme ve Çalışma Şartları faktörleri kritik faktörleri model tarafından belirlenmiştir. Karar ağaçları modelleri arasında en yüksek performans ölçüm değerlerine sahip olan C&R tree modeli yardımı ile elde edilen faktörler Çalışma Şartları (%41), Ücret/Ödüllendirme (%35), İş Güvencesi (%12), İş Arkadaşları (%11) ve Yönetimin Tavrı (%0) olarak model tarafından belirlenmiştir. Burada yönetimin tavrı en az öneme sahip faktördür. Diğer karar ağaçları modellerine bakıldığı zaman Çalışma Şartları ve Ücret/Ödüllendirme faktörleri en önemli faktörler olarak diğer modeller tarafından da öne çıkmaktadır.

Aynı şekilde, bilgi çalışanlarının kurumların finansal olmayan performansları üzerindeki etkileri DVM ve karar ağaçları algoritmaları kullanılarak incelenmiştir. Çalışmada kullanılan ankette, dört tane finansal olmayan performans değişkenleri mevcuttur. Bu değişkenlere ait değerlerin ilk önce ortalamaları alındı. Daha sonra, SPSS- Clementine yardımı ile kümeleme analizi yapıldı. Analiz sonucunda finansal olmayan performans algısı “başarılı” ve “başarısız” olacak şekilde homojen bir kümeleme yapıldı. Bu modelimizde, veri kümesi SPSS üzerinden SPSS-Clementine (v.12) modelleme programına aktararak veri kümesi üzerinde gerekli düzenlemeler yapıldı. DVM ve karar ağaçları makine öğrenme yöntemlerinden olduklarından, veri kümesinin %80’i eğitime kümesi (training set) için ve %20’si ise test kümesi (testing

set) için bölünmüştür. DVM ve karar ağaçları modellerinin her biri öncelikle eğitime kümesi üzerinde eğitildikten sonra test kümesi üzerinde test edilmiştir. Tüm modellerin performansları ikili sınıflandırma (Binary Classification) düğümü yardımı ile kıyaslanarak finansal performans incelemesinde elde edilen sonuçlara benzer şekilde elde edilmiştir. Destek vektör makinelerinin dört Kernel fonksiyonunun performansları incelendiği zaman RBF ve lineer DVM modellerinin %78,95 gibi en yüksek genel doğruluk oranlarına sahip modeller olduğu görülmektedir. Eğri altında kalan alan performans ölçütüne bakıldığı zaman RBM-DVM %89,7 ile en iyi model ve %89,5 ile yine lineer-DVM modeli en iyi ikinci DVM modeli olduğu görülmüştür. Karar ağaçları modelleri arasında en yüksek doğruluk oranı %79,83 ile CHAID modeli olmuştur. QUEST ve C5 karar ağacı modellerinin her ikisinin de doğruluk oranı %78 olarak bulunmuştur. Eğri altında kalan alan performans ölçümüne bakıldığı zaman yine CHAID karar ağacı modeli %83 ile yine en iyi model olmuştur. Karar ağacı modelleri ile DVM modellerinin performansları kıyaslandığı zaman genel doğruluk oranı ölçümü %79,83 ile yine CHAID en yüksek orana sahip olmasına karşın, alan altında kalan oran açısından RBF-DVM modeli en iyi performans ölçümüne sahiptir.

Bilgi çalışanlarının kurumların finansal olmayan performansları üzerindeki etkisi, RBF-DVM modeli yardımı ile incelendi. Buna göre, Ücret/Ödüllendirme faktörü en yüksek önem oranına sahip olduğundan (%46) en etken faktördür. Göreceli önem oranlarına göre sonraki en etken faktörler sırasıyla, Çalışma Şartları (%26), Yönetimin Tavrı (%16) olarak görülmüştür. Bunun yanı sıra, İş Arkadaşları (%10) ve İş Güvencesi (%2) en az etken faktörler olduğu RBF-DVM modeli tarafından elde edilmiştir.

Finansal olmayan performans ölçümlerinde karar ağaçları algoritmalarında en yüksek performans ölçümlerini veren CHAID modelidir. Bu modele göre bilgi çalışanlarında olması gereken kritik faktörlerin kurumların finansal olmayan performansları üzerindeki etkilerine bakıldığı zaman, çalışma şartları (%76) en yüksek göreceli öneme sahip bir faktör iken, ücret/ödüllendirme (%18) ikinci en etken

faktördür. İş arkadaşları (%3) ve iş güvencesi(%2) ile sonraki en iyi etken faktörler olarak model tarafından belirlenmiştir. Bu modelde yönetimin tavrı faktörü hiçbir şekilde etken bir faktör olarak listede yer almamaktadır. Her ne kadar CHAID karar ağacı modeli en yüksek değerli performans ölçümlerine sahip olsa da, diğer karar ağaçlarına da bakılabilir. Örneğin, C&R Tree karar ağacı modelinde bilgi çalışanlarında olması gereken kritik faktörler, finansal olmayan performans üzerindeki etkileri sırasıyla ücret/ödüllendirme (%50), çalışma şartları (%27), yönetimin tavrı (%17), iş arkadaşları (%6) ve iş güvencesi (%0) olmuştur. QUEST karar ağacı modelinde çalışma şartları (%56), ücret/ödüllendirme (40%), yönetimin tavrı (%2) ve iş arkadaşları (%2) olarak elde edilmiştir. Son olarak C 5.0 karar ağacında ise çalışma şartları (%57) ve ücret/ödüllendirme (%43) olarak model tarafından elde edilmiştir. Tüm modellerde görüldüğü gibi, çalışma şartları ve ücret/ödüllendirme faktörleri en yüksek göreceli önem oranlarına sahip kritik faktörler olarak karşımıza çıkmaktadır.

Bu tez çalışmasında bilgi çalışanlarında olması gereken kritik faktörler belirlenmiştir. Daha sonra bu faktörlerin kurumların finansal ve finansal olmayan performansları üzerindeki etkileri destek vektör makineleri ve karar ağaçları yöntemleri kullanılarak detaylı olarak incelenmiştir. Kurumlara yol gösterici olması için kurumların finansal ve finansal olmayan performansları için hangi özellikler üzerinde durulması konusunda yardımcı olmaktadır.

KAYNAKÇA

- Alpaydm, E. : Introduction to Machine Learning, MIT Press,2004.
- Anderson, : “Customer satisfaction and shareholder value”, **Journal of**
E.W., v.d. **Marketing**, vol. 68(4), 2004, pp. 172–185.
- Attneave, F. : Applications of Information Theory to Psychology, Holt, Rinehart and
Winston, 1959.
- Barlett, P.L. : “The sample complexity of pattern classification with neural
networks: The size of the weight is more important than the size of the
network”, **IEEE Transaction on Information Theory**, vol.44(3.2),
1998, pp.525-536.
- Bartlett, P.L., : “Rademacher and Gaussian complexities: Risk bounds and structural
Mendelson results”, **Journal of Machine Learning Research**, Vol.3 ,,2002, pp.
S. 463-482.
- Bellman, R.E. : Adaptive control processes: A Guided Tour, Princeton University
Press, 1961.
- Benjamini, : “Statistical Method for Data Mining”, **Data Mining and Knowledge**
Y.M. **Discovery Handbook**, 2nd Edition, Ed. by L. R. Oded Maimon, New
York, 2010, Springer Science +Business Media, LLC., pp. 523-540
- Bi,J., Zhang : “Column-generation boosting methods for mixture of kernels” **KDD**,
T., Bennett K. 2004, pp. 521-526.
- Blumer, A., : Learnability and the Vapnik-Chervonenkis Dimension, JACM, 1990.
v.d.
- Bowen., P., : ”Job satisfaction of South African quantity surveyors”, **Engineering,**
K. Cattell **Construction and Architectural Management**, vol. 15 no. 3, 2008,
pp.260-269.
- Bratko, I., : “Trading accuracy for simplicity in decision trees”, **Machine**
Bohanec M. **Learning**, vol.15, 1994, pp. 223-250.

- Brayfield, A.H., Crockett W.H. : “Employee attitudes and employee performance”, **Psychological Bulletin**, Vol. 52, 1955, pp. 396-424.
- Breiman, L. , Friedman J. , Olshen R., Stone C. : Classification and Regression Trees, Wadsworth Int. Group, 1984.
- Brown, Steven P., Lam Son K. : “A Meta-Analysis of Relationships Linking Employee Satisfaction to Customer Responses”, **Journal of Retailing**, vol.84(3), 2008, pp.243-255
- Brown, Steven P., Peterson R.A. : “Antecedents and Consequences of Salesperson Job Satisfaction: Meta-Analysis and Assessment of Causal Effects,” **Journal of Marketing Research**, vol. 30 (1), 1993, pp.63–77.
- Burke, R., J. Graham, F. Smith : “TQM IMPLEMENTATION Effects of reengineering on the employee satisfaction-customer satisfaction relationship”, **The TQM Magazine**, vol. 17 no. 4, 2005, pp.358-363.
- Chang, C.C., Lin, C.J. : “LibSVM: A library for support vector machines software”, (Çevirimiçi) <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>, 2001.
- Chapelle, O., Haffner P., and Vapnik V.N. : “Support Vector machines for histogram-based image classification”, **IEEE Trans. on Neural Networks**, Vol.10(3.5), 1999, pp.1055-1064.
- Chen, H.S. , C. Yang, J. Y. Siau, H. H. Wang : “The development of an employee satisfaction model for higher education”, **The TQM Magazine**, vol. 18 no. 5, 2006, pp.484-500
- Cherkassky, V., Mulier F. : Learning From Data: Concepts, Theory and Methods, Wiley, 1998
- Cherkassky, V., Shao X., Mulier F., Vapnik V.N. : “Model Complexity control for regression using VC generalization bounds”, **IEEE Transactions on Neural Networks**, vol.10, 1999, pp. 1075-1089.

- Christianini, N., Shawe-Taylor J. : An Introduction to Support Vector Machines and Other Kernel -Based Learning Methods, Cambridge University Press, 2000.
- Christopher, J.B. : “A Tutorial on Support Vector Machines for Pattern Recognition”, **Data Mining and Knowledge Discovery**, 2, 1998, pp. 121-167.
- Cole, M., Bruch, H. : “Organizational identity strength, identification, and commitment and their relationship with turnover intention: does organizational hierarchy matter?”, **Journal of Organizational Behavior**, Vol. 27, 2006, pp. 585-605.
- Collins, D. : “Knowledge Work or Working Knowledge? Ambiguity and Confusion in the Analysis of the Knowledge Age”, **Employee Relations**, vol. 19, number 1, 1997, pp. 41-47
- Cortes, C., Vapnik V. : “Support Vector Networks”, **Machine Learning**, vol.20, 1995, pp.273-297
- Cronbach, L.J. : “Coefficient Alpha and the Internal Structure of Tests”, **Psychometrika**, Vol. 16, 1951, pp.297-334.
- Cruz, N. M., V. : “The influence of employee motivation on knowledge transfer”, M. Perez, C. T. **Journal of Knowledge Management**, vol 13 no. 6, 2009, pp.478-490
Cantero
- Davenport, T. : “A Measurable Proposal”, **CIO**, vol. 16, issue 16, 2002, pp. 46-48.
- Davenport, T. : “Can you boost knowledge work’s impact on the bottom line?”, **Management Update**, Vol. 7 No. 11, 2002, pp. 3-5.
- Davenport, T., Thomas, R. and Cantrell, S. : “The mysterious art and science of knowledge-worker performance”, **MIT Sloan Management Review**, Vol. 44 No. 1, 2002, pp. 23-9.
- DePablos, P. : “Competencies and human resource management: implications for O., M. Lytras organizational competitive advantage”, **Journal of Knowledge Management**, vol. 12 no. 6, 2008, pp.48-56.
- Dickter, D., Roznowski M., Harrison D. : “Temporal tempering: an event history analysis of the process of voluntary turnovers”, **Journal of Applied Psychology**, Vol. 81, 1996, pp.705-16.

- Dietterich, T.G., Kearns M., Mansour Y. : “Applying the weak learning framework to understand and improve C4.5”, **Proceedings of the Thirteenth International Conference on Machine Learning**, San Francisco, 1996, Morgan Kaufmann, pp. 96-104.
- DiGiovanni, M. : “Developing Knowledge Workers on the Factory Floor”, **Development and Learning in Organizations**, vol. 18, number 4, 2004, pp. 9-17
- Dove, R. : “The knowledge worker”, **Automotive Manufacturing & Production**, Vol.110 No.6, 1998, pp.26-8
- Drucker, P. : **The Landmarks of Tomorrow**, New York, NY, Harper & Row, 1959.
- Drucker, P. : **Post-Capitalist Society**, New York, NY, HarperBusiness, 1993.
- Drucker, P. : **Adventures of a Bystander**, New Brunswick, NJ, Transaction Publishers, 1994.
- Drucker, P. : “Knowledge-worker productivity: the biggest challenge”, **California Management Review**, Vol. 41 No. 2, 1999, pp. 79-85.
- Drucker, P., : “The age of social transformation, **Atlantic Monthly**, 1994, pp.53-80.
- Duda, R.O., Hart P.E., Stork, D.G. : **Pattern Classification**, Wiley, 2001.
- Dura, C. : **Bilgi Toplumu**, T.C. Kültür Bakanlığı Yayınları, Ankara, 1990
- Ersel, B. : “Bilgi çağında çalışma ilişkileri ve beyin göçü”, **II. Ulusal Bilgi, Ekonomi ve Yönetim Kongresi**, Bildiriler Kitabı, Derbent-İzmit, 2003ss. 709-721
- Eskildsen, J. K., Dahlgaard, J. J. : “A causal model for employee satisfaction”, **Total Quality Management**, 11(8), 2000, pp. 1081–1094.

- Fayyad, U., Patetsky - Shapiro G., Smyth P., Uthurusamy R. : Advances in Knowledge Discovery and Data Mining, Cambridge, MA, AAAI/MIT Press, 1996.
- Fayyad, U., Piatetsky – Shapiro G., Smyth P. : “From Data Mining to Knowledge Discovery: An Overview”. Ed. by U.G.-S. Fayyad, **Advances in Knowledge Discovery and Data Mining**, 1996, AAAI/MIT Press. pp.1-30.
- Fayyad, U., Irani K.B. : “The attribute selection problem in decision tree generation”, **In Proceedings of Tenth National Conference on Artificial Intelligence**, Cambridge, MA, 1992, AAAI Press/MIT Press, pp. 104-110.
- Ferri, C., Flach P., Hernandez - Orallo J. : “Learning Decision Trees Using the Area Under the ROC Curve”, **Proceedings of the 19th International Conference on Machine Learning**, 2002, Morgan Kaufman, pp. 139-146.
- Field, A. : Discovering Statistics Using SPSS, London, Sage, 2005.
- Foote, D.A., T. Tang : “Job satisfaction and organizational citizenship behavior (OCB) Does team commitment make a difference in self-directed teams?”, **Management Decision** , vol. 46 no. 6, 2008, pp.933-947.
- Frawley, W., G. Piatetsky-Shapiro : **Discovery in Databases: An Overview**, Menlo Park, C.A., The AAAI/MIT Press, 1996
- Frawley, W.P.S. : “Knowledge Discovery in Databases: An overview”, **AI Magazine**, Vol.13, 1992, pp. 213-228.
- Friedman, J.H. : “A recursive partitioning decision rule for nonparametric classifiers”, **IEEE Trans. on Comp.**, C26, 1977, pp.404-408.
- Fukunaga, K. : Introduction to Statistical Pattern Recognition, San Diego, CA, Academic, 1990.
- Garrick, J., S. Clegg : “Knowledge Work and the New Demands of Learning”, **Journal of Knowledge Management**, vol. 4, number 4, 2000, pp. 279-286

- Gehrke, J. : “Decision Trees”, **The Handbook of Data Mining**, Ed. by Nong Ye, New Jersey, 2003, Lawrence Erlbaum Associates Inc, pp. 3-24.
- Gillo, M.W. : “A Honeywell 600 program for an automatised survey analysis”, **Behavioral Science**, vol.17, 1972, pp. 251-251.
- Gorunescu, Florin : **Data Mining: Concepts, Models and Techniques** (Intelligence System Library, Volume 12), Springer-Verlag Berlin Heidelberg, 2011.
- Graham, M. V., P. E. Messner : “Principals and job satisfaction”, **International Journal of Educational Management**, vol. 12 no. 5, 1998, pp.196-202.
- Gu, Z., R. C. Sensiu : “Drivers of job satisfaction as related to work performance in Macao casino hotels”, **International Journal of Contemporary Hospitality Management** , vol 21 no. 5, 2009, pp.561-578.
- Guns, W.D., Liisa V. : “Rethinking Knowledge Work: Creating Value through Idiosyncratic Knowledge”, **Journal of Knowledge Management**, Vol.1 No:4, 1998, pp. 287-293.
- Hamel, L. : Knowledge Discovery With Support Vector Machines, New Jersey, John Wiley & Sons Inc, 2009.
- Han, J., Kamber M. : Data Mining Concepts and Techniques, San Fransisco, Elsevier, 2006.
- Hancock, T.R., Jang T., Li M., Tromp J. : “Lower Bounds on Learning Decision Lists and Trees”, **Information and Computation**, vol. 126(2), 1996, pp.114-122.
- Hastie, Trevor, Tibshirani R., Friendman J. : The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd.Edition), New York, Springer Science+Business Media LLC, 2009.
- Haykin, S. : Neural Networks: A Comprehensive Foundation, Tsinghua University Press, 2001.
- Herbrich, R. : Learning Kernel Classifiers: Theory and Algorithms, Massachusetts, The MIT Press, 2002.

- Herzberg, F. : **The Motivation to Work**, John Wiley and Sons, New York, NY. 1959
- Herzberg, F., : Job Attitudes: Review of Research Opinions, Psychological Service of
Mausner B., Pittsburgh, Pittsburgh, PA., 1957.
Peterson
R.O., Capwell D
.F.
- Heskett, J. L. : The Service Profit Chain: How Leading Companies Link Profit and
v.d. Growth to Loyalty, Satisfaction, and Value, The Free Press, New
York, 1997.
- Horvath, D. : “Knowledge worker definition”,
(çevrimiçi) [http://http://searchcrm.techtarget.com/definition/knowledge-worker.html](http://searchcrm.techtarget.com/definition/knowledge-worker.html), 2006
- Hsu, C.H., Y. : “The development and empirical validation of the Employee
C. Wang Satisfaction Index model”, **Total Quality Management & Business
Excellence**, vol.19 no. 4, 2008, pp.353-366.
- Huang, K., : “Learning Large Margin Classifiers Locally and Globally”, **IEEE
Yang H., King Transactions on Neural Networks**, 19(2), 2008, pp. 260-272.
I., Lyu M.R.
- Hunter, L., : “Finding Relevant Biomolecular Features”, **ISMB**, 1993, pp.190-197.
Klein T.E.
- Hwang, J., Lay : “Nonparametrik multivariate density estimation: A comparative
S., Lippman A. study”, **IEEE Transaction on Signal Processing**, vol.42(10), 1994,
pp. 2795-2810.
- Hyafil L., : “Construction optimal binary decision trees is NP-complete”,
Rivest R.L. **Information Processing Letters**, vol.5(1), 1976, pp.15-17.
- Iaffaldano, : “Job satisfaction and job performance: a meta-analysis”,
M.T., **Psychological Bulletin**, Vol. 97, 1985, pp. 251-73.
Muchinsky
P.M.
- Islam, R., A. Z. : “Employee Motivation: A Malaysian Perspective”, **International
Ismail Journal of Commerce and Management** , vol. 18 no. 4, 2008, pp.
344-362.

- Janz, B., Colquitt J. : “Knowledge Worker Team Effectiveness”, **Personnel Psychology**, vol. 50, issue 4, , 1997, pp. 55-62.
- Jimenez, L.O., Landgrebe D.A. : “Supervised Classification in Hight- Dimensional Space: Geometrical, Statistical, and Asymptotical Properties of Multivariate Data”, **IEEE Transaction on Systems Mand, and Cybernetics - Part C: Applications and Reviews**, vol. 28, 1998, pp. 39-54.
- Joachims, T. : “Making Large-scale SVM learning practical”, **Advances in Kernel Methods - Support Vector Learning**, Ed. by Schölkopf B., Burges C., Smola A., MIT Press, 1999.
- Kaiser, H.F. : “An Index of Factorial Simplicity”, **Psycometrika**, vol. 39, 1974, pp. 31-36.
- Karauzum, A., N. Anafarta, S. Irmak : “Predictors of burnout among middle managers in the Turkish hospitality industry”, **International Journal of Contemporary Hospitality Management**, vol.20 no.2, 2008, pp.186-198.
- Kass, G.V. : “An exploratory technique for investigating large quantities of caregorical data”, **Applied Statistics**, vol. 29(2), 1980, pp.119-127.
- Kaur, Harleen, Siri K.Wasan : Empirical Study on Applicaitons of Data Mining Techniques in Healthcare, **Journal of Computer Science**, vol. 2(2), 2006, pp.194-200.
- Kearns, M., Mansour Y. : “On the boosting abilitiy of top-down decision tree learning algorithms”, **Journal of Computer and Systems Sciences**, vol. 58(1), 1999, pp. 109-128.
- Kelley, R.E. : The gold-collar worker: harnessing the brainpower of the New Work force. Reading, MA, Addison-Wesley, 1985.
- Kelley, R.E. : “Managing the new workforce”, **Machine Design**, 1990, pp.109-13.
- Kelloway, E. K. , Barling J. : “Knowledge work as organizational behavior”, **International Journal of Management Reviews**, vol. 2, 2000, pp.287-304.
- Koys, D. : “How the achievement of human-resources goals drives restaurant Performance”, **Cornell Hotel and Restaurant Administration Quarterly**, vol. 44 (1), 2003, pp.17–24.

- Lance, C : “Job performance as a moderator of the satisfaction”, **Journal of Organizational Behavior**, vol 9, 1988, pp.271-280.
- Lang, J. C. : “Managerial Concerns in Knowledge Management”, **Journal of Knowledge Management**, vol. 5, number 1, 2001, pp. 46-47
- Lawler, E.E. and Porter, L.W. : “The effect of performance on satisfaction”, **Industrial Relations**, Vol.7, 1967, pp. 20-8.
- Locke, Edwin A. : “The Nature and Causes of Job Satisfaction,” in **Handbook of Industrial and Organizational Psychology**, Ed. by Dunnette Marvin, Chicago, 1976, Rand McNally, , pp.1297–350.
- Locke, Edwin A. : “What is job satisfaction?”, **Organizational Behavior and Human Performance**, Vol. 4, 1969, pp. 309-36.
- Loh, W.Y., Shih X. : “Split selection methods for classification trees”, **Statistica Sinica**, vol. 7, 1997, pp.815-840.
- Loveday, M. : “Employee satisfaction and HR at ISS”, **Management Development Review**, vol.9 no. 4, 1996, pp. 9-12.
- Lopez de Mantras, R. : “A distance -based attribute selection measure for decision tree induction”, **Machine Learning**, vol.6, 1991
- Maimon, O., Lior Rokach : “Introduction To Supervised Methods”, **Data Mining and Knowledge Discovery Handbook**, Ed. by Lior Rokach and Oded Maimon, 2005, Springer Science+Business Media Inc
- Marques, D.P., Fernando J. G. Simon : “The effect of knowledge management practices on firm performance”, **Journal of Knowledge Management**, Vol. 10, No. 3, 2006, pp. 143-156
- Maltzer, K., B. Renzl : “Personality Traits, Employee Satisfaction and Affective Commitment”, **Total Quality Management & Business Excellence**, vol. 18 no.5, 2007, pp.589-598.
- Mehta, M., Rissanen J., Agrawal R. : “MDL-based decision tree pruning”, In Proceedings of the First International Conference on Knowledge Discovery in Databases and Data Mining, 1995.

- Mitchell, T. : **Machine Learning**, McGraw – Hill, 1997.
- Morgan, J.N., Messenger R.C. : THAID: a sequential search program for the analysis of nominal scale dependent variables, Ann arbor, MI, Institute for Social Research, University of Michigan, 1973.
- Morgan, J.N., Sonquist J.A. : “Problems in the analysis of survey data, and a proposal”, **J.Amer. Statist. Assoc.**, vol.58, 1963, pp.415-434.
- Murthy, S.K. : “Automatic Construction of Decision Trees from Data: A MultiDisciplinary Survey Data”, **Data Mining and Knowledge Discovery**, vol.2(4), 1998, pp.345-389.
- Narang, R., Alka Dwivedi : “Managing The Job Satisfaction of Knowledge Workers: An Empirical Investigation”, **Asia Pacific Journal of Business and Management**, Volume 1 (1), 2010, pp.1-14
- Nilsson, N. J. : Introduction To Machine Learning: An Early Draft Of A Proposed Textbook, Standford, CA, Standford University, 2005.
- Nisbet, R., Elder J., Miner G. : Handbook of Statistical Analysis and Data Mining Applications, Elsevier, 2009.
- Nunnally, J.C. : **Psychometric Theory**, McGraw-Hill, New York, 1978
- Olaru, C., Wehenkel L. : “A complete fuzzy decision tree technique”, **Fuzzy Sets and Systems**, vol.138(2), 2003, pp.221-254
- Olson, David., Delen Dursun : **Advanced Data Mining Techniques**, Springer-Verlag Berlin Heidelberg, 2008.
- Pathirage, C., Dilanthi A., Richard H. : “Tacit knowledge and organizational performance: construction industry perspective”, **Journal of Knowledge Management**, Vol. 11, No. 1, 2007, pp. 115-126
- Pekalska, E., Paclik, P., ve Duin, R.P.W : “A generalized kernel approach to dissimilarity - based classification”, **Journal of Machine Learning Research**, vol. 2, 2001, pp.175-211.
- Quinlan, J.R., Rivest R.L. : “Inferring decision trees using minimum description length principle”, **Information and Computations**, vol.80, 1989, pp.227-248.

- Quinlan, J.R. : **C4.5: Programs for Machine Learning**, Los Altos, Morgan Kaufmann, 1993.
- Quinlan, J.R. : “Decision Trees and Multivalued Attributes”. In J. R. (ed.), **Machine Intelligence**, v.11, Ed. by J. Richards, Oxford, England, 1988, Oxford Univ Press, pp.305-318.
- Quinlan, J.R. : “Introduction of decision trees”, **Machine Learning**, vol.1, 1986, pp.81-106.
- Quinlan, J.R. : “Simplifying decision trees”, **International Journal of Machine Studies**, vol.27, 1987, pp.221-234.
- Quinlan, J.R. : “Unknown attribute values in induction”, **Proceedings of the Sixth International Machine Learning Workshop**, Ed. by Segre A., Corneli, New York, 1989, Morgan Kaufman.
- Raudys, S. : “How good are support vector machines?”, **Neural Networks**, vol.13, 2000, pp.17-19.
- Rigollet, P. : “Generalization error bounds in semi - supervised classification under the cluster assumption”, **Journal of Machine Learning Research**, vol.8, 2007, pp.1369-1392.
- Rokach,L.,
Oded Maimon : “Decision Trees”, **Data Mining and Knowledge Discovery Handbook**, Ed. by Oded Maimon, Lior Rokach, 2005, Springer Science+Business Media Inc., pp.165-192.
- Rounds, E. : “A combined non-parametric approach to feature selection and binary decision tree design”, **Pattern Recognition**, vol.12, 1980, pp.313-317.
- Rychetsky,
Matthias : Algorithms and Architectures for Machine Learning Based on Regularized Neural Networks and Support Vector Approaches, Shaker Verlag, 2001.
- Saravanamut-
hu, K. : “Gold-Collarism in the Academy : The Dilemma in Transforming Bean-Counters into Knowledge Consultants”, **Critical Perspectives on Accounting**, vol. 15, 2004, pp. 580-589.
- Schölkopf, B.,
C.J.C. Burges,
A.J. Smola : Advances in Kernel Methods: Support Vector Learning, Cambridge, MA, MIT Press, 1999.

- Scott, D.W., : “Multidimensional Density Estimation”, **Handbook of Statistics 24: Data Mining and Data Visualization**, Ed. by C.R. Rao, E.J.Wegman, J.L.Solka, Elsevier B.V, 2005, pp.229-261.
- Stephan R.Sain
- Sewell, M. : Structural Risk Minimization. Retrieved from (Çevrimiçi) <http://www.svms.org/srm/>, 2006.
- Shawe-Taylor, : **Kernel Methods for Pattern Analysis**, New York, Cambridge University Press, 2004.
- J., Nello
Cristianini
- Sijtsema, P. : “Knowledge work productivity in distributed teams”, **Journal of Knowledge Management** , vol. 13 no. 6, 2009, pp.533-546.
- M., V.
Ruohomaki, M.
Vartiainen
- Sonquist, J.A., : Searching for Structure. Institute for Social Research, Ann Arbor, MI, Baker E.L., Univ. of Michigan, 1971.
- Morgan J.N.
- Spector, P.E. : Job Satisfaction: Application, Assessment, Causes, and Consequences, Thousand Oaks, CA., Sage,1997
- SPSS : Clementine® 12.0 Algorithms Guide. Integral Solutions Limited, 2007.
- Clementine®
- Steinberg, Dan : “CART: Classification and Regression Trees”, **The Top Ten Algorithms in Data Mining**, Ed. by, XindongWu, Vipin Kumar, Boca Raton FL, Chapman & Hall/CRC Taylor and Francis Group LLC, 2009, pp.180-199.
- Stevens, J.P. : Applied Multivariate Statistics for the Social Sciences, Hillsdale, NJ, Earlbaum, 1992.
- Stewart, : Intellectual Capital: The New Wealth of Organizations, , New York, Thomas A. Doubleday, 1997
- Sutton, C.D. : “Classification and Regression Trees, Bagging and Boosting”, **Handbook of Statistics 24: Data Mining and Data Visualization**, Ed. by Ed. by C.R. Rao, E.J.Wegman, J.L.Solka, Elsevier B.V., 2005, pp. 303-329.

- Tape, Thomas G. : “The Area Under an ROC Curve”,
(Çevrimiçi)<http://gim.unmc.edu/dxtests/roc3.htm>, 2011.
- Teas, K. R. : “A Test of a Model of Department Store Salespeoples’ Job Satisfaction”, **Journal of Reailing**, vol. 67, 1981, pp.3-25
- Tsang, I.W., Kocsor A., Kwok J.T. : Simpler Core Vector Machine with Enclosing Balls, ICML, 2007.
- Tsang, I.W., ve diğerleri: : “Core vector machines: Fast SVM training on very large data sets”, **Journal of Machine Learning Research**, Vol.6, 2005, pp.363-392.
- Tuzun, I. K. : “The impact of identification and commitment on job satisfaction”, **Management Research News**, vol. 32 no. 8, 2009, pp. 728-738.
- Ulmer, D., Syptak J.M., Marsland D.W. : “Job satisfaction: putting theory into practice”, **Family Practice Management**, 1999.
- UNESCO : **Statistical Yearbook**, Paris, UNESCO, 1993.
- Utgoff, P.E., Cluse, J.A. : **A Kolmogorov -Smirnov Metric for Decision Tree Induction**, Amherst, MA, University of Massachusetts, Department of Computer Science, 1996.
- Valiant, L. : “A Theory of the Learnable”, **Communications of the ACM**, vol.27, 1984, pp.1134-1142.
- Vapnik, V.N. : Estimation of Dependences Based on Empirical Data, New York, NY, Springer –Verlag, 1982.
- Vapnik, V.N. : **Statistical Learning Theory**, New York, NY, Wiley, 1998.
- Vapnik, V.N. : The Nature of Statistical Learning Theory, New York, Springer–Verlag, 1995.
- Vapnik, V.N., Chervonenkis A.Y : Teoriya raspoznavaniya obrazov: Statisticheskie problemy obucheniya. (Russian), **Theory of pattern recognition: Statistical problems of learning**, Moscow, Nauka, 1974.

- Vapnik, V.N., : “On the Uniform Convergence of Relative Frequencies of Events to
Chervonenkis Their Probabilities”, **Theory of Probability and Its Applications**,
A.Y. vol.16, 1971, pp.264-280.
- Vercellis, Carlo : Business Intelligence: Data Mining and Optimization for Decision
Making, West Sussex, U.K, John Wiley & Sons Ltd., 2009.
- Walpole, R.E. , : Probability and Statistics for Engineers and Scientists, Upper Saddle
K. Ye, R.H. Rivers, NJ, Prentice Hall, 2002.
Myers, S.L.
Myers
- Wegman, E.J., : “Statistical Data Mining”, **Handbook of Statistics 24: Data Mining
and Data Visualization**, Ed. by C.R. Rao, E.J.Wegman, J.L.Solka,
Jeffrey L. Elsevier B.V, 2005, pp.1-46.
Solka
- Weiss, H. : “Deconstructing job satisfaction: separating evaluations, beliefs and
affective experiences”, **Human Resource Management Review**,
Vol. 12, 2002pp. 173-94.
- Wikipedia : “Empirical risk minimization”,
(Çevrimiçi)http://en.wikipedia.org/wiki/Empirical_risk_minimization#cite_ref-0, 2011
- William W. : “Knowledge Workers” (Çevrimiçi)
Prince <http://www.referenceforbusiness.com/management/Int-Loc/Knowledge-Workers.html>, 2011
- Witten, I.H., : Data Mining: Practical Machine Learning Tools and Techniques, San
Frank E. Fransisco, CA, Elsevier Inc., 2005.
- Wolpert, D. : “The relationship between PAC, the statistical physics framework, the
(1995). Bayesian framework, and the VC framework”, **The Mathematics of
Generalization, The SFI Studies in the Sciences of Complexity**, Ed.
Wolpert, D.H.: By D.H. Wolpert, 1995, Addison Wesley, pp.117-214.
- Xue, H., S. : “Structural support vector machine”, **The Fifth International
Chen, Q. Yang Symposium on Neural Networks**, Part I, LNCS5263, 2008.

- Yee, R. W.Y., : “The impact of employee satisfaction on quality and profitability in
A. C. Yeung, high-contact service industries”, **Journal of Operations**
Cheng **Management**, vol 26, 2008, pp.651-668.
- Yeung, D.S., : “Structured large margin machines: Sensitive to data distributions”,
Wang D., v.d. **Machine Learning**, vol.68, 2007, pp.171-200.
- Wickramasingh : “Predictors of job satisfaction among IT graduates in offshore
e, V. outsourced IT firms”, **Personnel Review** , vol.38 no.4, 2009, pp. 413-431.
- Wiley, J.W. : “Customer satisfaction: a supportive work environment and its
financial costs”, **Human Resource Planning**, vol.14, 1991, pp.117-127.
- Zaim H., S. : "The Impacts of Tacit Knowledge on Organizations' Performance",
Zaim 6'th International Congress on Knowledge, Management and
Economy, İstanbul, Turkey, 2007, pp.94-102.
- Zaim, H., E. : “Performance of Knowledge Management Practices : a Casual
Tatoglu and S. Analysis", **Journal of Knowledge Management**, Vol. 11, No. 6,
Zaim 2007, pp. 54-67.
- Zeffane, R., M. : “Exploring the differential impact of job satisfaction on employee
Ibrahim, R. attendance and conduct”, **Employee Relations**, 2007, pp. 237-250.
ElMehairi

EKLER

EK-1: SPSS-CLEMENTINE v.12 Analiz Özet Bilgileri

Finansal Kurum Performans Algısı için İkili sınıflandırıcı modellerin özet bilgileri

Fields

Target
CLSTR_FIN
Inputs
Faktör_1
Faktör_2
Faktör_3
Faktör_4
Faktör_5

Build Settings

Use partitioned data: true
Partition: Partition
Rank models by: Overall accuracy
Rank models using: Test partition
Maximum models listed in summary report: 20
Percentile to use for lift calculation: 30
Fixed Cost: 5.0
Fixed Revenue: 10.0
Fixed Weight: 1.0
Use misclassification costs: false

Model Details

SVM 1

Analysis
Fields
Target
CLSTR_FIN
Inputs
Faktör_1
Faktör_2
Faktör_3
Faktör_4
Faktör_5

Build Settings

Use partitioned data: true
Partition: Partition
_variable_importance.LABEL: true
Calculate raw propensity scores: false
Calculate adjusted propensity scores: false
Mode: Simple
Append all probabilities (valid only for categorical targets): false
Stopping criteria: 1.0E-3
Kernel type: RBF
Regularization parameter (C): 10
Regression precision (epsilon): 0.1
RBF gamma: 0.1
Gamma: 1.0
Bias: 0.0
Degree: 3

Training Summary

Algorithm: SVM
Model type: Classification

C5 1

Analysis

Tree depth: 2

Fields

Target
CLSTR_FIN

Inputs
Faktör_4
Faktör_3

Build Settings

Use partitioned data: true
Partition: Partition
Calculate variable importance: true
Calculate raw propensity scores: false
Calculate adjusted propensity scores: false
Use weight: false
Output type: Decision tree
Group symbolics: false
Use boosting: false
Cross-validate: false
Mode: Simple
Favor: Accuracy
Expected noise (%): 0

Use misclassification costs: false
Training Summary
Algorithm: C5
Model type: Classification

C&R Tree 1

Analysis
Tree depth: 5
Fields
Target
CLSTR_FIN
Inputs
Faktör_1
Faktör_2
Faktör_3
Faktör_4
Faktör_5
Build Settings
Use partitioned data: true
Partition: Partition
Calculate variable importance: true
Calculate raw propensity scores: false
Calculate adjusted propensity scores: false
Use frequency: false
Use weight: false
Levels below root: 5
Mode: Simple
Use misclassification costs: false
Training Summary
Algorithm: C&R Tree
Model type: Classification

QUEST 1

Analysis
Tree depth: 3
Fields
Target
CLSTR_FIN
Inputs
Faktör_1
Faktör_2
Faktör_3

Faktor_4
 Faktor_5

Build Settings

- Use partitioned data: true
- Partition: Partition
- Calculate variable importance: true
- Calculate raw propensity scores: false
- Calculate adjusted propensity scores: false
- Use frequency: false
- Levels below root: 5
- Mode: Simple
- Use misclassification costs: false

Training Summary

- Algorithm: QUEST
- Model type: Classification

CHAID 1

Analysis

- Tree depth: 3

Fields

- Target
 - CLSTR_FIN
- Inputs
 - Faktor_2
 - Faktor_5

Build Settings

- Use partitioned data: true
- Partition: Partition
- Calculate variable importance: true
- Calculate raw propensity scores: false
- Calculate adjusted propensity scores: false
- Use frequency: false
- Use weight: false
- Levels below root: 5
- Mode: Simple
- Use misclassification costs: false

Training Summary

- Algorithm: CHAID
- Model type: Classification

EK-2: ANKET SORULARI

ÇALIŞAN MEMNUNİYETİ

Sayın Katılımcı, Bu anket, İstanbul Üniversitesi İşletme Fakültesi, Sayısal Yöntemler Anabilim Dalı doktora tez çalışmasına veri sağlamak üzere hazırlanmıştır. Ankette yer alan ifadelere vereceğiniz yanıtlarla, çeşitli konulara ilişkin görüşlerinizi aktararak bunların incelenmesine katkıda bulunacaksınız. Araştırma için kimliğinizin belirlenmesi önemli olmadığından, yanıtlamış olduğunuz anket araştırmacı tarafından bilimsel amaçla değerlendirilecektir. Katılımınız, ilgi ve yardımlarınız için şimdiden çok teşekkür ediyorum.

BİRİNCİ BÖLÜM: DEMOGRAFİK BİLGİLER

Doğum yılınız? *

Cinsiyetiniz? *

☐

Kadın

☐

Erkek

Yaşadığınız Şehir? *

Medeni Durumunuz? *

☐

Evli

☐

Bekar

Eğitim düzeyiniz: *

☐

Önlisans

☐

Lisans

☐

Yüksek Lisans

☐

Doktora

Çalıştığınız Sektör? *

☐

Özel

☐

Kamu

Çalıştığınız sektör türü: *

- ☐ Finans
☐ Sağlık
☐ İmalat
☐ Eğitim
☐ Bilişim
☐ Diğer:

Toplam çalışma yılınız: *

- ☐ 0-5 yıl
☐ 6-10 yıl
☐ 11-15 yıl
☐ 16-20 yıl
☐ 20 yıl ve üstü

Halen görev yaptığınız kurumdaki toplam çalışma yılınız: *

- ☐ 0- 5 yıl
☐ 6-10 yıl
☐ 11-15 yıl
☐ 16-20 yıl
☐ 20 yıl ve üstü

Kaç yıldır mesleğinizi yapıyorsunuz?: *

- ☐ 0- 5 yıl
☐ 6-10 yıl
☐ 11-15 yıl
☐ 16-20 yıl
☐ 20 yıl ve üstü

Çalıştığınız göreve isteyerek mi geldiniz? *

- ☐ Hayır
☐ Evet
☐ Kısmen

Mesleğinizle ilgili okul dışında mesleki eğitim aldınız mı? *

- ☐ Hayır
☐ Evet

Pozisyonunuz: *

- ☐ Üst düzey yönetici
☐ Orta düzey yönetici
☐ Operasyon yöneticisi
☐ Uzman
☐ Memur

İKİNCİ BÖLÜM: ÇALIŞAN MEMNUNİYETİ SORULARI

Bu bölümde yer alan 34 madde çalışan memnuniyeti ile ilgilidir. Burada sorular 5 ölçek üzerinden hazırlanmıştır. Bu ölçeklerde 5 ile “kesinlikle katılıyorum” ve 1 ile ise “kesinlikle karşı fikirdeyim” ifade edilmektedir. 1 ile 5 arasındaki rakamlar ise bu iki ifade arasında bir dereceyi temsil etmektedir. Bu değerlendirmede doğru veya yanlış bir cevap bulunmamakta olup tamamen sizin memnuniyet düzeyiniz belirlenmeye çalışılmaktadır. 1 kesinlikle hayır 2 sanırım hayır 3 belki 4 sanırım evet 5 kesinlikle evet

1)Çalıştığım kurumun bana değer verdiğini düşünüyorum. *

1	2	3	4	5
☐	☐	☐	☐	☐

2)Çalıştığım kurumdaki görevimi ve sağladığım katma değeri biliyorum. *

1	2	3	4	5
☐	☐	☐	☐	☐

3)Çalıştığım kurumda fikirlerim ve görüşlerim dikkate alınıyor ve onlara değer veriliyor. *

1	2	3	4	5
☐	☐	☐	☐	☐

4)İş güvencesine sahip olduğumu düşünüyorum. *

1	2	3	4	5
☐	☐	☐	☐	☐

5) Haklı bir gerekçe olmadıkça işime son verilmeyeceğini düşünüyorum. *

1	2	3	4	5
☐	☐	☐	☐	☐

6) Kurumda yükselme ve terfilerin adil biçimde gerçekleştiğini düşünüyorum. *

1	2	3	4	5
☐	☐	☐	☐	☐

7) Çalıştığım kurumda yükselme fırsatımın olduğuna inanıyorum. *

1	2	3	4	5
☐	☐	☐	☐	☐

8) Yöneticilerim, işimi daha iyi yapabilmem için bana yardımcı oluyorlar. *

1	2	3	4	5
☐	☐	☐	☐	☐

9) İşimi düzgün yapabilmem için gerekli yetkiye sahibim. *

1	2	3	4	5
☐	☒	☐	☐	☐

10) Kuruluşum, işimi düzgün yapabilmem için gerekli eğitim imkânlarını bana sağlamıştır. *

1	2	3	4	5
☐	☐	☐	☐	☐

11) İşimi düzgün yapabilmem için gerekli geri beslemeyi yöneticilerimden alabiliyorum (olumlu/olumsuz uyarılar ve teşvikler) *

1	2	3	4	5
☐	☐	☐	☐	☐

12) Çalıştığım kurumdan ayrılmayı düşünmüyorum. *

1	2	3	4	5
☐	☐	☐	☐	☐

13) İş arkadaşlarımla kolektif çalışma yapabiliyoruz. *

1	2	3	4	5

☐	☐	☐	☐	☐

14) İş arkadaşlarımla aramızdaki işbirliği ve yardımlaşma iyi düzeydedir. *

1	2	3	4	5

☐	☐	☐	☐	☐

15) İş arkadaşlarım takım çalışmasına yatkındır. *

1	2	3	4	5

☐	☐	☐	☐	☐

16) İş arkadaşlarımla müşterek kararlar alabiliyoruz. *

1	2	3	4	5

☐	☐	☐	☐	☐

17) Yöneticilere fikirlerimizi korkmadan söyleyebiliyoruz. *

1	2	3	4	5

☐	☐	☐	☐	☐

18) Yöneticiler hata yaptığımızda bizi tehdit etmeden ikaz ediyorlar. *

1	2	3	4	5

☐	☐	☐	☐	☐

19) Genel olarak yöneticilerimize saygı duyuyoruz. *

1	2	3	4	5

☐	☐	☐	☐	☐

20) Çalışanlar arasında güçlü bir iletişim sağlanmıştır. *

1	2	3	4	5

☐	☐	☐	☐	☐

21) İhtiyaç duyduğumuz kişiye (yöneticiler dahil) istediğimiz zaman ulaşabiliyoruz. *

1	2	3	4	5

☐	☐	☐	☐	☐

22) Ücretlerden memnunum. *

1	2	3	4	5

☐ ☐ ☐ ☐ ☐

23) Ücret politikasının adil olduğunu düşünüyorum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

24) Ücret dışı ödüllendirmelerden memnunum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

25) Genel olarak işimden memnunum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

26) Çalıştığım kurumdan memnunum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

27) Çalıştığım işyeri ortamının fiziki olarak çalışmaya uygun olduğunu düşünüyorum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

28) Çalıştığım işyeri ortamının sosyal açıdan (ilişkiler, kurum kültürü vs.) çalışmaya uygun olduğunu düşünüyorum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

29) Daha iyi çalışma konusunda yöneticilerim tarafından teşvik edildiğimi düşünüyorum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

30) Daha iyi çalışmam konusunda iş arkadaşlarım tarafından teşvik edildiğimi düşünüyorum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

31) Genel olarak ücret dışındaki prim ve ikramiyelerden memnunum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

32) Kurumun bana sağladığı sosyal imkanlardan memnunum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

33) Kurumun bana sağladığı sosyal yardımlardan (eğitim, sağlık, kira vs.) memnunum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

34) Çalıştığım kurumu diğer arkadaşlarıma da tavsiye ediyorum. *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

ÜÇÜNCÜ BÖLÜM: KURUMSAL PERFORMANS ALGISI

Bu bölüm Kurumunuzun genel performansı ile ilgili hazırlanmıştır. Burada sorular 5 ölçek üzerinden hazırlanmıştır. Bu ölçeklerde 5 ile “tamamen aynı fikirdeyim” ve 1 ile ise “kesinlikle karşı fikirdeyim” ifade edilmektedir. 1 ile 5 arasındaki rakamlar ise bu iki ifade arasında bir dereceyi temsil etmektedir. Bu değerlendirmede de doğru veya yanlış bir cevap bulunmamaktadır. Bu değerlendirmede tamamen sizin görev yapmakta olduğunuz kurum hakkındaki düşünceniz ve bu kurumun performansını nasıl algıladığınızı öğrenilmeye çalışılmaktadır.

1) Bence çalıştığım kurum genel olarak rakiplerine göre daha başarılıdır *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

2) Bence çalıştığım kurumun geliri rakiplerine göre daha yüksektir *

1 2 3 4 5

☐ ☐ ☐ ☐ ☐

3) Bence çalıştığım kurum rakiplerine göre daha hızlı büyümektedir *

1 2 3 4 5

☐	☐	☐	☐	☐

4) Bence çalıştığım kurum rakiplerine göre daha kârlıdır *

1	2	3	4	5

☐	☐	☐	☒	☐

5) Bence çalıştığım kurum rakiplerine göre daha yenilikçidir *

1	2	3	4	5

☐	☐	☐	☐	☐

6) Bence çalıştığım kurum rakiplerine göre daha kaliteli ürün ve hizmetler üretmektedir (sunmaktadır) *

1	2	3	4	5

☐	☐	☐	☐	☐

7) Bence çalıştığım kurum rakiplerine göre daha rekabetçi bir yapıya sahiptir *

1	2	3	4	5

☐	☐	☐	☐	☐

8) Bence çalıştığım kurumun yatırımlarının geri dönmesi rakiplerine göre daha yüksektir *

1	2	3	4	5

☐	☐	☐	☐	☐

ÖZGEÇMİŞ

Cemil Kuzey 01.02.1974 tarihinde Ardahan'da dünyaya geldi. İlkokul ile Ortaokul 1. ve 2. sınıfı Ardahan'da tamamladıktan sonra, Ortaokul son sınıf ve lise öğrenimini İstanbul Esenyurt Lisesinde 1992'de tamamlamıştır. Ondokuz Mayıs Üniversitesi Eğitim Fakültesi, Matematik bölümünden 1996'da mezun olmuştur. Daha sonra, Fatih Üniversitesi Fen Edebiyat Fakültesinde Uygulamalı Matematik alanında Yüksek Lisans programından 1999 yılında başarı ile mezun olmuştur. Yüksek lisans sonrası, 1999-2000 Eğitim Öğretim döneminde, İstanbul Üniversitesi İşletme fakültesi sayısal yöntemler anabilim dalında doktora programına başlayarak, doktora derslerini tamamladı. 2000-2008 yılları arasında Kanada'nın, Toronto şehrinde yaşadı. Bu dönemde, Lise düzeyinde genel matematik ve fen bilgisi derslerini bazı özel okullarda ve Toronto District School Board'a ait okullarda vermiştir. 2007-2008 döneminde Parkdale Collegiate Institute okulunda matematik dersleri vermiştir. Tekrar, 2008 eğitim öğretim döneminde, İstanbul Üniversitesi İşletme Fakültesi Sayısal Yöntemlere anabilim dalında doktora programına başlayarak akademik hayatında devam etmektedir.