

**T.C.
ERCIYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**HAVACILIK SEKTÖRÜNDE VERİ MADENCİLİĞİ İLE
FARKLI SINIFLANDIRMA TEKNİKLERİNİN
KARŞILAŞTIRMALI OLARAK UYGULANMASI**

**Tezi Hazırlayan
Feyza GÜRBÜZ**

**Tezi Yöneten
Prof. Dr. Hüseyin YAPICI
Yrd. Doç. Dr. Lale ÖZBAKIR**

**Makine Mühendisliği Anabilim Dalı
Doktora Tezi**

**Mart 2009
KAYSERİ**

**T.C.
ERCIYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**HAVACILIK SEKTÖRÜNDE VERİ MADENCİLİĞİ İLE
FARKLI SINIFLANDIRMA TEKNİKLERİNİN
KARŞILAŞTIRMALI OLARAK UYGULANMASI**

**Tezi Hazırlayan
Feyza GÜRBÜZ**

**Tezi Yöneten
Prof. Dr. Hüseyin YAPICI
Yrd. Doç. Dr. Lale ÖZBAKIR**

**Makine Mühendisliği Anabilim Dalı
Doktora Tezi**

**Bu çalışma Erciyes Üniversitesi Bilimsel Araştırma Projeleri Birimi tarafından
FBT-06-32 kodlu proje ile desteklenmiştir.**

**Mart 2009
KAYSERİ**

Prof. Dr. Hüseyin YAPICI danışmanlığında **Feyza GÜRBÜZ** tarafından hazırlanan “Havacılık Sektöründe Veri Madenciliği ile Farklı Sınıflandırma Tekniklerinin Karşılaştırılmalı Olarak Uygulanması” adlı bu çalışma, jürimiz tarafından Erciyes Üniversitesi Fen Bilimleri Enstitüsü Makine Mühendisliği Anabilim Dalında **Doktora** tezi olarak kabul edilmiştir.

06.03.2009

JÜRİ:

Başka: Prof. Dr. Hüseyin YAPICI

Üye: Doç. Dr. Şahin YILDIRIM

Üye: Doç. Dr. Cem SİNANOĞLU

Üye: Yrd. Doç. Dr. Berrin DENİZHAN

Üye: Yrd. Doç. Dr. Selda KAPAN ULUSOY

ONAY:

Bu tezin kabulü, Enstitü Yönetim Kurulunun 24.03.2009 tarih ve 2009/11-04 sayılı kararı ile onaylanmıştır.

24. / 03. / 2009


N. Ayyıldız
Prof. Dr. Nusret AYYILDIZ
Enstitü Müdürü

TEŞEKKÜR

“Havacılık Sektöründe Veri Madenciliği İle Farklı Sınıflandırma Tekniklerinin Karşılaştırmalı Olarak Uygulanması” konulu tez çalışmasının seçiminde, yürütülmesinde, sonuçlandırılmasında ve sonuçların değerlendirilmesinde maddi ve manevi destek ve yardımlarını esirgemeyen değerli hocam sayın Prof. Dr. Hüseyin YAPICI’ya teşekkür ederim.

Yapılan analizlerde her türlü yardımı yapan, zaman harcayan, emek veren, tecrübesini paylaşan hocam Yrd.Doç.Dr. Lale ÖZBAKIR’a teşekkür ederim.

Ayrıca tüm yorucu çalışmalar sırasında bana destek olan ve sabır gösteren eşime, aileme ve biricik kızıma da teşekkürü bir borç bilirim.

HAVACILIK SEKTÖRÜNDE VERİ MADENCİLİĞİ İLE FARKLI SINIFLANDIRMA TEKNİKLERİNİN KARŞILAŞTIRMALI OLARAK UYGULANMASI

Feyza GÜRBÜZ

Erciyes Üniversitesi, Fen Bilimleri Enstitüsü

Doktora Tezi, Mart 2009

Tez Danışmanları: Prof. Dr. Hüseyin YAPICI

Yrd. Doç. Dr. Lale ÖZBAKIR

ÖZET

Veri madenciliği, farklı bir çok alana başarı ile uygulanan bir tekniktir. Havacılık sektörü de bu alanlardan birisidir. Özellikle Türkiye’de gelişmekte olan hava yolu işletmelerinde de kullanılması mümkündür. Havacılık sektörü, çok büyük miktarlarda veri ve bilgi birikiminin olduğu bir alandır. Bu veriler pilot, bakım, uçuş, motor, kaza, parça veya tehir raporları şeklinde kayıt altına alınmaktadır. Bu çalışmada, Türkiye’de bir hava yolu işletmesinin güvenilirlik birimi veri tabanında yer alan sefer aksaklıkları raporları, parça raporları, pilot/bakım/kabin raporları ve gecikme raporları adı altında yer alan kayıtların veri madenciliği çalışması ile sınıflandırılması gerçekleştirilmiştir. Çalışmanın amacı, yaygın olarak kullanılan veri madenciliği tekniklerinin havacılık sektöründe de kullanılabileceğini ortaya koymaktır. Türkiye dışındaki havacılık sektöründe veri madenciliği teknikleri kullanılmaktadır. Ancak çalışmalar çoğunlukla teknik veriler üzerinde yapılmaktadır. Türkiye’de ki havacılık sektöründe ise işletmeler veri madenciliği konusunda çok fazla bilgi ve deneyim sahibi değildir. Bu çalışma ile veri madenciliği ilgili işletmeye tanıtılmış ve yaygın kullanımı için örnek teşkil etmiştir. Yapılan çalışma ile uçaklarda kullanılan parçaların, herhangi bir arıza oluşmadan önce düzeltici ve önleyici işlemlerin yapılması için ikaz seviyelerinin tespit edilmesine yönelik kural geliştirilmiştir. Sonuç olarak parçaların ikaz seviyelerini temsil edecek anlamlı bir kural elde edilmiş ve bulunan kurallar doğrulukları ve güvenilirlikleri bakımından test edilmiştir.

Anahtar Kelimeler: Veri madenciliği, Havacılık, Sınıflandırma, Kural Bulma

A COMPARATIVE APPLICATION OF DIFFERENT CLASSIFICATION TECHNIQUES WITH DATA MINING ON AVIATION INDUSTRY

Feyza GÜRBÜZ

Erciyes University, Graduate School of Natural and Applied Sciences

Ph.D. Thesis, March 2009

Thesis Supervisors: Prof. Dr. Hüseyin YAPICI

Assist. Prof. Lale ÖZBAKIR

ABSTRACT

Data mining methods have been successfully applied to different fields. Aviation industry is one of these fields. There is a large amount of knowledge and data accumulation in aviation industry. These data could be stored in the form of pilot reports, maintenance reports, incident reports, component reports or delay reports. This paper explains the data mining application on the component reports, incident reports and delay reports of an airline company. In this study, we applied data mining functions on the reports of reliability department of the airline company. The purpose of our application is to find out a rule about the warning level of the components to do correcting operation before any fault. And also find the relationships between the parameters. As a result we found out a meaningful rule that presents the formulation of the warning level of the components. The rules found are tested in terms of their accuracy and reliability. As a result we introduce the data mining techniques to the aviation company and this study can be an example for next applications on aviation industry.

Keywords: Data Mining, Aviation, Classification, Find Laws

İÇİNDEKİLER

KABUL VE ONAY.....	i
TEŞEKKÜR.....	ii
ÖZET.....	iii
ABSTRACT.....	iv
KISALTMALAR VE SİMGELER.....	v
TABLOLAR LİSTESİ.....	xiii
ŞEKİLLER LİSTESİ.....	xx
 1. BÖLÜM	
GİRİŞ.....	1
 2. BÖLÜM	
GENEL BİLGİLER.....	4
 3. BÖLÜM	
VERİ TABANLARINDAN BİLGİ KEŞFİ VE VERİ MADENCİLİĞİ.....	21
3.1. Giriş	21
3.2. Veri Tabanlarından Bilgi Keşfi	25
3.2.1. Veri Tabanlarından Bilgi Keşfi Süreci	27
3.3. Veri Madenciliği	28
3.3.1. Veri Madenciliği Nedir?	28
3.3.2. Veri Madenciliği Neden Gerekli?	30
3.3.3. Makine Öğrenimi ve Veri Madenciliği	31
3.3.4. İstatistik ve Veri Madenciliği	32
3.3.5. Veri Tabanı ve Veri Madenciliği	33
3.3.6. Veri Madenciliği Süreci	33
3.3.6.1.Problemin Tanımlanması	34
3.3.6.2.Verilerin Hazırlanması	34
3.3.6.3.Modelin Kurulması ve Değerlendirilmesi	36
3.3.6.4.Modelin Kullanılması	37
3.3.6.5.Modelin İzlenmesi ve Güncelleme	37

3.4. Veri Madenciliğinde Ön İşleme	38
3.4.1. Veri Temizleme	40
3.4.1.1.Kayıp Değerler	40
3.4.1.2.Gürültülü Veri	40
3.4.1.3.Uyumsuz Veri	41
3.4.2. Veri Birleştirme ve Dönüşümü	41
3.4.2.1.Veri Birleştirme	41
3.4.2.2.Veri Dönüşümü	42
3.4.2.3.Nitelik Oluşturma	42
3.4.3. Veri İndirgemesi	42
3.4.3.1.Veri Birleştirme	43
3.4.3.2.Boyut İndirgeme	43
3.4.3.3.Ver Sıkıştırma	44
3.4.3.4.Çokluk Azaltımı	44
3.4.3.5.Ayrıklaştırma ve Kavram Hiyerarşisi Üretme	44
3.5. Veri Madenciliği Fonksiyonları	45
3.5.1. Veri Madenciliği Modelleri	45
3.5.1.1.Sınıflama ve Regresyon Modelleri	46
3.5.1.2.Kümeleme Modelleri	46
3.5.1.3.Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler	47
3.6. Veri Madenciliği Hedefleri	47
3.6.1. Sınıflandırma	48
3.6.2. Kestirme	49
3.6.3. Öngörü	50
3.6.4. Benzerlik Tespiti	51
3.6.5. Gruplandırma	51
3.6.6. Tanımlama	52
3.7. Çıkarılan Bilgi Türleri	52
3.7.1. Eşleştirme Kuralları	52
3.7.2. Sınıflandırma Sıradüzeni	53
3.7.3. Sıra-Zaman Örüntüleri	53
3.7.4. Zaman Serileri Arası Bağımlılıklar	53
3.8. Veri Madenciliği Yaklaşımları	54

3.8.1. Kullanım Gdml Yaklaşım	54
3.8.2. Otomatik Yaklaşım	55
3.9. Veri Madencilięi Teknikleri	55
3.9.1. Sepet Analizi	55
3.9.2. Bellek Tabanlı Yntemler	57
3.9.3. Kmeleme	58
3.9.4. İlişkişel Analiz	58
3.9.5. Karar Ağaçları ve Kural Tretme	59
3.9.6. Yapay Sinir Ağları	60
3.9.7. Genetik Algoritmalar	61
3.9.8. Kaba Kme Yaklaşımı	62
3.10. Veri Madencilięinde Karşılaşılan Problemler	62
3.10.1. Veritabanı Boyutu	62
3.10.2. Grltl Veri	63
3.10.3. Boş Deęerler	63
3.10.4. Eksik Veri	63
3.10.5. Artık Veri	64
3.10.6. Dinamik Veri	64
3.10.7. Farklı Tipteki Verileri Ele Alma	65
3.11. Veri Madencilięini Etkileyen Eęilimler	65
3.11.1. Veri	65
3.11.2. Donanım	65
3.11.3. Bilgisayar Ağları	66
3.11.4. Ticari Eęilimler	66
3.11.5. Veri Madencilięi Hangi Tip Verilerle Çalışır?	66
3.12. Veri Madencilięi Uygulamaları	66
3.12.1. Uygulamada Kullanılan Sistemler	69
3.12.1.1. Ticari Veri Madencilięi Sistemi rnekleri	70
3.12.1.2. Dięer Ticari Yazılımlar	71
3.11.1.3.cretsiz Sistemler	72

4. BÖLÜM

KABA KÜME YAKLAŞIMI	74
4.1. Giriş	74
4.2. Kaba Kümeler	75
4.3. Kaba Küme Yaklaşımında Temel Kavramlar	76
4.3.1. Bilgi Sistemi	76
4.3.2. Karar Tablosu	77
4.3.3. Ayırt Edilemezlik Kavramı	78
4.3.4. Küme Yaklaşım Kavramı	79
4.3.5. Kaba Üyelik	80
4.3.6. Özellik İndirgeme	81
4.3.7. Karar Kurallarından Sonuç Çıkarma ve Karar Analizi	81
4.4. Kaba Küme Yaklaşımının Avantajları	82
4.5. Rosetta ile Nitelik İndirgeme	83
4.5.1. Genetik Algoritmalar ile Rosetta’da Nitelik İndirgeme	83

5. BÖLÜM

UYGULAMADA KULLANILAN VERİ MADENCİLİĞİ

PROGRAMI VE TEKNİKLER	85
5.1. Giriş	85
5.2. Uygulamada Kullanılan Veri Madenciliği Programları ve Teknikleri ..	85
5.2.1. Polyanalyst Programı.....	85
5.2.1.1. Temel PolyAnalyst Bileşenleri	86
5.2.1.2. PolyAnalyst Ara Yüzü	86
5.2.1.3. Veri Tipleri	88
5.2.1.4. Programda Yer Alan Algoritmalar	89
5.2.1.4.1. Bağımlılıkları Bulma	90
5.2.1.4.2. Karar Ağaçları	91
5.2.1.4.3. Kural Bulma Algoritması	98
5.2.1.4.4. PolyNet Tahminleyicisi	99
5.2.1.4.5. Lineer Regresyon	101
5.2.1.4.6. Bağlantı Analizi	103
5.2.2. Clementine İle Özellik Seçimi ve Aykırılıkların Tespiti	103

5.2.2.1.Özellik Seçimi	104
5.2.2.2.Aykırılıkların Tespiti	106
 6. BÖLÜM	
HAVACILIK SEKTÖRÜNDE VERİ MADENCİLİĞİ UYGULAMASI	107
6.1. Giriş	107
6.2. Değerlendirilen Veri Kümelerinin Tanıtılması	109
6.3. Önişleme Süreci	109
6.3.1. PolyAnalyst Aracı İle Yapılan Önişleme	109
6.3.1.1.Bağımlılıkları Bulma	110
6.3.1.2. Korelasyonların Belirlenmesinde Bağlantı Grafikleri	111
6.3.2. Kaba Kümeler Kullanarak Özellik İndirgeme	114
6.4. Uygulama ve Sonuçlar	114
6.4.1. Karar Ağaçları Analizi	116
6.4.2. Sonuçlar	120
6.5. Sonuç ve Tartışma	121
 7. BÖLÜM	
UYGULAMA	122
7.1. Giriş	122
7.2. Değerlendirilen Veri Kümelerinin Tanıtılması	123
7.2.1.Sefer Aksaklıkları Raporları	124
7.2.1.1.Sefer Aksaklıkları Raporlarında Yer Alan Parametreler ve Tanımları ...	127
7.2.1.2.Sefer Aksaklıkları Raporlarına Uygulanan Ön İşlemler	129
7.2.1.2.1. PolyAnalyst Aracı ile Bağımlılıkların Tespiti	130
7.2.1.2.2. Bağlantı Grafiği Kullanarak Korelasyonu Gösterme	132
7.2.1.2.3. Kaba Küme Yaklaşımı ile Etkili Parametrelerin Tespiti	133
7.2.1.2.4. Clementine Aracı ile Etkili Parametrelerin Tespiti	134
7.2.1.3.Analiz Süreci	134
7.2.1.4.Tehir Süresine Etki Eden Parametrelerin Tespit Edilmesi	140
7.2.1.5.Sonuçlar	140
7.2.2. Pilot/Kabin/Bakım (PMC)Raporları	142
7.2.2.1.Pilot Raporlarının Tutulmasında Etkili Olan Parametrelerin	

Gösterilmesi	143
7.2.2.2.Kabin Raporlarının Tutulmasında Etkili Olan Parametrelerin Gösterilmesi	144
7.2.2.3.Bakım Raporlarının Tutulmasında Etkili Olan Parametrelerin Gösterilmesi	145
7.2.2.4.Sonuçlar	147
7.2.3. Parça Raporlarının Analizi	149
7.2.3.1.Parçat Raporlarına Uygulanan Ön İşlemler	151
7.2.3.1.1. Bağımlılıkları Bulma Analizi ile Özellik Seçimi	152
7.2.3.1.2. Clementine ile Özellik Seçme Analizi	152
7.2.3.1.3. Kaba Kümeler ile Özellik Seçimi	155
7.2.3.1.4. Aykırılıkların Tespiti	161
7.2.3.2.Kural Bulma Algoritması İle İkaz Değerinin Tahmini	162
7.2.3.3.Lineer Regresyon Algoritması ile İkaz Değerinin Tahmin Edilmesi	178
7.2.3.4.Polynet Algoritması ile İkaz Değerinin Tahmin Edilmesi	184
7.2.3.5.Mtbf-Mtbr-Mtbur Parametrelerinin Kural Bulma Algoritması ile Tahmini	189
7.2.3.6.Sonuçlar	192
7.2.4. Gecikme Oranları	200
7.2.4.1.Gecikme Oranının Tahmini	200
7.2.4.1.1. Etkili Parametrelerin Belirlenmesi	200
7.2.4.1.2. Analiz Süreci	202
7.2.4.1.3. Sonuçlar	202
 8. BÖLÜM	
SONUÇ VE İRDELEME	204
 KAYNAKLAR	207
ÖZGEÇMİŞ	218

KISALTMALAR VE SİMGELER

FHY	: Federal Havacılık Yönetimi
VM	: Veri Madenciliği
VTBK	: Veri Tabanlarından Bilgi Keşfi
KK	: Kaba Kümeler
BS	: Bilgi Sistemi
U	: Boş olmayan sonlu nesneler kümesi.
A	: Bilgi sisteminde bulunan özelliklerin sonlu bir kümesi.
I(B)	: Ayırt edilmezlik ilişkisi
$B_x(X)$	: X'in alt yaklaşımı.
$B^x(X)$	: X'in üst yaklaşımı.
$BN_B(X)$	: Sınır bölgesi.
$\alpha_B(X)$	: Tahmin doğruluğu.
$\mu_x^B(X)$	: Kaba küme üyelik fonksiyonu.
GA	: Genetik Algoritma.
$f(B)$	: Uygunluk fonksiyonu.
S	: Ayırt edilebilirlik fonksiyonuna uyan nesnelerin kümesi.
α	: Alt küme maliyeti ve uygun bölge arasındaki ağırlığı.
ε	: Eşik değeri.
k	: Saklı liste genişliği.
P	: Olasılık değeri.
R^2	: Regresyon Katsayısı.
PMC	: Pilot/Bakım/Kabin raporları.
Rs	: Kaba kümeler ile analiz sonucunda elde edilen veri kümeleri.
Fd	: PolyAnalyst bağımlılıkları bulma analizi sonucunda elde edilen veri kümeleri.
Fs	: Clementine özellik seçimi analizi ile elde edilen veri kümeleri.
Q	: Uçaktaki parça miktarı.
H	: Toplam uçuş saati.
Fst	: Toplam söküm.

Fs1	: Son 12 aydaki planlı sökümmler.
Fs2	: Son 12 aydaki plansız sökümmler.
Fso	: Diğer sökümmler.
F	: Son 12 aydaki arızalı sökümmler.
Nf	: Son 12 aydaki faal plansız sökümmler.
F1	: Son 12 aydaki arızalı planlı sökümmler.
F2	: Son 12 aydaki arızalı plansız sökümmler.
Mtbr	: Sökümmler arasındaki ortalama süre.
Mtbur	: Plansız sökümmler arasındaki ortalama süre.
Mtbf	: Arızalar arasındaki ortalama süre.
Mtbf-3	: Son üç aylık arızalar arasındaki ortalama süre.
FF	: İkaz değeri.

TABLOLAR LİSTESİ

Tablo 4.1. Örnek bir bilgi sistemi	77
Tablo 4.2. Örnek bir karar tablosu.....	78
Tablo 5.1. Temel PolyAnalyst Bileşenleri.....	87
Tablo 5.2. PolyAnalyst’de veri tipleri.....	88
Tablo 5.3. Elektrik şirketine ait veri tabanından eğitim kümesi örnekleri.....	96
Tablo 5.4. Standart hata ve regresyon katsayısı formülleri.....	99
Tablo 6.1. Uygulamada kullanılmış veri kümesinde yer alan parametreler ve tanımları.....	110
Tablo 6.2. Analizlerde kullanılmak üzere ön işleme sonucunda elde edilen veri kümesi.....	112
Tablo 6.3. Bağımlılıkları bulma analizi sonuçlarına ait P değerleri ve ilgili parametreler.....	113
Tablo 6.4. Bağlantı grafiklerine ait korelasyon katsayıları ve ilgili parametre düzeyleri.....	113
Tablo 6.5. Kaba küme yaklaşımı ile elde edilen parametre kümeleri.....	115
Tablo 6.6. Karar ağaçları ile elde edilen tahmin kuralları.....	116
Tablo 6.7. Elde edilen kurallara ait sınıflandırma olasılığı ve sınıflandırma etkinlikleri.....	117
Tablo 6.8. Sınıflandırma olasılığı ve etkinliğine ait formüller.....	117
Tablo 6.9. Doğru sınıflandırılan kayıt sayıları kodlamaları.....	117
Tablo 6.10. Kaba küme yaklaşımı ile elde edilen veri kümelerinin karar ağacı analizine ait sınıflandırma etkinlikleri, sınıflandırma olasılıkları ve doğruluk değerleri.....	118
Tablo 6.11. Analiz sonucunda 2-3-4-7-8-16-19-23 no’lu veri kümeleri ile elde edilen kurallar.....	119
Tablo 7.1. Sefer aksaklıkları raporlarına ait parametreler ve özellikleri.....	126
Tablo 7.2. Teknik tehir kodları ve açıklamaları.....	129
Tablo 7.3. Arıza sonucu durumlar ve açıklamaları.....	130
Tablo 7.4. Sefer Aksaklıkları üzerinde yapılan ön işlemler.....	131
Tablo 7.5. Bağımlılıkları bulma algoritması ile belirlenen etkili parametreler.	132

Tablo 7.6. Bağlantı grafiklerinden elde edilen korelasyon katsayısı değerleri.....	133
Tablo 7.7. Kaba küme yaklaşımı ile elde edilen veri kümeleri.....	134
Tablo 7.8. Karar ağacı algoritması ile yapılan analiz sonucunda elde edilen kurallar.....	135
Tablo 7.9. Karar ağacı analizi ile elde edilen kurallara ait sınıflandırma olasılıkları, etkinlikleri ve doğrulukları.....	138
Tablo 7.10. Kural bulma algoritması ile yapılan analiz sonucunda elde edilen kural.....	138
Tablo 7.11. Kural Bulma tekniği ile yapılan analiz sonucunda elde edilen kurallara ait sınıflandırma olasılıkları, etkinlikleri ve doğrulukları.....	139
Tablo 7.12. Lineer regresyon algoritması ile yapılan analiz sonucunda elde edilen kurallar.....	139
Tablo 7.13. Lineer regresyon tekniği ile yapılan analiz sonucunda elde edilen kurallara ait sınıflandırma olasılıkları, etkinlikleri ve doğrulukları.....	139
Tablo 7.14. Tehir süresine etki eden parametreler.....	140
Tablo 7.15. PMC raporlarına ait parametrelerin özellikleri.....	143
Tablo 7.16. PMC Açıklamaları.....	143
Tablo 7.17. Uçuş Fazı Açıklamaları.....	144
Tablo 7.18. Pilot raporları ile ilgili olarak elde edilen etkili parametreler ve düzeylerinin korelasyon katsayıları.....	144
Tablo 7.19. Kabin raporları ile ilgili olarak elde edilen etkili parametreler ve düzeylerinin korelasyon katsayıları.....	145
Tablo 7.20. Bakım raporları ile ilgili olarak elde edilen etkili parametreler ve düzeylerinin korelasyon katsayıları.....	145
Tablo 7.21. Parça raporlarına ait parametrelerin özellikleri.....	150
Tablo 7.22. Parça raporlarında yer alan veri kümelerinin özellikleri.....	151
Tablo 7.23. Bağımlılık bulma analizi sonucunda elde edilen ilgili p değerleri.....	153
Tablo 7.24. Özellik seçme analizi ile tespit edilen etkin parametreler.....	154
Tablo 7.25. Birinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.....	155
Tablo 7.26. İkinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.....	156
Tablo 7.27. Üçüncü veri kümesi ile elde edilen indirgenmiş kaba kümeler.....	156
Tablo 7.28. Dördüncü veri kümesi ile elde edilen indirgenmiş kaba kümeler.....	157

Tablo 7.29. Beşinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.....	157
Tablo 7.30. Altıncı veri kümesi ile elde edilen indirgenmiş kaba kümeler.....	157
Tablo 7.31. Yedinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.....	158
Tablo 7.32. Sekizinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.....	158
Tablo 7.33. Dokuzuncu veri kümesi ile elde edilen indirgenmiş kaba kümeler..	159
Tablo 7.34. Onuncu veri kümesi ile elde edilen indirgenmiş kaba kümeler.....	159
Tablo 7.35. On birinci veri kümesi ile elde edilen indirgenmiş kaba kümeler....	160
Tablo 7.36. On ikinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.....	160
Tablo 7.37. On üçüncü veri kümesi ile elde edilen indirgenmiş kaba kümeler...	160
Tablo 7.38. On dördüncü veri kümesi ile elde edilen indirgenmiş kaba kümeler.	161
Tablo 7.39. On beşinci veri kümesi ile elde edilen indirgenmiş kaba kümeler...	161
Tablo 7.40. Veri kümelerinden çıkarılan aykırı kayıtlar sayısı ile eğitim ve test veri kümelerindeki kayıt sayısı.....	162
Tablo 7.41. Birinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.....	163
Tablo 7.42. Birinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	164
Tablo 7.43. İkinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.	164
Tablo 7.44. İkinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	165
Tablo 7.45. Üçüncü veri kümesinin kural bulma algoritması ile analizinin sonuçları.....	165
Tablo 7.46. Üçüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	166
Tablo 7.47. Dördüncü veri kümesinin kural bulma algoritması ile analizinin sonuçları.	166
Tablo 7.48. Dördüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.	167
Tablo 7.49. Beşinci veri kümesinin kural bulma algoritması ile analizinin Sonuçları.....	167
Tablo 7.50. Beşinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	167

Tablo 7.51. Altıncı veri kümesinin kural bulma algoritması ile analizinin sonuçları.....	168
Tablo 7.52. Altıncı veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	168
Tablo 7.53. Yedinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.....	169
Tablo 7.54. Yedinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	169
Tablo 7.55. Sekizinci veri kümesinin kural bulma algoritması ile analizinin sonuçlar.....	170
Tablo 7.56. Sekizinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	170
Tablo 7.57. Dokuzuncu veri kümesinin kural bulma algoritması ile analizinin sonuçları.....	171
Tablo 7.58. Dokuzuncu veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	172
Tablo 7.59. Onuncu veri kümesinin kural bulma algoritması ile analizinin Sonuçları.....	172
Tablo 7.60. Onuncu veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	173
Tablo 7.61. On birinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.....	173
Tablo 7.62. On birinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	174
Tablo 7.63. On ikinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.....	174
Tablo 7.64. On ikinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	175
Tablo 7.65. On üçüncü veri kümesinin kural bulma algoritması ile analizinin sonuçları.....	175
Tablo 7.66. On üçüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.....	176

Tablo 7.67. On dördüncü veri kümesinin kural bulma algoritması ile analizinin sonuçları.....	176
Tablo 7.68. On dördüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.	177
Tablo 7.69. On beşinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.....	177
Tablo 7.70. On beşinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.	178
Tablo 7.71. Birinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	179
Tablo 7.72. Birinci veri kümesinin lineer regresyon analizi sonucunda elde edilen en güvenilir ve doğrulukları yüksek kurallar.....	180
Tablo 7.73. İkinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	180
Tablo 7.74. Üçüncü veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	181
Tablo 7.75. Dördüncü veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	181
Tablo 7.76. Beşinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	181
Tablo 7.77. Altıncı veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	181
Tablo 7.78. Yedinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	182
Tablo 7.79. Sekizinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	182
Tablo 7.80. Dokuzuncu veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	182
Tablo 7.81. Onuncu veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	182
Tablo 7.82. On birinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	183

Tablo 7.83. On ikinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	183
Tablo 7.84. On üçüncü veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	183
Tablo 7.85. On dördüncü veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	183
Tablo 7.86. On beşinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.....	184
Tablo 7.87. Birinci veri kümesinin polynet algoritması ile analizinin sonuçları....	184
Tablo 7.88. İkinci veri kümesinin polynet algoritması ile analizinin sonuçları.....	185
Tablo 7.89. Üçüncü veri kümesinin polynet algoritması ile analizinin sonuçları....	185
Tablo 7.90. Dördüncü veri kümesinin polynet algoritması ile analizinin sonuçları..	185
Tablo 7.91. Beşinci veri kümesinin polynet algoritması ile analizinin sonuçları.....	185
Tablo 7.92. Altıncı veri kümesinin polynet algoritması ile analizinin sonuçları.....	186
Tablo 7.93. Yedinci veri kümesinin polynet algoritması ile analizinin sonuçları.....	186
Tablo 7.94. Sekizinci veri kümesinin polynet algoritması ile analizinin sonuçları...	186
Tablo 7.95. Dokuzuncu veri kümesinin polynet algoritması ile analizinin sonuçları.	187
Tablo 7.96. Onuncu veri kümesinin polynet algoritması ile analizinin sonuçları.....	187
Tablo 7.97. On birinci veri kümesinin polynet algoritması ile analizinin sonuçları..	187
Tablo 7.98. On ikinci veri kümesinin polynet algoritması ile analizinin sonuçları...	188
Tablo 7.99. Onüçüncü veri kümesinin polynet algoritması ile analizinin sonuçları...	188
Tablo 7.100. On dördüncü veri kümesinin polynet algoritması ile analizinin sonuçları.....	188
Tablo 7.101. On beşinci veri kümesinin polynet algoritması ile analizinin sonuçları.	188
Tablo 7.102. Mtb _f , Mtb _r ve Mtb _{ur} parametrelerine etki eden diğer parametreler ve ilgili P değerleri.....	189
Tablo 7.103. Mtb _f parametresinin tahmini için yapılan kural bulma analizi Sonuçları.....	190
Tablo 7.104. Mtb _{ur} parametresinin tahmini için yapılan kural bulma analizi sonuçları.....	190
Tablo 7.105. Mtb _r parametresinin tahmini için yapılan kural bulma analizi sonuçları.....	190

Tablo 7.106. Mtb _f parametresinin tahmini için yapılan lineer regresyon analizi sonuçları.....	190
Tablo 7.107. Mtb _{ur} parametresinin tahmini için yapılan lineer regresyon analizi sonuçları.....	191
Tablo 7.108. Mtb _r parametresinin tahmini için yapılan lineer regresyon analizi sonuçları.....	191
Tablo 7.109. Mtb _f parametresinin tahmini için yapılan polynet algoritması ile analizinin sonuçları.....	191
Tablo 7.110. Mtb _{ur} parametresinin tahmini için yapılan polynet algoritması ile analizinin sonuçları.....	191
Tablo 7.111. Mtb _r parametresinin tahmini için yapılan polynet algoritması ile analizinin sonuçları.....	191
Tablo 7.112. Kural bulma algoritması ile yapılan analiz sonuçları.....	193
Tablo 7.113. Kural bulma, Lineer regresyon ve polynet algoritmaları analizlerinden elde edilen en yüksek R ² değerleri.....	196
Tablo 7.114. Mtb _f , Mtb _r ve Mtb _{ur} parametrelerinin analizinden elde edilen R ² değerleri.....	199
Tablo 7.115. Mtb _f , Mtb _r ve Mtb _{ur} parametrelerinin analizinden elde edilen doğrulukları en yüksek kurallar.....	199
Tablo 7.116. Gecikme oranlarına ait kayıtlardaki parametreler ve tanımları.....	201
Tablo 7.117. Bağımlılıkları bulma algoritması ile belirlenen etkili parametreler.....	201
Tablo 7.118. Kaba küme yaklaşımı ile elde edilen veri kümeleri.....	202
Tablo 7.119. Sınıflandırma algoritmaları ile yapılan analiz sonucunda elde edilen R ² değerleri.....	203
Tablo 7.120. Gecikme oranlarının tahmini için elde edilen en güvenilir kurallar....	203

ŞEKİLLER LİSTESİ

Şekil 3.1. Veri tabanı sistemlerinin gelişimi.....	24
Şekil 3.2. Veri madenciliği ile disiplinler arası ilişki.....	25
Şekil 3.3. Veri tabanlarından bilgi keşfi süreci.....	27
Şekil 3.4. Tipik bir veri madenciliği sistemi.....	30
Şekil 3.5. Veri madenciliği süreci	34
Şekil 3.6. Veri madenciliği modelleri	38
Şekil 3.7. Verileri bir araya getirme örneği.....	43
Şekil 5. 1. PolyAnalyst temel bölümleri.....	88
Şekil 5.2. Temel ID3 algoritması.....	94
Şekil 5.3 Başlangıç karar ağacı kökü.....	97
Şekil 5.4. Son şekliyle karar ağacı diyagramı.....	98
Şekil 5.5. SKAT algoritması akış diyagramı.....	101
Şekil 5.6. Örnek bağlantı analizi grafiği.....	103
Şekil 5.7. Clementine örnek veri akış diyagramı.....	104
Şekil 5.8.Clementine kullanıcı ara yüzü.....	105
Şekil 5.9. Aykırı değerlerin tespit edilmesi ile ilgili örnek.....	106
Şekil 6.1. Federal Havacılık Yönetimi verileri ile yapılan veri madenciliği süreci.....	108
Şekil 7.1. Çalışma kapsamında yapılan veri madenciliği süreci.....	125
Şekil 7.2. Ata kodu yapısı.....	127
Şekil 7.3. PMC raporları ile uçuş fazı parametresi arasındaki ilişki.....	146
Şekil 7.4. PMC raporları ile filo parametresi arasındaki ilişki.....	146
Şekil 7.5. PMC raporları ile uçak kodu parametresi arasındaki ilişki.....	147
Şekil 7.6. PMC raporları ile ata kodu parametresi arasındaki ilişki.....	147

1. BÖLÜM

GİRİŞ

Bilgisayar sistemleri her geçen gün hem daha ucuzlamakta hem de güçleri artmaktadır. İşlemciler gittikçe hızlanmakta disklerin kapasiteleri artmaktadır. Artık bilgisayarlar daha büyük miktardaki veriyi saklayabilmekte ve daha kısa sürede işleyebilmektedirler [1].

Bilgisayarların ucuzlaması ile sayısal teknoloji daha yaygın olarak kullanılmaktadır. Veri doğrudan sayısal olarak toplanabilmekte ve saklanabilmektedir. Bunun sonucu olarak da detaylı ve doğru bilgiye ulaşılabilir.

Çok büyük ölçekli veriler, farklı alanlardaki büyük ölçekli veri tabanları içlerinde değerli verileri bulunduran bir veri madeni gibi düşünülebilir. Bu kadar büyük ölçekli veri tabanları arasında yararlı verilere ya da bilgiye erişim ihtiyacı doğmuştur. Bu verilerin analizi, bu analiz sonucunda daha anlamlı bilgi elde etme ve elde edilen bilgiyi yorumlama işi insan yeteneğini ve ilişkisel veri tabanlarının yapabileceklerini aşmaktadır. Bu ihtiyaçların sonucunda otomatik ve akıllı veri tabanı analizi için yeni kuşak teknikler doğmuştur. Bu teknikler öyle olmalıydılar ki, veriyi akıllı ve otomatikleşmiş şekilde işe yarar bilgiye dönüştürebilsin. Tüm bunların sonucunda veri madenciliği cevap olarak sunulmuş ve giderek önemini artıran bir araştırma alanı haline gelmiştir. Veri madenciliği, iş yönetimi, ürün kontrol sistemleri, pazarlama, mühendislik ile ilgili analizler gibi konularda kullanılmaya başlanmıştır.

Veritabanlarından Bilgi Keşfi ve Veri Madenciliği terimleri veritabanlarında aynı anlamı ifade eder. *Veritabanlarından bilgi keşfi*, veri içerisindeki geçerli, yeni, yararlı

ve sonuç olarak anlaşılabilir örüntülerin çıkarılması sürecidir [2]. Veriler bir veritabanına kaydedilir. Bilgi tabanındaki bilgi keşfi sürecinden sağlanan bilgi basit hesapla elde edilemez. Bilgi kullanıcı için yeni ve avantajlı olmalıdır. Bazı doğruluk dereceleriyle yeni veriye uygulanabilmelidir. Son olarak bilgi insanın anlayabileceği şekilde olmalıdır. Bunun yanında veri madenciliği terimi verideki kullanışlı örneklerin bulunmasını ifade eder. Veri madenciliği, verilerden modeller veya örnekler üretmek için veri analizleri ve keşif algoritmaları kullanmayı gerektirir.

Veritabanındaki bilgi keşfi süreci, birkaç adımdan oluşan etkileşimli ve iteratif bir süreçtir. Bu süreç, uygulama alanının öğrenilerek, uygulamanın amaçları doğrultusunda hedef veri setinin seçilmesi ile başlar. Daha sonra, gürültülü, boş ve tutarsız verilerin çıkarıldığı veri temizleme ve ön işleme basamağı gelir. Gerekli durumlarda veri, madenciliğe uygun bir forma dönüştürülür. Sonraki aşama olan *veri madenciliği*, zeki yöntemler aracılığıyla büyük miktarda veriden anlamlı bilgilerin çıkarılması sürecidir. Daha sonra, çıkarılan örüntüler, içlerinden yararlı olanların belirlenmesi için değerlendirilir. Veritabanlarında bilgi keşfinin son basamağı ise, elde edilen bilginin görüntüleme ve bilgi gösterimi yöntemleri kullanılarak kullanıcıya sunulmasıdır. Veri madenciliği modellerini gördükleri işlevlere göre sınıflama ve regresyon, kümeleme ve birliktelik kuralları olmak üzere üç ana başlık altında incelemek mümkündür [3]. Veri madenciliği, istatistik, makine öğrenimi, veri yönetimi ve veri tabanları, yapay zeka gibi alanların kesişimi esasına dayalı bir disiplindir [4].

Son yıllarda hava taşımacılığında meydana gelen hızlı gelişmeler ve artışlarla, uçuş ertelemeleri ve iptalleri ile birlikte kazalarda da hızlı bir artış gözlenmektedir [5]. Bunlarla birlikte tipik bir havayolu yılda yapısal ve metinsel veri içeren, raporlar toplamaktadır. Kuruluşlar çeşitli kaynaklardan bu tip havacılık güvenlik bilgilerini toplayabilmek amacıyla önemli yatırımlar yapmaktadırlar. Bu bilgilerin depolandığı veritabanları birçok kritik veri içermektedir; bununla birlikte bilginin analiz edilmesinde büyük problemler yaşanmaktadır. Havacılık endüstrisinde büyük hacimli karma yapıdaki verilerin zeki otomatik analizinde; yetenekli araçlara ihtiyaç duyulmaktadır. Bu araçlardan biri olan veri madenciliği, önceden bilinmeyen ve potansiyel olarak faydalı olabilecek, veri içinde gizli bilgilerin çıkarılmasına yarayan bir yöntemdir [1].

Bu çalışmada, bir havacılık sektörünün güvenilirlik birimi veri tabanında yer alan veriler incelenerek seçilen 4 farklı veri kaynağının, veri madenciliği sınıflandırma araçlarıyla analizleri gerçekleştirilmiştir. Analizler sonucunda, uçuşlarda meydana gelen sefer aksaklıkları ve uçaklarda kullanılan parçaların, herhangi bir arıza oluşmadan önce düzeltici ve önleyici işlemlerin yapılması için ikaz seviyelerinin tespit edilmesine yönelik kurallar mevcut veri madenciliği yazılımları yardımıyla geliştirilmiştir. Sonuç olarak anlamlı kurallar elde edilmiş ve bulunan kurallar doğrulukları ve güvenilirlikleri bakımından test edilmiştir. Yapılmış olan literatür çalışmalarına bakıldığında havacılık sektörüne ait sefer aksaklıkları ve parça raporlarına yönelik veri madenciliği çalışması daha önce yapılmamıştır.

Bu çalışmaları içeren tez yedi bölümden oluşmaktadır. 1. Bölüm’de çalışma konusu ve önemi izah edildikten sonra çalışmanın amacı ve kapsamı anlatılmıştır. 2. Bölüm’de veri madenciliği sınıflandırma tekniklerinin kullanım alanları ve biçimleri, veri madenciliğinde ön işleme ve kaba küme yaklaşımı ile havacılık sektöründe yapılmış veri madenciliği ile ilgili yapılmış çalışmalar özetlenmiştir. 3. Bölüm’de veri tabanlarından bilgi keşfi ve veri madenciliği konusu ayrıntılı olarak anlatılmıştır. 4. Bölüm’de veri ön işleme sürecinde kullanılan veri madenciliği yaklaşımlarından kaba küme yaklaşımı irdelenmiştir. 5. Bölüm’de uygulamada kullanılan veri madenciliği programı ve teknikleri ayrıntılı olarak verilmiştir. 6. Bölüm’de ise Federal Havacılık Yönetimi (FHY), kaza veri sisteminden alınmış kaza raporları üzerinde veri madenciliği çalışması yapılmış ve sonuçlar irdelenmiştir. 7. Bölüm’de ilgili hava yolu şirketine ait 4 farklı veri kaynağının veri madenciliği sınıflandırma ve ön işleme teknikleri ile analizi gerçekleştirilmiş ve yapılan çalışma ile Türkiye’de hava yolu şirketlerinde veri madenciliği çalışmalarının yapılabilirliği ortaya konulmuştur. Bu çalışmalarla ilgili sonuçlar ve öneriler 8. Bölüm’de verilmiştir.

2. BÖLÜM

GENEL BİLGİLER

2.1. Giriş

Veri madenciliği (VM) uygulama alanı çok geniş olan bir disiplindir. Bu bölümde genel olarak veri madenciliği ile sınıflandırma tekniklerinin hangi alanlarda ve nasıl kullanıldığı ile ilgili literatürde yapılmış olan çalışmalardan bahsedilmiştir. Ayrıca havacılık sektöründe veri madenciliği ile yapılmış olan çalışmalara da yer verilmiştir.

2.2. Literatür Araştırması

Günümüzde her gün terabaytlarca veri birçok organizasyon tarafından üretilmektedir. Bununla birlikte veri toplama ve üretme ile ilgili teknolojiler hızla gelişmektedir. Büyük miktarlarda veri içerisinden anlamlı bilgilerin çıkarılması için veri madenciliği teknikleri kullanılmaktadır [6-9]. Veri madenciliği metotları birçok alanda başarılı bir şekilde uygulanmaktadır. Şirketler veri madenciliğini gerçek hayat problemlerini çözmek için kullanmalarına karşın hala araştırma konusu olarak çalışılmaktadır [10]. İstatistik, yapay sinir ağları, karar ağacı yaklaşımları ve genetik algoritmalar gerçek hayat problemlerini çözmek için kullanılan veri madenciliği teknikleridir. Aşağıdaki paragraflarda çeşitli veri madenciliği tekniklerinin uygulama alanları ile ilgili yapılmış çalışmalardan bahsedilmiştir.

Hu ve Chen [11], 2008 yılında yaptıkları “Veri Tabanlarından Tipik Örneklerin Madenciliği” adlı çalışmalarında, “tipik çıkarım-TPM (Typical Patterns Mining)” adını verdikleri yeni bir bilgi çıkarımı türü geliştirmişlerdir. TPM algoritması veri

tabanlarından tipik örneklerin çıkarımı ve madenciliği için geliştirilmiştir. Deneysel sonuçlar, TPM algoritmasının etkili bir yöntem olduğunu ve tipik örneklerin orijinal veri kümesini uygun bir şekilde temsil edebileceğini göstermiştir.

Huang ve ark. [12], Taiwan’da bir mağazaya ait müşteri kayıtlarının sınıflandırılmasına yönelik, K-ortalamları, Bulanık c-ortalamları ve kova metodu ile kümeleme algoritmalarını gerçek kayıtlara uygulayarak elde edilen sonuçları karşılaştırmışlardır.

Ciarapica ve Giacchetta [13], İtalya’da 2002-2006 yılları arasında meydana gelen beş yıllık iş kazaları ile ilgili kayıtlara uyguladıkları analizde, neuro-fuzzy ağırları tekniğini verilerin sınıflandırılması için uygulayarak, bu tekniğin gerçek hayat problemlerine uygulanabilirliğini göstermişlerdir. Benzer şekilde Liao ve Perng [14], Taiwan’da bir işletmeden aldıkları 5 yıllık iş kazası verilerini analiz etmişlerdir. İş kazalarının karakterlerini tanımlamak için birliktelik kurallarından yararlanmışlardır.

Ghazavi ve Liao [15], yüksek boyutlu tıbbi verilerin, sadece işlem maliyetlerini azaltmak için değil aynı zamanda oluşturulan modelin kullanılabilirliğini artırmak amacıyla, özellik indirgeme yaparak bulanık modelleme ile veri madenciliği ile analizini gerçekleştirmişlerdir.

Bertolazzi ve ark. [16], büyük miktarlarda biyomedikal verilere lojik sınıflandırma ve özellik seçimi tekniklerini uygulamışlardır.

Yeh ve Lien [17], “Kredi kartı müşterilerinin tahmininde tahmin doğruluğuna göre veri madenciliği tekniklerinin karşılaştırılması” adlı çalışmalarında, veri madenciliği tekniklerinden, K-en yakın komşu, lojistik regresyon, diskriminant analizi, naive-bayes, yapay sinir ağırları ve sınıflandırma ağaçları tekniklerinin karşılaştırmasını yaparak, bunların içerisinde en doğru sınıflandırmayı yapay sinir ağı tekniğinin yaptığını göstermiştir.

Ke ve ark. [18], Çin’de bir motor üretici firmasından aldıkları kişisel gelişim eğitim verilerine, kümeleme analizi, karar ağaçları algoritması ve geri yayılım algoritması ile

yapay sinir ağıları metotlarını uygulamışlardır. Sınıflandırma algoritmaları sonuçlarına göre talep edilen bir kişisel gelişim modeli sunmuşlardır.

Piramuthu [19], “Karar ağaçları için girdi verisi” adlı çalışmasında, karar ağaçlarının girdi verilerinin karakterlerinin ve öğrenme performansı üzerindeki etkilerini araştırmışlardır. Sonuç olarak da veri üzerinde yapılacak çok küçük değişkenlerin bile karar ağacının performansını artıracağını göstermişlerdir.

Olafsson ve ark. [6], yöneylem araştırması ve veri madenciliği tekniklerinin kesişimine yönelik bir araştırma yapmışlardır. Çalışmanın amacı iki alan arasındaki kesişimin oranını tanımlamaktır. Veri madenciliğinde kullanılabilecek yöneylem araştırması teknikleri ile, yöneylem araştırmasında kullanılabilecek veri madenciliği tekniklerinden bahsetmişlerdir. Son olarak da bazı uygulamalardan bahsetmişlerdir.

Azadeh ve ark. [20], yaptıkları çalışmada, mevsimsel ve aylık olarak elektrik talebinin tahminine yönelik bulanık sistem, veri madenciliği ve zaman serileri bütünleşik bir sistem sunmuşlardır.

Li ve Jacob [21], çalışmalarında orijinal veri kümesini temsil edecek alt kümelerin seçiminde yeni bir metottan bahsetmişlerdir. Genetik algoritmalar araştırma/optimizasyon tekniğini kullanmışlardır.

Urtubia ve ark. [22], 24 farklı şarap imalatçısından elde ettiği şarap mayalanmasına yönelik problemlerle ilgili verileri, aykırı davranışların tespitinde veri madenciliği tekniklerinin nasıl kullanıldığını göstermek için analiz etmişlerdir. Çalışmada ilk olarak boyut indirgeme tekniklerinden olan Temel Bileşenler Analizi (PCA-Principal Component Analysis) uygulanmış daha sonra da K-ortalamları tekniği ile kümeleme analizi yapılmıştır. Elde edilen sonuçlar ümit verici olmuştur.

Kusiak ve Smith [23], veri madenciliği uygulamaları için uygun olan ürün ve üretim sistem tasarımı alanlarını ana hatlarıyla araştırmışlardır. Bu alanlardan bir tanesinin yenilik olduğunu ortaya koymuşlardır.

Gürbüz ve Özbakır [24], yaptıkları çalışmada, birden fazla çıktısı olan sistemlerde karar verme sürecini kısaltmak ve daha etkili kararlar verebilmek için veri madenciliği teknikleri kullanılarak sistemin karar verme mekanizması modellenmiştir. Bunun için, bir seramik üretimi yapan işletmeden ürünün kabul veya reddine karar vermemizde etkili olan veriler alınmış, bu veriler veri madenciliği teknikleri ile analiz edilerek ürünün kabulü için en uygun şartların belirlenmesine yardımcı olunmuştur.

Özbakır ve ark. [25], bir tekstil işletmesinde kalite iyileştirmeye yönelik bir veri madenciliği çalışması yapmışlardır. Veri madenciliği uygulamasında bir tekstil işletmesinin veri tabanından alınan kalite hatalarına yönelik veriler kullanılmıştır. Hata türü olarak işletmedeki elli dört hata türünden sırasıyla % 9,38, % 5,2 ve % 10 oranlarla en çok rastlanan kalite hatalarından duruş-kalkış, düğüm ve kafes hataları analiz kapsamına alınmıştır. Ele alınan hata türünde 100 m²'ye düşen hata oranını etkileyen parametrelerin belirlenmesi ve ilgili kuralların türetilmesi için SPSS-Clementine programının Yapay Sinir Ağları, C5.0 Düğümü ve C & R Düğümü (Sınıflandırma Ve Regresyon Ağacı Düğümü) modelleri kullanılmış ve bu modellerin sonuçları karşılaştırılmıştır.

Gürbüz ve Kulluk [26], 2007 yılında yaptıkları çalışmada, veri madenciliğinin uygulama alanlarından bahsederek bir imalat işletmesine ait veri kümeleri üzerinde gerçekleştirilmiş olan veri madenciliği çalışması anlatılmıştır. Ayrıca bir veri madenciliği yazılımı olan PolyAnalyst programı kısaca tanıtılmıştır. Uygulamada kullanılmış veriler bir tekstil fabrikasına ait veri tabanından alınmıştır. Veri tabanında yer alan kumaş üretimine ait kalite hata raporlarından pamuklama hatalarına ait veriler alınmış, programda kullanılabilecek formata uygun hale getirmek için gerekli ön işlem ve temizleme işlemleri uygulanmıştır. Analizdeki hedef değişken hata puanı ve yüz metrekaredeki hata parametreleridir. Amaç, bağımsız parametrelere bağlı olarak hata puanı ve yüz metrekaredeki hata oranını tahmin edecek bir kural oluşturmaktır. Analiz sonucunda da %96 doğrulukla bir tahmin kuralı elde edilmiştir.

Kulluk ve ark. [27] “Öğrenci Başarısını Etkileyen Faktörleri Belirlemeye Yönelik Bir Veri Madenciliği Uygulaması” başlıklı yaptıkları çalışmada, Erciyes Üniversitesi Mühendislik Fakültesi son sınıf öğrencilerine, öğrenci başarısına etki eden faktörler

üzerine uygulanan anket verileri kullanılmıştır. Ele alınan veri kümesinde beşi ikili, ikisi dilsel ve yirmi altısı kategorik olmak üzere toplam otuz üç girdi parametresi ve bir ikili çıktı parametresi mevcuttur. Analizlerde Clementine paket programının C5.0 modeli ile Weka paket programında PART ve NB-Ağacı olmak üzere üç model üzerinde analizler yapılarak sonuçla karşılaştırılmıştır.

Kusiak [28], tarafından yapılan “Veri Madenciliği: İmalat ve Hizmet Sektörü Uygulamaları” adlı çalışmada, makine öğrenimi ve veri madenciliğinin temel kavramları tanıtılmıştır. Makine öğrenimi algoritmaları, çok miktarda veriden karar verme sistemleri oluşturmak için bilgi çıkarır. Yapılan çalışmada imalat ve servis uygulamalarında karar verme için bilginin organize edilmesi ve uygulaması anlatılmıştır. Uygulamada karar verme yapıları olarak karar tabloları, karar haritaları ve atlaslar kullanılmıştır. Modern imalat ve servis organizasyonları için yeni bir paradigma önerilmiştir. Endüstride, tıpta... örnek veri madenciliği uygulamaları sunulmuştur. Alüminyum işleme, yarı iletken imalatında, elektronik imalatında, DNA üretiminde, biyoteknoloji ve kimya endüstrisinde, enerji üretiminde, tıbbi uygulamalarda ve ilaç sektöründeki uygulamalardan bahsedilmiştir.

Hsia ve ark. [29], 2006’da yaptıkları çalışmada Taiwan’da bir üniversitede eğitim kurslarının devam ettirilmesi ile ilgili olarak, kurs tercihlerinin ve kursa katılım oranlarının veri madenciliği ile analizini gerçekleştirmişlerdir. Çalışmada kullanılan veriler 2000-2005 yılları arasındaki 5 akademik yıla ait 1408 kayıttan oluşmaktadır. Bu veriler eğitim merkezlerinden alınmıştır. Verileri öncelikle; mühendislik, iş ve yönetim, bilgi yönetimi, yabancı dil ve diğer konular olmak üzere 5 gruba ayırmışlardır. Analiz aracı olarak karar ağaçları, bağlantı analizleri ve karar ormanları kullanılmıştır. Analiz sonucunda karar ağaçları ile katılımcıların kurs tercihleri, bağlantı analizi ile önerilen kurs kategorileri ile katılımcıların meslekleri arasındaki ilişkiler bulunmuş, karar ormanları ile de katılımcıların kurs tercihleri, meslekleri ile kursu tamamlama olasılıkları arasındaki ilişkiler ortaya konulmuştur. Elde edilen sonuçlar Chinkuo Teknoloji Üniversitesi’nde eğitim planlarının geliştirilmesinde referans olarak kullanılacaktır. Karar ağacı analizi sonucunda; mühendislik, iş ve yönetim ve bilgi yönetimi kurslarının en popüler kurslar olduğu, ayrıca imalat sektörünün büyük çoğunluğunun bilgi yönetimi kurslarına katıldığı, servis sektörünün ise en çok iş ve

yönetim kurslarına katıldığı ortaya çıkmıştır. Bağlantı analizi göstermiştir ki imalat sektöründen gelen katılımcılar ile kursu tamamlama arasında yüksek korelasyon vardır. Bir sonraki en yüksek korelasyon ise kurs tamamlama ile bilgi yönetimi arasındadır. Servis sektöründen katılanlar ise en az katılımı olanlardır. Son olarak karar ormanı analizi ile görüldü ki imalat sektörü daha çok bilgi yönetimi ve mühendislik kurslarına katılırken servis sektörü çoğunlukla iş ve yönetimi, yabancı dil ve diğer kursları tercih etmektedirler. Bu üç algoritma sonucu da birbirleri ile tutarlı sonuçlar vermiştir. İleriki çalışmalar da bu çalışmanın sonuçları diğer üniversiteler üzerinde test edilebilir.

2006 yılında Chang ve Chen [30], yaptıkları çalışmada sınıflandırma problemlerini çözmede veri madenciliği teknolojilerinin kullanımına örnek olarak üniversite kampüsündeki elektronik kütüphanelerdeki uygulamadan bahsetmiştir. Örnek çalışmadaki veri kümesi kampüsdeki dijital kütüphaneye ait 2000-2003 yılları arasındaki okuyucu verileri ve ödünç verme kayıtlarını içermektedir. Okuyucu bilgileri; okuyucu kimlik numarası, kimlik tipi, bölüm ve cinsiyetten oluşmaktadır. Ödünç verme kayıt bilgileri ise okuyucu kimlik numarası, süre... gibi milyonlarca kayıt içermektedir. Çalışmanın amacı, ödünç alma geçmişlerine göre okuyucuları sınıflandırmaktır. Gerekli veri ön işleme ve dönüşümleri yapıldıktan sonra veri madenciliği analizi ile veriler 5 kümeye ayrılmıştır. Her bir sınıf için farklı karakterler belirlenmiştir.

Proje karlılık değişkenlerinin belirlenmesinde veri madenciliği modeli adlı çalışmada Chang ve Leu [31], 1996-2001 yılları arasındaki 548 tamamlanmış proje verisi üzerinde bir sistem modeli ve veri madenciliği sürecini anlatmışlardır. Birçok mühendislik şirketi projelerini yürütmek için veri toplamak üzere birimler oluşturmaktadırlar. Fakat işlemsel değişkenler ile performans arasındaki ilişki ile genellikle ilgilenilmez ve analiz edilmez. Bu çalışmada proje karlılığında etkili olan parametrelerin tespit edilmesi ile ilgili adımlar ve veri madenciliği modeli sunulmaktadır. Girdi ve işlem parametrelerinin karlılıkla olan ilişkileri tespit edilmiştir. Verilen bilgi ve veriler temel alınarak önerilen bu model, şirketlere performanslarını geliştirmek ve işlemsel bilgi edinmelerinde yardımcı olacaktır.

Rubinov, ve ark. [32], 2005 yılında “Veri analizinde kümeler ve sınıflar” adlı çalışmalarında verilen sınıflardaki veri kümelerindeki sınıf ve kümeler arasındaki

ilişkileri tartışmışlardır. Elde edilen kümeler içinde sınıfların dağılımını farklı tekniklere dayanan farklı sınıflandırma metotları kullanarak incelemişlerdir. Aynı zamanda elde edilen kümelerin yapıları üzerinde de çalışmışlardır. Bu çalışmada elde edilen temel sonuçlardan birisi, kümeleme tekniklerinin doğruluğunu değerlendirmek için kavram saflığının her zaman kullanılamayacağıdır.

Aynı yıl Inniss [33], yaptığı çalışmada veri madenciliği uygulamalarında verideki bilinmeyen örüntü ve eğilimlerin ortaya çıkarılmasında sınıflandırma tekniklerinden olan danışmansız öğrenmenin yararlı bir teknik olduğu üzerinde durmuştur. Zaman serili verilerdeki sınıflara karar vermek için mevsimsel sınıflandırma tekniği geliştirmek için matematiksel programlama, istatistiksel teknikler ve metodolojiler birleştirmiştir. Bu tekniği, verimli trafik akış yönetiminde kullanılabilen olasılıklı varış kapasitelerinin dağılımına karar vermek için hava ve havacılık verilerine uygulamıştır. Genel olarak bu teknik mevsimsel tahmin ve planlama için kullanılabilir.

2005 yılında Mai [34], tarafından yapılan “Orman Veri Madenciliği için PolyAnalyst Uygulaması” adlı çalışmada gizli bilgi ve ilişkilerin çıkarımının veri madenciliğinin temel uygulamaları olduğundan bahsetmiştir. Polyanalyst veri madenciliğinde kullanıcı için dost olan bir yazılımdır. Kural çıkarımı, sınıflandırma ve tahmin, kümeleme modüllerini içermektedir. Bu çalışmada ormansız kalmaya etki edecek orman verilerine veri madenciliği uygulanmıştır. Dört temel faktör, yollar, köyler, insan nüfusu ve hayvan sürüleri, dikkate alınmıştır. Ormanlık bölgelerin uydu görüntülerinden elde edilmiş 300 veri ile çalışılmıştır. Amaç ormanların yok olmasına köylerin mi yoksa yollar mı sebep olduğunu tespit etmektir. Analiz için Polyanalyst programının lineer regresyon ve sınıflandırma algoritmaları uygulanmıştır.

Tseng ve ark. [35], CNC tezgahlarında işlem görmüş parçaların kabulünün tahminindeki kalite güvence problemini çözmek için yeni bir yaklaşım sunmuşlardır. Gerçek yüzey pürüzlülük değerinin tahmini yerine, işlem görmüş parçaların kalite kabulü üzerinde odaklanarak, sunulan kural bileşimi algoritmasını ve kural onaylama prosedürünü, işletmenin uzun yıllarda topladığı geçmiş verileri üzerinde test etmişlerdir. Elde ettikleri sonuçlar, geliştirilen tekniğin yüzey pürüzlülüğünün kabul edilebilirliğinin

tahmini açısından istatistiksel tekniklerden çok daha doğru sonuçlar verdiğini göstermiştir.

Hsu ve ark. [36], çalışmalarında sınıflandırmanın veri madenciliği içerisinde popüler tekniklerden birisi olduğundan bahsetmiştir. Sınıflandırmanın amacı veri kümeleri içinde farklı kesin grupların belirlenmesidir. Bir çok sınıflandırma algoritması yayınlanmıştır ancak bunlar genellikle sayısal ve kategorik veri ile sınırlandırılmıştır. Ancak çoğu veriler karışık, hem sayısal hem de kategoriktir. Bu çalışmada CAVE adı verilen varyans ve entropi tabanlı karışık veri kümelerini işlemeye uygun bir sınıflandırma algoritması sunulmuştur. CAVE'in etkinliği ve faydası bir çok gerçek ve deneysel veri setleri üzerinde kanıtlanmıştır.

Baykasoğlu ve Özbakır [37,38], veri madenciliği ve uygulamaları için sınıflandırma kurallarını keşfetmekte çoklu denklem programlama tabanlı yeni bir yaklaşım sunmuşlardır. Geliştirdikleri metodun performansını iki tıbbi veri kümesinde test etmişlerdir. Yöntemleri, doğruluk, kesinlik, esneklik, etkinlik ve kabiliyet yönünden memnun edicidir.

Canbulut ve ark. [39], dairesel cepli hidrostatik/hidrodinamik yataklarda sürtünme güç kaybına yönelik bir veri madenciliği çalışması yapmışlardır. Çalışmada hidrostatik/hidrodinamik yatakların performansını etkileyen parametre ve parametre düzeyleri, genetik programlama, yapay sinir ağları ve regresyon kullanarak belirlenmiş ve bu yöntemlerin etkinlikleri karşılaştırılmıştır.

Jothishankar ve ark. [40], basılmış devre levhalarının akış lehimlenmesi sırasındaki hataların oluşumuna etki eden faktörlerle birlikte problemi tanımlamışlardır. Önemli olan faktörleri belirlemek için gerekli olan veri madenciliği yaklaşımlarını sunmuşlardır. Sonuçlarını önerilen iyileştirmelerle birlikte ortaya koymuşlardır.

Last ve ark. [41], yeni algı-tabanlı “Otomatik Algı Ağı: APN (Automated Perceptions Network)” adını verdikleri metodu, yüksek derecede gürültülü veri kümelerinden özlü ve yorumlanabilir modellerin otomatik kurulum için geliştirmişlerdir. Geliştirdikleri metodu, iki yarı iletkenin ürünün gelir verilerinde değerlendirerek, veri madenciliği ve

bilgi keşfinde otomatik algıların gelecekte kullanımı ile ilgili mümkün talimatlar tanımlamışlardır.

2003 yılında Hyeon Bae ve ark. [42], birincil amacı mevcut veri madenciliği tekniklerini incelemek ve elektronik-tahmin ile elektronik-teşhis örneklerine bu metotları uygulamak olan bir çalışma yapmışlardır. Teknik olarak yapay sinir ağları, karar ağaçları ve bulanık mantığı kullanmışlardır.

Porzio ve Ragozini [43], yaptıkları çalışmada, yüksek otomasyonlu modern üretim süreçlerindeki büyük veri tabanlarında, veri madenciliği tekniğinin neden uygun olduğunu anlatmışlardır. Büyük ve yüksek boyutlu çevrim-dışı veri kümeleri için bir görsel veri madenciliği stratejisi geliştirmişlerdir. Stratejileri, kullanıcılara bağlantılı interaktif grafiksel araçlar sayesinde daha derin bir süreç anlayışı sağlamaktadır. Görsel veri madenciliği stratejilerini bir endüstriyel süreç durum çalışmasıyla örneklemişlerdir.

Adams [44], 2001 yılında yayınlanan makalesinde gerçek hayattaki üretim firmalarında uygulanan veri madenciliği tekniklerini kısaca anlatmış ve veri madenciliğinin firmalar için çok fazla veri ile başa çıkmakta ve anlamlı sonuçlar çıkararak geleceğe yönelik tahminler yapmakta ne kadar etkin bir yöntem olduğundan bahsetmiştir.

Kusiak ve Kurasek [45], çalışmalarında elektronik kalite hatalarının sebebini belirlemek için veri madenciliği çözüm yaklaşımını ortaya koymaktadırlar.

Lin ve McClean [46], şirketlerdeki ortaklıkların bozulmasının veri madenciliği ile tahminine yönelik bir çalışma yapmışlardır. Çalışmanın temelinde analizde kullanılacak özelliklerin seçimi yer almaktadır. Bunun için farklı teknikler kullanarak veri madenciliği sürecini gerçekleştirmişler ve doğruluklarına göre karşılaştırmalarını yapmışlardır.

Feelders ve ark. [47], “Veri Madenciliğinin Metodolojik ve Pratik Yönleri” adlı çalışmalarında veri madenciliği sürecinin farklı aşamalarını tanımlamışlar. Analizler üzerindeki baskın ilgiye rağmen, veri toplama ve ön-işleme en çok zaman alan aktivitelerdir ve en son başarıda önemli bir etkiye sahiptirler. Başarılı veri madenciliği

projeleri veri madenciliğinde uzmanların varlığına, firma verisine ve ilgili çalışma alanına gereksinim duyar. Tam otomatik veri analizlerinin çekici yönüne rağmen, verinin arkasındaki proseslerin bilgisi veri madenciliğinin birçok güçlüğünden kurtulmayı zorunlu kılmaktadır.

McDonald [48], süreç ürünlerini iyileştirmek için, entegre devre üretiminde ürün iyileştirmesi için veri madenciliği araçlarını kullanmışlardır. Veri madenciliğini, ince levhaların işlem gördüğü sıranın neden olduğu hataları belirlemek için süreç değişkenleri ile ürün arasındaki korelasyonu belirlemede ele almışlardır.

Milne ve ark. [49], geçmiş verileri kullanarak kağıt hatalarının tahmininde veri madenciliği tekniklerinden yararlanmışlardır. Böylece bu hatalarla ilgili doğru adımların atılabileceğini vurgulamışlardır. Kağıt bobinlerinin kalitesinin çevrim-içi olarak tahmin edilmesinde karar ağaçlarını, veri madenciliği aracı olarak da ReMind yazılımını kullanmışlardır. Yaptıkları uygulama sonucunda, iyi bobinlerin %96'sı, kötü bobinlerin ise %88'i doğru olarak tahmin edilmiştir.

1997'de Apte ve Weiss [50], yaptıkları çalışmada veri madenciliği uygulamalarında kural çıkarımı ve karar ağacının kullanımını açıklamaktadır. Desen tanıma, istatistik ve makine öğrenimi alanlarında geliştirilmiş sınıflandırma ve regresyon yöntemleri sembolik ve yorumlanabilir temsillerinde kullanıldığından, veri madenciliği açısından dikkate değer derecede ilgilidir. Sembolik çözümler, temelini teşkil eden mantık ve verinin içinde yer alan karar sınırları dahilinde yüksek bir anlayış derecesini sağlayabilir. Bu durum bu tahmin edici madencilik tekniklerini ticari ve endüstriyel veri madenciliği uygulamalarında cazip hale getirmektedir. Burada bazı ileri gelişmelerle birlikte kural madenciliği yöntemlerini ve karar ağacının ana başlıklarından bazıları temel olarak sunulmuştur.

Kim ve Lee [51], veri madenciliği teknikleri kullanarak, üretim sistemlerinin doğrusal olmayan tahmini ile ilgili çalışma yapmışlardır. Regresyon, hareketli ortalama, üstel düzeltme, sinir ağları ve durum tabanlı sebepleme gibi teknikleri kullanarak, girdi ve çıktı arasındaki ilişkileri belirlemişlerdir. Ayrıca gürültünün çeşitli örneklerinde öğrenme tekniklerinin gücünü ortaya koymuşlardır.

Stepien [52], geliřtirdiđi proses analizi paketiyle, istatistiksel proses kontrol tarafından üretilen bilgiyi yönetmek ve bu bilgileri karar üretmekte kullanmayı amaçlamıřtır. Ayrıca daha istikrarlı bir proses amaçlanmıřtır. Geliřtirilen paket, verilerin toplanabileceđi her tür üretim süreci için uygundur. Ayrıca bu paket bir veri madenciliđi aracı da içermektedir. Geliřtirilen yazılımın performansını test etmek için iki çimento iřletmesinde uygulama yapılmıřtır.

Çalıřma kapsamında uygulanan ve veri madenciliđi süreçlerinden birisi olan veri ön iřleme, veri analizinde önemli ařamalardan bir tanesidir. Çok çeřitli veri kaynaklarında düşük kalitede bir çok bilgi mevcuttur ve bir çok řirket bu bilgilerin yüksek karlı amaçlar için kullanılması amacıyla temizlenmiř formata nasıl çevrileceđi konusunda çalıřmaktadır. Bu amaçlar dođrultusunda iřlenmemiř ham verilerin temizlenmesi önemli bir ihtiyaç olmuřtur. Zhang ve ark.[53], çalıřmalarında veri analizinde ön iřleme sürecinin öneminden bahsetmiřler ve bu alanda bazı geliřmeleri tanıtmıřlardır. Son olarak da gelecek için bazı arařtırma ve geliřmeleri önermiřlerdir. Ařađıda da konu ile ilgili olarak yapılmıř bazı çalıřmalardan bahsedilmiřtir.

Özellik seçimi, örüntü tanıma, istatistik ve veri madenciliđi alanlarında yer alan aktif arařtırma konusudur. Özellik seçiminin temel fikri girdi verilerinde yer alan bazı özellikleri eleyerek yeni bir veri altkütmesi bulmaktır. Özellik seçimi sınıflandırma sonuçlarını önemli bir řekilde arttırmaktadır. Ayrıca görünmeyen özelliklerin modellenmesinde de iyi sonuçlar elde edilir [54].

Chen ve ark. [55], verilen veri kümesini temsil edebilecek en küçük özellikleri seçecek bir özellik seçme metodu sunmuřlardır. Önerilen yaklaşımın temeli kaba kümeler ve nokta esaslı tekniklere dayanmaktadır. İki ařamadan oluřmaktadır. İlk ařamada, veri kümesi nokta esaslı indeksleme ile matrislere ayrılır, ikinci ařamada ise uygun ve yeterli özellikler seçilir. Farklı veri kümeleri ile yapılan deney sonuçları önerilen metodun dođruluk ve etkinliđini göstermektedir.

Çalışma kapsamında kullanılan algoritmalarından bir diğeri olan kaba kümeler literatürde yaygın olarak farklı şekillerde kullanılmaktadır. Bunlardan bir kısmından aşağıdaki paragraflarda bahsedilmiştir.

Birçok gerçek yaşam probleminin kayıtlarında açıkça ifade edildiği gibi, kaba küme (KK) yaklaşımının uygulamada çok yararlı olduğunu kanıtlamıştır. KK yaklaşımının kullanılabileceği ana problemler; verilerin indirgenmesi, veri bağımlılıklarının keşfi, verilerin öneminin tahmini, verilerden karar algoritmalarının oluşturulması, verilerin yaklaşık sınıflaması, verilerdeki benzerlik ve farklılıkların keşfi, verilerdeki örüntülerin keşfi ve neden-sonuç ilişkilerinin keşfini kapsamaktadır [56-58] Bu tez çalışmasında KK yaklaşımı verilerin indirgenmesi sürecinde kullanılmıştır.

Yapılan literatür araştırmasına bağlı olarak KK yaklaşımının tıp, eczacılık, işletme, bankacılık, finans, mühendislik, meteoroloji, görüntü işleme, ses tanımlama, karar analizi, karakter tanımlama ve diğer gerçek yaşam problemlerinde başarıyla uygulandığı görülmüştür [59]. Bu çalışmaların bazılarında izleyen bölümde kısaca bahsedilmiştir.

Thangavel ve Pethalakshmi [60], yaptıkları incelemelerle kaba küme yaklaşımına dayanan boyut azaltımı hakkında bilgi vermişlerdir.

Leung ve ark. [61], yaptıkları çalışmada, bilgi azaltımı sürecinde gerçek değerli verilerin sınıflandırılmasında minimal özellik kümesiyle karar kurallarını seçerek sınıflandırma kuralı keşfeden bir yaklaşım önermişlerdir.

Tripathy ve ark. [62], 2007 yılında “Veri tabanlarından bilgi keşfinde akıllı bir yaklaşım: kaba küme” adlı çalışmalarında veri tabanlarından bilgi keşfinde kaba küme yaklaşımı uygulamalarının güncel durumundan bahsetmişlerdir. Çalışmanın temel amacı, bilgi çıkarımında kaba kümelerin nasıl etkili bir şekilde kullanıldığını göstermektir.

Yine 2007 yılında Zhu’nun [63], yaptığı “İlişkilere bağlı genelleştirilmiş kaba kümeler” adlı çalışmasında, kaba kümelerde isteğe bağlı ikili ilişkiler üzerinde durmuştur.

Çalışmada aynı zamanda alt ve üst yaklaşım operatörleri arasındaki ilişkilerin keşfinden de bahsedilmiştir.

2005 yılında Huang ve ark. [64], yaptığı çalışmada, kaba küme teorisi ve genetik algoritmanın avantajını kullanarak, hata tespiti çıkarım modeli öneren bir yaklaşım ortaya konulmuştur. Entegre yaklaşım 3 ana modülden oluşmaktadır: önışleme ve kesikleştirme, GA tabanlı indirgeme ve indirgenmiş özellik seçimi, kural oluşturma. Çalışmada kaba küme esasına göre çalışan Rosetta paket programı kullanılmıştır. Önerilen metot ile hata tanısında kullanılan şart niteliklerinin sayısı azaltılmış ve %80 sınıflandırma doğruluğu elde edilmiştir.

Busse ve ark. [65], eksik verilerden kural indirgemek için kaba küme yaklaşımını önermişlerdir. Karar tablolarında bazı nitelik değerleri eksik olabilir. Eksik nitelik değerlerinin kayıp (orijinal değer silinmiş) ve ilgisiz (orijinal değer ilgisiz) olmak üzere iki ana durumu incelenmiştir. LEM2 kural indirgeme algoritması kullanılmıştır.

Joseph ve Breault [66], 2001 yılında yaptıkları çalışmada, diyabetik veri tabanlarında veri madenciliği yaklaşımında kaba kümeleri incelemişlerdir. Bu çalışmada Rosetta paket programını tanıtmışlardır. Kaba küme yaklaşımı diyabetik veri tabanlarının analizinde oldukça etkili olmuştur.

Felix ve Ushio [67], 1999 yılında tutarsız ve eksik verilerle kaba kümeleri kullanarak kural indirimini gerçekleştirmişlerdir.

Pawlak [68], 1997’de yaptığı çalışmada kaba küme yaklaşımının temel kavramlarından ve mümkün uygulamalarından özetle bahsetmiştir. Ayrıca gelecekteki araştırma konularına da değinmiştir.

Bu çalışmada kapsamında havacılık sektöründe yapılmış olan veri madenciliği uygulamaları ile ilgili literatür araştırılmış, yapılan çalışmalar incelenmiş aşağıda kısaca bahsedilmiştir.

Gürbüz ve ark. [69], Türkiye’de bir hava yolu işletmesinin parça söküm raporları üzerinde veri madenciliği çalışması gerçekleştirmiştir. Çalışmanın amacı, uçaklarda kullanılan parçaların, herhangi bir arıza oluşmadan önce düzeltici ve önleyici işlemlerin yapılması için ikaz seviyelerinin tespit edilmesine yönelik kural geliştirmektir. Sonuç olarak parçaların ikaz seviyelerini temsil edecek anlamlı kurallar elde edilmiş ve bulunan kurallar doğrulukları ve güvenilirlikleri bakımından test edilmiştir.

Gürbüz ve Özbakır [70], 2008 yılında yaptıkları çalışmada farklı ön işleme ve özellik seçimi tekniklerini, Türkiye’deki bir havayolu şirketine ait parça raporlarına, veriyi anlamak ve temizlemek amacıyla uygulamışlardır. Veri azaltımı için regresyon analizi, aykırılıklar analizi, bağımlılıkları bulma ve kaba kümeler kullanılmıştır. Ayrıca parça ikaz değerinin tahmini için indirgenmiş veriler sınıflandırılmıştır. Bu amaçla, Polyanalyst, SPSS Clementine, Minitab ve Rosetta araçları kullanılmıştır. Sonuç olarak, aykırılıkların analizinden önce ve sonraki veri kümelerinin analiziyle yapılan sınıflandırma işlemlerinden, aykırılıkların analizinin ardından yapılan sınıflandırma ile sınıflandırma doğruluğunun arttığı görülmüştür.

Shur [71], insan hatasından kaynaklanan havacılık risklerini sınıflandırmak amacıyla, kaza ve güvenlik göstergelerini kullanarak yeni bir analitik metot geliştirmişlerdir. Bu çalışma için Taiwan Sivil Havacılık Yönetimi (CAA) tarafından geliştirilen, Uçuş Güvenliği Yönetim Bilgi Sistemi’nden aldıkları veri kümelerini kullanmıştır.

Dessureault ve ark. [72], Amerika’da oluşturulan kaza ile ilgili veri tabanının oluşturulması, veri ambarları ve bu verilerle yapılan analizlerden bahsetmişlerdir.

Solomon ve ark. [73], ölümlü kaza miktarlarını azaltarak, hava yolu trafik güvenliğini artırmada kırmızı-ışık-sinyalle kameraların veri madenciliği ile kontrol edilmesi ile ilgili bir çalışma yapmışlardır. Çalışmada 2000-2003 yılları arasında Maryland ve Washington DC üzerinde gerçekleşen ölümlü uçak kazaları ile ilgili veri kümesini kullanmışlardır. Veri madenciliği tekniklerinden karar ağaçlarını, gelecekle ilgili karar vermede; yapay sinir ağlarını geçmiş verilerle gelecek çıktılar tahmininde; sepet analizini değişkenler arasındaki ilişkinin şiddetinin gösterilmesinde ve k-ortalamalarını kümelerin hedef parametre üzerindeki etkilerini göstermek amacıyla kullanmışlardır.

Elde edilen sonuçlar, pilotların demografik durumlarının, çarpışma şeklinin ve kazanın günün hangi saatinde olduğunun ölümlü kazalar üzerinde etkili olduğunu göstermiştir.

Gürbüz ve ark. [74], yaptıkları çalışmada Türk Hava Yolları'na ait veri tabanında yer alan sefer aksaklıkları ile ilgili olan tehir raporlarını ele alarak bir veri madenciliği çalışması gerçekleştirmişlerdir. Tehir raporları kategorik, sayısal ve metinsel verilerden oluşmaktadır. Çalışmanın amacı, sefer aksaklıklarını azaltmak adına, kalkışlardaki gecikmelerin tehir olarak değerlendirilip değerlendirilmeyeceğine ilişkin bir kural çıkarmaktır. Verilerin analizi için bir veri madenciliği yazılımı olan PolyAnalyst 5.0 programı kullanılmıştır. Elde edilen kurallar, hem doğruluğu hem de güvenilirliği açısından analiz edilmiş ve sonuçlar anlamlı bulunmuştur.

Kulluk ve Gürbüz [75], yaptıkları çalışmada, Federal Havacılık Yönetmelikleri –FAR içinde yer alan Parts 121 ve Parts 135 Federal Düzen Kodlu –CFR, Amerikan hava nakliyecileri tarafından işletilen uçaklara ait kaza verilerine veri madenciliği tekniklerini uygulamışlardır. Çalışmanın amacı, havacılık güvenlik verilerinin analizinde veri madenciliği tekniklerini kullanarak havacılık güvenliğinin iyileştirilmesinde gizli olan sorunlu noktaların belirlenmesini sağlamak ve veri madenciliği tekniklerinin havacılık sektöründe uygulanabilir olduğunu ortaya koymaktır. Çalışma sonunda elde edilen sonuçların kullanıcılar tarafından anlaşılması ve yorumlanması kolaydır. Elde edilen kurallar yardımıyla geleceğe yönelik önlemler alınabilir ve olası kaza riski düşürülebilir. Sonuç olarak veri madenciliğinin havacılık güvenlik sektöründe etkin olarak kullanılabilecek kullanımı kolay, hızlı ve kaliteli sonuçlar üreten bir araç olduğu ortaya konulmuştur.

Tseng ve ark. [76], yaptıkları çalışmada, pilotların dikkatsizlikleri, rahatsızlıkları ve telaşları gibi faktörler ile oluşan kazalar arasındaki ilişkinin tanımlanması amacıyla veri madenciliği tekniklerini kullanmışlardır. Çalışmalarında 2000-2003 yılları arasında Maryland ve Washington DC üzerinde gerçekleşen ölümlü uçak kazaları ile ilgili veri kümesini kullanmışlardır. İlk olarak veri kümesini Kohonen ağları kullanarak gruplandırmışlar daha sonra da karar ağaçları ve yapay sinir ağı modellerini kullanarak kurallar oluşturmuşlardır.

2004 yılında Ananyan ve ark. [77], yaptıkları çalışmada havacılık sektörüne ait geçmiş kaza raporlarını kullanarak bir veri madenciliği çalışması yapmışlardır. Bu çalışmanın, hava yollarına, hava trafik kontrolcülerine, uçak üreticilerine ve havacılık güvenliği için çalışan firmalara yardımcı olacağını belirtmişlerdir.

2003’de, Bineid ve Fielding [78], yaptıkları çalışmada yolcu uçaklarının güvenli uçuşunu geliştirmek amacıyla yapılan tahmin yöntemini açıklamaktadırlar. Yöntem, sonraki düzeltici işlemleri ve maliyetlerini minimize etmek üzere uçak tasarımlarının ilk aşamalarında tehir oranlarını azaltmayı hedefleyen bir amaca sahiptir. Geliştirilen yöntem kısa mesafeli ve uzun mesafeli uçuşlar için olmak üzere iki model içermektedir. Üç saat ve daha az süreli uçuşlar kısa mesafeli, üç saatten fazla süren uçuşlar uzun mesafeli olarak nitelendirilmektedir. Çalışmada istatistiksel eşitlikler çıkarmak üzere her bir uçak sistemi için tehir oranları verisini ve tasarım/performans parametrelerini kullanmışlardır. Bu eşitlikler seçilen tasarım/performans parametresini belirli bir sistem için tehir oranları ile ilişkilendirmektedir. Modeller yüksek güvenilirlik vermiştir ve kavramsal tasarım aşamasında minimum girdi verisini kullanarak uygulanabilir. Tahmin yönteminin geliştirilmesi üç aşamada olmuştur: 1. Veri tabanı geliştirme, 2. Tahmin eşitliklerini türetme, 3. Modelin onaylanması. Geliştirilen yöntem tasarım aşamasında, uçak uçuş güvenliğinin tahminine olanak vermiş ve yüksek doğruluk oranları elde edilmiştir. Sonraki çalışmalarda daha fazla tehir oranları verisi ile eşitlikler geliştirilebilir.

Nazeri ve Zhang [79], 2002’de Ulusal Havacılık Sistemi (NAS) performansını etkileyen ağır hava şartlarının analizi ile ilgili olan “Ağır Hava Şartlarının Hava Sahası Sistem Performansı Üzerindeki Etkilerini Anlamak İçin Havacılık Verilerinin Madenciliği” adlı çalışmalarında; veri hazırlama ve özellik seçimi’nin ne kadar önemli olduğunu göstermişlerdir. Çalışmada hava ve hava trafiği verileri olmak üzere iki tip veri kullanılmıştır. Hava verileri ikili kodlar şeklinde temsil edilmiştir. Hava trafiği için ağır hava şartlarına sahip bir gün, her biri hava ve trafikle ilişkili yapılarla ağır hava bölgelerinin kümesi ile temsil edilmiştir. Bu verilere sınıflandırma, regresyon ve gruplama işlemleri uygulanmıştır. Başlangıç sonuçları ümit verici, teşvik edici olmuştur ve daha sonra sonuçlar geliştirilmiştir. Anlamlı sınıflandırma kuralları üretilmiştir ve hava trafik günleri için oluşturulan grupların Ulusal Havacılık Sistemi performansı ile

uygunluđu görölmüştür. Bu çalışmaya göre; hava trafiğindeki hızlı büyüme ile birlikte, uçuş tehir ve iptalleri de önemli ölçüde artmıştır. Son yıllarda uçuş tehirleri %58, uçuş iptalleri ise % 68 oranında artmıştır. Bundan dolayı uçuş tehir ve iptallerini azaltmak için Ulusal Havacılık Sistemi performansını geliştirmek kaçınılmaz olmuştur. Olumsuz hava koşullarının hava trafiğı ve uçuş tehirleri üzerinde temel etkileri vardır. Uygun stratejiler tehir ve iptallerin geliştirilmesini sağlayabilirler. Bundan dolayı uçuş esnasındaki hava durumunun Ulusal Havacılık Sistemi üzerindeki etkilerinin analizi zorlu ve ilginç bir veri madenciliğı uygulamasıdır.

Famili ve ark.nın [80], 1999 yılında yaptıkları “Uçak Parça Değişiminin Tahmininde Veri Madenciliğı” adlı çalışmalarında kısaca uçak motoru parçalarındaki problemleri tahmin etmek için model oluşturma üzerinde odaklanmışlardır. A320 tipi uçaktan alınan 19 tipte raporlar kullanılmıştır. Raporlar sayısal ve sembolik olarak 20 ile 150 arasında parametre içermektedir. Çalışmaya göre, yolcu uçakları gibi modern ve sensör-düzenli işlem ve bakıma sahip sistemler sayısal ve sembolik çok fazla veri oluşturmaktadır. Parçalardaki problemleri tahmin etmek için bu verilerin kullanılması tehirlerin azaltılması ve güvenliğin artırılması için yol gösterici olabilir. Çok büyük verilerden modelleri öğrenmek için bazı veri madenciliğı teknikleri kullanılmaktadır. Fakat uçağın bakım işlemi sırasında elde edilebilen verilerden istenilen modelin çıkarılması için bu tekniklerin kullanımı oldukça ilginçtir. Gösterilmek istenen güçlükler şunları içermektedir: veri toplama, veri etiketleme, veri ve model bütünleşmesi (entegrasyonu) ve model değerlendirme. Bu çalışma bu sorunları gösteren bir yaklaşım sunmaktadır. Aynı zamanda uçak parçalarının değişikliğı problemlerinin tahmini için model oluşturmak için bu yaklaşımın uygulama sonuçları da rapor edilmiştir.

3. BÖLÜM

VERİ TABANLARINDAN BİLGİ KEŞFİ-VERİ MADENCİLİĞİ

3.1. Giriş

Günümüz dünyasında işletmeler başarılı olabilmek için yeni ekonominin ve iş modellerinin kurallarına uygun hareket etmek durumundadırlar. Bilgisayar ve iletişim teknolojilerindeki gelişmeler çok daha fazla verinin hızlı toplanmasına, depolanmasına, işlenmesine ve bilgiye dönüştürülüp yeniden istenilen noktalara iletilmesine olanak sağlamıştır. Yeni teknolojilerden önemli ölçüde etkilenen ve dijital ekonomi olarak da adlandırılan bu yeni ekonominin kuralları gereği “bilgi” ve “zaman” boyutlarının önemi çok daha artmış ve dolayısıyla işletmelerin karar verme süreçlerinde doğru ve anlamlı bilgiye dayalı hızlı karar alma gereği her zamankinden daha fazla ön plana çıkmıştır.

Bilgisayar teknolojilerindeki hızlı gelişmeler ve verilerin dijital ortamda saklanmaya başlanması ile birlikte, yeryüzündeki bilgi miktarının her 20 ayda bir kendini iki katına çıkardığı günümüzde veri tabanlarının sayısı da benzer, hatta daha yüksek bir oranda artmaktadır. Yüksek kapasiteli işlem yapabilme gücünün ucuzlamasının bir sonucu olarak, veri saklama hem daha kolay olmuş, hem de verinin kendisi de ucuzlamıştır [81].

İşlemcilerin hızlanması, disk kapasitelerinin artması, bilgisayar ağlarındaki ilerleme sonucunda her bir bilgisayarın başka bilgisayarlardaki verilere ulaşma olanağı, bilgisayarların çok büyük miktarlardaki verileri saklayabilmesine ve daha kısa sürede işleyebilmesine imkan tanımaktadır.

Bilgisayar sistemlerindeki bu gelişmeyle birlikte kullanımı da bu ölçüde yaygınlaşmaktadır. Bununla birlikte işletmelerde üretilen sayısal bilgi miktarı artmış, buna paralel olarak veri tabanları daha fazla veriyi saklayabilecek boyutlara ulaşmış ve bilgisayar sistemlerindeki gelişme ile veriye ulaşma daha da kolaylaşmıştır. Bu sayede doğru ve daha detaylı bilgiye ulaşılması mümkün hale gelmekte fakat başka bir sorun ortaya çıkmaktadır. Bu sorun oluşan bu büyük sayısal veri yığınlarının yönetilmesi ve anlamlı hale getirilmesi sorunudur.

Bilgisayar sistemleri ile üretilen bu veriler tek başlarına değersizdirler. Çünkü çıplak gözle bakıldığında verilerin bir anlam ifade etmediği söylenebilir. Bu veriler belli bir amaç doğrultusunda işlendiği zaman anlamlı hale gelmektedir. İşte ham veriyi bilgiye veya anlamlı hale dönüştürme işi “**veri madenciliği**” ile yapılabilmektedir.

Büyük miktarlarda ve oldukça hızlı toplanan verilerin çeşitli analizler sonucunda anlamlı bilgilere dönüştürülmesi noktasında “veri madenciliği” süreci devreye girmektedir. Veri madenciliği tanımları incelendiğinde, bu tanımlardan ortak olan unsurlardan ilki “çok fazla” miktarlarda verinin veri ambarlarında tutulması ikincisi ise bu verilerden “anlamlı” bilgiler elde edilmesidir.

Veri madenciliği aslında, bilgi teknolojilerinin doğal gelişim sürecinin sonucu olarak da değerlendirilebilir. Veri tabanı sistemlerinin gelişim şeması Şekil 3.1’de gösterilmiştir [82]. Veri tabanı sistemlerinin gelişimi üç ayrı fonksiyon ile değerlendirilmektedir [1]:

- Veri toplama ve veri tabanı yaratma
- Veri yönetimi (veri depolama, veri düzeltme, bilgi keşfi)
- Veri analizi ve kavrama (veri ambarcılığı ve veri madenciliği gerektirir)

1. Veri Toplama: Veri tabanı ve bilgi teknolojileri 1960’lardan günümüze kadar dosya düzenleme sistemlerinden daha karmaşık ve güçlü olan veri tabanı sistemlerine doğru sistematik bir gelişim göstermiştir. Veri tabanı sitemlerindeki araştırma ve geliştirmeler 1970’lerden beri hiyerarşik yapılardan ilişkisel yapılara doğru ilerlemeye başlamıştır. İlişkisel veri tabanı sistemlerinin getirdiği

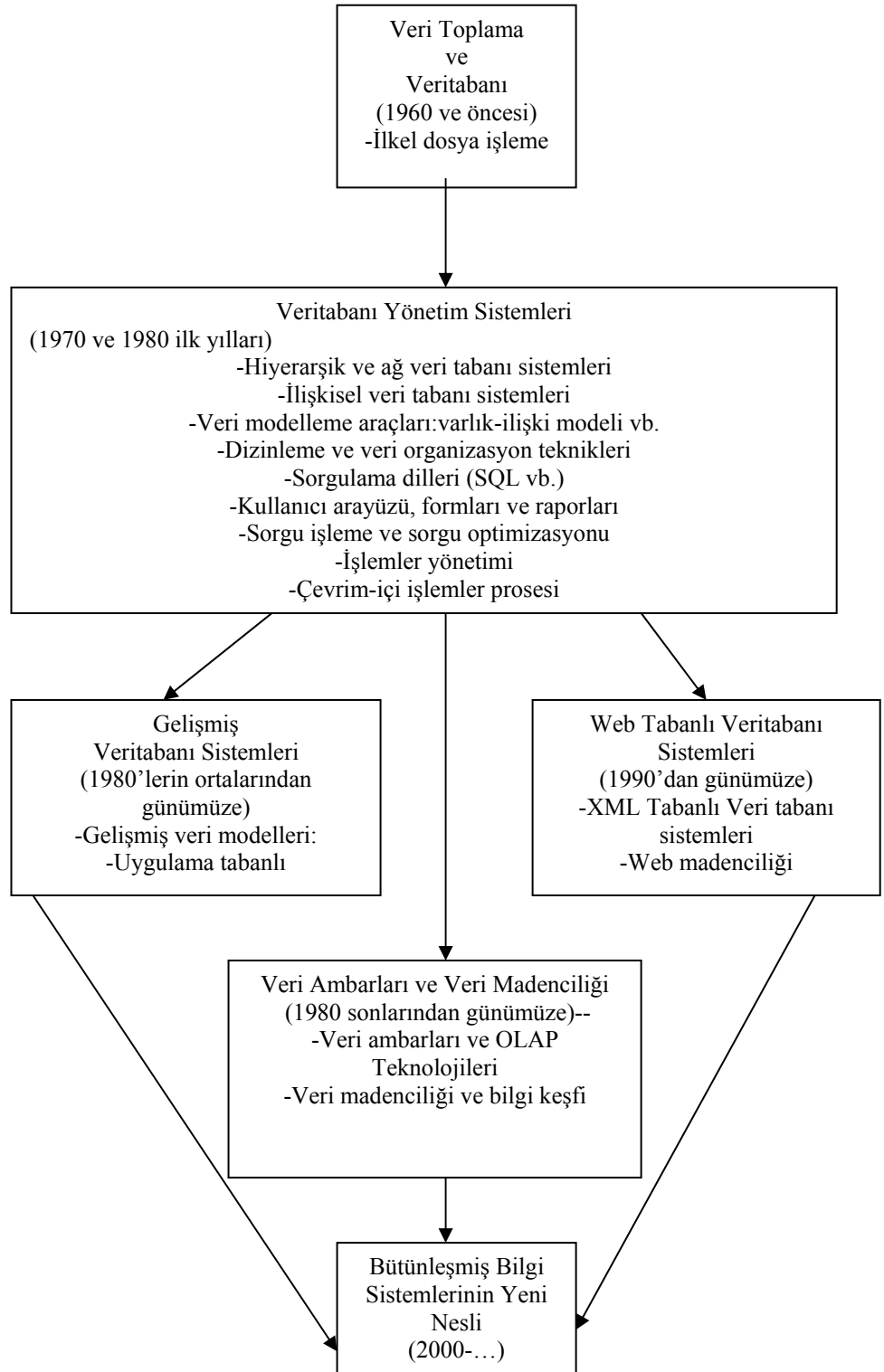
veri modelleme araçları, indeksleme ve veri organizasyon teknikleri ile tablolar birbirleri ile ilişkili olarak saklanmaya başlanmıştır. Bunlara ek olarak kullanıcılar, sorgulama dillerini, kullanıcı ara-yüzlerini kullanarak daha esnek ve daha rahat veri erişimi sağlamaya başlamışlardır [1].

Dünyada 1960'lı yıllarda veri toplama sistemleri, 1970'li yıllarda ise ilişkisel veri tabanları kullanılmaya başlanmıştır. 1980'lerde ise ilişkisel veri tabanlarının popüler olmaya başladığı görülmektedir. 1990 ve 2000'li yıllarda, bilgisayar sistemlerindeki teknolojik gelişmelere paralel olarak ilişkisel veri tabanlarında tutulan veri depoları kullanılmaya başlamıştır [1].

Çok büyük ölçekli veriler, farklı alanlardaki büyük ölçekli veri tabanları içlerinde değerli verileri bulunduran bir veri madeni gibi düşünülebilir. Bu büyüklükteki verilerin analizi, bu analiz sonucunda daha anlamlı bilgi elde etme ve elde edilen bilgiyi yorumlama işi insan yeteneğini ve ilişkisel veri tabanlarının yapabileceklerini aşmaktadır. Bu ihtiyaçların sonucunda otomatik ve akıllı veri tabanı analizi için yeni kuşak teknikler doğmuştur. Bu teknikler, veriyi akıllı ve otomatikleşmiş şekilde işe yarar bilgiye dönüştürebilecek yapıda olmalıdır. Tüm bunların sonucunda veri madenciliği cevap olarak sunulmuş ve giderek önemini artıran bir araştırma alanı haline gelmiştir [1].

2. **Veri Yönetimi:** Veri madenciliğinden veri tabanlarından bilgi keşfi diye de bahsedilir. Veri tabanlarından bilgi keşifleri ile yüksek-seviyeli bilgiler veritabanlarındaki ilişkili verilerden çekilmeye ve farklı açılardan değerlendirilmeye başlanmıştır. Büyük ölçekli veri tabanlarından bilgi çıkarmak, bir çok araştırmacı tarafından veri tabanı sistemlerindeki bir araştırma başlığı olarak fark edilmiştir. Veri madenciliği ile keşfedilen bilgi, bilgi yönetimi, karar mekanizmaları, kontrol sistemleri ve çevrim-içi veri takibi gibi birçok değişik uygulamada kullanılabilmektedir [1].

Veri madenciliği; veri ambarlarındaki tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarmak, bunlara karar vermek ve eylem planını gerçekleştirmek için kullanma sürecidir. Bu noktada, veri madenciliği



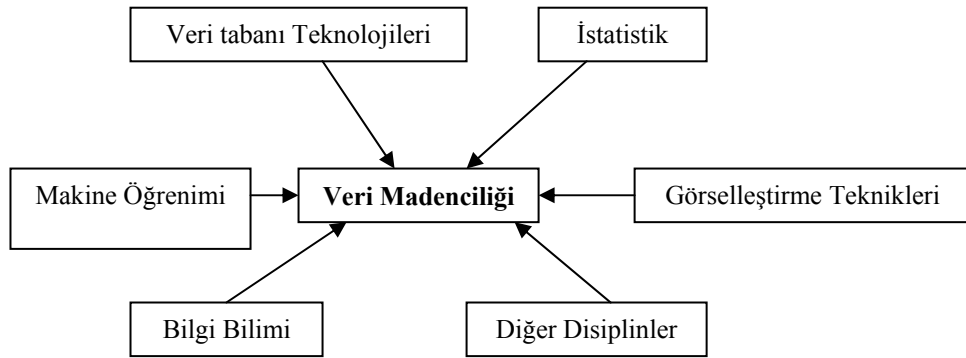
Şekil 3.1. Veri tabanı sistemlerinin gelişimi [1].

kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli olan bilgileri sağlamaya yarayan bir araçtır.

3. Veri Analizi: Veri madenciliği uygulamalarında kullanılan ve büyük miktarlarda verilerin tutulduğu veri ambarı projeleri de benzer şekilde son yıllarda önem kazanan projelerdir. Bilişim sektöründe veri ambarı ve veri madenciliğiyle ilgili gerek donanım gerek yazılım ürünleri üstünde çalışan işletmeler son yıllarda artış göstermektedir.

Veri madenciliği, istatistik, makine öğrenmesi, veri tabanı yönetimi gibi birçok alanla kesişimi olan çok disiplinli bir alandır [83]. Şekil 3.2 veri madenciliği ile disiplinler arası ilişkiyi göstermektedir.

Veri madenciliği astronomi, biyoloji, finans, pazarlama, sigorta, tıp ve birçok başka dalda uygulanmaktadır. Bununla birlikte günümüzde veri madenciliği teknikleri özellikle işletmelerde çeşitli alanlarda başarı ile kullanılmaktadır.



Şekil 3.2. Veri madenciliği ile disiplinler arası ilişki.

3.2. Veri Tabanlarından Bilgi Keşfi

Son yıllarda, veriyi toplama ve saklama kapasitesinde çok ani bir büyüme gerçekleşmektedir. Sonuç olarak da bir bilgisayarın işleyebileceği veriden daha fazlası üretilmektedir. Büyüme işlevleri cinsinden ifade edersek, veri saklama kapasitesi her 20 ayda bir tahmini ikiye katlanmaktadır. Buna karşılık ise, aynı peryotta, Moore kanununa

göre hesaplama gücü iki kat daha az büyümektedir. Bu aradaki fark, veriyi yakalama ve saklama oranının onu işleme ve kullanma yeteneğimizi çoktan geçtiğini göstermektedir [3] .

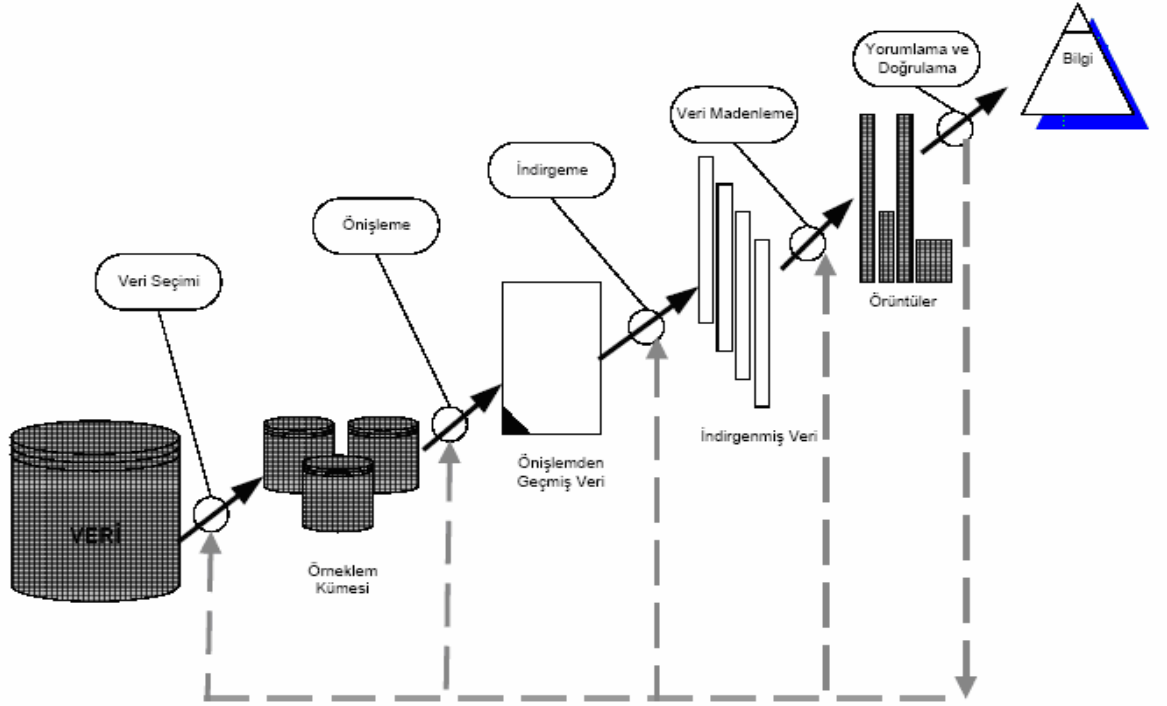
Bir başka deyişle, bir kısım veri nihai olarak bir daha hiç erişilmemek ve işlenmemek üzere saklanabilmektedir. Verinin biriktirilmesi ile birlikte onu yorumlamada ve kavramada yetersiz kalmamız, veri tabanı analizi için yeni nesil araçlara ve tekniklere ihtiyaç duyulduğunu ortaya çıkarmıştır. Bilgisayar sistemleri ile üretilen bu veriler özellikle veri tabanlarının bilgiyi sadece saklamak için tasarlandığı düşünüldüğünde tek başlarına anlamlı değildirler. Bu veriler ancak belli bir amaç doğrultusunda işlendiği zaman anlamlı hale gelmektedir. Bu nedenle veri tabanlarından bilgi keşfi (VTBK) önemli ve aktif bir araştırma alanı olmuştur. Buna bağlı olarak ham veriden bilgiye veya yararlı veriye ulaşabilmek için veri madenciliği çalışmaları günümüzde ön plana çıkmıştır.

Veriden anlamlı ilişkiler ve örüntüler çıkarma sürecine literatürde, veri madenciliği, bilgi çıkarımı, bilgi keşfi, veri arkeolojisi ve veri örüntü işleme gibi isimler verilmektedir [84]. Veri Madenciliği, bilgi keşfi sürecinin merkezini oluşturur ve genellikle bu iki terim birbirini yerine kullanılır.

Veri tabanlarından bilgi keşfi, veriden gerçekten anlamlı ve yararlı bilginin çıkarıldığı süreç olarak tanımlanmaktadır. Verinin nasıl saklanması, algoritmaların büyük veri kümelerine nasıl uyarlanması gerektiği, sonuçların nasıl yorumlanacağı ve görselleştirileceği, insan-makine etkileşiminin nasıl modelleneceği sorularının cevabını aramaktadır. Veriler bir veritabanına kaydedilir. Bilgi tabanındaki bilgi keşfi sürecinden sağlanan bilgi basit hesapla elde edilemez. Bilgi kullanıcı için yeni ve avantajlı olmalı aynı zamanda bazı doğruluk dereceleriyle yeni veriye uygulanabilmelidir. Bu anlamda veri madenciliği bu sürecin sadece bir kısmını oluşturmaktadır. Bilgi keşfi süreci; makine öğrenmesi, örüntü tanıma, istatistik, yapay zeka ve veri görselleştirmesi kavramlarının bir araya gelmiş şeklidir. Veri madenciliği, bilgi keşfi sürecinde makine öğrenmesi, örüntü tanıma ve istatistik tanımlamalarını bir arada kullanan bir aşamadır. Bilgi keşfi sürecinde amaç, büyük veri kümelerindeki düşük seviyedeki veriden yüksek seviyede bilgi çıkarımı sağlamaktır [84].

3.2.1. Veri Tabanlarından Bilgi Keşfi Süreci

Veritabanındaki Bilgi Keşfi süreci birkaç adımdan oluşan etkileşimli ve iteratif bir süreçtir. Şekil 3.3’de veri tabanlarından bilgi keşfi süreci gösterilmektedir. Genel olarak bu süreçler aşağıdaki gibi açıklanabilir [85].



Şekil 3.3. Veri tabanlarından bilgi keşfi süreci.

- 1- *Problemin Tanımlanması*: Bilgi keşfi projesinin amaçları ile hedefleri belirlenmeli ve tanımlanmalıdır. Bu amaçlar uygulanabilir olmalıdır. Aynı zamanda kullanılacak olan veriler de tanımlanmalıdır.
- 2- *Verilerin Hazırlanması*: Bu süreç kısaca verinin toplanması, temizlenmesi, seçilmesi ve dönüştürülmesi adımlarını kapsar.
- 3- *Veri Madenciliği*: Veri modellerinin çıkarılması için akıllı metotların uygulandığı önemli bir süreçtir. Veri madenciliği sadece modelleme ve keşif algoritmalarına ilişkin uygulamayı içeren ve ham verilerden değerli bilgiye kadar uzanan veri tabanlarında bilgi keşfi sürecinin özel bir evresini oluşturur [86]. Sınıflandırma, kural ve ağaçları, regresyon, eşleştirme, bağımlılık, kümeleme ... gibi sorguların işletilmesidir [3]. Veri madenciliği evresi, hem seçim yapmayı ve modelleme ile keşif tekniklerinin uygulamalarını, hem de

parametrelerin optimal değerlerine ayarlanmasını kapsamaktadır. Genellikle, aynı veri madenciliği hedefi için farklı birkaç teknik vardır ve veri madenciliği evresi, hem farklı tekniklerin uygulamalarını hem de sonuçların yerel değerlendirmesini ve karşılaştırmasını içerir [86].

- 4- *Yorumlama ve Doğrulama*: Elde edilen sonuçların değerlendirilmesi, uygulanması ve bilginin sunumu aşamalarından oluşmaktadır. Değerlendirme aşamasında, bilgiyi temsil eden doğru ve dikkate değer örnekler tanımlanır, modeli oluşturmada kullanılmayan bağımsız veri setleri üzerinde modelin doğruluğu test edilir. Modelin hassaslığı tayin edilerek, kullanılabilirliği için modele pilot çalışma uygulanır. Daha sonra, bir tahminleme modeli, yeni olaylar ve durumlar için sonuçları tahmin etmek amacıyla kullanılır. Böylece tahmin organizasyonel davranışları değiştirmek için kullanılacaktır. Son olarak da bilginin kullanıcılara sunumu için bilgi gösterimi ve bilgiyi görselleştirme teknikleri kullanılır.

Veri tabanlarından bilgi keşfi literatürde veri madenciliği ile eş anlamlı olarak kullanıldığından her iki süreçte benzer adımlar takip edilmektedir. Bu süreçler daha ayrıntılı olarak bölüm 3.3.6 “Veri madenciliği süreci” nde anlatılmıştır.

3.3. Veri Madenciliği

3.3.1. Veri Madenciliği Nedir?

Tutulması gereken verilerin miktarının ve çeşitliliğinin artması, analizlerin daha hızlı yapılmasının gerekliliği ve sonuçta anlamlı ve eyleme yönelik bilgiler ortaya çıkarılması, yeni ekonominin ve değişen piyasalardaki yoğun rekabet ortamının bir gereğidir. Milyonlarca bitlik verilerin depolandığı veri ambarlarının tutulduğu donanımlar ve bunlardan çok hızlı bir şekilde anlamlı bilgiler elde edebilmek için kullanılabilecek yazılımlar yeni teknolojinin sunduğu olanaklardır [87].

Basit bir tanım yapmak gerekir ise veri madenciliği, büyük ölçekli veriler arasından bilgiye ulaşma, bilgiyi madencilik işidir. Ya da bir anlamda büyük veri yığınları

içerisinden gelecek ile ilgili tahminde bulunabilmemizi sağlayabilecek bağıntıların bilgisayar programı kullanarak aranmasıdır [1].

Veri madenciliği sürecinde, kümeleme, veri özetleme, sınıflama kurallarının öğrenilmesi, bağımlılık ağlarının bulunması ve değişkenlik analizi gibi farklı bir çok teknik kullanılmaktadır.

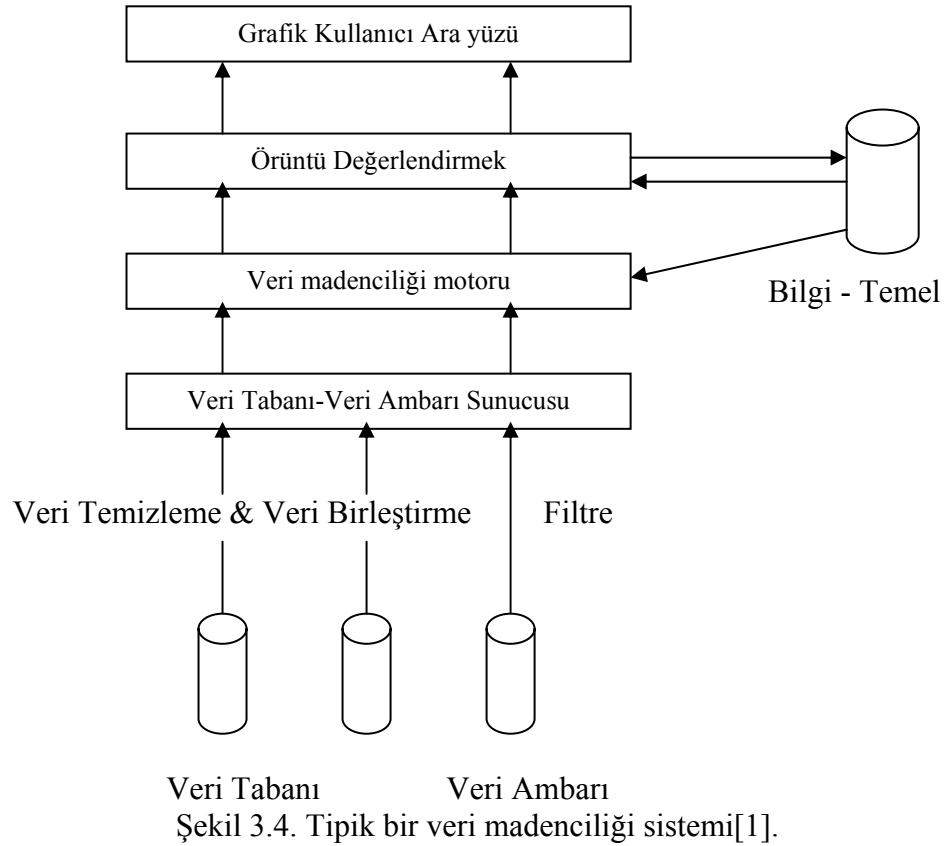
Veri madenciliği ile büyük veri yığınlarından oluşan veri tabanları sistemleri içerisinde gizli kalmış bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, veri tabanları teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır.

Veri madenciliği kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır. Veri madenciliği; analiste iş yapma aşamasında oluşan veriler arasındaki şablonları ve ilişkileri bulması için yardım etmektedir.

Tipik bir veri madenciliği sisteminin yapısı, Şekil 3.4’de gösterilen temel bileşenlere sahiptir [1,88].

- **Veri tabanı, veri ambarı veya diğer bilgi depoları:** Veri tabanlarının, veri ambarlarının, evrakların veya diğer bilgi depolarının biri veya bunlardan oluşan bir kümedir. Veri temizleme ve veri bütünleştirme teknikleri veriye uygulanabilir.
- **Veri tabanı ya da veri ambarı sunucusu:** Veri tabanı ya da veri ambarı sunucusu, kullanıcının veri madenciliği isteğine bağlı olarak ilgili veriyi getirmekle sorumludur.
- **Bilgi tabanı:** Ortaya çıkan örüntünün ilginçliğini değerlendirmek veya aramayı yönlendirmek için kullanılan alan bilgisidir.
- **Veri madenciliği motoru:** Veri madenciliği sisteminin temeli olan ve ideal olarak karakterize etme, birleştirme, sınıflandırma, kümeleme analizi, değerlendirme ve sapma analizi gibi işlerde fonksiyonel modüllerin bir kümesini içerir.

- **Örüntü değerlendirme modülü:** Aramayı ilginç örüntülere odaklamak için tipik olarak ilginçlik ölçümü uygulayan ve veri madenciliği modülleri ile etkileşimde olan bir bileşendir.
- **Grafiksel kullanıcı ara yüzü:** Bu modül kullanıcılar ve veri madenciliği sisteminin iletişimini sağlar.



3.3.2. Veri Madenciliği Neden Gerekli?

Günümüzün anahtar sözcükleri değişim ve yenilik, kaynağını stratejik olarak kullanılabilir hale gelen bilgiden almaktadır. İşletmeler kendi ürettikleri (zaten sahip oldukları) ve dışarıdan edindikleri bilgiyi anlamlı biçimde kullanabilmelerinin, hayatta kalabilmeleri için zorunlu olduğunu anlamışlardır. Doğru bilgiyi ihtiyaç duyduğu anda ve yerde anlaşılır bir şekilde, doğru kişilere ulaştırabilmek, günümüz bilgi işlem birimlerinin temel görevi haline gelmiştir [89].

Veri madenciliğinin bilgi endüstrisinde son yıllarda çektiği büyük ilginin temel sebebi büyük miktardaki verinin bulunabilirliği ve bu veriyi kullanışlı bilgiye çevirme ihtiyacıdır. Kazanılan bilgi iş yönetimi, üretim kontrol ve pazar analizinden mühendislik tasarımı ve bilgi keşfine kadar uzanan bir uygulama alanında kullanılabilir [1].

Örneğin eskiden süpermarketteki kasa basit bir toplama makinesinden ibaretti. Müşterinin o anda satın almış olduğu malların toplamını hesaplamak için kullanılırdı. Günümüzde ise kasa yerine kullanılan satış noktası terminalleri sayesinde bu hareketin bütün detayları saklanabilmektedir. Saklanan bu binlerce malın ve binlerce müşterinin hareket bilgileri sayesinde her malın zaman içindeki hareketleri ve eğer müşteriler bir müşteri numarası ile kodlanmışsa bir müşterinin zaman içindeki verilerine ulaşmak ve analiz etmek mümkün olmaktadır [1].

Süpermarket örneğimizde veri analizi yaparak her mal için bir sonraki ayın satış tahminleri çıkarılabilir; müşteriler satın aldıkları mallara bağlı olarak gruplanabilir; yeni bir ürün için potansiyel müşteriler belirlenebilir; müşterilerin zaman içindeki hareketleri incelenerek onların davranışları ile ilgili tahminler yapılabilir. Binlerce malın ve müşterinin olabileceği düşünülürse bu analizin gözle ve elle yapılamayacağı, otomatik olarak yapılmasının gerektiği ortaya çıkmaktadır.

3.3.3. Makine Öğrenimi ve Veri Madenciliği

İnsan ve makine uzman, veri madenciliğinde birbirini tamamlayan en önemli bileşenlerdendir. İnsan uzman, veri tabanının hazırlanması, problemlerin tanımlanması ve hedeflerin belirlenmesinde aktifken; makine uzman verileri, hedeflerle uyuşan örüntüleri saptamak amacıyla taramak ve bu örüntülere uygun kuralları öğrenmekten sorumludur [90].

Veri madenciliğinde olduğu gibi makine öğrenmesinde de, analize konu olan alana ait veriler, örnekler sisteme iletilir ve makine uzman ya da programın bu verilere ait örüntüyü ya da yapıyı öğrenmesi istenir. Makine öğrenmesi yöntemleri, büyük veri setlerinde rahatlıkla kullanılabilen ve daha çok bilgiyi kolaylıkla işlemeye yatkın güçlü yöntemlerdir [91].

Veri madenciliği ve makine öğrenmesi yöntemlerinin temel özellikleri şöyle özetlenebilir [90]:

- Her iki yöntem de kuramsal olmaktan çok deneye dayalı öğrenme/sınıflandırma problemleriyle ilgilenir.
- Verideki yapısal örüntüleri bulmak ve tanımlamak, bu veriyi açıklayacak aracı yani algoritmayı ve ilişkili kural kümelerini bulmak ve bu bilgiyi tahmin veya sınıflandırma amacıyla kullanmak makine öğrenmesinin asıl amacıdır.
- Veri seti örneklerden oluşmaktadır. Her iki yöntemin çıktısı, yeni örnekler üzerinde yapılan tahmin/sınıflandırma ya da öğrenilen kurallardır. Elde edilen bilginin, tanım alanı bağlamında kuramsallaştırılması önemli bir hedeftir.
- Makine öğrenmesi ve veri madenciliği yöntemlerinde, veriden bilgi çıkarma yarım ya da tam otomatik süreçlerdir.
- Karar destek sistemleri açısından, her iki yöntemde de, makine uzmanının bilgi işleme kapasitesine dayanılarak, insan uzmanının bu alandaki eksilikleri giderilmek ve rekabet avantajı sağlamak amaçlanmaktadır.

3.3.4. İstatistik ve Veri Madenciliği

Veri madenciliğiyle ilgili yazılım ürünleri ve uygulamalarında görülmektedir ki veri madenciliği esasen istatistiğin kullanıldığı bir tekniktir. Dolayısıyla işletmelerde önemli bir kullanım alanı bulması çok doğaldır [87].

Veri madenciliğinde vurgulanan unsurlar istatistiğin tanımı içinde zaten yer almaktadır. İstatistik verilerin toplanması, sınıflandırılması, özetlenmesi, grafik ve tablolarla sunulması, analiz edilerek ana kütle hakkında anlamlı bilgiler elde edilmesi ve yorumlar yapılmasıdır. Veri madenciliğinde ulaşılmak istenen amaç aslında istatistik biliminin amacı ile aynı doğrultudadır: verilerden bilgiyi keşfetmek. Zaten veri madenciliğinde kullanılan temel aracın istatistiksel yöntemler olduğu birçok tanımda ve uygulamada vurgulanmaktadır. Her ikisinde de temel olan öğeler veri ve bilgidir. Bu nedenle birbiriyle oldukça örtüşen konulardır. İstatistiğin amacı ana kütle hakkında anlamlı

bilgiler elde etmek ve yorum yapmak; veri madenciliğinin amacı da anlamlı bilgiler elde etmek ve bunu eyleme dönüştürecek kararlar için kullanmaktır [87].

3.3.5. Veri Tabanı ve Veri Madenciliği

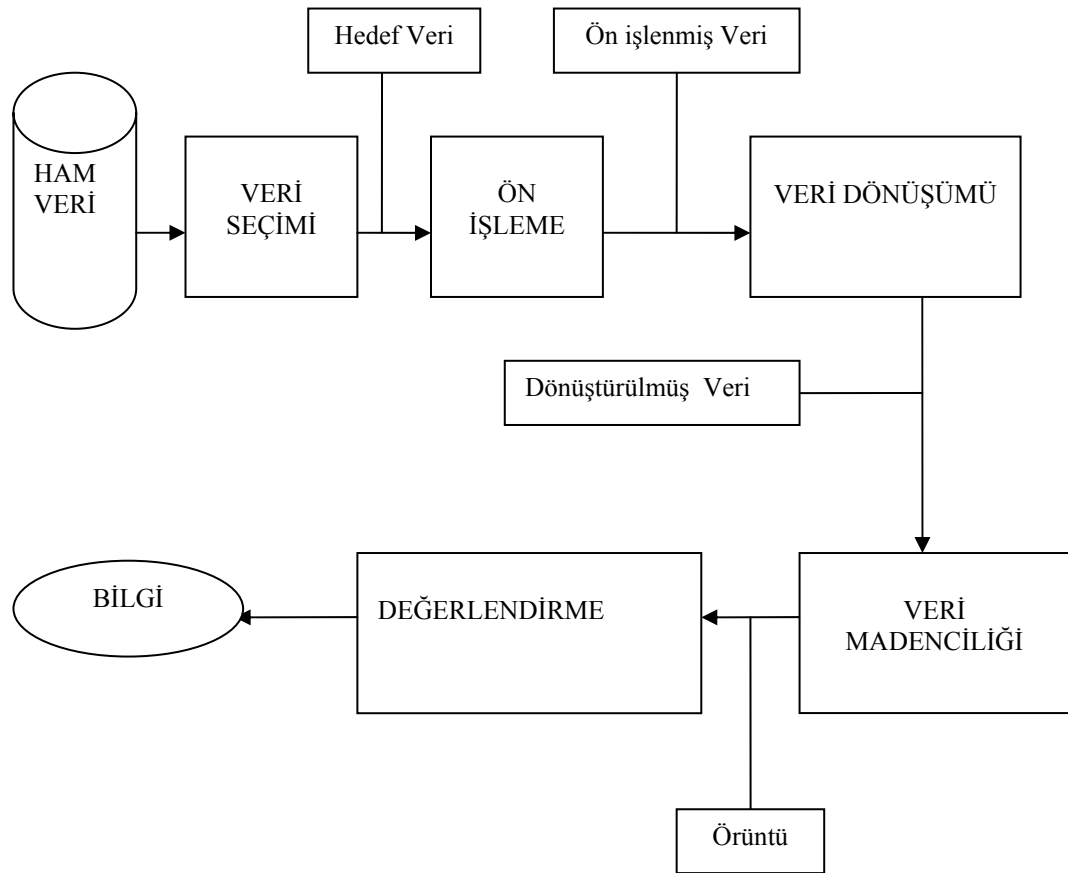
Veri madenciliği sorgularına girdi sağlamak amacıyla veri tabanı kullanılmaktadır. Veri tabanındaki sorgu cümlecikleri, veri madenciliğinin istediği örneklem kümesini elde etmek amacıyla kullanılmaktadır. Özellikle ilişkilendirme sorgusunda fazla miktarda veri tabanı sorgusu yapmak gerekmektedir. Veri madenciliği, veri tabananından farklıdır. Çünkü veri tabanında yer alan örüntüler için sorgular çalıştırılırken, veri madenciliğindeki sorgular genelde keşfe dayalı ve ortada olmayan örüntüleri ortaya çıkarmaya yöneliktir.

3.3.6. Veri Madenciliği Süreci

Veri madenciliği; veri ambarlarındaki tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarmak, bunları karar vermek ve eylem planını gerçekleştirmek için kullanma sürecidir. [87]. Veri madenciliği süreci kısaca şekil 3.5’de gösterilmiştir.

Veri madenciliği sürecinde izlenmesi gereken temel aşamalar şu şekildedir.

- Problemin Tanımlanması,
- Verilerin Hazırlanması,
- Modelin Kurulması ve Değerlendirilmesi,
- Modelin Kullanılması,
- Modelin İzlenmesi



Şekil 3.5. Veri madenciliği süreci [92].

3.3.6.1. Problemin Tanımlanması

Veri madenciliği çalışmalarında başarılı olmanın ilk şartı, uygulamanın hangi amaç için yapılacağına açık bir şekilde tanımlanmasıdır. İlgili işletme ya da kurum problem üzerine odaklanmış olmalı ve problem açık bir dille ifade edilerek, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Ayrıca yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir.

3.3.6.2. Verilerin Hazırlanması

Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır. Bu durum verilerin hazırlanması ve modelin kurulması aşamaları için, bir analistin veri keşfi sürecinin

toplamı içerisinde enerji ve zamanının % 70'ini harcamasına neden olmaktadır. Verilerin hazırlanması aşaması kendi içerisinde toplama, seçim, birleştirme ve temizleme, eksik verilerin telafisi ve dönüştürme adımlarından meydana gelmektedir.

- *Toplama:* Tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımdır. Çeşitli iç ve dış kaynaklardan gerekli veriler temin edilir, temsili verilere karar verilerek farklılıklar kodlanır. Bu kısımda, bütün bilgi veri madenciliğinin sonraki adımlarında kullanılacak girdiyi sağlamlaştıran basit bir veri kaynağına katılır. Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olacaktır. Bu nedenlerle, iyi sonuç alınacak modeller ancak iyi veriler üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir.
- *Seçim:* Bu adımda kurulacak modele bağlı olarak veri seçimi yapılır. Örneğin tahmin edici bir model için, bu adım bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır.
- *Birleştirme ve Temizleme:* Veri temizleme işlemi, veri uyumsuzlarına, aykırı verilere, gürültülü ve yanlış verilere, kayıp verilere ve belirsizliklere karar verme ve kontrol altına alma aşamasıdır. Gerçek, işlenmemiş veriler genellikle bozulmuş yani formatsız bir yapıya sahiptir. Kayıp değerler, bilinmeyen değerler veya yanlış değerler veri toplama ya da iletimi sırasında oluşabilecek hatalardandır. Verilerdeki bu tip hatalar ya da sonuçlardaki tutarsızlık, veri madenciliği uygulamalarını imkansız hale getirmekte ya da uygun olmayan sonuçların çıkmasına neden olmaktadır. Bundan dolayı, hatalı verilerle veri temizleme evresinde ilgilenmek gerekir [86]. Bu adımda farklı kaynaklardan toplanan verilerde bulunan ve bir önceki adımda belirlenen sorunlar mümkün olduğu ölçüde giderilerek veriler tek bir veri tabanında toplanır. Yanlış olduğu düşünülen veriler çıkarılır.
- *Eksik Verilerin Telafisi:* Geniş veri tabanlarında büyük bir ihtimalle, ölçülemeyen ya da doğru olmayan veriler içeren alanlar bulunacaktır. Modelin ve eldeki verinin durumuna göre içinde eksik alanlar barındıran ölçüm veya

numuneleri çıkarmak veya boş alanları uygun değerlerle doldurmakla bu sorun aşılabılır.

- *Dönüştürme*: Analizi kolaylaştırmak için verilerin ilk tanımlandığı şekliyle değil, yeni bir formatta modele dahil edilmesi mümkün olabilir. Ayrıca modelde kullanılan algoritma, verilerin gösteriminde önemli rol oynayacaktır. Örneğin bir uygulamada bir yapay sinir ağı algoritmasının kullanılması durumunda kategorik değişken değerlerinin evet/hayır olması; bir karar ağacı algoritmasının kullanılması durumunda ise örneğin gelir değişken değerlerinin yüksek/orta/düşük olarak gruplanmış olması modelin etkinliğini artıracaktır. Bu durumda sürekli verilerin kategorize edilmesi sıkça karşılaşılan bir durumdur [93] .

3.3.6.3. Modelin Kurulması ve Değerlendirilmesi

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılncaya kadar yinelenen bir süreçtir. Model kuruluş süreci denetimli (Supervised) ve denetimsiz (Unsupervised) öğrenimin kullanıldığı modellere göre farklılık göstermektedir.

Modeli değerlendirme sürecinde, genelde iki kriter söz konusudur, modelin doğrulanması ve anlaşılabilirliğidir. Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik (Simple Validation) 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, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır. (Doğruluk Oranı = 1 - Hata Oranı)

Önemli diğer bir değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, bir çok işletme uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Çok ender olarak yorumlanamayacak kadar karmaşıklaşılar da, genel

olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir.

3.3.6.4. Modelin Kullanılması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmin edilen envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir.

3.3.6.5. Modelin İzlenmesi ve Güncelleme

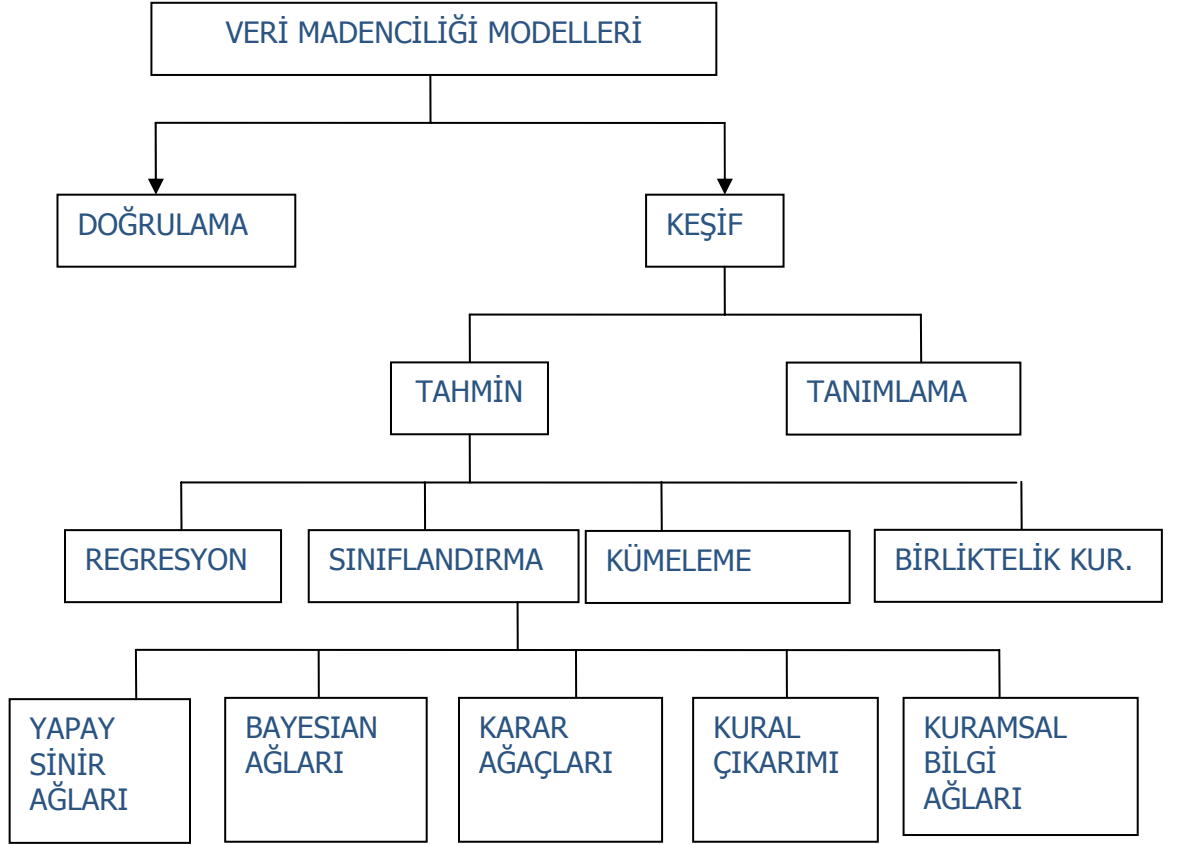
Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Güncelleme modelin yeni verilerle birlikte hala uygun olup olmadığına karar vermek için tekrar onaylanmasını gerektirir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir. Model ne olursa olsun, zaman içerisinde değişiklikler yapılabilir.

Veri madenciliği sürecinde, kümeleme, veri özetleme, sınıflama kurallarının öğrenilmesi, bağımlılık ağlarının bulunması ve değişkenlik analizi gibi farklı birçok teknik kullanılmaktadır.

Etkin bir veri madenciliği süreci için aşağıdaki hususlara dikkat edilmesi gerekmektedir:

- Veri gizliliği ve güvenliğinin sağlanması,
- Sonuçların yararlılık, kesinlik ve anlamlılık kıstaslarının sağlanması,
- Farklı tiplerdeki verileri ele alma,
- Farklı ortamlarda yer alan veri üzerinde işlem yapabilme,

- Veri madenciliği algoritmasının etkinliği ve ölçeklenebilirliği,
- Keşfedilen kuralların çeşitli biçimlerde gösterimi,
- Farklı birkaç sorgulama düzeyi ve etkileşimli veri madenciliği.



Şekil 3.6. Veri madenciliği modelleri [94].

3.4. Veri Madenciliğinde Ön İşleme

Verinin kalitesi, veri madenciliğinde en temel konulardan birisidir. Veriler hiçbir zaman ayıklanmış ve veri madenciliği için uygun formda değildir. Veri madenciliğinde güvenilirliğin artırılması için, veri ön işleme yapılmalıdır. Aksi halde hatalı olan girdiler bizi hatalı çıktıya götürecektir. Veri ön işleme, çoğu durumlarda yarı otomatik olan ve zaman isteyen bir veri madenciliği aşamasıdır. Bütün veri madenciliği işleminin en çok zaman alan aşaması bu aşamadır [95].

Verilerin sayısındaki artış ve buna bağlı olarak çok büyük sayıda verilerin ön işlemeden geçirilmesinin gerekliliği, otomatik veri ön işleme için etkin teknikleri önemli hale getirmiştir.

Veri ön işleme aşağıdaki sebeplerden dolayı verilere uygulanmaktadır [96]:

- 1- Veriler üzerinde herhangi bir analiz türünün uygulanmasını engelleyecek veri problemlerinin çözümü,
- 2- Verilerin doğasının anlaşılması ve anlamlı veri analizinin başarılması,
- 3- Verilen bir veri kümesinden daha anlamlı bilginin çıkarılması,

Çok sayıda uygulamada, veri ön işlemenin bir türünden daha fazlasına ihtiyaç duyulmaktadır. Veri ön işlemenin türünün belirlenmesi bu sebeple önemli bir aşamadır [97].

Çok sayıda veri ön işleme tekniği mevcuttur. Bunlardan biri olan veri temizleme verilerdeki gürültünün giderilmesi ve tutarsızlıkların düzeltilmesi için uygulanır. Bir diğer teknik olan veri birleştirme ise farklı kaynaklı verileri uygun bir veri tabanında birleştirir. Normalleştirme gibi veri dönüştürme yöntemleri uygulanabilir. Veri indirgeme de ise fazla olan bazı değişkenlerin atılması, birleştirilmesi veya kümeleme yolu ile veri büyüklüğünün azaltılması amaçlanır. Bu sözü edilen veri ön işleme teknikleri, veri madenciliğinden önce uygulanarak elde edilen sonuçların kalitesi artacak ve/veya veri madenciliği için harcanacak zaman azalacaktır [1]

Veri ön işleme teknikleri yukarıdaki paragraftan da anlaşılabilceği gibi şu şekilde sıralanabilir:

- 1- Veri Temizleme,
- 2- Veri Birleştirme ve Dönüştürme,
- 3- Veri İndirgeme,
- 4- Ayırıklaştırma ve Kavram Hiyerarşisi Üretme.

Yukarıda bahsedilen veri ön işleme teknikleri verilerin kalitesini artıracaktır. Kaliteli kararların alınmasında kaliteli veriler gerekliliği bir gerçektir. Bu sebeple, veri ön işleme süreci veri madenciliğinin önemli bir adımıdır. Bundan sonraki bölümlerde bu veri ön işleme teknikleri açıklanacaktır.

3.4.1. Veri Temizleme

Gerçek hayattaki veri kirlenmeye, eksik olmaya ve tutarsız olmaya meyillidir. Veri temizleme işlemleri kayıp değerleri doldurur, aykırı değerleri tanımlarken dış gürültüleri ve verideki uyumsuzluğu düzeltir [1].

3.4.1.1.Kayıp Değerler

Analizde kullanacağımız parametrelere ait tablo satırlarında bazı özelliklerin olmadığını düşünelim. Herhangi bir değişkene ilişkin eksik değerlerin doldurulması için farklı yollar vardır. Bunlardan bazıları şöyledir:

1. Tablo-satırını yok saymak,
2. Kayıp değerlerin el ile doldurulması,
3. Kayıp değerler için genel sabit kullanımı,
4. Kayıp değerleri doldurmak için ortalama özellikleri kullanmak,
5. Tablo-satırında verilen aynı sınıfa ait bütün değerler için ortalama özelliği kullanmak,
6. Kayıp değerleri doldurmak için en olası ortalama değerlerin kullanımı.

3.4.1.2. Gürültülü Veri

Veri temizleme tekniğinin kullanılması gereken bir diğer problem ise gürültülü verilerdir. Gürültü, ölçülen bir değerdeki hatadır. Verideki gürültüyü yok edebilmek için kullanılan bazı veri düzeltme teknikleri, kova metodu, histogram, kümeleme analizi ve egni uydurma (regresyon) gibi tekniklerdir [98].

3.4.1.3. Uyumsuz Veri

Bazı veri işlemleri (hareketleri) için veri kayıtları içinde uyumsuzluklar olabilir. Bazı veri uyumsuzlukları dış referanslar kullanılarak düzeltilebilir. Örneğin, veri girişinde oluşan hatalar kayıtların tekrar gözden geçirilmesi ile düzeltilebilir. Bu kodların uyumsuzluğunu düzeltmek için kullanılan rutin tasarımıdır. Bilinen veri sınırlarının ihlalinin saptanmasında bilgi mühendisliği araçları kullanılır. Örneğin, fonksiyonel sınırlamalarla çatışan değerlerin bulunması için nitelikler arası bilinen fonksiyonel bağımlılıklar kullanılabilir.

Veri bütünleştirme sebebi ile uyumsuzluk meydana gelebilir. Fazlalıklar atılmış olsa dahi bir veri tabanında verilen nitelikler bir başka veri tabanında farklı isimle olabilirler.

3.4.2. Veri Birleştirme ve Dönüşümü

Veri madenciliğinde genellikle farklı veri tabanlarındaki verilerin birleştirilmesi ve dönüştürülmesi gerekmektedir.

3.4.2.1. Veri Birleştirme

Farklı veri tabanlarındaki veriler tek bir çatı altında – ki bu genellikle bir veri ambarıdır- birleştirilirler. Farklı veri tabanlarındaki verilerin tek bir veri tabanında birleştirilmesiyle şema birleştirme hataları oluşur. Örneğin, bir veri tabanında girişler “tüketici-kimliği” şeklinde yapılmışken, bir diğerinde “tüketici-numarası” şeklinde olabilir. Bu tip şema birleştirme hatalarından kaçınmak için meta veriler kullanılır. Veri tabanları ve veri ambarları genellikle meta veriye sahiptirler.

Fazlalık diğer önemli bir konudur. Başka tablolarda bazı nitelikler fazlalık olabilir. Bazı fazlalıklar korelasyon analizi tarafından tesbit edilebilir. Örneğin, bu analiz kullanılan veri üzerinde verilen iki nitelikten birinin diğerini ne kadar güçlü biçimde gerektirdiği ölçütünü hesaplayan analizdir.

Çoklu kaynaktaki veri birleştirmesi dikkatlice yapılırsa sonuç veri setindeki veri azaltmaya ve fazlalıkların atılmasına ve uyumsuzlukların giderilmesine yardımcı olur. Bu sonradan yapılan madencilik işlemlerinin hızının ve doğruluğunun artmasına yardımcı olur.

3.4.2.2. Veri Dönüşümü

Veri dönüştürme ile veriler, veri madenciliği için uygun formlara dönüştürülürler. Veri dönüştürme; düzeltme, birleştirme, genelleştirme ve normalleştirme gibi değişik işlemlerden biri veya bir kaçını içerebilir. Veri normalleştirme en sık kullanılan veri dönüştürme işlemlerinden birisidir.

3.4.2.3. Nitelik Oluşturma

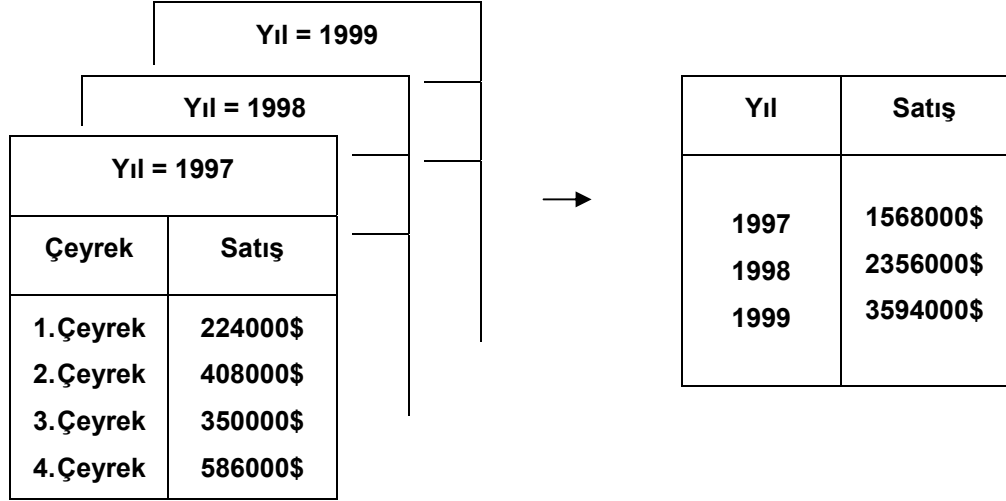
Orijinal niteliklerden daha önemli bilgi içeren ve veri madenciliği algoritmalarının başarısını artıracak yeni niteliklerin oluşturulmasıdır.

3.4.3. Veri İndirgemesi

Çok büyük miktarda verilerin analiz edilmesi uzun süreceğinden, veri bütünlüğünü bozmadan verinin tamamını temsil edecek az miktarda veri alınır ve analiz edilir. Bu işlemlerin sonucunun orijinale yakın olması beklenir. Beş şekilde yapılır:

1. Veri Birleştirme
2. Boyut indirgeme
3. Veri sıkıştırma
4. Çokluk azaltımı
5. Ayırıklaştırma ve Kavram Oluşturma.

3.4.3.1. Veri Birleştirme



Şekil 3.7. Verileri bir araya getirme örneği.

Veri birleştirme veya veri küpü yapılacak 1997-1999 yılları için çeyrek dönemlik satış tutarlarından oluşan bir veri kümesinin bulunduğunu varsayalım. Bu yıllar için yıllık satış tutarları tek bir tabloda toplandığında veri birleştirilmesi yapılmış olur. Sonuç olarak elde edilen veri kümesinin hacmi daha küçüktür fakat yapılacak analiz için bir bilgi kaybı söz konusu değildir. Veri küpleri ise çok değişkenli birleştirilmiş bilginin saklandığı küplerdir. Örneğin bir firmanın satış tutarları yıllar, satışı yapılan ürünler ve firmanın farklı satış yerleri için aynı küp üzerinde gösterilebilir. Veri küpleri özet bilgiye herhangi bir hesaplama yapmadan hızlı bir biçimde erişilmesini sağlarlar.

3.4.3.2. Boyut indirgeme

Veri madenciliği yapılacak veri kümesi bazen gereksiz olarak yüzlerce değişken içerebilir. Örneğin bir ürünün satışına ilişkin olarak düzenlenen bir veri kümesinde, tüketicilerin telefon numaraları gereksiz bir değişken olarak yer alabilir. Bu tür gereksiz değişkenler elde edilecek örüntüleri kalitesizleştirebileceği gibi veri madenciliği sürecinin yavaşlamasına da yol açacaktır.

Gereksiz değişkenlerin elenmesi amacıyla ileri veya geri yönlü olarak sezgisel seçimler yapılabilir. İleri yönlü sezgisel seçimde orijinal değişkenleri en iyi temsil edecek

değişkenler belirlenir. Ardından her bir değişken veya değişkenler grubunun, bu kümeye dahil edilip edilmeyeceği sezgisel olarak belirlenir. Geri yönlü sezgisel seçimde ise öncelikle değişkenlerin tüm kümesi ele alınır. Daha sonra gereksiz bulunan değişkenler kümeden dışlanarak, en iyi değişken kümesi elde edilmeye çalışılır. Nitelikler alt kümesi kullanılarak elde edilen sınıfların dağılımları gerçek dağılıma eşit ya da çok yakın olmalıdır.

3.4.3.3. Veri sıkıştırma

Analiz edilecek veri miktarını azaltmak için veri sıkıştırılır. Eğer veri sıkıştırma sonrası orijinal veriden herhangi bir bilgi kaybı oluşmuyorsa buna kayıpsız (lossless) sıkıştırma adı verilir. Ancak orijinal veriye yaklaşık bir değer elde ediliyorsa (belli bir oranda kayıp varsa) buna da kayıplı (lossy) sıkıştırma denir. Sıkıştırma işlemleri belirli algoritmalara göre yapılır. Sıklıkla kullanılan sıkıştırma yöntemleri: Wavelet Transformatları ve Temel Bileşen Analizi'dir.

3.4.3.4. Çokluk Azaltımı

Çokluk azaltımı ile ilgili teknikler, “Veri gösteriminin ‘daha küçük’ alternatif biçimlerini seçerek veri hacmini azaltabilir miyiz?” sorusuna yanıt arama amacını taşımaktadır. Bu teknikler parametrik ya da parametrik olmayan şeklinde ikiye ayrılır. Parametrik yöntemler içinde yer alan bir model, veriyi tahmin etmek üzere kullanılır; böylece gerçek veri yerine yalnızca veri parametreleri saklanır. Eğri uydurma, parametrik yöntemlere örnektir. Histogramlar, kümeleme ve örnekleme, parametrik olmayan yöntemlere örnektir.

3.4.3.5. Ayırıklaştırma ve Kavram Oluşturma

Ayırıklaştırma teknikleri, belirli bir sürekli özelliğin sırasını aralıklara bölerek, o özelliğin değerlerinin sayısını azaltmak için kullanılabilir. Bu yöntemler genellikle yenilemelidir. Bu nedenle sıralanması gereken ayırık değerler ne kadar az olursa, bu yöntemler de o kadar hızlı olmalıdır. Özellik değerlerinin kavram hiyerarşisi olarak

bilinen hiyerarşik bir bölümlenmesini sağlamak için, pek çok ayırıklaştırma tekniği yenilemeli olarak uygulanabilir.

Belirli bir sayısal özellik için bir kavram hiyerarşisi, o özelliğin bir ayırıklaştırmasını tanımlar. Kavram hiyerarşileri düşük-düzye kavramları (yaş özelliği için sayısal değerler gibi) yüksek-düzye kavramlara toplayarak (genç, orta yaş, yaşlı gibi) veriyi azaltmak için kullanılabilir.

Kavram hiyerarşilerinin elle tanımlanması, zor ve zaman harcayan bir görev olabilir. Neyse ki; pek çok hiyerarşi, veri tabanı şemasında dolaylı olarak yer alır ve şema tanımlama düzeyinde tanımlanabilir.

3.5. Veri Madenciliği Fonksiyonları

3.5.1. Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modelleri, tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında toplayabiliriz [99].

1. *Tahmin edici modellerde*; sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesi amaçlanmaktadır. 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.

2. *Tanımlayıcı modellerde*; ise karar vermeye rehberlik etmekte kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. 25 yaş altı bekar kişiler ile, 25 yaş üstü evli kişiler üzerinde yapılan ve ödeme performanslarını gösteren bir analiz tanımlayıcı modellere örnek olarak verilebilir.

Veri madenciliği modellerini gördükleri işlevlere göre,

- Sınıflama ve Regresyon
- Kümeleme
- Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

olmak üzere üç ana başlık altında incelemek mümkündür. Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir [100].

3.5.1.1. Sınıflama ve Regresyon Modelleri

Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan ve veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır. Ancak çok terimli lojistik regresyon gibi kategorik değerlerin de tahmin edilmesine olanak sağlayan tekniklerle, her iki model giderek birbirine yaklaşmakta ve bunun bir sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır. Sınıflama ve regresyon modellerinde kullanılan başlıca teknikleri şunlardır:

- Karar Ağaçları
- Yapay Sinir Ağları
- Genetik Algoritmalar
- En Yakın Komşu
- Bellek Tabanlı Nedenleme
- Lojistik Regresyon'dur.

3.5.1.2. Kümeleme Modelleri

Kümeleme modellerinde amaç, küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir. Başlangıç aşamasında veri tabanındaki

kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı bilinmemekte, konunun uzmanı olan bir kişi tarafından kümelerin neler olacağı tahmin edilmektedir.

3.5.1.3. Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak market sepet analizi adı altında veri madenciliğinde yaygın olarak kullanılmaktadır. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır.

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

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

Ardışık zamanlı örüntüler ise birbirleri ile ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır. Örneğin, X ameliyatı yapıldığında, 15 gün içinde % 45 ihtimalle Y enfeksiyonu oluşacaktır.

3.6. Veri Madenciliği Hedefleri

Pratik anlamdaki veri madenciliği genellikle sınırlı sayıda tekniğin, yine sınırlı sayıda koşullarda uygulanması ile gerçekleştirilir. Farklı sınıflandırmalar olsa da, güncel, akademik, ekonomik ve iş hayatındaki birçok sorun aşağıdaki altı genelleme ile tanımlanabilir.

- Sınıflandırma
- Kestirme
- Öngörü
- Benzerlik Tespiti
- Kümeleme
- Tanımlama

Tüm bu analizler için eşit oranda verimlilik sağlayacak kullanışlı bir veri madenciliği aracı ya da tekniği bulunmamaktadır. İstenilen amaca ve analizin türüne göre değişik teknikler kullanılmaktadır.

3.6.1. Sınıflandırma

En çok kullanılan veri madenciliği fonksiyonlarından birisi olan sınıflandırma işlemi, insan düşünce yapısına en yakın veri madenciliği işlemlerinden bir tanesidir. İnsanoğlu dünya üzerindeki maddeleri daha iyi anlamak ve başkalarına anlatmak için hemen hemen her şeyi sürekli sınıflandırmakta, kategorilere ayırmakta ve/veya derecelendirmektedir. Maddeleri elementlere, kedileri türlere, ülkeleri şehirlere, şehirleri semtlere vb. şeklinde devamlı kategorize etmektedir.

Veri madenciliğinde geçerli olan sınıflandırma işleminde ise amaç, yeni karşılaşılan bir girdinin özelliklerinin incelenip, bu girdinin daha önce tanımlanmış olan sınıflardan hangisine atanacağına karar vermektir [101].

Sınıflandırma işlemine örnek olarak verilebilecek bazı işlemler aşağıda sıralanmıştır:

- Kredi başvurularının düşük, orta ve yüksek riskli olarak sınıflandırması,
- Sigorta şirketlerini dolandırmak için işlenen sahteciliklerin tespit edilmesi ,
- Yeni gelen bir öğrencinin hangi sınıfta eğitim görmesi gerektiğine karar verilmesi.

Tüm bu örneklerde sınırlı sayıda sınıf bulunur ve beklenen de her bir kaydın bu sınıflardan birisine atanmış olmasıdır. Sınıflandırma için Karar Ağaçları, Bellek Tabanlı Sistemler, Yapay Sinir Ağları yöntemleri en uygun olan tekniklerdir. Ayrıca bazı durumlarda İlişkisel Analiz yöntemi de faydalı sonuçlar üretmektedir.

3.6.2. Kestirme

Bir önceki bölümde anlatılan sınıflandırma işlemi ayırık sonuçlar üretmektedir, yani bir kayıt ya belirli bir sınıfa aittir ya da değildir. Ancak kestirme işleminde ise sürekli değerler alabilen sonuçlar üretilmektedir. Kestirme işleminde, giriş verisine göre bilinmeyen sürekli bir değişken (örneğin gelir düzeyi, boy ya da kredi kartı bakiyesi gibi) için bir değere ulaşılır.

Uygulamada kestirme çoğunlukla sınıflandırma işlemlerini gerçeklemek için kullanılır, örneğin bir kredi kartı şirketi, hesap ekstrelerinde yer alan reklam alanını bir kayak takımı üreticisine kiralamak istediği zaman, kredi kartı şirketinin ekstre göndereceği tüm müşterilerini, kayak yapanlar ve yapmayanlar şeklinde ayıracak sınıflandırma modelini kurması gerekmektedir. Bir başka yaklaşım ise ekstre gönderilecek her müşterinin kayak yapma ihtimalinin hesaplanmasıdır. Bu ihtimal 0 ile 1 arasında herhangi bir değer alabilir. Şu aşamada sınıflandırma sorunu, uygun bir eşik değer seçimine indirgenmiştir. Eşik değere eşit ya da daha fazla "kayak yapma ihtimali"ne sahip müşteriler "kayak yapanlar", bu eşik değerinin altında bir değerde "kayak yapma ihtimali"ne sahip müşteriler ise "kayak yapmayanlar" olarak sınıflandırılmış olurlar [102].

Tahmin etme işleminin sınıflandırma işlemine göre bir diğer avantajı da her bir kaydın sahip olduğu değerlendirmeye göre tüm kayıtların sıralanabilir hale gelmesidir. Az önceki örnek üzerinden devam etmek gerekirse, kayak takımı üreticisinin 500 bin ekstreye reklam verebilecek sınırlı bir bütçesi olması durumunda eğer 1,5 milyon kredi kartı müşterisi "kayak yapanlar" olarak sınıflandırılmış ise seçilecek en basit yöntem bu 1,5 milyon kişi içerisinde rasgele müşterileri seçmek ve reklam içeren ekstreleri seçilen bu müşterilere göndermek olacaktır. Ancak eğer kredi kartı şirketinin her müşteriye ait bir kayak yapma ihtimali değerlendirmesi varsa, reklam içeren ekstreler,

kayak yapma ihtimali en yüksek olan 500 bin müşteriye gönderilerek en uygun çözüm sağlanmış olur.

Tahmin etme işleminin başka örnekleri aşağıda bulunmaktadır:

- Bir ailedeki çocuk sayısının tahmin edilmesi,
- Ailenin toplam gelirin hesaplanması,
- Müşterilerin bağlılık süresinin hesaplanması,
- Müşterilerin yeni banka hesap tiplerini seçme olasılığının bulunması.

3.6.3. Öngörü

Öngörü işlemi sınıflandırma veya kestirme işlemlerine çok benzer bir işlemdir. Ancak öngörü işleminde sınıflandırma, gelecek için tahmin edilen belirli bir davranışa ya da belirli bir değere göre yapılır. Öngörü işleminde yapılan sınıflandırmanın doğru olup olmadığını test etmenin tek yöntemi, bekle ve gör prensibidir.

Sınıflandırma ve değerlendirme işlemlerinde kullanılan hemen hemen her yöntem küçük değişikliklerle öngörü işlemi için de kullanılabilir hale gelirler. Bunun için, eski verilerden, gelecek için tahminde bulunulacak verinin daha önceden bilindiği bir eğitim kümesi hazırlamak gerekmektedir. Bu eski veriler, şu anda gözlenen davranışları açıklayacak bir modelin kurulmasında kullanılacaktır. Bu modelin herhangi bir veri kümesine uygulanması sonucu ortaya çıkan sonuç ise gelecekteki davranışlar için bir öngörü niteliğinde olacaktır.

Örneğin bir mağazadan yapılan alışverişlerde hangi malların daha çok bir arada alındığının ortaya çıkarılmasını amaçlayan Sepet Analizi Tekniği, küçük değişikliklerle gelecekte hangi malların bir arada satılacağını tahmin etmeyi ve bu amaca uygun olarak satışı daha da arttıracak hazırlıkların belirlenmesini sağlayacak şekilde uyarlanabilir.

Sepet analizi tekniği, bellek tabanlı teknikler, karar ağaçları ve yapay sinir ağları yöntemlerinin tümü öngörü işlemi için uygundur. Öngörü işlemi için kullanılacak olan yöntemin seçimi tamamen giriş verisinin özelliklerine, öngörülecek değerin tipine ve

yapılan öngörünün açıklanabilirlik özelliğinin ne derece önemli olduğuna doğrudan bağlıdır.

3.6.4. Benzerlik Tespiti

Benzerlik gruplaması ya da sepet analizi, hangi ürünlerin hangi ürünlerle daha çok satıldığının belirlenmesi amacı ile kullanılır. En yaygın örneği süper marketlerdeki alışveriş kayıtlarının incelenerek hangi satışlarda hangi ürünlerin daha çok beraber satıldıklarının saptanmasıdır. Perakende satış yapan mağazalar, benzerlik gruplaması işlemini daha çok ürünlerin raflardaki ya da kataloglardaki yerleşiminin, çok satılan ürünlerin birlikte gözükecek şekilde uygun olarak düzenlenmesi amacı ile kullanmaktadırlar. Aynı zamanda benzerlik gruplaması ile çapraz satış fırsatlarının belirlenmesi ya da hizmetlerde ve ürünlerde satışı artırıcı paketlerin, gruplandırmaların ve promosyonların yapılması sağlanabilmektedir [102].

Aynı zamanda benzerlik eşlemesi, verilerden kurallar çıkarmanın en basit yöntemidir. Örneğin kedi maması ile kedi kumu yeterli sayıda birlikte satılıyorlar ise aşağıdaki 2 ilişkisel kuralı türetebiliriz:

- Kedi maması alan bir müşterinin, kedi kumu da alma olasılığı P_1 'dir.
- Kedi kumu alan bir müşterinin, kedi maması da alma olasılığı P_2 'dir.

3.6.5. Gruplandırma

Gruplandırma işlemi, karışık yapıda olan bir popülasyonu, daha homojen birkaç alt gruba ya da kümeye bölümlenme işlemidir. Sınıflandırma ile gruplandırmayı birbirinden ayıran en önemli fark, gruplandırma işleminin, sınıflandırma işleminde olduğu gibi önceden belirlenmiş bir takım sınıflara göre bölümlenme yapmamasıdır.

Sınıflandırmada yeni gelen her bir kayıt, önceden sınıflandırılmış bir takım sınıflar üzerinde yapılan bir eğitim neticesinde ortaya çıkan bir modele göre, önceden belirlenmiş olan bir sınıfa atanmaktadır. Gruplandırma ya da demetleme işleminde ise önceden tanımlanmış sınıflar ya da numune sınıflar bulunmamaktadır. Kayıtların

gruplandırılması işlemi, kayıtların birbirlerine olan benzerliklerine göre yapılmaktadır. Oluşan sınıfların hangi anlamları taşıdığının belirlenmesi tamamen analizi yapana kalmıştır. Örneğin hasta kayıtlarından oluşan verilerin gruplanması sonucunda semptomlardan oluşan kümeler, değişik hastalıklara işaret edebilir [103].

Gruplama işlemi çoğunlukla bir başka veri madenciliği işlemi için bir ilk işlem olarak kullanılır. Örneğin gruplama işlemi bir pazar payı araştırması için bir ilk işlem olarak uygulanabilir. "Ne tip promosyonlar müşteriler tarafından rağbet görür?" sorusunun cevabını bulmayı kolaylaştırmak için herkes için tek bir model uygulamaktan vazgeçip, müşteriler alışveriş alışkanlıklarına göre gruplandırılırsa, her küme için "Bu kümedeki müşteriler hangi tip promosyonlara rağbet eder?" sorusunun cevabı çok daha kolay verilir.

3.6.6. Tanımlama

Veri madenciliği uygulamalarının amaçlarından birisi de, karmaşık bir veri tabanında neler olup bittiğini, verileri ilk elden oluşturan insanlar, ürünler ve işlemler bazında daha iyi anlamamızı sağlayacak açıklamalar bulmaktır. Bir davranış için verilen iyi bir tanımlama çoğu kez bu davranış için bir de açıklama getirmelidir. En azından iyi bir tanımlama, açıklamayı bulmak için nereden başlanması gerektiğini göstermelidir.

3.7. Çıkarılan Bilgi Türleri

Veri madenciliği ile elde edilen bilgiler genellikle çok basit olmayan, belirli düzeyde akılcılık gerektiren bilgilerdir. Bu bilgileri göstermek için kurallar, önerme mantığı, karar ağacı, semantik ağlar, nöron ağları ve sınıf sıra düzeni gibi çok çeşitli yöntemler kullanılabilir. Veri madenciliği ile çıkarılan başlıca bilgi türlerini aşağıdaki başlıklar altında sıralayabiliriz.

3.7.1. Eşleştirme Kuralları

"Eğer (girdi) => sonuç " biçiminde ifade edilebilecek bir kural, X değişkenler kümesinin her birleşimi için eğer öncel geçerli ise sonuç da geçerlidir' anlamını taşır.

Ancak öncel ile sonuç arasındaki ilişki kesin olmayan, olasılıklı bir ilişkidir. Dolayısıyla kuralın “eğer öncel geçerli ise, sonucun da geçerli olma olasılığı yüksektir” biçiminde anlaşılması daha doğrudur. Eşleştirme kurallarına örnek olarak ‘süper markette kereviz alan müşteriler havuç da alıyor’, ‘matematik dersini alan öğrenciler beden eğitimi ve spor dersini almıyor’ gibi kurallar verilebilir.

3.7.2. Sınıflandırma Sıradüzeni

Veri madenciliği, mevcut durumları ya da hareketleri inceleyerek bir sınıflandırma sıradüzeni oluşturabilir. Örneğin bir bankanın müşterileri, müşterilerin bugüne kadar kullandıkları krediler ve ödeme durumları ile ilgili bilgiler incelenerek öğrenim durumu, cinsiyet, yaş, gelir düzeyi gibi belirli niteliklerin birleşimi kullanılarak müşterilerin güvenilirlik düzeyini gösteren tek düzeyli bir sınıflandırma ya da çok düzeyli bir sınıflandırma sıradüzeni oluşturulabilir. Bu tür bilgiler kullanılarak yürütülecek uygulamalar, kurumların karşılaşılabileceği birçok olumsuz durumu önleyeceği, dolayısıyla kazançlarını arttıracığı için bu bilgiler kurumlar için çok değerli olabilir.

3.7.3. Sıra-Zaman Örüntüleri

Bu tür bilgiler belirli eylem ya da olayların birbirini izlediğini ve sıralı bir dizi oluşturduğunu ortaya koyan bilgilerdir. Örneğin “video-kamera satın alan müşterinin, ilk üç ay içinde hangi tüketim malzemelerini ya da ek gereçleri satın aldığı sonraki altı ayda neleri satın aldığı” bu tür bir bilgidir. Sıra örüntülerine bir diğer örnek olarak belirli bir ameliyat geçiren hastalarda ilk bir yıl içinde ne tür sorunların ortaya çıktığı, bu hastalarda ilk beş yılda hangi hastalıklara rastlandığı gibi bilgiler verilebilir.

3.7.4. Zaman Serileri Arası Bağımlılıklar

Birçok uygulamada belirli nitelik değerleri belirli aralıklarla tekrarlanarak zaman serileri oluşturulur. Günlük döviz kurları, günlük faiz oranları, günlük hisse senedi değerleri, haftalık satış miktarları, aylık yağış miktarları, yıllık nüfus değerleri zaman serisi örnekleridir. Veri madenciliği zaman serileri arasındaki belirli örüntüleri,

korelasyonları ya da farklı bağımlılıkları ortaya çıkaran bilgiler üretebilir. Bu tür bilgilere örmek olarak “iki ürünün satış eğrilerinin altı ay farkla birbirinin aynı olması”, ya da ‘faiz oranlarındaki hissedilir artışları izleyen iki gün içinde hisse senedi fiyatlarında hissedilir bir düşüş olması’ türü bilgiler sayılabilir.

3.8. Veri Madenciliği Yaklaşımları

Veri madenciliği için kullanılan belli başlı iki ana yöntem aşağıda anlatılmıştır.

3.8.1. Kullanım GÜdümlü Yaklaşım

Bu yaklaşımda bilginin çıkarılması için yürütülen işlemler kullanıcının güdümündedir. Örneğin eşleştirme kuralları bulmak isteyen kullanıcı, kurallarla ilgili hipotezler geliştirir. Geliştirdiği hipotezleri mevcut verileri kullanarak doğrulamaya çalışır. Eğer kural doğrulanmıyorsa, hipotezde değişiklikler yaparak doğrulanan bir kural bulmaya çalışır.

Örmek olarak öğrencilerin ÖSS sınavlarındaki başarıları ile üniversitelerdeki başarıları üzerinde çalışan ve bu iki başarı ile ilgili kural oluşturmaya çalışan bir kullanıcının “Üniversiteye girişte, ilgili puan sıralamasında ilk 10.000 arasında yer alan öğrencilerin üniversitedeki akademik ortalamaları tam notun % 80’inden yüksek oluyor hipotezini getirdiğini düşünelim. Bu hipotezin doğrulanması için hesaplanan destek ve uyum değerlerinin yeterli olmadığını varsayalım. Elde ettiği sonuçlar ışığında kullanıcı hipotezini bu kez “Üniversiteye girişte sayısal ağırlıklı puan sıralamasında ilk 10.000 arasında yer alan ve üniversitelerin mühendislik bölümlerinde okuyan öğrencilerin akademik ortalamaları tam notun % 80’inden yüksek oluyor” hipotezini geliştirebilir ve yapılan sınamalarda bu hipotez doğrulanabilir. Hipotezi doğrulanan kullanıcı yeni hipotezler geliştirerek yeni kurallar bulmaya çalışabilir.

Kullanıcı güdümlü veri madenciliği belki her alanda kullanılamaz. Ancak bu yaklaşımla, belirsizlik düzeyi düşük ortamlarda, konuya hakim uzman kullanıcılar tarafından geliştirilecek hipotezlerle çok çabuk sonuç elde edilebilir.

3.8.2. Otomatik Yaklaşım

Veri madenciliğinde otomatik bilgi çıkarma yaklaşımı yapay zeka ve makine öğrenme alanlarındaki çalışmalardan etkilenmiştir. Otomatik yaklaşımda kullanılan teknikler büyük ölçüde çıkarılacak bilginin türüne göre değişir. Örneğin sınıflandırma kuralı bulunmasında mevcut verilerden bir eğitim kümesi oluşturulur. Eğitim kümesindeki nesneler bilinen nesneler olduğu için bunların sınıflandırma bilgileri bellidir. Ya da kredi limitlerini belirlemek için müşterilerini sınıflandırmak isteyen bir bankada, sınıflandırma kuralı bulmaya çalışılırken eski müşterilerden oluşan bir eğitim kümesi oluşturulur. Bu kümedeki müşteriler uzun zamandır kredi kullandıkları için kullandıkları kredi miktarları ve ödeme bilgileri mevcuttur. Bu bilgilere göre eğitim kümesindeki müşterilerin sınıflandırma bilgileri belirlenebilir. Bundan sonra, uygun teknikler kullanılarak, müşterilerin eğitim durumu, gelir durumu, mesleği, yaşı, ve benzeri bilgilerinden sınıflandırma bilgisinin türetilmesini sağlayacak bir kural bulunmaya çalışılır. Sınıflandırma kuralı bulunmaya çalışılırken, önce tek niteliğe dayalı kurallar oluşturulup sınanır. Bunda başarı sağlanamazsa bu kez iki niteliğe dayalı kurallar oluşturulup sınanır ve çalışmalar tatmin edici bir sonuç elde edinceye kadar sürdürülür.

Otomatik kural bulma yaklaşımı uzman yazılımların kullanılmasını gerektiren ve uygulamaları çok uzun sürebilen bir yaklaşımdır. Ancak sonuçta kuruluşlar için çok değerli bilgiler elde edilebileceği için bedeli yüksek de olsa bu çalışmaların yapılması anlamlı olabilmektedir.

3.9. Veri Madenciliği Teknikleri

3.9.1. Sepet Analizi

Sepet analizi tekniği, adından da anlaşılacağı üzere, çok basit olarak hangi ürünlerin hangi ürünlerle satıldığını, hangi ürünlerin promosyona girmesi gerektiğini ve benzeri bilgileri ortaya çıkarır. Her ne kadar yoğunlukla satış alanında kullanılsa da, sepet analizi tekniği bunun dışında bir çok alanda daha kullanılmaktadır:

- Kredi kartları ile yapılan alışverişlerin işlenmesi ve müşterilerin yapacakları potansiyel harcama kalemlerinin bulunması,
- Cep telefonundaki opsiyonel hizmetlerin (GPRS, WAP, Telesekreter vb.) müşteriler tarafından tercih edilmelerine göre karı arttırmak için hangi ürünlerin birlikte kampanyaya girmesi gerektiğini belirlemek,
- Sigortacılıkta ortaya çıkan değişik tarzdaki suçların dolandırıcılık olup olmadığının belirlenmesi ve yapılacak soruşturmaya ışık tutulması,
- Hastaların sağlık kayıtlarından, önerilen çeşitli tedavi kombinasyonlarından doğan komplikasyonların öngörülmesi,

Sepet analizi çoğunlukla ticari anlam taşıyan verilerin var olduğu ancak bu veri üzerinde hangi örüntülerin aranılacağına bilinmediği durumlarda bir başlangıç noktası olarak kullanılır. Bu veri içerisindeki bazı kalıplar sayesinde kazancı arttırmak üzere bir takım aksiyonlara gidilebilir.

Ancak sepet analizi, geleceğe dönük tahminlerde pek başarılı değildir. Sepet analizinin temeldeki mantığında yer alan teknikler istatistik ve olasılıktan gelmektedir.

Sepet analizinin başarılı olduğu noktalar:

- Kolay ve anlaşılır sonuçlar üretir.
- Gözetimsiz veri madenciliği yöntemidir.
- Değişik boyutlardaki veriler üzerinde çalışabilir.
- Her ne kadar kayıtların sayısı ve kombinasyon seçimine göre işlem adedi artsa da sepet analizi için her adımda gerekli olan hesaplamalar diğer yöntemlere göre (genetik algoritmalar, yapay sinir ağları vb.) çok daha basittir.

Sepet analizinin başarısız olduğu noktalar:

- Sorunun boyutu büyüdükçe, gerekli hesaplamalar üstel olarak artmaktadır.
- Sepet analizinde kullanılacak doğru ürünlerin seçimi. Ürün gruplandırma (süt ürünleri, unlu mamuller vb.) biraz bilgi kaybı getirirse de analizin boyutlarını küçültebilir.

- Kayıtlarda çok az rastlanan ürünleri yok sayar. Sepet analizi tekniği, en doğru sonucu, tüm ürünlerin kayıtlar içinde yaklaşık aynı frekansta görüldüğü durumlarda üretmektedir.

3.9.2. Bellek Tabanlı Yöntemler

İnsanlar kararlarını genellikle daha önce yaşadıkları deneyimlere göre verirler. Örneğin doktorlar bir hastayı incelerken, daha önce tedavi ettiği benzer hastalığa yakalanmış hastalar üzerindeki deneyimlerini kullanırlar. Aynı şekilde bir sigorta eksper de bir olayın sahtekarlık olup olmadığını bulmak için daha önceki sahtekarlıklarla olan benzerlikleri göz önünde bulundurmaktadır. Gözetimli bir veri madenciliği tekniği olan bellek tabanlı yöntemler de benzer şekilde deneyimleri kullanmaktadır. Bu yöntemlerde, bilinen kayıtların bulunduğu bir veritabanı oluşturulur ve sistem yeni gelen bir kayda komşu olan diğer kayıtları belirler ve bu kayıtları kullanarak tahminde bulunur ya da bir sınıflandırma işlemi gerçekleştirir. Bellek yaklaşımli yöntemlerin en önemli özelliği veriyi olduğu gibi kullanabilme yeteneğidir. Diğer veri madenciliği tekniklerinin aksine bellek tabanlı yöntemler, kayıtların formatı yerine sadece iki işlemin varlığı ile ilgilenir: iki kayıt arasındaki uzaklığı belirleyen bir uzaklık fonksiyonu ve komşu kayıtları işleyerek bir sonuç üreten bir kombinasyon fonksiyonu [103].

Bellek tabanlı yöntemler sahtekarlık tespiti, müşteri tepkisinin tahmin edilmesi, klinik işlemlerde hastalara uygulanacak tedavi yöntemlerinin seçilmesi gibi bir çok alanda kullanılmaktadırlar.

Bellek tabanlı yöntemlerin güçlü olduğu noktalar şunlardır [102]:

- Kolayca anlaşılabilir sonuçlar üretir.
- Rasgele seçilen, hatta ilgisiz dahi olabilen verilere uygulanabilir.
- Analiz alanlarının çok olduğu durumlarda dahi etkin olarak çalışabilir.
- Eğitim kümesinin oluşturulması basittir.

Bellek tabanlı yöntemlerin zayıf olduğu noktalar da bulunmaktadır. Dezavantajları aşağıdaki gibidir [103]:

- Sınıflandırma ve kestirme işlemleri için kullanıldığında işlem maliyeti yüksektir.
- Eğitim kümesi için büyük miktarlarda yere ihtiyaç vardır.
- Üretilen sonuçlar seçilen uzaklık fonksiyonuna, kombinasyon fonksiyonuna ve komşu adedine doğrudan bağlıdır.

3.9.3. Kümeleme

Kümeleme belli sayıda nesneyi, belli sayıda değişkene bağlı olarak farklı gruplara atamak amacıyla tasarlanmış bir yöntemdir. Gruplar istatistiksel olarak birbirinden farklı ve kendi içlerinde homojen olarak oluşturulurlar.

Buradaki benzerlik, uzaklık ölçüleri, korelasyon ölçüleri ya da nitelik bilgileri dikkate alınarak belirlenir. Gruplar tüm özellikler dikkate alınarak oluşturulmaktadır.

3.9.4. İlişkisel Analiz

İş dünyası, insanların, mekanların ve bütün her şeyin birbirleri ile bağlantı kurduğu bir ilişkiler dünyasıdır. Havayolu şirketleri, kargo şirketleri ve benzeri firmalar şehirleri birbirine bağlar. Haberleşme şirketleri ile diğer müşteriler birbirleri ile telefon ve benzeri şekillerde bağlantı kurarlar. Her kredi kartı müşterisi belirli mağazaları ve lokantaları tercih eder. Benzer şekilde ilişkiler her alanda bolca bulunmaktadır ve bu ilişkiler birçok veri madenciliği tekniğinin kullanamayacağı kadar zengin bilgi içermektedir.

İlişkisel analiz, matematiğin bir alt alanı olan graf teorisi tabanlıdır. Yukarıda anlatılanlara rağmen ilişkisel analiz her tür sorunu çözemez ya da her türlü veri üzerinde uygulanamaz. İlişkisel analizi gerçeklemek üzere birkaç tane yazılım bulunmaktadır ve bu yazılımların çoğu hukuksal alanda özelleşmişlerdir. En basit

ilişkisel analiz aracı olarak ise ilişkisel veri tabanları üzerinde kullanılan SQL gösterilebilir.

İlişkisel analiz yönteminin uygulanmasında ortaya çıkan bir başka sorun ise olaylar arasındaki ilişkilerin ortaya çıkarılmasıdır. Telefon çağrılarının analizinde ilişkiler açıktır: Bir kişi bir başkasını arar ve burada çağrının kendisi ilişkidir. Ancak her durumda ilişkileri bulmak bu kadar kolay değildir, ilişkilerin bulunması için otomatikleşmiş bazı işlemlerin gerçekleştirilmesi gerekebilir.

3.9.5. Karar Ağaçları ve Kural Türetme

Karar ağaçları yöntemi en güçlü ve en yaygın sınıflandırma ve öngörü araçlarından birisidir. Ağaç yapılı yöntemlerin sık kullanılmasının nedeni ise yapay sinir ağlarının tersine ağaç yapılarının kuralları ifade edebilmesinden kaynaklanmaktadır. Kurallar insanların okuyup anlayabileceği herhangi bir dile çevrilebileceği gibi, bir veri tabanında belirli bir kategoriye düşen kayıtların getirilmesi için birer SQL cümlecisi haline de getirilebilir.

Tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağaçları, veri madenciliğinde, kuruluşlarının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri, güvenilirliklerinin daha iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip algoritmalar [104].

Ancak karar ağaçlarının tahmin için kullanıldığı durumlarda tahmin edilecek değişkenin sürekli değerler alması durumunda uygun sonuçlar üretilmemektedir [102].

Sık oynanan bazı oyunlarda olduğu gibi karar ağaçları da bir dizi soru sorup bunların cevapları doğrultusunda hareket ederek en kısa sürede sonuca gider. Karar ağaçları, sorduğu bir soruya gelen cevap ile soracağı diğer soruları belirler. Eğer sorular iyi seçilmiş olursa, yeni gelen bir kaydın sınıflandırılması işlemi, en az sayıda soru sorarak gerçekleştirilebilir.

Sorulacak sorular ve bu sorulara gelebilecek cevapların yönlendirdiği başka soruların bulunduğu bir ağaç yapısı olarak adlandırılan karar ağaçları ile değerlendirme yaparken yeni gelen bir kayıt, ağacın kökünden girer. Kökte test edilen bu yeni kayıt bu test sonucuna göre bir alt düğüme gönderilir. Bu süreç, yeni kayıt herhangi bir yaprak düğüme gelene kadar devam eder. Ağacın belirli bir yaprağına gelen bütün yeni kayıtlar aynı şekilde sınıflandırılırlar. Kökten her bir yaprağa giden sadece tek bir yol vardır. Bu yol, kayıtları sınıflandırmak için kullanılan bir kuralı tanımlamaktadır. Bazı yapraklar aynı sınıflandırmayı yapabilirler fakat her bir yaprak bu sınıflandırmayı farklı nedenlere dayanarak yaparlar [102].

3.9.6. Yapay Sinir Ağları

Yapay sinir ağlarının yaygın kullanımı, veri madenciliği ve karar destek sistemlerinde önceden kanıtlanmış başarılarından gelmektedir. Yapay sinir ağları, sınıflandırma, kümeleme ve kestirme amaçları ile kolaylıkla kullanılabilecek genel amaçlı ve güçlü araçlardır. Ekonomik alanlardan tıbbi konulara, değerli müşterilerin belirlenmesi için yapılan demetleme işlemlerinden kredi kartlarında sahtekarlıkların belirlenmesine kadar çok geniş bir alanda uygulama alanı bulmuştur.

Genel anlamda yapay sinir ağları, beynin bir işlevi yerine getirme yöntemini modellemek için tasarlanmış bir sistem olarak tanımlanabilir. Yapay sinir ağları, yapay sinir hücrelerinin birbirlerine çeşitli şekillerde bağlanmasından oluşurlar ve genellikle katmanlar şeklinde düzenlenir. Donanım olarak elektronik devrelerle ya da bilgisayarlarda yazılım olarak tasarlanabilir. Beynin bilgi işleme yöntemine uygun olarak yapay sinir ağları, bir öğrenme sürecinden sona bilgiyi toplama, hücreler arasındaki bağlantı ağırlıkları ile bu bilgiyi saklama ve genelleme yeteneğine sahip paralel dağılmış işlemcilerdir. Öğrenme süreci, arzu edilen amaca ulaşmak için yapay sinir ağları ağırlıklarının yenilenmesini sağlayan öğrenme algoritmalarını içerir [105].

Yapay sinir ağlarında kullanılan öğrenme algoritmaları veriden üniteler arasındaki bağlantı ağırlıklarını hesaplar. Yapay sinir ağları, sonuçları bilinen belirli veri kümeleri üzerinde algoritmaları çalıştırarak eğitirler. Bunun sonucunda yapay sinir ağının

içindeki ağırlık belirlenir. Bu ağırlıklar yeni gelen verilerin işlenmesinde kullanılır. Karmaşık veri tabanları üzerinde daha iyi sonuçlar üretebilirler [1].

Yapay sinir ağları, istatistiksel yöntemler gibi veri hakkında parametrik bir model varsaymaz, uygulama alanı daha geniştir ve bellek tabanlı yöntemler kadar yüksek işlem ve bellek gerektirmez [106]. Yapay sinir ağları; yapısı, bilgi işleme yöntemindeki farklılık ve uygulama alanları nedeniyle çeşitli bilim dallarının da kapsam alanına girmektedir [107].

3.9.7. Genetik Algoritmalar

Genetik algoritmalar da yapay sinir ağları ve bellek tabanlı yöntemler gibi biyolojik işlemlerden kaynağını almıştır. Biyolojik sistemlerin gelişim sürecini modelleyen genetik algoritma, ilk olarak Holland tarafından önerilmiştir [108]. Sezgisel bir yöntem olan genetik algoritma, problem için optimum sonucu bulamayabilir, ancak bilinen metotlarla çözülemeyen veya çözüm zamanı çok büyük olan problemlerde optimuma çok yakın çözümler vermektedir. Geçtiğimiz yıllar boyunca genetik algoritmalar yapay sinir ağlarının eğitilmesi, bellek tabanlı yöntemlerde kombinasyon fonksiyonunun oluşturulması gibi işlerde kullanılmışlardır.

Genetik algoritmalar açıklanabilir sonuçlar üretirler. Çok değişik tiplerdeki verileri işleme özelliğine sahip olan genetik algoritmalar optimizasyon amacı ile kullanılabilirler. Ayrıca genetik algoritmalar yapay sinir ağları ile ortaklaşa çalışarak başarılı sonuçlar üretmektedirler.

Genetik algoritmaların kullanılmalarında bazı sıkıntılar da yaşanmaktadır. Bunlardan en belirgin olanı karmaşık sorunların genetik kodlanmasının çok zor olmasıdır. Ayrıca optimal sonucun üretildiğine dair bir garanti de bulunmamaktadır. Son olarak genetik algoritmaların çalıştırılması çok ağır bir işlem yükü getirmektedir.

3.9.8. Kaba Küme Yaklaşımı

Kaba küme yaklaşımı, belirsizlik ve anlaşılmazlıkların ele alınması amacıyla, 1980'lerde Pawlak tarafından ortaya çıkarılmış yeni bir matematiksel araçtır [109]. Kaba küme teorisi tam veya doğru olmayan büyük miktardaki gürültülü veriler içerisindeki ilişkileri bulmak amacıyla sınıflandırma işleminde kullanılabilen bir yaklaşımdır. Verilen eğitim kümesi içerisindeki eşdeğer kümeler üzerine kurulmuş bir yaklaşım olup, özellik seçimi için de kullanılabilir [1]. Konu ile ilgili detaylı bilgi çalışmanın 4. bölümünde yer almaktadır.

3.10. Veri Madenciliğinde Karşılaşılan Problemler

Küçük veri kümelerinde hızlı ve doğru bir biçimde çalışan bir sistem, çok büyük veritabanlarına uygulandığında tamamen farklı davranabilir. Bir veri madenciliği sistemi tutarlı veri üzerinde mükemmel çalışırken, aynı veriye gürültü eklendiğinde kayda değer bir biçimde kötüleşebilir. Bu bölümde günümüz veri madenciliği sistemlerinin karşı karşıya olduğu problemler incelenecektir [3].

3.10.1. Veritabanı Boyutu

Veritabanı boyutları inanılmaz bir hızla artmaktadır. Pek çok makine öğrenimi algoritması oldukça küçük örneklemeleri ele alabilecek biçimde geliştirilmiştir. Aynı algoritmaların yüz binlerce kat büyük örneklemelerde kullanılabilmesi için azami dikkat gerekmektedir. Örneklem büyük olması, örüntülerin gerçekten var olduğunu göstermesi açısından bir avantajdır ancak böyle bir örneklemde elde edilebilecek olası örüntü sayısı çok büyüktür. Bu yüzden veri madenciliği sistemlerinin karşı karşıya olduğu en önemli sorunlardan biri veritabanı boyutunun çok büyük olmasıdır. Dolayısıyla veri madenciliği yöntemleri ya sezgisel bir yaklaşımla arama uzayını taramalıdır ya da örnekleme yatay/dikey olarak indirgemelidir.

Yatay indirgeme, nitelik değerlerinin önceden belirlenmiş genelleme sıra düzenine göre, bir üst nitelik değeri ile değiştirilme işlemi yapıldıktan sonra aynı olan çoklukların

çıkarılması işlemidir. Dikey indirgeme, artık niteliklerin indirgenmesi işlemidir. Özellik seçimi yöntemleri ya da nitelik bağımlılık çizelgesi uygulanarak yapılır.

3.10.2. Gürültülü Veri

Büyük veritabanlarında pek çok niteliğin değeri yanlış olabilir. Bu hata, veri girişi sırasında yapılan insan hataları veya girilen değerlerin yanlış ölçülmesinden kaynaklanır. Veri girişi ya da veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir. Ancak günümüzde kullanılan ticari ilişkisel veritabanları veri girişi sırasında oluşan hataları otomatik biçimde gidermek konusunda az bir destek sağlamaktadır. Hatalı veri gerçek dünya veritabanlarında ciddi problem oluşturabilir. Bu durum, bir veri madenciliği yönteminin kullanılan veri kümesinde bulunan gürültülü verilere karşı daha az duyarlı olmasını gerektirir. Eğer veri kümesi gürültülü ise sistem bozuk veriyi tanımalı ve ihmal etmelidir.

3.10.3. Boş Değerler

Bir veritabanında boş değer, birincil anahtarlar yer almayan herhangi bir niteliğin değeri olabilir. Boş değer tanımı gereği kendisi de dahil olmak üzere hiç bir değere eşit olmayan değerdir. Bir çoklukta eğer bir nitelik değeri boş ise o nitelik bilinmeyen ve uygulanamaz bir değere sahiptir. Bu durum ilişkisel veritabanlarında sıkça karşımıza çıkmaktadır. Bir ilişkide yer alan tüm çokluklar aynı sayıda niteliğe, niteliğin değeri boş olsa bile, sahip olmalıdır. Örneğin kişisel bilgisayarların özelliklerini tutan bir ilişkide bazı model bilgisayarlar için ses kartı modeli niteliğinin değeri boş olabilir.

3.10.4. Eksik Veri

Evrendeki her nesnenin ayrıntılı bir biçimde tanımlandığı ve bu nesnelerin alabileceği değerler kümesinin belirli olduğunu varsayalım. Verilen bir bağlamda her bir nesnenin tanımı kesin ve yeterli olsa idi, sınıflama işlemi basitçe nesnelerin alt kümelerinden faydalanılarak yapılardı. Bununla birlikte, veriler kurum ihtiyaçları göz önünde bulundurularak düzenlenip, toplandığından, mevcut veri gerçek hayatı yeterince yansıtmayabilir. Örneğin hastalığın tanısını koymak için kurallar sadece çok yaşlı

insanların belirtilerinin bulunduğu bir veri kümesi kullanılarak üretilseydi, bu kurallara dayanarak bir çocuğa tanı koymak pek doğru olmazdı. Bu gibi koşullarda bilgi keşfi modeli belirli bir güvenlik (ya da doğruluk) derecesinde tahmini kararlar alabilmelidir.

3.10.5. Artık Veri

Verilen veri kümesi, eldeki probleme uygun olmayan veya artık nitelikler içerebilir. Bu durum pek çok işlem sırasında karşımıza çıkabilir. Örneğin, eldeki problem ile ilgili veriyi elde etmek için iki ilişkiyi ortak nitelikler üzerinden birleştirirsek sonuç ilişkide kullanıcının farkında olmadığı artık nitelikler bulunur. Artık nitelikleri elemek için geliştirilmiş algoritmalar *özellik seçimi* olarak adlandırılır. Özellik seçimi, tümevarıma dayalı öğrenmede budama öncesi yapılan bir işlemdir. Başka bir deyişle, özellik seçimi, verilen bir ilişkinin içsel tanımını, dışsal tanımın taşıdığı bilgiyi bozmadan onu eldeki niteliklerden daha az sayıdaki niteliklerle (yeterli ve gerekli) ifadeleyebilmektir. Özellik seçimi yalnızca arama uzayını küçültmekle kalmayıp, sınıflama işleminin kalitesini de artırır.

3.10.6. Dinamik Veri

Kurumsal çevrim-içi veritabanları dinamiktir, yani içeriği sürekli olarak değişir. Bu durum, bilgi keşfi metotları için önemli sakıncalar doğurmaktadır. İlk olarak sadece okuma yapan ve uzun süre çalışan bilgi keşfi metodu bir veritabanı uygulaması olarak mevcut veri tabanı ile birlikte çalıştırıldığında mevcut uygulamanın da performansı ciddi ölçüde düşer. Diğer bir sakınca ise, veritabanında bulunan verilerin kalıcı olduğu varsayıp, çevrimdışı veri üzerinde bilgi keşif metodu çalıştırıldığında, değişen verinin elde edilen örüntülere yansımaları gerekmektedir. Bu işlem, bilgi keşfi metodunun ürettiği örüntüleri zaman içinde değişen veriye göre sadece ilgili örüntüleri yığmalı olarak güncelleme yeteneğine sahip olmasını gerektirir.

Aktif veritabanları tetikleme mekanizmalarına sahiptir ve bu özellik bilgi keşif metotları ile birlikte kullanılabilir.

3.10.7. Farklı Tipteki Verileri Ele Alma

Gerçek hayattaki uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik veya kategorik veri türleri değil, aynı zamanda tamsayı, kesirli sayılar, çoklu ortam verisi, coğrafiik bilgi içeren veri gibi farklı tipteki veriler üzerinde işlem yapılmasını gerektirir. Kullanılan verinin saklandığı ortam, düz bir kütük veya ilişkisel veritabanında yer alan tablolar olacağı gibi, nesneye yönelik veritabanları, çoklu ortam veritabanları, coğrafiik veritabanları vb. olabilir. Saklandığı ortama göre veri, basit tipte olabileceği gibi karmaşık veri tipleri de olabilir. Bununla birlikte veri tipi çeşitliliğinin fazla olması bir veri madenciliği algoritmasının tüm veri tiplerini ele alabilmesini olanaksızlaştırmaktadır. Bu yüzden veri tipine özgü veri madenciliği algoritmaları geliştirilmektedir.

3.11. Veri Madenciliğini Etkileyen Eğilimler

Temel olarak veri madenciliğini 5 ana harici eğilim etkiler. Bunlar:

3.11.1. Veri

Veri, veri madenciliğinin bu kadar gelişmesindeki en önemli etkidir. Son yıllarda sayısal verinin hızla artması, veri madenciliğindeki gelişmeleri hızlandırmıştır. Bu kadar fazla veriye bilgisayar ağları üzerinden erişilmektedir. Bu yüzden, verileri analiz etme yöntemleri ve teknikleri geliştirilmektedir.

3.11.2. Donanım

Veri madenciliği, sayısal ve istatistiksel olarak büyük veri kümeleri üzerinde yoğun işlemler yapmayı gerektirir. Gelişen bellek ve işlem hızı kapasitesi, birkaç yıl önce madencilik yapılamayan veriler üzerinde çalışmayı mümkün hale getirmiştir.

3.11.3. Bilgisayar Ağları

Yeni nesil internet, günümüzde kullanılan bilgisayar ağlarındaki hıza göre çok daha fazla bir sürat ve taşıma kapasitesine ulaşmıştır. Böyle bir bilgisayar ağı ortamı oluştuktan sonra, dağınık verileri analiz etmek ve farklı algoritmaları kullanmak mümkün olacaktır. Bundan 10 yıl önceki bilgisayar ağları teknolojisinde hayal edemediklerimizi artık kullanabiliyoruz. Buna bağlı olarak, veri madenciliğine uygun ağların tasarımı da yapılmaktadır.

3.11.4. 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. Çünkü veri madenciliği sayesinde müşterilerin ve müşteri faaliyetlerinin yarattığı fırsatlar daha kolay tespit edilebilmekte ve riskler daha açık görülebilmektedir.

3.11.5. Veri Madenciliği Hangi Tip Verilerle Çalışır?

Veri madenciliği birçok depolama birimi üzerinde uygulanabilmektedir. Bunlar, ilişkisel veritabanları, veri ambarları, geleneksel veri tabanları, gelişmiş veri tabanları ve dosyalar olabilir. Gelişmiş veri tabanı sistemleri arasında, nesneye yönelik, nesne ilişkisel, metin (text) veri tabanları, multi medya veri tabanları sayılabilir. Veri madenciliği tekniklerinin avantajları üzerinde uygulandığı depolama sistemlerine göre değişiklik gösterebilir [1].

3.12. Veri Madenciliği Uygulamaları

“Veri Madenciliği” veri içerisinden aranan bilgiye ulaşma işidir. Veri madenciliği, veritabanı sistemleri, veri ambarları, yapay öğrenme, istatistik, algoritma, veri görselliği, yapay sinir ağları gibi konuları kapsayan disiplinler arası bir alandır. Bu nedenle yeni bir disiplin olmasına karşın uygulama alanı oldukça geniştir.

Günümüzde veri madenciliğinin başlıca ilgi alanları olarak aşağıdakiler sayılabilir [99];

Pazarlama:

- Müşteri segmentasyonunda,
- Müşterilerin demografik özellikleri arasındaki bağlantıların kurulmasında,
- Çeşitli pazarlama kampanyalarında,
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında,
- Pazar sepeti analizinde,
- Çapraz satış analizlerinde,
- Müşteri değerlemede,
- Müşteri ilişkileri yönetiminde,
- Çeşitli müşteri analizlerinde,
- Satış tahminlerinde,

Bankacılık ve Borsa:

- Genel piyasa analizlerinde,
- Alım-satım stratejilerinin optimizasyonunda,
- Pazar segmentasyonunda,
- Müşteri kazanmada,
- Ana giderlerin azaltılmasında,
- Poliçe fiyatlarının belirlenmesinde,
- Hisse senedi fiyat tahminlerinde,
- Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında,
- Kredi kartı dolandırıcılıklarının tespiti,
- Müşteri segmentasyonunda,
- Kredi taleplerinin değerlendirilmesinde,
- Usulsüzlük tespiti,
- Risk analizlerinde,
- Risk yönetiminde,

- Hisse senedi fiyat tahmininde,
- Genel piyasa analizlerinde,
- Alım-satım stratejilerinin optimizasyonunda,

Sigortacılık:

- Yeni poliçe talep edecek müşterilerin tahmin edilmesinde,
- Sigorta dolandırıcılıklarının tespitinde,
- Riskli müşteri tipinin belirlenmesinde,

Perakendecilik:

- Satış noktası veri analizlerinde,
- Alış-veriş sepeti analizlerinde,
- Tedarik ve mağaza yerleşim optimizasyonunda,

Telekomünikasyon:

- Müşteri kazanma ve elde tutma analizlerinde,
- Kalite ve iyileştirme analizlerinde,
- Hile tespitlerinde,
- Hatların yoğunluk tahminlerinde,

Sağlık ve İlaç:

- Test sonuçlarının tahmininde,
- Ürün geliştirmede,
- Tıbbi teşhisde,
- Tedavi sürecinin belirlenmesinde,

Endüstri:

- Kalite kontrol analizlerinde,
- Lojistikte,
- Üretim süreçlerinin optimizasyonunda,

Bilim ve Mühendislik:

- Ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesinde,

Hava Yolları:

- Müşteri ihtiyaçlarını anlamada,
- Promosyonlarda,
- Uçuş rotalarının tespitinde,
- Bakım planlamada,
- Kaza analizlerinde,

Bilgisayar ve internet:

- Görüntü ve metin işlemede,
- Webdeki faydalı bilgiye ulaşmada,
- Elektronik ticarete müşteri portföyü ve alış-veriş eğilimleri hakkında bilgi toplamada uygulanmaktadır.

3.12.1. Uygulamada Kullanılan Sistemler

Veri madenciliği yeni bir alan olmasına karşın piyasada bu alanın kullanıldığı pek çok ürün mevcuttur. Her gün yeni özelliklerin, fonksiyonların, görsel tekniklerin eklendiği yeni ürünler ortaya çıkmaktadır. Kişiler ve organizasyonlar ihtiyaçlarına göre çözümlere yönelmektedirler. Bu çözümlerin seçiminde etkili olan kriterler aşağıdaki gibi sıralanabilir.

- Kullanılan veri tipleri,
- Donanım ve yazılıma yönelik sistem özellikleri,
- Kullanıcı bilgi ve yetenekleri,
- Kullanılan ve kullanılabilecek veri kaynakları,
- Veri analizi için kullanılan fonksiyonlar ve yöntemleri,
- Nihai çıktıda istenilen format ve görsellik yönü.

3.12.1.1. Ticari Veri Madenciliği Sistemi Örnekleri

Veri madenciliğine dair yazılımların sayısı her gün artmaktadır. Şu anda piyasada bulunan bazı yazılımlardan burada bahsedilmiştir.

Intelligent Miner

İlişkilendirme, sınıflama, tahmin modelleme, gruplama, sıralı desen analizi, regresyon analizi gibi fonksiyonları içeren IBM ürünü bir sistemdir. Ayrıca yapay sinir ağları algoritmaları, istatistik metotları, veri hazırlama ve görsel gösterim araçları gibi ek özellikler içerir. Ölçeklendirebilir olması ve IBM'in DB2 ilişkisel veritabanı sistemi ile sıkı birleşik olması onu diğerlerinden ayıran özelliklerdir.

Enterprise Miner

Regresyon, sınıflama ve istatistik analiz gibi fonksiyonları içeren SAS ürünü bir sistemdir. İstatistik analiz araçlarının çeşitliliği en önemli özelliğidir.

Mineset

İlişkilendirme, sınıflama gibi fonksiyonları ve istatistik ve görsel araçlar gibi özellikler sahip Silicon Graphics ürünü bir sistemdir. Veriyi pek çok formda gösterebilen grafiksel ara yüzü en önemli özelliğidir.

Clementine

Clementine, kural indirgeme, yapay sinir ağları, sınıflama ve görsel araçlar gibi özellikleri destekleyen, Integral Solutions ürünü bir sistemdir. Clementine tarihi bilgi içeren veriler için idealdir. Kredi kartı dolandırıcılığı tespiti, müşteri davranışı analizi, kimyasal maddelerin aşındırıcılık tahmini... gibi pek çok farklı alanda uygulanabilir.

DBMiner

OLAP analizi, ilişkilendirme, sınıflama ve gruplama algoritmalarını içeren DBMiner Technology ürünü bir sistemdir. Veri küp tabanlı çevrim içi çözümsel madencilik en önemli özelliğidir.

Weka

GNU lisansı altında açık kod bir yazılımdır ve bütünüyle ücretsizdir. %100 Java'da yazıldığı için platform bağımsızdır. Gelişmiş kullanıcı arabirimleri ile hızlı ve rahat uygulama geliştirmeye imkan verir. Barındırdığı geniş algoritma spektrumuyla kıyaslamalı uygulama geliştirmeye imkan sağlar.

Bu çalışmada PolyAnalyst paket programı kullanılmıştır. Program ile ilgili detaylı bilgi çalışmanın 5. bölümünde yer almaktadır.

3.12.1.2. Diğer Ticari Yazılımlar

- Affinium Model
- Alice d'Isoft
- ASA's ModelMAX
- BioComp i-Suite
- BrandWorks
- DataDetective
- Delta Miner
- DMM
- ESX
- EWA Systems
- Fair, Isaac Business Science
- GainSmarts
- GhostMiner
- Gornik
- Insightful Miner
- Kaidara Advisor
- Kensington Discovery Edition

- Kepler
- KnowledgeMiner
- KXEN
- LPA Data Mining
- Magnify PATTERN
- MCubiX from Diagnos
- Microsoft SQL Server 2000
- MERKUR Miner Plus
- Nuggets
- Oracle Data Mining (ODM)
- Partek
- Polyanalyst
- Predictive Data Mining Suite
- Previa
- Purple Insight MineSet
- Quadstone DecisionHouse
- Reel Two
- Salford Systems Data Mining Suite
- SPAD
- StarProbe
- Statistica Data Miner
- Teradata Warehouse Miner
- ThinkCRA
- WITNESS Miner
- XLMiner
- Xpertrule Miner
- Zoom 'n View

3.11.1.3. Ücretsiz Sistemler

- ADAM, Algorithm Development and Mining
- Databionic ESOM
- Gnome

- MiningMart
- MLC++
- Machine Learning in Java (MLJ)
- Orange
- StarProbe
- Superinduction
- TANAGRA
- YALE

4.BÖLÜM

KABA KÜME YAKLAŞIMI

4.1. Giriş

Gerçek hayatta veri hakkındaki bilgi, verinin neyi ifade ettiğini, nasıl açıklanacağını ve kullanılacağını tanımlar. Bu bilgi, denklemler, algoritmalar, karar tabloları veya ağaçlar kullanılarak ifade edilir. Birçok durumda bilgi mantık kuralları ile ifade edilmektedir. Kuralların kullanılmasıyla, daha kesin olmayan bilgi ifade edilirken, bilginin anlaşılması kolaylaşmaktadır.

Son yıllarda veri madenciliğinde özellikle belirsizliğin söz konusu olduğu veri kümelerinin analizinde kaba küme (KK) yaklaşımı kullanılmaya başlanmıştır. KK yaklaşımı, yaygın olarak kullanılan birçok istatistiksel yöntem gibi varsayımlara gerek duymadan özellikleri bilinen nesnelerin sınıflanmasına olanak sağlamaktadır ve bu yaklaşımla elde edilen sınıflama kuralları herkes tarafından kolaylıkla anlaşılabilir biçimdedir. Bu yaklaşım, sayısal veya nicel verilerin yanı sıra, kategorik (sembolik, nitel) değerlere sahip kesikli değişkenlerin de analize dahil edilmesini sağlamaktadır [110].

Birçok gerçek yaşam probleminin kayıtlarında açıkça ifade edildiği gibi, KK yaklaşımının uygulamada çok yararlı olduğunu kanıtlanmıştır. KK yaklaşımının kullanılabileceği ana problemler; verilerin indirgenmesi, veri bağımlılıklarının keşfi,

verilerin öneminin tahmini, verilerden karar algoritmalarının oluşturulması, verilerin yaklaşık sınıflaması, verilerdeki benzerlik ve farklılıkların keşfi, verilerdeki örüntülerin keşfi ve neden-sonuç ilişkilerinin keşfini kapsamaktadır [56-58]. Bu tez çalışmasında KK yaklaşımı verilerin indirgenmesi sürecinde kullanılmıştır.

Yapılan literatür araştırmasına bağlı olarak KK yaklaşımının tıp, eczacılık, işletme, bankacılık, finans, mühendislik, meteoroloji, görüntü işleme, ses tanımlama, karar analizi, karakter tanımlama ve diğer gerçek yaşam problemlerinde başarıyla uygulandığı görülmüştür [59].

4.2. Kaba Kümeler

Kaba küme (KK) yaklaşımı, 1980'lerin başlarında Zdzislaw Pawlak tarafından geliştirilen büyük miktarda ve eksik verilerden, belirsizliğin olduğu ortamlarda kullanılan matematiksel bir yaklaşımdır [111,57,112,113,114]. Diğer yöntemlerin aksine KK yaklaşımı sadece veri içerisindeki bilgiyi kullanır, istatistiksel parametrelere veya belirli varsayımlara dayanmamaktadır.

Kaba küme yaklaşımının önemi, bilgi kaynaklarındaki eksik veriler üzerinden örüntülerin keşfedilmesidir. Verilerin sınıflandırılması ve nitelik indirgemesi en önemli kullanım alanlarındandır [109].

Bu yaklaşım yapay zeka tekniklerinde karşılaşılan belirsiz ve şüpheli bilgilerle uğraşmakta kullanılmak üzere geliştirilmiş ve bilgi keşfinde karşılaşılan bir çok problemin çözülmesine temel sağlamıştır [115].

Kaba küme yaklaşımı, kümenin tek olarak elemanları ile tanımlandığı ve kümenin elemanları hakkında ilave hiçbir bilginin bulunmadığı klasik küme yaklaşımının aksine bir kümenin tanımlanması için başlangıçta evrenin elemanları hakkında bazı bilgilere gereksinim olduğu varsayımına dayanan bir yaklaşımdır. Nesneler, aynı bilgi ile nitelendiriliyorlarsa aynıdır veya ayırt edilemezler. Ortaya konulan ayırt edilememe ilişkisi, KK yaklaşımının temelini oluşturur. Bütün aynı nesnelerin kümesine temel küme denir ve bilginin temel taşı oluşturur. Temel kümelerin herhangi bir birleşimine

kesin küme adı verilir, aksi takdirde bir küme kabadır. Her KK'nin kesinlikle kümenin kendisinin ya da tümleyen kümesinin elemanları olarak sınıflandırılmayan elemanları (sınır hattı elemanları) vardır [116].

4.3. Kaba Küme Yaklaşımında Temel Kavramlar

Kaba küme yaklaşımının çıkış noktası, ilgilenilen her nesne ile bazı enformasyonun (veri, bilgi) bağdaştırılmasıdır. Nesnelerin yalnızca ilgili enformasyon ile görülebileceği varsayımıyla bilginin taneli bir yapısı olduğu görüşü ortaya çıkmaktadır. Kaba küme yaklaşımında belirsizlik, evrenin bazı elemanları hakkında bilgi olmayışından kaynaklanır.

KK yaklaşımında herhangi bir belirsiz kavram, bir kesin kavramlar çifti ile tanımlanır ve bunlara belirsiz kavramın alt ve üst yaklaşımları adı verilir. Alt yaklaşım (lower approximation) kesin olarak kavrama ait olan bütün nesnelerden oluşur. Üst yaklaşım (upper approximation) ise kavrama ait olması muhtemel bütün nesneleri içerir. Alt ve üst yaklaşımlar arasındaki fark ise sınır bölgesini oluşturmaktadır [58].

4.3.1. Bilgi Sistemi

Kaba küme yaklaşımı ile veri analizinin başlangıcı, bilgi sistemi olarak adlandırılmış olan veri kümesidir. Bilgi sistemi, tablo ile temsil edilmektedir. Bu tabloda her bir satır nesneyi; her bir sütun ise her nesne için ölçülebilen veya insan deneyimi (uzmanlığı) ile elde edilen bir niteliği temsil etmektedir [111-59-117-118]. Bir bilgi sistemi (BS) aşağıdaki gibi tanımlanır:

$$BS = (U, A) \quad (4.1)$$

Burada U ; boş olmayan sonlu nesneler kümesini, A ise bilgi sisteminde bulunan özelliklerin sonlu bir kümesini temsil etmektedir.

$$U = \{x_1, x_2, x_3, \dots, x_n\} \quad (4.2)$$

$$A = \{a_1, a_2, a_3, \dots, a_n\} \quad (4.3)$$

Her bir özellik $a \in A$ için,

$f_a: U \rightarrow V_a$ bir bilgi fonksiyonunu tanımlar. V_a a 'nın değer kümesidir. Tablo 4.1'de bir bilgi sistemi örneği verilmiştir.

Tablo 4.1. Örnek bir bilgi sistemi.

U	Yaş	Kilo
x_1	16-30	41-50
x_2	16-30	31-40
x_3	31-45	51-60
x_4	31-45	51-60
x_5	46-65	61-70
x_6	16-30	61-70
x_7	46-65	61-70

Kaba küme yaklaşımının terminolojisi kullanılarak bu veri setini bir bilgi sistemi olarak düşündüğümüzde:

Nesne uzayı; $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$,

Özellik seti; $A = \{Yaş, Kilo\}$,

Özelliklerin alanı ise; $V_1 = \{16-30, 31-45, 46-65\}$, $V_2 = \{31-40, 41-50, 51-60, 61-70\}$.

Bilgi sistemi tablosundan, x_3 ve x_4 ile x_5 ve x_7 nesnelerin almış olduğu özellik değerlerinin aynı olduğu görülmektedir. Bu örnekler mevcut özelliklere göre ayırt edilemez yani benzerdirler.

4.3.2. Karar Tablosu

Birçok uygulamada bilinen bir sınıflandırma çıktısı vardır. Karar niteliği olarak adlandırılan bu ayırt edici nitelik, deney sonrasında bir bilgi edinilmesini sağlar. Bilgi sistemine bir uzman tarafından ilave edilen karar niteliği sonucunda karar tablosu

oluşur. Karar tablosu $A = \{U, a \cup \{d\}\}$ biçiminde gösterilir ve burada $d \notin A$ olup karar özelliğini belirtir. Tablo 4.2’de örnek bir karar tablosu sunulmaktadır.

Bir karar sistemi, bilgi sisteminin özel bir türü olduğundan bilgi sistemleri için geliştirilen yaklaşım karar sistemlerine de uygulanabilir [111].

Tablo 4.2. Örnek bir karar tablosu.

U	Yaş	Kilo (kg.)	Ağırlıklar
x_1	16-30	41-50	Normal
x_2	16-30	31-40	Düşük
x_3	31-45	51-60	Normal
x_4	31-45	51-60	Düşük
x_5	46-65	61-70	Normal
x_6	16-30	61-70	Yüksek
x_7	46-65	61-70	Normal

Karar tablosundan hem x_3 ve x_4 nesneleri hem de x_5 ve x_7 nesneleri aynı şart özelliklerine göre tamamen aynı değerlere sahiptirler. Fakat (x_3, x_4) örnek çifti farklı karar sonuçlarına sahip iken (x_5, x_7) örnek çifti ise aynı sonuç değerlerine sahiptir. Bu şekilde karar tablosundaki tanımlamalar birleştirilerek kural biçimi elde edilmiş olacaktır. “Eğer yaş 46-65 arasında ve kilo 61-70 kg arasında ise normal ağırlıktadırlar” kuralı karar tablosundan kolaylıkla yazılabilir.

4.3.3. Ayırt Edilemezlik Kavramı

Kaba küme yaklaşımının başlangıç noktası, ilgilenilen nesneler hakkında bilgiler tarafından meydana getirilen ayırt edilmezlik ilişkisidir. Ayırt edilmezlik ilişkisi, bilgi eksikliği yüzünden sahip olduğumuz bilgilerle bazı nesneleri ayırt edemediğimiz gerçeğini ortaya koymayı amaçlamaktadır. Ayırt edilmezlik ilişkisi genel matematiksel ifadeler kullanılarak formüle edilebilir [68].

Ayırt edilmezlik ilişkisi $I(B)$ notasyonu ile gösterilir ve aşağıdaki şekilde tanımlanır.

$$I(B) = \{(x,y) \in UXU : f_a(x) = f_a(y), \forall a \in B\} \quad (4.4)$$

Ayırt edilemezlik ilişkisi, bir eşitlik ilişkisi olarak tanımlanmaktadır. $I(B)$ ilişkisinin eşitlik sınıflarının tamamını içeren aile, U/B olarak gösterilmektedir. Herhangi (x) elemanına sahip olan eşitlik sınıfı ise $B(x)$ olarak gösterilmektedir. Eğer (x,y) , $I(B)$ ye ait ise x ve y ayırt edilemezdir. $I(B)$ ilişkisinin eşitlik sınıfları B temel kümeleri olarak adlandırılmaktadır. Temel kümeler, kaba kümeleme yaklaşımında gerçekte ilgili bilgilerin temel kavramıdır.

Tablo 4.2’de verilen örnek karar tablosu için ayırt edilemezlik ilişkisinin nasıl tanımlanacağını gösterecek olursak; bu tabloda boş olmayan alt kümeler $\{Yaş\}, \{Kilo\}, \{Yaş, Kilo\}$ şeklindedir.

$$I(\{Yaş\}) = \{(x_1, x_2, x_6), (x_3, x_4), (x_5, x_7)\}$$

$$I(\{Kilo\}) = \{(x_1), (x_2), (x_3, x_4), (x_5, x_6, x_7)\}$$

$$I(\{Yaş\}, \{Kilo\}) = \{(x_1), (x_2), (x_3, x_4), (x_5, x_7), (x_6)\}$$

ve karar özelliği (U/D) için üç farklı ağırlık sonucuna göre (Normal, Düşük, Yüksek) temel küme:

$$I(\{Ağırlıklar\}) = \{(x_1, x_3, x_5, x_7), (x_2, x_4), (x_6)\} \text{ şeklinde elde edilir.}$$

4.3.4. Küme Yaklaşım Kavramı

Kaba küme yaklaşımının temel fikri $I(B)$ ikili bağıntısını kullanarak kümelerin yaklaşımını oluşturmaktır. Eğer X , A özelliklerini kullanarak kesin bir biçimde tanımlanamıyorsa alt ve üst yaklaşımlarla gösterilirler [111].

Tablo 4.2’de yer alan örnek karar tablosundaki x_3 ve x_4 nesnelerinin mevcut özellikleri ile tam olarak hangi kümeye ait olduklarına karar verilemez. Bu noktada KK yaklaşımı devreye girer. Bu iki nesne tam olarak tanımlanamaz ancak, hangi sınırlar arasında

oldukları belirlenebilir. Eğer sınır bölgesi boş değilse kümeleme işlemi kaba olarak adlandırılır. Bu kavramlar şu şekilde ifade edilirler:

$A = (U, A)$ bir bilgi sistemi, $B \subseteq A$ ve $X \subseteq U$ olsun. X 'in alt yaklaşımı $B_x(X)$ ve üst yaklaşımı $B^x(X)$ şeklinde gösterilir. Bu yaklaşımlarda şu şekilde tanımlanır:

$$B_x(X) = \{x \in U : B(x) \subseteq X\}$$

$$B^x(X) = \{x \in U : B(x) \cap X = \emptyset\}$$

$BN_B(X) = B^x(X) - B_x(X)$ ise sınır bölgesi ve $U - B^x(X)$ ise çıkış bölgesi olarak adlandırılmaktadırlar.

$B_x(X)$ 'deki nesneler B 'deki bilgi temeline göre kesin olarak X 'in elemanları iken, $B^x(X)$ 'deki nesneler B 'deki bilgi temeline göre sadece X 'in muhtemel elemanlarıdır.

KK yaklaşımında $\alpha_B(X)$ olarak adlandırılan tahmin doğruluğu aşağıdaki notasyonla hesaplanır. $|X|$, X 'in önemini göstermektedir. $X = \emptyset$ ve $0 \leq \alpha_B(X) \leq 1$. Eğer $\alpha_B(X) = 1$ ise X kümesi B 'ye göre kesin kümedir. Tersî söz konusu ise yani $\alpha_B(X) < 1$ ise X , B 'ye göre kaba kümedir [68].

$$\alpha_B(X) = \frac{|B_*(X)|}{|B^*(X)|} \quad (4.5)$$

4.3.5. Kaba Üyelik

Klasik küme teorisinde bir eleman bir kümeye ya aittir ya da ait değildir. Üyelik fonksiyonu klasik küme için 1 veya 0 değerlerini alır. KK yaklaşımında ise, üyelik fonksiyonu X kümesi ile X 'in ait olduğu denklik sınıfı arasındaki nispi örtüşmenin derecesini ölçer ve aşağıdaki şekilde tanımlanır [111].

$$\mu_x^B(X) = \frac{|X \cap B(X)|}{|B(X)|} \text{ ve } \mu_x^B(X) \in [0,1] \quad (4.6)$$

Burada dikkat edilmesi gereken üyelik fonksiyonunun, bulanık üyelik fonksiyonundan farklı olarak değerinin tahmin edilmeyip eldeki verilerden hesaplanarak elde edilmesidir. Kaba üyelik fonksiyonu, yaklaşımları ve kümenin sınır bölgesini tanımlamak için aşağıda gösterildiği biçimde kullanılır.

$$B_x(X) = \{x \in U : \mu_x^B(X) = 1\} \quad (4.7)$$

$$B^x(X) = \{x \in U : \mu_x^B(X) > 0\} \quad (4.8)$$

$$BN_B(X) = \{x \in U : 0 < \mu_x^B(X) < 1\} \quad (4.9)$$

4.3.6. Özellik İndirgeme

Veri tablolarında çoğu zaman gereksiz veriler bulunabilmektedir. Bu gereksiz verilerin çıkarılıp çıkarılmayacağı sorunuyla karşılaşılabilir. Bir veri tablosundan bazı verilerin, veri tablosunun temel özellikleri değişmeden atılabilmesi nitelik seçimini ifade etmektedir.

Küçük nitelik kümeleri kullanılarak aynı bağımlılık derecesi ve yaklaşım kesinliği bulunabilir. Bu fikri daha iyi ifade etmek için B , A kümesinin bir alt kümesi ve a niteliği B 'ye ait olsun; eğer $I(B) = I(B - \{a\})$ ise a niteliği B 'de gereksizdir, aksi halde a niteliği B 'de gereklidir. B bütün nitelikleri gerekli ise bağımsızdır denir. Eğer B bağımsız ve $I(B') = I(B)$ ise B 'nin indirgenmiş B' altkümesidir. İndirgenmiş özellikler kümesi, atılabilir özellikler içermeyen alt kümelerdir. Bu kümelerin birkaç önemli özelliği vardır. Bunlardan en önemlilerinden birisi çekirdek (öz-core) özelliğidir. $Core(B) = \cap Red(B)$ şeklinde ifade edilir. $Red(B)$; B 'nin tüm indirgenmiş kümeleridir [116].

4.3.7. Karar Kurallarından Sonuç Çıkarma ve Karar Analizi

KK yaklaşımının en önemli özelliklerinden birisi görsel gözlemlerin yapılamadığı durumlarda, yeni nesnelerin sınıflandırılması için verilen bilgi sisteminden kurallar üretilmesidir. Kurallar “eğer şart veya şartlar ise karar” şeklinde gösterilmektedir. Bir karar kuralı daima tek ve kısmen bağımsız bilgiler içermektedir [59].

Karar analizlerinin amacı, iki temel soruyu yanıtlamaktır. Bunlardan birincisi; kararın alınması sırasında içinde bulunulan koşullara göre kararın niteliğini açıklamak, ikincisi ise; özel koşullar altında nasıl bir karar verileceğine bir çözüm sunmaktır. KK yaklaşımı bu sorulara cevap vermek için matematiksel araçlar sunmaktadır ve bu tür problemlerin analizine uygundur [68, 120, 121].

4.4. Kaba Küme Yaklaşımının Avantajları

Aşağıda bahsedilen özellikler KK'lere önemli uygulama avantajları sunmaktadır [111, 68]:

- 1- Büyük hacimli ve herhangi türdeki verilerin ele alınması yeteneği, birçok alt sistemin birçok değişkenler tarafından yönetildiği, alt sistemler arasında ilişki olan ve birbirlerinin performansını etkileyen, bütün sistem için birkaç optimal çözümün olduğu mühendislik analiz ve modellemesinde çok yararlıdır. Tipik uygulama alanları süreç kontrolü, mühendislik tasarımı ve modellemesi, ağ haberleşme sistemlerinin analizi ve spektrum analizidir.
- 2- Doğrusal olmayan veya sürekli olmayan fonksiyonel bağıntıları modelleme yeteneği, karmaşık, çok boyutlu örüntüleri nitelendirmek için güçlü bir yöntem sağlar. Üretilen kurallar ve kullanılan özellikler gereğinden fazla olmadığından örüntüler çok özlü, güçlü ve sağlamdır. Bu özellikler, finansal modelleme, dolaysız pazarlama, fikir oylaması ve kredi modelleri gibi uygulamalarda benzersiz üstünlükler sağlar.
- 3- KK'ler belirli olmayan (non-deterministic) sistemleri tanımlayabilir ve karakterize edebilirler. Karar vermede olasılıklı enformasyonu içerirler. Bu özellikler teşhise yönelik uygulamalarda (tıbbi teşhis gibi) veya mevcut enformasyonun hedef kararları tam olarak nitelendirmek için uygun olmadığı durumlarda uzman sistemlerin oluşturulması için çok yararlıdır.
- 4- KK'lerin bilgiyi anlaşılması kolay mantık örüntüleri olarak göstermeleri; bilginin denetlenmesi ve geçerliliğinin ortaya konulması gereken yerlerde veya kararların kurallara ve destekleyen durumlara uyumlu olduğu yerlerde bu metodun başarılı olmasını sağlamıştır. Keza, örüntülerin hem mantıksal hem de olasılıklı bilgiyi kapsamayı, keşfedilen bilgiye ilave bir bakış açısı sağlar.

5- KK'ler eksik deęerleri yerine koymadan tam olmayan verilerle alıřma, farklı indirgenmiř kmeler arasında geiř, daha ucuz veya alternatif lm kmeleri kullanma ve benzeri yeteneęi, enformasyonun iletilmesi ve izlenmesinde bařarsızlık ihtimali olduęu veya kritik grevler iin gereęinden fazla kontrol stratejileri tasarlama veya ncelik deęiřkenleri tanımlamada harici hedeflerin dahil edilmesine gereksinim olduęu durumlarda ok avantajlıdır.

4.5. Rosetta ile Nitelik İndirgeme

Rosetta (A Rough Set Toolkit for Analysis of Data); veri tablolarını kaba kmeler yaklařımıyla analiz etmeye yarayan bir programdır. Bu program tm veri madencilięi ve bilgi geliřtirme srelerinde kullanılmaktadır. Rosetta bir doktora projesi erevesinde geliřtirilmiř ve dnyadaki tm kullanıcıların kullanımına aık olan bir programdır [121]. Bu alıřmada rosetta programında yer alan veri indirgeme aralarından birisi olan genetik algoritmalar kullanılmıřtır. Genetik algoritma dıřında yer alan Johnson, Holte 1 ve Rses algoritmaları da ayrıca denenmiř ancak bulunan veri kmelerinin Genetik algoritma ile yapılan indirgemeler sonucunda elde edilen veri kmelerinin birer alt kmesi olduęu tespit edilmiřtir. Bu nedenden dolayı bu tez kapsamında sadece genetik algoritmalar ile yapılmıř olan veri indirgemenin bahsedilmiřtir.

4.5.1. Genetik Algoritmalar ile Rosetta'da Nitelik İndirgeme

İndirgenmiř veri kmelerini bulmak iin, Rosetta kaba kme yaklařımı ile problem zmlerinde ok gl bir aratır [121, 122]. Genetik Algoritma (GA) seeneęi Rosetta ile nitelik indirgemenin kullanılan bir yntemdir.

GA tabanlı indirgemenin bir bilgi tablosu sisteme girdi olarak verilmektedir. GA optimizasyonu, girdi deęiřkenlerinin indirgenmesi iin uygulanır ve ıktı olarak indirgenmiř veri kmelerinden oluřan bilgi tablosu oluřturulur. Rosetta paket programında GA tabanlı nitelik indirgemenin nasıl yapıldıęı ařaęıda anlatılmıřtır.

Vinterbo ve Øhrn tarafından minimal uygun kümelerin (minimal hitting sets) hesaplanabilmesi için bir GA tanımlanmıştır [123]. Algoritma hem maliyet bilgisini hem de yaklaşık çözümleri desteklemektedir. Algoritmanın uygunluk fonksiyonu $f(B)$ aşağıdaki şekilde tanımlanmıştır [124].

$$f(B) = (1-\alpha) \times \frac{\cos t(A) - \cos t(B)}{\cos t(A)} + \alpha \times \min \left\{ \varepsilon, \frac{||S \text{ in } \delta | S \cap B \neq \emptyset||}{|\delta|} \right\} \quad (4.10)$$

Burada,

S : ayırt edilebilirlik fonksiyonuna uyan nesnelerin kümesi,

α : parametresi alt küme maliyeti ve uygun bölge (hitting fraction) arasındaki ağırlığı,

ε : eşik değeri, yaklaşık çözümlerin durumuyla alakalıdır.

A 'nın B alt kümeleri evrimsel aramaya göre bulunmaktadır. En düşük ε 'na sahip olanlar bir saklı listede toplanır. Saklı listenin genişliği k olarak tanımlanmaktadır. Yaklaşık çözümler iki parametre ε ve k 'ya göre kontrol edilirler. Tüm uygunluk fonksiyonu hesaplama iterasyonlarından sonra, hedef nesneye göre indirgemelerin minimum alt kümesi veri kümesinden bulunmaktadır [124].

5. BÖLÜM

UYGULAMADA KULLANILAN VERİ MADENCİLİĞİ PROGRAMI VE TEKNİKLER

5.1. Giriş

Uygulamada veri madenciliği tekniklerinden sınıflandırma ve ön işleme aracı olarak, karar ağaçları, lineer regresyon, yapay sinir ağları, bağımlılık analizi, bağlantı grafikleri, özellik seçme analizi ve kaba kümeler kullanılmıştır. Bu amaçlara yönelik olarak veri madenciliği yazılımlarından olan, PolyAnalyst, SPSS Clementine ve Rosetta araçları kullanılmıştır. Bunlardan kaba küme ve Rosetta programı bölüm 4’de açıklanmıştır. Diğer araçlar ve teknikler ise aşağıdaki bölümlerde sırasıyla yer almıştır.

5.2. Uygulamada Kullanılan Veri Madenciliği Programları ve Teknikleri

5.2.1. PolyAnalyst Programı

Bu çalışmanın bir parçasında, veri madenciliği yazılımlarından birisi olan, Megaputer Intelligence Inc. kuruluşu tarafından geliştirilmiş, PolyAnalyst programı kullanılmıştır. Otomatik veri analizi için birçok metodu barındıran güçlü bir veri madenciliği sistemidir. Program, temel veri madenciliği fonksiyonları olan sınıflandırma, kümeleme, karar ağacı, sepet analizi gibi karar destek araçlarıyla metin analizi araçlarını hazır olarak kullanıcıya sunmakla birlikte birçok fonksiyonunu da esnek kullanım için barındırmaktadır. Bahsi geçen bu fonksiyonları yerine getirmek için program, lineer regresyon ve yapay sinir ağları gibi yöntemlerin yanı sıra, kendine özgü birtakım

algoritmaları da kullanmaktadır. Genel olarak bütün araçların çalışma prensibi benzerdir. İlgili veri kümesi üzerinde istenilen analiz için gerekli araç eğitim veri kümesi ile analiz edilir ve analiz sonucunda elde edilen kurallar test veri kümesi ile test edilir. Kurallar, ilgili parametreler kümesinden oluşan basit eşitlikler ve yüksek dereceli polinomlardan oluşan matematiksel ifadelerdir. Genel test yöntemi, mevcut veri kümesinin rasgele olarak test ve eğitim kümeleri olarak iki bölüme ayrılarak, eğitim kümesi ile modelin öğretilmesi, test kümesi ile de sonuçların test edilmesi şeklinde gerçekleştirilmektedir.

5.2.1.1. Temel PolyAnalyst Bileşenleri

Programa ait temel bileşenler tablo 5.1’de gösterilmiştir.

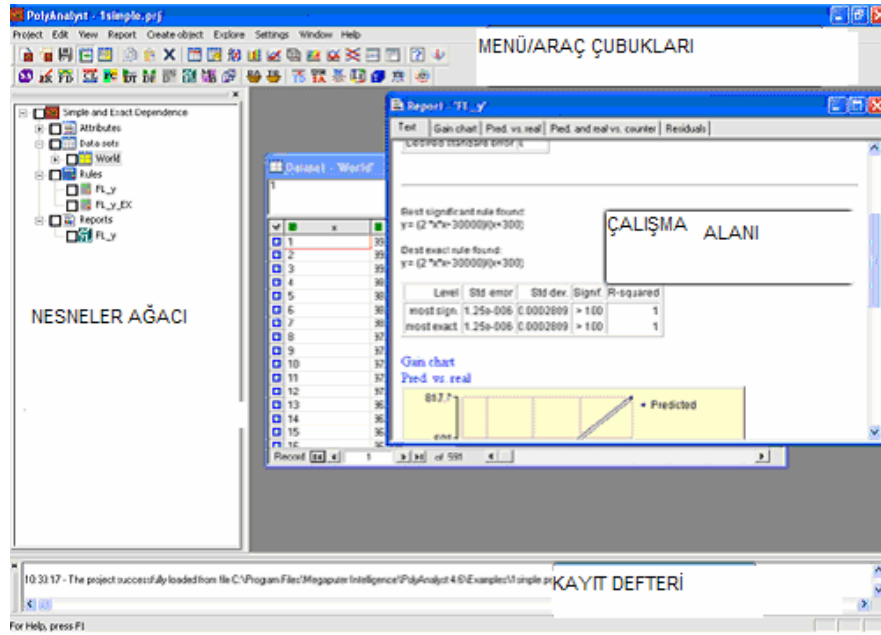
5.2.1.2. PolyAnalyst Ara Yüzü

PolyAnalyst ara yüzü, standart Microsoft Windows iletişim kutuları ve ekranları ile birlikte isteğe göre uyarlanabilir şekildedir. Bazı temel bileşenleri şu şekildedir [125]:

- Menü ve Araç Çubukları: Pencerenin üst kısmında bulunan bazı menüler ve araç çubuklarıdır. Araç çubukları farklı menülere ait kısa yolları içermektedir.
- Nesneler Ağacı: Ekranın sol bölümünde bulunan farklı projede yer alan farklı nesnelere ait ağaçtır. Nesneler, veri kümesi, kurallar, raporlar... olabilir. Bu ağaç nesnelerin yönetilmesine yardımcı olmaktadır. Her bir nesne içerik menüsüne sahiptir.
- Kayıt Defteri: PolyAnalyst içinde yapılan işlemlerle ilgili temel bilgileri içerir ve ekranın alt bölümünde bulunan kısımdır.
- Çalışma Alanı: Ekranın sağ tarafında bulunan ve büyük bölümü kapsayan kısımdır. Nesnelerin ayrıntılarını içeren bölümdür. Örneğin, veri setine ait satır ve sütunların tamamını bu ekranda görülebilmektedir. Aynı şekilde kural ve raporlara da bu alandan ulaşılabilir.

Tablo 5.1. Temel PolyAnalyst bileşenleri.

Proje	Her bir proje belirli bir analiz için bütün parçaları içerir. Projeler, verileri, kuralları, grafikleri ve raporları içeren bütün çalışmaları daha sonra erişilebilecek şekilde kaydeder.
Veri Kümesi	Bir veri kümesi verilerden oluşan bir tablodur. Exceldeki tek bir çalışma sayfasına benzer şekildedir. Sütun ve satırlardan oluşmaktadır. Her bir sütun özel bir tipi içerir. Bu tipler sayılar, kelimeler, evet/hayır, tarihler ve kategorik veriler olabilir. Alt veri kümeleri içeren, birleşim ve kesişimlerden oluşan bir çok veri kümeleri oluşturulabilir.
Kural	Kurallar PolyAnalyst'in sembolik kural dilinden oluşan kurallardır. Kurallar kullanıcılar tarafından oluşturulabilir veya analiz sonucunda oluşturulur. Kurallar PolyAnalyst kural editörü sayesinde tekrar değiştirilebilir. Kurallar veri kümeleri üzerinde yeni sonuçlar elde edebilmek için çalıştırılabilir.
Grafik	Grafikler, diyagramlar ve şemalar değiştirilebilir şekilde kullanılabilirler.
Sonuç Raporu	Raporlar PolyAnalyst içindeki farklı analiz araçlarını çalıştırma sonucunda elde edilen sonuçları içermektedir.
Analiz Aracı	Analizler, makine öğrenme algoritmaları, istatistiksel araçlar... gibi farklı isimlere sahiptirler. Tanımlama, tahminleme, sınıflandırma gibi farklı amaçlar için çeşitli araçlar içermektedir.



Şekil 5.1. PolyAnalyst temel bölümleri.

5.2.1.3. Veri Tipleri

Aşağıda Tablo 5.2’de PolyAnalyst’de yer alan temel veri tipleri açıklanmaktadır:

Tablo 5.2. PolyAnalyst’de veri tipleri.

İsim	Açıklama	Örnekler
Nümerik	Gerçek sayıların her hangi bir tipi	0, 1, 1.2323, 0.444444
Tam Sayı	Tam sayıların herhangi bir çeşidi	0,1,2,-4
Kategorik	Kategorik olarak nitelendirilen sayılar	1, 2, 3
İkili	Doğru/Yanlış	(0,1) veya (doğru ,yanlış) veya (evet,hayır)
Dizi	255 karaktere kadar	1, 2, 3, bayan,erkek
Tarih	Bir Tarih	01/01/1900, 12/31/9999
Metinsel	Herhangi bir karakterler kümesi	Bir paragraf yazı,bir metnin tamamı veya bir kısmı

5.2.1.4. Programda Yer Alan Algoritmalar

Programda kullanılan veri madenciliği algoritmaları mevcut verilerin sınıflandırması amacıyla kullanılmışlardır. Sınıflandırmada temel fikir; bir örneğin (çalışılan veri kümesindeki bir kayıt) üye olduğu sınıfı, o örneğin sahip olduğu tahminleyici nitelik değerlerine göre tahmin etmektir [126]. Başka bir ifadeyle, veri kümesindeki elemanların ortak özelliklerinin bulunması ve sınıflama kurallarına göre bu elemanların farklı sınıflara dahil edilmesi işlemidir [127].

Veri sınıflandırma, veri tabanındaki mevcut nesne kümelerinin ortak özelliklerini tanımlayan ve bu nesneleri farklı sınıflara kategorize eden bir öğrenme sürecidir. Sınıflamada eğitim kümesi olarak adlandırılan örnekler göz önünde bulundurulur. Sınıflandırmanın görevi, öğrenme verisindeki mevcut yapıları kullanarak, her bir sınıf için bir tanımlama veya model kurmaktır. Her bir sınıfın kurulan modeli, sınıfın öğrenme verisine konulmamış kalan diğer verisine, yani test verisine uygulanır [128].

Veri madenciliğinde artan bir ilgiye sahip olan sınıflandırma alanı iki kategoriye ayrılabilir; kural tabanlı sınıflandırma ve kural tabanlı olmayan sınıflandırma [129].

- Kural tabanlı sınıflandırma, verileri analiz ederek doğrudan anlaşılabilir bilgi türeten ve uzman tarafından keşfedilen bilginin doğruluğunun ve faydasının sıklıkla değerlendirilebildiği yaklaşımdır.
- Kural tabanlı olmayan sınıflandırma, bir kara kutu gibi çalışır. İçsel nedenlerden ve etkilerden dolayı kurallar saklı olmasına rağmen kullanıcı genellikle nihai çıktıya erişebilir.

Programda yer alan algoritmalar, bağımlılıkları bulma ve bağlantı analizi algoritmaları ön işleme araçları olarak kullanılmıştır. Sınıflandırma algoritmaları olarak da kural bulma, polynet tahminleyicisi, karar ağaçları ve lineer regresyon araçları kullanılmıştır.

5.2.1.4.1. Bağımlılıkları Bulma

Bağımlılık modelleri değişkenler üzerindeki bağımlılıkları tanımlar. Yapısal ve sayısal olmak üzere iki seviyede gerçekleşirler. Yapısal modeller yerel olarak hangi değişkenlerin bağımlı olduklarını belirler. Sayısal modeller ise sayısal ölçülerle bağımlılıkların kuvvetini gösterir. Bağımlılık modellerine örnek olarak olasılık grafikleri ve bulanık grafikler verilebilir. Diğer bağımlılık modeli yaklaşımlarına istatistiksel analizleri (korelasyon katsayısı, faktör analizi... gibi) ve yapay sinir ağları kullanılarak yapılan duyarlılık analizleri örnek gösterilebilir [101].

Bu çalışmada PolyAnalyst programı algoritmalarından birisi olan bağımlılıkları bulma algoritması kullanılmıştır. Bu algoritma genel olarak ön işleme aracı olarak tasarlanmış, açıklanabilir kuralları olmayan, muhtemel en kuvvetli bağıntıları veya aykırılıkları tespit eden bir algoritmadır. Veri kümesi içerisindeki ilişkileri ve bağımlılıkları hızlı bir şekilde bulmak için kullanılır. Değişkenler arasındaki bağlantının ne kadar güçlü olduğunu gösterir. Bağımlılıkları bulma ile veri kümesinde bulunan kötü veriler elimine edilebilir. Hedef değişkeni sayısal veya ikili (binary) bir parametre olmalıdır. Çıktı olarak, elde edilen raporlarda ilk olarak hedef değişkenin tahmininde etkisi olan en önemli parametreleri vermektedir. Daha sonra ilgili bağımlılıkların şiddetini ifade eden P değeri ile keşfedilen bağımlılıklara uyan veri sayısını göstermektedir.

Çıktı olarak elde edilen değerler içerisinde yer alan P değerine göre bağımlılığın şiddetine karar verilmektedir. Bu değer sıfıra ne kadar yakınsa parametreler arasında o kadar güçlü bağımlılık olduğu sonucu çıkarılmaktadır. PolyAnalyst bağımlılıkları bulma modülü Arnavac adlı algoritmayı kullanır. Bu algoritma veri kümesi içerisindeki karışık parametrelerin, hedef değişken üzerindeki etkisini ölçme esasına dayanır. Arnavac, maksimum bağımlılığa sahip parametre kümesini belirlemek için farklı stratejiler kullanmaktadır [125].

İncelenecek olan veri kümesi, $[N \times M+1]$ boyutlarında olsun. Her bir örnek arama uzayında bir noktaya denk gelir. Burada N toplam kayıt sayısını, $M+1$, ise değişken sayısını ifade etmektedir. Değişkenlerden bir tanesi “bağımlı değişken”, geriye kalan M tanesi de “bağımsız değişken” olsun. R^{M+1} uzayında, bağımsız değişkenler “ x_i ”

$i=1,\dots,M$; bağımlı değişken de “ y ” ile ifade edilsin. Algoritmanın temeli R^M uzayında yer alan değişkenlerin değerlerini $\{E_i\}$ alanlarına bölerek, y değişkeninin her bir E_i alanındaki dağılımının karşılaştırılması esasına dayanır. Y_i , bir E_i kümesinde yer alan bağımlı değişkenin aldığı bütün değerler olsun. Eğer y ’nin bir çok kümedeki dağılımı, toplam veri kümesindeki dağılımından kesin bir farklılık gösteriyorsa, “ y ” bağımlı değişkeni, o kümedeki bağımsız değişkenlere bağımlı kabul edilir. Bu prosedür her bir E_i bölgesine sırayla uygulanarak bütün bağımsız değişkenler kümesi test edilir. Bölünmüş kümelerde yer alan bütün veri kümesine fr -alt kümesi denir. vfr ise fr ’nin N ile bölünmesinden elde edilir. Bağımlılığın şiddetini gösteren P değeri S_{fr} ile ifade edilir ve $S_{fr} = b(N_{vfr}, N, A)$ ifadesi ile hesaplanır. Burada $b(k, K, p)$; binominal dağılımın K örnek sayılı ve p olasılığıyla kuyruk alan olasılığıdır. A da şu şekilde hesaplanır. $A = [(1/N_{nbölüm}) * \sum L_i]$ ’dir. L_i ise, bütün veri kümesi içerisindeki y değişkeninin i bölgesinde yer alan kayıt sayısıdır [130].

5.2.1.4.2. Karar Ağaçları

Karar Ağaçları, sınıflama ve regresyon modellerinde kullanılan başlıca tekniklerden bir tanesidir. Bir karar ağacı ağaç yapısına benzer şekilde kökten yapraklara doğru bir kuralı içeren akış diyagramıdır [131]. Ağaç yapısındaki her düğüm, bir nitelik üzerindeki test işlemini, her dallanma test işleminin sonucunu temsil eder. En son olarak da ağaç sınıflar ile son bulur [1]. Karar ağaçları ile veri kümesi otomatik olarak sınıflandırılmaktadır [29].

İstatistiksel yöntemlerde veya yapay sinir ağlarında veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılması ve yorumlanması oldukça zordur. Karar ağaçları ise oluşturulduktan sonra kökten yaprağa doğru inilerek, *EĞER <Koşul> O HALDE <Tahmin (Sınıf)>* şeklinde kurallar kolayca yazılabilir.

Kurallar genellikle bilginin ifade edilmesinde kullanılır ve insan tarafından kolayca anlaşılabilir. Kural öğrenme, eğitim örnekleri kümesinden kural türetme sürecidir. Kural öğrenmede bir çok algoritma, bir örneği daha önceden belirlenmiş sınıflardan birisine

atamak için kural araştırması yapar [132]. Bu araştırmayı yaparken AQ (Ancestral Quest), CN2, ID3 ve C4.5 gibi algoritmaları kullanırlar. PolyAnalyst programına ait karar ağaçları aracı aşağıdaki bölümlerde anlatılan “ID3” algoritmasını kullanmaktadır. Özellik seçimi ölçüsü olarak da “Information Gain (IG)” bölme kriterini kullanmaktadır.

Karar ağaçları sınıflandırma problemlerinin çözümünde en iyi çalışan algoritmadır. Bu metotlar aynı zamanda sayısal bağımlılıkların keşfi için de uygulanabilirler. Ancak sınıflandırma problemlerinin çözümüne göre daha az etkindir. Bu yüzden PolyAnalyst karar ağaçları algoritması sadece sınıflandırma işlemleri için kullanılmıştır. Problemi çözümleme sürecinde sınıflandırma algoritması olarak kullanılan karar ağaçları, örneklerin çoklu kategorilere ayrılması işlemini gerçekleştirmektedir. Karar ağaçları PolyAnalyst programının büyük miktarda veri ile çalışan en hızlı algoritmasıdır. Hedef değişkenleri kategorik veya ikili (binary) özellikli olabilir. Çıktı olarak sınıflandırma istatistikleri, karar ağacı diyagramı ve gerçek/tahmin edilen verileri gösteren tabloları vermektedir [125].

- **ID3 Algoritması**

ID3 algoritması ilk olarak J. Quinlan Ross tarafından Sydney Üniversitesi’nde geliştirilmiştir [133]. ID3 makine öğrenmesi ve bilişim teorisine dayanarak, verilen örnekler içerisinde en ayırıcı olan değişkeni bulan bir algoritmadır [134]. ID3 algoritması, sabit örnek kümesinden karar ağacı türeten bir algoritmadır. Nihai karar ağacı gelecekteki örneklerin sınıflarını tahmin etmek için kullanılır. ID3 algoritması karar ağacında, her bir düğüm hedef-olmayan bir niteliği, düğümler arasındaki her yay ise niteliğin olası bir değerine karşılık gelir. Ağacın bir yaprağı, bu yapraktan köke kadar ki yolda tanımlanan kayıtlar için hedef niteliklerin beklenen değerini belirler. Karar ağacında her bir düğüm kökten başlayarak yol üzerinde henüz dikkate alınmamış olan nitelikler arasından en çok bilgi sağlayan hedef-olmayan nitelik ile ilişkilendirilebilir. Bu “iyi” bir karar ağacının nasıl olduğunu gösterir.

ID3 algoritmasında, ayırtırmaya rehberlik edecek bir bilgi ölçüsü (entropi) kullanılmıştır [135]. Entropi, eldeki bilginin sayısallaştırılmasıdır. ID3 diğer

değişkenler içinde sınıflamada en ayırıcı özelliğe sahip değişkeni bulurken entropi kavramından yararlanır [136]. Entropi kavramı Claude Shannon tarafından ilk kez Bilgi Teorisi içinde tanımlanmıştır [137]. ID3 algoritması, veri tabanı bölümlenmeden önce, doğru sınıflandırma yapmak için gelen bilgiyle, veritabanı bölündükten sonra doğru sınıflandırma için gelen bilgi arasındaki farkı kullanarak, öncelikli düğüme ve dallanmalara karar verir. Bu aradaki fark ise kazanım (gain) olarak adlandırılır [138]. ID3 algoritması, bilgi kazancını maksimize eden bir karakteristiği kök düğüm olarak seçer. Her düğüm için ilgili alt setin bilgi kazancı hesaplanarak, bu düğüme göre alt setlere ayırıştırma işlemine, her düğümde tek sınıf kalıncaya kadar devam edilir [135]. Bilgi, entropi ve kazanç kavramları ile ilgili notasyonlardan önce ID3 algoritmasının şekil 5.2’de gösterilen temel stratejisini verecek olursak:

- Eğitim kümesinde yer alan örnekleri temsil eden başlangıç düğümü ile başlar (1. adım).
- Eğer bütün örnekler aynı sınıftaysa, düğüm bir yaprak olur ve sınıf adıyla etiketlenir (2. ve 3. adım).
- Aksi halde, entropi temelli, bilgi kazancı olarak bilinen değişken seçme metodu ile örnekleri sınıflara ayırır (6.adım). Bu değişken düğümde test değişkeni olur (7. adım).
- Test değişkeninin her bir değeri için bir dal oluşturulur ve örnekler sırayla bölümlenmiş olur (8-10. adım).
- Algoritma, aynı süreci karar ağacını oluşturmak için, tekrarlı olarak her bir parçadaki örneklere uygular. Bir değişken bir düğümde yer aldıysa, devamında gelen yeni dallarda tekrar yer alması gerekmez (13. adım).
- Bu tekrarlamalı süreç aşağıdakilerden bir tanesinin gerçekleşmesi durumuna kadar devam eder;
 - (a) Bütün örnekler verilen düğüm için aynı sınıfa aitse (2.ve 3. adım), veya
 - (b) Geriye bölümlenecek örneklerle dolu bir değişken yoksa (4. adım). Bu durumda çoğunluğu seçme işlemi uygulanır (5.adım). Çoğunluğun olduğu sınıfın ismi verilen bir yaprak oluşturulur.
 - (c) Test değişkeninin a_i durumu için hiçbir örnek yoksa (11. adım). Bu durumda örnekler içinde çoğunluğun yer aldığı sınıfla yeni bir yaprak oluşturulur (12. adım).

S kümesi s örnekten oluşan bir küme olsun. Hedef değişken m sabit değerden oluştuğunu farz edersek, C_i ($i=1, \dots, m$) sabit sınıfları temsil etsin. s_i ; C_i sınıfında S kümesinden yer alan örnek sayısıdır. Mevcut S örnek kümesini sınıflandırmak için gerekli olan bilgi;

$$I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m p_i \log_2(p_i), \quad (5.1)$$

şeklindedir.

- (1) Bir başlangıç N düğümü oluştur;
- (2) Eğer tüm örnekler aynı sınıfdaysa, C , o halde
- (3) N düğümüne yaprak düğümü olarak geri dön ve C sınıfı olarak adlandır;
- (4) Eğer değişken listesi boş ise
- (5) N düğümüne yaprak düğümü olarak geri dön ve örnekler içinde en yaygın olan sınıfla etiketlendir;
- (6) değişken listesinde yer alan değişkenlerden en yüksek kazanç değerine sahip değişkeni test değişkeni olarak seç;
- (7) N düğümünü test değişkeninin ismini ver;
- (8) test değişkeninin bilinen a_i değerlerinin her biri için
- (9) N kökünden test değişkeninin a_i durumu için bir dal oluştur;
- (10) s_i, a_i durumunun olduğu örnekler kümesi olsun;
- (11) Eğer s_i kümesi boş ise;
- (12) örnekler içindeki en yaygın olan örneklerle adlandırılmış bir yaprak ekle;
- (13) Aksi halde, karar ağacı tarafından oluşturulmuş düğümü ekle.

Şekil 5.2. Temel ID3 algoritması.

Burada p_i , C_i kümesine ait olan rasgele seçilmiş örneklerin olasılıklarıdır ve (s_i/s) şeklinde hesaplanır. Bilgi bitlerle kodlandığı için 2 tabanlı logaritma kullanılmaktadır.

A değişkeni v adet değer içersin $\{a_1, a_2, \dots, a_v\}$. A değişkeni, S kümesini v kadar alt kümeye bölmek için kullanılabilir, $\{S_1, S_2, \dots, S_v\}$, burada S_j , S kümesindeki A değişkeninin a_j değerlerini içermektedir. Eğer A , test değişkeni olarak seçilirse bu alt kümeler dallarda yer alacaktır. s_{ij} , S_j altkümesinde yer alan C_i sınıfına ait eleman sayısı olsun. Entropi veya beklenen bilgi aşağıdaki notasyonla bulunur:

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{mj}}{s} I(s_{1j}, \dots, s_{mj}) \quad (5.2)$$

S_j için,

$$I(s_{1j}, s_{2j}, \dots, s_{mj}) = - \sum_{i=1}^m p_{ij} \log_2(p_{ij}) \quad (5.3)$$

Burada, $p_{ij} = \frac{s_{ij}}{|S_j|}$, S_j kümesindeki bir örneğin C_i sınıfına ait olma olasılığıdır.

A değişkeni için istenilen kazanç değeri $G(A)$ ise şöyle hesaplanır:

$$G(A) = I(s_1, s_2, \dots, s_m) - E(A), \quad (5.4)$$

Algoritma, her bir değişken için kazanç değerini hesaplamaktadır. En yüksek kazançta sahip değişken test değişkeni olarak belirlenir. Bir kök oluşturulur ve değişken ismiyle etiketlenir. Dallar her bir değişken değeri için oluşturulur ve örnekler bu şekilde bölümlenir.

Algoritmanın işleyişinin açıklanabilmesi için basit bir örnek ele alınmıştır. Tablo 5.3’de bir elektrik şirketinin müşteri veri tabanından alınan bilgiler yer almaktadır. Hedef sınıf değişkeni iki değer alan (evet/hayır) müşterilerin bilgisayar alıp almadığıdır ($m=2$). C_1 sınıfı bilgisayar alanlar, C_2 sınıfı ise bilgisayar almayanlar olsun. Bilgisayar alan 9, almayan 5 örnek mevcuttur [1]. Kazanç değerini hesaplamadan önce 5.1 eşitliği kullanılarak sınıflandırma için beklenen bilgi hesaplanır.

$$I(s_1, s_2) = I(9, 5) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.940$$

Daha sonra her bir değişken için entropi değerini hesaplamamız gerekmektedir. İlk olarak *yaş* değişkenini ele alalım. Yaş sütunu için tablodan, her bir örneğin aldığı evet/hayır dağılımı dikkate alınmalıdır.

Yaş = “≤30” için:

$$s_{11} = 2, s_{21} = 3 \quad I(s_{11}, s_{21}) = 0.971,$$

Yaş = “31...40” için:

$$s_{12} = 4, s_{22} = 0 \quad I(s_{12}, s_{22}) = 0,$$

Yaş = “>40” için:

$$s_{13} = 3, s_{23} = 2 \quad I(s_{13}, s_{23}) = 0.971.$$

5.2 eşitliğini kullanarak yaş değişkeni için entropiyi aşağıdaki gibi hesaplarız:

$$E(\text{yaş}) = \frac{5}{14} I(s_{11}, s_{21}) + \frac{4}{14} I(s_{12}, s_{22}) + \frac{5}{14} I(s_{13}, s_{23}) = 0.694.$$

Tablo 5.3. Elektrik şirketine ait veri tabanından eğitim kümesi örnekleri.

Müşteri No	Yaş	Gelir	Öğrenci	Kredi Oranı	Sınıf (bilgiyasar alan ve almayan)
1	≤30	yüksek	hayır	orta	hayır
2	≤30	yüksek	hayır	çok iyi	hayır
3	31...40	yüksek	hayır	orta	evet
4	>40	orta	hayır	orta	evet
5	>40	düşük	evet	orta	evet
6	>40	düşük	evet	çok iyi	hayır
7	31...40	düşük	evet	çok iyi	evet
8	≤30	orta	hayır	orta	hayır
9	≤30	düşük	evet	orta	evet
10	>40	orta	evet	orta	evet
11	≤30	orta	evet	çok iyi	evet
12	31...40	orta	hayır	çok iyi	evet
13	31...40	yüksek	evet	orta	evet
14	>40	orta	hayır	çok iyi	hayır

Yaş için kazanç değeri ise 5.4 eşitliği ile bulunur.

$$G(\text{yaş}) = I(s_1, s_2) - E(\text{yaş}) = 0.246$$

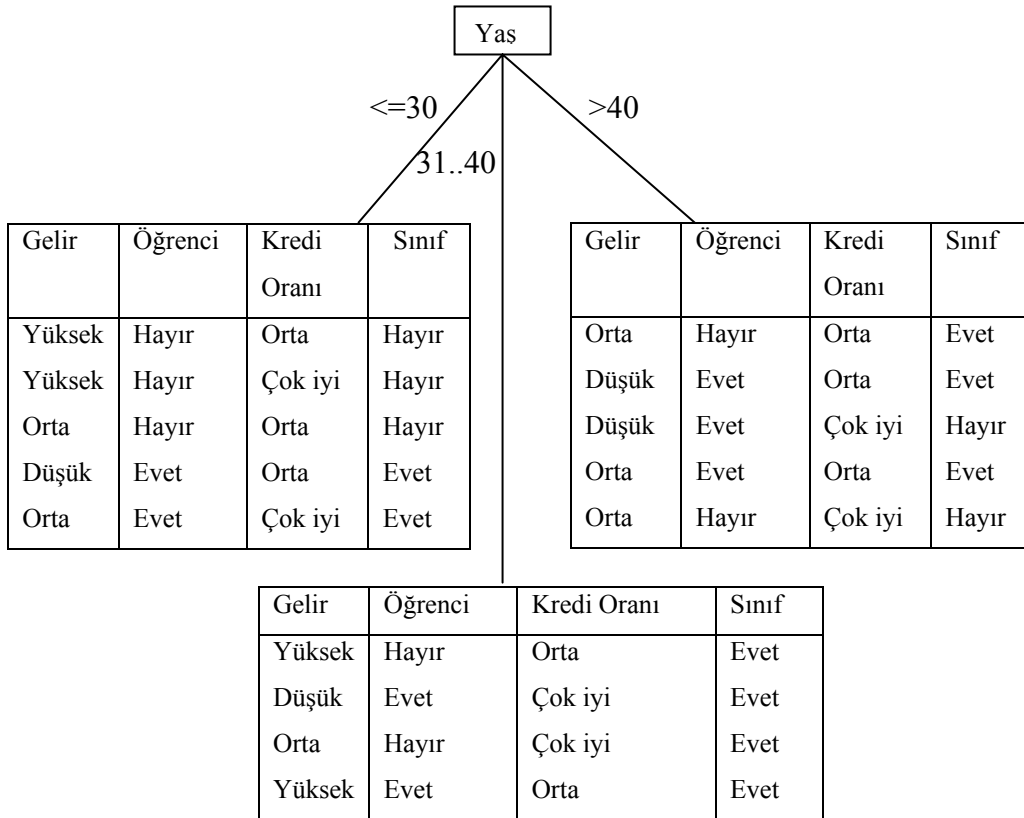
Benzer şekilde “gelir”, “öğrenci” ve “kredi oranı” değişkenlerine ait kazanç değerleri de hesaplanarak aşağıdaki gibi bulunmuştur.

$$G(\text{gelir}) = 0.029,$$

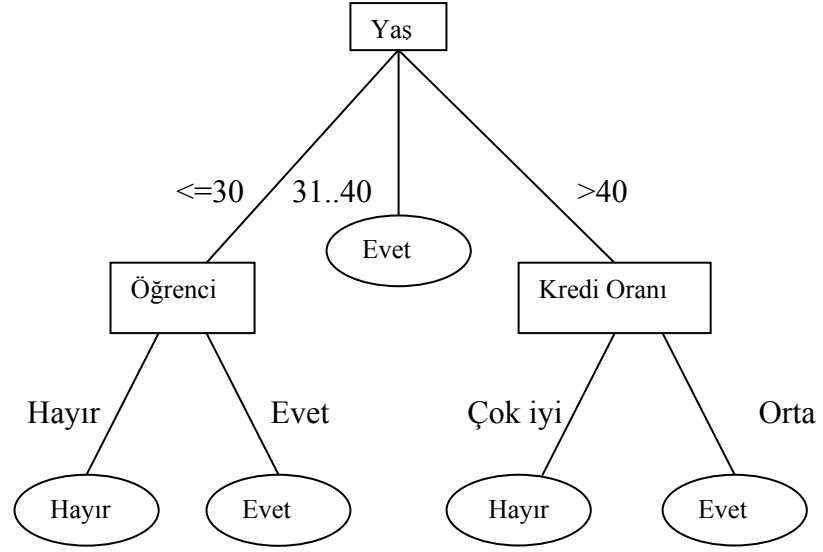
$$G(\text{öğrenci}) = 0.151$$

$$G(\text{kredi oranı}) = 0.048.$$

Kazanç değeri en yüksek değişken “yaş” olduğu için test değişkeni olarak seçilmiştir. Karar ağacının ilk başlangıcı olarak bir kök ve oluşturulmuş ve “yaş” olarak etiketlenilerek, şekil 5.3’de gösterildiği gibi, aldıkları değerlere göre dallandırılmıştır. “Yaş” değişkeninden sonra en yüksek kazanç değerine sahip “öğrenci” değişkeni ile devam edilmiş ve “kredi oranı” ile karar ağacı sonlandırılmıştır. Karar ağacının son şekli şekil 5.4’de gösterilmiştir.



Şekil 5.3 Başlangıç karar ağacı kökü.



Şekil 5.4. Son şekliyle karar ağacı diyagramı.

5.2.1.4.3. Kural Bulma Algoritması

Kural bulma algoritması bilgi çıkarımında, PolyAnalyst içerisindeki en güçlü algoritmalarından birisidir. Amacı, SKAT (Symbolic Knowledge Acquisition Technology) algoritmasını kullanarak veri kümesi içerisindeki çok boyutlu lineer olmayan ilişkileri otomatik olarak keşfetmek ve bu ilişkileri temsil eden, kullanıcı tarafından anlaşılması ve analizi kolay, matematiksel formüllerle kurallar oluşturmaktadır. En temel avantajı, verideki gizli ilişkileri ortaya çıkarabilme kabiliyetidir. Bir tahmin kuralı elde etmek istenilen durumlarda kullanılabilecek mükemmel bir araçtır. Hedef değişkeni sayısal bir parametre olmalıdır. Çoklu regresyon mantığına dayanarak çalışır. Lineer regresyonun zayıf sonuçlar verdiği veriler üzerinde uygulanır. Çıktı olarak; en uygun kural ve en kesin kural olmak üzere iki tip kural üretir. En uygun kural, istatistiksel olarak uygun olan bir modeli temsil eder. En kesin kural ise en doğru modeli temsil eder. En uygun kural istatistiksel olarak güvenilirlik testlerinden geçmiş olan kuralken, en kesin kural tam olarak test edilmemektedir. Yani başka veri kümesi için tam doğru olmayabilir. Bu yüzden uygulamalarda en uygun kuralın kullanılması gerektiği önerilmektedir. Kuralların doğruluğu ve etkinlikleri regresyon katsayısına (R^2) göre değerlendirilmektedir. Bu katsayı $[0,1]$ aralığında değer

almaktadır. 1'e yaklaştıkça doğruluğu artmaktadır. Regresyon katsayısı, standart hataya göre hesaplanmaktadır. İlgili formüller tablo 5.4'de verilmiştir [125].

Kural bulma algoritmasının temeli, Megaputer Inc. tarafından geliştirilen ve şekil 5.5'de gösterilen Symbolic Knowledge Acquisition Technology (SKAT) algoritmasına dayanmaktadır. SKAT algoritması, evrimsel programlama metodlarını kullanarak, doğrusal olmayan ilişkileri daha iyi temsil eden, yüksek dereceli rasyonel ifadeler ve matematiksel yapılar oluşturan bir algoritmadır. Yapay sinir ağlarının aksine SKAT, çıktı olarak okunabilir sembolik ifadeler verir, böylece insanlar kurulan modelin nasıl çalıştığını kolayca anlayabilirler. Bu evrimsel algoritma veriyi temsil eden en iyi ifadeleri bulmak için "Generalizing Transformations (GT)" fikrinden yararlanır [125,130,139].

Tablo 5.4. Standart hata ve regresyon katsayısı formülleri.

Standart Hata (ρ)	Regresyon Katsayısı (R^2)
$\rho = \sqrt{\frac{\sum_{i=1}^N (p_i - P_i)^2}{(N-1)\delta}}$ N=Veri kümesi içerisinde yer alan kayıt sayısı. $P_i = p_i$ değişkenin ortalaması $\delta = p_i$ değişkenin dağılımı	$R^2 = (1 - \rho^2)$

5.2.1.4.4. PolyNet Tahminleyicisi

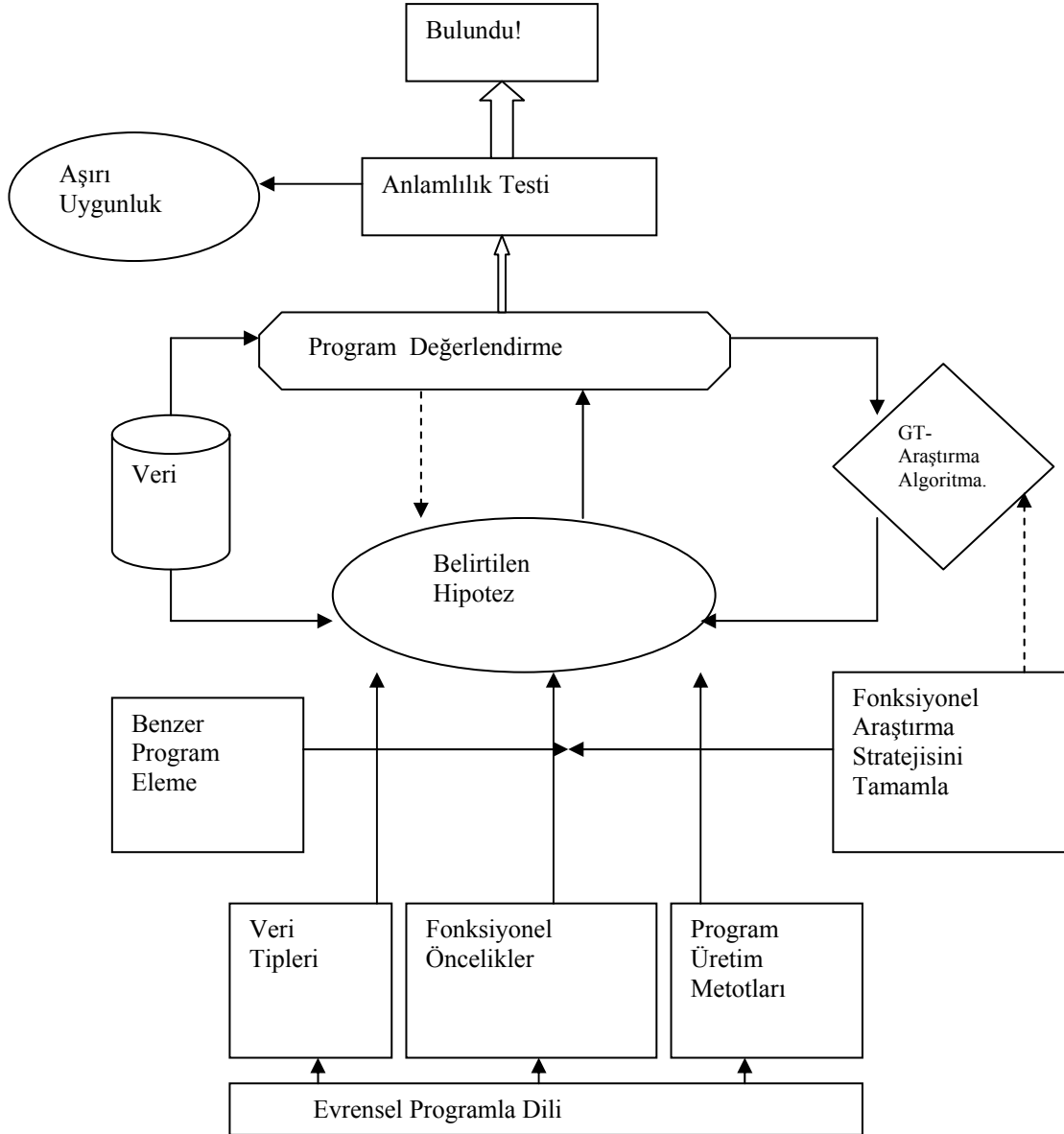
Polynet tahminleyicisi, PolyAnalyst programının yapay sinir ağı aracıdır. Sayısal bir hedef değişkenin tahmininde kullanılır. Kategorik, karakter dizisi ve ikili değişkenler de analizde kullanılabilir ancak hedef değişken olamazlar. Büyük veri kümeleri ile çalışabilir. Düğüm noktalarından oluşan bir yapay sinir ağı oluşturur. Her bir nokta diğer parametrelere bağlı olarak bir parametrenin tahmini için bir kural içerir. Eğer sinir ağı basitse kurallar sembolik olarak temsil edilebilir. Ancak çok karmaşık bir yapıya sahipse bir kara kutu gibi çalışır ve sembolik bir kural elde edilemez. Kuralların doğruluğu ve etkinlikleri kural bulma modülündeki gibi regresyon katsayısına göre değerlendirilmektedir. Bu katsayı [0,1] aralığında değer almaktadır. 1'e yaklaştıkça

doğruluğu artmaktadır. Regresyon katsayısının hesaplanması ile ilgili formüller tablo 5.4’de verilmiştir [125].

Polynet tahminleyicisi, çok terimli (polinom) yapay sinir ağları olarak bilinen, “Group Method of Data Handling-GMDH” adlı algoritmayı kullanmaktadır. Bu algoritma, veri madenciliği ve bilgi keşfi, tahmin, sistem modelleme, optimizasyon ve örüntü tanıma alanlarında çok yaygın olarak kullanılmaktadır. GMDH algoritması, veri içindeki ilişkilerin otomatik olarak bulunmasına, model veya ağın optimum yapısının seçilmesine ve mevcut algoritmaların doğruluğunu artırmaya imkan tanır [140].

Polynet tahminleyicisi, bulunan ilişkileri, iki girdisi olan hiyerarşik bir ağ ile temsil eder. Mevcut ağ, birkaç ara katmana sahip olabilir. En alt katmandaki düğümler, incelenen veri kümesindeki değişkenlere uyan özel düğümlere bağlıdır. En üst katman ise temel yapının çıktısı olarak nitelendirilen tek bir düğüme bağlıdır. Ağın çıktı değerini hesaplamak için yapılan hesaplamalar, en alt katmandan en üst katmana doğru yapılmaktadır. Her düğüm kendi ile ilgili bir polinoma sahiptir. Bu polinomlar iki girdi düğümü ile alakalı iki değişken içermektedir ve ikinci veya üçüncü dereceden polinomlardır [125].

Keşif süresince polynet tahminleyicisi, sırasıyla şu adımları takip ederek bir ağ oluşturmaktadır. İlk olarak, başlangıç değişkenleri ile bütün olası kombinasyonlar sınanır. Her bir kombinasyon, 2 ve 3. dereceden polinom şeklinde bir regresyon modeli oluşturmak için kullanılır. Daha sonra algoritma en doğru ve aynı zamanda birbirleriyle en zayıf korelasyona sahip modelleri seçer. Bu kombinasyonlar daha sonraki modellerin oluşturulmasında kullanılır. Oluşturulan her model F testi ile test edilirler. Bu süreç, daha önce oluşturulan modellerden daha anlamlı bir model elde edilemeyinceye kadar devam eder. En doğru model ağın çıktı düğümü olarak seçilir. Elde edilen modelin doğrulanması, diğer algoritmalarda olduğu gibi rasgele seçilmiş olan veri kümesi üzerinde yapılır [125].



Şekil 5.5. SKAT algoritması akış diyagramı (139).

5.2.1.4.5. Lineer Regresyon

Veriler arasındaki korelasyonu bulmak için en yaygın kullanılan yaklaşımlardan birisi lineer regresyondur [101]. Lineer regresyon istatistiksel tahminlemede en iyi bilinen, hızlı ve en eski metotlardan birisidir. Lineer regresyon y değişkeninin x değişkenine bağımlılığını bulmak için x, y eksenleri üzerinde bir doğru çizme prosesi olarak tanımlanabilir. PolyAnalyst aracının lineer regresyon algoritması, adım adım lineer regresyon algoritması temeliyle çalışmaktadır. Hedef değişkeni sayısal olan her türlü veri kümesi ile çalışabilmektedir. Kategorik ve dizi parametreler, “eğer” ifadesi ile

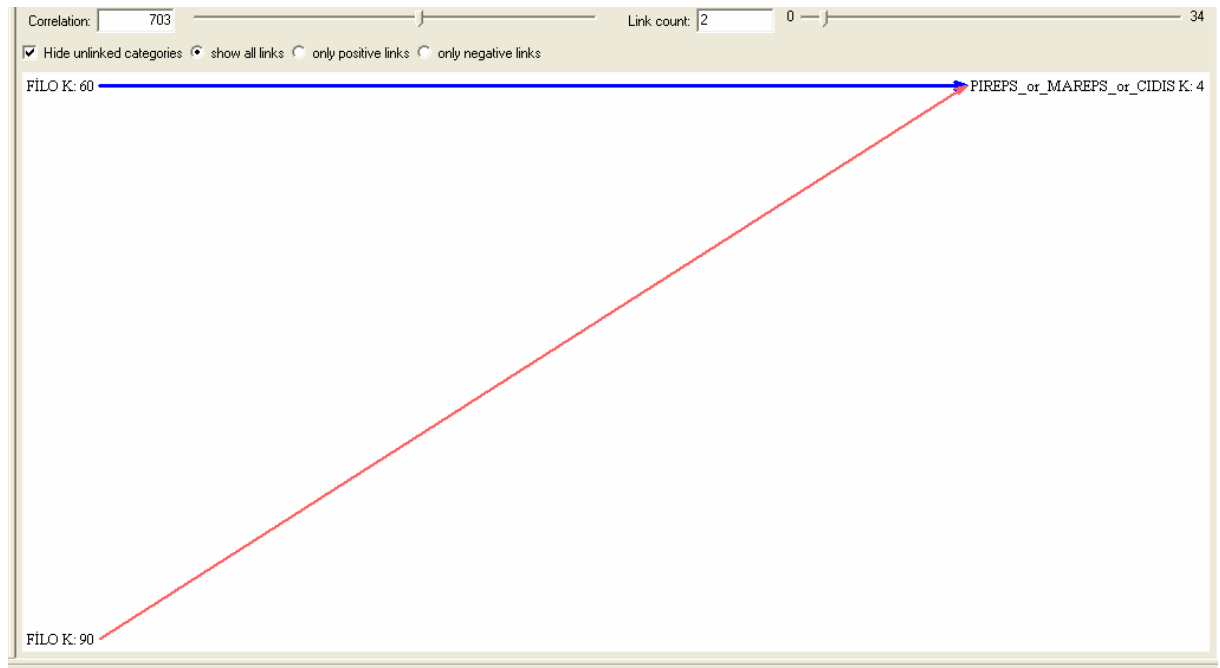
regresyona dahil edilmektedirler. Büyük miktardaki veri kümesi için hızlı ve uygun bir algoritmadır. Tahminleme de kullanıldığı gibi, ön işleme aracı olarak da kullanılabilir. Polynet tahminleyicisinde olduğu gibi F testini kullanmaktadır. Çıktı olarak lineer regresyon algoritması bir tahmin kuralı verir. Elde edilen kuralın doğruluğunun değerlendirilmesinde ise, kural bulma ve polynet tahminleyicisinde olduğu gibi regresyon katsayısına bakılmaktadır. Bu katsayı 0 ile 1 arasında her hangi bir sayıdır ve 1'e yaklaştıkça doğruluğu artmaktadır. Regresyon katsayısının hesaplanması ile ilgili formüller tablo 5.4'de verilmiştir [125].

PolyAnalyst lineer regresyon algoritması ile ilgili ayrıntılar aşağıdaki gibidir:

- 1- İncelenen veri kümesinde kategorik ve ikili değişkenler de olabilir. İkili değerler 0 ve 1 olmak üzere sayısal olarak dikkate alınır. Her bir kategorik değişken ise otomatik olarak ikili değişkenler kümesi olarak kodlanırlar. İncelenen veri kümesinde yer alan kategorik A değişkeninin 3 farklı değeri "a", "b" ve "c" olsun. Böylece $A="a"$, $A="b"$ ve $A="c"$ olmak üzere yeni ikili değişkenler oluşturulmuş olur. Bu yeni değişkenler 0 ve 1 değerlerini içeren sayısal değişkenler olarak nitelendirilirler.
- 2- Regresyon modelinin oluşturulması artan bir süreçtir. İlk olarak sadece tek boyutlu modeller oluşturulur. Bunların içerisinde en doğru model seçilir. Daha sonra ilk modele yeni bir değişken eklenerek iki boyutlu modeller oluşturulur. Yeni değişkenli iki boyutlu modelin en iyisi seçilir. Seçilen bütün regresyon modellerinin, regresyon katsayılarının Fisher istatistik değerlerine göre testten geçmiş olmaları gerekmektedir. Eğer bir terimin f oranı değeri istenilen düzeyin altında ise bu terim modelden çıkarılır. Yeni terim ekleme prosesi modele eklenecek yeni bir terim kalmayana kadar devam eder.
- 3- Oluşturulan modelin kesinliği iki farklı metotla doğrulanır. İlk metot modeldeki her bir terimin f oranı değerlerinin kontrol edilmesi esasına dayanır. İkinci metot ise rasgele test etme işlemidir. Bu süreçte PolyAnalyst, rasgele seçtiği veri kümeleri üzerinde en iyi regresyon modelini test ederek öğrenir. En iyi doğruluk değerini veren regresyon modeli seçilerek sunulur.

5.2.1.4.6. Bağlantı Analizi

Bağlantı analizi bütün kategorik ve ikili değişkenler arasındaki karmaşık korelasyonu gösterir. Bu korelasyonu bağlantı grafiği ile gösterir. Çok boyutlu korelasyon analizi esasına göre çalışan bir algoritmadır. Hedef değişken gerektirmemektedir. Sadece analiz edilecek değişkenlerin belirlenmesi yeterlidir. Ancak değişkenler, kategorik, ikili ve metinsel olmalıdır. Değişkenler arasındaki pozitif ve negatif korelasyonları gösterir. Korelasyon katsayısı $[-1,1]$ arasında bir değer almaktadır. Bu değer 1'e yaklaştıkça parametrenin etkinliği artmaktadır. Kırmızı çizgiler pozitif korelasyonları, mavi çizgileri ise negatif korelasyonları göstermektedir. Rengin yoğunluğu aynı zamanda korelasyonun şiddetini de ifade eder. Veri madenciliği sürecinde bağlantı analizi ilk adım olabilir [125]. Şekil 5.6'da bağlantı analizi sonucunda elde edilmiş olan ve negatif/pozitif ilişkileri gösteren örnek bir grafik verilmiştir.



Şekil 5.6. Örnek bağlantı analizi grafiği.

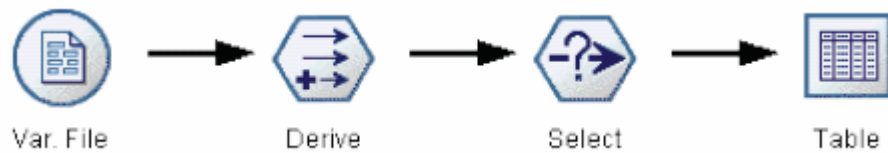
5.2.2. Clementine ile Özellik Seçimi ve Aykırılıkların Tespiti

SPSS Clementine, karar verme sürecini geliştirmek için tahminleyici modeller kurmaya yarayan bir veri madenciliği aracıdır. Kolay kullanımı ile veri içerisindeki gizli ve kıymetli ilişkilerin keşfinde yardımcı olur. Basitçe Clementine ile veri analizi süreci üç

aşamadan oluşur. İlk olarak, verinin okutulması, daha sonra veriyi işlemlerle çalıştırmak ve son olarak veriyi tanımlama aşamalarıdır. Bu ardışık işlemler süreci, “veri akışı-data stream” olarak bilinir. Şekil 5.7’de örnek bir veri akışı gösterilmiştir [141]. Clementine; sınıflandırma, tahmin, bölümleme, ilişki bulma gibi bir çok veri madenciliği tekniğini şekil 5.8’de gösterilen görsel programlama ara yüzü ile kullanıcıya sunmaktadır. Bu ara yüz, verinin analizi sürecinde diyagramları çizmek için kullanılmaktadır. Her bir işlem bir düğüm veya simge ile ifade edilmektedir ve bir düğüm birbirlerine bağlı bir şekilde veri akış diyagramını oluşturmaktadır. Clementine’in içerdiği modelleme teknikleri şu şekilde sınıflandırılabilirler [141]:

- Tahminleme modelleri; karar ağaçları, yapay sinir ağları ve istatistiksel modelleri içermektedir.
- Kohonen, K-ortalamları (k-means) ve İki-Adım (Two-step) kümeleme metotlarıdır.
- Birliktelik kuralları metodu.
- Eleme yapma modelleri ise, özellik seçimi (feature selection) ve aykırılıkların tespiti (anomaly detection)’dir.

Bu çalışma kapsamında, ön işleme aracı olarak, özellik seçimi ve aykırılıkların tespiti modelleri kullanılmıştır. Bu metotlar, mevcut veri içerisindeki, modelleme için kullanılabilecek en ilişkili özellikler ile veride yer alan ve bilinmeyen aykırı kayıtları tespit etmede kullanılmıştır.



Şekil 5.7. Clementine örnek veri akış diyagramı.

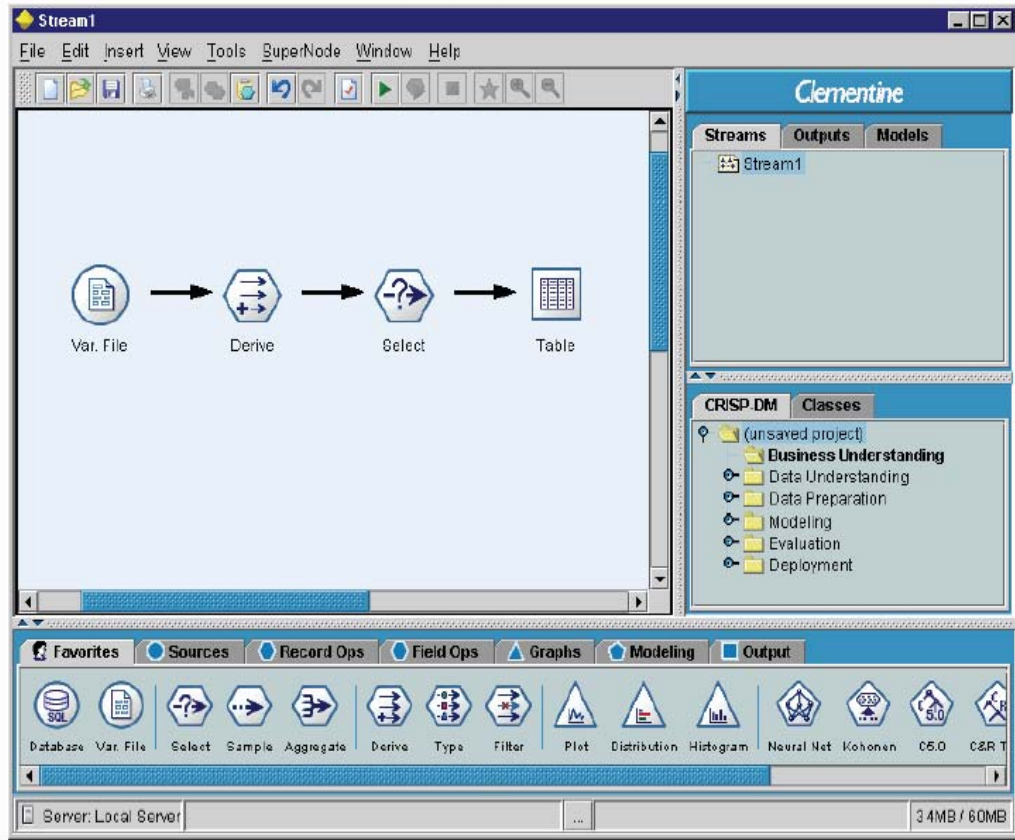
5.2.2.1. Özellik Seçimi

Veri madenciliği problemleri, tahminleyici olarak kullanılabilme potansiyeli içeren yüzlerce, hatta binlerce alan içermektedirler. Sonuç olarak hangi alan ve değişkenlerin modele dahil edileceği konusu üzerine, büyük zaman ve güç harcanması gerekmektedir.

Seenekleri azaltmak iin, zellik seimi, analiz srecinde en nemli olanları belirlemede kullanılabilir.

zellik seimi  adımda gerekleřtirilir.

- 1- *Tarama*, nemsiz ve problemli kayıt ve olayları kaldırır.
- 2- *nceliklere gre sıralama*, nem derecelerine gre geriye kalan tahminleyicileri sıralar.
- 3- *Seme*, sonraki modelde kullanılmak zere zellikler alt kmesini belirler.



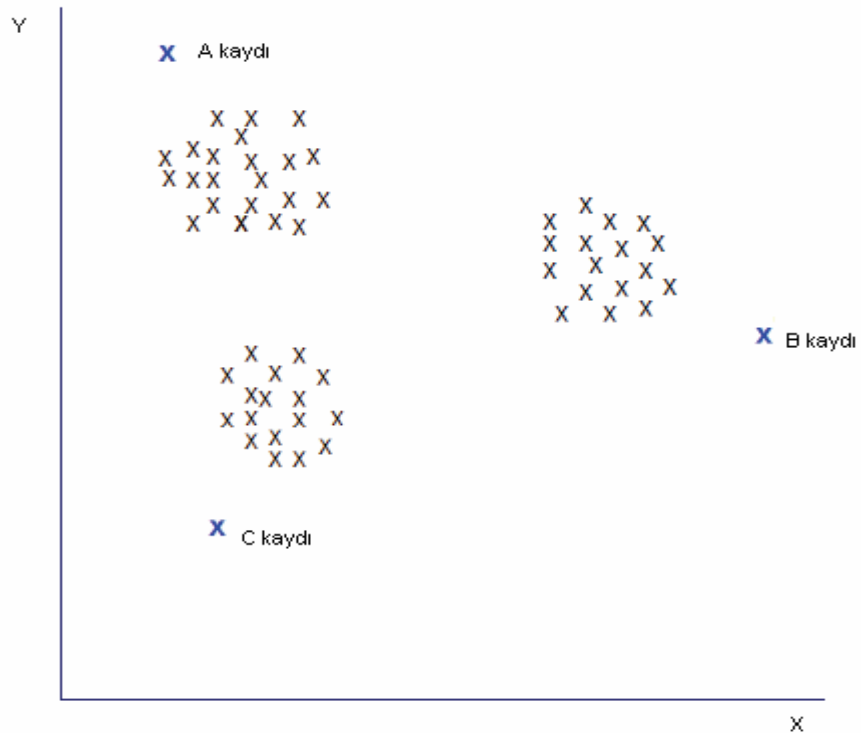
řekil 5.8.Clementine kullanıcı ara yz.

ok fazla veri ile řiřmiř durumda olan organizasyonlar zellik seme sayesinde modelleme srecinde basitlik ve hzdan dolayı yarar saėlarlar. Etkili kayıtlar zerine daha ok dikkat ekerek, daha hzlı iřlem sresi ile daha basit, daha doėru ve anlaşılır modellerle alıřabilirler. Ayrıca daha sonraki veri toplama srecinde de toplanan veri miktarında azaltma da yapılabilir [141].

5.2.2.2. Aykırılıkların Tespiti

Aykırılıkların tespiti modelleri, veri içerisinde yer alan aykırı veya olağan dışı kayıtların belirlenmesinde kullanılır. Diğer modellerden farklı olarak, “normal” olayların özelliklerini tespit eder. Bu metot eğitim kümesi gerektirmeyen bir danışmansız öğrenme metodudur.

Aykırılıkları belirleyen geleneksel modeller genellikle aynı anda bir veya iki değişken üzerinde yoğunlaşırken, Clementine aykırılıkların tespiti modeli, kümeleri veya benzer grupları belirlemek için büyük miktarlarda değişkenleri inceleyebilir. Her bir kayıt kendi içinde bulunduğu gruptaki diğer kayıtlarla karşılaştırılarak olası aykırılıklar belirlenir. Şekil 5.8’de gösterilen örnekte olduğu gibi A, B ve C kayıtları belirlenmiş olan 3 gruba da benzer özellikler taşımadığı için aykırı değerler olarak belirlenmektedir. Her bir kayıta, ait olduğu sınıfla ilgili olarak gruptan sapma oranını gösteren bir aykırılık indeksi atanır. Bu indeksin değeri ne kadar yüksekse, kayıt grup ortalamasından o kadar sapma gösteriyor demektir. Aykırılık indeksi “1-1.5” dan daha düşük değerler alırsa bu kayıtlar aykırı değer sayılmazlar. Ancak aykırılık indeksi 2’nin üstünde bir değer alırsa, ilgili kayıtlar aykırı değerler olarak tespit edilirler [141].



Şekil 5.9. Aykırı değerlerin tespit edilmesi ile ilgili örnek.

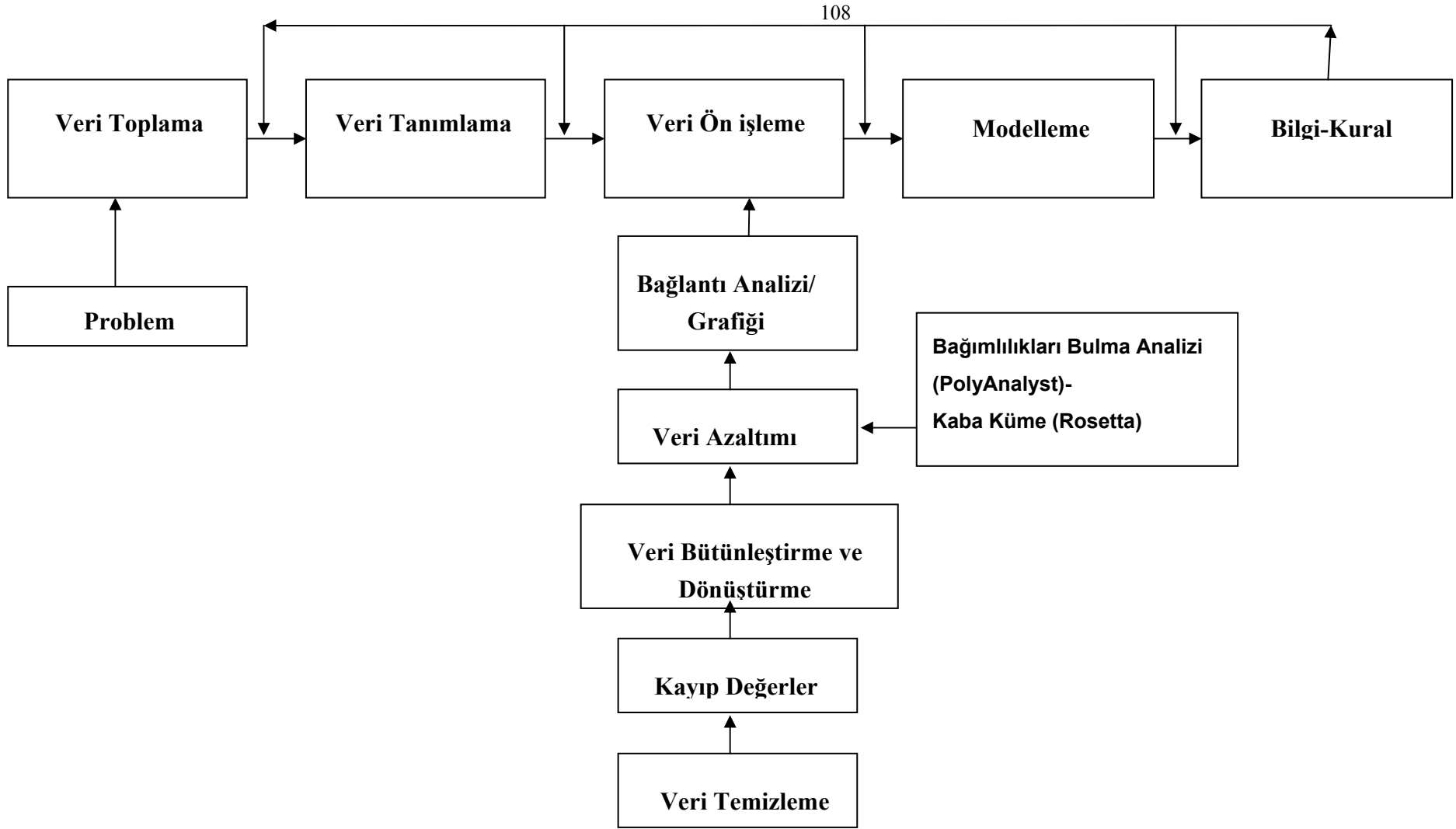
6.BÖLÜM

HAVACILIK SEKTÖRÜNDE VERİ MADENCİLİĞİ UYGULAMASI

6.1. Giriş

Güvenlik ve risk havacılık sektöründe her zaman önemli konular olmuştur. Hava taşımacılığında meydana gelen artışlar, uçuş ertelemelerini, uçuş iptallerini ve kazaları da beraberinde getirmektedir. Bu kazalar da insan yaralanması ve ölümleri ile sonuçlanmaktadır.

Veri madenciliği teknikleri pek çok farklı alanda başarılı şekilde uygulanmaktadır. Bu alanlardan bir tanesi de havacılık sektörüdür. Havacılık sektöründe oldukça büyük miktarlarda veri birikimi söz konusudur. Bu veriler, pilot raporları, bakım raporları, kaza raporları veya tehir raporları şeklinde tutulmaktadır. Bu çalışmada, Federal Havacılık Yönetimi (FHY), kaza veri sisteminden alınmış kaza raporları üzerinde veri madenciliği çalışması yapılmıştır. Bu veriler 2000 ve 2006 yılları arasında gerçekleşmiş olan sivil uçuşlara ait kayıtları içermektedir. Veri kümesinde yer alan parametre sayısını azaltmak için kaba küme yaklaşımı uygulanmıştır. Çalışmanın amacı, kazalardaki ölüm oranını etkileyen parametrelerin belirlenmesidir. Ölümle sonuçlanan kazalardaki parametreler arasındaki ilişkileri ve kuralları bulmak için sınıflandırma araçları ve karar ağaçları kullanılmıştır. Sonuç olarak ölümle sonuçlanan kazalar ile ilgili bazı kurallar ve ilişkiler elde edilmiştir. Elde edilen kurallar doğrulukları ve güvenilirlikleri açısından test edilmiş ve kabul edilebilir düzeylerde olduğu görülmüştür. Şekil 6.1’de analiz süreci özetle gösterilmiştir.



Şekil 6.1. Federal Havacılık Yönetimi verileri ile yapılan veri madenciliği süreci.

6.2. Değerlendirilen Veri Kümelerinin Tanıtılması

Çalışmada kullanılmış olan veriler Federal Havacılık Yönetiminin internet sitesinden alınmıştır. Veri tabanı 1978 yılından itibaren gerçekleşmiş olan genel ve ticari hava taşımacılığı kazalarına ait kayıtları içermektedir. Bu çalışmada 2000 ve 2006 yılları arasındaki kayıtlar kullanılmıştır. Bu veriler tablo 6.1’de gösterilmiş olan parametreleri içermektedir [142].

Tablo 6.1’de görüldüğü gibi, kaza raporları 24 parametre içermektedir. Çalışmanın amacı bu parametrelerin ölümle sonuçlanan kazalar üzerindeki etkilerini ortaya koymaktır. Bu amaca yönelik yapılan çalışmalar aşağıda açıklanmıştır.

6.3. Önışleme Süreci

Bu süreçte etkili parametrelerin tespit edilmesinde iki farklı önışleme aracı kullanılmıştır. Bunlardan ilki PolyAnalyst adlı veri madenciliği aracını kullanarak yapılan önışlemedir. İkincisi ise kaba küme yaklaşımı ile Rosetta adlı aracı kullanarak yapılan özellik indirgeme işlemidir.

6.3.1. PolyAnalyst Yazılımı İle Yapılan Önışleme

Önışleme sürecinde mevcut 24 parametre ile analizler gerçekleştirilmiştir. Veri madenciliği sürecine geçmeden önce yapılacak olan analizin daha etkili olması amacıyla kullanılacak olan parametrelerin doğru olarak tespit edilmesi gerekmektedir. Bu amaçla veri madenciliği yazılımı olan PolyAnalyst programına ait bağımlılıkları bulma ve bağlantı analizleri modülleri ön işleme aracı olarak kullanılmıştır.

Tablo 6.1. Uygulamada kullanılmış veri kümesinde yer alan parametreler ve tanımları.

Parametre Adı	Açıklama
Rapor Numarası	Olay rapor numarası.
Olay Tarihi	Olayın meydana geldiği tarih
Şehir	Olayın meydana geldiği şehir.
Ülke	Olayın meydana geldiği ülke.
Havaalanı Adı	Olay meydana geldiği yer ile bağlantılı olan havalimanı.
Olay Tipi	Olay ve uçuş operasyonu tipi.
Uçak Hasar Tipi	Olay sonucunda meydana gelen hasar tipi.
Uçuş Fazı	Olay sırasındaki uçuş operasyonu fazı.
Uçağın Üretici Firması	Olaya karışan uçağın üretici firması.
Uçağın Modeli	Olaya karışan uçağın modelinin tanımı.
Uçak Seri Numarası	Olaya karışan uçağın seri numarası.
Operatör	Operatörün adı.
İlk Uçuş Fazı	Olay sırasında uçağın ilk hareket tipi.
Uçuş Tipi	Uçuşun ne amaçla yapıldığı.
Uçuş Planı	Planlanan uçuş tipi.
Kayıt Numarası	Uçağa firma tarafından verilen kayıt numarası.
Ölüm Sayısı	Olay sonucunda meydana gelen ölüm sayısı.
Yaralanma Sayısı	Olay sonucunda meydana gelen ölüm dışı yaralanmalar.
Motorun Yapımcı Firması	Olaya karışan uçağın yapımcı firması.
Motor Modeli	Uçağın motorunun model numarası.
Motor Model Grubu	Uçağın motorunun hangi model grubundan olduğu.
Motor Sayısı	Uçaktaki motor sayısı.
Pilot Sertifikası	Uçağın pilotunun sahip olduğu sertifika.
Pilotun toplam faal uçuş saati	Pilotun uçuş esnasında faal olarak kaç saat görev aldığı.

6.3.1.1. Bağımlılıkları Bulma

Bağımlılıkları bulma analizi, mevcut 24 parametre ile yapılmıştır. Hedef değişkenimiz kaza sonucunda meydana gelen ölümlerdir. Bağımlılıkları bulma analizi ile kaza

sonucunda meydana gelen ölüm sayısına etkisi olan parametreler ve etki düzeyleri belirlenmiştir. Yapılan analizde ölümle sonuçlanan kazalara etkili olan 8 parametre tespit edilmiştir. Bu parametreler tablo 6.2’de gösterildiği gibi {olay tarihi, uçak hasar tipi, olay tipi, uçuş tipi, uçuş planı, motor sayısı, pilot sertifikası, ülke}’dir. Etki derecelerini tespit etmek üzere daha önce bölüm 5’de anlatıldığı gibi P değerleri dikkate alınmıştır. P değeri ilgili bağımlılıkların şiddetini ifade etmektedir. Bu değer sıfıra ne kadar yakınsa parametreler arasında o kadar güçlü bağımlılık olduğu sonucu çıkarılmaktadır. P değeri 10^{-7} değerinin altında ise 0 (sıfır) kabul edilir. Analiz sonucunda elde edilen P değerleri incelenerek eşik değeri kabul edilen 10^{-7} değerinin üzerinde olan 7 parametre seçilmiştir. Ayrıca diğer 17 parametre içerisinde 10^{-7} değerine en yakın olan değer olan $4.185e^{-006}$ P değeri de kabul edilerek ilgili parametre etkili parametreler içerisine dahil edilmiştir. Elde edilen P değerleri ve ilgili parametreler tablo 6.3’de gösterilmiştir.

Elde edilmiş olan parametre kümesi, bölüm 6.4’de anlatıldığı şekilde veri madenciliği sürecinde uygulanmış olan analizlerde kullanılmıştır.

6.3.1.2. Korelasyonların Belirlenmesinde Bağlantı Grafikleri

Bağımlılıkları bulma analizi ile elde edilmiş olan sonuçların doğrulanması ve ilgili parametrelerin hangi düzeylerinin daha etkili olduğunun tespit edilmesi için bağlantı grafikleri kullanılmıştır. Bağlantı grafikleri ikili ve kategorik parametreler arasındaki ilişkileri hızlı bir şekilde göstermek için kullanılmaktadır. İlişkinin negatif ya da pozitif olduğu da belirlenmektedir. [125].

Bağlantı grafiklerinde, parametreleri bağlayan çizginin yoğunluğu ilişkinin yoğunluğunu ifade eder. Kalın çizgi güçlü bağlantıyı; ince çizgi ise zayıf bağlantıyı temsil eder. İlgili parametreler arasındaki bağlantıyı ölçmek için analizlerde korelasyon katsayısı kullanılmıştır. Korelasyon katsayısı iki değişken arasındaki istatistiksel ilişkiyi ifade eder. Negatif ve pozitif olabilir. Korelasyon katsayısı $[-1,1]$ arasında bir sayıdır. Katsayı 1’e ne kadar yakınsa, değişkenler o kadar ilişkilidir.

Tablo 6.2. Analizlerde kullanılmak üzere ön işleme sonucunda elde edilen veri kümesi.

Parametre Adı	Parametre Tipi	Açıklama
Olay Tarihi	Tarih	01-Ocak-2000 ve 31-Aralık-2006 tarihleri arasında herhangi bir gün.
Ülke	Dizi	Ülkeyi temsil eden kod.
Olay Tipi	Kategorik	1 (Hava taşımacılığı kazası) 2 (Genel havacılık kazası)
Hasar Tipi	Kategorik	1 (Parçalanma) 2 (Önemsiz) 3 (Hasar Yok) 4 (Önemli) 5 (Bilinmeyen)
Uçuş Tipi	Kategorik	1 (Tarımsal) 2 (Ticari Hava Taşımacılığı) 3 (Günlük Hava Taksiciliği) 4 (Yabancı Hava Taşımacılığı) 5 (Genel İşlem Uçuşları) 6 (Paraşüt Atlama) 7 (Part 125 İşlemleri) 8 (Pilot Okulları) 9 (Dış Yol Operasyonu) 10 (Işıklıdırma Uçuşu)
Uçuş Planı	Kategorik	1 (Araç-Gereç Uçuş Kuralı) 2 (Bilinmeyen) 3 (Genel Uçuş Kuralı) 4 (Özel Uçuş Kuralları) 5 (Genel Uçuş Takip Kuralı) 6 (Yok)
Motor Sayısı	Sayısal	Uçaktaki motor sayısı.
Pilot Sertifikası	Kategorik	1 (Havayolu Taşımada Pilotu) 2 (Havayolu Taşımada Pilot Eğitmeni) 3 (Özel Pilot) 4 (Özel Pilot Eğitmeni) 5 (Sertifikaasız Pilot) 6 (Gizli-Kişiyeye Özel Pilot) 7 (Gizli-Kişiyeye Özel Pilot Eğitmeni) 8 (Spor-Eğlence Pilotu) 9 (Özel Amaçlı Pilot) 10 (Öğrenci) 11 (Bilinmeyen/Yabancı)
Kaza sonucu ölüm var mı?	İkili	Evet/Hayır

Analizimizde tespit edilen etkili parametrelerden {Hasar Tipi, Uçuş Tipi, Olay Tipi, Uçuş Planı, Pilot Sertifikası} parametreleri ikili veya kategorik özelliklerde olduğundan bağlantı grafiklerinin oluşturulmasında 8 parametreden sadece 5 tanesi kullanılmıştır. Bağlantı grafikleri ile ilgili parametrelerin düzeyleri ile ölümle sonuçlanan kazalar

arasındaki ilişkilere bakılmıştır. Tablo 6.4’de en yüksek korelasyon katsayısına sahip 8 parametre düzeyi korelasyon katsayıları ile birlikte gösterilmiştir.

Tablo 6.3. Bağımlılıkları bulma analizi sonuçlarına ait P değerleri ve ilgili parametreler.

Parametre Adı	P-Değeri
Olay Tarihi	$4.375e^{-123}$
Hasar Tipi	$4.375e^{-123}$
Uçuş Tipi	$4.375e^{-123}$
Olay Tipi	$1.166e^{-018}$
Uçuş Planı	$1.166e^{-018}$
Motor Sayısı	$1.166e^{-018}$
Pilot Sertifikası	$1.504e^{-007}$
Ülke	$4.185e^{-006}$

Tablo 6.4. Bağlantı grafiklerine ait korelasyon katsayıları ve ilgili parametre düzeyleri.

Parametre ve Düzeyi	Korelasyon Katsayısı
Hasar Tipi 2 (Önemsiz)	1
Hasar Tipi 3 (Hasar Yok)	0.73184
Uçuş Tipi 5 (Genel İşlem Uçuşları)	- 0.64763
Uçuş Tipi 6 (Paraşüt Atlama)	1
Olay Tipi 1 (Hava taşımacılığı kazası)	-1
Olay Tipi 2 (Genel havacılık kazası)	1
Uçuş Planı 6 (Yok)	1
Pilot Sertifikası 5 (Sertifikaless Pilot)	1

Tablo 6.4’de görüldüğü gibi önemsiz olarak kabul edilen hasar tiplerinin, paraşütle atlama uçuşlarının, genel havacılık uçuşları kazalarının, sertifikasız pilotların ve uçuş planına sahip olmayan uçuş, parametre düzeylerinin ölümle sonuçlanan kazalar üzerinde daha etkili oldukları tespit edilmiştir.

6.3.2. Kaba Kümeler Kullanarak Özellik İndirgeme

Birçok gerçek yaşam probleminin kayıtlarında açıkça ifade edildiği gibi, KK yaklaşımının uygulamada çok yararlı olduğunu kanıtlanmıştır. 4. bölümde belirtildiği gibi KK yaklaşımının kullanılabileceği ana problemler; verilerin indirgenmesi, veri bağımlılıklarının keşfi, verilerin öneminin tahmini, verilerden karar algoritmalarının oluşturulması, verilerin yaklaşık sınıflaması, verilerdeki benzerlik ve farklılıkların keşfi, verilerdeki örüntülerin keşfi ve neden-sonuç ilişkilerinin keşfini kapsamaktadır [56-58]. Bu tez çalışmasında KK yaklaşımı verilerin indirgenmesi sürecinde kullanılmıştır. İndirgenmiş özelliklerin tespiti için kaba küme yaklaşımıyla problemlerin çözümünde güçlü kabul edilen bir araç olan Rosetta kullanılmıştır [124]. Rosetta ile veri indirgeme seçeneklerinden genetik algoritma uygulanmıştır.

Rosetta aracı kullanılarak kaba küme yaklaşımıyla 23 adet indirgenmiş veri kümeleri elde edilmiştir. Bu veri kümeleri PolyAnalyst kullanılarak karar ağaçları ile yapılan analiz için girdi olarak kullanılmışlardır. Elde edilen indirgenmiş veri kümeleri tablo 6.5’de gösterilmiştir.

6.4. Uygulama ve Sonuçlar

Bu çalışmanın amacı, kazalardaki ölüm oranını azaltmak adına etkili olan parametrelerin tespit edilerek, sınıflandırma kuralı çıkarılmasıdır.

Veri madenciliği çalışması için PolyAnalyst programında yer alan karar ağaçları algoritması kullanılmıştır. Elde edilen kuralların, doğrulukları ve güvenilirlikleri kontrol edilmiş ve sonuçlar anlamlı bulunmuştur.

Karar ağaçları, çoklu kategoriye sahip durumların sınıflandırma işlemlerinde kullanılırlar. PolyAnalyst programının, çok büyük miktarda veri ile çalışabilen en hızlı algoritması karar ağaçlarıdır. Her türlü veri çeşidi ile çalışabilen bir algoritmadır [125].

Tablo 6.5. Kaba küme yaklaşımı ile elde edilen parametre kümeleri.

Küme Numarası	Parametre Kümeleri
1	{Olay Tarihi, Hasar Tipi, Pilot Sertifikası}
2	{Ülke, Olay Tipi, Uçuş Tipi, Uçuş Planı, Motor Sayısı}
3	{Olay Tarihi, Hasar Tipi, Uçuş Tipi}
4	{Olay Tarihi, Uçuş Tipi, Motor Sayısı}
5	{Olay Tarihi, Uçuş Planı, Motor Sayısı}
6	{Ülke, Olay Tipi, Hasar Tipi, Uçuş Planı, Motor Sayısı, Pilot Sertifikası}
7	{Ülke, Olay Tipi, Hasar Tipi, Uçuş Tipi, Motor Sayısı, Pilot Sertifikası}
8	{Ülke, Hasar Tipi, Uçuş Tipi, Uçuş Planı, Motor Sayısı, Pilot Sertifikası}
9	{Olay Tarihi, Olay Tipi, Hasar Tipi}
10	{Olay Tarihi, Olay Tipi, Pilot Sertifikası}
11	{Olay Tarihi, Uçuş Planı, Pilot Sertifikası}
12	{Olay Tarihi, Hasar Tipi, Uçuş Planı}
13	{Olay Tarihi, Olay Tipi, Motor Sayısı}
14	{Olay Tarihi, Ülke, Olay Tipi}
15	{Olay Tarihi, Ülke, Motor Sayısı}
16	{Olay Tipi, Hasar Tipi, Uçuş Tipi, Uçuş Planı, Motor Sayısı, Pilot Sertifikası}
17	{Olay Tarihi, Hasar Tipi, Motor Sayısı}
18	{Olay Tarihi, Motor Sayısı, Pilot Sertifikası}
19	{Olay Tarihi, Ülke, Uçuş Tipi}
20	{Olay Tarihi, Ülke, Hasar Tipi}
21	{Olay Tarihi, Olay Tipi, Uçuş Planı}
22	{Olay Tarihi, Olay Tipi, Hasar Tipi, Motor Sayısı}
23	{Ülke, Olay Tipi, Hasar Tipi, Uçuş Tipi, Uçuş Planı, Pilot Sertifikası}

6.4.1. Karar Ağaçları Analizi

Analizde kullanılan veri kümesi test ve eğitim veri kümeleri olarak ikiye bölünmüştür. Bütün verinin %30'u test verisi olarak, geriye kalan %70'i ise eğitim verisi olarak rasgele seçilmiştir. Eğitim kümesi ile analizler gerçekleştirilmiş; test kümesi ile elde edilen kuralların doğruluğu test edilmiştir. Hedef değişkeni daha önce de belirtildiği gibi kaza sonucunda ölüm olup olmadığıdır.

İlk olarak, bağımlılıkları bulma analizi ile elde edilen ve tablo 6.2'de gösterilmiş olan veri kümesi kullanılarak karar ağaçları ile analiz gerçekleştirilmiştir. Analiz sonucunda elde edilen kurallar tablo 6.6'da verilmiştir. Kurallara ait sınıflandırma olasılık ve etkinlikleri ile doğrulukları da tablo 6.7'de gösterilmiştir. İlgili formülasyonlar da tablo 6.8 ve tablo 6.9'da verilmiştir.

Tablo 6.6. Karar ağaçları ile elde edilen tahmin kuralları.

- EĞER Uçuş Tipi=1 VEYA Uçuş Tipi=2 VEYA Uçuş Tipi= 3 VEYA Uçuş Tipi=4 VEYA Uçuş Tipi=7 VEYA Uçuş Tipi=9 O HALDE kaza sonucu ölüm= HAYIR
- EĞER Uçuş Tipi=5 VE Olay Tarihi<25.03.2002 VE Hasar Tipi= 2 O HALDE kaza sonucu ölüm= HAYIR
- EĞER Uçuş Tipi=5 VE Olay Tarihi<25.03.2002 VE Hasar Tipi= 1 O HALDE kaza sonucu ölüm= EVET
- EĞER Uçuş Tipi=5 VE Olay Tarihi<25.03.2002 VE Hasar Tipi= 3 VE Uçuş Planı= 1 VEYA Uçuş Planı= 2 O HALDE kaza sonucu ölüm= HAYIR
- EĞER Uçuş Tipi=5 VE Olay Tarihi<25.03.2002 VE Hasar Tipi= 3 VE Uçuş Planı= 6 O HALDE kaza sonucu ölüm= EVET
- EĞER Uçuş Tipi=5 VE Olay Tarihi>= 25.03.2002 O HALDE kaza sonucu ölüm= HAYIR
- EĞER Uçuş Tipi=6 VEYA Uçuş Tipi=10 O HALDE kaza sonucu ölüm = EVET

Tablo 6.7. Elde edilen kurallara ait sınıflandırma olasılığı ve sınıflandırma etkinlikleri.

	Sınıflandırma Olasılığı (%)	Sınıflandırma Etkinliği (%)	Doğruluk (%)
Eğitim	99.61	79.03	99.4
Test	99.44	76.92	99.4

Tablo 6.8. Sınıflandırma olasılığı ve etkinliğine ait formüller*.

Toplam Sınıflandırma Hatası	$100 \times \frac{N_{yan0} + N_{yan1}}{N_{yan0} + N_{dog0} + N_{yan1} + N_{dog1}}$
Sınıflandırma Olasılığı	%100- Toplam Sınıflandırma Hatası
Sınıflandırma Etkinliği	$100 \times \frac{N_{dog0} + N_{dog1} - \max(N_{dog0} + N_{yan0} + N_{tn0}, N_{dog1} + N_{yan1} + N_{tn1})}{\min(N_{dog0} + N_{yan0} + N_{tn0}, N_{dog1} + N_{yan1} + N_{tn1})}$

Tablo 6.9. Doğru sınıflandırılan kayıt sayıları kodlamaları

Gerçek/Tahmin	0	1	Tanımsız (tn)
0	N_{dog0}	N_{yan0}	N_{tn0}
1	N_{yan1}	N_{dog1}	N_{tn1}

Bağımlılıkları bulma algoritması ile elde edilen veri kümesinin analizinin ardından, kaba küme yaklaşımı ile elde edilen ve tablo 6.5’de gösterilen 23 adet veri kümesi de tek tek karar ağaçları metodu ile analiz edilmiştir. Bu süreç sonunda elde edilen kurallara ait sınıflandırma olasılıkları, sınıflandırma etkinlikleri ve doğrulukları tablo 6.10’da verilmiştir.

Tablo 6.10’daki sonuçlar incelendiğinde 2-3-4-7-8-16-19-23 no’lu veri kümelerinin dikkate alınması gerektiği görülmektedir. Bu veri kümeleri ile elde edilen kurallar tablo 6.11(a) ve tablo 6.11 (b)’de gösterilmiştir.

Tablo 6.10. Kaba küme yaklaşımı ile elde edilen veri kümelerinin karar ağacı analizine ait sınıflandırma etkinlikleri, sınıflandırma olasılıkları ve doğruluk değerleri.

No	Eğitim		Test		Eğitim Doğruluğu %	Test Doğruluğu %
	Sınıflandırma Olasılığı (%)	Sınıflandırma Etkinliği (%)	Sınıflandırma Olasılığı (%)	Sınıflandırma Etkinliği (%)		
1	98.21	33.87	98.79	50.0	98.2	98.8
2	99.39	74.19	99.44	76.92	99.3	99.4
3	99.52	75.81	99.44	76.92	99.3	99.4
4	99.52	75.81	99.44	76.92	99.3	99.4
5	98.46	20.97	98.51	30.77	97.8	98.3
6	98.12	30.65	98.42	34.62	98.1	98.4
7	99.39	74.19	99.44	76.92	99.3	99.4
8	99.39	74.19	99.44	76.92	99.3	99.4
9	98.21	33.87	98.89	53.85	98.2	98.9
10	98.30	37.10	98.70	46.15	98.3	98.7
11	98.30	37.10	98.70	46.15	98.3	98.7
12	98.03	27.42	98.33	30.77	98	98.3
13	98.03	27.42	98.14	23.08	98	98.1
14	97.64	12.90	97.22	15.38	97.6	97.2
15	98.03	27.42	98.14	23.08	98	98.1
16	99.39	74.19	99.44	76.92	99.3	99.4
17	98.08	29.03	98.42	34.62	98.1	98.4
18	98.34	35.48	98.88	42.31	98.3	98.6
19	99.48	77.42	99.44	76.92	99.4	99.4
20	98.21	33.87	98.89	53.85	98.2	98.9
21	97.86	17.74	98.33	19.23	97.8	98.1
22	98.12	27.42	98.51	34.62	98	98.4
23	99.39	74.19	99.44	76.92	99.3	99.4

Tablo 6.11.(a) Analiz sonucunda 2-3-4-7-8-16-19-23 no’lu veri kümeleri ile elde edilen kurallar.

Veri Kümesi Numarası	Kurallar
2	- EĞER Uçuş Tipi=1 VEYA Uçuş Tipi=2 VEYA Uçuş Tipi= 3 VEYA Uçuş Tipi=4 VEYA Uçuş Tipi=7 VEYA Uçuş Tipi=9 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=5 VE Uçuş Planı= 1 VEYA Uçuş Planı= 2 VEYA Uçuş Planı= 3 VEYA Uçuş Planı= 6 O HALDE kaza sonucu ölüm=HAYIR - EĞER Uçuş Tipi=6 VEYA Uçuş Tipi=10 O HALDE kaza sonucu ölüm = EVET
3	- EĞER Uçuş Tipi=1 VEYA Uçuş Tipi=2 VEYA Uçuş Tipi= 3 VEYA Uçuş Tipi=4 VEYA Uçuş Tipi=7 VEYA Uçuş Tipi=9 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=5 VE Hasar Tipi=2 VEYA Hasar Tipi=3 O HALDE kaza sonucu ölüm=HAYIR - EĞER Uçuş Tipi=5 VE Hasar Tipi=1 O HALDE kaza sonucu ölüm= EVET - EĞER Uçuş Tipi=5 VE Hasar Tipi=2 VEYA Hasar Tipi=3 VEYA Hasar Tipi=4 VEYA Hasar Tipi=5 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=6 VEYA Uçuş Tipi=10 O HALDE kaza sonucu ölüm = EVET
4	- EĞER Uçuş Tipi=1 VEYA Uçuş Tipi=2 VEYA Uçuş Tipi= 3 VEYA Uçuş Tipi=4 VEYA Uçuş Tipi=7 VEYA Uçuş Tipi=9 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=5 VE Hasar Tipi=2 VEYA Hasar Tipi=3 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=5 VE Hasar Tipi= 1 O HALDE kaza sonucu ölüm= EVET - EĞER Uçuş Tipi=6 VEYA Uçuş Tipi=10 O HALDE kaza sonucu ölüm = EVET
7	- EĞER Uçuş Tipi=1 VEYA Uçuş Tipi=2 VEYA Uçuş Tipi= 3 VEYA Uçuş Tipi=4 VEYA Uçuş Tipi=7 VEYA Uçuş Tipi=9 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=5 VE Hasar Tipi=1 VEYA Hasar Tipi=2 VEYA Hasar Tipi=3 VEYA Hasar Tipi=4 VEYA Hasar Tipi=5 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=6 VEYA Uçuş Tipi=10 O HALDE kaza sonucu ölüm = EVET
8	- EĞER Uçuş Tipi=1 VEYA Uçuş Tipi=2 VEYA Uçuş Tipi= 3 VEYA Uçuş Tipi=4 VEYA Uçuş Tipi=7 VEYA Uçuş Tipi=9 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=5 VE Hasar Tipi=1 VEYA Hasar Tipi=2 VEYA Hasar Tipi=3 VEYA Hasar Tipi=4 VEYA Hasar Tipi=5 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=6 VEYA Uçuş Tipi=10 O HALDE kaza sonucu ölüm = EVET

Tablo 6.11. (devamı) (b) Analiz sonucunda 2-3-4-7-8-16-19-23 no'lu veri kümeleri ile elde edilen kurallar.

16	- EĞER Uçuş Tipi=1 VEYA Uçuş Tipi=2 VEYA Uçuş Tipi= 3 VEYA Uçuş Tipi=4 VEYA Uçuş Tipi=7 VEYA Uçuş Tipi=9 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=5 VE Hasar Tipi=1 VEYA Hasar Tipi=2 VEYA Hasar Tipi=3 VEYA Hasar Tipi=4 VEYA Hasar Tipi=5 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=6 VEYA Uçuş Tipi=10 O HALDE kaza sonucu ölüm = EVET
19	- EĞER Uçuş Tipi=1 VEYA Uçuş Tipi=2 VEYA Uçuş Tipi= 3 VEYA Uçuş Tipi=4 VEYA Uçuş Tipi=7 VEYA Uçuş Tipi=9 O HALDE kaza sonucu ölüm= HAYIR - EĞER Uçuş Tipi=6 VEYA Uçuş Tipi=10 O HALDE kaza sonucu ölüm = EVET
23	- EĞER Uçuş Tipi=1 VEYA Uçuş Tipi=2 VEYA Uçuş Tipi= 3 VEYA Uçuş Tipi=4 VEYA Uçuş Tipi=7 VEYA Uçuş Tipi=9 O HALDE kaza sonucu ölüm= HAYIR -EĞER Uçuş Tipi=5 VE Hasar Tipi=3 VE Uçuş Planı=1 VEYA Uçuş Planı=2 VEYA Uçuş Planı=3 VEYA Uçuş Planı=6 O HALDE kaza sonucu ölüm=HAYIR - EĞER Uçuş Tipi=5 VE Hasar Tipi =1 VEYA Hasar Tipi =2 VEYA Hasar Tipi =4 VEYA Hasar Tipi =5 O HALDE kaza sonucu ölüm=HAYIR - EĞER Uçuş Tipi=6 VEYA Uçuş Tipi=10 O HALDE kaza sonucu ölüm = EVET

6.4.2. Sonuçlar

Analiz sonucunda elde edilen ve Tablo 6.11’de verilen yüksek doğruluklu kurallar incelendiğinde ortaya çıkan sonuçlar şu şekildedir. Uçuş tipi 6 ve 10 olan paraşüt atlama ve ışıktandırma uçuşları, diğer parametrelerden bağımsız olarak %99.4 doğrulukla ölümle sonuçlanmaktadır. Diğer taraftan uçuş tipinin 1,2,3,4,7 ve 9 olduğu durumlarda %99.4 doğrulukla ölümlü kazalar olmamaktadır. Uçuş tipinin 5 olduğu genel uçuşlarda ise parçalanmaların olduğu kazalarda %99.4 doğrulukla ölümler meydana gelmektedir. Belli bir uçuş planı olmayan uçuşlar da, herhangi bir hasar olmasa dahi %99.4 oranında ölümle sonuçlanmaktadır. Hasar tipi 2,4 veya 5 ve uçuş tipinin 5 olduğu kazaların ölümle sonuçlanmadığı da %99.4 doğrulukla elde edilen kurallardan çıkarılmıştır. Elde edilen kurallara ait olasılık, etkinlik ve doğruluk değerlerine bakıldığında sonuçların anlamlı olduğu görülmektedir.

Tablo 6.7’de görüldüğü gibi 8 parametre, {Olay Tarihi, Hasar Tipi, Uçuş Tipi, Olay Tipi, Uçuş Planı, Motor Sayısı, Pilot Sertifikası, Ülke} ile elde edilen sınıflandırma

sonuçları en yüksek olasılık, etkinlik ve doğruluk değerlerini vermektedir. Ancak kaba kümeler kullanılarak sadece 3 parametre ile {Olay Tarihi, Ülke, Uçuş Tipi}, aynı doğruluk, yakın olasılık ve etkinlik değerleri elde edilmiştir.

Yapılan çalışma ile karar ağaçları, kaba kümeler ve bağımlılıkları bulma algoritmaları veri içerisindeki etkili ve doğru bilginin bulunmasında kullanılabilir etkin metotlardır. Elde edilen sonuçlar kullanıcının kolayca anlayabileceği ve yorumlayabileceği şekildedir. Ayrıca geleceğe yönelik tahminler için de kullanılabilirler.

Mevcut verilerin analizi sonucunda, paraşüt ile atlama ile ilgili uçuşlarda daha dikkatli olunması gerektiği sonucu çıkmıştır. Uçuşlarda ticari pilotlar değil de sertifikalı pilotlar yer almalıdır.

6.5. Sonuç ve Tartışma

Çalışmanın bu bölümünde, havacılık sektöründeki kaza verileri üzerinde farklı veri madenciliği tekniklerinin uygulaması gerçekleştirilmiştir. Çalışmadaki temel amaç, kaba küme yaklaşımının etkinliği ve testteki doğruluğunu kanıtlamaktır. İlk aşamada, mevcut veri kümesindeki gürültü ve gereksiz kayıtlar temizlenmiş ve girdi ile çıktı parametreleri arasındaki ilişkiler tespit edilmiştir. İkinci aşamada ise, kaza verilerinden örüntüler çıkarmak için karar ağaçları algoritması veriler üzerinde uygulanmıştır.

KK'ler ile tespit edilen nitelik sayısı az olmasına rağmen, karar ağaçları ile birlikte kullanıldığında yüksek sınıflandırma olasılığı, etkinliği ve test doğruluğu vermektedir. Yani KK yaklaşımı, tercih edilebilir sınıflandırma kuralları ve anlamlı sonuçlar elde etmeyi sağlayan etkin parametrelerin tespit edilmesini sağlamıştır.

7. BÖLÜM

UYGULAMA

7.1. Giriş

Daha önceki bölümde bahsedildiği gibi veri madenciliği teknikleri farklı alanlarda başarılı şekilde uygulanmaktadır. Bu alanlardan bir tanesi de havacılık sektörüdür. Havacılık sektöründe oldukça büyük miktarlarda veri birikimi söz konusudur. Bu veriler, pilot raporları, bakım raporları, kaza raporları veya tehir raporları şeklinde tutulmaktadır. Bu bölümde, bir hava yolu şirketinin güvenilirlik biriminde yapılan veri madenciliği çalışması kapsamlı şekilde anlatılacaktır. Bu çalışma kapsamında, uygulamanın yapıldığı hava yolu şirketinin yetkili kişileri ile yapılan görüşmeler sonucunda ilgili şirketin güvenilirlik biriminde veri madenciliği çalışmasının yapılması uygun görülmüştür. Güvenilirlik birimi veri tabanında yer alan kayıtlar incelenmiş ve analizde kullanılabilecek olanlar tespit edilmiştir. Yapılan veri madenciliği çalışmasında, meydana gelen sefer aksaklıklarının teknik tehir olarak değerlendirilip değerlendirilmeyeceği, kayıt altına alınan pilot, bakım ve kabin raporlarında ne tip kayıtların tutulduğunun tespit edilmesi, parça ikaz seviyelerinin ve filo bazında gerçekleşen gecikme oranlarının tespitine yönelik sınıflandırma kurallarının oluşturulması amaçlanmıştır. Bu amaçla yapılan çalışma süreci şekil 7.1’de özetlenmiştir. Analiz sürecinde sınıflandırma kurallarının oluşturulması için karar ağaçları, kural bulma, lineer regresyon, polynet gibi sınıflandırma teknikleri ile çeşitli veri ön işleme araçları kullanılmıştır.

Veri madenciliğinin ilk sürecinde veri tabanında yer alan veriler incelenerek, problemlerin tanımlandıktan sonra verilerin analiz sürecine hazırlanması amacıyla temizlenmesi, birleştirilmesi, dönüştürülmesi gibi ön işleme teknikleri uygulanmıştır.

Ön işleme sonucunda analize uygun forma getirilen veriler çeşitli veri madenciliği teknikleri ile analiz edilmiştir.

Kolaylık açısından veri işleme süreci MS Excell programında gerçekleştirilmiştir. Her veri kümesi önce kaynak dosyalardan MS Excell'e aktarılarak gerekli görsel, matematiksel düzeltmeler yapılmıştır.

Bunu takiben verilerin temizlenmesi, başka değerlerle değiştirilmesi, ayrıştırılması ya da kategorize edilmesi için excell fonksiyonlarından istifade edilmiştir.

Programın zorunlu kıldığı ya da kolaylık için başvuru alan veri işleme sürecinden sonra problem gerektiği şekilde programa aktararak, bazı durumlarda veriyi okumak için kalıcı olan ve olmayan değişiklikler program üzerinden gerçekleştirilmiştir.

Ayrıca genellikle çözüm sürecinin başında bir kazanç getirebileceği düşüncesiyle PolyAnalyst programının istatistiksel özetleme modülü ve bağımlılıkları bulma modülü ön işleme araçları olarak kullanılmıştır. İstatistiksel özetleme işlemi matematiksel bir yönü olmayan ancak tüm veriyi özetleyerek tablo ve görsel öğelerle kullanıcıya bilgi sunan bir özelliğe sahiptir.

7.2. Değerlendirilen Veri Kümelerinin Tanıtılması

Uygulama için ilgili hava yolu şirketinin güvenilirlik birimi veri tabanından alınan 6 grup veri incelenmiştir. Bunlar;

- Sefer Aksaklıkları Raporları
- Pilot/Kabin/Bakım (PMC) Raporları
- Parça Raporları
- Parça Söküm Raporları
- Gecikme Oranları
- Motor Parametreleri

Bu verilerin incelenmesi sonucunda 6 tip veriden sadece 4 tanesinin (Sefer Aksaklıkları

Raporları, Pilot/Kabin/Bakım (PMC) Raporları Parça Raporları, Gecikme Oranları) analiz sürecinde kullanılabileceğine karar verilmiştir. Bu verilere ait tanımlamalar ve analiz süreci her bir veri kümesi için sırasıyla aşağıdaki bölümlerde anlatılacaktır.

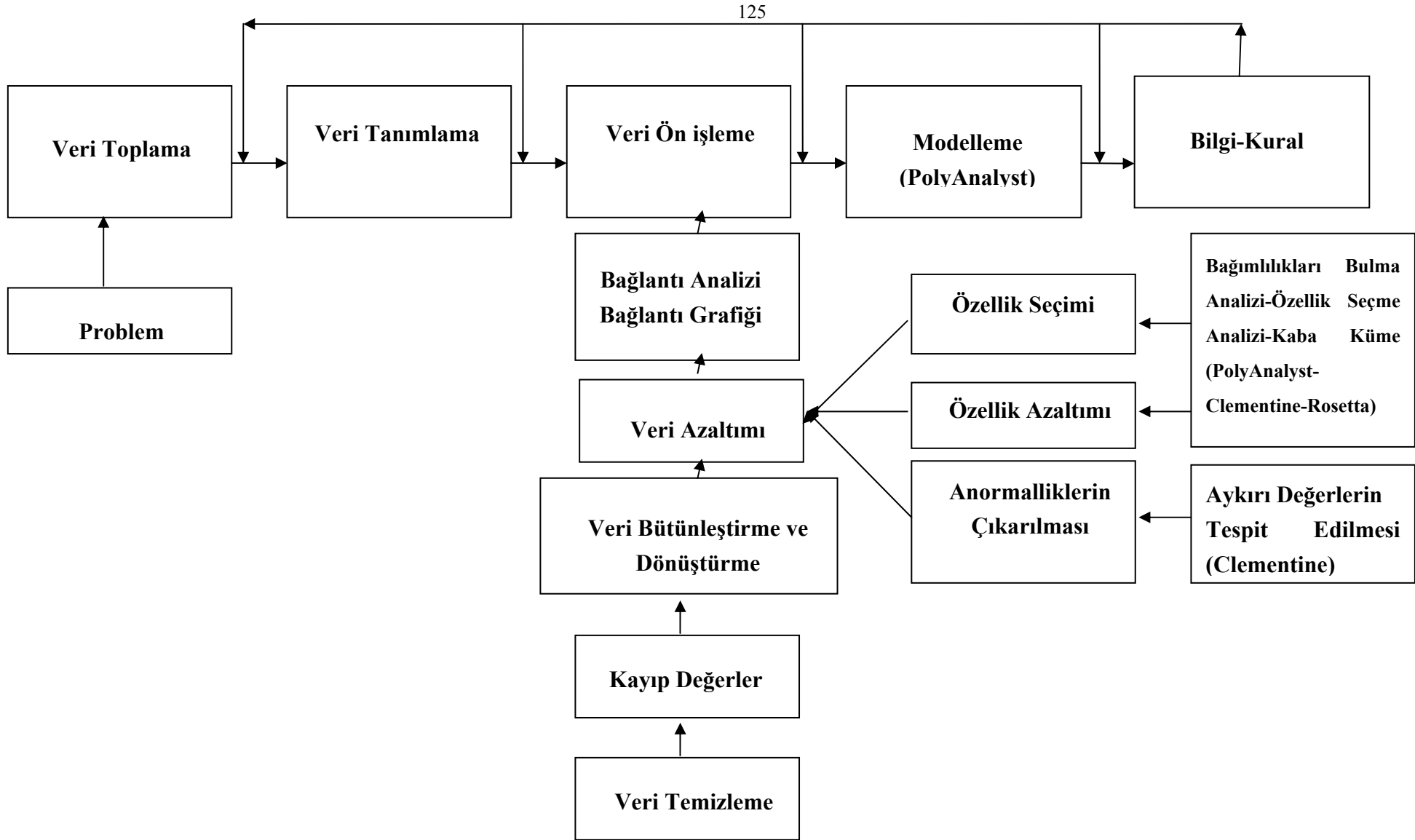
7.2.1. Sefer Aksaklıkları Raporları

Uygulama amacıyla analiz edilmek üzere ele alınan verilerden birincisi sefer aksaklıkları ile ilgili olarak tutulan raporlardır.

Teknik nedenler sonucunda oluşan sefer aksama bilgileri (tehirler, iptaller, havadan geri dönüş ve sapmalar, ramp'a geri dönüş v.s) Hat Bakım Arıza Kontrol Şefliği tarafından sisteme girilir. Sefer Aksaklığı raporunda, aksaklığın nedeni ve hangi sistem veya parçanın arızalandığı açık bir şekilde belirtilir.

Sefer aksaklıkları raporu veri tabanından ilk alındığı aşamada 65 parametre içermektedir. Verilerin uzman kişilerle birlikte değerlendirilmesi sonucunda, boş ve tekrarlı parametrelerin çıkarılması ve veri birleştirme işlemleri sonucunda, analizde kullanılacak 16 parametre seçilmiştir. Programdaki analizlere bu 16 parametre ile başlanmıştır. Veri kümesi, 6207 adet kayıt içermektedir. Bu kayıtların üçte birine karşılık gelen 2069 adet kayıt, test verisi olarak rasgele seçilmiştir. Geriye kalan 4138 adet kayıt ise eğitim verisi olarak alınmıştır.

Analizde kullanılan parametreler ve ilgili tanımlar özetle tablo 7.1'de gösterilmiştir.



Şekil 7.1.Çalışma kapsamında yapılan veri madenciliği süreci.

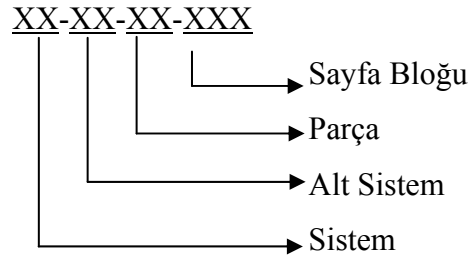
Tablo 7.1. Sefer aksaklıkları raporlarına ait parametreler ve özellikleri.

Parametre Adı	Tipi	Veri çeşidi	Tanımı
Ata Kodu	Bağımsız	38 farklı kategorik veri	Sırasıyla uçağa ait sistem, alt sistem, parça ve sayfa bloğu'nu tanımlar.
Tarih	Bağımsız	03.04.2006 şeklinde tarih	Uçağın uçuş tarihini göstermektedir.
Kalkış Limanı	Bağımsız	72 farklı dizi veri	Uçağın hareket ettiği hava limanı.
Varış Limanı	Bağımsız	108 farklı dizi veri	Uçağın indiği hava limanıdır.
Tarife Geliş Saati	Bağımsız	14:00:00 şeklinde saat, dakika, saniye şeklindedir.	Uçağın kalkış için havalimanına geldiği saattir.
Tarife Kalkış Saati	Bağımsız	14:00:00 şeklinde saat, dakika, saniye şeklindedir.	Uçağın kalkış saatidir.
Tehir Kodu	Bağımsız	9 farklı kategorik veri	Uçak sefer aksaklıklarının kaynağı.
Tehir Süresi	Bağımsız/Hedef	0-835 dakika arasında sürekli veri	Uçak hareketinin ertelenme süresidir.
Arızalı Uçak	Bağımsız	92 farklı dizi veri	Tehir yapan uçak kodu.
Arıza Sonucu Durum	Bağımsız	10 farklı kategorik veri	Ertelenen uçuş sonrasında uçakla ilgili gerçekleşen olay.
İşlem Kodu	Bağımsız	13 farklı kategorik veri	Yapılmış olan ilgili işlemin kodu.
Filo	Bağımsız	7 farklı kategorik veri	Uçağın hangi filoya ait olduğunu gösteren koddur.
Teknik Olay	Bağımsız	Evet/Hayır olarak tanımlı ikili (binary) veri	Gecikme dışında önemli bir teknik olay olup olmadığıdır.
Uçuş Esnasında Motor Durması	Bağımsız	Evet/Hayır olarak tanımlı ikili (binary) veri	Uçuş esnasında motorun durup durmadığıdır.
Bakım Çıkışı Gecikmesi	Bağımsız	Evet/Hayır olarak tanımlı ikili (binary) veri	Bakımdan kaynaklanan bir gecikme olup olmadığıdır.
Teknik Tehir	Hedef (Bağımlı)	Evet/Hayır olarak tanımlı ikili (binary) veri	Gecikmenin teknik tehir olması ya da olmamasıdır.

7.2.1.1. Sefer Aksaklıkları Raporlarında Yer Alan Parametreler ve Tanımları

Bu bölümde sefer aksaklıkları raporlarında yer alan parametrelerin açıklaması sırasıyla yapılmıştır.

- **Ata Kodu:** Her bir uçak tipi için aynı standart dokümanlara sahip olmak amacıyla Amerikan Hava Ulaşım Birliği (Air Transport Association of America-Ata), imalatçıların dokümanlarının birbirleriyle daha uyumlu hale getirilmesini sağlamıştır. Bu amaçla hazırlanan Ata kodlama yapısında, her bir sistem tipi için bir numara tanımlanmıştır. Ata kodları, iki haneli üçlü grup ve devamı üç haneli tek gruptan oluşmak üzere, aşağıda belirtilen şekilde, sırasıyla uçağa ait sistem (kodu/system), alt sistem (section/subsystem), parça (subject/unit), sayfa bloğu (page block)'nu tanımlar.



Şekil 7.2. Ata Kodu yapısı.

Buna bir örnek olarak:

52 Kapı

52-11 Yolcu kapısı

52-11-02 Yolcu kapısı kolu

52-11-02-401 Yolcu kapı kolu için söküm/takım prosedürü verilebilir.

Ata numaralandırma yapısındaki ilk iki hane sistemi tanımlamak üzere tüm imalatçılar tarafından ortak kullanılır. Alt sistem ve parça bir imalatçıdan diğerine veya tasarım farklılıklarından dolayı aynı imalatçıya ait bir uçak modelinden diğerine farklılık gösterebilmektedir. Son grup olan sayfa blokları bilgisi tüm imalatçılar için aynı olmaktadır.

Hava yolu şirketi veri tabanında yer alan Ata-kodu parametresi 4 haneli olarak

alt sistem ve sistemi temsil etmektedir. Analizde sadece ilk iki hane olan sistemle ilgili parametreler kullanılmıştır.

- Tarih: Uçağın uçuş tarihini göstermektedir. (Örnek: 21.01.2006)
- Kalkış ve Varış Limanı: Uçağın hareket ettiği ve indiği hava limanıdır. (IST=İSTANBUL)
- Teknik Tehir Kodu: Uçak sefer aksaklıklarının sebepleri şirket içerisinde, kaynağına göre sınıflandırılmıştır. Bunlardan bazıları, yolcu ve bagaj kaynaklı, uçak hasarlanması, hava, meydan ve idari otoriteler, teknik ve uçak techizatı şeklindedir. Çalışmada incelenen raporlar teknik ve uçak techizatına bağlı olan tehirlerdir. Bu veriler kategorik veri olarak analizde yer almıştır. Teknik tehir kodları ve açıklamaları tablo 7.2’de gösterilmiştir.
- Tehir Süresi: Uçağın hareketinin ne kadar süre ertelendiğidir. Dakika cinsinden hesaplanmıştır.
- Arızalı Uçak (Uçak Kodu): Tehir yapan uçağın kodudur.
- Arıza Sonucu Durum: Ertelenen uçuş sonucunda uçağın durumu olarak nitelendirilen bu veri 11 kategorik durumu içermektedir. Bu durumlar ve açıklamaları tablo 7.3’de gösterilmiştir.
- İşlem Kodu: Yapılmış olan ilgili işlemin kodudur.
- Tarife Geliş ve Kalkış Saati: Uçağın havaalanına geliş ve havaalanından kalkış saatini ifade eder.
- Filo: Uçağın hangi filoya ait olduğunu gösteren koddur.
- Uçuş Esnasında Motor Durması: Uçuş esnasında motorun duruşu ile ilgili bilgiyi verir.
- Bakım Çıkışı Gecikmesi: Bakımdan kaynaklanan gecikme durumunu gösterir.
- Teknik Olay: Gecikme dışında önemli bir teknik olay olup olmadığıdır.
- Tehir Hesabına Katılacak mı?: Gecikmenin teknik tehir olarak dikkate alınıp alınmayacağını ifade etmektedir.

Tablo 7.2. Teknik tehir kodları ve açıklamaları.

Teknik tehir kodu	Açıklamalar
41	Uçak arızaları
42	Periyodik bakım, bakımdan geç teslim
43	Periyodik harici bakım, özel kontroller ve /veya normal periyodik bakım dışında kalan ek çalışmalar
44	Yedek parça ve bakım teçhizatı, mevcut olmaması veya arızalı durumda olması
45	Arızalı uçak için yedek parçaların diğer bir istasyona taşınması
46	Teknik nedenlerden dolayı uçak değişikliği
47	Yedek uçak, teknik nedenlerle planlanmış yedek uçağın mevcut olmaması
48	Belirlenmiş kabin ayırımı / sınıflara göre koltuk düzenlenmesi

7.2.1.2. Sefer Aksaklıkları Raporlarına Uygulanan Ön İşlemler

Veri ön işleme sürecinde yapılan ilk işlem gereksiz olduğu düşünülen parametreler ile boş-bilgi bulunmayan- ve tekrarlı (içerik aynı fakat parametre isimleri farklı olan) parametrelerin çıkarılması olmuştur. Bu adımdan sonra mevcut verilerde yer alan eksik bilgiler ilgili birimdeki uzmanların desteği ile tamamlanmıştır. Daha sonra arıza sonucu durum parametresi adı altında altı farklı parametre birleştirilerek tek bir parametre olarak düzenlenmiştir. Bu adımlardan sonra raporda yer alan 65 parametre 16 parametreye indirilmiştir. Daha sonra 16 parametreye ait eksik bilgiler uzman kişilerin yardımı ile tamamlanmıştır. Eksik bilgilerin tamamlanması esnasında eski kayıtlar incelenmiş, paralel bilgilere göre uzman kişilerin öneri ve fikirleri de alınmıştır. Bu parametreler ve her birine yapılan ön işlemler tablo 7.4’de gösterilmiştir.

Tablo 7.3. Arıza sonucu durumlar ve açıklamaları.

Arıza Sonucu Durum	Açıklama
1	Cancellation: İptal
2	Delay: Geç kalkma, rötar.
3	Delay ve Diversion: Gecikme ve başka bir yere gitme.
4	Delay ve Air Turn Back: Gecikme ve geri dönüş.
5	Delay ve Return To Gate: Gecikme ve kalkıştan dönme.
6	Diversion: Başka bir yere gitme.
7	Air Turn Back: Geri dönüş.
8	Return To Gate: Kalkıştan dönüş.
9	Substitute Aircraft: Uçak değişimi.
10	Substitute Aircraft ve Air Turn Back: Uçak değişimi ve geri dönüş.
11	Substitute Aircraft ve Return To Gate: Uçak değişimi ve kalkıştan dönüş.

İlk ön işleme sürecinin ardından, mevcut 16 parametrenin sefer aksaklıklarının teknik tehir olarak değerlendirilip değerlendirilmeyeceğinde etkilerinin olup olmadığı paket programlardan PolyAnalyst, Clementine ve Rosetta araçları kullanılarak ön işleme sürecine devam edilmiştir. Her bir ön işleme süreci sırasıyla aşağıda anlatılmıştır.

7.2.1.2.1. PolyAnalyst Aracı ile Bağımlılıkların Tespiti

Sefer aksaklıkları raporlarının analizinde hedef değişken olarak gecikmelerin teknik tehir olarak değerlendirilip değerlendirilmeyeceği ile ilgili olan parametre tespit edilmiştir. Hedef parametre ile ilgili kural oluşturma sürecine geçmeden önce, analiz sürecinde kullanılacak hedef parametreye etkisi olan parametrelerin belirlenmesi gerekmektedir. Bunun için PolyAnalyst programında yer alan bağımlılıkları bulma (Find Dependencies) algoritması kullanılmıştır. Bağımlılıkları bulma algoritması veri kümesi içerisindeki bağımlılıkları ve ilişkileri hızlı bir şekilde tespit etmek için kullanılır. Çıktı olarak elde edilen değerler içerisinde yer alan P değerine göre bağımlılığın şiddetine karar verilmektedir. Bu değer sıfıra ne kadar yakınsa parametreler

arasında o kadar güçlü bağımlılık olduğu sonucu çıkarılmaktadır. P değeri 10^{-7} değerinden küçük olursa sıfır (0) olarak kabul edilmiştir [125].

Tablo 7.4. Sefer Aksaklıkları üzerinde yapılan ön işlemler.

Parametre Adı	Yapılan ön işlem
Ata Kodu	Dört haneden iki haneye indirilmiş, eksik bilgiler tamamlanmıştır.
Tarih	Formatı programda kullanılmaya uygun şekle getirilmiştir.
Arızalı Uçak	Eksik bilgiler kısmen tamamlanmıştır.
Sefer No	Eksik bilgiler kısmen tamamlanmıştır.
Kalkış Limanı	Eksik bilgiler kısmen tamamlanmıştır.
Varış Limanı	Eksik bilgiler kısmen tamamlanmıştır.
Tarife Geliş Saati	Formatı programda kullanılmaya uygun şekle getirilmiştir.
Tarife Kalkış Saati	Formatı programda kullanılmaya uygun şekle getirilmiştir.
Teknik Tahir Kodu	Dört Haneli kodlar iki haneli olarak yeniden düzenlenmiş ve eksik bilgiler kısmen tamamlanmıştır.
Tahir Süresi	Dakika cinsinden yeniden hesaplanarak sayısal formatta yazılmıştır.
Teknik Tahir	Eksik bilgiler kısmen tamamlanmıştır.
İşlem Kodu	Metinsel olarak alınan işlem kodları sınıflandırılarak kategorik olarak yeniden yazılmıştır.
Filo	Mevcut filolar yeniden kodlanmıştır ve eksik bilgiler kısmen tamamlanmıştır.
Arıza Sonucu Durum	Sefer aksaklığı sonucunda uçuşun gecikmesi, iptali, uçağın geri dönmesi.... gibi durumlar tek bir parametrede birleştirilmiştir.

Yapılan analiz sonucunda elde edilen P değerlerine bağlı olarak sefer aksaklıklarının teknik tehir olarak değerlendirilip değerlendirilmeyeceği konusuna etkili olan parametre kümesi, {Tehir süresi, Uçuş esnasında motor durması, Ata kodu, Tehir kodu, Teknik olay, Kalkış limanı, Arıza sonucu durum, Bakım çıkışı gecikmesi}olarak tespit edilmiştir. İleriki veri madenciliği analizlerinde bu parametreler kullanılacaktır. Bu analize ait P değerleri ve ilgili parametreler tablo 7.5’de gösterilmiştir

Tablo 7.5. Bağımlılıkları bulma algoritması ile belirlenen etkili parametreler.

Parametre adı	P-değeri
Tehir Süresi	0
Uçuş Esnasında Motor Durması	0
Ata Kodu	$1.662 e^{-20}$
Tehir Kodu	$4.989 e^{-14}$
Teknik Olay	$4.989 e^{-14}$
Kalkış Limanı	$2.383 e^{-13}$
Arıza Sonucu Durum	$9.804 e^{-7}$
Bakım Çıkışı Gecikmesi	$9.804 e^{-7}$

Bir sonraki aşamada gerçekleştirilen veri madenciliği analiz sürecinde tablo 7.5’de verilen parametre kümesi kullanılmıştır.

7.2.1.2.2. Bağlantı Grafiği Kullanarak Korelasyonu Gösterme

Parametrelerin özel değerleri arasındaki korelasyonun karmaşık örneklerini görsel olarak açığa çıkarmak amacıyla bağlantı grafiği oluşturulmuştur. Bağlantı grafiğinin görsel çıktısı araştırılan verinin gizli yapısının daha iyi anlaşılmasına yardımcı olmuştur ve ileriki aşamalar için temel oluşturmuştur. Bağlantı analizi çok boyutlu korelasyon analizi esasına bağlı olarak çalışmaktadır [125].

Korelasyon, iki nesne arasındaki bağlantının kuvvetini ifade etmektedir. Bu değer $[-1,1]$ arasında yer alan pozitif veya negatif bir değer olabilir. Korelasyon bağlantı grafiği incelendiğinde tablo 7.6’da gösterildiği gibi tehirlere değerlendirilmesinde etkili olan parametrelerden, tehir kodunun 43 ve 46 değerleri; arıza sonucu durum parametresinin

1,2,7 ve 9 değerleri; uçaklardan JGH kodlu uçağın; kalkış limanının TLV (Ben Gruion), IST (İstanbul), ESB(Ankara) ve MXP (Malsenpa) hava limanları; Ata kodu 27 değerlerinin önemli olduğu görülmüştür. Ayrıca teknik olaylar ve bakım çıkışı gecikmesi durumları ile de büyük ancak negatif bir etki olduğu görülmüştür.

Tablo 7.6. Bağlantı grafiklerinden elde edilen korelasyon katsayısı değerleri.

Parametre Adı ve Değeri	Korelasyon Katsayısı
Tehir Kodu 43	-0.99
Tehir Kodu 46	0.92
Arıza sonucu durum 9	0.99
Arıza sonucu durum 2	-0.78
Arıza sonucu durum 1	-0.63
Arıza sonucu durum 7	-0.63
Arızalı Uçak JGH	-0.99
Kalkış Limanı IST	-0.99
Kalkış Limanı TLV	0.97
Kalkış Limanı ESB	-0.66
Kalkış Limanı MXP	0.60
Ata kodu 27	0.58
Bakım Çıkışı Gecikmesi Evet	-0.99
Teknik olay Evet	-0.99

7.2.1.2.3. Kaba Küme Yaklaşımı ile Etkili Parametrelerin Tespiti

Ön işleme sürecinde yer alan uygulamalardan bir diğeri kaba küme yaklaşımı kullanılarak çıktı parametresine etkisi olan veri kümelerinin tespit edilmesi sürecidir. Bu aşamada uygulama için kaba küme yaklaşımı esasına dayanan Rosetta adlı paket program kullanılmıştır. Analiz sonucunda elde edilen etkili veri kümeleri (Rs) tablo 7.7’de gösterilmiştir.

Tablo 7.7. Kaba küme yaklaşımı ile elde edilen veri kümeleri.

Sıra No	Veri Kümesi
Rs1	{ata kodu, varış limanı, tehir süresi, arızalı uçak, filo}
Rs2	{ata kodu, varış limanı, tehir süresi, arızalı uçak, işlem kodu}
Rs3	{kalkış limanı, varış limanı, tehir kodu, tehir süresi, arızalı uçak, işlem kodu}
Rs4	{kalkış limanı, varış limanı, tehir süresi, arızalı uçak, işlem kodu, filo, arıza sonucu durum, bakım çıkışı gecikmesi}
Rs5	{kalkış limanı, varış limanı, tehir süresi, işlem kodu, arıza sonucu durum, teknik olay, havada motor durması}
Rs6	{ata kodu, kalkış limanı, tehir kodu, tehir süresi, arızalı uçak, işlem kodu, filo, arıza sonucu durum, bakım çıkışı gecikmesi}

PolyAnalyst programı yardımıyla elde edilen etkili parametrelerle benzer şekilde kaba kümeler yardımıyla elde edilmiş olan veri kümeleri de analiz sürecinde ayrı ayrı değerlendirilmiştir.

7.2.1.2.4. Clementine Aracı ile Etkili Parametrelerin Tespiti

SPSS Clementine özellik seçimi algoritması sonucunda sadece tehir süresi etkili parametre olarak tespit edilmiştir. Tehir süresine bağlı olarak lineer regresyon algoritması kullanılarak %78.35 doğruluk ile aşağıdaki kural elde edilmiştir.

Eğer tehir süresi ≤ 15 ise teknik tehir değildir.

Eğer tehir süresi > 15 ise teknik tehirdir.

7.2.1.3. Analiz Süreci

Sefer aksaklıkları raporlarının veri madenciliği teknikleri ile analizinde PolyAnalyst programına ait karar ağacı ve sınıflandırma algoritmaları (kural bulma, lineer regresyon ve polynet) kullanılmıştır. Analiz için ön işleme sürecinde elde edilen veri kümeleri sırasıyla ele alınmıştır.

Karar ağaçları tekniği ile yapılmış olan analiz sonucunda her bir girdi değişkenleri ile elde edilen kurallar tablo 7.8 (a-b)'de verilmiştir. Analiz sonuçlarına ait sınıflandırma olasılıkları, etkinlikleri ve doğrulukları ise tablo 7.9'da gösterilmiştir.

Tablo 7.8. (a) Karar ağacı algoritması ile yapılan analiz sonucunda elde edilen kurallar.

Veri Kümesi	Elde Edilen Kural
Rs1	Eğer tehir süresi<15.5 ise ve işlem kodu=1 ise teknik tehirdir. Eğer tehir süresi<15.5 ise ve işlem kodu $\neq$ 1 ise teknik tehir değildir. Eğer tehir süresi $\geq$ 15.5 ise teknik tehirdir.
Rs 2	Eğer tehir süresi<15.5 ise teknik tehir değildir. Eğer tehir süresi $\geq$ 15.5 ise ve Ata kodu=5 ise teknik tehir değildir. Eğer tehir süresi $\geq$ 15.5 ise ve Ata kodu $\neq$ 5 ise teknik tehirdir
Rs 3	Eğer tehir süresi<15.5 ise ve işlem kodu=1 ise teknik tehirdir. Eğer tehir süresi<15.5 ise ve işlem kodu $\neq$ 1 ise teknik tehir değildir. Eğer tehir süresi $\geq$ 15.5 ise ve tehir kodu=40 veya 41 veya 2 veya 43 veya 44 veya 46 veya 47 veya 48 ise teknik tehirdir. Tehir süresi bilinmiyor ve kalkış limanı= IST veya ADB veya ESB ise teknik tehir değildir.
Rs 4	Eğer tehir süresi<15.5 ise ve bakım çıkışı gecikmesi veya bakım çıkışı gecikmesi değil ise teknik tehir değildir. Eğer tehir süresi<15.5 ise ve bakım çıkışı gecikmesi bilinmiyor ise ve filo=5 veya 3 veya 6 veya 1 veya 4 veya 7 ise teknik tehir değildir. Eğer tehir süresi $\geq$ 15.5 ise ve bakım çıkışı gecikmesi değil ise ve arıza sonucu durum= 3 veya 10 veya 5 veya 6 veya 1 veya 13 veya 4 veya 12 ise teknik tehirdir. Eğer tehir süresi $\geq$ 15.5 ise ve bakım çıkışı gecikmesi değil ise ve arıza sonucu durum= 2 veya 11 ise teknik tehir değildir. Eğer tehir süresi $\geq$ 15.5 ise ve bakım çıkışı gecikmesi ise teknik tehir değildir.

Tablo 7.8. (devam) (b) Karar ağacı algoritması ile yapılan analiz sonucunda elde edilen kurallar.

Rs 5	Eğer tehir süresi<15.5 ise ve işlem kodu=1 ise teknik tehirdir. Eğer tehir süresi<15.5 ise ve işlem kodu=8 veya 11 veya 4 veya 5 veya 12 veya 9 veya 6 veya 2 veya 13 veya 7 veya 10 ise teknik tehir değildir. Eğer tehir süresi>=15.5 ise ve işlem kodu=8 veya 11 ise teknik tehirdir. Eğer tehir süresi>=15.5 ise ve işlem kodu bilinmiyor ve arıza sonucu durum=3 veya 10 veya 5 veya 6 veya 1 veya 13 veya 4 veya 12 ise teknik tehirdir. Eğer tehir süresi>=15.5 ise ve işlem kodu=4 ise ve teknik olay veya değilse teknik tehirdir. Eğer tehir süresi>=15.5 ise ve işlem kodu=5 veya 6 veya 7 veya 2 veya 13 veya 10 veya 9 veya 3 veya 1 ise teknik tehirdir. Eğer tehir süresi>=15.5 ise ve işlem kodu=12 ise ve arıza sonucu durum=3 veya 10 veya 5 veya 6 ise teknik tehirdir. Eğer tehir süresi>=15.5 ise ve işlem kodu=12 ise ve arıza sonucu durum=2 ise teknik tehir değildir.
Rs 6	Eğer tehir süresi<15.5 ise ve bakım çıkışı gecikmesi değil ise teknik tehir değildir. Eğer tehir süresi<15.5 ise ve bakım çıkışı olup olmadığı bilinmiyor ve filo=5 veya 3 veya 6 veya 1 veya 4 veya 7 ise teknik tehir değildir. Eğer tehir süresi<15.5 ise ve bakım çıkışı gecikmesi ise teknik tehir değildir. Eğer tehir süresi>=15.5 ise ve tehir kodu=40 ise ve arıza sonucu durum=3 veya 5 veya 6 veya 1 veya 4 ise teknik tehirdir. Eğer tehir süresi>=15.5 ise ve tehir kodu=40 ise ve arıza sonucu durum=2 ise teknik tehir değildir. Eğer tehir süresi>=15.5 ise ve tehir kodu=41 veya 46 ise ve bakım çıkışı gecikmesi değil veya bilinmiyorsa teknik tehirdir. Eğer tehir süresi>=15.5 ise ve tehir kodu=46 ise ve bakım çıkışı

	gecikmesi ise teknik tehir değildir.. Eğer tehir süresi≥ 15.5 ise ve tehir kodu=42 veya 43 veya 44 veya 45 veya 48 veya 47 ise teknik tehirdir. Eğer tehir süresi bilinmiyor ve bakım çıkışı gecikmesi değilse veya bakım çıkışı gecikmesi olup olmadığı bilinmiyorsa teknik tehirdir.
PolyAnalyst (Fd)	Eğer tehir süresi< 15.5 ise teknik tehir değildir. Eğer tehir süresi≥ 15.5 ise ve tehir kodu=40 ve arıza sonucu durum=3 veya 5 veya 6 veya 1 veya 4 ise teknik tehirdir. Eğer tehir süresi≥ 15.5 ise ve tehir kodu=40 ve arıza sonucu durum=2 ise teknik tehir değildir. Eğer tehir süresi≥ 15.5 ise ve tehir kodu=45 ise teknik tehir olup olmadığı bilinmiyor. Eğer tehir süresi≥ 15.5 ise ve tehir kodu=46 ve bakım çıkışı gecikmesi değil ise teknik tehirdir. Eğer tehir süresi≥ 15.5 ise ve tehir kodu=46 ve bakım çıkışı gecikmesi ise teknik tehir değildir. Eğer tehir süresi≥ 15.5 ise ve tehir kodu=41 veya 42 veya 43 veya 44 veya 48 veya 47 ise teknik tehirdir.
Clementine (Fs)	Eğer tehir süresi≤ 15 ise teknik tehir değildir. Eğer tehir süresi> 15 ise teknik tehirdir.

Çalışma kapsamında kullanılan veri madenciliği tekniklerinden sınıflandırma araçları olan kural bulma algoritması, lineer regresyon ve polynet algoritmaları ile her bir veri kümesi ile analizler gerçekleştirilmiştir. Analizde kullanılan sınıflandırma araçları ile ilgili açıklamalar bölüm 5.2.1.4’de ayrıntılı olarak yapılmıştır. Lineer regresyon algoritması ile anlamlı ve doğrulukları yüksek olmayan kurallar elde edilmiştir. Sadece Clementine ile yapılmış olan özellik seçimi ile elde edilmiş veri kümesi ile anlamlı bir kural elde edilmiştir. Polynet algoritması ile yapılmış olan analizler sonucunda ise hiçbir kural elde edilememiştir. Kural bulma algoritması ile bütün veri kümelerinin analizi sonucunda tek bir kural bulunmuştur. Bu kural ve ilgili sonuçlar tablo 7.10 ve 7.11’de verilmiştir.

Tablo 7.9. Karar ağacı analizi ile elde edilen kurallara ait sınıflandırma olasılıkları, etkinlikleri ve doğrulukları.

Veri Kümesi	Eğitim		Test		Eğitim Doğruluğu (%)	Test Doğruluğu (%)
	Sınıflandırma Olasılığı (%)	Sınıflandırma Etkinliği (%)	Sınıflandırma Olasılığı (%)	Sınıflandırma Etkinliği (%)		
Rs1	97.14	92.79	97.32	93.09	97.14	97.32
Rs 2	97.20	92.79	97.18	92.57	97.14	97.12
Rs 3	96.77	91.28	96.85	91.75	96.54	96.78
Rs 4	96.83	91.85	97.05	92.23	96.77	96.98
Rs 5	96.71	91.56	96.85	91.71	96.66	96.78
Rs 6	96.89	92.00	97.05	92.40	96.83	97.05
PolyAnalyst (Fd)	96.89	92.00	97.05	92.40	96.89	97.05
Clementine (Fs)	96.63	91.49	96.98	92.23	96.63	96.98

Tablo 7.10. Kural bulma algoritması ile yapılan analiz sonucunda elde edilen kural.

Veri Kümesi	Elde Edilen Kural
Rs1, Rs2, Rs3, Rs4, Rs5, Rs6, Fd, Fs	Eğer $(0.00219117 * \text{tehir süresi} + 0.501149) > 0.5351$ ise teknik tehirdir aksi halde teknik tehir değildir.

Tablo 7.11. Kural Bulma tekniği ile yapılan analiz sonucunda elde edilen kurallara ait sınıflandırma olasılıkları, etkinlikleri ve doğrulukları.

No	Eğitim		Test		Eğitim Doğruluğu (%)	Test Doğruluğu (%)
	Sınıflandırma Olasılığı (%)	Sınıflandırma Etkinliği (%)	Sınıflandırma Olasılığı (%)	Sınıflandırma Etkinliği (%)		
Rs1	96.68	89.33	97.04	91.19	95.77	96.58
Rs 2	96.68	89.33	97.04	91.19	95.77	96.58
Rs 3	96.68	89.33	97.04	91.19	95.77	96.58
Rs 4	98.36	63.59	98.31	63.56	85.58	85.86
Rs 5	96.68	89.33	97.04	91.19	95.77	96.58
Rs 6	96.68	89.33	97.04	91.19	95.77	96.58
PolyAnalyst (Fd)	96.68	89.33	97.04	91.19	95.77	96.58
Clementine (Fs)	96.68	89.33	97.04	91.19	95.77	96.58

Tablo 7.12. Lineer regresyon algoritması ile yapılan analiz sonucunda elde edilen kurallar.

Veri Kümesi	Elde Edilen Kural
Rs1, Rs2, Rs3, Rs4, Rs5, Rs6, Fd	Güvenilir ve anlamlı kurallar elde edilememiştir.
Fs	Eğer $(0.00219117 * \text{tehir süresi} + 0.501149) > 0.5351$ ise teknik tehirdir aksi halde teknik tehir değildir. (Kural Bulma Algoritması ile elde edilen kuralla aynı kural bulunmuştur.)

Tablo 7.13. Lineer regresyon tekniği ile yapılan analiz sonucunda elde edilen kurallara ait sınıflandırma olasılıkları, etkinlikleri ve doğrulukları.

Veri Kümesi	Eğitim		Test		Eğitim Doğruluğu (%)	Test Doğruluğu (%)
	Sınıflandırma Olasılığı (%)	Sınıflandırma Etkinliği (%)	Sınıflandırma Olasılığı (%)	Sınıflandırma Etkinliği (%)		
Fs	96.68	89.33	97.04	91.19	95.77	96.58

7.2.1.4.Tehir Süresine Etki Eden Parametrelerin Tespit Edilmesi

Yukarıda anlatılan çalışmaların yanı sıra mevcut parametreler içerisinde yer alan tehir süresi parametresine etkisi olan parametreler de tespit edilmiştir. Polyanalyst-bağımlılıkları bulma algoritması ile tespit edilen parametreler tablo 7.14’de gösterildiği gibi {uçuş esnasında motor durması, tarife kalkış saati, teknik olay, teknik tehir kodu, filo, kalkış limanı, arızalı uçak, ata kodu, işlem kodu} parametrelerinden oluşan veri kümesidir.

Tablo 7.14. Tehir süresine etki eden parametreler.

Parametre adı	P-değeri
Uçuş Esnasında Motor Durması	0
Tarife Kalkış Saati	$4.089 e^{-24}$
Teknik Olay	$4.089 e^{-24}$
Tehir Kodu	$1.455 e^{-13}$
Ata Kodu	$7.719 e^{-12}$
Filo	$4.178 e^{-8}$
Kalkış Limanı	$3.721 e^{-7}$
İşlem Kodu	$1.593 e^{-7}$
Arızalı Uçak	$9.896 e^{-7}$

7.2.1.5.Sonuçlar

Yapılan analizlerden karar ağacı ile yapılmış olan analiz sonuçlarına göre en yüksek doğruluk değerlerini veren Rs1, Rs2, Rs6 ve fd veri kümeleri olmuştur. Bu veri kümelerinin karar ağacı analizi sonucunda elde edilen kurallara göre, {ata kodu, işlem kodu, arıza sonucu durum, tehir süresi, bakım çıkışı, filo, tehir kodu} parametrelerinin teknik tehirlerle etkili olduğu sonucu çıkarılmıştır. Bu parametrelerden ise en etkin olanı tehir süresidir.

Sınıflandırma algoritmaları ile yapılan analizlerin sonucunda bütün veri kümelerinden tek bir anlamlı kural elde edilmiştir. Bu kurala göre ise sadece tehir süresi teknik tehir

ile ilgili olarak belirlenmiştir.

Karar ağacı ve sınıflandırma algoritmalarından elde edilen kuralların test doğrulukları dikkate alınarak, sonuçların %97.32 ve %96.98 birbirine çok yakın değerler olduğu tespit edilmiştir. Bu durumda daha basit bir kural tercih edilirse sınıflandırma algoritmalarından kural bulma algoritması ile elde edilen kural; daha karışık bir kural tercih edilirse karar ağaçları ile elde edilen kurallar kullanılabilir.

7.2.2. Pilot/Kabin/Bakım (PMC)Raporları

Analizde kullanılacak bir diğer rapor, PMC (Pilot/Bakım/Kabin) raporlarıdır. Pilot raporları, uçuş ekibi tarafından girilen olay ve arızaları, bakım raporları, uçak bakım yetkili teknisyenleri tarafından uçağın kaydedilen olay ve arızaları, kabin raporları ise kabin görevlisinin yaptığı kayıtları içermektedir. Pilot raporları, uçuş ekibi tarafından tecrübe edilen, uçağın güvenilirliği ile ilgili direkt bir gösterge olması nedeniyle en önemli veri kaynakları arasında yer alır.

PMC raporlarında, Ocak.2000 tarihi ile Mayıs.2006 tarihleri arasına ait toplam 125.653 adet kayıt bulunmaktadır. Bu raporlar ilk olarak farklı iki kaynak şeklinde işletmeden alınmış ve tek bir dosya olarak birleştirilmiştir. Analiz sürecinde, pilot, bakım veya kabin raporlarının parametrelerin en çok hangi düzeyleriyle ilgili olarak tutulduğunu göstermeye yönelik analizler yapılmıştır. Ayrıca her bir değişkenin diğer bir değişkenle olan bağımlılığında etkili düzeyleri tespit edilmiştir. Bu süreçte de, bağlantı analizleri ve grafiklerinden faydalanılmıştır.

Bu raporda yer alan toplam 12 parametre uzmanlarla birlikte yapılan ön işlemler sonucu boş ve gereksiz parametrelerin çıkarılmasıyla, özellikleri tablo 7.15’de gösterilen 9 parametreye indirilmiştir. Bunlar;

- Ata Kodu: Tutulan raporun uçağın hangi bölümüne ait olduğunu göstermektedir.
- Tarih: Arızanın kaydedildiği tarihi göstermektedir.
- Uçak Kodu: Uçağın kuyruk bölümünde yazılı olan numaradır.
- Filo: Uçağın ait olduğu filoyu göstermektedir.
- İstasyon: Arızanın hangi istasyonda kaydedildiğini göstermektedir.
- Sebebi: Metinsel olarak yazılan açıklamaları içermektedir. Arızanın sebebidir.
- Yapılan İşlem: Arızayı gidermek için yapılan işlemin ne olduğunu gösteren metinsel olarak yazılan açıklamaları içermektedir.
- Pireps (P) or Mareps (M) or Cidis (C): Eğer arızayı pilot kaydederse Pilot Raporu (Pirep), bakımda arıza bulunursa Bakım Raporu (Marep), arızayı kabin görevlileri kaydederse Kabin Raporu (Cidis) oluyor. Yapılan

kodlamalar tablo7.16’da gösterilmiştir.

- Uçuş Fazı: Arızanın hangi uçuş fazında meydana geldiğini göstermektedir. İlgili kodlamalar tablo7.17’de gösterilmiştir.

Tablo 7.15. PMC raporlarına ait parametrelerin özellikleri.

Parametre Adı	Özellikleri
Ata Kodu	47 farklı değer içeren kategorik bir veri
Tarih	Tarih
Uçak Kodu	112 farklı değerde dizi
Filo	9 farklı kategorik veri
İstasyon	84 farklı dizi
Sebebi	Metin
Yapılan İşlem	Metin
Pireps (P) or Mareps (M) or Cidis (C)	Kategorik veridir.
Uçuş Fazı	12 farklı değerden oluşan kategorik veri

Uzmanların yardımıyla yapılan özellik indirgeme işleminin ardından, mevcut 9 parametre içerisinden, ata kodu, filo, pmc ve uçuş fazı parametreleri kategorik olarak yeniden kodlanmıştır.

Tablo 7.16. PMC Açıklamaları.

P (3)	Pilot Raporu
M (4)	Bakım Raporu
C (1)	Kabin Raporu

7.2.2.1. Pilot Raporlarının Tutulmasında Etkili Olan Parametrelerin Gösterilmesi

Pilot Raporları, uçuş ekibi tarafından girilen olay ve arızalarının tutulduğu raporlardır. Pilot raporlarında etkili parametrelerin belirlenmesinde bağlantı grafikleri kullanılmıştır. Analizlerde pilot raporları “3” rakamı ile temsil edilmektedir. Şekil 7.3, 7.5 ve 7.6’dan görüldüğü gibi pilot raporları en çok “9” ve “12” kodlu, periyodik bakım ve transit kontrol uçuş fazı aşamaları; JLB ve JDK kodlu uçaklar ve “25” numaralı ata kodu ile temsil edilen malzeme, mefruşat ve donatım ile ilgili olarak raporlanmıştır. Bunlarla ilgili olarak elde edilen korelasyon katsayıları tablo 7.18’de verilmiştir.

Tablo 7.17. Uçuş Fazı Açıklamaları.

Uçuş Fazı	Açıklama
MS (8)	Motor Başlangıcı
CL (2)	Tırmanma
AP (1)	Yaklaşma
TA (10)	Taksi
CR (3)	Gezinti
FA (5)	Son yaklaşma
TK (11)	Kalkış
DE (4)	İniş süresince
LA (7)	İniş
TR (12)	Transit Kontrol
KN (6)	Konaklama
PB (9)	Periyodik Bakım

Tablo 7.18. Pilot raporları ile ilgili olarak elde edilen etkili parametreler ve düzeylerinin korelasyon katsayıları.

Parametre ve Düzeyi	Korelasyon Katsayısı
Uçuş Fazı-9	-0.830
Uçuş Fazı-12	0.662
Uçak Kodu-JLB	0.568
Uçak Kodu-JDK	-0.692
Ata Kodu-25	-0.999

7.2.2.2. Kabin Raporlarının Tutulmasında Etkili Olan Parametrelerin Gösterilmesi

Kabin raporları, uçağın kabin görevlisinin yaptığı kayıtları içermektedir. Analizimizde kabin raporları “1” rakamı ile temsil edilmektedir. Şekil 7.5 ve şekil 7.6’da görüldüğü gibi kabin raporları en çok JLJ kodlu uçakla ve “25” numaralı ata kodu ile temsil edilen malzeme, mefruşat ve donatım ile ilgili olarak raporlanmıştır. Bunlarla ilgili olarak elde edilen korelasyon katsayıları tablo 7.19’da verilmiştir.

Tablo 7.19. Kabin raporları ile ilgili olarak elde edilen etkili parametreler ve düzeylerinin korelasyon katsayıları.

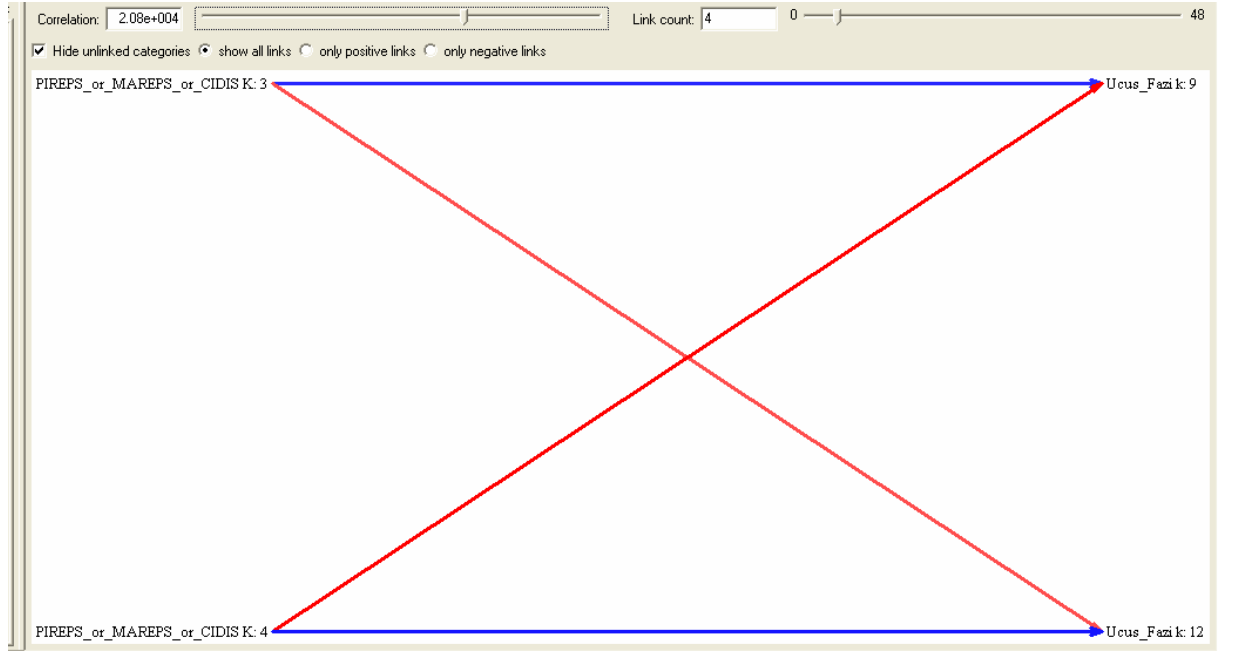
Parametre ve Düzeyi	Korelasyon Katsayısı
Uçak Kodu-JLJ	-0.632
Ata Kodu-25	0.849

7.2.2.3. Bakım Raporlarının Tutulmasında Etkili Olan Parametrelerin Gösterilmesi

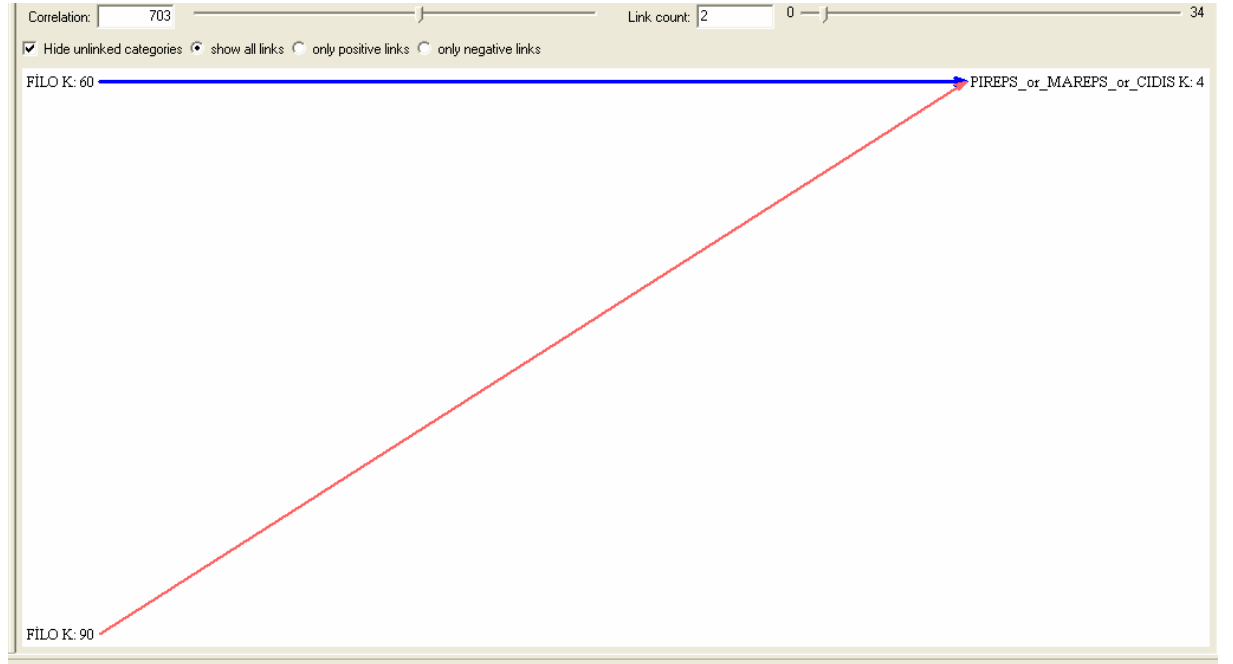
Bakım Raporları, uçak bakım yetkili teknisyenleri tarafından uçağın kaydedilen olay ve arızalarının kaydedildiği raporlardır. Analizlerde bakım raporları “4” rakamı ile temsil edilmektedir. Şekil 7.3, 7.4 ve 7.5’den görüldüğü gibi pilot raporları en çok “9” ve “12” kodlu, periyodik bakım ve transit kontrol uçuş fazı aşamaları; “60” ve “90” kodlu filolar ve JLB ve JDK kodlu uçaklar ile ilgili olarak raporlanmıştır. Bunlarla ilgili olarak elde edilen korelasyon katsayıları tablo 7.20’de verilmiştir.

Tablo 7.20. Bakım raporları ile ilgili olarak elde edilen etkili parametreler ve düzeylerinin korelasyon katsayıları.

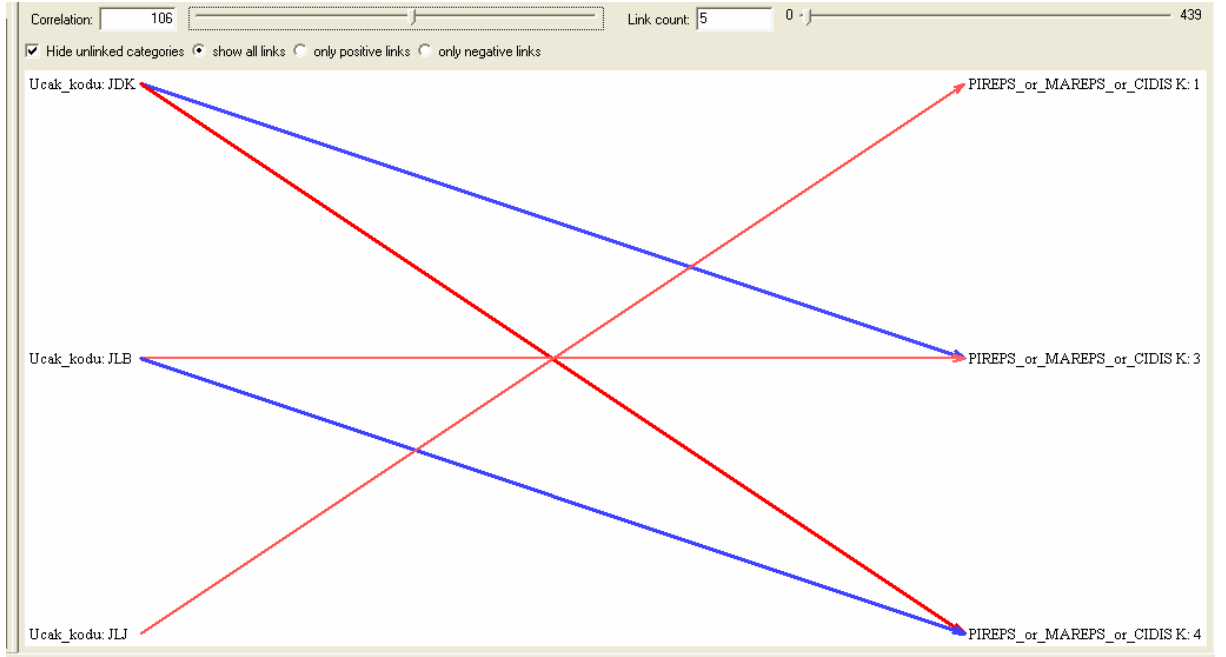
Parametre ve Düzeyi	Korelasyon Katsayısı
Uçuş Fazı-9	0.999
Uçuş Fazı-12	-0.909
Uçak Kodu-JLB	-0.723
Uçak Kodu-JDK	1
Filo-60	-0.996
Filo-90	0.561



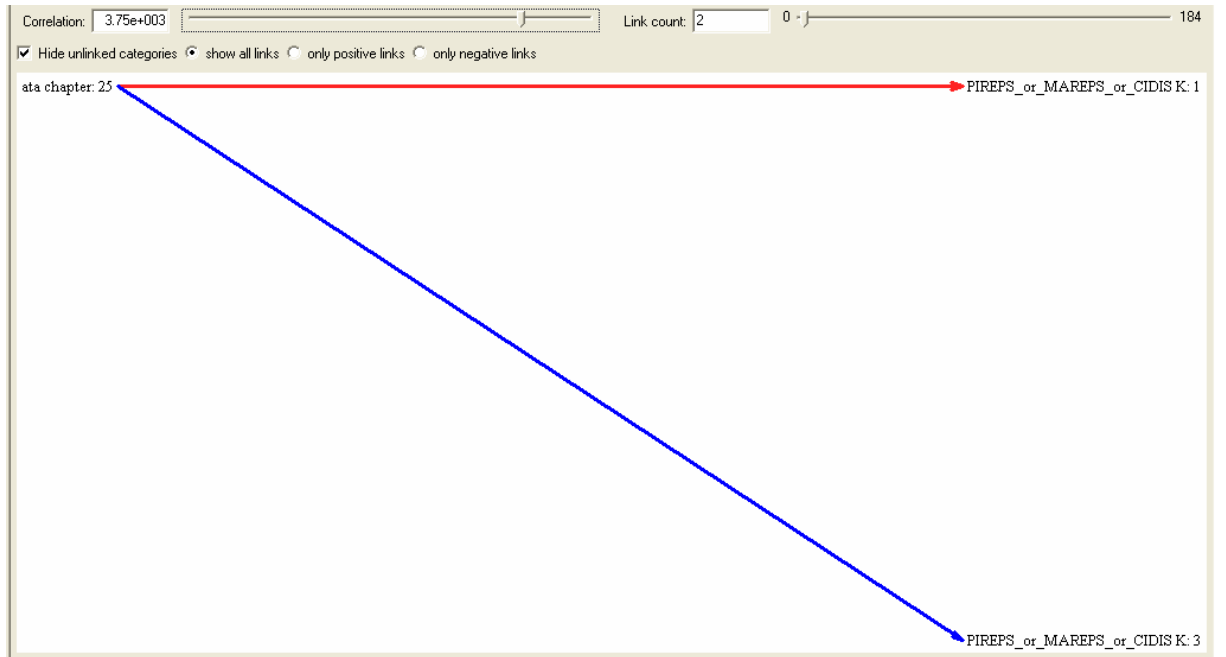
Şekil 7.3. PMC raporları ile uçuş fazı parametresi arasındaki ilişki.



Şekil 7.4. PMC raporları ile filo parametresi arasındaki ilişki.



Şekil 7.5. PMC raporları ile uçak kodu parametresi arasındaki ilişki.



Şekil 7.6. PMC raporları ile ata kodu parametresi arasındaki ilişki.

7.2.2.4. Sonuçlar

Yapılan analizlerden görüldüğü gibi işletmede, uçuş esnasında oluşan problem veya arıza olarak çok önemli sorunlar çıkmadığı sadece malzeme ve tefrişattan kaynaklanan sorunların olduğu görülmüştür. Uçuşlarda kullanılan 112 farklı uçak tipinden sadece 3

tanesi, filolardan ise sadece 2 tanesi ile ilgili sorunlar olduğu görülmektedir. Problemlerin olduğu uçuş aşamaları ise uçakların periyodik bakımı veya transit geçitlerde yapılan kontroller esnasında görülerek kaydedildiği de yine analizler sonucunda gözlenmiştir. Bu sonuçlara göre işletmede uçuş esnasında çok önemli arızalar veya problemlerle karşılaşmamaktadır. Yani güvenli uçuşlar gerçekleştiriliyor diyebiliriz.

7.2.3. Parça Raporlarının Analizi

Parça raporlarında, söküm raporlarından daha ayrıntılı olarak sökümün planlı mı yoksa plansız mı olduğu, sökümün arızadan kaynaklanıp kaynaklanmadığı, parçanın uçaktaki miktarı gibi aşağıda belirtilen parametrelerle birlikte daha ayrıntılı kayıtlar içerdiği tespit edilmiştir. Parça raporlarında yer alan 31 parametre uzmanlarla birlikte yapılan değerlendirme sonucunda gereksiz ve eksik bilgilerin çıkarılmasıyla 19 parametreye indirgenmiştir. Tablo 7.21’de özellikleri gösterilen bu parametreler ve tanımları şöyledir:

- Ata Kodu: Sırasıyla uçağa ait sistem (Kodu/system), alt sistem (section/subsystem), parça (subject/unit), sayfa bloğu (page block)’nu tanımlar.
- Parça Kodu: Hava yolu şirketinin her bir parçaya verdiği koddur.
- Filo: Parçanın hangi filoya ait olduğunu göstermektedir.
- Tanım : Parçanın adını açıklamaktadır.
- Qpa (Q): Parçadan bir uçakta kaç tane olduğunu göstermektedir.
- H (H): Parçanın kaç saat uçtuğunu belirtir.
- Toplam Söküm (Fs Total-Fst): Son 12 ayda belirtilen Parçanın kaç kere söküldüğünü gösterir.
- Söküm1 (Fs1): Son 12 aydaki sökümlerden kaç tanesinin planlı olarak yapıldığını belirtir.
- Söküm2 (Fs2): Son 12 aydaki sökümlerden kaç tanesinin plansız olarak yapıldığını belirtir.
- Diğer Sökümler (Fs Other-Fso): Son 12 aydaki sökümün diğer söküm sebepleridir.
- Arıza (Fail-F): Son 12 aydaki sökümün kaç tanesinin arızalı olduğunu belirtir.
- Faal (Not Fail-Nf): Son 12 aydaki plansız sökümün kaç tanesinin faal olduğunu gösterir.
- Arıza1 (Fail1-F1): Son 12 aydaki planlı sökümün kaç tanesinin arızalı olduğunu gösterir.
- Arıza2 (Fail2-F2): Son 12 aydaki plansız sökümün kaç tanesinin arızalı olduğunu gösterir.

- Mtbr: Sökümler arasındaki ortalama süre.
- Mtbur: Plansız sökümler arasındaki ortalama süre.
- Mtbf: Arızalar arasındaki ortalama süre.
- Mtbf-3: Son üç aylık arızalar arasındaki ortalama süre.
- FF (Falert): İkaz değeri.

Tablo 7.21. Parça raporlarına ait parametrelerin özellikleri.

Parametre Adı	Özellikleri
Ata Kodu	Kategorik bir veridir.
Parça Kodu	Dizi tipinde alınmıştır.
Filo	Kategorik bir veridir.
Tanım	Dizi tipinde alınmıştır.
Q	Sayısal nitelikli veridir.
H	Sayısal nitelikli veridir.
Fst	Sayısal nitelikli veridir.
Fs1	Sayısal nitelikli veridir.
Fs2	Sayısal nitelikli veridir.
Fso	Sayısal nitelikli veridir.
F	Sayısal nitelikli veridir.
Nf	Sayısal nitelikli veridir.
F1	Sayısal nitelikli veridir.
F2	Sayısal nitelikli veridir.
Mtbr	Sayısal nitelikli veridir.
Mtbur	Sayısal nitelikli veridir.
Mtbf	Sayısal nitelikli veridir.
Mtbf-3	Sayısal nitelikli veridir.
FF-Falart	Sayısal nitelikli veridir.

İşletmeye ait Parça raporları, 2005 yılına ait 12 aylık kayıtlar, 2006 yılı aralık ayına ait kayıtlar, 2005 yılı toplam kayıtlar ve iki yıllık aralık ayı kayıtları olmak üzere 15 adet veri kümesi şeklinde analiz edilmiştir. Analiz sürecinde ikaz değerinin, sökümler arasındaki ortalama sürenin (mtbf), plansız sökümler arasındaki ortalama sürenin (mtbur) ve arızalar arasındaki ortalama sürenin (mtbf) tahminine yönelik veri

madenciliği fonksiyonları kullanılmıştır.

Her bir veri kümesine ait özellikler tablo 7.22’da gösterilmiştir.

Tablo 7.22. Parça raporlarında yer alan veri kümelerinin özellikleri.

Veri Kümesi	Ait Olduğu Dönem	Dizi Değişken Sayısı	Kategorik Değişken Sayısı	Sürekli Değişken Sayısı	Kayıp Değer	Kayıt Sayısı
1	Ocak	2	2	15	Evet	4272
2	Şubat					4249
3	Mart					4276
4	Nisan					4439
5	Mayıs					4438
6	Haziran					4526
7	Temmuz					4398
8	Ağustos					4429
9	Eylül					4381
10	Ekim					4367
11	Kasım					4396
12	Aralık					4460
13	Bir yıllık toplam kayıtlar	2	2	15	Evet	52977
14	Önceki yıla ait Aralık					5192
15	İki yıllık Aralık kayıtları					9652

7.2.3.1. Parça Raporlarına Uygulanan Ön İşlemler

Yapılan ilk işlem raporlarda yer alan gereksiz ve boş olan 10 parametrenin çıkarılması

olmuştur. Daha sonra sırasıyla ata kodu parametresi iki haneli olarak tekrar kodlanmış ve filo parametresi kategorik veri olması için yeniden düzenlenerek, toplam söküm parametresinin, planlı, plansız ve diğer sökümünün toplamından oluşmasından dolayı toplam söküm parametresinin tekrar kullanılması uygun olmadığından veri kümesinden çıkarılmıştır. Benzer şekilde arızalı parça sayısının planlı ve plansız arızaların toplamından oluşmasından dolayı arıza parametresi de veri kümesinden çıkarılmıştır. Bu işlemlerden sonra yapılan ön işleme süreci iki aşamada gerçekleştirilmiştir. Birinci aşamada PolyAnalyst-Bağımlılıkları Bulma, Clementine-Özellik Seçimi ve Rosetta-Kaba Küme araçları kullanılarak özellik seçimi gerçekleştirilmiştir. Özellik seçiminin ardından ikinci aşamada her bir veri kümesine Clementine-aykırılıkların tespiti uygulanarak gürültülü kayıtlar çıkarılmıştır.

7.2.3.1.1. Bağımlılıkları Bulma Analizi ile Özellik Seçimi

Ön işleme sürecinin ilk adımı, ikaz değerine etkisi olan parametrelerin bağımlılıkları bulma analizi ile tespit edilmesidir. Her bir veri kümesi için tekrarlanan işlemler sonucunda elde edilen P değerlerine göre aralarında ilişki olup olmadığına karar verilmiştir. İlgili P değerleri $[0,1]$ aralığında bir değer almaktadır. Bu değer 0'a yaklaştıkça ilişkinin kuvveti artmaktadır. Analiz sonucunda elde edilen P değerleri her bir veri kümesi için tablo 7.23 (a) ve tablo 7.23 (b)'de gösterilmiştir.

7.2.3.1.2. Clementine ile Özellik Seçme Analizi

Ön işleme sürecinin ikinci adımında bağımlılıkları bulma analizine alternatif olarak özellik seçme analizi ile yine ikaz değerine etki eden parametrelerin tespit edilmesi amaçlanmıştır. Özellik seçme analizi ile belirlenen etkin parametreler yine her bir veri kümesi için tablo 7.24'de sırasıyla gösterilmiştir.

Tablo 7.23. (a) Bağımlılık bulma analizi sonucunda elde edilen ilgili P değerleri.

Veri Kümesi	Etkili Parametreler	P Değerleri
1	{Mtb _f , F ₂ , Mtb _{ur} , Q, Fs ₂ , Nf, Mtb _r , H, Fs ₁ }	{2.062 e ⁻⁷⁰ , 1.551 e ⁻³⁹ , 1.551 e ⁻³⁹ , 1.005 e ⁻²⁶ , 1.005 e ⁻²⁶ , 1.088 e ⁻¹⁹ , 1.088 e ⁻¹⁹ , 1.577 e ⁻¹³ , 1.577 e ⁻¹³ }
2	{F ₁ , Mtb _f , Q, H, F ₂ , Mtb _{ur} , Fs ₁ , Fso, Mtb _r , Mtb _f -3 }	{3.28 e ⁻¹⁵ , 3.28 e ⁻¹⁵ , 1.528 e ⁻¹⁴ , 1.528 e ⁻¹⁴ , 1.528 e ⁻¹⁴ , 1.528 e ⁻¹⁴ , 1.131 e ⁻⁹ , 1.131 e ⁻⁹ , 1.131 e ⁻⁹ , 1.131 e ⁻⁹ }
3	{Fs ₂ , F ₂ , Mtb _r , Mtb _{ur} , Fs ₁ , Mtb _f , Q, H, Fso, Nf, Mtb _f -3}	{1.345 e ⁻⁵² , 1.345 e ⁻⁵² , 1.345 e ⁻⁵² , 1.345 e ⁻⁵² , 1.119 e ⁻³⁶ , 1.119 e ⁻³⁶ , 5.909 e ⁻¹³ , 5.909 e ⁻¹³ , 7.907 e ⁻⁸ , 7.907 e ⁻⁸ }
4	{Fs ₁ , Fs ₂ , F ₂ , F ₁ , Q, H, Mtb _{ur} , Mtb _f , Nf, Mtb _r , Mtb _f -3}	{2.992 e ⁻⁴⁴ , 2.992 e ⁻⁴⁴ , 2.992 e ⁻⁴⁴ , 2.992 e ⁻⁴⁴ , 2.106 e ⁻³⁸ , 2.106 e ⁻³⁸ , 2.106 e ⁻³⁸ , 2.106 e ⁻³⁸ , 2.306 e ⁻¹⁷ , 2.306 e ⁻¹⁷ , 2.306 e ⁻¹⁷ }
5	{F ₁ , Mtb _{ur} , H, Mtb _f , Fs ₂ , F ₂ , Nf, Mtb _{ur} , Mtb _f -3, Q, Fs ₁ }	{1.634 e ⁻⁵² , 1.634 e ⁻⁵² , 2.749 e ⁻⁴³ , 2.749 e ⁻⁴³ , 2.557 e ⁻³³ , 2.557 e ⁻³³ , 1.018 e ⁻²⁵ , 1.018 e ⁻²⁵ , 1.033 e ⁻¹⁹ , 5.601 e ⁻⁹ , 0.0005073}
6	{Fs ₁ , Mtb _{ur} , Q, H, Fs ₂ , Mtb _f , Nf, F ₂ , F ₁ , Mtb _r , Mtb _f -3}	{7.004 e ⁻⁴³ , 7.004 e ⁻⁴³ , 5.776 e ⁻³⁹ , 5.776 e ⁻³⁹ , 5.776 e ⁻³⁹ , 5.776 e ⁻³⁹ , 2.821 e ⁻²⁷ , 2.821 e ⁻²⁷ , 2.821 e ⁻²⁷ , 2.821 e ⁻²⁷ , 1.033 e ⁻¹⁹ }
7	{H, F ₂ , Mtb _{ur} , Mtb _f , Fs ₂ , Fso, F ₁ , Mtb _r , Fs ₁ , Mtb _f -3}	{1.647 e ⁻⁴⁸ , 1.647 e ⁻⁴⁸ , 1.647 e ⁻⁴⁸ , 1.647 e ⁻⁴⁸ , 1.721 e ⁻³¹ , 1.721 e ⁻³¹ , 1.721 e ⁻³¹ , 1.721 e ⁻³¹ , 1.172 e ⁻¹⁰ , 1.172 e ⁻¹⁰ }
8	{H, Fs ₂ , F ₂ , Mtb _{ur} , Fs ₁ , F ₁ , Mtb _f , Nf, Mtb _r , Mtb _f -3}	{8.712 e ⁻⁴⁶ , 8.712 e ⁻⁴⁶ , 8.712 e ⁻⁴⁶ , 8.712 e ⁻⁴⁶ , 2.289 e ⁻³⁶ , 2.289 e ⁻³⁶ , 2.289 e ⁻³⁶ , 3.539 e ⁻¹⁷ , 3.539 e ⁻¹⁷ , 3.539 e ⁻¹⁷ }
9	{F ₂ , Mtb _r , Mtb _{ur} , Mtb _f , H, Fs ₁ , Fs ₂ , F ₁ , Fso, Nf, Mtb _f -3}	{7.567 e ⁻⁴⁶ , 7.567 e ⁻⁴⁶ , 7.567 e ⁻⁴⁶ , 7.567 e ⁻⁴⁶ , 2.722 e ⁻⁴² , 2.722 e ⁻⁴² , 2.722 e ⁻⁴² , 2.722 e ⁻⁴² , 6.763 e ⁻⁹ , 6.763 e ⁻⁹ , 6.763 e ⁻⁹ }
10	{F ₂ , Mtb _r , Mtb _f , Mtb _{ur} , Q, H, Fs ₁ , Fs ₂ , Fso, Nf, F ₁ , Mtb _f -3 }	{2.857 e ⁻⁴⁷ , 2.857 e ⁻⁴⁷ , 2.857 e ⁻⁴⁷ , 2.857 e ⁻⁴⁷ , 2.285 e ⁻³¹ , 2.285 e ⁻³¹ , 2.285 e ⁻³¹ , 2.285 e ⁻³¹ , 1.063 e ⁻⁸ , 1.063 e ⁻⁸ , 1.063 e ⁻⁸ }
11	{Fs ₂ , Nf, Mtb _r , Mtb _f , Fs ₁ , F ₂ , Mtb _{ur} , H, Mtb _f -3}	{5.334 e ⁻⁴⁴ , 5.334 e ⁻⁴⁴ , 5.334 e ⁻⁴⁴ , 5.334 e ⁻⁴⁴ , 7.332 e ⁻⁴³ , 7.332 e ⁻⁴³ , 7.332 e ⁻⁴³ , 9.702 e ⁻¹⁵ , 9.702 e ⁻¹⁵ }

Tablo 7.23. (devam) (b) Bağımlılık bulma analizi sonucunda elde edilen ilgili P değerleri.

12	{Fs1, Fs2, Mtbur, F2, Mtbr, Mtbf, Q, H, F1, Nf, Mtbf}	{ $9.722 \cdot 10^{-45}$, $9.722 \cdot 10^{-45}$, $9.722 \cdot 10^{-45}$, $1.06 \cdot 10^{43}$, $1.06 \cdot 10^{43}$, $1.06 \cdot 10^{43}$, $8.873 \cdot 10^{-41}$, $8.873 \cdot 10^{-41}$, $8.873 \cdot 10^{-41}$, $4.336 \cdot 10^{-10}$, $4.336 \cdot 10^{-10}$ }
13	{Q, Mtbur, Fs1, Mtbf, Fs2, Fso, Nf, F2, Mtbr, H}	{ $1.075 \cdot 10^{-113}$, $1.075 \cdot 10^{-113}$, $6.512 \cdot 10^{-112}$, $6.512 \cdot 10^{-112}$, $9.484 \cdot 10^{-88}$, $9.484 \cdot 10^{-88}$, $3.179 \cdot 10^{-47}$, $3.179 \cdot 10^{-47}$, $3.179 \cdot 10^{-47}$, $1.844 \cdot 10^{-28}$ }
14	{Fs1, Mtbur, Q, Fs2, F2, Mtbr, F1, Mtbf, H, Fso, Nf, Mtbf-3}	{ $5.89 \cdot 10^{-39}$, $5.89 \cdot 10^{-39}$, $2.253 \cdot 10^{-34}$, $2.253 \cdot 10^{-34}$, $2.253 \cdot 10^{-34}$, $2.253 \cdot 10^{-34}$, $1.238 \cdot 10^{-26}$, $1.238 \cdot 10^{-26}$, $1.236 \cdot 10^{-13}$, $1.236 \cdot 10^{-13}$, $1.236 \cdot 10^{-13}$ }
15	{Fs2, F2, Mtbf, Nf, Mtbur, Q, Fs1, Mtbr, H, Fso, F1, Mtbf-3}	{ $8.419 \cdot 10^{-77}$, $8.419 \cdot 10^{-77}$, $8.419 \cdot 10^{-77}$, $2.067 \cdot 10^{-68}$, $2.067 \cdot 10^{-68}$, $1.028 \cdot 10^{-27}$, $1.028 \cdot 10^{-27}$, $1.028 \cdot 10^{-27}$, $1.579 \cdot 10^{-14}$, $1.579 \cdot 10^{-14}$, $1.579 \cdot 10^{-14}$, $1.579 \cdot 10^{-14}$ }

Tablo 7.24. Özellik seçme analizi ile tespit edilen etkin parametreler.

Veri Kümesi	Etkili Parametreler
1	{Mtbf, Mtbf-3, Mtbur, Mtbr, H, Fs1, Q}
2	{Mtbr, Mtbur, Mtbf, Q}
3	{Mtbf, Mtbf-3, Mtbur, Mtbr, Fs1, F1}
4	{Mtbf, Mtbur, Mtbf-3, H, Mtbr, Fs1, F1}
5	{Mtbf, Mtbf-3, Mtbur, Mtbr, H, Fs1, Q}
6	{Mtbf, Mtbf-3, Mtbur, Mtbr}
7	{Mtbf-3, Mtbf, Mtbur, Mtbr, H}
8	{Mtbf-3, Mtbf, Mtbur, Mtbr, H}
9	{Mtbf, Mtbf-3, Mtbur, Mtbr, H}
10	{Mtbf, Mtbur, H, Mtbr, Mtbf-3, Fs1, F1}
11	{Mtbf, Mtbur, H, Mtbr, Mtbf-3, F1, Fs1}
12	{Mtbf, Mtbr, Mtbur, H, F2}
13	{Mtbf, Mtbur, Mtbf-3, Mtbr, H, Fs1, Q, F1}
14	{Mtbf, Mtbur, H, Mtbf-3, Mtbr, F1, Fs1}
15	{Mtbur, Mtbf, Mtbf-3, H, Mtbr, , F1, Fs1, Q}

7.2.3.1.3. Kaba Kümeler ile Özellik Seçimi

İkaz değeri üzerinde etkisi olan parametrelerin tespit edilmesine yönelik yapılan analizlerden üçüncü ve sonuncusu olan kaba kümeler yardımı ile özellik seçimi sonucunda elde edilen indirgenmiş veri kümeleri, her bir veri kümesi için sırasıyla tablo 7.25-7.39 arasında gösterilmiştir.

Tablo 7.25. Birinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fso, F2, Mtbr, Mtbur}
2	{Q, Fso, Mtbr, Mtbur, Mtbfb}
3	{Q, Fs1, F2, Mtbr, Mtbur}
4	{Q, Fs1, Mtbr, Mtbur, Mtbfb}
5	{Q, Fs1, Fso, F2, Mtbr}
6	{Q, Fs1, Fso, Mtbr, Mtbfb}
7	{Q, Fs1, Fs2, F2, Mtbr}
8	{Q, Fs2, Fso, Mtbr, Mtbfb}
9	{Q, Fs1, Fs2, Mtbr, Mtbfb}
10	{Q, Fs2, Fso, F2, Mtbr}
11	{Q, H, Fs1, Fso, Mtbur, Mtbfb}
12	{Q, H, Fs1, Fs2, Fso, Nf, Mtbfb}
13	{Q, H, Fs1, Fs2, Fso, F2, Mtbur}

Tablo 7.26. İkinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Fs2, Fso, Mtbr, Mtbfr, Mtbfr-3}
2	{Fs2, Fso, F2, Mtbr, Mtbfr-3}
3	{Fso, F2, Mtbr, Mtbfr, Mtbfr-3}
4	{Fs1, Fso, F2, Mtbr, Mtbfr-3}
5	{Q, Fs1, Fs2, Mtbr, Mtbfr, Mtbfr-3}
6	{Q, Fs1, Fso, Mtbr, Mtbfr, Mtbfr-3}
7	{Q, Fs1, Mtbr, Mtbfr, Mtbfr-3}
8	{Q, Fso, Nf, Mtbr, Mtbfr, Mtbfr-3}
9	{Q, Fs1, F2, Mtbr, Mtbfr, Mtbfr-3}
10	{Q, Fs1, Fs2, F2, Mtbr, Mtbfr-3}
11	{H, Fso, Nf, F2, Mtbr, Mtbfr-3}
12	{Q, Fso, Nf, F2, Mtbr, Mtbfr-3}
13	{Q, Fso, Mtbr, Mtbfr, Mtbfr-3}
14	{Fso, Nf, F2, Mtbr, Mtbfr, Mtbfr-3}
15	{Q, H, Fs1, Fs2, Fso, Mtbfr, Mtbfr-3}
16	{Q, H, Fs1, Fs2, Fso, F2, Mtbfr-3}
17	{Q, H, Fs1, Fso, F2, Mtbfr, Mtbfr-3}
18	{Q, H, Fs1, Fso, Nf, Mtbfr, Mtbfr-3}

Tablo 7.27. Üçüncü veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fs2, Fso, Nf, F1, Mtbr, Mtbfr, Mtbfr-3}
2	{Q, Fso, Nf, F1, Mtbr, Mtbfr, Mtbfr-3}
3	{Q, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbfr-3}
4	{Fs1, Fs2, Fso, Nf, F2, F1, Mtbr, Mtbfr-3}
5	{Q, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtbfr-3}
6	{Q, Fs1, Nf, F2, F1, Mtbr, Mtbfr, Mtbfr-3}
7	{Q, Fs1, Nf, F1, Mtbr, Mtbfr, Mtbfr-3}
8	{Q, Fs1, Fso, Nf, F1, Mtbr, Mtbfr, Mtbfr-3}
9	{Q, Fs1, Fs2, Nf, F1, Mtbr, Mtbfr, Mtbfr-3}
10	{Q, Fso, Nf, F2, F1, Mtbr, Mtbfr, Mtbfr-3}
11	{Q, FS2, Fso, Nf, F2, F1, Mtbr, Mtbfr-3}
12	{Fs1, Fs2, Fso, Nf, F1, Mtbr, Mtbfr, Mtbfr-3}

Tablo 7.28. Dördüncü veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtbf-3}
2	{Q, Fs1, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
3	{Q, Fs2, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
4	{Q, Fso, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
5	{Q, Fs1, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
6	{Q, Fs1, Fs2, Nf, F1, Mtbr, Mtbf, Mtbf-3}
7	{Q, Fs2, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
8	{Q, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
9	{Q, Fso, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
10	{Q, Fs1, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}

Tablo 7.29. Beşinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fs1, Nf, Mtbr, Mtbur, Mtbf-3}
2	{Q, Fs2, Fso, Nf, Mtbr, Mtbf-3}
3	{Q, Fso, Nf, Mtbr, Mtbur, Mtbf-3}
4	{Q, Fs1, Fs2, Nf, Mtbr, Mtbf-3}
5	{Q, Fs1, Fso, Nf, Mtbr, Mtbf-3}

Tablo 7.30. Altıncı veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fso, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
2	{Q, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
3	{Q, Fs2, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
4	{Q, Fso, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
5	{Q, Fs1, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
6	{Q, H, Fs1, Fso, Nf, F2, F1, Mtbur, Mtbf-3}
7	{Q, H, Fs1, Fso, Nf, F1, Mtbur, Mtbf, Mtbf-3}
8	{Q, H, Fs1, Fs2, Fso, Nf, F1, Mtbf, Mtbf-3}
9	{Q, H, Fs2, Fso, Nf, F2, F1, Mtbr, Mtbf-3}

Tablo 7.31. Yedinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fso, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
2	{Q, Fs1, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
3	{Q, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
4	{Q, Fs2, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
5	{Q, Fs1, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
6	{Q, Fs1, Fs2, Nf, F1, Mtbr, Mtbf, Mtbf-3}
7	{Q, Fso, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
8	{Q, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtbf-3}
9	{Q, Fs1, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
10	{Q, Fs2, Fso, Nf, F2, F1, Mtbr, Mtbf-3}

Tablo 7.32. Sekizinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fs1, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
2	{Q, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
3	{Q, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtbf-3}
4	{Q, Fs2, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
5	{Q, Fso, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
6	{Q, Fs1, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
7	{Q, Fs2, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
8	{Q, Fs1, Fs2, Nf, F1, Mtbr, Mtbf, Mtbf-3}
9	{Q, Fs1, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
10	{Q, Fso, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}

Tablo 7.33. Dokuzuncu veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
2	{Q, Fs1, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
3	{Q, Fso, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
4	{Q, Fs2, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
5	{Q, Fso, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
6	{Q, Fs2, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
7	{Q, H, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtbf-3}
8	{Q, H, Fs1, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
9	{Q, H, Fs1, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
10	{Q, H, Fs1, Fs2, Nf, F1, Mtbr, Mtbf, Mtbf-3}

Tablo 7.34. Onuncu veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, H, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
2	{Q, H, Fs1, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
3	{Q, H, Fso, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
4	{Q, H, Fso, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
5	{Q, H, Fs1, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
6	{Q, H, Fs1, Fs2, Nf, F1, Mtbr, Mtbf, Mtbf-3}
7	{Q, H, Fs2, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
8	{Q, H, Fs2, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
9	{Q, H, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtbf-3}
10	{Q, H, Fs1, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}

Tablo 7.35. On birinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, H, Fso, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
2	{Q, H, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtbf-3}
3	{Q, H, Fs1, Fs2, Nf, F1, Mtbr, Mtbf, Mtbf-3}
4	{Q, H, Fs2, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
5	{Q, H, Fs1, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
6	{Q, H, Fs1, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
7	{Q, H, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
8	{Q, H, Fso, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
9	{Q, H, Fs2, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
10	{Q, H, Fs1, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}

Tablo 7.36. On ikinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fso, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
2	{Q, Fs1, Fs2, F2, F1, Mtbr, Mtbf, Mtbf-3}
3	{Q, Fs, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
4	{Q, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtb-3}
5	{Q, Fs1, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
6	{Q, Fs1, Fso, F2, F1, Mtbr, Mtbf, Mtbf-3}
7	{Q, Fso, F2, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
8	{Q, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
9	{Q, Fs1, F2, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
10	{Q, Fs2, Fso, F2, F1, Mtbr, Mtbf, Mtbf-3}

Tablo 7.37. On üçüncü veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, H, Fso, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
2	{Q, H, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
3	{Q, H, Fs1, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
4	{Q, H, Fs2, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
5	{Q, H, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtbf-3}

Tablo 7.38.On dördüncü veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fs1, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
2	{Q,Fs2, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
3	{ Q, Fs2, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
4	{ Q, Fs1, Fs2, Nf, F1, Mtbr, Mtbf, Mtbf-3}
5	{ Q, Fs1, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
6	{ Q, Fso, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
7	{ Q, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtbf-3}
8	{ Q, Fso, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}
9	{ Q, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
10	{ Q, Fs1, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}

Tablo 7.39. On beşinci veri kümesi ile elde edilen indirgenmiş kaba kümeler.

Kaba Küme-Rs Numarası	İndirgenmiş Veri Kümesi
1	{Q, Fs1, Fso, Nf, F2, F1, Mtbr, Mtbf-3}
2	{Q, Fs1, Fs2, Nf, F2, F1, Mtbr, Mtbf-3}
3	{Q, Fs1, Fso, Nf, F1, Mtbr, Mtbf, Mtbf-3}
4	{Q, Fs1, Fs2, Nf, F1, Mtbr, Mtbf, Mtbf-3}
5	{Q, Fs1, Nf, F1, Mtbr, Mtbur, Mtbf, Mtbf-3}
6	{Q, Fs1, Nf, F2, F1, Mtbr, Mtbur, Mtbf-3}

7.2.3.1.4. Aykırılıkların Tespiti

Aykırılıkların tespiti modeli, aykırı değerler ve olağan dışı farklı olayları tanımlamak için kullanılmaktadır [141]. Bu çalışmada da, aykırılıkların tespiti modeli, bahsi geçen 15 adet veri kümesinde yer alan anormalliklerin-aykırı kayıtların tespit edilerek veri madenciliği sürecinden önce veri kümesinden çıkarılması amacıyla kullanılmıştır. Tablo 7.40’da her bir veri kümesinden çıkarılan aykırı kayıtların sayısı sırasıyla verilmiştir.

Tablo 7.40. Veri kümelerinden çıkarılan aykırı kayıtlar sayısı ile eğitim ve test veri kümelerindeki kayıt sayısı.

Veri Kümesi	Aykırı kayıt sayısı	Eğitim Kümesi Kayıt Sayısı	Test Kümesi Kayıt Sayısı
1	33	3180	1059
2	0	3187	1062
3	0	3207	1069
4	43	3297	1099
5	70	3276	1092
6	31	3147	1348
7	0	3299	1099
8	0	3322	1107
9	0	3286	1095
10	0	3276	1091
11	0	3297	1099
12	0	3345	1115
13	523	30000	22454
14	0	3635	1557
15	0	6757	2895

7.2.3.2. Kural Bulma Algoritması ile İkaz Değerinin Tahmini

Ön işleme süreci tamamlanan her bir veri kümesi kural bulma algoritması ile analiz edilerek ilgili döneme ait ikaz seviyelerinin tahmin edilmesine yönelik kurallar elde edilmiştir. Elde edilen kuralların güvenilirlik ve doğruluklarına analiz sonucunda elde edilen R^2 değerlerine göre karar verilmiştir. R^2 değeri 1'e ne kadar yakınsa kuralın doğruluğu ve güvenilirliği o kadar yüksek kabul edilmiştir. Hem eğitim hem de test sonunda elde edilen en yüksek R^2 değerlerine sahip kurallar işletme tarafından kullanılabilir olarak düşünülmüş ve önerilmiştir. Her bir veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen eğitim ve test R^2 değerleri ile elde edilen kurallar sırasıyla tablo 7.41-7.70 arasında gösterilmiştir.

Tablo 7.41. Birinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1	0.8757	0.253177
Rs2	0.9012	0.515817
Rs3	0.8784	0.299643
Rs4	0.9008	0.516292
Rs5	0.7917	0.530210
Rs6	0.9008	0.516292
Rs7	0.8185	0.464924
Rs8	0.9008	0.516292
Rs9	0.9008	0.516292
Rs10	0.7307	0.275958
Rs11	0.895	0.464243
Rs12	0.8969	0.586386
Rs13	0.8571	0.478644
Fd	0.9053	0.481894
Fs	0.9001	0.431387

Birinci veri kümesinin kural bulma algoritması ile analizi sonundaki en yüksek eğitim ve test R^2 değerlerine bağlı olarak, Rs2-4-6-8-9 indirgenmiş veri kümeleri ile özellik seçme analizi ile elde edilen veri kümelerinden elde edilen ve tablo 7.42’de gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.42.Birinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs2	$FF = (0.617227 * Q^2 * MTBF + 1.14886e^{-12} * Q^2 * MTBF^3 - 1.45284e^{-6} * Q^2 * MTBF^2 - 1.87477e^{-17} * MTBF^4 + 1580.04) / (Q^2 - 8.52188e^{-6} * Q * MTBF + 0.571867 * Q)$
Rs4-Rs6-Rs8-Rs9	$FF = (0.62589 * Q * MTBF + 1.16394e^{-12} * Q * MTBF^3 - 1.4148e^{-6} * Q * MTBF^2 - 2.44595e^{-12} * MTBF^3) / (Q - 8.48828e^{-6} * MTBF + 0.571867)$
Fs	$FF = (1.0534 * MTBR * Q * MTBF + 6.18635e^{-18} * MTBR * Q * MTBF^4 - 6.01097e^{-12} * MTBR * Q * MTBF^3 + 94512.3 * MTBF) / (MTBR * Q + 0.0977098 * MTBR * Q^2 + 172992)$

Tablo 7.43.İkinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1	0.7566	0.355814
Rs2	0.7418	0.466001
Rs3	0.7277	0.593886
Rs4	0.685	0.447113
Rs5	0.5608	0.506533
Rs6	0.6888	0.368099
Rs7	0.9414	0.577004
Rs8	0.7923	0.136450
Rs9	0.7094	0.426950
Rs10	0.6061	0.481575
Rs11	0.5415	0.464661
Rs12	0.6849	0.52786
Rs13	0.781	0.28887
Rs14	0.7059	0.495780
Rs15	0.633	0.50018
Rs16	0.6536	0.605595
Rs17	0.666	0.41925
Rs18	0.666	0.29457
Fd	0.7071	0.555277
Fs	0.5818	0.320069

İkinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R^2 değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs16 indirgenmiş veri kümesi ile bağımlılıkları bulma analizi ile elde edilen veri kümelerinden elde edilen ve tablo 7.44’de gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.44. İkinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs16	$FF = (-0.371851 * FSO * H * Q^2 + 3.87859e^{-6} * FSO * H^2 * Q^2 + 1.78169 * FSO * H * Q + 2101.67 * Q^2) / (FSO * Q * F2 - 2.6654 * FSO * Q + 17.3882 * FSO + 1.03702 * FSO^2)$
Fd	$FF = (0.0919908 * MTBF + 3.61687e^{-6} * MTBF^2 * F2^2 - 2.80311e^{-7} * MTBF^2 * Q + 3642.83 * Q - 1341.17 * Q^2) / (F2 * Q - 7.83467e^{-5} * F1 * MTBF - 0.0866749 * F1)$

Tablo 7.45. Üçüncü veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1	0.4405	0.358058
Rs2	0.4307	0.350007
Rs3	0.4405	0.358058
Rs4	0.4405	0.358058
Rs5	0.4405	0.358058
Rs6	0.5338	0.427255
Rs7	0.2453	0.213003
Rs8	0.4405	0.358058
Rs9	0.4405	0.358058
Rs10	0.5263	0.420519
Rs11	0.4405	0.358058
Rs12	0.4405	0.358058
Fd	0.5664	0.506212
Fs	0.3581	0.297831

Üçüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R^2 değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs6-10 indirgenmiş veri kümeleri ile bağımlılıkları bulma analizi ile elde edilen veri kümelerinden elde edilen ve tablo 7.46’da gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.46.Üçüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs6	$FF=(7.31643e^{-7} * MTBR^3 - 3.76749e^6 - 7.82386e^{-7} * F1 * MTBR^4) / (MTBR - 2707.64 * F1 * MTBR - 11287.8)$
Rs10	$FF=(7.31643e^{-7} * MTBR^3 - 3.76749e^6 - 0.0939585 * F1 * MTBR^3) / (MTBR - 2707.64 * F1 * MTBR - 11287.8)$
Fd	$FF=(2.01177e^{-6} * MTBUR^2 * MTBR^2 - 1.71993e^6 + 74.2212 * MTBR) / (MTBUR * MTBR + 0.00958202 * MTBR - 2270.18 + 0.00531958 * Q * MTBR + 11681.1 * Q^2 * MTBUR)$

Tablo 7.47.Dördüncü veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1	0.6646	0.365105
Rs2	0.3912	0.173857
Rs3	0.2458	0.171947
Rs4	0.2563	0.125733
Rs5	0.4047	0.203526
Rs6	0.4048	0.203740
Rs7	0.2443	0.116416
Rs8	0.2443	0.224331
Rs9	0.2416	0.125115
Rs10	0.4043	0.202811
Fd	0.4335	0.195247
Fs	0.4245	0.195290

Dördüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R^2 değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs1 indirgenmiş veri kümesi ile elde edilen ve tablo 7.48’de gösterilen kuralın daha güvenilir olduğu görülmüştür.

Tablo 7.48. Dördüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs1	$FF = (9.89786e^{-7} * MTBR * FS1 * MTBF3^3 + 7.15407e^{-7} * MTBR^3 * MTBF3 - 2121.82 * MTBR) / (MTBF3^2 - 2.12104e^6 * FS1 + 53.2268 * MTBR * FS1)$

Tablo 7.49. Beşinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1	0.7652	0.723692
Rs2	0.7522	0.676260
Rs3	0.7624	0.259613
Rs4	0.5862	0.531956
Rs5	0.779	0.171672
Fd	0.6832	0.617257
Fs	0.4182	0.348828

Beşinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R² değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs2 indirgenmiş veri kümeleri ile bağımlılıkları bulma analizi ile elde edilen veri kümelerinden elde edilen ve tablo 7.50’de gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.50. Beşinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs2	$FF = (-0.0183299 * FSO^6 * MTBR^2 - 11.405 * FSO^4 * MTBR^2 + 1.45125 * FSO^5 * MTBR^2 + 26.4848 * FSO^3 * MTBR^2 - 2.94895e^{11} * FSO - 5.58974e^6 * FSO^2 * MTBR + 2.98425e^{11} - 7163.59 * Q^2 * MTBR) / (FSO^5 * MTBR - 0.000145312 * FSO^7 * MTBR - 1.70741e^7 * FSO^2)$
Fd	$FF = (0.179029 * H * Q^2 + 1.93647e^{-11} * H^3 * Q^2 - 99.021 * Q^3) / (Q - 0.124929 * FS1 - 17.4953 * FS1^3 + 0.000134368 * FS1^2 * H - 0.36372)$

Tablo 7.51. Altıncı veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1	0.6892	0.634033
Rs2	0.5585	0.505334
Rs3	0.674	0.541465
Rs4	0.6062	0.468318
Rs5	0.5501	0.436167
Rs6	0.5726	0.476796
Rs7	0.5731	0.478486
Rs8	0.556	0.444641
Rs9	0.5538	0.444403
Fd	0.6877	0.53945
Fs	0.6148	0.142423

Altıncı veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R^2 değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs3-4 indirgenmiş veri kümeleri ile bağımlılıkları bulma analizi ile elde edilen veri kümelerinden elde edilen ve tablo 7.52’de gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.52. Altıncı veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs3	$FF = (3522.49 * Q^2 + 0.208272 * Q * MTBF - 13.2259 * MTBF - 270917 + 2.04969e^{-15} * MTBF^3 * MTBF^3 + 4.35308 * MTBF^3) / (Q^2 + 7.7232e^{-6} * Q^2 * MTBF - 35.8738 * Q + 222.92 * F1)$
Rs4	$FF = (2.05514e^{-5} * Q^2 * MTBF - 3 * MTBR * Q - 1.23799 * Q^2 * MTBF^3 * Q + 14.7463 * Q^5 - 6.18331e^{-11} * Q^2 * MTBF^3 * MTBR^2 * Q - 0.252923 * Q^2 * MTBR * Q + 3.12268e^6 * Q) / (Q^4 - 37.082 * Q^3 + 0.000424333 * Q^2 * MTBR + 1192.39 * Q - 9536.77 * F2)$
Fd	$FF = (1.40228e^{-6} * MTBR * MTBF * Q + 31197.7 * Q - 127220 + 560810 * F1 - 1.51884e^{-5} * MTBR^2 - 14.2632 * MTBF) / (Q^2 + 4.54859e^{-5} * MTBR * Q - 36.6653 * Q + 157.768 * F1)$

Tablo 7.53. Yedinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1	0.6107	0.543290
Rs2	0.4069	0.284862
Rs3	0.3353	0.237162
Rs4	0.1947	0.120061
Rs5	0.2	0.141911
Rs6	0.5651	0.368808
Rs7	0.6153	0.426167
Rs8	0.6409	0.579326
Rs9	0.5134	0.438498
Rs10	0.6093	0.436167
Fd	0.4733	0.425221
Fs	0.5201	0.155182

Yedinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R² değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs8 indirgenmiş veri kümesi ile elde edilen ve tablo 7.54’de gösterilen kuralın daha güvenilir olduğu görülmüştür.

Tablo 7.54. Yedinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs8	$FF = (4.34669e^{-7} * FS1 * MTBR^2 * MTBF3 * MTBR * F1 + 586.328 * FS1 * MTBR + 631339 * FS1 * MTBF3 * F1^2) / (FS1 * MTBR - 13696.7 * FS1 - 1282.34 * MTBF3 + 9.14969e^7 * F1^2)$

Tablo 7.55. Sekizinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1	0.6365	0.412947
Rs2	0.712	0.644326
Rs3	0.7623	0.332229
Rs4	0.4974	0.123161
Rs5	0.5443	0.363361
Rs6	0.4617	0.292790
Rs7	0.3723	0.298584
Rs8	0.5634	0.447209
Rs9	0.3734	0.327690
Rs10	0.4107	0.269727
Fd	0.6051	0.340111
Fs	0.4945	0.258056

Sekizinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R² değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs2 indirgenmiş veri kümesi ile bağımlılıkları bulma analizi ile elde edilen veri kümelerinden elde edilen ve tablo 7.56’da gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.56. Sekizinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs2	$FF = (1.06245e^{-11} * MTBR^2 * MTBF3^2 + 5758.86 * MTBF3 - 17.968 * FS1 * MTBR * MTBF3 - 625372 * FS1 * MTBR + 5.75216e^{10} * FS1 - 16.4371 * NF * MTBR^2) / (MTBF3 - 0.00331098 * NF * MTBR * MTBF3 + 4.0291e^7 * FS1 - 0.00427892 * FS1 * MTBR^2)$
Fd	$FF = (2.57441e^{-13} * MTBF3^2 * MTBUR^2 - 1.97757e^{-7} * MTBF3 * MTBUR^2 + 9486.22 * MTBF3 - 179.454 * FS1^2 * MTBF3 - 2.5299e^6) / (MTBF3 - 1.89967e^{-6} * MTBF3 * MTBUR - 2.07839e^{-5} * FS1 * MTBF3^2 + 23607.2 * FS1)$

Tablo 7.57. Dokuzuncu veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Rs2	0.5702	0.395884
Rs3	0.4092	0.220406
Rs4	0.1548	0.121642
Rs5	0.1541	0.119608
Rs6	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Rs7	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Rs8	0.5863	0.349288
Rs9	0.5857	0.363503
Rs10	0.5564	0.313641
Fd	0.4361	0.237734
Fs	0.4401	0.246994

Dokuzuncu veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R^2 değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs2-8-9 indirgenmiş veri kümeleri, bağımlılıkları bulma analizi ile elde edilen veri kümeleri ve özellik seçme analizi ile elde edilen ve tablo 7.58’de gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.58. Dokuzuncu veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs2	$FF=(1.88477e^{-5}*MTBF^2*FS1^3+0.00115183*F1*MTBF^2*FS1)/(MTBF*FS1^2-2.58694e^{-11}*MTBF^3*FS1-0.00576541*NF*MTBF*FS1^2-0.00655651*NF*MTBF+953.31*FS1)$
Rs8	$FF=(-2.55891*Q*MTBUR*H+1728.8*Q)/(Q-0.000785848*Q^3-8.79685*Q*MTBUR+107.565*F1*H+0.72685)$
Rs9	$FF=(-1.75125e^{-6}*MTBUR^2*H^2-0.0071005*MTBUR^3+545.457*MTBUR*H+2082.98*H-2.43002e^7+2.678e^{-6}*MTBUR*H^3-0.0185117*H^2)/(MTBUR*HOURS-0.627298*MTBUR^2-0.00827328*Q*H+10298.4)$
Fd	$FF=(2.04138e^{-7}*F1*MTBF*H-0.0454321*MTBUR-1.82817e^{-7}*MTBF*H+25.5549)/(F1-6.44792e^{-6}*MTBUR)$
Fs	$FF=(1.55527e^{-8}*MTBUR^2*MTBF^2*H+9527.88*MTBUR*MTBF^2+9.09072e^7*MTBUR*MTBF-12500*MTBUR^2*MTBF-1.93524e^{-20}*MTBUR^3*MTBF*H^3)/(MTBUR^2*MTBF-1.28673*MTBUR^3+4888.38*MTBUR^2+1.19144e^8*MTBF)$

Tablo 7.59. Onuncu veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1	0.4661	0.428785
Rs2	0.5622	0.520561
Rs3	0.4724	0.407417
Rs4	0.5632	0.510395
Rs5	0.6156	0.556952
Rs6	0.5274	0.478179
Rs7	0.461	0.419318
Rs8	0.461	0.419318
Rs9	0.461	0.419318
Rs10	0.4694	0.438095
Fd	0.5046	0.400690
Fs	0.5405	0.469565

Onuncu veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R^2 değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs5 indirgenmiş veri kümesi, bağımlılıkları bulma analizi ve özellik seçme analizi ile elde edilen veri kümelerinden elde edilen ve tablo 7.60'da gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.60. Onuncu veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs5	$FF = (103098 * Q^2 + 3.28536e^{-5} * H * MTBF3 * Q^2 - 12.1517 * H * Q^2 + 0.73854 * FS1^2 * H * Q^3 + 11.3897 * MTBF3) / (Q^3 - 36.3397 * Q^2 + 7.29008e^{-7} * MTBF3 * Q^4 + 1.59272 * FS1^2 * Q^3 + 937.177 * FS1^2)$
Fd	$FF = (-5.09983e^{-8} - 1.03594e^{-5} * H^3) / (Q * H - 1.07636e^{-6} - 127.418 * Q^2 + 0.486513 * FSO * MTBF)$
Fs	$FF = (-4.65753e^{-8} * H * MTBF^2 + 9.12847e^{-14} * H^2 * MTBF^2 + 6492.21 * MTBF - 1.17865e^{-7} * MTBF^3 + 0.0517087 * MTBF^2 - 165570) / (MTBF - 1.85906e^{-6} * MTBF^2 + 8821.27 * FS1 - 0.06729 * MTBR * FS1)$

Tablo 7.61. On birinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1	0.4944	0.402011
Rs2	0.4636	0.378325
Rs3	0.3944	0.34608
Rs4	0.4993	0.391969
Rs5	0.4944	0.402011
Rs6	0.5868	0.476995
Rs7	0.5891	0.444020
Rs8	0.6023	0.496966
Rs9	0.4702	0.199378
Rs10	0.5377	0.450043
Fd	0.5216	0.388918
Fs	0.4758	0.377612

On birinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R^2 değerlerine bağlı olarak, kaba küme analizi ile elde edilen

Rs6-7-10 indirgenmiş veri kümeleri ile özellik seçme analizi ile elde edilen veri kümelerinden elde edilen ve tablo 7.62’de gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.62.On birinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs6	$FF=(1.82925e^{-7} * MTBUR * H^2 + 1150.94 * H - 0.0124328 * FS1 * MTBUR * H - 26.5496 * MTBUR) / (H - 8.04861e^{-7} * Q * H^2 - 0.0336326 * MTBUR - 0.19453 * FS1^2 * MTBUR)$
Rs7	$FF=(210.827 * FSO * Q - 3184.95 * Q) / (FSO * Q - 35.9946 * FSO + 1.31176 * F2 * Q^2)$
Rs10	$FF=(0.147193 * H * MTBF^2 - 1.127e^{-7} * H^3 * MTBF - 11067.1 * H^2 - 0.784435 * H^2 * MTBF + 6.96222e^6 * MTBF) / (HOURS * MTBF - 5.49076 * H^2 + 0.103387 * FSO * H^2 - 0.196268 * FSO * MTBF^2)$
Fs	$FF=(-0.06615 * MTBUR * MTBF^2 + 812820 * MTBUR * MTBF + 1.08036e^{-6} * MTBUR * MTBF^3) / (MTBF * MTBUR - 0.000173517 * MTBUR * MTBF^2 + 92.6728 * MTBUR * MTBF + 0.473401 * FSO * MTBF^2 + 158726 * FS1)$

Tablo 7.63.On ikinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1	0.4117	0.108920
Rs2	0.6952	0.626394
Rs3	0.4441	0.407906
Rs4	0.2981	0.271294
Rs5	0.3775	0.289844
Rs6	0.6952	0.626394
Rs7	0.4257	0.267100
Rs8	0.6631	0.574827
Rs9	0.5157	0.265730
Rs10	0.565	0.516580
Fd	0.5399	0.382750
Fs	0.4907	0.357712

On ikinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R^2 değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs2-6 indirgenmiş veri kümelerinin analizinden elde edilen ve tablo 7.64’de gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.64. On ikinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs2	$FF=(0.000140474*MTBF3^2*FS1^2+7486.52*MTBF3-204.07*MTBF+6.2219e^6*FS1+0.000971172*MTBF^2)/(MTBF3+2557.76*FS1-0.0324177*MTBR*FS1)$
Rs6	$FF=(0.000140474*MTBF3^2*FS1^2+7486.52*MTBF3-204.07*MTBF+6.2219e^6*FS1+0.000971172*MTBF^2)/(MTBF3+2557.76*FS1-0.0324177*MTBR*FS1)$

Tablo 7.65. On üçüncü veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1	0.5404	0.347683
Rs2	0.4645	0.141504
Rs3	0.6291	0.526918
Rs4	0.574	0.459127
Rs5	0.6459	0.569531
Fd	0.7414	0.529343
Fs	0.6721	0.441232

On üçüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R^2 değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs5 indirgenmiş veri kümesi ile bağımlılıkları bulma analizi ile elde edilen veri kümelerinden elde edilen ve tablo 7.66’da gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.66. On üçüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs5	$FF=(0.0432074*H*Q-573.257*Q+0.40997*H)/(FS2-7.03202e^{-7}*H*Q*FS2-2.03142e^{-6}*H*FS2)$
Fd	$FF=(-0.564227*MTBF^2+0.936223*H*MTBF+5.60333*H^2*-3.34255e^{-9}*F2-5.88686*H*MTBF*F2-283962*H-1.58617e^{10})/(MTBF*F2^2-2.44e^6*F2)$

Tablo 7.67. On dördüncü veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1	0.5054	0.30279
Rs2	0.5106	0.224054
Rs3	0.4639	0.259289
Rs4	0.6132	0.480113
Rs5	0.5801	0.291283
Rs6	0.4962	0.214060
Rs7	0.6108	0.476011
Rs8	0.4709	0.265835
Rs9	0.5649	0.385751
Rs10	0.5581	0.381175
Fd	0.4515	0.247210
Fs	0.6284	0.475825

On dördüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R² değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs4-7 indirgenmiş veri kümeleri ile bağımlılıkları bulma analizi ve özellik seçme analizi ile elde edilen veri kümelerinden elde edilen ve tablo 7.68’de gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.68. On dördüncü veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs4	$FF = (0.000105719 * MTBF3 * FS1^6 - 71.5363 * MTBF3^2 * FS1 - 7501.99 * MTBF3^2 + 2856.48 * FS1) / (FS1^4 + 8.701 * MTBF3 * FS1^2 - 37710.1 * MTBF3 - 1028.13 * MTBF * FS1)$
Rs7	$FF = (2032.1 * MTBF3 * FS1 - 26.0796 * MTBF3^2) / (MTBF3 * FS1 - 150.233 * MTBF3 + 2272.18 * FS1)$
Fd	$FF = 1.74336 * F1 * NF^3 * FSO^2 / (NF^2 * FSO - 0.89217 * F1 * FSO + 0.0182821 * FSO^2 - 4.80664 * F1)$
Fs	$FF = (0.0125045 * FS1^3 * MTBF3 - 1.15248e^{-8} * FS1 * MTBF3^3 + 3392.94 * FS1 - 2814.26 * MTBF3) / (FS1^2 - 0.711778 * MTBF3 + 1.96383e^{-8} * FS1 * MTBF3^2 + 2.364e^{-6} * MTBF3^2)$

Tablo 7.69. On beşinci veri kümesinin kural bulma algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1	0.1739	Düşük olduğundan hesaplanmamıştır.
Rs2	0.4507	0.369911
Rs3	0.293	Düşük olduğundan hesaplanmamıştır.
Rs4	0.2579	Düşük olduğundan hesaplanmamıştır.
Rs5	0.3436	Düşük olduğundan hesaplanmamıştır.
Rs6	0.2964	Düşük olduğundan hesaplanmamıştır.
Fd	0.4469	0.402990
Fs	0.5637	0.434264

On beşinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R² değerlerine bağlı olarak, kaba küme analizi ile elde edilen Rs2 indirgenmiş veri kümesi ile elde edilen veri kümelerinden elde edilen ve tablo 7.70’de gösterilen kuralların daha güvenilir olduğu görülmüştür.

Tablo 7.70. On beşinci veri kümesinin kural bulma algoritması ile analizi sonucunda elde edilen en güvenilir ve doğrulukları en yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs2	$FF = (-0.195429 * F1 * MTBF3^3 + 417584 * MTBF3^2 + 11541.9 * FS1^3 * MTBF3 + 1.16843e^{11}) / (MTBF3^2 - 12474.6 * FS1 * MTBF3 + 2.61788e^7 * FS1 + 0.142504 * F1 + 5015.66 * FS1 * F1 + 1.65365e^6 * MTBF3)$

Kural bulma algoritması ile elde edilen R^2 değerleri ve kurallar karşılaştırıldığında, uzman kişilerin de düşünceleri alınarak, yüksek eğitim ve test R^2 değerleri ile elde edilen anlamlı kuralların tercih edilmesine karar verilmiştir. Kullanılabilirliği mümkün olan kurallar, tablo 7.112’de özetlenmiştir. En yüksek R^2 değerleri 1. veri kümesi ile elde edilmiş ve buna bağlı olarak oluşturulan kuralların da, ikaz değerinin “mtbf” ve “Q” parametrelerine bağlı olduğu ortaya çıkmıştır.

7.2.3.3. Lineer Regresyon Algoritması ile İkaz Değerinin Tahmin Edilmesi

Bu tez çalışması kapsamında kullanılan veri madenciliği araçlarından PolyAnalyst programının sınıflandırma algoritmasından birisi de lineer regresyon algoritmasıdır. İkaz değerinin tahmini için kural bulma algoritmasına alternatif olarak ve sonuçların karşılaştırılması amacıyla lineer regresyon algoritması tahminleme aracı olarak kullanılmıştır. Analizler sonucunda elde edilen eğitim ve test R^2 değerleri ve buna bağlı olarak hava yolu işletmesine önerilebilecek düzeyde olan kurallar sırasıyla tablo 7.71 ile tablo 7.86 arasında gösterilmiştir.

Tablo 7.71. Birinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1	0.6206	0.540574
Rs2	0.7526	0.726291
Rs3	0.5982	0.582352
Rs4	0.7044	0.661696
Rs5	0.5213	0.229243
Rs6	0.7424	0.746718
Rs7	0.3088	0.308691
Rs8	0.7341	0.782058
Rs9	0.7044	0.661696
Rs10	0.5244	0.225653
Rs11	0.7624	0.721034
Rs12	0.7424	0.775237
Rs13	0.6092	0.593576
Fd	0.7044	0.661696
Fs	0.7624	0.721034

Birinci veri kümesinin lineer regresyon algoritması ile analizi sonucunda elde edilen en yüksek eğitim ve test R^2 değerlerine bağlı olarak kaba küme analizi ile elde edilen Rs2-4-6-8-9-11-12 indirgenmiş veri kümeleri ile bağımlılıkları bulma analizi ve özellik seçme analizi sonucunda elde edilen veri kümesinden elde edilen kuralların daha güvenilir olduğu görülmüştür. Bu veri kümelerinden elde edilen kurallar tablo 7.72’de gösterilmiştir.

Tablo 7.72. Birinci veri kümesinin lineer regresyon analizi sonucunda elde edilen en güvenilir ve doğrulukları yüksek kurallar.

Veri Kümesi	Elde Edilen Kural
Rs2	$FF = -1539.54*Q + 1049.69*FSO - 0.323476*MTBUR + 0.828843*MTBF$
Rs4	$FF = -0.172405*MTBR + 0.576954*MTBF$
Rs6	$FF = -1623.82*Q - 726.878*FS1 + 1024.70*FSO + 0.578379*MTBF$
Rs8	$FF = -1689.91*Q + 877.088*FSO + 0.571760*MTBF$
Rs9	$FF = -0.172405*MTBR + 0.576954*MTBF$
Rs11	$FF = -1461.30*Q - 791.389*FS1 + 1217.61*FSO - 0.336987*MTBUR + 0.846787*MTBF$
Rs12	$FF = -1623.82*Q - 726.878*FS1 + 1024.70*FSO + 0.578379*MTBF$
Fs	$FF = -1461.30*Q - 791.389*FS1 + 1217.61*FSO - 0.336987*MTBUR + 0.846787*MTBF$
Fd	$FF = -0.172405*MTBR + 0.576954*MTBF$

Tablo 7.73-tablo 7.86'dan görüldüğü gibi birinci veri kümesi dışında diğer 14 veri kümesinin lineer regresyon algoritması ile analizinden yüksek R^2 değerlerine sahip sonuçlar elde edilememiştir. R^2 değerlerine bağlı olarak analiz sonucunda elde edilen kurallar güvenilir ve doğru kabul edilemeyeceğinden gösterilmemiştir.

Tablo 7.73. İkinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs5-Rs6-Rs7-Rs8-Rs13-Rs14-Fd-Fs	0.07244	Düşük olduğundan hesaplanmamıştır.
Rs2- Rs3- Rs4- Rs9- Rs10- Rs11- Rs12	0.04799	Düşük olduğundan hesaplanmamıştır.
Rs15	0.04592	Düşük olduğundan hesaplanmamıştır.
Rs16	0	Düşük olduğundan hesaplanmamıştır.
Rs17-Rs18	0.05276	Düşük olduğundan hesaplanmamıştır.

Tablo 7.74.Üçüncü veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1- Rs3- Rs4- Rs5- Rs8- Rs9- Rs11- Rs12	0.143	Düşük olduğundan hesaplanmamıştır.
Rs2- Rs6- Rs7- Rs10- Fd- Fs	0.1501	Düşük olduğundan hesaplanmamıştır.

Tablo 7.75. Dördüncü veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1-Rs7-Rs8	0.02169	Düşük olduğundan hesaplanmamıştır.
Rs2-Rs4	0.1803	0.101581
Rs3-Rs6-Rs9-Rs10	0.1916	0.123566
Rs5	0.1965	0.122355
Fd-Fs	0.2028	0.116927

Tablo 7.76.Beşinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1-Rs2-Rs3-Rs4-Rs5	0.0476	Düşük olduğundan hesaplanmamıştır.
Fd-Fs	0.06595	Düşük olduğundan hesaplanmamıştır.

Tablo 7.77.Altıncı veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1-Rs3-Rs5	0.08937	Düşük olduğundan hesaplanmamıştır.
Rs2-Rs4-Rs9	0	Düşük olduğundan hesaplanmamıştır.
Rs6-Fd	0.1039	Düşük olduğundan hesaplanmamıştır.
Rs7	0.1289	Düşük olduğundan hesaplanmamıştır.
Rs8	0.1289	Düşük olduğundan hesaplanmamıştır.
Fs	0.1304	Düşük olduğundan hesaplanmamıştır.

Tablo 7.78.Yedinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs2-Rs3-Rs4-Rs5-Rs6-Rs7-Rs8-Rs9-Rs10	0	Düşük olduğundan hesaplanmamıştır.
Fd-Fs	0.09395	Düşük olduğundan hesaplanmamıştır.

Tablo 7.79.Sekizinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs5-Rs6-Rs7-Rs8-Fd-Fs	0.0772	Düşük olduğundan hesaplanmamıştır.
Rs2-Rs3-Rs4	0.05515	Düşük olduğundan hesaplanmamıştır.
Rs9-Rs10	0.07482	Düşük olduğundan hesaplanmamıştır.

Tablo 7.80.Dokuzuncu veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs2-Rs3-Rs4-Rs5-Rs6-Rs7-Rs8-Rs9-Rs10	0	Düşük olduğundan hesaplanmamıştır.
Fd-Fs	0.04543	Düşük olduğundan hesaplanmamıştır.

Tablo 7.81.Onuncu veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs2-Rs3-Rs4-Rs5-Rs6-Rs7-Rs8-Rs9-Rs10-Fd	0	Düşük olduğundan hesaplanmamıştır.
Fs	0.06134	Düşük olduğundan hesaplanmamıştır.

Tablo 7.82.On birinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1-Rs3- Rs4-Rs5- Rs10-Fd-Fs	0.09339	Düşük olduğundan hesaplanmamıştır.
Rs2-Rs6- Rs7-Rs8-Rs9	0	Düşük olduğundan hesaplanmamıştır.

Tablo 7.83.On ikinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1-Rs2-Rs3-Rs4-Rs4-Rs6- Rs7-Rs8-Rs9-Rs10-Fd	0	Düşük olduğundan hesaplanmamıştır.
Fs	0.04559	Düşük olduğundan hesaplanmamıştır.

Tablo 7.84.Onüçüncü veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1-Rs3	0.1999	0.120252
Rs2	0.1218	0.039505
Rs4-Rs5	0.1216	0.033399
Fd	0.2964	0.089567
Fs	0.2083	0.037889

Tablo 7.85. On dördüncü veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1-Rs2-Rs3-Rs4-Rs4-Rs6- Rs7-Rs8-Rs9-Rs10-Fd	0	Düşük olduğundan hesaplanmamıştır.
Fs	0.08806	Düşük olduğundan hesaplanmamıştır.

Tablo 7.86.On beşinci veri kümesinin lineer regresyon algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1-Rs2-Rs3-Rs4-Rs5-Rs6-Fd	0.004811	Düşük olduğundan hesaplanmamıştır.
Fs	0.0243	Düşük olduğundan hesaplanmamıştır.

Tablo 71-86'dan görüldüğü gibi lineer regresyon ile yapılan analizlerden yüksek R² değerlerine sahip kurallar sadece birinci veri kümesi ile elde edilmiştir. Bu R² değerleri kural bulma ve polynet algoritmaları ile elde edilen en iyi sonuçlar ile tablo 7.113'de karşılaştırılmıştır.

7.2.3.4.Polynet Algoritması ile İkaz Değerinin Tahmin Edilmesi

PolyAnalyst programının sınıflandırma algoritmalarından bir diğeri olan polynet algoritması da analiz sürecinde kullanılmıştır. Ancak lineer regresyon ile analizlerde olduğu gibi tablo 7.87-tablo 7.101'den de görüldüğü gibi elde edilen R² değerlerine göre polynet algoritması ile analizlerden de doğru ve güvenilir kabul edilebilecek kurallar elde edilememiştir.

Tablo 7.87. Birinci veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R ²	
	Eğitim	Test
Rs1-Rs2- Rs5-Rs6-Rs8-Rs10-Fs	0.5917	0.406042
Rs3-Rs4-Rs7	0.3381	0.301732
Rs9	0.3784	0.226237
Rs11-Rs12-Rs13	0.4667	0.265778
Fd	0.3784	0.226237

Tablo 7.88. İkinci veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs2-Rs6-Rs8-Rs13	0.1976	Düşük olduğundan hesaplanmamıştır.
Rs3-Rs4-Rs5-Rs7-Rs9-Rs10-Rs11-Rs12-Rs14-Rs15-Rs16-Rs18-Fd-Fs	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Rs17	0.09506	Düşük olduğundan hesaplanmamıştır.

Tablo 7.89. Üçüncü veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1- Rs2- Rs3-Rs4-Rs5-Rs6-Rs7-Rs8-Rs9-Rs10-Rs11-Rs12-Fs	0.3128	0.034219
Fd	0.2701	0.072640

Tablo 7.90. Dördüncü veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs7-Rs8	0.2384	0.168358
Rs2-Rs4	0.2548	0.046344
Rs3-Rs5-Rs6-Rs9-Rs10-Fd-Fs	0.2693	0.293127

Tablo 7.91. Beşinci veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs4	0.07034	Düşük olduğundan hesaplanmamıştır.
Rs2-Rs3	0.09859	Düşük olduğundan hesaplanmamıştır.
Rs5	0.2028	Düşük olduğundan hesaplanmamıştır.
Fd	0.1523	Düşük olduğundan hesaplanmamıştır.
Fs	0.08085	Düşük olduğundan hesaplanmamıştır.

Tablo 7.92. Altıncı veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs2-Rs3-Rs4-Rs5-Rs9-Fd-Fs	0.3325	0.31817
Rs6	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Rs7-Rs8	0.3298	0.183288

Tablo 7.93. Yedinci veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs7	0.1585	Düşük olduğundan hesaplanmamıştır.
Rs2	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Rs3-Rs4-Rs6-Rs8-Rs10	0.1565	Düşük olduğundan hesaplanmamıştır.
Rs5	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Rs9	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Fd-Fs	0.1653	Düşük olduğundan hesaplanmamıştır.

Tablo 7.94. Sekizinci veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs2-Rs3-Rs4-Rs5-Rs6-Rs7-Rs8-Rs9-Rs10-Fd-Fs	0.3193	0.28661

Tablo 7.95.Dokuzuncu veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1	0.1989	Düşük olduğundan hesaplanmamıştır.
Rs2-Rs4-Rs5	0.1625	Düşük olduğundan hesaplanmamıştır.
Rs3-Rs6	0.1621	Düşük olduğundan hesaplanmamıştır.
Rs7	0.1444	Düşük olduğundan hesaplanmamıştır.
Rs8-Rs9-Fd-Fs	0.1692	Düşük olduğundan hesaplanmamıştır.
Rs10	0.1682	Düşük olduğundan hesaplanmamıştır.

Tablo 7.96.Onuncu veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1	0.2377	Düşük olduğundan hesaplanmamıştır.
Rs2	0.121	Düşük olduğundan hesaplanmamıştır.
Rs3-Rs10	0.2887	Düşük olduğundan hesaplanmamıştır.
Rs4-Rs5-Fs	0.124	Düşük olduğundan hesaplanmamıştır.
Rs6-Rs8	0.2872	Düşük olduğundan hesaplanmamıştır.
Rs7	0.1052	Düşük olduğundan hesaplanmamıştır.
Rs9	0.246	Düşük olduğundan hesaplanmamıştır.
Fd	0.2939	Düşük olduğundan hesaplanmamıştır.

Tablo 7.97. On birinci veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs2-Rs5-Fd-Fs	0.1518	Düşük olduğundan hesaplanmamıştır.
Rs3	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Rs4-Rs7-Rs8-Rs9-Rs10	0.3221	Düşük olduğundan hesaplanmamıştır.
Rs6	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.

Tablo 7.98. On ikinci veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs5	0.1339	Düşük olduğundan hesaplanmamıştır.
Rs2-Rs6-Rs9-Rs10	0.1346	Düşük olduğundan hesaplanmamıştır.
Rs3	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Rs4-Rs8	0.1111	Düşük olduğundan hesaplanmamıştır.
Rs7	Düşük olduğundan hesaplanmamıştır.	Düşük olduğundan hesaplanmamıştır.
Fd-Fs	0.1385	Düşük olduğundan hesaplanmamıştır.

Tablo 7.99. On üçüncü veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs4-Fd	0.3131	0.271923
Rs2-Rs3-Rs5	0.1935	0.125209
Fs	0.3168	0.316552

Tablo 7.100. On dördüncü veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs4-Rs7	0.3132	0.023380
Rs2-Rs3-Rs5-Rs6-Rs8-Rs9-Fd	0.3495	0.020613
Rs10	0.4373	0.049248
Fs	0.4232	0.017344

Tablo 7.101. On beşinci veri kümesinin polynet algoritması ile analizinin sonuçları.

Veri Kümesi	R^2	
	Eğitim	Test
Rs1-Rs2	0.1574	Düşük olduğundan hesaplanmamıştır.
Rs3	0.2935	Düşük olduğundan hesaplanmamıştır.
Rs4	0.3081	Düşük olduğundan hesaplanmamıştır.
Rs5-Fd-Fs	0.3236	0.296875
Rs6	0.3166	0.149414

Tablo 87-101'dan görüldüğü gibi polynet algoritması ile yapılan analizlerden yüksek R^2 değerlerine sahip kurallar sadece birinci veri kümesi ile elde edilmiştir. Bu R^2 değerleri lineer regresyon ve kural bulma algoritmaları ile elde edilen en iyi sonuçlar ile tablo 7.113'de karşılaştırılmıştır.

7.2.3.5.Mtbf-Mtbr-Mtbur Parametrelerinin Kural Bulma Algoritması ile Tahmini

İşletmeden konu ile ilgili uzmanların da önerileriyle mtbf, mtbr ve mtbur parametrelerinin de tahminine yönelik bir çalışmanın uygun olacağına karar verilmiştir. Bunun için aylık verilerden ziyade bir yıllık toplam kayıtlarla yapılacak analiz sonuçlarının daha etkili olacağı tespit edilmiştir. Bundan dolayı 13. veri kümesi olan bir yıllık toplam veri kümesi mtbf, mtbur ve mtbr parametrelerinin tahmini için kullanılmıştır. Bu parametrelerin tahmini için ikaz değerinin tahmininde kullanılan algoritmalar-kural bulma, lineer regresyon, polynet-benzer şekilde kullanılmıştır. Analiz sürecinden önce ilgili parametrelere etkisi olan diğer parametreler bağımlılıkları bulma analizi ile tespit edilmiştir. Çıktı olarak elde edilen değerler içerisinde yer alan P değerine göre bağımlılığın şiddetine karar verilmektedir. Bu değer sıfıra ne kadar yakınsa parametreler arasında o kadar güçlü bağımlılık olduğu sonucu çıkarılmaktadır. Bağımlılıkları bulma analizi sonuçları her bir parametre için tablo 7.102'de verilmiştir.

Tablo 7.102. Mtbf, Mtbr ve Mtbur parametrelerine etki eden diğer parametreler ve ilgili P değerleri.

Hedef Değişken	Etkili Parametreler	P Değeri
Mtbf	{Q, H, Fs2, F2, Fs01, Fso, Nf }	{0, 0, 0, 0, $1.74 \cdot 10^{-73}$, $1.74 \cdot 10^{-73}$, $1.74 \cdot 10^{-73}$ }
Mtbr	{Q, H, Fs2, Fs1, Fso, Nf, F2 }	{0, 0, 0, 0, 0, 0, 0 }
Mtbur	{Q, H, Fso, Fs1, Fs2, Nf, F1, F2 }	{0, 0, 0, 0, 0, 0, $1.433 \cdot 10^{-46}$ }

Bağımlılıkları bulma analizi sonucu tespit edilen etkili veri kümeleri ile mtbf, mtbr ve mtbur parametrelerinin tahmini için PolyAnalyst programının sınıflandırma araçları olan kural bulma, lineer regresyon ve polynet algoritmaları ayrı ayrı kullanılmıştır. Bu analizlerden kural bulma algoritması ile elde edilen sonuçlar tablo 7.103, tablo 7.104 ve tablo 7.105'de sırasıyla verilmiştir.

Tablo 7.103. Mtbf parametresinin tahmini için yapılan kural bulma analizi sonuçları.

R ²	Eğitim	0.986
	Test	0.967142
Elde Edilen Kural	$MTBF = (0.86335 * FSO * H * Q * F2^2 - 0.0491792 * FSO * H * Q^2 * F2 + 3565.7 * F2^3) / (FSO * F2^3 - 1.0159e^{-6} * FSO * H * F2 - 0.0545782 * FSO * Q * F2^2 - 2.49417e^{-6} * FSO * Q^3 + 7.60222e^{-6} * FSO^3)$	

Tablo 7.104. Mtbur parametresinin tahmini için yapılan kural bulma analizi sonuçları.

R ²	Eğitim	0.9825
	Test	0.977781
Elde Edilen Kural	$MTBUR = (0.99417 * FSO * H * Q^2 * FS2 - 4070.48 * FSO * Q^2 * FS2 + 11090.1 * FSO * Q * FS2 - 6558.61 * FS2) / (FSO * Q * FS2^2 + 0.0428274 * FSO * Q - 0.0008993 * FSO^2 * Q)$	

Tablo 7.105. Mtbr parametresinin tahmini için yapılan kural bulma analizi sonuçları.

R ²	Eğitim	0.9781
	Test	0.973309
Elde Edilen Kural	$MTBR = (-0.0147531 * H - 1333.91 * Q + 2146.98 + 1.01111 * Q * H) / (FSO + 0.984477 * FS2 + 1.01027 * FS1 + 0.0130213 * Q)$	

Benzer şekilde lineer regresyon ve polynet algoritması ile mtbf, mtbr ve mtbur parametrelerinin tahminine yönelik yapılan analiz sonuçları tablo 7.106-tablo 7.111’de gösterilmiştir.

Tablo 7.106. Mtbf parametresinin tahmini için yapılan lineer regresyon analizi sonuçları.

R ²	Eğitim	0.1201
	Test	0.094578
Elde Edilen Kural	$MTBF = +1756.71 * Q + 0.667136 * H + 270.734 * FS1 - 88.8877 * FSO - 408.020 * F2$	

Tablo 7.107. Mtbur parametresinin tahmini için yapılan lineer regresyon analizi sonuçları.

R^2	Eğitim	0.1162
	Test	0.098840
Elde Edilen Kural	MTBUR=+1731.66*Q+0.686087*H-439.231*F2	

Tablo 7.108. Mtbr parametresinin tahmini için yapılan lineer regresyon analizi sonuçları.

R^2	Eğitim	0.1878
	Test	0.179858
Elde Edilen Kural	MTBR=+4112.75+807.373*Q+0.516511*H-218.660*FS1-138.278*FSO -322.353*F2	

Tablo 7.109.Mtbf parametresinin tahmini için yapılan polynet algoritması ile analizinin sonuçları.

R^2	Eğitim	0.5077
	Test	0.206389
Elde Edilen Kural	MTBF=(13842.2 +H*(0.4234+H*(-3.89143e ⁻⁸ -1.3565e ⁻¹³ *H))+FS1*(FS1*(16.9066-0.0196852*FS1+2.14114e ⁻⁵ *H)+H*(-0.0144595+8.65003e ⁻⁹ *H)))	

Tablo 7.110.Mtbur parametresinin tahmini için yapılan polynet algoritması ile analizinin sonuçları.

R^2	Eğitim	0.4691
	Test	0.274876
Elde Edilen Kural	MTBUR=(16330.7+H*(0.333049+H*(-4.68991e ⁻⁸ -1.04744e ⁻¹³ *H))+FS1*(FS1*(11.1851-0.0126726*FS1)+H*(-0.0080916+6.76206e ⁻⁹ *H)))	

Tablo7.111.Mtbr parametresinin tahmini için yapılan polynet algoritması ile analizinin sonuçları.

R^2	Eğitim	0.3267
	Test	-
Elde Edilen Kural	Kullanılabilir bir kural elde edilememiştir.	

Kural bulma, lineer regresyon ve polynet algoritmaları ile yapılan analizlerde, “mtbf, mtbur ve mtbr” parametrelerinin tahmininde kural bulma algoritmasının daha etkin olduğu görülmüştür. Elde edilen kuralların da anlamlı olduğu ilgili şirketin uzmanları tarafından onaylanmıştır.

7.2.3.6.Sonuçlar

Yapılan bu çalışmada amacımız uçaklarda kullanılan parçaların, herhangi bir arıza oluşmadan önce düzeltici ve önleyici işlemlerin yapılması için ikaz seviyelerinin tespit edilmesine yönelik kural geliştirmektir. On beş adet veri kümesinin analizi ile hava yolu şirketine aylık olarak ikaz seviyesinin belirlenmesinde kullanabilecekleri kurallar ile yıllık olarak genel bir kural önerilmektedir. Hava yolu şirketinin Aralık ayı için ayrı bir analiz istemesi üzerine, aralık ayına ait iki ayrı kuralla beraber iki yıllık aralık ayı toplam kayıtlarına ait kurallar da elde edilmiştir. Ayrıca mtbf, mtbr ve mtbur parametrelerinin hesaplanmasında kullanılan formüller de tahmin edilerek alternatif formüller önerilmiştir.

İkaz değerinin tahmini için yapılan analiz sonucunda elde edilen R^2 değerlerine bağlı olarak her bir veri kümesinden kural bulma algoritması ile elde edilen ve hava yolu şirketine önerilen formüllerin tamamı tablo 7.112 (a-b-c)’de gösterilmiştir

Tablo 7.112. (a) Kural bulma algoritması ile yapılan analiz sonuçları.

Ocak	R^2	Eğitim	0.9008
		Test	0.516292
	Elde Edilen Kural (Rs4-Rs6-Rs8-Rs9)	$FF=(0.62589*Q*MTBF+1.16394e^{-12}*Q*MTBF^3-1.4148e^{-6}*Q*MTBF^2-2.44595e^{-12}*MTBF^3)/(Q-8.48828e^{-6}*MTBF+0.571867)$	
Şubat	R^2	Eğitim	0.7071
		Test	0.555277
	Elde Edilen Kural (Fd)	$FF=(0.0919908*MTBF+3.61687e^{-6}*MTBF^{2^2}*F2^{2^2}-2.80311e^{-7}*MTBF^2*Q+3642.83*Q-1341.17*Q^2)/(F2*Q-7.83467e^{-5}*F1*MTBF-0.0866749*F1)$	
Mart	R^2	Eğitim	0.5664
		Test	0.506212
	Elde Edilen Kural (Fd)	$FF=(2.01177e^{-6}*MTBUR^2*MTBR^2-1.71993e^6+74.2212*MTBR)/(MTBUR*MTBR+0.00958202*MTBR-2270.18+0.00531958*Q*MTBR+11681.1*Q^2*MTBUR)$	
Nisan	R^2	Eğitim	0.6646
		Test	0.365105
	Elde Edilen Kural (Rs1)	$FF=(9.89786e^{-7}*MTBR*FS1*MTBF3^3*+7.15407e^{-7}*MTBR^3*MTBF3-2121.82*MTBR)/(MTBF3^2-2.12104e^6*FS1+53.2268*MTBR*FS1)$	
Mayıs	R^2	Eğitim	0.7522
		Test	0.676260
	Elde Edilen Kural (Rs2)	$FF=(-0.0183299*FSO^6*MTBR^2-11.405*FSO^4*MTBR^2+1.45125*FSO^5*MTBR^2+26.4848*FSO^3*MTBR^2-2.94895e^{11}*FSO-5.58974e^6*FSO^2*MTBR+2.98425e^{11}-7163.59*Q^2*MTBR)/(FSO^5*MTBR-0.000145312*FSO^7*MTBR-1.70741e^7*FSO^2)$	
Haziran	R^2	Eğitim	0.6877
		Test	0.53945
	Elde Edilen Kural (Fd)	$FF=(1.40228e^{-6}*MTBR*MTBF*Q+31197.7*Q-127220+560810*F1-1.51884e^{-5}*MTBR^2-14.2632*MTBF)/(Q^2+4.54859e^{-5}*MTBR*Q-36.6653*Q+157.768*F1)$	

Tablo 7.112. (devam) (b) Kural bulma algoritması ile yapılan analiz sonuçları.

Temmuz	R^2	Eğitim	0.6409
		Test	0.579326
	Elde Edilen Kural (Rs8)	$FF=(4.34669e^{-7} * FS1 * MTBR^2 * MTBF3 * MTBR * F1 + 586.328 * FS1 * MTBR + 631339 * FS1 * MTBF3 * F1^2) / (FS1 * MTBR - 13696.7 * FS1 - 1282.34 * MTBF3 + 9.14969e^7 * F1^2)$	
Ağustos	R^2	Eğitim	0.712
		Test	0.644326
	Elde Edilen Kural (Rs2)	$FF=(1.06245e^{-11} * MTBR^2 * MTBF3^2 + 5758.86 * MTBF3 - 17.968 * FS1 * MTBR * MTBF3 - 625372 * FS1 * MTBR + 5.75216e^{10} * FS1 - 16.4371 * NF * MTBR^2) / (MTBF3 - 0.00331098 * NF * MTBR * MTBF3 + 4.0291e^7 * FS1 - 0.00427892 * FS1 * MTBR^2)$	
Eylül	R^2	Eğitim	0.5857
		Test	0.363503
	Elde Edilen Kural (Rs9)	$FF=(-1.75125e^{-6} * MTBUR^2 * H^2 - 0.0071005 * MTBUR^3 + 545.457 * MTBUR * H + 2082.98 * H - 2.43002e^7 + 2.678e^{-6} * MTBUR * H^3 - 0.0185117 * H^2) / (MTBUR * HOURS - 0.627298 * MTBUR^2 - 0.00827328 * Q * H + 10298.4)$	
Ekim	R^2	Eğitim	0.6156
		Test	0.556952
	Elde Edilen Kural (Rs5)	$FF=(103098 * Q^2 + 3.28536e^{-5} * H * MTBF3 * Q^2 - 12.1517 * H * Q^2 + 0.73854 * FS1^2 * H * Q^3 + 11.3897 * MTBF3) / (Q^3 - 36.3397 * Q^2 + 7.29008e^{-7} * MTBF3 * Q^4 + 1.59272 * FS1^2 * Q^3 + 937.177 * FS1^2)$	
Kasım	R^2	Eğitim	0.5868
		Test	0.476995
	Elde Edilen Kural (Rs6)	$FF=(1.82925e^{-7} * MTBUR * H^2 + 1150.94 * H - 0.0124328 * FS1 * MTBUR * H - 26.5496 * MTBUR) / (H - 8.04861e^{-7} * Q * H^2 - 0.0336326 * MTBUR - 0.19453 * FS1^2 * MTBUR)$	
Aralık	R^2	Eğitim	0.6952
		Test	0.626394
	Elde Edilen Kural (Rs2-Rs6)	$FF=(0.000140474 * MTBF3^2 * FS1^2 + 7486.52 * MTBF3 - 204.07 * MTBF + 6.2219e^6 * FS1 + 0.000971172 * MTBF^2) / (MTBF3 + 2557.76 * FS1 - 0.0324177 * MTBR * FS1)$	

Tablo 7.112. (devam) (c) Kural bulma algoritması ile yapılan analiz sonuçları.

Bir yıllık toplam	R^2	Eğitim	0.6459
		Test	0.569531
	Elde Edilen Kural (Rs5)	$FF=(0.0432074*H*Q-573.257*Q+0.40997*H)/(FS2-7.03202e^{-7}*H*Q*FS2-2.03142e^{-6}*H*FS2)$	
Geçmiş dönem Aralık	R^2	Eğitim	0.6284
		Test	0.475825
	Elde Edilen Kural (Fs)	$FF=(0.0125045*FS1^3*MTBF3-1.15248e^{-8}*FS1*MTBF3^3+3392.94*FS1-2814.26*MTBF3)/(FS1^2-0.711778*MTBF3+1.96383e^{-8}*FS1*MTBF3^2*+2.364e^{-6}*MTBF3^2)$	
İki dönemlik Aralık Toplam	R^2	Eğitim	0.4507
		Test	0.369911
	Elde Edilen Kural (Rs2)	$FF=(-0.195429*F1*MTBF3^3+417584*MTBF3^2+11541.9*FS1^3*MTBF3"+1.16843e^{11})/(MTBF3^2-12474.6*FS1*MTBF3+2.61788e^7*FS1+0.142504*F1+5015.66*FS1*F1+1.65365e^6*MTBF3)$	

Tablo 7.112'den görüldüğü gibi ikaz değerinin tespit edilmesine yönelik 15 farklı veri kümesinden 15 farklı kural elde edilmiştir. Elde edilen kurallar uygulanan karma veri madenciliği tekniklerinden faydalanılarak bulunmuştur.

Kural bulma algoritmasının, lineer regresyon ve yapay sinir ağı aracı olan polynet algoritmalarına olan üstünlüğünü göstermek amacıyla tablo 7.113'de her bir veri kümesinin kural bulma, lineer regresyon ve polynet algoritmaları ile analizinden elde edilen en yüksek R^2 değerleri gösterilmiştir. Tablodan da görüldüğü gibi her bir veri kümesi için kural bulma algoritması ile analiz sonuçları en güvenilir sonuçlardır.

Tablo 7.113. (a) Kural bulma, Lineer regresyon ve Polynet algoritmaları analizlerinden elde edilen en yüksek R^2 değerleri.

Ocak	R^2 Kural Bulma	Eğitim	0.9008
		Test	0.516292
	R^2 Lineer Regresyon	Eğitim	0.7624
		Test	0.721034
	R^2 Polynet	Eğitim	0.5917
		Test	0.406042
Şubat	R^2 Kural Bulma	Eğitim	0.7071
		Test	0.555277
	R^2 Lineer Regresyon	Eğitim	0.07244
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.1976
		Test	Düşük olduğundan hesaplanmamıştır.
Mart	R^2 Kural Bulma	Eğitim	0.5664
		Test	0.506212
	R^2 Lineer Regresyon	Eğitim	0.1501
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.3128
		Test	0.034219
Nisan	R^2 Kural Bulma	Eğitim	0.6646
		Test	0.365105
	R^2 Lineer Regresyon	Eğitim	0.2028
		Test	0.116927
	R^2 Polynet	Eğitim	0.2693
		Test	0.293127
Mayıs	R^2 Kural Bulma	Eğitim	0.7522
		Test	0.676260
	R^2 Lineer Regresyon	Eğitim	0.06595
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.2028
		Test	Düşük olduğundan hesaplanmamıştır.

Tablo 7.113. (devam) (b) Kural bulma, Lineer regresyon ve Polynet algoritmaları analizlerinden elde edilen en yüksek R^2 değerleri.

Haziran	R^2 Kural Bulma	Eğitim	0.6877
		Test	0.53945
	R^2 Lineer Regresyon	Eğitim	0.1304
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.3325
		Test	0.31817
Temmuz	R^2 Kural Bulma	Eğitim	0.6409
		Test	0.579326
	R^2 Lineer Regresyon	Eğitim	0.09395
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.1653
		Test	Düşük olduğundan hesaplanmamıştır.
Ağustos	R^2 Kural Bulma	Eğitim	0.712
		Test	0.644326
	R^2 Lineer Regresyon	Eğitim	0.07482
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.3193
		Test	0.28661
Eylül	R^2 Kural Bulma	Eğitim	0.5857
		Test	0.363503
	R^2 Lineer Regresyon	Eğitim	0.04543
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.1989
		Test	Düşük olduğundan hesaplanmamıştır.
Ekim	R^2 Kural Bulma	Eğitim	0.6156
		Test	0.556952
	R^2 Lineer Regresyon	Eğitim	0.06134
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.2939
		Test	Düşük olduğundan hesaplanmamıştır.

Tablo 7.113. (devam) (c) Kural bulma, Lineer regresyon ve Polynet algoritmaları analizlerinden elde edilen en yüksek R^2 değerleri.

Kasım	R^2 Kural Bulma	Eğitim	0.5868
		Test	0.476995
	R^2 Lineer Regresyon	Eğitim	0.09339
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.3221
		Test	Düşük olduğundan hesaplanmamıştır.
Aralık	R^2 Kural Bulma	Eğitim	0.6952
		Test	0.626394
	R^2 Lineer Regresyon	Eğitim	0.04559
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.1385
		Test	Düşük olduğundan hesaplanmamıştır.
Bir yıllık toplam	R^2 Kural Bulma	Eğitim	0.6459
		Test	0.569531
	R^2 Lineer Regresyon	Eğitim	0.2964
		Test	0.089567
	R^2 Polynet	Eğitim	0.3168
		Test	0.316552
Geçmiş dönem Aralık	R^2 Kural Bulma	Eğitim	0.6284
		Test	0.475825
	R^2 Lineer Regresyon	Eğitim	0.08806
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.4373
		Test	0.049248
İki dönemlik Aralık Toplam	R^2 Kural Bulma	Eğitim	0.4507
		Test	0.369911
	R^2 Lineer Regresyon	Eğitim	0.0243
		Test	Düşük olduğundan hesaplanmamıştır.
	R^2 Polynet	Eğitim	0.3236
		Test	0.296875

Benzer şekilde mtbf, mtbr ve mtbur parametrelerinin tahmini için yapılan analizlerden elde edilen en yüksek R^2 değerleri tablo 7.114’de gösterilmiştir. Tablo 7.114’de verilen en yüksek R^2 değerlerine göre tahmin için kullanılabilecek kurallar, kural bulma algoritması ile elde edilen kurallardır. Bu kurallar da tablo 7.115’de verilmiştir.

Tablo 7.114. Mtb_f, Mtb_r ve Mtb_{ur} parametrelerinin analizinden elde edilen R² değerleri.

MTBF	R ² Kural Bulma	Eğitim	0.986
		Test	0.967142
	R ² Lineer Regresyon	Eğitim	0.1201
		Test	0.094578
	R ² Polynet	Eğitim	0.5077
		Test	0.206389
MTBR	R ² Kural Bulma	Eğitim	0.9781
		Test	0.973309
	R ² Lineer Regresyon	Eğitim	0.1878
		Test	0.179858
	R ² Polynet	Eğitim	0.3267
		Test	Düşük olduğundan hesaplanmamıştır.
MTBUR	R ² Kural Bulma	Eğitim	0.9825
		Test	0.977781
	R ² Lineer Regresyon	Eğitim	0.1162
		Test	0.098840
	R ² Polynet	Eğitim	0.4691
		Test	0.274876

Tablo 7.115. Mtb_f, Mtb_r ve Mtb_{ur} parametrelerinin analizinden elde edilen doğrulukları en yüksek kurallar.

Parametre	Elde Edilen Kural
Mtb _f	$MTBF = (0.86335 * FSO * H * Q * F_2^2 - 0.0491792 * FSO * H * Q^2 * F_2 + 3565.7 * F_2^3) / (FSO * F_2^3 - 1.0159e^{-6} * FSO * H * F_2 - 0.0545782 * FSO * Q * F_2^2 - 2.49417e^{-6} * FSO * Q^3 + 7.60222e^{-6} * FSO^3)$
Mtb _{ur}	$MTBUR = (0.99417 * FSO * H * Q^2 * FS_2 - 4070.48 * FSO * Q^2 * FS_2 + 11090.1 * FSO * Q * FS_2 - 6558.61 * FS_2) / (FSO * Q * FS_2^2 + 0.0428274 * FSO * Q - 0.0008993 * FSO^2 * Q)$
Mtb _r	$MTBR = (-0.0147531 * H - 1333.91 * Q + 2146.98 + 1.01111 * Q * H) / (FSO + 0.984477 * FS_2 + 1.01027 * FS_1 + 0.0130213 * Q)$

7.2.4. Gecikme Oranları

Gecikme oranları her bir filo için ata kodu bazında meydana gelen gecikme oranlarını gösterir. Kayıtlar 9 parametre içermektedir. Bu 9 parametre 13 parametre içerisinde seçilmiştir. Parametreler ve tanımları Tablo 7.116'de gösterilmiştir. Bu veriler 238 kayıt içermektedir. Amacımız maksimum kalkış ağırlığı, toplam yolcu sayısı, motor sayısı gibi tasarım parametrelerine bağımlı olarak her bir filonun ata kodu bazında gecikme oranlarını tahmin etmektir. Ayrıca parametreler arasında ilişkiler de program yardımıyla tanımlanmıştır.

İşletmede gecikme oranları ile ilgili olarak bir bütün halinde mevcut raporlar bulunmamaktadır. Bunun için öncelikle gecikme oranlarının tahmini için gerekli parametreler tespit edilmiştir. Bu parametrelere ait veriler farklı tablolardan birleştirilerek veri kaynağı oluşturulmuştur. Birleştirme işlemi filo bazında ata kodlarına bağlı olarak yapılmıştır. Ancak kayıt sayısı veri madenciliği çalışması açısından yeterli değildir. Bu analiz işletmeye örnek olması açısından yapılmıştır. Bu çalışmadan sonra aşağıda tablo 7.116'de yer alan parametrelerle ilgili kayıtlar birleştirilerek kayıtlarının tutulması önerilmiştir.

İlgili parametrelerin tespit edilerek tek bir dosyada yeniden oluşturulmasının ardından ata kodu ve filo parametreleri tekrar kategorik olarak düzenlenmiştir. Amacımız kayıtlarda yer alan parametrelerin gecikme oranına olan etkilerini tespit ederek, gecikme oranının tahminine yönelik bir kural oluşturmaktır.

7.2.4.1. Gecikme Oranının Tahmini

7.2.4.1.1. Etkili Parametrelerin Belirlenmesi

Gecikme oranının tahmininden önce etkili parametrelerin tespit edilmesi gerekmektedir. Bunun için, PolyAnalyst-Bağımlılıkları Bulma ve Rosetta-Kaba Küme araçları kullanılarak özellik seçimi gerçekleştirilmiştir. Bağımlılıkları bulma analizi ile tablo 7.117'de etki değerleri de gösterilen parametrelerin gecikme oranını etkilediğine karar verilmiştir. Aynı şekilde kaba kümeler ile yapılan özellik seçimi işleminde de tablo

7.118’de gösterilen veri kümelerinin ayrı ayrı etkili olduğu tespit edilmiştir. Tahmin sürecinde bu veri kümeleri ayrı ayrı test edilmiştir.

Tablo 7.116. Gecikme oranlarına ait kayıtlardaki parametreler ve tanımları.

Parametre Adı	Tanımı	Birimi
Ata Kodu	Uçağın alt sistemlerini gösteren kategorik bir parametredir.	-
Filo	Uçakların ait olduğu filoyu gösteren kategorik veridir.	-
Gecikme Oranı	Filoya ait gecikme oranlarını gösterir. 0-1 arasında sayısal bir veridir.	-
MTW	Maksimum kalkış ağırlığıdır. Sayısal bir veridir.	Kilogram (kg.)
NP	Uçağın toplam yolcu sayısını gösterir.	-
THR	Uçak motoruna ait maksimum itkiyi gösterir.	Libre
NE	Motor sayısına bağlı olarak motor tipini göstermektedir.	-
MPR	Uçağın menzilini göstermektedir.	Kilometre (km.)
FC	Kilogram/saat cinsinden uçağın yakıt tüketimini gösterir.	Kilogram/saat (kg/h)

Tablo 7.117. Bağımlılıkları bulma algoritması ile belirlenen etkili parametreler.

Parametre adı	P-değeri
Ata Kodu	$2.355 e^{-10}$
NE, MPR	$1.811 e^{-7}$
Filo, MTW, NP, THR, FC	$8.651 e^{-7}$

Tablo 7.117’den de görüldüğü gibi uzmanlarla birlikte karar verilen ve gecikme oranına etkisi olduğu düşünülen parametrelerin tamamı bağımlılıkları bulma algoritması tarafından da etkili bulunmuştur.

Tablo 7.118. Kaba küme yaklaşımı ile elde edilen veri kümeleri.

Sıra No	Veri Kümesi
Rs1	{Ata kodu, MPR}
Rs2	{Ata kodu, FC}
Rs3	{Ata kodu, THR}
Rs4	{Ata kodu, MTW}
Rs5	{Ata kodu, Filo}
Rs6	{Ata kodu, NP}

7.2.4.1.2. Analiz Süreci

Gecikme oranları raporlarının veri madenciliği teknikleri ile analizinde PolyAnalyst programına ait sınıflandırma algoritmaları (kural bulma, lineer regresyon ve polynet) kullanılmıştır. Analiz için ön işleme sürecinde elde edilen veri kümeleri sırasıyla ele alınmıştır.

7.2.4.1.3. Sonuçlar

Sınıflandırma analizi sonucunda elde edilen ve tablo 7.119’da gösterilen R^2 değerlerine göre tahmin için yeterli ve güvenilir sonuçlar elde edilememiştir. Ancak veri kümesinin yetersiz olduğunu dikkate alarak mevcut sonuçların karşılaştırması yapılacak olursa; diğer veri kümelerinin analiz sonuçlarında da görüldüğü gibi bağımlılıkları bulma analizi ile elde edilen veri kümesinin kural bulma algoritması analizi ile doğrulukları en yüksek tahmin kuralları elde edilmiştir. Ayrıca kaba kümeler ile elde edilen veri kümelerinden sadece “Rs6” veri kümesi ile yapılan kural bulma analizi sonucunda “fd” veri kümesinin analizi sonucuna en yakın doğrulukta tahmin kuralı bulunmuştur. Bu kurallar tablo 7.120’de gösterilmiştir.

Tablo 7.119. Sınıflandırma algoritmaları ile yapılan analiz sonucunda elde edilen R^2 değerleri.

Veri Kümesi	R^2 (Kural Bulma)	R^2 (Polynet)	R^2 (Lineer Regresyon)
Rs1	0.04087	Düşük olduğundan hesaplanmamıştır.	0.04086
Rs 2	0.08892	Düşük olduğundan hesaplanmamıştır.	0.08892
Rs 3	0.007782	Düşük olduğundan hesaplanmamıştır.	0.007782
Rs 4	0.0527	Düşük olduğundan hesaplanmamıştır.	0.0527
Rs 5	-	Düşük olduğundan hesaplanmamıştır.	0.1772
Rs 6	0.3064	Düşük olduğundan hesaplanmamıştır.	0.16675
PolyAnalyst (Fd)	0.3931	0.184	0.1772

Tablo 7.120. Gecikme oranlarının tahmini için elde edilen en güvenilir kurallar.

Veri Kümesi	Kural Bulma Analizi İle Elde Edilen Kurallar
Rs 6	Gecikme Oranı = $(0.00359472 \cdot NP \cdot NP + 35.2644 \cdot \text{Eğer } (\sim 29 = \text{"AtaKodu"}, 1, 0.294495) - 265.699) / (NP - 268.762)$
PolyAnalyst (Fd)	Gecikme Oranı = $(0.836909 \cdot \text{Eğer } (\sim 29 = \text{"AtaKodu"} \text{ veya } \sim 73 = \text{"AtaKodu"}, 1, -15228.2) \cdot NE - 1.41655) / (\text{Eğer } (\sim 29 = \text{"AtaKodu"} \text{ veya } \sim 73 = \text{"AtaKodu"}, 1, -15228.2) - 0.853899)$

Tablo 7.120'den de görüldüğü gibi gecikme oranlarında etkili olan parametreler {atakodu, NP, NE} parametreleridir.

8. BÖLÜM

SONUÇ VE İRDELEME

Bu tez çalışmasında, bir hava yolu şirketi veri tabanlarında yer alan kaza raporları, sefer aksaklıkları, pilot/bakım/kabin raporları, parça raporları ve gecikme raporlarının analizinde veri madenciliği teknikleri kullanılmıştır. Daha sonra kural bulma, lineer regresyon, karar ağaçları ve yapay sinir ağları gibi teknikler kullanılarak sınıflandırmaları yapılmıştır. İlk aşamada, mevcut veri kümesindeki gürültülü ve gereksiz kayıtlar temizlenmiş ve girdi ile çıktı parametreleri arasındaki ilişkiler tespit edilmiştir. İkinci aşamada ise, verilerinden örüntüler çıkarmak için karar ağaçları, kural bulma, lineer regresyon ve yapay sinir ağları veriler üzerinde uygulanmıştır.

Altıncı bölümde Federal Havacılık Yönetimi (FHY), kaza veri sisteminden alınmış kaza raporları üzerinde veri madenciliği çalışması yapılmıştır. Sonuç olarak ölümle sonuçlanan kazalar ile ilgili bazı kurallar ve ilişkiler elde edilmiştir. Elde edilen kurallar doğrulukları ve güvenilirlikleri açısından test edilmiş ve kabul edilebilir düzeylerde olduğu görülmüştür. {Olay Tarihi, Hasar Tipi, Uçuş Tipi, Olay Tipi, Uçuş Planı, Motor Sayısı, Pilot Sertifikası, Ülke} parametreleri kullanılarak yapılan analiz sonucunda elde edilen sınıflandırma sonuçları en yüksek olasılık, etkinlik ve doğruluk değerlerini vermektedir. Yapılan çalışma ile karar ağaçları ve bağımlılıkları bulma algoritmaları veri içerisindeki etkili ve doğru bilginin bulunmasında kullanılabilir etkin metotlar olduğu görülmüştür. Elde edilen sonuçlar kullanıcının kolayca anlayabileceği ve yorumlayabileceği şekildedir. Ayrıca geleceğe yönelik tahminler için de kullanılabilirler. Sonuç olarak elde edilen kurallara bağlı olarak, paraşütle atlama ve ışıktandırma uçuşları, diğer parametrelerden bağımsız olarak %99.4 doğrulukla ölümle

sonuçlanmaktadır. Mevcut verilerin analizi sonucunda, paraşüt ile atlama ile ilgili uçuşlarda daha dikkatli olunması gerektiği sonucu çıkmıştır. Uçuşlarda ticari pilotlar değil de sertifikalı pilotlar yer almalıdır.

Yedinci bölümde sırasıyla, sefer aksaklıkları, pilot/bakım/kabin raporları, parça raporları ve gecikme oranlarının sınıflandırma teknikleri ile analizleri yapılmıştır.

Sefer aksaklıkları raporlarının karar ağacı ile analizi sonucunda elde edilen kurallara göre, {ata kodu, işlem kodu, arıza sonucu durum, tehir süresi, bakım çıkışı, filo, tehir kodu} parametrelerinin teknik tehirlere etkili olduğu sonucu çıkarılmıştır. Bu parametrelerden ise en etkin olanı tehir süresidir.

Sınıflandırma algoritmaları ile yapılan analizlerin sonucunda bütün veri kümelerinden tek bir anlamlı kural elde edilmiştir. Bu kurala göre ise sadece tehir süresi teknik tehir ile ilgili olarak belirlenmiştir.

Karar ağacı ve sınıflandırma algoritmalarından elde edilen kuralların test doğrulukları dikkate alınarak, sonuçların %97.32 ve %96.98 birbirine çok yakın değerler olduğu tespit edilmiştir. Bu durumda daha basit bir kural tercih edilirse sınıflandırma algoritmalarından kural bulma algoritması ile elde edilen kural; daha karışık bir kural tercih edilirse karar ağaçları ile elde edilen kurallar kullanılabilir.

Pilot/bakım/kabin raporlarının analizlerinden görüldüğü gibi işletmede, uçuş esnasında oluşan problem veya arıza olarak çok önemli sorunlar çıkmadığı sadece malzeme ve tefrişattan kaynaklanan sorunların olduğu görülmüştür.

Parça raporlarının kural bulma, lineer regresyon ve polynet algoritmaları ile analizi sonucunda, kural bulma algoritması ile daha etkin ve güvenilir sonuçlar elde edilmiştir. Ayrıca ön işleme tekniklerinde kullanılan kaba kümeler, bağımlılıkları bulma ve özellik seçimi analizlerinden elde edilen karma veri kümelerinin analizi ile daha etkin sonuçlar bulunmuştur.

Gecikme oranlarının analizi için ise özellikle veri madenciliği çalışması için uygun formatta veri kümesinin oluşturulması üzerine yoğunlaşarak örnek veri kümesi oluşturulmuştur.

Bu tez çalışması ile veri madenciliği teknikleri havacılık verilerine uygulanarak, ülkemizde havacılık sektöründeki veri birikimlerinin veri madenciliği ile değerlendirilebileceği ortaya konmuştur. Son zamanlarda, teknolojiye meydana gelen gelişmelere rağmen, hava taşımacılığında kazalar da artmıştır. Veri madenciliğinin bu sektörde kullanımının yaygınlaşmasıyla oluşabilecek aksaklıkların önceden tahmin edilmesi mümkün olabilecektir. Böylece, önceden yapılabilecek kestirimci bakımla muhtemel kazalar ve maddi kayıpların azaltılmasına yönelik çalışmalar veri madenciliği ile geliştirilebilecektir.

Ülkemizde son yıllarda hava taşımacılığı sektörü yolcu taşıma alanında çok iyi konumda olup kazaların giderilmesi yönünde çok yönlü çalışmalar mevcut değildir. Bu tezde yapılan çalışmalar, havacılık sektöründe gelecekte yapılabilecek çalışmalara örnek teşkil edebilecektir. Daha sonraki çalışmalarda, yazılım ve donanım destekleri ile hava yolu şirketine ait farklı birimlerdeki veri tabanlarının veri madenciliği ile analizleri gerçekleştirilebilir. Ayrıca sayısal ve kategorik verilerle birlikte metinlerin de analizleri yapılabilir.

KAYNAKLAR

1. Jiawei, H., Kamber, M., Data Mining: Concepts and Techniques, University of Simon Fraser, 2001.
2. Wong, M. L., Leung, K. S., Data Mining Using Grammar Based Genetic Programming & Applications, Genetic Programming, Vol.3, ISBN: 0-7923-7746-X, 2000.
3. Sever, H., Oğuz, B., Veri Tabanlarında Bilgi Keşfine Formal Bir Yaklaşım Kısım 1: Eşleştirme Sorguları ve Algoritmalar, Bilgi Dünyası 3(2): 173-204, Ekim, 2002.
4. Karakaş, M., Veri Madenciliği Üzerine, <http://www.bilgiyonetimi.org>.
5. Nazeri, Z., Jianping, Z., Mining aviation data to understand impacts of severe weather on airspace system performance, IEEE, Proceedings of the International Conference on Information Technology, 2002.
6. Olafsson, S., Li, X., Wu, S., Operations Research and Data Mining, European Journal of Operational Research, 187, 1429-1448, 2008.
7. Lee, S. J., Siau, K., A Review of Data Mining Techniques, Industrial Management&Data Systems, 101/1, 41-46, MCB University Press, 2001.
8. Read, B. J., Data Mining and Science? Knowledge Discovery in Science as Opposed to Business, Ercim Workshop Proceedings - No. 00/W002, 12th Ercim Workshop On Database Research, Amsterdam, 2-3 November 1999.
9. Forcht, K. A., Cochran, K., Using Data Mining and Datawarehousing Techniques, Industrial Management&Data Systems, 99/5, 189-196, MCB University Press, 1999.
10. Editorial of Engineering Applications of Artificial Intelligence 19, 361-362, Recent Advances in Data Mining, 2006.
11. Hu, H.L., Chen, Y.L., Mining Typical Patterns From Databases, Information Sciences 178, 3683-3696, 2008.
12. Huang, S.C., Chang, E.C., Wu, H.H., A Case Study of Applying Data Mining Techniques in an Outfitter's Customer Value Analysis, Expert Systems With Applications, article in press, 2008.
13. Cirapica, F. E., Giacchetta, G., Classification and Prediction of Occupational Injury Risk Using Soft Computing Techniques: An Italian Study, Safety Science, article in press, 2008.

14. Liao, C.W., Perng, Y.H., Data Mining for Occupational Injuries in the Taiwan Construction Industry, *Safety Science* 46, 1091-1102, 2008.
15. Ghazavi, S.N., Liao, T. W., Medical Data Mining by Fuzzy Modeling With Selected Features, *Artificial Intelligence in Medicine* 43, 195-206, 2008.
16. Bertolazzi, P., Felici, G., Festa, P., Lancia, G., Logic Classification and Feature Selection for Biomedical Data, *Computers and Mathematics With Applications*, 55, 889-899, 2008.
17. Yeh, I.C., Lien, C.H., The Comparisons of Data Mining Techniques for The Predictive Accuracy of Probability of Default Credit Card Clients, *Expert Systems With Applications*, article in press, 2008.
18. Ke, M.C., Lin, W.T., Wang, S.T., Wang, W.S., Planning of Educational Training Courses by Data Mining: Using China Motor Corporation as an Example, *Expert Systems With Applications*, article in press, 2008.
19. Piramuthu, S., Input Data for Decision Trees, *Expert Systems With Applications* 34, 1220-1226, 2008.
20. Azadeh, A., Saberi, M., Ghaderi S.F., Gitiforouz, A., Ebrahimipour, V., Improved Estimation of Electricity Demand Function By Integration of Fuzzy System and Data Mining Approach, *Energy Conversion and Management* 49, 2165-2177, 2008.
21. Li, X.B., Jacob, V. S., Adaptive Data Reduction for Large-Scale Transaction Data, *European Journal of Operational Research* 188, 910-924, 2008.
22. Urtubia, A., Perez-Correa, R.J., Soto, A., Pszczolkowski, P., Using Data Mining Techniques to Predict Industrial Wine Problem Fermentations, *Food Control* 18, 1512-1517, 2007.
23. Kusiak, A., Smith, M., Data Mining in Design of Products and Production Systems, *Annual Reviews in Control* 31, 147-156, 2007.
24. Gürbüz, F., Özbakır, L., Seramik Üretiminde Veri Madenciliği, ÜAS'07, Üretim Araştırmaları Sempozyumu, 15-17 Kasım, Ankara, 2007.
25. Özbakır, L., Baykasoğlu, A., Kulluk, S., Gürbüz, F., Bir Tekstil İşletmesinde Kalite İyileştirmeye Yönelik Veri Madenciliği Uygulaması, 7. Ulusal Üretim Araştırmaları Sempozyumu (ÜAS'07), 15-17 Kasım, Ankara, 2007.
26. Gürbüz, F., Kulluk, S., İşletme Performansına Etki Eden Faktörlerin Belirlenmesinde Veri Madenciliği Yaklaşımı, I. Üniversite Sanayi İşbirliği Sempozyumu, USIS'2007, 5-7 Haziran, Sakarya, 2007.

27. Kulluk, S., Özbakır, L., Gürbüz, F., Baykasoğlu, A., Uzun, G., Öğrenci Başarısını Etkileyen Faktörleri Belirlemeye Yönelik Bir Veri Madenciliği Uygulaması, 13. Ulusal Ergonomi Kongresi, 6-8 Aralık, Kayseri, 2007.
28. Kusiak, A., Data Mining: Manufacturing and Service Applications, International Journal of Production Research, Vol. 44, Nos. 18-19, 15 September-1 October 2006.
29. Hsia, T. C., Shie, A. J., Chen, L. C., Course Planning of Extension Education to Meet Market Demand by Using Data Mining Techniques-An Example of Chinkuo Technology University in Taiwan, Taiwan, 2006.
30. Chang, C. C., Chen, R. S., Using Data Mining Technology to Solve Classification Problems, A Case Study of Campus Digital Library, Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan, 2006.
31. Chang, A. S., Leu, S. S., Data Mining Model For Identifying Project Profitability Variables, International Journal of Project Management, 24, 199-206, 2006.
32. Rubinov, M.A., Soukhorokova, N.V., Ugon, J., Classes and Clusters in Data Analysis, CIAO and School of Information Technology and Mathematical Sciences, University of Ballarat, Australia, 2005.
33. Inniss, T. R., Seasonal Clustering Technique for Time Series Data, European Journal of Operational Research, USA, 2005.
34. Mai, C. K., PolyAnalyst Application for Forest Data Mining,, VNRVJ Institute of Engg. & Technology Hyderabad, India, IEEE, 2005.
35. Tseng, B., Kwon, Y., Yalçın, E., Feature-based Rule Induction in Machining Operation Using Rough Set Theory for Quality Assurance, Robotics and Computer-Integrated Manufacturing, 21:6, 559-567, 2005.
36. Hsu, C. C., Chen, Y. C., Mining Of Mixed Data With Application To Catalog Marketing, Expert Systems With Applications, Taiwan, 2005.
37. Baykasoğlu, A., Özbakır, L., A New Multiple Expression Programming Based Approach to Discover Classification Rules For Data Mining Applications, CIE 35, 35th International Conference on Computers and Industrial Engineering, 19-22, İstanbul, 2005.
38. Baykasoğlu, A., Özbakır, Mepar-Miner: Multi-Expression Programming For Classification Rule Mining, European Journal of Operational Research 183, 767-784, 2007.

39. Canbulut, F., Gürbüz, F., Kulluk, S., Özbakır, L., Dairesel Cepli Hidrostatik/Hidrodinamik Yataklarda Sürtünme Güç Kaybı, YA/EM 2005, 4-6 Temmuz, Koç Üniversitesi, İstanbul, 2005.
40. Jothishankar, M. C., Wu, T., Johnie, R., Shiau, J.Y., Case Study: Applying Data Mining to Defect Diagnosis, Journal of Advanced Manufacturing Systems, 3:1- 69-83, 2004.
41. Last, M., Kandel, A., Discover Useful and Understandable Patterns in Manufacturing Data, Robotics and Autonomous Systems, 49, 137-152, 2004.
42. Hyeon Bae, Sungshin kim, Yeajin Kim, Man Hyung Lee, Kwang Bang Woo, E-Prognosis and Diagnosis for Process Management Using Data Mining and Artificial Intelligence, IEEE, 2003.
43. Porzio, G. C., Ragozini, G., Visially Mining Off-Line Data for Quality Improvement, Quality and Reliability Engineering International, 19, 273-283, 2003.
44. Larry Adams, Mining The World of Quality Data, Quality, 40, 8, ABI/INFORM Global, pp. 36, Aug 2001.
45. Kusiak, A., Kurasek, C., Data Mining of Printed-Circuit Board Defects, IEEE, Transactions on Robotics and Automation, 17:2, April, 2001.
46. Lin, F.Y., McClean, S., A Data Mining Approach To The Prediction of Corporate Failure, Knowledge-Based Systems 14, 189-195, 2001.
47. Feelders, A, Daniels, H., Holsheimer, M., Methodological And Practical Aspects Of Data Mining, Information & Management 37, 271-281, 2000.
48. McDonald, C. J., New Tools for Yield Improvement in Integrated Circuit Manufacturing: Can They be Applied to Reliability?, Microelectronics Reliability, 39, 731-739, 1999.
49. Milne, R., Durummond, M., Renoux, P., Predicting Paper Making Defects On-Line Using Data Mining, Knowledge-Based Systems, 11, 331-338, 1998.
50. Apte, C., Weiss, S., Data mining with decision trees and decision rules, Future Generation Computer Systems, 1997.
51. Kim, S. H., Lee, C. M., Nonlinear Prediction of Manufacturing Systems Through Explicit and Implicit Data Mining, Computers Ind. Eng., 33:3-4. 461-464, 1997.
52. Stepien, D., Refractory Management Using Fact-Based Analysis For the Cement Klin, IEEE/PCA Cement Industry Technical Conference, XXXIX Conference Record, 359-369, Hershey, PA, USA, 1997.

53. Zhang, S., Zhang, C., Yang, Q., Data Preparation For Data Mining, *Applied Artificial Intelligence*, 17:375–381, 2003.
54. Kim, Y., Street, W. N., Menczer, F., Feature Selection In Data Mining, University Of Iowa, Usa.
55. Chen, W., Tseng, S., Hong, T., An Efficient Bit-Based Feature Selection Method Expert Systems With Applications 34, 2858–2869, 2008.
56. Pawlak, Z., Grzymala-Busse, J., Slowinski, R., Ziarko, W., “Rough Sets”, *Communications of the ACM*, C. 38, S.11, s 89-95,1995.
57. Pawlak, Z., “Rough Sets Theoretical Aspects of Reasoning About Data”, Kluwer Academic Publishers, Boston, 1-30, 1991.
58. Pawlak, Z., Slowinski, R., “Rough Set Approach to Multi-attribute Decision Analysis”, *European Journal of Operational Research*, 72: 443-459, 1994.
59. Pawlak, Z., *Rough Sets and Intelligent Data Analysis*, Information Sciences, Vol. 147, pp. 1-12, 2002.
60. Thangavel, K., Pethalakshmi, A., Dimensionality Reduction Based on Rough Set Theory: A Review, *Accepted Manuscript, Applied Soft Computing*, 2008.
61. Leung, Y., Fischer, M. M., Wu, W. Z., Mi, J. S., A Rough Set Approach For The Discovery of Classification Rules in Interval-Valued Information Systems, *International Journal of Approximate Reasoning*, 47, 233-246, 2008.
62. Tripathy, H. K., Tripathy, B. K., Das, P. K., An Intelligent Approach of Rough Set in Knowledge Discovery Databases, *International Journal of Computer Science and Engineering*, Vol. 2, No. 1, 2007.
63. Zhu, W., Generalized Rough Sets Based on Relations, *Information Sciences*, 177, 4997-5011, 2007.
64. Huang, C. L., Li, T. S., Peng, T.K., A Hybrid Approach Of Rough Set Theory and Genetic Algorithm for Fault Diagnosis, *Int. J. Adv. Manuf. Tehnol.*, 27: 119-127, 2005.
65. Busse, J., Siddhaye, S., Rough Set Approaches to Rule Induction From Incomplete Data, *Proceedings of the IPMU'2004, The 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Italy, July 4-9, vol 2, 923-930, 2004.

66. Joseph, L., Breault, Data Mining Diabetic Databases: Are Rough Sets a Useful Addition, Proc. 33rd Symposium on the Interface, Computing Science and Statistics, Fairfax, 1-10, 2001.
67. Felix, R., Ushio, T., Rule Induction From Inconsistent and Incomplete Data Using Rough Sets, Department of Systems and Human Science, Graduate School of Engineering Science, Osaka University, Japan, Available From: <http://ieeexplore.ieee.org/iel5/6569/17562/00815540.pdf?arnumber=815540>.].
68. Pawlak, Z., Rough Set Approach to Knowledge-Based Decision Support, European Journal of Operational Research, 99, 48-57, 1997.
69. Gürbüz, F., Özbakır, L., Yapıcı, H., Türkiye’de Bir Havayolu İşletmesine Ait Parça Söküm Raporlarına İlişkin Veri Madenciliği Uygulaması, Gazi Üniversitesi, Mühendislik Mimarlık Dergisi, basım aşamasında, 2008.
70. Gürbüz, F., Özbakır, L., Feature Selection and Preprocessing Application On Component Reports Of An Airline Company in Turkey, HDM-2008, International Conference on Multivariate Statistical Modeling And High Dimensional Data Mining, 19-23 Haziran, Erciyes Üniversitesi, Kayseri, Türkiye, 2008.
71. Shyur, H. J., A quantitative model for aviation safety risk assesment, Computers&Industrial Engineering, Vol. 54, No 1, pp. 34-44, 2008.
72. Dessureault, S., Sinuhaji, A., Coleman, P., Data mining mine safety data, Mining Engineering, Littleton:Vol. 59, Iss. 8; pg. 64, 7 pgs., 2007.
73. Solomon, S., Nguyen, H., Liebowitz, J., Agresti, W., Using data mining to improve traffic safety programs, Industrial Management & Data Systems, Vol. 106 No. 5, pp. 621-643, 2006.
74. Gürbüz, F., Özbakır, L., Kulluk, S., Havayolu İşletmeleri İçin Tehir Raporlarında Veri Madenciliği Uygulaması, YA/EM 2007, Yöneylem Araştırması/Endüstri Mühendisliği XXVII. Ulusal Kongresi, 2-4 Temmuz, İzmir, 2007.
75. Kulluk, S., Gürbüz, F., Havacılık Uçuş Güvenliğinde Bir Veri Madenciliği Uygulaması, Hasem’06 VI. Havacılık Sempozyumu, 12-14 Mayıs, Nevşehir, 2006.
76. Tseng, W. S., Nguyen, H., Liebowitz, J., Agresti, W., Distractions and motor vehicle accidents:Data mining application on fatality analysis reporting system (FARS) data files, Industrial Management&Data Systems, Vol. 105, No. 9, pp 1188-1205, 2005.

77. Ananyan, S., Goodfellow, M., Application of PolyAnalyst With IATA Steades Data, Global Aviation Information Network, 2004.
78. Bineid, M., Fielding, J. P., Development of a Civil Aircraft Dispatch Reliability Prediction Methodology, Aircraft Engineering and Aerospace Technology, Volume 75, Number 6, pp. 588-594, 2003.
79. Nazeri, Z., Jianping, Z., Mining Aviation Data to Understand Impacts of Severe Weather On Airspace System Performance, Proceedings of the International Conference on Information Technology, IEEE, 2002.
80. Letourneau, S., Famili, F., Matwin, S., Data Mining For Prediction of Aircraft Component Replacement, August, 1999.
81. Vahaplar, A., İnceoğlu, M. M., Veri Madenciliği ve Elektronik Ticaret, <http://inet-tr.org.tr/inetconf7/eposter/inceoglu.doc>, 2002.
82. Hand, David; Heikki Mannila; Padhraic Smyth; Principles of Data Mining, The Mit Press, England 2001.
83. Feelders, A., Daniels, H., Knowledge Discovery in Practice, Proceedings of the International Conference Machine Learning and Applications, Workshop 08: Data Mining in Economics, Marketing and Finance, Yunanistan, s. 1., 1999.
84. Biçen, P., Veri Madenciliği: Sınıflandırma ve tahmin Yöntemlerini Kullanarak bir Uygulama, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, İstanbul, 2002.
85. Chengqi, Z., Shichao, Z., Association Rule Mining: Models and Algorithms, Springer-Verlag Berlin Heidelberg, New York, 2002.
86. Sarıkan, B., Veri Madenciliği ve İstatistik, Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara, 2003.
87. Özmen, Ş., İş Hayatı Veri Madenciliği ile İstatistik Uygulamalarını Yeniden Keşfediyor, Marmara Üniversitesi, İngilizce İşletme Bölümü.
88. Yurtsever, U., Veri Madenciliği ve Uygulaması, Yüksek Lisans Tezi, Sakarya Üniversitesi, Sakarya, 2002.
89. Gezder, Ü., Veri Ambarcılığı ve Türkiye Uygulaması, Yüksek Lisans Tezi, Atatürk Üniversitesi, Erzurum, 2002.
90. Çakır, M., Firma Başarısızlığının Dinamiklerinin Belirlenmesinde Makine Öğrenmesi Teknikleri: Ampirik Uygulamalar ve Karşılaştırmalı Analiz, Uzmanlık Yeterlilik Tezi, Türkiye Cumhuriyeti, Merkez Bankası, İstatistik Genel Müdürlüğü, Ankara, Aralık, 2005.

91. Witten, I.H., Frank, E., Data Mining Practical Machine Learning Tools and Techniques With Java Implementations, Morgan Kaufman Publishers, Inc, San Fransisco, CA, 2000.
92. Chang, C. C., Chen, R. S., Using Data Mining Technology to Solve Classification Problems, A Case Study of Campus Digital Library, Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan, 2006.
93. Larose, T. Daniel, Discovering Knowledge in Data : An Introduction to Data Mining, s. 63, John Wiley & Sons, Inc. Publishes, New Jersey, 2005.
94. Analysis of Data Mining Techniques For Customer Segmentation and Predictive Modeling-A Case Study By Rupasri R. Kadammi Doktora Tezi, Bangalore Üniversitesi, India, 2002.
95. Baykal, A., Veri Madenciliği Öğrenci Verileri Üzerinde Uygulamaları, Dicle Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Bölümü, Doktora Tezi, Diyarbakır, Haziran, 2003.
96. Oğuzlar, A., Veri Ön İşleme, Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Sayı: 21, Temmuz-Aralık, ss. 67-76, 2003.
97. Famili, A., Shen W., Weber R., Simoudis, E., Data Preprocessing and Intelligent Data Analysis , Intelligent Data Analysis, 1, USA, pp.3-23.,1997.
98. Goharian, N. and Grossman, D., “Data Preprocessing”, İnternet Adresi; <http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM-reprocessing.pdf>, 2003.
99. Eker, H., Veri Madenciliği veya Bilgi Keşfi, http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538, erişim tarihi, 24.12.2008.
100. Özmetin, R., Müşteri ilişkileri (CRM) yönetiminde veri madenciliği metodolojisi ve bir gıda firmasına uygulama çalışmaları, Yüksek Lisans Tezi, Sakarya Üniversitesi, Sakarya, 2002.
101. Zhang, C., Zhang, S., Association Rule Mining, Models and Algorithms, Springer, 2002.
102. Tantığ, A.C., Veri madenciliği ve demetleme, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2002.
103. Mailvaganam, H., Text Mining, http://www.dwreview.com/Data_mining/text_mining.html.

104. Akpınar, H., Veritabanlarında Bilgi Keşfi ve Veri Madenciliği, İstanbul Üniversitesi, İşletme Fakültesi Dergisi, 1-22, Nisan, 2000.
105. Haykin, S., Neural Networks, Macmillan College Publishing Company, 696, USA, 1994.
106. Aynekin, G., İnternek İçerik Madenciliğinde Yapay Sinir Ağları ve Bir Uygulama, Yüksek Lisans Tezi, Uludağ Üniversitesi, Bursa, 2006.
107. Elmas, Ç., Yapay Sinir Ağları, Seçkin Yayıncılık, Ankara, 192, 2003.
108. Holland, J. H., Adaption in Natural and Artificial Systems, University of Michigan Pres, Ann Arbor, 1975.
109. Khoo, L. P., Tor, S. B., Zhai, L. Y., A Rough-Set-Based Approach for Classification and Rule Induction, The International Journal of Advanced Manufacturing Technology 15: 438-444, Springer-Verlag, 1999.
110. Khoo, L. P., Tor, S. B., Zhai, L. Y., A rough set-based approach for classification and rule induction, Advanced Manufacturing Technology, 15:438-444, Springer-Verlag, 1999.
111. Binay, H. S., Yatırım kararlarında kaba küme yaklaşımı, Doktora Tezi, Ankara Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara, 2002.
112. Pawlak, Z., "Rough sets", International Journal of Computer and Information Sciences, 11:341-356. 1982.
113. Marek, W., Pawlak, Z., Rough Sets and Information Systems, Fundamenta Informaticae 17, 105-115, 1984.
114. Hassan, Y., Tazaki, E., Emergent rough set data analysis, Kybernetes, Vol. 34 No. 6, pp. 869-887, 2005.
115. Yin, X., Zhou, Z., Li, N., Chen, S., An Approach for Data Filtering Based on Rough Set Theory, Wang, G. Lu, H. Lu (Ed), Computer Science 2118, Berlin, Springer-Verlag, pp. 367-374, 2001.
116. Aydoğan-Kızılkaya, E., Veri madenciliğinde sınıflandırma problemleri için evrimsel algoritma tabanlı yeni bir yaklaşım: Rough-Mep Algoritması, Doktora Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 2008.
117. Walczak, B., Massart, D. L., Rough Sets Theory, Chemometrics and Intelligent Laboratory Systems, Vol. 47, pp. 1-16, 1999.

118. Guo, J. Y., Rough Set-Based Approach to Data Mining, Ph. D. Thesis, Case Western Reserve University, Department of Electrical Engineering and Computer Science, 2003.
119. Goh, C., Law, R., Incorporating the Rough Set Theory Into Travel Demand Analysis, *Tourism Management*, Vol.24, pp 511-517, 2003.
120. Pawlak, Z., Rough Sets, Decision Algorithms and Bayes' Theorem *European Journal of Operational Research*, Vol. 136, pp. 181-189, 2002.
121. Øhrn, A., Discernibility and Rough Sets in Medicine: Tools and Applications. PhD thesis, Norwegian University of Science and Technology, Department of Computer and Information Science, NTNU report 1999-
<http://www.idi.ntnu.no/~aleks/thesis/>.
122. Telçeken, S., Kaba Kümeler Teorisi Yardımı ile Büyük Veri Topluluklarının Analizi, Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi, Ağustos, 2003.
123. Vinterbo, S., Øhrn, A., Minimal Approximate Hitting Sets and Rule Templates, *International Journal of Approximate Reasoning*, 25, 123-143, 2000.
124. Norwegian University of Science and Technology, ROSETTA version 1.4.4.1. Available from: www.idi.ntnu.no/~alex/rosetta, 2001.
125. User Manuel of PolyAnalyst 5, April 2005.
126. Carvalho, D.R., Freitas, A. A., A Genetic Algorithm For Discovering Small-Disjunct Rules in Data Mining, *Applied Soft Computing* 2, 75-88, 2002.
127. Takac, A., Cellular Genetic Programming Algorithm Applied to Classification Task, Institute of Informatics, Faculty of Mathematics, Physics and Informatics Coneius University, Brastislava, Slovakia, 2002.
128. Folino, G., Pizzuti, C., Spezzano, G., A Cellular Genetic Programming Approach to Classification, *Proc. Of the Genetic and Evolutionary Computation Conference GECCO99*, Morgan Kaufmann, 1015-1020, Orlando, Florida, 1999.
129. Tan, K. C., Yu, Q., Ang, J. H., A Dual Objective Evolutionary Algorithm for Rules Extraction in Data Mining, *Computational Optimization and Applications*, 34, 273-294, 2006.
130. Kiselev, M. V., PolyAnalyst 2.0: Combination of Statistical Data Preprocessing and Symbolic KDD Technique, *Proceedings of ECML-95 Workshop on Statistics*,

- Machine Learning and Knowledge Discovery in Databases, Heraklion, Greece, pp 187-192, 1995.
131. Chien, C., Wang, W., Cheng, J., Data Mining For Yield Enhancement in Semiconductor Manufacturing and an Empirical Study, Taiwan, 2006.
 132. Wong, M.L., Leung, K. S., Data Mining Using Gramer Based Genetic Programming and Applications, Kluwer Academic Publishers, London, 1999.
 133. Quinlan, J. R., Induction of Decision Trees, Journal of Machine Learning, 81-106, 1986.
 134. Mitchell, T., Machine Learning, London, McGraw-Hill International, 52-81, 1997.
 135. Akgöbek, Ö., Öztemel, E., Endüktif Öğrenme Algoritmalarında Kural Türetme Yöntemleri ve Performanslarının Karşılaştırılması, SAU, Fen Bilimleri Enstitüsü Dergisi, 10. Cilt, 1. Sayı, 2006.
 136. Dunham, M. H., Data Mining Introductory and Advanced Topics, Prentice Hall, Pearson Education Inc., New Jersey, 2003.
 137. Aydoğan, F., E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirimi, Yüksek Lisans Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Ankara, 2003.
 138. Silahtaroglu, G., Veri Madenciliğinde Kümeleme Analizi ve Öğretim Başarısının Değerlemesine İlişkin Bir Uygulama, Doktora Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı, İstanbul, 2004.
 139. Kiselev, M.V., Arseniev, S.B., Discovery Of Numerical Dependencies In Form Of Rational Expressions, Computer Patient Monitoring Laboratory at National Research Center of Surgery, per.Abriksovski, 2, Moscow 119874, RUSSIA.
 140. www.gmdh.net, erişim tarihi: Aralık 2008.
 141. Clementine 10.0 User's Guide, 2005.
 142. www.asias.faa.gov, erişim tarihi: Mayıs 2006.

ÖZGEÇMİŞ

Feyza Gürbüz 1977 yılında Kayseri’de doğdu. İlk öğrenimini Kayseri’de tamamladı. 1989 yılında Kayseri Sami Yangın Anadolu Ticaret Lisesi’ni kazanarak orta ve lise öğrenimini burada tamamladı. 1995 yılında kazandığı Erciyes Üniversitesi Endüstri Mühendisliği Bölümü’nden 1999 yılında mezun oldu. 2000’de Erciyes Üniversitesi Mühendislik Fakültesi Endüstri Mühendisliği Bölümü’nde uzman olarak göreve başladı. Aynı yıl Erciyes Üniversitesi Fen Bilimleri Enstitüsü Makine Mühendisliği Bölümü’nde yüksek lisans eğitimine başladı. Yüksek lisans eğitimini 2003’de tamamladıktan sonra aynı yıl Erciyes Üniversitesi Fen Bilimleri Enstitüsü Makine Anabilim dalında doktora eğitimine başladı. Halen Erciyes Üniversitesi Mühendislik Fakültesi Endüstri Mühendisliği Bölümü’nde görevine devam etmektedir. Feyza Gürbüz evli olup bir kız çocuğu annesidir.

Adres :Erciyes Üniversitesi Mühendislik Fakültesi
Endüstri Mühendisliği Bölümü
38039-KAYSERİ

Telefon : 0 352 437 49 01 – 32453

Faks : 0 352 437 57 84

E-Mail : feyza@erciyes.edu.tr

