

**TC
YILDIZ TEKNİK ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANA BİLİM DALI
İŞLETME YÖNETİMİ YÜKSEK LİSANS PROGRAMI
YÜKSEK LİSANS TEZİ**

**İSTATİKSEL ANALİZ YÖNTEMİ OLARAK
VERİ MADENCİLİĞİNDE CHAID ALGORİTMASI VE
TÜRKİYE’DE İŞGÜCÜ PİYASASININ DURUMUNUN
VE BUNUN NEDENLERİNİN BELİRLENMESİNE
İLİŞKİN BİR UYGULAMA**

**EBRU YILMAZ
09713023**

**TEZ DANIŞMANI
Yrd. Doç. Dr. ESRA ÖZTAMAN**

**İSTANBUL
2012**

**TC
YILDIZ TEKNİK ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANA BİLİM DALI
İŞLETME YÖNETİMİ YÜKSEK LİSANS PROGRAMI
YÜKSEK LİSANS TEZİ**

**İSTATİKSEL ANALİZ YÖNTEMİ OLARAK
VERİ MADENCİLİĞİNDE CHAID ALGORİTMASI VE
TÜRKİYE’DE İŞGÜCÜ PİYASASININ DURUMUNUN
VE BUNUN NEDENLERİNİN BELİRLENMESİNE
İLİŞKİN BİR UYGULAMA**

**EBRU YILMAZ
09713023**

**TEZ DANIŞMANI
Yrd. Doç. Dr. ESRA ÖZTAMAN**

**İSTANBUL
2012**

TC
YILDIZ TEKNİK ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANA BİLİM DALI
İŞLETME YÖNETİMİ YÜKSEK LİSANS PROGRAMI
YÜKSEK LİSANS TEZİ

**İSTATİKSEL ANALİZ YÖNTEMİ OLARAK
VERİ MADENCİLİĞİNDE CHAID ALGORİTMASI VE
TÜRKİYE'DE İŞGÜCÜ PİYASASININ DURUMUNUN
VE BUNUN NEDENLERİNİN BELİRLENMESİNE
İLİŞKİN BİR UYGULAMA**

**EBRU YILMAZ
09713023**

Tezin Enstitüye Verildiği Tarih: 18.09.2012
Tezin Savunulduğu Tarih: 22.10.2012

Tez Oy birliği / ~~Oy çokluğu~~ ile başarılı bulunmuştur.

Unvan Ad Soyad
Tez Danışmanı : Yrd. Doç. Dr. Esra ÖZTAMAN
Jüri Üyeleri : Yrd. Doç. Dr. Arzu Karaman Akgül
Yrd. Doç. Dr. Serdar Bozkurt

İmza
Öztan
Karaman
Bozkurt

**İSTANBUL
2012**

ÖZ

İSTATİKSEL ANALİZ YÖNTEMİ OLARAK VERİ MADENCİLİĞİNDE CHAID ALGORİTMASI VE TÜRKİYE’DE İŞGÜCÜ PİYASASININ DURUMUNUN VE BUNUN NEDENLERİNİN BELİRLENMESİNE İLİŞKİN BİR UYGULAMA

Ebru YILMAZ
Eylül, 2012

Bu tez çalışmasının temel amacı, Türkiye’de işgücü piyasası durumunun 2009 ve 2010 yılları için, analiz edilerek ortaya konması ve analiz sonuçlarına bakılarak bu durumun nedenlerinin ve neler yapılabileceğinin değerlendirilmesidir. Bu amaçla yöntem olarak, büyük veri yığınlarında güncel bir karar destek analiz yöntemi olan veri madenciliği analizi kullanılmıştır. Çalışma kapsamında, TÜİK’ten temin edilen 2009 ve 2010 yılları Hanehalkı İşgücü Anketi verilerine ayrı ayrı Veri Madenciliği Analizi uygulanmış ve sonuçlar özetlenmiştir. Bu çalışma kapsamında, 15 yaş ve üstü olan grup yani aktif nüfus analize dahil edilmiştir. Analizde hedef değişken olarak “İşgücü Durumu”; bağımsız değişkenler ise “cinsiyet, referans yakınlık durumu, son bitirilen okul, öğrenim devam etme durumu, medeni durum, bir yıl önceki işgücü durumu, NUTS1 (İstatistiki Bölge Birimleri Sınıflaması) ve Kır-Kent” olarak belirlenmiştir. SPSS Clementine 11.1 paket programı ile CHAID algoritması kullanılarak veriler analiz edilmiştir. Yapılan analiz sonuçları değerlendirilmiş, çalışma kapsamına alınan 2009 yılı 366.273 ve 2010 yılı için 384.846 kişinin bir önceki yıla göre takip eden yılda, işgücüne katılım oranının yükselmiş, işsizlik değerinin ise azalmış olduğu belirlenmiştir. İstidamda olma oranı ülkemizde çok düşük olup 2009 yılında ancak % 39,95 iken 2010 yılında % 41,4 olarak bulunmuştur. İşgücü durumunu belirlemede en etkili değişken, bir yıl önceki iş gücü durumudur. İşgücüne katılım oranının en yüksek olduğu bölge Marmara, Batı Anadolu, Batı Karadeniz iken, işsizliğin en yüksek olduğu bölge İstanbul ve Doğu Anadolu’ dur. İşgücü piyasası açısından dezavantajlı gruplar kadınlar, özürllüler veya hastalar olurken, cinsiyet değişkeni açısından kadınların işgücüne katılma oranının, ülke bazında erkeklerin ancak 1/3’ü kadar olduğu görülmüştür. İşgücüne katılma oranının düşük olmasının temelinde, özellikle köyden kente göç olgusu ve eğitim düzeyinin etkili olduğu sonucuna varılmıştır.

Anahtar Kelimeler: Veri Madenciliği Analizi, CHAID, İşgücü Piyasası, İşsizlik.

ABSTRACT

THE CHAID ALGORITHM IN DATA MINING AS A STATISTICAL ANALYSIS METHOD AND AN APPLICATION INCLUDING DETERMINATION TO THE SITUATION OF LABOUR MARKET AND ITS REASONS IN TURKEY

Ebru YILMAZ
September, 2012

The main purpose of this study is to determine to the situation of labour market in Turkey, both in 2009 and 2010, and to evaluate the reasons and what to do by considering analysis results. For this purpose, data mining, which is used in large data sets as a current decision support analysis method is used as the application method. In the scope of of this study, Data Mining Analysis is applied separately to the Household Labour Survey sets for 2009 and 2010 years, which are provided from TÜİK. 15 years and over, in other words active population is included to the analysis. “Labour Status”, as target (dependent) variable; “gender, proximity to the reference state, the last completed school, continuing education, marital status, labour status a year earlier, NUTS1 (Nomenclature of Territorial Units for Statistics) and rural-urban area is identified as independent variables. These data are analysed as using CHAID algorithm by SPSS Clementine program version 11.1. Finally, results of the analysis are evaluated and it’s found that the labour participation rate increased and unemployment value is reduced for 366.273 persons in 2009 and 384.846 persons in 2010, in consecutive year over the previous year. The employment rate is found as very low in our country while only % 39,95 in 2009 and % 41,4 in 2010. The most effective variable to explain the labour status is the labour status a year earlier. Labour force participation rate is discovered as highest in the region of Marmara, Western Anatolia, the Western Black Sea, while Istanbul and the Eastern Anatolia region have the highest unemployment rate. Disadvantaged groups in the labour market appear women, disabled persons or patients; besides, in terms of gender, labour force participation rate of women in the whole country is detected only a third of per men. Finally, it is concluded that especially the rural-urban migration and level of education is very effective on the basis of being a low labour participation rate.

Keywords: Data Mining Analysis, CHAID, Labour Situation, Unemployment.

ÖNSÖZ

Tez çalışmam boyunca değerli katkı ve desteği ile beni yönlendiren tez danışmanım Sayın Yrd. Doç. Dr. Esra ÖZTAMAN'a teşekkürü bir borç bilirim.

Çalışmamın uygulama bölümünde bana sağladığı katkılarından dolayı Sayın Arş. Gör. Ömer BİLEN'e şükranlarımı sunarım.

Hayatımın her anında olduğu gibi tez çalışması boyunca da destek ve anlayışını benden esirgemeyen canım babam, annem ve ablam Dr. Funda YILMAZ'a; yüksek lisans öğreniminde çok şey paylaştığım arkadaşım Nur HÜZMELİ'ye minnettarım.

İstanbul; Eylül 2012

Ebru YILMAZ

İÇİNDEKİLER

	Sayfa No.
TEZ ONAY SAYFASI	
ÖZ.....	iii
ABSTRACT	iv
ÖNSÖZ	v
İÇİNDEKİLER	vi
TABLO LİSTESİ.....	ix
ŞEKİL LİSTESİ.....	x
KISALTMALAR	xii
1. GİRİŞ	1
2. VERİ MADENCİLİĞİ	5
2.1. Veri Madenciliği	5
2.2. Veri Madenciliği ve İş Zekası	8
2.3. Veri Madenciliği Kullanım Nedeni	9
2.4. Veri Madenciliği ve Diğer Disiplinler.....	10
2.5. Veri Madenciliği ve Veri Ambarı Arasındaki Fark	11
2.6. Veri Madenciliğini Etkileyen Etmenler	12
2.7. Veri Madenciliğindeki Problemler	13
2.8. Veri Madenciliği Kullanım Alanları ve Uygulamalar	14
2.9. Veri Madenciliğinde Kullanılan Paket Programlar	15
3. VERİ TABANINDA BİLGİ KEŞFİ SÜRECİNDE VERİ MADENCİLİĞİ. 17	
3.1. Araştırma Probleminin Tanımlanması	22
3.2. Veri Madenciliği Veritabanının Kurulması.....	22
3.2.1. Veri Kaynaklarının Belirlenmesi	22
3.2.1.1. Veri tabanı Sistemleri	23
3.2.1.2. OLAP ve Veri Ambarları	23
3.2.2. Veri Tanımlama (Veri Anlama Aşaması)	24
3.3. Veri Seçimi	25
3.4. Veri Kalitesini İyileştirme Ve Ön Hazırlık Süreçleri.....	26
3.5. Model Oluşturma	31
3.5.1. Veri Madenciliği Modelleri.....	31
3.5.2. Sınıflandırma (Classification)	33
3.5.2.1. Karar Ağaçları (Decision Trees).....	34

3.5.2.1.1. Karar Ağacı Oluşturmak Dağınıklık Kriterleri	46
3.5.2.2. Yapay Sinir Ağları (Artificial Neural Networks).....	48
3.5.2.3. Genetik Algoritmalar (Genetic Algorithms)	49
3.5.2.4. K-En Yakın Komşu (K-Nearest Neighbor).....	49
3.5.2.5. Bellek Temelli Nedenleme (Memory Based Reasoning)	50
3.5.2.6. Bayes Sınıflandırması	51
3.5.2.7. Naive Bayes	52
3.5.2.8. Zaman Serisi Analizi.....	53
3.5.3. Regresyon.....	53
3.5.3.1. Doğrusal Regresyon.....	54
3.5.3.2. Lojistik Regresyon	55
3.5.4. Kümeleme	56
3.5.4.1. Bölme Yöntemleri (Partitioning Methods)	59
3.5.4.2. Hiyerarşik Yöntemler (Hierarchical Methods)	60
3.5.4.3. Yoğunluk Tabanlı Yöntemler	60
3.5.4.4. Izgara Tabanlı Yöntemler	61
3.5.4.5. Model Tabanlı Yöntemler.....	61
3.5.5. Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler	62
3.5.5.1. Birliktelik Kuralları.....	62
3.5.5.2. Ardışık Zamanlı Örüntüler.....	65
3.6. Model İçi Değerlendirme Süreci	65
3.6.1. Basit Geçerlilik.....	66
3.6.2. Çapraz Geçerlilik.....	67
3.6.3. Bootstrap	68
3.7. Modelin Değerlendirilmesi ve Raporlama	69
3.7.1. Risk Matrisi	70
3.7.2. Kaldıraç Grafiği.....	71
3.7.3. Alıcı Çalışma Karakteristik Grafiği (ROC)	72
3.8. Modelin Yönetimi	73

4. UYGULAMA..... 74

4.1. Araştırmanın Problemi	74
4.2. Araştırmanın Amacı ve Önemi	75
4.3. Araştırmanın Sınırlılıkları	76
4.4. Araştırmanın Varsayımları.....	76
4.5. Bilgi Ve Veri Toplama Yöntemi.....	76
4.6. Veri Analiz Tekniği.....	78
4.6.1. Veri Ön İşlem Süreci.....	78
4.6.2. Değişkenlerin Açıklanması	79
4.6.3. Modelin Kurulması ve Değerlendirilmesi	88
4.6.4. Algoritmanın Uygulanması	92
4.6.4.1. 2009 YILI.....	92
4.6.4.2. 2010 YILI.....	94
4.6.5. 2009 YILI Ana Dalları	97
4.6.5.1. Karar Ağacı Birinci Anadalı	97
4.6.5.2. Karar Ağacı İkinci Anadalı	99
4.6.5.3. Karar Ağacı Üçüncü Anadalı	100

4.6.5.4. Karar Ağacı Dördüncü Anadalı	102
4.6.5.5. Karar Ağacı Beşinci Anadalı	104
4.6.5.6. Karar Ağacı Altıncı Anadalı	109
4.6.5.7. Karar Ağacı Yedinci Anadalı	112
4.6.5.8. Karar Ağacı Sekizinci Anadalı	114
4.6.6. 2010 YILI Ana Dalları	115
4.6.6.1. Karar Ağacı Birinci Anadalı	115
4.6.6.2. Karar Ağacı İkinci Anadalı	117
4.6.6.3. Karar Ağacı Üçüncü Anadalı	118
4.6.6.4. Karar Ağacı Dördüncü Anadalı	120
4.6.6.5. Karar Ağacı Beşinci Anadalı	122
4.6.6.6. Karar Ağacı Altıncı Anadalı	125
4.6.6.7. Karar Ağacı Yedinci Anadalı	128
4.6.6.8. Karar Ağacı Sekizinci Anadalı	129
4.7 Analiz Genel Değerlendirmesi	130
5. SONUÇ	133
KAYNAKÇA	140
EKLER	147
Ek 1. 2009 Yılı İşgücü Durumu İçin Elde Edilen Karar Ağacı Görünümü .	147
Ek 2. 2010 Yılı İşgücü Durumu İçin Elde Edilen Karar Ağacı Görünümü .	148
ÖZGEÇMİŞ.....	149

TABLO LİSTESİ

Tablo 2.1: Veri Madenciliği Tarihçesi	6
Tablo 3.1: Sınıflandırma Matrisi ve CHAID Tahmin Edici Modeli	43
Tablo 3.2: Optimum K ve Optimum Uzaklıklar	50
Tablo 3.3: ROC Matrisi.....	72
Tablo 4.1: 2009 Yılı ve 2010 Yılı Cinsiyet Frekans Tablosu	81
Tablo 4.2: 2009 Yılı ve 2010 Yılı Eğitim Durumu Frekans Tablosu	82
Tablo 4.3: 2009 Yılı ve 2010 Yılı Eğitime Devam Etme Durumu Frekans Tablosu	83
Tablo 4.4: 2009 Yılı Ve 2010 Yılı Medeni Durum Frekans Tablosu	83
Tablo 4.5: 2009 Yılı Ve 2010 Yılı Referans Yakınlık Durumu Frekans Tablosu	84
Tablo 4.6: 2009 Yılı ve 2010 Yılı İstatistiki Bölge Birimleri Frekans Tablosu	85
Tablo 4.7: 2009 Yılı ve 2010 Yılı Kır-Kent Frekans Tablosu	86
Tablo 4.8: 2009 Yılı ve 2010 Yılı Bir Yıl Önceki İşgücü Durumu Frekans Tablosu	86
Tablo 4.9: 2009 Yılı Ve 2010 Yılı İşgücü Durumu Frekans Tablosu.....	87
Tablo 4.10: 2009 Yılı İşgücü Göstergesi Özet Tablosu	87
Tablo 4.11: 2010 Yılı İşgücü Göstergesi Özet Tablosu	88

ŞEKİL LİSTESİ

Şekil 2.1: Veri Madenciliği ve İş Zekası	8
Şekil 2.2: Veri Madenciliği İlgi Alanları	10
Şekil 2.3: Veri Ambarı Bileşenleri ve Veri Madenciliği İlişkisi	11
Şekil 3.1: Veri Tabanlarında Bilgi Keşfi Süreci	18
Şekil 3.2: Veriden Bilgi Keşfi	19
Şekil 3.3: Basit Bir Veri Madenciliği Süreci	20
Şekil 3.4: CRISP – DM Veri Madenciliği Uygulama Süreci	21
Şekil 3.5: Veri Küpü Örneği	24
Şekil 3.6: Veri Ön Hazırlık	30
Şekil 3.7: Veri Madenciliği Modelleri	32
Şekil 3.8: Karar Ağacı Genel Yapısı	35
Şekil 3.9: Karar Ağacı Örneği	36
Şekil 3.10: Bonferroni Yaklaşımı Ortalama Vektörü	39
Şekil 3.11: Doğrudan Mail Tanıtımına Cevap Verenlerin CHAID Analizi Çıktısı ..	42
Şekil 3.12: Çok Katmanlı İleri Yayımlı Bir Yapay Sinir Ağı Örneği	48
Şekil 3.13: Doğrusal Regresyon	54
Şekil 3.14: Veri Birimleri ve Kümeleme	57
Şekil 3.15: Bir Simülasyon Çalışmasında Bölme Yöntemi Aşamaları	59
Şekil 3.16: Basit Geçerlilik Ölçümünde Veri Organizasyonu	67
Şekil 3.17: Risk Matrisi Örneği	71
Şekil 3.18: Kaldıraç Grafiği	71
Şekil 4.1: İş Gücü Durumu İçin Veri Kaynakları	79
Şekil 4.2: Uygulamada Kullanılan Hedef ve Açıklayıcı Değişkenler	81
Şekil 4.3: İşgücü Durumu Veri Setinin Eğitim Ve Test Olarak Gruplandırılması	89
Şekil 4.4: 2009 Yılı Sınıflandırma Algoritmalarının Doğruluk Yüzdeleri	90
Şekil 4.5: 2010 Yılı Sınıflandırma Algoritmalarının Doğruluk Yüzdeleri	90
Şekil 4.6: 2009 Yılı Ve 2010 Yılı CHAID Algoritması Modelin Test Yüzdeleri	91
Şekil 4.7: CHAID Algoritması Sonucu Elde Edilen Karar Ağacının Başlangıç Nodu 2009	92
Şekil 4.8: Elde Edilen 8 Alt Sınıfın Gösterimi 2009 Yılı	93
Şekil 4.9: Elde Edilen 8 Alt Sınıfın Gösterimi 2009 Yılı (Devam)	94
Şekil 4.10: CHAID Algoritması Sonucu Elde Edilen Karar Ağacının Başlangıç Nodu 2010	94
Şekil 4.11: Elde Edilen 8 Alt Sınıfın Gösterimi 2010 Yılı	95
Şekil 4.12: Elde Edilen 8 Alt Sınıfın Gösterimi 2010 Yılı (Devam)	96
Şekil 4.13: 2009 Karar Ağacı Birinci Anadalı	97
Şekil 4.14: 2009 Karar Ağacı İkinci Anadalı	99
Şekil 4.15: 2009 Karar Ağacı Üçüncü Anadalı	100
Şekil 4.16: 2009 Karar Ağacı Dördüncü Anadalı	102
Şekil 4.17: 2009 Karar Ağacı Beşinci Anadalı	104
Şekil 4.18: 2009 Karar Ağacı Altıncı Anadalı	109
Şekil 4.19: 2009 Karar Ağacı Yedinci Anadalı	112

Şekil 4.20: 2009 Karar Ağacı Sekizinci Anadalı.....	114
Şekil 4.21: 2010 Karar Ağacı Birinci Anadalı.....	115
Şekil 4.22: 2010 Karar Ağacı İkinci Anadalı	117
Şekil 4.23: 2010 Karar Ağacı Üçüncü Anadalı	118
Şekil 4.24: 2010 Karar Ağacı Dördüncü Anadalı.....	120
Şekil 4.25: 2010 Karar Ağacı Beşinci Anadalı.....	122
Şekil 4.26: 2010 Karar Ağacı Altıncı Anadalı.....	125
Şekil 4.27: 2010 Karar Ağacı Yedinci Anadalı	128
Şekil 4.28: 2010 Karar Ağacı Sekizinci Anadalı.....	129

KISALTMALAR

BMKP	: Birleşmiş Milletler Kalkınma Programı
BT	: Bilgi Teknolojileri
BTN	: Bellek Temelli Nedenleme
C&RT	: Sınıflandırma ve Regresyon Ağacı
CHAID	: Otomatik Ki-Kare Etkileşim Belirleme Analizi
CRISP-DM	: Cross Industry Process for Data Mining
DB	: Data Base
DM	: Data Mining
EUROSTAT	: Avrupa Birliği İstatistik Ofisi
KNN	: K-En Yakın Komşu Algoritmesi
NUTS1	: İstatistiksel Amaçlı Bölge Sınıflaması Düzey-1
NUTS2	: İstatistiksel Amaçlı Bölge Sınıflaması Düzey-2
OLAP	: Çevrimiçi Analitik İşleme
ROC	: Alıcı Çalışma Karakteristik Grafiği
SPSS	: Statistical Package For The Social Sciences
TÜİK	: Türkiye İstatistik Kurumu
VA	: Veri Ambarları
VM	: Veri Madenciliği
VTBK	: Veri Tabanlarında Bilgi Keşfi

1. GİRİŞ

İşsizlik sorunu, bugün gelişmiş ve gelişmekte olan tüm toplumların ortak ve en önemli sorunlarından biridir. Gelişmişlik farkı gözetmeksizin hemen her ülke, işsizlik ve neden olduğu sosyal sorunlarla artık daha sık karşılaşmaktadır. Türkiye gelişmekte olan bir ülke olup; gelişmişlik yolundaki uyum sürecinde en sorunlu alanlardan biri kuşkusuz istihdam konusudur. Nüfusun sürekli artması, eksik istihdam oranının yükselişi ve işgücünün etkin bir şekilde kullanılamaması gibi temel nedenlerden dolayı işgücü piyasasında istenen seviye yakalanamamıştır. Ülkemizde işsizlik sorunu ciddiyetini korumasına rağmen, bunun sebepleri tam olarak anlaşılabilmiş değildir.

İşsizlik aslında, modern çağın ekonomik ve sosyal bir olgusu olarak gelişmiş ve gelişmekte olan tüm ekonomiler için önemli bir sorun kaynağı konumundadır. İşsizlik bireylerin ekonomik olarak refah seviyelerini etkilerken, aynı zamanda ciddi sosyal ve psikolojik etkiler yaratan bir olgudur. Ayrıca işsizlik birey açısından olduğu kadar toplum açısından da önemli bir potansiyel ekonomik kayıp yaratmaktadır. Nihayet, işsiz kitleler, toplumu huzursuz edebileceği gibi, politik açıdan da demokratik rejimleri tehdit etme potansiyeline sahiptirler.

İşsizliğin karar alıcılar arasında yakın zamana kadar genellikle sadece bir büyüme sorunu olarak algılanmış olması, bu probleme karşı kapsamlı ve bütüncül bir strateji geliştirilmesini önlemiştir. Bu noktadan hareketle, istikrarlı ekonomik büyüme performansının ülkemizde işsizliğin gerilemesine yeterince katkı sağlayamamış olması, gerçekte işgücü talebinin beklenildiği ölçüde artmamış olması ve işsizliği belirleyen etkenler gibi makro ve bireysel sebeplerin mikro düzeyde tekrar ele alınması ihtiyacı, bu çalışmaya başlanmasının temellerini oluşturmaktadır.

Son yıllarda konunun ciddiyetinin devam ettiği görülmektedir. Bu bağlamda literatürde yer alan çalışmalar incelenmiş; ancak Türkiye’de istihdam piyasasını ve işgücü profilini belirlemeye yönelik gerek tez çalışmaları gerek akademik makaleler mevcut olsa da, bu konudaki etkenlerin her yıl değişmesinden dolayı güncel olan

bulguya rastlanmamıştır. Buradan yola çıkarak, işgücü durumunun özellikle son yaşanan 2008 krizi sonrası 2009 ve 2010 yıllarındaki seyrinin ortaya konması; istihdam, işsizlik ve işgücüne katılmama oranları belirlenerek güncel istihdam piyasasına ilişkin sebep ve sonuç değerlendirmesinin yapılması ve bu değerlendirmenin gelecekte işgücü durumu ile ilgili kararlar alınmasında yetkililere ışık tutması amaçlanmıştır. Dolayısı ile bu çalışmanın ana hipotezi, Türkiye’de işgücü durumunun 2009 ve 2010 yılları kıyaslaması yapılarak bulunacak farklılık ve benzerliklerine göre son iki yıldaki seyrinin ortaya konması, istihdamda olanlara ve işsizlere ilişkin değerlendirmenin yapılması olarak belirlenmiştir. Bu amacın gerçekleştirilebilmesi için, son yıllarda yaygınlaşan ve büyük hacimli veri yığınları içerisinden karar alabilmek için potansiyel olarak faydalı olabilecek, uygulanabilir ve anlamlı bilgilerin çıkarılmasına yardımcı olan bir metodoloji olan “Veri Madenciliği” analiz yöntemi olarak kullanılmıştır. Çalışmanın, akademik olarak bu alanda ayrıntılı araştırma yapmak isteyenler için bir yol gösterici niteliğinde olması amaçlanmıştır.

Tez çalışması, konunun önemi ve açıklanan amaçlar kapsamında giriş ve sonuç bölümleri dışında 3 ayrı ana bölümden oluşmaktadır.

Tezin birinci bölümünde, araştırmanın uygulama kısmının analiz yöntemi olarak seçilen Veri Madenciliğinin teorik bilgilendirilmesi ve tarihsel gelişimi anlatılmıştır. Organizasyonların veri ambarlarındaki çok önemli bilgilere yoğunlaşmasına yardımcı olan, çok büyük veri tabanlarında saklı, akıllı bilgiyi ortaya çıkaran kullanıcı dostu yeni bir teknoloji olması ve karar vermenin her geçen gün daha da zorlaştığı rekabet ortamında, analizcilerin karşılaştıkları büyük veri yığınları ile başa çıkmalarındaki en başarılı yöntemlerden biri olması nedeni ile Veri Madenciliği analiz metodu olarak kullanılmıştır. Ayrıca, veri madenciliğinin çıkış noktası olan İş Zekası ile bağlantısı anlatılmıştır. Veri madenciliğinin, veriyi toplamak ve yönetmekten çok, analiz yapmayı ve tahmini içermesinden dolayı veri ambarı ile veri madenciliği arasındaki ayırmadan kısaca bahsedilmiştir. Bu bölümün ilerleyen safhalarında analizin kullanıcı açısından sunduğu ve dikkate alınması gereken problemler temel düzeyde incelenmiştir. Veri madenciliğinin öneminin daha iyi anlaşılması için kullanım nedenleri ve kullanıldığı alanlar detaylı bir şekilde incelenmiştir. Bu bölümde son olarak, bugün veri madenciliğini bilgisayar ortamında kullanma olanağı sunan yaygın ve güncel paket programlar anlatılmıştır.

İkinci bölümde; uygulamanın doğru ve tutarlı yapılması bakımından anlamlı olması için uygulamada takip edilen sırasıyla Veri Madenciliği süreci temel ve alt adımları teorik olarak detaylı bir şekilde ele alınmıştır. Burada, Veri Madenciliği sürecinden bahsetmeden önce aktarılması gereken temel bir kavram olarak Veri Tabanlarında Bilgi Keşfi (VTBK) ele alınmıştır, çünkü veri madenciliği, veriden bilgiye doğru keşif sürecinde en önemli kesimi oluşturmaktadır. Farklı akademik kaynaklarda veri madenciliği sürecinin farklı aşamaları görülse de genelde en çok kabul edileni, detaylı süreç tanımı uluslararası düzeyde standart olarak kabul edilmiş, CRISP-DM (Cross Industry Process for Data Mining) ile yapılmaktadır. Tezin bu bölümünün ana başlıkları genel olarak bu sırayla devam etmektedir. Bu bağlamda; sırasıyla araştırma problemin tanımlanması, veri madenciliği veri tabanının kurulması ve veri kaynaklarının belirlenerek verinin seçiminin nasıl yapıldığı anlatılmıştır. Bundan sonraki aşamada, çok sayıda veri ile uğraşan veri madenciliğinin, bu verilerde kaliteyi sağlama ve bunun için gerekli olan ön hazırlık süreçleri ele alınmıştır. Bu ön süreçler; veri temizleme, veri bütünleştirme, veri dönüşümü ve veri azaltma alt başlıkları olup gerek bu alanda uygulama yapmak isteyenlere gerek teorik olarak bilgilenmek isteyenlere yardımcı olabilecek bilgiler şeklinde derlenmiştir. Bu bölümde, bir sonraki ana başlık ve süreç adımı, modelin oluşturulmasıdır ve burada veri madenciliği modelleri detaylı bir şekilde ele alınmıştır. Veri madenciliği modelleri, tahmin edici (denetimli) ve tanımlayıcı (denetimsiz) modeller olmak üzere temelde ikiye ayrılırken bu modeller de kendi içlerinde alt kırılımlar göstermektedir. Tezin bir sonraki bölümü olan uygulamada, tahmin edici modellerden bir sınıflama modeli olan Karar Ağacı modeli uygulanmıştır. Bu nedenle, karar ağacının başlıca algoritmaları olan CHAID, C4.5, C&RT, ID3 ve QUEST kavramlar ele alınmıştır. Bu kavramlar içerisinde de CHAID algoritması, teorik olarak ayrıntılı incelenmiştir. Modelin oluşturulmasından sonraki adım önce model içi değerlendirme sonra da modelin bütünüyle değerlendirilmesi şeklindedir. Son olarak da modelin tamamlanması için raporlama ve modelin yönetimi adımları ele alınarak bu bölüm tamamlanmıştır.

Tez çalışmasının son bölümü olan uygulama kısmında ise analiz yönetimi olarak seçilen veri madenciliği yöntemi, çalışmanın hedef kitlesi yani ana kütlesi tüm ülke kapsamı ile birebir örtüştüğü için, örneklem olarak TÜİK'ten temin edilen 2009 ve 2010 yılları Hanehalkı İşgücü Anketi verilerine uygulanmıştır. Bu nedenle bu

bölümde sırasıyla araştırmanın problemine, amacına ve önemine ayrıntılı olarak yer verilmiştir. Bundan sonra ise araştırmanın sınırları çizilmiş, bu sınırlar dahilinde çalışmanın varsayımlarından da bahsedilmiştir. Veri toplama yöntemi olarak TÜİK Hanehalkı İşgücü Anketi ve verilerin nasıl düzenlendiği hakkında açıklayıcı bilgi ve tez çalışması için de gerekli olan bazı işgücü tanımları anlatılmıştır. Bu kullanılan veriler ön işlemlerden geçirildikten sonra analizde kullanılacak olan değişkenler belirlenip hedef (bağımlı) ve bağımsız değişkenler ayrımı yapılmıştır. Bu aşamadan sonra, SPSS Clementine 11.1 paket programı ile 2009 ve 2010 yılları için veri analiz edilmiştir. Sonrasında, *eğitim (training)* ve *test (testing)* olmak üzere partition işlemcisi ile iki örnek gruba ayrılmıştır. Böylece, doğruluk yüzdelerine bakılarak CHAID algoritmasının kullanılmasına karar verilmiştir. Hedef değişken “İşgücü Durumu” için uygulanan CHAID algoritması sonucu 2009 ve 2010 yılları için elde edilen karar ağaçları ayrı ayrı “bir sonraki açıklayıcı değişken, dallanma ve sınıf sayıları, başlangıç düğümleri ve hedef değişkenin seçeneklerine göre olasılık dağılımları” anlatılmıştır. Uygulama ile ilgili açıklamaların ardından, uygulanan analizin sonuçları şekiller ve frekans tabloları açıklamaları ile verilmiştir. Buradaki son aşama, bir anlamda uygulama kısmının özeti ya da modelin raporu olarak da belirtilebilecek analiz genel değerlendirmesi yapılmıştır. Tez çalışmasının alt hipotezleri olan 8 karar ağacı ana dalı yani bir önceki yıldaki işgücü durumuna göre iki yıl arasında fark olup olmadığı ve farklılıkların nasıl olduğu sayısal olarak 2009 ve 2010 yıllarını birbiri ile kıyaslayarak bu bölüm sonlandırılmıştır.

Sonuç bölümünde ise tez çalışmasında uygulama yöntemi olarak öğrenilmek istenen Veri Madenciliği analizinin önemi ve çalışma konusu ile bağlantısı belirtilmiş ve analiz verilerine yer verilmiştir. Bu verilerden yola çıkarak işsizlik sorunu ile ilgili alınabilecek kararlara ışık tutması açısından bu iki yılın analiz sonuçları birbiriyle detaylı olarak karşılaştırılmıştır. Her bir kıyaslama sonucu görülen benzerlik ya da farkın hedef değişkeni en iyi açıklayan bir yıl önceki işgücü durumu açısından sebepleri belirlenmeye çalışılmış bu noktada verilecek çözüm önerileri hem literatür taramaları hem devlet tarafından işgücü piyasasına ve işsizliğe ilişkin kaynaklar dahilinde ele alınmıştır. Tüm sonuç ve yorumlar, bu tez çalışmasından ileride faydalanabilecek tüm karar yetkililerine ve akademik çalışma yapmak isteyenlere güncel bir kaynak olması amacı ile yapılarak tamamlanmıştır.

2. VERİ MADENCİLİĞİ

2.1. Veri Madenciliği

İlk bilgisayarın hayatımıza girmesinden itibaren bilgisayarlar her geçen gün biraz daha gelişmiş ve değişmiştir. Kullanıcı gereksinimleri doğrultusunda değerli verilerin depolanması isteği ile veri tabanları ortaya çıkmıştır. Günümüzde hemen hemen bütün organizasyonlar bilgisayar teknolojisi ve internetin de etkisi ile kayıtlarını elektronik veri tabanlarına taşımışlardır. Örneğin, devlet kuruluşları ve özel işletmelerde oluşturulan veri tabanları kuruluşların faaliyetlerini sürdürmeleri için gerekli olan verileri depolamaktadırlar. Bu veri tabanlarının boyutları terabaytlarla ifade edilmektedir. Zamanla veri tabanlarında gizli, stratejik ve politik değere sahip olabilecek, işe yarar örüntüler ve şablonlar elde etmenin mümkün olacağı düşüncesi yayılmıştır¹.

İnsanoğlu geçmişten bugüne her zaman verileri yorumlayıp bilgi edinmeye çalışmıştır ve bunun için çeşitli donanımlar oluşturmuştur. Bu donanımlar bilginin taşınmasını sağlamıştır. Zamanla her alanda bilgi toplanmaya başlanmış ve kronolojik olarak gelişimi aşağıdaki tabloda özetlenmiştir.

¹ Süreyya Çankırı ve diğ., “Organizasyonlarda Bilgi Yönetimi Sürecinde Veri Madenciliği Yaklaşımı”, **Bilgi Çağında Varoluş: “Fırsatlar ve Tehditler” Sempozyumu 01-02 Ekim 2009** (İstanbul: Yeditepe Üniversitesi,2011): 149.

Tablo 2.1: Veri Madenciliği Tarihçesi

Gelişim Adımları	Cevaplanan Karar Problemi	Kullanılabilen Teknolojiler	Ürün Sağlayıcıları	Karakteristikler
Veri Toplama (1960'lar)	"Benim toplam karım geçen 5 yılda ne kadardı?"	Bilgisayarlar, Teypler, Diskler	IBM, CDC	Geriye dönük , statik veri dağıtımı
Veri Erişimi (1980'ler)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	İlişkisel Veritabanları, SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Kayıt düzeyinde geriye dönük, dinamik veri dağıtımı
Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	OLAP, Çok Boyutlu Veritabanı Sistemleri, Veri ambarları	Pilot, Comshare, Arbor, Cognos, Microstrategy	Çoklu düzeylerde, geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	"Gelecek ay Boston'daki birim satışlar muhtemelen ne olabilir, niçin?"	İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veritabanları	Pilot, Lockheed, IBM, SGI, SPSS, SAS, Microsoft vs.	Geleceğe dönük ,proaktif enformasyon dağıtımı

"Veri Madenciliğine Giriş", İnternet: Hacettepe Üniversitesi, <http://yunus.hacettepe.edu.tr/~hcingi/ist376a/6Bolum.doc> [02.11.2011], 1.

Veri Madenciliği, 60'lı yıllarda veri analizi sorunlarının bilgisayar ile çözülmeye başlanması ile ortaya çıkmış olup, "Veri Madenciliği" ismi 90'lı yıllarda bilgisayar mühendisleri tarafından ortaya atılmıştır².

Veri madenciliği büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak ilişki ve kuralların aranmasıdır³.

Veri madenciliği, büyük hacimli veri yığınları içerisinde karar alabilmek için potansiyel olarak faydalı olabilecek, uygulanabilir ve anlamlı bilgilerin çıkarılmasına verilen addır. Veri madenciliği geniş anlamda veri analiz teknikleri bütünüdür ve tek başına bir çözüm değildir. Mevcut problemleri çözmek, kritik kararları almak veya geleceğe yönelik tahminleri yapmak için gerekli olan bilgileri elde etmeye yarayan bir araçtır. Ortaya çıkarılması hedeflenen bilgiler; üstü kapalı, çok net olmayan, önceden bilinmeyen, daha önce keşfedilmemiş ancak potansiyel olarak kullanışlı anlamlı ve kritik bilgilerdir. Yani Veri Madenciliğini veri dağları altındaki elmasları, altınları, külçeleri özel yazılımlar ile keşfetmek olarak tanımlamak mümkündür⁴.

² Age, 149.

³ Yılmaz Argüden, Burak Erşahin, **VERİ MADENCİLİĞİ Veriden Bilgiye, Masraftan Değere**, 1.bs. (İstanbul: ARGE Danışmanlık Yayınları, 2008), 9.

⁴ Age, 15.

Veri Madenciliği ile ilgili literatürde geçen diğer bazı tanımlamalar şöyledir:

Veri tabanı, yapay zeka ve istatistik dünyasını birbirine bağlayan son zamanlarda çıkmış bir alandır⁵.

Veri madenciliği büyük hacimli verilerden öz bilginin çıkarılması sürecidir. Bir başka ifade ile veri madenciliği, büyük ve karmaşık verilerde beklenmeyen patikaların, değerli yapıların ve ilginç ilişkilerin keşfedilmesi bilimidir⁶.

VM, birleşik verilerdeki gizli bilgileri bulmak ve iş uzmanlığını arttırmak amacıyla yapılan yeni bir karar destek analiz işlemidir⁷.

Veri madenciliği, veri tabanı kullanıcıları için veri tabanlarındaki beklenmeyen ilginç, değerli ilişkilerin bulunmasını hedefleyen süreç olarak tanımlanır. Bu yüzden veri madenciliği, veri tabanından anlamlı örüntüler veya kurallar elde etmek için geniş bir araştırma alanı olarak görülmektedir⁸.

Veri Madenciliği, geniş veri gruplarındaki bilinmeyen, geçerli desenler ve ilişkileri keşfetmek için gereken, geliştirilmiş veri analiz araçlarının kullanımını içeren bir analiz tekniğidir. Bu araçlar, istatistiksel metotları, matematiksel algoritmaları ve yapay öğrenme metotlarını (gözlem doğrultusunda performanslarını geliştiren algoritmalar gibi, örneğin sinir ağları veya karar ağaçları) içermektedir. Sonuç olarak, veri madenciliği, veriyi toplamak ve yönetmekten çok, analiz yapmayı ve tahmini içermektedir⁹.

Özetle, veri madenciliği veriden bilgi üreterek, ortalama kararlar yerine özgün kararlar verilmesini destekleyen, satışları, karlılığı, yenilikçiliği ve kaynak kullanımında etkinliği artıran önemli bir yönetim aracıdır. Organizasyonların veri

⁵ E Y. Lindell, R. Israel, ve B. Pinkas, “Privacy Preserving Data Mining”, **Journal of Cryptology** (2000), 1’den aktaran Süreyya Çankırı ve diğ., “Organizasyonlarda Bilgi Yönetimi Sürecinde Veri Madenciliği Yaklaşımı”, **Bilgi Çağında Varoluş: “Fırsatlar ve Tehditler” Sempozyumu 01-02 Ekim 2009** (İstanbul: Yeditepe Üniversitesi,2011): 149.

⁶ Selim Tüzüntürk, “VERİ MADENCİLİĞİ VE İSTATİSTİK”, **Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi**, c. 29. s. 1(2010) : 67.

⁷ Ali Sait Albayrak, Şebnem Koltan Yılmaz, “Veri Madenciliği: Karar Ağacı Algoritmaları Ve İMKB Verileri Üzerine Bir Uygulama”, **Süleyman Demirel Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Dergisi**, c.14, s.1 (2009):36.

⁸ **Data Mining: Statistics And More, The American Statistician**, 2.bs.(1998), s: 112-118’den Aktaran Derya Ay, İbrahim Çil, “Migros Türk A.Ş.’ de Birlikte Kurallarının Yerleşim Düzeni Planlamada Kullanılması”, **Endüstri Mühendisliği Dergisi**, c.21,s.2 (2010):15.

⁹ Jeffrey W. Seifert, “Data Mining: An Overview”, **CRS Report for Congress**, 16 Kasım 2004 (Federation of American Scientists, Congressional Research Service, 2004): 4.

ambarlarındaki çok önemli bilgilere yoğunlaşmasına yardımcı olan, çok büyük veri tabanlarında saklı, akıllı bilgiyi ortaya çıkaran yeni bir teknolojidir.

2.2. Veri Madenciliği ve İş Zekası

Veri Madenciliği tanımının ortaya atıldığı günden bu güne verilerden bir bakışta anlaşılmayan ancak işletme için kimi durumlarda hayati öneme sahip bilgilerin çıkartılmasına ilişkin değişik tanımlar ve algoritmalar geliştirilmiştir. Örneğin Karar Destek Sistemleri, İş Zekası gibi kavramları birbirinden kesin sınırlarla ayırmak pek mümkün olmamaktadır. Diğer terimlere göre yakın zamanda ortaya atılmış İş Zekası kavramı, Veri Madenciliğini kapsayan bir terim olarak kullanılmaktadır¹⁰.

Terimlerin anlamları üzerine ayrıntılı bilgi vermeden önce, genel bir görünümü şu şekilde ortaya koyabiliriz.



Şekil 2.1: Veri Madenciliği ve İş Zekası

E.M. Awad, H.M. Ghaziri, **Knowledge Management**. (Prentice Hall, 2004), 328'den aktaran Süreyya Çankırı ve diğ., "Organizasyonlarda Bilgi Yönetimi Sürecinde Veri Madenciliği Yaklaşımı", **Bilgi Çağında Varoluş: "Fırsatlar ve Tehditler" Sempozyumu 01-02 Ekim 2009** (İstanbul: Yeditepe Üniversitesi, 2011): 147-178.

¹⁰ Fatih Mehmet Güleç, "Kurumsal Verilerin Yapay Zekâ Modelleri İle İşlenmesi İçin Modelleme Aracı Alt Yapı Tasarım Ve Gerçekleştirimi" (Yüksek Lisans Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, 2007), 4.

İş zekası terimi, işte karar vermeyi destekleyen ve bilgi teknolojilerini temel alan bütün süreçler, teknikler ve araçlar için genel anlamda kullanılan bir ifadedir¹¹. Bilgi yönetimi konusundaki çalışmalar ve gelişmeler bugünün yeni kavramlarından birisi olan İş Zekası kavramının ortaya çıkmasına ve gelişmesine katkı sağlamıştır. Bilgi yönetimi ile veri madenciliğinin kesiştiği bir diğer nokta da burasıdır: Veri madenciliği, İş Zekası'nın bir bileşeni olarak kabul edilir¹².

2.3. Veri Madenciliği Kullanım Nedeni

Günümüzde bankacılık, sigorta ve borsa gibi birçok alandaki bütün bilgiler bilgisayar sistemleri sayesinde çeşitli veri tabanlarına kaydedilmektedir. Bütün bunlara bir de uydulardan elde edilen veriler katıldığından, ulaşılan veri hacminin inanılmaz boyutlara vardığı açık bir gerçektir. Veri madenciliği, eldeki yapısız veriden, anlamlı ve kullanışlı bilgiyi çıkarmaya yönelik çalışmalarının bütünü olmuştur. Sonuç olarak yıllar ilerledikçe ortaya çıkan veri yığınlarına bir düzen verme, başka bir anlamda potansiyel olarak kullanışlı bilgi haline getirme amacıyla veri madenciliği ortaya çıkmıştır¹³.

Veri madenciliği, özellikle kar ve pazar payı elde edebilmek için yoğun rekabetin yaşandığı pazarlama alanında ön plana çıkmaktadır. Hangi müşteri, hangi ürünü, ne zaman satın alabilir, kimler tedarikçilerinden vazgeçmekte ve bu tur müşterileri vazgeçirmek / geri kazanmak için neler yapılabilir, ürünün değerini yitirmesine hangi değişkenler neden olmaktadır, vb. soruların cevapları veri yığınlarının altındadır ve cevapları bulabilmek için veri madenciliği çözümleri gereklidir.

Veri Madenciliği ile şirketler önceden bilinmeyen bilgileri ortaya çıkararak karar verme süreçlerini iyileştirirler. Veri madenciliği teknikleri kullanarak; maliyetleri azaltmak, gelirleri artırmak, verimliliği artırmak, yeni fırsatları ortaya çıkarmak, yeni keşifler yapmak, emek yoğun faaliyetleri otomatikleştirmek, sahtekârlıkları belirlemek ve müşteri deneyimini geliştirmek mümkündür¹⁴.

¹¹ Ali Serhan Koyuncugil, Nermin Özgülbaş, “ Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları”, **Bilişim Teknolojileri Dergisi**, c.2,s.2,(Mayıs 2009): 25.

¹² E.M. Awad, H.M. Ghaziri, **Knowledge Management**. (Prentice Hall, 2004), 328'den aktaran Süreyya Çankırı ve diğ. , “Organizasyonlarda Bilgi Yönetimi Sürecinde Veri Madenciliği Yaklaşımı”, **Bilgi Çağında Varoluş: “Fırsatlar ve Tehditler” Sempozyumu 01-02 Ekim 2009** (İstanbul: Yeditepe Üniversitesi,2011): 164.

¹³ “Veri Madenciliğine Giriş”, İnternet: Hacettepe Üniversitesi, <http://yunus.hacettepe.edu.tr/~hcingi/ist376a/6Bolum.doc> [02.11.2011], 2.

¹⁴ Argüden, **age**, 17.

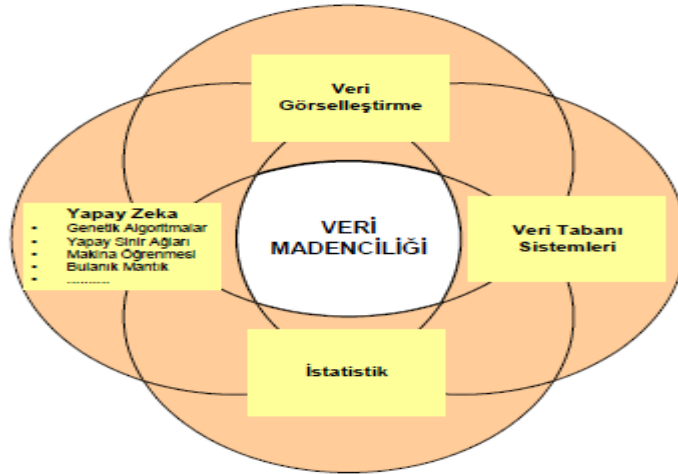
Özetle, veri madenciliği iki gereksinimden ortaya çıkmaktadır:

- Toplanan çok miktarda verinin işlenme ihtiyacı,
- Artan rekabette doğru karar verebilme yetkinliğini artırmak ihtiyacı.

2.4. Veri Madenciliği ve Diğer Disiplinler

Veri madenciliği çok disiplinli bir çalışma alanıdır. Veri tabanı teknolojisi, istatistik, yapay zeka, matematik, yönetim bilişim sistemleri veri madenciliğinde birlikte çalışmaktadır.

Veri madenciliğini istatistiksel bir yöntemler serisi olarak görmek mümkün olabilir. Ancak veri madenciliği, geleneksel istatistikten birkaç yönde farklılık gösterir. Veri madenciliğinde amaç, kolaylıkla mantıksal kurallara ya da görsel sunumlara çevrilebilecek nitel modellerin çıkarılmasıdır. Şekil 2.2’de de gösterildiği gibi veri madenciliği sahası, istatistik, yapay zeka, veri tabanları ve veri görselleştirme gibi alanlar ile yakından ilişkilidir¹⁵.



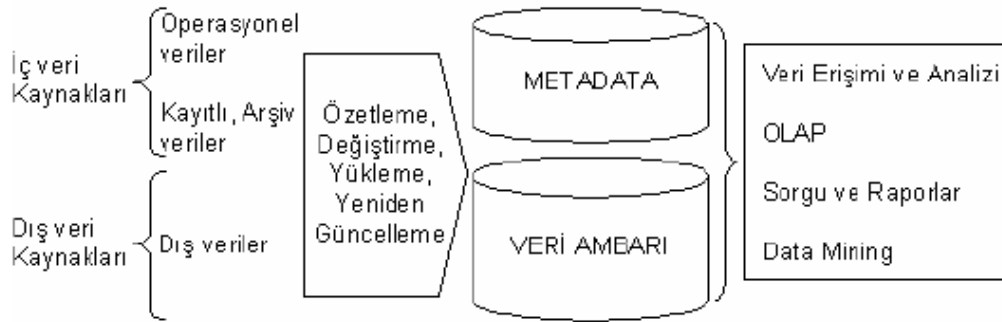
Şekil 2.2: Veri Madenciliği İlgili Alanları

Adil Baykasoğlu , “ Veri Madenciliği Ve Çimento Sektöründe Bir Uygulama”,**Gaziantep Üniversitesi Dergisi**:2.

¹⁵ Adil Baykasoğlu , “ Veri Madenciliği Ve Çimento Sektöründe Bir Uygulama”,**Gaziantep Üniversitesi Dergisi**:2

2.5. Veri Madenciliği ve Veri Ambarı Arasındaki Fark

Veri ambarı, yönetimin karar verme sürecini destekleyen, konu tabanlı, tümleşik, zamanla değişen, kalıcı verilerin topluluğudur. Bu amaca ulaşmak için, çeşitli programlar aracılığıyla operasyonel ortamdaki veri ambarı verisini barındıran veri tabanından ve kullanıcılara veri sağlayan sistemlerden veriler çıkarılır¹⁶. Veri ambarı ilişkili verilerin sorgulana bilindiği ve analizlerinin yapılabildiği bir depodur. Bir veri ambarı, analizler ve sorgular için kullanılabilir, bütünleşmiş bilgi deposudur. Veri ve bilgiler, üretildiklerinde heterojen kaynaklardan elde edilirler¹⁷.



Şekil 2.3: Veri Ambarı Bileşenleri ve Veri Madenciliği İlişkisi

Ümit Burak Usgurlu, “Veritabanı, Veri Madenciliği, Veri Ambarı, Veri Pazarı”
<http://mail.baskent.edu.tr/~20394676/0302/bil483/HW2.pdf> [16.01.2012].

Veri Ambarları, (VA), farklı veri tabanlarında tutulan verilerin bir araya getirilmesi, toplanması ile oluşturulan bütünleştirilmiş veri depolarıdır. VA kurumlarda var olan, depolanan verileri yönetmek için, VM ise daha çok karar destek sistemleri için verilerden anlamlı bilgiler çıkartarak bu bilgileri üst yönetime hızlı ve doğru bir biçimde sunmak amacı ile kurulurlar¹⁸.

Veri Ambarlarında veriler içerisinde çok sayıda eksik, hatalı, gürültülü veri mevcuttur ve bu veriler üzerinde kullanıcılar sorgular yaparak raporlar oluşturabilir. VM ise bu tutarsız, hatalı, gürültülü verilerle çalışmadığından verilerin mutlaka ön

¹⁶ Ahmet Cüneyt Baysal, “Bayi Değerlendirmesi İçin Veri Madenciliği Uygulaması” (Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, 2008), 6.

¹⁷ Ümit Burak Usgurlu, “Veri Tabanı, Veri Madenciliği, Veri Ambarı, Veri Pazarı”
<http://mail.baskent.edu.tr/~20394676/0302/bil483/HW2.pdf> [16.01.2012].

¹⁸ Türkiye Bilişim Derneği Kamu Bilgi İşlem Merkezleri Yöneticileri Birliği Kamu Bilişim Platformu, **Kamuda Karar Destek Sistemlerinin Kullanımı ve Bir Model Önerisi**(İstanbul, 2010), 68.

işlemden geçirilmesi gerekmektedir¹⁹. Her ne kadar veri madenciliği için veri ambarı olması zorunlu olan bir yapı değilse de veri madenciliği uygulamalarının verimliliğini arttırmaktadır.

2.6. Veri Madenciliğini Etkileyen Eğilimler

Temel olarak veri madenciliğini 5 ana harici eğilim etkiler²⁰ :

- a) Veri: Veri madenciliğinin bu kadar gelişmesindeki en önemli etkidir. Son yirmi yılda sayısal verinin hızla artması, veri madenciliğindeki gelişmeleri hızlandırmıştır. Bu kadar fazla veriye bilgisayar ağları üzerinden erişilmektedir. Diğer yanda bu verilerle uğraşan bilim adamları, mühendisler ve istatistikçilerin sayısı hala aynıdır. O yüzden, verileri analiz etme yöntemleri ve teknikleri geliştirilmektedir.
- b) Donanım: Gelişen bellek ve işlem hızı kapasitesi sayesinde, birkaç yıl önce madencilik yapılamayan veriler üzerinde çalışmayı mümkün hale getirmiştir.
- c) Bilgisayar Ağları: İnternet kullanım hızının artması, günümüzde kullanılan bilgisayar ağlarındaki hızın 100 katından daha fazla bir sürat ve taşıma kapasitesi demektir. Böyle bir bilgisayar ağı ortamı oluştuktan sonra, dağıtık verileri analiz etmek ve farklı algoritmaları kullanmak mümkün olacaktır. Buna bağlı olarak, veri madenciliğine uygun ağların tasarımı da yapılmaktadır.
- d) Bilimsel Hesaplamalar: Veri madenciliği ve bilgi keşfi, bu 3 metodu birbirine bağlamada önemli rol almaktadır: teori, deney ve simülasyon.
- e) Ticari Eğilimler: Günümüzde ticaret ve işler çok karlı olmalı, daha hızlı ilerlemeli ve daha yüksek kalitede servis ve hizmet verme yönünde olmalı, bütün bunları yaparken de minimum maliyeti ve en az insan gücünü göz önünde bulundurmalıdır. Bu tip hedef ve kısıtların yer aldığı iş dünyasında veri madenciliği, temel teknolojilerden biri haline gelmiştir.

¹⁹ Betül Demirel, “Veri Madenciliğinde Chaid Algoritmasının Sosyal Güvenlik Kurumu Veri Tabanına Uygulanması”, (Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, 2010), 42.

²⁰ Alper Vahaplar, Mustafa Murat İnceoğlu, “Veri Madenciliği ve Elektronik Ticaret”, www.bayar.edu.tr/bilisim/dokuman/inceoglu.doc Benzer [05.12.2011].

2.7. Veri Madenciliğindeki Problemler

Veri madenciliği araçları, güçlü çıktılar elde etmesine rağmen tek başına verimli uygulamalar değildir. Tam anlamı ile başarılı olmak için, veri madenciliği donanımlı teknik ve analitik uzmanlara ihtiyaç duymaktadır. Yani, genel olarak VM'deki sınırlamalar teknolojiden çok, öncelikle veri veya kişi temellidir.

Her ne kadar Veri Madenciliği, ilişkileri veya veri desenlerini açıklamaya yardımcı olsa da, kullanıcıya bu desenlerin önemini veya verinin değerini anlatmaz. Bu kararlar, kullanıcı tarafından alınmalıdır. Benzer bir şekilde, analiz sonucunda keşfedilen veri örüntülerinin geçerliliği de, gerçek dünyadaki koşullara bağlılığına göre kıyaslanmalıdır.

Veri Madenciliğindeki bir diğer problem de, değişkenler ve / veya davranışlar arasındaki ilişkiyi tanımlarken ortaya çıkmaktadır. Bu ilişkiler, her zaman nedensel bir ilişki olmak zorunda değilse biraz karmaşık olabilmektedir. Örneğin, tıpkı uçuş planlanmadan evvel, hava yolu biletlerini satın almaya eğilim gibi bir davranışın desenini tanımlayan bir uygulama, gelir, eğitim seviyesi ve internet kullanımı gibi özellikler ile ilgili olabilir²¹.

Veri madenciliği girdi olarak ham veriyi sağlamak üzere veri tabanlarına dayanır. Bu da veri tabanlarının dinamik, eksiksiz, geniş ve net veri içermemesi durumunda sorunlar doğurur. Diğer sorunlar da verinin konu ile uyumsuzluğundan doğabilir²².

Sınıflandırmak gerekirse başlıca sorunlar şunlardır ²³:

- *Veri Tabanı Boyutu*: Veri tabanı boyutları inanılmaz bir hızla artmaktadır. Veri tabanı algoritması çok sayıda küçük örnekleme ele alabilecek biçimde geliştirilmiştir. Aynı algoritmaların yüzlerce kat büyük örneklemelerde kullanılabilmesi için çok dikkat gerekmektedir. Örneklemin büyük olması, tahminlerin doğruluğu açısından bir avantaj olsa da dikkatsizlikten kaynaklanacak hatalar göz ardı edilemez.
- *Aykırı Veri*: Veri girişi veya veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir. Verilerde ne kadar çok gürültü varsa o derece güvenilir sonuçlara ulaşmak zorlaşacaktır. Bu gürültüler geleceğe dair tahminlerin

²¹ Seifert, **age**, 6.

²² Vahaplar, **age**, [05.12.2011].

²³ "Veri Madenciliğine Giriş", **age**, 3.

doğruluğunun azalmasına neden olur. Gürültülü verilerin teşhis edilmesi amacıyla histogram, kümeleme analizi ve regresyon kullanılır.

- *Eksik Veri:* Veri kümesinin büyüklüğünden ya da doğasından kaynaklanmaktadır. Eksik veriler olduğunda yapılması gerekenler şunlardır:
 - Eksik veri içeren kayıt veya kayıtlar çıkarılabilir.
 - Değişkenin ortalaması eksik verilerin yerine kullanılabilir.
 - Var olan verilere dayalı olarak en uygun değer kullanılabilir.

Veri madenciliğinde çeşitli problem tipleri çözümlenmeye çalışılır. Çözümlenmeye çalışılan bu problem tiplerine bir veya birkaç istatistiksel teknik kullanılarak çözümler üretilir.

2.8. Veri Madenciliği Kullanım Alanları ve Uygulamaları

VM, çok geniş bir kullanım alanına sahiptir. Bu alanlar ve uygulama örnekleri başlıklar halinde aşağıda ifade edilmiştir²⁴:

- Bilgi teknolojisi: Bilgisayarların, yazılımların ve Web portallarının değerlendirilmesi.
- Projeler: Projelerin, yatırımların değerlendirilmesi, ürün portföy değerlendirmesi.
- Sigortacılık: Müşteri kaybı sebeplerinin belirlenmesi, usulsüzlüklerin önlenmesi, Yeni poliçe talep edecek müşterilerin tahmin edilmesi.
- Telekomünikasyon: Hile tespiti, hatların yoğunluk tahminleri.
- Borsa: Hisse senedi fiyat tahmini, genel piyasa analizleri.
- Tıp: Tıbbi teşhis, uygun tedavi sürecinin belirlenmesi, risk değerlendirmesi.
- Bilim ve Mühendislik: Ampirik veriler üzerinde modeller kurularak bilimsel ve teknik problemlerin çözümlenmesi.
- Endüstri: Kalite kontrol, lojistik, üretim süreçlerinin en uygun tasarımı, ürün tasarımı.
- Bankacılık: Risk analizleri ve usulsüzlük tespiti.
- İşletme: İş partner seçimi, işletmenin performans değerlendirilmesi.

²⁴ Türkiye Bilişim Derneği Kamu Bilgi İşlem Merkezleri Yöneticileri Birliği Kamu Bilişim Platformu, age, 84.

- Personel Yönetimi: Personel değerlendirme, eğitim planlama.
- Finansal Planlama ve Varlık Değerlendirme: Nakit akışı planlama ve analizi, zaman serileri analizleri (finansal oranlar, trend analizi vb.).
- Kaynak Planlama: Kaynak tahsisi, kaynak ve harcamaların özetlenip karşılaştırılması.
- Rekabet: Rakiplerin ve pazar şartlarının takip edilmesi, müşterilerin sınıflara ayrılması ve buna göre fiyatlandırmanın yapılması, fiyatlandırma stratejilerinin çok rekabetçi bir ortamda belirlenmesi.

Türkiye’de son yıllarda yapılmış çalışmalar incelendiğinde, çok sayıda çalışma yapıldığı görülmektedir. Bu çalışmalardan bazıları kronolojik olarak şöyle sıralanabilir: Bozkır, Gök ve Sezer (2008) Üniversite öğrencilerinin internet kullanımı üzerine karar ağaçları, kümeleme ve birliktelik kuralları analizleri uygulamaları, Ata ve Seyrek (2009) imalat firmaları üzerine karar ağaçları ve sinir ağları analizleri uygulaması, Albayrak ve Yılmaz (2009) İMKB verileri üzerine karar ağaçları analizi uygulaması, Özkul ve Pektekin (2009) muhasebe yolsuzluklarının tespiti, Gürbüz, Özbakır ve Yapıcı (2009) havayolu işletmesine ait raporlar ile ilgili uygulama yapmıştır²⁵.

2.9. Veri Madenciliğinde Kullanılan Paket Programlar

VM’nde kullanılan birçok program vardır. Bunlar arasında en çok tercih edilen programların başında SPSS Clementine ve SAS Enterprise Manager programları gelmektedir²⁶.

- *SPSS Clementine*: SPSS’in VM metodolojisi olarak kabul ettiği CRISP DM (Cross Industry Standart Processing For Datamining) % 50’nin üzerinde bir kullanıma sahiptir. İnternet kayıtlarına ve elde edilen verilere, gelişmiş VM teknikleri uygulayarak, kullanıcılar ile birebir ilişki kurmayı sağlayacak öngörüler elde edilebilir.
- *SAS (Statistical Analysis Software) Enterprise Miner*: Şirketlerin çok büyük veri yığınlarından kritik bilgileri elde etmelerini sağlayan VM çözümlerinde dünyada önemli bir yere sahip olan SAS, veri üzerinde değil bilgi üzerinde

²⁵ Tüzüntürk, **age**, 70.

²⁶ Demirel, **age**, 68.

düşünme ve strateji geliştirme avantajını bir adım öteye taşıyarak SAS Enterprise Miner programını geliştirmiştir.

- *Statistica Data Miner*: VM için parametrik istatistik ve makine öğrenimi kombinasyonu zengin algoritmalar sunan bir programdır. Grafiksel programlama ara yüzüne sahiptir, tüm ortak VM görevleri için araçlar sunar.
- *DBMiner*: DBMiner'i diğer programlardan ayıran en önemli özelliği, OLAP yöntemleriyle VM algoritmalarını bütünleşik hale getirmiş olmasıdır.

3. VERİ TABANINDA BİLGİ KEŞFİ SÜRECİNDE VERİ MADENCİLİĞİ

Veri Madenciliği Uygulama Süreci:

Veri Madenciliği sürecinden bahsetmeden önce aktarılması gereken temel bir kavram Veri Tabanlarında Bilgi Keşfi (VTBK)'dir.

Veri tabanında bilgi keşfi, verilerin doğru, faydalı ve anlaşılır modeller ve kalıplar elde etmede kullanılan özel bir süreçtir²⁷.

Veri tabanlarında bilgi keşfi; verilerden

- Doğru
- Yeni
- Faydalı
- Anlaşılır

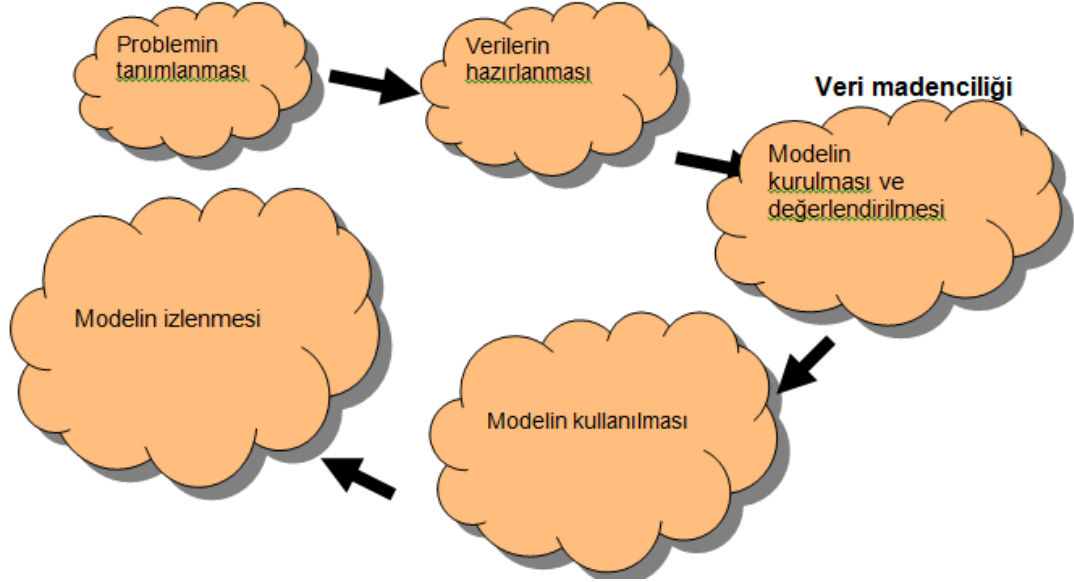
modeller ve kalıplar elde etmede kullanılan özel bir süreçtir.

Bu tanımdaki “özel bir süreç” ifadesinden de anlaşılacağı gibi Veri Tabanlarında Bilgi Keşfi Süreci (VTBK), karmaşık bir işlemdir. “Model elde etmek” ile kastedilen ise, verilere model uydurmak, verilerden yapı ortaya çıkarmak veya genel bir ifadeyle, veri kümesine yüksek dereceli açıklama getirmektir. “Süreç” ise, VTBK'nin birçok adımdan ve çeşitli iterasyonlardan oluştuğunu göstermektedir. Bilgi keşfinin belirli bir güven düzeyinde, yeni veriler için geçerli olması, süreç sonunda elde edilen bilginin ise, iş konusunda çeşitli avantajlara olanak sağlayacak şekilde faydalı ve anlaşılır olması gerekmektedir.

Geleneksel sorgu veya raporlama araçlarının veri yığınları karşısında yetersiz kalması, **Veri Tabanlarında Bilgi Keşfi (VTBK)** adı altında, sürekli ve yeni arayışlara neden olmaktadır. Şekil 3.1'de görülen VTBK süreci içerisinde, modelin kurulması ve değerlendirilmesi aşamalarından meydana gelen **Veri Madenciliği** en

²⁷ “Veri Madenciliğine Giriş”, **age**, 4.

önemli kesimi oluşturmaktadır. Veri madenciliği buradaki modelin kurulması ve modelin değerlendirilmesi aşamalarından meydana gelmektedir²⁸.



Şekil 3.1: Veri Tabanlarında Bilgi Keşfi Süreci

“Veri Madenciliğine Giriş”, **age**, 5.

VTBK süreci, veri tabanlarını kullanarak veri tabanlarında istenilen seçim, ön işleme, alt örnekleme, dönüşüm, örüntülerin açığa çıkarılması için Veri Madenciliği yöntemlerinin (algoritmalarının) uygulanması ve açığa çıkarılan örüntülerin tanımlanması için Veri Madenciliği ürünlerinin yorumlanmasını içermektedir. VTBK sürecinin, VM bileşeni, veriden hangi örüntülerin aktarılıp, dikkate alınacağını algoritmik anlamda ifadesi olarak değerlendirilmelidir.

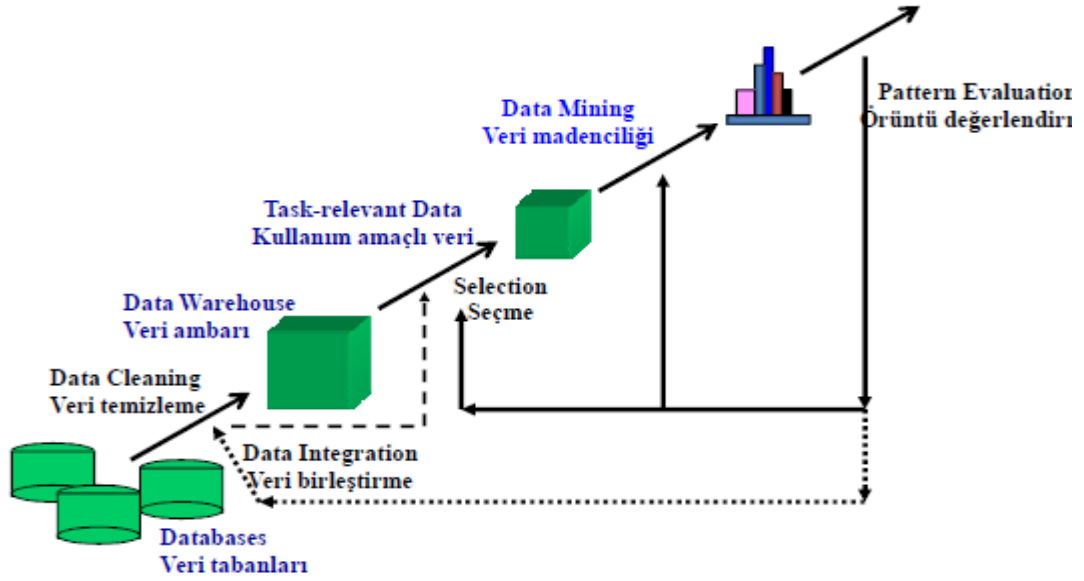
Veri Madenciliği gözlenen veriye model uydurmayı veya gözlenen verideki örüntüleri tanımlamayı gerektirmektedir. Model uydurma, bilgi çıkarımı rolünü üstlenmektedir: Modelin, kullanışlı veya ilginç keşifsel bilgiye işaret edip etmediği, tamamıyla interaktif VTBK sürecinin subjektif insan yargısına tipik olarak ihtiyaç duyduğu bir parçasıdır²⁹. Teoride veri madenciliği bilgi keşfi işleminin bir parçasıdır. Pratikte veri madenciliği ve veriden bilgi keşfi aynı anlamda kullanılır³⁰.

²⁸ “Veri Madenciliğine Giriş”, **age**, 5.

²⁹ Koyuncu, **age**, 22.

³⁰ Suat Özdemir, “Veri Madenciliği Ders Notları” <http://ceng.gazi.edu.tr/~ozdemir/> [16.01.2012].

VERİDEN BİLGİ KEŞFİ



Şekil 3.2: Veriden Bilgi Keşfi

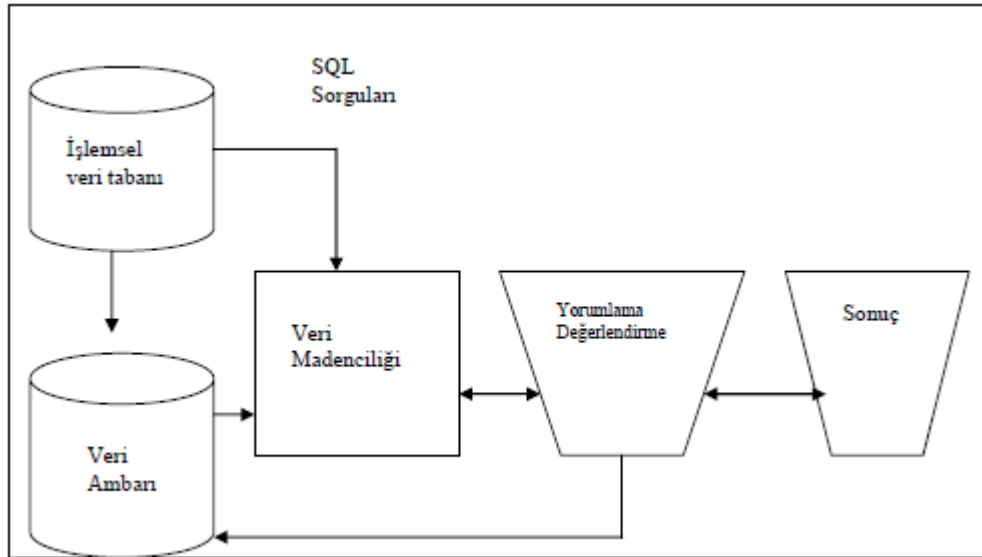
Suat Özdemir, “Veri Madenciliği Ders Notları” <http://ceng.gazi.edu.tr/~ozdemir/> [16.01.2012].

Veri madencisinin geleneksel yöntemlerde olduğunun aksine başlangıçta herhangi bir amacı ya da varmak istediği bir kavram yoktur. Yapılacak analizlerden sonra elde edilen verilerin bir istatistikçi gözü ile incelenip daha önceden düşünülmemiş kavramların ortaya çıkarılması, başarılı bir VM süreci olarak kabul edilmektedir³¹.

Şekil 3.3’de veri madenciliği süreci görsel olarak özetlenmektedir. Burada işlemsel veri tabanı, veri ambarı, veri madenciliği, yorumlama değerlendirme ve sonuç arasında çeşitli bağlar bulunmaktadır³².

³¹ Albayrak, **age**, 35.

³² Tüzüntürk, **age**, 68.



Şekil 3.3: Basit Bir Veri Madenciliği Süreci

Selim Tüzüntürk, “VERİ MADENCİLİĞİ VE İSTATİSTİK”, *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, c. 29. s. 1(2010): 68.

Veri madenciliği süreci en basit şekli ile dört adımdan oluşmaktadır:

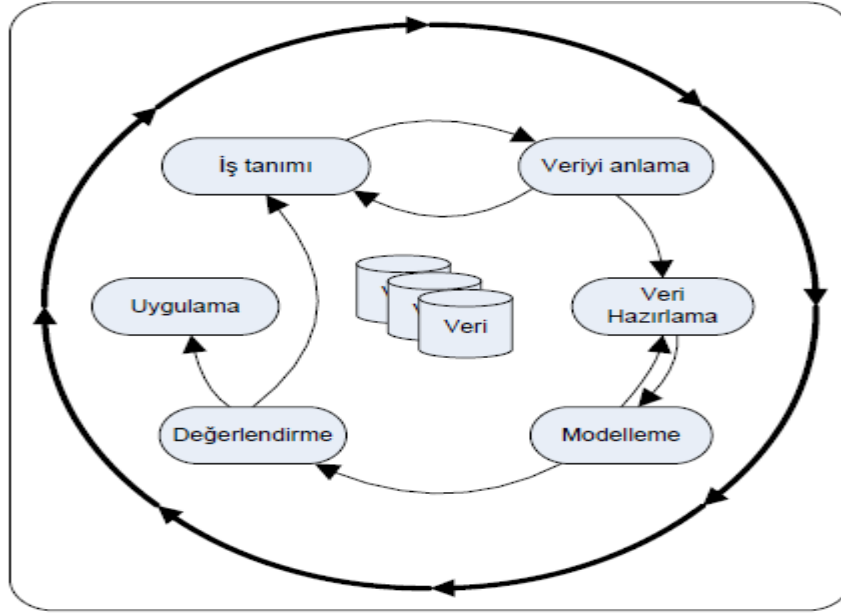
1. Veri Seçmek
2. Veri İşlemek
3. Veri Eğilimlerini / Desenlerini Belirlemek
4. Bilgiye Ulaşmak

Detaylı süreç tanımı ise veri madenciliğinin uluslararası düzeyde standardı olarak kabul edilmiş, CRISP-DM (Cross Industry Process for Data Mining) ile yapılmaktadır³³. Veri madenciliği projelerinin hızlı, daha verimli ve daha az maliyetli gerçekleştirilmesi için geliştirilmiş olan bu süreç altı adımdan oluşmaktadır³⁴:

- (1) Araştırma Probleminin Tanımlanması (Business Understanding)
- (2) Verileri Tanıma Aşaması (Data Understanding)
- (3) Veri Hazırlama Aşaması (Data Preperation)
- (4) Modelleme Aşaması (Modelling)
- (5) Değerlendirme Aşaması (Evaluation)
- (6) Uygulama Aşaması (Deployment)

³³ Argüden, **age**, 20.

³⁴ Albayrak, **age**, 36.



Şekil 3.4: CRISP – DM Veri Madenciliği Uygulama Süreci

Sinan Aydın, “Veri Madenciliği Ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama”(Doktora Tezi, Eskişehir Anadolu Üniversitesi Sosyal Bilimler Enstitüsü Eylül 2007), 16.

Bir diğer veri madenciliği uygulama süreci Two Crows şirketi tarafından önerilmiştir. Two Crows şirketi bankacılık, sigortacılık, telekomünikasyon, perakendecilik, devlet uygulamaları, danışmanlık ve enformasyon sistemleri için veri madenciliği uygulama adımlarını tanımlayan raporun üçüncü sürümünü 1999 yılında yayınlamıştır. Bu teknik rapora göre uygulama adımları aşağıdaki gibi sıralanmıştır ve bugün genel olarak bu sıra takip edilmektedir³⁵.

1. Problemin tanımlanması
2. Veri madenciliği veri tabanının oluşturulması
3. Verinin incelemesi
4. Model için veri hazırlama
5. Modelin oluşturulması
6. Modelin değerlendirilmesi
7. Modelin uygulanması ve sonuçların izlenmesi

Bu ana başlıklar alt başlıklar halinde ve detaylı olarak tezin takip eden bölümlerinde incelenmektedir.

³⁵ **Introduction to Data Mining and Knowledge Discovery** (Two Crows Corp., 2005), 22, www.trocrows.com [14.12.2011].

3.1. Araştırma Probleminin Tanımlanması

Uygulamanın hangi işletme amacı için yapılacağını açık bir şekilde tanımlanmasıdır. İlgili işletme amacı işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalıdır³⁶.

Bu aşama, veri madenciliği sürecinin en önemli aşamasıdır. Araştırma probleminin (konusunun) tanımlanması aşaması araştırmanın amacını, mevcut durumun değerlendirilmesini, veri madenciliğinin amaçlarını ve proje planlama sürecinin belirlenmesini kapsamaktadır³⁷.

İş ve iş ortamını anlama da denilen bu aşamada, ilk adım veri madenciliği çalışmasının hangi amaç için yapılacağını net olarak tanımlanmasıdır. Amaç, cevap aranan sorunun üzerine odaklanmalı, net biçimde ifade edilmeli ve sonuç değerlendirme kriterleri tanımlanmalıdır³⁸.

3.2. Veri Madenciliği Veri tabanının Kurulması

Veri madenciliği veri tabanının kurulması aşaması, ilk adımda tanımlanan problemin çözümüne ilişkin analizin ihtiyaç duyduğu özellik ve nitelikteki verinin hazırlanması olarak ifade edilebilir. Veri madenciliği veri tabanının kurulması aşaması, ilk adımda tanımlanan problemin çözümüne ilişkin analizin ihtiyaç duyduğu özellik ve nitelikteki verinin hazırlanması olarak ifade edilebilir. Bu aşama veri kaynaklarının belirlenmesi, veri tanımlama, veri seçme, veri kalitesi ve ön hazırlık süreçleri, veri madenciliği veri tabanının yüklenmesi ve bakımı görevlerinin yerine getirilmesi ile tamamlanır³⁹.

3.2.1. Veri Kaynaklarının Belirlenmesi

Veri kaynaklarının belirlenmesi, madenciliği yapılacak olan veri kaynaklarının tanımlanmasıdır. Veriler pek çok farklı kaynaktan elde edilebilir. Verilerin saklanması ve yönetilmesinde basit dosya sistemlerinden veri tabanları ve veri ambarlarına uzanan birçok farklı yöntem ve teknoloji bulunmaktadır.

³⁶ Tüzüntürk, **age**, 68.

³⁷ Albayrak, **age**,36.

³⁸ Argüden, **age**, 20.

³⁹ Sinan Aydın, “Veri Madenciliği Ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama”(Doktora Tezi, Eskişehir Anadolu Üniversitesi Sosyal Bilimler Enstitüsü Eylül 2007), 18.

3.2.1.1 Veri tabanı Sistemleri

Veri tabanı sistemleri günümüzde veri saklama ve yönetiminde önemli bir yere sahiptir. Bir veri tabanı sistemi veri tabanı dosyaları ve veri tabanı yönetim sisteminden oluşur.

1. İlişkisel Veri tabanları: İlişkisel veri tabanlarının temel elemanı tablolardır. Veri tabanı içinde her tablonun sadece kendine ait bir ismi vardır. Tablolar arası ilişki de anahtar öznitelikler ya da sütunlar yardımı ile sağlanmaktadır⁴⁰.

İlişkisel veri tabanlarında veri madenciliği uygulandığında eğilimler veya veri örüntüleri araştırılarak daha fazla bilgi elde etmek mümkün olur. Örneğin yeni müşterilerin kredi riskini tahmin etmek için müşterilerin gelir, yaş ve önceki kredi bilgilerine dayanarak veri madenciliği analizi gerçekleştirilebilir. İlişkisel veri tabanları en popüler ve en zengin enformasyon depolarından biridir ve bu nedenle veri madenciliğinde kullanılan en yaygın veri kaynaklarıdır⁴¹.

2. Hareket Veri Tabanları: Her bir hareketin ya da işlemin bir satır veya kayıtlarla ifade edildiği veri tabanlarıdır. Hareket veri tabanlarında her bir işlem benzersiz bir tanımlayıcı ile ifade edilirken hareketi oluşturan nesnelere ait tablolar bulunmaktadır. Hareketi oluşturan nesnelerin özelliklerini barındıran bu tablolar da aynı veri tabanında saklanmaktadır. Bir markette yapılan satışların depolandığı veri tabanı bu tür bir veri tabanına örnektir⁴².

Bir hareket veri tabanı ile ilişkisel veri tabanı arasındaki temel fark, kullandıkları veri modelidir. Yani verinin sunuş şekli bu farkı oluşturur.

3.2.1.2. OLAP Ve Veri Ambarları

Veri madenciliğinin ihtiyaç duyduğu ve veri ambarında depolanan veriler farklı yapılarıdaki veri kaynaklarının bir araya getirilmesinden oluşturulur. Bu nedenle veri ambarlarında depolanan veriler veri madenciliği kaynağı olarak kullanılabilir. Veri madenciliği ve veri ambarları birbirini tamamlayıcıdır. Veri ambarları ve OLAP

⁴⁰ Cem Sütçü, “İstatiksel Veri Sistemleri ve Basın Sektöründe bir Karar Destek Sistemi Uygulaması”(Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, 1995), 56.

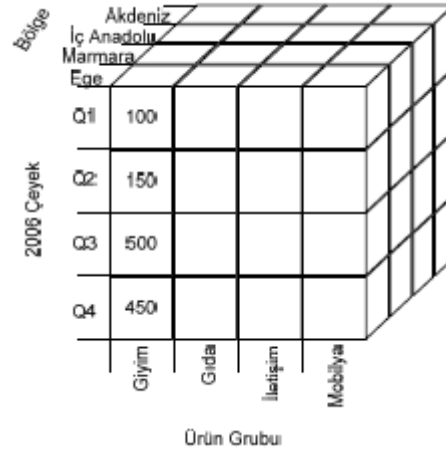
⁴¹ Aydın, age, 21.

⁴² Age, 21.

(Çevrimiçi Analitik İşleme) araçları çok boyutlu veri modeline dayalıdır. Çok boyutlu veri modeli, verileri bir küp şeklinde ele alır.

Sorgulama ve raporlama aracının gerçekleştirimi öncesinde bu aracın arka planda besleneceği verilerin tutulduğu veri ambarının ve bu ambar üzerinden üretilecek çok boyutlu küp tanımlarının tasarlanması ve oluşturulması gerekmektedir⁴³.

Bir veri küpü verilerin çok boyutlu olarak yapılandırılmasını ve görüntülenmesini sağlar. Örneğin bir şirketin satışlarına ilişkin veri küpü oluşturulduğunu düşünelim. Veri küpünün ana temasının satış tutarları olduğunu, boyutlarının ise satış bölgesi, zaman ve ürün grubu olduğunu varsayalım. Bu durumda Şekil 3.5’de görüldüğü gibi bir veri küpü oluşturulabilir. Bir veri küpünün ihtiyaca göre üçten fazla boyutu da tanımlanabilir. Oluşturulan veri küpleri ihtiyaç duyulan bilgiye göre OLAP sistemi tarafından sorgulanır⁴⁴.



Şekil 3.5: Veri Küpü Örneği

Aydın, age, 21.

3.2.2 Veri Tanımlama (Veri Anlama Aşaması)

Veri tanımlama ya da “veri anlama aşaması” olarak belirlenmiş bu aşamada; veriyi anlamak için ilk yapılması gereken verinin toplanmasıdır.

⁴³ Selda Düzgünoğlu, “Hacettepe Üniversitesi Veri Ambarı Ve Olap Teknolojilerinden Yararlanılarak Karar Destek Amaçlı Raporlama Aracı Gerçekleştirimi”, (Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı, 2006), 57.

⁴⁴ Aydın, age, 23.

Veri madenciliği için verinin tek bir tabloda toplanması en uygunu olacaktır⁴⁵. Veri anlama aşaması veri toplamakla başlamaktadır. Daha sonra benzer verileri bir araya getirme, veri niteliklerini tanımlama, verileri keşfetme, gizli bilgileri sınıflandırma ile sürece devam etmektedir⁴⁶.

Veri toplama işlemi tamamlandıktan sonra, veriler test ve analiz veri seti olarak iki gruba ayrılır. Genellikle yapılan uygulamalarda verilerin %80'i analiz %20'si ise test verisi olarak ayrılır⁴⁷.

Proje kaynaklarında tanımlanmış olan başlangıç verilerinin toplanmasından sonraki adım, toplanan verinin tanımlanması ve ihtiyaçları karşılama yeterliliğinin değerlendirilmesi aşamasıdır.

Veriyi Keşfetmek ile anlatılmak istenen, başlangıç aşamasında toplanan veriler ile başlangıç hipotezlerinin oluşturulması, limitli bir şekilde veriden çıkarımlar yapılması aşamasıdır. Bu aşamada sonuca yönelik bilgilerin elde edilmesinden daha çok çalışmanın gerçekleştirilebilmesi için veri anlamında eksikliklerin tespit edilmesi amaçlanır⁴⁸.

3.3. Veri seçimi

Yapılacak olan analiz ile ilgili olan verileri belirleme aşaması “veri seçimi” dir. Yapılacak analizde kullanılacak verilerin belirlenmesi aşamasıdır⁴⁹. Veri madenciliği veri tabanı hazırlamada veri tanımlama aşamasından sonra madenciliği yapılacak verinin alt kümesi seçilir. Bu aşamada veri tabanını örnekleme veya tahmin edici değişkenleri seçme işlemi değil gereksiz veya ihtiyaç duyulmayan verinin analiz dışı bırakılmasıdır⁵⁰.

Verilerin değerlendirilmesi sırasında verinin hedefler ile ilişkisine, kalitesine, teknik limitlere dikkat etmek gerekir. Verilerin değişken sayısı kadar kullanılan kayıt sayısı

⁴⁵ Gültekin Arabacı, “ Veri Madenciliğinde Apriori, Tahminci Apriori Ve Tertius Algoritmalarının Weka Ve Yale Programları İle Karşılaştırılması Ve Bir Uygulama” (Yüksek Lisans Tezi, İstanbul Ticaret Üniversitesi Sosyal Bilimler Enstitüsü, 2007), 10.

⁴⁶ Albayrak, **age**, 36.

⁴⁷ Adem Tekerek, “Veri Madenciliği Süreçleri Ve Açık Kaynak Kodlu Veri Madenciliği Araçları”, <http://ab.org.tr/ab11/bildiri/29.pdf> [16.12.2011].

⁴⁸ Argüden, **age**, 21

⁴⁹ Tekerek, **age**, 2.

⁵⁰ Aydın, **age**, 26.

da önemlidir. Gereğinden az veri, çalışmayı eksik bırakabileceği gibi, gereğinden fazla veri, veri kirliliğine ve sürecin uzamasına neden olabilecektir⁵¹.

3.4. Veri Kalitesini İyileştirme Ve Ön Hazırlık Süreçleri

Veri kaynaklarında yer alan verilerin her zaman mükemmel olması mümkün değildir. Veri madenciliği teknikleri ve algoritmalarının çoğu verilerdeki kusurları göz ardı edebilmesine rağmen veri kalitesini anlamak ve iyileştirmek konusuna odaklanmak veri madenciliği çıktı kalitesini artırır. Bu da zaten VM analistlerinin hedeflediği sonuçtur.

Veri kalitesi kavramı verideki gürültü ve aykırı değerler, eksik, tutarsız veya tekrarlı verilerin varlığı ile ölçülebilir⁵². Veri kalitesi, veri madenciliğinde anahtar bir konudur. Veri madenciliğinde güvenilirliğin artırılması için, veri ön işleme yapılmalıdır⁵³.

Ayrıca, veri kalitesi, verinin doğruluğunu ve bütünlüğünü de ifade eder. Veri kalitesi, analiz edilen verinin tutarlılığı ve yapısından da etkilenebilir. Bir diğer hem veri kalitesini hem de analiz sonuçlarını etkileyen kavram, gizlilik kavramıdır. Giderek önemi artan bir konu olan veri gizliliği, veri madenciliği uygulamalarının orijinal amaçlarını, gerçek projelere uygun olarak yerine getirmesi açısından çok önemlidir⁵⁴.

Veri hazırlama aşaması, ham veriden başlayarak son veriye kadar yapılması gereken bütün düzenlemeleri içermektedir⁵⁵. Her veri analizi işi yeni bir veri setlerinin toplanması, betimlenmesi ve temizlenmesiyle başlar. Bu süreçten sonra, veriler analiz edilebilir ve sonuçlara ulaşılır⁵⁶.

Verilerin hazırlanması aşaması şu adımlardan oluşmaktadır:

- *Verilerin Temizlenmesi:* Veriler, toplanma şekillerine ve toplandıkları kaynaklara bağlı olarak gürültülü ya da tutarsız olabilirler, eksik özellik değerlerine sahip olabilirler. Bu tip veriler, işlem sonrasında yapılan tahminlerde yanlışlara sebep

⁵¹ Argüden, **age**, 22.

⁵² Aydın, **age**, 26.

⁵³ Tüzüntürk, **age**, 68.

⁵⁴ Seifert, **age**, 16.

⁵⁵ Albayrak, **age**, 36.

⁵⁶ Tüzüntürk, **age**, 69.

olabilirler⁵⁷. Veri temizleme işleminin amacı, veriler içindeki uygun olmayan veya hatalı girilmiş verileri ayıklamaktır⁵⁸.

Birkaç nitelik için eksik veri girişi varsa bazı yöntemlerle eksik veriler tamamlanır. Eksik nitelik değeri taşıyan veriler, göz ardı edilebilir, rastgele değerlerle doldurulabilir, bu niteliğe özgü global bir değerle doldurulabilir, o niteliğe ait ortalama değerle doldurulabilir, aynı sınıfa ait verilerin nitelik değerlerinin ortalamasıyla doldurulabilir ya da veri kümesinin o niteliğe ait olasılığı en yüksek değeriyle doldurulabilir⁵⁹. Regresyon yöntemi kullanılarak eldeki eksik olmayan tüm verilere bir regresyon denklemi ve regresyon katsayıları elde edilerek, kayıp olan veriler tahmin edilebilir. Eğer eksik veri çok ise bu kaydın silinmesi gerekir⁶⁰.

Verilerimiz içerisinde tutarsız ya da gerçeğe aykırı olan “gürültülü veri” diye adlandırılan hatalı veriler de bulunabilir. Gürültü, veri madenciliği tekniği ile analiz etmek istediğimiz verilerdeki beklenen değerlerden sapan aykırı değerler veya hatalardır. Gürültülü veri büyük veri tabanları ve veri ambarlarında karşılaşılan yaygın problemlerdendir⁶¹.

Gürültülü veriler, bölmeleme, demetleme ya da eğri uydurma gibi metotlarla düzeltilebilir. Bölmeleme işlemi eldeki verinin sıralanarak eşit bölmelere ayrılması ve her bölmenin kendisine ait ortalama ya da uç değerlerle ifade edilmesidir. Böylece verilerdeki hata miktarlarının minimize edilmesi amaçlanır. Demetleme ile benzer veriler aynı demette olacak şekilde gruplanır ve dışarıda kalan veriler göz ardı edilir. Böylece muhtemel yanlış ölçümler ayıklanmış olur. Eğri uydurma metodu ile ise nitelik değeri diğer niteliklere bağlı olarak belli bir fonksiyona uydurulur. Bu fonksiyon kullanılarak nitelik değerindeki tutarsızlıklar giderilir⁶².

Kayıp veriler dışında, gürültülü verilerin tespiti içinde kullanılan yöntemlerden birkaçı kümeleme analizi, histogram ve regresyon yöntemi olarak sayılabilir⁶³.

⁵⁷ Gökçe Kolay, “İşletmelerde Bilgi Sistemleri Verimliliğini Arttırmada Veri Madenciliği Yöntemi: Bir Simülasyon Çalışması” (Yüksek Lisans Tezi, Zonguldak Karaelmas Üniversitesi Sosyal Bilimler Enstitüsü, 2006), 35.

⁵⁸ Tekerek, **age**, 3.

⁵⁹ Cengiz Coşkun, “Veri Madenciliği Algoritmalarının Karşılaştırılması”(Yüksek Lisans Tezi, Dicle Üniversitesi, Fen Bilimleri Enstitüsü, 2010), 12.

⁶⁰ Demirel, **age**, 23.

⁶¹ Aydın, **age**, 30.

⁶² Coşkun, **age**, 14.

⁶³ Demirel, **age**, 23.

- *Verilerin Birleştirilmesi:* Bu aşamada farklı veri tabanlarındaki verilerin tek bir çatı altında (veri ambarında) birleştirilmesi işlemidir. Veriye sahip olma, verinin işlenmesi, iletilmesi ve depolama kapasitesindeki ilerlemeler işletmelerin sahip oldukları çeşitli veri tabanlarının veri ambarlarında birleştirilmesine olanak tanımıştır⁶⁴. Veri entegrasyonu, aslında veri ambarlama olarak da adlandırılabilir. Fakat veri ambarı kurmadan da veri entegrasyonu yapılabilir. Veri ambarı kalıcı bir çözüm iken, veri entegrasyonu anlık olarak da yapılabilir. Kalıcı bir veri ambarı kurmak ya da tek başına veri entegrasyonu yapmak arasındaki seçimi, kullanılacak veri madenciliği uygulaması, bu uygulamanın çalışma sıklığı, anlık veri transferinin hızı ve maliyeti, transfer edilecek verinin büyüklüğü gibi faktörler belirler⁶⁵.

Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olmaktadır. Ancak burada çok dikkatli ve titiz davranmak gereklidir.

Veri bütünleştirmede şema birleştirilmesi, fazla veri sorunları ve veri değer karmaşalarının belirlenmesi ve çözülmesi olmak üzere üç temel konu ön plana çıkar.

İki farklı kaynaktan gelen verilerin eşleştirilmesi için aynı varlıklar belirlenerek veriler şemalar yardımıyla birleştirilir. Şema birleştirme işleminde hataları engellemek için meta veri kullanılabilir⁶⁶.

Veri tabanlarının birleştirilmesi ile birlikte farklı tablolardan gelen aynı niteliğe sahip değişkenlerin fazlalıkları VA'nın gereksiz yere büyümesine sebep olmaktadır. Örneğin; veri tabanının birinde tüketici_id olarak geçen kolon adı diğer veri tabanında tüketici_numarası olarak adlandırılmış olabilir. Dolayısıyla yapılan birleştirme işlemi ile birlikte aynı değişken çoklanmış olmakta ve şema birleştirme hatası meydana gelmektedir. Yapılması gereken, yukarıda bahsedilen tüketici_id ile katsayı elde edilirse değişkenlerden birinin veri tabanından çıkartılması şeklinde olmalıdır⁶⁷.

Veri bütünleştirmede üçüncü önemli konu veri değer karmaşıklığının belirlenmesi ve çözülmesidir. Farklı veri kaynaklarından gelen özellik değerleri ölçekleme, birim

⁶⁴ Tüzüntürk, **age**, 69.

⁶⁵ Kolay, **age**, 38.

⁶⁶ Aydın, **age**, 32.

⁶⁷ Demirel, **age**, 24.

sistemi veya gösterimdeki farklılıklar yüzünden birbirlerinden farklı olabilirler. Örneğin ağırlık özelliği farklı kaynaklarda farklı birim sistemiyle depolanmış olabilir. Veri bütünleştirme işlemlerinde verinin bu tür heterojenliği dikkate alınmalıdır⁶⁸.

- *Verilerin Dönüştürülmesi:* Veri seti oluşturulduktan sonra kullanılacak modele göre anlam değişikliği yapmayacak format düzenlemelerinin yapılması aşamasıdır⁶⁹. VM’nde birçok algoritma kullanılmakta olup, bu algoritmaların bazıları kesikli verilerle bazıları ise sürekli verilerle çalışmaktadır. Dolayısıyla kullanacağımız algoritmaya göre bu verilerin bir dönüşüme tabi tutulması gerekmektedir⁷⁰. Dönüştürme işlemi, verilerin veri madenciliği için uygun formlara dönüştürülmesidir. Veri dönüştürme; düzeltme, birleştirme, genelleştirme ve normalleştirme gibi işlemlerin bir veya bir kaçını içerir⁷¹.

Düzeltilme; bölmeleme, kümeleme ve regresyon gibi teknikler kullanılarak verilerdeki gürültünün temizlenmesidir. Standartlaştırma veya normalizasyonun amacı sayısal veri değerlerinin küçük bir bölgede yer alması için ölçeklenmesidir. Normalleştirilmiş veriler sınıflama için kullanılan yapay sinir ağları algoritmalarının öğrenme aşamasının hızlanmasına yardım edecektir. Kümeleme gibi mesafe ölçümlerine dayalı algoritmalarda normalleştirilmiş verilerin kullanılması faydalı olacaktır⁷².

Genelleme ise düşük düzeydeki verinin kavram hiyerarşisi kullanılarak daha yüksek seviyeye dönüştürülmesidir. Yaş vb. verinin süreklilik gösterdiği durumlarda ilgili kolon belirlenmiş gruplara ayrılır (Çocuk, genç, orta yaşlı, yaşlı, çok yaşlı)⁷³.

- *Verilerin İndirgenmesi:* Her zaman çok büyük veri kümeleri ile çalışmak hem zaman yönünden hem de maliyet yönünden sıkıntılı olduğundan, bu veriyi en iyi şekilde temsil edebilecek bir örneklem ile çalışmak daha doğru bir yöntem olmuştur. Veri indirgeme sırasında sonucu değiştirmeyecek şekilde kullanılan temel yöntemler aşağıdaki gibidir:

⁶⁸ Aydın, **age**, 32.

⁶⁹ Argüden, **age**, 23.

⁷⁰ Demirel, **age**, 24.

⁷¹ Tüzüntürk, **age**, 69.

⁷² Aydın, **age**, 33.

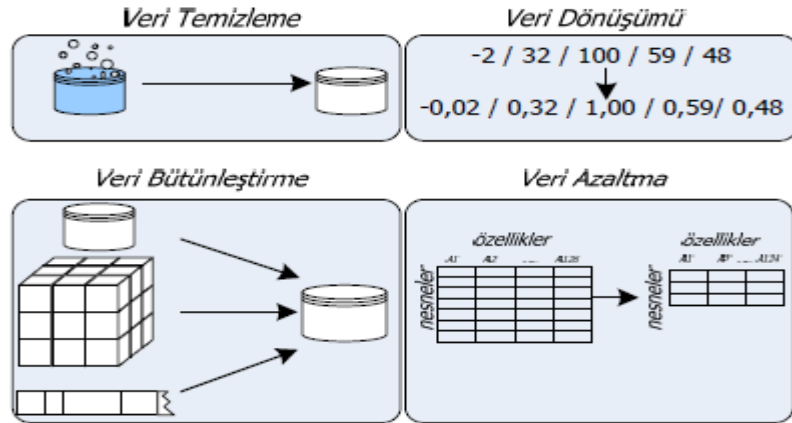
⁷³ Arabacı, **age**, 12.

Veri Küpü Birleştirme: Veriyi indirgeme aşamasında verilen çok boyutlu veri küpleri biçiminde dönüştürmek söz konusu olabilir. Veriler arasında bir seçme işlemi yapılarak, gereksiz veriler veri tabanından çıkarılır ve boyut azaltılması sağlanabilir⁷⁴. Aylık satış fiyatlarının yıllık temelde daha küçük veri seti haline dönüştürülmesi örnek olarak verilebilir.

Boyut Azaltma: Gereksiz olan özelliklerin azaltılması pek çok veri madenciliği algoritmasının daha verimli çalışmasını ve veri madenciliği algoritmaları için gerekli olan işlemci süresi ve hafızasını azaltır⁷⁵.

Veri Sıkıştırma: Veri sıkıştırma yönteminde ise; orijinal verileri en iyi şekilde temsil edecek olan veriler veri şifreleme veya veri dönüşümü yöntemi ile sıkıştırılarak veri indirgenir⁷⁶.

Büyük Sayıların Azaltılması: Veri hacmi parametrik veya parametrik olmayan yöntemler kullanılarak azaltılır. Parametrik yöntemlerde gerçek veri yerine sadece veri parametreleri saklanır ve sıkıştırılan veriyi tahmin etmek için bir model kullanılır. Parametrik olmayan yöntemlere ise veri indirgeme yöntemlerine histogramlar, kümeleme ve örnekleme gösterilebilir⁷⁷.



Şekil 3.6: Veri Ön Hazırlık

Aydın, age, 27.

⁷⁴“Veri Madenciliği”, www.egitim.aku.edu.tr/verimadenciligi.ppt [11.12.2011].

⁷⁵ Aydın, age, 34.

⁷⁶ Demirel, age, 26.

⁷⁷ Aydın, age, 34.

Sonuç olarak, veri ön hazırlığının hedefi, veri ambarı kullanarak, ya da var olan veri tabanlarını veri madenciliğine elverişli hale getirerek, veri madenciliği uygulamalarının performansını arttırmaktır.

3.5. Model Oluşturma

Bu aşama uygun modelleme tekniğinin seçimi, test tasarımının üretimi, model geliştirme ve tahmin işlemlerini içermektedir. Uygun modellerin seçilip uygulanmasıyla birlikte parametreler en uygun değişkenlere dönüştürülmektedir. VM, her problem tipi için farklı yöntemler içermektedir⁷⁸. Model kurma veri madenciliğinin çekirdeğidir. Modeli doğru bir şekilde kurabilmek için yapılacak projenin amacı çok iyi bir şekilde kavranmış olmalıdır. Her amaç ile ilgili birden fazla algoritma mevcuttur. Bu durumda eldeki veriler üzerinde uygun algoritmaların hepsi çalıştırılır ve en doğru sonucu veren algoritma kullanılır⁷⁹.

3.5.1. Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modeller, tahmin edici (*Predictive*) ve tanımlayıcı (*Descriptive*) olmak üzere iki ana başlık altında incelenmektedir⁸⁰.

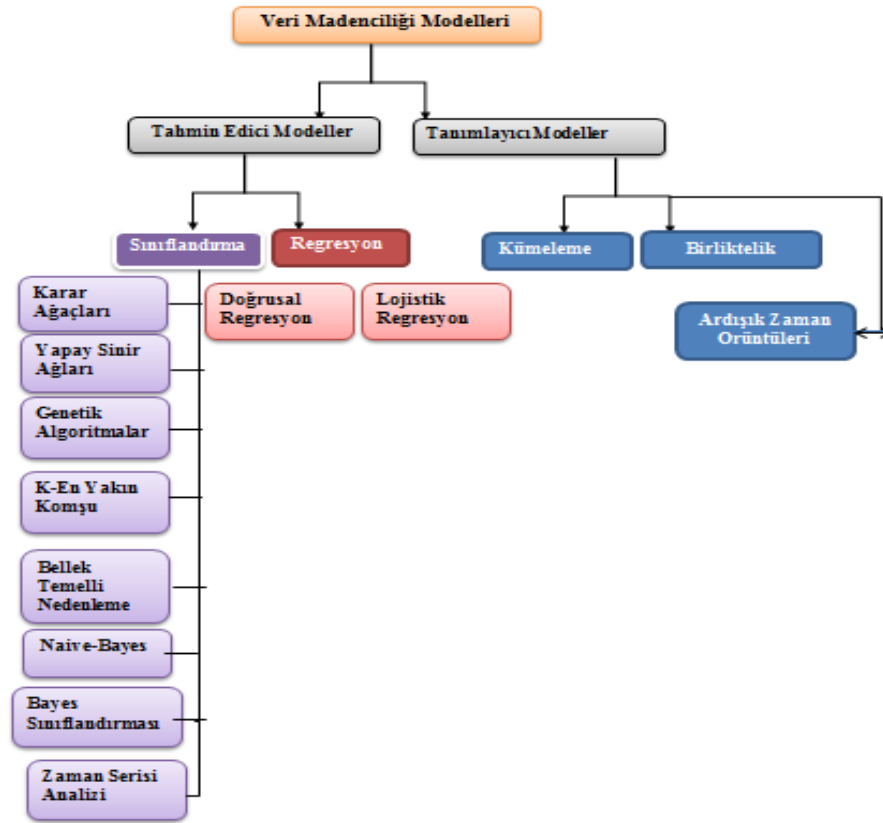
Literatürdeki ilgili kaynaklar incelendiğinde Veri Madenciliği yöntemlerini denetimli ve denetimsiz olmak üzere iki ana kategoriye ayırmak da mümkündür. “Veri Madenciliğinde iyi tanımlanmış veya kesin bir hedef olduğunda denetimli (supervised) ifadesi kullanılır. Elde edilmesi istenen sonuç için özel bir tanımlama yapılmamışsa veya belirsizlik söz konusu ise denetimsiz (unsupervised) ifadesi kullanılır”⁸¹. Esasında içerikleri aynı olup sadece tanımsal olarak değişen denetimli modeller, tahmin edici; denetimsiz modeller de tanımlayıcı modeller grubunu ifade etmektedir.

⁷⁸ Albayrak, **age**, 37.

⁷⁹ Tekerek, **age**, 3.

⁸⁰ Serhat Özkes, “Veri Madenciliği Modelleri Ve Uygulama Alanları”, **İstanbul Ticaret Üniversitesi Dergisi**, c.3, s.3 (2003): 67.

⁸¹ Koyuncugil, **age**, 26.



Şekil 3.7: Veri Madenciliği Modelleri

A) Tahmin edici modeller:

Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır⁸². Bu tahmin, kronolojik zamana göre gelecek durumun tahminini içermekle birlikte, bir ya da daha çok değişkenin başka değişkenler cinsinden tahmin edilmesini olanaklı kılan ilişkiler de sunabilmektedir⁸³. Örneğin, bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken değeri ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır.

⁸²N. Zhong, L. Zhou, "Methodologies for Knowledge Discovery and Data Mining", Third Pacific-Asia Conference, April 26-28, 1999 (China, 1999): 1-3' den aktaran Özkes, **age**, 67.

⁸³ Türkiye Bilişim Derneği Kamu Bilgi İşlem Merkezleri Yöneticileri Birliği Kamu Bilişim Platformu, **age**, 80.

Model kurulurken geçmiş deneyimlerde, faktörlerin aldığı değerlere göre elde edilen sonuçlar girdi olarak kullanılır. Tahmin edilen sonuçların kalitesi (ne kadar iyi tahmin edildiği) tahmin edilen sonuç kadar önemlidir. Çoğunlukla tahmin edilen sonuç ile birlikte, bu sonucun kalitesine yönelik; güvenlik aralığı, olasılığı, vb. değerleri belirlenir⁸⁴.

Sınıflama ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin eden modelleri kurabilen iki veri analiz yöntemidir. Sınıflama kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır⁸⁵.

3.5.2. Sınıflandırma (Classification)

En yaygın uygulanan VM tekniklerinden biri olan sınıflama, sınıfı tanımlanmış mevcut verilerden yararlanarak sınıfı belli olmayan verilerin sınıfını tahmin etmek için kullanılan VM modelidir. Sınıflama iki adım içeren bir işlemdir. Birinci adımda tahmin için kullanılacak bir model oluşturulmaktadır. İkinci adımda, oluşturulan bu model sınıfı belli olmayan veriler üzerinde uygulanarak sınıflar tahmin edilmektedir⁸⁶. Başka bir ifade ile yeni bir nesnenin niteliklerini inceleme ve bu nesneyi önceden tanımlanmış bir sınıfa atamaktır⁸⁷. Sınıflandırmaya örnek olarak kredi kartı başvurularını düşük, orta ve yüksek risk grubu olarak ayırmak veya segmentasyon yapmak gösterilebilir.

Verilerin sınıflandırılması için belirli bir süreç izlenir. Öncelikle var olan veri tabanının bir kısmı eğitim amacıyla kullanılarak sınıflandırma kurallarının oluşturulması sağlanır. Daha sonra bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir. Temeli öğrenme algoritmasına dayanır. Veri tabanından rastgele seçilecek olan örnek veriler üzerinde test edilerek bir sınıflandırma modeli oluşturulur. Test verileri üzerinde sınıflandırma kuralları belirlenir ve örnek veri test edilir. Elde edilen modelin doğruluğu kabul edilirse, bu model diğer veriler üzerinde de uygulanır⁸⁸.

⁸⁴ Argüden, **age**, 37.

⁸⁵ Çankırı, **age**, 155.

⁸⁶ Albayrak, **age**, 38.

⁸⁷ Türkiye Bilişim Derneği Kamu Bilgi İşlem Merkezleri Yöneticileri Birliği Kamu Bilişim Platformu, **age**, 79.

⁸⁸ Demirel, **age**, 50.

Bir deneme kümesi modelin doğruluğunu belirlemek için kullanılır. Genellikle verilen veri kümesi öğrenme ve deneme kümesi olarak ikiye ayrılır. Öğrenme kümesi modeli oluşturulmasında, deneme kümesi modelin doğrulanmasında kullanılır. Başka bir örnek olarak “Genç kadınlar küçük araba satın alır, yaşlı, zengin erkekler büyük, lüks araba satın alır” örüntüsü elde edilebilir. Amaç bir malın özellikleri ile müşteri özelliklerini eşlemektir. Bu örneğe bakıldığında, bir otomobil satıcısı şirket geçmiş müşteri hareketlerinin analizi ile yukarıdaki gibi iki kural bulursa genç kadınların okuduğu bir dergiye reklam verirken küçük modelinin reklamını verir⁸⁹.

Sınıflama modellerinde kullanılan başlıca teknikler şunlardır⁹⁰:

- 1 - Karar Ağaçları (Decision Trees)
- 2- Yapay Sinir Ağları (Artificial Neural Networks)
- 3- Genetik Algoritmalar (Genetic Algorithms)
- 4- K-En Yakın Komsu (K-Nearest Neighbor)
- 5- Bellek Temelli Nedenleme (Memory Based Reasoning)
- 6- Naive-Bayes
- 7- Bayes Sınıflandırması
- 8- Zaman Serisi Analizi

3.5.2.1. Karar Ağaçları (Decision Trees)

Karar ağacı metodu, ilk olarak Grochtmann ve Grimm tarafından, bir kara kutu test tekniği olarak geliştirilmiştir. Temel konsepti, bir programın giriş etki alanını, parçalanmış farklı sınıf setlerine bölmek ve daha sonra test aşamalarını formüle etmek için her bir seti aynı sınıf içinde tekrar birleştirmektir. Bu metot aşağıdaki adımları içermektedir⁹¹:

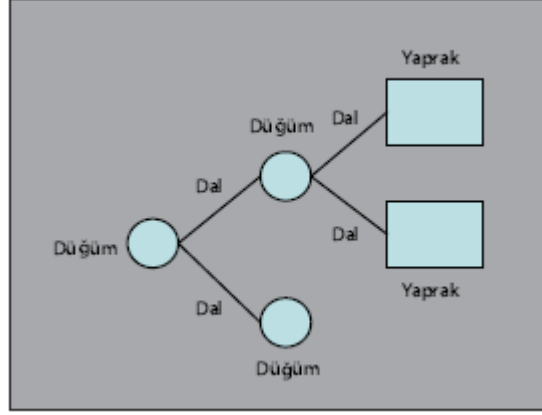
1. Tanımlamalardaki tüm sınıflandırmaları ve onların ilgili sınıflarını tanımla.
2. Sınıflandırmalardaki veri setlerinden ve ilgili sınıflarından, sınıf ağacı yapısını oluştur.
3. Sınıflandırıcı ağacın birleştirme tablosunu oluştur.
4. Birleştirme tablosundaki sınıflardan tüm uygun sınıfları tanımla. Sınıflardaki her bir kombinasyon, bir test aşamasını ifade eder.

⁸⁹ Argüden, **age**, 37.

⁹⁰ Özekes, **age**, 68.

⁹¹ T.Y. Chen, P.L. Poon, “Construction Of Classification Trees Via The Classification-Hierarchy Table”, **Information and Software Technology**, c. 39 (1997): 889.

Karar ağacı, adında belirtildiği şekilde ağaç görünümünde bir tekniktir. Karar düğümleri (kök düğüm), dallar ve yapraklardan oluşur.



Şekil 3.8: Karar Ağacı Genel Yapısı

Argüden, **age**, 47.

Karar ağaçları, kurgulanmasının, yorumlanmasının ve veri tabanları ile entegrasyonun kolaylığı nedeniyle en yaygın kullanılan öngörü yöntemlerinden yani sınıflandırma tekniklerinden biridir. Güvenilirliklerinin iyi olması da bir başka tercih edilme nedenidir. Karar ağaçlarının hedefi bağımlı değişkendeki farklılıkları maksimize edecek şekilde veriyi sıralı bir biçimde parçalarına (farklı gruplara) ayırmaktır⁹². Ağaç, bağımlı değişkenin varyansı, grupların arasında maksimum ve grup içinde minimum olacak şekilde inşa edilmiştir⁹³.

Karar ağaçları esas itibarı ile kural çıkarma algoritmalarıdır. Bu algoritmalar bir veri kümesinden kullanıcıların çok kolay anlayabileceği (IF-THEN) türündeki kuralları bir ağaç yapısında türetebilirler⁹⁴.

Karar ağacının araştırdığımız sınıfı sağlayan başlangıç düğümüne kök düğüm, ara safhalardaki düğümlere yaprak düğüm, ağacın bittiği düğümlere ise son düğüm denir. Her bir düğümdeki gözlem sayısı düğümün büyüklüğünü ifade ederken, ağaçtaki dallanma sayısı ağacın derinliğini gösterir. Şekil 3.9’ daki örnekte ağacın derinliği 4’ tür. Ağacın bir ürünü alan kişilerin özelliklerini araştırdığını düşünelim.

⁹² Argüden, **age**, 47.

⁹³ Susan Chiu, Domingo Tavella, **Data Mining and Market Intelligence for Optimal Marketing Returns**, 1.edition (USA: Elsevier, 2008), 157.

⁹⁴ Baykasoğlu, **age**, 6.

Bu durumda ağacı kısaca şöyle yorumlayabiliriz. Bu kişiler için en önemli özellik yaş olarak gösterilmiş, dallanma bu değişkenden başlamıştır. Bir sonraki düğümde ise 30 yaş altı grup için çocuk sayısının önemli bir gösterge olduğu, 30 yaş üstü için ise firmada yaptığı cironun ayırıcı bir özellik olduğunu görebiliriz. Aynı yorumlar diğer düğümler için de geçerli olmaktadır⁹⁵.



Şekil 3.9: Karar Ağacı Örneği

Mine Çelik, “Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama” Yüksek Lisans Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı, 2009), 22.

Karar ağacı algoritmaları analitik metotlardır ve kullanılan algoritma türüne bağlı olarak veri kümesini ardışık bölünmeler gerçekleştirerek iki veya daha fazla dala ayırarak karar ağacı yapısını oluşturur. Ağaç yapısı, kök düğüm ile başlar ve her bir ayrılma sonucu yaprak düğüm (iç düğüm) oluşumu ile devam eder. Her bir iç düğüm, sonucu olasılıkları ile temsil eden bir kararı gösterir⁹⁶. Karar ağaç tabanlı bir model, veriye uygulandığında; her kayıt, o kayıt yaprak ya da son düğüme ulaşana kadar, bir dizi test tarafından belirlenmiş yol boyunca ağaçta akar. Burada kesikli ya da sürekli

⁹⁵ Mine Çelik, “Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama” Yüksek Lisans Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı, 2009), 23.

⁹⁶ Ramesh Konda, “Predicting Machining Rate in Non-Traditional Machining using Decision Tree Inductive Learning” (Doctorate of Philosophy, Nova Southeastern University, Computer Information Systems Graduate School of Computer and Information Sciences, April 2010), 14.

bir deęer ile sonulandırılır. Bir karar aęacı, kararların her bir dęümde yapıldığı ve bir dizi kurallara evrildiğı yukarıdan ařağıya inen bir aęaç yapısında olur⁹⁷.

Bir karar aęacı algoritmasının prensipte görevi, veriyi özyinelemeli olarak alt veri gruplarına dallanma yaparak bölmektir. Temel olarak iki adımdan oluşur. Birinci adım aęacın oluşturulması, dięer adım ise veri tabanındaki her bir kaydın bu aęaca uygulanarak verilerin sınıflandırılmasıdır⁹⁸.

Test verisine uygulanan bir modelin doęruluęu, yaptığı doęru sınıflamanın test verisindeki tüm sınıflara oranıdır. Her test örneğinde bilinen sınıf, model tarafından tahmin edilen sınıf ile karşılaştırılır. Eęer modelin doęruluęu kabul edilebilir bir deęer ise model, sınıfı bilinmeyen yeni verileri sınıflama amacıyla kullanılabilir⁹⁹. Ortaya ıkan karar aęacı, yeni örneklerle doęruluęunu deęerlendirmek için sonra test veri setine uygulanır. Bir karar aęacı eęitim setine uygulandığında, yeni bir veri onun sınıflandırma doęruluęunu azaltabilir. O zaman aęaç, gerek hayattaki uygulamada daęıtılmadan önce, budanmış olmalıdır¹⁰⁰.

Sınıflandırma aęaçları, yaygın olarak, ilaç, bilgisayar bilimleri, botanik ve psikoloji gibi farklı alanlarda kullanılır. Bu aęaçlar, özellikle kendilerinin grafiksel olarak görüntülenmesini saęlayıp, sadece sayısal bir yorum bile olsa, bunun yorumlanmasını kolaylařtırmaktadır¹⁰¹.

Başlıca Karar Aęacı algoritmaları; CHAID, C4.5, CART, ID3, SLIQ, SPRINT ve QUEST olup tez kapsamında ele alınmış ve bilgi verilmiştir.

CHAID Algoritması (Otomatik Ki-Kare Etkileşim Belirleme Analizi):

alışmamız kapsamında, CHAID algoritması yöntemi tercih edilmektedir. Bu nedenle, algoritmayı detaylı bir şekilde ele almak yerinde olacaktır.

⁹⁷ Jong Hwan Suh, Chung Hoon Park, Si Hyun Jeon, “Applying text and data mining techniques to forecasting the trend of petitions filed to e-People”, **Expert Systems with Applications**, c. 37 (2010): 7259.

⁹⁸ Demirel, **age**, 53.

⁹⁹ Özekes, **age**, 69.

¹⁰⁰ Suh, **age**, 7259.

¹⁰¹ M. Ramaswami, R. Bhaskaran, “A CHAID Based Performance Prediction Model in Educational Data Mining”, **International Journal of Computer Science Issues**, c. 7, s. 1 (2010):14

CHAID' in açılımı Ki-Kare(x^2) Otomatik Etkileşim Detektörü olarak tanımlanabilir. Ki- kare olmasının sebebi, algoritmanın cross- tab ve istatistiksel anlamlılık oranları ile çalışmasıdır. Ağaç diyagramı çizilirken, en önemli ilişkiler seçilir¹⁰².

Hedef değişkene göre istatistiksel olarak önemli bir fark olmadığında, herhangi bir düğümdeki (node) en iyi ayırmayı belirlemek için tahmin edici değişkenin uygun kategori çiftleri birleştirilir. Bu birleşme hiçbir önemsiz çift kalmayana kadar devam eder. Bu değişkenlerden en iyi açıklayıcı değişken seçilir ve düğüm ayrılır. CHAID analizi, kategorik değişkenlere ilişkin veri kümesini ve bağımlı değişkeni en iyi açıklayabilecek şekilde ayrıntılı homojen alt gruplara böler¹⁰³.

Chaid algoritması, sadece nominal veya sıralı kategori belirleyicilerini kabul eder. Belirleyiciler sürekli olduğunda, bir sonraki algoritmayı kullanmadan önce sıralı belirleyicilere dönüştürülür. Her bir tahmin edici değişken X için, anlamlı olmayan kategorileri birleştirir. Her bir son X kategorisi, eğer X, düğümü bölmek için kullanılırsa, küçük bir düğüm ile sonlanır. Birleştirme adımı ayrıca, bölücü adımlarda kullanılan düzenlenmiş p değerini hesaplar¹⁰⁴.

Aslında, CHAID bir popülasyonu farklı ve kesin segmentlere yinelemeli olarak parçalayan (bölen) bir tekniktir. Düğümler olarak adlandırılan bu segmentler, (kategorik) sorumlu değişkenin varyasyonun segmentler içinde minimize ve segmentlerin arasında maksimize edildiği düğümlerdir. Popülasyonun başlangıç düğümünü iki ya da daha fazla düğüme ayırdıktan sonra, (bağımsız ya da tahmin edici değişkenin değeri olarak tanımlanır), bölme prosesi her bir düğüm için tekrarlanır. Ta ki durdurma kuralları ile karşılaşırsa; örneğin sınıf değeri aynı çıkarsa ya da bölmede tek bir nesne kalırsa, proses sona erer. CHAID tahmin etme modelinin çıktısı, hiyerarşik ağaç yapısı formunda gözüktür. Bu yapıya göre, kök popülasyondur ve dallar segmentleri bağlar.

CHAID tahmin etme model yapısında temel adım, sınıflandırmanın ilgili kriterlerini seçmektir. Kriter seçim tekniklerinin amacı, hesaplama zamanının azalmasına yardım eder ve modelin öngörü doğruluk oranını artırır. Ki- Kare, bir kriter

¹⁰² Ray Hoare, "Using CHAID For Classification Problems", **New Zealand Statistical Association 2004 Conference** (2004): 1-5.

¹⁰³ Ömer Satıcı, Zeki Akkuş, Aykut Alp, "Tıp Fakültesi Öğretim Elemanlarının Teknolojiye İlişkin Tutumlarının CHAID Analizi İle İncelenmesi", **Dicle Tıp Dergisi**, c. 36. s. 4 (2009): 269.

¹⁰⁴ Mevlüt Türe A, Füsün Tokatlı, İmran Kurt, "Using Kaplan–Meier Analysis Together With Decision Tree Methods (C&RT, CHAID, QUEST, C4.5 And ID3) In Determining Recurrence-Free Survival Of Breast Cancer Patients", **Elsevier**, c. 36(2009): 2020.

oluşumunun, sınıf değerinin bağımsızlığı olduğunu kabul ederse; beklenen dağılımdan sapmayı ölçen temel istatistik testi olmaktadır. Pearson ki- kare (χ^2), kriter seçiminde yaygın olarak kullanılan bir metottur ve ki- kare istatistiği ile ölçerek kriterleri tek tek değerlendirir¹⁰⁵.

Adımsal olarak uygulanan benzer kategorileri birleştirme işlemi, değişkenler arasında daha fazla birleştirme sağlanamayacağına istatistiksel olarak karar verilinceye kadar devam eder. Değişkenlerin bölünmeye uygun olup olmadığına, Bonferroni düzeltilmiş p değeri kullanılarak karar verilir¹⁰⁶.

Bonferroni yaklaşımı, her bir grubun ortalama vektörlerin genel ortalama vektörlerinden farkları bulunduğundan sonra, bu farkların sıfır olup olmadığına dayanır.

Genel ortalama vektörü $\bar{x}$ ve her grubun i . değişkenine göre ortalama vektörleri aşağıdaki gibi gösterilir:

$$\bar{x} = \begin{bmatrix} \bar{x}_{1.} \\ \bar{x}_{2.} \\ . \\ . \\ \bar{x}_{p.} \end{bmatrix} \quad \bar{x}_{.i} = \begin{bmatrix} \bar{x}_{1i} \\ \bar{x}_{2i} \\ . \\ . \\ \bar{x}_{pi} \end{bmatrix}$$

Şekil 3.10: Bonferroni Yaklaşımı Ortalama Vektörü

k . grup ile e grup i . değişken ortalamaları arasındaki ortalama farkları arasındaki $1-\alpha$ güven aralığı aşağıdaki biçimde hesaplanır:

$$(d_{ki} - d_{ei}) = (\bar{x}_{ki} - \bar{x}_{ei}) \pm t \left(\frac{\alpha}{pg(g-1)}, (N-g) \right) \sqrt{\left(\frac{1}{n_k} + \frac{1}{n_e} \right) \frac{w_{ii}}{N-g}} \quad (3.1)$$

¹⁰⁵ Ramaswami, , age, 15.

¹⁰⁶ İsmet Doğan, “Holştayn Irkı İneklerde Süt Verimine Etki Eden Faktörlerin CHAID Analizi İle İncelenmesi”, **Afyon Ankara Üniversitesi Veterinerlik Fakülte Dergisi**, c. 50, (2003): 66.

Burada $N=n_1 + n_2 + \dots + n_g$, p değişken sayısı, g grup sayısı, ve W_{jj} , W matrisinin köşegen elemanlarıdır. W matrisi, gruplar içi değişimi gösterir ve;

$$W = \sum_{i=1}^{g n_i} \sum_{j=1}^{g n_i} (X_{ij} - X_j) \cdot (X_{ij} - X_i) \quad (3.2)$$

g = grup sayısı, n_i = i . gruptaki veri sayısı,

İfadesi ile hesaplanır. Her bir değişken için, gruplar ikiye ayrılarak dikkate alınır ve eşitlik kullanılarak i . değişken için elde edilen aralığın sıfır değerini içerip içermediği kontrol edilir. Eğer sıfır değeri belirlenen aralıkta yer alıyorsa, ilgili gruplar arasında anlamlı bir farklılık olmadığı, aksi durumda grupların farklı oldukları şeklinde yorumlanır.

CHAID algoritması, 3 bölümden oluşur: Birleştirme, Bölme ve Durdurma¹⁰⁷. Karar ağacının oluşturulmasında kullanılan CHAID algoritmasının işleyişi aşağıda adımsal olarak ifade edilmiştir.

1.Adım: Her bir tahmin edici değişken X için, X 'in, Y hedef değişkenini dikkate alan en az öneme sahip kategori çiftini bul (en büyük p değerine sahip)¹⁰⁸.

Yöntem, Y 'nin ölçüm düzeyine bağlı olarak p değerlerini hesaplayacaktır.

- Eğer Y sürekli ise F testini kullan.
- Eğer Y nominal (isimsel) ise, X 'in kategorileri satırlarda ve Y 'nin kategorileri sütunlarda olacak biçimde iki yönlü çapraz tablo düzenle. *Pearson ki-kare testini* veya *olabilirlik oranı testini* kullan.
- Eğer Y sıralı ise, bir Y birliktelik modeli uygula. Olabilirlik oranı testini kullan.

2.Adım: En büyük p değerine sahip X 'in kategori çifti için, p değerini önceden belirlenmiş alfa düzeyi *birleş* α ile kıyasla.

- Eğer p değeri *birleş* α 'den küçük ise bu çifti bir tek kategori altında birleştir.

¹⁰⁷ Lukas Kleppin, Roland Pesch, Winfried Schro, "CHAID Models On Boundary Conditions Of Metal Accumulation In Mosses Collected In Germany in 1990, 1995 and 2000", **Atmospheric Environment Elsevier**, c. 42 (2008): 5222.

¹⁰⁸ Demirel, **age**, 59.

X 'in yeni kategori kümesi için süreci Adım 1'den başlat.

- Eğer p değeri α 'den büyük ise Adım 3'e git.

3.Adım: X 'in ve Y 'nin kategori kümesi için uygun Bonferroni çarpanını kullanarak, düzeltilmiş p değerini hesapla¹⁰⁹.

4.Adım: En küçük düzeltilmiş p değerine sahip X tahmin edici değişkenini seç (en önemli olan). Bunun p değerini önceden tanımlanmış alfa düzeyi α ile kıyasla.

- Eğer p değeri, α değerinden küçük veya eşit ise düğümü X 'in kategori kümesini temel alarak böl.
- Eğer p değeri, α değerinden büyük ise düğümü bölme. Bu düğüm uç düğümdür. Ağaç büyütme sürecini durma kuralları görülene kadar sürdür.

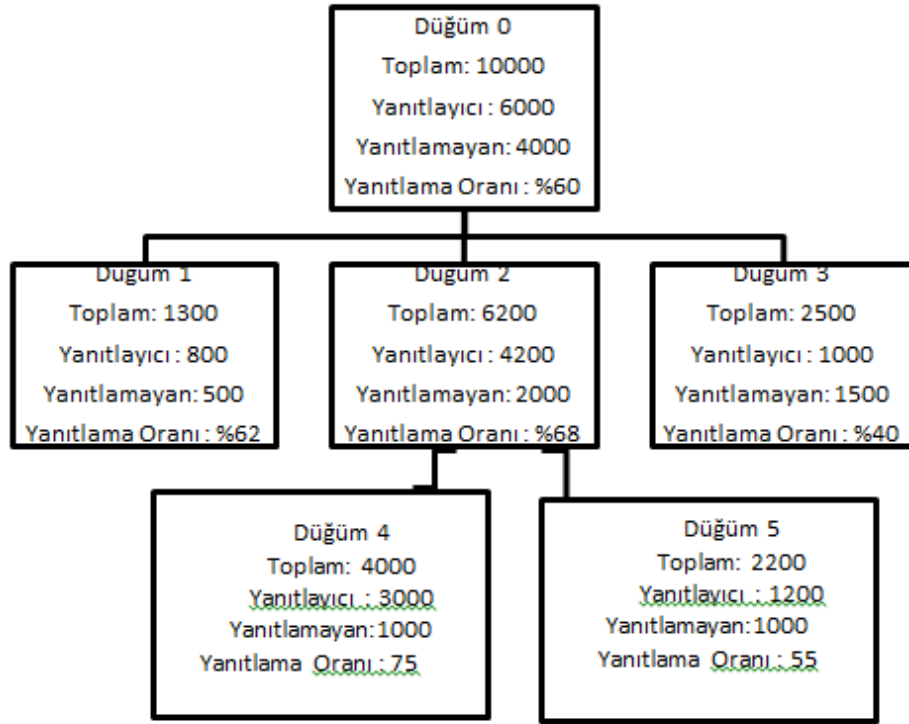
CHAID algoritmasını daha iyi anlamak için, doğrudan mail tanıtımına yüksek oranda ilişkin yanıt veren üyelerin nitelikleri ile ilgili bir örnek verilebilir. Bağımlı değişken, “Evet” veya “Hayır” olarak mail yolu ile tanıtımına ilişkin “var” ya da “yok” kavramına sahip iki cevaplı bir değişkendir. Bir pazar örneğinde, 10.000 birey olduğunu varsayalım. Bunlardan 6000 tanesi cevap vermiş, 4000 tanesi de cevap vermemiş olsun. Neticede, doğrudan mail ile tanıtıma yanıt verme oranı % 60 olur. Cevaplayıcı niteliklerine ilişkin 3 ayrı bağımsız değişken olduğunu var sayalım: Yaş, Gelir, Cinsiyet. CHAID algoritması tüm değerleri; yaş, gelir ve cinsiyete göre, karar ağacının kırımlarını yapmak için önemli χ^2 istatistikleri ile böler¹¹⁰.

Aşağıdaki şekil (1, 3, 4, 5 düğümleri) 4 kırımlı olan ağacı yani CHAID çıktısını göstermektedir. Yaş ve cinsiyetin, mail ile tanıtıma yanıt verme ile ilişkin olduğu görülmektedir. 25 ve 45 yaş arasındaki kadın grubu, aynı yaş grubundaki erkek bireyler ile beraber en duyarlı gruplardır. 45 yaş üzeri grup ise cinsiyetine bakılmaksızın, yanıt vermeye en az meyilli gruptur¹¹¹.

¹⁰⁹ K. P. Soman, S. Diwakar, “Insight into Data Mining Theory and Practice”, (New Delhi, PHI Learning Pvt. Ltd. , 2006),72.

¹¹⁰ Chiu, **age**, 158

¹¹¹ **Age**, 159.



Şekil 3.11: Doğrudan Mail Tanıtımına Cevap Verenlerin CHAID Analizi Çıktısı

Susan Chiu, Domingo Tavella, **Data Mining and Market Intelligence for Optimal Marketing Returns**, 1.bs. (USA: Elsevier, 2008), 157.

CHAID analizi, değişkenlerdeki varyasyonu gruplar arası maksimum ve gruplar içi minimum olacak şekilde farklı alt bölümlere tekrarlı olarak ayırarak işlem yapar. Bu işlemler gerçekleşirken CHAID analizi bazı varsayımları dikkate almadan da işlem yapabilir. Buna CHAID analizinin avantajları olarak bakılabilir¹¹².

Diğer karar ağacı algoritmalarından farklı olarak, CHAID'in avantajı sonuçları okuması ve yorumlamasının kolay olmasıdır. Ağaç tabanlı segmentasyon tekniği olarak, nominal (yazılı), sıralı ve sürekli değişkenler arasındaki etkileşimi tespit edebilir ve sonuçları hiyerarşik ağaç yapısında göstermesidir. Varsayılan olarak çok yönlü bölünmeyi kullandığı için ve küçük boyutlu örnekler hızla küçülebileceği ve anlamlı sonuçlar vermeyeceği için etkin çalışması için büyük örnekler gereklidir¹¹³.

Aşağıdaki tablo 3.1'de bir lisedeki mezuniyet için tanımlanmış notların, hangi aralığa düştüğünü görmek için yapılandırılmış bir CHAID modeli görülmektedir.

¹¹² Satıcı, **age**, 269.

¹¹³ Kleppin, **age**, 5221.

Yapılan işlemler sonucu, 11 terminal düğüm yani, 11 sınıf kuralı bulunmuştur. “Lise Mezuniyet Seviyesi = A ” için belirlenen kural; önceki akademik başarı ve orta öğretim seviyesindeki akademik başarı, yükseköğretim düzeyindeki akademik başarının üzerinde etkiye sahiptir anlamını göstermektedir. Bu matristen 7 tane mezuniyet not sınıf kategori çıktısına göre öğrenci sayısını gösterdiği, sadece A,B ve C grupları için modelin %50’den yüksek doğrulukta çıktı gösterdiği söylenebilir¹¹⁴.

Tablo 3.1: Sınıflandırma Matrisi ve CHAID Tahmin Edici Modeli

Lise Mezuniyet Seviyesi- Not		Beklenen							Tahmin Doğruluğu %
		O	A	B	C	D	E	F	
Gözlenen	O		33	4					0.00
	A		90	61	8				56.60
	B		46	149	62				58.00
	C		4	87	105				53.60
	D			21	46	1			0.00
	E			1	2				0.00
	F			21	31				0.00

Ramaswami, **age**, 17.

CHAID analizi ile elde edilecek bir regresyon denklemi, bilinen klasik varsayımlardan (normallik, doğrusallık, homojenlik vb.) bağımsız tutulmaktadır. Çünkü güçlü bir öteleme algoritması ile bütün olan evren kararlı alt düğümlere (node) bölünebilmektedir. Bu işlem ayrıca verilerin dağılımında normalliği ve homojenliği sağlayabilmektedir. Ayrıca CHAID analiziyle sürekli ve kategorik veriler, aynı anda modele dahil edilebilmektedir¹¹⁵.

Sonuç olarak, ağaç yapısı ve kolay kural çıkarımına olanak tanınmasıyla oldukça yararlı bir tekniktir. Bu bağlamda karar ağaçlarının daha çok tıp, endüstri ve

¹¹⁴ Ramaswami, **age**, 17.

¹¹⁵ Engin Üngüren, Hulusi Doğan, “Beş Yıldızlı Konaklama İşletmelerinde Çalışanların İş Tatmin Düzeylerinin Chaid Analiz Yöntemiyle Değerlendirilmesi”, **C.Ü. İktisadi ve İdari Bilimler Dergisi**, c. 11, s. 2 (2010), 42.

mühendislik bilimlerinde yaygın olarak kullanıldığı bilinmektedir. Ülkemizde ise karar ağacı algoritmalarının kullanımı oldukça yenidir.

C&RT (Sınıflandırma ve Regresyon Ağacı):

C&RT (CART) algoritmaları, bağımlı değişkenin kategorik olduğu durumlarda sınıflandırma, sürekli olduğu durumlarda tahminleme modeli kuran bir karar ağacı algoritmasıdır. C&RT algoritmaları için birincil amaç, mümkün olan en iyi doğruluğu olan modeli kurabilmektir. En iyi doğruluk ise minimum maliyetli tahminler yapılmasını içerir. C&RT her bir adımda tahminin doğruluğuna en fazla katkısı olan ayrımı yaparak ilerler. Ayrım ölçütü olarak Gini indeksi veya x^2 gibi ölçütler kullanır¹¹⁶.

C&RT, dallanması sürecinde, tekrarlanan ikili bölümlmeye göre tahminleme yapar. C&RT' nin klasik doğrusal ve lojistik regresyon algoritmalarına göre potansiyel avantajı, parametrik istatistik varsayımlarına bağlı olmayan yani parametrik olmayan ve doğrusal olmayan bir metot olmasıdır. Bundan dolayı, C&RT, belirli bir çıktıyı tahmin etmek için en önemli tahmin edici değişkenler arasındaki ilişkiyi göz önüne almadan çok sayıda değişken arasından, tahmin edici değişkeni seçebilir¹¹⁷.

ID3:

ID3, sayısal olmayan değerlere sahip veri kümelerini sınıflandırmak için etkili bir ayrıştırma ağacına sahip olan sentez bir prosedürü olan karar ağacı algoritmasıdır. ID3 yaklaşımı, etiketli verinin kullanımını sağlar ve tüm etiketli örneklerin düzenli bir şekilde sınıflandırılmasına kadar, özelliklerin nasıl incelenebileceğine karar verir.

ID3, kesin şartlar altında geçerlidir ve kapsamı dışındaki alanlarda geçersizdir. ID3 yaklaşımı için gerekli olan 2 kesin sınır şöyledir¹¹⁸:

1. Sadece kategorik veri kullanır.
2. Kısmi bilgi ile baş edemez.

¹¹⁶ Çelik, age, 30.

¹¹⁷ Maria Chudzinska, Danuta Baralkiewicz, "Application Of ICP-MS Method Of Determination Of 15 Elements In Honey With Chemometric Approach For The Verification Of Their Authenticity", **Food And Chemical Toxicology**, v.49 (2011): 2745.

¹¹⁸ Sukumar Chakraborty, "Rid3: An Id3-Like Algorithm For Real Data", **Information Sciencies**, c. 96 (2000), 272.

C4.5:

C4.5 algoritması, ID3 algoritmasının bir uzantısıdır ve budama metodolojisi ile sayısal nitelikleri, kayıp değerleri ve gürültülü verileri işlemeyi kapsayan “böl ve yönet” yaklaşımını içermektedir. Bölme düğüm stratejisi, bilgi kazanım oranını hesaplamaya dayanmaktadır. Temel fikir, kök düğümünden bu düğüme olan yolda henüz dikkate alınmamış nitelikler arasında, her bir düğümün en bilgilendirici ilgili nitelik ile ilgili bir soruyu tutmasıdır¹¹⁹.

SLIQ:

SLIQ algoritması hem sayısal hem de kategorik verilerin sınıflandırılmasında kullanılabilir. Sayısal verilerin değerlendirilmesindeki maliyeti azaltmak için ağacın oluşturulması sırasında önceden-sıralama tekniği kullanılır. Sayısal verilerle işlem yapılırken en iyi dallara ayırma kriterini bulmak için verileri sıraya dizme önemli bir faktördür. SLIQ algoritmasında kullanılan teknik ise, verileri sıralama işlemini her düğümde yapmak yerine öğrenme verileri sadece bir kere ve o da ağacın büyüme aşamasının başlangıcında yapılarak gerçekleştirilir¹²⁰.

QUEST:

Hesapladığı sınıflandırma ağacının güvenilirliği ve etkinliğini arttırmak için, bir dizi inovatif özellikler ve kuadratik diskriminant analizi kullanan bir sınıflandırma algoritmasıdır. En önemli özelliği hızlı ve tarafsız olmasıdır. Yani, tahmin edici değişkenler onlarca farklı seviyede olduğu zaman özelliklerle aynıymış gibi yaklaşır¹²¹.

SPRINT:

SPRINT algoritması, her bir değişken için ayrı bir liste kullanarak sıraya dizme işlemini sadece bir kez yapar. Bu yönüyle SLIQ algoritmasına benzer ve sözü edilen diğer algoritmalarından ayrılır. Ancak farklı veri yapıları kullanarak SLIQ algoritmasından ayrılır¹²².

¹¹⁹ Jao-Hong Cheng, Huei-Ping Chen, Yi-Min Lin, “A Hybrid Forecast Marketing Timing Model Based On Probabilistic Neural Network, Rough Set and C4.5”, **Expert Systems With Applications**, c. 37 (2010): 1815.

¹²⁰ Şenol Zafer Erdoğan, “Sınıflandırma Teknikleri ve Algoritmaları”, http://Akademik.Maltepe.Edu.Tr/~Senole/Data_Mining/Bolum3.Pdf [22.12.2012].

¹²¹ Hierarchical Nature of Classification Trees
<http://www.statsoft.com/textbook/classification-trees/#character1> [22.12.2012].

¹²² Erdoğan, **age**, 25.

3.5.2.1.1. Karar Ağacı Oluşturmak İçin Kullanılan Dağımlılık (Ayrırma) Kriterleri

Ağacın dallandırılmasını devam ettirecek değişken seçilirken belirli kriterler gözetilmektedir. Kriterler tek değişkenli ve çok değişkenli ayırma kriterleri olarak iki grupta sınıflandırılabilir.

Tek Değişkenli Ayırma Kriterleri:

Kazanım Oranı: Kazanım oranı, kazanım teorisinin başlangıcından gelen kirlilik ölçütü entropiyi kullanan kirlilik temelli bir kriterdir¹²³. Bilgi kazanımı, entropi ilkesine dayanır. Bu yöntem ayırma yöntemi olarak seçildiğinde, algoritma, entropiyi azaltan çözümler üretir. Bilgi (X), X veri kümesindeki herhangi bir kaydı sınıflandırmak için beklenen bilgi; Bilgi_k (X), X veri kümesindeki bir kaydı sınıflandırırken her bir değişken için ihtiyaç duyulan bilgi olmak üzere bilgi kazanımı, açıklanan bu iki değer farkı ile ifade edilmektedir¹²⁴.

$$\text{Kazanım (X)} = \text{Bilgi}_k(X) - \text{Bilgi (X)} \quad (3.3)$$

Değişkenlerin, o ağaca o noktada dallandırma yapıldığında kazandıracakları bilgi hesaplanarak hangi değişkene göre ilerleneceğine karar verilir. Zira en fazla bilgi veren, kazanımı en yüksek olan değişken ile ilerlenir.

Entropi: Karar ağacı algoritmasında ana seçim ağaçtaki her düğümde hangi değişkenin seçileceğidir. Değişkenin değerinin ölçüsü “kazanım oranı yani information gain” adı verilen istatistiki bir değer ile belirlenir. Information gain’i dikkatlice tanımlamak için information teorisinde genel olarak kullanılan ve bir ölçü tanımlayan “entropy” adı verilen bir tanımlama kullanılır. Dalların yapılacağı niteliği bulmak için kullanılan Entropy, örneklerin keyfi olarak toplanmasının kirliliğini (gürültüsünü) karakterize eder. Verilen S topluluğu hedef değişkeninin pozitif ve negatif örneklerini içerir. Formül 1’de S topluluğunun entropy değeri gösterilmiştir.

¹²³ Oded Maimon, Lior Rokach , **Data Mining and Knowledge Discovery Handbook**, (İsrail: Springer, 2005), 170.

¹²⁴ Çelik, **age**, 24.

$$Entropy(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus} \quad (3.4)$$

$p_{\oplus}$, S'deki pozitif örneklerin S'deki bütün örneklere oranıdır ve $p_{\ominus}$, S'deki negatif örneklerin S'deki bütün örneklere oranıdır. Bir değişkenin etkinlik ölçüsü information gain terimi ile kullanılmaktadır¹²⁵.

Entropi, veri tabanındaki verinin sayısallaştırılması, verideki belirsizliğin ölçülmesidir ve 0–1 arasında değer alır. Her değişken için entropi kavramından yararlanarak değişkenlerdeki belirsizlik ölçülür. Sonraki aşamada ise her değişken için kazançlar hesaplanır ve hangi kazanç yüksek ise dallanma o değişkenden yapılır¹²⁶.

Gini: Gini indeksi, hedef özelliğin değerleri arasındaki olasılık dağılımları arasındaki ayrılmayı ölçen kirlilik temelli bir kriterdir. Yaklaşımı bilgi kazanımına benzer. X veri kümesindeki herhangi bir kaydı sınıflandırmak için gini indeksi aşağıdaki şekilde ifade edilir¹²⁷.

$$Gini(X) = 1 - \sum_{i=1}^m p_i^2 \quad (3.5)$$

p_i , X veri kümesindeki bir kaydın, L kümesine ait olma olasılığını gösterir.

Tüm değişkenlerin gini değeri hesaplandıktan sonra, minimum indekse sahip değişken seçilerek karar ağacı dallandırılır. Minimum indeks değeri, maksimum safsızlık anlamına gelir¹²⁸.

Çok Değişkenli Ayırma Kriterleri: Çok değişkenli ayırma kriterlerinde en uygun parçalayan değişkeni bulmak için, sonuca deneme yanılma yolu ile giden, doğrusal programlama, doğrusal diskriminant analizi gibi yöntemler kullanılmaktadır¹²⁹.

¹²⁵ Baha Vural Kök, Necati Kuloğlu, “Sollama Esnasında Taşıt Ve Yol İle İlgili Faktörlerin Karar Ağacı Yöntemiyle İrdelenmesi”, *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi* c. 21, s. 2, (2005): 183.

¹²⁶ Demirel, *age*, 54.

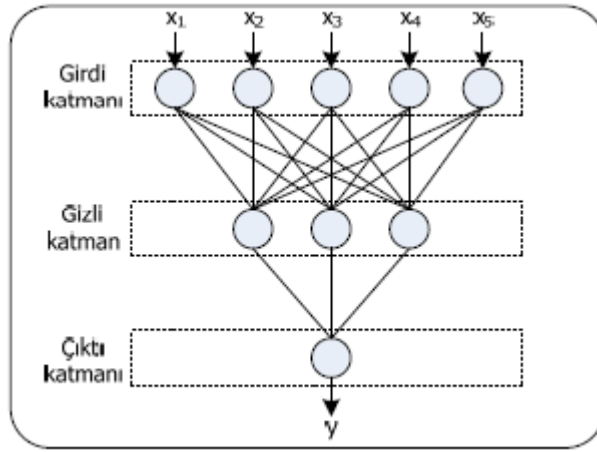
¹²⁷ Maimon, *age*, 170.

¹²⁸ Çelik, *age*, 26.

¹²⁹ Maimon, *age*, 186.

3.5.2.2. Yapay Sinir Ağları (Artificial Neural Networks)

Yapay sinir ağları yöntemi, gerçek sinir ağı yapısından esinlenilerek ortaya çıkarılmıştır. Her bir girdinin bir ağırlığı vardır ve çıktı esnasında, bu ağırlıklı girdilerin lineer olmayan kombinasyonu alınarak, bir transfer fonksiyonundan geçirilir. Veri madenciliğinde kullanılan yapay sinir ağı yöntemleri, genellikle bir veya birden fazla gizli tabakaya sahiptir¹³⁰. Yapay Sinir Ağları algoritması, yapay sinir ağı sistemlerinin yapısının bağlantı ağırlıklarının eş zamanlı optimizasyonu için tasarlanmıştır. Bağlantı ağırlıklarının ayarını iletmek için daha elverişli olduğundan, doğrudan kodlama yaklaşımı aday çözümleri temsil etmek için kullanılır¹³¹.



Şekil 3.12: Çok Katmanlı İleri Yayımlı Bir Yapay Sinir Ağı Örneği

Pang-Ning Tan, M. Steinbach, V.Kumar, **Introduction to Data Mining** (USA: Pearson Education, 2006)'den aktaran Sinan Aydın, "Veri Madenciliği Ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama"(Doktora Tezi, Eskişehir Anadolu Üniversitesi Sosyal Bilimler Enstitüsü Eylül 2007), 52.

Yapay sinir ağları yöntemi, çok güçlü bir tahmin modelleme tekniğidir. Bu gücün bir kısmını, hemen her alanda uygulanabilir olmasından almaktadır. Veri madenciliği ve karar destek sistemlerindeki bir çok başarılı uygulamalarla geçerliliğini ispatlamış bir yöntemdir. Ancak, yapay sinir ağları, çok karmaşık modeller üretmektedir. Bu durum, yapay sinir ağlarının en büyük dezavantajıdır. Modelin kendisi rakamsal değerlerin karmaşık hesaplamaları ile uğraştığı için, girdilerinin de rakamsal

¹³⁰ Erkan Kıyak, "Crisp-Dm Yöntemi Bilimi Kullanılarak Deniz Kuvvetleri Verisi Üzerinde Veri Madenciliği Sınıflandırma Tekniklerinin Karşılaştırılması" (Yüksek Lisans Tezi, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, 2006), 57.

¹³¹ Marco Castellani, Hefin Rowlands, "Evolutionary Artificial Neural Network Design and Training For Wood Veneer Classification", **Engineering Applications Of Artificial Intelligence**, v. 22 (2009): 735.

olmasını bekler. Girdinin veya çıktının rakamsal olmaması durumunda, verinin uygun şekilde dönüştürülmesi gerekmektedir. Birçok veri madenciliği yazılımı, bu dönüşümü otomatik olarak yapabilecek yeteneğe sahiptir¹³².

3.5.2.3. Genetik Algoritmalar (Genetic Algorithms)

Genetik algoritmalar yapay zekanın evrimsel hesaplama tekniğinin bir parçasını oluşturmaktadır. Genetik Algoritmalar, benzer bir şekilde çalışan arama ve en iyileme yöntemidir. Problemlere tek bir çözüm üretmek yerine farklı çözümlerden oluşan bir çözüm kümesi üretir. Çözüm kümesindeki çözümler birbirinden tamamen bağımsızdır¹³³.

Genetik algoritmalar, bir popülasyon oluşturmak için araştırma alanındaki aile kromozomlardan rastgele seçim ile başlar. Burada oluşturulmak istenen popülasyon, seçim, yeniden kombine etme ve mutasyonu içeren genetik proses oluşumunu modelleyen genetik işlemcileri uygulayarak daha iyi kromozomları ifade etmektedir¹³⁴. Yeni popülasyon oluşturulması için seçilen bireyler, çözüme yakınlık derecelerine göre belirlenirler. Çözüme yakın bireylerin seçilmesi ve bu seçimden yeni bireylerin oluşturulması ile daha iyi sonuçlar üretilmesi olasıdır. Bu istenen çözüm sağlanıncaya kadar devam ettirilir¹³⁵.

3.5.2.4. K-En Yakın Komşu (K-Nearest Neighbor)

K- En Yakın Komşu algoritması (KNN), parametrik olmayan ve denetimli öğrenmeye dayalı en eski ve en basit sınıflandırma metotlarından birisidir. Amaç, yeni bir örnek görüntülendiğinde, mevcut eğitim veri setindeki en yakın “k” örneğini bulmak ve en benzer sınıfa göre ortadaki numuneyi sınıflandırmaktır. Genellikle yakınlık, Öklid uzaklığı ile tanımlanır¹³⁶.

¹³² Kıyak, **age**, 30.

¹³³ Demirel, **age**, 56.

¹³⁴ Elisaveta G. Shopova , Natasha G. Vaklieva-Bancheva , “BASIC—A Genetic Algorithm For Engineering Problems Solution”, **Computers and Chemical Engineering**, c. 30 (2006): 1295.

¹³⁵ Güleç, **age**, 20.

¹³⁶ Mehmet Acı, Mutlu Avcı, “K Nearest Neighbor Reinforced Expectation Maximization Method”, **Expert Systems with Applications**, c. 38 (2011): 12586.

Tablo 3.2: Optimum K ve Optimum Uzaklıklar

<i>Veri Seti</i>	<i>K</i>	<i>β</i>
İris	1-14	1
Şarap	1-20	3
Meme Kanseri	1-13	3
Cam	1-5	2
Maya	1-4	2

Acı, age, 12590.

Sınıflandırma yapılırken veri tabanındaki her bir kaydın diğer kayıtlarla olan uzaklığı 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 veri tabanındaki 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 oluşturur¹³⁷.

Genel olarak KNN algoritmasında aşağıdaki adımlar uygulanmaktadır¹³⁸ :

1. Uygun bir mesafe ölçüm uzayı belirle, (Euclid mesafe ölçümü en çok kullanılandır.)
2. Birbirine en yakın k adet noktayı belirle.
3. Belirlenen grubun en çok rastlandığı sınıfı belirle.
4. Bu gruba belirlenen sınıfın ismini ata.

3.5.2.5. Bellek Temelli Nedenleme (Memory Based Reasoning)

Bellek Temelli Nedenleme (BTN), insan davranışını otomatik olarak taklit etmeye çalışır. Belirli olayların geçmişini, kuralları oluşturmak için tecrübeden yararlanan bir sistemde dolaylı değil doğrudan karar vermek için kullanır. Bu algoritma, 2 adımlı bir prosedüre sahiptir: İlk olarak, tecrübelerdeki benzer durumları tanımlama ve ikinci olarak, bu durumlardaki bilgiyi yeni durumlara uygulayabilme. BTN,

¹³⁷ Erdoğan, age, 34.

¹³⁸ Shopova, age, 12587.

genellikle sayısal olmayan verilere daha uygundur. İki gözlemin farklılığını tayin etmek için uzaklık ölçüsü ve cevaba ulaşmak için komşu noktalarındaki sonuçları birleştiren bir kombinasyon fonksiyonu gereklidir¹³⁹.

Bellek tabanlı veya örnek tabanlı (memory-based, instance-based methods; case-based reasoning) bu yöntemler istatistikte 1950’li yıllarda önerilmiş olmasına rağmen o yıllarda gerektirdiği hesaplama ve bellek yüzünden kullanılamamış ama günümüzde bilgisayarların ucuzlaması ve kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla, kullanılabilir olmuştur¹⁴⁰.

Bu uygulama, büyük veri setlerinde veri madenciliği yapmak için kullanılmaktadır. Bu metot, sıklıkla belirsiz veri madenciliğinde başlangıç noktası olarak kullanılır. Buna en iyi örnek pazar kümeleme analizi olarak gösterilebilir¹⁴¹.

3.5.2.6. Bayes Sınıflandırması

Belirsiz durumlarda tahmin yapmak, sınıflandırma yapmak için kullanılır. Eldeki verilerin belirlenmiş olan sınıflara ait olma olasılıklarını öngörür. İstatistikteki Bayes teoremine dayanır. Bu teorem; belirsizlik taşıyan herhangi bir durumun modelinin oluşturularak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler doğrultusunda belli sonuçlar elde edilmesine olanak sağlar¹⁴². Başlangıçta bir amacımız olduğu zaman Bayes Teoremi ile bu inanç revize edebilir. Başlangıç amacı ile kastedilen birincil olasılık ve revize edilen inanç ise arkadaki yani gerçekleşmesi istenen olasılıktır. Bayes teoremi, aşağıdaki matematiksel formüle göre inancın gerçekleşmesi için hesaplama yapar¹⁴³:

$$P(X|Y) = P(Y|X) \times P(X) / P(Y) \quad (3.6)$$

Bu ifade; Y’nin gerçekleşmesi halinde X’ in gerçekleşme ihtimalinin ne olduğunu belirtmektedir. Bu değeri bulabilmek için “X” in gerçekleştiği durumlarda Y’nin

¹³⁹ “ 13.5.1.3. Memory Based Reasoning”, <http://fedc.wiwi.hu-berlin.de/xplore/ebooks/html/csa/node206.html#SECTION03015130000000000000> [25.12.2011].

¹⁴⁰ Çankırı, **age**, 154.

¹⁴¹ “ 13.5.1.3. Memory Based Reasoning”, **age**, 1.

¹⁴² Argüden, **age**, 52.

¹⁴³ Jyotindra N. Dharwa Dr. Ashok R. Patel, “A Data Mining with Hybrid Approach Based Transaction Risk Score Generation Model (TRSGM) for Fraud Detection of Online Financial Transaction”, **International Journal of Computer Applications**, c. 16, s. 1 (2011): 22.

gerçekleşme ihtimali” ile X' in gerçekleşme ihtimalini çarpmak ve bunu Y ’nin gerçekleşme ihtimaline bölmek gereklidir.

Belirsizlik taşıyan durumlarda karar verme konusunda çok kullanışlıdır. En önemli zafiyeti değişkenler arası ilişkinin modellenmiyor olması ve değişkenlerin birbirinden tamamen bağımsız olduğu varsayımıdır¹⁴⁴.

3.5.2.7. Naive-Bayes

Naive-Bayes sınıflandırması, sınıf tahmininin Bayes teoremine dayalı olarak yapıldığı bir sınıflandırma metodudur. Her bir örneğin m ayrıık değerli niteliklerden ve C sonlu serisinden oluştuğu bir dizi eğitim verisini düşünün. Naive- Bayes bilinmeyen bir örneğin sınıfını, en olası çıktıyı hesaplamak için mevcut veri setini kullanarak olasılıkla tahmin edebilir¹⁴⁵.

Naive Bayes kolay uygulanabilir olduğu kadar üstün performansıya da metin veri madenciliği çalışmalarında en çok kullanılan metotlardan biri haline gelmiştir. Metotta önce tüm eğitim verisindeki metinlerde kullanılan kelimelerden bir sözlük oluşturulur. Daha sonra her bir kelimenin her bir sınıftaki tekrar sayıları(frekansı) bulunur. Sınıflandırılması istenen yeni bir metin önceden geldiğinde oluşturulan sözlükte var olan kelimelerin her bir sınıftaki frekansları bulunur. Bir metnin C sınıfına dahil olma olasılığı C sınıfının eğitim setindeki oranıyla, metnin içindeki her bir kelimenin C sınıfına ait olma olasılıkları çarpılarak bulunur¹⁴⁶. Bu olası sınıf C_{NB} aşağıdaki formül ile hesaplanmaktadır¹⁴⁷:

$$C_{NB} = \arg \max_{c \in C} p(c) \prod_{i=1}^n p(a_i|c) \quad (3.7)$$

¹⁴⁴ Argüden, **age**, 53.

¹⁴⁵ Damrongrit Setsirichok ve diğ. , “Classification Of Complete Blood Count And Haemoglobin Typing Data by a C4.5 Decision Tree, A Naïve Bayes Classifier and A Multilayer Perceptron For Thalassaemia Screening”, **Biomedical Signal Processing and Control**, c. 250 (2011): 5.

¹⁴⁶ M. Fatih Amasyalı, Tülay Yıldırım, “Otomatik Haber Metinleri Sınıflandırma”, www.ce.yildiz.edu.tr/en/mygetfile.php?id=268 [25.12.2011].

¹⁴⁷ Setsirichok, **age**, 5.

3.5.2.8. Zaman Serisi Analizi

Zaman değişkeniyle ilişkili bir değişken hakkında, elde edilen gözlem değerlerini zamana göre sıralanmış olarak gösteren serilere “zaman serileri” denir.

Zaman serisi analizlerinde değişkenler tek tek ele alınır. Değişkenlerin zaman içerisinde oluşan değerlerinin ki buna veri üretme süreci (data generating process) denir, belirli bir kalıba (modele) göre meydana geldiği varsayılır ve uygun kalıbın (modelin) belirlenmesi amaçlanır. Böylece değişkenin gelecekte alacağı değerlerin sağlıklı bir biçimde öngörülebileceği varsayılır¹⁴⁸.

Veri tabanı üzerinden elde edilen verilerle zaman serileri oluşturulur. Oluşturulan zaman serileri ile gerekli çözümler yapılır ve öngörü işlemi ile geleceğe yönelik tahminlerde bulunulur. Bu noktada veri madenciliğinden yararlanılması, istenilen verilere daha kolay bir şekilde ulaşılmasını sağlar. Bu, aynı zamanda maliyet ve zamandan da tasarruf sağlanacağı anlamına gelmektedir¹⁴⁹.

3.5.3.Regresyon

Sınıflandırma kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır Bağımlı değişkenin kategorik olması durumunda Sınıflandırma Ağacı (Classification Tree), sürekli değişken olması durumunda da Regresyon Ağacı (Regression Tree) kullanılmaktadır. 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 istatistiğidir¹⁵⁰.

Regresyon, bir veya daha fazla değişken ile bir bağımlı değişken arasındaki ilişkiyi ortaya çıkarmak için kullanılan bir istatistik yöntemidir. Doğrusal regresyon ve lojistik regresyon, kredi kartının tespiti, otomobil sigortası ve kurumsal dolandırıcılıklarının saptanmasında bir regresyon tekniği olarak kullanılır¹⁵¹.

¹⁴⁸ Tüzüntürk, **age**, 81.

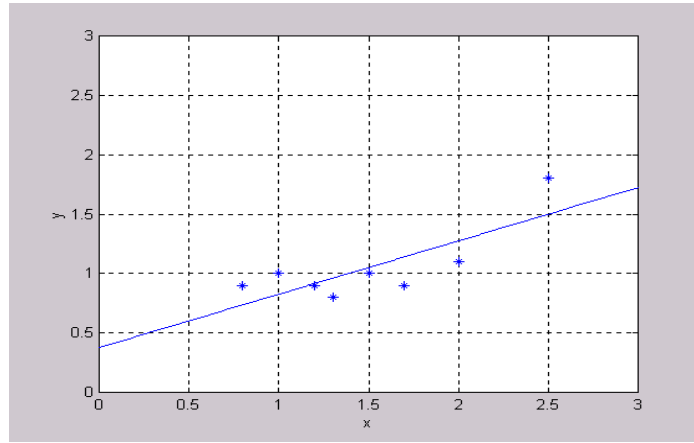
¹⁴⁹ “Veri Madenciliğine Giriş”, İnternet: Hacettepe Üniversitesi”, **age**, 31.

¹⁵⁰ Murat Kayı, Murat Boysan, “Bilişsel Yatkınlık İle Depresyon Düzeyleri İlişkisinin Sınıflandırma ve Regresyon Ağacı Analizi İle İncelenmesi”, **Hacettepe Üniversitesi Eğitim Fakültesi Dergisi**, c.34 (2008): 168.

¹⁵¹ E.W.T. Ngai ve diğ., “The Application Of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and An Academic Review Of Literature”, **Decision Support Systems**, c. 50 (2011): 562.

3.5.3.1. Doğrusal Regresyon

Basit doğrusal regresyon analizinde bir bağımlı ve bir bağımsız değişken söz konusu iken, çoklu doğrusal regresyon analizinde ise bir bağımlı değişken varken iki ya da daha fazla bağımsız değişken vardır ve her iki analizde de değişkenler arasında doğrusal bir ilişki vardır. Amacı: açıklayıcı değişkenlerin bilinen ya da değişmeyen değerlerine dayanarak bağımlı değişkenin ortalama değerini tahmin etmektir¹⁵².



Şekil 3.13: Doğrusal Regresyon

Ethem Alpaydın, Intelligent Data Mining,
www.cmpe.boun.edu.tr/~ethem/files/papers/veri-maden_2k.ppt [27.12.2011].

Değişkenler arasında bulunduğu varsayılan gerçek doğrusal ilişki, Y bağımlı değişken, X bağımsız değişken olmak üzere aşağıdaki doğrusal regresyon modeli elde edilir. Ana kütle için bu denklem şu şekilde yazılabilir¹⁵³:

$$Y_i = \alpha + \beta x_i + \epsilon \quad (3.8)$$

Burada ϵ ile gösterilen değer hata (error) terimidir.

¹⁵² “Korelasyon Analizi”, http://maden.karaelmas.edu.tr/hocalar/hamitaydin/Sunu_6ci_hafta.pdf [27.12.2011].

¹⁵³ Beril Sipahi, E. Serra Yurtkoru, Murat Çinko, **Sosyal Bilimlerde SPSS’le Veri Analizi**, 3. Bs. (İstanbul: Beta Yayınları, 2010), 154.

3.5.3.2. Lojistik Regresyon

Doğrusal regresyonda Y (açıklanan) iki değer alan (yani; 0 ve 1) gösterge değişkeni olarak tanımlandığında bunlara ilişkin hata terimlerinin ϵ (e_i) beklenen değeri sıfır, olmayacaktır. Bunun bir sonucu olarak varsayımlardan sapma durumunda elde edilen tahminler en iyi doğrusal ve sapmasız tahmin ediciler olmayacaktır. Bu yetersizlik sınıflandırma analizlerinde doğrusal regresyonun kullanılmasını engellemektedir. Bu nedenle lojistik regresyon, sınıflandırma analizlerinde sık kullanılan yöntemlerden biridir. Lojistik regresyon, çok değişkenli normal dağılım varsayımına ihtiyaç göstermediğinden bu tür uygulamalarda üstünlük sağlamaktadır. Ayrıca sınıf üyeliğine ilişkin olasılıkları belirlemek özelliği de vardır. Lojistik regresyonun varsayımlarından biri doğrusal olasılık fonksiyonunun, hata terimlerinin dağılımının lojistik dağılıma uymasıdır¹⁵⁴.

Lojistik regresyon ve sınıflandırma ağaçları, bağımlı değişkenin herhangi bir varsayımı olmadan kategorik bağımsız değişkenlerin sınıf ilişkilerini tespit ederken kullanılmaktadır. Bu metotlar, genellikle öğrenme uygulamalarında, bilgisayar bilimlerinde veri madenciliğinde ve sınıflandırma modellerinde kullanılmaktadır¹⁵⁵.

B) Tanımlayıcı Modeller

Sadece veri madenciliği teknikleri aracılığı ile stoklanmış veriden yararlı desen ve birliktelikleri elde etmek mümkündür. Kümeleme ve birliktelik gibi veri madenciliği teknikleri, gelecek tahminler için anlamlı desenleri bulma konusunda kullanılabilir. Örneğin, satış işlemlerindeki davranış önemlidir.

Bu tür bir analizde kullanılan genel terim, Pazar Sepet Analizi olarak bilinmektedir. Bugün perakende sektöründe oldukça kullanılmaktadır¹⁵⁶.

¹⁵⁴ “Veri Madenciliği”, www.deu.edu.tr/userweb/k.yaralioglu/dosyalar/ver_mad.doc [27.12.2011].

¹⁵⁵ Handan Ankaralı Çamdeviren ve diğ., “Comparison Of Logistic Regression Model and Classification Tree: An Application To Postpartum Depression Data”, **Expert Systems with Applications**, c. 32 (2007): 987.

¹⁵⁶ Aurangzeb Khan, Baharum Baharudin, Khairullah Khan, “Mining Customer Data For Decision Making Using New Hybrid Classification Algorithm”, **Journal Of Theoretical And Applied Information Technology**, c. 27, s. 1 (2011): 54.

3.5.4. Kümeleme

Çok değişkenli istatistiksel tekniklerden birisi olan kümeleme analizi, grup sayısı bilinmeyen ve gruplandırılmamış verilerin benzerliklerine göre sınıflandırılması amacıyla kullanılmaktadır. Kümeleme analizi verilerin birimlere veya değişkenlere göre birbirlerine benzerlikleri bakımından ayrık kümelerde toplanmasını sağlayan bir tekniktir¹⁵⁷.

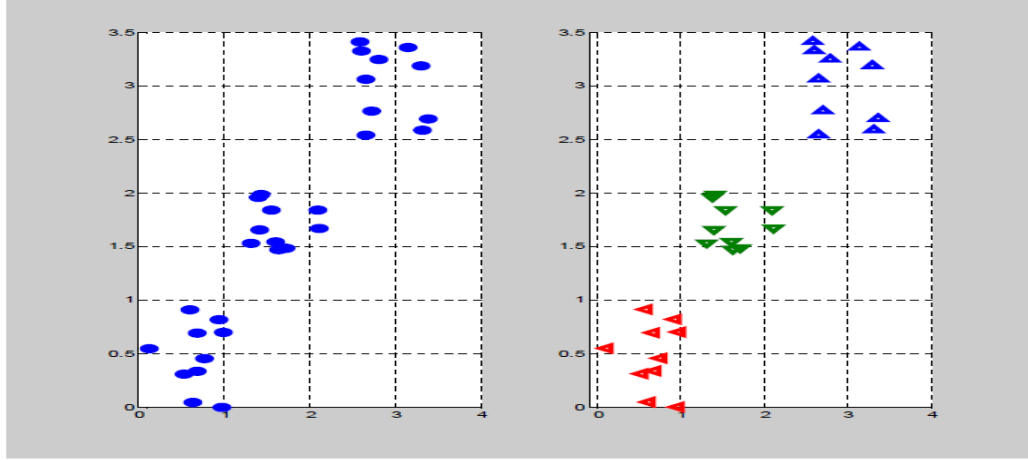
Kümeleme Analizi özellikle son yıllarda popüler olan, çok değişkenli istatistiksel yöntemlerden biridir. Bu yöntem, özellikle bilim ve iş alanında, birçok durumda uygulanabilen, en etkili ve en kolay yorumlanabilen bir yöntem olma özelliği taşımaktadır. Kümeleme Analizinin genel amacı, gruplandırılacak verileri, benzerliklerine göre alt sınıflara ayırarak açıklamaktır. Başka bir ifade ile, çalışmada yer alan tüm değişkenler itibari ile bireyler veya nesneler arasındaki benzerlikler esas alınarak, benzer bireylerin aynı gruplarda veya kümelerde toplanması ve yeni bir bireyin hangi gruba dahil olduğunun tahmin edilmesi Kümeleme Analizinin esasını teşkil etmektedir.

Kümeleme Analizi, bireylerin ya da nesnelerin sınıflandırılmasını ayrıntılı bir şekilde açıklamak amacıyla geliştirilmiştir. Bu amaca yönelik olarak, ele alınan örnekte yer alan varlıklar aralarındaki benzerliklere göre gruplara ayrılır, daha sonra bu gruplara dahil edilen bireylerin profili ortaya konur. Bir başka ifade ile bu analizde, özellikle amaçlanan şey, öncelikle ele alınan örnekte gerçekte var olduğu bilinen grupları, varlıklar (birey ya da nesne) arasındaki benzerliklere dayanan az sayıdaki karşılıklı özel grupları oluşturmak, daha sonra bu gruplara giren varlıkların profilini ortaya koymaktır¹⁵⁸.

Çoğunlukla, sosyal bilimler, tıp ve ziraat olmak üzere mühendislik bilimlerinde yaygın olan kümeleme analizi, çok değişkenli varyans analizi, lojistik regresyon analizi gibi diğer çok değişkenli analizlerle de yakın ilişkili olan bir tekniktir. Kümeleme analizi, ölçü birimlerinin ve ölçme tekniklerinin değişik olmasından dolayı benzerlik ya da farklılıkların ortaya konmasında da kümeleme analizi uygulanmaktadır. Şekil 3.14’ de basit bir kümeleme işlemi gösterilmiştir.

¹⁵⁷ Zeki Çakmak, Nevin Uzgören, Gülnur Keçek, “Kümeleme Analizi Teknikleri İle İllerin Kültürel Yapılarına Göre Sınıflandırılması Ve Değişimlerinin İncelenmesi”, 3, <http://sbe.dpu.edu.tr/12/15-36.pdf> [30.12.2012].

¹⁵⁸ İsmet Doğan, “Kümeleme Analizi ile Seleksiyon”, **Türk Vet Anim Sciences**, c. 26 (2002) : 48.



Şekil 3.14: Veri Birimleri ve Kümeleme

Şebnem Koltan Yılmaz, Said Patır, “Kümeleme Analizi Ve Pazarlamada Kullanımı”, *Akademik Yaklaşımlar Dergisi*, c. 2, s. 1 (2011): 101.

Kümeleme analizinin genel uygulama alanları; benzer örüntülerin tanınması, uzaysal (spatial) veri analizi, görüntü işleminde, ekonomi bilimlerinde (özellikle piyasa araştırmalarında) ve internet uygulamalarında (dokümanların sınıflandırılması ve benzer giriş profillerinin tespiti ve gruplandırılması) şeklindedir¹⁵⁹.

Kümeleme analizi, yaygın olarak anketler ve test panellerindeki çok değişkenli veriler ile çalışırken yapılan pazar araştırmalarında kullanılır. Pazar araştırmacıları, tüketicilerin genel popülasyonunu, pazar segmentlerine atamak ve aralarındaki birliktelik ilişkisini anlamak için kümeleme analizini kullanırlar¹⁶⁰.

Kümeleme, bir veri setini gruplara ayırmayı ve bölmeyi, böylece bir gruptaki her bir noktanın birbirine benzerliğini ve diğer gruptaki noktalardan farklılığını ifade eden bir veri madenciliği tekniğidir. Kümeleme modelleri, hedef nitelikleri kullanmaz, yerine benzer özellikleri kullanarak bölümlemeyi sağlar.

Kümeleme analizi benzerlik-mesafe ölçümü gibi öznellik kavramları içermektedir. Farklı kümeleme analiz teknikleri, farklı sonuçlar vermekte ve hangisinin doğru olduğu bilinmemektedir. Fakat sınıflandırma problemlerinde karar ağacı başarısız olabilirken, bu analizin her zaman bir sonuç sağlaması gibi güçlü bir özelliği

¹⁵⁹ Yakut Gevrekçi, “Hayvancılık Verilerinde Kümeleme Analizi Uygulamalarına Bir Örnek”, http://4uzbk.sdu.Edu.Tr/4UZBK/POSTER/BGP/4UZBKP_066.pdf [17.01.2012].

¹⁶⁰ Khan, age, 55.

vardır¹⁶¹. Bu algoritmada temel fikir, her bir küme için k sentroitlerini (mesafeler) tanımlamaktır. Bu sentroitler, farklı lokasyonların farklı sonuçlara sebep olmasından dolayı doğru olarak yerleştirilmelidir. Dolayısı ile en iyi seçim, her birinin bir değerinden mümkün olduğunca uzak olarak yerleştirilmesidir¹⁶².

Kümeleme işlemi, belirlenen amaca göre, iki gözlemin veya iki değişkenin benzerlik(yakınlık) veya uzaklık ölçüsüne bakılarak yapılır. Yakınlık bir çeşit mesafe ölçüsüdür ve genellikle Öklid mesafesi kullanılır¹⁶³.

$$Ed_{ij} = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2} \quad (3.9)$$

Kümelemede ölçümü anlamlı kılan ve uygun birleşmeleri sağlayan WGS'dir ve aşağıdaki gibi ifade edilir¹⁶⁴:

$$WGS = \frac{\sum (R_i - R)^2 \cdot N_i}{N} \quad (3.10)$$

R, cevap verme oranı; N ise toplam varlık sayısıdır.

Sınıflandırma, eğitim verisinde ilgili her verinin hangi sınıflarda yer aldığı bilgisinin verilmesi ile yeni verilerin hangi sınıfta yer aldığının bulunmasıdır. Kümeleme de ise, modelden kümelerin bulunması beklenmektedir¹⁶⁵.

Daha önce de belirtildiği gibi, kümeleme teknikleri veri setlerinden “küme” adı verilen homojen grupları oluşturmak için kullanılır. Temel olarak 4 kullanım alanı şöyledir¹⁶⁶:

¹⁶¹ Barak Aviad, Gelbard Roy, “Classification By Clustering Decision Tree-like Classifier Based on Adjusted Clusters”, **Expert Systems with Applications**, c. 38 (2011): 8221.

¹⁶² Khan, **age**, 56.

¹⁶³ Çakmak, **age**, 4.

¹⁶⁴ Aviad, **age**, 8222.

¹⁶⁵ Güleç, **age**, 55.

¹⁶⁶ Karl Kopper, **The Linux Enterprise Cluster: Build A Highly Available Cluster with Commodity Hardware and Free Software**, (San Francisco: No Starch Press, 2005), 9.

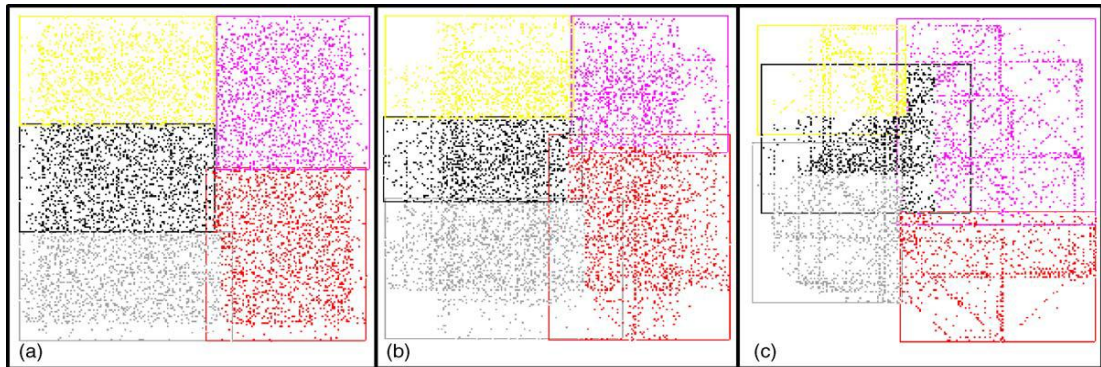
1. Tipoloji veya sınıflandırmanın gelişimi için
2. Varlıkları gruplama için kullanışlı kavramsal şemaları araştırmak için
3. Veri keşfi sırasında hipotez oluştururken
4. Hipotez testi veya veri tiplerini belirlemek için

Kümeleme modellerinde kullanılan başlıca teknikler (yöntemler) sırasıyla aşağıdaki başlıklar halinde incelenecektir.

3.5.4.1. Bölme Yöntemleri (Partitioning Methods)

Bölme yöntemlerinde, n veri tabanındaki nesne sayısı ve k oluşturulacak küme sayısı olarak kabul edilir. Bölme algoritması n adet nesneyi, k adet kümeye böler ($k \leq n$). Kümeler tarafsız bölme kriteri olarak nitelendirilen bir kritere uygun oluşturulduğu için aynı kümedeki nesneler birbirlerine benzerken, farklı kümedeki nesnelerden farklıdırlar¹⁶⁷.

Bölme yöntemleri; sistemdeki tüm maddeleri içeren bölgelerdeki yakın optimal bölümü bulmayı amaçlar. Öyle ki, bunun için iki koşulu gerçekleştirmek gereklidir: Bölgelerin sınırlarına yakın ajanları minimize etmek ve her bölgedeki ajanların sayısını eşitlemektir. Burada ilk element farklı bölümlene tekniklerini karşılaştırarak tüm metotlar tarafından bölümlerin kalitesini ölçümlemek için homojen kriter oluşturmak amacı ile kullanılır¹⁶⁸.



Şekil 3.15: Bir Simülasyon Çalışmasında Bölme Yöntemi Aşamaları: Başlangıç, Gelişme ve Sonuç

Vigueras ve diğ., “A Comparative Study Of Partitioning Methods For Crowd Simulations”, *Applied Soft Computing*, c. 10 (2010): 228.

¹⁶⁷ Özekes, age, 72.

¹⁶⁸ G. Vigueras ve diğ., “A Comparative Study Of Partitioning Methods For Crowd Simulations”, *Applied Soft Computing*, c. 10 (2010): 226.

3.5.4.2 Hiyerarşik Yöntemler (Hierarchical Methods)

Hiyerarşik teknikler, nesneler arasındaki uzaklıkların matrisi ile başlamaktadır. Bütün nesneler tek bir nesnenin yer aldığı gruplarda tek başına başlamakta yani analizin başında her bir nesne tek başına bir küme olarak düşünülmekte ve birbirine yakın gruplar birleştirilmekte, birleştirilme işlemi ardarda tekrarlanarak sürdürülmektedir¹⁶⁹.

Hiyerarşik Küme Analizi adı verilen istatistik yöntemini kullanarak benzer özelliğe sahip objeleri veya olayları gruplandırarak incelemek mümkündür. Analize konu olan her olay birbirine benzeme ya da benzememe eğiliminde olduğundan, kümeleme analizi yöntemiyle birbirine benzer homojen grupların tespiti matematiksel olarak yapılabilir. Diğer istatistik tekniklerinde olduğu gibi bu yöntemde de analizler sırasında hangi veri tiplerinin kullanılacağı, mesafenin hangi matematiksel metotla hesap edileceği ve hangi kriterler kullanılarak grup sayısının belirleneceğine önceden karar verilmelidir¹⁷⁰.

3.5.4.3 Yoğunluk Tabanlı Yöntemler (Density-Based Methods)

Yoğunluk-tabanlı kümeleme yaklaşımı ise hiyerarşik olmayan bir kümeleme yaklaşımıdır. Yoğunluk-tabanlı kümeleme algoritması verinin yoğun olduğu alanı küme olarak aldığı için, rassal şekillere sahip kümeler bulmada sorun yaşamaz¹⁷¹. Algoritma, üç aşamadan oluşur: Ekleme aşaması; komşuluk mesafelerini bulur ve yoğun birimlerin ayıklanmış mesafelerini özetler; ayıklama aşaması, çıkarma ve yığınlaşmış mesafeleri inceler ve başlangıç kümelerini oluşturur ve son aşama ise birleştirme aşaması kümelerin hiyerarşisini tamamlar. Bu yöntem tüm veri setinin

¹⁶⁹ Yasemin Koldere Akın, “Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi”, (Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, 2008), 10.

¹⁷⁰ Murat Karabulut, Mehmet Gürbüz, Ersin Kaya Sandal, “Hiyerarşik Kluster (Küme)Tekniği Kullanılarak Türkiye’de İllerin Sosyo-Ekonomik Benzerliklerinin Analizi, Hiyerarşik Kluster (Küme)Tekniği Kullanılarak Türkiye’de İllerin Sosyo-Ekonomik Benzerliklerinin Analizi”, **Ankara Üniversitesi Dergisi**, c.2, s.2 (2004): 4.

¹⁷¹ Çağatan Taşkın, Gül Gökay Emel, “Veri Madenciliğinde Kümeleme Yaklaşımları ve Kohonen Ağları ile Perakendecilik Sektöründe Bir Uygulama”, **Süleyman Demirel Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Dergisi**, c.15, s.3 (2010): 399.

yüksek kaliteli kümeler şeklinde bölümlenmesini ve böylece istenen veriye ulaşılmasını sağlar¹⁷².

Yoğunluğa Dayalı Kümeleme Teknikleri, farklı olarak kümeleri oluştururken nesneler arasındaki uzaklıkları kullanmamakta; bu teknikler nesnelerin yoğunluklarına göre kümeleri oluşturmaktadır. Özetle, yoğunluğa dayalı kümeleme teknikleri, düzensiz kümeleri bulmada veya gürültülü ve aykırı değerler mevcut olduğunda sıkça kullanılmaktadır¹⁷³.

3.5.4.4 Izgara Tabanlı Yöntemler (Grid-Based Methods)

Izgara tabanlı kümeleme metotları, veri uzayını sonlu sayıda karelere bölerek kümeleme işlemlerinin tümünün gerçekleştirileceği bir ızgara yapısı elde etmektedir. Bu yaklaşımın en büyük avantajı, performansının sadece ızgara çözünürlüğüne diğer bir deyişle ızgaradaki kare sayısına bağlı olup veri tabanının büyüklüğünden bağımsız olmasıdır. Bu nedenle genellikle çok sayıda veri noktası içeren yüksek yoğunluklu veri tabanlarında kullanılır¹⁷⁴.

3.5.4.5 Model Tabanlı Yöntemler (Model-Based Methods)

Bu metodolojide, normal dağılımların sonlu karışımı, maksimum olabilirlik tahmini algoritması kullanılarak veriye uygulanır, grupların numarası Bayes model seçimi kullanılarak seçilir ve eğer sabit kümeleme istenirse her piksel, kendisinin bir sonraki grubuna atanır¹⁷⁵. Model tabanlı yöntemlerde, her kümenin bir modele uyduğu varsayılır ve bu modellere uyan verileri gruplandırılır¹⁷⁶.

Model tabanlı prosedürler, öncelikle değişkenlerin bağımsız alt kümelerinden birden fazla küme yapısı tespiti için geliştirilmiştir. Herhangi bir model seçimine izin veren bu metodun genel kural olarak kabul edilir. Bu metodun kullanılabilirliği simülasyon ve gerçek örnekler üzerinde uygulanabilir olması ile ortaya çıkmaktadır¹⁷⁷.

¹⁷² Morteza Haghir Chehrehgani, Hassan Abolhassani , Mostafa Haghir Chehrehgani, “Improving Density-Based Methods For Hierarchical Clustering Of Web Pages”, **Data & Knowledge Engineering**, c. 37 (2008): 33.

¹⁷³ Akın, **age**, 105.

¹⁷⁴ Meral Demiralay, “Hiyerarşik Kümeleme Metotları ile Veri Madenciliği Uygulamaları” (Yüksek Lisans Tezi, Maltepe Üniversitesi Mühendislik-Mimarlık Fakültesi, 2005) , 25.

¹⁷⁵ Giuliano Galimberti, Gabriele Soffritti, “Model-Based Methods to Identify Multiple Cluster Structures in A Data Set”, **Computational Statistics and Data Analysis**, c. 52 (2007): 520.

¹⁷⁶ Halil Kaya, Kemal Köymen, “Veri Madenciliği Kavramı Ve Uygulama Alanları”, **Doğu Anadolu Bölgesi Araştırmaları** (2008): 162.

¹⁷⁷ Fionn Murtagha, Adrian E. Raftery, Jean-Luc Starck, “Bayesian Inference For Multiband Image Segmentation via Mmodel-Based Cluster Trees”, **Image and Vision Computing**, c. 23 (2005): 588.

3.5.5. Birliktelik Kuralları (Association Rules) ve Ardışık Zamanlı Örüntüler (Sequential Patterns)

Birliktelik Kuralları ile Ardışık Zamanlı Örüntü modelleri, kümeleme algoritmasındaki modeller gibi veri madenciliğinin tanımlayıcı modelleri olup iki metot arasındaki fark aşağıdaki birkaç örnek ile belirtilmiştir.

Birliktelik kuralları aşağıda sunulan örneklerde görüldüğü gibi eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır¹⁷⁸.

- *Müşteriler bira satın aldığı anda, % 75 olasılıkla patates cipsi de alırlar,*
- *Düşük yağlı peynir ve yağsız yoğurt alan müşteriler, %85 olasılıkla diyet süt de satın alırlar.*

Ardışık zamanlı örüntüler ise aşağıda sunulan örneklerde görüldüğü gibi birbirleri ile ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır.

- *X ameliyatı yapıldığında, 15 gün içinde % 45 olasılıkla Y enfeksiyonu oluşacaktır,*
- *İMKB endeksi düşerken A hisse senedinin değeri % 15'den daha fazla artacak olursa, üç iş günü içerisinde B hisse senedinin değeri % 60 olasılıkla artacaktır.*

3.5.5.1. Birliktelik Kuralları

Toplanan ve depolanan verinin her geçen gün gittikçe büyümesi yüzünden, şirketler veri tabanlarındaki birliktelik kurallarını ortaya çıkarmak istemektedirler. Büyük miktardaki mesleki işlem kayıtlarından ilginç birliktelik ilişkilerini keşfetmek, şirketlerin karar alma işlemlerini daha verimli hale getirmektedir. Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır. Bu işlem, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelikleri bularak müşterilerin satın alma alışkanlıklarını analiz eder¹⁷⁹.

¹⁷⁸ Çankırı, age, 159.

¹⁷⁹ Özekes, age, 75.

Pazar sepeti analizi olarak da bilinen bu uygulamada, hangi ürünlerin birlikte satılma eğiliminde olduğu bilgisini vererek müşteri davranışları, stok kontrol, satış stratejileri gibi analizlerde fayda sağlamaktadır¹⁸⁰.

Birliktelik kuralları, büyük miktardaki verinin niteliklerini veya öğelerinin ilişkilerini ve potansiyel birlikteliklerini keşfetmek için kullanılır. Birliktelik kurallarının keşfi veri madenciliğindeki önemli bir adımdır. Kesin bir işlemi içeren işlemlerin oranı çok yüksek olduğu zaman, birçok birliktelik kuralı, işlemsel veri tabanından üretilebilir¹⁸¹.

Birliktelik kurallarını oluşturmak için aşağıdaki iki koşulu yani daha büyük, eşit veya en az destek eşiği kuralını sağlamalıdır¹⁸².

Destek ($X \rightarrow Y$) $\geq$ en küçük destek

Güven ($X \rightarrow Y$) $\geq$ en küçük güven

Birliktelik kuralı bulma algoritmaları tipik olarak iki aşamada çalışırlar; sık öge kümelerinin bulunması ve ikincisi bunlardan önemli kuralların çıkartılması. Teknik olarak birinci aşama çok daha zordur ve ikinci aşama birincisi bittikten sonra kolaylıkla yapılabilir. Birinci aşamanın zorluğu teorik bir zorluk olmayıp, pratik bir zorluktur; çünkü olası tüm öge alt kümelerinin (veri tabanında hiç desteği olmayanlar için saymak gerekmez) destek değeri, veri tabanında sayılmalıdır. Anti-monotonluk özelliği kullanılsa dahi, bu aşama büyük veri tabanlarında uzun süren bir işlemdir¹⁸³. En önemlisi, birliktelik kuralları uygulanması sonucu öğelerin arasındaki ilişki düzeyinin güvenilirliğinin yüksek olması için doğru modelin geliştirilmesi gerekmektedir. Literatürde birliktelik sorgusu için çok sayıda algoritma önerilmiştir¹⁸⁴. Bunlar arasında en bilineni ise Apriori algoritmasıdır.

¹⁸⁰ Ay, age, 16.

¹⁸¹ Chia-Wen Lia, Yeng-Horng Perng, Tsung-Lung Chiang, “Discovery Of Unapparent Association Rules Based On Extracted Probability”, **Decision Support Systems**, c. 47 (2009): 354.

¹⁸² Age, 355.

¹⁸³ Gülistan Özdemir Özdoğan, Osman Abul, Ali Yazıcı, “Paralel Veri Madenciliği Algoritmaları” http://hpc.etu.edu.tr/docs/Basarim09_gooetal.pdf [03.01.2012].

¹⁸⁴ Zailani Abdullah ve diğ., “Mining Significant Association Rules From Educational Data Using Critical Relative Support Approach”, **Procedia - Social and Behavioral Sciences** c. 28 (2011): 98.

Apriori Uygulaması:

Literatürde birliktelik kuralı çıkaran değişik algoritmalar bulunmaktadır. Apriori algoritması, birliktelik kuralı çıkarım algoritmaları içerisinde en fazla bilinen algoritmadır. Bu algortmada sık geçen öge kümelerini bulmak için birçok kez veri tabanını taramak gerekir. İlk taramada bir elemanlı minimum desteği sağlayan sık geçen öge kümeleri bulunur. İzleyen taramalarda bir önceki taramada bulunan sık geçen öge kümeleri aday kümeler adı verilen yeni potansiyel sık geçen öge kümelerini üretmek için kullanılır. Aday kümelerin destek değerleri tarama sırasında hesaplanır ve aday kümelerinden minimum destek sağlayan kümeler o geçişte üretilen sık geçen öge kümeleri olur. Sık geçen öge kümeleri bir sonraki geçiş için aday küme olurlar. Bu süreç yeni bir sık geçen öge kümesi bulunmayana kadar devam eder¹⁸⁵.

Eğer bir nesne küme sık geçen ise bu kümenin bütün altkümeleri de sık geçen nesne küme olmalıdır mantığı, bu algortmayı özetlemektedir. Bunun için yöntemin Birleştirme ve Budama adı verilen işlemlerini kapsayan adımları vardır¹⁸⁶.

1. Birleştirme Adımı: Minimum destek değerine göre sık geçen k ögeli nesne kümeleri bulmak için k ögeli aday nesne kümeler, k 'nıncı geçişte sık geçen nesne aday kümelerinin kendi arasında birleştirilmesiyle oluşturulur. Apriori, bir işlemdeki nesnelerin veya nesne kümelerin alfabetik sıraya göre sıralandığını varsayar.

2. Budama Adımı: k ögeli aday nesne kümeler çok büyük boyutta olabilir ve çok ağır hesaplamalar gerektirebilir. Bu nesne kümenin boyutunu azaltmak için Apriori özelliği kullanılır. Herhangi bir sık olmayan $(k-1)$ -nesne küme sık geçen nesne kümenin altkümesi olamaz. Bundan dolayı, aday k -nesne kümenin bütün $(k-1)$ ögeli altkümeleri k 'nıncı geçişte sık geçen nesne aday kümesinde yoksa bu aday sık geçen nesne küme olamaz ve k ögeli aday nesne kümeden çıkarılır.

¹⁸⁵ Murat Karabatak, Melih Cevdet İnce, "Apriori Algoritması İle Öğrenci Başarısı Analizi", http://Www.Emo.Org.Tr/Ekler/24f4c5eef7ec01c_Ek.Pdf [03.01.2012].

¹⁸⁶ Ay, age, 18.

3.5.5.2. Ardışık Zamanlı Örüntüler

Wojciechowski ardışık örüntüleri şu şekilde tanımlamıştır: "Ardışık örüntüler meydana gelen bir dizi olay içinde en sıklıkla rastlanan alt dizilerdir." Ardışık zaman örüntüleri alanında kullanılmak üzere geliştirilmiş algoritma ve bu algoritmaları kullanan yazılımlar, örüntülerin belirlenme işlemini basitleştirmek ve hızlandırmak için kullanıcıya bazı kısıtlar koyması için olanak tanırırlar¹⁸⁷.

Ardışık örüntülü madencilik, minimum destek eşik değerli bir ardışık veri tabanındaki ardışık örüntülerin tüm grubunu bulmak için geliştirilmiştir. Verinin geniş büyümesi, büyük veriler arasındaki anlamlı örüntüleri bulmak için yol gösterir. Ancak ardışık zamanlı örüntülerin en büyük dezavantajı, ardışık örüntülü büyük veri gruplarının üretilmesi ve hesaplama zamanının çok yüksek olmasıdır. Dahası, bunu azaltmak için geliştirilen algoritmalar zamanı kısaltmış da olsa güvenilirlikleri çok yüksek değildir¹⁸⁸.

Gerçek hayatta, ardışık örüntüler, bazı alanlarda diğer birçok uygulamadan daha işe yarar. Örneğin, web sayfası uygulamalarında verilerin yüklenmesinde, müşteri satın alma davranış analizinde, proje takım koordinasyonunda, perakende yönetiminde biyomedikal veri analizlerinde ve genetik DNA yapısının çözümlenmesinde kullanılmaktadır¹⁸⁹.

3.6. Model İçi Değerlendirme Süreci

Çözülmesi gereken problem için en uygun modelin bulunabilmesi çok sayıda modelin oluşturularak test edilmesi ile mümkündür. Bu nedenle bir veri kümesi için model oluşturma en iyi modele ulaşıncaya kadar tekrarlanan bir süreçtir. Model oluşturma süreci, tahmin edici (denetimli) ve tanımlayıcı (denetimsiz) modellerine göre farklılık gösterir. Bu tür modellerin geçerlilik yöntemleri modelin tipine göre farklılık gösterir¹⁹⁰.

Denetimli öğrenmede, tahmin edici olarak adlandırılan veri madenciliği modellerinde seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada

¹⁸⁷ Erdoğan, **age**, 13.

¹⁸⁸ Unil Yun, "A New Framework For Detecting Weighted Sequential Patterns In Large Sequence Databases", **Knowledge-Based Systems**, c. 21 (2008): 110.

¹⁸⁹ Chieh-Yuan Tsai, Yu-Chen Shieh, "A Change Detection Method For Sequential Patterns", **Decision Support Systems**, c. 46 (2009): 501.

¹⁹⁰ Aydın, **age**, 55.

verinin bir kısmı modelin öğrenilmesi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenilmesi, öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi belirlenir.

Denetimsiz öğrenmede, tanımlayıcı olarak adlandırılan veri madenciliği modellerinde örneğin kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır¹⁹¹.

Model içi doğruluğunun test edilmesinde çeşitli geçerlilik yöntemleri kullanılır.

3.6.1. Basit Geçerlilik

Bir modelin doğruluğunun test edilmesinde kullanılan en basit algoritma basit geçerlilik testidir.

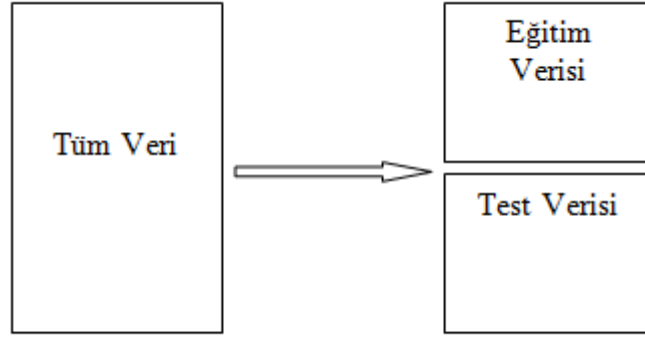
Bu yöntemde tipik olarak verilerin % 5 ile % 33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı bulunurken doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı (Doğruluk Oranı = 1 - Hata Oranı) hesaplanır¹⁹².

Doğrulanması gereken tüm hesaplamalar için verinin bu şekilde iki gruba ayrılmasında seçimi tesadüfi olarak yapan yöntemler kullanılmalıdır. Böylece eğitim ve test veri kümeleri modeli oluşturan veriyi temsil edecektir. Bu şekildeki bir bölümlenme, modelin gelecek performanslarındaki hata oranını da azaltmış olacaktır¹⁹³.

¹⁹¹ Özlem Terzi ve diğ. , “Veri Madenciliği Süreci Kullanılarak Güneş Işınımı Tahmini”, **SDU International Technologic Science**, c. 3, s. 2 (2011): 33.

¹⁹² Terzi, **age**, 33.

¹⁹³ Nong Ye, **The Handbook of Data Mining**, (USA: Lawrence Erlbaum Assoc., Inc., 2003), 435.



Şekil 3.16: Basit Geçerlilik Ölçümünde Veri Organizasyonu

Nong Ye, **The Handbook of Data Mining**, (USA: Lawrence Erlbaum Assoc. , Inc., 2003), 435.

3.6.2. Çapraz Geçerlilik

Sınırlı miktarda veriye sahip olunması durumunda, kullanılabilecek diğer bir algoritma, çapraz geçerlilik testidir. Bu yöntemde veri kümesi rastgele iki eşit parçaya ayrılır. İlk aşamada bir parça üzerinde model eğitimi ve diğer parça üzerinde test işlemi; ikinci aşamada ise ikinci parça üzerinde model eğitimi ve birinci parça üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır¹⁹⁴.

Bu denklemde, e_C çapraz geçerlilik hatası, e_i ortalama hata ve k , çapraz geçerlilikte veri grubunun rassal olarak bölümlendiği sayı olmak üzere i . veri bölümü için oran hesabı aşağıdaki gibi yapılır¹⁹⁵.

$$e_C = \frac{1}{k} \sum_{i=1}^k e_i \quad (3.11)$$

Her verinin bir kez eğitim ve bir kez de test olarak kullanıldığı çapraz geçerlilik yönteminin geliştirilmiş hali n -katlı çapraz geçerlilik yöntemidir. Bu yöntemde veri kümesi tesadüfi olarak eşit n adet gruba ayrılır. Verilerin n gruba bölünmesinin ardından bir grubun test olarak belirlendiği ve kalan $n-1$ grubun da eğitim için kullanıldığı n adet basit geçerlilik süreci tekrarlanır. Böylece her grup bir kez test

¹⁹⁴ Terzi, **age**, 33.

¹⁹⁵ Ye, **age**, 436.

için kullanılmış olur. Elde edilen n adet bağımsız hata oranının ortalaması, oluşturulan modelin hata oranı olarak kullanılır¹⁹⁶.

3.6.3. Bootstrap

Bootstrapping, küçük veri kümeleri için modelin hata düzeyinin tahmininde kullanılan bir başka tekniktir. Bootstrap metodu ilk olarak Efron (1979) tarafından parametrik olmayan bir yaklaşım olarak tanıtılmıştır. Bootstrap tahmininde, verinin bilinmeyen bir F olasılık dağılımından elde edildiği varsayımından hareket edilmektedir. $\sigma(F)$, bilinmeyen F örnekleme dağılımının bir fonksiyonu olan standart hata Q' yı göstermektedir. Standart hatanın bootstrap tahmini, basit olarak $\sigma(F)$ olarak da ifade edilebilmektedir¹⁹⁷.

$$se_{boot}(Q) = \sigma(F) \quad (3.12)$$

Bootstrap veri kümesinin büyüklüğünün tespitinde modelin hata oranını en iyi temsil edecek oran 0,632 olarak belirlenmiştir. Tahmini bootstrap hatası aşağıdaki şekilde ifade edilmektedir¹⁹⁸.

$$\frac{1}{k} = \sum_{i=1}^k [0,632 \times \text{test hatası} + 0,368 \times \text{eğitim hata}] \quad (3.13)$$

Bootstrap metot, amaca göre uygulanırken bootstrap örnekleme B'nin belirlenmesi önem taşımaktadır. Buna göre iki kural göze çarpar¹⁹⁹.

1. Küçük sayıda bootstrap örneklemleri örneğin, B=25 konu için genellikle aydınlatıcı iken B=50'nin ancak standart hatanın iyi tahminini vermekte yeterli olduğu ifade edilmektedir.

¹⁹⁶ Aydın, **age**, 57.

¹⁹⁷ Mustafa Kemal Beşer, "Zaman Serilerinde Bootstrap Çözümlemeleri ve Türkiye'de Tanzi Etkisine Uygulaması" (Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, 2006), 19.

¹⁹⁸ Ye, **age**, 436.

¹⁹⁹ Meltem Şengün, "Ucal, Ekonometrik Model Seçim Kriterleri Üzerine Kısa Bir İnceleme", C.Ü. **İktisadi Ve İdari Bilimler Dergisi**, c. 7,s. 2 (2006) 50.

2. Bir standart hatanın tahmin edilmesi için B=200'den fazla örnekleme çok fazla ihtiyaç duyulmazken, güven aralıklarının belirlenmesinde B'nin daha büyük değerleri istenmektedir.

Bootstrap model seçim kriterinin performansı uygulamada diğer bilinen bazı seçim kriterlerine göre parametrik olarak sınındığında daha fazla üstünlük sergilediği görülmektedir. Çapraz-geçerlilik ve bootstrap metot birlikte veya karşılaştırmalı olarak model seçiminde kullanılırken, doğrusal regresyon modeli kurulurken, ekstra yeniden örneklem oluşturulmasının getireceği maliyet ve diğer regresyon türlerinde kullanışlılığı konularına kesinlikle dikkat edilmelidir²⁰⁰.

3.7. Modelin Değerlendirilmesi ve Raporlama

Bu aşamaya gelindiğinde kurulmuş bir model vardır. Bu aşama, modelin nihai olarak sunulmasından önce modelin yoğun olarak değerlendirilmesi ve iş hedefleri ile uyumlu olup olmadığının kontrol edilmesini amaçlar. Kapsanmamış (açıkta kalmış) konu olup olmadığı değerlendirilmelidir²⁰¹.

Veri kümeleriyle ilgili olarak elde edilen modellerin değerlendirilmesinde tercih edilen bir modelin uygulanması ile sağlanacak faydaların da karşılaştırılması gereklidir. Bunun için değerlendirme amaçlı olarak kullanılan bazı yöntemler mevcuttur. Bunlardan Risk Matrisi, Kaldıraç Grafiği ile Alıcı Çalışma Karakteristik Grafiği (ROC) en yaygın olanlarıdır.

Modelin tamamlanmış olması projenin esas sonucu değildir. Modelin sonuçlarının sunulması için genellikle gerçek verilerden örnekler gerekmektedir. Bunun için raporlama yapılması gerekmektedir. Raporlama veri madenciliği bulgularını göstermek için önemli bir dağıtım kanalıdır. Birçok veri madenciliği aracı, örneğin paket programlar elde edilen modelden kullanıcıların daha önceden tanımladığı raporları göstermek için gerekli araçlara sahiptir²⁰². Yapılan çalışmanın başkaları tarafından da tekrarlanabilirliğini sağlamak ve sonuçlarını karar vericilere aktarabilmek üzere hazırlanan rapor ise “final raporudur”²⁰³.

²⁰⁰ Age, 51.

²⁰¹ Argüden, age, 24.

²⁰² Tekerek, age, 3.

²⁰³ Argüden, age, 25.

3.7.1. Risk Matrisi

Risk yönetim yaklaşımı, çeşitli türdeki risk odaklı tarihi istatistik veriye dayalı tipik yarı kantitatif bir değerlendirme aracıdır. Çok kesin olmamasına rağmen ve matematiksel temelden tam olarak yararlanmasa da, matrisin yapısının yorumlanışının kolay olması ile imalat, hizmet sektöründe ve diğer endüstrilerde bugün veri değerlendirmesinde yaygın olarak kullanılmaktadır²⁰⁴.

Bir risk matrisini oluşturmak için bazı temel kurallar takip edilmelidir²⁰⁵:

- Risk matrisini oluşturmak için iki girdi değişken gereklidir. Çıktı risk indeksine ise sadece sonuçların şiddeti ve olasılığına göre karar verilir.
- Sonuçların şiddeti, olasılığı ve çıktı indeksi sırasıyla kalitatif tanımlar ve ölçekler ile farklı düzeyde ayrılabilir. Çıktı sonuçları matris içinde konumları itibari ile farklı bölgelerdedir.
- Matrisin çıktıları için hesaplama prosesi mantıksal olarak şöyledir:

$$\text{Risk} = \text{Şiddet} \times \text{Olasılık}$$

Bu metot ile öncelikle bir olayın gerçekleşme ihtimali ile gerçekleşmesi takdirinde sonucunun derecelendirilmesi ve ölçümü yapılır. Model kurulduktan sonra, gerçek veriler ve gelecekteki veriler kullanılarak elde edilen çeşitli senaryolarda kullanılan bir yöntemdir²⁰⁶.

²⁰⁴ Huihui Ni, An Chen, Ning Chen, “Some Extensions On Risk Matrix Approach”, **Safety Science**, c. 48 (2010): 1270.

²⁰⁵ **Age**, 1272.

²⁰⁶ “Risk Değerlendirme Karar Matrisi”, http://www.isguv.com/risk_degerlendirme.htm [10.01.2012].

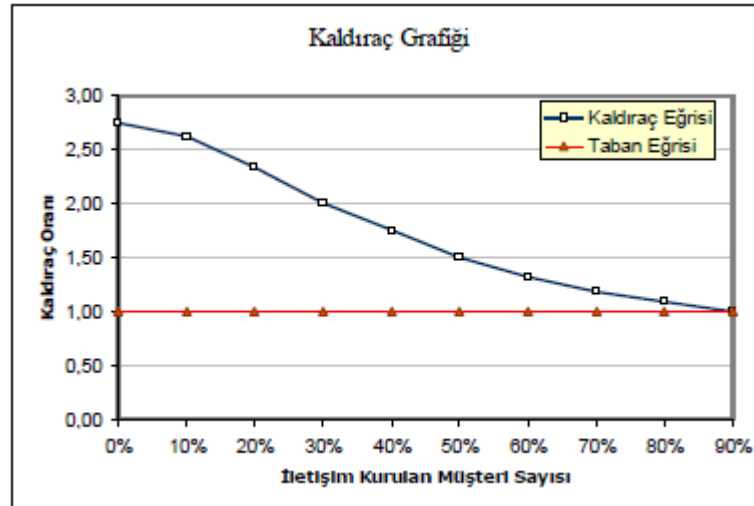
İHTİMAL	ŞİDDET				
	1 (Çok Hafif)	2 (Hafif)	3 (Orta Derece)	4 (Ciddi)	5 (Çok Ciddi)
1 (Çok Küçük)	Anlamsız 1	Düşük 2	Düşük 3	Düşük 4	Düşük 5
2 (Küçük)	Düşük 2	Düşük 4	Düşük 6	Orta 8	Orta 10
3 (Orta Derece)	Düşük 3	Düşük 6	Orta 9	Orta 12	Yüksek 15
4 (Yüksek)	Düşük 4	Orta 8	Orta 12	Yüksek 16	Yüksek 20
5 (Çok Yüksek)	Düşük 5	Orta 10	Yüksek 15	Yüksek 20	Tolere Edilemez 25

Şekil 3.17: Risk Matrisi Örneği

“Risk Değerlendirme Karar Matrisi”, http://www.isguv.com/risk_degerlendirme.htm [10.01.2012].

3.7.2. Kaldıraç Grafiği

Kaldıraç grafiği, eğitim veri kümesine dayalı olarak gerçekleştirilen modelin test veri kümesinde başarılı olarak yaptığı tahminlerin oranını artıran ya da azalan sırada grafiğe yansıtmasıdır. Kaldıraç eğrisi ile taban eğrisi arasındaki alan ne kadar büyükse model o kadar iyidir. Alternatif modellerden elde edilen tahmin değerleri aynı grafik üzerine çizilerek karşılaştırılabilir²⁰⁷.



Şekil 3.18: Kaldıraç Grafiği

Aydın, age, 60.

²⁰⁷ Aydın, age, 59.

3.7.3. Alıcı Çalışma Karakteristik Grafiği (ROC)

Sınıflandırma sistemlerinin analizinde en eski ve bugün en çok kullanılan performans tekniklerinden biri de Alıcı Çalışma Karakteristik Grafiğidir (ROC) . Bu grafik, kurulan bir modelin doğruluğunu ölçmek için kullanılan diğer bir yöntemdir.

En yaygın kullanılan ROC grafiklerinden birisi; bir örnek sınıf ve doğru-yanlış sınıflandırması yaparak, çıktı özelliği ikili (binary) ve parametrik olmayan sınıflama modellerinin değerlendirilmesinde kullanılmaktadır²⁰⁸.

Alıcı çalışma karakteristik grafiği (ROC), sınıflandırıcıları performanslarına göre görselleştirme, organize etme ve seçme için kullanılan bir tekniktir. ROC grafikleri, sınıflandırıcıların isabet oranları ve hata oranları arasındaki ilişkiyi göstermek için kullanılmaktadır²⁰⁹.

Sınıflandırıcı ve bir örneği göz önüne aldığımızda, 4 olası çıktı vardır. Eğer bir örnek pozitifse, pozitif doğru olarak; eğer yanlışsa yanlış negatif olarak adlandırılır. Eğer bir örnek negatifse, negatif doğru olarak; eğer pozitifse pozitif yanlış olarak adlandırılır. Oluşan bu matris 2 x 2 matrisi olup, ortaya çıkan sonuçlar (0,1) aralığında yer alır. Bir ROC grafiği YP oranını yatay eksen, DP oranını dikey eksen gösterir²¹⁰.

Tablo 3.3: ROC Matrisi

Y	Doğru Pozitif	Yanlış Pozitif
N	Yanlış Negatif	Doğru Negatif
	P	n

Christine M.Schubert, Steven N. Thorsen, Mark E. Oxley, “The ROC Manifold For Classification Systems”, **Pattern Recognition**, c. 44 (2011): 350.

²⁰⁸ Christine M.Schubert, Steven N. Thorsen, Mark E. Oxley, “The ROC Manifold For Classification Systems”, **Pattern Recognition**, c. 44 (2011): 350.

²⁰⁹ Tom Fawcett, “An Introduction to ROC Analysis”, **Pattern Recognition Letter**, c. 27 (2006): 861.

²¹⁰ **Age**, 862.

3.8. Modelin Yönetimi

Her bir veri madenciliği modeli bir yaşam döngüsüne sahiptir. Bazı uygulamalarda işler, özellikler durağandır ve modelin yeniden eğitilmesine gerek yoktur. Fakat birçok iş özellikleri sık sık değişir. Yeni veriler geldikçe modelin yeniden eğitilmesine gerek vardır. Yani bir model kurulduktan sonra eğer çok sık olarak veri setinde değişiklik yapılıyorsa model sık sık güncellenmelidir²¹¹.

²¹¹ Tekerek, **age**, 3.

4. UYGULAMA

4.1. Araştırmanın Problemi

İşsizlik ülkemizde ve dünyada en önemli sorun olarak gündemin ilk sıralarında yer almaktadır. Gelişmişlik farkı gözetmeksizin hemen her ülke, işsizlik ve neden olduğu sosyal sorunlarla artık daha sık karşılaşmaktadır.

Türkiye 2009 yılında dünya genelinde yaşanan ekonomik krizden olumsuz etkilenmiş, kriz nedeniyle gerek 2009 yılı gerek takip eden 2010 yılında büyüme ve istihdam alanında başarısız bir dönem geçirmiştir. Ülkemizde kalıcı ve uzun süreli işsizlik sorunu halen ciddiyetini korumaktadır.

Bugün ekonominin en önemli sorunlarından olan işsizlik, gerek bu durumu yaşayanlar için büyük bir sıkıntı olurken gerek ülke ekonomisi açısından da var olan işgücünün tam kullanılamamasına sebep olmaktadır. Ayrıca, ülkenin ekonomisi için olduğu kadar, gelişmiş bir toplum olmak için temel olan huzurlu ve refah ortamını oluşturmak için de işsizlik hızla ele alınıp çözülmesi gereken sorunlardan biridir.

Türkiye ekonomisindeki işgücü ile ilgili bilgiler ağırlıklı olarak Türkiye İstatistik Kurumu (TÜİK) tarafından gerçekleştirilen “Hanehalkı İşgücü Anketi” sonuçlarına dayanmaktadır. Çünkü hanehalkı İşgücü Anketi ile ekonomik politikayı belirleyenlerin ve ülke ekonomisini değerlendirenlerin kullanımı için nüfusun işgücü açısından durumunun açıklıkla ortaya konması; istihdamda olanlara ve işsizlere ilişkin ayrıntılı bilgi derlenip bunların kamuoyuna sunulması amaçlanmaktadır. Bu nedenle, çalışmanın hedef kitlesi ile birebir örtüştüğü için uygulama kapsamında, TÜİK’ten temin edilen 2009 ve 2010 yılları Hanehalkı İşgücü Anketi verileri kullanılmaktadır.

Bu tez çalışmasında, 2009 yılı ve 2010 yılı hane halkı işgücü anket verilerine Veri Madenciliği algoritmalarından CHAID yöntemi uygulanmaktadır. Böylece, Türkiye’de işgücü durumu belirlenerek işsiz ve istihdamda olan kişilere ilişkin genel bir profil çizilerek, işsizlik sorunun çözümü için önlem ve kararlar alınmasına yardımcı olunacaktır.

4.2. Araştırmanın Amacı ve Önemi

Bu çalışmanın temel amacı ile, 2009 ve 2010 yılları için ülkemizde işgücü piyasasının durumu ve çalışma potansiyeli olan grubun (aktif nüfus) işgücü profili belirlenmeye çalışılmıştır. Bu iki yıl için kıyaslama yapılarak, mevcut işsizlik sorunun sebeplerinin bulunması ve gelecekte işgücü durumu ile ilgili kararlar alınmasında kamu ve özel sektördeki yetkililere ışık tutması ve daha sonra çıkarılacak olan kanunların bu yapılan çalışmada öngörülen analiz sonuçları ışığında karar destek amaçlı olarak katkıda bulunması amaçlanmıştır.

Bu tez çalışmasında, karar vermenin her geçen gün daha da zorlaştığı rekabet ortamında, analizcilerin karşılaştıkları büyük veri yığınları ile başa çıkmalarındaki en başarılı yöntemlerden biri olan Veri Madenciliği Analizi yöntemi ile örnek bir uygulama gerçekleştirilmiştir. Daha çok ticari sektörlerde yoğun olarak kullanılan VM'nin Türkiye genelinde işgücü durumunu ilgilendiren bir kamu kurumunda kullanılmasıyla farklı bir bakış açısı getirilmeye çalışılmıştır.

Çalışmanın diğer temel amacı, işgücü durumu konusuyla ilgili daha ayrıntılı araştırma yapmak isteyenler için bir yol gösterici niteliğinde olması yönündedir. Bu noktada giderek daha çok tercih edilen bir analiz yöntemi olan veri madenciliğinin öğrenilmesi, bu alanda çalışacaklara da kaynaklık edecektir. Bu noktada, akademik literatüre katkı yapması hedeflenmiştir.

Türkiye'nin gelişmiş ülkeler grubunda yer alması için ekonomide sürdürülebilir bir büyümeyi gerçekleştirmesi şarttır. Bunu sağlanması için işgücüne ihtiyaç bulunmaktadır. Ülkedeki işgücünün nitelik ve niceliği önemlidir. Somut bir örnekle belirtmek gerekirse, 2007 yılı İlerleme Raporu'nda istihdam politikası ile ilgili olarak fazla ilerleme kaydedilmediği belirtilmiştir. İstihdamı artırıcı ve işsizliği önleyici etkin politikaların belirlenmesi ve uygulanması gereği, dahası bunun sürekli olarak güncellenmesi bu açıdan önem kazanmaktadır. Bunun için uygun analizlerin ve doğru yorumların yapılması ihtiyaçtan çok bir zorunluluk haline gelmektedir.

4.3. Araştırmanın Sınırlılıkları

Konu Sınırlaması: Ankette yer alan tüm değişkenlerin veri madenciliği uygulamasına dahil edilmesi analiz açısından uygun olmadığı ve anlamlı sonuçlar vermediğinden, kişinin işgücü durumunu belirlemede yardımcı olabilecek bazı değişkenler seçilmiştir. Bu seçimde, geçmiş yıllarda yapılan literatürdeki tez çalışmaları referans alınmıştır. Bu değişkenler 11 adet olup bunlardan 9 tanesi analizde kullanılmıştır.

Uygulama Sınırlaması: Araştırmanın uygulama bölümüne kaynaklık eden veri seçimi, en güncel yapıyı oluşturması açısından 2009 ve 2010 yılı ile kısıtlanmıştır. Tez çalışmasına, 2011 yılında başladığı için bu zaman periyodunda 2011 yılı verileri henüz tamamlanmamıştır. Bu nedenle analize dahil edilmemiştir.

4.4. Araştırmanın Varsayımları

Araştırma kapsamında gerçekleştirilen veri madenciliği analizi için girdi ve çıktı kümesi seçimi işleminin, yetkili kişilerin karar vermeleri ve işgücü politikaları oluşturması açısından etkin sonuç verdikleri varsayılmaktadır.

Ayrıca, uygulamada kullanılan ankete cevap verenlerin Türkiye’ nin tamamını temsil ettiği varsayılmaktadır.

4.5. Bilgi Ve Veri Toplama Yöntemi

Uygulama kapsamında, TÜİK’ten temin edilen 2009 ve 2010 yılları Hanehalkı İşgücü Anketi verileri kullanılmıştır.

Hanehalkı İşgücü Anketi

Tüm Türkiye genelinde kurumsal olmayan nüfusla (Aktif Nüfus) TÜİK’in Bölge Müdürlüğü birimleri tarafından alan uygulaması gerçekleştirilen ankete ilişkin veriler, TÜİK Merkez Teşkilatındaki ilgili birim tarafından birleştirilmekte, analiz edilmekte ve yayınlanmaktadır. Anketin esas amacı; ülkedeki işgücünün yapısını ortaya koymak, istihdam edilenlerin; iktisadi faaliyet, meslek (ya da tuttuğu iş), işteki durum ve çalışma süresi, işsizlerin ise; iş arama süresi ve aradıkları meslek (ya da iş) ve benzer özellikleri hakkında bilgi derlemektir.

Ankette örnekleme yöntemi olarak tesadüfî, 2 aşamalı, tabakalı, küme örnekleme yöntemi uygulanmaktadır. Örneklem birimi ise ‘adreste yaşayan hane halkıdır ve anketin coğrafi kapsamı Türkiye sınırları içinde bulunan hanelerde yaşayan tüm fertlerdir. 20.001 ve daha fazla nüfuslu yerleşim yerleri KENT olarak tanımlanmıştır. 20.000 ve daha az nüfuslu yerleşim yerleri ise KIR olarak tanımlanmıştır.

Örnekleme sonucu elde edilen veri setinden Türkiye genelini temsil edecek değerlere ulaşabilmek için ağırlıklandırma işlemi gerçekleştirilmektedir. Kullanılacak nihai ağırlıklar birçok etkenin birleşiminden oluşmaktadır. Bunlar arasında tasarım ağırlıkları, cevapsızlık düzeltmeleri, dışsal dağılım kontrolleri ve nihai katlama faktörü sayılabilir.

Veri Derleme Yöntemi

Anket, anketörler tarafından yüz yüze görüşme yöntemi ile uygulanmakta ve veriler doğrudan dizüstü bilgisayarlara kaydedilmektedir.

İşsizlik rakamları ile ilgili sıkça kullanılan bazı kavram ve tanımları şu şekildedir:

Kurumsal Olmayan Nüfus: Okul, yurt, otel, çocuk yuvası, huzurevi, hastane, hapisane, kışla ya da orduevinde ikamet edenler dışında kalan nüfustur.

Kurumsal Olmayan Çalışma Çağındaki Nüfus: Kurumsal olmayan nüfus içerisindeki 15 ve daha yukarı yaştaki nüfustur.

Hanehalkı: Aralarında akrabalık bağı bulunsun ya da bulunmasın aynı konutta veya aynı konutun bir bölümünde yaşayan, temel ihtiyaçlarını birlikte karşılayan, hanehalkı hizmet ve yönetimine iştirak eden bir veya birden fazla kişiden oluşan topluluktur.

Referans Kişi: Hanehalkının sosyo-ekonomik durumu ve hanede yaşayan tüm fertlerin kişisel özellikleri hakkında en doğru bilgiye sahip, hanenin yönetim veya geçiminden sorumlu yetişkin hanehalkı üyesidir.

İstihdam Edilenler: Referans döneminde iş başında olanlar ile çeşitli nedenlerle iş ile bağlantısı devam ettiği halde iş başında bulunamayan kişilerdir.

İşsiz: Referans döneminde istihdam edilmeyen, iş aramak için iş bulma kanallarından herhangi birini kullanan ve iş bulması halinde işe başlayabilecek durumda olan kişiler işsizdir.

İşgücü göstergelerinin elde edilmesinde kullanılan formüller aşağıda yer almaktadır:

$\text{İŞGÜCÜ} = \text{İSTİHDAM (İstihdam edilen nüfus)} + \text{İŞSİZ (İşsiz nüfus)}$

$\text{İSTİHDAM ORANI} = \text{İSTİHDAM} / 15+ \text{ Yaştaki Kurumsal Olmayan Nüfus}$

$\text{İŞSİZLİK ORANI} = \text{İŞSİZ} / \text{İŞGÜCÜ}$

4.6. Veri Analiz Tekniği

2009 ve 2010 yılları Hanehalkı İşgücü Anketi verilerine ayrı ayrı Veri Madenciliği Analizi yöntemlerinden Karar Ağacı metodu uygulanmıştır.

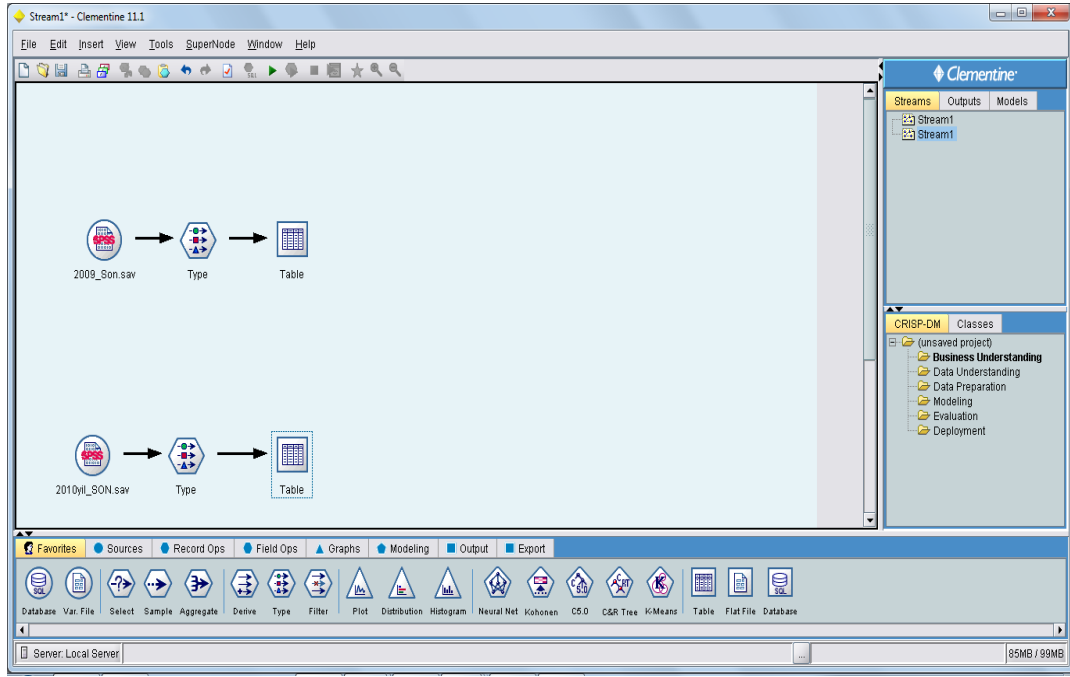
4.6.1. Veri Ön İşlem Süreci

- **Veri Temizleme**

Veri temizliği, üçüncü bölümde ifade edildiği üzere VM sürecinin en önemli aşamalarından birisidir. Anket soru kâğıdında yer alan istihdamla ilgili sorular hanedeki yalnız 15 yaş ve üzerindeki tüm fertlere yani hedef kitle olan Kurumsal Olmayan Çalışma Çağındaki Nüfus'a (Aktif Nüfus) sorulduğu için, 15 yaş ve altı grup için istihdam ile ilgili bilgiler yer almamaktadır. Bu çalışma kapsamındaki hedef; işgücü durumunun analiz edilmesi olduğundan dolayı “Durum” değişkeni = “4” olanlar 15 yaş altı olduğu için analize dahil edilmemiştir ve bu gruba ait veri kayıtları silinmiştir.

- **Veri Birleştirme**

Gerek 2009 yılı gerekse 2010 yılına ait veriler ayrı ayrı kaynaklarda bulunmaktadır. 2009 ve 2010 yılları birbiri ile karşılaştırılacağı için birleştirilmemiştir. Bu SPSS dosyaları veri madenciliğinde aynı çalışma sayfasında (Stream) kullanılmış ve gerekli analizler yapılmıştır.



Şekil 4.1: İş Gücü Durumu İçin Veri Kaynakları

- **Veri Dönüştürme**

Veri temizlemeden sonraki adım veri dönüştürmedir. Yapılandırma durumu için anket veri dönüştürme işleminde 2009 yılında not sayfası (text) olarak verilen veriler önce Excel, arkasından SPSS formatına dönüştürülerek anket sonuçlarının aynı formatta verilmesi sağlanmıştır.

4.6.2. Değişkenlerin Açıklanması

Uygulamada kullanılan bağımlı / hedef değişken, işgücü durumu olup istihdam (1), işsiz (2) ve işgücüne dahil olmama (3) şeklinde üç seviyeli kategorik bir değişken olarak alınmıştır. Bağımsız / açıklayıcı değişkenler ise aşağıdaki gibidir:

- **Cinsiyet**

1-Erkek / 2-Kadın

- **Referans kişiye yakınlık durumu**

1-Referans kişi / 2-Eşi / 3-Çocuğu / 4-Gelini veya damadı / 5-Torunu /

6-Kendi veya eşinin anne/babası / 7-Diğer akrabalar / 8-Akraba olmayanlar

- En son bitirilen okul

0-6 yaşından küçük olanlar / 1-Bir okul bitirmeyen / 2-İlkokul(5 yıl) /

3-Ortaokul, mesleki ortaokul ve ilköğretim (8 yıl) / 4-Genel lise /

5-Mesleki veya teknik lise / 6-Yüksekokul, fakülte ve üzeri

- Bir öğrenim kurumuna devam etme durumu

1-Evet / 2-Hayır

- Medeni durum

1-Hiç evlenmemiş / 2-Evli / 3-Boşandı / 4-Eşi öldü

- Bir yıl önceki işgücü durumu

1-Bir işte çalışıyordu / 2-Şu anki işinde çalışıyordu / 3-Emekliydi / 4-İş arıyordu /

5-Ev işleriyle meşguldü / 6-Eğitim/öğretimine devam ediyordu / 7-Özürlü veya hastaydı / 8-Askerdi / 9-Yaşlı (65 ve + yaşta) / 10-Diğer

- NUTS1 (İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey)

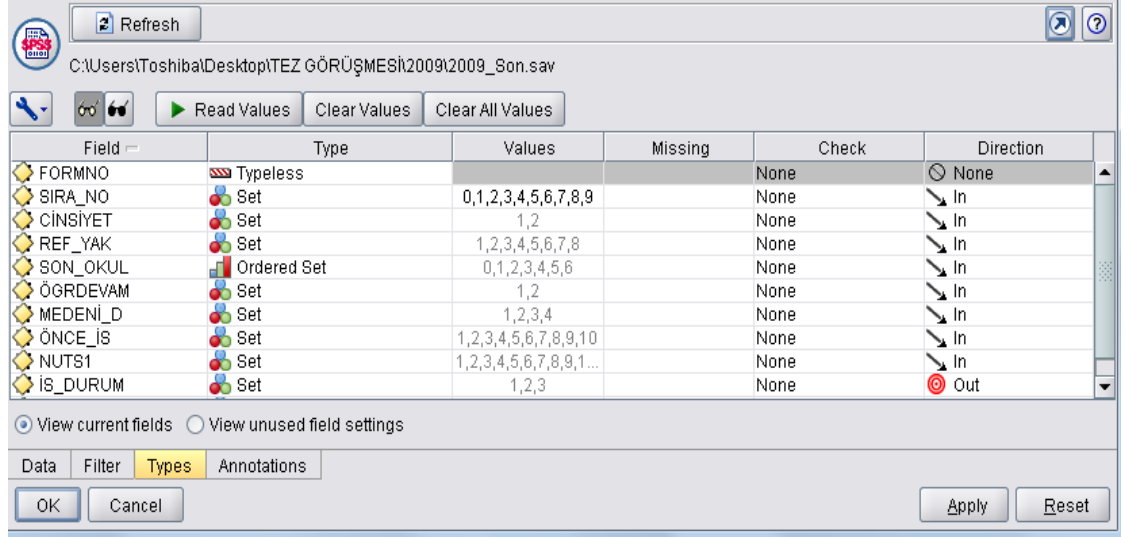
1- İstanbul (İstanbul) / 2- Batı Marmara (Tekirdağ-Balıkesir) / 3- Ege (İzmir-Aydın-Manisa) / 4- Doğu Marmara (Bursa-Kocaeli) / 5- Batı Anadolu (Ankara-Konya) / 6- Akdeniz (Antalya-Adana-Hatay) / 7- Orta Anadolu (Kırıkkale-Kayseri) / 8- Batı Karadeniz (Zonguldak-Kastamonu-Samsun) / 9- Doğu Karadeniz (Trabzon) / 10- KuzeyDoğu Anadolu (Erzurum-Ağrı) / 11- Orta Doğu Anadolu (Malatya-Van) / 12- Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin)

- İşgücü durumu

1-İstihdamda / 2-İşsiz / 3-İşgücüne dahil değil

- Kır-kent

1-Kır / 2-Kent



Şekil 4.2: Uygulamada Kullanılan Hedef ve Açıklayıcı Değişkenler

Uygulamada kullanılan hedef değişken ve açıklayıcı değişkenlere ilişkin frekans tabloları aşağıda özetlenmektedir.

Tablo 4.1’de görüldüğü üzere, 2009 yılında 503.329 kişiyle gerçekleştirilen ankete cevap verenlerin %48,7’si erkek ve %51,3’ü kadındır. 2010 yılında ise 522.171 kişiyle gerçekleştirilen ankete cevap verenlerin % 48,8’i erkek ve % 51,2’si kadındır.

Tablo 4.1: 2009 Yılı ve 2010 Yılı Cinsiyet Frekans Tablosu

	2009		2010	
Cinsiyet	Frekans	Yüzde	Frekans	Yüzde
Erkek	245.120	48,7	255.053	48,8
Kadın	258.209	51,3	267.118	51,2
TOPLAM	503.329	100,0	522.171	100,0

Tablo 4.2’de görüldüğü gibi, 2009 yılında 6 yaş ve üzerindeki kişilere sorulan en son bitirilen okul sorusundan elde edilen verilere göre, Türkiye genelinde görüşülen 503.329 kişinin %30,7 gibi çok büyük bir oranı bir okul bitirmemiştir ve %10,3’ü de okur-yazar bile değildir. Yine görüşülen kişilerin %27,1 gibi büyük bir bölümü ilkokul mezunudur. Lise dengi bir okul bitiren kişilerin oranı ise %12’dir. Görüşülen

kişiler içinde eğitime devam edenlerin oranı %22,9'dur. 2010 yılında; 522.171 kişinin okul bitirmememe oranı %29,9'a düşmektedir. Ortaokul, mesleki ortaokul ve ilköğretim, lise ve yüksekokul bitirenlerin yüzdesi artmaktadır. Genel olarak okuryazarlık oranı ve eğitimin kalitesinin arttığı söylenebilir.

**Tablo 4.2: 2009 Yılı ve 2010 Yılı Eğitim Durumu (Son Bitirilen Okul)
Frekans Tablosu**

	2009		2010	
Eğitim Durumu	Frekans	Yüzde	Frekans	Yüzde
6 yaşından küçük olanlar	50.442	10,0	50.604	9,7
Bir okul bitirmeyen	154.770	30,7	156.296	29,9
İlkokul (5 yıl)	136.317	27,1	139.702	26,8
Ortaokul, mesleki ortaokul ve ilköğretim (8 yıl)	69.805	13,9	76.529	14,7
Genel lise	34.771	6,9	36.447	7,0
Mesleki veya teknik lise	25.890	5,1	26.894	5,2
Yüksekokul, fakülte ve üzeri	31.334	6,2	35.699	6,8
TOPLAM	503.329	100,0	522.171	100,0

2009 Yılında eğitime devam etme durumu %22,9 iken 2010 yılında bu oran %23,1'e yükselmiştir.

Tablo 4.3: 2009 Yılı ve 2010 Yılı Eğitime Devam Etme Durumu Frekans Tablosu

	2009		2010	
Eğitime Devam Etme Durumu	Frekans	Yüzde	Frekans	Yüzde
Evet	115.159	22,9	120.519	23,1
Hayır	337.728	67,1	351.048	67,2
6 yaşından küçük olanlar	50.442	10,0	50.604	9,7
TOPLAM	503.329	100,0	522.171	100,0

Tablo 4.4' de 2009 yılında ankete katılanların %17'si hiç evlenmemiş, %49,3'ünün evli, %1,2'sinin boşanmış, %4,5'inin eşi ölmüş ve %27,2'sinin 15 yaşından küçük olduğu görülmektedir. 2010 yılına göre, ankete katılan kişilerin %18'i hiç evlenmemiş, %49,7'si evli, %1,3'ünün boşanmış, %4,7'sinin eşi ölmüş ve %26,3'ünün 15 yaşından küçük olduğu görülmektedir.

Tablo 4.4: 2009 Yılı Ve 2010 Yılı Medeni Durum Frekans Tablosu

	2009		2010	
Medeni Durum	Frekans	Yüzde	Frekans	Yüzde
Hiç evlenmedi	89.407	17,8	137.325	18,0
Evli	248.167	49,3	94.193	49,7
Boşandı	6.126	1,2	259.561	1,3
Eşi öldü	22.573	4,5	6.773	4,7
15 yaşından küçük olanlar	137.056	27,2	24.319	26,3
TOPLAM	503.329	100,0	522.171	100,0

Referans kişilerin frekans tablosuna bakıldığı zaman, ankete katılan kişilerin hem 2009 hem 2010 yılında çekirdek aile yapısını oluşturanların ağırlıkta olduğu görülmektedir.

Tablo 4.5: 2009 Yılı Ve 2010 Yılı Referans Kişiye Yakınlık Durumu Frekans Tablosu

	2009		2010	
Referans kişiye yakınlık durumu	Frekans	Yüzde	Frekans	Yüzde
Referans kişi	135.891	27,0	143.871	27,6
Eşi	110.247	21,9	115.455	22,1
Çocuğu	209.644	41,7	214.123	41,0
Gelini veya damadı	10.263	2,0	10.687	2,0
Torunu	18.392	3,7	19.875	3,8
Kendi veya eşinin anne/babası	97.25	1,9	91.62	1,8
Diğer akrabalar	7.240	1,4	6.942	1,3
Akraba olmayanlar	1.927	,4	2.056	,4
TOPLAM	503.329	100,0	522.171	100,0

Tablo 4.6’da görüldüğü gibi 2009 yılında ankete cevap verenlerin %26,2’si Marmara Bölgesi’nde, %13,4’ünün Ege Bölgesi’nde, % 14,5’inin İç Anadolu, % 11,3 Akdeniz, % 12,1 Karadeniz, %11,2 Doğu Anadolu ve %11,3 Güney Doğu Anadolu Bölgesi’nde ikamet etmektedir. 2010 yılına bakılırsa, %25,5’i Marmara Bölgesi’nde, %12,1’inin Ege Bölgesi’nde, %18,4’ünün İç Anadolu, %11,1 Akdeniz, %11,9 Karadeniz, %10,9 Doğu Anadolu ve % 10,1 Güney Doğu Anadolu Bölgesi’nde ikamet etmektedir.

Tablo 4.6: 2009 Yılı ve 2010 Yılı İstatistiki Bölge Birimleri Frekans Tablosu

	2009		2010	
Düzey 1	Frekans	Yüzde	Frekans	Yüzde
İstanbul (İstanbul)	59.359	11,8	56.720	10,9
Batı Marmara (Tekirdağ-Balıkesir)	29.099	5,8	31.992	6,1
Ege (İzmir-Aydın-Manisa)	67.331	13,4	63.039	12,1
Doğu Marmara (Bursa-Kocaeli)	43.654	8,7	44.440	8,5
Batı Anadolu (Ankara-Konya)	48.475	9,6	65.093	12,5
Akdeniz (Antalya-Adana-Hatay)	57.010	11,3	58.202	11,1
Orta Anadolu (Kırıkkale-Kayseri)	24.792	4,9	30.630	5,9
Batı Karadeniz (Zonguldak-Kastamonu-Samsun)	43.491	8,6	37.464	7,2
Doğu Karadeniz (Trabzon)	17.373	3,5	24.752	4,7
Kuzey Doğu Anadolu (Erzurum-Ağrı)	25.081	5,0	29.332	5,6
Orta Doğu Anadolu (Malatya-Van)	30.955	6,2	27.832	5,3
Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin)	56.709	11,3	52.675	10,1
TOPLAM	503.329	100,0	522.171	100,0

Tablo 4.7’de görüldüğü gibi 2009 yılında ankete cevap verenlerin %31,1 kırsal kesimde, %68,5’i kentsel kesimde ikamet etmektedir. Bu oranlar 2010 yılında kırsal alan için %30,1’e düşerken kentsel alanda %69,9’a yükselmektedir.

Tablo 4.7: 2009 Yılı ve 2010 Yılı Kır-Kent Frekans Tablosu

	2009		2010	
Kır-Kent	Frekans	Yüzde	Frekans	Yüzde
Kır (nüfusu 20 000 ve daha az olan yerleşim yerleri)	158.568	31,5	157.389	30,1
Kent (nüfusu 20 001 ve daha fazla olan yerleşim yerleri)	344.761	68,5	364.782	69,9
TOPLAM	503.329	100,0	522.171	100,0

Bir yıl önceki işgücü durumuna bakıldığında mevcut işinden memnun olanların 2010 yılında 2009 yılına göre arttığı gözlemlenmektedir. 2010 yılında emekliye ayrılanların yüzdesi artarken, ev işleriyle meşgul olan kesimden işgücüne dahil olduğu söylenebilir. İş arayanların yüzdesinin düşmesi ise eğitime devam etme durumundaki artış ya da işe girdikleri yönünde açıklanabilir.

Tablo 4.8: 2009 Yılı ve 2010 Yılı Bir Yıl Önceki İşgücü Durumu Frekans Tablosu

	2009		2010	
Bir yıl önceki işgücü durumu	Frekans	Yüzde	Frekans	Yüzde
Bir işte çalışıyordu	28.305	5,6	29.153	5,6
Şu anki işinde çalışıyordu	114.960	22,8	122.770	23,5
Emekliydi	28.061	5,6	30.359	5,8
İş arıyordu	25.624	7,0	24.300	6,3
Ev işleriyle meşguldü	108.720	21,6	109.763	21,0
Eğitim/öğretimine devam ediyordu	33.513	6,7	37.375	7,2
Özürlü veya hastaydı	16.528	3,3	17.193	3,3
Askerdi	2.612	0,5	2.756	0,5
Yaşlı (65 ve daha yukarı yaş)	8.559	1,7	9.292	1,8
15 yaşından küçük olanlar	139.059	25,3	137.325	24,6
TOPLAM	503.329	100,0	522.171	100,0

2009 yılında 503.329 kişiyle gerçekleştirilen ankette 15 yaş ve üzerindeki kişiler arasında istihdamda olanların sayısı 145.934'tür ve oranı %29,0'dur. 2010 Yılında ise 159.343 kişi ile bu oran %30,5'e yükselirken işsizlik oranının %3,9'a gerilediği işgücüne dahil olmayan grubun oranında ise bir değişme olmadığı gözlemlenmektedir.

Tablo 4.9: 2009 Yılı Ve 2010 Yılı İşgücü Durumu Frekans Tablosu

	2009		2010	
İşgücü durumu	Frekans	Yüzde	Frekans	Yüzde
İstihdam	145.934	29,0	159.343	30,5
İşsiz	22.420	4,5	20.483	3,9
İşgücüne dahil olmayan	197.919	39,3	205.020	39,3
15 yaşından küçük olanlar	137.056	27,2	137.325	26,3
TOPLAM	503.329	100,0	522.171	100,0

Tablo 4.10'daki değerler ışığında, 2009 yılında ankete cevap veren 15 yaş üzerindeki 366.273 kişi için hesaplanan işgücü göstergeleri şöyledir:

Tablo 4.10: 2009 Yılı İşgücü Göstergesi Özet Tablosu

İşgücü = İstihdam + İşsiz	145934 + 22420	168354
15+ Yaştaki Kurumsal Olmayan Nüfus = İşgücü + İşgücüne Dahil Olmayanlar	168354+197919	366.273
İstihdam Oranı = İstihdam / Aktif Nüfus	145934/366273	%39,8
İşsizlik Oranı = İşsiz / İşgücü	22420/168354	%13,3
İşgücüne Katılım Oranı = İşgücü / 15+Yaştaki Kurumsal Olmayan Nüfus	168354 /366273	%46,0

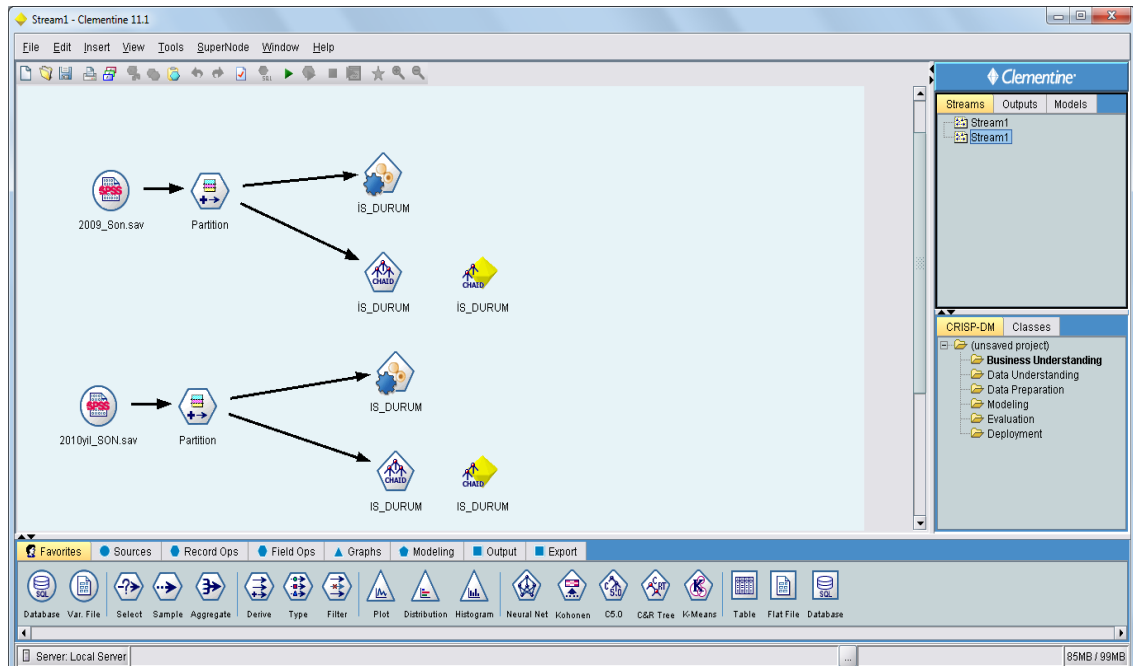
Tablo 4.11'deki deęerler ışığında, 2010 yılında ankete cevap veren 15 yaşı üzerindeki 384.846 kişı için hesaplanan işgücü göstergeleri şöyledir:

Tablo 4.11: 2010 Yılı İşgücü Göstergesi Özet Tablosu

İşgücü = İstihdam + İşsiz	159343 + 20483	179826
15+ Yaştaki Kurumsal Olmayan Nüfus = İşgücü + İşgücüne Dahil Olmayanlar	179826+205020	384846
İstihdam Oranı = İstihdam / Aktif Nüfus	159343/ 384346	%41,4
İşsizlik Oranı = İşsiz / İşgücü	20483/179826	%11,3
İşgücüne Katılım Oranı = İşgücü / 15+Yaştaki Kurumsal Olmayan Nüfus	179826/384846	%46,7

4.6.3. Modelin Kurulması ve Deęerlendirilmesi

Veri ön işlemleri yapıldıktan sonra analiz için hedef kitle belirlenmiştir. Bu aşamadan sonra, SPSS Clementine 11.1 paket programı ile 2009 yılı için 364.270 veri ve 2010 yılı için 384.823 veri analiz edilmiştir. Sonrasında, *eđitim (training)* (%70) ve *test (testing)* (%30) olmak üzere partition işlemcisi ile iki örnek gruba ayrılmıştır.



Şekil 4.3: İşgücü Durumu Veri Setinin Eğitim (*Traning*) Ve Test (*Testing*) Olarak Gruplandırılması

Verilerin % 70' ini oluşturan eğitim (training) veri setinde farklı sınıflandırma algoritmalarından C&R Tree, CHAID, Lojistik Regresyon, QUEST, Neural Net, Decision List algoritmaları denenerek model geliştirilmiş ve oluşturulan modellerin hepsi test (testing) veri seti ile test edilmiştir.

Eğitim (training) veri seti üzerinde uygulanan sınıflandırma algoritmaları sonucunda kurulan modellerin doğruluk yüzdeleri 2009 yılı ve 2010 yılı için Şekil 4.4 ve 4.5’ de görülmektedir.

Generate	Model	Build time (mins)	Max Profit	Max profit occurs in (%)	Lift (Top 30%)	Overall accuracy (%)	No. fields used	Area under curve
☐	CHAD 1	< 1	876.734,157	54	1,817	89,2	7	0,985
☐	Logistic regression 1	< 1	880.111,5	53	1,816	87,546	9	0,984
☐	Neural net 1	< 1	882.252,872	54	1,807	89,29	9	0,984
☐	C&R Tree 1	< 1	876.666,271	54	1,76	89,2	6	0,974
☐	QUEST 1	< 1	872.937,531	54	1,806	89,2	8	0,982
☐	Decision List 1	< 1	143.588	100	1,076	43,638	4	0,534

Şekil 4.4: 2009 Yılı Sınıflandırma Algoritmalarının Doğruluk Yüzdeleri

Generate	Model	Build time (mins)	Max Profit	Max profit occurs in (%)	Lift (Top 30%)	Overall accuracy (%)	No. fields used	Area under curve
☐	CHAD 1	< 1	909.054,194	54	1,831	89,859	6	0,983
☐	Logistic regression 1	< 1	912.025	53	1,828	88,865	9	0,982
☐	Neural net 1	< 1	913.130	53	1,814	89,702	9	0,98
☐	QUEST 1	< 1	906.276,421	53	1,819	89,859	8	0,961
☐	C&R Tree 1	< 1	902.997,911	53	1,77	89,859	4	0,949
☐	Decision List 1	< 1	125.995	100	1,072	44,799	4	0,532

Şekil 4.5: 2010 Yılı Sınıflandırma Algoritmalarının Doğruluk Yüzdeleri

Bu algoritmalar arasında CHAD algoritması, training (eğitim) verisinin analizinde, her iki yıl için de tahmin etme doğruluğunun daha yüksek çıkmıştır. Bu nedenle

çalışmanın amacına uygun olarak, değişkenler arasındaki ilişkinin ötesinde, değişkenlerin içerisindeki veri grupları arasındaki ilişkilerin en alt detayına kadar araştırılmasına olanak sağladığından yapılan çalışmanın amacına uygun veri madenciliği algoritması olarak belirlenmiştir.

Test verileri model işlemcisinden geçirildiğinde doğruluk değeri bir miktar düşer, eğer doğruluk değeri geniş ölçüde azalıyorsa, *eğitim* verisinde modelin aşırı öğrenme (overfit) eğiliminde olduğunun bir göstergesi olabilir. Eğer doğruluk çok az bir miktarda düşüyorsa bu modelin gelecekte iyi çalışacağını kanıtlar. İşgücü durumu için uygulanan algoritmanın sonucu da bu modelin iyi çalıştığının bir göstergesi olmaktadır.

Kurulan modelin doğruluğunun değerlendirilmesinde basit geçerlilik testi kullanılmış olup α birleştir = α böl = 0.05 ve ağaç derinliği 5 olarak alınmıştır. CHAID algoritması sonucu kurulan modelin test risk yüzdeleri her iki yıl için de aşağıda verilmiştir.

Test Sample

Risk		
Sample	Estimate	Std. Error
Training	,098	,001
Test	,099	,001
Growing Method: CHAID Dependent Variable: IS_DURUM		

Test Sample

Risk		
Sample	Estimate	Std. Error
Training	,101	,001
Test	,101	,001
Growing Method: CHAID Dependent Variable: is_durum		

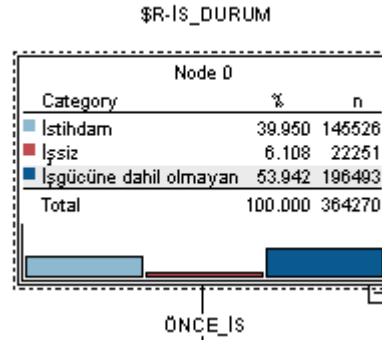
**Şekil 4.6: 2009 Yılı Ve 2010 Yılı
CHAID Algoritması Sonucu Kurulan Modelin Test Yüzdeleri**

4.6.4. Algoritmanın Uygulanması

Hedef değişken İşgücü Durumu için uygulanan CHAID algoritması sonucu 2009 ve 2010 yılları için elde edilen karar ağacı ve dallarının görünümüleri Ek 1 ve Ek 2’de gösterilmektedir.

4.6.4.1. 2009 YILI:

CHAID algoritması sonucunda elde edilen karar ağacının başlangıç nodu (düğümü) aşağıda Şekil 4.7’de gösterilmektedir.



Şekil 4.7: CHAID Algoritması Sonucu Elde Edilen Karar Ağacının Başlangıç Nodu Gösterimi 2009 Yılı

Dallanmanın (bölümlemenin) başladığı nodda toplam 364.270 kişi arasından istihdam’da olan 145.526 (%39,95), işsiz olan 22.251 kişi (%6,108) ve işgücüne dahil olmayan 196.493 (%53,942) kişi bulunmaktadır.

İşgücü durumu olarak tanımlanan hedef değişken için dallara ayırmada önceki iş durumu yani Önce_İs açıklayıcı değişkeni önemli bulunmuştur. İlk dallanma, önceki iş durumu değişkenine göre başlamakta ve kişilerin önceki iş durumunda verdikleri cevaplara göre kollara/dallara ayrılmaktadır.

İşgücü durumu için, önemli bulunan Önce_İs açıklayıcı değişkenine göre Şekil 4.8’den anlaşılacağı üzere 8 alt sınıf meydana gelmiştir.

1. Önce_İs = Alt Sınıf Bir İşte Çalışıyordu

Node 1		
Category	%	n
Istihdam	52.167	14766
İşsiz	25.681	7269
İşgücüne dahil olmayan	22.152	6270
Total	7.770	28305



2. Önce_İs= Şu anki işinde Çalışıyordu

Node 8		
Category	%	n
Istihdam	100.000	114960
İşsiz	0.000	0
İşgücüne dahil olmayan	0.000	0
Total	31.559	114960



3. Önce_İs = Emekliydi

Node 9		
Category	%	n
Istihdam	1.707	479
İşsiz	1.012	284
İşgücüne dahil olmayan	97.281	27298
Total	7.703	28061



4. Önce_İs= “İş Arıyordu” ; “Askerdi”

Node 17		
Category	%	n
Istihdam	37.804	9687
İşsiz	43.826	11230
İşgücüne dahil olmayan	18.369	4707
Total	7.034	25624



5. Önce_İs= “Ev İşleriyle Meşguldü”

Node 23		
Category	%	n
Istihdam	2.769	3011
İşsiz	1.632	1774
İşgücüne dahil olmayan	95.599	103935
Total	29.846	108720



Şekil 4.8: Elde Edilen 8 Alt Sınıfın Gösterimi 2009 Yılı

6. Önce_İs= “Eğitim/Öğretimine Devam ediyordu”

Node 46		
Category	%	n
Istihdam	7.185	2408
İşsiz	4.616	1547
İşgücüne dahil olmayan	88.199	29558
Total	9.200	33513

7. Önce_İs= “Özürlü veya Hastaydı”

Node 55		
Category	%	n
Istihdam	1.240	205
İşsiz	0.877	145
İşgücüne dahil olmayan	97.882	16178
Total	4.537	16528

8. Önce_İs= “Yaşlı(65 ve daha yukarı yaş)”

Node 60		
Category	%	n
Istihdam	0.117	10
İşsiz	0.023	2
İşgücüne dahil olmayan	99.860	8547
Total	2.350	8559

Şekil 4.9: Elde Edilen 8 Alt Sınıfın Gösterimi 2009 Yılı (Devam)

4.6.4.2. 2010 YILI:

CHAID algoritması sonucunda elde edilen karar ağacının başlangıç nodu (düğümü) aşağıda Şekil 4.10’da gösterilmektedir.

Node 0		
Category	%	n
1.000	41.403	159330
2.000	5.322	20482
3.000	53.274	205011
Total	100.000	384823

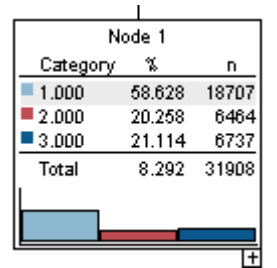
Şekil 4.10: CHAID Algoritması Sonucu Elde Edilen Karar Ağacının Başlangıç Nodu Gösterimi 2010 Yılı

Dallanmanın (bölümlenmenin) başladığı nodda toplam 384.823 kişi bulunmaktadır. Bunların arasından istihdamda olan 159.330 (%41.403), işsiz olan 20.482 kişi (%5.322) ve işgücüne dahil olmayan 205.011 (%53.274) kişi bulunmaktadır.

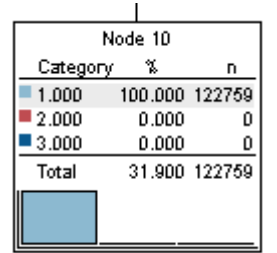
İşgücü durumu olarak tanımlanan hedef değişken için dallara ayırmada 2009 yılındaki gibi önceki iş durumu yani Önce_İs açıklayıcı değişkeni önemli bulunmuştur. İlk dallanma, önceki iş durumu değişkenine göre başlamakta ve kişilerin önceki iş durumunda verdikleri cevaplara göre kollara/dallara ayrılmaktadır.

İşgücü durumu için, önemli bulunan Önce_İs açıklayıcı değişkenine göre şekil 4.11'den anlaşılacağı üzere 8 alt sınıf meydana gelmiştir.

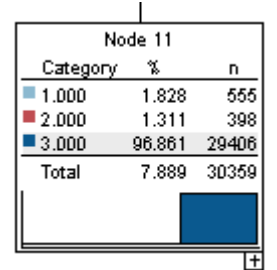
1. Alt sınıf Önce_İs= “Bir İşte Çalışıyordun” ;“Askerdi”



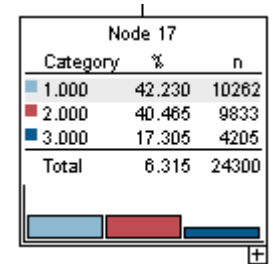
2. Alt sınıf Önce_İs= “Şu anki işinde çalışıyordun”



3. Alt sınıf Önce_İs= “Emekliydi”



4. Alt sınıf Önce_İs= “İş Arıyordu”



Şekil 4.11: Elde Edilen 8 Alt Sınıfın Gösterimi 2010 Yılı

5. Alt sınıf Önce_İs= “Ev işleriyle meşguldü”

Node 23		
Category	%	n
1.000	3.244	3561
2.000	1.765	1937
3.000	94.991	104260
Total	28.522	109758

6. Alt sınıf Önce_İs= “Eğitim/öğretimine devam ediyordu”

Node 39		
Category	%	n
1.000	8.241	3235
2.000	4.445	1745
3.000	87.314	34276
Total	10.201	39256

7. Alt sınıf Önce_İs= “Özürlü veya hastaydı”

Node 53		
Category	%	n
1.000	1.390	239
2.000	0.611	105
3.000	97.999	16847
Total	4.467	17191

8. Alt sınıf Önce_İs= “Yaşlı (65 veya daha yukarı yaş)”

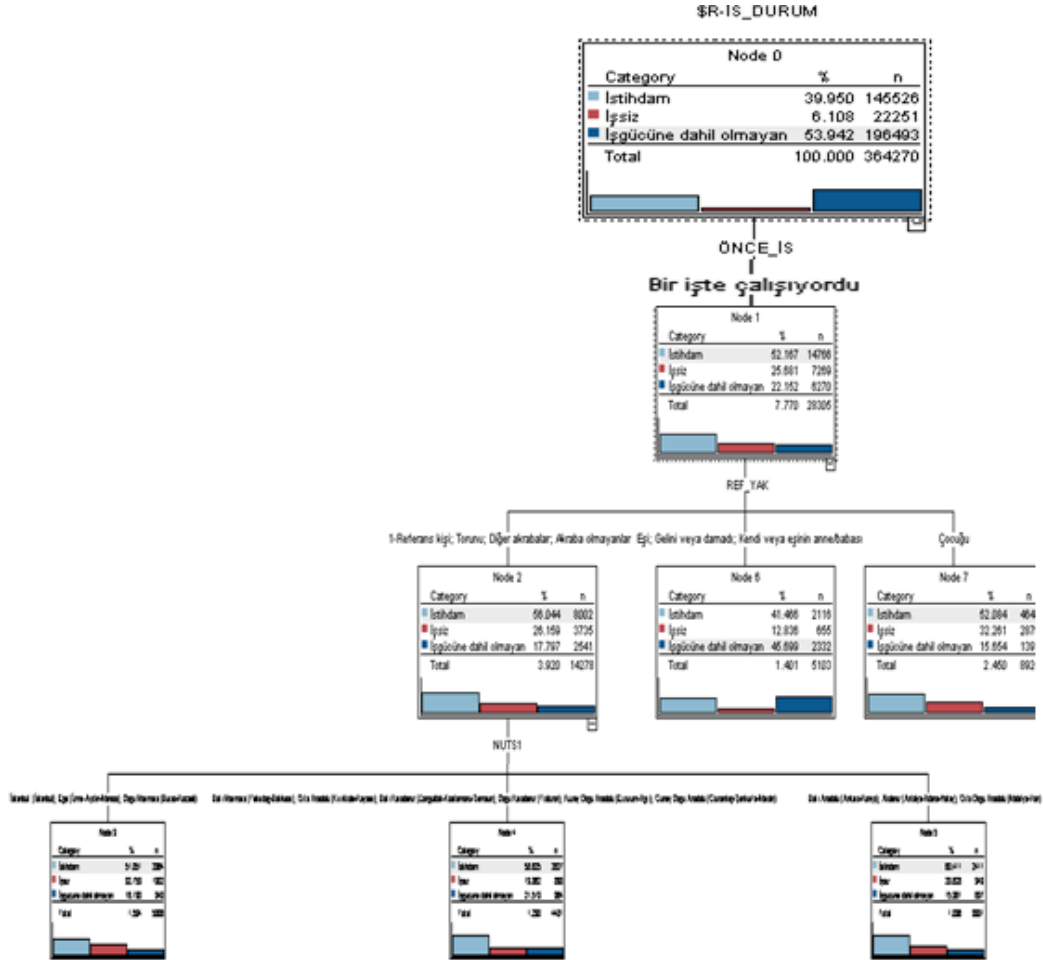
Node 56		
Category	%	n
1.000	0.129	12
2.000	0.000	0
3.000	99.871	9280
Total	2.415	9292

Şekil 4.12: Elde Edilen 8 Alt Sınıfın Gösterimi 2010 Yılı (Devam)

Başlangıç düğümü ile oluşan karar ağacı ana dalları aşağıdaki şekillerde başlıklar halinde sırasıyla 2009 ve 2010 Yılları için ifade edilmektedir.

4.6.5. 2009 YILI Ana Dalları

4.6.5.1. Karar Ağacı Birinci Anadalı



Şekil 4.13: 2009 Karar Ağacı Birinci Anadalı

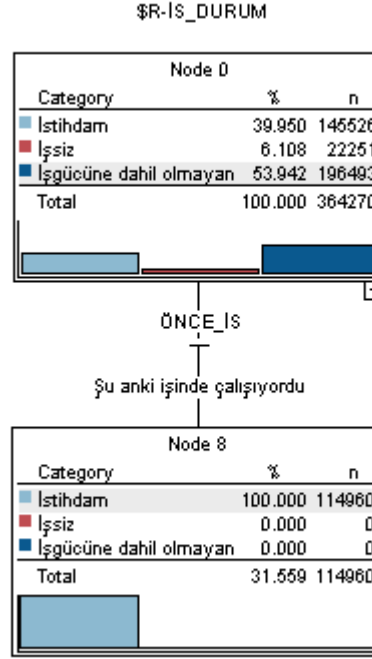
Şekil 4.13'ten görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), bir işte çalışıyordu, grubunda olan kişi sayısı 28.305 olup, bu kişilerin 14.766'sı (% 52.17) istihdam edilirken, 7.269'u ($\approx$ %26) işsiz grubunda olup kalan 6.270 kişi ise işgücüne dahil olmamaktadır. Önceki işgücü durumu "bir işte çalışıyordu" olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *referans kişiye yakınlık* olup;

- Referans kişiye yakınlık durumu; referans kişinin kendisi, torunu, diğer akrabaları ve akrabası olmayan 14.278 kişinin 8.002'si (%56) istihdam edilirken, 3.735'i (%26,2) işsiz olup 2.541 kişi ise (%18) işgücüne dahil olmamaktadır.
- Referans kişiye yakınlık durumu; eşi, gelini veya damadı, kendi veya eşinin anne / babası olan 5.103 kişinin 2.116'sı (%41,2) istihdamda iken, 655'i (%13) işsiz ve 2.332'si (% 46) işgücüne dahil olmamaktadır.
- Referans kişiye yakınlık durumu; çocuğu olan 8.924 kişinin 4.648' i (%52) istihdam edilirken, 2.879'u (%32,3) işsiz ve 1.397'si (%16) işgücüne dahil olmamaktadır.

Referans kişiye yakınlık durumu; referans kişinin kendisi, torunu, diğer akrabaları ve akrabası olmayan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1'i İstanbul (İstanbul), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli) olan 5.806 kişinin 2.964'ü (%51,1) istihdam edilirken, 1.902'si (%33) işsiz olup 940'ı (%17) işgücüne dahil olmamaktadır.
- NUTS1'i Batı Marmara (Tekirdağ-Balıkesir), Orta Anadolu (Kırıkkale-Kayseri), Batı Karadeniz (Zonguldak-Kastamonu-Samsun), Doğu Karadeniz (Trabzon), Kuzey Doğu Anadolu (Erzurum-Ağrı), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan 4.481 kişinin 2.627'si (%59) istihdam edilmekte, 890'ı (%20) işsiz olup 964'ü (%22) işgücüne dahil olmamaktadır.
- NUTS1'i Batı Anadolu (Ankara-Konya), Akdeniz (Antalya-Adana-Hatay), Orta Doğu Anadolu (Malatya-Van) olan 3.991 kişinin 2.411'i (%60,4) istihdam edilmekte, 943'ü (%24) çalışmamakta olup 637'si (%16) işgücüne dahil olmamaktadır.

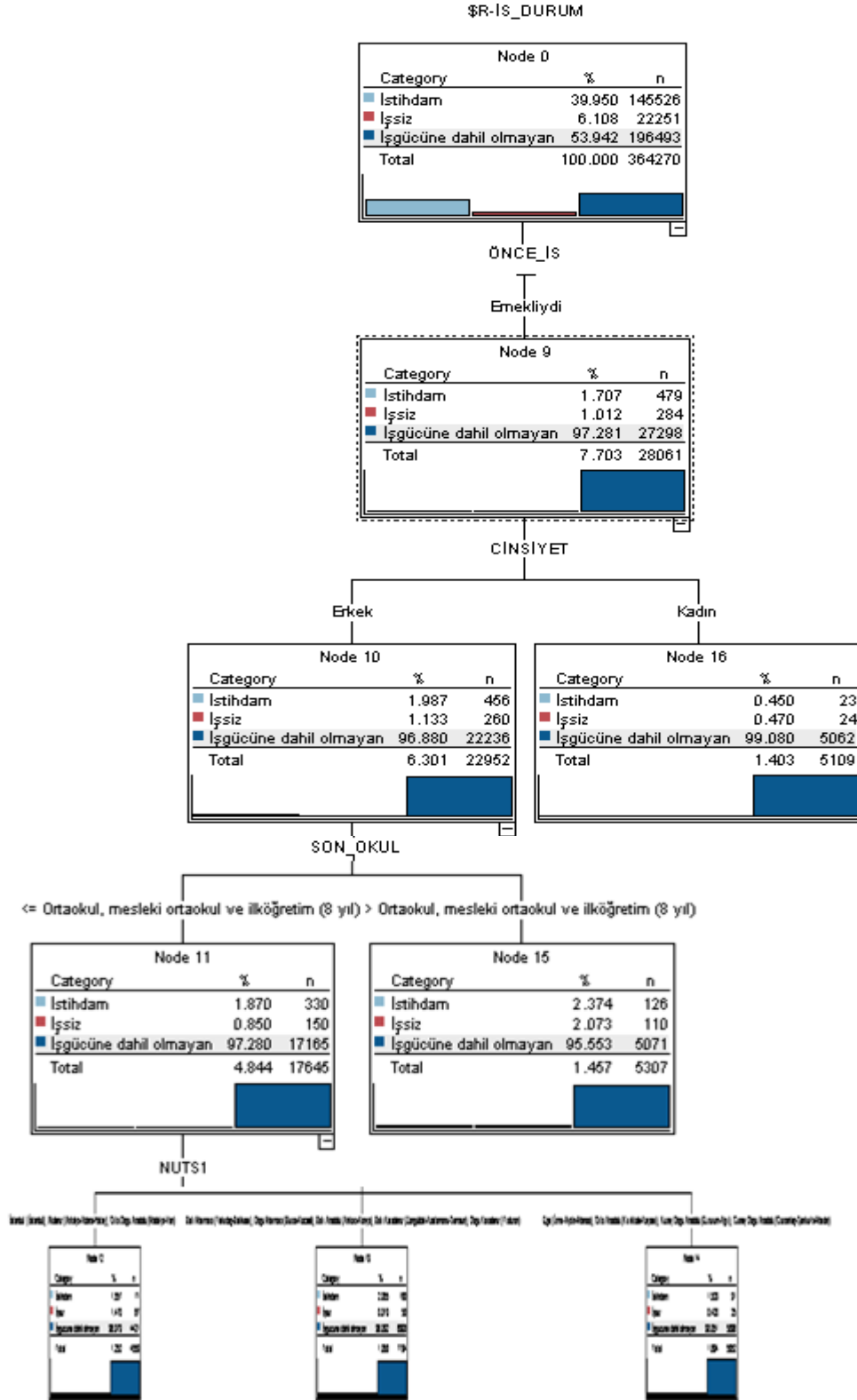
4.6.5.2. Karar Ağacı İkinci Anadalı



Şekil 4.14: 2009 Karar Ağacı İkinci Anadalı

Şekil 4.14’te görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), şu anki işinde çalışıyordu grubunda olan kişi sayısı 114.960 olup, bu kişilerin 114.966’sı yani tamamı (%100) istihdam edilmektedir.

4.6.5.3. Karar Ağacı Üçüncü Anadalı



Şekil 4.15: 2009 Karar Ağacı Üçüncü Anadalı

Şekil 4.15'ten görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), emekliydi olan kişi sayısı 28.061 olup, bu kişilerin 479'u (%2) istihdam edilirken, 284'ü (%1) işsiz grubunda olup kalan çoğunluk olan 27.298 (%97) kişi ise işgücüne dahil olmamaktadır. Önceki İşgücü Durumu “emekliydi” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *cinsiyet* olup;

- Cinsiyeti erkek olan 22.952 kişinin 456'sı (%2) istihdam edilirken, 260'ı (%1) işsiz grubunda olup kalan çoğunluk olan 22.236'sı (%97) ise işgücüne dahil olmamaktadır.
- Cinsiyeti kadın olan 5.109 kişinin 23'ü (%1) istihdam edilirken, 24'ü (%1) işsiz grubunda olup ve 5.062'si (%99) işgücüne dahil olmamaktadır.

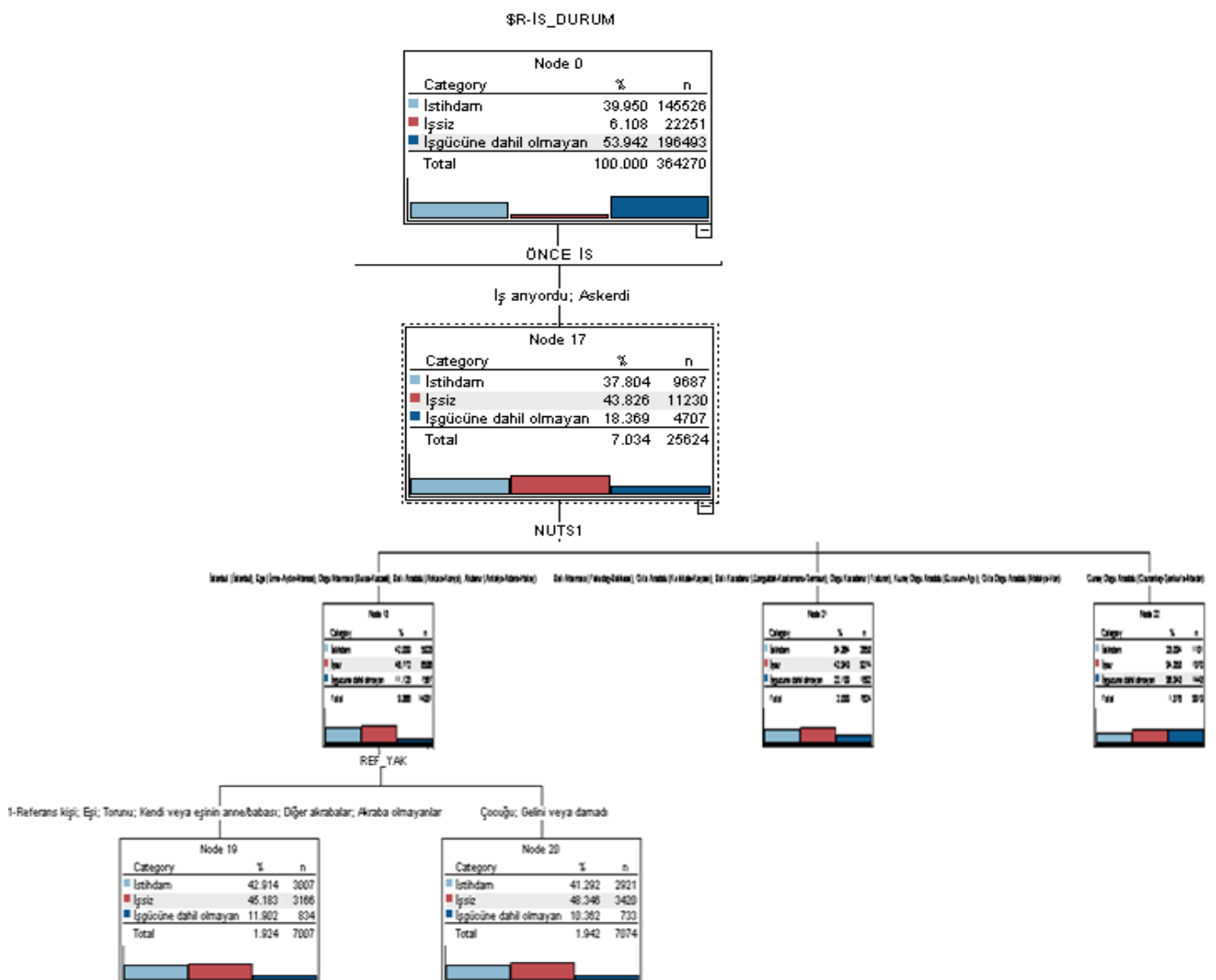
Cinsiyeti erkek olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken “*en son bitirilen okul*” olup;

- En son bitirdiği okul düzeyi, “Ortaokul, mesleki ortaokul ve ilköğretim (8 yıl)” düzeyine eşit ya da bu düzeyden daha az olan 17.645 kişinin 330'u (%2) istihdam edilirken, 150'si (%1) işsiz olup, 17.165'i (%97) işgücüne dahil olmamaktadır.
- En son bitirdiği okul düzeyi, “Ortaokul, mesleki ortaokul ve ilköğretim (8 yıl)” düzeyinden daha çok olan 5.307 kişinin 126'sı (%2) istihdam edilirken, 110'u (%2) işsiz olup, 5.071'i (%96) işgücüne dahil olmamaktadır.

En son bitirdiği okul düzeyi, “Ortaokul, mesleki ortaokul ve ilköğretim (8 yıl)” düzeyine eşit ya da bu düzeyden daha az olan olan kişilerde, dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistikî Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1'i İstanbul (İstanbul), Akdeniz (Antalya-Adana-Hatay), Orta Doğu Anadolu (Malatya-Van) olan 4.421 kişinin 71' i (%2) istihdam edilirken, 67'si (%2) işsiz olup 4.421' i (%97) işgücüne dahil olmamaktadır.
- NUTS1'i Batı Marmara (Tekirdağ-Balıkesir), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya), Batı Karadeniz (Zonguldak-Kastamonu-Samsun), Doğu Karadeniz (Trabzon) olan 7.134 kişinin 168' i (%)

- NUTS1'i Ege (İzmir-Aydın-Manisa), Orta Anadolu (Kırıkkale-Kayseri), Kuzey Doğu Anadolu (Erzurum-Ağrı), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan 5.952 kişinin 91'i (%2) istihdam edilmekte, 25' i (%0,4) işsiz olup 5.836' s ı (%98) işgücüne dahil olmamaktadır.



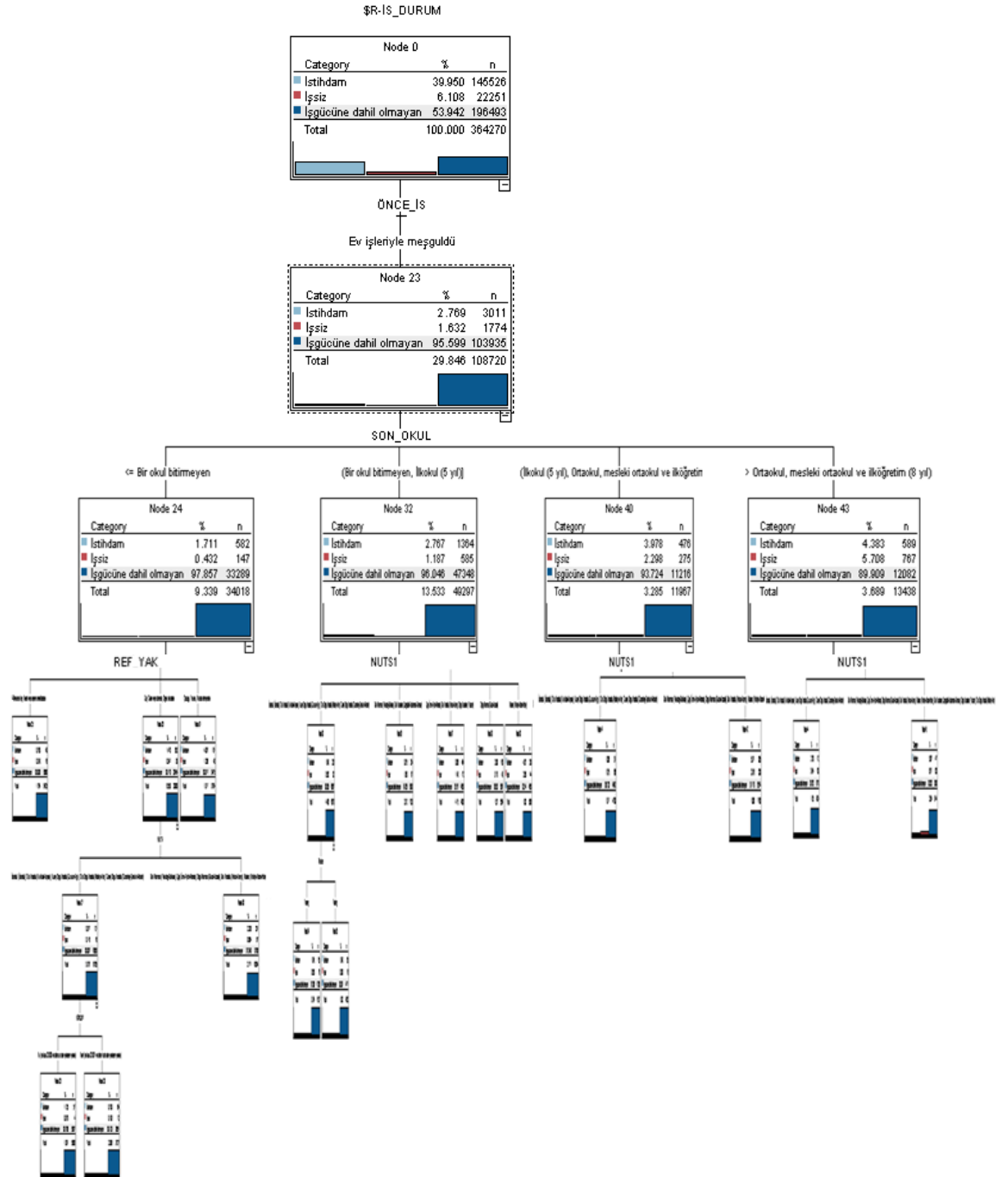
Şekil 4.16'dan görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs) “iş arıyordu; askerdi” olan gruptaki toplam kişi sayısı 25.624 olup, bu kişilerin 9.687’si (%38) İstihdam edilirken, 11.230’u (%44) işsiz grubunda olup 4.707’si (%18) kişi ise işgücüne dahil olmamaktadır. Önceki işgücü durumu “iş arıyordu; askerdi” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1’i İstanbul (İstanbul), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya), Akdeniz (Antalya-Adana-Hatay) olan 14.081 kişinin 5.928’i (%42) istihdam edilirken, 6.586’sı (%47) işsiz olup 1.567’si (%11) işgücüne dahil olmamaktadır.
- NUTS1’i Batı Marmara (Tekirdağ-Balıkesir), Orta Anadolu (Kırıkkale-Kayseri), Batı Karadeniz (Zonguldak-Kastamonu-Samsun), Doğu Karadeniz (Trabzon), Kuzey Doğu Anadolu (Erzurum-Ağrı), Orta Doğu Anadolu (Malatya-Van) olan 7.624 kişinin 2.658’i (%35) istihdam edilmekte, 3.274’ü (%43) işsiz olup 1.692’si (%22) işgücüne dahil olmamaktadır.
- NUTS1’i Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan 3919 kişinin 1.101’i (%28) istihdam edilmekte, 1.370’ i (%35) işsiz olup 1.440’ı (%37) işgücüne dahil olmamaktadır.

NUTS1 yani İstatistiki Bölge Birimleri Sınıflamasında İstanbul (İstanbul), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya), Akdeniz (Antalya-Adana-Hatay) grubunda olan kişilerde, dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *referans kişiye yakınlık durumu* olup;

- Referans Kişiye Yakınlık Durumu; Referans Kişinin kendisi, eşi, torunu, kendi veya eşinin anne/ babası, diğer akrabaları ve akrabası olmayan 7.007 kişinin 3.007’si (%43) istihdam edilirken, 3.166’sı (%45) işsiz olup 834 kişi ise (%12) işgücüne dahil olmamaktadır.
- Referans Kişiye Yakınlık Durumu; çocuğu, gelini veya damadı olan 7.074 kişinin 2.921’i (%41,2) istihdamda iken, 3.420’si (%48) işsiz ve 733’ü (%10) işgücüne dahil olmamaktadır.

4.6.5.5. Karar Ağacı Beşinci Anadalı



Şekil 4.17: 2009 Karar Ağacı Beşinci Anadalı

Şekil 4.17’den görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), ev işleriyle meşguldü grubunda olan kişi sayısı 108.720 olup, bu kişilerin 3.011’i (%3) İstihdam edilirken, 1.774’ü (%2) işsiz grubunda olup kalan çoğunluk olan 103.935 kişi ise işgücüne dahil olmamaktadır.

Önceki İşgücü Durumu “ev işleriyle meşguldü” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *en son bitirilen okul* olup;

- En son bitirdiği okul düzeyi, “bir okul bitirmeyen” düzeyine eşit ya da bu düzeyden daha az olan 34.018 kişinin 582’si (%2) istihdam edilirken, 147’si (%0,4) işsiz olup, 33.289’u (%98) işgücüne dahil olmamaktadır.
- En son bitirdiği okul düzeyi, “bir okul bitirmeyenden daha üst ve İlkokul (5 yıl) düzeyine eşit ya da daha alt düzeyde” olan 49.297 kişinin 1.364’ü (%3) istihdam edilirken, 585’i (%1) işsiz olup, 47.348’i (%96) işgücüne dahil olmamaktadır.
- En son bitirdiği okul düzeyi, “ilkokul (5 yıl) düzeyinden daha üst düzeyde ve ortaokul, mesleki ortaokul ve ilköğretim (8 yıl) düzeyine eşit ya da daha alt düzeyde” olan 11.967 kişinin 476’sı (% 4) İstihdam edilirken, 275’i (%2) işsiz olup, 11.216’sı (%94) işgücüne dahil olmamaktadır.
- En son bitirdiği okul düzeyi, “ortaokul, mesleki ortaokul ve ilköğretim (8 yıl) düzeyinden daha yukarı düzeyde” olan 13.438 kişinin 589’u (%4) İstihdam edilirken, 767’si (%6) işsiz olup, 12.082’si (%90) işgücüne dahil olmamaktadır.

A) En son bitirdiği okul düzeyi; “Bir okul bitirmeyen” düzeyine eşit ya da bu düzeyden az olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *referans kişiye yakınlık (REF_YAK)* olup;

- Referans kişiye yakınlık durumu; referans kişinin kendisi, kendi veya eşinin anne / babası olan 6.425 kişinin sadece 45’i (%1) istihdam edilirken, 16’sı (%0,2) işsiz olup 6.360 kişi ise (%99) işgücüne dahil olmamaktadır.
- Referans kişiye yakınlık durumu; referans kişinin eşi, gelini veya damadı, diğer akrabaları olan 23.889 kişinin 352’si (%2) istihdam edilirken, 82’si (%0,3) işsiz olup 23.454 kişi ise (%98) işgücüne dahil olmamaktadır.
- Referans kişiye yakınlık durumu; referans kişinin çocuğu, torunu olan ve akrabası olmayanların oluşturduğu gruptaki 3.704 kişinin 181’i (%5) istihdam edilirken, 48’i (%1) işsiz olup 3.475 kişi ise (%94) işgücüne dahil olmamaktadır.

Referans kişiye yakınlık durumu; “referans kişinin eşi, gelini veya damadı, diğer akrabaları olan” kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistikî Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1’i İstanbul (İstanbul), Orta Anadolu (Kırıkkale-Kayseri), Kuzey Doğu Anadolu (Erzurum-Ağrı), Orta Doğu Anadolu (Malatya-Van), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan 13.795 kişinin 121’i (%1) istihdam edilirken, 16’sı (%0,2) işsiz olup 13.658’i (%99) işgücüne dahil olmamaktadır.
- NUTS1’i Batı Marmara (Tekirdağ-Balıkesir), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya), Orta Anadolu (Kırıkkale-Kayseri), Akdeniz (Antalya-Adana-Hatay), Batı Karadeniz (Zonguldak-Kastamonu-Samsun), Doğu Karadeniz (Trabzon) olan 10.094 kişinin 231’i (%2) istihdam edilmekte, 67’si (%1) işsiz olup 9.796’sı (%97) işgücüne dahil olmamaktadır.

NUTS1’i İstanbul (İstanbul), Orta Anadolu (Kırıkkale-Kayseri), Kuzey Doğu Anadolu (Erzurum-Ağrı), Orta Doğu Anadolu (Malatya-Van), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *Kır-Kent* olup;

- Kır-Kent değişkeninde; yerleşim birimi olarak Kırsal kesimde (nüfusu 20.000 ve daha az olan yerleşim yerleri) yaşayan kişi sayısı 5.068 olup bunun 57’si (%1) istihdam edilirken çoğunluk olan 5.007’si (%99) işgücüne dahil olmamaktadır.
- Kır-Kent değişkeninde; yerleşim birimi olarak Kentsel kesimde (nüfusu 20.001 ve daha fazla olan yerleşim yerleri) yaşayan kişi sayısı 8.727 olup 8.651’i ($\approx$ %100) işgücüne dahil olmayıp istihdamda olan kişi sayısı yok denecek kadar azdır.

B) En son bitirdiği okul düzeyi; “Bir okul bitirmeyenden daha üst ve İlkokul (5 yıl) düzeyine eşit ya da daha alt düzeyde” olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistikî Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1'i İstanbul (İstanbul), Orta Anadolu (Kırıkkale-Kayseri), Kuzey Doğu Anadolu (Erzurum-Ağrı), Orta Doğu Anadolu (Malatya-Van), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan 16.379 kişinin 220'si (%1) istihdam edilirken, 92'si (%1) işsiz olup 16.067'si (%98) işgücüne dahil olmamaktadır.
- NUTS1'i Batı Marmara (Tekirdağ-Balıkesir), Batı Karadeniz (Zonguldak-Kastamonu-Samsun), olan 7.329 kişinin 284'ü (%4) istihdam edilmekte, 137'si (%2) işsiz olup 6.908'i (%95) işgücüne dahil olmamaktadır.
- NUTS1'i Ege (İzmir-Aydın-Manisa), Batı Anadolu (Ankara-Konya), Doğu Karadeniz (Trabzon) olan 14.989 kişinin sadece 464'ü (%4) istihdam edilmekte ve 14.353'ü (%96) işgücüne dahil olmamaktadır.
- NUTS1'i Doğu Marmara (Bursa-Kocaeli) olan 5.564 kişinin 163'ü (%3) istihdam edilmekte, 5.361'i (%96) işgücüne dahil olmamaktadır.
- NUTS1'i Akdeniz (Antalya-Adana-Hatay) olan 5.036 kişinin sadece 233'ü (%5) istihdam edilmekte ve 4.659'u (%93) işgücüne dahil olmamaktadır.

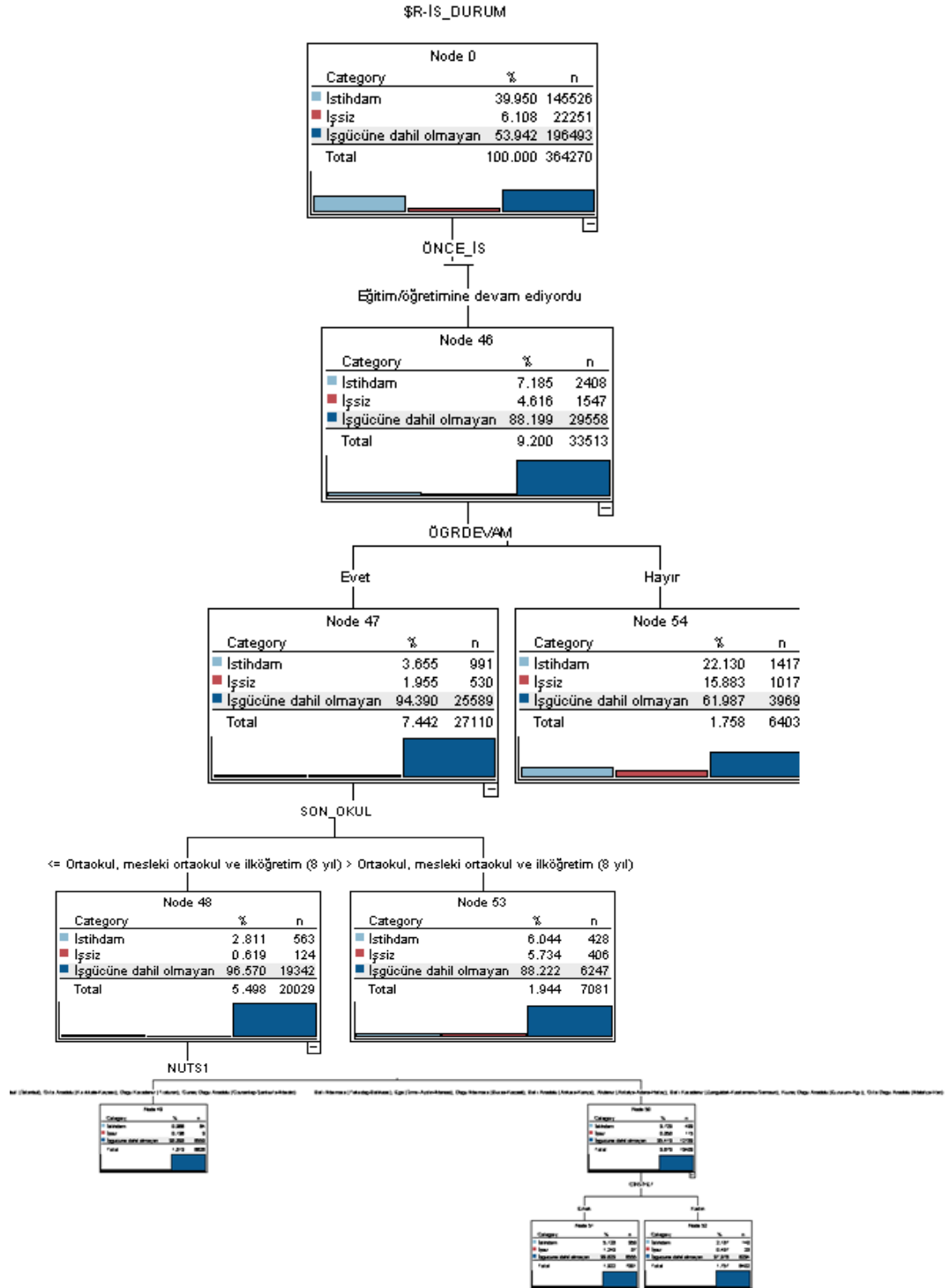
C) En son bitirdiği okul düzeyi; “İlkokul (5 yıl) düzeyinden daha üst düzeyde ve Ortaokul, mesleki ortaokul ve ilköğretim (8 yıl) düzeyine eşit ya da daha alt düzeyde” olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1'i İstanbul (İstanbul), Orta Anadolu (Kırıkkale-Kayseri), Kuzey Doğu Anadolu (Erzurum-Ağrı), Orta Doğu Anadolu (Malatya-Van), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan 4.799 kişinin 91'i (%2) istihdam edilirken, 66'sı (%1) işsiz olup 4.642'si (%97) işgücüne dahil olmamaktadır.
- NUTS1'i Batı Marmara (Tekirdağ-Balıkesir), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya), Akdeniz (Antalya-Adana-Hatay), Batı Karadeniz (Zonguldak-Kastamonu-Samsun), Doğu Karadeniz (Trabzon) olan 7.168 kişinin 385'ü (%5) istihdam edilmekte, 209'u (%3) işsiz olup 6.574'ü (%92) işgücüne dahil olmamaktadır.

D) “Ortaokul, mesleki ortaokul ve ilköğretim (8 yıl) düzeyinden daha yukarı düzeyde” olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1’i İstanbul (İstanbul), Orta Anadolu (Kırıkkale-Kayseri), Kuzey Doğu Anadolu (Erzurum-Ağrı), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan 4.799 kişinin 91’i (%2) istihdam edilirken, 66’sı (%1) işsiz olup 4.642’si (%97) işgücüne dahil olmamaktadır.
- NUTS1’i Batı Marmara (Tekirdağ-Balıkesir), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya), Akdeniz (Antalya-Adana-Hatay), Batı Karadeniz (Zonguldak-Kastamonu-Samsun), Doğu Karadeniz (Trabzon) olan 9.414 kişinin 477’si (%5) istihdam edilmekte, 628’i (%7) işsiz olup 8.309’u (%89) işgücüne dahil olmamaktadır.

4.6.5.6. Karar Ağacı Altıncı Anadalı



Şekil 4.18: 2009 Karar Ağacı Altıncı Anadalı

Şekil 4.18'den görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), eğitim/öğretimine devam ediyordu grubunda olan kişi sayısı 33.513 olup, bu kişilerin 2.408'i (%7) İstihdam edilirken, 1.547'si (%5) işsiz grubunda olup kalan çoğunluk

olan 29.558 kiři (%88) ise iřgücüne dahil olmamaktadır. Önceki iřgücü durumu “eđitim/ öđretimine devam ediyordu” olan kiřiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı deđiřken *bir öđrenim kurumuna devam edip etmediđi* olup;

- Bir öđrenim kurumuna devam edenlerin toplam sayısı 27.110 olup 991’i (%4) istihdam edilirken 530’u (%2) iřsiz olup 25.589’u (%62) iřgücüne dahil olmamaktadır.
- Bir öđrenim kurumuna devam etmeyenlerin toplam sayısı 6.403 olup 1.417’si (%22) istihdam edilirken 1,017’si (%16) iřsiz olup 3.969’u (%62) iřgücüne dahil olmamaktadır.

Bir öđrenim kurumuna devam edip etmediđi “Evet” olan kiřiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı deđiřken *son bitirdiđi okul düzeyi* olup;

- En son bitirdiđi okul düzeyi, “ortaokul, mesleki ortaokul ve ilköđretim (8 yıl)” düzeyine eřit ya da bu düzeyden daha az olan 20.029 kiřinin sadece 563’ ü (%3) istihdam edilirken çođunluđu oluřturan 19.342 (%97) kiři ise iřgücüne dahil olmamaktadır.
- En son bitirdiđi okul düzeyi, “ortaokul, mesleki ortaokul ve ilköđretim (8 yıl)” düzeyinden daha yukarı olan 7.081 kiřinin 428’i (%6) istihdamda, 406’sı (% 6) iřsiz ve 6247’si ise iřgücüne dahil olmamaktadır.

En son bitirdiđi okul düzeyi, “ortaokul, mesleki ortaokul ve ilköđretim (8 yıl)” düzeyine eřit ya da bu düzeyden daha az olan kiřilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı deđiřken *NUTS1 yani İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey* olup;

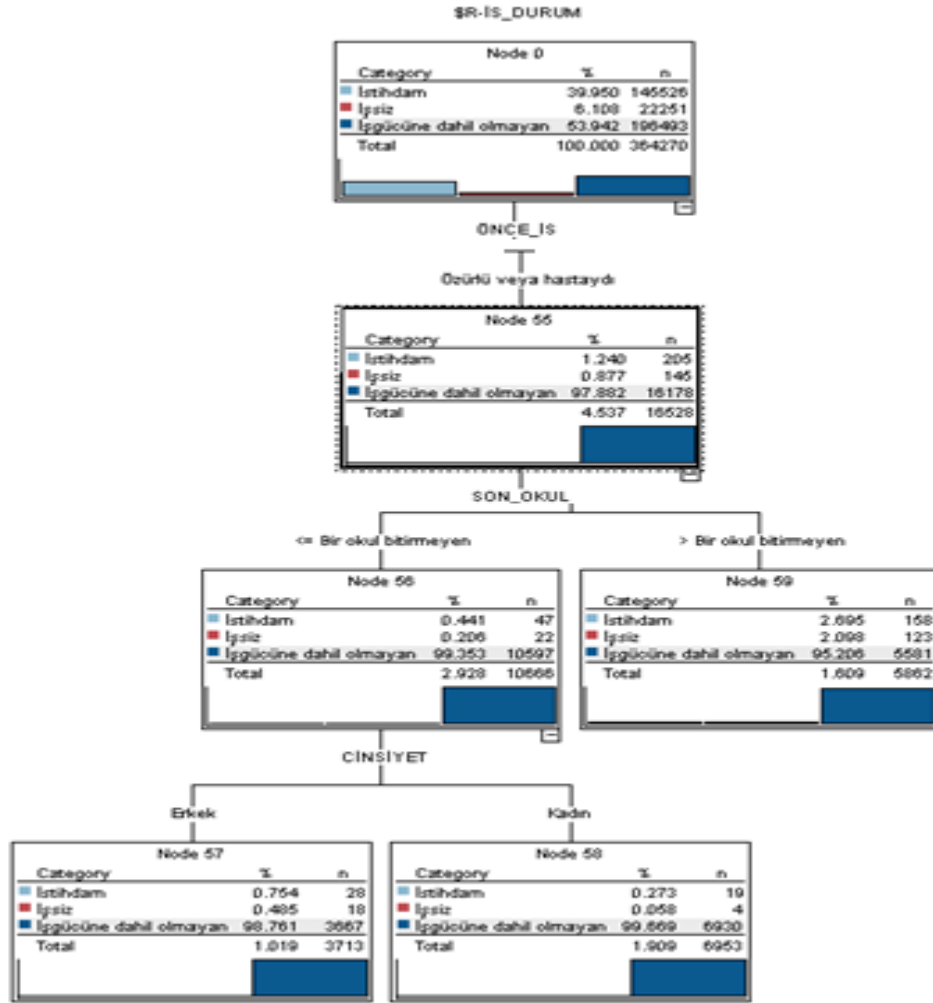
- NUTS1’i İstanbul (İstanbul), Orta Anadolu (Kırıkkale-Kayseri), Dođu Karadeniz (Trabzon) ve Güney Dođu Anadolu (Gaziantep-řanlıurfa-Mardin) olan 6.553 kiřinin %1’i ancak istihdam edilirken 6.553’ü (%99) iřgücüne dahil olmamaktadır.

- NUTS1'i Batı Marmara (Tekirdağ-Balıkesir), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya), Akdeniz (Antalya-Adana-Hatay), Batı Karadeniz (Zonguldak-Kastamonu-Samsun), Orta Doğu Anadolu (Malatya-Van), Kuzey Doğu Anadolu (Erzurum-Ağrı), olan 13.403 kişinin 499'u (%4) istihdam edilmekte, 12.789 (%96) işgücüne dahil olmamaktadır.

NUTS1'i Batı Marmara (Tekirdağ-Balıkesir), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya), Akdeniz (Antalya-Adana-Hatay), Batı Karadeniz (Zonguldak-Kastamonu-Samsun), Orta Doğu Anadolu (Malatya-Van), Kuzey Doğu Anadolu (Erzurum-Ağrı) olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *cinsiyet* olup;

- Cinsiyeti erkek olan 7.001 kişinin 359'u (%5) istihdamda, %1'i işsiz ve 6.555'i (%94) ise işgücüne dahil olmamaktadır.
- Cinsiyeti kadın olan 6.402 kişinin 140'ı (%2) istihdamda, %1'i işsiz ve 6.234'ü (%97) ise işgücüne dahil olmamaktadır.

4.6.5.7. Karar Ağacı Yedinci Anadalı



Şekil 4.19: 2009 Karar Ağacı Yedinci Anadalı

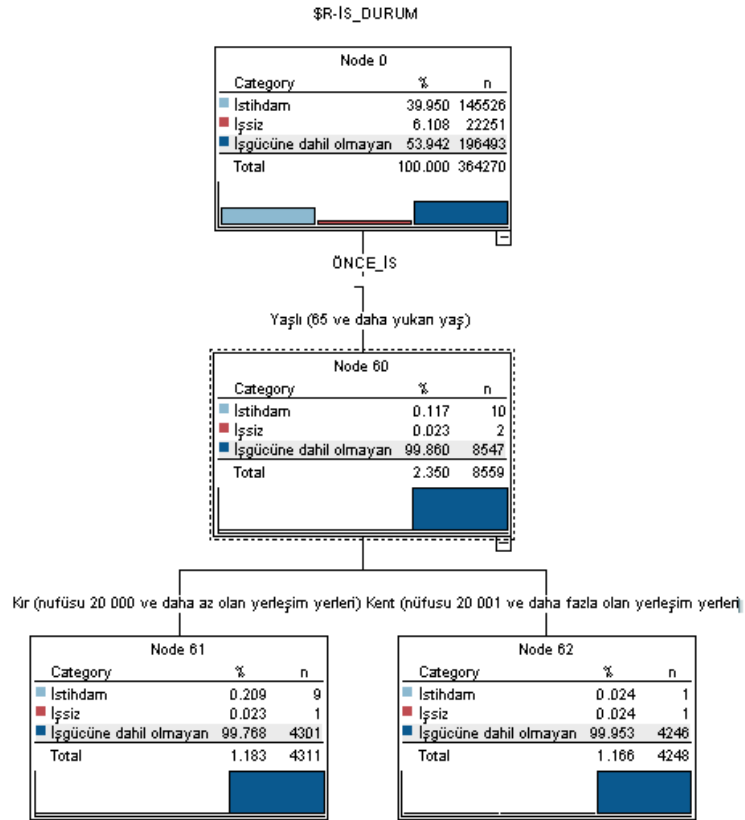
Şekil 4.19'dan görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), özürliydü veya hastaydı grubunda olan kişi sayısı 16.528 olup, bu kişilerin sadece 205'i (%1) İstihdam edilirken, 145'i (%1) işsiz grubunda olup kalan çoğunluk olan 16.178 kişi (%98) ise işgücüne dahil olmamaktadır. Önceki işgücü durumu "özürliydü veya hastaydı" olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *en son bitirdikleri okul düzeyi* olup;

- En son bitirilen okul düzeyi; bir okul bitirmeyen veya daha alt düzeyde olan 10.066 kişinin neredeyse tamamı olan 10.597'si (%99) işgücüne dahil olmamaktadır.
- En son bitirilen okul düzeyi; bir okul bitirmeyen düzeyinden yukarı olan yani en az İlkokul (5 yıl) düzeyinde olan 5,862 kişinin 158'i (%3) istihdam edilirken, %2'si işsiz olup 5.581'i (%95) işgücüne dahil olmamaktadır.

En son bitirilen okul düzeyi; “bir okul bitirmeyen veya daha alt düzeyde olan” 3.713 kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı cinsiyet olup;

- Cinsiyeti erkek olan 3.713 kişinin 3.667'si (%99) neredeyse tamamına yakını işgücüne dahil olmamaktadır.
- Cinsiyeti kadın olan 6.953 kişinin 6.930'u (%99) neredeyse tamamına yakını işgücüne dahil olmamaktadır.

4.6.5.8. Karar Ağacı Sekizinci Anadalı



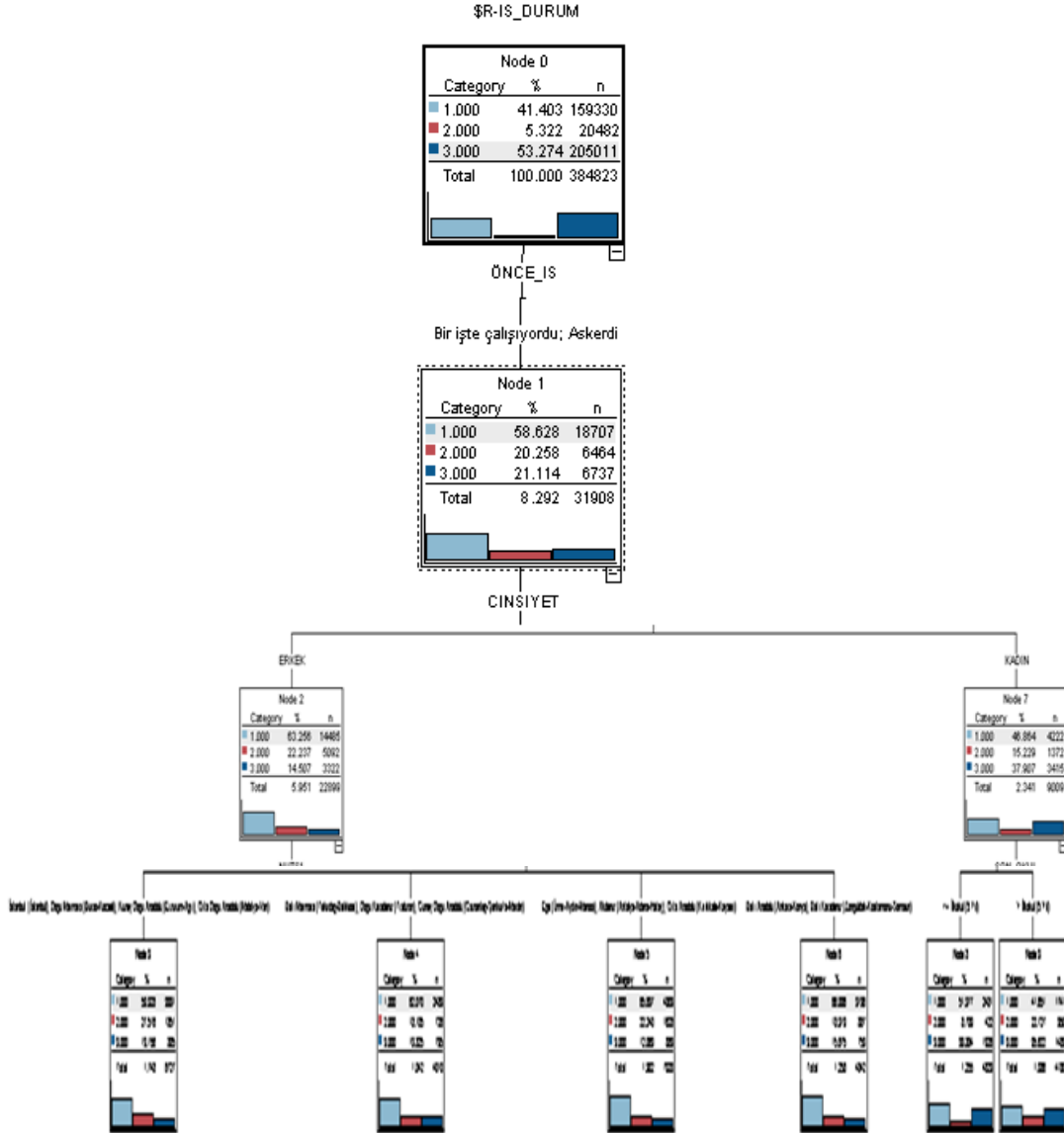
Şekil 4.20: 2009 Karar Ağacı Sekizinci Anadalı

Şekil 4.20’den görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), yaşlı (65 ve daha yukarı yaş) grubunda olan kişi sayısı 8.559 olup, bu kişilerin sadece (%0,1) istihdam edilirken, kalan çoğunluk olan 8.547 kişi (%99) ise işgücüne dahil olmamaktadır. Önceki işgücü durumu “yaşlı (65 ve daha yukarı yaş)” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *kır-kent* olup;

- Kır-Kent değişkeninde; yerleşim birimi olarak kırsal kesimde (nüfusu 20.000 ve daha az olan yerleşim yerleri) yaşayan kişi sayısı 4.311 olup bunun çoğunluk olan 4.301’i ($\approx$ % 100) işgücüne dahil olmamaktadır.
- Kır-Kent değişkeninde; yerleşim birimi olarak kentsel kesimde (nüfusu 20.001 ve daha fazla olan yerleşim yerleri) yaşayan kişi sayısı 4.248 kişi olup 4.246’sı ($\approx$ %100) işgücüne dahil olmayıp istihdamda olan kişi sayısı yok denecek kadar azdır.

4.6.6. 2010 YILI Ana Dalları

4.6.6.1. Karar Ağacı Birinci Anadalı



Şekil 4.21: 2010 Karar Ağacı Birinci Anadalı

Şekil 4.21’den görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), bir işte çalışıyordu, askerdi grubunda olan kişi sayısı 31.908 olup, bu kişilerin 18.707’si (%59) istihdam edilirken, 6.464’ü (%20) işsiz grubunda olup kalan 6.737 (%21) kişi ise işgücüne dahil olmamaktadır. Önceki işgücü durumu “bir işte çalışıyordu ve

askerdi” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *cinsiyet* olup;

- Cinsiyeti erkek olan 22.899 kişinin 14.485’i (% 63) istihdam edilirken, 5.092’si (% 22) işsiz grubunda olup 3.322’ si (% 16) ise işgücüne dahil olmamaktadır.
- Cinsiyeti kadın olan 9.009 kişinin 4.322’si (%47) istihdam edilirken, 1.372’ si (%15) işsizdir ve 3.415’i (%38) İşgücüne dahil olmamaktadır.

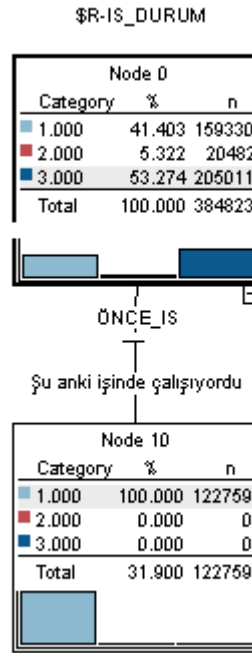
A) Cinsiyeti erkek olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1’i İstanbul (İstanbul), Doğu Marmara (Bursa-Kocaeli), Kuzey Doğu Anadolu (Erzurum-Ağrı), Orta Doğu Anadolu (Malatya-Van) olan 6.727 kişinin 3.991’i (%59) istihdam edilirken, 1.851’i (%28) işsiz olup 885’i (%13) işgücüne dahil olmamaktadır.
- NUTS1’i Batı Marmara (Tekirdağ-Balıkesir), Doğu Karadeniz (Trabzon), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan 4.010 kişinin 2.489’u (%62) istihdam edilmekte, 726’sı (%18) işsiz olup 795’i (%20) işgücüne dahil olmamaktadır.
- NUTS1’i Ege (İzmir-Aydın-Manisa), Orta Anadolu (Kırıkkale-Kayseri), Akdeniz (Antalya-Adana-Hatay), olan 7.320 kişinin 4.809 (%66) istihdam edilmekte, 1.628’i (%22) çalışmamakta olup 883’ü (%12) işgücüne dahil olmamaktadır.
- NUTS1’i Batı Anadolu (Ankara-Konya), Batı Karadeniz (Zonguldak-Kastamonu-Samsun) olan 4.842 kişinin 3.196’sı (%66) istihdam edilmekte, 887’ si (%18) çalışmamakta olup 759’u (%16) işgücüne dahil olmamaktadır.

B) Cinsiyeti kadın olan kişilerde ise dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken “*en son bitirilen okul*” olup;

- En son bitirdiği okul düzeyi, “ilkokul (5 yıl) düzeyine eşit ya da bu düzeyden daha az olan 4.829 kişinin 2.481’i (%51) istihdam edilirken, 422’si (%9) işsiz olup, 1.926’sı (%40) işgücüne dahil olmamaktadır.
- En son bitirdiği okul düzeyi, “ilkokul (5 yıl) düzeyinden daha yukarı” olan 4.180 kişinin 1.741’i (%42) istihdam edilirken, 950’si (%23) işsiz olup, 1.489’u (%36) işgücüne dahil olmamaktadır.

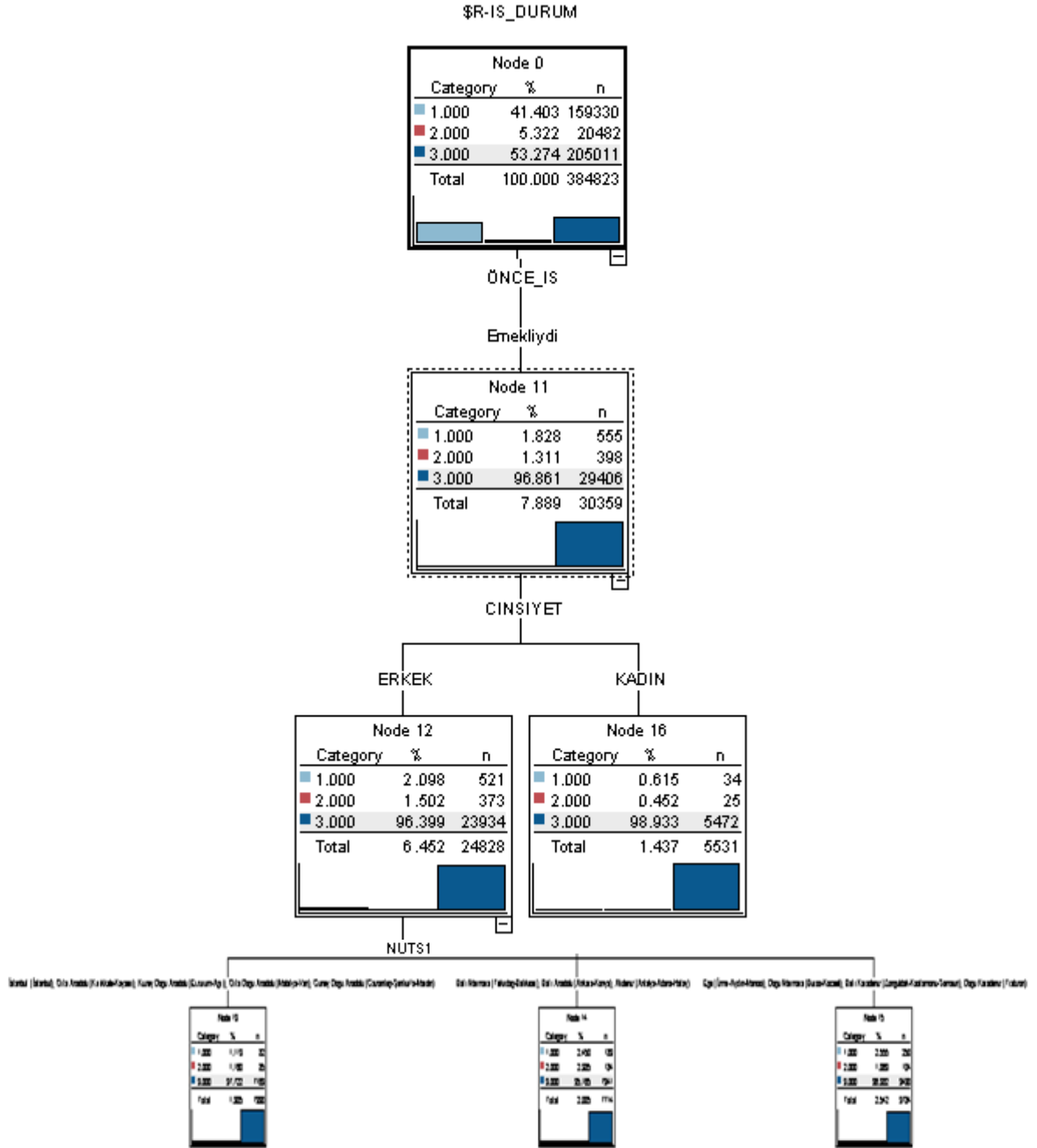
4.6.6.2. Karar Ağacı İkinci Anadalı



Şekil 4.22: 2010 Karar Ağacı İkinci Anadalı

Şekil 4.22’den görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), şu anki işinde çalışıyordu grubunda olan kişi sayısı 122.759 olup, bu kişilerin tamamı (%100) istihdam edilmektedir.

4.6.6.3. Karar Ağacı Üçüncü Anadalı



Şekil 4.23: 2010 Karar Ağacı Üçüncü Anadalı

Şekil 4.23'ten görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), emekliydi olan kişi sayısı 30.359 olup, bu kişilerin yalnız 555'i (%2) istihdam edilirken, kalan çoğunluk olan 29.406 (%97) kişi ise işgücüne dahil olmamaktadır. Önceki işgücü

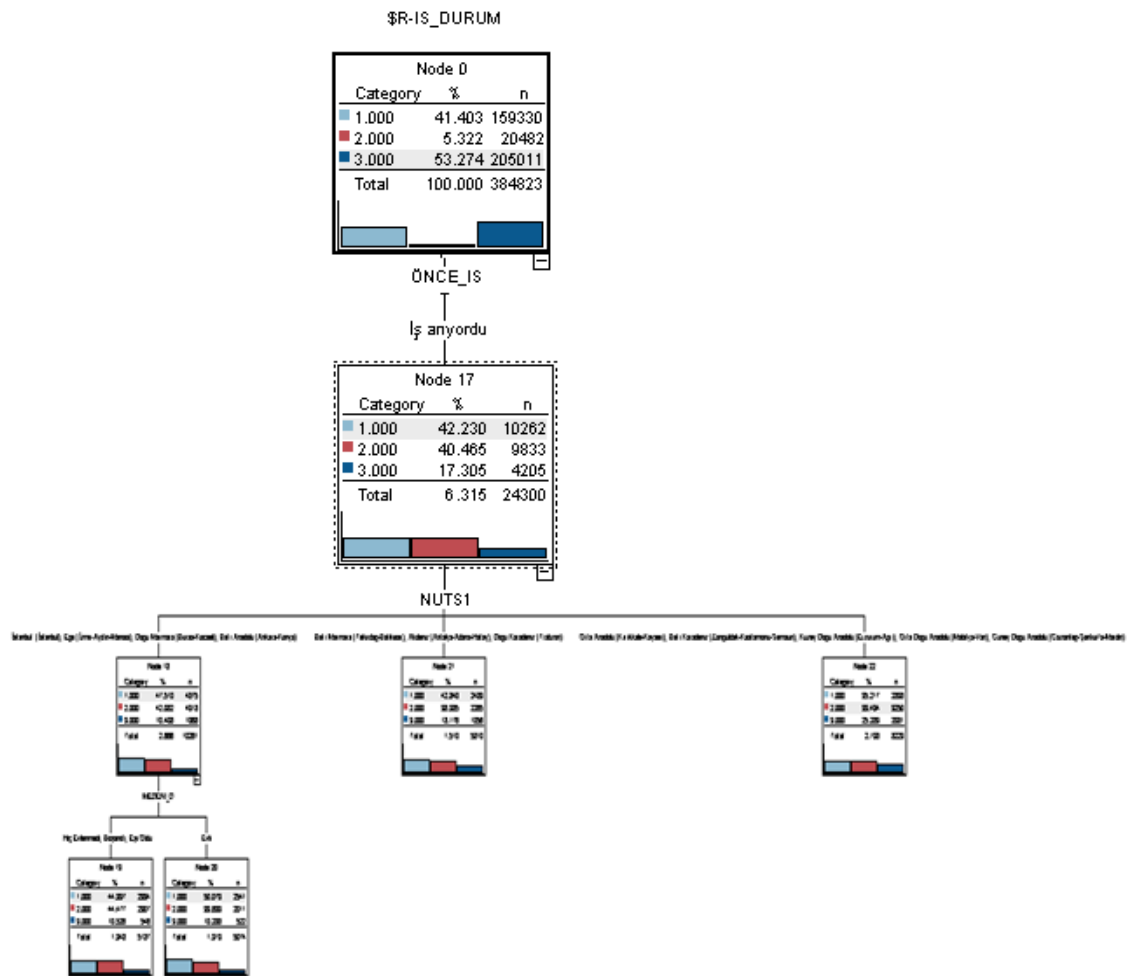
durumu “emekliydi” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *cinsiyet* olup;

- Cinsiyeti erkek olan 24.828 kişinin 521’i (%2) istihdam edilirken kalan çoğunluk olan 23.934’ü (%96) ise işgücüne dahil olmamaktadır.
- Cinsiyeti kadın olan 5.472 kişinin 23’ü (% 1) istihdam edilirken 5.472’si (%99) neredeyse tamamına yakını işgücüne dahil olmamaktadır.

Cinsiyeti erkek olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1’i İstanbul (İstanbul), Orta Anadolu (Kırıkkale-Kayseri), Kuzey Doğu Anadolu (Erzurum-Ağrı), Orta Doğu Anadolu (Malatya-Van), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan 7.330 kişinin %1’i istihdam edilirken, %1 işsiz olup çoğunluk olan 7.163’ü ise (%98) işgücüne dahil olmamaktadır.
- NUTS1’i Batı Marmara (Tekirdağ-Balıkesir), Batı Anadolu (Ankara-Konya), Akdeniz (Antalya-Adana-Hatay) olan 7.714 kişinin %2’si istihdam edilmekte olup 7.341’i (%95) işgücüne dahil olmamaktadır.
- NUTS1’i Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Doğu Karadeniz (Trabzon), Batı Karadeniz (Zonguldak-Kastamonu-Samsun) olan 9.784 kişinin %3’ü istihdam edilmekte, 9.430’u (%96) işgücüne dahil olmamaktadır.

4.6.6.4. Karar Ağacı Dördüncü Anadalı



Şekil 4.24: 2010 Karar Ağacı Dördüncü Anadalı

Şekil 4.24'ten görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), iş arıyordu olan gruptaki toplam kişi sayısı 24.300 olup, bu kişilerin 10.262'si (%42) İstihdam edilirken, 9.833'ü (%41) işsiz grubunda olup 4.205'i (%17) kişi ise işgücüne dahil olmamaktadır. Önceki işgücü durumu “iş arıyordu” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1'i İstanbul (İstanbul), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya) olan 10.261 kişinin 4.875'i

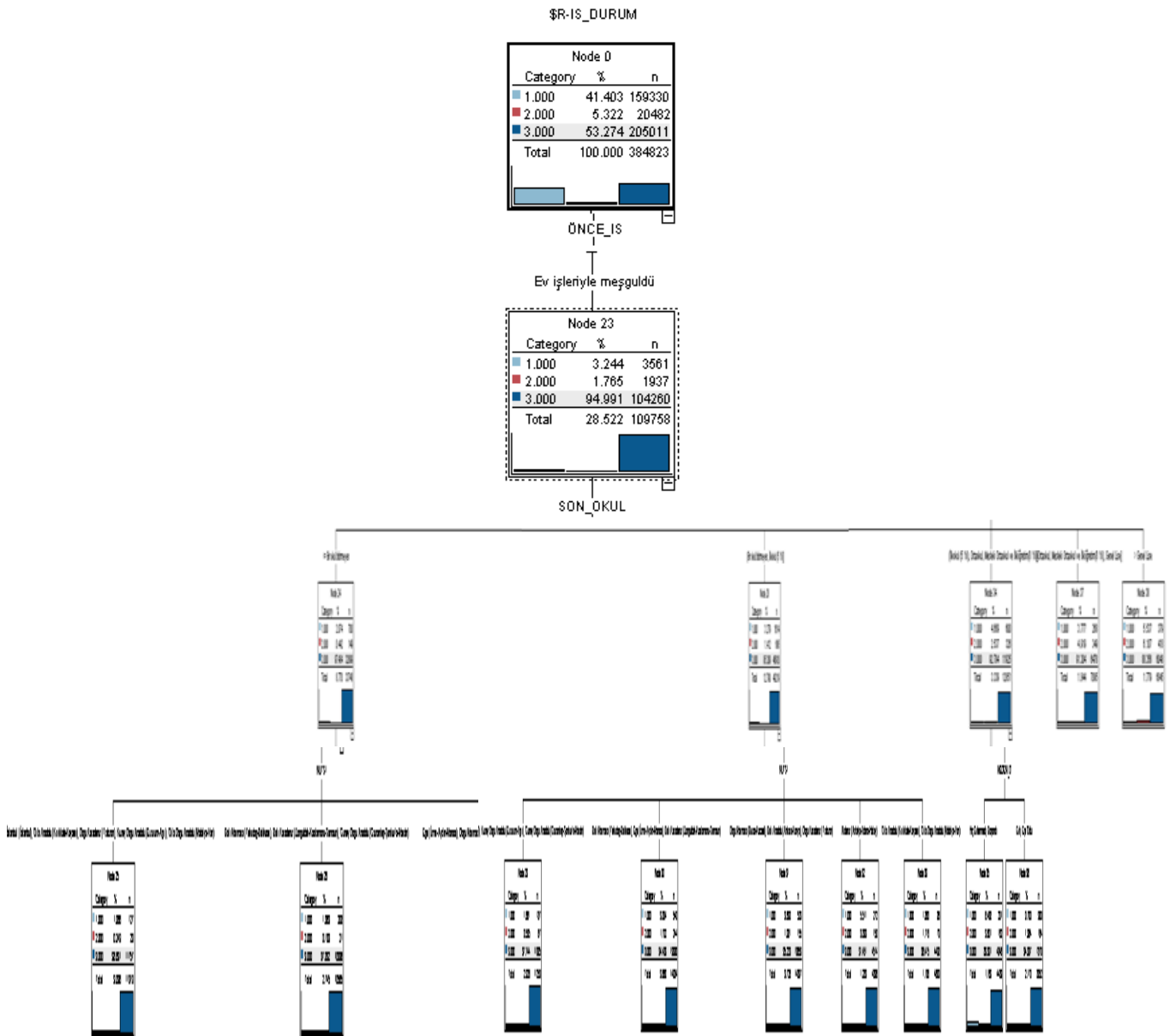
(%48) istihdam edilirken, 4.318'i (%42) işsiz olup 1.068'i (%10) işgücüne dahil olmamaktadır.

- NUTS1'i Batı Marmara (Tekirdağ-Balıkesir), Akdeniz (Antalya-Adana-Hatay), Doğu Karadeniz (Trabzon) olan 5.810 kişinin 2.489'u (%43) istihdam edilmekte, 2.265'i (%39) işsiz olup 1.056'sı (%18) işgücüne dahil olmamaktadır.
- NUTS1'i Orta Anadolu (Kırıkkale-Kayseri), Batı Karadeniz (Zonguldak-Kastamonu-Samsun), Kuzey Doğu Anadolu (Erzurum-Ağrı), Orta Doğu Anadolu (Malatya-Van), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin) olan 8.229 kişinin 2.898'i (%35) istihdam edilmekte, 3.250'si (%39) işsiz olup 2.081'i (%25) işgücüne dahil olmamaktadır.

NUTS1'i İstanbul (İstanbul), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya) olan kişilerde dallara ayırmada bir sonraki açıklayıcı değişken *medeni durum* olup;

- Medeni durumu; “hiç evlenmedi; boşandı; eşi öldü” olan 5.187 kişinin 2.334'ü (%45) istihdam edilmekte, 2.307'si (%44) işsiz olup 546'sı (%11) işgücüne dahil olmamaktadır.
- Medeni durumu; “evli” olan 5.074 kişinin 2.541'i (%50) istihdam edilmekte, 2.011'i (%40) işsiz olup 522'si (%10) işgücüne dahil olmamaktadır.

4.6.6.5. Karar Ağacı Beşinci Anadalı



Şekil 4.25: 2010 Karar Ağacı Beşinci Anadalı

Şekil 4.25'ten görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), ev işleriyle meşguldü grubunda olan kişi sayısı 109.758 olup, bu kişilerin 3.561'i (%3) İstihdam edilirken, 1.937'si (%2) işsiz grubunda olup kalan çoğunluk olan 104.260 kişi ise işgücüne dahil olmamaktadır. Önceki işgücü durumu “ev işleriyle meşguldü” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *en son bitirilen okul* olup;

- En son bitirdiği okul düzeyi, “bir okul bitirmeyen” düzeyine eşit ya da bu düzeyden daha az olan 32.899 kişinin 700 (%2) istihdam edilirken, 147’si (%0,4) işsiz olup, 32.889’u (%97) işgücüne dahil olmamaktadır.
- En son bitirdiği okul düzeyi, “Bir okul bitirmeyenden daha üst ve ilköğretim (5 yıl) düzeyine eşit ya da daha alt düzeyde” olan 49.219 kişinin 1.614’ü (%3) istihdam edilirken, 695’ i (%1) işsiz olup, 46.910’u (%95) işgücüne dahil olmamaktadır.
- En son bitirdiği okul düzeyi, “ilköğretim (5 yıl) düzeyinden daha üst düzeyde ve ortaokul, mesleki ortaokul ve ilköğretim (8 yıl) düzeyine eşit ya da daha alt düzeyde” olan 12.851 kişinin 600’ü (% 5) istihdam edilirken, %3 işsiz olup, kalan 11.925’ i (%93) işgücüne dahil olmamaktadır.
- En son bitirdiği okul düzeyi, “ortaokul, mesleki ortaokul ve ilköğretim (8 yıl) ve genel lise düzeyine eşit ya da daha alt düzeyde” olan 7.095 kişinin 268’i (%4) istihdam edilirken, %5 işsiz olup, kalan 6.478’i (%91) işgücüne dahil olmamaktadır.
- En son bitirdiği okul düzeyi, “genel lise düzeyinden daha yukarı ” olan 6.845 kişinin 379’u (%6) istihdam edilirken, %6’sı işsiz olup, kalan 6.048’i (%88) işgücüne dahil olmamaktadır.

A) En son bitirdiği okul düzeyi, “Bir okul bitirmeyen” düzeyine eşit ya da bu düzeyden daha az olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey* olup;

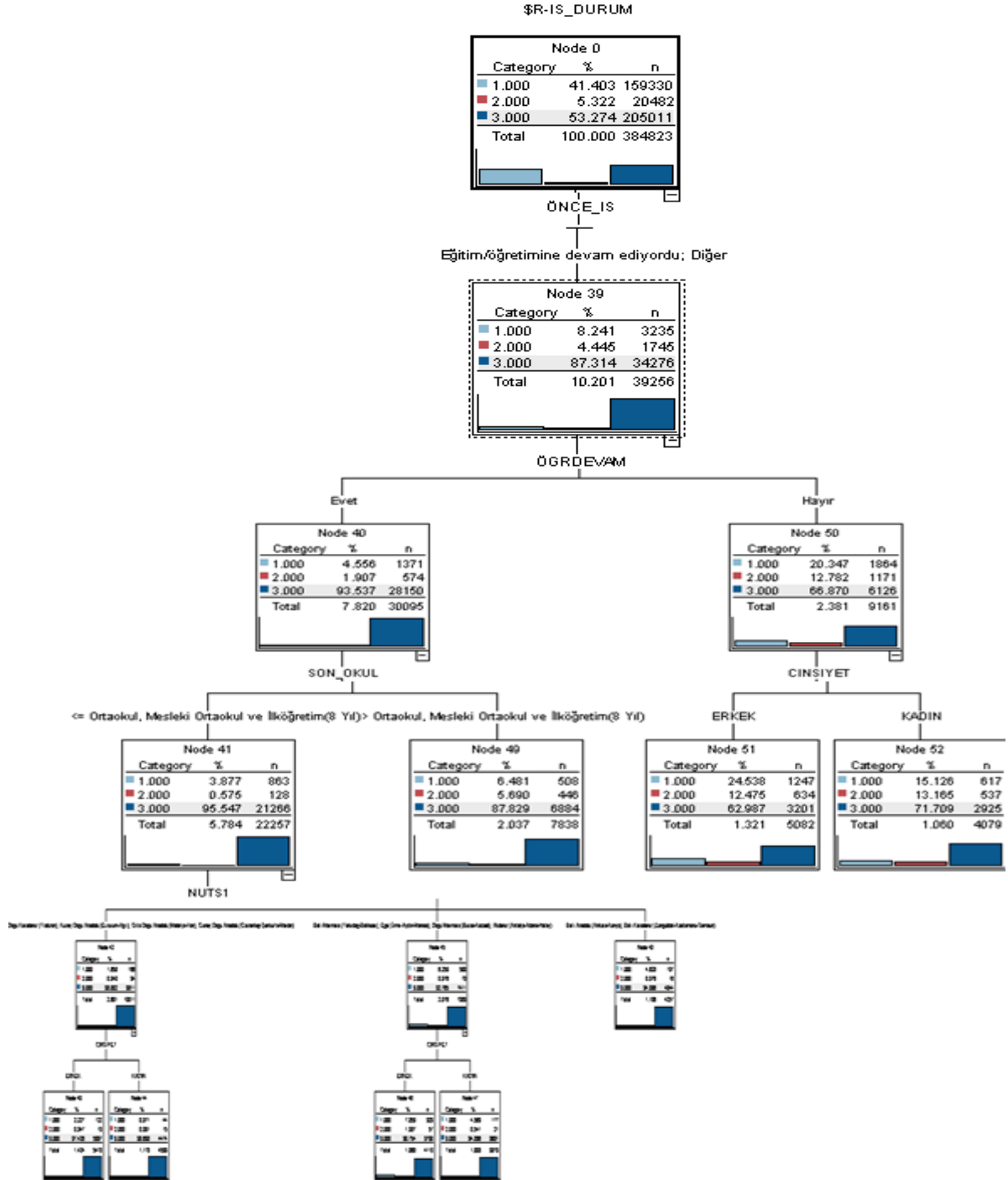
- NUTS1’i İstanbul (İstanbul), Orta Anadolu (Kırıkkale-Kayseri), Doğu Karadeniz (Trabzon), Kuzey Doğu Anadolu (Erzurum-Ağrı), Orta Doğu Anadolu (Malatya-Van), olan 11.913 kişinin %1 istihdam edilirken, %99’u işgücüne dahil olmamaktadır.
- NUTS1’i Batı Marmara (Tekirdağ-Balıkesir), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin), Batı Anadolu (Ankara-Konya), olan 10.565 kişinin %2’si istihdam edilirken % 99’u çoğunluğu işgücüne dahil olmamaktadır.

- Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Orta Anadolu (Kırıkkale-Kayseri), Akdeniz (Antalya-Adana-Hatay), Batı Karadeniz (Zonguldak-Kastamonu-Samsun) olan 11.270 kişinin %3'ü istihdam edilirken %96' sı işgücüne dahil olmamaktadır.

B) En son bitirdiği okul düzeyi, “ortaokul, mesleki ortaokul ve ilköğretim (8 yıl) ve genel lise düzeyine eşit ya da daha alt düzeyde ” olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *medeni durum* olup;

- Medeni durumu; hiç evlenmedi, boşandı olan 4.499 kişi (%1) olup bunun %6'sı istihdam edilirken, %4'ü işsiz olup %10'u ise işgücüne dahil olmamaktadır.
- Medeni durumu; evli, eşi öldü olan çoğunluk 8.352 kişi (%2) olup bunun %4'sı istihdam edilirken, %2'si işsiz olup %94'ü ise işgücüne dahil olmamaktadır.

4.6.6.6. Karar Ağacı Altıncı Anadalı



Şekil 4.26: 2010 Karar Ağacı Altıncı Anadalı

Şekil 4.26'dan görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), eğitim/ öğretime devam ediyordu, Diğer grubunda olan kişi sayısı 39.256 (%10) olup, bu kişilerin 3.235'i (%8) istihdam edilirken, 4.445'i (%4) işsiz grubunda olup kalan çoğunluk olan 34.276 kişi (%87) ise işgücüne dahil olmamaktadır. Önceki işgücü durumu “eğitim/ öğretime devam ediyordu; diğer” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *bir öğrenim kurumuna devam edip etmediği* olup;

- Bir öğrenim kurumuna devam edenlerin toplam sayısı 30.095 olup 1.371'i (%5) istihdam edilirken 574'ü (%2) işsiz olup 28.150'si (%93) işgücüne dahil olmamaktadır.
- Bir öğrenim kurumuna devam etmeyenlerin toplam sayısı 9.161 olup 1.864'ü (%20) istihdam edilirken 1.171'i (%12) işsiz olup 6.126'sı (%67) işgücüne dahil olmamaktadır.

A) Bir öğrenim kurumuna devam edip etmediği “evet” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *son bitirdiği okul düzeyi* olup;

- En son bitirdiği okul düzeyi, “ortaokul, mesleki ortaokul ve ilköğretim (8 yıl)” düzeyine eşit ya da bu düzeyden daha az olan 22.257 kişinin sadece %3'ü istihdam edilirken çoğunluğu oluşturan 21.266 (%96) kişi ise işgücüne dahil olmamaktadır.
- En son bitirdiği okul düzeyi, “ortaokul, mesleki ortaokul ve ilköğretim (8 yıl)” düzeyinden daha yukarı olan 7.838 kişinin 508' i (%6) istihdamda, 406' sı (%6) işsiz ve % 88'i ise işgücüne dahil olmamaktadır.

En son bitirdiği okul düzeyi, “ortaokul, mesleki ortaokul ve ilköğretim (8 yıl)” düzeyine eşit ya da bu düzeyden daha az olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *NUTS1 yani İstatistiki Bölge Birimleri Sınıflaması, 1. Düzey* olup;

- NUTS1'i İstanbul (İstanbul), Orta Anadolu (Kırıkkale-Kayseri), Doğu Karadeniz (Trabzon), Kuzey Doğu Anadolu (Erzurum-Ağrı), Güney Doğu

Anadolu (Gaziantep-Şanlıurfa-Mardin), Orta Doğu Anadolu (Malatya-Van) olan 10.001 kişinin (%3) 9.811'inin (%98) işgücüne dahil olmamaktadır.

- NUTS1'i Batı Marmara (Tekirdağ-Balıkesir), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya), Akdeniz (Antalya-Adana-Hatay), Batı Karadeniz (Zonguldak-Kastamonu-Samsun) olan 13.403 kişinin 499' u (%4) istihdam edilmekte, 12.789 (%96) işgücüne dahil olmamaktadır.
- NUTS1'i Batı Anadolu, Batı Karadeniz (Zonguldak-Kastamonu-Samsun) olan 4.257 kişinin % 4'ü istihdam edilmekte, %95'i işgücüne dahil olmamaktadır.

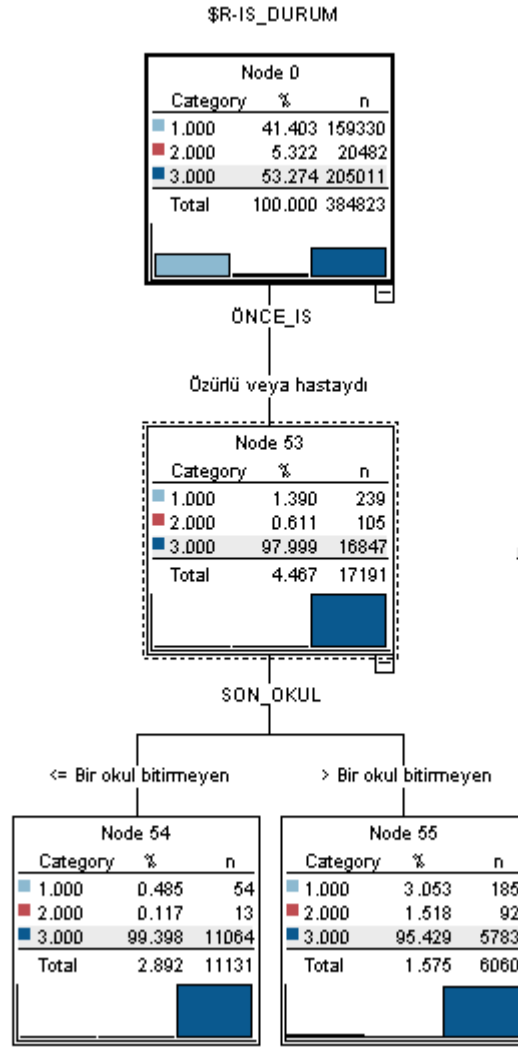
NUTS1'i İstanbul (İstanbul), Orta Anadolu (Kırıkkale-Kayseri), Doğu Karadeniz (Trabzon), Kuzey Doğu Anadolu (Erzurum-Ağrı), Güney Doğu Anadolu (Gaziantep-Şanlıurfa-Mardin), Orta Doğu Anadolu (Malatya-Van), Batı Marmara (Tekirdağ-Balıkesir), Ege (İzmir-Aydın-Manisa), Doğu Marmara (Bursa-Kocaeli), Batı Anadolu (Ankara-Konya), Akdeniz (Antalya-Adana-Hatay), Batı Karadeniz (Zonguldak-Kastamonu-Samsun) olan kişilerde dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *cinsiyet* olup;

- Cinsiyeti erkek olan 6.546 kişinin %2'si istihdamda, %1'i işsiz olup kalan çoğunluk (% 97) ise işgücüne dahil olmamaktadır.
- Cinsiyeti kadın olan 8.412 kişinin %3'ü istihdamda, %1'i işsiz ve kalan çoğunluk %96'sı ise işgücüne dahil olmamaktadır.

B) Bir öğrenim kurumuna devam edip etmediği “hayır” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *cinsiyet* olup;

- Cinsiyeti erkek olan 5.082 kişinin 1.247'si istihdamda olup (%25), 634'ü işsiz olup (%12) 3.201'i ise (%63) işgücüne dahil olmamaktadır.
- Cinsiyeti kadın olan 4.079 kişinin 617'si istihdamda (%15), 537'si işsiz (%13) ve kalan çoğunluk %71'i ise işgücüne dahil olmamaktadır.

4.6.6.7. Karar Ağacı Yedinci Anadalı

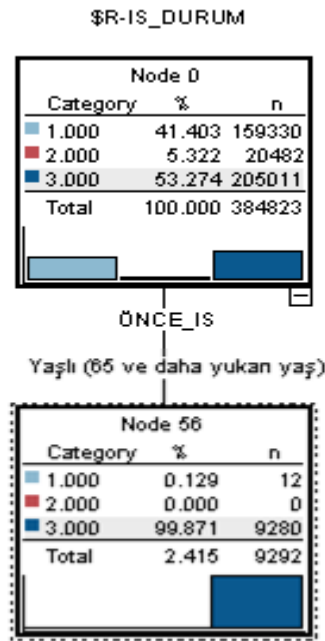


Şekil 4.27: 2010 Karar Ağacı Yedinci Anadalı

Şekil 4.27’den görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), özürlüydü veya hastaydı grubunda olan kişi sayısı 17.191 olup, bu kişilerin sadece 239’u (%1) İstihdam edilirken, 105’i (%1) işsiz grubunda olup kalan çoğunluk olan 16.847 kişi (%98) ise işgücüne dahil olmamaktadır. Önceki işgücü durumu “özürlüydü veya hastaydı” olan kişiler için dallara ayırmada önemli bulunan bir sonraki açıklayıcı değişken *en son bitirdikleri okul düzeyi* olup;

- En son bitirilen okul düzeyi; bir okul bitirmeyen veya daha alt düzeyde olan 11.131 kişinin neredeyse tamamı olan 11.064'ü (%99) işgücüne dahil olmamaktadır.
- En son bitirilen okul düzeyi; bir okul bitirmeyen düzeyinden yukarı olan yani en az ilkokul (5 yıl) düzeyinde olan 6.060 kişinin 185' i (%3) istihdam edilirken, % 2'si işsiz olup 5.783'ü (%95) işgücüne dahil olmamaktadır.

4.6.6.8. Karar Ağacı Sekizinci Anadalı



Şekil 4.28: 2010 Karar Ağacı Sekizinci Anadalı

Şekil 4.28'den görüldüğü gibi; bir yıl önceki işgücü durumu (Önce_İs), yaşlı (65 ve daha yukarı yaş) grubunda olan kişi sayısı 9.292 olup, bu kişilerin sadece (%0,1) istihdam edilirken, kalan çoğunluk olan 9280 kişi (%99) ise neredeyse tamamına yakını işgücüne dahil olmamaktadır.

4.7. Analiz Genel Değerlendirmesi

Analizin sonuçları karar ağacının ana dallarına göre değerlendirilirse;

Karar ağacının birinci anadalına göre yani önceki yıl bir işte çalışmakta olanlarda; işgücü durumundan sonra etkili olan değişken 2009 yılında Referans Kişiyne Yakınlık iken, 2010 yılında Cinsiyettir. 2009 yılına göre referans kişinin kendisi, torunu, akrabaları ve akraba olmayanlarının oluşturduğu grubun istihdam oranı %56 ile referans değişkenindeki en yüksek gruptur. Çocuklarında ise bu değer %52 gibi yüksek bir değer olup, işsizlik oranı %32 olmakla beraber işgücüne katılım oranı en yüksek bu gruptadır. Referans kişinin kendisi, torunu, akrabaları ve akraba olmayanlarında bir sonraki işgücü durumu belirleyicisi bölge düzeyi (NUTS1) olup, 12 bölge içerisinde en yüksek istihdam oranı %60 ile Batı Anadolu, Akdeniz ve Orta Doğu Anadolu' dur. İşsizlik oranı (%33) ise Türkiye' nin en gelişmiş bölümleri olan İstanbul, Ege ve Doğu Marmara' dadır. 2010 yılına göre ise, cinsiyeti erkek olanlarda hem istihdam hem de işsizlik değeri, kadınlara göre daha yüksektir. Yani işgücüne katılım oranı (%86) daha yüksektir. Erkeklerde işgücü durumunu belirleyen değişken bölge düzeyi olup 12 bölge için de hem istihdam değeri hem de işgücüne katılım oranı birbirine yakın değerlerdedir. Kadınların ise bir işte çalışmasında ya da iş aramasında en önemli değişken en son bitirdikleri okul düzeyi olup ilköğretim düzeyi veya daha düşük eğitim durumunda olanların istihdam değeri (%51), ilköğretimden daha yüksek eğitim alan / almakta olan kadınlarınkinden daha yüksek iken ikinci grupta işsizlik daha yüksek orandadır.

Karar ağacının ikinci ana dalına göre; hem 2009 hem 2010 yılında, bir önceki işgücü durumu şu anki işinde çalışanların, bir sonraki yılda tamamının istihdamda olması mevcut işten memnun olunduğunu ve bu alanlarda işgücünün devamlılığının sağlandığını gösterebilir.

Karar ağacının üçüncü ana dalına göre; bir önceki yıl emekli olanların işgücü durumunu, her iki yıl da cinsiyet durumları belirlemektedir. Emekli olan erkeklerin oranı, kadınlara göre daha yüksek olmakla beraber; büyük çoğunluğu (%96) işgücüne dahil olmamaktadır. Buna karşın, 2010 yılında işgücüne katılım oranı, 2009 yılına göre nispi olarak arttığı söylenebilir, çünkü bir sonraki sene istihdam oranı artmaktadır. Ayrıca 2009 yılında, emekli olanların büyük çoğunluğunun son bitirdikleri okul düzeyinin ortaokul, mesleki ortaokul ve ilköğretim düzeyinden

düşük olmaktadır. Cinsiyetten sonra, önceki yıl emekli olanlarda işgücü durumunu bölge düzeyi belirlemektedir. 2009 yılında hem istihdamın hem de işgücüne katılım oranının en yüksek olduğu bölge Batı ve Doğu Marmara, Batı Anadolu, Batı ve Doğu Karadeniz iken, işsizliğin en yüksek olduğu bölge İstanbul, Akdeniz ve Orta Doğu Anadolu' dur. 2010 yılında istihdamın ve işgücünün yine en yüksek olduğu bölge aynı kalırken, işsizliğin en yüksek olduğu bölge de burasıdır.

Karar ağacının dördüncü ana dalına göre; bir önceki yıl iş aramakta ya da asker olanların hem 2009 yılında hem 2010 yılında gerek istihdam gerekse işsizlik değerlerinin en yüksek olduğu bölgeler; İstanbul, Ege, Doğu Marmara ve Batı Anadolu bölümleridir. Bununla beraber, işgücüne katılım oranı 2009 yılında %82 iken, 2010 yılında %83'e yükselmiştir. Tüm bölgeler içinde işgücüne katılım oranının en düşük olduğu bölge ise her iki yıl için de Güneydoğu Anadolu olup, 2010 yılında, bir önceki seneye göre genel olarak Güneydoğu ve Doğu Anadolu bölgelerinde işgücüne dahil olmayan gruptaki kişi sayısı azalmakta ve istihdamda olanlarda artış görülmektedir. Bunun yanı sıra; 2009 yılında işgücüne katılım oranının en yüksek olduğu bölgelerde; evden sorumlu olan referans fert, eşi ve çalışma potansiyeli olan diğer bireyler ile sorumluluğu daha az olan referans fert dışındakilerin (çocuğu, gelini / damadı) istihdam oranları genel olarak yakın olmaktadır. 2010 yılında ise, işgücüne katılım oranlarında anlamlı bir fark olmamasına karşın medeni durumları bakımından istihdam oranlarında farklılık vardır. Hiç evlenmemiş ya da boşanmış, eşi ölmüş olanlar %45 oranı ile istihdamda iken, evli olanlar %51 oranıyla istihdamdadır.

Karar ağacının beşinci ana dalına göre; 2009 ve 2010 yıllarında, önceki yıl ev işleriyle meşgul olanların, işgücü durumunu belirleyen birincil değişken son bitirdikleri okul düzeyidir. 2009 yılında, istihdamın ve işsizlik değerinin dolayısı ile işgücüne katılım oranının en yüksek olduğu düzey ortaokul, mesleki ortaokul ve ilköğretimi bitirenlerdir. 2010 yılında bu oran genel lise düzeyinden daha yukarı bir eğitim bitirenlerde en yüksektir. Ayrıca, her iki yıl da bir okul düzeyini bitirmeyenlerde işgücüne dahil olmama değeri en yüksektir. Her iki yıl için, bir okulu bitirmeyenlerden İstanbul, Orta Doğu Anadolu ve Kuzey Doğu Anadolu bölgelerinde yaşayanların işgücüne katılım oranı en düşük olup 2009 yılında bu bölgelerde genel olarak Kır-Kent ayrımında istatistiksel olarak kentsel yerlerde işgücüne katılımın daha düşük olduğu göze çarpmaktadır. Bir sonraki grupta ise

ilkokula devam eden ya da tamamlayanların işgücüne katılım oranı en düşük İstanbul, Kuzey Doğu Anadolu ve Güney Doğu Anadolu bölgeleri iken her iki yıl da istihdamın daha yüksek olduğu bölge Akdeniz' dir. İlköğretime devam eden ya da tamamlayanlarda ise 2009 yılında aynı durum yaşanırken, 2010 yılında bu gruptakilerde aile yapısına sahip olmayanların işgücüne katılma oranı daha yüksektir.

Karar ağacının altıncı ana dalına göre; 2009 ve 2010 yıllarında, bir önceki yıl eğitim / öğretime devam edenlerde işgücü durumunu öğretime devam etme durumu belirleyici rol oynamaktadır. 2009 yılında eğitime devam edenlerin sadece %19'u çalışmakta iken %76'sı işgücüne dahil olmamaktadır. Oysa ki, etmeyenlerde %42'si istihdam edilirken %51'i işgücüne dahil olmamaktadır. Her iki yıl da öğretime devam edenlerin işgücüne katılım oranı, edenlere göre, daha düşük olmakla beraber 2010 yılında bu oran bir önceki yıla göre nispeten azalmıştır. Ayrıca bir sonraki kırıma bakılırsa, 2010 yılında, eğitime devam etmeyen kadınların işgücüne katılım oranının erkeklerle göre daha düşük olduğu, yani daha az işgücüne dahil oldukları görülmektedir. Hem 2009 hem 2010 yılında, eğitime devam edenlerde son bitirdikleri okul düzeyi ortaokul ve ilköğretim ya da daha düşük düzey olanlarda Batı ve Doğu Marmara, Ege, Batı Anadolu, Akdeniz, Batı Karadeniz ve genel olarak Doğu Anadolu bölgelerinde işgücüne katılım oranı daha yüksek iken özellikle bu bölgelerde erkeklerde istihdam edilme değeri kadınlara göre daha yüksektir.

Karar ağacının yedinci ana dalına göre; önceki yıl işgücü durumu özürlü / hasta olanların işgücüne katılım oranı her iki yıl da düşüktür. Bu grubun işgücü durumunu son bitirdikleri okul düzeyi belirlemektedir. Buna göre; bir okul bitirmeyenlerde işgücüne dahil olmama değeri daha yüksek iken, 2009 yılında özellikle bu grubun kadınlarda daha yüksektir.

Karar ağacının sekizinci ana dalına göre; 65 yaş ve üzerindeki yaşlarda her iki yıl da işgücüne dahil olmayan grup büyük çoğunluktadır. Bununla beraber, 2009 yılında kır-kent ayrımı daha etkin rol oynamaktadır. Kırdaki yaşayan yaşlı grubu, ilerleyen yaşına rağmen çalışmayı sürdürürken, kentte yaşayanlar zamanla işgücünden çıkmaktadır.

5. SONUÇ

İlerleyen teknoloji ve sunduğu olanaklar sayesinde, gerek iş gerek eğitim ve gerekse pek çok alanda ulaşılan veri hacminin inanılmaz boyutlara vardığı açık bir gerçektir. Toplanan çok miktarda verinin işlenmesi; artan rekabette doğru ve hızlı karar verebilme yetkinliğini artırmak açısından günümüzde son derece önemli bir konudur. Artan rekabet ortamında alınan kararların riski de büyümektedir ve yöneticilerin kararlarında sezgilerinden çok bilgisayar ve istatistiksel destek kaçınılmaz olmaktadır. Yıllar ilerledikçe ortaya çıkan veri yığınlarına bir düzen verme, başka bir anlamda potansiyel olarak kullanışlı bilgi haline getirme amacıyla veri madenciliği ortaya çıkmıştır.

Veri madenciliği, eldeki yapısız veriden, anlamlı ve kullanışlı bilgiyi çıkarmaya yönelik çalışmalarının bütünü olmuştur. Veri madenciliği, veriden bilgi üreterek ortalama kararlar yerine özgün kararlar verilmesini destekleyen, satışları, karlılığı, yenilikçiliği ve kaynak kullanımında etkinliği artıran önemli bir yönetim aracıdır. Organizasyonların veri ambarlarındaki çok önemli bilgilere yoğunlaşmasına yardımcı olan, çok büyük veri tabanlarında saklı, akıllı bilgiyi ortaya çıkaran yeni bir teknolojidir. Veri madenciliği, veriyi toplamak ve yönetmekten çok, analiz yapmayı ve tahmini içermektedir

Bu tez çalışmasında, Veri Madenciliği Analizi yöntemi ile örnek bir uygulama gerçekleştirilmiştir. Uygulama kapsamında, TÜİK'ten temin edilen 2009 ve 2010 yılları Hanehalkı İşgücü Anketi verilerine ayrı ayrı Veri Madenciliği Analizi uygulanmış ve sonuçlar özetlenmiştir. Anketin coğrafi kapsamı Türkiye sınırları içinde bulunan hanelerde yaşayan tüm fertlerdir. Ancak, bu çalışma kapsamındaki hedef; işgücü durumunun analiz edilmesi olduğundan dolayı 15 yaş ve üstü olan grup yani aktif nüfus analize dahil edilmiştir. Analiz sonuçları; Türkiye'de çalışma potansiyeli olan grubun mevcut yıl ve bir önceki yıla ilişkin istihdam, işsizlik ve işgücüne dahil olmama durumları için istatistiksel bakımından anlamlı çıkmıştır.

Dolayısı ile, bu çalışma bir sonraki dönemde işgücü durumunun potansiyelin nasıl olacağı ve neler yapılması konusunda karar yetkililerine yardımcı olabilecektir.

İşsizlik sorunu ile ilgili alınabilecek kararlara ışık tutması açısından bu iki yılın analiz sonuçları birbiriyle karşılaştırılmıştır. Elde edilen Veri Madenciliği analizinde en anlamlı ağaç modelini CHAID algoritması verdiği için, analiz CHAID algoritması üzerinden gerçekleştirilmiştir ve sonuçları aşağıda özetlenmiştir.

Analiz sonucuna göre hem 2009 hem 2010 yılında kişilerin işgücü durumunda en etkili değişken bir yıl önceki iş gücü durumudur. Analiz sonuçları göstermektedir ki; bir yıl önceki iş durumu, aktif nüfusun ilerideki işgücünün en etkin belirleyicidir. Bir önceki yıla göre takip eden yılda, işgücüne katılım oranı yükselmiş, işsizlik değeri ise azalmıştır. İstihdam oranı arttığı ölçüde işsizlik oranının azalması, iş arayanların işe girmiş olduklarının yani istihdama dahil olduklarını desteklemektedir.

2009 yılında, bir yıl önceki iş gücü durumu bir işte çalışmakta olanların %52'si istihdamda iken, şu anki işinde çalışmakta olanların %100'ü yani tamamı, emekli olanların %2'si, iş arayan ya da asker olanların %38'i, ev işleriyle meşgul olanların %3'ü, öğrenimine devam etmekte olanların %7'si, özürlü veya hasta olanların %1'i ve yaşlı (65 ve daha yukarı yaş) grubunda olanların sadece % 1'i istihdamdadır. İşsiz olma oranlarına bakıldığında bir işte çalışmakta olanlarda %26, emekli olanların %1'i, iş arayan ya da asker olanların %44'ü, ev işleriyle meşgul olanların %2'si, öğrenimine devam etmekte olanların %5'i, özürlü veya hasta olanların %1'i işsizdir. İşgücüne dahil olmama durumuna bakıldığında ise bir işte çalışmakta olanlarda bu oran %22, emekli olanlarda %97, iş arayan ya da asker olanların %18, ev işleriyle meşgul olanlarda %96, öğrenimine devam etmekte olanlarda % 88, özürlü veya hasta olanlarda %98 ve Yaşlı (65 ve daha yukarı yaş) grubunda olanlarda ise neredeyse tamamına yakındır. Bu durumda 2009 yılında işgücüne katılım oranı en yüksek olan grup “şu anki işinde çalışmakta” olup (%100) bunu %81 oranı ile iş arayan ya da asker olanlar ve %78 ile “bir işte çalışanlar” izlemektedir. Öte yandan işgücüne katılmama oranı yaşlı, emekli ve özürlü ya da hasta olanlarda aynı değerlerde olup (%98) en yüksektir.

2010 yılında, bir yıl önceki iş gücü durumu bir işte çalışmakta olanların %59'u istihdamda iken, şu anki işinde çalışmakta olanların %100'ü yani tamamı, emekli olanların % 2' si, iş arayan ya da asker olanların %42'si, ev işleriyle meşgul olanların

%3'ü, öğrenimine devam etmekte olanların % 8'i, özürlü veya hasta olanların % 1'i ve Yaşlı (65 ve daha yukarı yaş) grubunda olanların sadece % 1'i istihdamdadır. İşsiz olma oranlarına bakıldığında bir işte çalışmakta olanlarda % 20, emekli olanların %1'i, iş arayan ya da asker olanların % 40' ı, ev işleriyle meşgul olanların % 3' ü, öğrenimine devam etmekte olanların % 8' i, özürlü veya hasta olanların % 1'i işsizdir. İşgücüne dahil olmama durumuna bakıldığında ise bir işte çalışmakta olanlarda bu oran % 21, emekli olanlarda % 97, iş arayan ya da asker olanlarda % 17, ev işleriyle meşgul olanlarda % 95, öğrenimine devam etmekte olanlarda % 87, özürlü veya hasta olanlarda % 98 ve yaşlı (65 ve daha yukarı yaş) grubunda olanlarda ise yine neredeyse tamamına yakındır. Dolayısı ile 2010 yılında işgücüne katılım oranı en yüksek olan grup “şu anki işinde çalışmakta” olup (%100) bunu % 82 oranı ile iş arayan ya da asker olanlar ve % 79 ile “bir işte çalışanlar” izlemektedir. Öte yandan işgücüne katılmama oranı yaşlı, emekli ve özürlü ya da hasta olanlarda aynı değerlerde olup (% 99) en yüksektir.

Bir yıl önceki işgücü durumu bir işte çalışmakta olanların oranı 2010 yılında yükselirken, her iki yıl da önceki yıl çalışanların işgücü durumunu cinsiyet değişkeni belirlemektedir. Cinsiyeti erkek olanlarda hem istihdam hem de işsizlik değeri, kadınlara göre daha yüksektir. Yani işgücüne katılım oranı daha yüksektir. Genel olarak, 2009 yılında erkeklerin istihdamda olmaları oranı % 60 iken kadınların istihdamda olma oranları sadece % 22'dir. İşgücüne dahil olmama oranlarına bakıldığında kadınlarda bu oran % 75 iken erkeklerde % 31'dir. 2010 yılında, genel olarak oranlar her iki grup için aynı kalırken, 2010 yılında kadınların işgücüne daha çok dahil olma eğiliminde oldukları görülmüş ve bu oran % 23'e yükselmiştir. Erkeklerde işgücü durumunu belirleyen bir sonraki değişken bölge düzeyi olup 12 bölge için hem istihdam değeri hem de işgücüne katılım oranı birbirine yakın değerlerdedir. Kadınların ise bir işte çalışmasında ya da iş aramasında en önemli değişken en son bitirdikleri okul düzeyi olup ilkokul düzeyi veya daha düşük eğitim durumunda olanların istihdam değeri (%51), ilkokuldan daha yüksek eğitim alan / almakta olan kadınlarınkinden daha yüksek iken ikinci grupta işsizlik daha yüksek orandadır. Kadınların iş aramama nedenlerine bakıldığında ailevi ve kişisel nedenlerin çoğunlukta olduğu belirtilebilir.

Ayrıca her iki sene de, analiz yapılan yılda bir yıl önceki işgücü durumuna göre mevcut işine devam edenlerin işgücüne katılım oranı en yüksek çıkmıştır. Bu durum

kişilerin çalıştıkları işten memnun olma potansiyelinin arttığını gösterebildiği gibi makro düzeyde bakıldığında ilgili yıldaki krizde iş piyasasının sürekliliği gerektirdiği bir durum olarak da belirtilebilir.

Bir önceki yıl emekli olanların işgücü durumunu, her iki yıl da cinsiyet durumları belirlemektedir. Emekli olan erkeklerin oranı, kadınlara göre daha yüksek olmakla beraber; büyük çoğunluğu (%96) işgücüne dahil olmamaktadır. Buna karşın, 2010 yılında işgücüne katılım oranının, 2009 yılına göre nispen arttığı söylenebilir, çünkü bir sonraki sene istihdam oranı artmaktadır. Cinsiyetten sonra, önceki yıl emekli olanlarda işgücü durumunu bölge düzeyi belirlemektedir. 2009 yılında hem istihdamın hem de işgücüne katılım oranının en yüksek olduğu bölge Batı ve Doğu Marmara, Batı Anadolu, Batı ve Doğu Karadeniz iken, işsizliğin en yüksek olduğu bölge İstanbul, Akdeniz ve Orta Doğu Anadolu' dur. 2010 yılında istihdamın ve işsizliğin en yüksek olduğu bölge değişmemiştir.

Bir önceki yıl iş aramakta ya da asker olanların hem 2009 yılında hem 2010 yılında gerek istihdam gerekse işsizlik değerlerinin en yüksek olduğu bölgeler; İstanbul, Ege, Doğu Marmara ve Batı Anadolu bölümleridir. Bununla beraber, işgücüne katılım oranı 2009 yılında %82 iken, 2010 yılında %83'e yükselmiştir. Tüm bölgeler içinde işgücüne katılım oranının en düşük olduğu bölge ise her iki yıl için de Güneydoğu Anadolu olmaktadır. Bunun yanı sıra; 2009 yılında işgücüne katılım oranının en yüksek olduğu bölgelerde; evden sorumlu olan referans fert, eşi ve çalışma potansiyeli olan diğer bireyler ile sorumluluğu daha az olan referans fert dışındakilerin istihdam oranları genel olarak yakın olmaktadır. Buna karşın; her iki grupta da işgücüne dahil olan kesimin daha az kişi olup işsizlik değerinin artması, mevcut durumda iş aramakta veya askerde oldukları için çalışmadıklarını gösterebilir. 2010 yılında ise, işgücüne katılım oranlarında anlamlı bir fark olmamasına karşın medeni durumları bakımından istihdam edilme oranları farklıdır. Hiç evlenmemiş ya da boşanmış, eşi ölmüş olanlar % 45 oran ile istihdamda iken, evli olanlar % 51 oranıyla istihdamdadır. Evli olup bir önceki yıl iş arayanların, mevcut yılda bir işe yerleştiklerini veya askerlik sonrasında işe devam ettiklerini göstermektedir.

Önceki yıl ev işleriyle meşgul olanların, işgücü durumunu son bitirdikleri okul düzeyi belirlemektedir. 2009 yılında, istihdamın ve işsizlik değerinin dolayısı ile işgücüne katılım oranının en yüksek olduğu düzey ortaokul, mesleki ortaokul ve

ilköğretimi bitirenlerdir. 2010 yılında bu oran genel lise düzeyinden daha yukarı bir eğitim bitirenlerde en yüksektir. Ayrıca, her iki yıl da bir okul düzeyini bitirmeyenlerde işgücüne dahil olma değeri en yüksektir. Bu durumda, eğitim düzeyinin yükselmesinin işgücüne dahil olmakta etkili olduğu söylenebilir. Her iki yıl için, bir okulu bitirmeyenlerden İstanbul, Orta Doğu Anadolu ve Kuzey Doğu Anadolu bölgelerinde yaşayanların işgücüne katılım oranı en düşük olup 2009 yılında bu bölgelerde genel olarak kır-kent ayrımında istatistiksel olarak kentsel yerlerde işgücüne katılımın daha düşük olduğu göze çarpmaktadır. Bunun nedeninin kentsel alanlarda nitelikli işgücüne ihtiyacın daha fazla olması ve eğitim düzeyinin düşük olmasının bu konuda dezavantaj oluşturduğu söylenebilir. Bir sonraki grupta ise ilkokula devam eden ya da tamamlayanların işgücüne katılım oranı en düşük İstanbul, Kuzey Doğu Anadolu ve Güney Doğu Anadolu bölgeleri iken her iki yıl da istihdamın daha yüksek olduğu bölge Akdeniz' dir. İlköğretime devam eden ya da tamamlayanlarda ise 2009 yılında aynı durum yaşanırken, 2010 yılında bu gruptakilerde aile yapısına sahip olmayanların işgücüne katılma oranı daha yüksektir. Bu durumun da daha çok iş aramaya yönelttiği düşünülebilir.

2009 ve 2010 yıllarında, bir önceki yıl eğitim / öğretimine devam edenlerde öğretimine devam etme durumu, işgücü durumunu belirleyici olmaktadır. Her iki yıl da öğretimine devam edenlerin işgücüne katılım oranı, devam etmeyenlere göre, daha düşüktür. Bununla beraber, 2010 yılında bu oran bir önceki yıla göre nispeten azalmıştır. Öte yandan, 2010 yılında, eğitime devam etmeyen kadınların işgücüne katılım oranının erkeklere göre daha düşük olduğu, yani işgücüne daha az dahil oldukları görülmektedir. Bu da, erkeklerin eğitime devam ediyor olmalarına rağmen işgücüne katılımın hızla artması, hayata atılma ihtiyaçlarının daha yüksek olduğunu göstermektedir.

Önceki yıl işgücü durumu özürlü / hasta olanların işgücüne katılım oranı her iki yıl da düşüktür. Bu grubun, ilerisi için işgücü durumunu son bitirdikleri okul düzeyi belirlemektedir. Buna göre; bir okul bitirmeyenlerde işgücüne dahil olmama değeri daha yüksek iken, 2009 yılında bu değer kadınlarda daha yüksektir. Bununla beraber, 2010 yılında okul bitirenlerin istihdama katılma değerinin 2009 yılına göre yükselmesi, bu grupta eğitimin işgücüne katılmada olumlu bir etki yarattığını göstermektedir.

65 yař ve üzerindeki yařlarda her iki yıl da iřgücüne dahil olmayan grup büyük çoğunluktaadır. Bununla beraber, 2009 yılında kıır-kent ayrımı daha etkin rol oynamaktadır. Kıırda yařayan yařlı grubu, ilerleyen yařına raėmen alıřmayı sürdürürken, kentte yařayanlar zamanla iřgücünden çıkmaktadır. Bunda özellikle kırsal alandaki sektörlerde, özellikle tarım ve hayvancılık, kendi iřlerinde alıřma imkanına sahip olma etkili olmaktadır.

Bu alıřmada yapılan veri madenciliėi analizi kapsamında, 2009 ve 2010 Yılları Hanehalkı İřgücü Anketi verileri göstermektedir ki; istihdamda olma oranı ülkemizde düşük olup 2009 yılında ancak % 39,95 iken 2010 yılında % 41,4'tür. İřgücüne katılma oranının düşük olmasının temelinde, tarım istihdamının azalması sonucu köyden kente gö olgusu yatmaktadır. Çünkü kentsel alanlarda Eėitim düzeyi arttıka istihdamda olma oranı artmakta, iřsiz olma oranı azalmaktadır. Aynı řekilde iřgücüne katılım oranları da eėitim düzeyi arttıka artmaktadır.

İřgücü piyasası aısından dezavantajlı gruplar kadınlar, özürölüler veya hastalardır.

Kadınların iřgücüne katılma oranı, ölke bazında erkeklerin ancak üçte biri kadardır. İřgücüne dahil olmayanların çoğunluėunu ev kadınları oluşturmaktadır. Kıırdan kente gö sürecinde kıırda iřgücü içinde yer alan kadınlar kentte iřgücü dışında görünmektedir. Diėer taraftan, kentlerde iřlerin belirli beceriler gerektirmesi buna karřın kadınların eėitim düzeyinin düşükölüėü iřgücü dışında kalmayı kalıcı hale getirmektedir. Bu durum, özellikle kentsel yerlerde yařayan kadınlar olmak üzere iřgücüne katılımın bir an önce artırılması gereėini ortaya koymaktadır. Eėitim seviyesi düşük kadınlar genelde katma deėeri yüksek olmayan iřlerde düşük ücretle alıřmaktadırlar. Ücretlerin düşük olması ise kadınların birçoėunun alıřmak yerine ev kadını olarak kalmayı tercih etmelerine sebep olmaktadır.

Kırsal kesimde, yařayan kadının ise eėitim imkanlarına ulaşması ve yararlanması sadece buralardaki eėitim imkanlarının kente göre daha az olmasından deėil, aynı zamanda bazı kültürel, sosyal ve geleneksel sebeplerle daha da zorlařmaktadır. Kadınların iřgücüne katılım oranlarını arttırmak için ise öncelikle eėitim düzeylerinin artırılması ve iř hayatında başarılı olabilecekleri konusunda yüreklendirilmesi gerekmektedir.

Erkeklerde ise onların eėitime devam etmelerine raėmen alıřmaya ihtiya duymalarına neden olan ekonomik sıkıntıların azaltılması ve daha çok iř imkanı

sunulması için ülkenin genel ekonomi düzeyinin iyileştirilmesi, yatırımların artırılarak iş olanakların artırılması gerekmektedir.

Özürülerin veya hastaların işgücüne katılma oranı oldukça düşüktür. Bunun nedeni ülkemizdeki işyerlerinde uygun ortamın yaratılamaması ve özürü işgücüne olan talebin oldukça sınırlı olmasıdır. İş yerlerinin ölçeğine bağımlı kalmadan sayıca ve çalışabilecekleri alanların daha detaylı olarak düzenlenmesi hem işe hem topluma kazandırılması açısından gereklidir.

Son olarak, işsizliği azaltmada en etkin çözüm istihdam dostu ve sürekli eğitimi destekleyen bir işgücü piyasasını oluşturmaktır. Bunun için değışen işgücü piyasasının gereklerine kısa sürede uyum sağlayabilecek niteliklere sahip işgücünün, yaşam boyu eğitimini sağlayan kurumsal düzenlemeler yapılması yerinde olacaktır. Ayrıca eğitime devam edenler ve erken yaşta işte çalışmak durumunda olanlar için, işsizliği önlemek amacı ile yarım gün işin yasallaşması ve yaygınlaştırılması yararlı olacaktır. Bu alanda, iş amaçlı kursların ve ardından da sertifikaların verilmesi faydalı olabilir. Kırsal kesimde çalışanların oranının artırılması, yani tarımın desteklenmesi özellikle kırsal kesimde işgücüne katılım oranını arttıran adımlardan biri olacaktır. Ülkemizde kalıcı ve uzun süreli işsizlik sorunu ciddiyetini korumaktadır. Dolayısı ile işsizlikle toplumca bir mücadele bir an önce başlamak zorundadır. Kaybedilen her yıl işsizliğin tetikleyeceği toplumsal sorunları daha da ağırlaştıracak ve ülkemizi ileri taşımayacaktır.

KAYNAKÇA

- “ 13.5.1.3. Memory Based Reasoning”.
<http://fedc.wiwi.huberlin.de/xplore/ebooks/html/csa/node206.html#SECTION03015130000000000000> [25.12.2011].
- Acı, Mehmet, Mutlu Avcı. “K Nearest Neighbor Reinforced Expectation Maximization Method”. **Expert Systems with Applications**. c. 38 (2011): 12586.
- Albayrak, Ali Sait, Şebnem Koltan Yılmaz. “Veri Madenciliği: Karar Ağacı Algoritmaları Ve İMKB Verileri Üzerine Bir Uygulama”. **Süleyman Demirel Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Dergisi**. c.14. s.1 (2009):36.
- Alpaydın, Ethem, “Intelligent Data Mining”.
www.cmpe.boun.edu.tr/~ethem/files/papers/veri-maden_2k.ppt [27.12.2011].
- Amasyalı, M. Fatih, Tülay Yıldırım. “Otomatik Haber Metinleri Sınıflandırma”.
www.ce.yildiz.edu.tr/en/mygetfile.php?id=268 [25.12.2011].
- Arabacı, Gültekin. “ Veri Madenciliğinde Apriori, Tahminci Apriori Ve Tertius Algoritmalarının Weka Ve Yale Programları İle Karşılaştırılması Ve Bir Uygulama”. Yüksek Lisans Tezi. İstanbul Ticaret Üniversitesi Sosyal Bilimler Enstitüsü, 2007.
- Argüden, Yılmaz, Burak Erşahin. **VERİ MADENCİLİĞİ Veriden Bilgiye, Masraftan Değere**. 1. bs. İstanbul: ARGE Danışmanlık Yayınları, 2008).
- Aviad, Barak, Gelbard Roy. “Classification By Clustering Decision Tree-like Classifier Based on Adjusted Clusters”. **Expert Systems with Applications**. c. 38 (2011): 8221.
- Awad, E.M., H.M. Ghaziri. **Knowledge Management**. Prentice Hall, 2004 (Aktaran: Çankırı, Süreyya ve Diğ.. “Organizasyonlarda Bilgi Yönetimi Sürecinde Veri Madenciliği Yaklaşımı”. **Bilgi Çağında Varoluş: “Fırsatlar ve Tehditler” Sempozyumu, 01-02 Ekim 2009**. İstanbul: Yeditepe Üniversitesi, 2011: 147-178).
- Aydın, Sinan. “Veri Madenciliği Ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama. Doktora Tezi. Eskişehir Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eylül 2007.
- Baykasoğlu, Adil. “Veri Madenciliği Ve Çimento Sektöründe Bir Uygulama”. **Gaziantep Üniversitesi Dergisi**: 2.
- Baysal, Ahmet Cüneyt. “Bayi Değerlendirmesi İçin Veri Madenciliği Uygulaması”. Yüksek Lisans Tezi. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, 2008.

- Beşer, Mustafa Kemal. “Zaman Serilerinde Bootstrap Çözümlemeleri ve Türkiye’de Tanzi Etkisine Uygulaması”. Doktora Tezi. Marmara Üniversitesi Sosyal Bilimler Enstitüsü, 2006.
- Castellani, Marco, Hefin Rowlands. “Evolutionary Artificial Neural Network Design and Training for Wood Veneer Classification”. **Engineering Applications Of Artificial Intelligence**. c. 22 (2009): 735.
- Chakraborty, Sukumar. “Rid3: An Id3-Like Algorithm For Real Data”. **Information Sciencies**. c. 96 (2000): 272.
- Chehreghani, Morteza, Haghir Hassan Abolhassani, Mostafa Haghir Chehreghani. “Improving Density-Based Methods For Hierarchical Clustering Of Web Pages”. **Data & Knowledge Engineering**. c. 37 (2008): 33.
- Chen, T.Y., P.L. Poon. “Construction Of Classification Trees Via The Classification-Hierarchy Table”. **Information and Software Technology**. c. 39 (1997): 889.
- Cheng, Jao-Hong, Huei-Ping Chen, Yi-Min Lin. “A Hybrid Forecast Marketing Timing Model Based on Probabilistic Neural Network, Rough Set and C4.5”. **Expert Systems With Applications**. c. 37 (2010): 1815.
- Chiu, Susan, Domingo Tavella. **Data Mining and Market Intelligence for Optimal Marketing Returns**. 1.edition. USA: Elsevier, 2008.
- Chudzinska, Maria, Danuta Baralkiewicz. “Application Of ICP-MS Method Of Determination Of 15 Elements in Honey With Chemometric Approach For The Verification Of Their Authenticity”. **Food And Chemical Toxicology**. c.49 (2011): 2745.
- Coşkun, Cengiz. “Veri Madenciliği Algoritmalarının Karşılaştırılması”. Yüksek Lisans Tezi. Dicle Üniversitesi Fen Bilimleri Enstitüsü, 2010.
- Çakmak, Zeki, Nevin Uzgören, Gülnur Keçek. “Kümeleme Analizi Teknikleri İle İllerin Kültürel Yapılarına Göre Sınıflandırılması Ve Değişimlerinin İncelenmesi”. <http://sbe.dpu.edu.tr/12/15-36.pdf> [30.12.2012].
- Çamdeviren, Handan Ankaralı ve Diğ. “Comparison Of Logistic Regression Model And Classification Tree: An Application To Postpartum Depression Data”. **Expert Systems with Applications**. c. 32 (2007): 987.
- Çelik, Mine. “Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama”. Yüksek Lisans Tezi. İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı, 2009.
- Data Mining: Statistics And More, The American Statistician**. 2. bs. 1998 (Aktaran: Ay, Derya, İbrahim Çil. “Migros Türk A.Ş.’ de Birlikte Kurallarının Yerleşim Düzeni Planlamada Kullanılması”. **Endüstri Mühendisliği Dergisi**. c. 21. s. 2 (2010): 15).

- Demiralay, Meral. “Hiyerarşik Kümeleme Metotları ile Veri Madenciliği Uygulamaları”. Yüksek Lisans Tezi. Maltepe Üniversitesi Mühendislik-Mimarlık Fakültesi, 2005.
- Demirel, Betül. “Veri Madenciliğinde Chaid Algoritmasının Sosyal Güvenlik Kurumu Veri Tabanına Uygulanması”. Yüksek Lisans Tezi. Gazi Üniversitesi Fen Bilimleri Enstitüsü, 2010.
- Dharwa, Jyotindra N., Ashok R. Patel. “A Data Mining with Hybrid Approach Based Transaction Risk Score Generation Model (TRSGM) for Fraud Detection of Online Financial Transaction”. **International Journal of Computer Applications**. c. 16. s. 1 (2011): 22.
- Doğan, İsmet. “Holştayn Irkı İneklerde Süt Verimine Etki Eden Faktörlerin CHAID Analizi İle İncelenmesi”. **Afyon Ankara Üniversitesi Veterinerlik Fakülte Dergisi**. c. 50 (2003): 66.
- Doğan, İsmet. “Kümeleme Analizi ile Seleksiyon”. **Türk Vet Anim Sciences**. c. 26 (2002) : 48.
- Düzgünoğlu, Selda. “Hacettepe Üniversitesi Veri Ambarı Ve Olap Teknolojilerinden Yararlanılarak Karar Destek Amaçlı Raporlama Aracı Gerçekleştirimi”. Yüksek Lisans Tezi. Bilgisayar Mühendisliği Anabilim Dalı, 2006.
- Erdoğan, Şenol Zafer. “Sınıflandırma Teknikleri ve Algoritmaları”. http://Akademik.Maltepe.Edu.Tr/~Senole/Data_Mining/Bolum3.Pdf [22.12.2012].
- Fawcett, Tom. “An Introduction to ROC Analysis”. **Pattern Recognition Letter**. c. 27 (2006): 861.
- Galimberti, Giuliano, Gabriele Soffritti. “Model-Based Methods to Identify Multiple Cluster Structures in A Data Set”. **Computational Statistics and Data Analysis**. c. 52 (2007): 520.
- Gevrekçi, Yakut. “Hayvancılık Verilerinde Kümeleme Analizi Uygulamalarına Bir Örnek”. http://4uzbk.sdu.Edu.Tr/4UZBK/POSTER/BGP/4UZBKP_066.pdf [17.01.2012].
- Güleç, Fatih Mehmet. “Kurumsal Verilerin Yapay Zeka Modelleri İle İşlenmesi İçin Modelleme Aracı Alt Yapı Tasarım Ve Gerçekleştirimi”. Yüksek Lisans Tezi. Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, 2007.
- “Hierarchical Nature of Classification Trees”. <http://www.statsoft.com/textbook/classification-trees/#character1> [22.12.2012].
- Hoare, Ray. “Using CHAID For Classification Problems”. **New Zealand Statistical Association 2004 Conference**, 2004: 1-5.
- Introduction to Data Mining and Knowledge Discovery**. Two Crows Corp., 2005. www.trocrows.com [14.12.2011].

- Karabulut, Murat. Mehmet Gürbüz, Ersin Kaya Sandal. “Hiyerarşik Kluster (Küme)Tekniği Kullanılarak Türkiye’de İllerin Sosyo-Ekonomik Benzerliklerinin Analizi”. **Ankara Üniversitesi Dergisi**. c. 2, s. 2 (2004): 4.
- Kaya, Halil, Kemal Köymen. “Veri Madenciliği Kavramı Ve Uygulama Alanları”. **Doğu Anadolu Bölgesi Araştırmaları** (2008): 162.
- Kayri, Murat, Murat Boysan. “Bilişsel Yatkınlık İle Depresyon Düzeyleri İlişkisinin
- Khan, Aurangzeb, Baharum Baharudin, Khairullah Khan. “Mining Customer Data For Decision Making Using New Hybrid Classification Algorithm”. **Journal Of Theoretical And Applied Information Technology**. c. 27. s. 1 (2011): 54.
- Kıyak, Erkan. “Crisp-Dm Yöntemi Bilimi Kullanılarak Deniz Kuvvetleri Verisi Üzerinde Veri Madenciliği Sınıflandırma Tekniklerinin Karşılaştırılması”. Yüksek Lisans Tezi. Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, 2006.
- Kleppin, Lukas, Roland Pesch, Winfried Schro. “CHAID Models On Boundary Conditions Of Metal Accumulation in Mosses Collected in Germany in 1990, 1995 and 2000”. **Atmospheric Environment Elsevier**. c. 42 (2008): 5222.
- Kolay, Gökçe. “İşletmelerde Bilgi Sistemleri Verimliliğini Arttırmada Veri Madenciliği Yöntemi: Bir Simülasyon Çalışması”. Yüksek Lisans Tezi. Zonguldak Karaelmas Üniversitesi Sosyal Bilimler Enstitüsü, 2006.
- Koldere Akın, Yasemin. “Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi”. Doktora Tezi. Marmara Üniversitesi Sosyal Bilimler Enstitüsü, 2008.
- Koltan Yılmaz, Şebnem Said Patır. “ Kümeleme Analizi Ve Pazarlamada Kullanımı”. **Akademik Yaklaşımlar Dergisi**. c. 2. s. 1 (2011): 101.
- Konda, Ramesh. “Predicting Machining Rate in Non-Traditional Machining using Decision Tree Inductive Learning”. Doctorate of Philosophy. Nova Southeastern University Computer Information Systems Graduate School of Computer and Information Sciences, April 2010.
- Kopper, Karl. **The Linux Enterprise Cluster: Build A Highly Available Cluster with Commodity Hardware and Free Software**. San Francisco: No Starch Press, 2005.
- Koyuncugil, Ali Serhan, Nermin Özgülbaş. “Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları”. **Bilişim Teknolojileri Dergisi**. c. 2. s. 2 (Mayıs 2009): 25.
- “Korelasyon Analizi”.
http://maden.karaelmas.edu.tr/hocalar/hamitaydin/Sunu_6ci_hafta.pdf
[27.12.2011].
- Kök, Baha Vural, Necati Kuloğlu, “Sollama Esnasında Taşıt Ve Yol İle İlgili Faktörlerin Karar Ağacı Yöntemiyle İrdelenmesi”. **Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi**. c. 21. s. 2 (2005): 183.

- Lia, Chia-Wen, Yeng-Horng Perng, Tsung-Lung Chiang. "Discovery Of Unapparent Association Rules Based On Extracted Probability". **Decision Support Systems**. c. 47 (2009): 354.
- M.Schubert, Christine, Steven N. Thorsen, Mark E. Oxley. "The ROC Manifold For Classification Systems". **Pattern Recognition**. c. 44 (2011): 350.
- Maimon, Oded, Lior Rokach. **Data Mining and Knowledge Discovery Handbook**, Israel: Springer, 2005.
- Murtagha, Fionn, Adrian E. Raftery, Jean-Luc Starck. "Bayesian Inference For Multiband Image Segmentation via Mmodel-Based Cluster Trees". **Image and Vision Computing**. c. 23 (2005): 588.
- Ngai, E.W.T. ve Diğ.. "The Application Of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and An Academic Review Of Literature". **Decision Support Systems**. c. 50 (2011): 562.
- Ni, Huihui, An Chen, Ning Chen. "Some Extensions On Risk Matrix Approach". **Safety Science**. c. 48 (2010): 1270.
- Özdemir Özdoğan, Gülistan, Osman Abul, Ali Yazıcı. "Paralel Veri Madenciliği Algoritmaları". http://hpc.etu.edu.tr/docs/Basarim09_gooetal.pdf [03.01.2012].
- Özdemir, Suat. "Veri Madenciliği Ders Notları". <http://ceng.gazi.edu.tr/~ozdemir/> [16.01.2012].
- Özekes, Serhat. "Veri Madenciliği Modelleri Ve Uygulama Alanları". **İstanbul Ticaret Üniversitesi Dergisi**. c. 3. s.3 (2003): 67.
- Ramaswami, M., R. Bhaskaran. "A CHAID Based Performance Prediction Model in Educational Data Mining". **International Journal of Computer Science Issues**. c. 7. s. 1 (2010):14.
- "Risk Değerlendirme Karar Matrisi". http://www.isguv.com/risk_degerlendirme.htm [10.01.2012].
- Satıcı, Ömer, Zeki Akkuş, Aykut Alp. "Tıp Fakültesi Öğretim Elemanlarının Teknolojiye İlişkin Tutumlarının CHAID Analizi İle İncelenmesi". **Dicle Tıp Dergisi**. c. 36. s. 4 (2009): 269.
- Setsirichok, Damrongrit ve Diğ.. "Classification Of Complete Blood Count And Haemoglobin Typing Data by A C4.5 Decision Tree, A Naive Bayes Classifier And A Multilayer Perceptron For Thalassaemia Screening". **Biomedical Signal Processing And Control**. s. 250 (2011): 5.
- Shopova, Elisaveta G., Natasha G. Vaklieva-Bancheva. "BASIC—A Genetic Algorithm For Engineering Problems Solution". **Computers And Chemical Engineering**. c. 30 (2006): 1295.

- Sınıflandırma Ve Regresyon Ağacı Analizi İle İncelenmesi”. **Hacettepe Üniversitesi Eğitim Fakültesi Dergisi**. c. 34 (2008): 168.
- Sipahi, Beril, E. Serra Yurtkoru, Murat Çinko. **Sosyal Bilimlerde SPSS’le Veri Analizi**. 3. bs. İstanbul: Beta Yayınları, 2010.
- Soman, K. P., S. Diwakar. **“Insight into Data Mining Theory and Practice”**. New Delhi: PHI Learning Pvt. Ltd., 2006.
- Suh, Jong Hwan, Chung Hoon Park, Si Hyun Jeon. “Applying Text and Data Mining Techniques to Forecasting The Trend Of Petitions Filed to E-People”. **Expert Systems with Applications**. c. 37 (2010): 7259.
- Sütçü, Cem. “İstatiksel Veri Sistemleri ve Basın Sektöründe bir Karar Destek Sistemi Uygulaması”. Doktora Tezi. Marmara Üniversitesi Sosyal Bilimler Enstitüsü, 1995.
- Şengün, Meltem. “Ucal, Ekonometrik Model Seçim Kriterleri Üzerine Kısa Bir İnceleme”. **C.Ü. İktisadi Ve İdari Bilimler Dergisi**. c. 7. s. 2 (2006): 50.
- Taşkın, Çağatan, Gül Gökay Emel. “Veri Madenciliğinde Kümeleme Yaklaşımları ve Kohonen Ağları ile Perakendecilik Sektöründe Bir Uygulama”. **Süleyman Demirel Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Dergisi**. c. 15, s. 3 (2010): 399.
- Tekerek, Adem. “Veri Madenciliği Süreçleri Ve Açık Kaynak Kodlu Veri Madenciliği Araçları”. <http://ab.org.tr/ab11/bildiri/29.pdf> [16.12.2011].
- Terzi, Özlem ve Diğ. “Veri Madenciliği Süreci Kullanılarak Güneş Işınımı Tahmini”. **SDU International Technologic Science**. c. 3. s. 2 (2011): 33.
- Tsai, Chieh-Yuan, Yu-Chen Shieh, “A Change Detection Method For Sequential Patterns”. **Decision Support Systems**. c. 46 (2009): 501.
- Türe, Mevlüt A, Füsün Tokatlı, İmran Kurt. “Using Kaplan–Meier Analysis Together With Decision Tree Methods (C&RT, CHAID, QUEST, C4.5 And ID3) In Determining Recurrence-Free Survival Of Breast Cancer Patients”. **Elsevier**. c. 36 (2009): 2020.
- Türkiye Bilişim Derneği Kamu Bilgi İşlem Merkezleri Yöneticileri Birliği Kamu Bilişim Platformu. **Kamuda Karar Destek Sistemlerinin Kullanımı ve Bir Model Önerisi**. İstanbul, 2010.
- Tüzüntürk, Selim. “VERİ MADENCİLİĞİ VE İSTATİSTİK”. **Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi**. c. 29. s. 1(2010) : 67.
- Usgurlu, Ümit Burak. “Veritabanı, Veri Madenciliği, Veri Ambarı, Veri Pazarı” . <http://mail.baskent.edu.tr/~20394676/0302/bil483/HW2.pdf> [16.01.2012].
- Üngüren, Engin, Hulusi Doğan. “Beş Yıldızlı Konaklama İşletmelerinde Çalışanların İş Tatmin Düzeylerinin Chaid Analiz Yöntemiyle Değerlendirilmesi”. **C.Ü. İktisadi Ve İdari Bilimler Dergisi**. c. 11. s. 2 (2010): 42.

Vahaplar, Alper, Mustafa Murat İnceoğlu. “Veri Madenciliği ve Elektronik Ticaret”. www.bayar.edu.tr/bilisim/dokuman/inceoglu.doc Benzer [05.12.2011].

“Veri Madenciliği”. www.deu.edu.tr/userweb/k.yaralioglu/dosyalar/ver_mad.doc [27.12.2011].

“Veri Madenciliği”. www.egitim.aku.edu.tr/verimadenciligi.ppt [11.12.2011].

Viguera, G. ve Diğ. “A Comparative Study Of Partitioning Methods For Crowd Simulations”. **Applied Soft Computing**. c. 10 (2010): 226.

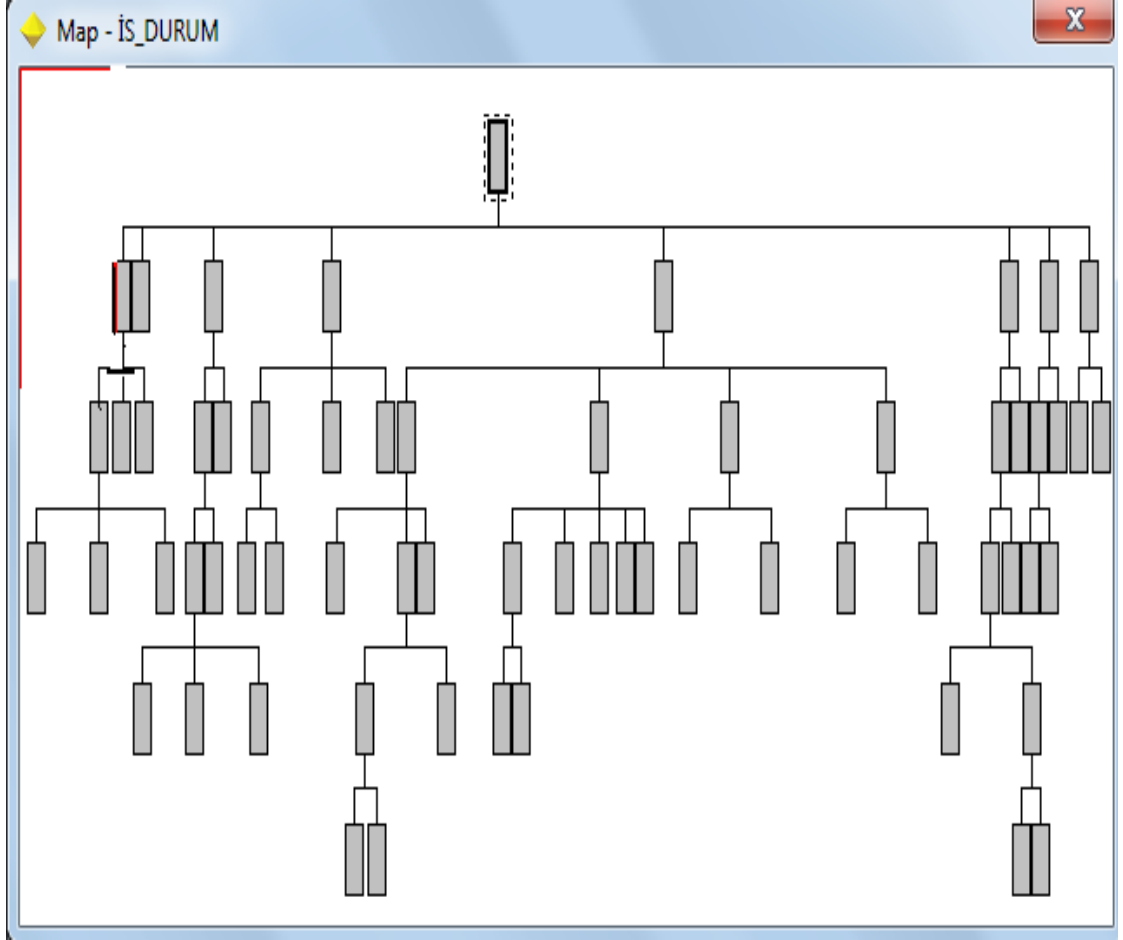
W. Seifert, Jeffrey. “Data Mining: An Overview”. **CRS Report for Congress, 16 Kasım 2004**. Federation of American Scientists, Congressional Research Service, 2004: 4.

Ye, Nong. **The Handbook of Data Mining**. USA: Lawrence Erlbaum Assoc. Inc., 2003.

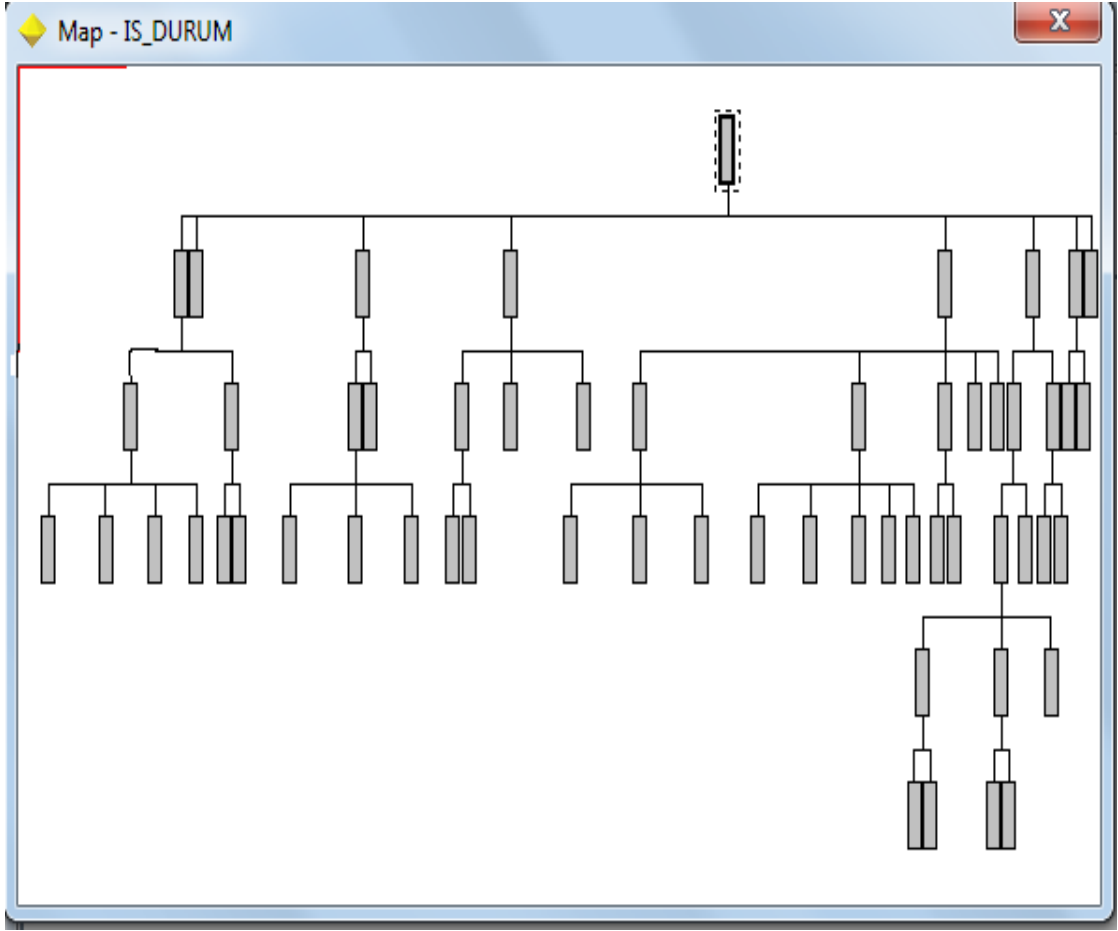
Yun, Unil. “A New Framework For Detecting Weighted Sequential Patterns In Large Sequence Databases”. **Knowledge-Based Systems**. c. 21 (2008): 110.

EKLER

Ek 1. 2009 Yılı İşgücü Durumu İçin Elde Edilen Karar Ağacının Görünümü



Ek 2. 2010 Yılı İşgücü Durumu İçin Elde Edilen Karar Ağacının Görünümü



ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı : Ebru Yılmaz
Uyruğu : T.C.
Doğum Tarihi : 02.10.1984
Doğum Yeri : İstanbul
Medeni Hali : Bekar
E-posta : ebruyilmaz04@gmail.com

Eğitim

Derece Eğitim Birimi Mezuniyet Tarihi Derecesi

Yüksek Lisans Yıldız Teknik Üniversitesi 2012 3,42/4
İşletme Yönetimi

Lisans Yıldız Teknik Üniversitesi 2008 3,78/4
Endüstri Mühendisliği

Lise Vefa Anadolu Lisesi 2003 4,49/5

İş Tecrübesi

Mercedes Benz-Türk It İş Analisti Aralık 2009 - Temmuz 2011
Beko Elektronik A.Ş. İş Geliştirme Proje Stajyeri Mart 2007-Mayıs 2008
THY Teknik A.Ş Atölye Stajı Haziran-Ağustos 2006

Yabancı Dil

İngilizce (İleri Düzey)
Almanca (Temel Düzey)