

T.C
MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE REGRESYON AĞAÇLARI İLE
SINIFLANDIRMA VE BİR UYGULAMA

DOKTORA TEZİ
Gülser DONDURMACI

İstatistik Anabilim Dalı
İstatistik Programı

Tez Danışmanı: Prof. Dr. Nalan CİNEMRE

NİSAN 2011

ÖZET

Bilgisayar teknolojisindeki gelişmelerle birlikte üretilen hem sayısal hem de sayısal olmayan bilgi miktarının arttığı, veri tabanlarının daha fazla veriyi saklayabilecek boyutlara ulaştığı, ayrıca veriye ulaşmanın giderek kolaylaştığı görülmektedir. Veri tabanı sistemlerinin artan kullanımı ve hacimlerindeki olağanüstü artış, ister istemez elde edilen bu büyük hacimli verilerden organizasyonların nasıl yararlanacağı konusunu gündeme getirmiştir. Büyük hacimli veri kümelerinden değerli olan bilginin ve gizli örüntülerin ortaya çıkarılması olayına “Veri Madenciliği” adı verilmektedir.

Bilgisayarların gücündeki artış ve fiyatlarının düşmesi, veri madenciliği kapsamındaki tekniklerin de çeşitlenmesini sağlamıştır. Veri madenciliğinin amaçları genellikle sınıflandırma, kümeleme, tahmin öngörü ve benzer gruplama olarak sıralanmaktadır. Amaçlardan biri olan kümeleme, istatistiksel veri analizi, örüntü tanıma vb, birçok alanda oldukça sık kullanılmaktadır. Veritabanlarındaki verilerin gruplar veya kümeler altında toplanarak, benzer özelliklere sahip nesnelerin bir araya gelmesini sağlayan sınıflandırma algoritmaları veri madenciliği alanında büyük öneme sahiptir. Bu tekniklerden biri de ağaç tabanlı sınıflandırma yöntemleri arasında yer alan “Regresyon Ağaçları”dır. Regresyon ağaçları, sayısal değerlerin kullanılabildiği ve sonuç olarak karar vericilere karar vermelerinde kullanabilecekleri karar kurallarının üretilbildiği bir veri madenciliği yöntemidir.

Bu çalışma üç bölümden oluşmuştur. Birinci bölümde veri madenciliği süreci, verilerin hazırlanması, nitelik seçimi, sınıflandırma, modelin değerlendirilmesi konuları açıklanmıştır. İkinci bölümde öğrenme kavramı, karar ağaçlarının elde edilme süreci, karar kuralları ve karar ağaçlarında entropiye dayalı bölünme, regresyon ağaçları ile sınıflandırma esasları ele alınmıştır. Üçüncü bölümde IMKB 30 grubuna dahil menkul kıymetlere ilişkin günlük kapanış fiyatları, menkul kıymet teknik analizlerinde yaygın biçimde kullanılan teknik göstergeler, altın fiyatlarındaki değişimler, dolar kurundaki değişimler ve bazı yurtdışı borsa göstergeleri göz önüne alınarak karar ağaçlarının oluşturulması ve bu ağaçlara dayalı olarak karar kurallarının elde edilmesi sağlanmıştır.

Analize dahil edilen niteliklerin sayısını azaltmak amacıyla “Çoklu Doğrusal Regresyon Modelleri” düzenlenmiş, çok sayıdaki modele giren değişkenlerden istatistiksel bakımdan anlamlı olanlar dikkate alınarak bu değişkenleri içeren regresyon ağaçları oluşturulmuştur. Karar kuralları da bu regresyon ağaçlarının yorumlanmasıyla oluşturulmuştur. Bu deneysel işlemlerin sonucunda, bazı menkul kıymetlere ait regresyon modellerinden elde edilen niteliklerin seçilmesiyle daha düşük ortalama hataya sahip regresyon ağaçlarının elde edilebileceği anlaşılmıştır.

SUMMARY

It is seen that the amount of digital information produced has increased, the databases reached heights to store more data and it became easier reach data with advances in computer technology. Increased use of data base systems and an extraordinary increase in volumes raised the issue of how organizations benefit from the data obtained. The revelation of knowledge and hidden patterns in large-volume datasets are called "Data Mining". The increase in the power of computers and decrease in prices have also diversified the techniques within the scope of data mining. One of these techniques is "Regression Trees" that is among the tree-based classification methods. Regression trees are a data mining method where numerical values can be used and decision rules are produced as a result.

This study consisted of three parts. In the first part, the mining process, data preparation, quality selection, classification, assessment of model has been described. The second chapter discussed the concept of learning, the process of acquisition of decision trees, decision rules and the entropy-based division of decision trees, regression trees and classification principles. In the third chapter, the creation of decision trees and provision of decision rules based on these trees have been provided considering ISE, the daily market prices regarding securities included in ISE 30 group, technical indicators widely used in technical analysis of securities, changes in gold prices, changes in dollar and some foreign stock market indicators.

In order to reduce the number of attributes included in the analysis, Multiple Linear Regression Analysis models were found, regression trees including these variables considering the variables in the model and decision rules were created as a result. As a result of these experimental procedures, it has been understood that regression trees with less average error in the selection of variables obtained from the Multiple Linear Regression model derived from securities were revealed.

TEŞEKKÜR

Değerli katkı ve yardımlarından dolayı danışmanım Sayın Prof. Dr. Nalan CİNEMRE'ye, desteğini benden esirgemeyen eşim Adnan DONDURMACI'ya, katkılarından dolayı kızım Dilay DONDURMACI'ya, Deniz BİLMİŞ ÇUTUK'a ve IMKB verilerini elde etmem konusunda gösterdiği yardımlardan dolayı IMKB Eğitim Müdürü Dr. Murat KAYACAN' a sonsuz teşekkür ederim.

Nisan 2011

Gülser Dondurmacı

İÇİNDEKİLER

ÖZET	i
SUMMARY	iii
TEŞEKKÜR	iv
İÇİNDEKİLER	v
ŞEKİL LİSTESİ	ix
TABLO LİSTESİ	xi
KISALTMALAR LİSTESİ	xii
BÖLÜM I : SINIFLANDIRMA SÜRECİ.....	1
1.1. GİRİŞ	1
1.2. VERİ MADENCİLİĞİ SÜRECİ	3
1.2.1. VERİ TABANLARINDA BİLGİ KEŞFİ.....	3
1.2.2. VERİ MADENCİLİĞİ SÜRECİ.....	5
1.3. VERİ HAZIRLAMA.....	6
1.3.1. VERİ TEMİZLEME	7
1.3.2. VERİ BÜTÜNLEŞTİRME	8
1.3.3. VERİ İNDİRGEME	9
1.3.4. VERİ DÖNÜŞTÜRME.....	11
1.3.4.1. Ondalık Ölçekleme.....	11
1.3.4.2. Min-Max Normalizyonu	12
1.3.4.3. Z-score Standartlaştırma	12
1.5. VERİ MADENCİLİĞİ YÖNTEMİNİN BELİRLENMESİ	13
1.5.1. SINIFLANDIRMA.....	14
1.5.2. KÜMELEME	14
1.5.3. BİRLİKTELİK KURALLARI	14
1.6. SONUÇLARIN SUNUM VE DEĞERLENDİRİLMESİ	15
1.7. NİTELİK SEÇİMİ	16
1.7.1. ENTROPİ TABANLI NİTELİK SEÇİMİ.....	16
1.7.2. TEMEL BİLEŞENLER ANALİZİ	18
1.7.3. DEĞİŞKENLERİN STANDARDİZE EDİLMESİ.....	19

1.7.3.1. Korelasyon Matrisi ile Temel Bileşenlerin Bulunması	20
1.7.3.2. Temel Bileşenlerin Sayısını Belirlemek.....	21
1.8. VERİ MADENCİLİĞİNDE SINIFLANDIRMA	23
1.9. SINIFLANDIRMA MODELLERİ	25
1.9.1. KARAR AĞAÇLARI İLE SINIFLANDIRMA.....	26
1.9.2. BAYES SINIFLANDIRMA YÖNTEMİ.....	28
1.9.3. DESTEK VEKTÖR MAKİNALARI İLE SINIFLANDIRMA	31
1.9.3.1. Doğrusal Destek Vektör Makineleri	31
1.9.3.2. Doğrusal Olmayan Sınıflandırıcılar	35
1.10. MODEL PERFORMANSININ ROC ANALİZLERİ İLE	36
DEĞERLENDİRİLMESİ	
1.10.1. KONTENJANS TABLOSU	36
1.10.2. ROC EĞRİLERİ	37
BÖLÜM II: KARAR AĞAÇLARI VE SINIFLANDIRMA.....	40
2.1. ÖĞRENME.....	40
2.1.1. MAKİNE ÖĞRENMESİ.....	40
2.1.2. ÖĞRETMENLİ ÖĞRENME YÖNTEMLERİ.....	43
2.2. KARAR AĞAÇLARININ ELDE EDİLMESİ SÜRECİ	46
2.2.1. ÖRNEKLERİN KULLANIMI	48
2.2.2. ÖĞRENME SÜRECİ	49
2.2.3. KARAR AĞAÇLARININ ELDE EDİLMESİ	51
2.2.4. KARAR AĞAÇLARININ AVANTAJLARI VE DEZAVANTAJLARI	53
2.3. KARAR KURALLARI VE SINIFLANDIRICI DOĞRULUĞU	54
2.3.1. KARAR KURALLARININ BELİRLENMESİ.....	54
2.3.2. KURALLARIN GEÇERLİLİĞİNİ DOĞRULAMA	55
2.3.2.1. Sınıflandırıcı Doğruluğu.....	56
2.3.2.2. Doğruluk Değeri.....	56
2.3.2.3. Doğruluğun Artırılması	62
2.3.3. KURALLARIN UYGULANMASI VE TAHMİN	66
2.4. KARAR AĞAÇLARINDA ENTROPİYE DAYALI BÖLÜNME	67
2.4.1. ENTROPİ	67
2.4.2. KAZANÇ ÖLÇÜTÜ VEYA ID3 ALGORİTMASI	69

2.4.3. KAZANÇ ORANI VEYA C4.5 ALGORİTMASI.....	71
2.5. SINIFLANDIRMA VE REGRESYON AĞAÇLARI	73
2.5.1. TWOİNG ALGORİTMASI	73
2.5.2. GİNİ ALGORİTMASI.....	75
2.5.3. REGRESYON AĞAÇLARI	76
2.6. ÇOKLU DOĞRUSAL REGRESYON MODELİ.....	76
2.6.1. BASİT REGRESYON MODELİ	76
2.6.2. ÇOKLU DOĞRUSAL REGRESYON.....	79
2.6.3. TEMEL VARSAYIMLAR	80
2.6.4. EN KÜÇÜK KARELER TAHMİNLERİ.....	80
2.6.5. ORTALAMA VE VARYANS.....	81
2.6.6. STANDART HATA	83
2.6.7. KORELASYON KATSAYISI:	84
2.7. REGRESYON AĞAÇLARI İLE SINIFLANDIRMA	85
2.7.1. REGRESYON AĞAÇLARI VE EN KÜÇÜK KARELER REGRESYONU	85
2.7.2. HATA ÖLÇÜMLERİ VE HATA TAHMİNLERİ.....	88
2.7.3. AĞAÇ YAPILI REGRESYON	89
BÖLÜM III: İSTANBUL MENKUL KIYMETLER BORSASI İÇİN UYGULAMA.....	94
3.1. GİRİŞ	94
3.2. VERİ KÜMESİ.....	95
3.3. NİTELİKLER	96
3.3.1. GÜNLÜK KAPANIŞ FİYATLARI	96
3.3.2. TEKNİK ANALİZ GÖSTERGELERİ.....	97
3.3.2.1. Göreceli Güç İndeksi	97
3.3.2.2. Stokastik Osilator.....	98
3.3.2.3. MACD.....	99
3.3.2.4. Momentum	100
3.3.2.5. CCI	100
3.3.3. DİĞER NİTELİKLER	101
3.4. İMKB 30 MENKUL KIYMETLERİ İÇİN REGRESYON AĞACI MODELİ.....	102
3.4.1. EĞİTİM VE TEST KÜMESİNİN BELİRLENMESİ	102
3.4.2. REGRESYON AĞACI MODELLERİ	102

3.4.3. ÇOKLU DOĞRUSAL REGRESYON MODELLERİ.....	103
3.4.4.ÇDR MODELİNDE YER ALAN NİTELİKLERİN REGRESYON AĞAÇLARINDA KULLANIMI.....	104
3.4.5. KARAR KURALLARI.....	107
3.5. SONUÇLAR VE DEĞERLENDİRME.....	108
EKLER.....	110
EK 1: İMKB 30 İÇİN ÇDR MODELLERİ	110
EK-2: İMKB 30 İÇİN REGRESYON AĞAÇLARI.....	117
EK 3: İMKB 30 İÇİN KARAR KURALLARI.....	132
KAYNAKLAR.....	147
ÖZGEÇMİŞ.....	152

ŞEKİL LİSTESİ

BÖLÜM I

Şekil 1.1. Veri Tabanlarında Bilgi Keşfi Süreci.	4
Şekil 1.2. Veri Madenciliği Süreci.	6
Şekil 1.3. Veri Bütünleştirme.	9
Şekil 1.4. Veri İndirgeme.	11
Şekil 1.5. Kümeleme Sonuçlarını Gösteren Bir Dendrogram.	16
Şekil 1.6. Eğim Grafiği	23
Şekil 1.7. Eğitim Verileri İle Bir Veri Madenciliği Algoritmasının Kullanılması.	24
Şekil 1.8. Modelin Uygulanması.	25
Şekil 1.9. Karar Ağacı	27
Şekil 1.10. Bir Bayes Ağı.....	31
Şekil 1.11. İki Boyutlu Uzayda Verilerin Doğrusal İki Sınıfa Ayrılması.....	32
Şekil 1.12. Sınıflandırıcı Çoklu Düzlemin Elde Edilişi.....	33
Şekil 1.13. Destek Vektörler.	34
Şekil 1.14. Roc Uzayı.....	38
Şekil 1.15 Roc Eğrisinin Görünümü.	39

BÖLÜM II

Şekil 2.1 Öğretmensiz Öğrenme.	41
Şekil 2.2 Öğretmenli Öğrenme.....	42
Şekil 2.3. Sınıflandırma Ağaçlarında Öğrenme Süreci.	50
Şekil 2.4. Bir Karar Ağacının Görünümü.....	52
Şekil 2.5. Karar Kurallarının Elde Edilme Süreci.	525
Şekil 2.6. Holdout Yöntemi İle Doğruluğun Tahmin Edilmesi.....	58
Şekil 2.7. K-Katlı Çapraz Doğrulama Eğitim Ve Test Kümelerinin Belirlenmesi.....	59
Şekil 2.8. Bootstrap Yönteminde Eğitim Ve Test Kumesinin Seçilmesi İşlemi.	61
Şekil 2.9. Bagging Yöntemi İle Kestirim.	63
Şekil 2.10. Karar Kurallarının Yeni Veriye Uygulanması.	67
Şekil 2.11. Tahmin Edilen Doğru İçin Kalıntılar.	78
Şekil 2.12. Karar Ağacında İkili Bölünmeler.	89
Şekil 2.13. Regresyon Yüzeyinin Histogram Biçimindeki Görünümü.	91

EKLER

Şekil E.1. Aefes İçin Regresyon Ağacı.	117
Şekil E.2. Akbnk İçin Regresyon Ağacı.....	117

Şekil E.3. Akgrt İçin Regresyon Ağacı	118
Şekil E.4. Asyab İçin Regresyon Ağacı	118
Şekil E.5. Bagfs İçin Regresyon Ağacı	119
Şekil E.6. Dohol İçin Regresyon Ağacı	119
Şekil E.7. Dyhol İçin Regresyon Ağacı	120
Şekil E.9. Ereql İçin Regresyon Ağacı	121
Şekil E.10. Garan İçin Regresyon Ağacı	121
Şekil E.11. Halkb İçin Regresyon Ağacı	122
Şekil E.12. Isgyo İçin Regresyon Ağacı	122
Şekil E.13. Isbtr İçin Regresyon Ağacı	123
Şekil E.15. Krdmd İçin Regresyon Ağacı	124
Şekil E.16. Sahol İçin Regresyon Ağacı	124
Şekil E.17. Skbnk İçin Regresyon Ağacı	125
Şekil E.18. Sıse İçin Regresyon Ağacı	125
Şekil E.19. Petkm İçin Regresyon Ağacı	126
Şekil E.20. Tavhl İçin Regresyon Ağacı	126
Şekil E.23. Tkfen İçin Regresyon Ağacı	128
Şekil E.24. Toaso İçin Regresyon Ağacı	128
Şekil E.25. Tskb İçin Regresyon Ağacı	129
Şekil E.26. Ttkom İçin Regresyon Ağacı	129
Şekil E.27. Tuprps İçin Regresyon Ağacı	130
Şekil E.28. Tcell İçin Regresyon Ağacı	130
Şekil E.29. Vkbmk İçin Regresyon Ağacı	131

TABLO LİSTESİ

BÖLÜM I

Tablo 1.1. Varyans Ve Birikimli Varyans Değerleri.....	22
Tablo 1.2 Kontenjans Tablosu.	36

BÖLÜM II

Tablo 2.1. Sınıflandırma Ağacına Esas Oluşturacak Veri.....	49
Tablo 2.2. Karar Ağacı Bileşenleri.....	52

BÖLÜM III

Tablo 3.1. İmkb 30 Grubuna Dahil Şirketler.....	94
Tablo 3.2. İmkb 30 Verileri Tarih Aralığı.....	95
Tablo 3.3. Araştırmada Kullanılan Diğer Nitelikler.....	101
Tablo 3.4. Regresyon Ağacı Modellerinin Ortalama Hata Ve Rms Hata Değerleri.....	103
Tablo 3.5. Çdr Modellerinin Ortalama Hata Ve Rms Hata Değerleri.....	104
Tablo 3.6. Yeni Regresyon Ağacı Modellerinin Ortalama Hata Ve Rms Hata Değerleri.....	106
Tablo 3.7. Sonuç Olarak Elde Edilen Regresyon Ağacı Modelleri.....	107

KISALTMALAR LİSTESİ

KDD	: Veritabanlarında bilgi keşfi (Knowledge Discovery in Databases)
OLAP	: Çevrimiçi Analitik İşlem (On-Line Analytical Processing)
CART	: Sınıflandırma ve regresyon ağaçları (Classification and Regression Trees)
TP	: Doğru pozitiflerin sayısı
FP	: Yanlış pozitiflerin sayısı
FN	: Yanlış negatiflerin sayısı
TN	: Doğru negatiflerin sayısı
TPR	: Doğru Pozitif Oranı
FPR	: Yanlış pozitif Oranı
ROC	: Alıcı İşlem Karakteristiği (Receiver Operating Characteric)
RSI	: Göreceli Güç Endeksi
STO	: Skotastik Osilatör
MACD	: Hareketli Ortalamaların Yakınsaması
CCI	: Emtia Kanalı Endeksi
RMS	: Kök Ortalama Kareli Hata (Root Mean Square)
ÇDR	: Çoklu Doğrusal Regresyon

BÖLÜM I : SINIFLANDIRMA SÜRECİ

1.1. GİRİŞ

Veri madenciliği daha çok, veritabanlarında bilgi keşfi (KDD-Knowledge Discovery in Databases) şeklindeki daha geniş kapsamda ele alınmaktadır. Veri madenciliği araştırıcı ve tekrarlayan bir süreçtir. Yani veri analizi sırasında, yeni bilgiler ortaya çıkabilir veya yeni hipotezler ortaya atılabilir. Bu da alternatif soruların düşünülmesine neden olabilir. Verinin altındaki bilgiyi ortaya çıkarmak için veri madenciliği, araştırmacının konu ile ilgili deneyimiyle birlikte kullanılmalıdır. Veri madenciliği süreci genellikle uygulamadaki iki temel soruna cevap bulmayı amaçlamaktadır. Bunlardan birincisi mevcut veritabanından tahminler yapmak (tahmin edici veya kestirici model), ikincisi ise veriden elde edilecek davranışları betimlemektir (tanımlayıcı bilgi).

Özetle ifade etmek gerekirse veri madenciliği, veriler arasındaki bilginin ortaya çıkarılması işlemidir. Geleneksel araştırma yöntemlerine göre en belirgin özelliği, özellikle büyük hacimli veri tabanları üzerinde yararlı olabilecek, anlamlı ve değeri olan bilgiye ulaşmaya çalışılmasıdır. Buna karşılık, klasik istatistiksel teknikler veri madenciliği yöntemlerinin temel dayanağı olmaya devam etmektedir (Olson ve Delen, 2008, s.5). Veri madenciliği için firmaların çok büyük miktarlarda veri içeren, veri tabanlarındaki kayıtları analiz edilerek, amaca uygun işlemlerle bilgi keşfi yapılmaktadır.

Veri madenciliğinin amaçları genellikle sınıflandırma, kümeleme, tahmin-öngörü ve benzer gruplama olarak sıralanmaktadır. Amaçlardan biri olan sınıflandırma istatistiksel veri analizi, örüntü tanıma vb, birçok alanda oldukça sık kullanılmaktadır. Veritabanlarındaki verilerin sınıflandırılmasında kullanılan algoritmalar çok çeşitlidir ve veri madenciliği alanında oldukça büyük öneme sahiptir. Ancak sınıflandırma algoritmalarının seçimi hem mevcut veri türüne hem de amaç ve uygulamaya bağlıdır.

Özellikle veri sayısı arttıkça aşırı hesaplamalardan dolayı performans sorunlarının ortaya çıkması kaçınılmazdır. Özellikle ilişkisel veri tabanları üzerinde bu çözümlemeleri yapmak zorlaşır. Bu tür verileri çözümleyebilmek için hem yeni veri

tabanı kavramlarına hem de yeni çözümleme yöntemlerine gereksinim duyulmaktadır. Veriyi yönetmek için “Veri Ambarı” ve verileri çözümleyerek “yararlı bilgiye” erişilmesini sağlamak için “Veri Madenciliği” kavramları ortaya atılmıştır (Özkan, 2008, s.38).

1991 yılında ilk kez William H. Inmon tarafından ortaya atılan veri ambarı, yönetimin kararlarını desteklemek amacı ile çeşitli kaynaklardan elde ettikleri bilgileri zaman degiskeni kullanarak veri toplama olarak tanımlanmaktadır. Kısaca birçok veritabanından alınarak birleştirilen verilerin toplandığı depolardır. Veri ambarlarının özelliği kullanıcılara farklı detay düzeyleri sağlayabilmesidir. Detayın en alt düzeyi arşivlenen kayıtların kendisi ile ilgili iken, daha üst düzeyler zaman gibi daha fazla bilginin toplanması ile ilgilidir. Veri ambarları ciddi yatırımlar gerektirmekte ve uygulanması bir yıl veya daha uzun zaman almaktadır (Murray ve Glick, 1999, s.260).

Veri ambarlarında bulunan verilerin en önemli özelliği, temiz ve dönüşümden geçmiş olmaları olduğu için veri ambarına kaydedilecek verinin öncelikle temizlenmesi, gerekli dönüşümlerin yapılması ve verilerin birleştirilmesi işlemlerinin gerçekleştirilmesi gerekmektedir. Böylece veri ambarında saklanan veriler kullanılmaya hazır hale getirildiğinden kullanıcılar amaçlarına uygun veriye, veri ambarlarına gerekli sorgulamaları yaparak ulaşmaktadırlar. İşte veri madenciliği bu noktada devreye girmektedir. Veri madenciliğine geçmeden önce veri ambarlarıyla ilgili önemli birkaç noktanın vurgulanması uygun olur.

Veri ambarları, verilerin üzerine yazmak ve verilerde değişiklik yapmak için değil sadece okumaya yönelik olarak oluşturulmaktadır. Bu nedenle veri ambarındaki veriler, analiz yapmayı kolaylaştıran bir formatta tutulmaktadır. Burada analiz; sorgular, raporlar, karar destek sistemleri veya istatistiki hesapları kapsamaktadır. Veri ambarında tutulan bilgi şu özelliklere sahiptir (Thuraisingham, 2000, s.55):

- Özetlenmiş operasyonel veri içerir.
- Ayı bir sistemde tutulur.
- Sadece ayıklanmış ve güvenilir veri kullanılır.
- Veriler çok boyutlu bir halde tutulur.

Veri ambarlarında 2 boyutlu operasyonel sistemlerin tersine çok boyutlu bilgi saklanabilir. Boyuttan kasıt zaman, ürün, dağıtım kanalı, marka, model, fabrika vb. olabilir. Boyutlar veri ambarının hazırlandığı işletmenin yapısına göre tayin edilir. Çok boyutlu veritabanı mantığı bir verinin her boyuta göre incelenebilmesidir.

Veri madenciliği konusunda çeşitli tanımlar yapılmaktadır. Basit bir tanım yapmak gerekirse, veri madenciliği, büyük ölçekli veriler arasından “değeri olan” bir bilgiyi elde etme işidir. Bu sayede veriler arasındaki ilişkileri ortaya koymak ve gerektiğinde ileriye yönelik kestirimlerde de bulunmak mümkündür. Bunun anlamı, veri madenciliği bir kurumda üretilen tüm verilerin belirli yöntemler kullanarak var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi su yüzüne çıkarma süreci olarak değerlendirilebilir. Bu açıdan bakıldığında, veri madenciliği işinin kurumların karar destek sistemleri için önemli bir yere sahip olabileceği söylenebilir (Özkan, 2008, s.38).

1.2. VERİ MADENCİLİĞİ SÜRECİ

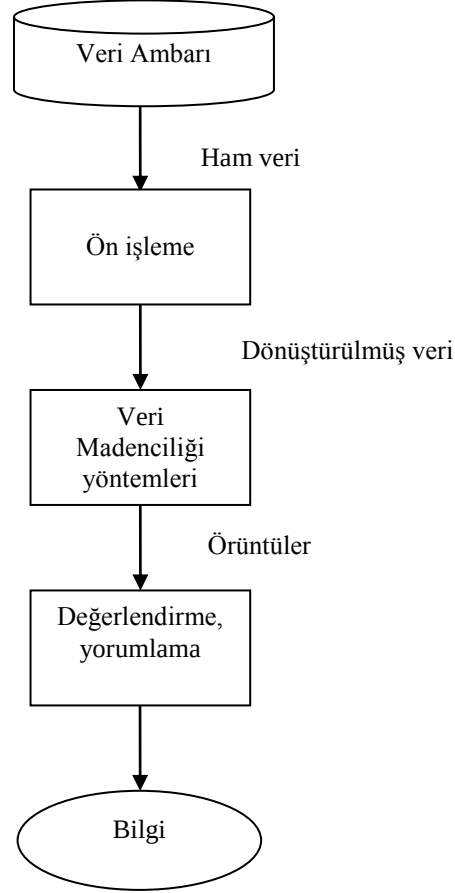
Veri madenciliği karar destek sistemlerinin kullandığı araçlardan biridir. Veri madenciliğinin tanımı farklı kaynaklardan araştırıldığında, yapılmış olan tanımların birbirine çok benzediği görülmektedir. Tüm kaynaklarda ortak olan nokta, veri madenciliğinin veri yığınları içerisinde, anlamlı ve faydalı ilişki ve modellerin keşfi süreci olduğudur. Bu işlemin “keşif” odaklı doğası sebebiyle kimi kaynaklar veri madenciliğini “veri tabanlarında bilgi keşfi süreci” olarak da adlandırmışlardır.

Veri madenciliği daha çok, veritabanlarında bilgi keşfi (KDD-Knowledge Discovery in Databases) şeklindeki daha geniş kapsamda ele alınmaktadır. Veri madenciliği sürecinde, özellikle veri ambarlarında tutulan çok çeşitli veriler kullanılarak, keşfedilmemiş, anlamlı ve yararlı bilgiler ortaya çıkartılır ve bu bilgiler karar verme işlemlerinde kullanılır.

1.2.1. Veri Tabanlarında Bilgi Keşfi

Veri tabanlarında bilgi keşfi, veri ambarlarında biriken verilere erişerek yorumlanması, verinin modellenmesi ve sonuçlar üretilmesi üzerine odaklanmıştır. Veri tabanlarında bilgi keşfi süreci, veri ambarı oluşturulması ve bu bilgilere

dayanarak veri madenciliği işlemlerinin yerine getirilmesi aşamalarını kapsar (Mitra ve Acharya, 2003, s.4). Bu süreç Şekil 1.1'deki gibi gösterilir (Mitra ve Acharya, 2003, s.5).



Şekil 1.1. Veri Tabanlarında Bilgi Keşfi Süreci.

Şekil 1.1’de açıklandığı gibi veri madenciliği veri tabanlarında bilgi keşfi süreci içinde bir adım olarak karşımıza çıkmaktadır. Veri madenciliği işlemlerini yerine getirirken çoğunlukla özel bir veri tabanından yararlanılır. Bu tür veri tabanı yapısına “veri ambarı” adı verilmektedir. Yukarıda açıklandığı gibi, veri ambarı, kurumların karar destek amacıyla kullandıkları verileri içerir. Veri ambarında verilere hızlı erişim ana amaçtır ve veri ambarı büyük miktardaki veriyi hızlı biçimde sorgulayacak araçlara sahiptir. Ayrıca kuruma ilişkin eski verileri de saklaması karar destek sistemlerinin oluşturulmasında önemli bir rol oynamaktadır.

Şekil 1.1’de görüldüğü gibi, veri tabanlarında bilgi keşfi sürecinin ikinci adımı “ön işleme” dir. Veri tabanlarındaki veya veri ambarındaki veriyi incelemeksizin doğrudan

doğruya kullanmak doğru bir yol değildir. Veri madenciliği algoritmalarının uygulanması öncesinde veri üzerinde bazı ön işlemlerin yerine getirilmesi söz konusu olabilir. Etkili bir veri madenciliğini olumsuz olarak etkileyen en büyük neden bilgi

eksikliğidir. Genellikle veritabanları veri madenciliği ile doğrudan ilgisi olmayan nedenlerden dolayı oluşturulur. Bu nedenle birçok bilgi eksik birçoğu da gereksiz olabilir. Bir diğer neden ise yanlış verilerin veritabanında yer almasıdır. Bu sorunları ortadan kaldırmak için ise veri ambarları kullanılmaktadır. Veri ambarlarındaki veriler etkili ve ölçeklenebilir veri madenciliği ile birleştğinde veriler doğru ve etkili bir şekilde kullanılmaktadır. Bu ön işlemler arasında niteliklerin seçimi, kayıp verilerin temin edilmesi, nitelik türlerinin dönüştürülmesi gibi yöntemler sayılabilir.

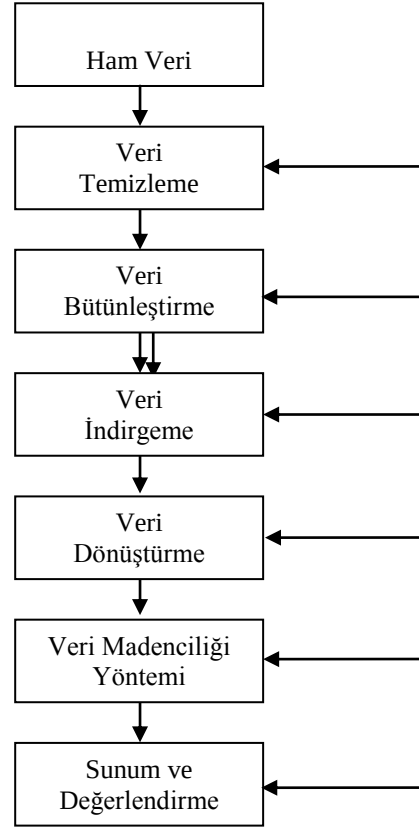
Ön işleme işlemlerinin uygulanmasının ardından, veri analiz için hazır hale getirilir. Veri madenciliği yöntemlerinin bu işlenmiş veriye uygulanmasıyla sonuçlar elde edilir. Sonuçlar analiz edilerek veri içindeki örüntüler açığa çıkarılır.

1.2.2. Veri Madenciliği Süreci

Veri hazırlama ile ilgili işlemleri veri madenciliği kavramı içinde düşünürsek, veri madenciliğinin bir süreç olarak değerlendirmesi gerekir. Veri madenciliği süreci aşağıda belirtilen adımları içermektedir (Han ve Kamber, 2000, s.7):

- Veri temizleme
- Veri bütünleştirme
- Veri indirgeme
- Veri dönüştürme
- Veri madenciliği algoritmasını uygulama
- Sonuçları sunum ve değerlendirme

Veri madenciliği sürecinin aşamaları Şekil 1.2’de ifade edilmiştir. Burada ham veri, henüz herhangi bir işleme tabi tutulmamış (ön işlem öncesi) veridir. Ham veri, sırasıyla temizleme, bütünleştirme, indirgeme ve dönüştürme gibi işlemlerden geçtikten sonra, veri madenciliği algoritması uygulanarak sonuçlar elde edilmektedir.



Şekil 1.2. Veri Madenciliği Süreci.

Eğer veri madenciliği verisi bir veri ambarından sağlanıyorsa, bu işlemlere gerek kalmayabilir. Bu tür işlemler veri ambarı hazırlanırken yerine getirilir. Eğer veri madenciliği için veri ambarı kullanılmıyorsa, veri temizleme, bütünleştirme, indirgeme ve dönüştürme aşamalarının ayrıca yerine getirilmesi gerekir.

1.3. VERİ HAZIRLAMA

Veri tabanları içindeki verinin ve bu veriye dayalı olarak elde edilen veri madenciliği sonuçlarının kalitesinin artırılması, veriyi analize hazırlarken dikkat edilmesi gereken en önemli noktadır. Veri madenciliği işlemlerini kolaylaştırmak ve verimliliği arttırmak için veri tabanındaki veriler, bir “ön işleme” aşamasından geçirilir (Han ve Kamber, 2006, s.289). Veri madenciliği öncesinde verinin hazırlanması süreci olarak da kabul edilen bu işlemler özellikle veri tabanındaki bozuk değerleri ve veriler arasındaki tutarsızlıkları kaldırmayı amaçlamaktadır.

1.3.1. Veri Temizleme

Veri temizleme aşaması, eksik değerleri tamamlamak, aykırı değerleri belirleyerek gürültüyü azaltmak ve verilerdeki tutarsızlıkları gidermek için kullanılan birçok tekniği kapsamaktadır.

Kirli veriler veri madenciliğinde karmaşıklığa neden olmakta ve sonuçlardan kuşku duyulmasına, güvensizliğe neden olmaktadır. Bu nedenle uygulamaya geçmeden önce veri ön işleme adımı olan veri temizleme işlemlerinden sonra verilerin kullanılması gerekmektedir. Veri analizlerine veriyi temizleyerek ve uygun olmayan verilerin ayıklanması suretiyle başlanır. Bu tür işlemlerin ardından veri analiz edilir ve sonuçlar elde edilir (Dasu ve Johnson, 2003; s.1; Maimon, 2007, s.22).

Veri temizleme sürecinde, kayıp olan değerlerin yerine yenilerinin bulunması, gürültülü (noisy) ve uyumsuz verilerin ayıklanması gibi işlemlere yer verilir. Veri tabanındaki kayıp verilerin telafi edilebilmesi için bazı yollara başvurulur. Kayıp veri sorununun nasıl halledileceği konusuna geçmeden önce kayıp veri sorununu yaratan nedenler üzerinde durulmasında yarar vardır.

Kayıp veriler birçok nedenden dolayı oluşmaktadır. Örneğin; müşteri bilgileri her zaman mevcut olmayabilir ya da bu bilgilere ulaşılamayabilir. Bazı veriler önemsiz görüldüğünden veri girişi sürecinde göz ardı edilmiş olabilir. Bazı veriler ise kayıtlı verilerle uyumlu olmadıkları düşünülerek silinmiş olabilir. Ayrıca geçmişin kayıtları veya verilerdeki değişiklikler dikkate alınmamış olabilir. Veri kümesini kayıp değerlerden arındırmak için kullanılan yöntemlerden en önemlileri ve yaygın biçimde kullanılanları şunlardır (Han ve Kamber, 2006, s.289):

- Kayıp verilerin yer aldığı kayıtlar tümüyle yok edilir. Bu şekilde yok edilen sorunlu kayıtların sayısı fazla ise bu yöntem kullanışlı olmaz.
- Kayıp değerler için genel bir sabit kullanılır. Bu yöntemde “BİLİNMIYOR” veya “∞” gibi sabit bir değer tanımlanabilir. Ancak veri madenciliği programları bu değeri ortak bir değer gibi algılayabilir. Bu nedenle çok tercih edilen bir yaklaşım değildir.

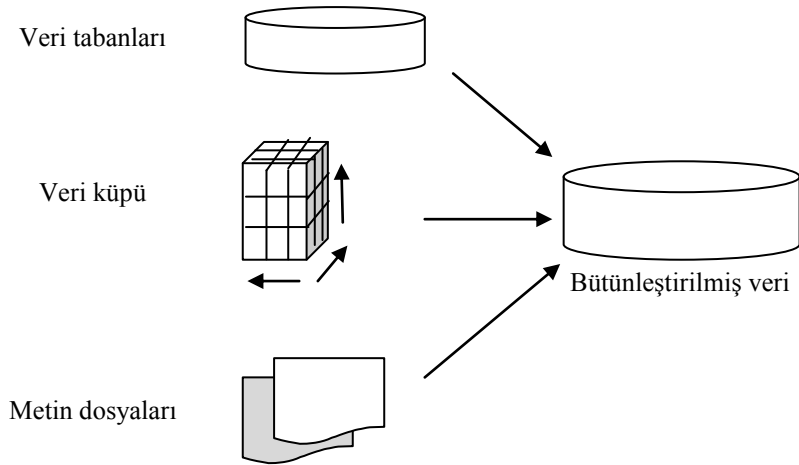
- Sayısal değerlere sahip nitelikler için hesaplanacak ortalama değer kayıp değerler yerine kullanılabilir.
- Kayıp değer yerine aynı sınıfa ait tüm örnekler için değişkenin ortalaması kullanılabilir.
- Kayıp değer mevcut verilere dayanarak en uygun değer kullanılarak tamamlanabilir. En uygun değer belirlenebilmesi için regresyon veya karar ağacı gibi teknikler kullanılabilir. Örneğin veri setindeki diğer müşteri verilerini kullanarak gelir ile ilgili kayıp değerleri öngörmek için bir karar ağacı oluşturmak gibi.

1.3.2. Veri Bütünleştirme

Veri madenciliği uygulamalarının kullandığı veri ambarının oluşturulması esnasında değişik kaynaklardan (veritabanları, veri küpleri, metin dosyaları, vb.) elde edilen veriler arasında uyum sağlamak için verilerin bütünleştirilmesi işlemlerinin yerine getirilmesi beklenir. Bu adım, bütün veri kaynaklarını karakteristiklerine, özelliklerine, ve toplanma seviyelerini eşleştirerek tutarlılık sağlayarak bütünleştirilir.

Farklı kaynaklardaki verilerin bütünleştirilmesi, örneğin bir kaynakta müşteri numarasının bir başka kaynaktan gelen müşteri numarası ile aynı tanıma sahip olması amacıyla yapılır. Veri tabanları ve veri ambarlarında veri türlerinin tanımlandığı metadata bu sorunu çözmede yardımcı olur (Han ve Kamber, 2006, s.67). Başka örnek de, bir bankanın farklı veritabanları üzerinde kredi kartları ile ilgili hesaplara sahip olması üzerine verilebilir. Bu hesaplarda müşteri isimleri farklı biçimlerde kaydedilmiş olabilir (Sumathi ve Sivanandam, 2006, s.11). Özetle ifade etmek gerekirse veri bütünleştirilmesi veri kalitesi ile ilgili sorunların çözülmesi için kullanılan bir yöntemdir (Berti, 2007, s.101).

Veri bütünleştirme aşamaları Şekil 1.3' de gösterilmiştir (Han ve Kamber, 2006, s.67). Farklı veritabanlarının, veri küpü ve kurumun ürettiği metin dosyaları veri bütünleştirme işlemine tabi tutularak veri madenciliği için hazır hale getirilir. Veri küpü çok boyutlu bir veritabanıdır. OLAP (On-Line Analytical Processing) veri tabanı olarak da adlandırılır.



Şekil 1.3. Veri Bütünleştirme.

1.3.3. Veri İndirgeme

Bilindiği gibi veri madenciliği uygulamalarının bazılarında milyarlarca kayıt ve binlerce nitelik ile karşılaşmak mümkün olabilmektedir. Bu durumda verinin temel özelliklerini kaybetmeden, bir kısmını atarak azaltma yoluna gidilip gidilmeyeceği sorunu ile karşılaşılır (Stepaniuk, 2008, s.43). Bu sorunu azaltmanın yolu olan birleştirmede önemli noktalardan biri de veri indirgemedir.

Değişkenlerdeki tutarsızlıklar elde edilen veri setinde fazlalıklara neden olabilir. Bu fazlalıkları tespit etmek için korelasyon analizinden yararlanılmaktadır. Örneğin korelasyon analizi ile müşteri kimliği ile müşteri numarası arasında yüksek bir ilişki bulunursa değişkenlerden biri veri deposundan çıkarılarak indirgeme yapılır. Veri madenciliği yöntemleri hesaplama ağırlıklı olduğundan, veri indirgemesiyle doğal olarak maliyetlerin azalması sağlanabilmektedir. Veri indirgeme işlemleri sonucunda aşağıda belirtilen yararlar elde edilir (Maimon ve Rokach, 2007, s.95):

- Öğrenme (model) maliyetleri azalır.
- Öğrenme (model) performansı artar.
- İlgisiz boyut sayısı azalır.
- Gereksiz boyut sayısının azaltılması.

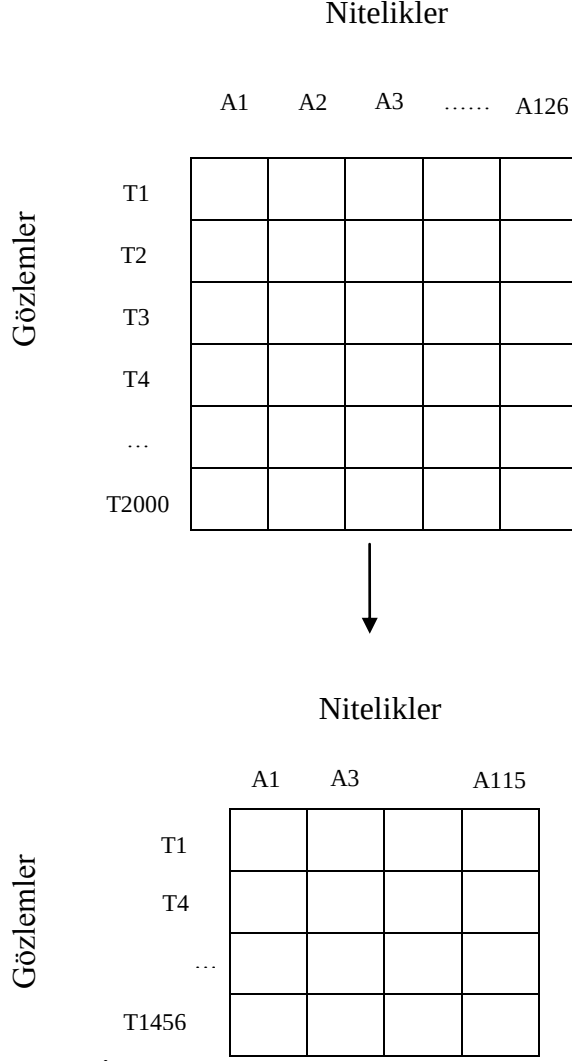
İlgisiz ve gereksiz boyut sayısı sorununun üstesinden gelmek için;

- Nitelik seçimi,
- Kayıt seçimi.

işlemlerine başvurulabilir.

Veri madenciliği araştırmasında hangi niteliklerin kullanılacağına karar verme işlemi “nitelik seçimi” olarak isimlendirilir. Birbirini andıran niteliklerin hepsini kullanmak yerine benzer nitelikleri gözardı edip diğerlerinden anlamlı derecede farklı olanları dikkate almak yeterli olabilir.

Bir veri madenciliği uygulamasında $A_1, A_2, \dots, A_{126}$ biçiminde 126 adet nitelik olduğu varsayalım. 126 nitelikten oluşan veriyi analiz etmekte zorluklar ortaya çıkabilir. Bunun için nitelik sayısının azaltılıp azaltılamayacağı araştırılır. Şekil 1.4’de görüldüğü gibi nitelik sayısının 115’e düşürülmesiyle büyük bir avantaj yakalanmış olur. Bazı niteliklerin gözardı edilmesi bazı kayıtların da göz ardı edilmesi sonucunu doğurur. Böyle olmasa bile bazı kayıtların dikkate alınması gerekmeyebilir. Aynı şekilde gösterildiği gibi orijinalinde $T_1, T_2, \dots, T_{2000}$ biçiminde 2000 kayıttan oluşan veri 1456 adet kayıta indirilebilir. 126×2000 boyutundaki bir veri kümesi yerine 115×1456 boyutundaki bir veri kümesiyle çalışmanın avantajları aşikardır.



Şekil 1.4. Veri İndirgeme.

1.3.4. Veri Dönüştürme

Veri birleştirmede önemli olan üçüncü ve son nokta ise ölçekleme ve kodlamalardaki farklılıklardır. Örneğin, ağırlık bir sistemde kilogram ile kaydedilmişken başka bir sistemde paund ile kaydedilmiş olabilir. Verilerin bu şekilde heterojen bir yapı oluşturmaları veri bütünlüğü açısından büyük tehlikeler oluşturmaktadır. Veri dönüştürmede kullanılan yaklaşımlar aşağıda açıklanmıştır.

1.3.4.1. Ondalık Ölçekleme

Ondalık ölçekleme, sayısal değerlerin -1 ile 1 arasında yer almalarını sağlayacak biçimde dönüştürülmesine karşılık gelir. Bu yöntemde orijinal veri değeri ondalık

noktanın yeri değiştirilerek dönüştürülür. v niteliğinin değeri $v(i)$ olmak üzere, i durumu için aşağıdaki oranla hesaplanır.

$$v'(i) = \frac{v(i)}{10^k} \quad (1.1)$$

Burada k , $\max(|v'(i)|) < 1$ koşulunu sağlayan en küçük değerdir. Veri serisi içinden önce $\max(|v'(i)|)$ elde edilir. Bu işlemin ardından ondalık noktası maximum mutlak değeri 1 den küçük oluncaya kadar hareket ettirilir ve k elde edilir. Dolayısıyla hareket edecek ondalık kısım, değişkenin maksimum mutlak değerine bağlıdır. Örneğin, en büyük değer 455, en küçük değer -834 ise maksimum mutlak değer .834 olarak belirlenmiş olur. O halde veriler 1000 e bölünecektir ve böylece $k = 3$ için diğer değerler de dönüştürülmüş olacaktır.

1.3.4.2. Min-Max Normalizyonu

Min-maks normalizasyonu veya min-maks normalleştirme orijinal veri üzerinde doğrusal bir dönüşüm yapar. X gözlem matrisi olmak üzere, min-max normalizasyonu,

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1.2)$$

dönüşümü ile elde edilir (Larose, 2005, s.36).

Bu bağıntıda yer alan $\min(X)$ değeri X gözlem değerlerinin en küçük olanını, $\max(X)$ ise en büyük olanını ifade etmektedir.

1.3.4.3. Z-score Standartlaştırma

Verileri dönüştürmek amacıyla kullanılan bir diğer yöntem z-score standartlaştırma yöntemi olarak bilinir. İstatistiksel veri dönüştürme teknikleri arasında yer alan ve en yaygın biçimde kullanılan bu yöntem, ele alınan verinin ortalama ve standart sapma değerlerini kullanır. X gözlem değerleri olmak üzere,

$$X^* = \frac{X - \bar{X}}{\sigma_X} \quad (1.3)$$

bağıntısı yardımıyla dönüştürme işlemi gerçekleştirilir. Burada $\bar{X}$ değeri X gözlem değerlerinin aritmetik ortalamasını, σ_X ise standart sapmasını ifade etmektedir.

Farklı tekniklerle gerçekleştirilebilen normalizasyon işlemi, veri boyutunun küçültülmesi amacıyla kullanılabileceği gibi, verilerle gerçekleştirilecek işlemlerin uygun aralıklara normalize edilmiş değerlerle yapılarak işlemlerin daha hızlı gerçekleştirilip ve daha anlamlı ve kolay yorumlanabilir sonuçlar almak amacı ile de kullanılabilir (Khemka, 2003, ss.12-13).

1.5. VERİ MADENCİLİĞİ YÖNTEMİNİN BELİRLENMESİ

Veri madenciliğinin bu kadar geniş bir alana yayılmasının sebebi ise veritabanı yönetimi, istatistik ve bilgisayar bilimi disiplinlerinin birleşmesinden oluşmasıdır. Dolayısıyla kullanılabilecek çok sayıda yöntem ve çok sayıda algoritma vardır. Bu durumda, uygulamada veri madenciliği yöntemlerinden hangisinin kullanılacağına belirlenmesi uygulamanın başarısı bakımından kilit öneme sahiptir. Örneğin uygulama için karar ağaçlarıyla nalize karar verilmişse pek çok algoritma (ID3, C4.5, CART yöntemleri, Bayes ağları ya da Destek Vektör makinaları, vb.) arasından amaca en uygun olanının seçilmesi esastır. Benzer biçimde veritabanlarında toplanan çok fazla sayıdaki verilerden dolayı, bir kümeleme analizi yapılacak ise ona uygun veri madenciliği yöntemlerine başvurulur. Veri madenciliği konusunda çok sayıda yöntem ve algoritma geliştirilmiştir. Bu yöntemlerin bir çoğu istatistiksel tabanlıdır. Bazı yöntemler ise yapay zeka teknikleri arasındadır (Olson ve Delen, 2008, s.4). Veri madenciliği yöntemleri temel olarak aşağıdaki şekilde gruplandırılabilir:

- Sınıflandırma
- Kümeleme
- Birliktelik kuralları

1.5.1. Sınıflandırma

İnsanlar verileri daima sınıflandırdıkları, kategorize ettikleri ve derecelendirdikleri için Sınıflandırma hem veri madenciliğinin temeli olarak hem de veri hazırlama aracı olarak sıkça kullanılmaktadır. Sınıflandırmanın esas amacı yeni bir nesnenin özelliklerini açıklamak ve bu yeni nesnenin daha önceden tanımlanmış gruplardan birine atamasını yapmaktır. Sınıflandırılmamış veriye uygun bir model uygulanabilir. Veri madenciliğinin sınıflandırma grubu içerisinde en sık kullandığı teknik karar ağaçlarıdır. Aynı zamanda lojistik regresyon, diskriminant analizi, sinir ağları ve fuzzy (bulanık) kümeleri de kullanılmaktadır.

Sınıflandırmayla örneğin, bir bankadan kredi alan müşterilerin risk durumu belirlendikten sonra, bu bilgiler kullanılarak yeni başvurularda müşterinin hangi risk durumuna uyduğu kestirilebilir. Çalışmanın 1.6. kısmında sınıflandırma konusu ayrıntılı biçimde açıklanacaktır.

1.5.2. Kümeleme

Veri kümesindeki gözlem satırları göz önüne alınarak, bu gözlemler arasındaki gruplaşmaları belirleme işlemidir (Jain ve diğerleri, 1988, s.1). Kümelemenin amacı heterojen olan anakütleyi homojen gruplara veya kümelere ayırmaktır. Kümeleme birimleri önceden tanımlanmış sınıfları baz alarak gruplamaz. Yani kümeleme gruplamada benzerlikleri kullanırken, sınıflandırma önceden tanımlanmış sınıflar modelini temel alır. Kümeleme, veri madenciliği sürecinden önce veri hazırlama esnasında da kullanılabilir. k-means ve toplayıcı algoritma (agglomerative algorithm) teknikleri ile onların varyasyonları kümeleme yöntemleri olarak kullanılmaktadır. Özetle ifade etmek gerekirse, kümeleme işleminde sınıflandırmada olduğu gibi hedef niteliğe ihtiyaç yoktur. Örneğin, alışveriş mağazalarında müşterilerin gruplara ayrılması bir kümeleme işlemidir.

1.5.3. Birliktelik Kuralları

Birliktelik kuralları madenciliği, büyük veri yığınları arasındaki ilginç ilişkileri ya da ilişki bağıntılarını yakalar. Birliktelik kuralını uygulamadaki amaç, veritabanındaki ilgili veri gruplarından, " $A_1 \wedge A_2 \wedge \dots \wedge A_m \Rightarrow B_1 \wedge B_2 \wedge \dots \wedge B_n$ " (A_m ve B_n değer

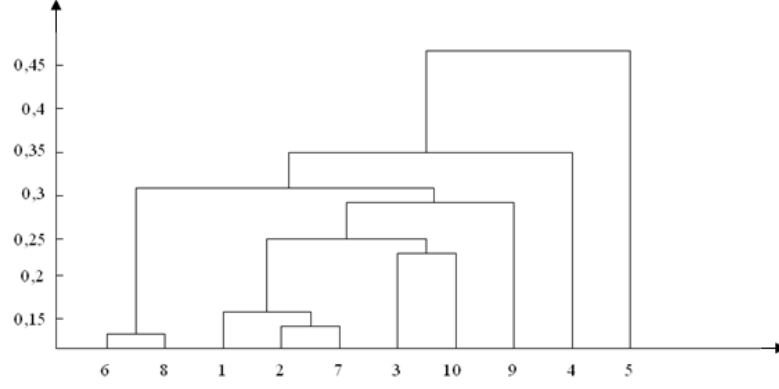
kümelerini ifade eder) şeklinde birliktelik kuralları çıkarmaktır. İlişki kuralları analizinin uygulandığı tipik bir alan sepet analizi alanıdır. Bu tür uygulamalar, müşterilerin sepetlerinde yer alan farklı ürünlerden yola çıkarak; müşterilerin alışveriş alışkanlıklarının yakalanabilmesini hedefler.

Birliktelik kuralları, iki olayın aynı anda ortaya çıkma durumunu ölçen kümeleme analizinde kullanılan eşleştirme yöntemine benzer. Bu ilişkileri değerlendirirken genellikle kullanılan iki ölçüt vardır. İlki olan “güven”, B olayı olduğu zaman A olayının da olma olasılığı olarak hesaplanır. “Destek” terimi ise, kesişen olayların ortaya çıkma sayısının toplam olay sayısı içindeki yüzdesini ifade eder. Bir ilişkiyi kural olarak belirlemek için eldeki olay kümesinin önemli bir kısmı tarafından desteklenmesi gerekir. Bu nedenlerden dolayı, $X \Rightarrow Y$ ilişki kuralı kullanıcı tarafından minimum değeri belirlenmiş güvenilirlik ve destek eşik değerlerini sağlayacak biçimde üretilir.

İlişki kurallarını tespit etmek için büyük veri yığınları üzerinde, tekrar eden sorgular yapılması gerektiğinden, işlem miktarı çok fazla olur ve performans konusunda iyileştirme yapmak gerekliliğinin ortaya çıkması muhtemeldir (Han ve Kamber, 2001, ss.225-250; Hair vd., 1998, s.681).

1.6. SONUÇLARIN SUNUM VE DEĞERLENDİRİLMESİ

Veri madenciliği yöntemlerinin uygulanmasının ardından elde edilen sonuçların karar vericilere sunulması gerekmektedir. Ayrıca sonuçların şekil veya uygun grafiklerle desteklenmesi beklenir. Örneğin merkeze dayalı ayırıcı tekniklerde olduğu gibi bazı kriterler kullanarak nesnelerin (veya veri setinin) hiyerarşik bir yapı içinde ayrılmasını sağlar. bir hiyerarşik kümeleme modeli uygulanmış ise sonuçlar dendrogram adı verilen özel grafiklerle sunulur (Özkan, 2008, s.44). Dendrogramın ağaca benzer bir grafik olduğu aşağıdaki şekilden görülebilir.



Şekil 1.5. Kümeleme Sonuçlarını Gösteren Bir Dendrogram.

Şekil 1.5’deki dendrogramda dikey eksen kümeler arasındaki uzaklığı, yatay eksen ise kümeleri oluşturan gözlem değerlerini ifade etmektedir. Şekil üzerinde görüldüğü gibi (6, 8), (2, 7) ve (1, 2, 7) veriye uygun olarak elde edilen üç adet örnek kümedir.

1.7. NİTELİK SEÇİMİ

Veri madenciliği başta olmak üzere, araştırmalarda nitelik (özellik) seçimi önemli bir konudur. Çok sayıda nitelik ile çalışmak yerine, aynı sonuçları üretebilecek daha az sayıda nitelik kullanılması tercih edilmektedir. Bu nedenle çeşitli nitelik seçme türleri arasında veri yapısına en uygun olanı seçilmelidir. Nitelik seçiminde kullanılan belli başlı yaklaşımlar (algoritmalar) aşağıda açıklanmıştır.

1.7.1. Entropi Tabanlı Nitelik Seçimi

Niteliklerin sıralanması amacıyla kullanılabilen ve entropi ölçümüne dayalı bir nitelik seçim yöntemidir. Bu yöntem hem kategorik hem de sayısal niteliklere uygulanabilmektedir. Tüm nitelikleri sayısal olan veri kümelerinde Entropi tabanlı nitelik seçimi için öncelikle benzerlik ölçüsünden yararlanılır. Nitelikleri önem sırasına göre dizilerek seçim yapmak gerektiğinde hesaplanan benzerlik ölçütü S_{ij} ile gösterilir. i ve j farklı örnekleri göstermek üzere benzerlik ölçütü,

$$S_{ij} = e^{-\alpha D_{ij}} \quad (1.4)$$

ile ifade edilir (Kantardzic, 2003, s.47).

Bu eşitlikte S_{ij} , i ve j örnekleri arasındaki benzerlik ölçütü; D_{ij} , i ve j örnekleri arasındaki normalleştirilmiş Öklid uzaklıkları ve α , bir parametre olarak ifade edilmektedir. α parametresi ise,

$$\alpha = -\frac{\ln 0.5}{D} \quad (1.5)$$

şeklinde ifade edilebilir. Burada D, örnekler arasındaki ortalama uzaklığı ifade etmektedir. Uygulamada $\alpha = 0.5$ olarak kabul edilmektedir. Normalize edilmiş D_{ij} Öklid uzaklığı, X_i ve X_j örneği için şu şekilde hesaplanır:

$$D_{ij} = \left[\sum_{k=1}^n \left(\frac{x_{ik} - x_{jk}}{\max_k - \min_k} \right)^2 \right]^{1/2} \quad (1.6)$$

n : Boyutların sayısı.

$\max_k$: k. boyutun maksimum değeri.

$\min_k$: k. boyutun minimum değeri.

Tümü sayısal olmayan veri kümelerin benzerlik hesaplarında ise yukarıda verilen benzerlik ölçütündeki Öklid uzaklık bağıntısı yerine “Hamming uzaklık bağıntısı” kullanılır. Hamming uzaklığı, i ve j örnekleri için şöyle ifade edilir (Estivill-Castro ve Houle, 2002, s.36; Baldi 2001, s.156):

$$S_{ij} = \frac{\sum_{i=1}^m |x_{ik} - x_{jk}|}{m} \quad (1.7)$$

Burada $x_{ik} = x_{jk}$ ise $|x_{ik} = x_{jk}|$ ifadesinin değeri 1, aksi halde 0'dır. Niteliklerin toplam sayısı ise m'dir. Özelliklerin bazıları sayısal ise, bu durumda ayrıklaştırma (discretization) işlemi uygulanarak sayısal olmayan nitelikler kategorik niteliklere dönüştürülebilirler.

Veri kümesinde en az önemli olan niteliği belirlemek amacıyla entropi kavramından yararlanılır. N örnekten oluşan bir veri kümesinde kullanılacak entropi bağıntısı aşağıdaki gibi ifade edilir:

$$E = - \sum_{i=1}^{N-1} \sum_{j=i+1}^N [S_{ij} \log_2 S_{ij} + (1 - S_{ij}) \log_2 (1 - S_{ij})] \quad (1.8)$$

Burada S_{ij} , X_i ile X_j örneği arasındaki benzerlik ölçütüdür. Tüm nitelikleri içeren F kümesinde nitelikleri sıralayarak seçme işlemi yapılabilmesi için aşağıda belirtilen algoritma uygulanır:

- F'deki bir nitelik ele alınır. Bu nitelik dışındaki tüm niteliklerin her biri için ayrı ayrı E değeri hesaplanır.
- Tüm nitelikler için (a) tekrarlanır.
- En küçük E entropi ölçütüne sahip olan nitelik F kümesinden çıkarılır. Artık F'nin eleman sayısı 1 azalmıştır.
- Yeni F kümesi göz önüne alınarak (a) ya dönülür.
- Geriye bir nitelik kalıncaya kadar işlemlere devam edilir.
- Nitelikler bu şekilde önem sırasına göre listelenir.

1.7.2. Temel Bileşenler Analizi

Veri madenciliğinde veri sıkıştırma amacıyla kullanılan Temel Bileşenler Analizi istatistikte yaygın biçimde kullanılan temel bileşenler analizi tabanlıdır. Temel bileşenler analizi, çok değişkenli analizin en eski ve en çok bilinen tekniğidir. İlk kez

tarafından kullanılan teknik daha sonra bağımsız olarak Hotelling tarafından geliştirilmiştir. Temel bileşenler analizi ile değişkenlerin bileşkeleri diyebileceğimiz daha az sayıda yeni değişkenler ya da temel bileşenler oluşturulur. Temel bileşenler birbirinden bağımsızdırlar. Böylece değişkenler arası bağımlılık yapısı da ortadan kaldırılmaktadır. Ayrıca temel bileşenler analizi çoklu doğrusal bağlantı problemine literatürde önerilen çözümlerden biridir.

Aralarında korelasyon bulunan, yani ilişkili değişkenlerin oluşturduğu bir veri kümesini, aralarında korelasyon bulunmayan, yani ilişkisiz ve daha az sayıdaki değişkenden oluşan bir veri kümesine dönüştürme işlemine “Temel bileşenler analizi” (Principal Component Analysis) adı verilmektedir. Bir veri kümesinde değişkenler arasında önemli derecede kovaryans veya korelasyon varsa, orijinal veriler yerine, veri kümesinin doğrusal bileşenlerinden oluşan daha az sayıda yapay değişkenin oluşturduğu veri kümesi kullanılır (Özdamar, 2004, s.213).

P sayıdaki nitelikten oluşan $X = [x_1, x_2, \dots, x_p]$ veri kümesini göz önüne alalım. X gözlem matrisindeki P değişkeninin bileşenlerini bulmak için, kovaryans matrisi ve korelasyon matrisi kullanılır. Bu matrislerin özdeğerleri ve özvektörleri hesaplanarak doğrusal bileşenler elde edilir.

Özdeğerlere karşılık gelen özvektörler birbirlerine diktir. Yani bağımsız oldukları anlaşılır. Bu özellikten yararlanılarak temel bileşenler, özdeğerlerin büyüklüğü göz önüne alınarak seçilir (Özdamar, 2004, s.216). Temel bileşenler analizi ile ulaşılması istenilen ilk sonuç; $X_1, X_2, \dots, X_p$ gibi p tane değişkeni, önemli bir bilgi kaybına neden olmaksızın, bu değişkenleri temsil edebilen daha az sayıda değişkene indirgemek ve değişkenlere etki eden genel nedensel faktörleri elde etmektir. Daha sonra indirgenmiş yeni değişkenler ile çalışmanın amacı doğrultusunda çeşitli sonuçlara ulaşılabacaktır.

1.7.3. Değişkenlerin Standardize Edilmesi

X gözlem verilerinde değişkenler arasında yüksek korelasyonlar varsa, değişkenler çok farklı birimlerle ifade edilmişlerse ya da değişkenlerin değerlerindeki değişimler çok büyük ise, bu durumda değişkenlerin standardize edilmesi yoluna gidilir.

X gözlem değerlerini standardize etmek üzere, ortalama ve standart sapma değerlerinden yararlanılır. Z standardize edilmiş matris olsun.

$$Z = \begin{bmatrix} \frac{x_{11} - \bar{x}_1}{\sigma_{11}^2} & \frac{x_{12} - \bar{x}_1}{\sigma_{11}^2} & \dots & \frac{x_{1p} - \bar{x}_1}{\sigma_{11}^2} \\ \frac{x_{21} - \bar{x}_2}{\sigma_{22}^2} & \frac{x_{22} - \bar{x}_2}{\sigma_{22}^2} & \dots & \frac{x_{2p} - \bar{x}_2}{\sigma_{22}^2} \\ \dots & \dots & \dots & \dots \\ \frac{x_{n1} - \bar{x}_p}{\sigma_{pp}^2} & \frac{x_{n2} - \bar{x}_p}{\sigma_{pp}^2} & \dots & \frac{x_{np} - \bar{x}_p}{\sigma_{pp}^2} \end{bmatrix} \quad (1.9)$$

Matrislerle ifade edildiğinde, aşağıdaki $V^{1/2}$ köşegen matrisi, μ değeri x_i 'nin ortalaması olmak üzere,

$$Z = (V^{1/2})^{-1}(x - \mu) \quad (1.10)$$

yazılabilir (Larose, 2006, s.4). Standardize edilmiş verilerin kovaryans matrisi ve korelasyon matrisi birbirine eşittir.

$$Corr(Z) = Cov(Z) \quad (1.11)$$

1.7.3.1. Korelasyon Matrisi ile Temel Bileşenlerin Bulunması

Standardize edilmemiş verilerden elde edilen kovaryans matrisinin özdeğerleri,

$$\det(S - \lambda I) = 0 \quad (1.12)$$

denklemini çözülerek elde edilir. Büyüklük sırasına dizilen $\lambda_1, \lambda_2, \dots, \lambda_m$ özdeğerleri yardımıyla E öz vektörü hesaplanır. Temel bileşenler,

$$Y = E^T X \quad (1.13)$$

şeklinde ifade edilir. Bu durumda birinci temel bileşen,

$$y_1 = e_1^T X = e_{11}x_1 + e_{12}x_2 + \dots + e_{1m}x_m \quad (1.14)$$

şeklinde ifade edilir. Benzer şekilde diğer temel bileşenler bulunur. Standardize edilmiş veriler kullanılacak ise, önce Z standardize veri matrisinin hesaplanmış olması gerekir. Burada $R = \text{Corr}(Z) = \text{Cov}(Z)$ olduğu bilinmektedir. Bu matrisin özdeğerlerini bulmak için,

$$\det(Z - \lambda I) = 0 \quad (1.15)$$

denklemini çözülür. Elde edilen λ özdeğerleri sıraya dizilerek E özdeğer matrisi bulunur. Temel bileşenleri elde etmek için,

$$Y = E^T Z \quad (1.16)$$

bağıntısından yararlanılır. Bu durumda birinci temel bileşen,

$$y_1 = e_1^T Z = e_{11}z_1 + e_{12}z_2 + \dots + e_{1m}z_m \quad (1.17)$$

şeklinde ifade edilir.

1.7.3.2. Temel Bileşenlerin Sayısını Belirlemek

Temel bileşenlerin sayısını belirlemek için çeşitli yöntemlere başvurulur. Bu yöntemlerden en yaygın kullanılanları oran testi ile eğim testidir. Bu iki test aşağıda açıklanmıştır.

Oran Testi: Temel bileşenler analizinde, temel bileşen sayısına karar vermek için kullanılan yöntemlerden biri oran testidir. Elde edilen λ_i özdeğerleri göz önüne alınarak, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ olmak üzere m sayıda özdeğer kullanılarak,

$$R = \left(\sum_{i=1}^m \lambda_i \right) / \left(\sum_{i=1}^n \lambda_i \right) \quad (1.18)$$

oranı hesaplanır. Bu oran bir " R^* " eşik değeriyle karşılaştırılır. Eğer $R > R^*$ ise belirlenen "m" sayıda özdeğere sahip özellik temel bileşen olarak kabul edilir.

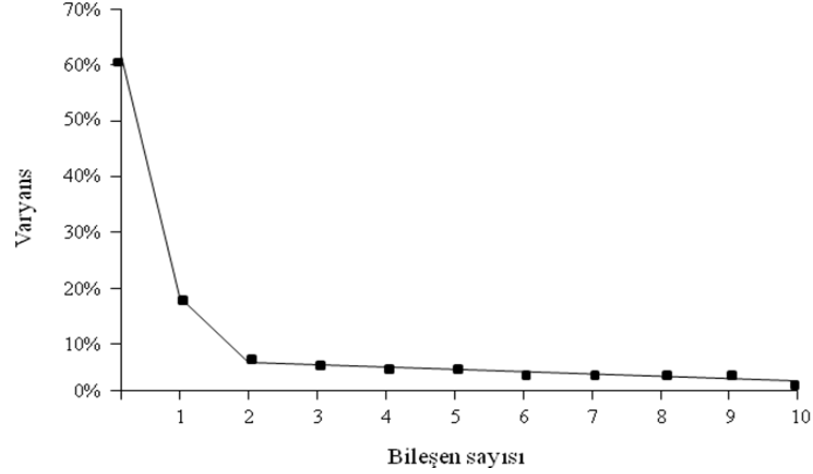
Eğim Testi: R veya S matrisinin özdeğerleri büyükten küçüğe doğru sıralandıktan sonra, bunlar bir grafik üzerinde gösterilir ve eğimlerine bakılır. Eğimin sabitleştiği ya da çok küçük azalma değerlerine sahip olduğu noktaya kadar olan özdeğerler temel bileşen olarak işleme alınır, diğerleri göz ardı edilir (Özdamar, 2004, s.229).

Örnek olması bakımından her bir temel bileşenin genel varyansı açıklama oranının hesaplandığını ve bunların büyükten küçüğe doğru Tablo 1.1'in "Varyans" başlıklı sütununda gösterildiği gibi olduklarını varsayalım. Bu durumda birikimli varyans değerleri Tablo 1.1'in son sütununda gösterildiği gibi elde edilirler. Bu değerlerle çizilen eğim grafiği Şekil 1.6'da gösterilmiştir. Grafikte dikey eksen temel bileşenlere bağlı olarak varyans miktarlarını, yatay eksen ise temel bileşenleri ifade etmektedir. Grafiğe bağlı olarak temel bileşen sayısı belirlenir.

Tablo 1.1. Varyans Ve Birikimli Varyans Değerleri.

TB	Varyans (%)	Birikimli Varyans (%)
1	60.2	60.2
2	19.0	79.2
3	4.7	83.9
4	4.0	87.9
5	3.2	91.1
6	2.9	94.0
7	2.0	96.0
8	1.7	97.7
9	1.4	99.1
10	0.9	100

Tablo 1.1’deki veriler esas olmak üzere bileşen sayısı ile onların varyansları arasındaki ilişkiyi yansıtan eğrinin incelenmesinin ortaya koyduğu gibi eğrinin başlangıcındaki ilk iki doğrunun eğimlerinin birbirlerinden önemli ölçüde farklı oldukları kalan sekiz doğrunun eğimlerinin birbirlerine aşağı yukarı eşit oldukları görülebilir. Dolayısıyla bu örnek için temel bileşenlerin sayısı ikidir (Witten ve Frank, 2005, s.308).



Şekil 1.6. Eğim Grafiği.

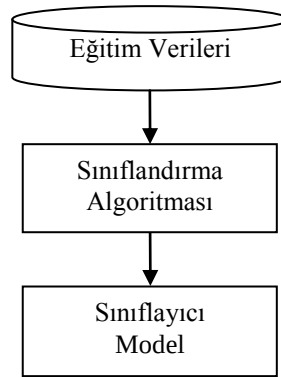
1.8. VERİ MADENCİLİĞİNDE SINIFLANDIRMA

Veri kümesindeki verilerin ortak özelliklerini ortaya çıkarmak ve örüntüleri (patterns) belirlemek veri madenciliğinde büyük önem taşımaktadır. Veri sınıflandırması, bazı özelliklerinin taşıdığı değerlere göre veri yığınlarını sınıflandırmaktır. Bu tarz bir sınıflandırma modeli kurabilmek için, büyük bir veritabanındaki özellik kümelerini içeren örnek veritabanı öğrenme grubu (training set) olarak belirlenir. Önce öğrenme grubundaki veriler analiz edilir ve varolan özellikler kullanılarak bir model ya da kesin tanımlar, yani sınıflandırma kuralları geliştirilir. Daha sonra bu tanım veya modeller, gelecekteki verileri sınıflandırmak için ya da veritabanındaki sınıflar için daha iyi sınıflandırma kuralları geliştirmek için veritabanı üzerinde kullanılır. Örneğin, bir araba satıcısı için müşterilerini tercih ettikleri modele göre sınıflandırmak, satış elemanlarının kime yöneleceğini ve yeni modelleri içeren katalogların hangi müşterilere gönderileceğini bu sınıflandırma doğrultusunda belirlemek, satıcının ticari fırsatlarını arttıracaktır (Chen, Han ve Yu, 1996, ss.19-21). Başka bir örnek olarak bir alışveriş merkezi yöneticileri ortalama alışveriş miktarından daha fazla alışveriş yapan

müşterileri “varlıklı”, diğerlerini ise “yoksul” olarak birbirlerinden ayırt etmek isteyebilir. Bu ayrım sonucunda alışveriş merkezine yeni bir düzenleme getirerek satışların artmasını sağlayabilir. Burada açığa çıkan en önemli husus bir veri kümesi içinde ortak özellikleri ya da farklılıkları ortaya koyacak biçimde sınıflandırma yapılabilmesidir (Özkan, 2008, s.51).

Sınıflandırma tek adımda yerine getirilen bir işlem olmayıp, bir süreç olarak göz önüne alınması gerekir. Genel olarak sınıflandırma sürecinin iki adımdan oluştuğu söylenebilir (Han ve Kamber, 2006, s.287).

Sınıflandırma sürecinin ilk adımını sınıflayıcı modelin elde edilmesi aşaması oluşturur. Sınıflandırmada kullanılacak modelin belirlenmesinde kullanılabilecek pek çok algoritma vardır. Bu algoritmalarından hangisinin en uygun olduğu hem mevcut veri türüne hem de amaç ve uygulamaya bağlıdır. Uygulamada tanımlayıcı ve açıklayıcı bir araç olarak karar ağaçları kullanılacaksa öncelikle eldeki veri kümesinin iki alt kümeye ayrılması işlemi gerçekleştirilir. Bunun için orijinal veri kümesinden bir kısmı eğitim verisi adı verilen kümeyi oluşturmak üzere seçilir. Karar ağaçları bu kümedeki veriler kullanılarak eğitilir. Eğitim için uygun bir sınıflandırma algoritmasının belirlenmiş olması esastır. Eğitim süreci tamamlandığında karar vermede kullanılacak olan karar kuralları türetilmiş olur. Sınıflandırma modelinin elde edilmesi süreci Şekil 1.7’de gösterilmiştir.

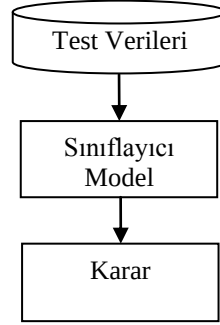


Şekil 1.7. Eğitim Verileri İle Bir Veri Madenciliği Algoritmasının Kullanılması.

Sınıflayıcı modelin elde edilmesinin ardından, bu modelin test edilmesine geçilir. Kısaca bu aşama belirli ölçütlerle modelin sonuçlarının veya birden fazla model söz

konusuysa modellerin karşılaştırılmasının yapıldığı aşamadır. Test için eğitim kümesinin seçildiği veri tabanından bir miktar veri (test kümesi) daha alınarak model bu yeni veriler üzerinde denenir.

Deneme sonucu tatminkar, yani model test verileri üzerinde başarı ile çalışmışsa söz konusu modelin gelecekte de kullanılmasına karar verilir. Eğer model yeterince başarılı bulunmamışsa sınıflandırma modelinin ve/veya modelin oluşturulmasında kullanılan algoritmanın değiştirilmesine ve yeni bir model oluşturulmasına karar verilebilir. Bu işlemler başarılı bir model elde edilinceye kadar sürdürülür. Modelin performansının belirlenmesi amacıyla gerçekleştirilen işlemler aşağıdaki şekilde özetlenmiştir.



Şekil 1.8. Modelin Uygulanması.

1.9. SINIFLANDIRMA MODELLERİ

En yaygın uygulanan veri madenciliği tekniklerinden biri olan sınıflama veya sınıflandırmada veri kümelerini sınıflandıran model oluşturmak için önceden sınıflandırılmış örnekler kullanılır. Özellikle hile tespiti ve kredi riski uygulamaları bu tip analizlere en uygun uygulamalardır. Bu yöntem yaygın olarak karar ağaçları ve nöron ağlarını esas alan algoritmaları kullanır. Karar ağaçları dışında kalan belli başlı sınıflandırma teknikleri şunlardır:

- Yapay Sinir Ağları (Artificial Neural Networks)
- Genetik Algoritmalar (Genetic Algorithms)
- K- En Yakın Komşu (K-Nearest Neighbour)
- Bellek Temelli Nedenleme (Memory Based Reasoning)

- Naive - Bayes
- Lojistik Regresyon (Logistic Regression)

Karar ağaçları yöntemleri, temel anlamda hedef (bağımlı) değişkeni, tahmin edici değişkenlere göre ayırma mantığına dayansa da, bünyesinde değişik amaçlara hizmet eden birbirinden farklı algoritmalara sahiptir (Koyuncugil, 2006, s.72). Karar ağacı oluşturmak için geliştirilen bu algoritmalar arasında CHAID (Chi-Squared Automatic Interaction Detector), Exhaustive CHAID, CRT (Classification and Regression Trees), ID3, C4.5, MARS (Multivariate Adaptive Regression Splines), QUEST (Quick, Unbiased, Efficient Statistical Tree), C5.0, SLIQ (Supervised Learning in Quest), SPRINT (Scalable Parallelizable Induction of Decision Trees) yer almaktadır.

Sınıflandırma modelleri pek çok alanda uygulanabilmektedir. Bunlardan birkaçını şu şekilde sıralamak mümkündür (Mitra ve Acharya, 2003, s.18):

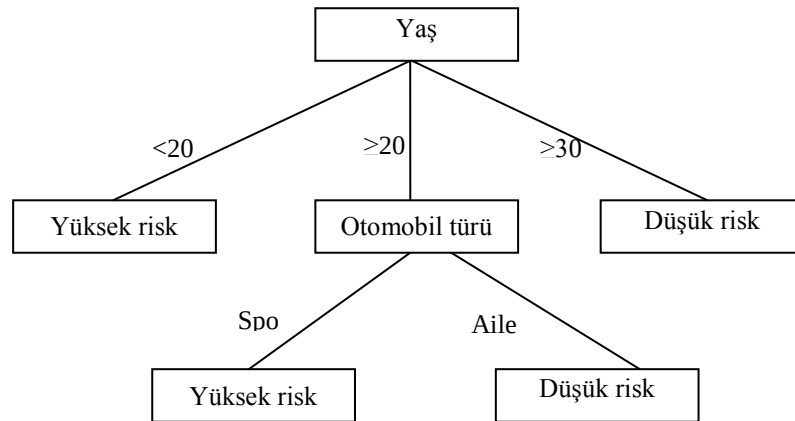
- Bankalarda veya imza duyarlı belge işlemlerinde imzanın belirlenmesinde (uyuyor, uymuyor),
- Güvenlik uygulamalarında sayısal parmak izinin belirlenmesinde (karşılaşıyor, karşılaşmıyor),
- Kredi kartı değerlendirmelerinde (iyi, kötü)
- Banka konumunu göz önünde bulundurarak müşteri kalitesi ve iş olanakları (iyi, vasat, zayıf)
- Kan örneklerinin dijital görüntüsü içindeki şüpheli hücrelerin algılanması (evet, hayır).

1.9.1. Karar Ağaçları ile Sınıflandırma

Karar ağaçları tekniği, kural çıkarmanın basitliği, sonuçlarının görselliği ve yorumlanmasının kolay olması nedenleriyle son yıllarda yaygın olarak uygulanan veri madenciliği tekniklerinden birisidir. Karar ağaçları, sınıfları bilinen örnek veriden tümevarım yöntemiyle öğrenilen ağaç şekilli bir karar yapısı çeşididir (Sun ve Li, 2008, s.2).

Bir karar ağacı, basit karar verme adımları uygulanarak, büyük miktarlardaki kayıtları, çok küçük kayıt gruplarına bölerek kullanılan bir yapıdır. Her başarılı bölme işlemiyle, sonuç gruplarının üyeleri bir diğeriyle çok daha benzer hale gelmektedir (Berry, 2004, s.194). Bierman ve Friedman tarafından 1973 yılında önerilmiş olan karar ağaçları değişkenleri parçalayarak bir ağaç oluşturma esasına dayanmaktadır.

Karar ağacında, tanımlanan sorunun cevabı gruplara ayrılmaktadır. Soruya verilecek bir ölçüt belirlendikten sonra kümeler arasındaki riski maksimize edecek şekilde cevaplar bölünmektedir. En iyi bölünmeyi bulmak için her soruda bu işlem tekrar edilmektedir. Bir soru için grup oluşturulduktan ve gruplar arasındaki risk maksimize edildikten sonra oluşan iki grup için aynı işlemler devam ettirilmektedir. Bu işlemler istatistik olarak anlamlı bir fark bulunana kadar devam ettirilip istatistik olarak anlamlı bir fark bulunmadığı durumlarda ise son verilmektedir. Ayrıştırma işlemi tamamlandıktan sonra ise o grup içerisinde yer alan gözlemlerin oranına göre grup değerlendirilmektedir (Thomas, 2000, s.159). Hayvanlar alemi bir ana düğüm olarak kabul edilirse, bunu memeliler, kuşlar, sürüngenler biçiminde bölümlere ayırmak mümkündür. Bir sonraki adımda her bir grup, başka alt guruplara bölünerek şematik biçimde ifade edilebilmektedir (Berry ve Linoff, 1997, s.167). Bir akış çizelgesi formunda olan karar ağaçları niteliklerin test işleminin yapıldığı düğümlerden ve test işleminin sonucuna göre oluşan dallardan meydana gelen bir yapıdır. Sonuç olarak ağaç sınıflar ile son bulmaktadır. Karar ağacı öğrenmeye yukarıdan aşağıya doğru başlar. Ağaç üzerindeki düğümler test noktalarıdır. Şekil 1.9’da bir karar ağacı örneği verilmiştir.



Şekil 1.9. Karar Ağacı.

Karar ağaçlarını kullanan sınıflandırma modelleri çalışmanın ikinci bölümünde daha ayrıntılı biçimde ele alınarak incelenecektir.

1.9.2. Bayes Sınıflandırma Yöntemi

Bayes sınıflandırma yöntemi, Thomas Bayes tarafından geliştirilen, koşullu olasılıkların hesaplanmasında kullanılan Bayes teoremine (ters olasılık ilkesi) dayanır. Söz konusu teorem bir olayın ortaya çıkmasında birden fazla bağımsız nedenin etkili olması durumunda, bu nedenlerden herhangi birinin o olayı meydana getirme olasılığını hesaplamada kolaylık sağlar.

Bayesyen sınıflandırma tekniği, elde bulunan, hali hazırda sınıflanmış verileri kullanarak yeni bir verinin mevcut, sınıflardan herhangi birine girme olasılığını hesaplayan bir yöntemdir. Bayes kuralına dayalı geliştirilmiş algoritma ve sınıflandırma teknikleri bu adla anılır.

İlk kez Thomas Bayes'in ölümünden sonra yayımlanan çalışmasında gün ışığına çıkan Bayes olasılık denklemi aşağıda gösterilmiştir. Bayes denklemine göre, durumu (D) verilen olayın (h) olasılığının (1.19) denkleminde, $P(h)$; olayın, durum belirtilmeden önceki olasılığını, $P(h/D)$ ise olayın, durum belirtildikten sonraki olasılığını göstermektedir (Mitchell, 1997, s.156).

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (1.19)$$

İstatistikte yaygın biçimde kullanılan Bayes Teoremi, veri madenciliğinde sınıflandırma işlemlerinde de yaygın biçimde kullanılmaktadır. (1.19)'dan hareket ederek Sade Bayes sınıflandırıcısı (Naive Bayes Classifier) ya da kısaca “Bayes sınıflandırıcısı” adı verilen kavrama ulaşılır.

X'in sınıf üyeliği bilinmeyen bir veri örneği olduğunu kabul edelim. Bu veri örneği $X = \{x_1, x_2, \dots, x_n\}$ nitelik değerlerinden oluşsun. $C_1, C_2, \dots, C_n$ biçiminde ifade edilen n tane sınıf değeri olduğu varsayılarak sınıfı belirlenecek olan örneğe ilişkin olarak,

$$P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)} \quad (1.20)$$

olasılıkları yazılabilir. Hesaplama yükünü azaltmak için $P(C_i|X)$ yerine, x_i değerlerinin birbirinden bağımsız olduğu varsayılarak aşağıdaki bağıntı kullanılabilir (Han ve Kamber, 2006, s.312),

$$P(C_i|X) = \prod_{k=1}^n P(x_k|C_i) \quad (1.21)$$

Seçilen örneğin hangi sınıfa ait olduğunu bulmak için yukarıdaki hesaplanan değerler içinden en büyük olanı seçilir. Grubu bilinmeyen birimin sınıflandırılması için (1.22) nolu denklemle ifade edilen olasılığın hesaplanması yeterlidir (Mitchell, 1997, s.157):

$$\arg \max_{C_i} \{P(X|C_i)P(C_i)\} \quad (1.22)$$

Yukarıdaki denklemle açıklanan yöntem “en büyük sonsal sınıflandırma yöntemi (Maximum A Posteriori classification = MAP)” olarak bilinir. Bu açıklamalar bağlı olarak, Bayes sınıflandırıcısı aşağıdaki bağıntı ile tanımlanır (Mitra ve Acharya, 2003, s.197).

$$C_{\text{MAP}} = \arg \max_{C_i} \prod_{k=1}^n P(x_k|C_i) \quad (1.23)$$

Bayes sınıflandırıcı konusunu kaptmadan önce Bayes teoremine bağlı olarak geliştirilen Bayes ağları kavramı üzerinde duralım. Bayes ağları Bayes teoremi temelli bir sınıflandırma yöntemidir (Jensen, 2001, s.21). Bir Bayes ağın model yapısı yön verilmiş döngüsel olmayan grafikdir ve model parametreleri koşullu olasılık fonksiyonlarındaki sayısal değerlerdir.

Bayes ağları iki önemli özelliğe sahiptirler:

- Değişkenler arasındaki koşullu olasılık ilişkilerini bir grafik yapısı üzerinde tutar.
- Değişkenler arasındaki ortak olasılık dağılımının kompakt bir gösterimidir.

Üç değişken söz konusu olduğunda bu değişkenler arasındaki ortak olasılık dağılımı (1.24) nolu denklemle açıklanır.

$$P(X_1, X_2, X_3) = P(X_2|X_1)P(X_3|X_1)P(X_1) \quad (1.24)$$

(1.24) eşitliği n değişken için genelleştirildiğinde,

$$P(X_1, X_2, \dots, X_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i | \text{ebeveynleri}) \quad (1.25)$$

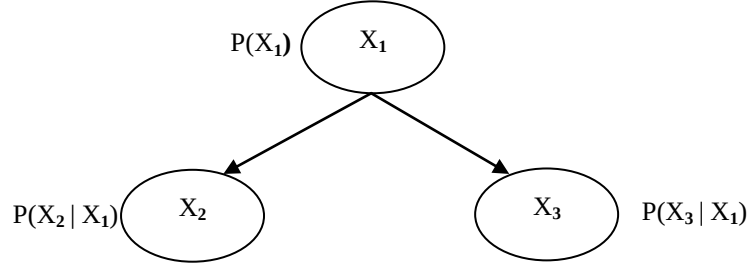
elde edilir (Jensen, 2001, s.21, Mitchell, 1997, s.185).

Bayes ağlarının kurulması sürecinde uzman görüşleri çok önemlidir. Modelin kurulmasında ve önsel olasılıkların belirlenmesin araştırmayı yapan uzmanların görüşlerine başvurulması ihmal edilmemelidir (Heckerman, 1995, s.1). Ağ kurulduktan sonra, ağa soru yöneltilerek sonuçlar elde edilir.

Genel anlamda, belirsizlik taşıyan bir sistemin Bayes ağına göre modelini oluşturmak için aşağıdaki adımlar gerçekleştirilir.

- Sistemde önemli olan tüm değişkenler, düğümler şeklinde oluşturulur.
- Sistemdeki değişkenlerin birbirleriyle olan ilişkisi, düğümlerin bağlantısı şeklinde tanımlanır.
- Önceden bilinen ya da deneyimler sonucu elde edilmiş bilgiler ışığında koşullu olasılık değerleri belirlenir.
- Bu belirlemelerin ardından ağın çizimi gerçekleştirilir.

Bu aşamaların dikkate alınmasıyla oluşturulan bir Bayes ağı Şekil 1.10'da gösterilmiştir. Ağın çiziminde (1.24) nolu eşitlik dikkate alınmıştır. Dolayısıyla Şekil 1.10 (1.24) nolu denklemin grafik gösterimidir



Şekil 1.10. Bir Bayes Ağı

1.9.3. Destek Vektör Makinaları ile Sınıflandırma

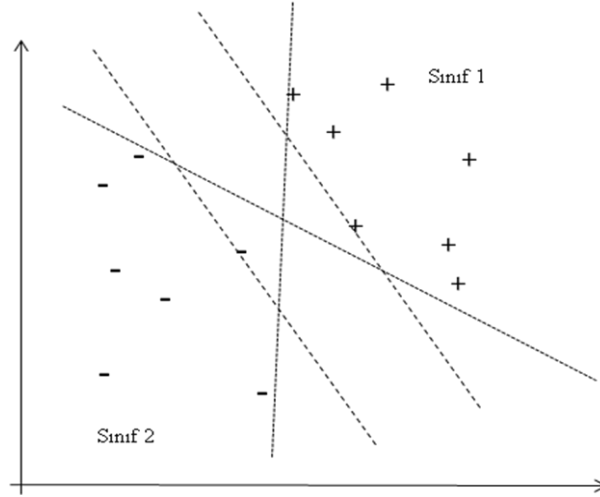
Vapnik tarafından geliştirilen Destek Vektör Makineleri istatistiksel öğrenme teorisi alanında ortaya çıkmış bir öğrenme yöntemidir. Doğrusal olmayan örnek uzayını, örneklerin doğrusal olarak ayrılabilceği bir yüksek boyuta aktararak, farklı örnekler arasındaki maksimum sınırın bulunması esasına dayanan Destek Vektör Makinelerinde karşılaşılabilecek iki durum, verilerin doğrusal olarak ayrılacakları bir yapıda olması veya doğrusal olarak ayrılamayan yapıda olmasıdır. Ancak, gerçek yaşam problemlerinin büyük çoğunluğu birçok farklı bileşenden oluşan problemlerdir ve doğrusal bir yapı halinde karşımıza çıkmazlar. Doğrusal olarak ayrılmış olan veriler arasında maksimum sınırın doğrudan doğruya bulunması oldukça kolay olmakla birlikte, doğrusal olarak ayrılamayan veriler öncelikle doğrusal olarak ayrılacakları farklı bir uzaya aktarılmalıdırlar. Bu bölüm altında bu işlemlerin nasıl gerçekleştirildiği açıklanacaktır.

1.9.3.1. Doğrusal Destek Vektör Makineleri

Destek Vektör Makineleri sınıflandırıcıları, ilk aşamada verinin daha iyi ayırt edilebileceği yüksek boyutlu bir öznitelik uzayına iz düşürülmesini ve burada optimizasyon yöntemi aracılığı ile sınıfları ayırt eden bir optimum ayırma çokludüzlemin (optimum separating hyperplane) bulunmasını sağlar.

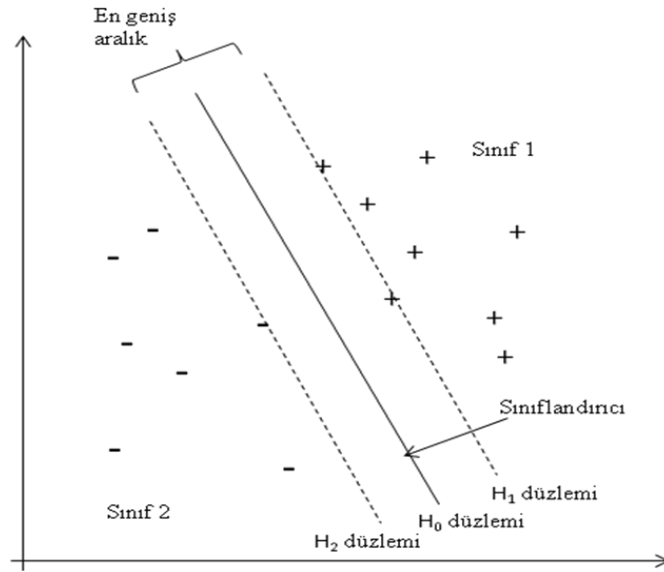
İki boyutlu bir uzayda verileri doğrusal olarak birbirinden ayırmak mümkün olabilir. Bu durum Şekil 1.11'de gösterilmiştir. Bu grafikte amaç pozitif (+) ve negatif (-) olan

noktaları birbirinden ayırmaktır. Verileri iki farklı gruba ayırmak üzere, bu noktalar arasından bir değil, birden fazla doğru geçirmek mümkündür. Eğer iki boyutlu uzay yerine çok boyutlu uzay ile ilgileniliyorsa bu durumda iki veri grubu arasından doğru/doğrular değil çokludüzlemler geçirmek gerekecektir. Burada önemli olan geçirilen doğru ya da çokludüzlemlerden hangisinin dikkate alınması gerektiğinin belirlenmesidir. Destek vektör makineleri çerçevesinde de üzerinde durulan nokta bu soruya cevap bulmaya dayanmaktadır.



Şekil 1.11. İki Boyutlu Uzayda Verilerin Doğrusal İki Sınıfa Ayrılması

Sınıflandırmada çokludüzlem kullanımı Şekil 1.12’de gösterilmiştir. Şeklin ortaya koyduğu gibi sınıflandırmada kullanılabilecek çok sayıda çokludüzlem vardır. Bu çokludüzlemlerden aralarındaki uzaklık maksimum olan iki çokludüzlemin seçilmesi gerekir. Şekil 1.12’de de gösterildiği gibi tanımlanan iki çokludüzlem H_1 ve H_2 olsun. Söz konusu bu iki çokludüzlemin ortasından geçen ve her ikisine birden paralel bir diğer çokludüzlem H_0 olsun. H_1 ve H_2 çokludüzlemleri üzerindeki gözlemler “destek vektör” adını alır. En uygun ayırma düzlemi H_0 düzlemidir. Bu düzlem sınıflandırıcı görevini yapar. (Kecman, 2001, s.149).



Şekil 1.12. Sınıflandırıcı Çoklu Düzlemin Elde Edilişi.

Sınıflandırıcı çoklu düzlemin (H_0) belirlenmesinin ardından sıra bu düzlem üzerindeki noktaların belirlenmesine gelir. $\langle w, x \rangle = \sum_{i=1}^n w_i x_i$ olmak üzere H_0 düzlemi üzerindeki noktalar cinsinden (1.26)'daki gibi ifade etmek mümkündür.

$$H_0: \langle WX \rangle + b = 0 \quad (1.26)$$

Burada W ağırlık vektörü $W = \{w_1, w_2, \dots, w_n\}$ ile ifade edilir ve b ise sabit sayı olarak eşitlikte yer alır. Sınıflandırıcı çoklu düzlemin üst tarafında kalan noktalar için, yani $y_i = +1$ değerleri için $\langle w, x \rangle + b > 0$; benzer biçimde çoklu düzlemin alt tarafında kalan $y_i = -1$ noktaları için $\langle w, x \rangle + b < 0$ yazılabilir. Bu iki eşitsizlik yardımıyla (Kecman, 2001, s.150),

$$y_i(\langle w, x \rangle + b) - 1 \geq 0 \quad (1.27)$$

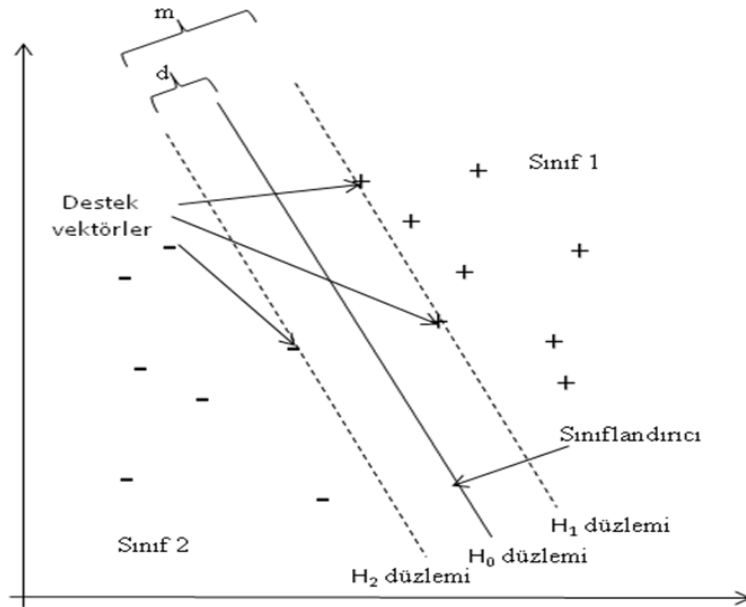
elde edilir. p , çoklu düzlem üzerinde bir nokta olmak üzere, destek vektörlerle H_0 çoklu düzlemi arasındaki uzaklık “ d ” olarak ve aşağıdaki eşitlikteki gibi ifade edilir.

$$d = \frac{|w \cdot x_p \pm b|}{\|w\|} \quad (1.28)$$

x_1 destek vektörü ile H_0 çoklu düzlemi arasındaki uzaklık, (1.27) ve (1.28)'den

hareketle, $d = \frac{|\langle w, x_1 + b \rangle|}{\|w\|} = \frac{1}{\|w\|}$ ile elde edilir. Böylece x_1 destek vektörü ile H_2 çoklu düzlemi arasındaki uzaklık, $m = 2d = \frac{2}{\|w\|}$ olarak belirlenir. Bu açıklamalar Şekil 1.13'deki gibi grafik üzerinde gösterilmiştir.

H_1 ve H_2 çoklu düzlemleri arasındaki uzaklığın maksimum olması amaçladığına göre, $\|w\|$ ifadesinin minimize edilmesi söz konusu olacak ve problem, bir quadratik (kareli, yani doğrusal olmayan) programlama modeline dönüşecektir.



Şekil 1.13. Destek Vektörler.

Kuadratik formda amaç H_1 ve H_2 çoklu düzlemleri arasındaki boşluğun maksimize edilmesidir. $\|w\|$ ifadesinin minimum olması m 'nin maksimize edilmesi demektir.

Bu durumda kuadratik programlama modelinin amaç fonksiyonu, $\langle w, x \rangle = \sum_{i=1}^n w_i x_i$ biçiminde ifade edilir. (1.27) nolu bağıntıyla açıklanan eşitsizlik kuadratik modelin kısıtlayıcı fonksiyonları olmak üzere kuadratik programlama modeli genel gösterimle aşağıdaki gibi açıklanır.

$$\text{Min} : \quad \langle w, x \rangle \quad (1.29)$$

$$\text{Kısıtlar :} \quad y_i(\langle w, x \rangle + b) \geq 1$$

(1.29) ile açıklanan doğrusal olmayan bu problem doğrusal olmayan programlama problemlerinin çözümünde kullanılan Lagrange çarpanları yardımıyla çözümlenebilir (Cristianini, 2000, s.96).

$$f(x) = \text{sign}(\langle w^*, x \rangle + b^*) \quad \{x_i\} \in SV \quad (1.30)$$

Çözümde elde edilen Lagrange çarpanlarının büyük çoğunluğunun değeri sıfır olacaktır. Sıfır değerli olmayanlar destek vektörleridir. SV destek vektörlerini ifade etmek üzere, sadece destek vektörleri göz önüne alınmak şartıyla sınıflandırma (1.30) dikkate alınarak gerçekleştirilir (Kecman, 2001, s.157). Burada w^* ve b^* optimal düzlemin parametreleridir.

1.9.3.2. Doğrusal Olmayan Sınıflandırıcılar

Bilindiği gibi, gerçek hayatta gürültü ve veri sınıflarının doğrusal olmaması nedenleriyle doğrusal olarak ayrılabilir veri genelde bulunmaz. Böyle durumlarda, verinin daha büyük boyutlu uzaylara dönüştürülerek sınıflandırılması gerekir. Bu tür geçişleri sağlamak üzere Çekirdek Fonksiyonları kullanılır. Çekirdek fonksiyonları sınıflandırmanın daha yüksek boyutlarda yapılabilmesini sağlar ve uygulamaların

büyük çoğunluğunda birçok çekirdek fonksiyonu iyi sonuçlar verir. Sık kullanılan çekirdek fonksiyonları doğrusal, polinomial, radyal tabanlı ve sigmoid fonksiyonlarıdır.

$K(x_i, x_j)$ çekirdek fonksiyonu sınıflandırıcı, SV destek vektör değerleri, α_i 'ler ise Lagrange katsayıları olmak üzere çekirdek tabanlı sınıflandırıcı,

$$f(x) = \text{sign} \sum_{i \in SV} y_i \alpha_i K(x_i, x_j) \quad (1.31)$$

biçiminde tanımlanır (Kecman, 2001, s.173).

1.10. MODEL PERFORMANSININ ROC ANALİZLERİ İLE DEĞERLENDİRİLMESİ

Model oluşturulduktan sonra modelin sonuçlarının değerlendirilmesi ve elde edilen sonuçların öneminin yorumlanması gereklidir. Veri kümeleriyle ilgili olarak elde edilen modellerin değerlendirilmesi için sadece modellerin kendi aralarında karşılaştırılması değil aynı zamanda tercih edilen bir modelin uygulanması ile sağlanacak faydaların da karşılaştırılması gereklidir. Bir sınıflama modelinin değerlendirilmesinde kontenjans tablosu ve ROC Grafiği yaygın olarak kullanılır.

1.10.1. Kontenjans Tablosu

Modelin başarımını ölçmek için gerçek sınıf değerleri ile öngörülen sınıf değerleri arasında aşağıdaki tabloda belirtildiği gibi kontenjans tablosu düzenlenir (Witten ve Frank, 2005, s.172).

Tablo 1.2 Kontenjans Tablosu.

	Öngörülen pozitifler	Öngörülen negatifler
Gerçek pozitifler	TP	FN
Gerçek negatifler	FP	TN

Tablo 1,2'de TP doğru pozitiflerin sayısını, FP yanlış pozitiflerin sayısını, FN yanlış negatiflerin sayısını ve TN doğru negatiflerin sayısını ifade etmektedir. Model ile ilgili

olarak hazırlanan bu kontenjans tablosundan yararlanılarak, model başarısında kullanılmak üzere aşağıdaki değerler hesaplanır (Witten ve Frank, 2005, s.172):

$$\text{Doğruluk} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1.32)$$

$$\text{Hata oranı} = \frac{FN + FP}{TP + TN + FP + FN} \quad (1.33)$$

$$\begin{aligned} \text{Duyarlılık} &= \frac{\text{Doğru sınıflandırılmış pozitif örnek sayısı}}{\text{Pozitif sınıflandırılmış örneklerin sayısı}} \\ &= \frac{TP}{TP + FN} \end{aligned} \quad (1.34)$$

$$\begin{aligned} \text{Doğru Pozitif Oranı} &= \frac{\text{Doğru sınıflandırılmış pozitif örnek sayısı}}{\text{Pozitif örneklerin sayısı}} \\ &= \frac{TP}{TP + FN} \end{aligned} \quad (1.35)$$

Buna göre, söz gelimi iki model karşılaştırılmak istendiğinde hangi modelin doğru pozitif oranı ve duyarlılığı yüksekse o modelin daha iyi bir sınıflandırıcı olduğu kabul edilerek, sınıflandırmada onun kullanılmasına karar verilir. Duyarlılık ve doğru pozitif oranı arasında ters orantı olduğu gözden kaçırılmamalıdır.

1.10.2. ROC Eğrileri

Alıcı işletim karakteristiği (orijinal adıyla; Receiver Operating Characteric- ROC) analizi, sınıflandırma modelinin duyarlılık ve belirliliğini ölçmek için kullanılan bir yaklaşımdır. Bunun için teşhisin duyarlılık ve belirliliği arasındaki ilişkiyi tanımlayacak ROC eğrileri kullanılır.

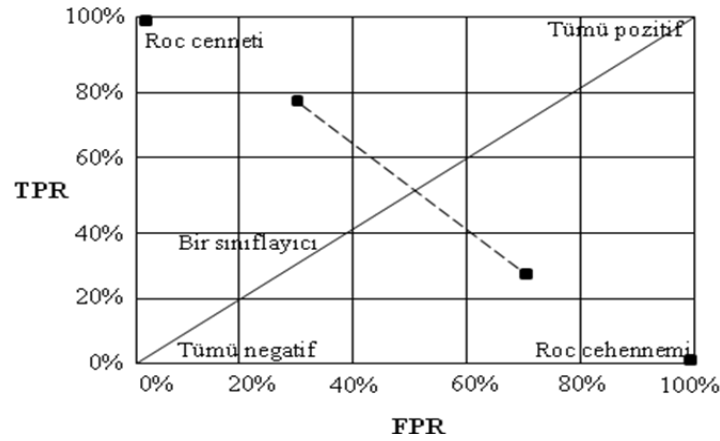
Bir ROC grafiği FP (yanlışlıkla pozitif) oranını yatay ekseninde, TP (doğrulukla pozitif) oranını düşey ekseninde gösterir. Bu grafiğin (0,1) noktası tüm pozitif ve negatif

durumların doğru olarak tahmin edildiği mükemmel bir sınıflamanın gerçekleştiğini (ROC Cenneti) ifade eder. (0, 0) noktası tüm durumların negatif, (1,1) noktası tüm durumların pozitif ve (1, 0) noktası tüm durumların hatalı tahmin edildiğini (ROC Cehennemi) ifade eder (bkz. Şekil 1.14). Dolayısıyla düşey eksene ve üst sınıra yakın olan eğriler başarılı bir testi ifade ederken, 45° gibi eğime sahip eğriler başarısız bir testi gösterir. Böylece ROC eğrileri incelenerek testin başarısı belirlenebilir. Başarılı bir testte eğrilerin altında kalan alanın büyük olması beklenir. ROC grafiği bir sınıflama algoritmasının pozitif durumları doğru olarak belirleme yeteneği ile hatalı olarak sınıflanmış negatif durumların sayısı arasındaki ödünleşmeyi incelemek için kullanılan görsel bir araçtır. ROC eğrilerinin çizimi için gerekli olan doğru pozitif oranı TPR (y ekseni) ile yanlış pozitif oranı FPR (x ekseni) aşağıdaki biçimde ifade edilir (Witten ve Frank, 2005, s.173).

$$TPR = \frac{TP}{TP + FN} \quad (1.36)$$

$$FPR = \frac{FP}{TN + FP} \quad (1.37)$$

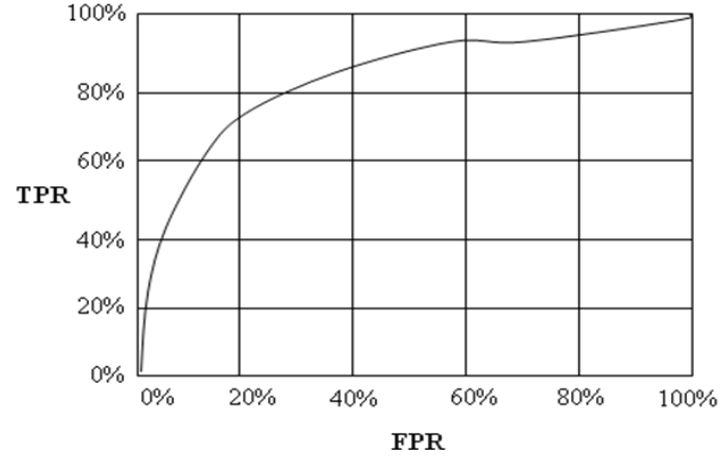
ROC eğrileri TP ve FP değerleri kullanılarak çizilir.



Şekil 1.14. ROC Uzayı.

Özellikle risk taşıyan iki veya daha fazla sistemin karşılaştırılmasında çok faydalı olan muhtemel bir ROC eğrisi Şekil 1.14’de gösterilmiştir. Eğri altındaki alan 1’e oldukça

yakın olduğundan eğrisi çizilen modelin sınıflandırmada kullanılabileceği anlamına gelmektedir.



Şekil 1.15 ROC Eğrisinin Görünümü.

BÖLÜM II: KARAR AĞAÇLARI VE SINIFLANDIRMA

2.1. ÖĞRENME

Canlıların yaşamını sürdürebilmeleri için çevre hakkında bilgi sahibi olmaları gerekir. Bilgi öğrenilerek kazanılır. Psikologlara göre öğrenme, doğuştan gelen davranışların ve eğilimlerin dışında, çevredeki etkileşimler yoluyla davranışların oluşturulması ya da değiştirilmesi sürecidir.

Öğrenme tekrar ve yaşantılar sonucu davranışlarda meydana gelen kalıcı bir değişimdir. Ayrıca öğrenme davranışta bir değişikliktir; bu değişiklik kötü veya iyi olabilir. Örneğin, sigara ve içki içimi, anormal davranışlarda bulunma, bazı şeylerden korkmak ya da hoşlanmak birer öğrenmedir. Bu değişiklikler tekrar ve yaşantılar sonucu meydana gelir. Örneğin, bir çocuk sobaya elini dokundurduğunda, eli yandığı için, bir daha o sobaya dokunmayacağını öğrenir. Öğrenme devamlı bir değişimdir ve bu değişimler süreklilik göstererek, zaman içinde meydana gelir (Ünlü ve Hakan, 2001, s.36). Genel anlamda değerlendirildiğinde, öğrenme bir sistemin çevreye uyum sağlayabilmek için kendini yeniden düzenlemesi olarak ifade edilir (Nabiyev, 2003, s.563).

2.1.1. Makine Öğrenmesi

Öğrenme sadece canlılar için değil, yapay olarak tasarlanmış sistemler (öğrenme sistemleri) için de söz konusudur. Örneğin, canlıların beyin yapısını kopya ederek oluşturulan “Yapay Sinir Ağları” yardımıyla bir öğrenme işlevi yerine getirilebilir. Makine öğrenmesi (machine learning) adı altında ortaya konan kavram, yapay sinir ağları gibi, bilgisayarlar yardımıyla yapay öğrenme sistemlerinin yaratılmasını amaçlar (Öztemel, 2003, 31). Makine öğrenmesi, bilgisayarın kendisine sunulan verileri kullanarak nasıl davranılacağı konusunda çözüm üretme süreci olarak değerlendirilir. Veri madenciliğinde makine öğrenmesi yöntemlerinden sıkça yararlanılır (Witten, 2005, s.7).

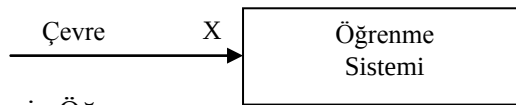
Bilgisayara belirli bir konu hakkında örneklerin yüklenmesi ve bu örneklerle dayanarak bir öğrenme modelinin geliştirilmesi söz konusudur. Burada “öğrenme” kelimesi ile “makine öğrenmesi” kavramı kastedilmektedir. Makine öğrenmesi konusunda çok

sayıda yöntem geliştirilmiştir ve bu yöntemler “Yapay Zeka” adı verilen bilim dalının temelini oluşturmaktadır. Söz konusu yöntemlerin birçoğu yoğun hesaplama işlemi gerektirdiğinden bilgisayarların yeteneklerinden yararlanılması gerekir. Makine öğrenmesi olayına genel açıdan bakıldığında, iki tür temel öğrenmeden söz edilebilir.

- Öğretmenli öğrenme
- Öğretmensiz öğrenme

Öğretmenli öğrenmede, öğretmen sisteme olay ile ilgili olarak girdi ve çıktıları örnekler yoluyla tanıtır. Girdiler ile çıktılar arasındaki ilişkiler belirlenerek öğrenme olayı gerçekleştirilir (Öztemel, 2003, s.25). Öğretmenli öğrenme, bilinen girdi-çıkıtı örneklerinden hareketle, bilinmeyen bir bağımlılığı tahmin etmek için kullanılır. Sınıflandırma ve regresyon bu tür genel tüme varımcı (inductive) öğrenme tarafından desteklenir. Öğretmenli öğrenme, bir öğretmen-uygunluk fonksiyonunun var olduğunu varsayar. “Öğretmenli” terimi, eğitim örnekleri için çıktı değişkenlerinin bilindiğini belirtmektedir

Öğretmensiz öğrenmede ise sadece girdi değerleri kullanılır. Bu nedenle bir öğretmene gerek yoktur. Bu öğrenme biçiminde değişkenler arasındaki ilişkileri sistem kendi kendine öğrenmek zorundadır (Öztemel, 2003, s.25). Şekil 2.1 üzerinde görüldüğü gibi, çevreden gelen X bilgileri öğrenme sisteminin doğrudan bir girdisidir. Öğrenme sistemi çevreden aldığı bilgilerle kendini düzenler. Sistem dışarıdan aldığı bilgileri kıyaslar, sınıflandırır ve kendi kendine öğrenir.



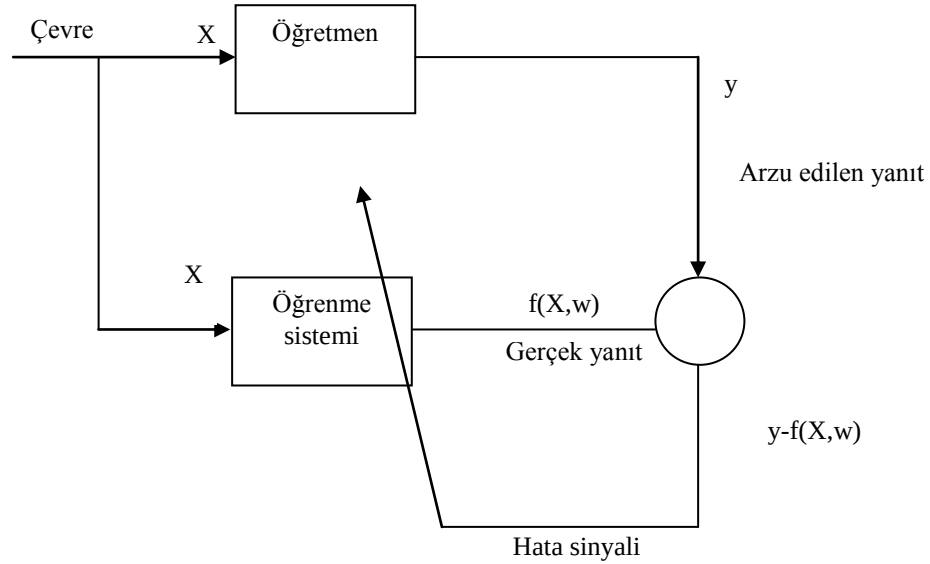
Şekil 2.1 Öğretmensiz Öğrenme.

Teorik açıdan bakıldığında, öğretmenli ve öğretmensiz öğrenme sadece modelin nedensel yapısı açısından farklılık gösterir. Öğretmenli öğrenmede model bir gözlem kümesinin etkisini tanımlar. Gözlem kümesini oluşturan gözlemlere “girdi” denir. Model gözlemlerin bir diğer kümesini de içerir. Buna da “çıkıtı” adı verilir. Bir başka deyişle, girdi sürecin başlangıcını, çıkıtı ise nedensel zincirin sonunu tanımlar. Modeller girdi ve çıktılar arasında arabulucu değişkenleri içerebilir. Öğretmenli ve

öğretmensiz öğrenme arasındaki temel farkı, Şekil 2.1 ve Şekil 2.2 ‘yi karşılaştırarak görmek mümkündür.

Öğretmenli öğrenme sistemini tanıtan Şekil 2.2’de görüldüğü gibi, öğrenme sistemi çevreden gelen bilgileri ele alarak değerlendirmektedir. Sistemin sonuç olarak neyi elde etmesi gerektiği öğretmen tarafından “y” olarak atanır. Bu arzu edilen yanıtı ifade eder.

Öğrenme sistemi, sisteme giren bilgileri kullanarak öğretmen tarafından verilen “y” değerlerini elde etmeye çalışır ve amaca ulaşmak için “w” ağırlıklarını belirler. Öğrenme sisteminin ürettiği $f(X,w)$ ile “y” yanıt değerleri karşılaştırılır. Karşılaştırılan değerler arasında fark olması durumunda üretilen hata sinyali öğrenme sistemine iletilir.



Şekil 2.2 Öğretmenli Öğrenme.

Bu durumda, öğretmenli öğrenme bir girdi kümesi, çıktı kümesi ve bunlar arasındaki ilişkiyi belirleyen bir fonksiyondan ibarettir. Bu açıdan bakıldığında, girdi kümesine eğri uydurma, bir fonksiyon için basit bir öğretmenli öğrenme örneği olarak gösterilebilir (Nilson, 2010, s.6).

2.1.2. Öğretmenli Öğrenme Yöntemleri

Makine öğrenmesi çerçevesinde öğretmenli öğrenme ile ilgili yöntemler geliştirilmiştir. Çalışmanın temelini oluşturan sınıflandırma konusunda en çok kullanılan öğretmenli öğrenme yöntemleri aşağıda sıralanmaktadır (Caruana ve Mizil, 2006, s.162) :

- Destek Vektör Makinası
- Yapay Sinir Ağları
- Lojistik Regresyon
- Bayes Sınıflandırma yöntemi
- En yakın k-komşu algoritması
- Karar Ağaçları

Destek Vektör Makinası: Kesim 1.7.3’de açıklandığı gibi istatistiksel öğrenme teorisi alanında ortaya çıkmış bir öğrenme yöntemi olan destek vektör makinası doğrusal olmayan örnek uzayını, örneklerin doğrusal olarak ayrılabilceği bir yüksek boyuta aktararak, farklı örnekler arasındaki maksimum sınırın bulunması esasına dayanır. Destek vektör makinesi örüntü tanımada gittikçe gözde hale gelen bir öğretmenli genelleme yöntemidir.

Yapay Sinir Ağları: Yapay sinir ağları, insan beyninin özelliklerinden olan öğrenme yolu ile yeni bilgiler türetebilme, yeni bilgiler oluşturabilme ve keşfedebilme gibi yetenekleri herhangi bir yardım almadan otomatik olarak gerçekleştirmek amacı ile geliştirilen bilgisayar sistemleridir. Bu yetenekleri geleneksel programlama yöntemleri ile gerçekleştirmek oldukça zordur veya mümkün değildir. O nedenle, yapay sinir ağları bilim dalının, programlanması çok zor veya mümkün olmayan olaylar için geliştirilmiş adaptif bilgi işleme ile ilgilenen bir bilgisayar bilim dalı olduğu söylenebilir.

Yapay sinir ağları, olayların örneklerine bakmakta, onlardan ilgili olay hakkında genellemeler yapmakta, bilgiler toplamakta ve daha sonra hiç görmediği örnekler ile karışılışınca öğrendiği bilgileri kullanarak o örnekler hakkında karar verebilmektedir.

1990'lı yıllardan beri bilgisayarların öğrenmesini sağlayan Yapay Sinir Ağları teknolojisinde oldukça hızlı bir gelişme görüldü. Yapay sinir ağları, insan beyninin özelliklerinden olan öğrenme yolu ile yeni bilgiler türetebilme, yeni bilgiler oluşturabilme ve keşfedebilme gibi yetenekleri herhangi bir yardım almadan otomatik olarak gerçekleştirmek amacı ile geliştirilen bilgisayar sistemleri olduklarından hem yeni gelişmelere neden oldular hem de nasıl çalıştığı bilinmeyen insan beyni hakkında yapılan araştırmalara da önemli katkılar sağladılar ve sağlamaktadırlar. Yapay sinir ağları öğretmenli öğrenmenin yanı sıra öğretmensiz öğrenme konusunda da uygulanabilen algoritmalara sahiptir.

Lojistik Regresyon Analizi: Lojistik regresyon analizi, son dönemlerde özellikle sosyal bilimler alanında kullanımı yaygınlaşan bir yöntemdir. Neden sonuç ilişkilerinin ortaya konulması amacıyla yapılan çoğu sosyo-ekonomik araştırmada, incelenen değişkenlerden bazıları olumlu-olumsuz, başarılı-basarisiz, evet-hayır, memnun-memnun değil şeklinde iki düzeyli (kategorili) verilerden oluşmaktadır. Lojistik regresyon analizi, bağımlı değişkenin kategorik ve bağımsız değişkenlerin karışık ölçekli olması durumunda belirli bir dağılım varsayımına bağlı kalmaksızın bağımlı değişken ile bağımsız değişkenler arasındaki neden sonuç ilişkisinin belirlenmesinde ve aynı zamanda bağımsız değişkenlerin etkilerine dayanarak verilerin sınıflandırılmasında kullanılan bir yöntemdir. Bağımlı değişkenin 0 ve 1 gibi iki veya ikiden fazla düzey içeren kesikli bir değişken olması durumunda da uygulanabilir olmasının yanında, matematiksel olarak esnekliği ve kolay yorumlanabilirliği, bu yönüne olan ilgiyi arttırmaktadır.

Bayes Sınıflandırma Yöntemi: Kesim 1.7.2'de açıklandığı üzere Bayes sınıflandırma yöntemi Bayes teoremine dayalı bir sınıflandırma tekniğidir. Bayes teoreminden faydalanan Bayes sınıflayıcıları verilen bir örneğin özel bir sınıfa ait olma olasılığı gibi sınıf üyelik olasılıklarını tahmin edebilirler. Bayes sınıflandırma tekniği öğretmenli öğrenme yöntemleri arasındadır.

K-En Yakın Komşu Algoritması: Bu teknikte tüm örneklem bir örüntü uzayında saklanır. Algoritma, bilinmeyen bir örneklemin hangi sınıfa dahil olduğunu belirlemek için örüntü uzayını araştırarak bilinmeyen örnekleme en yakın olan k örneklemini bulur.

bilinmeyen örneklem, k en yakın komşu içinden en çok benzediği sınıfa atanır. K-en yakın komşu algoritması, aynı zamanda, bilinmeyen örneklem için bir gerçek değerin tahmininde de kullanılabilir (Toktaş ve Demirhan, 2004, s.5).

En yaygın biçimde kullanılan algoritmalarından birisidir. Sınıflandırma yapılırken veritabanındaki her bir kaydın diğer kayıtlarla olan uzaklığını hesaplar. Ancak, bir kayıt için diğer kayıtlardan sadece k adedi göz önüne alınır. Algoritmanın isminden de anlaşılacağı gibi, bu k adet kayıt, başka bir deyişle veritabanındaki nokta; mesafesi hesaplanan noktaya diğer kayıtlara nazaran en yakın olan kayıtlardır. Bu yöntem coğrafi bilgi sistemlerinde çok kullanılır; belirlenen bir noktaya en yakın şehir, istasyon vs belirlenmesi aslında k-en yakın komşu algoritmasının temelini oluşturur.

Algoritmada k değeri önceden seçilir; değerinin yüksek olması birbirlerine benzemeyen noktaların bir araya toplanmasına, çok küçük seçilmesiyle birbirine benzediği, yani aynı sınıfın noktaları oldukları halde; bazı noktaların ayrı sınıflara konmasına ya da o tür noktalar için ayrı sınıfların açılmasına neden olur. K-en yakın komşu algoritmasında i ve j noktaları arasındaki uzaklık, p nitelik, k'da göz önüne

alınacak komşu sayısını göstermek üzere, $d(i, j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$ öklid uzaklık formülü ile tanımlanır.

Öklit uzaklığı çok sık kullanılsa da, bir değişkenin ölçü birimine çok duyarlı olduğu için uzaklıkların sıralanması üzerinde önemli etkilere sahiptir. Ölçü birimine bağımlılık giderilebilmekle birlikte, bunun giderilmesi de diğer bazı problemlere neden olmaktadır.

Karar Ağacı Öğrenmesi: Tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağaçları, veri madenciliğinde hem kuruluşlarının hem de yorumlanmalarının kolay olması, veri tabanı sistemlerine kolayca entegre edilebilmeleri, güvenilirliklerinin daha iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip olan bir tekniktir (Olafsson, 2008, s.1436). Büyük veri tabanlarının kullanıldığı pek çok sınıflama probleminde ve karmaşık ya da hata içeren bilgilerde karar ağaçları yararlı bir çözüm olmaktadır (Aitkenhead, 2008, s.18). Tümevarımcı çıkarsama konusundaki

en yaygın kullanılan yöntemlerden birisidir (Mitra, 2003, s.181). Karar ağaçları öğrenmesi, bir kesikli hedef fonksiyonu yaklaştırması yöntemidir. Burada sözü edilen yöntem karar ağacı ile gösterilen bir öğrenme fonksiyonunu ortaya koyar. Öğrenme sürecini tamamlamış ağaçlar konuşma diline yakın “if-then” kuralları biçiminde de ifade edilebilir. Tümevarımcı çıkarım yöntemleri arasında yer alan bu öğrenme yöntemleri tıptan kredi kartı risk analizlerine kadar pek çok uygulamada başarıyla uygulanmıştır (Mitchell, 1997, s.53). Bu çalışmanın konusunu oluşturan Regresyon Ağaçları yöntemi de bir karar ağacı öğrenmesi olarak değerlendirilir.

2.2. KARAR AĞAÇLARININ ELDE EDİLMESİ SÜRECİ

Karar ağaçları, sınıfları bilinen örnek veriden tümevarım yöntemiyle öğrenilen ağaç şekilli bir karar yapısı çeşididir (Sun ve Li, 2008, s.2). Bir karar ağacı, basit karar verme adımları uygulanarak, büyük miktarlardaki kayıtları, çok küçük kayıt gruplarına bölerek kullanılan bir yapıdır. Her başarılı bölme işlemiyle, sonuç gruplarının üyeleri bir diğeriyle çok daha benzer hale gelmektedir (Berry, 2004, s.194).

Karar ağaçları Bierman ve Friedman tarafından 1973 yılında önerilmiş olup değişkenleri parçalayarak bir ağaç oluşturmaya dayanmaktadır. Karar ağacında, tanımlanan sorunun cevabı gruplara ayrılmaktadır. Soruya verilecek bir ölçüt belirlendikten sonra kümeler arasındaki riski maksimize edecek şekilde cevaplar bölünmektedir. En iyi bölünmeyi bulmak için her soruda bu işlem tekrar edilmektedir. Bir soru için grup oluşturulduktan ve gruplar arasındaki risk maksimize edildikten sonra oluşan iki grup için aynı işlemler devam ettirilmektedir. Bu işlemler istatistik olarak anlamlı bir fark bulunana kadar devam ettirilip istatistik olarak anlamlı bir fark bulunmadığı durumlarda ise son verilmektedir. Ayırıştırma işlemi tamamlandıktan sonra ise o grup içerisinde yer alan gözlemlerin oranına göre grup değerlendirilmektedir (Thomas, 2000, s.159).

Büyük hacimli verileri kullanarak ona uygun bir karar ağacının elde edilmesi amacıyla birçok yöntem geliştirilmiştir. Regresyon ağaçları yöntemi de bunlardan biridir. Bağımsız değişken/değişkenler ile bağımlı değişken arasındaki ilişkiyi inceleyen regresyon analizi (basit ve çoklu), bir takım varsayımların (doğrusallık, normallik, homojenlik, toplanabilirlik gibi) yerine getirilmesinden sonra uygulanabilen bir test

istatistiğidir. Varsayımların yerine gelmemesi durumunda bilinen bir takım dönüştürme (transformation) işlemlerine tabi tutulacak olan veri kümesi, uygun hale getirilmeye çalışılmaktadır. Bu da veri setindeki orijinal değerlerin ya logaritmik dönüşümleri, ya da karekök gibi dönüşüm yöntemleri ile yapılabilir. Parametrik yöntemlerde her ne kadar dönüştürme metotları kullanılarak ön koşul varsayımlar yerine getirilmeye çalışılıyorsa da, yapılan analiz dahilinde veri setine ilişkin yanlış (biased) sonuçlar elde edilmesi söz konusu olabilmektedir. Bu nedenle alternatifin olmadığı durumlarda başvurulması verileri dönüştürme yoluna gidilmesi istatistiki açıdan daha doğrudur. Bahsedilen avantajlı yönleriyle çoklu regresyon analizine alternatif sayılabilecek ve çoklu regresyon analizinin gerektirdiği bir takım varsayımları taşımayan Sınıflandırma ve Regresyon Ağacı'nın (Classification and Regression Tree; CART) kullanılması uygun olur.

Sınıflama ve regresyon ağacı, sadece bağımlı değişken ile bağımsız değişken arasındaki ilişkinin yapısını araştırmakla kalmaz; aynı zamanda bağımsız değişkenlerin birbirleri ile olan etkileşimlerini de ortaya koymaya çalışır. Sınıflama ve regresyon ağacı analizinin kullandığı güçlü algoritma, bağımsız değişkenlerin bağımlı değişkenle ilişkisini değerlendirmede ve model içindeki etkileşim yapısını çözümlemede önemli avantajları mevcuttur. Sınıflama ve regresyon ağacının sahip olduğu algoritma, benzerlik gösteren (similarity) değişkenlerin aynı ağaç düğümünde toplanmasına dayalı olup, bütün oluşturduğu alt dalları (sub-branches) bağımlı değişken olan kök düğüme bağlamayla son bulmaktadır.

Kullanılan veri çeşidine göre parametrik veya parametrik olmayan yöntemler grubunda yer alabilen regresyon ağacı yönteminde, bağımsız değişkenlerin kesikli veya sürekli olması önem arz etmediği gibi, parametrik testlerin önemsendiği normallik, homojenlik ve doğrusallık gibi varsayımları da dikkate almamaktadır. Araştırmacılara sağladığı bu avantajlarla birlikte elde ettiği sonuçların diyagram biçimindeki sunumu, tabloyla gösterimin ötesinde diğer araştırmacılar için açıklayıcı olmaktadır (Kayri ve Boysan, 2008, ss.168-170).

2.2.1. Örneklerin Kullanımı

Sınıflandırma ağaçlarının oluşturulabilmesi için öncelikle girdi verisine, yani örneklere gereksinim vardır. Örnekler sayısal ya da kategorik niteliklerden oluşabilir. Bunun yanı sıra bir adet hedef niteliğe gereksinim vardır. Hedef nitelik çıktı değerlerini içerir.

Sınıflandırma sürecinde sınıflandırma ağacının eğitiminde kullanılacak verinin seçimi gerekir. Temizlenerek kullanılmaya hazır hale getirilmiş tüm veri ağacın eğitilmesinde kullanılmaz. Aksi takdirde elde edilecek ağacın doğruluğunu test etme olanağı yoktur. Dolayısıyla, ilgili veri kümesi ikiye bölünür. Bu kümelerden biriyle (eğitim kümesi) ağaç oluşturulurken, diğeri (test kümesi) ağacın test edilmesi işleminde kullanılır. Böylelikle bu karar mekanizmasına ne kadar güvenileceği bilgisi elde edilmiş olur. Bu ikili bölme yönteminin yanında başka yöntemlerde araştırılmıştır. Çünkü veri kümesinin bölünecek kadar büyük olmaması karar ağacını test etmede önemli bir sorunlar yaratmaktadır.

Çok daha büyük bir veri kümesinin alt kümesi olduğu kabul edilen Tablo 2.1'deki veri, bir eğitim kümesi olarak kabul edilebilir. Bu veri kümesi, kasko sigortası yaptıranların cinsiyetleri, meslekleri ve yaşları ile olan ilişkisini karar ağacı yoluyla çözümlemek üzere kullanılacaktır.

Sekiz adet gözlemden oluşan Tablo 2.1'deki bu veri kümesinde “Cinsiyet”, “Meslek” ve “Yaş¹” birer kategorik nitelik olarak yer almaktadır. Hedef ise, yine kategorik olan “Kasko” isimli niteliktir. Kasko niteliğinin kasko yaptıranlar için “Evet”, yaptırmayanlar için “Hayır” olmak üzere iki değeri vardır.

¹ Yaş aslında kategorik olmayan bir özellik olmakla birlikte burada yaşlı, genç ayrımına gidildiği için kategorik hale dönüştürülmüştür.

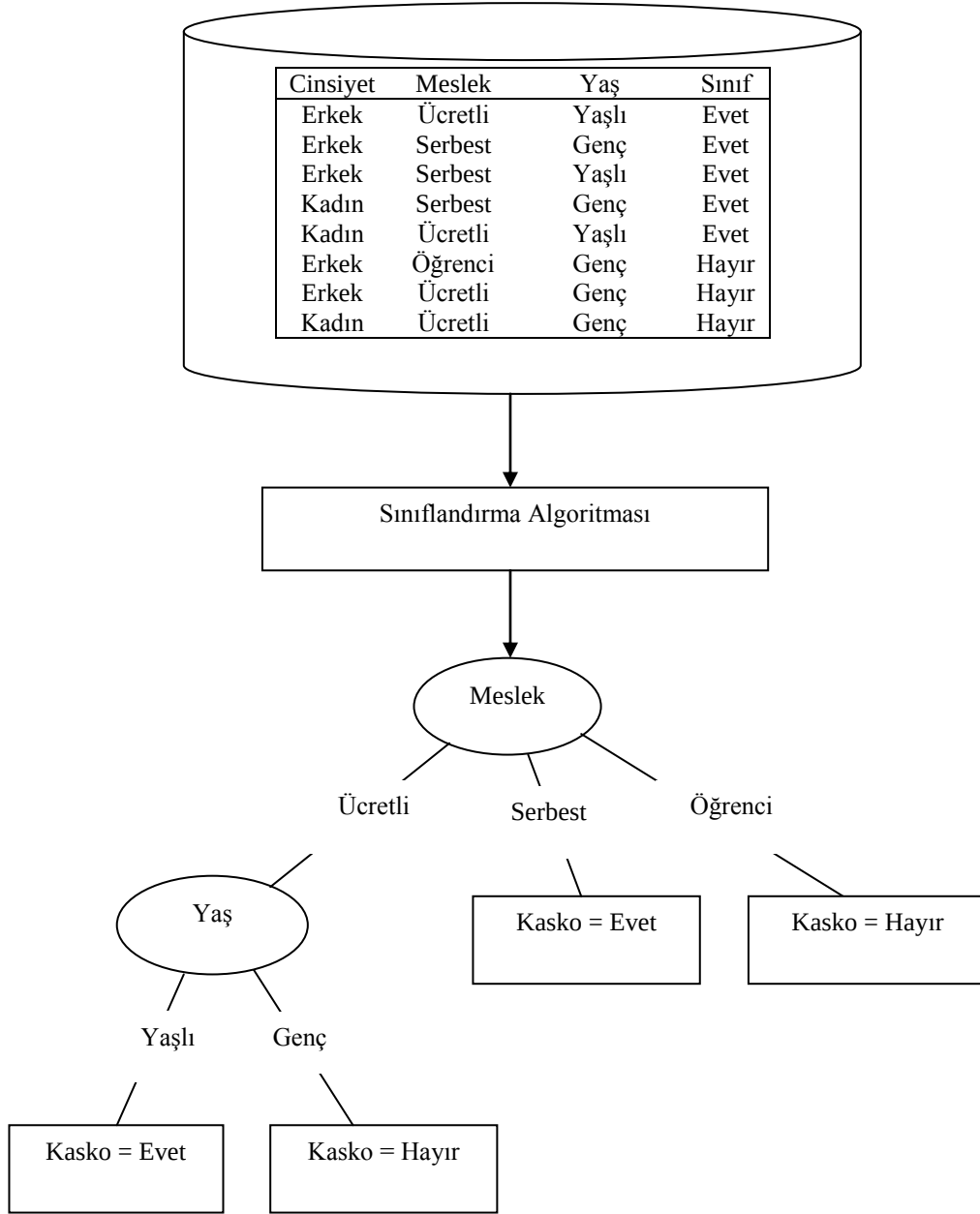
Tablo 2.1. Sınıflandırma Ağacına Esas Oluşturacak Veri.

Cinsiyet	Meslek	Yaş	Kasko
Erkek	Ücretli	Yaşlı	Evet
Erkek	Serbest	Genç	Evet
Erkek	Serbest	Yaşlı	Evet
Kadın	Serbest	Genç	Evet
Kadın	Ücretli	Yaşlı	Evet
Erkek	Öğrenci	Genç	Hayır
Erkek	Ücretli	Genç	Hayır
Kadın	Ücretli	Genç	Hayır

2.2.2. Öğrenme Süreci

Eğitim için ayrılan veri kümesi ile önce sınıflandırma ağacı ve onun bir sonucu olarak karar kuralları türetilir. Karar ağacı eğitim verisinden hareketle bir öğrenme olayı sonucunda oluşturulur.

Karar ağaçlarının oluşturulması için çok sayıda veri madenciliği algoritması (C5.0, CHAID, ID.3, C4.5, QUEST, vb.) geliştirilmiştir. Karar ağaçları oluşturulurken kullanılan algoritmanın ne olduğu önemlidir. Kullanılan algoritmaya göre ağacın şekli değişebilir. Değişik ağaç yapıları da farklı sınıflandırma sonuçları verecektir. Kök denilen ilk düğümü oluşturan A_i 'nin farklı olması, en uçtaki yaprağa ulaşılırken izlenecek yolu ve dolayısıyla sınıflandırmayı da değiştirecektir. Gerek kök düğümün gerekse de bundan sonraki her bir düğümün belirlenmesinde en büyük kriter o noktadan dallara ayrıldığında veritabanının geri kalan kısmı belli eşit parçalara bölünmüş olmasıdır. Örneğin veritabanında bulunan cevap evet /hayır gibiyse iki eşit parçaya, evet/hayır/belki gibi üç değişkenliyse mümkün olduğunca üç eşit parçaya bölünmesi istenmektedir. Burada amaç, en kısa yoldan istenen yanıt ya da sınıfa ulaşmaktır. Bu algoritmalar yardımıyla bir karar ağacı oluşturulduğunda öğrenme olayı da gerçekleşmiş olur. Şekil 2.3 bu süreci göstermektedir.



Şekil 2.3. Sınıflandırma Ağaçlarında Öğrenme Süreci.

Karar kuralları, karar ağacının kökten başlayarak yapraklara doğru yorumu yapılarak elde edilir. Elde edilen karar kuralları karar verme sürecinde kullanılır. Şekil 2.3’de açıklandığı gibi, bir veri kümesine sınıflandırma algoritması uygulanarak karar ağacı elde edilmektedir. Eğitilmiş olan bu karar ağacı kullanılarak yeni bir örnek hakkında karar verilebilir.

2.2.3. Karar Ağaçlarının Elde Edilmesi

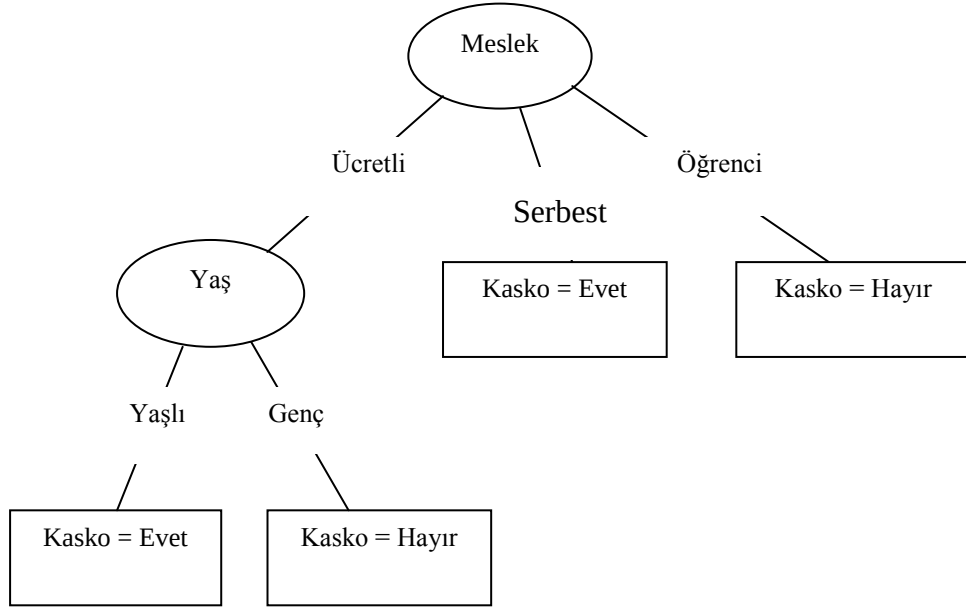
Bir karar sürecinde seçenekler, olaylar ve bunlara bağlı olarak elde edilecek sonuçların grafik biçiminde gösterilmesi sonucu ortaya çıkan grafiğe “karar ağacı” denir. Karar ağaçları veri madenciliğinde sınıflandırma uygulamalarında yaygın biçimde kullanılmaktadır. İsimden de anlaşılacağı gibi, karar ağaçları kök, dallar ve yapraklardan oluşan bir ağaç görünümündedir.

Karar ağaçları ile sınıflandırma süreci kök düğümden başlar. Ağacın her bir düğümü bir niteliği tanımlar. Düğümlerden niteliğin her bir değerine göre dallar ayrılır. Bu dallanma süreci yapraklar elde edilinceye dek devam eder. Karar ağacı bilginin en iyi gösterim biçimlerinden birisi olmasına karşın, bazı durumlarda karmaşık ve yorumlaması oldukça zor olan bir yapıya sahip olabilir (Koza, 2002, s.9). Bir karar ağacında sadece bir kök düğüm olabilir. Diğer simgeler için bir sayı sınırlaması yoktur.

Şekil 2.4’deki karar ağacı Tablo 2.1’de verilen örneklerle dayalı olarak oluşturulmuştur. Karar ağacının birinci düğümü (kök düğüm) “Meslek” olarak seçilmiştir. Bu seçim rasgele yapılmaz. Uygun bir veri madenciliği algoritması yardımıyla seçim işlemleri yürütülür. Kök düğümden söz konusu niteliğin değerleri olan “Ücretli”, “Serbest” ve “Öğrenci” dalları ayrılmaktadır.

Şekil 2.4’de yer alan karar ağacında, kök düğümden çıkan her bir dal için yeniden bir seçim işlemi uygulanarak uygun bir düğüm elde edilmiştir. Örneğin bu ağaç üzerinde “Ücretli” dalına bağlı olarak “Yaş” düğümü elde edilmiştir. Benzer biçimde “Yaş” düğümünden itibaren bu niteliğin değerleri olan “Yaşlı” ve “Genç” dalları ayrılmaktadır. Sonuç olarak dikdörtgen biçiminde ifade edilen yapraklara ulaşılmaktadır. Böylece karar ağacı elde etme süreci sona ermiştir. Örnekteki tüm niteliklerin karar ağacında yer alması beklenmemelidir. Nitekim örnek verideki “Cinsiyet” niteliği karar ağacında yer almamıştır.

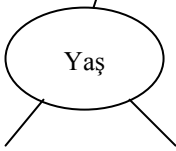
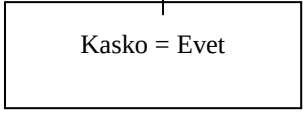
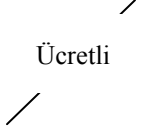
Yapraklar karar sınıflarını belirlemektedir (Witten, 2005, s.62). Düğümler ise test işlemine tabi tutulan nitelikleri göstermektedir (Quinlan, 1993, s.5). Karar ağaçlarında yer alan bazı bileşenlerin ne anlama geldiği Tablo 2.2’ de açıklanmıştır.



Şekil 2.4. Bir Karar Ağacının Görünümü.

Tablo 2.2’de verildiği gibi karar ağaçlarında düğümler elipslerle, yapraklar dikdörtgenlerle, dallarsa düğüm ve yaprakları birleştiren çizgilerle gösterilir.

Tablo 2.2. Karar Ağacı Bileşenleri.

Bileşen	İşlevi
	Düğüm
	Yaprak
	Nitelik değerine göre bir dal

2.2.4. Karar Ağaçlarının Avantajları Ve Dezavantajları

Her tekniğin kendine göre avantaj ve dezavantajları vardır. Karar ağaçlarını, diğer yöntemlere göre daha avantajlı kılan bazı özellikler şu şekilde sıralanabilir:

- Karar ağaçları çok sayıda nitelik sayısına ve çok büyük hacimli veri kümelerine kolayca uygulanabilir. Bunun sonucunda daha hızlı uygulama geliştirme olanağı elde edilmiş olur Elde edilen sonuçların yorumu, diğer yöntemlere göre daha kolay ve kullanışlıdır (Mitra, 2003, s.187).
- Parametrik olmayan yöntemler arasında olması nedeniyle diğer çok değişkenli tekniklerin gerektirdiği istatistiksel varsayımlar söz konusu değildir.
- İlişkilerin yönünü ve önem sırasını görsel olarak ortaya koyması bir diğer avantajdır (Albayrak ve Yılmaz, 2009, s.42).
- Hatalı verilere sahip veri kümelerinin yanı sıra eksik değerler içeren veri kümeleri ile çalışma olanağı sağlar (Maimon ve Rokach, 2005, s.183).
- Sürekli değere sahip niteliklerin yanı sıra kategorik niteliklerin de kullanılması mümkündür. Karmaşık ilişkiler ağaç yapısı ile daha anlaşılır biçimde ortaya konulmaktadır (Myatt, 2007, s.141).

Karar ağaçlarının dezavantajlarını da göz önüne almak gerekir. Bunlar aşağıdaki gibi sıralanabilir.

- Büyük hacimli verilerin kullanıldığı ve çok sayıda sayısal değişkenin yer aldığı uygulamalarda, karar ağaçlarının elde edilmesi diğer sınıflandırma yöntemlerine göre daha fazla hesaplama maliyeti gerektirebilir (Myatt, 2007, s.141).
- Karar ağaçları için geliştirilen bazı algoritmalar (ID3 gibi) sadece kategorik niteliklerle çalışır (Maimon ve Rokach, 2005, s.185). Bu husus, söz konusu algoritmanın bir sınırlaması olarak değerlendirilebilir.

2.3. KARAR KURALLARI VE SINIFLANDIRICI DOĞRULUĞU

2.3.1. Karar Kurallarının Belirlenmesi

Karar kurallarının türetilmesi işlemine kök düğümden başlanır ve tüm dallar ve yapraklar değerlendirilir. Böylece karar kuralları eksiksiz biçimde tanımlanmış olur. Karar ağacının en ucundaki yaprak sayısı kadar karar kuralı çıkartılacağı aşıkardır (Witten 2005, s.62). Şekil 2.4'deki karar ağacı esas alınarak türetilen karar kuralları aşağıda sıralanmıştır.

Kural 1: (Cinsiyet ne olursa olsun)

EĞER MESLEK = Ücretli VE

YAŞ = Yaşlı İSE

KASKO = EVET;

Kural 2: (Cinsiyet ne olursa olsun)

EĞER MESLEK = Ücretli VE

YAŞ = Genç İSE

KASKO = Hayır;

Kural 3: (Cinsiyet ve yaş ne olursa olsun)

EĞER MESLEK = Serbest İSE

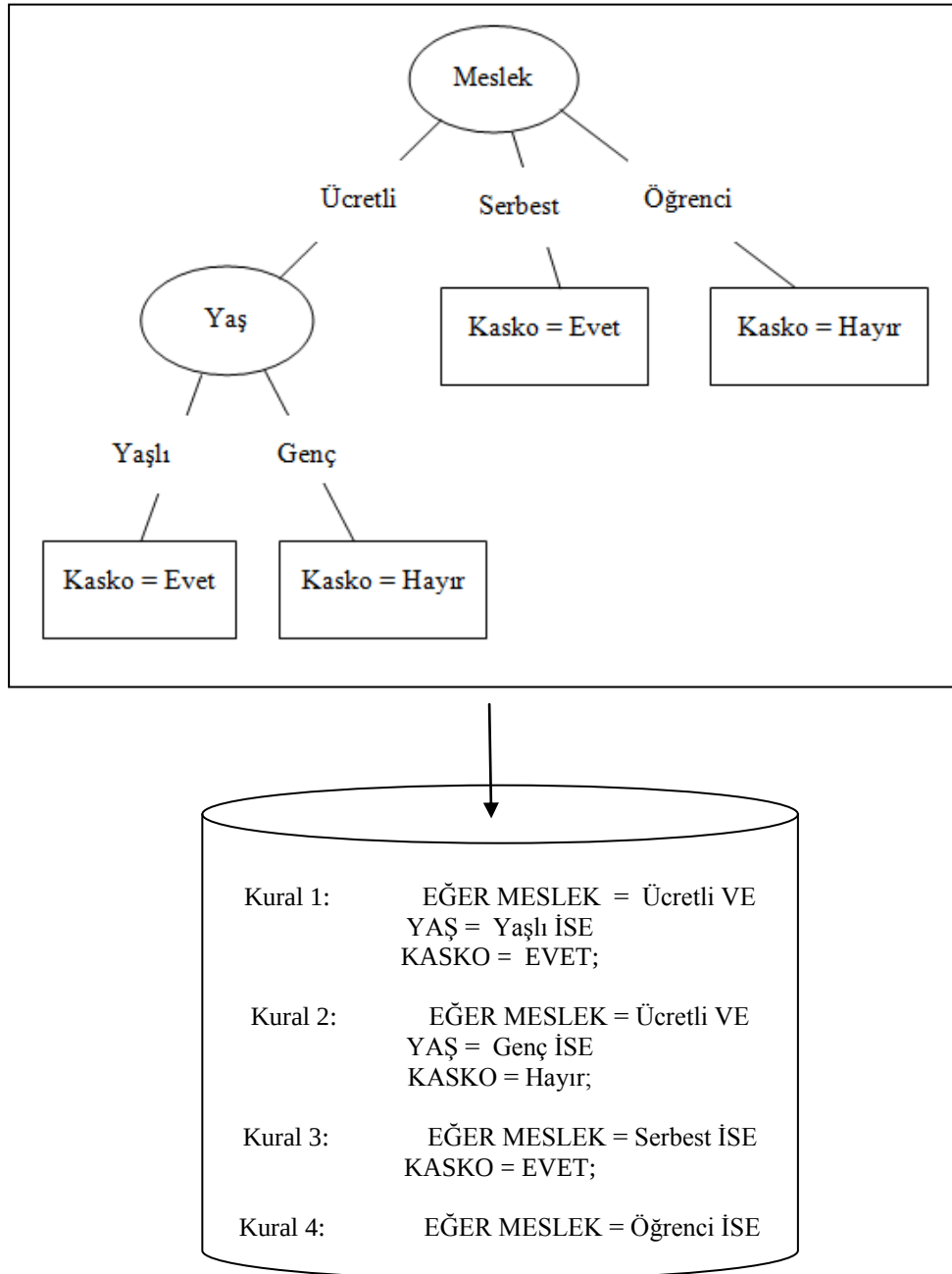
KASKO = EVET;

Kural 4: (Cinsiyet ve yaş ne olursa olsun)

EĞER MESLEK = Öğrenci İSE

KASKO = Hayır;

Karar kuralları bilgisayar programlama dillerinde kullanılan “IF ... THEN ... ELSE ...” yapısına benzemektedir. Bu yapı, elde edilen kuralların yeni veriye uygulanmasında kolaylık sağlar. Şekil 2.5 karar kurallarının elde edilme sürecini ifade etmektedir.



Şekil 2.5. Karar Kurallarının Elde Edilme Süreci.

2.3.2. Kuralların Geçerliliğini Doğrulama

Karar ağacının eğitilmesi, yani öğrenme işlemi eğitim kümesi ile yapılır. Eğer elde edilen karar kuralları test kümesi için geçerli değilse, bu kuralların kestirim işinde kullanılması söz konusu olamaz. Kestirim yapabilmek için belirlenen karar ağacının ve buna bağlı karar kurallarının başarılı olduğunun kanıtlanması gerekir.

2.3.2.1. Sınıflandırıcı Doğruluğu

Eğitim kümesi verisi ile eğitilen ağacın geçerliliği test edildikten sonra, sonuçlar olumlu ise bir sonraki adımda karar ağacı yeni değerlerin kestirilmesinde (Şekil 2.5’de açıklandığı gibi) kullanılır.

Sürekli değerlerin kestirimini yapmak için istatistiksel yöntemlerden geniş ölçüde yararlanılır. Tek değişkenli veya çok değişkenli regresyon modelleri bu tür kestirim modelleri oluşturmak için kullanılabilir. Örneğin, yeni çıkacak bir ürüne olan talebi tahmin etmek için geçmiş yılların ücret seviyesi bilgilerinden yararlanılarak bir regresyon modeli oluşturulabilir. Bu model yardımıyla, bu ürün için belirlenen bir fiyatın satışlar üzerindeki etkisi ortaya konulabilir.

Sınıflandırma işlemlerinde, eğitim verisi üzerinde bir sınıflandırma algoritması uygulanarak sınıflandırıcıya ulaşılır. Ancak doğal olarak sınıflandırıcı doğruluğunun belirlenmesi önem taşımaktadır. Sınıflandırma işlemi sınıflandırıcı yardımıyla yapılır. Veri kümesini kullanarak bir sınıflandırıcı elde edildikten sonra bu sınıflandırıcının doğruluğunun sınanması gerekir. Bu amaç için birçok yöntem geliştirilmiştir. Holdout, Çapraz-doğrulama ve Bootstrap yaygın biçimde kullanılan doğrulama yöntemleridir (Han ve Kamber, 2006, s.359).

2.3.2.2. Doğruluk Değeri

Bir sınıflandırma modelinin doğruluğunun belirlenmesi katedilen süreçteki en önemli adımlardan birisidir. Model, eğitim verisine dayalı olarak oluşturulduktan sonra sınıfları veya test veritabanı değerlerini tahmin etmek için kullanılır. Test verisi üzerinde modelin çalıştırılmasından sonra elde edilen sonuçlar değeri bilinen gerçek verilerle karşılaştırılır. Yanlış tahmin edilen veya sınıflanan örnek sayısının toplam test örnek sayısına oranı modelin hata oranını verir. Benzer olarak doğru sınıflanan veya tahmin edilen verinin toplam test örnek sayısına oranı da doğruluk oranını verir. Bir başka şekilde ifade edilirse, “doğruluk oranı = 1 - (hata oranı)” şeklinde yazılabilir (Akpınar, 2000, s.6).

“ ℓ ” sayıda gözlemden oluşan eğitim kümesi $\mathcal{L}$, “ t ” sayıda gözlemden oluşan test kümesi $\mathcal{T}$ olmak üzere, veri kümesindeki sınıf değerlerini içeren nitelik y_i ile hedef nitelik dışında kalan nitelikler X_i ile ifade edilsin. Sınıflandırma modeli ise, X_i gözlem değerlerini kullanarak y_i değerlerini kestirmek üzere $X_i \in \mathcal{T}$ için $f(X_i)$ biçiminde elde edilen bir fonksiyondur. Bu fonksiyon sınıflandırma algoritması tarafından belirlenen A modelini temsil etmektedir. A modeli için $L(y_i, f(X_i))$ kayıp fonksiyonu şu şekilde yazılabilir (Vercellis, 2009, s.227):

$$L(y_i, f(X_i)) = \begin{cases} 0, & y_i = f(X_i) \text{ ise,} \\ 1, & y_i \neq f(X_i) \text{ ise.} \end{cases} \quad (2.1)$$

Bu durumda A modelinin test kümesi ile ilgili doğruluk değeri $acc_A(\mathcal{T})$ şu şekilde ifade edilir:

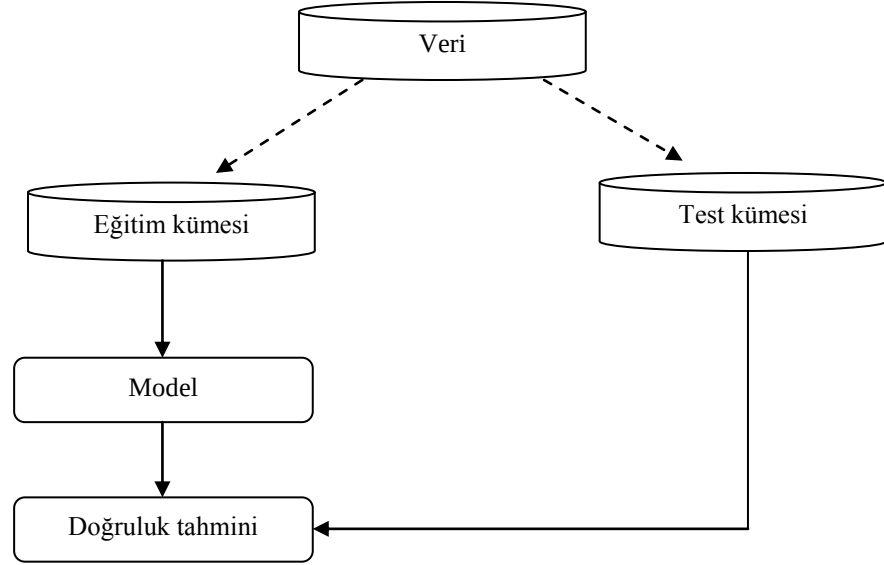
$$acc_A(\mathcal{T}) = 1 - \frac{1}{t} \sum_{i=1}^t L(y_i, f(X_i)) \quad (2.2)$$

Bazı durumlarda alternatif performans göstergesi olarak, sınıflandırma algoritmasının $\mathcal{T}$ verilerine uygulanması sonucunda ortaya çıkan hata oranı ($err_A(\mathcal{T})$) hata oranı kullanılır:

$$err_A(\mathcal{T}) = 1 - acc_A(\mathcal{T}) = \frac{1}{t} \sum_{i=1}^t L(y_i, f(X_i)) \quad (2.3)$$

Holdout Yöntemi İle Doğrulama: Holdout yönteminde veri, Şekil 2.6’da gösterildiği biçimde eğitim ve test kümesi olarak iki ayrı kümeye bölümlendirilir. Arkasından eğitim kümesinde bir sınıflandırma modeli türetilir ve performansı test kümesinden

ölçülür. Eğitim ve test amaçlı olarak seçilen verinin oranı değişmekle birlikte genellikle verinin 2/3'ü sınıflandırıcının elde edilmesi, kalan 1/3'ü test amacıyla kullanılır. Sınıflandırıcının doğruluğu, test kümesinden elde edilen modelin doğruluğundan çıkarılabilir.



Şekil 2.6. Holdout Yöntemi İle Doğruluğun Tahmin Edilmesi.

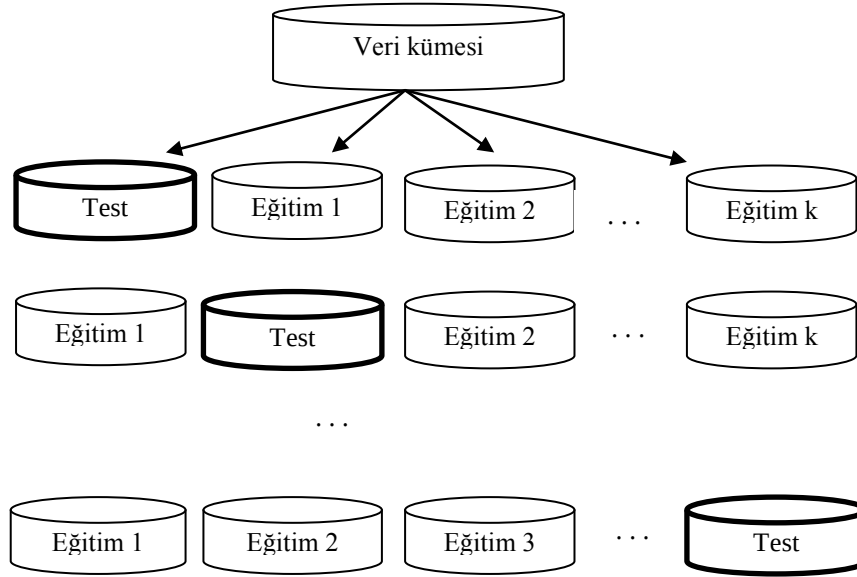
Holdout yönteminin birkaç iyi bilinen sınırlaması vardır. İlki, az etiketlenen örnekler için uygundur, çünkü bazı kayıtlar test için ayrılmıştır. Sonuçta, çıkarılan model, tüm etiketlenen örnekler eğitim için kullanıldığındaki gibi iyi olmayacaktır. İkincisi model eğitim ve test kümelerinin birleşimine iyi derecede bağlı olabilir. Eğitim kümesi ne kadar küçükse modelin değişkenliği o kadar büyük olacaktır. Diğer yandan eğitim kümesi çok büyük olursa, daha küçük olan test kümesinden hesaplanarak beklenen doğruluk daha az güvenilir olacaktır.

Holdout yöntemi sınıflandırıcının performans tahmini için birkaç kez tekrarlanabilir. Yöntemin tekrarlanması sonucunda “tekrarlı holdout yöntemi” veya “tekrarlı örnekleme” adı verilen yöntem elde edilir. Her bir tekrarda, ℓ gözlemden oluşan bir rasgele $\mathcal{L}_k$ eğitim örneği elde edilir. Burada $\mathcal{T}_k = D - \mathcal{L}_k$ olmak üzere $acc_A(\mathcal{L}_k)$ doğruluk değeri belirlenir. Tekrarlardan oluşan sürecin sonunda A modelinin

acc_A doğruluk değeri örneklerin doğruluk değerleri ortalaması alınarak elde edilir (Vercellis, 2009, s.228). Tekrar sayısı r olmak üzere doğruluk değeri şöyledir.

$$acc_A = \frac{1}{r} \sum_{i=1}^r acc_A(\mathcal{T}_k) \quad (2.4)$$

K-Katlı Çapraz Doğrulama: Model doğruluğunu belirlemede kullanılan bir diğer yöntem “k-katlı çapraz doğrulama” adıyla bilinmektedir. Bu yöntemde gözlem kümesi analiz başlangıcında $s_1, s_2, \dots, s_k$ biçiminde tesadüfi k alt kümeye ayrılır. Uygulamada genellikle $k = 10$ seçilir ve alt kümelerin hemen hemen aynı boyutta olmasına özen gösterilir. Alt kümeler bu şekilde belirlendikten sonra, test kümesinin yeri değiştirilmek suretiyle analizler adım adım yapılır.



Şekil 2.7. K-katlı Çapraz Doğrulamada Eğitim ve Test Kümelerinin Belirlenmesi.

Şekil 2.7’de, birinci adımda s_1 kümesi test $s_2, \dots, s_k$ alt kümeleri ise eğitim kümeleri olarak seçilir. Sınıflandırma modeli için bu eğitim ve test verileri kullanılarak doğruluk değerleri bulunur. İkinci adımda ise s_2 kümesi test, $s_1, s_3, \dots, s_k$ alt kümeleri

ise eğitim kümeleri olarak seçilir ve benzer biçimde modelin doğruluğu hesaplanır. Bu şekilde işlemler k defa tekrarlanır (Han ve Kamber, 2006, s.364).

K-katlı çapraz doğrulama yöntemi, i. iterasyonda $\mathcal{L}_i$ eğitim kümesini, $\mathcal{T}_i$ test kümesini göstermek üzere matematiksel olarak şu şekilde ifade edilebilir (Vercellis, 2009, s.229):

$$\mathcal{T}_i = s_i \quad \mathcal{L}_i = \bigcup_{j \neq i} s_j \quad (2.5)$$

Sınıflandırma modelinin doğruluğu; k tekrar sonucu toplam kayıp olarak hesaplanan hata miktarının, toplam gözlem sayısına bölünmesiyle elde edilir (Han ve Kamber, 2006, s.364). Bu yöntemin en önemli sakıncası maliyetin k'ya bağlı olarak artmasıdır.

Bootstrap: Örnekleme, anakütle parametrelerinin tahmin edilmesinde kullanılan istatistiksel bir yöntemdir. Anakütleden çekilen örneklerin anakütleyi en iyi biçimde temsil edebilmesi için anakütlenin çerçevesi hakkında eksiksiz ve doğru bilgiye ihtiyaç vardır. Örnek seçiminde çeşitli yöntemlere başvurulabilir. Bunlar arasında en sık kullanılan ve olasılık yöntemlerine dayalı “Basit Rasgele Örnekleme”, “Tabakalı Örnekleme”, “Kümelere Göre Örnekleme”, “Sistemik Örnekleme” sayılabilir. Bütün bu yöntemler parametrik örnekleme yöntemleridir.

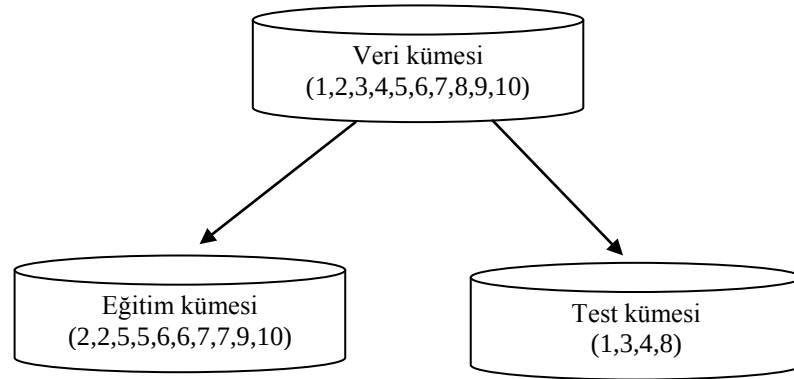
Parametrik istatistiksel yöntemlerin uygulanamadığı durumlarda test edilen hipotezdeki istatistiğin anlamlılığını ortaya koymak için kullanılan tekrarlı örnekleme yöntemleri içinde en iyi bilinen yöntemler “Bootstrap” ve “Jackknife” yöntemleridir.

1979 yılında Bradley Efron tarafından ortaya atılan Bootstrap örnekleme bilgisayar tabanlı bir yöntemdir (Efron, 1994, s.10). Genellikle küçük veri kümeleri ile çalışmak söz konusu olduğunda tercih edilen Bootstrap yönteminde, gözlenen veriler içinden her biri gözlenen verilerle aynı büyüklükte “B” adet iadeli örneklem oluşturulur. Yeni veri kümesinin her biri için bir istatistik (Aritmetik Ortalama, Mod, Medyan, vb.) hesaplanır. Hesaplanan istatistikler bir Bootstrap dağılım oluşturur. Amaç az sayıda

gözlemden çok sayıda veri türeterek bir dağılım oluşturmak ve bu dağılım yardımıyla anakütle parametrelerini tahmin edebilmektir.

N adet gözlemden oluşan veri kümesini $(x_1, x_2, \dots, x_N)$ ile ifade edelim. Bu veri kümesinden iadeli (yerine koyarak) seçimle her biri N birimli B adet $(x_1^*, x_2^*, \dots, x_N^*)$ bootstrap örneklemleri oluşturulur ve oluşturulan örneklemler eğitim kümesi olarak kabul edilir. Eğitim kümesine dahil edilmeyen gözlemler ise test kümesini oluştururlar (Witten, 2005, s.152).

On gözlemden oluşan veri kümesini “1, 2, 3, 4, 5, 6, 7, 8, 9, 10” ile gösterelim. Bu veri kümesinden iadeli seçimle belirlenen bootstrap örneklemlerinden “2, 2, 5, 5, 6, 6, 7, 7, 9, 10” olarak belirlenebilir. İşte bu örneklem bir eğitim kümesi oluşturur. Bu kümenin dışında kalan birimlerin oluşturduğu (1, 3, 4, 8) kümesi test kümesi olarak kabul edilir. Bu örnek Şekil 2.8’de gösterilmiştir.



Şekil 2.8. Bootstrap Yönteminde Eğitim Ve Test Kümesinin Seçilmesi İşlemi.

Bootstrap yöntemleri arasında en yaygın kullanılanı “.632 bootstrap” adıyla bilinmektedir. Yönteme ismini veren Tesadüfi olarak örneklenen Bootstrap veri kümesinin büyüklüğünün tespitinde modelin hata oranını en iyi temsil edecek oranın 0,632 olarak belirlenmiş olmasıdır. Bu yöntem, daha düşük varyansa sahip olması nedeniyle diğerlerine göre daha avantajlıdır (Fitzmaurice ve diğerleri, 239). Bu yöntemde N elemandan oluşan bir veri kümesinin verildiği varsayılır. Bu veri kümesindeki gözlemlerin 0.632’si bootstrap’ta eğitim kümesini, kalan 0.368’i test

kümesini oluşturur. $\mathcal{L}$ eğitim kümesi için her gözlemin seçilme olasılığı $\frac{1}{N}$ olarak hesaplanır. Bu durumda seçilmemiş olanların seçilme olasılığı $\left(1 - \frac{1}{N}\right)$ 'ye eşit olur. Seçim işlemi N kez tekrarlanır. Böylece bütün seçim süreci sonunda seçilmemiş olanların seçilme olasılığı $\left(1 - \frac{1}{N}\right)^N$ biçimindedir. Eğer N büyükse olasılık $e^{-1} \cong 0.368$ 'dir. Böylece tüm verinin 0.368'i test kümesinde yer alır; kalan 0.632'si eğitim kümesini oluşturur (Witten, 2005, s.153).

Bir M modelinin eğitim kümesi için hesaplanan doğruluğu $acc_M(\mathcal{L})$, test kümesi için hesaplanan doğruluğu $acc_M(\mathcal{T})$ ise, modelin doğruluğu,

$$acc_M = (0.632)acc_M(\mathcal{T}) + (0.368)acc_M(\mathcal{L}) \quad (2.6)$$

biçiminde elde edilir. Yukarıda açıklanan süreç k kez tekrarlanır. M modelinin doğruluk değeri yukarıda açıklanan sürecin k kez tekrarlanmasıyla aşağıdaki gibi hesaplanır (Han ve Kamber, 2006, s.365).

$$acc_M = \sum_{i=1}^k (0.632)acc_{M_i}(\mathcal{T}) + (0.368)acc_{M_i}(\mathcal{L}) \quad (2.6)$$

Burada $acc_{M_i}(\mathcal{L})$ değeri, model M'nin i'inci iterasyonda bootstrap eğitim kümesi ile elde edilen doğruluğunu, $acc_{M_i}(\mathcal{T})$ ise aynı modelin i'inci iterasyondaki test kümesi ile elde edilen doğruluk değerini ifade etmektedir.

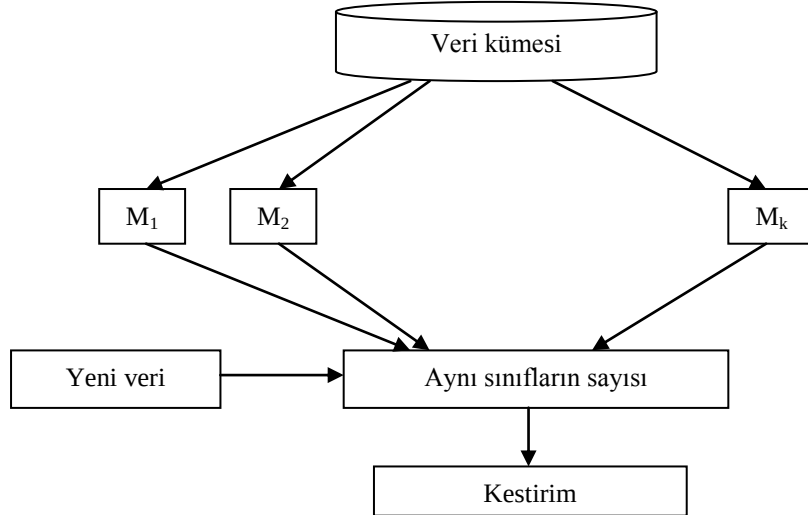
2.3.2.3. Doğruluğun Artırılması

Elde edilen sınıflandırıcının doğruluğunu artırmak üzere birçok yöntem geliştirilmiştir. Bunlar arasında “Bagging” ve “Boosting” yöntemleri sayılabilir. Bu yöntemler daha

önce elde edilen sınıflandırıcıların birleştirilmesi esasına dayandıklarından birleştirme (ensemble) yöntemleri başlığı altında incelenirler (Han ve Kamber, 2006, s.366). Birleştirme yöntemleri, gruplama yöntemlerini ya da kombinasyon modellerini kullanan yöntemlerdir. Öğrenme süreci sonunda elde edilen $M_1, M_2, \dots, M_k$ biçimindeki k model birleştirilerek model M^* yaratılır. Yeni veriyle kullanılmak üzere modelleri birleştirmek için oylama stratejileri uygulanır. Bagging ve boosting, kestirimin yanısıra sınıflandırma amacıyla da kullanılabilen yöntemlerdir (Han ve Kamber, 2006, s.366).

Bagging (Bootstrap Aggregating): 1994 yılında L. Breiman tarafından ortaya atılan Bagging yöntemi sınıflandırma ve regresyondaki kestirimin varyansını azaltmak amacıyla kullanılmaktadır. Bagging yönteminde, veri kümesine bootstrap uygulanır ve eğitim kümesi oluşturulur. Bu sürecin başlangıcıdır ve bunu diğer adımlar (Sınıflandırma modelinin bulunması ve diğerleri) izler.

Şekil 2.9’da gösterildiği gibi, model elde etme işlemi k kez tekrarlanır ve $M_1, M_2, \dots, M_k$ ile gösterilen k sayıda model elde edilir. Hangi sınıfta olduğu bilinmeyen yeni veriye bu modeller tek tek uygulanır ve sınıf kestirimleri elde edilir. Elde edilen bu sınıf değerleri sayılır ve en sık tekrarlanan sınıf belirlenerek veri sınıfı olarak atanır.



Şekil 2.9. Bagging Yöntemi İle Kestirim.

Boosting: Boosting de Bagging yöntemi gibi, bir öğrenme yönteminin doğruluğunu artırmak ve daha güçlü sınıflandırıcıları elde etmek amacıyla başvurulan bir yöntemdir (Han ve Kamber, 2006, s.366).

Boosting yönteminin çalışma biçimi şu şekilde bir olaya benzetilebilir: Bir yatırım şirketinin beş adet danışmanı olduğunu ve her danışmanın bir gün sonraki hisse senetleri kapanış fiyatları ile ilgili tahminler yaptığını varsayalım. Her danışman için doğru tahmin sayılarına bağlı olarak ağırlıklar belirlenmekte ve son tahmin ağırlıklı tahminlerin bir kombinasyonu olmaktadır.

Boosting konusunda çok sayıda algoritma geliştirilmiştir. Bunlar arasında Adaboost , LPBoost, TotalBoost, BrownBoost, MadaBoost ve LogitBoost algoritmaları sayılabilir. Bu algoritmalar içinde en sık kullanılanı Adaboost algoritmasıdır. Adaboost algoritmasında D'nin eğitim kümesi olduğunu varsayalım. Eğitim kümesinde “d” sınıf niteliğini gösterebilir. Her defasında bir sınıflandırıcı elde etmek üzere “k” değeri raundların, yani adımların sayısını gösterebilir. İadeli örnekleme ile $i = 1, 2, \dots, k$ için D_i örnekleri bulunur ve her örnek için M_i modeli elde edilir. Modellerin her biri için bir hata oranı hesaplanır. M_i modeli yanlış sınıflandırdığı için, D_i 'deki gözlemlerin her birinin ağırlığı toplanır (Han ve Kamber, 2006, s.368).

$$err(M_i) = \sum_j^d w_j err(X_j) \quad (2.7)$$

Bu bağıntıda $err(X_j)$ ifadesi X_j gözlemlerinin sınıflandırılma hatasıdır. Eğer yanlış sınıflandırılma söz konusu ise $err(X_j) = 1$, aksi halde 0 olur (Clarke ve diğr., 2009, s.318). Eğer M_i sınıflandırıcısının performansı düşükse, sınıflandırma hatası 0.5'i aşar ve bu durumda M_i 'den vazgeçilir. Vazgeçilen M_i yerine, yeni bir M_i 'den yeni bir D_i eğitim kümesi oluşturularak işlem tekrarlanır. M_i 'nin hata oranı eğitim kümesi ağırlıklarının güncellenmesini etkiler. Denemedeki bir gözlem doğru bir şekilde sınıflandırılmışsa, gözlemin ağırlığı,

$$\frac{err(M_i)}{1 - err(M_i)} \quad (2.8)$$

ile çarpılır. Doğru bir şekilde sınıflandırılmış tüm örneklerin ağırlığı güncellenir ve bütün örnekler, toplamlar değişmesin diye normalleştirilir. Sonuç olarak, yanlış sınıflandırılmış örneklerin ağırlıkları artırılır ve yukarıda açıklandığı gibi doğru sınıflandırılmış örneklerin ağırlıkları azaltılır. Bir sınıflandırıcının hata oranı ne kadar düşük olursa doğruluğu o derece yüksektir. Her bir M_i sınıflandırıcısının oylama ağırlığı w_i olmak üzere aşağıdaki gibi hesaplanır (Lin ve diğerleri, 2006, s.238):

$$w_i = \log \frac{1 - err(M_i)}{err(M_i)} \quad (2.9)$$

Yukarıdaki açıklamalar çerçevesinde Adaboost algoritması aşağıdaki gibi özetlenebilir. D eğitim kümesini, d sınıf etiketli eğitim kümesini “k” ise raundların sayısını ifade etmektedir. Algoritmanın çıktısı birleşik bir modeldir (Han ve Kamber, 2006, s.369).

Adaboost Algoritması:

- [1] D’deki her örneğin ağırlığını 1/d olarak başlatın.
- [2] i = 1 den başlamak üzere aşağıdaki işlemleri k defa tekrarlayın:
- [3] D_i ’yi elde edin. Bunun için D eğitim kümesinde iadeli örnekleme yöntemini kullanın.
- [4] D_i eğitim kümesini kullanarak M_i modelini elde edin.
- [5] $err(M_i) = \sum_j^d w_j err(X_j)$ hata oranını hesaplayın.
- [6] Eğer $err(M_i) > 0.5$ ise, o zaman;
- [7] 1/d’ye göre ağırlıkları yeniden başlatın.
- [8] 3.adıma geri dönün ve tekrar deneyin.
- [9] Doğru bir şekilde sınıflandırılmış D_i ’deki her satır için şunu yapın,
- [10] $\frac{err(M_i)}{1 - err(M_i)}$ ’ye göre satırın ağırlığını çarpın ve ağırlıkları güncelleyin.
- [11] Her satırın ağırlığını normalleştirin;

Birleşik modeli sınıflandırmada kullanmak için aşağıdaki algoritma kullanılır:

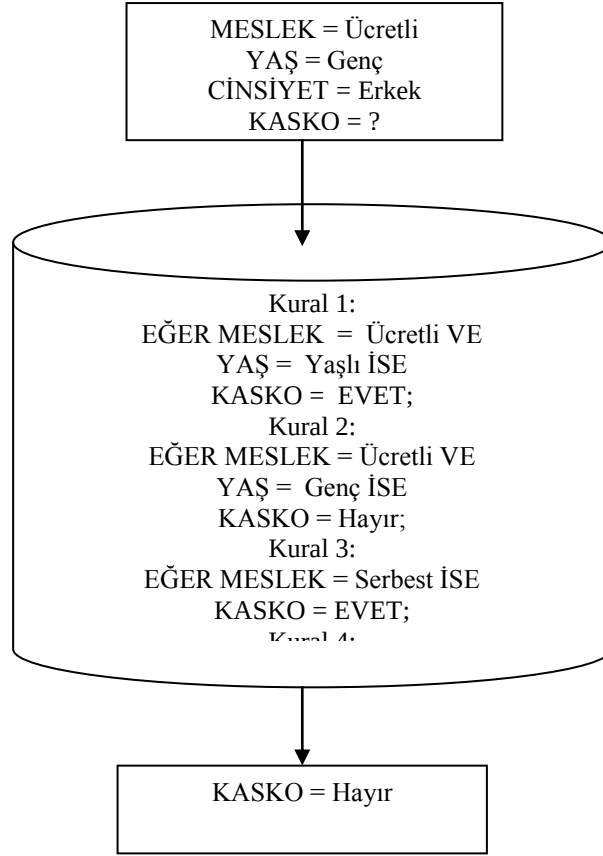
- [1] 0'a göre her sınıfın ağırlığını başlatın;
- [2] $i = 1$ den başlamak üzere her sınıflandırıcı için aşağıdaki işlemleri k defa tekrarlayın:
 - [3] Sınıflandırıcının oylama ağırlığını hesaplayın.
$$w_i = \log \frac{1 - err(M_i)}{err(M_i)}$$
 - [4] M_i den X için sınıf kestirimini alın.
$$c = M_i(X)$$
 - [5] c sınıfı ağırlığına w_i 'yi ekleyin.
 - [6] En büyük ağırlıklı sınıfı elde edin.
 - [7] Algoritmayı sonlandırın.

2.3.3. Kuralların Uygulanması ve Tahmin

Kuralların doğruluğu test edildikten sonra bu kurallar kestirim amacıyla kullanıma açılır. Örneğin, Şekil 2.10 üzerinde görüldüğü gibi MESLEK = Ücretli, YAŞ = Genç, CİNSİYET = Erkek nitelik değerlerine sahip yeni bir gözlem ile karşılaşıldığında, elde edilen karar kuralları kullanılarak sınıflandırma işlemi, yani kestirim işlemi gerçekleştirilir.

Bu varsayımlar karşısında KASKO niteliğinin değeri ne olacaktır? Bunun için 2.3.1' de elde edilen kurallar göz önüne alınarak, “Kural 2” nin amaca uygun olduğu anlaşılır. Söz konusu kuralın sonucu “KASKO = Hayır” dır ve sorunun doğru yanıtı da budur.

Veri kümesinde CİNSİYET isimli bir nitelik yer almasına karşılık, karar kuralları içinde bu nitelik yer almamıştır. Bu duruma, söz konusu niteliğin içerdiği değerlerin sınıflandırma üzerinde etkili olmaması neden olmaktadır.



Şekil 2.10. Karar Kurallarının Yeni Veriye Uygulanması.

2.4. KARAR AĞAÇLARINDA ENTROPİYE DAYALI BÖLÜNME

Sınıflandırma ağaçlarının oluşturulması esnasında hangi nitelikten başlayarak bölünmenin gerçekleştirileceği en önemli noktadır. Bölünmenin başlayacağı niteliğin seçilmesi amacıyla birçok yöntem geliştirilmiştir. Bunlar arasında yer alan kazanç ölçütü ve kazanç oranı, entropi tabanlı bölünme konusunda sıkça kullanılan ölçütlerdir.

2.4.1. Entropi

Felsefe, teoloji ve bilimde de yeralan entropi sözcüğü “enerji” gibi korunabilme özelliği olmayan bir sistemdeki rasgelelik, belirsizlik ve düzensizliğin ölçüsü olarak tanımlanabilir. Kısaca entropi bir sistemdeki belirsizliğin ölçüsüdür. S bir kaynak olsun. Bu kaynağın $\{m_1, m_2, \dots, m_n\}$ olmak üzere n mesaj üretilebildiğini varsayalım. Tüm mesajlar birbirinden bağımsız olarak üretilmektedir ve m_i mesajlarının üretilme olasılıkları p_i 'dir. $P = \{p_1, p_2, \dots, p_n\}$ olasılık dağılımına sahip mesajları üreten S kaynağının entropisi;

$$H(S) = - \sum_{i=1}^n p_i \log_2(p_i) \quad (2.10)$$

olarak ifade edilir.

İlk olarak istatistiksel fizikte kullanılan bu kavram, bilişim teorisine, bu teorinin yaratıcısı kabul edilen Claude Elwood Shannon tarafından uyarlanmıştır. Öz bilgi (self information) olarak adlandırdığı bir nicelikten bahseden Shannon'a göre bir A mesajının taşıdığı bilgi, aşağıdaki bağıntıda ifade edildiği gibi, onun gerçekleşme olasılığına bağlıdır ve bu olasılığın x tabanına göre eksi işaretli logaritması ile ifade edilir (Mackay, 2003, s.32).

$$I(A) = \log_x \frac{1}{P(A)} = -\log_x P(A) \quad (2.11)$$

Bilginin birimi logaritmanın tabanına bağlıdır. Eğer taban 2 ise birim bit'tir, e ise birim nat'tır, 10 ise birim hartley'dir. Bilgisayar dünyasında 0 ve 1'lerle yani bit'lerle çalıştığımız için, logaritmanın tabanını 2 kabul edeceğiz: Bu bağıntıya göre, iletilen mesajın olasılığı 1/8 ise, mesajın ilettiği bilgi $-\log_2 \left(\frac{1}{8} \right) = 3$ bit olarak ifade edilir.

Bu formülde verilen öz bilgi, o bilgiyi ifade edebilmek için kaç bit kullanılması gerektiğini gösterir. Bu eşitlik bize, yüksek olasılığa sahip mesajların düşük bilgi içerdiğini, düşük olasılığa sahip mesajların ise yüksek bilgi içerdiğini göstermektedir.

Entropi tanımının ortaya koyduğu gibi,

- Örnekler aynı sınıfa aitse entropi = 0
- Örnekler sınıflar arasında eşit dağılmışsa entropi = 1
- Örnekler sınıflar arasında rastgele dağılmışsa $0 < \text{entropi} < 1$

olur.

2.4.2. Kazanç Ölçütü veya ID3 Algoritması

Bölünmenin nereden başlayacağı konusunda kullanılan yöntemler arasında ID3 ve onun daha gelişkin biçimi olan C4.5 sayılabilir (Duda, 2000, s.412).

ID3 algoritması, öğrenme süreci sonunda bir karar ağacı üretir. Karar ağacı şeklinde kural üreten algoritmaların en iyilerinden biri olarak kabul edilmektedir. Örnek kümesinde geçen karakteristiklerin bilgi kazancını hesaplayarak verilen örnekler kümesini alt kümelere ayırır ve karar ağacını bu şekilde oluşturur (Bilgi kazancının hesaplanması sırasında bağıl frekans dikkate alınır). Bu algoritma verilen bir örneği sınıflandırmak için gerekli olan kural veya kuralları minimum sayıda test yaparak oluşturmayı amaçlamaktadır. ID3 algoritması, bilgi kazancı en fazla olan karakteristiği seçerek bu karakteristiği ilk aşamada ağacın kökü, diğer aşamalarda ise bir düğüm (alt set için kök) olarak alır ve buna göre karar ağacını dallandırır.

Bu algoritma ile ağaç türetme süreci şu şekilde işler: T eğitim kümesinin alt kümeleri $T_1, T_2, \dots, T_n$ olsun. C_j sınıflarına ait S altkümesindeki sınıflara ait durumların sayısını $freq(C_j, S)$ ile gösterelim. S'deki elemanların sayısını $|S|$ biçiminde ifade edelim. S kümesinden rasgele seçilen bir birimin C_j sınıfına ait olduğu varsayımı altında bu mesajın olasılığı,

$$\frac{freq(C_j, S)}{|S|} \quad (2.12)$$

biçimindedir. Böylece iletilen bilgi (2.10) bağıntısı göz önüne alınarak,

$$-log_2 \left(\frac{freq(C_j, S)}{|S|} \right) \quad (2.13.)$$

şeklinde elde edilir. Bu mesajdan beklenen bilgiyi bulmak için aşağıdaki bağıntıya başvurulabilir.

$$info(S) = - \sum_{j=1}^k \left(\frac{freq(C_j, S)}{|S|} \right) \log_2 \left(\frac{freq(C_j, S)}{|S|} \right) \quad (2.14)$$

(2.14) nolu bağıntısı T eğitim kümesine uygulandığında, info(T) değeri T deki sınıfları belirlemek için gerek duyulan bilginin ortalama miktarını belirler (Witten, 2005, s.102). Bu değer S kümesindeki entropi olarak da bilinmektedir (Baldi 2001, s.357). X niteliğinin değerine göre T eğitim kümesi alt kümelere bölündükten sonra her bir bölünme için benzer bir ölçü kullanılabilir. Bu durumda beklenen bilgi aşağıdaki gibi hesaplanır:

$$info_X(T) = \sum_{i=1}^n \frac{|T_i|}{|T|} info(T_i) \quad (2.15)$$

Bu durumda, X niteliği için kazanç miktarı aşağıdaki bağıntı yardımıyla hesaplanır.

$$gain(X) = info(T) - info_X(T) \quad (2.16)$$

Kazanç kriteri, bu bilgi kazancını maximize edecek niteliği seçme durumudur. Kriterin nasıl kullanıldığına örnek olması bakımından Tablo 2.1'deki eğitim kümesini dikkate alalım. Hedef niteliği olan “Kasko”, “Evet” ve “Hayır” değerlerine sahip durumdadır. Entropi tanımından yararlanarak,

$$-\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} = 0.954 \quad (2.17)$$

elde edilir. Bölünme işlemine hangi nitelikten başlanacağını belirlemek için her bir nitelik için entropi hesaplamaları yapılır. Önce “Cinsiyet” niteliği ile başlandığı varsayılırsa, bu niteliğin “Erkek” değeri için,

$$-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}=0.971 \quad (2.18)$$

“Kadın” değeri için,

$$-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}=0.918 \quad (2.19)$$

entropisi elde edilir.

Kadın ve Erkek kategorileri için hesaplanan entropi değerlerine bağlı olarak cinsiyet niteliğinin seçilmesiyle elde edilecek bilgi kazancı aşağıdaki gibi hesaplanır.

$$0.954 - \left[\frac{5}{8}(0.971) + \frac{3}{8}(0.918) \right] = 0.954 - 0.951 = 0.003 \quad (2.20)$$

Diğer nitelikler için de benzer hesaplamalar yapıldıktan sonra en büyük bilgi kazancını sağlayacak nitelik, birinci düğüm olarak belirlenmiş olur.

2.4.3. Kazanç Oranı veya C4.5 Algoritması

C4.5 öğrenme algoritması Quinlan tarafından geliştirilmiştir. Bu algoritma da tıpkı ID3 algoritması gibi yukarıdan aşağıya doğru karar ağacı üretir. Günümüzde en çok bilinen ve en çok kullanılan algoritma olarak kabul edilmektedir. C4.5 Algoritması ID3 algoritmasının bütün özelliklerini taşıyan bir algoritmadır. Ayrıca ID3 için yapılan açıklamalarda yer alan kavramlara yeni kavramlar eklenmiştir. Bölünme-Dağılma Bilgisi (Split-Info), özelliklerin kayıp değerleriyle baş edilmesi, sayısal özellik değerlerinin hesaplara katılması bu kavramlardan en önemlileridir.

ID3 algoritması karar ağacının her düğümü için bilgi kazancı değerlerini hesaplar. C4.5 ise bilgi kazancı ile beraber alt sette yer alan karakteristiklerin bilgi kazanç oranlarını da hesaplayarak bilgi kazanç oranı en yüksek olan karakteristiği düğüm noktası olarak seçer. Karar ağacının her dalı sadece bir tek sınıfa karşılık gelinceye kadar işlemleri sürdürür. C4.5 (Kazanç Oranı temelli) algoritması karar ağacı

oluşturma prosedürü, düğüm noktalarının belirlenmesi dışında ID3 (Bilgi Kazancı temelli) ile aynıdır. Daha sonra karar ağacını kural setine dönüştürür.

Bir kategorik özelliğin olası değer çeşitliliği ne kadar yüksek olursa o özelliğin bilgi kazancı gereksiz bir şekilde yüksek çıkar ve bu durum ağacın doğruluğunu kötü bir şekilde etkiler. Bu tip özellikler işe yaramadıkları gibi bilgi kazancı yüksek özelliklerin de önüne geçip veride gizlenmiş kuralların çıkarılmasına engel teşkil ederler.

Karar ağaçlarının oluşturulması esnasında bölünmenin nereden başlayacağı konusunda (2.16)'da ifade edilen kazanç miktarı kavramı kullanılabilir. Daha duyarlı sonuçlar elde etmek için bu ifade yerine, “Kazanç oranı” adı verilen bir ifade tercih edilebilir. Bunun için “bölünme bilgisi” (split information) adı verilen kavramdan yararlanılır. Bölünme bilgisi ile ilgili açıklamalar aşağıda verilmiştir.

A bir özellik, A_i bu özelliğin değerleri; T_i , A_i özelliğinin bu veride kaç kez tekrarlandığı, T ise ele alınan olay sayısını temsil etsin. Bu durumda bölünme bilgisi (2.21) ile ifade edilir. (Mitchell, 1997, s.73):

$$split\ info(X) = - \sum_{i=1}^n \frac{|T_i|}{|T|} \log_2 \left(\frac{|T_i|}{|T|} \right) \quad (2.21)$$

Bu bölünme bilgisinin tüm özelliklerin bilgi kazanç formülüne bölen olarak eklenmesiyle elde edilen oran kazanç oranı olarak ifade edilir. Bu durumda A özelliğinin kazanç oranı;

$$Kazanç\ oranı = \text{bilgi kazancı}(A) / \text{bölünme bilgisi}(A)$$

şeklinde hesap edilir (Mitchell, 1997, s.74). Kazanç oranı (2.21) ile bağlantılı olarak aşağıdaki eşitlikle açıklanır.

$$gain\ ratio(X) = \frac{gain(X)}{split\ info(X)} \quad (2.22)$$

2.5. SINIFLANDIRMA VE REGRESYON AĞAÇLARI

Karar ağaçlarının elde edilmesi amacıyla çok sayıda sınıflandırma algoritması geliştirilmiştir. Her bir sınıflandırma modelinin performansı farklıdır (Gey, 2005, s.658). ID3 ve C4.5 gibi ağaç yapılı sınıflandırma modelleri dışında “Sınıflandırma ve Regresyon Ağaçları” (Classification And Regression Trees = CART) adı verilen yöntemler yer almaktadır. CART ağaçları “ikili ağaç” (binary tree) yapısındadır. Yani CART ağaçlarında herhangi bir düğümden sadece iki dal çıkmaktadır. Çalışmanın bu bölümünde CART kapsamında Twoing, Gini ve Regresyon ağaçları algoritmaları açıklanacaktır.

2.5.1. Twoing Algoritması

Bu algoritma eğitim kümesinin her adımında iki parçaya ayrılması esasına dayanır (Duda, 2000, s.401). T eğitim kümesinin tümünü göstermek üzere, küme iki parçaya bölündüğünde, ayrılan parçalardan t_L sol taraftakini, t_R ise sağ taraftakini ifade eder.

Eğitim kümesindeki kayıtların sayısı N olmak üzere sol taraf için P_L olasılıkları şu şekilde hesaplanır (Larose, 2005, s.110):

$$P_L = \frac{s_{L1}}{N} \quad (2.23)$$

Burada s_{L1} t_L kümesindeki her bir nitelik değerinin ilgili nitelik değerleri içindeki tekrar sayısını, N ise eğitim kümesindeki kayıtların sayısını göstermektedir. Aşağıda verilen $P(j|t_L)$ koşullu olasılık ifadesi ise bir j sınıf değerinin sol taraftaki bölünmede yer alma olasılığını ifade etmektedir.

$$P(j|t_L) = \frac{s_{L2}}{s_{L3}} \quad (2.24)$$

Bu bağıntıda s_{L2} t_L kümesindeki kayıtların j sınıfları sayısını, s_{L3} ise, t_L kümesindeki her bir nitelik değerinin ilgili nitelik değerleri arasındaki tekrar sayısını ifade etmektedir. Sağ taraftaki t_R kümesi için de benzer tanımlamalar yapılabilir.

$$P_R = \frac{s_{R1}}{N} \quad (2.25)$$

Bağıntıdaki s_{R1} , t_R kümesindeki her bir nitelik değerinin ilgili nitelik değerleri içindeki tekrar sayısını göstermektedir. $P(j|t_R)$ ifadesi ise bir j sınıf değerinin sağ taraftaki bölünmede bulunma olasılığını göstermektedir.

$$P(j|t_R) = \frac{s_{R2}}{s_{R3}} \quad (2.26)$$

Burada s_{R2} t_R kümesindeki kayıtların j sınıfları sayısını gösterirken s_{R3} t_R kümesindeki her bir nitelik değerinin ilgili nitelik değerleri arasındaki tekrar sayısını ifade etmektedir.

Hangi niteliğin seçileceği konusunda karar verebilmek için “uygunluk ölçütünün” değerine bakılır. $\emptyset(s|t)$, t düğümündeki bölünmelerin uygunluk ölçütü olsun. Söz konusu uygunluk ölçütü,

$$\emptyset(s|t) = 2P_L P_R \sum_{j=1}^n |P(j|t_L) - P(j|t_R)| \quad (2.27)$$

şeklinde hesaplanır.

$\emptyset(s|t)$ değerleri hesaplandıktan sonra bunlardan en büyük olanı ($\max \emptyset(s|t)$) seçilir.

En büyük uygunluk değerine sahip olan satır, bölünmenin başlatılacağı niteliği belirler.

2.5.2. Gini Algoritması

CART kapsamında ele alınan sınıflandırma yöntemlerinden birisi de Gini Algoritması'dır. Bir T kümesi içindeki i sınıflarının nispi frekans (Relative Frequency) sayısı veya bir başka deyişle i sınıf değerlerinin T kümesinin eleman sayısına oranı p_i olmak üzere, Gini ölçütü,

$$gini(T) = 1 - \sum_{i=1}^k p_i^2 \quad (2.28)$$

olarak ifade edilir (Abbas 2002, s.80).

Gini ölçütünden yararlanarak ortaya atılan Gini sınıflandırma algoritması, aynen Twoing algoritmasında olduğu gibi nitelik değerlerinin sol ve sağda olmak üzere iki bölüme ayrılması esasına dayanmaktadır. Öncelikle her nitelik değerleri ikili olacak biçimde gruplanır ve bu yolla elde edilen sol ve sağ bölünmelere karşılık gelen sınıf değerleri gruplandırılır. Veri kümesindeki her bir nitelik ile ilgili sol ve sağ taraftaki bölünmeler için $gini_L$ ve $gini_R$ değerleri aşağıdaki gibi hesaplanır:

$$gini_L = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_L|} \right)^2 \quad (2.29)$$

$$gini_R = 1 - \sum_{i=1}^k \left(\frac{R_i}{|T_R|} \right)^2 \quad (2.30)$$

Burada; k sınıfların sayısını, T bir düğümdeki örnekleri, $|T_L|$ sol taraftaki örneklerin sayısını, L_i sol tarafta i kategorisindeki örneklerin sayısını, $|T_R|$ sağ taraftaki örneklerin sayısını ve R_i sağ tarafta i kategorisindeki örneklerin sayısını göstermektedir. Gini algoritmasında $gini_L$ ve $gini_R$ değerleri hesaplandıktan sonra bölünmenin başlayacağı niteliği seçmek üzere Gini ölçütü kullanılır. Eğitim kümesindeki satır sayısı n olmak üzere her j niteliği için Gini ölçütü aşağıdaki şekilde hesaplanır:

$$gini_j = \frac{1}{n} (|T_L|gini_L + |T_R|gini_R) \quad (2.31)$$

Hesaplanan tüm ölçütler arasında ($\min gini_j$) seçilir ve bu değere sahip olan nitelik bölünmenin başlayacağı düğüm olarak belirlenir.

2.5.3. Regresyon Ağaçları

İkili bölünme yöntemini kullanan bir diğer CART yöntemi “regresyon ağaçları” (Regression Trees) adıyla bilinmektedir. Çalışmanın uygulamasında kullanılan bu yöntem, Breiman ve arkadaşları tarafından yapılan çalışmalar sonucunda ortaya çıkmıştır. Regresyon ağaçlarının esasları (2.7) ‘ de açıklanmıştır.

2.6. ÇOKLU DOĞRUSAL REGRESYON MODELİ

2.6.1. Basit Regresyon Modeli

X ve Y değişkenleri arasındaki ilişkinin şekli, basit regresyon modeli biçiminde ortaya konabilir. Burada X bağımsız değişkenleri, Y ise bağımlı değişkeni ifade etmektedir. Basit regresyon, göz önüne alınan bir olayda değişkenler arasındaki ilişkiyi ortaya koyar. Basit regresyon modeli şu şekilde ifade edilir (Chatterjee 2006, s.28):

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (2.32)$$

Regresyon modelinde aşağıda belirtilen varsayımların geçerli olduğu kabul edilir (Demaris 2004, s.45):

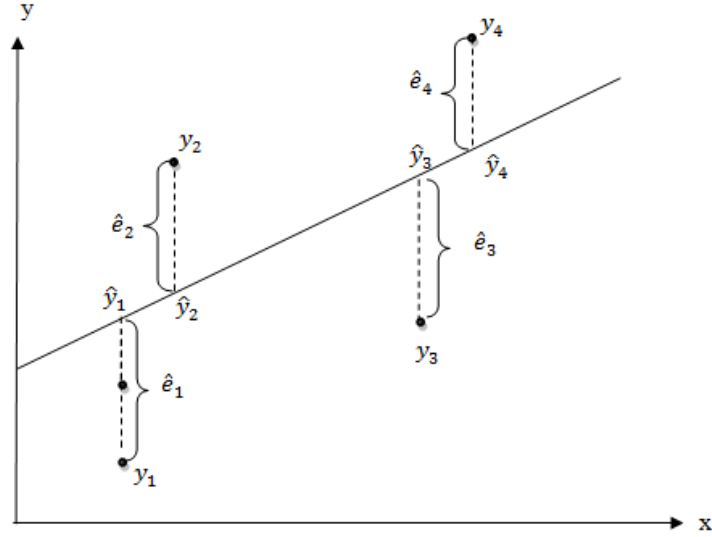
$$E(e_i) = 0$$

$$var(e_i) = \sigma^2 \quad (2.33)$$

$$cov(e_i, e_j) = 0 \quad i \neq j \quad (2.34)$$

Modelde yer alan e_i ifadesi bilinmeyen tesadüfi bir değer, β_0 , β_1 ve σ^2 ise bilinmeyen parametrelerdir. Parametrelerin tahmini EKK ile yapılır. Bunun için modelde yer alan kalıntılardan (residuals), yani e_i değerlerinden yararlanılır. Şekil 2.11 de en küçük kareler doğrusu, gerçek değerler, tahmini değerler ve kalıntıları gösterilmektedir. β_0, β_1 ve σ^2 parametrelerinin tahminleri sırasıyla $\hat{\beta}_0, \hat{\beta}_1$ ve $\hat{\sigma}^2$ biçiminde gösterilecek olursa, kalıntıların tahmini şu şekilde ifade edilebilir (Rawling ve diğer. 1998, s.3):

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad i = 1, 2, \dots, n \quad (2.35)$$



Şekil 2.11. Tahmin Edilen Doğru İçin Kalıntılar.

β_0 ve β_1 parametrelerinin tahmin edicileri sırasıyla $\hat{\beta}_0$ ve $\hat{\beta}_1$ 'dir. Kalıntıların kareli ortalaması (RSS) ise aşağıdaki şekilde ifade edilir (Chatterjee 2006, s.29):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (2.36)$$

Kalıntıları minimize edecek $\hat{\beta}_0$ ve $\hat{\beta}_1$ değerleri, β_0 ve β_1 parametrelerinin tahmin edicileridir. Minimizasyon problemini çözmek üzere aşağıdaki eşitliklerin çözüme kavuşturulması gerekir (Weisberg 2005, s.279):

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (2.37)$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (2.38)$$

Bu eşitlik düzenlenerek şu şekilde bir eşitliğe ulaşılır:

$$\beta_0 n - \beta_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \quad (2.39)$$

$$\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \quad (2.40)$$

Bu eşitliklere regresyon analizlerinde “normal eşitlikler” adı verilmektedir ve bu eşitlikler çözülerek, $\hat{\beta}_0$ ve $\hat{\beta}_1$ parametre tahminleri elde edilir:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (2.41)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n (\bar{x})^2} \quad (2.42)$$

2.6.2. Çoklu Doğrusal Regresyon

Tek bir bağımlı değişkene karşılık, birden fazla bağımsız değişken söz konusu ise, model çoklu regresyon modeli olarak ifade edilir.

Bağımsız değişken y olmak üzere, $x_2, x_3, \dots, x_k$ gibi k sayıda bağımlı değişken ve u hata terimi arasında doğrusal bir ilişki bulunduğunu varsayalım. Bu durumda değişkenler arasındaki ilişki,

$$y = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + u_i \quad (i = 1, 2, \dots, n) \quad (2.43)$$

biçiminde ifade edilir. Değişkenleri aşağıda gösterildiği biçimde matrislerle ifade etmek mümkündür (Weisberg, 2005, s.54).

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{21} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{2n} & \dots & x_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad (2.44)$$

Bu matris gösterimlerine dayanarak (1) ifadesi aşağıdaki şekilde yazılır:

$$Y = X\beta + u \quad (2.45)$$

2.6.3. Temel Varsayımlar

Çoklu Doğrusal Regresyon Modelinin elde edilmesinde aşağıda sıralanan varsayımların kabul edilmesi gerekmektedir (Greene, 2003, s.16):

- $E(u) = 0$
- $E(uu') = \sigma^2 I_n$
- X sabit sayılar kümesidir.
- X matrisinin rankı $k < n$ 'dir.

2.6.4. En Küçük Kareler Tahminleri

β 'nin tahminlerini gösteren satır vektör $\hat{\beta}$ 'dir ve

$$\hat{\beta} = \{\beta_0, \hat{\beta}_1, \dots, \hat{\beta}_k\} \quad (2.46)$$

şeklinde ifade edilir.

$Y - X\hat{\beta}$ satır vektörü ise e ile ifade edildiğinde, kalıntıların karesi,

$$\sum_{i=1}^n e_i^2 = e'e \quad (2.47)$$

biçiminde matrislerle ifade edilebilir. Bu durumda $e'e$ şu şekilde hesaplanır.

$$e'e = (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) \quad (2.48)$$

$$= Y'Y - 2\tilde{\beta}'X'Y + \tilde{\beta}'X'X \quad (2.49)$$

Kalıntıların kareleri toplamını minimize eden $\tilde{\beta}$ değeri, en küçük kareler yöntemi uygulanarak elde edilir (Greene, 2003, s.20).

$$\frac{\partial(e'e)}{\partial\tilde{\beta}} = -2X'Y - 2X'X\tilde{\beta} \quad (2.50)$$

(2.50) ifadesi sıfıra eşitlenerek,

$$X'X\tilde{\beta} = X'Y \quad (2.51)$$

elde edilir ve buradan,

$$\tilde{\beta} = (X'X)^{-1}X'Y \quad (2.52)$$

sonucu bulunur.

2.6.5. Ortalama ve Varyans

$Y = X\beta + u$ ifadesi $\tilde{\beta} = (X'X)^{-1}X'Y$ 'de yerine yazılırsa,

$$\hat{\beta} = (X' X)^{-1} X' (X \beta + u) \quad (2.53)$$

bulunur. Bu ifadenin her iki yanının beklenen değeri,

$$E(\hat{\beta}) = E(\beta) + E[(X' X)^{-1} X' u] \quad (2.54)$$

alınarak düzenlendiğinde, $E(\beta) = \beta$ olduğuna göre,

$$E(\hat{\beta}) = \beta + (X' X)^{-1} X' E(u) \quad (2.55)$$

ve ayrıca $E(u) = 0$ varsayımından hareketle,

$$E(\hat{\beta}) = \beta \quad (2.56)$$

elde edilir. Bu durum en küçük kareler tahmincisinin yansız olduğunu gösterir. (2.53)'deki bağıntıdan,

$$(\hat{\beta} - \beta) = (X' X)^{-1} X' u \quad (2.57)$$

yazılabilir. Buradan $\hat{\beta}$ 'nin varyansı,

$$\begin{aligned} \text{Var}(\hat{\beta}) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E[(X' X)^{-1} X' u u' X (X' X)^{-1}] \end{aligned} \quad (2.58)$$

elde edilir. (2.58)'den,

$$\text{Var}(\hat{\beta}) = (X' X)^{-1} E(u u') X (X' X)^{-1} \quad (2.59)$$

yazılır. $E(u u') = \sigma^2 I_n$ varsayımından hareket edilerek,

$$\text{Var}(\hat{\beta}) = \sigma^2 (X' X)^{-1} \quad (2.60)$$

elde edilir.

2.6.6. Standart Hata

$Y = X \beta + u$ ifadesi (2.52)' de yerine yazılırsa,

$$e = [I_n - X(X' X)^{-1} X'] u \quad (2.61)$$

elde edilir. Burada $D = I_n - X(X' X)^{-1} X'$ tanımlanacak olursa, D simetrik idempotent matrisi için, $D' = D$ ve $D^2 = D$ özelliklerinden dolayı,

$$e'e = u' D' D u$$

$$= u' [I_n - X(X' X)^{-1} X'] u$$

yazılır. Her iki tarafın beklenen değerleri alınarak ve $(X' X)^{-1} (X' X) = I_k$ özelliği kullanılarak,

$$E(e'e) = (n - k) \sigma^2 \quad (2.62)$$

elde edilir. Buradan hata varyansının yansız bir tahmincisi olarak,

$$s^2 = \frac{e'e}{n - k} \quad (2.63)$$

bulunur. $e'e$ yerine (2.49)'daki değeri kullanılarak,

$$e'e = Y'Y - \hat{\beta}'X'Y \quad (2.64)$$

elde edilir. O halde sonuç olarak,

$$S^2 = \frac{Y'Y - \hat{\beta}'X'Y}{n - k} \quad (2.65)$$

bulunur. (2.65) den $S = \sqrt{S^2}$ tahminin standart hatası olarak elde edilir.

2.6.7. Korelasyon katsayısı:

Değişkenler arasındaki ilişkinin gücünü ve yönünü belirlemek üzere korelasyon katsayısı kullanılır. Korelasyon katsayısını elde etmek için “açıklanabilen kareler toplamı” ve “kalıntı kareler toplamı” kullanılır. Açıklanabilen kareler toplamı (AKT) şu şekilde ifade edilir:

$$AKT = \sum y_i^2 - \sum e_i^2 \quad (2.66)$$

Matrisler kullanılarak, (2.64) bağıntısı yardımıyla,

$$AKT = \hat{\beta}'X'Y - \frac{(\sum y)^2}{n} \quad (2.67)$$

bulunur. AKT ifadesinden hareketle R^2 çoklu korelasyon katsayısı şu şekilde elde edilir:

$$R^2 = \frac{\hat{\beta}' X' Y - \frac{(\sum y)^2}{n}}{Y' Y - \frac{(\sum y)^2}{n}} \quad (2.68)$$

Korelasyon katsayısı -1 ile +1 arasında değişir. Bu sınır değerlere yaklaştıkça değişkenler arasındaki ilişki o derece güçlüdür. Eğer korelasyon katsayısı sıfıra yakın ise iki değişken arasında doğrusal bir ilişkinin olmadığına hükmedilir.

2.7. REGRESYON AĞAÇLARI İLE SINIFLANDIRMA

2.7.1. Regresyon Ağaçları ve En Küçük Kareler Regresyonu

Regresyon ağaçları ile ilgili çalışmalarda genel olarak Y “yanıt değişkeni”, X ise “ölçüm vektörü” olarak isimlendirilmektedir. Bir kestirim kuralı veya x reel değerlerini alan bir kestirimci $d(x)$ fonksiyonu biçiminde ifade edilebilir. $d(x)$, x 'e bağlı gerçek değerli bir fonksiyondur. Regresyon analizi, $\mathcal{L}$ öğrenme örneğinden başlayarak, $d(x)$ tahmin edicisinin elde edilmesi olayına karşılık gelen genel bir terimdir. Bir kestirimci yaratmanın iki amacı olabilir:

- Yanıt değişkenlerinin daha sonra ortaya çıkacak ölçüm vektörlerine uygun biçimde kestirilmesi,
- Ölçüm değişkenleri ile yanıt değişkenleri arasındaki yapısal ilişkinin ortaya konulması.

n adet $(x_1, y_1), \dots, (x_n, y_n)$ gözlemlerinden oluşan bir $\mathcal{L}$ öğrenme örneğinin bir $d(x)$ kestirimcisini bulmak için kullanıldığını varsayalım. Bu durumda kestirimin doğruluğunun nasıl belirleneceği sorusu ortaya çıkmaktadır. Eğer N_2 sayıda çok büyük bir $(x'_1, y'_1), \dots, (x'_{N_2}, y'_1)$ test örneği mevcut ise, $d(x)$ 'in doğruluğu aşağıdaki bağıntıdan hareketle hesaplanan “ortalama hata” ile ölçülebilir (Breiman, 1998, s.222).

$$\frac{1}{N_2} \sum_{n=1}^{N_2} |y'_n - d(x'_n)| \quad (2.69)$$

Burada $d(x'_n)$ ifadesi $n = 1, \dots, N_2$ olmak üzere y'_n nin bir tahmincisidir. Bu ölçü “en küçük mutlak sapma” regresyonuna dayanır. Ancak hesaplamaları daha kolay bir hale getirmek için, doğruluk ölçüsü olarak klasik regresyonda olduğu gibi “hata kareleri ortalaması” kavramına başvurulur. Hata kareleri ortalaması aşağıdaki şekilde ifade edilebilir:

$$\frac{1}{N_2} \sum_{n=1}^{N_2} (y'_n - d(x'_n))^2 \quad (2.70)$$

2.71, d tahmin edicisinin $R^*(d)$ hata kareleri ortalamasını tanımlar (Breiman, 1998, s.222):

$$R^*(d) = E(Y - d(X))^2 \quad (2.71)$$

Yani $R^*(d)$ ifadesi, Y nin kestirimcisi olarak $d(X)$ i kullanan “beklenen hata kareleri” dir. $R^*(d)$ ifadesi d sınıflandırıcısının hatalı sınıflandırma oranını belirtmek için kullanılır. Burada kestirimcinin doğruluğu için de aynı notasyon tercih edilir. $R^*(d)$ ‘yi minimize eden d_E kestirimcisi,

$$d_E(x) = E(y|X = x) \quad (2.72)$$

biçiminde ifade edilir. Başka bir deyişle, $d_E(x)$, verilen bu X ölçüm vektörü için yanıtın koşullu beklentisidir. Aşağıdaki teorem bu hususu aydınlatmaktadır.

Teorem 2.1 : Y sürekli tesadüfî değişken olmak üzere, hata karelerinin beklenen değerini, yani $E(Y - a)^2$ ifadesini minimize eden bir a sabiti, Y'nin beklenen değerine eşittir (Breiman, 1998, s.223).

Teoremin ispatı şu şekilde yapılabilir:

$$\emptyset(k) = E(Y - k)^2 = \int_{-\infty}^{+\infty} (y - k)^2 f(y) dy \quad (2.73)$$

olsun.

$$E(Y) = \int_{-\infty}^{+\infty} y f(y) dy \quad (2.74)$$

olduğu bilinmektedir. (2.73) dan şu ifade elde edilir:

$$\emptyset(k) = \int_{-\infty}^{+\infty} (y^2 - 2yk + k^2) f(y) dy \quad (2.75)$$

$\int_{-\infty}^{+\infty} f(y) dy = 1$ olduğuna göre,

$$\emptyset(k) = \int_{-\infty}^{+\infty} y^2 f(y) dy - 2k \int_{-\infty}^{+\infty} y f(y) dy + k^2 \quad (2.76)$$

elde edilir. $\emptyset(k)$ fonksiyonunu k 'ya göre minimize etmek için, k ya göre kısmi türevi alınarak sıfıra eşitlenir.

$$\frac{\partial \emptyset(k)}{\partial k} = 0 \quad (2.77)$$

$$0 - 2 \int_{-\infty}^{+\infty} yf(y) dy + 2k = 0 \quad (2.78)$$

O halde k için,

$$k = \int_{-\infty}^{+\infty} yf(y) dy \quad (2.79)$$

sonucu elde edilir. (2.79)'nin sağ tarafının (2.74) ile açıklanan $E(Y)$ 'ye eşit olduğu görülebilir.

2.7.2. Hata Ölçümleri ve Hata Tahminleri

$(x_1, y_1), \dots, (x_N, y_N)$ den oluşan bir $\mathcal{L}$ öğrenme örneği, hem $d(x)$ kestirimcisinin elde edilmesinde hem de $R^*(d)$ hatasının tahmininde kullanılır. $R^*(d)$ 'yi tahmin etmek için birkaç yol vardır. Yerine koyma (resubstitution) hatası şu şekildedir (Breiman, 1998, s.223):

$$R(d) = \frac{1}{N} \sum_n (y_n - d(x_n))^2 \quad (2.80)$$

$R^{ts}(d)$ test örnek tahminleri, $\mathcal{L}$ kümesi $\mathcal{L}_1$ ve $\mathcal{L}_2$ biçiminde bölünerek aşağıda gösterildiği biçimde elde edilir:

$$R^{ts}(d) = \frac{1}{N_2} \sum_{(x_n, y_n) \in \mathcal{L}_2} (y_n - d(x_n))^2 \quad (2.81)$$

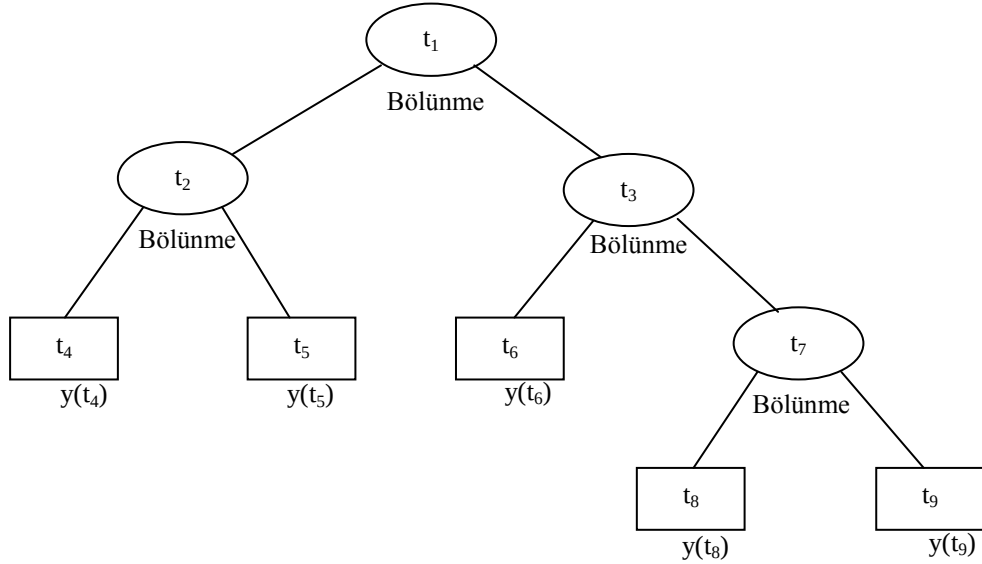
$R^{cv}(d)$ çapraz doğrulama tahmini ise, $\mathcal{L}$ yi aynı sayıda olay içeren v sayıda $\mathcal{L}_1, \dots, \mathcal{L}_v$ alt kümelere bölerek elde edilir. Her bir $v, v = 1, \dots, V$ için aynı ağaç oluşturma işlemi

$d^{(V)}(X)$ tahmini alınarak $\mathcal{L} - \mathcal{L}_V$ öğrenme örneğine uygulanır. Buradan çapraz doğrulama verisi için aşağıdaki ifade yazılabilir:

$$R^{CV}(d) = \frac{1}{N} \sum_V \sum_{(x_n, y_n) \in \mathcal{L}_V} (y_n - d^{(V)}(x_n))^2 \quad (2.82)$$

2.7.3. Ağaç Yapılı Regresyon

Ağaç yapılı tahmin, ağaç yapılı sınıflandırmaya dayanır. Şekil 2.12’de açıklandığı gibi, terminal düğümlerinde, $y(t)$ yanıtları sabit değerdedir (Breiman, 1998, s.228).



Şekil 2.12. Karar Ağacında İkili Bölünmeler.

Kestirimci $d(x)$, her terminal düğümü boyunca sabit olduğundan Şekil 2.12’deki ağaç, Şekil 2.13’de olduğu gibi regresyon yüzeyinin bir histogram tahmini olarak da gösterilebilir (Breiman, 1998, s.229).

Bir $\mathcal{L}$ öğrenme örneği ile başlayarak bir ağaç kestirimini belirlemek için gerekli üç unsur vardır. Bunlar;

- Her ara düğümde bir bölünmeyi seçmek için bir yol bulunması

- Bir düğümün terminal olduğu zamanı belirlemek için bir kural tanımlanması
- Her t terminal düğümüne, bir $y(t)$ değerini atamak için bir kural oluşturulmasıdır.
- Bu unsurlar incelendiğinde, sınıflandırmada olduğu gibi, düğüm belirleme kuralının kolayca ortaya konulacağı anlaşılır. Süreç $R^*(d)$ için yeniden yerine koyma tahmini ile başlar ve tahmin aşağıdaki şekildedir (Breiman, 1998, s.230):

$$R(d) = \frac{1}{N} \sum_n (y_n - d(x_n))^2 \quad (2.83)$$

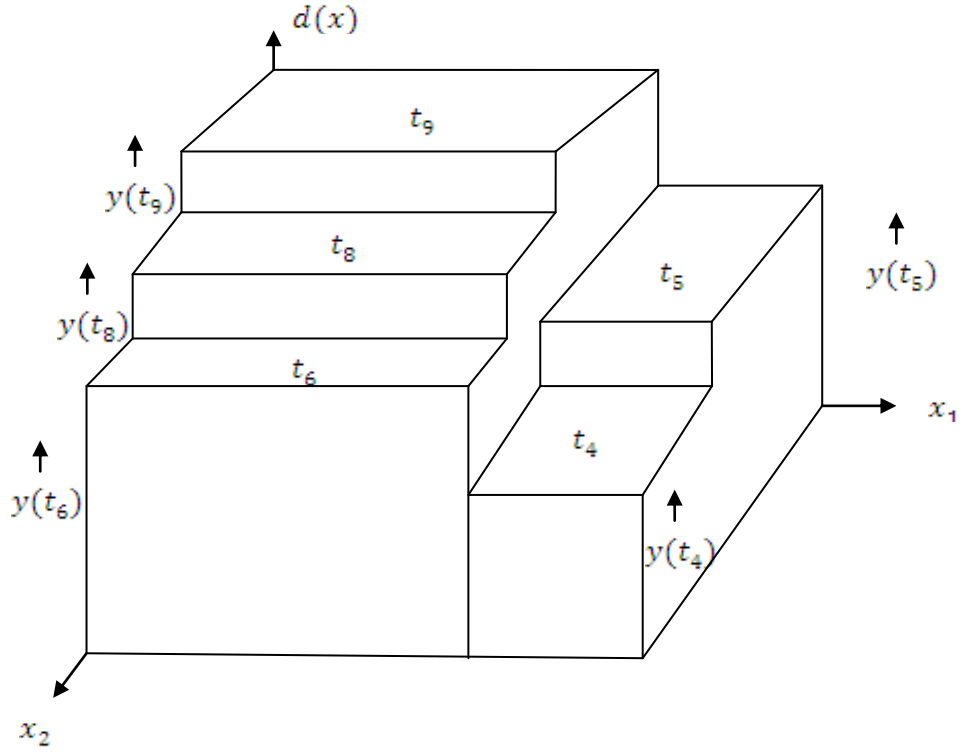
Önerme 2.1 : $R(d)$ 'yi minimize eden $\bar{y}(t)$ değeri, t'ye ait tüm (x_n, y_n) noktalarıyla ilgili y_n değerlerinin ortalamasıdır.

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{x_n \in t} y_n \quad (2.84)$$

Burada $N(t)$, t nin içindeki olayların toplam sayısıdır. Bu önermenin ispatı

$\sum_n (y_n - a)^2$ 'yi minimum yapan a'yı belirlemeye dayanır. Söz konusu a değerini elde etmek için $\sum_n (y_n - a)^2$ ifadesi a'ya göre türevi alınarak sıfıra eşitlenir.

$$-2 \sum_{n=1}^N y_n + 2aN = 0$$



Şekil 2.13. Regresyon Yüzeyinin Histogram Biçimindeki Görünümü.

Önerme 2.2 : $R(d)$ 'yi minimize eden $\bar{y}(t)$ değeri, t 'ye ait tüm (x_n, y_n) noktalarıyla ilgili y_n değerlerinin ortalamasıdır.

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{x_n \in t} y_n \quad (2.85)$$

Burada $N(t)$, t nin içindeki olayların toplam sayısıdır. Bu önermenin ispatı

$\sum_n (y_n - a)^2$ 'yi minimum yapan a 'yı belirlemeye dayanır. Söz konusu “ a ” değerini elde etmek için $\sum_n (y_n - a)^2$ ifadesi a 'ya göre türevi alınarak sıfıra eşitlenir.

$$-2 \sum_{n=1}^N y_n + 2aN = 0$$

Eşitlik çözülürse aşağıdaki a değeri elde edilir:

$$a = \frac{1}{N} \sum_{n=1}^N y_n \quad (2.86)$$

Buradan bir düğümdeki herhangi bir kestirim değeri yerine $a = \bar{y}(t)$ 'nin kullanılabileceği anlaşılmaktadır. Böylece (2.83)'de verilen $R(d)$ için aşağıdaki $R(T)$ notasyonu kullanılabilir (Breiman, 1998, s.230).

$$R(T) = \frac{1}{N} \sum_{t \in \mathcal{T}} \sum_{x_n \in t} (y_n - \bar{y}(t))^2 \quad (2.87)$$

Bu durumda t düğümü için aşağıdaki ifade yazılabilir:

$$R(t) = \frac{1}{N} \sum_{x_n \in t} (y_n - \bar{y}(t))^2 \quad (2.88)$$

Böylece (2.87)

$$R(T) = \sum_{t \in \mathcal{T}} R(t) \quad (2.89)$$

ile ifade edilir.

Tanım 2.1 : t nin en iyi s^* bölünmesi , $R(T)$ yi en fazla azaltan bölünmedir. Daha doğrusu, s bölünmesi t_L ve t_R gibi iki bölünmeye neden oluyorsa aşağıdaki ifade yazılabilir (Breiman, 1998, s.231):

$$\Delta R(s, t) = R(t) - R(t_L) - R(t_R) \quad (2.90)$$

Bu durumda,

$$\Delta R(s^*, t) = \max_{s \in S} \Delta R(s, t) \quad (2.91)$$

en uygun bölünmeyi tanımlar. Böylece regresyon $R(T)$ deki azalışı maximize etmek üzere iteratif olarak bölünme kuralları elde edilmiş olur. Bir adım daha ileri giderek,

$p(t) = \frac{N(t)}{N}$ olasılığını göz önüne alalım. Bir t düğümü için örnek varyansı,

$$s^2(t) = \frac{1}{N(t)} \sum_{x_n \in t} (y_n - \bar{y}(t))^2 \quad (2.92)$$

biçiminde ifade edilebilir. O halde $R(t) = s^2(t)p(t)$ ve buradan,

$$R(t) = \sum_{t \in T} s^2(t) p(t) \quad (2.93)$$

$s^2(t)$, t düğümündeki y_n değerlerinin bir örnek varyansıdır. En iyi bölünmenin aşağıdaki ağırlıklı varyansı minimize etmesi beklenir.

$$p_L s^2(t_L) + p_R s^2(t_R) \quad (2.94)$$

Sonuç olarak, regresyon ağaçları yönteminde en uygun bölünmenin nereden başlayacağı hususunda karar vermek üzere (2.93) bağıntısı kullanılır.

$Min(p_L s^2(t_L) + p_R s^2(t_R))$ koşuluna uyan bölünmeler seçilerek adım adım karar ağacı oluşturulur.

BÖLÜM III: İSTANBUL MENKUL KIYMETLER BORSASI İÇİN UYGULAMA

3.1. GİRİŞ

“İMKB 30”, İstanbul Menkul Kıymetler Borsası (İMKB)’nda işlem gören ve yatırım ortaklıkları dışında kalan şirketlerden en büyük 30 tanesinin karlılığa gören artan ya da azalan göstergesidir. İMKB’deki şirketlerin bu gruba dahil olabilmesi için piyasa değerlerinin, likiditelerinin ve sektörel temsil yeteneklerinin yüksek olması gerekmektedir. İMKB 30’a dahil şirketlerin listesi Tablo 3.1’de verildiği gibidir.

Tablo 3.1. İMKB 30 Grubuna Dahil Şirketler.

Menkul	Şirket Adı
AEFES	Anadolu Efes Biracılık A.Ş.
AKBNK	Akbank T.A.Ş.
AKGRT	Aksigorta A.Ş.
ASYAB	Asyabank
BAGFS	Bağfaş Bandırma Gubre Fabrikaları A.Ş.
DOHOL	Doğan Holding A.Ş.
DYHOL	Doğan Yayın Holding A.Ş.
ENKA	ENKA İnşaat ve Sanayi A.Ş.
EREGL	Ereğli Demir ve Çelik Fabrikaları T.A.Ş.
GARAN	T. Garanti Bankası A.Ş.
HALKB	T. Halk Bankası A.Ş.
ISBTR	T. İş Bankası A.Ş.
ISGYO	İş Gayrimenkul Yatırım Ortaklığı A.Ş.
KCHOL	Koç Holding A.Ş.
KRDMD	Kardemir Karabük Demir Çelik Sanayi Ve Ticaret A.Ş.
SAHOL	Sabancı Holding A.Ş.
SKBNK	Şekerbank T.A.Ş.
SISE	Türkiye Şişe ve Cam Fabrikaları A.Ş.
PETKM	Petkim Petrokimya Holding A.Ş.
TAVHL	Tav Havalimanları Holding A.Ş.
TEBNK	Türk Ekonomi Bankası A.Ş.
THY	Türk Hava Yolları A.O.
TKFEN	Tekfen Holding A.Ş.
TOASO	Tofaş Türk Otomobil Fabrikası A.Ş.
TCELL	TURKCELL İletişim Hizmetleri A.Ş.
TUPRPS	TÜPRAS-Türkiye Petrol Rafinerileri A.Ş.
TSKB	Türkiye Sınai Kalkınma Bankası A.Ş.
TTKOM	Türk Telekomünikasyon A.Ş.
VKBNK	Türkiye Vakıflar Bankası T.A.O.
YKBNK	Yapı ve Kredi Bankası A.Ş.

Çalışmanın bu bölümünde regresyon ağaçları yardımıyla, 1990-2010 yılları arasındaki İMKB 30'da yeralan şirketlerin günlük kapanış fiyatları, menkul kıymet teknik analizlerinde yaygın biçimde kullanılan teknik göstergeler, altın fiyatlarındaki değişimler, dolar kurundaki değişimler ve bazı yurtdışı borsa göstergeleri kullanılarak sınıflandırma modellerinin oluşturulması ve karar kurallarının neler olduğunun belirlenmesi amaçlanmaktadır. Bu amaçla öncelikle şirketlere ait göstergelerden hangilerinin gerçekte önemli oldukları ÇDR modelleri ile belirlenmiş ve daha sonra sınıflandırma ve kural çıkarma işlemleri gerçekleştirilmiştir.

İşlemler XLMiner 3.2.6 kullanılarak gerçekleştirilmiştir. XLMiner yazılımı Ms Excel elektronik tablosu üzerine eklenebilen bir yazılım olup, verinin hazırlanması ve regresyon ağaçlarının oluşturulması amacıyla birçok araca sahiptir.

3.2. VERİ KÜMESİ

Çalışmada kullanılacak İMKB 30 kapsamındaki şirketlerin günlük kapanış fiyatları zaman aralığı 1990-2010 olmak üzere İMKB Eğitim Müdürlüğü'nden temin edilmiştir. Ancak tatil günlerine rastlayan günlerde İMKB'de işlem görülmediği için veri üretilmemiştir. İMKB 30 menkul kıymetlerinin veri kümesi Tablo 3.2 de belirtilen tarih aralıklarına sahiptir.

Tablo 3.2. İMKB 30 Verileri Tarih Aralığı.

Menkul Kıymet	Başlangıç	Bitiş	Menkul Kıymet	Başlangıç	Bitiş
AEFES	24.07.2000	27.10.2010	SAHOL	08.07.1997	27.10.2010
AKBNK	01.08.1990	27.10.2010	SKBNK	10.04.1997	27.10.2010
AKGRT	02.01.1997	27.10.2010	SISE	01.08.1990	27.10.2010
ASYAB	12.05.2006	27.10.2010	PETKM	01.08.1990	27.10.2010
BAGFS	01.08.1990	27.10.2010	TAVHL	23.02.2007	27.10.2010
DOHOL	21.06.1993	27.10.2010	TEBNK	28.02.2000	27.10.2010
DYHOL	06.08.1998	27.10.2010	THY	15.01.1991	27.10.2010
ENKA	01.08.1990	27.10.2010	TKFEN	23.11.2007	27.10.2010
EREGL	01.08.1990	27.10.2010	TOASO	01.07.1991	27.10.2010
GARAN	01.08.1990	27.10.2010	TCELL	11.07.2000	27.10.2010
HALKB	09.05.2007	27.10.2010	TUPRPS	30.05.1991	14.10.2009
ISBTR	01.08.1990	27.10.2010	TSKB	01.08.1990	27.10.2010
ISGYO	09.12.1999	27.10.2010	TTKOM	15.05.2008	27.10.2010
KCHOL	01.08.1990	27.10.2010	VKBNK	18.11.2005	27.10.2010
KRDMD	08.06.1998	27.10.2010	YKBNK	01.08.1990	27.10.2010

Verilerde 1 Ocak 2005 tarihinden itibaren TL den YTL'ye geçiş söz konusu olduğu için bu tarihten önceki tüm verilere bu geçişe uygun biçimde dönüşüm uygulanmıştır.

3.3. NİTELİKLER

Çalışmada, menkul kıymet teknik analizlerinde kullanılan bazı göstergeler model nitelikler olarak ele alınmıştır. Söz konusu nitelikler aşağıda sıralandığı gibidir.

- a) İMKB ye dahil İMKB 30 menkul kıymetleri günlük kapanış fiyatları.
- b) Menkul kıymetin kapanış fiyatlarına dayalı olarak hesaplanan teknik analiz göstergeleri:
 - Göreceli Güç İndeksi
 - Stokastik Osilatör
 - MACD
 - Momentum
 - CCI
- c) Günlük dolar kuru
- d) Günlük altın fiyatları
- e) Günlük Dow Jones indeksi
- f) Günlük NIKKEI indeksi
- g) Günlük S&P indeksi

İzleyen alt kesimlerde yukarıda sıralanan niteliklerin kısa açıklamalarına ve formülasyonlarına yer verilecektir. Çalışmada günlük kapanış fiyatı hedef nitelik geri kalanlar ise açıklayıcı nitelikler olarak ele alınmıştır.

3.3.1. Günlük Kapanış Fiyatları

Çalışmada oluşturulacak regresyon ağaçları modellerinde hedef nitelik olarak günlük fiyat değişimleri (KAP%) ele alınmıştır. KAP% hedef niteliği Eşitlik 3.1 yardımı ile hesaplanır.

$$\%KAP = \frac{KAP_t - KAP_{t-1}}{KAP_t} \quad (3.1)$$

Formülde KAP_t ifadesi o andaki kapanış fiyatını, KAP_{t-1} ise bir gün önceki kapanış fiyatını göstermektedir.

3.3.2. Teknik Analiz Göstergeleri

3.3.2.1. Göreceli Güç İndeksi

Göreceli Güç Endeksi (RSI), bir hisse senedinin incelenen dönem içindeki kapanış değerlerini bir önceki günle karşılaştırarak, menkul kıymetin gücünü bulmaya çalışmaktadır. Bu indeks ilk olarak J. Welles Wilder tarafından kullanılmıştır. RSI'nin amacı, menkul kıymetin değeri göz önüne alınarak, bir dönem içinde bir önceki güne göre yükselen günlerle, bir önceki güne göre düşen günleri birbiri ile karşılaştırmak ve böylece gelecekteki değerleri ile ilgili öngörülerde bulunabilmektir (Özekşi, 2002, s.249). RSI grafik oluşumlarını tanımlamakta yararlı olmasının yanı sıra menkul kıymet değerinin kısa dönem için pahalı ya da ucuz olduğu seviyeyi belirlemekte kullanılır. Ayrıca geleneksel grafik metod ve şekilleri ile görülmesi zor olan destek ve direnç seviyelerini çoğu kez daha önce ve daha belirgin biçimde gösterdiği belirlenmiştir (Aybars, 1997, s.57).

RSI hesaplanması için 14 günlük kapanış değerleri gerekir. Hesaplama bir defa yapıldıktan sonra ondördüncü günden itibaren sadece bir önceki günün kapanış değerini kullanmak yeterlidir. RSI'nın hesaplanmasında kullanılan bağıntı Eşitlik 3.2 verildiği gibidir.

$$RSI = 100 - \frac{100}{1 + RS} \quad (3.2)$$

Eşitlikte yer alan RS değeri, N adet günün pozitif kapanış değerlerinin ortalamasının, N adet günlük negatif kapanış değerlerinin ortalamasına bölünmesi ile elde edilir (Aybars, 1997, s.58). N adet günün pozitif kapanış ortalaması U ve N adet günün negatif kapanış ortalaması ise D ile gösterilirse RS Eşitlik 3.3'deki gibi ifade edilebilir.

$$RS = \frac{U}{D} \quad (3.3)$$

Eşitlik 3.2 ve 3.3 yardımıyla RSI Eşitlik 3.4'teki gibi elde edilir.

$$RSI = 100 - \frac{100}{1 + \frac{U}{D}} = \frac{100 U}{U + D} \quad (3.4)$$

3.3.2.2. Stokastik Osilator

Menkul kıymet teknik analizlerinde kullanılan bir diğer gösterge Stokastik Osilatör (STO) adıyla bilinir. STO, belirlenen bir dönem için kapanış fiyatlarını günlük en yüksek ve en düşük seviyelerine göre ölçen ve fiyatların dönüm noktalarını belirleyen bir göstergedir (Aybars, 1997, s.66). STO bugünkü kapanış ile n-günlük periyodun en yüksek ya da en düşük değerleri arasındaki ilişkiyi gösterir. RSI dan farklı olarak günlük kapanış değerlerinin yerine K ve D ile gösterilen iki eğri kullanılır ve bu eğrilerin birbirlerine göre konumlarına bakılarak fiyatların aşırı yüksek yada düşük olduğu konusunda karar verilir (Person, 2004, s.145). Diğer bir deyişle, periyot içindeki en yüksek ve en düşük seviyelere olan uzaklığına göre fiyatın değerlendirilmesidir (Özekşi, 2002, s.265). Stokastik, grafik üzerinde bir düz bir de kesikli çizgiyle gösterilir. Bunlardan düz çizgi ile gösterileni K, kesikli çizgi ile gösterilen ise D'dir (Özekşi, 2002, s.266).

Stokastik eğrilerin elde edilmesi için üç değişkene ihtiyaç vardır. Söz konusu değişkenler Ham Değer (HD), K ve D'dir. Periyot (gün sayısı) 20 olarak seçilirse HD Eşitlik 3.5.'de verildiği gibi hesaplanır.

$$HD = \frac{100(\text{son günün kapanış değeri} - 20 \text{ gün kapanışlarının en düşüğü})}{(20 \text{ gün kapanışlarının en yükseği} - 20 \text{ gün kapanışlarının en düşüğü})} \quad (3.5)$$

HD hesabı için 20 günlük en yüksek, en düşük ve kapanış fiyat hareketleri bilinmelidir. HD hesaplandıktan sonra K ve D hesaplanır. K, HD'nin üç günlük ortalaması D ise K'nın üç günlük ortalamasıdır. Bu değerlerin hesaplanmasından sonra

takip eden günlerin değerleri için (3.6) ve (3.7) formülleri kullanılır (Aybars, 1997, s.67). Yeni K değeri K_t , önceki K değeri ise K_{t-1} biçiminde; yeni D değeri D_t , önceki D değeri ise D_{t-1} ifade edilirse, K_t ve D_t 'nin hesaplanması sırasıyla Eşitlik 3.6. ve 3.7'de verildiği gibidir.

$$K_t = \frac{2}{3}K_{t-1} + \frac{1}{3}HD \quad (3.6)$$

$$D_t = \frac{2}{3}D_{t-1} + \frac{1}{3}K_t \quad (3.7)$$

STO, 0 ile 100 arasındaki değerlerden oluşur. Bu skala üzerine 20 ve 80 değerlerinden çizilen yatay çizgiler STO referans değerleridir. 20 ve 80 değerleri stokastik hesaplamasında genel kabul gören değerlerdir. Bir senedin stokastik değerinin 20 referans değerine ulaşması ya da altına inmesi aşırı satıma ve bir süre sonra toparlanma olacağına, 80 referans değerine ulaşması ya da üzerine çıkması ise aşırı alıma ve yakın gelecekteki gevşemeye işaret eder (Özekşi, 2002, s.269).

3.3.2.3. MACD

MACD, iki hareketli ortalamanın farkından yola çıkarak orta vadeli alım satım sinyalleri üreten ve piyasadaki fiyat hareketlerinin ne yöne doğru eğilimli olduğu sorusu ile ilgilenen bir göstergedir. MACD 12 ve 26 günlük üssel hareketli ortalamanın farkı alınarak hesaplanır. MACD'deki hareketin şiddetini belirleyen, bu iki üssel ortalamanın birbirlerinden uzaklaşma hızlarıdır (Person, 2004, s.143). MACD'nin hesaplanması Eşitlik 3.8'de verildiği gibidir.

$$MACD = \begin{array}{cc} 12 \text{ günlük üssel} & - \quad 26 \text{ günlük üssel} \\ \text{hareketli ortalama} & \text{hareketli ortalama} \end{array} \quad (3.8)$$

12 günlük üssel hareketli ortalama 26 günlük üssel hareketli ortalamadan büyükse gösterge değeri sıfır ekseninin üzerinde, küçükse olması durumunda ise eksenin

altında bir değere sahip olur. Eşitlik durumunda ise gösterge değeri sıfıra eşit olur (Weir, 2006, s.130).

3.3.2.4. Momentum

Momentum (MOM), belli bir zaman dilimi içerisinde fiyatların ne yönde, ne miktarda ve ne şiddette hareket ettiğini açıklayan bir göstergedir (Weir, 2006, s.135). Bugünkü kapanış ile, n gün önceki kapanışı karşılaştıran MOM, o güne kıyasla şu an nerede bulunulduğunu belirler. Eşitlik 3.9'da nasıl hesaplanacağı görülen MOM, n gün öncesine göre ne oranda değişme gösterildiği ile ilgilenir.

$$MOM = \frac{\text{Son günün kapanışı}}{\text{n gün önceki kapanış}} (100) \quad (3.9)$$

Yukarıdaki formülde MOM'u hesaplayabilmek için Son günün kapanışı belirlenen periyot kadar önceki günün kapanışına bölünür. 100 referans değerinden hızla uzaklaşan momentum değerleri, fiyatların ve trendin değişiminin MOM'da gösterdiği yönde artarak devam ettiğine, ancak aşırı alım ya da satım bölgesinde olduğundan bir dip ya da tepe oluşumuna hazır olunması gerektiğine aksi takdirde yavaşlayan bir MOM trendin sonlanmak üzere olduğuna işaret eder (Özekşi, 2002, s.290). MOM'da 12-14 günlük periyotlar daha olumlu sonuçlar verir (Özekşi, 2002, s.289).

3.3.2.5. CCI

Menkul kıymet fiyatının kendi basit hareketli ortalamasının ne tarafında ve ne kadar uzağında olduğunu belirliyeabilmek ve fiyatların gelecekteki olası hareketleri için öngöründe bulunabilmek için CCI (Commodity Channel Index) göstergesi kullanılır. Bu gösterge D. Lambert tarafından geliştirilmiştir (Özekşi, 2002, s.237).

CCI hesaplanmasında fiyat yerine ortalama fiyat kavramı kullanılır. Fiyat yerine kullanılan rakam, o günün en yüksek, en düşük ve kapanış değerlerinin toplamalarının üçe bölünmesi ile bulunan ortalama fiyattır (Özekşi, 2002, s.238). CCI'nın formülasyonu Eşitlik 3.10 verildiği gibidir.

$$CCI = \frac{\text{Ortalama Fiyat} - \text{Ortalama Fiyatlar Hareketli ortalaması}}{(0,015)\text{Sapma Ortalaması}} \quad (3.10)$$

Formülde yer alan “ortalama fiyat” aşağıdaki şekilde hesaplanır.

$$\text{Ortalama Fiyat} = \frac{\text{Günün en yükseği} + \text{Günün en düşüğü} + \text{Kapanış}}{3} \quad (3.11)$$

3.3.3. Diğer Nitelikler

Çalışmada, oluşturulacak modellere ele alınan menkul kıymetlerin günlük kapanış fiyatları ve yukarıda açıklamalarına yer verilen teknik göstergelerinin yanısıra, Tablo 3.3’de verilen nitelikler de dahil edilmiştir.

Tablo 3.3. Araştırmada Kullanılan Diğer Nitelikler.

Nitelik	Anlamı
%DOL	Dolar kurundaki değişme
%S&P	Standard & Poor uluslararası menkul kıymetler indeksindeki değişme
%DOW	ABD menkul kıymetler indeksi Dow Jones’taki değişme
%NIK	Japon menkul kıymetler indeksi Nikkei’deki değişme
%ALT	Altın fiyatlarındaki değişme

Tablo 3.3’deki yüzde değişmeler, t zamanı göstermek üzere sırasıyla aşağıda belirtilen eşitliklere göre hesaplanır.

$$\%DOL = \frac{DOL_t - DOL_{t-1}}{DOL_t} \quad (3.12)$$

$$\%S\&P = \frac{S\&P_t - S\&P_{t-1}}{S\&P_t} \quad (3.13)$$

$$\%DOW = \frac{DOW_t - DOW_{t-1}}{DOW_t} \quad (3.14)$$

$$\%NIK = \frac{NIK_t - NIK_{t-1}}{NIK_t} \quad (3.15)$$

$$\%ALT = \frac{ALT_t - ALT_{t-1}}{ALT_t} \quad (3.16)$$

3.4. İMKB 30 MENKUL KIYMETLERİ İÇİN REGRESYON AĞACI MODELİ

3.4.1. Eğitim ve Test Kümesinin Belirlenmesi

İMKB 30'daki şirketlere ait regresyon ağacı modellerini oluşturmak üzere Tablo 3.2 de ayrıntıları verilen veri kümesi kullanılmıştır. Söz konusu modellerin oluşturulması ve test edilmesi için ele alınan veri skümesinin %66,66'sı eğitim ve geriye kalan %33,34'ü ise test kümesi olmak üzere iki parçaya ayrılmıştır. Eğitim ve test kümelerinin oluşturulması tesadüfi olarak gerçekleştirilmiştir.

3.4.2. Regresyon Ağacı Modelleri

Çalışmada, İMKB 30'daki şirketlere ilişkin oluşturulan regresyon ağacı modellerinin başarılarını değerlendirmek üzere ortalama hata ve RMS hata değerlerinden yararlanılmıştır. Gerek eğitim kümesi gerekse test kümesinin başarıları bu hata değerlerine göre değerlendirilmiştir. Daha sonra aynı veriler kullanılarak oluşturulan ÇDR modellerinin başarıları da ortalama hata ve RMS hata değerlerine göre incelenmiştir.

Eğitim ve test kümeleri için İMKB 30'da işlem gören şirketlere ilişkin oluşturulan regresyon ağacı modellerinin ortalama hata ve RMS hata değerleri Tablo 3.4'de verildiği gibidir.

Tablo 3.4. Regresyon Ağacı Modellerinin Ortalama Hata ve RMS Hata Değerleri.

Menkul Kıymet	Nitelikler	Eğitim Verisi		Test Verisi	
		Ortalama Hata	RMS Hata	Ortalama Hata	RMS Hata
AEFES	CCI, STO	-1,30024E-08	0,152044293	-0,012809757	0,396677658
AKBNK	CCI, MACD, STO	8,69826E-09	0,338316411	0,00921957	0,144009861
AKGRT	CCI, STO	2,42291E-08	0,22991636	-0,00284684	0,197993257
ASYAB	CCI, MACD, DOW%	-2,7248E-09	0,061912484	0,004681016	0,0245632
BAGFS	CCI, MACD	3,96515E-08	0,229721111	0,004536148	0,115478281
DOHOL	CCI, MACD	-1,11849E-08	0,662034956	0,012859322	0,118665711
DYHOL	CCI, MOM, RSI	-2,78884E-08	0,137243847	-0,010881803	0,400539249
ENKA	CCI, MACD	-3,64964E-08	0,426817461	0,008419117	0,22836976
EREGL	CCI, MACD, STO	-1,14114E-08	0,264594265	-0,05245633	2,318952577
GARAN	CCI, STO	-3,59928E-09	0,198603326	0,006203455	0,112176101
HALKB	CCI, MACD, DOL%, DOW%	-1,76056E-09	0,026745338	-0,00097506	0,026470658
ISBTR	CCI, MOM, MACD, ALT%	-1,11513E-08	0,495931134	-0,044654512	2,236397327
ISGYO	CCI, MACD, DOW%	-1,89732E-08	0,400008773	0,005505779	0,148209493
KCHOL	CCI, STO, MACD	-2,97119E-08	0,613597927	0,004863023	0,538794958
KRDMD	CCI, STO	4,22397E-08	0,249352508	0,005381312	0,170403921
SAHOL	CCI, STO	-1,69182E-08	0,300279264	-0,012465459	0,778702582
SKBNK	CCI, STO	-2,74157E-08	0,120674552	-0,005559353	0,162722976
SISE	CCI, STO, MACD	3,0012E-08	0,201944467	0,000531768	0,129529962
PETKM	CCI, MACD, MOM	-6,0024E-08	0,201771456	-0,01913408	0,94447262
TAVHL	CCI, MACD, STO	1,93165E-08	0,129332213	-0,02596803	0,261691718
TEBNK	CCI, STO, MACD	-8,39832E-09	0,338196068	0,003068805	0,244864695
THY	CCI, STO, MACD	3,1719E-08	0,138392381	-0,00881473	0,309406131
TKFEN	CCI, STO, NIK%-DOL%	2,73109E-08	0,025618105	-0,00035184	0,031284393
TOASO	CCI, MACD, STO	-1,0989E-08	0,110270155	-0,00678119	0,30318927
TCELL	CCI, STO, NIK%	-1,53121E-08	0,360910008	-0,008943998	0,445905118
TUPRPS	CCI, STO, MACD	-1,44201E-08	0,229766163	0,005924537	0,107676595
TSKB	CCI, MACD	1,38387E-08	0,437800235	-0,01195646	0,718476877
TTKOM	CCI, MACD, DOW%, DOL%	1,00756E-08	0,015631671	-0,00126611	0,019662303
VKBNK	CCI, NIK%, DOL%, DOW%	-9,82801E-09	0,023697388	-0,002603027	0,05197582
YKBNK	CCI, MACD, STO	-2,54873E-08	0,285798761	0,007218207	0,118890942

3.4.3. Çoklu Doğrusal Regresyon Modelleri

İMKB 30’da yer alan şirketler için belirlenen hedef ve açıklayıcı nitelikler kullanılarak ÇDR modelleri oluşturulmuştur. ÇDR modellerinde yer alan parametrelerin anlamlılıklarını sınamak üzere p değerinden yararlanılmıştır. Buna göre $p < 0,05$ koşulunu sağlayan nitelikler önemli nitelikler olarak belirlenmiştir. İMKB 30’da yer alan şirketler için anlamlı bulunan nitelikler ve oluşturulan ÇDR modellerinin ortalama hata ve RMS hata değerleri Tablo 3.5’de verildiği gibi olup modellerin ayrıntılarına EK 1’de yer verilmiştir.

Tablo 3.5. ÇDR Modellerinin Ortalama Hata ve RMS Hata Değerleri.

Menkul Kıymet	Nitelikler	Eğitim Verisi		Test Verisi	
		Ortalama Hata	RMS Hata	Ortalama Hata	RMS Hata
AEFES	MACD, CCI	-7,07825E-12	0,149125323	-0,012448646	0,373081011
AKBNK	STO, CCI	6,4213E-08	0,322353263	0,009180733	0,135733982
AKGRT	CCI, MOM	1,7657E-07	0,215280672	-0,00075516	0,176476152
ASYAB	CCI, STO, DOW%	1,89011E-07	0,042968207	0,005106298	0,042542977
BAGFS	STO, MOM, CCI	-7,28775E-08	0,21348991	-0,000110785	0,11576537
DOHOL	STO, CCI	-2,61362E-07	0,643068945	0,006069984	0,147228922
DYHOL	CCI	1,00716E-08	0,137433898	-0,01021094	0,384406591
ENKA	CCI, MACD	5,18476E-09	0,422449654	0,011652097	0,228257912
EREGL	RSI, STO, MOM, CCI, DOL%, DOW%	3,2744E-07	0,240044554	-0,04647415	2,27894308
GARAN	RSI, STO, MACD, MOM, CCI, DOL%	3,18693E-07	0,161745186	0,008648967	0,124680004
HALKB	STO, MOM, CCI, DOL%, DOW%, ALT%	7,31966E-08	0,023857182	-0,00033539	0,023542502
ISBTR	STO, MOM, CCI	2,71424E-07	0,476342198	-0,04135305	2,240501609
ISGYO	MACD, MOM, CCI	1,72779E-07	0,33789261	0,010359333	0,20761895
KCHOL	STO, CCI	1,09733E-07	0,60134164	0,005896856	0,537227797
KRDMD	RSI, MACD, MOM, CCI, DOL%	-3,66679E-08	0,204695347	0,003317217	0,177053375
SAHOL	RSI, MOM, CCI	8,45647E-08	0,24331014	-0,02714176	0,753000002
SKBNK	RSI, STO, MOM, CCI, DOL%, DOW%	-1,37266E-09	0,089287182	-0,003775952	0,16122723
SISE	RSI, STO, MOM, CCI	2,10115E-07	0,186669424	0,003625504	0,113294397
PETKM	MOM, CCI	4,65661E-07	0,204035442	-0,01822885	0,924800188
TAVHL	RSI, STO, MACD, MOM, CCI, S&P%, DOW%	2,14074E-07	0,109019281	-0,02783755	0,208287154
TEBNK	STO, CCI, DOL%	-8,78828E-09	0,323817818	0,000441134	0,199634034
THY	STO, CCI	1,08516E-07	0,133318765	-0,0081587	0,304937074
TKFEN	STO, CCI, DOL, S&P%, DOW%, NIK%, ALT%	1,32998E-07	0,022994229	-0,00269161	0,024736964
TOASO	RSI, MOM, CCI	-6,01654E-07	0,106863372	-0,006547508	0,303339679
TCELL	MACD, MOM, CCI	5,25842E-08	0,305551684	-3,81777E-05	0,451005868
TUPRPS	STO, MOM, CCI	-6,07094E-08	0,213353061	0,005399403	0,117156491
TSKB	MACD, CCI	4,44585E-09	0,422835643	-0,01066479	0,714942263
TTKOM	STO, CCI, S&P%, DOW%, ALT%	9,06748E-08	0,014687713	-0,00099463	0,017118321
VKBNK	RSI, CCI, DOL%, DOW%	1,49024E-07	0,02468415	-0,00319876	0,042081385
YKBNK	MACD, MOM, CCI	3,26995E-07	0,267188274	0,010226208	0,134361686

3.4.4.ÇDR Modelinde Yer Alan Niteliklerin Regresyon Ağaçlarında Kullanımı

Bu alt kesimde çalışmada ele alınan tüm açıklayıcı nitelikler kullanılarak oluşturulan regresyon ağacı modellerinde bu niteliklerin hepsini kullanmaktansa sadece ÇDR modelleri ile anlamlı bulunan açıklayıcı değişkenlerin kullanılması sonucunda regresyon ağaçlarının yeniden düzenlenmesinin ne ölçüde geçerli olacağı araştırılmıştır. Bu amaçla her bir şirket için oluşturulan ÇDR modeli sonucunda elde edilen anlamlı niteliklerin ilgili şirketin regresyon ağacını yeniden oluşturmak üzere açıklayıcı nitelikler olarak ele alınması sağlanmıştır. ÇDR modelinden elde edilen niteliklerin, daha düşük ortalama hataya sahip regresyon ağaçlarının elde edilmesinde

bir nitelik seçim yöntemi olarak kullanılabileceği düşüncesinden hareketle aşağıda uygulama adımları verilen algoritma geliştirilmiştir.

Algoritma: Regresyon ağaçlarında ortalama hatayı azaltma işlemi.

Girdi: Açıklayıcı nitelikler, menkul kıymet kapanış fiyatlarındaki değişimler.

Çıktı: Regresyon ağacı.

Yöntem: Algoritmanın adımları aşağıdaki gibidir.

Adım 1: Tüm nitelikleri kullanarak RA_1 regresyon ağacını oluşturun.

Adım 2: Tüm nitelikler için ÇDR modelini oluşturun.

Adım 3: ÇDR modelinde $p < 0.05$ koşuluna uygun nitelikleri seçin.

Adım 4: ÇDR modelinden seçilen nitelikleri kullanarak RA_2 regresyon ağacını oluşturun.

Adım 5: Eğer (RA_1 ortalama hatası $<$ RA_2 ortalama hatası) ise RA_1 regresyon ağacı modelini, aksi takdirde RA_2 modelini kabul edin.

Adım 6: Algoritmayı sonlandırın.

Algoritmanın işletilmesi sonucunda oluşturulan yeni regresyon ağacı modellerinin ortalama hata ve RMS hata değerleri Tablo 3.6'daki gibidir.

Tablo 3.6. Yeni Regresyon Ağacı Modellerinin Ortalama Hata ve RMS Hata Değerleri.

Menkul Kıymet	Nitelikler	Eğitim Verisi		Test Verisi	
		Ortalama Hata	RMS Hata	Ortalama Hata	RMS Hata
AEFES	MACD, CCI	-1,59574E-08	0,152057042	-0,01268829	0,396688924
AKBNK	STO, CCI	-1,55969E-08	0,338316898	0,009085846	0,144024018
AKGRT	CCI, MOM	4,36123E-08	0,229920936	-0,0028715	0,197995281
ASYAB	CCI, DOW%	-2,7248E-09	0,061939862	0,00482912	0,024676737
BAGFS	CCI	3,60469E-09	0,229745479	0,004285447	0,115543958
DOHOL	STO, CCI	-1,04858E-09	0,662044499	0,013066506	0,118629488
DYHOL	CCI	5,47809E-09	0,13739334	-0,01098385	0,400585812
ENKA	CCI, MACD	-3,64964E-08	0,426817461	0,008419117	0,22836976
EREGL	CCI, STO	-1,32132E-08	0,264606331	-0,052460324	2,318954423
GARAN	CCI, STO	-3,59928E-09	0,198603326	0,006203455	0,112176101
HALKB	STO, CCI, DOL%, DOW%	5,28169E-09	0,02675574	-0,00080933	0,026480568
ISBTR	STO, MOM, CCI	-5,12357E-09	0,49594771	-0,044651386	2,236403563
ISGYO	MACD, CCI	-3,73884E-08	0,400010179	0,005428334	0,148206399
KCHOL	STO, CCI	-2,13085E-08	0,613601096	0,004941978	0,538795159
KRDMD	MACD, CCI	4,86248E-08	0,249356165	0,00537049	0,170399007
SAHOL	RSI, CCI	-9,14495E-10	0,300289575	-0,012249023	0,778703645
SKBNK	CCI, STO	-2,74157E-08	0,120674552	-0,005559353	0,162722976
SISE	CCI, STO	3,15126E-08	0,201959369	0,000528136	0,129582561
PETKM	MOM, CCI	-4,20168E-08	0,201794853	-0,018985313	0,944479569
TAVHL	CCI, MACD, STO	1,93165E-08	0,129332213	-0,02596803	0,261691718
TEBNK	STO, CCI, DOL%	1,10978E-08	0,338203301	0,003046282	0,244853571
THY	STO, CCI	-4,43383E-09	0,138436559	-0,008677753	0,309425641
TKFEN	STO, CCI, DOL, NIK%	2,73109E-08	0,025618105	-0,00035184	0,031284393
TOASO	MOM, CCI	-1,72684E-08	0,110290517	-0,00687549	0,303195291
TCELL	MACD, CCI	4,71143E-08	0,360922123	-0,00879575	0,445911463
TUPRPS	STO, CCI	-2,63323E-08	0,229800962	0,005825323	0,107793926
TSKB	MACD, CCI	1,38387E-08	0,437800235	-0,01195646	0,718476877
TTKOM	STO, CCI, DOW%	2,267E-08	0,015731754	-0,00085856	0,019204341
VKBNK	CCI, DOL%, DOW%	-1,84275E-08	0,023701875	-0,002840275	0,052196522
YKBNK	MACD, CCI	-2,30885E-08	0,285801666	0,007151257	0,11890213

Herhangi bir ayırım gözetmeksizin belirlenen tüm açıklayıcı nitelikleri kullanmak suretiyle elde edilen regresyon ağacı modellerinin ortalama hata ve RMS hata değerlerinin yer aldığı Tablo 3.4'deki sonuçlar ile ÇDR modelleri sonucunda anlamlı bulunan açıklayıcı niteliklere göre yeniden oluşturulan regresyon ağaçlarına ilişkin hata değerlerinin yer aldığı Tablo 3.6'daki sonuçlar karşılaştırıldığında, ÇDR yönteminin bir nitelik seçim işlemi olarak uygulanması durumunda, bazı şirketler için daha düşük ortalama hataya sahip regresyon ağacı modellerine ulaşıldığı görülmüştür. Her iki tabloda yer alan sonuçların, en düşük hataya sahip regresyon ağacı modeli asıl model olacak şekilde seçilmek suretiyle birleştirilmesi Tablo 3.7'deki gibidir. EK 2'de seçilen modellerin ağaç diyagramlarına yer verilmiştir.

Tablo 3.7. Sonuç Olarak Elde Edilen Regresyon Ağacı Modelleri.

Menkul Kıymet	Nitelikler	Eğitim Verisi		Test Verisi	
		Ortalama Hata	RMS Hata	Ortalama Hata	RMS Hata
AEFES	CCI, STO	-1,30024E-08	0,152044293	-0,012809757	0,396677658
AKBNK	CCI, MACD, STO	8,69826E-09	0,338316411	0,00921957	0,144009861
AKGRT	CCI, STO	2,42291E-08	0,22991636	-0,00284684	0,197993257
ASYAB	CCI, MACD, DOW%	-2,7248E-09	0,061912484	0,004681016	0,0245632
BAGFS	CCI	3,60469E-09	0,229745479	0,004285447	0,115543958
DOHOL	CCI, STO	-1,04858E-09	0,662044499	0,013066506	0,118629488
DYHOL	CCI	5,47809E-09	0,13739334	-0,01098385	0,400585812
ENKA	CCI, MACD	-3,64964E-08	0,426817461	0,008419117	0,22836976
EREGL	CCI, MACD, STO	-1,14114E-08	0,264594265	-0,05245633	2,318952577
GARAN	CCI, STO	-3,59928E-09	0,198603326	0,006203455	0,112176101
HALKB	CCI, MACD, DOL%, DOW%	-1,76056E-09	0,026745338	-0,00097506	0,026470658
ISBTR	CCI, MOM, STO	-5,12357E-09	0,49594771	-0,044651386	2,236403563
ISGYO	CCI, MACD, DOW%	-1,89732E-08	0,400008773	0,005505779	0,148209493
KCHOL	CCI, STO	-2,13085E-08	0,613601096	0,004941978	0,538795159
KRDMD	CCI, STO	4,22397E-08	0,249352508	0,005381312	0,170403921
SAHOL	CCI, RSI	-9,14495E-10	0,300289575	-0,012249023	0,778703645
SKBNK	CCI, STO	-2,74157E-08	0,120674552	-0,005559353	0,162722976
SISE	CCI, STO, MACD	3,0012E-08	0,201944467	0,000531768	0,129529962
PETKM	CCI, MOM	-4,20168E-08	0,201794853	-0,018985313	0,944479569
TAVHL	CCI, MACD, STO	1,93165E-08	0,129332213	-0,02596803	0,261691718
TEBNK	CCI, STO, MACD	-8,39832E-09	0,338196068	0,003068805	0,244864695
THY	CCI, STO	-4,43383E-09	0,138436559	-0,008677753	0,309425641
TKFEN	CCI, STO, NIK%, DOL%	2,73109E-08	0,025618105	-0,00035184	0,031284393
TOASO	CCI, MACD, STO	-1,0989E-08	0,110270155	-0,00678119	0,30318927
TCELL	CCI, STO, NIK%	-1,53121E-08	0,360910008	-0,008943998	0,445905118
TUPRPS	CCI, STO, MACD	-1,44201E-08	0,229766163	0,005924537	0,107676595
TSKB	CCI, MACD	1,38387E-08	0,437800235	-0,01195646	0,718476877
TTKOM	CCI, MACD, DOW%, DOL%	1,00756E-08	0,015631671	-0,00126611	0,019662303
VKBNK	CCI, NIK%, DOL%, DOW%	-9,82801E-09	0,023697388	-0,002603027	0,05197582
YKBNK	CCI, MACD	-2,30885E-08	0,285801666	0,007151257	0,11890213

3.4.5. Karar Kuralları

EK 2’de verilen her bir şirkete ilişkin regresyon ağacı modelinden yararlanarak şirketler için karar kuralları elde edilmiştir. elde edilen karar kuralları EK 3’te verildiği gibidir. Bu kurallar aynı zamanda sınıflayıcı işlevini de yerine getirmektedir.

Örneğin, AEFES menkul kıymeti için elde edilen ve EK 3’te yer alan karar kuralları kullanılarak bir sınıflandırma işlemi yapılabilir. Belirlenen bir tarihte örneğin, CCI = 30, MACD = 3,2 olarak hesaplanmış ise, söz konusu menkul kıymet için belirlenen kurallar arasında yer alan aşağıdaki kuralın kullanılabileceği anlaşılır.

KURAL6: EĞER CCI>-86,07 İSE VE
CCI > -0,875 İSE VE
CCI ≤ 84,55 İSE VE
CCI ≤ 35,3321 İSE VE
MACD > 1,77 İSE
KAP%= 0,003372

Kural içinde yer alan koşullar sınıflandırma işlemini yerine getirmektedir. Bu kurala göre kapanış fiyatlarının belirlenmesinde CCI ve MACD nitelikleri etkili olmaktadır. İlk dört koşul CCI ile ilgilidir. Bu koşullar göz önüne alındığında, $CCI = 30$ için $-0,875 < CCI \leq 35,3321$ koşulunun ve $MACD > 1,77$ koşullarının geçerli olduğu anlaşılmaktadır.

Bu koşulların tümünü karşılayan kapanış fiyatlarındaki değişim yüzdesi 0,003372 olarak belirlenmiştir. Regresyon ağaçlarında karar niteliği sürekli değerlerden oluştuğu için sınıf değeri de sürekli değer olarak kestirilmiştir.

3.5. SONUÇLAR VE DEĞERLENDİRME

Bu çalışmada İMKB 30 kapsamındaki 30 şirkete ilişkin günlük kapanış fiyatları esas alınarak regresyon ağaçları modelleri oluşturulmuş ve bu modellerin performansını artırma yolları araştırılmıştır. Çalışmada, veri madenciliği sınıflandırma ve kestirim yöntemleri arasında yer alan regresyon ağaçları yöntemi ile klasik istatistik konuları arasında yer alan ÇDR yöntemine yer verilmiştir. Çözümlemeler sonucunda şu sonuçlara ulaşılmıştır:

- a) Çalışmanın ilk aşamasında KAP% hedef nitelik olmak üzere RSI, STO, MACD, MOM, CCI, %DOL, %S&P, %DOW, %NIK, %ALT nitelikleri kullanılarak regresyon ağaçları elde edilmiştir. Bu regresyon ağaçlarından daha az ortalama hataya sahip başka regresyon ağaçlarının olup olmadığı araştırılmıştır.
- b) ÇDR yöntemi ile nitelik seçimi yapılarak, ÇDR modeline girebilen nitelikler ile regresyon ağacı modeli yeniden oluşturulmuştur. Elde edilen BAGFS, DOHOL, DYHOL, ISBTR, KCHOL, SAHOL, PETKM, THY, YKBNK menkul kıymetlerine ilişkin regresyon ağaçlarında ortalama hatada bir azalma belirlenmiştir. Bu durumdan hareketle, başlangıçtaki regresyon ağacı modeli yerine, ÇDR modelinden elde edilen niteliklerin yer aldığı daha düşük ortalama hataya sahip ikinci regresyon ağacı modelinin kullanılabileceği anlaşılmış ve bu sonucu elde etmeye yönelik bir algoritma geliştirilmiştir.

- c) ÇDR yöntemi ile nitelik seçimi yapılmasının ardından, ÇDR modeline girebilen nitelikler ile regresyon ağacı modeli yeniden oluşturulduğunda herhangi bir İMKB 30 menkul kıymetine ait RMS hatada bir azalma gözlemlenmemiştir. Bu durumda hatayı azaltmak için ÇDR yöntemiyle nitelik seçimi yapılmasının anlamlı olmadığı ortaya çıkmıştır.
- d) İMKB 30 şirketlerinin tümünde regresyon ağaçlarının CCI niteliğini içerdiği ortaya çıkmıştır. Teknik analizlerde menkul kıymetlerin kapanış fiyatlarının tahmininde kullanılan CCI niteliğinin bu şekilde tüm modellere girmesi beklenen bir sonuç olmuştur.
- e) Menkul kıymetlerin %57'sinin STO, %47'sinin MACD, %17'sinin %DOW, %13'ünün %DOL, %10'unun %NIK, %7'sinin MOM ve %3'ünün RSI niteliğini içerdiği belirlenmiştir. %S&P ve %ALT niteliklerinin regresyon ağacı modellerine giremediği ve böylece sınıflandırma ve kestirim sürecinde etkili olmadığı anlaşılmıştır. %DOL, %DOW ve %NIK endeks değişimlerinin, yani dış borsalardaki gelişmelerin İMKB menkul kıymetleri kapanış fiyatlarında etkili olması ve bazı menkul kıymetlerde modele girmesi iktisadi açıdan beklenen bir sonuç olmuştur.

EKLER

EK 1: İMKB 30 İÇİN ÇDR MODELLERİ

1. AEFES İÇİN ÇDR MODELİ

$$\text{KAP\%} = -0,02077989 + 0,00606698 \text{ MACD} + 0,00034123 \text{ CCI} \\ (0,00802214) \quad (0,00259483) \quad (0,000028)$$

$$R^2 = 0,075512603 \quad S = 0,1492577 \\ \text{RMS Hata} = 0,149125323 \quad \text{Ortalama Hata} = -7,07825\text{E-}12$$

2. AKBNK İÇİN ÇDR MODELİ

$$\text{KAP\%} = 0,06021575 - 0,00112392 \text{ STO} + 0,00072783 \text{ CCI} \\ (0,0171403) \quad (0,00030409) \quad (0,00003748)$$

$$R^2 = 0,114340303 \quad S = 0,32249841 \\ \text{RMS Hata} = 0,322353263 \quad \text{Ortalama Hata} = 6,4213\text{E-}08$$

3. AKGRT İÇİN ÇDR MODELİ

$$\text{KAP\%} = 0,05478835 - 0,0061724 \text{ MOM} + 0,00053821 \text{ CCI} \\ (0,01043943) \quad (0,00009158) \quad (0,00002622)$$

$$R^2 = 0,160082704 \quad S = 0,21542306 \\ \text{RMS Hata} = 0,215280672 \quad \text{Ortalama Hata} = 1,7657\text{E-}07$$

4. ASYAB İÇİN ÇDR MODELİ

$$\text{KAP\%} = 0,03676795 - 0,00074197 \text{ STO} + \\ (0,0051463) \quad (0,00009187) \\ 0,00038829 \text{ CCI} + (0,00001292) \quad 0,35123721 \text{ DOW\%} \\ (0,10477236)$$

$$R^2 = 0,568200335 \quad S = 0,04308577 \\ \text{RMS Hata} = 0,042968207 \quad \text{Ortalama Hata} = 1,89011\text{E-}07$$

5. BAGFS İÇİN ÇDR MODELİ

$$\text{KAP\%} = -0,01494196 - 0,00050094 \text{ STO} + \\ (0,01347474) \quad (0,00019463) \\ 0,00032239 \text{ MOM} + 0,00034028 \text{ CCI} \\ (0,00010317) \quad (0,00001391)$$

$$R^2 = 0,164407896 \quad S = 0,21361829 \\ \text{RMS Hata} = 0,21348991 \quad \text{Ortalama Hata} = -7,28775\text{E-}08$$

6. DOHOL İÇİN ÇDR MODELİ

$$\text{KAP\%} = 0,10157688 - 0,00216308 \text{ STO} + 0,00098333 \text{ CCI} \quad (0,00006983) \\ (0,03890783) \quad (0,00069138)$$

$$R^2 = 0,066725847 \quad S = 0,64340633 \\ \text{RMS Hata} = 0,643068945 \quad \text{Ortalama Hata} = -2,61362\text{E-}07$$

7. DYHOL İÇİN ÇDR MODELİ

$$\text{KAP\%} = -0,00516338 + 0,00017675 \text{ CCI} \\ (0,00306967) \quad (0,00001915)$$

$$R^2 = 0,040721496 \quad S = 0,13750239 \\ \text{RMS Hata} = 0,137433898 \quad \text{Ortalama Hata} = 1,00716\text{E-}08$$

8. ENKA İÇİN ÇDR MODELİ

$$\text{KAP\%} = -0,03708779 + 0,00469702 \text{ MACD} + 0,00045489 \text{ CCI} \\ (0,00995192) \quad (0,00159495) \quad (0,00003846)$$

$$R^2 = 0,045706392 \quad S = 0,42264253 \\ \text{RMS Hata} = 0,422449654 \quad \text{Ortalama Hata} = 5,18476\text{E-}09$$

9. EREGL İÇİN ÇDR MODELİ

$$\text{KAP\%} = 0,10508797 + 0,00113749 \text{ RSI} - 0,00078382 \text{ STO} + \\ (0,01596742) \quad (0,00027158) \quad (0,0025301) \\ 0,00130124 \text{ MOM} + 0,00057574 \text{ CCI} + 0,70645726 \text{ DOL\%} + \\ (0,00008844) \quad (0,00002031) \quad (0,31588054) \\ 0,14517656 \text{ DOW\%} \\ (0,02704395)$$

$$R^2 = 0,211608064 \quad S = 0,240297726 \\ \text{RMS Hata} = 0,240044554 \quad \text{Ortalama Hata} = 3,2744\text{E-}07$$

10. GARAN İÇİN ÇDR MODELİ

$$\text{KAP\%} = 0,23466095 + 0,00090241 \text{ RSI} - 0,0013645 \text{ STO} - 0,0168403 \text{ MACD} \\ (0,01406767) \quad (0,0001924) \quad (0,00017066) \quad (0,00783464) \\ - 0,0020649 \text{ MOM} + 0,00087186 \text{ CCI} + 0,56740677 \text{ DOL\%} \\ (0,00011672) \quad (0,00002011) \quad (0,021880691)$$

$$R^2 = 0,374525885 \quad S = 0,16191526 \\ \text{RMS Hata} = 0,161745186 \quad \text{Ortalama Hata} = 3,18693\text{E-}07$$

11. HALKB İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,00875793 - 0,00047213 \text{ STO} + 0,0003247 \text{ MOM} + \\ & (0,01233628) (0,00007435) (0,00012875) \\ & 0,00025863 \text{ CCI} 0,27219644 \text{ DOL\%} + 0,57578993 \text{ DOW\%} + \\ & (0,00001462) (0,08904546) (0,05872695) \\ & 0,14285855 \text{ ALT\%} \\ & (0,0670371) \end{aligned}$$

$R^2 = 0,51663874$ $S = 0,02400556$
RMS Hata = 0,023857182 Ortalama Hata = 7,31966E-08

12. ISBTR İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,07286902 - 0,00199709 \text{ STO} - 0,0001642 \text{ MOM} + \\ & (0,02082499) (0,00042931) (0,00003846) \\ & 0,00110347 \text{ CCI} \\ & (0,00006122) \end{aligned}$$

$R^2 = 0,090383064$ $S = 0,47662956$
RMS Hata = 0,476342198 Ortalama Hata = 72,71424E-07

13. ISGYO İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,13007078 - 0,42564878 \text{ MACD} - 0,00061158 \text{ MOM} + \\ & (0,01168813) (0,0570794) (0,00013147) \\ & 0,00045975 \text{ CCI} \\ & (0,00001808) \end{aligned}$$

$R^2 = 0,297671808$ $S = 0,33827037$
RMS Hata = 0,33789261 Ortalama Hata = 1,72779E-07

14. KCHOL İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,08795778 - 0,00205935 \text{ STO} + 0,00094012 \text{ CCI} \\ & (0,03130972) (0,00056425) (0,00006675) \end{aligned}$$

$R^2 = 0,057059631$ $S = 0,60161251$
RMS Hata = 0,60134164 Ortalama Hata = 1,09733E-07

15. KRDMMD İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,05839618 + 0,00091554 \text{ RSI} - 0,22573426 \text{ MACD} - \\ & (0,01446245) (0,00026675) (0,03182055) \\ & 0,0008334 \text{ MOM} + 0,000041659 \text{ CCI} + 0,60310262 \text{ DOL\%} \\ & (0,00006461) (0,00001268) (0,36063954) \end{aligned}$$

$$\begin{aligned} R^2 = 0,357260457 & \quad S = 0,20499763 \\ \text{RMS Hata} = 0,204695347 & \quad \text{Ortalama Hata} = -3,66679\text{E-}08 \end{aligned}$$

16. SAHOL İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,08814565 - 0,0012492 \text{ RSI} - 0,00030457 \text{ MOM} + \\ & (0,01713482) (0,00030771) (0,00006675) \\ & 0,00076476 \text{ CCI} \\ & (0,00002167) \end{aligned}$$

$$\begin{aligned} R^2 = 0,365271122 & \quad S = 0,24353296 \\ \text{RMS Hata} = 0,24331014 & \quad \text{Ortalama Hata} = 8,45647\text{E-}08 \end{aligned}$$

17. SKBNK İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,10480124 + 0,00039626 \text{ RSI} + 0,00046065 \text{ STO} - \\ & (0,00752441) (0,0001202) (0,00011004) \\ & 0,0010373 \text{ MOM} + 0,00052723 \text{ CCI} + 0,44231457 \text{ DOL\%} + \\ & (0,0000437) (0,00001153) (0,15897435) \\ & 0,43336916 \text{ DOW\%} \\ & (0,14567037) \end{aligned}$$

$$\begin{aligned} R^2 = 0,499322657 & \quad S = 0,08942796 \\ \text{RMS Hata} = 0,089287182 & \quad \text{Ortalama Hata} = -1,37266\text{E-}09 \end{aligned}$$

18. SISE İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,04514087 + 0,0006656 \text{ RSI} - 0,00056394 \text{ STO} - \\ & (0,01350936) (0,00021457) (0,00018856) \\ & 0,00056377 \text{ MOM} + 0,00038612 \text{ CCI} \\ & (0,00011053) (0,00001425) \end{aligned}$$

$$\begin{aligned} R^2 = 0,190760204 & \quad S = 0,18680964 \\ \text{RMS Hata} = 0,186669424 & \quad \text{Ortalama Hata} = 2,10115\text{E-}07 \end{aligned}$$

19. PETKM İÇİN ÇDR MODELİ

$$\text{KAP\%} = -0,08804107 + 0,00077389 \text{ MOM} + 0,0000644 \text{ CCI} \\ (0,01367833) \quad (0,00012847) \quad (0,00001265)$$

$$R^2 = 0,023523978 \quad S = 0,20412736 \\ \text{RMS Hata} = 0,204035442 \quad \text{Ortalama Hata} = 4,65661\text{E-}07$$

20. TAVHL İÇİN ÇDR MODELİ

$$\text{KAP\%} = -0,01165095 - 0,00082632 \text{ RSI} - 0,00201652 \text{ STO} + \\ (0,0296384) \quad (0,000031468) \quad (0,00030208) \\ 0,04094522 \text{ MACD} + 0,00130838 \text{ MOM} + 0,00079511 \text{ CCI} - \\ (0,012977) \quad (0,00023649) \quad (0,00005615) \\ 0,73873121 \text{ S\&P\%} + 0,60675555 \text{ DOW\%} \\ (0,19896482) \quad (0,1632385)$$

$$R^2 = 0,356475856 \quad S = 0,10967308 \\ \text{RMS Hata} = 0,109019281 \quad \text{Ortalama Hata} = 2,14074\text{E-}07$$

21. TEBNK İÇİN ÇDR MODELİ

$$\text{KAP\%} = 0,05659601 - 0,00125048 \text{ STO} + 0,00070754 \text{ CCI} + \\ (0,01724903) \quad (0,00030352) \quad (0,00003669) \\ 1,47385383 \text{ DOL\%} \\ (0,50431931)$$

$$R^2 = 0,102068458 \quad S = 0,32401225 \\ \text{RMS Hata} = 0,323817818 \quad \text{Ortalama Hata} = -8,78828\text{E-}09$$

22. THY İÇİN ÇDR MODELİ

$$\text{KAP\%} = 0,00738914 - 0,00024116 \text{ STO} + 0,00031301 \text{ CCI} \\ (0,00673119) \quad (0,00012764) \quad (0,00001511)$$

$$R^2 = 0,1347239 \quad S = 0,13338701 \\ \text{RMS Hata} = 0,133318765 \quad \text{Ortalama Hata} = 1,08516\text{E-}07$$

23. TKFEN İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,01707774 - 0,00031781 \text{ STO} + 0,0002327 \text{ CCI} + \\ & (0,00378825) (0,00006864) (0,00001387) \\ & 0,31728965 \text{ DOL\%} - 0,11274907 \text{ S\&P\%} + 0,29984024 \text{ DOW\%} + \\ & (0,11562901) (0,05581708) (0,06750228) \\ & 0,17396079 \text{ NIK\%} + 0,22914799 \text{ ALT\%} \\ & (0,05944917) (0,06977896) \end{aligned}$$

$$\begin{aligned} R^2 &= 0,4979569702 & S &= 0,02318993 \\ \text{RMS Hata} &= 0,022994229 & \text{Ortalama Hata} &= 1,32998\text{E-}07 \end{aligned}$$

24. TOASO İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & -0,0432539 + 0,00025853 \text{ RSI} + 0,0002545 \text{ MOM} + \\ & (0,00668093) (0,00010927) (0,00004845) \\ & 0,00019308 \text{ CCI} \\ & (0,00001001) \end{aligned}$$

$$\begin{aligned} R^2 &= 0,141272375 & S &= 0,10693054 \\ \text{RMS Hata} &= 0,106863372 & \text{Ortalama Hata} &= -6,01654\text{E-}07 \end{aligned}$$

25. TCELL İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,4753204 - 0,02175616 \text{ MACD} - 0,0045755 \text{ MOM} + \\ & (0,03412458) (0,01166538) (0,00030541) \\ & 0,00098698 \text{ CCI} \\ & (0,00003636) \end{aligned}$$

$$\begin{aligned} R^2 &= 0,305082045 & S &= 0,3059122 \\ \text{RMS Hata} &= 0,305551684 & \text{Ortalama Hata} &= 5,25842\text{E-}08 \end{aligned}$$

26. TUPRPS İÇİN ÇDR MODELİ

$$\begin{aligned} \text{KAP\%} = & 0,00585048 - 0,00099708 \text{ STO} + 0,00037699 \text{ MOM} + \\ & (0,01845159) (0,00021323) (0,0001746) \\ & 0,00059189 \text{ CCI} \\ & (0,00002429) \end{aligned}$$

$$\begin{aligned} R^2 &= 0,171178394 & S &= 0,21348695 \\ \text{RMS Hata} &= 0,213353061 & \text{Ortalama Hata} &= -6,07094\text{E-}08 \end{aligned}$$

27. TSKB İÇİN ÇDR MODELİ

$$\text{KAP\%} = 0,03278211 - 0,19212124 \text{ MACD} + 0,00053457 \text{ CCI} \\ (0,01203331) \quad (0,04254613) \quad (0,00003135)$$

$$R^2 = 0,080613097 \quad S = 0,42302659 \\ \text{RMS Hata} = 0,422835643 \quad \text{Ortalama Hata} = 4,44585\text{E-}09$$

28. TTKOM İÇİN ÇDR MODELİ

$$\text{KAP\%} = 0,01195403 - 0,00022541 \text{ STO} + 0,00016327 \text{ CCI} - \\ (0,0027875) \quad (0,00004954) \quad (0,00000965) \\ 0,10007128 \text{ S\&P\%} + 0,18961039 \text{ DOW\%} + 0,11839493 \text{ ALT\%} \\ (0,03574177) \quad (0,04193242) \quad (0,05100165)$$

$$R^2 = 0,505575043 \quad S = 0,01479998 \\ \text{RMS Hata} = 0,014687713 \quad \text{Ortalama Hata} = 9,06748\text{E-}08$$

29. VKBNK İÇİN ÇDR MODELİ

$$\text{KAP\%} = -0,00952947 + 0,00019462 \text{ RSI} + 0,00012394 \text{ CCI} + \\ (0,00276554) \quad (0,00005113) \quad (0,00000887) \\ 0,26479378 \text{ DOL\%} + 0,55790257 \text{ DOW\%} \\ (0,0898831) \quad (0,05759098)$$

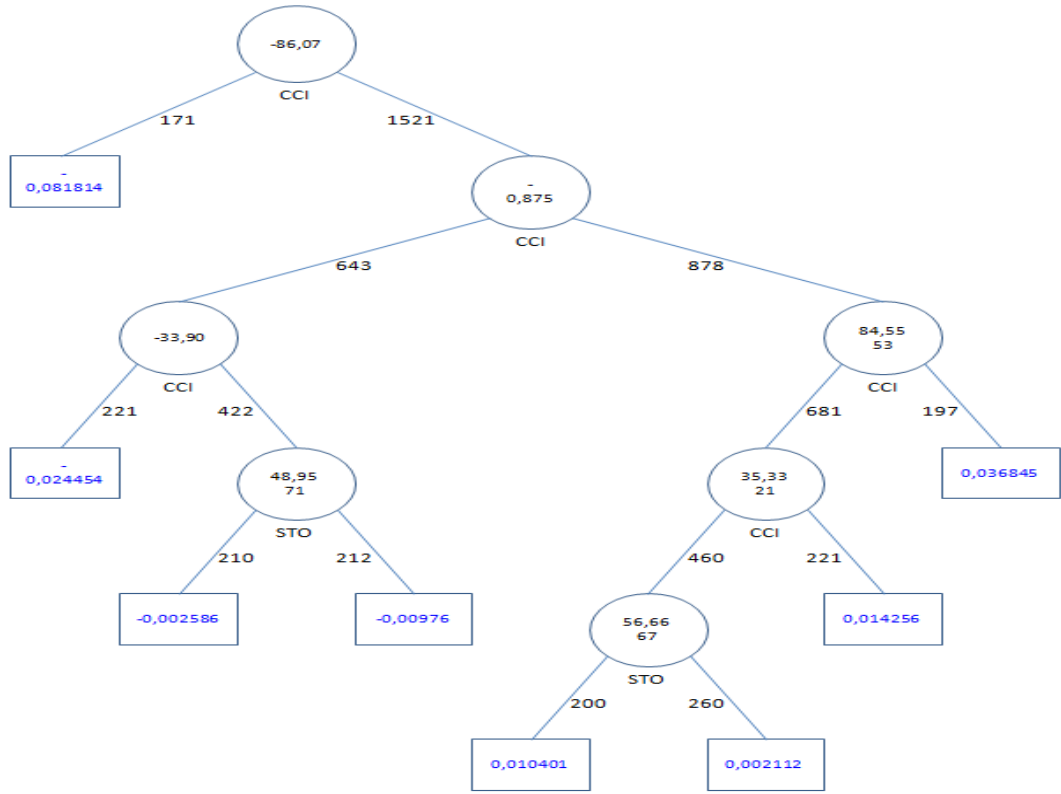
$$R^2 = 0,335987321 \quad S = 0,02476031 \\ \text{RMS Hata} = 0,02468415 \quad \text{Ortalama Hata} = 1,49024\text{E-}07$$

30. YKBNK İÇİN ÇDR MODELİ

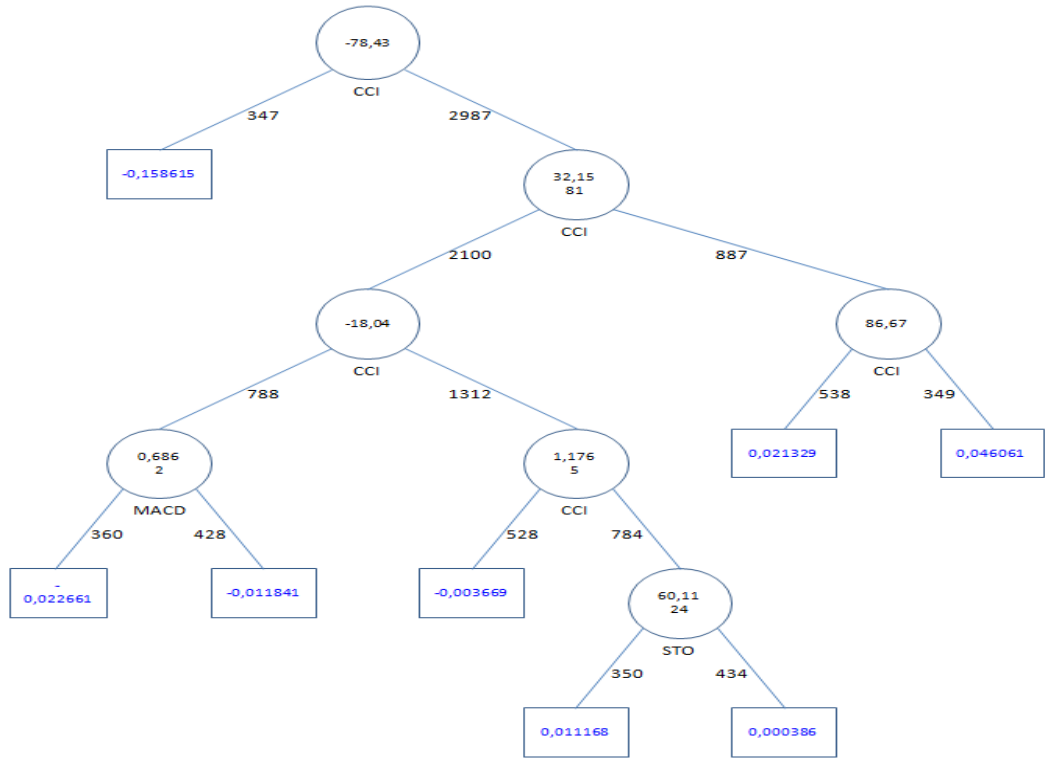
$$\text{KAP\%} = 0,15113258 - 0,15335406 \text{ MACD} - 0,00098618 \text{ MOM} + \\ (0,01875037) \quad (0,01376896) \quad (0,00017612) \\ 0,00037848 \text{ CCI} \\ (0,00001594)$$

$$R^2 = 0,1642349 \quad S = 0,26734865 \\ \text{RMS Hata} = 0,267188274 \quad \text{Ortalama Hata} = 3,26995\text{E-}07$$

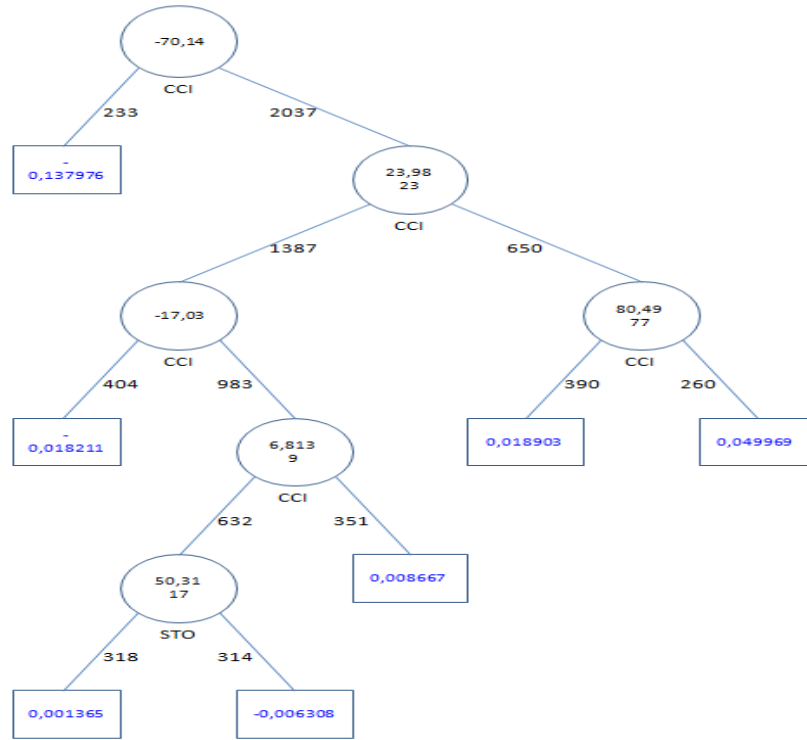
EK-2: İMKB 30 İÇİN REGRESYON AĞAÇLARI



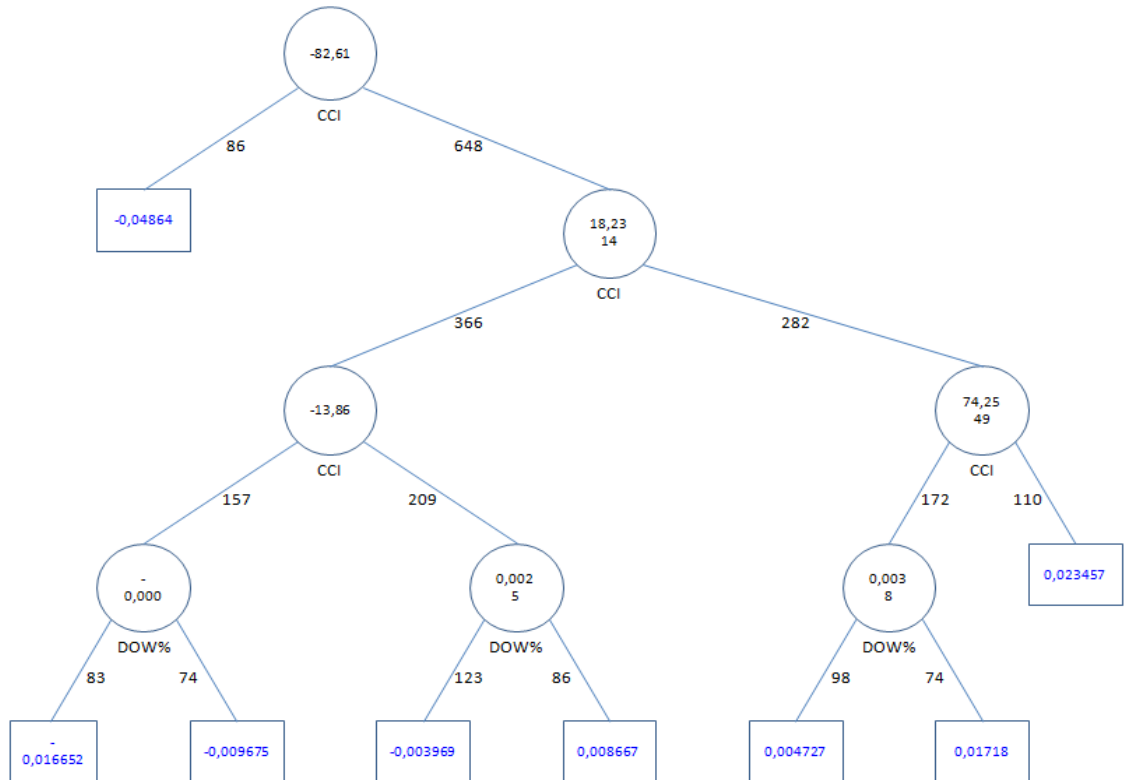
Şekil E.1. AEFES İçin Regresyon Ağacı.



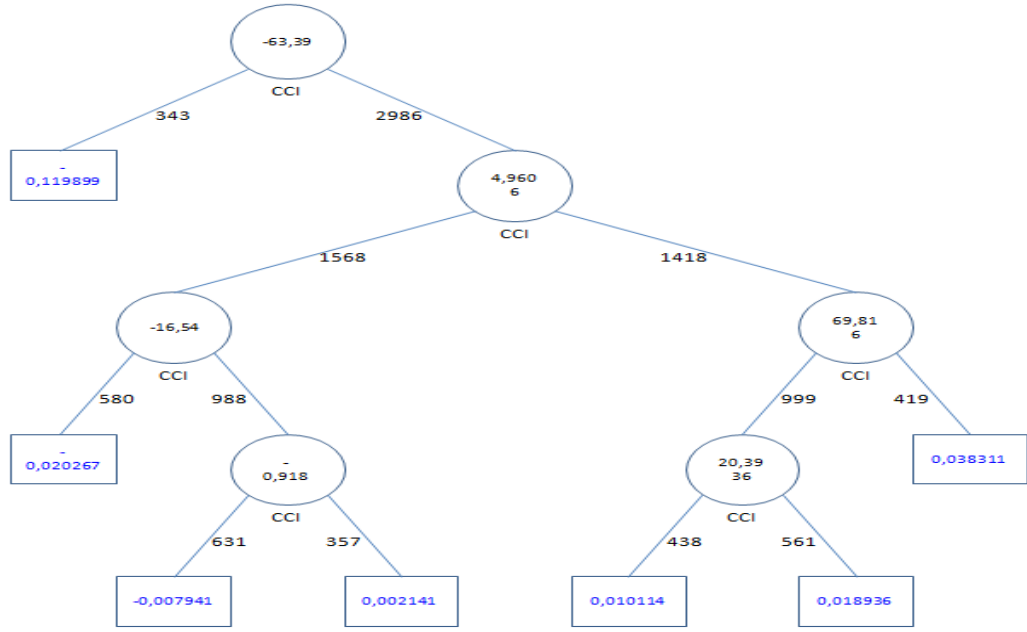
Şekil E.2. AKBNK İçin Regresyon Ağacı.



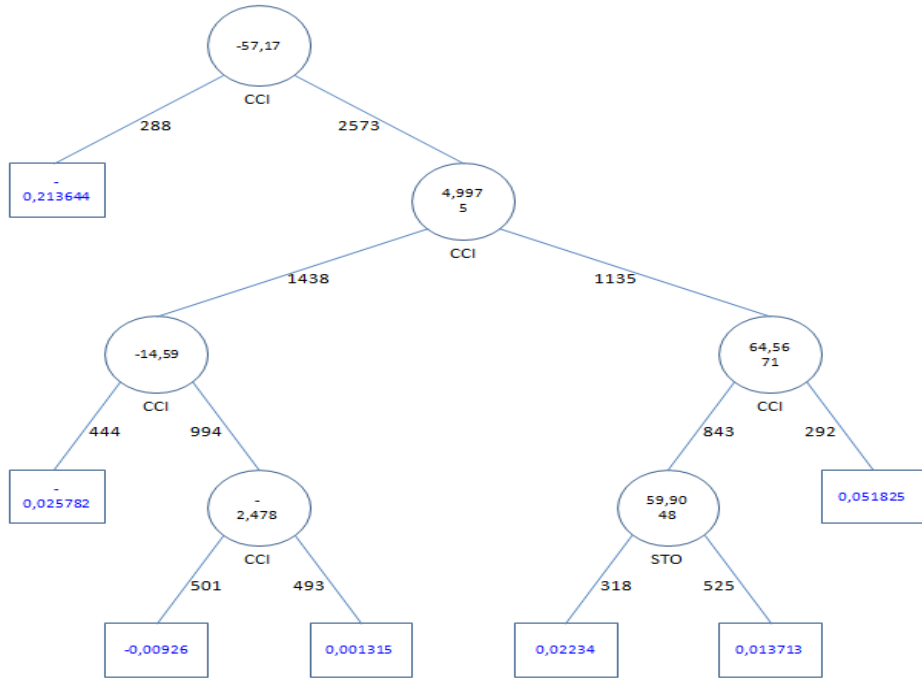
Şekil E.3. AKGRT İçin Regresyon Ağacı.



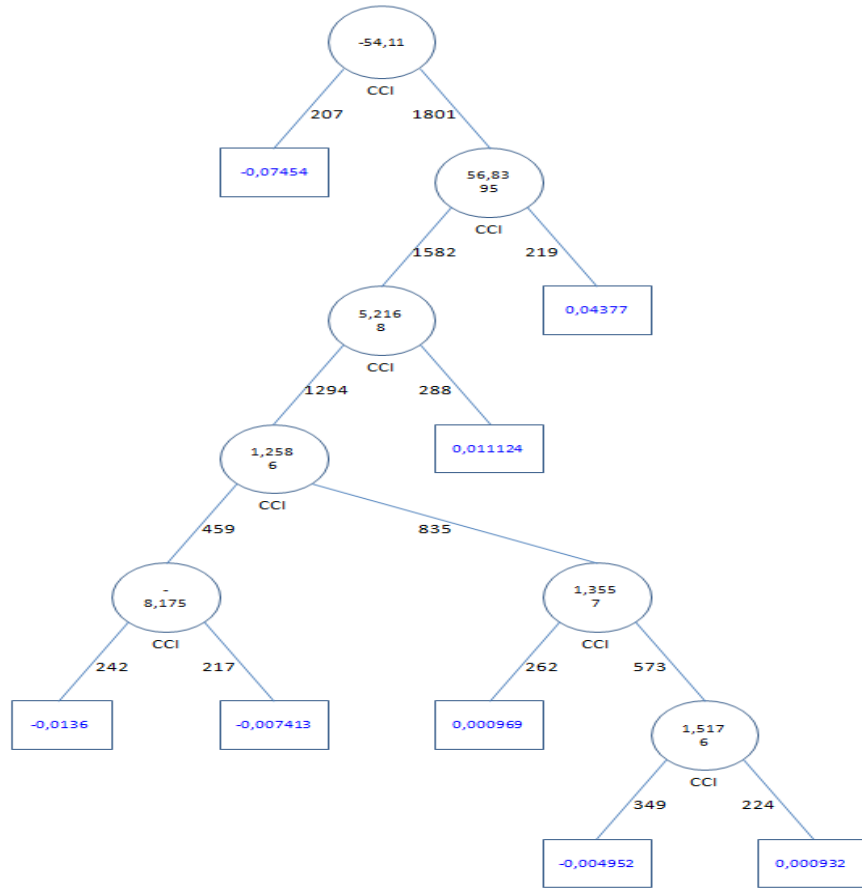
Şekil E.4. ASYAB İçin Regresyon Ağacı.



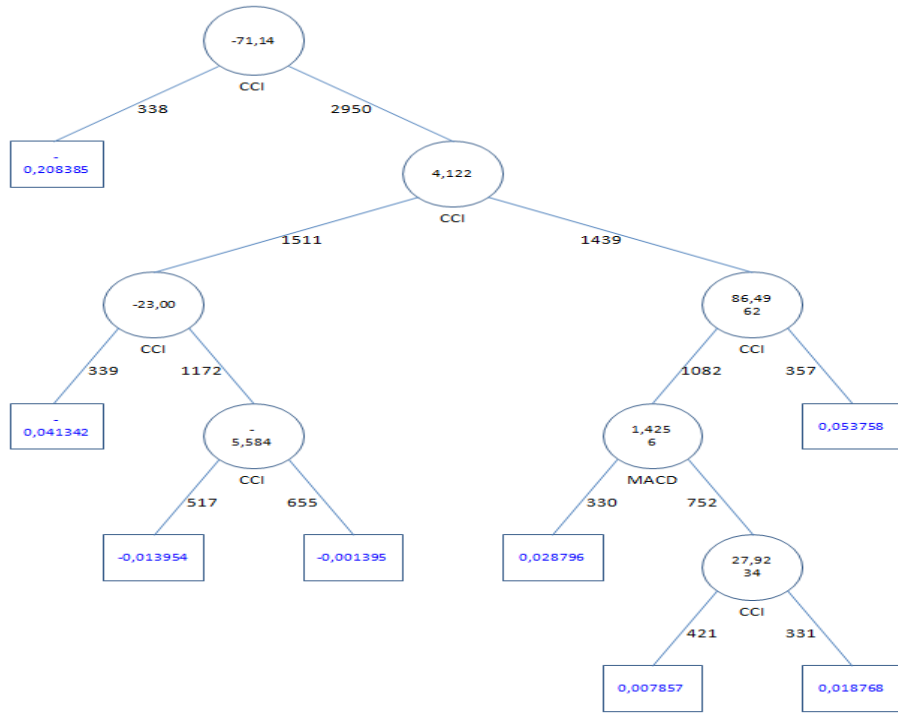
Şekil E.5. BAGFS İçin Regresyon Ağacı.



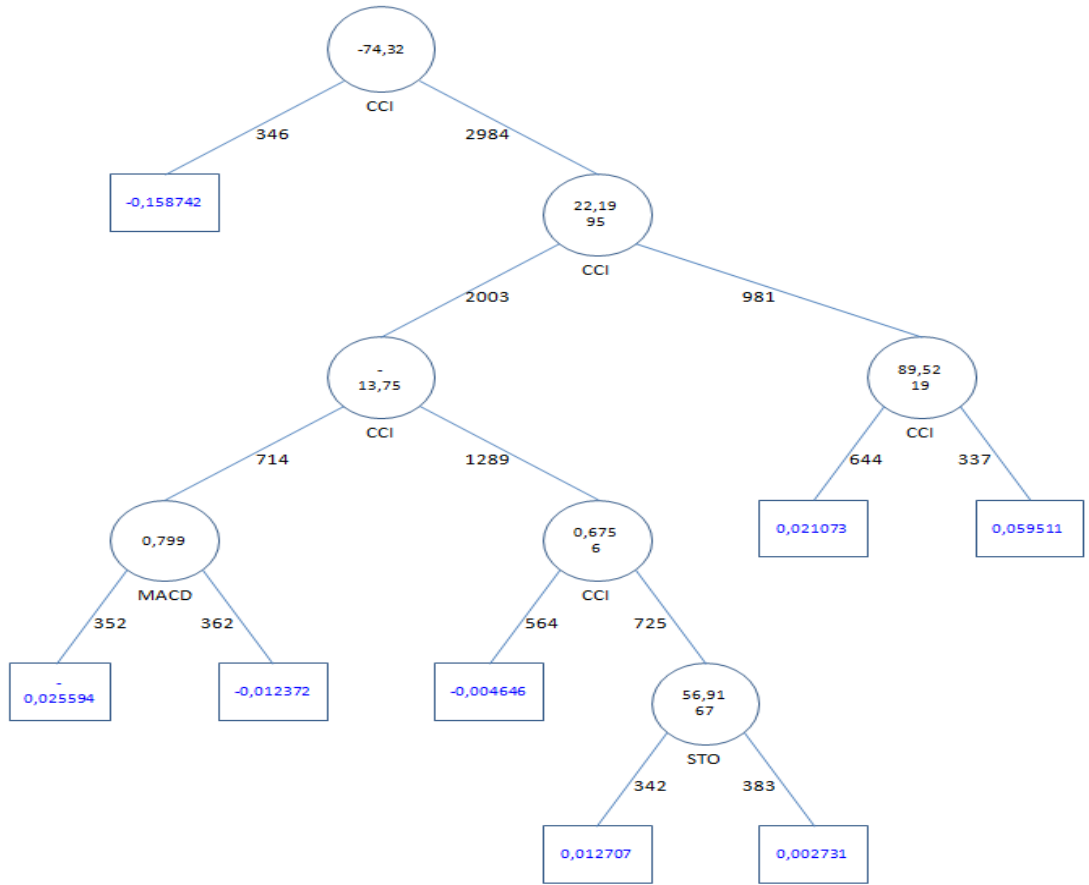
Şekil E.6. DOHOL İçin Regresyon Ağacı.



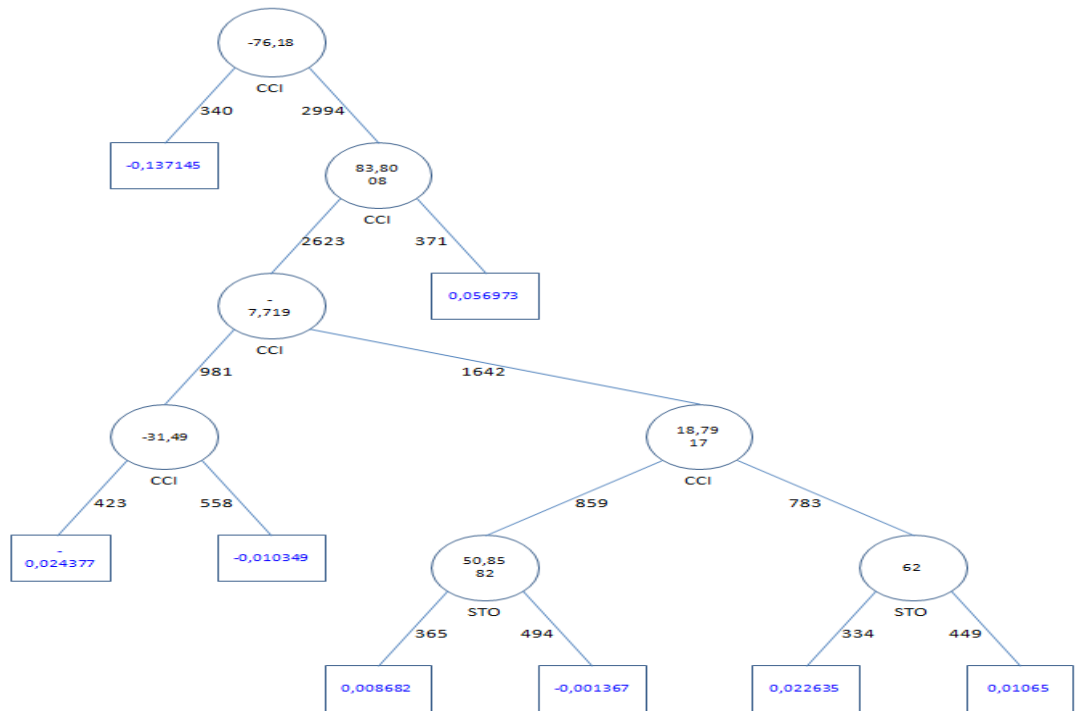
Şekil E.7. DYHOL İçin Regresyon Ağacı.



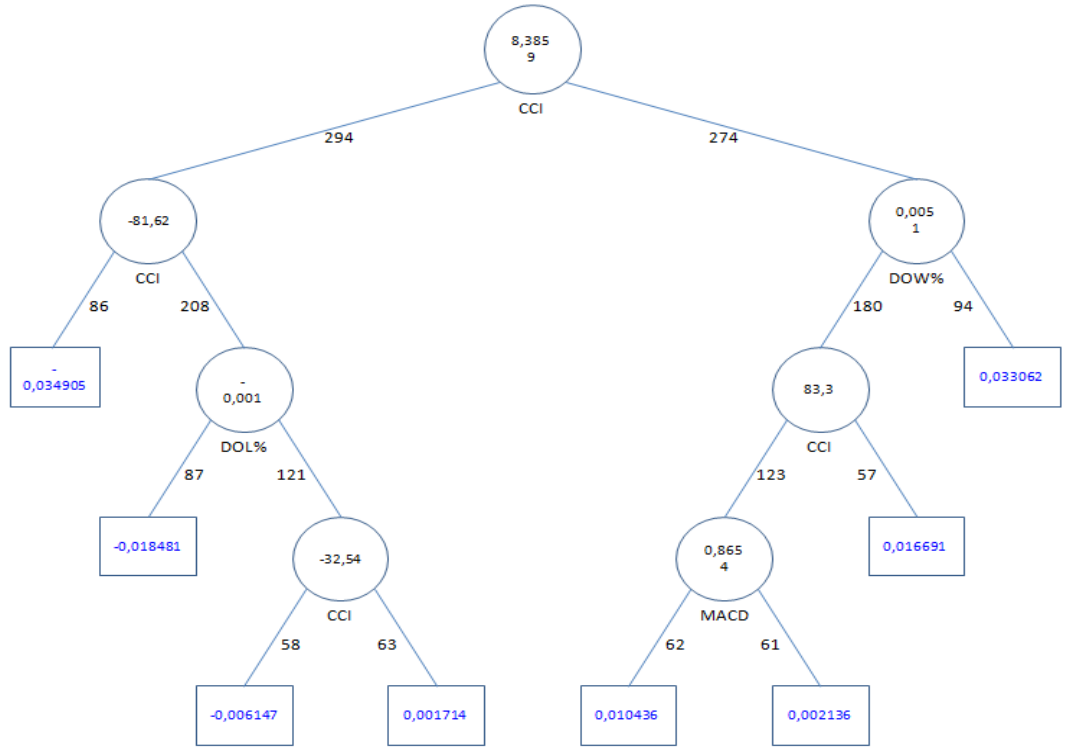
Şekil E.8. ENKA İçin Regresyon Ağacı.



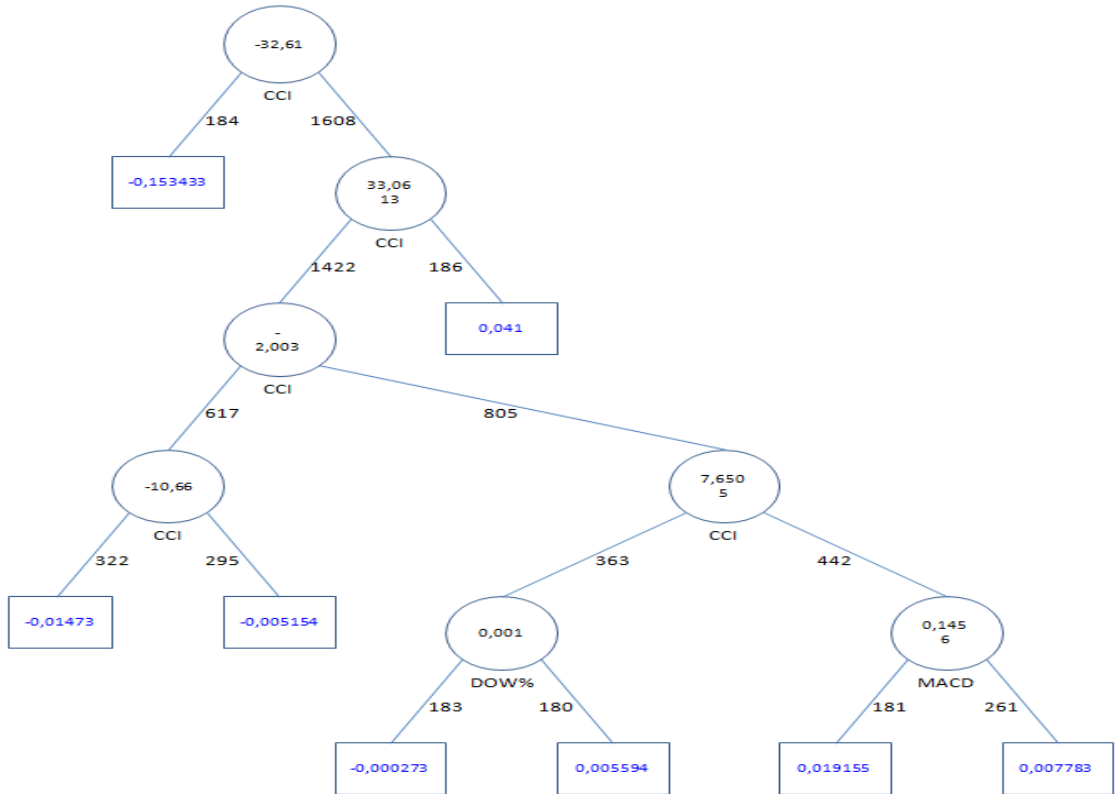
Şekil E.9. EREGL İçin Regresyon Ağacı.



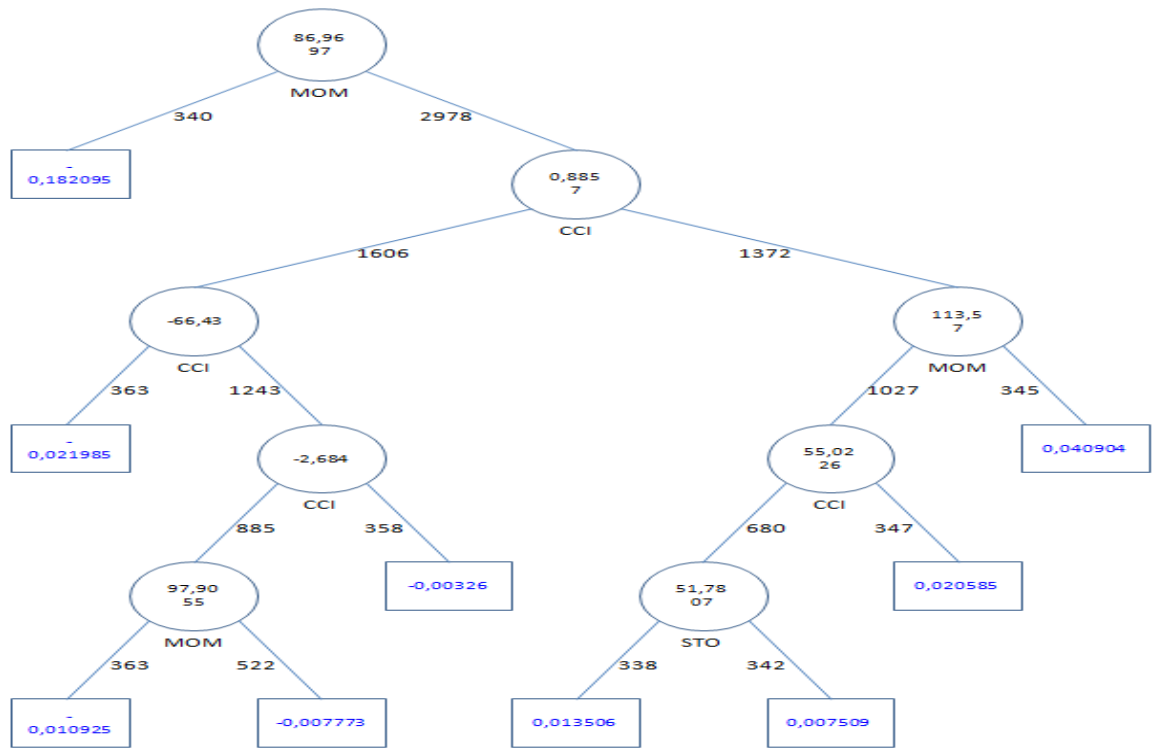
Şekil E.10. GARAN İçin Regresyon Ağacı.



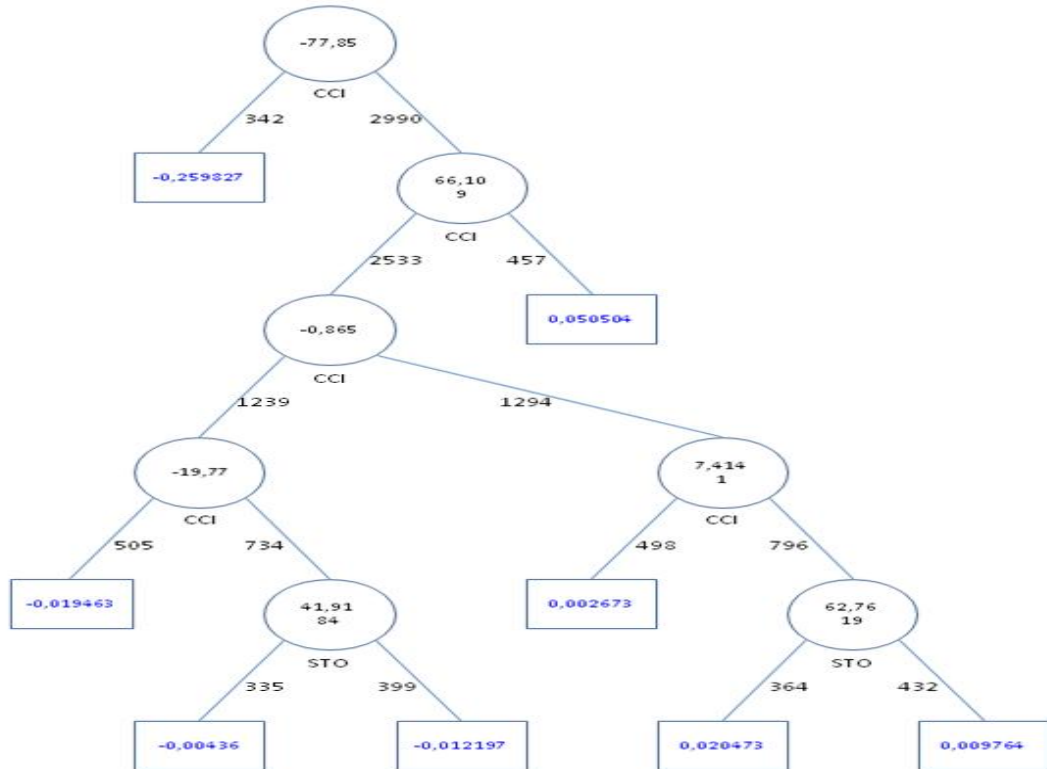
Şekil E.11. HALKB İçin Regresyon Ağacı.



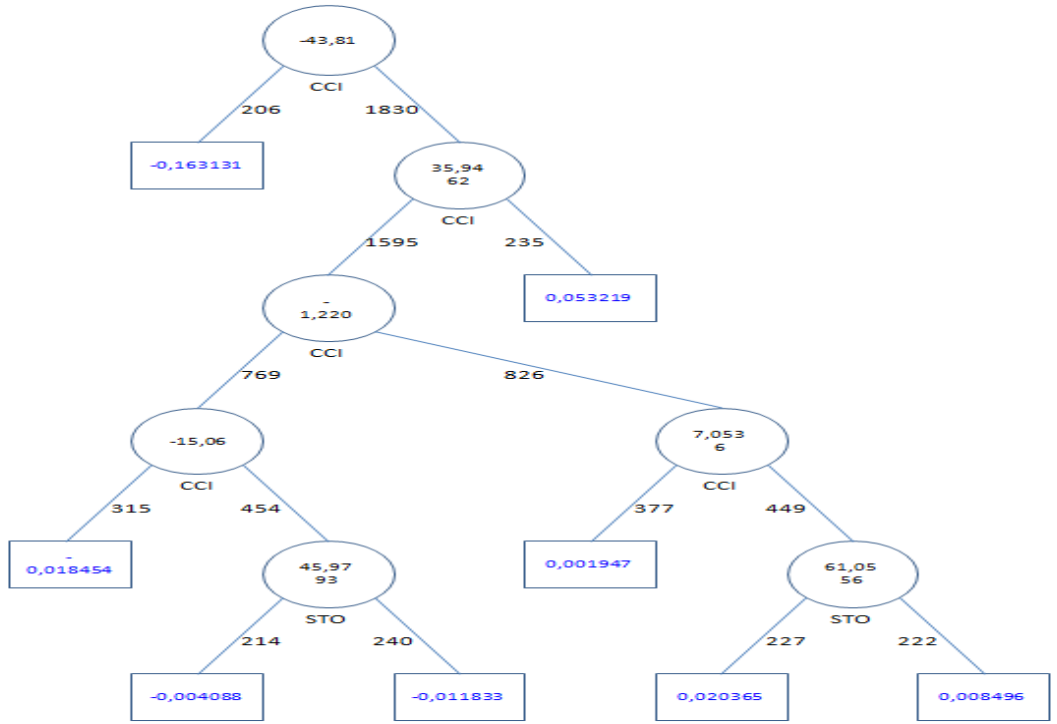
Şekil E.12. ISGYO İçin Regresyon Ağacı.



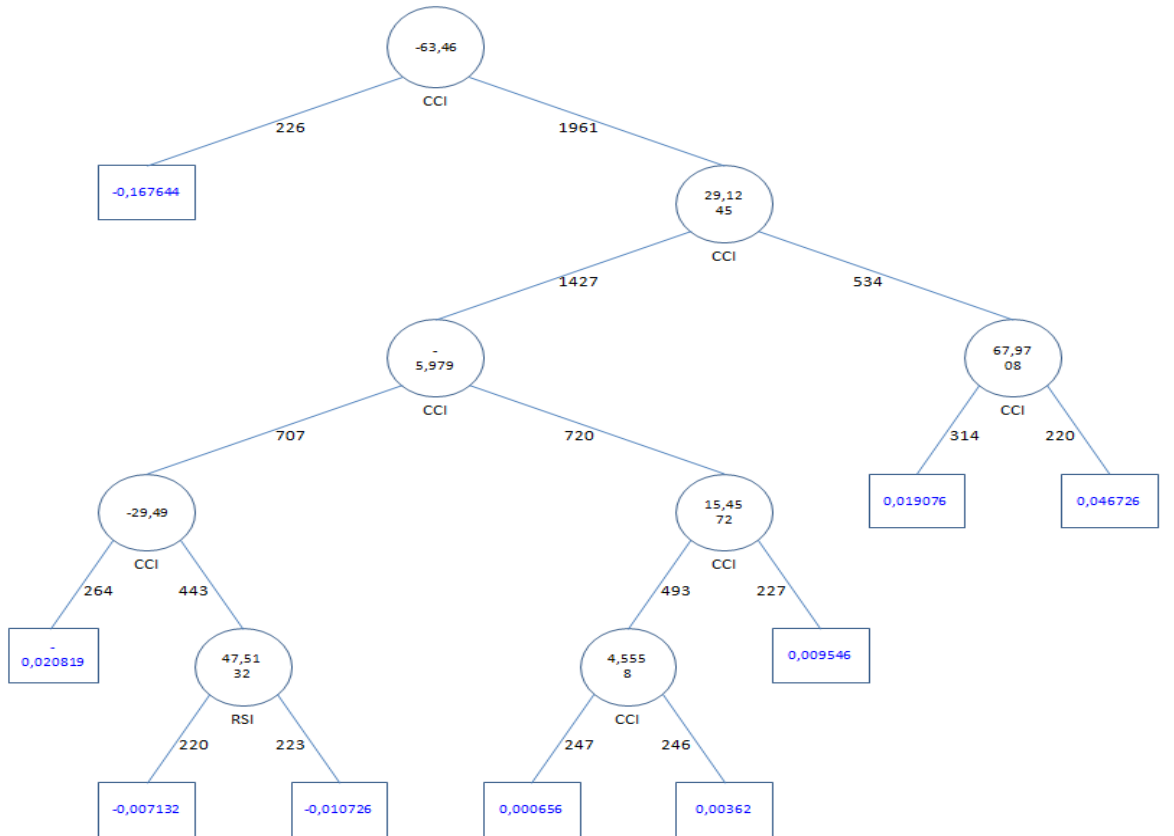
Şekil E.13. ISBTR İçin Regresyon Ağacı.



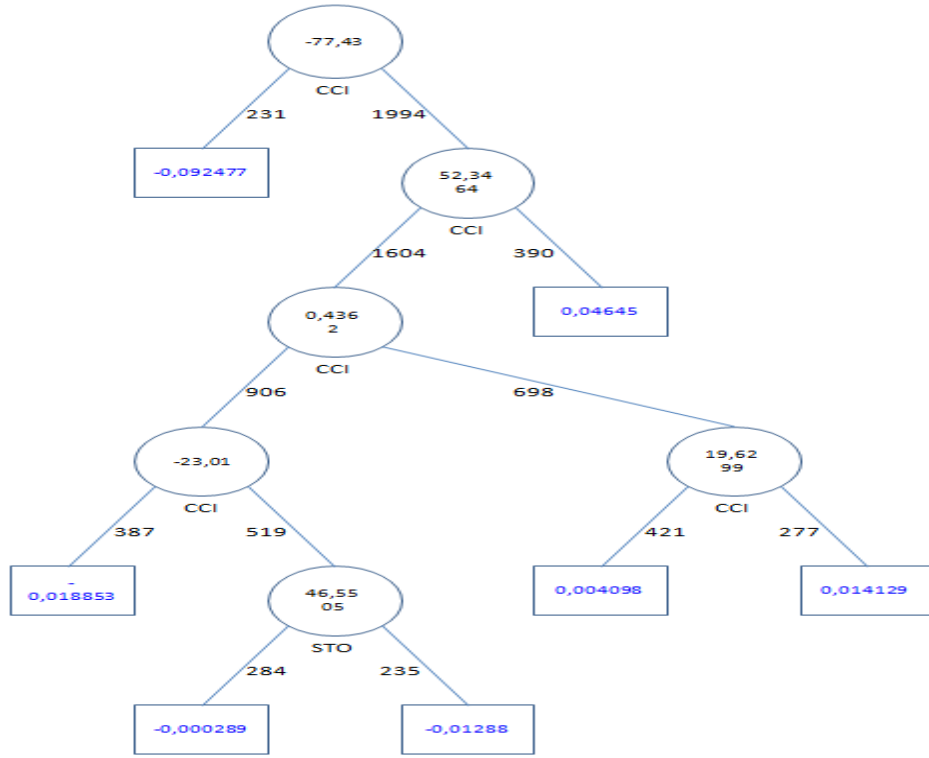
Şekil E.14. KCHOL İçin Regresyon Ağacı.



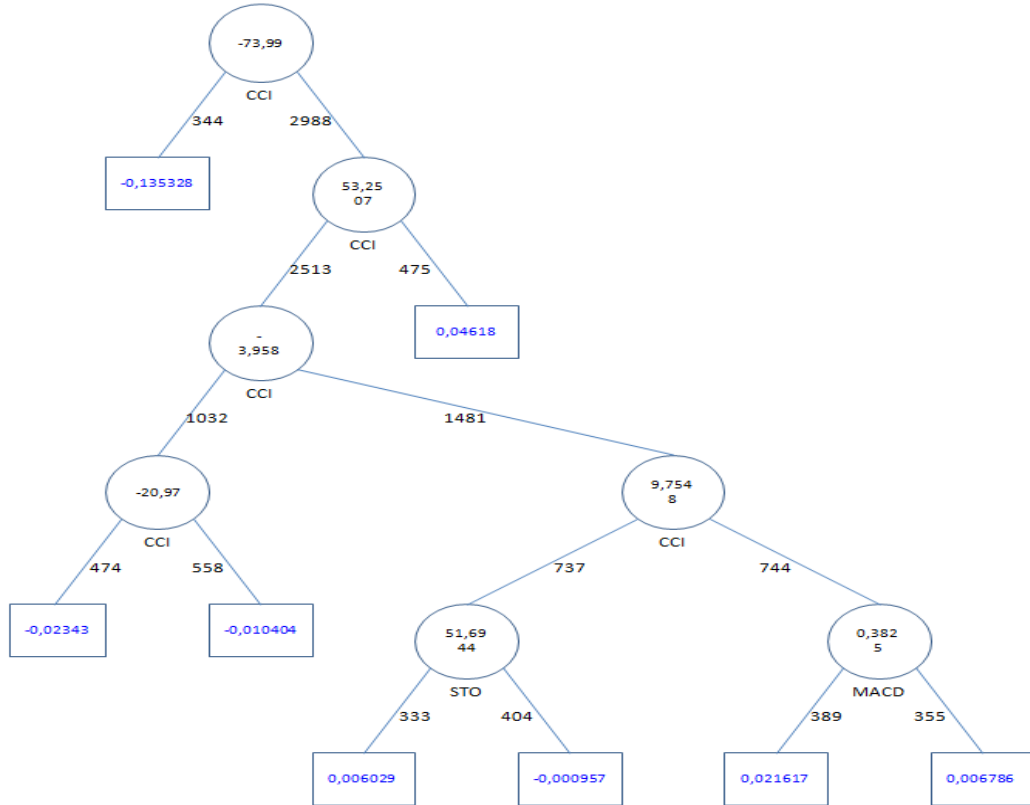
Şekil E.15. KRDMD İçin Regresyon Ağacı.



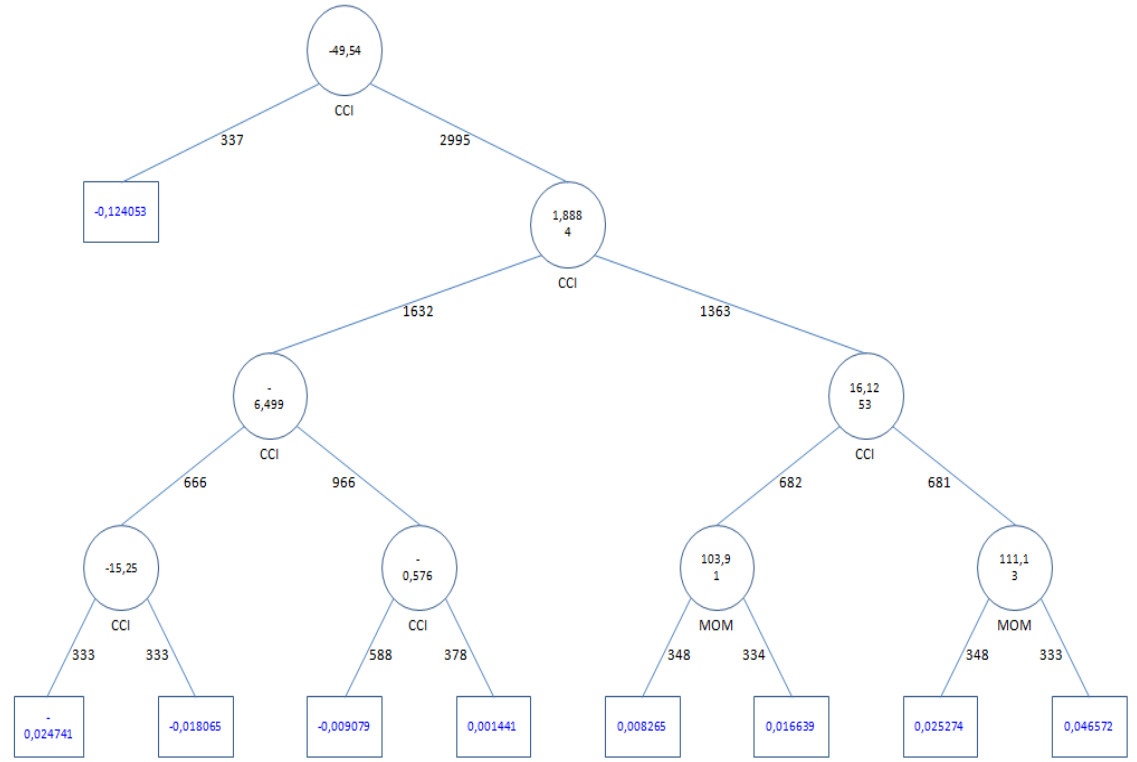
Şekil E.16. SAHOL İçin Regresyon Ağacı.



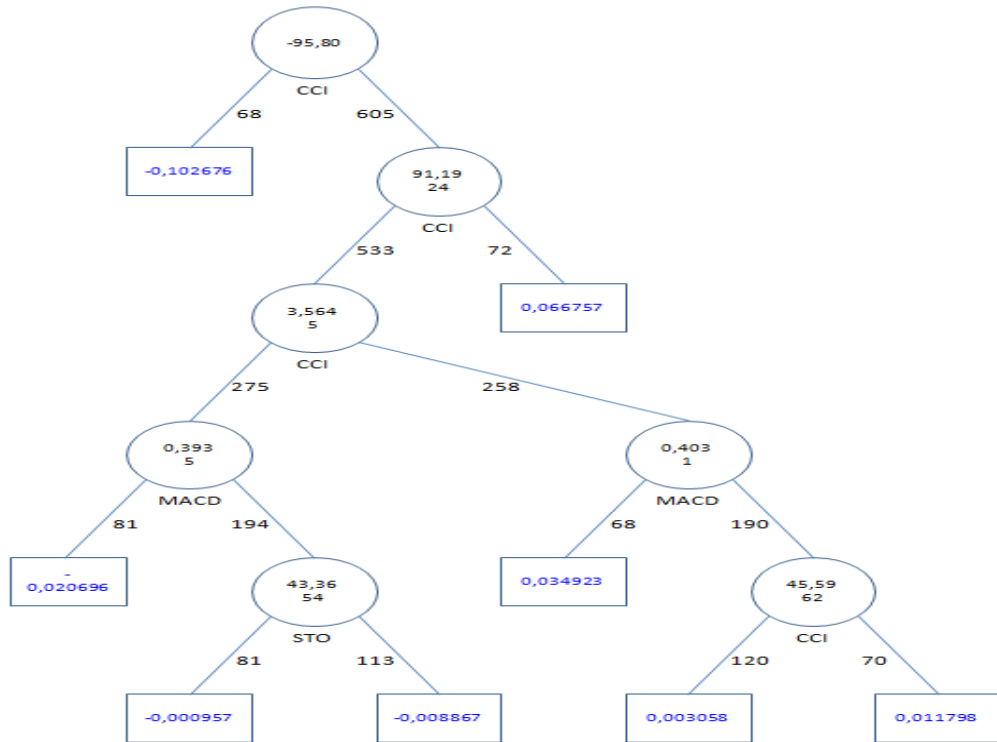
Şekil E.17. SKBNK İçin Regresyon Ağacı.



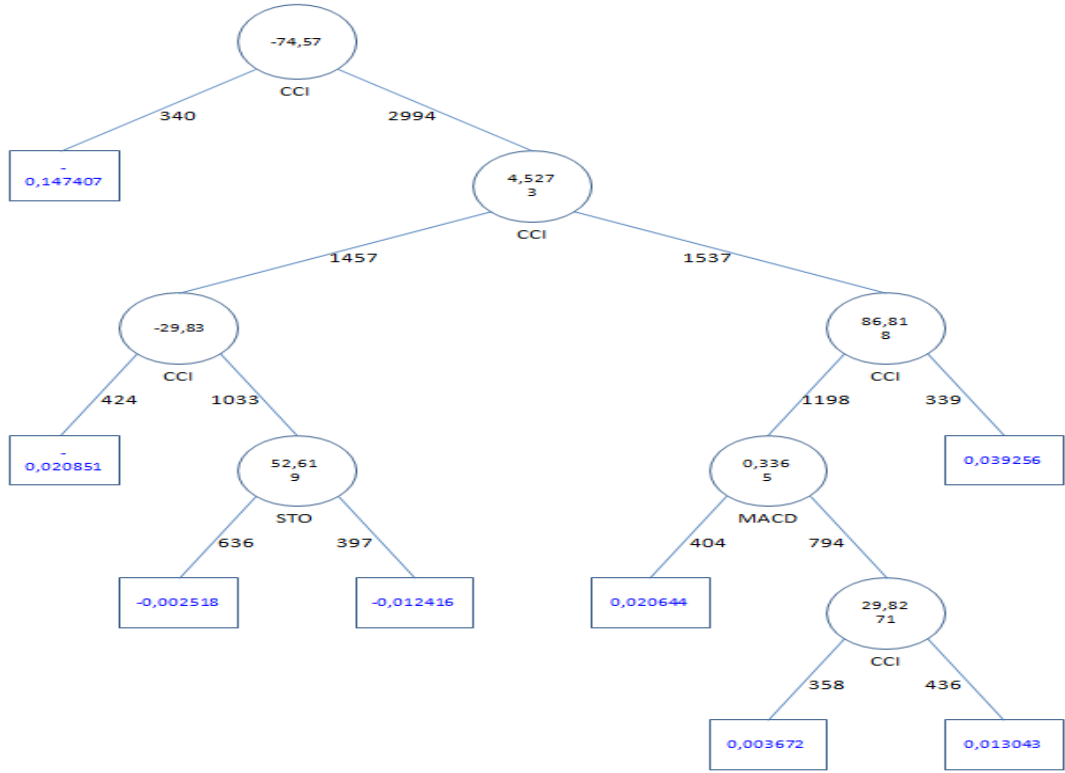
Şekil E.18. SISE İçin Regresyon Ağacı.



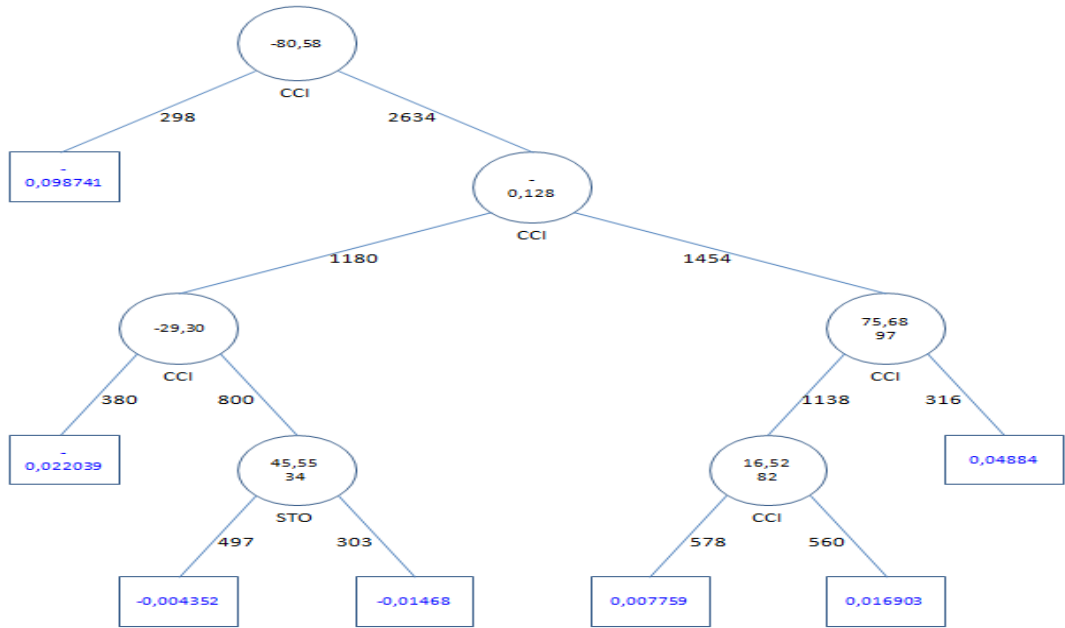
Şekil E.19. PETKM İçin Regresyon Ağacı.



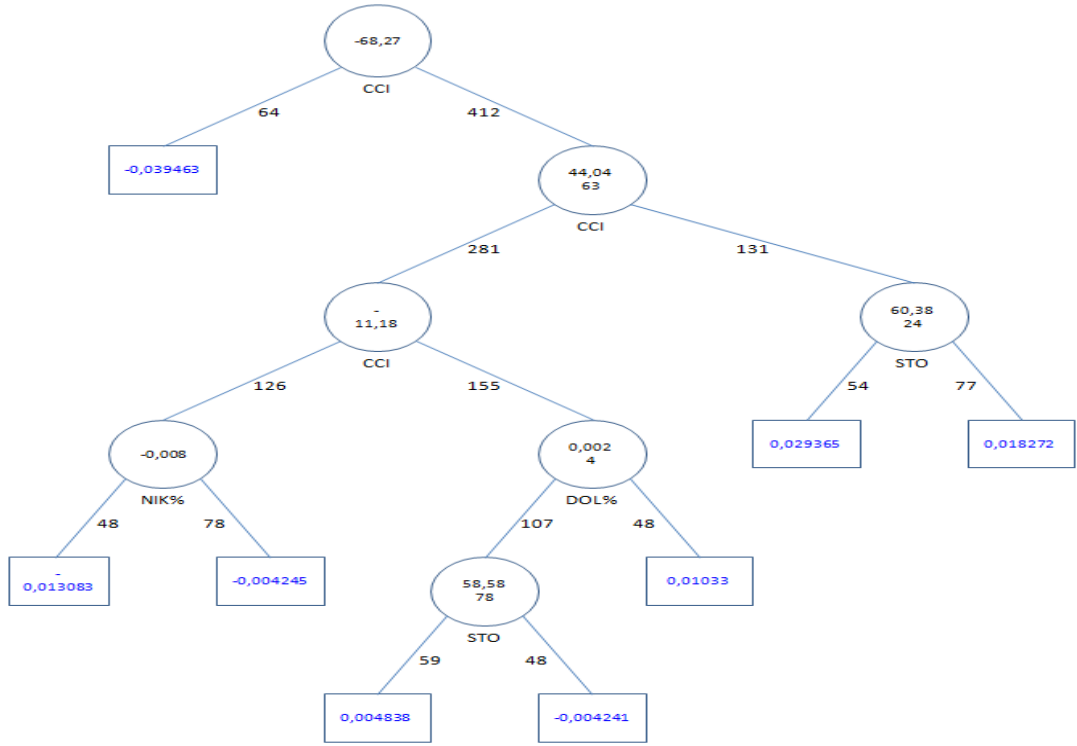
Şekil E.20. TAVHL İçin Regresyon Ağacı.



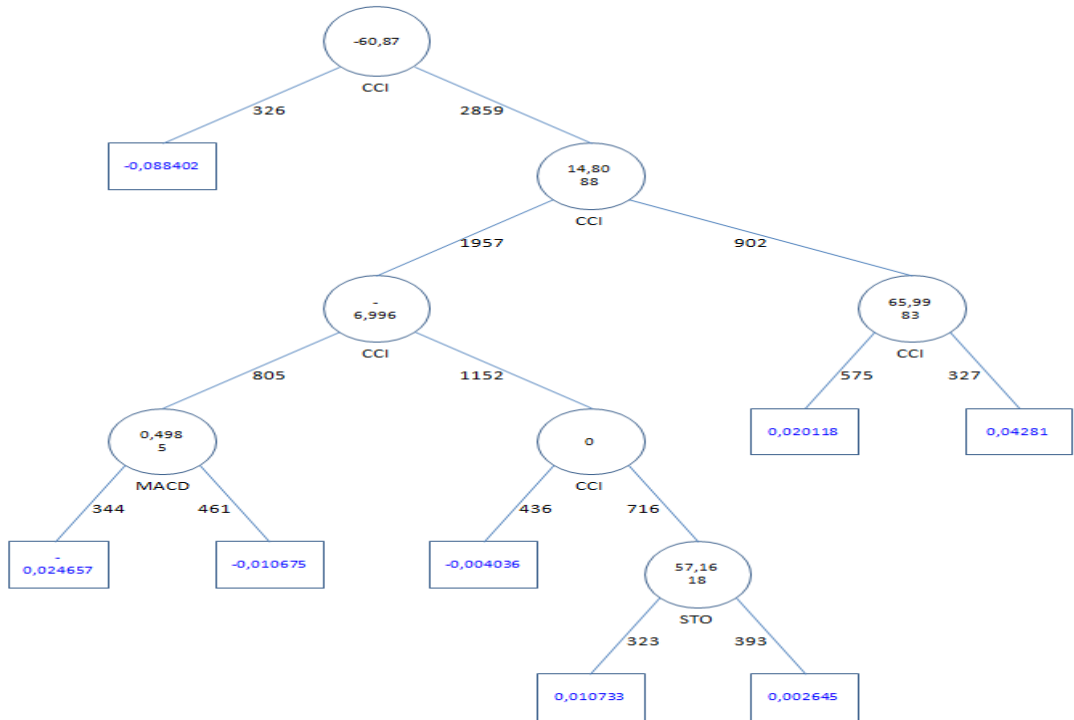
Şekil E.21. TEBNK İçin Regresyon Ağacı.



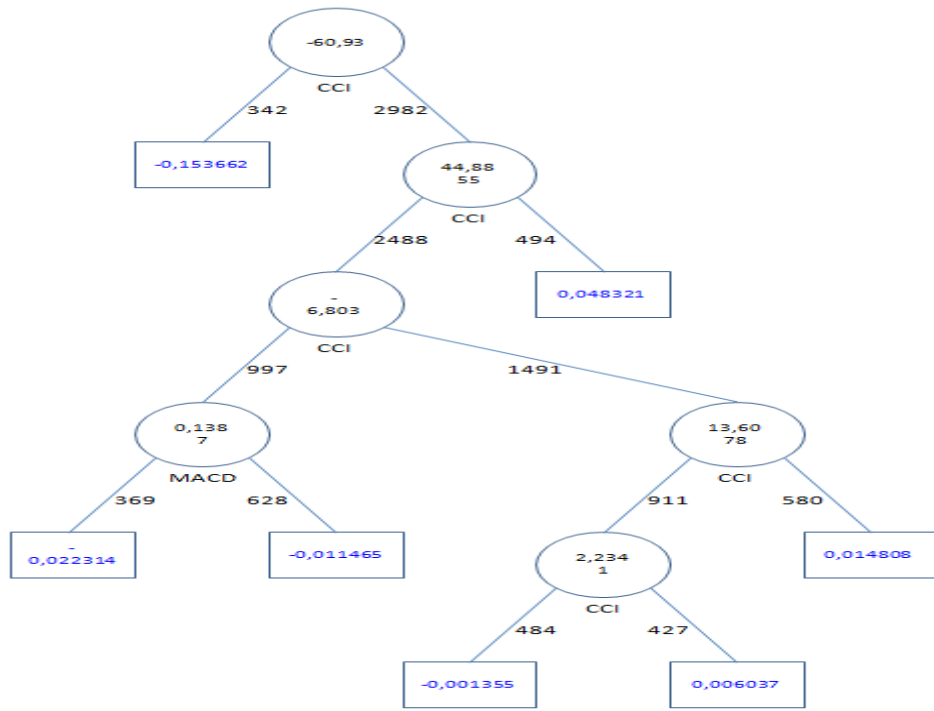
Şekil E.22. THY İçin Regresyon Ağacı.



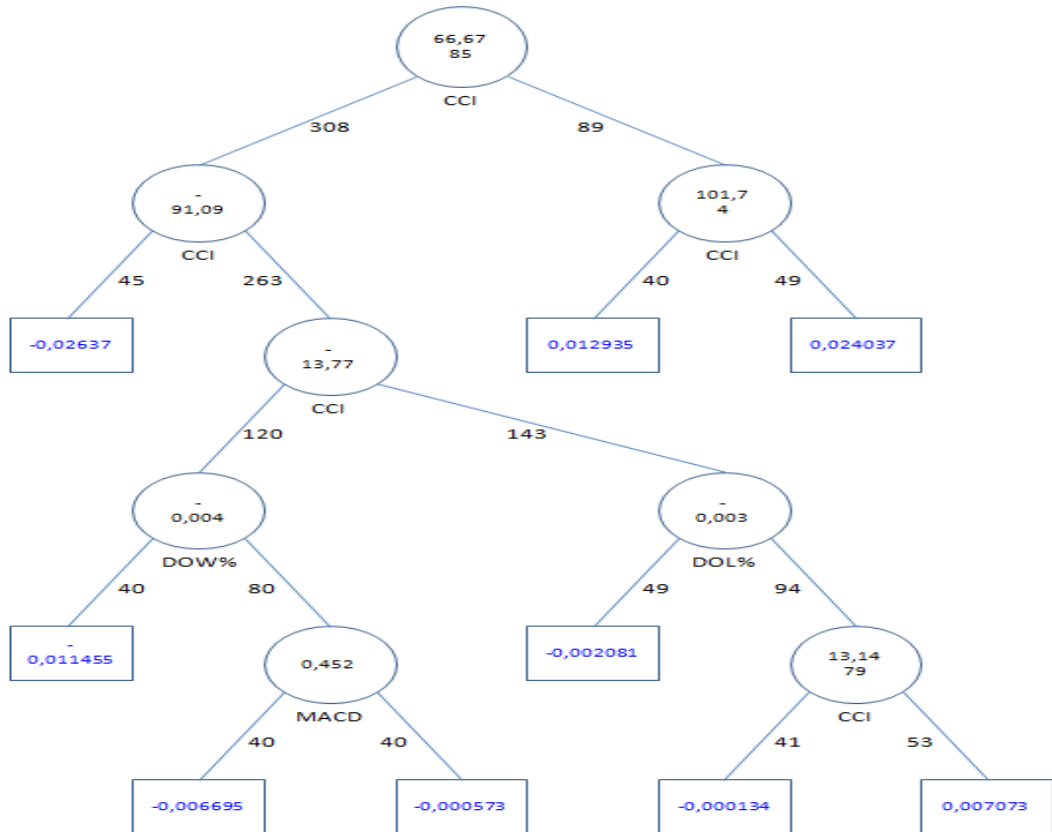
Şekil E.23. TKFEN İçin Regresyon Ağacı.



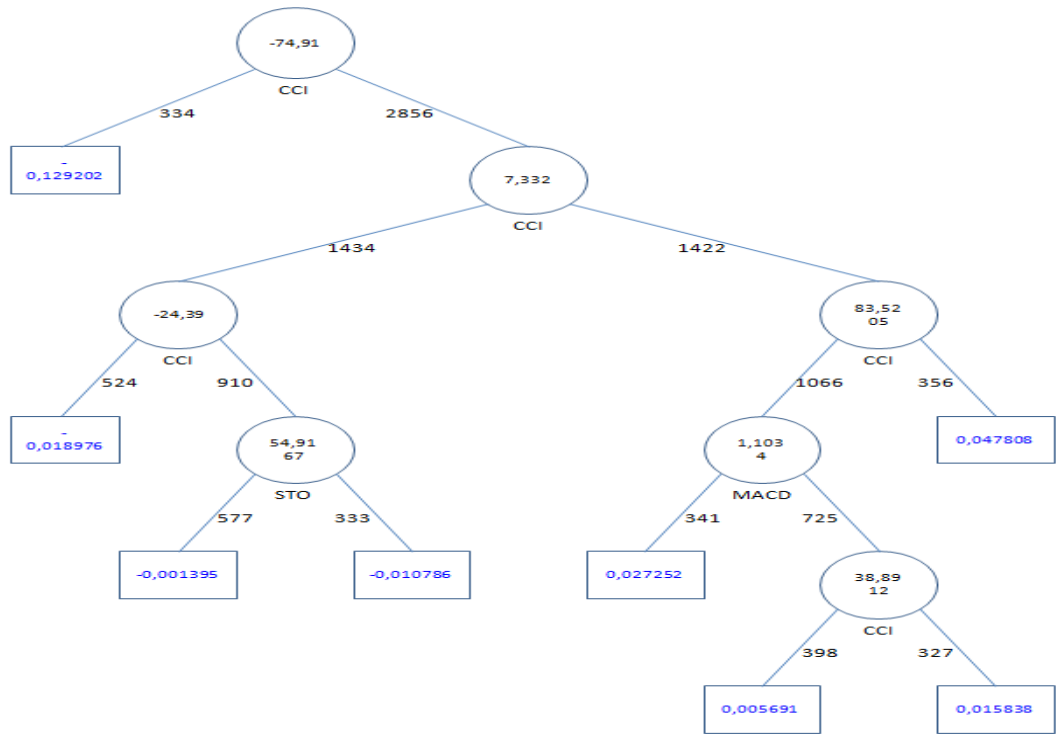
Şekil E.24. TOASO İçin Regresyon Ağacı.



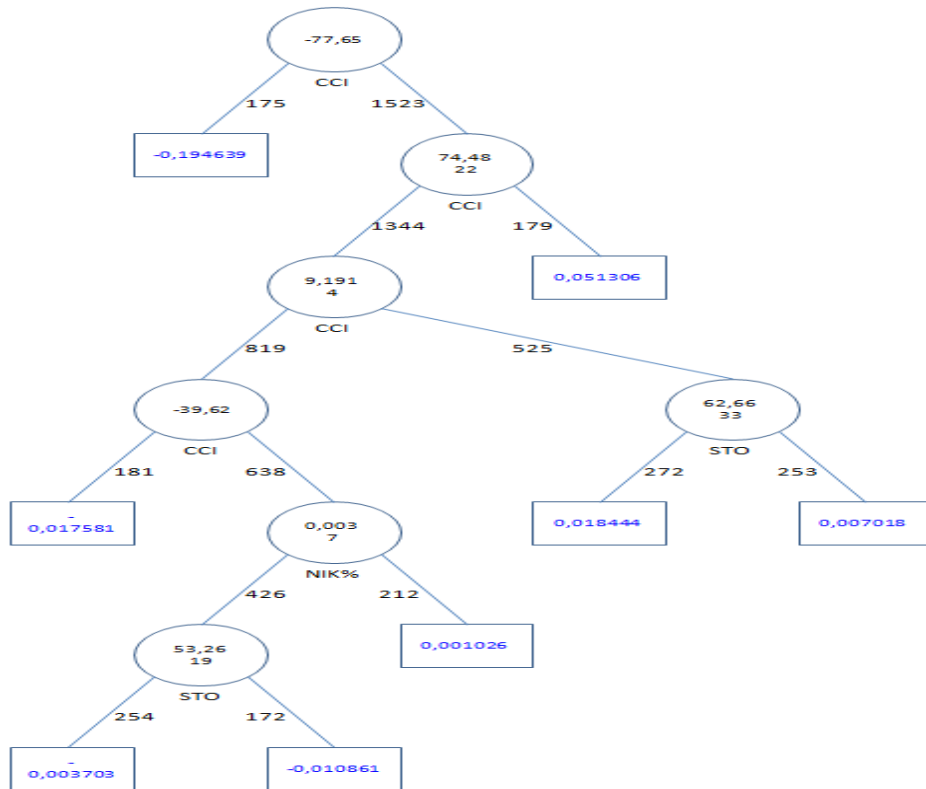
Şekil E.25. TSKB İçin Regresyon Ağacı.



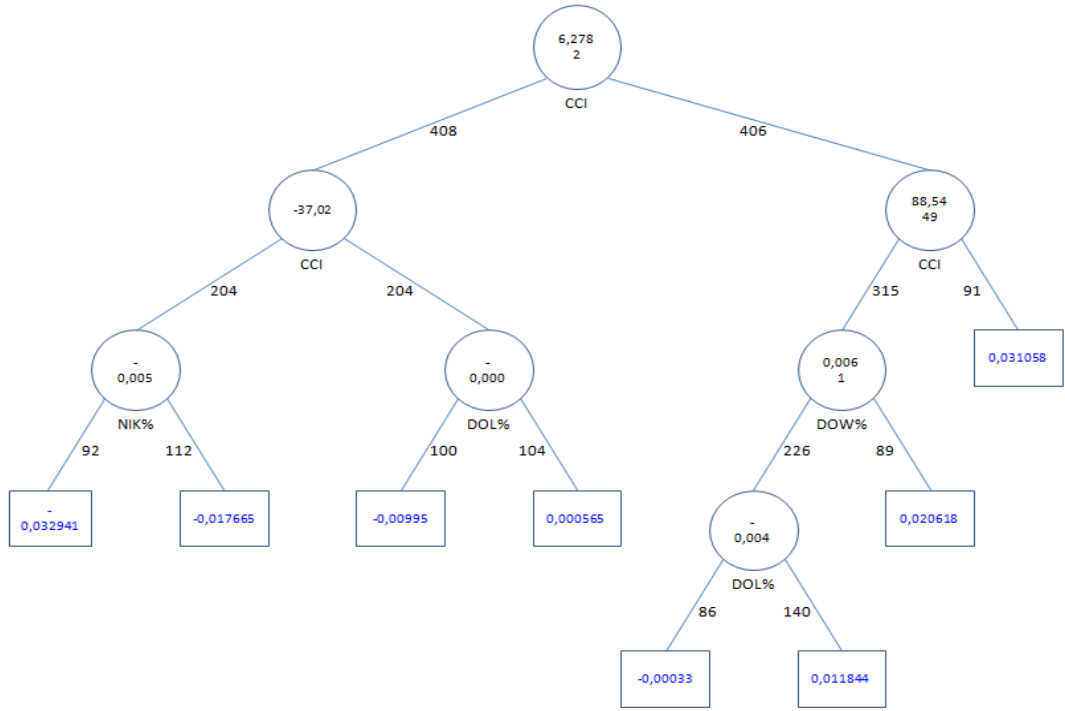
Şekil E.26. TTKOM İçin Regresyon Ağacı.



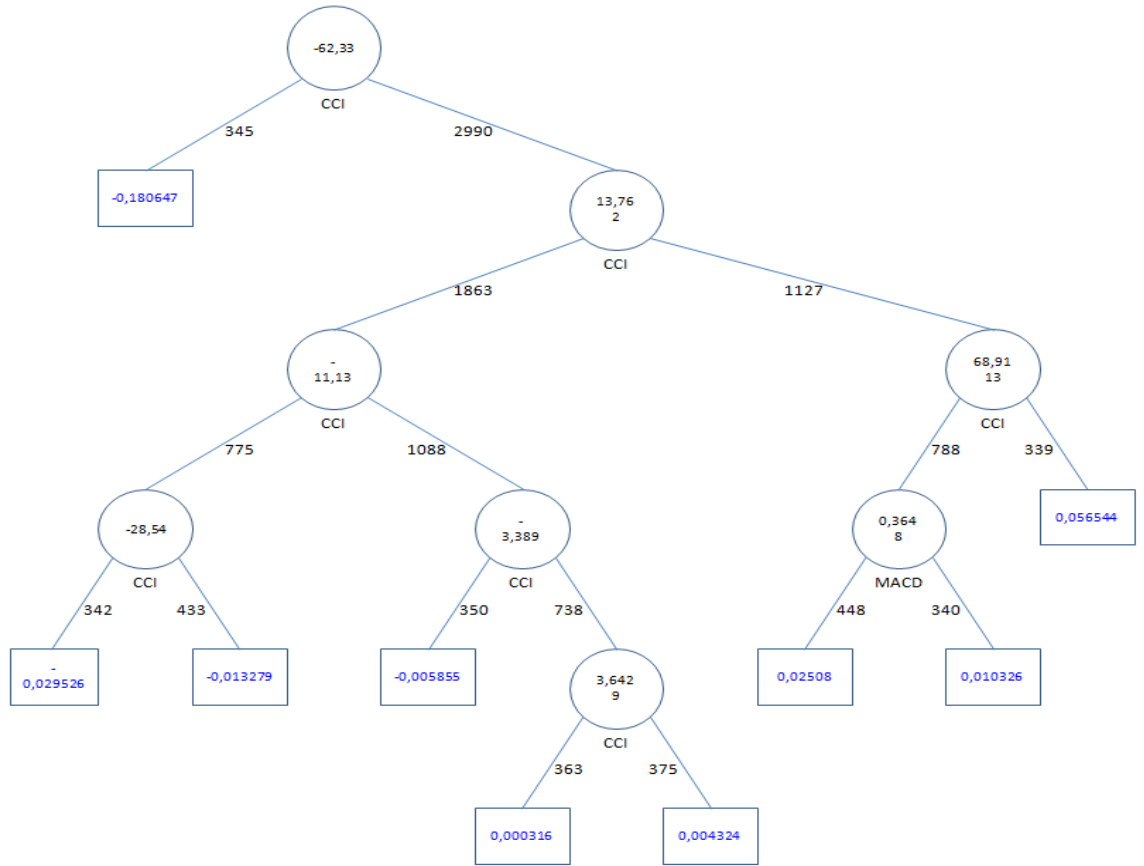
Şekil E.27. TUPRPS İçin Regresyon Ağacı.



Şekil E.28. TCELL İçin Regresyon Ağacı.



Şekil E.29. VKBNK İçin Regresyon Ağacı.



Şekil E.30. YKBNK İçin Regresyon Ağacı.

EK 3: İMKB 30 İÇİN KARAR KURALLARI

1. AEFES İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -86,07$ İSE
 $KAP\% = 0,081834$

KURAL2: EĞER $CCI > -86,07$ İSE VE
 $CCI \leq -0,875$ İSE VE
 $CCI \leq -33,90$ İSE
 $KAP\% = 0,024454$

KURAL3: EĞER $CCI > -86,07$ İSE VE
 $CCI \leq -0,875$ İSE VE
 $CCI > -33,90$ İSE VE
 $STO \leq 48,9571$ İSE
 $KAP\% = -0,002586$

KURAL4: EĞER $CCI > -86,07$ İSE VE
 $CCI \leq -0,875$ İSE VE
 $CCI > -33,90$ İSE VE
 $STO > -48,9571$ İSE
 $KAP\% = -0,00976$

KURAL5: EĞER $CCI > -86,07$ İSE VE
 $CCI > -0,875$ İSE VE
 $CCI \leq 84,55$ İSE VE
 $CCI \leq 35,3321$ İSE VE
 $STO \leq 56,6667$ İSE
 $KAP\% = 0,010401$

KURAL6: EĞER $CCI > -86,07$ İSE VE
 $CCI > -0,875$ İSE VE
 $CCI \leq 84,55$ İSE VE
 $CCI \leq 35,3321$ İSE VE
 $STO > 56,66,67$ İSE
 $KAP\% = 0,002112$

KURAL7: EĞER $CCI > -86,07$ İSE VE
 $CCI > -0,875$ İSE VE
 $CCI \leq 84,55$ İSE VE
 $CCI > 35,3321$ İSE
 $KAP\% = 0,014256$

KURAL8: EĞER $CCI > -86,07$ İSE VE
 $CCI > -0,875$ İSE VE
 $CCI > 84,55$ İSE
 $KAP\% = 0,036845$

2. AKBNK İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -78,43$ İSE
 $KAP\% = -0,158615$

KURAL2: EĞER $CCI > -78,43$ İSE
 $CCI \leq 32,1581$ İSE VE
 $CCI \leq -18,04$ İSE VE
 $MACD \leq 0,6862$ İSE
 $KAP\% = -0,022661$

KURAL3: EĞER $CCI > -78,43$ İSE
 $CCI \leq 32,1581$ İSE VE
 $CCI \leq -18,04$ İSE VE
 $MACD > 0,6862$ İSE
 $KAP\% = -0,011841$

KURAL4: EĞER $CCI > -78,43$ İSE
 $CCI \leq 32,1581$ İSE VE
 $CCI > -18,04$ İSE VE
 $CCI \leq 1,1765$ İSE
 $KAP\% = -0,003669$

KURAL5: EĞER $CCI > -78,43$ İSE
 $CCI \leq 32,1581$ İSE VE
 $CCI > -18,04$ İSE VE
 $CCI > 1,1765$ İSE
 $STO \leq 60,1124$ İSE
 $KAP\% = 0,011168$

KURAL6: EĞER $CCI > -78,43$ İSE
 $CCI \leq 32,1581$ İSE VE
 $CCI > -18,04$ İSE VE
 $CCI > 1,1765$ İSE
 $STO > 60,1124$ İSE
 $KAP\% = 0,000386$

KURAL7: EĞER $CCI > -78,43$ İSE
 $CCI > 32,1581$ İSE VE
 $CCI \leq 86,67$ İSE
 $KAP\% = 0,021329$

KURAL8: EĞER $CCI > -78,43$ İSE
 $CCI > 32,1581$ İSE VE
 $CCI > 86,67$ İSE
 $KAP\% = 0,046061$

3. AKGRT İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -70,14$ İSE
KAP%= -0,137976

KURAL2: EĞER $CCI > -70,14$ İSE VE
 $CCI \leq 23,9823$ İSE VE
 $CCI \leq -17,03$ İSE
KAP%= -0,018211

KURAL3: EĞER $CCI > -70,14$ İSE VE
 $CCI \leq 23,9823$ İSE VE
 $CCI > -17,03$ İSE VE
 $CCI \leq 6,8139$ İSE VE
 $STO \leq 50,3117$ İSE
KAP%= 0,001365

KURAL4: EĞER $CCI > -70,14$ İSE VE
 $CCI \leq 23,9823$ İSE VE
 $CCI > -17,03$ İSE VE
 $CCI \leq 6,8139$ İSE VE
 $STO > 50,317$ İSE
KAP%= -0,006308

KURAL5: EĞER $CCI > -70,14$ İSE VE
 $CCI \leq 23,9823$ İSE VE
 $CCI > -17,03$ İSE VE
 $CCI > 6,8139$ İSE
KAP%= 0,008667

KURAL6: EĞER $CCI > -70,14$ İSE VE
 $CCI > 23,9823$ İSE VE
 $CCI \leq 80,4977$ İSE
KAP%= 0,018903

KURAL7: EĞER $CCI > -70,14$ İSE VE
 $CCI > 23,9823$ İSE VE
 $CCI > 80,4977$ İSE
KAP%= 0,049969

4. ASYAB İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -82,61$ İSE
KAP%= 0,04864

KURAL2: EĞER $CCI > -82,61$ İSE
 $CCI \leq 18,2314$ İSE VE
 $CCI \leq -13,86$ İSE VE
 $DOW\% \leq -0,000$ İSE
KAP%= 0,0166523

KURAL3: EĞER $CCI > -82,61$ İSE
 $CCI \leq 18,2314$ İSE VE
 $CCI \leq -13,86$ İSE VE
 $DOW\% > -0,000$ İSE
KAP%= -0,009675

KURAL4: EĞER $CCI > -82,61$ İSE
 $CCI \leq 18,2314$ İSE VE
 $CCI > -13,86$ İSE VE
 $DOW\% \leq 0,0025$ İSE
KAP%= -0,003969

KURAL5: EĞER $CCI > -82,61$ İSE
 $CCI \leq 18,2314$ İSE VE
 $CCI > -13,86$ İSE VE
 $DOW\% > 0,0025$ İSE
KAP%= 0,008667

KURAL6: EĞER $CCI > -82,61$ İSE
 $CCI > 18,2314$ İSE VE
 $CCI \leq 74,2549$ İSE VE
 $DOW\% \leq 0,0038$ İSE
KAP%= 0,004727

KURAL7: EĞER $CCI > -82,61$ İSE
 $CCI > 18,2314$ İSE VE
 $CCI \leq 74,2549$ İSE VE
 $DOW\% > 0,0038$ İSE
KAP%= 0,01718

KURAL8: EĞER $CCI > -82,61$ İSE
 $CCI > 18,2314$ İSE VE
 $CCI > 74,2549$ İSE
KAP%= 0,023457

5. BAGFS İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -63,39$ İSE
KAP% = -0,119899

KURAL2: EĞER $CCI > -63,39$ İSE
 $CCI \leq 4,9606$ İSE VE
 $CCI \leq -16,54$ İSE
KAP% = -0,020267

KURAL3: EĞER $CCI > -63,39$ İSE
 $CCI \leq 4,9606$ İSE VE
 $CCI > -16,54$ İSE VE
 $CCI \leq -0,918$ İSE
KAP% = -0,007941

KURAL4: EĞER $CCI > -63,39$ İSE
 $CCI \leq 4,9606$ İSE VE
 $CCI > -16,54$ İSE VE
 $CCI > -0,918$ İSE
KAP% = 0,002141

KURAL5: EĞER $CCI > -63,39$ İSE
 $CCI > 4,9606$ İSE VE
 $CCI \leq 69,816$ İSE VE
 $CCI \leq 20,3936$ İSE
KAP% = 0,010114

KURAL6: EĞER $CCI > -63,39$ İSE
 $CCI > 4,9606$ İSE VE
 $CCI \leq 69,816$ İSE VE
 $CCI > 20,3936$ İSE
KAP% = 0,018936

KURAL7: EĞER $CCI > -63,39$ İSE
 $CCI > 4,9606$ İSE VE
 $CCI > 69,816$ İSE
KAP% = 0,038311

6. DOHOL İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -57,17$ İSE
KAP% = -0,213644

KURAL2: EĞER $CCI > -57,17$ İSE
 $CCI \leq 4,9975$ İSE VE
 $CCI \leq -14,59$ İSE
KAP% = -0,025782

KURAL3: EĞER $CCI > -57,17$ İSE
 $CCI \leq 4,9975$ İSE VE
 $CCI > -14,59$ İSE VE
 $CCI \leq -2,478$ İSE
KAP% = -0,00926

KURAL4: EĞER $CCI > -57,17$ İSE
 $CCI \leq 4,9975$ İSE VE
 $CCI > -14,59$ İSE VE
 $CCI > -2,478$ İSE
KAP% = 0,001315

KURAL5: EĞER $CCI > -57,17$ İSE
 $CCI > 4,9975$ İSE VE
 $CCI \leq 64,5671$ İSE VE
 $STO \leq 59,9048$ İSE
KAP% = 0,02234

KURAL6: EĞER $CCI > -57,17$ İSE
 $CCI > 4,9975$ İSE VE
 $CCI \leq 64,5671$ İSE VE
 $STO > 59,9048$ İSE
KAP% = 0,013713

KURAL7: EĞER $CCI > -57,17$ İSE
 $CCI > 4,9975$ İSE VE
 $CCI > 64,5671$ İSE
KAP% = 0,051825

7. DYHOL İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -54,11$ İSE
KAP% = -0,07454

KURAL2: EĞER $CCI > -54,11$ İSE VE
 $CCI \leq 56,8395$ İSE VE
 $CCI \leq 5,2168$ İSE VE
 $CCI \leq 1,2586$ İSE VE
 $CCI \leq -8,175$ İSE
KAP% = 0,0136

KURAL3: EĞER $CCI > -54,11$ İSE VE
 $CCI \leq 56,8395$ İSE VE
 $CCI \leq 5,2168$ İSE VE
 $CCI \leq 1,2586$ İSE VE
 $CCI > -8,175$ İSE
KAP% = -0,007413

KURAL4: EĞER $CCI > -54,11$ İSE VE
 $CCI \leq 56,8395$ İSE VE
 $CCI \leq 5,2168$ İSE VE
 $CCI > 1,2586$ İSE VE
 $CCI \leq 1,3557$ İSE
KAP% = 0,000969

KURAL5: EĞER $CCI > -54,11$ İSE VE
 $CCI \leq 56,8395$ İSE VE
 $CCI \leq 5,2168$ İSE VE
 $CCI > 1,2586$ İSE VE
 $CCI > 1,3557$ İSE VE
 $CCI \leq 1,5176$ İSE
KAP% = -0,004952

KURAL6: EĞER $CCI > -54,11$ İSE VE
 $CCI \leq 56,8395$ İSE VE
 $CCI \leq 5,2168$ İSE VE
 $CCI > 1,2586$ İSE VE
 $CCI > 1,3557$ İSE VE
 $CCI > 1,5176$ İSE
KAP% = 0,000932

KURAL7: EĞER $CCI > -54,11$ İSE VE
 $CCI \leq 56,8395$ İSE VE
 $CCI > 5,2168$ İSE
KAP% = 0,011124

KURAL8: EĞER $CCI > 54,11$ İSE VE
 $CCI > 56,8395$ İSE
KAP% = 0,04377

8. ENKA İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -71,14$ İSE
KAP% = -0,208385

KURAL2: EĞER $CCI > -71,14$ İSE VE
 $CCI \leq 4,122$ İSE VE
 $CCI \leq -23,00$ İSE
KAP% = -0,041342

KURAL3: EĞER $CCI > -71,14$ İSE VE
 $CCI \leq 4,122$ İSE VE
 $CCI > -23,00$ İSE VE
 $CCI \leq -5,584$ İSE
KAP% = -0,013954

KURAL4: EĞER $CCI > -71,14$ İSE VE
 $CCI \leq 4,122$ İSE VE
 $CCI > -23,00$ İSE VE
 $CCI > -5,584$ İSE
KAP% = -0,001395

KURAL5: EĞER $CCI > -71,14$ İSE VE
 $CCI > 4,122$ İSE VE
 $CCI \leq 86,4962$ İSE VE
 $MACD \leq 1,4256$ İSE
KAP% = 0,028796

KURAL6: EĞER $CCI > -71,14$ İSE VE
 $CCI > 4,122$ İSE VE
 $CCI \leq 86,4962$ İSE VE
 $MACD > 1,4256$ İSE VE
 $CCI \leq 27,9234$ İSE
KAP% = 0,007857

KURAL7: EĞER $CCI > -71,14$ İSE VE
 $CCI > 4,122$ İSE VE
 $CCI \leq 86,4962$ İSE VE
 $MACD > 1,4256$ İSE VE
 $CCI > 27,9234$ İSE
KAP% = 0,018768

KURAL8: EĞER $CCI > -71,14$ İSE VE
 $CCI > 4,122$ İSE VE
 $CCI > 86,4962$ İSE
KAP% = 0,053758

9. EREGL İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -74,32$ İSE
KAP%= -0,158742

KURAL2: EĞER $CCI > -74,32$ İSE VE
 $CCI \leq 22,1995$ İSE VE
 $CCI \leq -13,75$ İSE VE
 $MACD \leq 0,799$ İSE
KAP%= -0,025594

KURAL3: EĞER $CCI > -74,32$ İSE VE
 $CCI \leq 22,1995$ İSE VE
 $CCI \leq -13,75$ İSE VE
 $MACD > 0,799$ İSE
KAP%= -0,012372

KURAL4: EĞER $CCI > -74,32$ İSE VE
 $CCI \leq 22,1995$ İSE VE
 $CCI > -13,75$ İSE VE
 $CCI \leq 0,6756$ İSE
KAP%= -0,004646

KURAL5: EĞER $CCI > -74,32$ İSE VE
 $CCI \leq 22,1995$ İSE VE
 $CCI > -13,75$ İSE VE
 $CCI > 0,6756$ İSE VE
 $STO \leq 56,9167$ İSE
KAP%= 0,012707

KURAL6: EĞER $CCI > -74,32$ İSE VE
 $CCI \leq 22,1995$ İSE VE
 $CCI > -13,75$ İSE VE
 $CCI > 0,6756$ İSE VE
 $STO > 56,9167$ İSE
KAP%= 0,002731

KURAL7: EĞER $CCI > -74,32$ İSE VE
 $CCI > 22,1995$ İSE VE
 $CCI \leq 89,5219$ İSE VE
KAP%= 0,021073

KURAL8: EĞER $CCI > -74,32$ İSE VE
 $CCI > 22,1995$ İSE VE
 $CCI > 89,5219$ İSE VE
KAP%= 0,059511

10. GARAN İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -76,18$ İSE
KAP%= -0,137145

KURAL2: EĞER $CCI > -76,18$ İSE VE
 $CCI \leq 83,8008$ İSE VE
 $CCI \leq -7,719$ İSE VE
 $CCI \leq -31,49$ İSE
KAP%= -0,024377

KURAL3: EĞER $CCI > -76,18$ İSE VE
 $CCI \leq 83,8008$ İSE VE
 $CCI \leq -7,719$ İSE VE
 $CCI > -31,49$ İSE
KAP%= -0,010349

KURAL4: EĞER $CCI > -76,18$ İSE VE
 $CCI \leq 83,8008$ İSE VE
 $CCI > -7,719$ İSE VE
 $CCI \leq 18,7917$ İSE VE
 $STO \leq 50,8582$ İSE
KAP%= 0,008682

KURAL5: EĞER $CCI > -76,18$ İSE VE
 $CCI \leq 83,8008$ İSE VE
 $CCI > -7,719$ İSE VE
 $CCI \leq 18,7917$ İSE VE
 $STO > 50,8582$ İSE
KAP%= -0,001367

KURAL6: EĞER $CCI > -76,18$ İSE VE
 $CCI \leq 83,8008$ İSE VE
 $CCI > -7,719$ İSE VE
 $CCI > -18,7917$ İSE VE
 $STO \leq 62$ İSE
KAP%= 0,022635

KURAL7: EĞER $CCI > -76,18$ İSE VE
 $CCI \leq 83,8008$ İSE VE
 $CCI > -7,719$ İSE VE
 $CCI > -18,7917$ İSE VE
 $STO > 62$ İSE
KAP%= 0,01065

KURAL7: EĞER $CCI > -76,18$ İSE VE
 $CCI > 83,8008$ İSE
KAP%= 0,056973

11. HALKB İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq 8,3859$ İSE VE
 $CCI \leq -81,62$ İSE
 $KAP\% = -0,034905$

KURAL2: EĞER $CCI \leq 8,3859$ İSE VE
 $CCI > -81,62$ İSE VE
 $DOL\% \leq -0,001$ İSE
 $KAP\% = -0,018481$

KURAL3: EĞER $CCI \leq 8,3859$ İSE VE
 $CCI > -81,62$ İSE VE
 $DOL\% > -0,001$ İSE VE
 $CCI \leq -32,54$ İSE
 $KAP\% = -0,006147$

KURAL4: EĞER $CCI \leq 8,3859$ İSE VE
 $CCI > -81,62$ İSE VE
 $DOL\% > -0,001$ İSE VE
 $CCI > -32,54$ İSE
 $KAP\% = 0,001714$

KURAL5: EĞER $CCI > 8,3859$ İSE VE
 $DOW\% \leq 0,0051$ İSE VE
 $CCI \leq 83,3$ İSE VE
 $MACD \leq -0,8654$ İSE
 $KAP\% = 0,010436$

KURAL6: EĞER $CCI > 8,3859$ İSE VE
 $DOW\% \leq 0,0051$ İSE VE
 $CCI \leq 83,3$ İSE VE
 $MACD > 0,8654$ İSE
 $KAP\% = 0,002523$

KURAL7: EĞER $CCI > 8,3859$ İSE VE
 $DOW\% \leq 0,0051$ İSE VE
 $CCI > 83,3$ İSE
 $KAP\% = 0,016691$

KURAL8: EĞER $CCI > 8,3859$ İSE VE
 $DOW\% > 0,0051$ İSE
 $KAP\% = 0,033062$

12. ISGYO İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -32,61$ İSE
 $KAP\% = -0,153433$

KURAL2: EĞER $CCI > -32,61$ İSE VE
 $CCI \leq 33,0613$ İSE VE
 $CCI \leq -2,003$ İSE VE
 $CCI \leq -10,66$ İSE
 $KAP\% = -0,01473$

KURAL3: EĞER $CCI > -32,61$ İSE VE
 $CCI \leq 33,0613$ İSE VE
 $CCI \leq -2,003$ İSE VE
 $CCI > -10,66$ İSE
 $KAP\% = -0,005154$

KURAL4: EĞER $CCI > -32,61$ İSE VE
 $CCI \leq 33,0613$ İSE VE
 $CCI > -2,003$ İSE VE
 $CCI \leq 7,6505$ İSE VE
 $DOW\% \leq 0,01$ İSE
 $KAP\% = -0,000273$

KURAL5: EĞER $CCI > -32,61$ İSE VE
 $CCI \leq 33,0613$ İSE VE
 $CCI > -2,003$ İSE VE
 $CCI \leq 7,6505$ İSE VE
 $DOW\% > 0,01$ İSE
 $KAP\% = 0,005594$

KURAL6: EĞER $CCI > -32,61$ İSE VE
 $CCI \leq 33,0613$ İSE VE
 $CCI > 7,6505$ İSE VE
 $MACD \leq 0,1456$ İSE
 $KAP\% = 0,019155$

KURAL7: EĞER $CCI > -32,61$ İSE VE
 $CCI \leq 33,0613$ İSE VE
 $CCI > 7,6505$ İSE VE
 $MACD > 0,1456$ İSE
 $KAP\% = 0,007783$

KURAL8: EĞER $CCI > -32,61$ İSE VE
 $CCI > 33,0613$ İSE
 $KAP\% = 0,041$

13. ISBTR İÇİN KARAR KURALLARI

KURAL1: EĞER $MOM \leq 86,9697$ İSE
KAP%= -0,182095

KURAL2: EĞER $MOM > 86,9697$ İSE VE
 $CCI \leq 0,8857$ İSE VE
 $CCI \leq -66,43$ İSE
KAP%= -0,021985

KURAL3: EĞER $MOM > 86,9697$ İSE VE
 $CCI \leq 0,8857$ İSE VE
 $CCI > -66,43$ İSE VE
 $CCI \leq -2,684$ İSE VE
 $MOM \leq 97,9055$ İSE
KAP%= -0,010925

KURAL4: EĞER $MOM > 86,9697$ İSE VE
 $CCI \leq 0,8857$ İSE VE
 $CCI > -66,43$ İSE VE
 $CCI \leq -2,684$ İSE VE
 $MOM > 97,9055$ İSE
KAP%= -0,007773

KURAL5: EĞER $MOM > 86,9697$ İSE VE
 $CCI \leq 0,8857$ İSE VE
 $CCI > -66,43$ İSE VE
 $CCI > -2,684$ İSE
KAP%= -0,00326

KURAL6: EĞER $MOM > 86,9697$ İSE VE
 $CCI > 0,8857$ İSE VE
 $MOM \leq 113,57$ İSE VE
 $CCI \leq 55,0226$ İSE VE
 $STO \leq 51,7807$ İSE
KAP%= 0,013506

KURAL7: EĞER $MOM > 86,9697$ İSE VE
 $CCI > 0,8857$ İSE VE
 $MOM \leq 113,57$ İSE VE
 $CCI \leq 55,0226$ İSE VE
 $STO > 51,7807$ İSE
KAP%= 0,007509

KURAL8: EĞER $MOM > 86,9697$ İSE VE
 $CCI > 0,8857$ İSE VE
 $MOM \leq 113,57$ İSE VE
 $CCI > 55,0226$ İSE
KAP%= 0,020585

KURAL9: EĞER $MOM > 86,9697$ İSE VE
 $CCI > 0,8857$ İSE VE
 $MOM > 113,57$ İSE
KAP%= 0,040904

14. KCHOL İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -77,85$ İSE
KAP%= -0,259827

KURAL2: EĞER $CCI > -77,85$ İSE VE
 $CCI \leq 66,109$ İSE VE
 $CCI \leq -0,865$ İSE VE
 $CCI \leq -19,77$ İSE
KAP%= -0,019463

KURAL3: EĞER $CCI > -77,85$ İSE VE
 $CCI \leq 66,109$ İSE VE
 $CCI \leq -0,865$ İSE VE
 $CCI > -19,77$ İSE VE
 $STO \leq 41,9184$ İSE
KAP%= -0,00436

KURAL4: EĞER $CCI > -77,85$ İSE VE
 $CCI \leq 66,109$ İSE VE
 $CCI \leq -0,865$ İSE VE
 $CCI > -19,77$ İSE VE
 $STO > 41,9184$ İSE
KAP%= -0,012197

KURAL5: EĞER $CCI > -77,85$ İSE VE
 $CCI \leq 66,109$ İSE VE
 $CCI > -0,865$ İSE VE
 $CCI \leq 7,4141$ İSE
KAP%= 0,002673

KURAL6: EĞER $CCI > -77,85$ İSE VE
 $CCI \leq 66,109$ İSE VE
 $CCI > -0,865$ İSE VE
 $CCI > 7,4141$ İSE VE
 $STO \leq 62,7619$ İSE
KAP%= 0,020473

KURAL7: EĞER $CCI > -77,85$ İSE VE
 $CCI \leq 66,109$ İSE VE
 $CCI > -0,865$ İSE VE
 $CCI > 7,4141$ İSE VE
 $STO > 62,7619$ İSE
KAP%= 0,009764

KURAL8: EĞER $CCI > -77,85$ İSE VE
 $CCI > 66,109$ İSE VE

KAP%= 0,050504

15. KRDMMD İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -43,81$ İSE
KAP%= -0,163131

KURAL2: EĞER $CCI > -43,81$ İSE VE
 $CCI \leq 35,9462$ İSE VE
 $CCI \leq -1,220$ İSE VE
 $CCI \leq -15,06$ İSE
KAP%= -0,018454

KURAL3: EĞER $CCI > -43,81$ İSE VE
 $CCI \leq 35,9462$ İSE VE
 $CCI \leq -1,220$ İSE VE
 $CCI > -15,06$ İSE VE
 $STO \leq 45,9793$ İSE
KAP%= -0,004088

KURAL4: EĞER $CCI > -43,81$ İSE VE
 $CCI \leq 35,9462$ İSE VE
 $CCI \leq -1,220$ İSE VE
 $CCI > -15,06$ İSE VE
 $STO > 45,9793$ İSE
KAP%= -0,011833

KURAL5: EĞER $CCI > -43,81$ İSE VE
 $CCI \leq 35,9462$ İSE VE
 $CCI > -1,220$ İSE VE
 $CCI \leq 7,0536$ İSE
KAP%= 0,001947

KURAL6: EĞER $CCI > -43,81$ İSE VE
 $CCI \leq 35,9462$ İSE VE
 $CCI > -1,220$ İSE VE
 $CCI > 7,0536$ İSE VE
 $STO \leq 61,0556$ İSE
KAP%= 0,020365

KURAL7: EĞER $CCI > -43,81$ İSE VE
 $CCI \leq 35,9462$ İSE VE
 $CCI > -1,220$ İSE VE
 $CCI > 7,0536$ İSE VE
 $STO > 61,0556$ İSE
KAP%= 0,008496

KURAL8: EĞER $CCI > -43,81$ İSE VE
 $CCI > 35,9462$ İSE
KAP%= 0,053219

16. SAHOL İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -63,46$ İSE
KAP%= -0,167644

KURAL2: EĞER $CCI > -63,46$ İSE VE
 $CCI \leq 29,1245$ İSE VE
 $CCI \leq -5,979$ İSE VE
 $CCI \leq -29,49$ İSE
KAP%= -0,020819

KURAL3: EĞER $CCI > -63,46$ İSE VE
 $CCI \leq 29,1245$ İSE VE
 $CCI \leq -5,979$ İSE VE
 $CCI > -29,49$ İSE VE
 $RSI \leq 47,5132$ İSE
KAP%= -0,007132

KURAL4: EĞER $CCI > -63,46$ İSE VE
 $CCI \leq 29,1245$ İSE VE
 $CCI \leq -5,979$ İSE VE
 $CCI > -29,49$ İSE VE
 $RSI > 47,5132$ İSE
KAP%= -0,010726

KURAL5: EĞER $CCI > -63,46$ İSE VE
 $CCI \leq 29,1245$ İSE VE
 $CCI > -5,979$ İSE VE

$CCI \leq 4,5558$ İSE
KAP%= 0,000656

KURAL6: EĞER $CCI > -63,46$ İSE VE
 $CCI \leq 29,1245$ İSE VE
 $CCI > -5,979$ İSE VE
 $CCI \leq 15,4572$ İSE VE
 $CCI > 4,5558$ İSE
KAP%= 0,00362

KURAL7: EĞER $CCI > -63,46$ İSE VE
 $CCI \leq 29,1245$ İSE VE
 $CCI > -5,979$ İSE VE
 $CCI > 15,4572$ İSE
KAP%= 0,009546

KURAL8: EĞER $CCI > -63,46$ İSE VE
 $CCI > 29,1245$ İSE VE
 $CCI \leq 67,9708$ İSE
KAP%= 0,019076

KURAL9: EĞER $CCI > -63,46$ İSE VE
 $CCI > 29,1245$ İSE VE
 $CCI > 67,9708$ İSE
KAP%= 0,046726

CCI≤15,4572 İSE VE

17. SKBNK İÇİN KARAR KURALLAR:

KURAL1: EĞER CCI≤-77,43 İSE
KAP%= -0,092477

KURAL2: EĞER CCI>-77,43 İSE VE
CCI≤52,3464 İSE VE
CCI≤-0,4362 İSE VE
CCI≤-23,01 İSE
KAP%= -0,018853

KURAL3: EĞER CCI>-77,43 İSE VE
CCI≤52,3464 İSE VE
CCI≤-0,4362 İSE VE
CCI>-23,01 İSE VE
STO≤46,5505 İSE
KAP%= -0,000289

KURAL4: EĞER CCI>-77,43 İSE VE
CCI≤52,3464 İSE VE
CCI≤-0,4362 İSE VE
CCI>-23,01 İSE VE
STO>46,5505 İSE
KAP%= -0,01288

KURAL5: EĞER CCI>-77,43 İSE VE
CCI≤52,3464 İSE VE
CCI>-0,4362 İSE VE
CCI≤19,6299 İSE
KAP%= 0,004098

KURAL6: EĞER CCI>-77,43 İSE VE
CCI≤52,3464 İSE VE
CCI>-0,4362 İSE VE
CCI>19,6299 İSE
KAP%= 0,014129

KURAL7: EĞER CCI>-77,43 İSE VE
CCI>52,3464 İSE
KAP%= 0,04645

18. SİSE İÇİN KARAR KURALLARI

KURAL1: EĞER CCI≤-73,99 İSE
KAP%= -0,135328

KURAL2: EĞER CCI>-73,99 İSE VE
CCI≤53,2507 İSE VE
CCI≤-3,958 İSE VE
CCI≤-20,97 İSE
KAP%= -0,02343

KURAL3: EĞER CCI>-73,99 İSE VE
CCI≤53,2507 İSE VE
CCI≤-3,958 İSE VE
CCI>-20,97 İSE
KAP%= -0,010404

KURAL4: EĞER CCI>-73,99 İSE VE
CCI≤53,2507 İSE VE
CCI>-3,958 İSE VE
CCI≤9,7548 İSE VE
STO≤51,6944 İSE
KAP%= 0,006029

KURAL5: EĞER CCI>-73,99 İSE VE
CCI≤53,2507 İSE VE
CCI>-3,958 İSE VE
CCI≤9,7548 İSE VE
STO>51,6944 İSE
KAP%= -0,000957

KURAL6: EĞER CCI>-73,99 İSE VE
CCI≤53,2507 İSE VE
CCI>-3,958 İSE VE
CCI>9,7548 İSE VE
MACD≤0,3825 İSE
KAP%= 0,021617

KURAL7: EĞER CCI>-73,99 İSE VE
CCI≤53,2507 İSE VE
CCI>-3,958 İSE VE
CCI>9,7548 İSE VE
MACD>0,3825 İSE
KAP%= 0,000786

KURAL8: EĞER CCI>-73,99 İSE VE
CCI>53,2507 İSE
KAP%= 0,04618

19. PETKM İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -49,54$ İSE
 $KAP\% = -0,124053$

KURAL2: EĞER $CCI > -49,54$ İSE VE
 $CCI \leq 1,8884$ İSE VE
 $CCI \leq -6,499$ İSE VE
 $CCI \leq -15,25$ İSE
 $KAP\% = -0,024741$

KURAL3: EĞER $CCI > -49,54$ İSE VE
 $CCI \leq 1,8884$ İSE VE
 $CCI \leq -6,499$ İSE VE
 $CCI > -15,25$ İSE
 $KAP\% = -0,018065$

KURAL4: EĞER $CCI > -49,54$ İSE VE
 $CCI \leq 1,8884$ İSE VE
 $CCI > -6,499$ İSE VE
 $CCI \leq -0,576$ İSE
 $KAP\% = -0,009079$

KURAL5: EĞER $CCI > -49,54$ İSE VE
 $CCI \leq 1,8884$ İSE VE
 $CCI > -6,499$ İSE VE
 $CCI > -0,576$ İSE
 $KAP\% = 0,001441$

KURAL6: EĞER $CCI > -49,54$ İSE VE
 $CCI > 1,8884$ İSE VE
 $CCI \leq 16,1253$ İSE VE
 $MOM \leq 103,91$ İSE
 $KAP\% = 0,008265$

KURAL7: EĞER $CCI > -49,54$ İSE VE
 $CCI > 1,8884$ İSE VE
 $CCI > 16,1253$ İSE VE
 $MOM > 103,91$ İSE
 $KAP\% = 0,016639$

KURAL8: EĞER $CCI > -49,54$ İSE VE
 $CCI > 1,8884$ İSE VE
 $CCI > 16,1253$ İSE VE
 $MOM \leq 111,13$ İSE
 $KAP\% = 0,025274$

KURAL9: EĞER $CCI > -49,54$ İSE VE
 $CCI > 1,8884$ İSE VE
 $CCI > 16,1253$ İSE VE
 $MOM > 111,13$ İSE
 $KAP\% = 0,046572$

20. TAVHL İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -95,80$ İSE
 $KAP\% = -0,102676$

KURAL2: EĞER $CCI > -95,80$ İSE VE
 $CCI \leq 91,1924$ İSE VE
 $CCI \leq 3,5645$ İSE VE
 $MACD \leq 0,3935$ İSE
 $KAP\% = -0,020696$

KURAL3: EĞER $CCI > -95,80$ İSE VE
 $CCI \leq 91,1924$ İSE VE
 $CCI \leq 3,5645$ İSE VE
 $MACD > 0,3935$ İSE VE
 $STO \leq 43,3654$ İSE
 $KAP\% = -0,000957$

KURAL4: EĞER $CCI > -95,80$ İSE VE
 $CCI \leq 91,1924$ İSE VE
 $CCI \leq 3,5645$ İSE VE
 $MACD > 0,3935$ İSE VE
 $STO > 43,3654$ İSE
 $KAP\% = -0,008867$

KURAL5: EĞER $CCI > -95,80$ İSE VE
 $CCI \leq 91,1924$ İSE VE
 $CCI > 3,5645$ İSE VE
 $MACD \leq 0,4031$ İSE
 $KAP\% = 0,034923$

KURAL6: EĞER $CCI > -95,80$ İSE VE
 $CCI \leq 91,1924$ İSE VE
 $CCI > 3,5645$ İSE VE
 $MACD > 0,4031$ İSE VE
 $CCI \leq 45,5962$ İSE
 $KAP\% = 0,003058$

KURAL7: EĞER $CCI > -95,80$ İSE VE
 $CCI \leq 91,1924$ İSE VE
 $CCI > 3,5645$ İSE VE
 $MACD > 0,4031$ İSE VE
 $CCI > 45,5962$ İSE
 $KAP\% = 0,011798$

KURAL8: EĞER $CCI > -95,80$ İSE VE
 $CCI > 91,1924$ İSE VE
 $KAP\% = 0,066757$

21. TEBNK İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -74,57$ İSE
KAP%= -0,147407

KURAL2: EĞER $CCI > -74,57$ İSE VE
 $CCI \leq 4,5273$ İSE VE
 $CCI \leq -29,83$ İSE VE
KAP%= -0,020851

KURAL3: EĞER $CCI > -74,57$ İSE VE
 $CCI \leq 4,5273$ İSE VE
 $CCI > -29,83$ İSE VE
 $STO \leq 52,619$ İSE
KAP%= -0,002518

KURAL4: EĞER $CCI > -74,57$ İSE VE
 $CCI \leq 4,5273$ İSE VE
 $CCI > -29,83$ İSE VE
 $STO > 52,619$ İSE
KAP%= -0,012416

KURAL5: EĞER $CCI > -74,57$ İSE VE
 $CCI > 4,5273$ İSE VE
 $CCI \leq 86,818$ İSE VE
 $MACD \leq 0,3365$ İSE
KAP%= 0,020644

KURAL6: EĞER $CCI > -74,57$ İSE VE
 $CCI > 4,5273$ İSE VE
 $CCI \leq 86,818$ İSE VE
 $MACD > 0,33650,0017$ İSE VE
 $CCI \leq 29,8271$ İSE
KAP%= 0,013483

KURAL7: EĞER $CCI > -74,57$ İSE VE
 $CCI > 4,5273$ İSE VE
 $CCI \leq 86,818$ İSE VE
 $MACD > 0,3365$ İSE VE
 $CCI > 29,8271$ İSE
KAP%= 0,013043

KURAL8: EĞER $CCI > -74,57$ İSE VE
 $CCI > 4,5273$ İSE VE
 $CCI > 86,818$ İSE
KAP%= 0,039256

22. THY İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -80,58$ İSE
KAP%= -0,098741

KURAL2: EĞER $CCI > -80,58$ İSE VE
 $CCI \leq -0,128$ İSE VE
 $CCI \leq -29,30$ İSE VE
KAP%= -0,022039

KURAL3: EĞER $CCI > -80,58$ İSE VE
 $CCI \leq -0,128$ İSE VE
 $CCI > -29,30$ İSE VE
 $STO \leq 45,5534$ İSE
KAP%= -0,004352

KURAL4: EĞER $CCI > -80,58$ İSE VE
 $CCI \leq -0,128$ İSE VE
 $CCI > -29,30$ İSE VE
 $STO > 45,5534$ İSE
KAP%= -0,01468

KURAL5: EĞER $CCI > -80,58$ İSE VE
 $CCI > -0,128$ İSE VE
 $CCI \leq 75,6897$ İSE VE
 $CCI \leq 16,5282$ İSE
KAP%= 0,007759

KURAL6: EĞER $CCI > -80,58$ İSE VE
 $CCI > -0,128$ İSE VE
 $CCI \leq 75,6897$ İSE VE
 $CCI > 16,5282$ İSE
KAP%= 0,016903

KURAL7: EĞER $CCI > -80,58$ İSE VE
 $CCI > -0,128$ İSE VE
 $CCI > 75,6897$ İSE
KAP%= 0,04884

23. TKFEN İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -68,27$ İSE
KAP% = -0,039463

KURAL2: EĞER $CCI > -68,27$ İSE VE
 $CCI \leq 44,0463$ İSE VE
 $CCI \leq -11,18$ İSE VE
 $NIK\% \leq -0,008$ İSE
KAP% = -0,013083

KURAL3: EĞER $CCI > -68,27$ İSE VE
 $CCI \leq 44,0463$ İSE VE
 $CCI \leq -11,18$ İSE VE
 $NIK\% > -0,008$ İSE
KAP% = -0,004245

KURAL4: EĞER $CCI > -68,27$ İSE VE
 $CCI \leq 44,0463$ İSE VE
 $CCI > -11,18$ İSE VE
 $DOL\% \leq 0,0024$ İSE VE
 $STO \leq 58,5878$ İSE
KAP% = 0,004838

KURAL5: EĞER $CCI > -68,27$ İSE VE
 $CCI \leq 44,0463$ İSE VE
 $CCI > -11,18$ İSE VE
 $DOL\% \leq 0,0024$ İSE VE
 $STO > 58,5878$ İSE
KAP% = -0,004241

KURAL6: EĞER $CCI > -68,27$ İSE VE
 $CCI \leq 44,0463$ İSE VE
 $CCI > -11,18$ İSE VE
 $DOL\% > 0,0024$ İSE
KAP% = 0,01033

KURAL7: EĞER $CCI > -68,27$ İSE VE
 $CCI \leq 44,0463$ İSE VE
 $STO \leq 60,3824$ İSE
KAP% = 0,029365

KURAL8: EĞER $CCI > -68,27$ İSE VE
 $CCI \leq 44,0463$ İSE VE
 $STO > 60,3824$ İSE
KAP% = 0,018272

24. TOASO İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -60,87$ İSE
KAP% = -0,088402

KURAL2: EĞER $CCI > -60,87$ İSE VE
 $CCI \leq 14,8088$ İSE VE
 $CCI \leq -6,996$ İSE VE
 $MACD \leq 0,4985$ İSE
KAP% = -0,024657

KURAL3: EĞER $CCI > -60,87$ İSE VE
 $CCI \leq 14,8088$ İSE VE
 $CCI \leq -6,996$ İSE VE
 $MACD > 0,4985$ İSE
KAP% = -0,010675

KURAL4: EĞER $CCI > -60,87$ İSE VE
 $CCI \leq 14,8088$ İSE VE
 $CCI > -6,996$ İSE VE
 $CCI \leq 0$ İSE
KAP% = -0,004036

KURAL5: EĞER $CCI > -60,87$ İSE VE
 $CCI \leq 14,8088$ İSE VE
 $CCI > -6,996$ İSE VE
 $CCI > 0$ İSE VE
 $STO \leq 57,1618$ İSE
KAP% = 0,010733

KURAL6: EĞER $CCI > -60,87$ İSE VE
 $CCI \leq 14,8088$ İSE VE
 $CCI > -6,996$ İSE VE
 $CCI > 0$ İSE VE
 $STO > 57,1618$ İSE
KAP% = 0,002645

KURAL7: EĞER $CCI > -60,87$ İSE VE
 $CCI > 14,8088$ İSE VE
 $CCI \leq 65,9983$ İSE
KAP% = 0,020118

KURAL8: EĞER $CCI > -60,87$ İSE VE
 $CCI > 14,8088$ İSE VE
 $CCI > 65,9983$ İSE
KAP% = 0,04281

25. TSKB İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -60,93$ İSE
KAP%= -0,153662

KURAL2: EĞER $CCI > -60,93$ İSE VE
 $CCI \leq 44,8855$ İSE VE
 $CCI \leq -6,803$ İSE VE
 $MACD \leq 0,1387$ İSE
KAP%= -0,022314

KURAL3: EĞER $CCI > -60,93$ İSE VE
 $CCI \leq 44,8855$ İSE VE
 $CCI \leq -6,803$ İSE VE
 $MACD > 0,1387$ İSE
KAP%= -0,011465

KURAL4: EĞER $CCI > -60,93$ İSE VE
 $CCI \leq 44,8855$ İSE VE
 $CCI > -6,803$ İSE VE
 $CCI \leq 13,6078$ İSE VE
 $CCI \leq 2,2341$ İSE
KAP%= -0,001355

KURAL5: EĞER $CCI > -60,93$ İSE VE
 $CCI \leq 44,8855$ İSE VE
 $CCI > -6,803$ İSE VE
 $CCI \leq 13,6078$ İSE VE
 $CCI > 2,2341$ İSE
KAP%= 0,006037

KURAL6: EĞER $CCI > -60,93$ İSE VE
 $CCI \leq 44,8855$ İSE VE
 $CCI > -6,803$ İSE VE
 $CCI \leq 13,6078$ İSE
KAP%= 0,014808

KURAL7: EĞER $CCI > -60,93$ İSE VE
 $CCI > 44,8855$ İSE
KAP%= 0,048321

26. TTKOM İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq 66,6785$ İSE VE
 $CCI \leq -91,09$ İSE VE
KAP%= -0,02637

KURAL2: EĞER $CCI \leq 66,6785$ İSE VE
 $CCI > -91,09$ İSE VE
 $CCI \leq -13,77$ İSE VE
 $DOW\% \leq -0,04$ İSE
KAP%= -0,11455

KURAL3: EĞER $CCI \leq 66,6785$ İSE VE
 $CCI > -91,09$ İSE VE
 $CCI \leq -13,77$ İSE VE
 $DOW\% > -0,04$ İSE VE
 $MACD \leq 0,452$ İSE
KAP%= -0,006695

KURAL4: EĞER $CCI \leq 66,6785$ İSE VE
 $CCI > -91,09$ İSE VE
 $CCI \leq -13,77$ İSE VE
 $DOW\% > -0,04$ İSE VE
 $MACD > 0,452$ İSE
KAP%= -0,000573

KURAL5: EĞER $CCI \leq 66,6785$ İSE VE
 $CCI > -91,09$ İSE VE
 $CCI > -13,77$ İSE VE
 $DOL\% \leq -0,003$ İSE
KAP%= -0,002081

KURAL6: EĞER $CCI \leq 66,6785$ İSE VE
 $CCI > -91,09$ İSE VE
 $CCI > -13,77$ İSE VE
 $DOL\% > -0,003$ İSE VE
 $CCI \leq 13,1479$ İSE
KAP%= -0,000134

KURAL7: EĞER $CCI \leq 66,6785$ İSE VE
 $CCI > -91,09$ İSE VE
 $CCI > -13,77$ İSE VE
 $DOL\% > -0,003$ İSE VE
 $CCI > 13,1479$ İSE
KAP%= 0,007073

KURAL8: EĞER $CCI > 66,6785$ İSE VE
 $CCI \leq 101,74$ İSE
KAP%= 0,012935

KURAL9: EĞER $CCI > 66,6785$ İSE VE
 $CCI > 101,74$ İSE

KAP%= 0,024037

27. TUPRPS İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -74,91$ İSE
KAP%= -0,129202

KURAL2: EĞER $CCI > -74,91$ İSE VE
 $CCI \leq 7,332$ İSE VE
 $CCI \leq -24,39$ İSE
KAP%= -0,018976

KURAL3: EĞER $CCI > -74,91$ İSE VE
 $CCI \leq 7,332$ İSE VE
 $CCI > -24,39$ İSE VE
 $STO \leq 54,9167$ İSE
KAP%= -0,001395

KURAL4: EĞER $CCI > -74,91$ İSE VE
 $CCI \leq 7,332$ İSE VE
 $CCI > -24,39$ İSE VE
 $STO > 54,9167$ İSE
KAP%= -0,010786

KURAL5: EĞER $CCI > -74,91$ İSE VE
 $CCI > 7,332$ İSE VE
 $CCI \leq 83,5205$ İSE VE
 $MACD \leq 1,1034$ İSE
KAP%= 0,027252

KURAL6: EĞER $CCI > -74,91$ İSE VE
 $CCI > 7,332$ İSE VE
 $CCI \leq 83,5205$ İSE VE
 $MACD > 1,1034$ İSE VE
 $CCI \leq 38,8912$ İSE
KAP%= 0,005691

KURAL7: EĞER $CCI > -74,91$ İSE VE
 $CCI > 7,332$ İSE VE
 $CCI \leq 83,5205$ İSE VE
 $MACD > 1,1034$ İSE VE
 $CCI > 38,8912$ İSE
KAP%= 0,015838

KURAL8: EĞER $CCI > -74,91$ İSE VE
 $CCI > 7,332$ İSE VE
 $CCI > 83,5205$ İSE
KAP%= 0,047808

28. TCELL İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -77,65$ İSE
KAP%= -0,194639

KURAL2: EĞER $CCI > -77,65$ İSE VE
 $CCI \leq 74,4822$ İSE VE
 $CCI \leq -9,1914$ İSE VE
 $CCI \leq -39,62$ İSE
KAP%= -0,017581

KURAL3: EĞER $CCI > -77,65$ İSE VE
 $CCI \leq 74,4822$ İSE VE
 $CCI \leq -9,1914$ İSE VE
 $CCI > -39,62$ İSE VE
 $NIK \% > 0,0037$ İSE VE
 $STO \leq 53,2619$
KAP%= 0,003703

KURAL4: EĞER $CCI > -77,65$ İSE VE
 $CCI \leq 74,4822$ İSE VE
 $CCI \leq -9,1914$ İSE VE
 $CCI > -39,62$ İSE VE
 $NIK \% > 0,0037$ İSE VE
 $STO > 53,2619$ İSE
KAP%= -0,010861

KURAL5: EĞER $CCI > -77,65$ İSE VE
 $CCI \leq 74,4822$ İSE VE
 $CCI \leq -9,1914$ İSE VE
 $CCI > -39,62$ İSE VE
 $NIK \% > 0,0037$ İSE VE
KAP%= 0,001026

KURAL6: EĞER $CCI > -77,65$ İSE VE
 $CCI \leq 74,4822$ İSE VE
 $CCI > -9,1914$ İSE VE
 $STO \leq 62,6633$ İSE
KAP%= 0,018444

KURAL7: EĞER $CCI > -77,65$ İSE VE
 $CCI \leq 74,4822$ İSE VE
 $CCI > -9,1914$ İSE VE
 $STO > 62,6633$ İSE
KAP%= 0,007018

KURAL8: EĞER $CCI > -77,65$ İSE VE
 $CCI > 74,4822$ İSE
KAP%= 0,051306

29. VKBNK İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq 6,2782$ İSE VE
 $CCI \leq -37,02$ İSE VE
 $NIK\% \leq -0,005$ İSE
 $KAP\% = -0,032941$

KURAL2: EĞER $CCI \leq 6,2782$ İSE VE
 $CCI \leq -37,02$ İSE VE
 $NIK\% > -0,005$ İSE
 $KAP\% = -0,017665$

KURAL3: EĞER $CCI \leq 6,2782$ İSE VE
 $CCI > -37,02$ İSE VE
 $DOL\% \leq -0,000$ İSE
 $KAP\% = -0,00995$

KURAL4: EĞER $CCI \leq 6,2782$ İSE VE
 $CCI > -37,02$ İSE VE
 $DOL\% > -0,000$ İSE
 $KAP\% = -0,00565$

KURAL5: EĞER $CCI > 6,2782$ İSE VE
 $CCI \leq 88,5449$ İSE VE
 $DOW\% \leq 0,0061$ İSE VE
 $DOL\% \leq -0,004$ İSE
 $KAP\% = -0,00033$

KURAL6: EĞER $CCI > 6,2782$ İSE VE
 $CCI \leq 88,5449$ İSE VE
 $DOW\% \leq 0,0061$ İSE VE
 $DOL\% > -0,004$ İSE
 $KAP\% = 0,011844$

KURAL7: EĞER $CCI > 6,2782$ İSE VE
 $CCI \leq 88,5449$ İSE VE
 $DOW\% > 0,0061$ İSE
 $KAP\% = 0,020618$

KURAL8: EĞER $CCI > 6,2782$ İSE VE
 $CCI > 88,5449$ İSE
 $KAP\% = 0,031058$

30. YKBNK İÇİN KARAR KURALLARI

KURAL1: EĞER $CCI \leq -62,33$ İSE
 $KAP\% = -0,180647$

KURAL2: EĞER $CCI > -62,33$ İSE VE
 $CCI \leq 13,762$ İSE VE
 $CCI \leq -11,13$ İSE VE
 $CCI \leq -28,54$ İSE
 $KAP\% = -0,029526$

KURAL3: EĞER $CCI > -62,33$ İSE VE
 $CCI \leq 13,762$ İSE VE
 $CCI \leq -11,13$ İSE VE
 $CCI > -28,54$ İSE
 $KAP\% = -0,013279$

KURAL4: EĞER $CCI > -62,33$ İSE VE
 $CCI \leq 13,762$ İSE VE
 $CCI > -11,13$ İSE VE
 $CCI \leq -3,389$ İSE
 $KAP\% = -0,005855$

KURAL5: EĞER $CCI > -62,33$ İSE VE
 $CCI \leq 13,762$ İSE VE
 $CCI > -11,13$ İSE VE
 $CCI > -3,389$ İSE VE
 $CCI \leq 3,6429$ İSE
 $KAP\% = 0,000316$

KURAL6: EĞER $CCI > -62,33$ İSE VE
 $CCI \leq 13,762$ İSE VE
 $CCI > -11,13$ İSE VE
 $CCI > -3,389$ İSE VE
 $CCI > 3,6429$ İSE
 $KAP\% = 0,004324$

KURAL7: EĞER $CCI > -62,33$ İSE VE
 $CCI > 13,762$ İSE VE
 $CCI \leq 68,9113$ İSE VE
 $MACD \leq 0,3648$ İSE
 $KAP\% = 0,02508$

KURAL8: EĞER $CCI > -62,33$ İSE VE
 $CCI > 13,762$ İSE VE
 $CCI \leq 68,9113$ İSE VE
 $MACD > 0,3648$ İSE
 $KAP\% = 0,010326$

KURAL9: EĞER $CCI > -62,33$ İSE VE
 $CCI > 13,762$ İSE VE
 $CCI > 68,9113$ İSE
 $KAP\% = 0,056544$

KAYNAKLAR

1. Albayrak, A.S., Yılmaz, Ş.K., “Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine bir Uygulama”, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 2009, C.14, S.1 s.31-52.
2. Aitkenhead, M. J. (2008); “A Co-Evolving Decision Tree Classification,” Expert Systems with Applications, Cilt 34, Sayı 1, s. 18-25.
3. Aybars A., “Modern Teknik Analiz”, Dünya Yayıncılık,1997.
4. Baldi, P., Brunak, S., “Bioinformatics : the machine learning approach”, The MIT Press, 2001.
5. Berry., M., Linoff., G., “Data mining techniques: for marketing, sales, and customer relationship management”, Wiley, 1997.
6. Berti, L., “Measuring and Modelling Data Quality for Quality-Awareness in Data Mining”, (in: “Quality Measures in Data Mining” Ed: Hamilton, H.J), Springer, 2007.
7. Bhavani Thuraisingham, ‘A Primer for Understanding and Applying Data Mining’, IT Professional Technology Solutions for the Enterprise, Vol.2, No:6, (2000), s.55.
8. Bishop, C.M., “Pattern Recognition and Machine Learning”, Springer, 2006.
9. Breiman, L., “Bagging predictors”, Machine Learning 24, 123–140, 1996.
10. Breiman, L., Freidman, J.H., Olshen, R.A., Stone, C.J., “Classification And Regression Trees”, Chapman&Hall, 1998.
11. Caruana, R., Mizil, A.N., “An Empirical Comparison of Supervised Learning Algorithms”, Proceedings of the 23rd international conference on Machine learning, 2.161-168, 2006.
12. Chatterjee, S., Hadi, A.S., “Regression Analysis by Example”, Wiley, 2006.
13. Chen, Ming-Syan, Jiawei Han, Philip S. Yu; 1996 , Data Mining: An Overview from Database Perspective, IEEE Transactions on Knowledge and Data Engineering, 8:866-883, December 1996
14. Cios, K., W., Pedrycz, Swiniarski, K., "Data Mining Methods for Knowledge Discovery", Kluwer, 1998.
15. Clarke, B., Fokou'e, E., Zhang, H.H., “Principles and Theory for Data Mining And Machine Learning”, Springer, 2009.

16. Cristianini, N., "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.
17. Dasu, T., Johnson, T., "Exploratory data mining and data cleaning", Wiley, 2003.
18. DeMaris, A., "Regression with social data : modeling continuous and limited response variables", Wiley, 2004.
19. Duda, R.O., Hart, P.E., Stork, D.G., "Pattern Classification", Wiley, 2000.
20. Efron, B., Tibshirani R. J., "An Introduction to the Bootstrap", Chapman & Hall/CRC, 1994.
21. Estivill-Castro, Houle, M.E., "The Discovery of Interesting Nuggets Using Heuristic Techniques", (İçinde: Abbass, H.A., Sarker, R.A., Newton, C.S.,(Eds), "Data mining : a heuristic approach"), Idea Group Publishing, 2002.
22. Fawcett, T., (2004), "ROC Graphs: Notes and Practical Considerations for Researchers", Technical report, Palo Alto, USA: HP Laboratories.
23. Fawcett, T., "An Introduction to ROC Analysis", Pattern Recognition Letters, 27:861-874, 2005.
24. Fitzmaurice, G.M., Krzanowski, W. J., Hand, D. J., "A Monte Carlo study of the 632 bootstrap estimator of error rate", Journal of Classification, Volume 8, Number 2, 239-250, 1991.
25. Flasch, P., Blockeel, H., Ferri, C., Hernandez, J., Struyf, J., "Introduction to ROC analysis and its Applications", (ed: Mladenović, D.), "Data mining and decision support: integration and collaboration", Kluwer Academic Publishers, 2003.
26. Freund, Y., Schapire, R.E "Game Theory, On-Line Prediction and Boosting", Annual Workshop on Computational Learning Theory, pp.325-332, 1996.
27. Gey, S., Nadelec, E., "Model Selection for CART Regression Trees", IEEE Transactions On Information Theory, 2005, VOL 51; NUMB 2, pages 658-670.
28. Greene, W.H., "Econometric Analysis", Prentice Hall, 2003).
29. Hair, Joseph F., Ronald L. Tahtam, Rolph E. Anderson, William Black; 1998, Multivariate Data Analysis, New Jersey: Prentice Hall.
30. Haldun Akpınar, "Veritabanlarında Bilgi Keşfi ve Veri Madenciliği," İşletme Fakültesi Dergisi, (Cilt No: 29, Sayı No:1: 1-22, Nisan 2000). s.6.
31. Han, J., Kamber, M., "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2006.

32. Han, J., Kamber, M., "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2006.
33. Heckerman, D., "A Tutorial On Learning Bayesian Networks", Technical Report, Microsoft Research, 1995.
34. Iglesia, B., Rayward-Smith, V.J., "The Discovery of Interesting Nuggets Using Heuristic Techniques", (İçinde: Abbass, H.A., Sarker, R.A., Newton, C.S.,(Eds), "Data mining : a heuristic approach"), Idea Group Publishing, 2002.
35. Jain, A.K., Dubes, R.C., "Algorithms for Clustered Data", Prentice Hall, 1998.
36. Jensen, F.V., "Bayesian Networks and Decision Graphs", Springer, 2001.
37. Kantardzic, M., "Data Mining: Concepts, Methods, and Algorithms", Wiley, 2003.
38. Kecman, V., "Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models ", MIT Press, 2001.
39. Khemka, Alok; 2003, "A Collaborative Predictive Data Mining Model," Yüksek Lisans Tezi, Faculty of University of Missouri-Kansas City, Missouri
40. Kmenta, "Basic Econometrics", Tokyo, 1978.
41. Koza, J., "Data mining using grammar based genetic programming and applications", Kluwer Academic Publishers, 2002.
42. Larose, D.T., "Data mining methods and models", Wiley, 2006.
43. Larose, D.T., "Discovering Knowledge in Data: An introduction to Data Mining", Wiley, 2005.
44. Lenic, M., Povalej, P., Kokol, P., "Impact of Purity Measures on Knowledge Extraction in Decision Trees", (In: Lin, T.Y., Ohsuga, S., Churn-Jung Liao, C., Hu, X. (Eds), "Foundations and Novel Approaches in Data Mining"), Springer, 2006.
45. Liu, H. Motoda, H., "Computational methods of feature selection", CRC Press, 2007:
46. Mackay, D.J.C., "Information Theory, Inference, and Learning Algorithms", Cambridge University Press, 2003.
47. Maimon O., Rokach, L., "Decision Trees", (İçinde: Maimon O., Rokach, L., (Eds.) "Data Mining And Knowledge Discovery Handbook"), Springer, 2005.
48. Maimon, O., Rokach, L., "Data Mining And Knowledge Discovery Handbook", Spring, 2007.

49. Mitchell, T.M., "Machine Learning", McGraw-Hill, 1997.
50. Mitra, S., Acharya, T., "Data Mining Multimedia, Soft Computing and Bioinformatics", Wiley, 2003.
51. Murray J. Mackinnon ve Ned Glick, 'Data Mining and Knowledge Discovery in Databases- An Overview', J.Statistics., Vol.41, No.3, (1999), s.260.
52. Nabiyev, V.V., "Yapay Zeka", Seçkin Yayınevi, 2003.
53. Nilson, N.J., "Introduction to Machine Learning", <http://ai.stanford.edu/~nilsson/mlbook.html>, Ocak 2011.
54. Nong, Y., (Ed) "The Handbook Of Data Mining", Gehrke, J., "Decision Trees", Lawrence Erlbaum Associates, Publishers, 2003.
55. Olafsson, Sigurdur, Xiaonan Li ve Shuning Wu (2008); "Operations Research and Data Mining," European Journal of Operational Research, Cilt 187, Sayı 3, s.1429-1448.
56. Olson, D.L., Delen, D., "Advanced Data Mining Techniques", Springer, 2008.
57. Özdamar, K., "Paket Programlar İle İstatistiksel Veri Analizi (Cilt 2)", Kaan Kitapevi, 2004.
58. Özekşi A., "Kazanma Sanatı", Alfa Yayınevi, 2005.
59. Özkan, Y., "Veri Madenciliği Yöntemleri", Papatya yayınları, 2008.
60. Öztemel, E., "Yapay Sinir Ağları", Papatya Yayıncılık, 2003.
61. Paliouras, G., Karkaletsis, V., Spyropoulos, C.D., "Machine learning and its applications", Springer, 1998.
62. Person, J.L., "A Complete Guide to Technical Trading Tactics", Wiley, 2004.
63. Quinlan, J.R., "C4.5: Programs For Machine Learning", Morgan Kaufmann, 1993.
64. Quinlan, J.R., "Improved Use of Continuous Attributes in C4.5", Journal of Artificial Intelligence Research 4, 1996
65. Quinlan, J.R., "Induction of Decision Trees", Machine Learning, 1: 81-106, 1986.
66. Rawlings, J.O., Pentula, S.G., Dickey, D.A., "Applied regression analysis: a research tool", Springer, 1998.
67. Shannon, C.E., "A Mathematical Theory of Communication", The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

68. Stepaniuk , J., “Rough - Granular Computing in Knowledge Discovery and Data Mining”, Springer, 2008.
69. Sumathi,S., Sivanandam, S.N.,” Introduction to Data Mining and its Applications”, Wiley, 2006.
70. Sun, Jie ve Hui Li (2008); “Data Mining Method for Listed Companies, Financial Distress Prediction,” Knowledge-Based Systems, Cilt 21, Sayı 1, s. 1-5.
71. Thomas, Lyn. C. (2000); “A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumer,” International Journal of Forecasting, Cilt 16, Sayı 2, s. 149-172.
72. Toktas, Peral ve Melek Basak Demirhan (2004); “Risk Analizinde Veri Madenciliği Uygulamalar,” YA/EM Yöneylem Arastırması/Endüstri Mühendisligi- XXIV Ulusal Kongresi, 15- 18 Haziran, Adana.
73. Ünlü S., Hakan, A.,”Psikoloji”, Anadolu Üniversitesi Yayını, 2001.
74. Vapnik, V.N., "The Nature of Statistical Learning Theory", Springer, 1999.
75. Vapnik, V.N., “Statistical Learning Theory”, Wiley Inter-Sciences, 1998.
76. Vercellis, “Business intelligence : Data Mining and Optimization for Decision Making”, Wiley, 2009.
77. Weir, D. J., “Timing the market : how to profit in the stock market using the yield curve, technical analysis, and cultural indicators”, Wiley, 2006.
78. Weisberg, S., “Applied Linear Regression”, Wiley, 2005.
79. Witten,I.H., Frank, E., , “Data mining : practical machine learning tools and techniques”, Morgan Kaufmann Publishers, 2005.
80. Yohannes, Y., Hoddinott, J., “Classification And Regression Trees: An Introduction (Technical Guide #39)”, International Food Policy Research Institute, 1999.

ÖZGEÇMİŞ

Adı Soyadı : Gülser DONDURMACI

Doğum Yeri : İstanbul

Doğum Yılı : 05.02.1957

Medeni Hali : Evli

Ünvanı: Öğretim Görevlisi

Yabancı Dil : İngilizce

Eğitim Durumu:

Derece	Alan	Üniversite	Yıl
Lisans	Jeofizik Mühendisliği	İstanbul Üniversitesi	1981
Y. Lisans	İstatistik	Marmara Üniversitesi	1992

Ulusal Hakemli Dergilerde Yayımlanan Makaleler:

- [1] “Oracle7 Workgroup Server Nesneleri”, BYTE Türkiye, Aralık 1998.
- [2] “Oracle Mimarisi”, BYTE Türkiye, Kasım 1998.

Ulusal Bilimsel Toplantılarda Sunulan Ve Bildiri Kitabında Basılan Bildiriler:

- [1] Özkan, Y., Dondurmacı, G., “Elektronik tablolara dayalı ve dağıtıma yönelik bir Karar Destek Sistemi”, 8. Türkiye Bilgisayar Kongresi.
- [2] Özkan, Y., Dondurmacı, G., “Karar Destek Sistemleri Geliştirme Süreci ve Uygulamaları”, Bilişim’94, İstanbul 1994.
- [3] Özkan, Y., Dondurmacı, G., “İnternet Üzerindeki Üniversite ve bir Uygulama”, 22. Ulusal Bilişim Kurultayı, 2003.

Diğer Yayınlar:

- [1] Özkan, Y., Dondurmacı, G., “Oracle7 Workgroup Server”, ANT Eğitim Dizisi, 1996.
- [2] Özkan, Y., Dondurmacı, G., “Windows 95 için Access 7.0”, Alfa Yayınevi, 1996
- [3] Özkan, Y., Dondurmacı, G., “Access 97”, Alfa Yayınevi, 1997.
- [4] Özkan, Y., Dondurmacı, G., “Oracle Power Objects 2.0”, Sistem, 1997.
- [5] Özkan, Y., Dondurmacı, G., “Access 2000”, Alfa Yayınevi, 2000.
- [6] Özkan, Y., Dondurmacı, G., “Excel 2000”, Alfa Yayınevi, 2000.
- [7] Özkan, Y., Dondurmacı, G., “Javascript ile Web Programlama”, Alfa Yayınevi, 2001.

- [8] Özkan, Y., Dondurmacı, G., “Uygulamalı Excel 2002”, Alfa Yayınevi,2002.
- [9] Özkan, Y., Dondurmacı, G., “Uygulamalı Excel 2002 Fonksiyonları”, Alfa Yayınevi, 2002.
- [10] Dondurmacı G, “BİLGİSAYAR, Ahmet Yesevi Üniversitesi Hazırlık sınıfları Ders Kitabı, Ahmet Yesevi Üniversitesi Yayınları.(Kazakça), (2002)
- [11] Dondurmacı G, Visual Basic ile programlama, Alfa Yayınevi, (2003)
- [12] Dondurmacı G, ASP.Net ile Web programlama, Alfa Yayınevi, (2004)

İdari Görevler:

- Uzaktan Eğitim Merkezi Koordinatör Yardımcısı(Zirve Üniversitesi 2010-)
- Bilgisayar Programcılığı Bölüm Başkanı (Ahmet Yesevi Üniversitesi 2002-2010)
- Bilgisayar Programcılığı Bölümü Akademik Değerlendirme Kurul Üyesi (Ahmet Yesevi Üniversitesi 2002-2010)
- Uzaktan Eğitim Başkan Yardımcısı (Ahmet Yesevi Üniversitesi 2005-2009)

Bilimsel ve Mesleki Kuruluşlara Üyelikler:

- Türkiye Bilişim Derneği
- İnternet Kullanıcıları Derneği

Verdiği Dersler:

- Bilgisayar Kullanımına Giriş
- Web Programlama I (HTML, Java Script)
- Görsel Programlama (Visual Basic)
- Web Programlama II (ASP .Net)
- Veri Tabanı Yönetim Sistemleri
- İleri Veri Tabanı Sistemleri
- İşletmelerde Karar Destek Sistemleri
- Bilişim Sistemleri Semineri
- Uygulama