

**VERİ MADENCİLİĞİNİN TIPTA KULLANIMI VE BİR UYGULAMA :
HEMODİYALİZ HASTALARI İÇİN RİSK SEVİYELERİNE GÖRE RİSK
FAKTÖRLERİNİN ETKİLEŞİMLERİNİN İNCELEMESİ**

Yunus Yesari ALTINTAŞ

**YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**OCAK 2010
ANKARA**

**VERİ MADENCİLİĞİNİN TIPTA KULLANIMI VE BİR UYGULAMA :
HEMODİYALİZ HASTALARI İÇİN RİSK SEVİYELERİNE GÖRE RİSK
FAKTÖRLERİNİN ETKİLEŞİMLERİNİN İNCELENMESİ
(Yüksek Lisans Tezi)**

Yunus Yesari ALTINTAŞ

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
Ocak 2010**

ÖZET

Veri madenciliği, verilerden keşfedilecek desenler yardımıyla yeni bilgiler elde etme amacıyla çok farklı disiplinlerde kullanılan çeşitli metotlardan oluşmaktadır. Tıp alanındaki verinin büyüklüğü ve hayati önem taşıması, Veri madenciliğinin bu alanda da uygulanmasını gerekli kılmıştır. Bu tezde Veri Madenciliğinin Tıp alanında kullanımı incelenmiştir. Uygulama çalışması için özel bir diyaliz merkezinde 12 ay ve daha uzun süre hemodiyaliz tedavisi görmüş hastalara ait veriler bir araya getirilerek bir veri kümesi oluşturulmuştur. Bu veri kümesi üzerinde ESTARD ve WEKA yazılımları kullanılarak sınıflama, kümeleme ve karar ağacı algoritmaları çalıştırılmış, elde edilen karar kuralları uzman desteğiyle incelenerek risk seviyelerine göre risk faktörlerinin etkileşimlerinden yararlı desenlere ulaşılmaya çalışılmıştır.

Bilim Kodu : 906.1.071
Anahtar Kelimeler : Veri Madenciliği, Tıbbi bilişim, Kronik Böbrek Yetmezliği, Hemodiyaliz
Sayfa Adedi : 222
Tez Yöneticisi : Prof. Dr. Hadi GÖKÇEN

**THE USAGE OF DATA MINING IN MEDICINE AND
AN APPLICATION: ANALYSIS OF RISK FACTORS' INTERACTIONS
ACCORDING TO RISK LEVELS FOR HEMODIALYSIS PATIENTS
(M.Sc. Thesis)**

Yunus Yesari ALTINTAŞ

**GAZİ UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY
January 2010**

ABSTRACT

Data mining is composed of several methods used within very different fields in order to gain new information from discovered data patterns. For the field of medicine, the amount of medical data and its vital importance make the usage of data mining in medicine a requirement. In this thesis, the usage of Data mining in the field of Medicine is analysed. For the application study, a data set formed by collecting data of dialysis patients treated for 12 months or more from a private dialysis center. By running classification, clustering and decision tree algorithms on this data set in ESTARD and WEKA softwares, decision rules are obtained as the results. These decision rules are analysed with the expert support and tried to gain useful patterns from interactions of risk factors according to risk levels.

Science Code : 906.1.071

**Key Words : Data mining, Medical informatics, Chronic kidney
failure, Hemodialysis**

Page Number : 222

Adviser : Prof. Dr. Hadi GÖKÇEN

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren Tez Danışmanı Hocam Prof. Dr. Hadi GÖKÇEN Bey'e, yine kıymetli tecrübelerinden faydalandığım hocam Prof. Dr. Emin ACAR Bey'e, uygulama çalışması için yol gösteren ve her aşamada yardımcı olan Refik Saydam Hıfzıssıhha Merkezi Daire Tabibi Dr. Eyüp ÖZEREN Bey ile Ankara Numune Eğitim ve Araştırma Hastanesi Başhekim Yardımcısı ve 3. Biyokimya Klinik Şefi Doç. Dr. Metin YILDIRIMKAYA Bey'e, firmasının kapılarını açan ve uygulama sürecinde her tür desteği veren Özel CAN Diyaliz Merkezi sahibi ve Özel Diyaliz Merkezleri Derneği (DİADER) Yönetim Kurulu Üyesi Sn. Ömer ASLANTAŞ Bey'e, uzman desteği sağlayan Özel Can Diyaliz Merkezi Hemodiyaliz hekimlerinden Dr. Mustafa Mahir ÜLGÜ Bey ile Ankara Numune Eğitim ve Araştırma Hastanesi Nefroloji Klinik Şefi Dr. Ali Rıza ODABAŞ Bey ve Hemodiyaliz Hekimi Dr. Hadim AKOĞLU Bey'e, verilere erişim ve ulaşım konularında yardımcı olan tüm Can Diyaliz çalışanlarına, manevi destekleriyle beni hiçbir zaman yalnız bırakmayan çok değerli aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	xi
ŞEKİLLERİN LİSTESİ	xiii
RESİMLERİN LİSTESİ.....	xv
SİMGELER VE KISALTMALAR.....	xvi
1. GİRİŞ	1
2. VERİ MADENCİLİĞİNE GENEL BAKIŞ.....	2
2.1. Veri Madenciliği ile İlişkili Bilim Dalları.....	5
2.2. Veri Madenciliği Süreci.....	6
2.2.1. İşin (Problemin) belirlenmesi ve verilerin anlaşılması.....	8
2.2.2. Verilerin toplanması.....	9
2.2.3. Verilerin hazırlanması.....	22
2.2.4. Modelin kurulması ve uygulanması.....	27
2.2.5. Modelin yorumlanması ve elde edilen sonuçların sunumu..	31
2.3. Veri Madenciliği Sistemleri.....	31
2.3.1. Tipik bir veri madenciliği sisteminin yapısı.....	31
2.3.2. Veri madenciliği sistemlerinin sınıflandırılması.....	33
2.4. Veri Madenciliğinde Kullanılan Metotlar.....	35
2.5. Veri Madenciliğinin Kullanım Alanları.....	36

	Sayfa
2.6. Veri Madenciliği Yazılımları.....	37
2.7. Veri Madenciliğinin Geleceği.....	41
2.7.1. Potansiyel veri madenciliği uygulama alanları.....	41
2.7.2. Veri madenciliğinden doğan yeni kavramlar.....	43
3. VERİ MADENCİLİĞİ METOTLARI.....	45
3.1. Veri Madenciliği Metotlarının Sınıflandırılması.....	45
3.2. Regresyon Tabanlı Modeller.....	47
3.2.1. Doğrusal regresyon.....	48
3.2.2. Çoklu regresyon.....	50
3.3. Sınıflandırma Modelleri.....	52
3.3.1. Karar ağaçları ve karar ağacı algoritmaları.....	53
3.3.2. Yapay sinir ağları.....	64
3.3.3. Saf Bayes sınıflaması.....	69
3.3.4. Diğer sınıflama metotları.....	70
3.4. Kümeleme Analizi.....	72
3.5. Birliktelik Kuralları.....	77
3.6. Diğer Veri Madenciliği Metotları.....	79
3.7. Veri Madenciliği Metotlarının Kullanım.....	81
4. VERİ MADENCİLİĞİNİN TIP ALANINDA KULLANIMI.....	83
4.1. Tıbbi Verilerin Yapısı.....	84
4.2. Tıp Alanında Veri Madenciliği Uygulamaları.....	86
4.2.1. Tıp literatürü üzerinde metin madenciliği uygulamaları.....	86
4.2.2. Hastane bilgi sistemi üzerinde veri madenciliği çalışmaları..	91

	Sayfa
4.3.3. Genetik alanında veri madenciliği çalışmaları.....	93
4.3.4. Eczacılık alanında veri madenciliği çalışmaları.....	95
4.3.5. Tıp alanında kullanılan veri madenciliği tekniklerine göre çalışmalar.....	96
4.3.6. Hastalık tiplerine göre yapılan veri madenciliği çalışmaları..	103
4.3.7. Hemodiyaliz ile ilgili yapılan veri madenciliği çalışmaları....	106
5. BİR UYGULAMA : HEMODİYALİZ HASTALARI İÇİN RİSK SEVİYELERİ- NE GÖRE RİSK FAKTÖRLERİNİN ETKİLEŞİMLERİNİN İNCELENMESİ.....	108
5.1. Problemin Belirlenmesi ve Verilerin Anlaşılması.....	108
5.2. Hastalık Bilgisi (Son aşama kronik böbrek hastalığı ve hemodiyaliz tedavisi.....	110
5.3. Verilerin Toplanması.....	116
5.4. Verilerin Hazırlanması.....	127
5.4.1. Eksik verilerin tamamlanması.....	127
5.4.2. Veri kümesine eklenen özellikler.....	129
5.4.3. Dönüştürülen özellikler.....	130
5.5. Modelin Kurulması.....	133
5.5.1. Metot seçimi.....	135
5.5.2. Yazılım seçimi.....	135
5.6. Uygulama ve Elde Edilen Sonuçlar.....	141
5.6.1. Risk skorunun hesaplanması.....	141
5.6.2. Anahtar özellik için sınıf sayısının tespiti.....	148
5.6.3. Anahtar özellik tahmininde etkili özelliklerin tespiti.....	154
5.6.4. Karar ağaçlarının oluşturulması.....	155

	Sayfa
5.6.5. Elde edilen sonuçlar.....	159
6. SONUÇ VE ÖNERİLER.....	161
6.1. Bulguların Değerlendirilmesi.....	161
6.2. Öneriler.....	165
KAYNAKLAR.....	167
EKLER.....	181
EK-1. Özel Can Diyaliz Merkezi arşiv görünümü.....	182
EK-2. Veri toplamada kullanılan formlar.....	183
EK-3. Veri kümesini oluşturmada kullanılan çizelgeler.....	189
EK-4. CMS HSS ESRD Modeli risk skoru hesaplama çizelgeleri.....	204
EK-5 Uygulama çıktı raporları.....	211
ÖZGEÇMİŞ.....	222

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 2.1 SQL sorgusu sonuç tablosu.....	16
Çizelge 2.2 İşlemsel veri tabanı için temel tablo yapısı.....	18
Çizelge 2.3 Veri yapılarına göre yazılımlar.....	21
Çizelge 2.4 Ticari veri temizleme yazılımları ve üretici şirketler.....	25
Çizelge 2.5 Risk Matrisi.....	29
Çizelge 2.6 Veri madenciliği yazılımlarının genel özellikleri	38
Çizelge 2.7 Veri madenciliği yazılımlarının teknik özellikleri	39
Çizelge 2.8 Bazı veri madenciliği yazılımlarının güçlü ve zayıf nitelikleri.....	40
Çizelge 3.1 Golf örneği için hedef-olmayan nitelikler.....	54
Çizelge 3.2 Golf örneği için öğrenme verisi.....	55
Çizelge 3.3 a ve b niteliklerinin değerleri.....	70
Çizelge 3.4 Veri madenciliği teknikleri ve kullanılabileceği alanlar.....	81
Çizelge 3.5 Farklı tipteki veri madenciliği tekniklerini nitelikleri bakımından karşılaştırması.....	82
Çizelge 4.1 RCC ve ICD-9-CM standartları arasındaki ilişki.....	85
Çizelge 4.2 Sağlık Bilimlerinde kullanılan Metin Madenciliği uygulama araçları.....	90
Çizelge 5.1 Kronik böbrek hasarı evreleri.....	112
Çizelge 5.2 Çalışma kapsamındaki hastaların halen tedavi görme durumlarına göre dağılımı.....	120
Çizelge 5.3 Eksik veri içeren özelliklere ait boş veri oranı bilgisi, boş hücrelere doldurulan veriler ve boş hücrelerin IDleri.....	127
Çizelge 5.4 Eksik hemodiyaliz seansı verileri.....	128

Çizelge	Sayfa
Çizelge 5.5 Veri özellikleri ve bunlara ilişkin bilgiler.....	132
Çizelge 5.6 Nominal hale getirilen özelliklere ilişkin etiketleme.....	138
Çizelge 5.7 CMS HCC ESRD Modeli Demografik Katsayılar.....	143
Çizelge 5.8 CMS HCC ESRD Modeli Klinik Katsayılar	144
Çizelge 5.9 Veri kümesindeki hastalar için kullanılacak Katsayılar	147
Çizelge 5.10 Her sınıf sayısı için ilgili risk seviyesine giren kayıtların nominal etiketlemesi ve sınıflara göre dağılım oranları.....	153
Çizelge 5.11 Her sınıf için elde edilen kümeleme performans sonuçları.	154
Çizelge 5.12 Veri kümesinde yer alan özelliklere ilişkin istatistikler.....	157
Çizelge 5.13 İki Risk seviye sınıflı Risk skoru özelliği için Karar ağacı algoritmalarının sınıflama performansları.....	158

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 1.1 Veri tabanı sistemi teknolojisinin gelişim.....	4
Şekil 2.1 Veri madenciliği ile diğer disiplinler arası ilişki.....	5
Şekil 2.2 Veri madenciliği süreci (geniş anlamda).....	7
Şekil 2.3 Veri madenciliği süreci adımları.....	7
Şekil 2.4 ASCII formatında bir metin dosyası örneği.....	13
Şekil 2.5 Excel hesap tablosu örneği.....	13
Şekil 2.6 Tablo yapısı ve Birincil ve İkincil anahtarların tablolar arası ilişki kurmada rolü.....	15
Şekil 2.7 Veri ambarının tipik çerçevesi.....	16
Şekil 2.8 OLAP yapısı.....	17
Şekil 2.9 Tipik bir veri madenciliği sisteminin yapısı.....	33
Şekil 2.10 Veri madenciliği haritası.....	42
Şekil 3.1 Veri madenciliği metotları.....	45
Şekil 3.2 Doğrusal Regresyon modelinin grafiksel gösterimi.....	49
Şekil 3.3 Çoklu Regresyon modelinin grafiksel gösterimi.....	51
Şekil 3.4 Golf örneği için Karar ağacı örneği.....	55
Şekil 3.5 ID3 Algoritması.....	60
Şekil 3.6 Gerçek ve yapay nöron modeli.....	65
Şekil 3.7 Tipik bir Çok Katmanlı Algılayıcının yapısı.....	66
Şekil 3.8 Bir ÇKA modelinin Girdi-Çıktı hesaplama süreci.....	67
Şekil 3.9 Genetik algoritma süreci.....	72
Şekil 3.10 k-means yöntemiyle kümeleme örneği.....	75
Şekil 3.11 Veri tiplerine göre kullanılan Görselleştirme teknikleri.....	80

Şekil	Sayfa
Şekil 4.1 Tıbbi kılavuzların Kural tabanlı oluşturulması süreci	89
Şekil 5.1 Hemodiyaliz süreci.....	115
Şekil 5.2 Uygulama süreci adımları.....	133
Şekil 5.3 ESTARD Data Miner ekran görünümü.....	136
Şekil 5.4 WEKA Kullanıcı Ara yüzü.....	137
Şekil 5.5 WEKA Explorer bölümünün ekran görünümü.....	138
Şekil 5.6 Veri kümesinin .arff dosya formatında temsili görünümü...	140
Şekil 5.7 ESTARD Data Miner veri değişkenliği oranları ekran görüntüsü I.....	149
Şekil 5.8 ESTARD Data Miner veri değişkenliği oranları ekran görüntüsü II.....	149
Şekil 5.9 ESTARD Data Miner sınıflama için özellik seçimi ekran görüntüsü	150
Şekil 5.10 ESTARD Data Miner RSCORE iki sınıf seviyesi için veri dağılımı.....	151
Şekil 5.11 ESTARD Data Miner RSCORE üç sınıf seviyesi için veri dağılımı	151
Şekil 5.12 ESTARD Data Miner RSCORE dört sınıf seviyesi için veri dağılımı	152
Şekil 5.13 ESTARD Data Miner RSCORE beş sınıf seviyesi için veri dağılımı	152
Şekil 5.14 J4.8 algoritması tarafından oluşturulan karar ağacı.....	158

RESİMLERİN LİSTESİ

Resim		Sayfa
Resim 5.1	Özel Can Diyaliz Merkezi Binası - Giriş.....	117
Resim 5.2	Özel Can Diyaliz Merkezi Binası - Genel görünüm.....	118
Resim 5.3	Fresenius 4008B model Hemodiyaliz makinesi.....	118

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler

Açıklama

ϵ

Hata terimi

σ^2

Varyans

Kısaltmalar

Açıklama

OLAP

Çevrimiçi Analitik İşleme

CHAID

Kikare Otomatik Etkileşim Bulucu

CART

Sınıflandırma ve Regresyon Ağaçları

VKİ

Vücut Kitle İndeksi

NRI

Beslenme Risk İndeksi

CMS HCC ESRD

Diyaliz Hastaları için Risk Değerlendirme Modeli

ÇKA

Çok Katmanlı Algılayıcı

YSA

Yapay Sinir Ağları

HDL

Yüksek Yoğunluklu Lipoprotein

LDL

Düşük Yoğunluklu Lipoprotein

EPO

Eritropoetin

ICD

Uluslar arası Hastalık Sınıflandırması

HCC

Hiyerarşik Bir arada Durumlar

1 GİRİŞ

Günümüz dünyasında teknolojinin hızlı gelişimi ile herhangi bir faaliyetten elde edilen artan miktardaki verilerin toplanması, saklanması ve işlenmesi gibi problemler ortaya çıkmıştır. Bu problemlere gelişen bilgisayar teknolojisi yardımıyla ucuzlayan ve kapasitesi artan donanımlar ve veri tabanları alanındaki gelişmelerle çözüm aranmıştır [1].

Paket programların kullanılması ve barkod okuyucuların yaygınlaşması gibi gelişmeler ile verilerin veri tabanına istenen formatta ve çok daha hızlı girilmesi sağlanmıştır. Verilerin veri tabanların saklanmaya başlanması ile yeryüzündeki bilgi miktarı her 20 ayda bir iki katına çıkmakta ve her geçen gün olağanüstü bir hızla artmaya devam etmektedir [2].

Veri miktarı arttıkça bu verilerin insanlarca anlaşılması zorlaşmış; veriler içerisinde saklı kalmış yararlı bilginin elde edilmesi ve Karar Destek Sistemi çerçevesinde kullanımı, herhangi bir araç kullanmaksızın imkansız hale gelmiştir. Bu probleme bir çözüm olarak “Veri Madenciliği” (Data Mining) kavramı geliştirilmiştir.

Bu çalışmada Veri Madenciliğinin Tıp alanında kullanımı incelenmiş ve bu alanda bir uygulama yapılması amaçlanmıştır.

Uygulama çalışması kapsamında, hemodiyaliz hastalarına ait veriler (demografik veriler, tedavi ve aylık test verileri) ve her hasta için hesaplanacak ölüm riski skoru içeren bir veri kümesi hazırlanacak, bu küme üzerinde veri madenciliği teknikleri kullanılması ile elde edilecek karar ağaçları / birliktelik kuralları uzman desteğiyle incelenerek risk seviyelerine göre desenler elde edilmeye çalışılacaktır.

2 VERİ MADENCİLİĞİNE GENEL BAKIŞ

Shalvi ve DeClaris'e (1998) göre "Veri Madenciliği, belirli bir alanda ve belirli bir amaç için toplanan veriler arasındaki gizli kalmış ilişkilerin (desenlerin, modellerin vb.) ortaya konulmasıdır" [3].

Veri Madenciliğinin daha detaylı bir tanımı şöyledir [4] :

"Veri madenciliği, depolanmış yüksek miktardaki veriden istatistiksel ve matematiksel teknikler gibi desen tanımlayıcı teknolojiler kullanarak anlamlı ve yeni ilişkiler, desenler ve trendler keşfetme prosesidir."

Madencilik terimi, topraktan küçük ve değerli maden parçalarının çıkarılması ile yüksek miktarda veri içinden değerli bilginin çıkarılması arasındaki benzerlik nedeniyle kullanılmıştır [5].

Veriler içinde desen aramak insan hayatının başlamasından beri devam eden eski bir süreçtir. Avcılar hayvanların göç etme davranışlarında, çiftçiler ekin gelişiminde, politikacılar seçmen görüşlerinde tahmin edici modeller aramışlardır. Bilim insanlarının işi ise veriler içinde fiziksel dünyanın nasıl çalıştığını anlatan modeller aramak ve bulacağı modeller yardımıyla ortaya çıkacak yeni durumlarda neler olacağını tahmin edecek teoriler geliştirmektir [6].

Veri tabanı kavramı ve veri tabanı sistemi teknolojisi geliştirilme sürecinde verilerin belli bir amaca yönelik düzenli olarak toplanmasını, saklanmasını ve işlenmesini sağlayarak anlamlı bilgi eldesi için gerekli altyapıyı sağlamıştır.

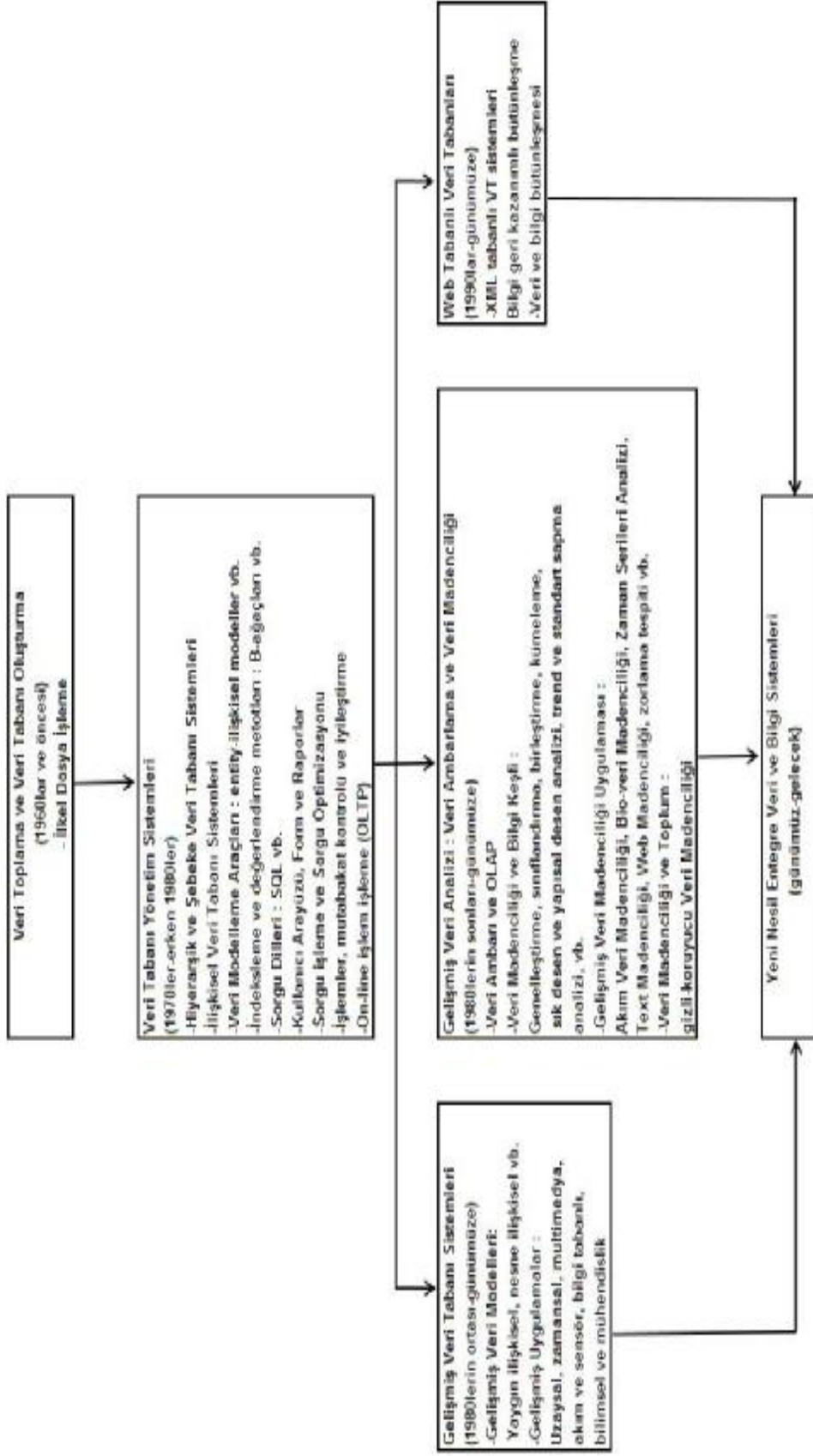
Şekil 1.1'de Veri tabanı sistemi teknolojisinin gelişimi verilmiştir.

1960'lı yıllarda verilerin ilkel dosyalara işlenmesi ile başlayan verileri toplama girişimleri sonraki 20 yıllık süreçte önce verilerin sınıflara ayrılması ardından bu sınıflar arasında ilişkisel bağlantılar kurulması ile veri tabanı kavramının ortaya çıkmasını sağlamıştır. Veri tabanı kavramı ardından kullanıcı ile veri tabanı arasındaki etkileşimi sağlayan arayüzler geliştirilmiş, 1980'lerin

ortasından itibaren çok farklı yapı ve boyutlarda verileri işleyebilecek gelişmiş veri tabanı sistemleri tasarımı çalışmaları günümüzde de devam etmektedir.

Geliştirilen veri tabanı sistemlerinde çok büyük miktar ve boyutlarda veriler toplanması, bu verilerden etkin ve yeni bilgiler eldesinde zorluklara yol açmıştır. 1990'lardan itibaren gelişen İnternet ve bilgisayar teknolojisi devasa boyutlara ulaşan verilerin analizinde işlemci gücü yardımıyla yeni matematiksel tekniklerin kullanılabilmesini dolayısıyla veri analizinin kolaylaşmasını sağlamıştır. Veri analizi ile çok miktarda bilgi eldesi bu bilgilerin ne kadar anlamlı olduğu sorusuna önem kazandırmıştır. Veri Madenciliği kavramı bu ortamda ortaya atılmış ve bu başlık altında anlamlı bilgi eldesi için büyük boyutlu ve çok farklı yapıdaki veri kümelerinde desen araştırma teknikleri geliştirilmiştir.

Özetle veri madenciliği kavramının gelişimi ve bu konudaki uygulamalarının yaygınlaşması bilgi teknolojisinin gelişiminin bir sonucudur [5].



Şekil 1.1 Veri tabanı sistemi teknolojisinin gelişimi [5]

2.1. Veri Madenciliği ile İlişkili Bilim Dalları

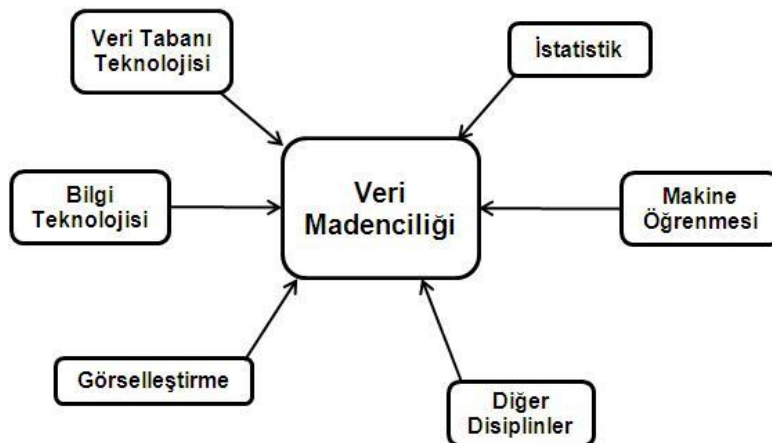
Veri Madenciliği, Veri tabanı Sistemleri, İstatistik, Makine Öğretimi, İnsan-Makine Etkileşimi, Veri Görselleştirme gibi farklı disiplinlerden faydalanan disiplinlerarası bir alandır.

Kullanılan Veri Madenciliği yaklaşımına bağlı olarak, Veri Madenciliği'nde farklı disiplinlerde kullanılan tekniklerden de yararlanılmaktadır. Bu tekniklere örnek olarak Sinirsel Ağlar, Bulanık Küme Teorisi, Tümevarımsal Mantık Programlama, yüksek performanslı hesaplamalar verilebilir.

İşlenen verinin yapısına bağlı olarak Veri Madenciliği, Uzaysal Veri Analizi, Resim Analizi, Sinyal İşleme, Bilgisayar Grafikleri, Web teknolojisi vb. tekniklerden de faydalanmaktadır.

Farklı disiplinlerden elde edilen veriler üzerinde yapılan Veri madenciliği uygulamalarında farklı modeller kurulması gerekliliğini ortaya çıkarmıştır. Veri madenciliği için farklı modelleme gerektiren disiplinlere Ekonomi ve Biyoinformatik gibi disiplinler örnek verilebilir.

Şekil 2.1'de Veri Madenciliği'nin Diğer disiplinlerle ilişkisi verilmiştir.



Şekil 2.1 Veri madenciliği ile diğer disiplinler arası ilişki [5]

2.2. Veri Madenciliği Süreci

Dar anlamıyla Veri Madenciliği Verilerden Bilgi Keşfi (Knowledge Discoveri from Data – KDD) sürecinin içinde bir adımdır. Bu yaklaşıma göre Verilerden Bilgi Keşfi Süreci birbirini takip eden aşağıdaki adımlardan oluşur:

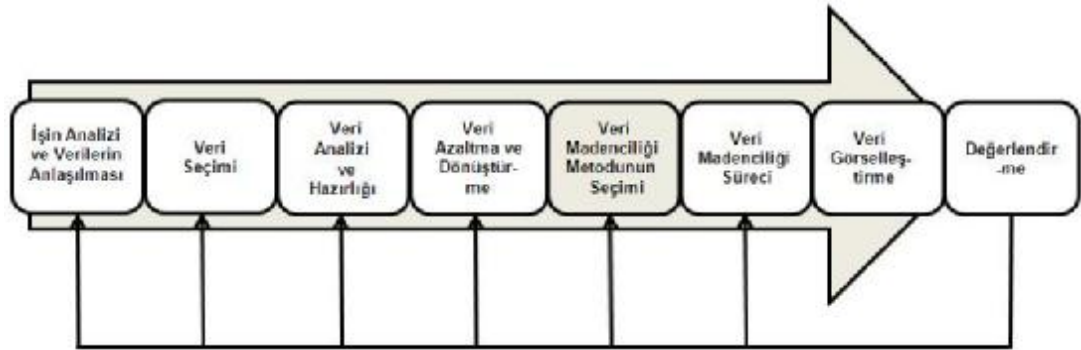
1. *Verilerin Temizlenmesi* (Tutarsız verilerin çıkarılması)
2. *Verilerin Bütünleştirilmesi* (Farklı kaynaklardan elde edilen verilerin bir araya getirilmesi)
3. *Verilerin Seçilmesi* (Veri tabanından analiz yapılacak verilerin seçilmesi)
4. *Verilerin Dönüştürülmesi* (Verilerin madencilik tekniklerinin gerektirdiği uygun yapıya getirmesi)
5. *Veri Madenciliği* (Veri desenleri elde etmek için akıllı tekniklerin kullanıldığı temel işlem)
6. *Desen Değerlendirme* (İlginçlik ölçütlerine göre gerçek ilginç desenlerin tanımlanması)
7. *Bilgi Sunumu* (Uygulama sonuçlarının görselleştirme ve betimleme teknikleri yardımıyla kullanıcıya sunumu) [5].

Geniş anlamda ele alındığında Veri Madenciliği daha karmaşık bir süreç olarak görülebilir ve daha detaylı bir bölümlenme ile aşağıdaki adımlardan oluşur :

- i) İşin (Problemin) Analizi ve Verilerin Anlaşılması
- ii) Veri Seçimi
- iii) Veri Analizi ve Hazırlığı
- iv) Veri Azaltma ve Dönüştürme
- v) Önemli Özelliklerin (Attributes) Seçimi
- vi) Veri Kümesi Boyutunun Küçültülmesi
- vii) Normalleştirme
- viii) Birleştirme
- ix) Veri Madenciliği Metodunun Seçimi

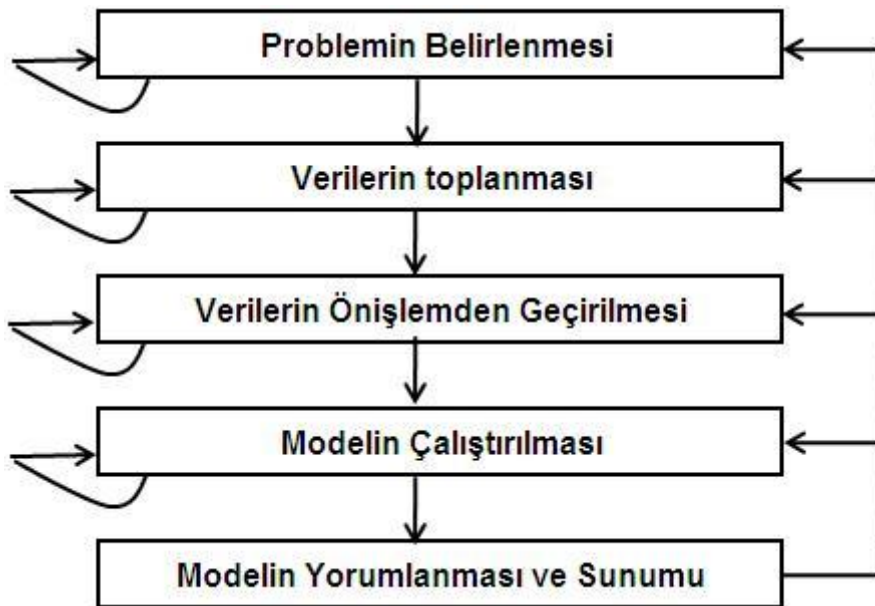
- x) Veri Madenciliği Süreci
- xi) Görselleştirme
- xii) Değerlendirme
- xiii) Bilgi Kullanımı ve Sonuçların Hedefe Uygun olarak Değerlendirilmesi

Şekil 2.2'de Veri madenciliği sürecinin geniş anlamda adımları gösterilmiştir.



Şekil 2.2 Veri madenciliği süreci (geniş anlamda)

Bu iki yaklaşım da dikkate alınarak Veri Madenciliği süreci genel deneysel prosedür olarak Şekil 2.3'te belirtilen 5 adımdan oluşur:



Şekil 2.3 Veri madenciliği süreci adımları [7].

2.2.1. İşin (Problemin) belirlenmesi ve verilerin anlaşılması

Veri Madenciliği uygulamasının ilk ve temel aşaması işin (problemin) belirlenmesi ve verilerin anlaşılmasıdır.

Bu adımda işin (problemin) bilinmeyen bağımlılıklara göre değişkenleri belirlenir ve sonuçta bir model kurmak için veriler arası ilişkilerden hipotez ya da hipotezler oluşturulmaya çalışılır.

Veri Madenciliğini uygulamasının başarısında işin (problemin) iyi tespit edilmesi, uygulamada hangi veri madenciliği tekniğinin seçileceği kararından daha önemlidir. Bu nedenle belirtilen 10 Altın Kurala uyulması işin (problemin) belirlenmesi ve verilerin anlaşılması adımının en iyi şekilde gerçekleştirilmesini sağlayacaktır :

- i) İş (problem) elle tutulur faydalar sağlayacak şekilde açıkça tanımlanmalı.
- ii) Muhtemel sonuç (beklenti) belirlenmeli
- iii) Elde edilecek sonucun nasıl kullanılacağı belirlenmeli.
- iv) İş (problem) ve veriler olabildiğince anlaşılmalı.
- v) İş (problem) modele dönüştürülmeli
- vi) Varsayımlar belirlenmeli.
- vii) Model döngüsel olarak iyileştirilmeli.
- viii) Model mümkün olan en basit (kolay anlaşılır) hale getirilmeli.
- ix) Modelin kararsızlığı (girdilerdeki küçük bir değişimin çıktılar üzerinde çok etkili olduğu kritik alanları) tanımlanmalı.
- x) Modelin belirsizliği (veri kümesinde modelin düşük güvenilirlikte çıktılar ürettiği kritik alan ve aralıklar) tanımlanmalı.

Bu kurallardan bazıları (iv-v-vi-vii-x) verilerin anlaşılması ile ilgili olduğu kadar verilerin hazırlanması ile de ilgilidir.

Modelin döngüsel iyileştirilmesi, verilerin hazırlanması, modelin kurulması, çıkacak sonuçların değerlendirilmesi ve buna göre modelde değişiklik yapılması şeklinde ifade edilebilir. Bu kuralın uygulanması sırasında modeldeki değişimlere bağlı olarak veri kümesinin de değişecektir [8].

İşin (problemin) ve verilerin niteliğine bağlı olarak ilgili alanda uzman desteğine ihtiyaç olabilir. Bu durumda ilgili alanda bir uzman ile veri madenciliği uzmanının işbirliği gereklidir. Başarılı Veri Madenciliği uygulamalarında bu işbirliğinin sadece bu adımda değil bütün süreç boyunca sürdüğü görülür [7].

2.2.2. Verilerin toplanması

Bu adım verilerin nasıl oluşturulacağı ve toplanacağı ile ilgilidir.

Verilerin oluşturulması sürecinde İki farklı yaklaşım vardır. Verilerin oluşturulması süreci uzman (modelci) kontrolünde yapılırsa Tasarlanmış Deney; uzman kontrolü olmadan yapılırsa Gözlemsel Yaklaşım olarak adlandırılır. Çoğu veri madenciliği uygulamasında veri oluşturma süreci Gözlemsel yaklaşım ile yapılır.

Modelin kurulmasında, test edilmesinde ve uygulanmasında kullanılan verilerin aynı bilinmeyen örnekleme dağılımından gelmesi önemlidir.

Verilerin oluşturulması ile ilgili olarak Veri Yapısı, Verilerin toplanması ile ilgili olarak Veri Kaynakları ve adımın geneli ile ilgili olarak Veri Kalitesi ayrı başlıklar altında incelenecektir.

Veri yapısı

Veri Madenciliği uygulaması yapabilmek için verilerin aşağıdaki şartları sağlamalıdır :

- * Tüm veri tek bir tabloda yer almalıdır. (Bu tabloya Veri Kümesi (Data Set) denir.)
- * Her satır bir bağımsız varlığa (entity) ait olmalıdır. Veri kümesi bir iş ile ilgili ise her satır bir müşteriye ait olması, bu çalışmadaki uygulamada ise her satır bir hemodiyaliz hastasına ait olması gibi.
- * Tek değere sahip sütunlar çıkarılmalıdır.
- * Her sütun için farklı değere sahip sütunlar çıkarılmalıdır.
- * Tahmin edici modellemede hedef sütun tanımlanmalı ve tüm eşdeğer sütunlar veri kümesinden çıkarılmalıdır [9].

Sütunlarda yer alan veri niteliklerinden her birine Veri Madenciliği terminolojisinde Özellik (Attribute) denir. Bundan sonraki kısımda veri nitelikleri için özellik tabiri kullanılacaktır.

Her özellik taşıdığı niteliğin adının kısaltması ile adlandırılır.

Her özellik aynı zamanda birer değişkendir. Değişkenler veri yapılarına göre Kesikli (Discrete) değişkenler ve Sürekli (Continuous) değişkenler olmak üzere ikiye ayrılır :

a) *Kesikli (Discrete) Değişkenler* kendi içinde Nominal (Kategorik) değişkenler ve Ordinal (Sıralı) değişkenler olarak ikiye ayrılır.

Nominal (Kategorik) Değişkenler, aralarında sıra, ölçü ve uzaklık ölçüsü olmayan veriler için kullanılır. Örneğin Cinsiyet özelliği için Erkek veya Kadın, Hastalık özelliği için Sarılık, Kızamık, Kolera vb. gibi.

Nominal değişkenlerin rakamlarla gösterimi (1=Erkek, 2=Kadın vb.) bazı sayısal algortmaların uygulanmasından hatalara yol açtığı için tercih edilmez. Bunun yerine nominal değişkenlerin gösteriminde en uygun yol kısa adlarla gösterilmeleridir. Örnek : E=Erkek, K=Kadın gibi.

Ordinal (Sıralı) Değişkenler, aralarında sıra ölçüsü olan veriler için kullanılır. Örneğin kredi kartı ödemesi için risk seviyesi özelliği için Düşük, Orta, Yüksek risk seviyesi. Yüksek-Orta-Düşük arasında sıralı bir ilişki vardır ancak ölçüsel bir ilişki yoktur. Bu tip değişkenlerin de rakamlarla gösterimi uygun değildir.

Bazı ordinal değişkenler ölçü ilişkili olarak gelebilir. Genellikle bir olayla ilgili frekans sıklığını belirten değişkenlerdir. Örneğin bir müşterinin aylık kredi kartı kullanma sıklığı için 0-5, 6-10, 11-15 vb. Bu durumda en iyi gösterim sıralı rakamsal gösterimdir. (0-5=1, 6-10=2, 11-15=3 vb.) Ancak bu durum iki nedenle veri madenciliği algoritmaları için zararlı olabilir : (1) verinin dağınık yapısını azaltarak tahmin modelinin doğruluğunu azaltır ve (2) ölçülen orijinal büyüklüğün ordinal yapısını değiştirir.

b) Sürekli (Continuous) Değişkenler, aralarında sıra, ölçü, uzaklık gibi tüm ilişkilere sahip veriler için kullanılır. Kullanımı ve yorumlanması en kolay değişkenlerdir. Özelliğe göre pozitif, negatif ya da 0 değeri alabilirler ve ondalıklı şekilde yazılabilirler.

Özel bir durum olarak zaman değerleri kendi özgün yapılarında kullanılamaz. Veri madenciliğinde zaman verilerini kullanabilmek için en etkin yol bu değerleri zaman periyodu ölçüsü şeklinde dönüştürmektir. Bu dönüşüm tarih veya zaman değeri ile sabit bir referans değeri arasındaki fark bulunarak gerçekleştirilir [6].

Kullanıcı topladığı verilerin hangi değişken şeklinde ifade edileceğini tercih edebilir. Sürekli değişkenlerin kategorik değişken haline dönüştürmek mümkündür. Bu dönüştürme işlemi, eldeki verilerin kategorik anlamının önemine, özellik verilerinin çeşitliliğine bağlı olarak tercih edilir.

Veri kaynakları

Veriler, sırasıyla ilk tasarım, toplama ve değerlendirme, depolama ve yönetim, işleme ve uygulama ve son olarak kullanım adımlarından oluşan bir hayat çevrimine sahiptir. Depolama ve yönetim, Veri Madenciliği açısından bu hayat çevriminin en önemli adımıdır.

Veri Madenciliği uygulaması için veri sağlayan kaynaklar, farklı depolama ve yönetim yaklaşımları sonucunda tarihsel olarak basit metin dosyalarından (text files) veri tabanlarına, veri ambarlarından gelişmiş veri tabanı sistemlerine doğru gelişim göstermiştir. Gelişim adımları alt başlıklar halinde kısaca tanıtılmıştır.

a) Metin Dosyaları

Metin dosyaları için en yaygın format Amerikan Ulusal Standartlar Enstitüsü (ANSI) tarafından oluşturulan ASCII (American Standard Code for Information Interchange)'dir. Alfabetik, numerik ve özel 128 karakterden oluşur. Her karakter 7 bitlik ikili anahtar ile kodlanır. (7 adet 0 veya 1'li kod) Unix ve DOS tabanlı işletim sistemleri metin dosyaları için ASCII kodunu kullanır.

Windows NT ve daha gelişmiş Windows XP gibi işletim sistemleri ASCII formatının gelişmiş versiyonu (16 bitlik ikili anahtar) olan Unicode formatını kullanır.

Bir başka metin dosyası formatı ise IBM tarafından OS/390 gibi büyük işletim sistemleri için geliştirilen EBCDIC (Extended Binary-Coded Decimal Interchange Code) formatıdır. Bu format 8 bitlik ikili anahtar ile 256 karakteri kodlayabilmektedir. Şekil 2.4'te ASCII formatında bir metin dosyası örneği verilmiştir.

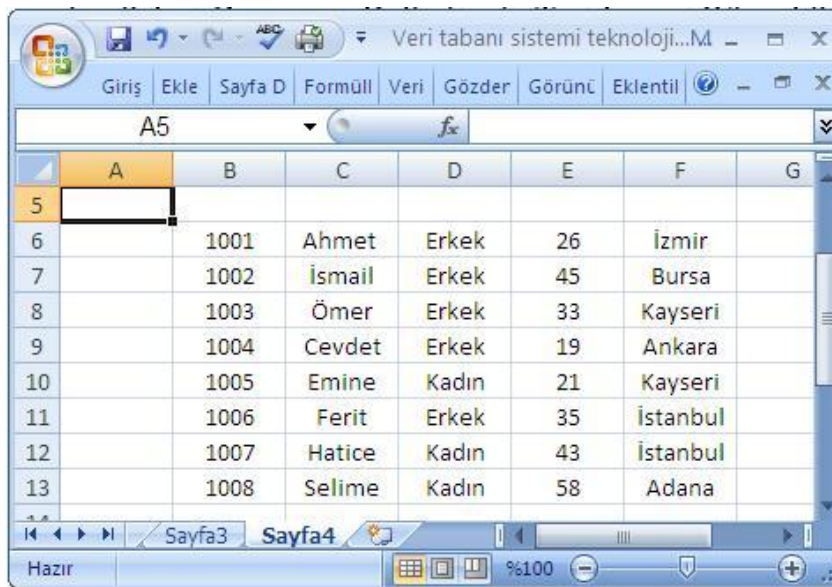


Dosya	Düzen	Biçim	Görünüm	Yardım
1001	Ahmet	Erkek	26	İzmir
1002	İsmail	Erkek	45	Bursa
1003	Ömer	Erkek	33	Kayseri
1004	Cevdet	Erkek	19	Ankara
1005	Emine	Kadın	21	Kayseri
1006	Ferit	Erkek	35	İstanbul
1007	Hatice	Kadın	43	İstanbul
1008	Selime	Kadın	58	Adana

Şekil 2.4 ASCII formatında bir metin dosyası örneği

b) Hesap tablosu (Spreadsheet) Dosyaları

Bir başka veri depolama formatı ise hesap tablosu (Spreadsheet)'dur. 1961'de Prof. Richard Mattessich tarafından bilgisayara uyarlanmıştır. Hesap tablosu formatında veriler sütun ve satırlardan oluşan iki boyutlu tablo formatında depolanırlar. Veriler üzerinde matematiksel ve istatistiksel tekniklerin kullanımı ile metin dosyasından daha çok işlem yapılabilir. Grafik desteği yardımıyla verilerin sunum olanakları da oldukça gelişmiştir. Şekil 2.5'te bir Excel hesap tablosu örneği verilmiştir.



	A	B	C	D	E	F	G
5							
6		1001	Ahmet	Erkek	26	İzmir	
7		1002	İsmail	Erkek	45	Bursa	
8		1003	Ömer	Erkek	33	Kayseri	
9		1004	Cevdet	Erkek	19	Ankara	
10		1005	Emine	Kadın	21	Kayseri	
11		1006	Ferit	Erkek	35	İstanbul	
12		1007	Hatice	Kadın	43	İstanbul	
13		1008	Selime	Kadın	58	Adana	

Şekil 2.5 Excel hesap tablosu örneği

İlk hesap tablosu yazılımı olan VisiCalc 1978'de geliştirilmiş, ardından 1983'te Lotus 1-2-3 piyasaya çıkmıştır. 1987'de ise gelişmiş grafiksel olanakları ile Microsoft Excel ortaya çıkmıştır. Günümüzde Excel, Appleworks, FileMaker, Corel Quattro Pro gibi yazılımlar kullanılmaktadır.

Bu basit veri depolama formatlarının ardından Veri tabanı sistemlerinin gelişmesi ile Veri Madenciliği uygulamasının önü açılmış, büyük miktarda veriye erişim probleminin ortadan kalkması ile veriden bilgi eldesi potansiyeli hızla artmıştır.

c) Veri Tabanları

Veri tabanı sistemleri, Veri tabanı dosyaları ve Veri tabanı yönetim sisteminden oluşur. Veri tabanı dosyaları, orijinal veriler, indisler, veritabanı mantıksal verilerini içerir. Veri tabanı yönetim sistemi'nin görevi verilerin yönetimi ve işlem performansını artırmaktır. Veri tabanı yönetim sisteminin araçları, Veri tabanı yöneticisi, İşlem Müdürü ve Sorgu İşlemcisi'nden oluşur. Veri tabanı kullanıcıdan gelen komutları bu araçlar yardımıyla alır [8].

Veri tabanı sistemlerinin kullanılmaya başlanmasından sonra Veri Madenciliği için verilerin elde edildiği kaynaklar dört ana başlık altında incelenebilir :

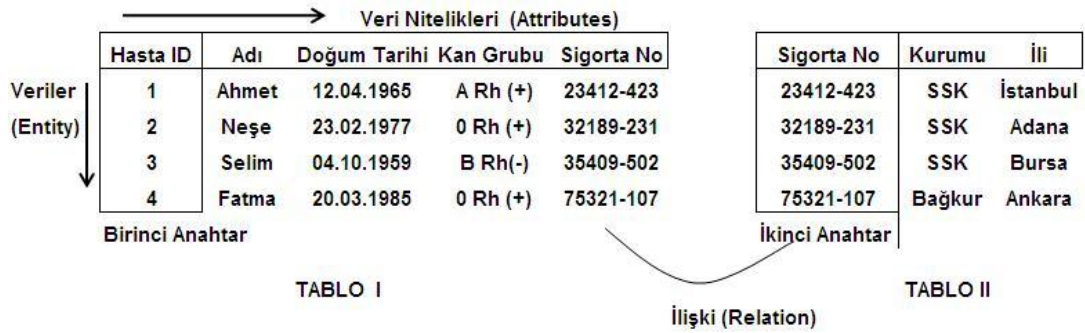
i) İlişkisel veri tabanları :

İlişkisel veri tabanları, ilk kez 1970'te Ted Codd tarafından veri tabanı sistemlerinin verileri birbirleri ile olan ilişkilerini dikkate alarak tablo şeklinde göstermesi önerisinin geliştirilmesi ile ortaya çıkmıştır.

İlişkisel Veri tabanı sistemlerinde kullanılan ilişkisel model, birimler (Entities), özellikler (attributes) ve birimler arasındaki ilişki olmak üzere üç ana bileşenden oluşur. Bir birim, varolan ve ayırt edilen herhangi bir şey olarak tanımlanabilir. Bu birim kayıt edilerek bir veri birimine dönüşür. Veri

birimlerinin özellikleri (attributes) vardır. İlişki bir ya da daha çok veri kümesi arasındaki bağlantıdır. Tablosal gösterimde veri birimleri yatay düzlemde sıralanan özelliklere dikey düzlemde kaydedilir. Böylece veri sütunları oluşur. Tablo yapısında her bir kaydı tek olarak tanımlamak için kullanılan alana Birincil Anahtar (Primary key) adı verilir. Bir başka veri tablosunun ayırt edici alanına ise İkincil ya da Yabancı Anahtar adı verilir. Böylece ilişkisel veri tabanında birden çok tablo arasında ilişkisel bağlantı kurulabilmektedir.

Şekil 2.6'da Örnekte bir tablo yapısı ve Birincil ve İkincil anahtarların tablolar arası ilişki kurmada rolü görülmektedir.



Şekil 2.6 Tablo yapısı ve Birincil ve İkincil anahtarların tablolar arası ilişki kurmada rolü

İlişkisel bir veri tabanında tüm veriler tablolarda depolanır ve veriler ile ilgili tüm işlemler tablolar üzerinde yapılır ve işlem sonuçları tablo şeklinde sonuçlanır. İlişkisel tabloları kullanan çeşitli diller geliştirilmiştir. En çok kullanılan dil SQL (Structured Query Language , Standart Sorgu Dili) dir. SQL dili kullanılarak verilerin güncellenmesinden veri tabanından veri istemeye kadar pek çok işlem basit komut satırları ile yapılabilir.

SELECT komutu ile istenilen tablodan verilen özelliklere uygun veri istenebilir. Örneğin; Sigorta No 32189-231 olan hastanın adı ve doğum tarihini isteyelim.

```
SELECT Adı, Doğum_Tarihi
FROM TABLO I
WHERE Sigorta_No = "32189-231"
```

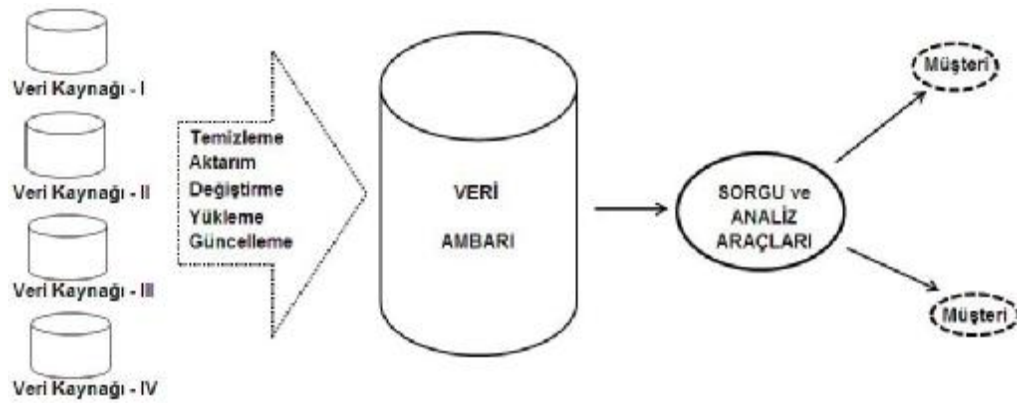
Sonuç Çizelge 2.1'deki şekilde gelecektir.

Çizelge 2.1 SQL sorgusu sonuç tablosu

Sigorta No	Adı	Doğum Tarihi
32189-231	Neşe	23.02.1977

ii) Veri ambarları :

Veri Ambarı, birçok kaynaktan toplanan verilerin birleştirilmiş tek bir şemada depolanması ile oluşan bilgi depolarıdır. Veri ambarları, veri temizleme, veri aktarımı, veri birleştirme, veri yükleme ve periyodik olarak veri güncelleme prosesleri ile inşa edilir. Şekil 2.7'de veri ambarının tipik çerçevesi verilmiştir.



Şekil 2.7 Veri ambarının tipik çerçevesi [5]

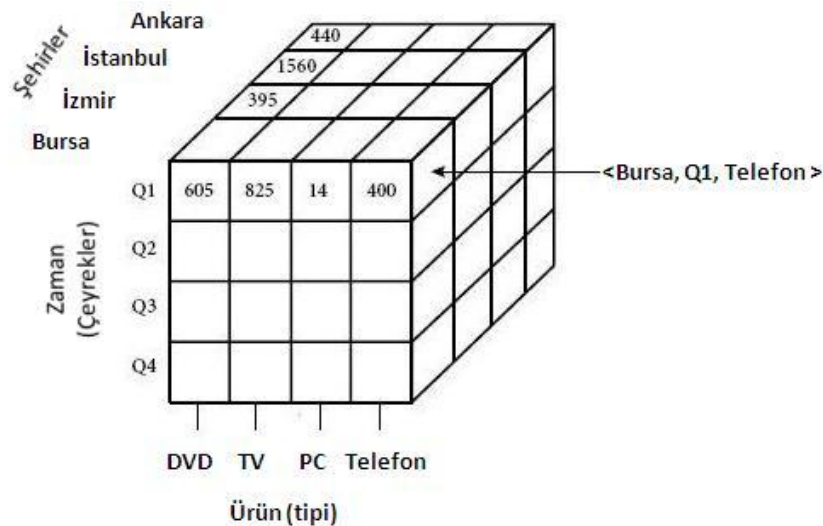
Karar vermeye yardımcı olmak için veri ambarındaki veriler müşteri, tedarikçi, aktivite, parça gibi belli konular etrafında organize edilir. Tarihsel bir

perspektif sağlamak için veriler belli bir periyodu kapsayacak şekilde (son 5-10 yıl vb.) depolanır ve özetlenir.

Veri ambarı genellikle çok boyutlu veri tabanı yapısında modellenir. Her boyut bir özelliğe veya özellikler kümesine bağlanır ve her hücre bir veri değeri içerir. Bir veri ambarının fiziksel yapısı yukarıda belirtilen ilişkisel veri deposu şeklinde veya çok boyutlu veri küpü şeklinde olabilir.

Çok boyutlu veri görünümüne ve özetlenmiş verinin işlenmesine olanak tanınması sebebiyle veri ambarları, Çevrimiçi Analitik İşleme (On-line Analytical Processing, OLAP) için çok uygundur. OLAP çok boyutlu veri tabanında veri analizini destekleyen sorgu bazlı metottur. OLAP'ta veri, çok boyutlu bir uzay üzerinde tanımlanan ve bir çok boyutları olan küpler biçiminde gösterilir. Her bir boyut bir araya toplanmış bir kümeden oluşmaktadır. OLAP ile çok boyutlu veriler içerisinde derinlemesine farklı boyut analizlerinin yapılması sağlanmaktadır.

Şekil 2.8'de Çok boyutlu veri küpü örneği yer almaktadır.



Şekil 2.8 OLAP yapısı [6]

Veri Madenciliği, istatistiksel teknikler gibi geleneksel teknikler dışında veri ambarlamayı kapsayan çeşitli veri tabanı teknolojilerini ve bu veri tabanları

üzerinde çalışan OLAP (Çevrimiçi Analitik İşleme) gibi çok boyutlu metot, yazılım paketleri ve sorgu dillerinin geliştirilmesi konuları da kapsar.

iii) İşlemsel veri tabanları :

İşlemsel veri tabanları her kaydın bir işlemi temsil ettiği veri dosyalarından oluşur. Her işlemin içeriğini oluşturan bir madde listesi vardır. Bu tip veri tabanları yapılan her işlemi içeriği olan liste ile birlikte kaydedilmesini gerektirir. Depolanan veriler kullanılarak Müşteri Sepet analizi yardımıyla maksimum işlem sayısına ulaşacak liste içeriği belirlenmeye çalışılır. Çizelge 2.2'de işlemsel veri tabanı için temel tablo yapısı gösterilmiştir.

Çizelge 2.2 İşlemsel veri tabanı için temel tablo yapısı [5]

İşlem ID	Madde Listesi
T100	I1,I3,I8,I16
T200	I2,I8
*****	****

iv) Gelişmiş veri tabanı sistemleri :

Kullanılan verilerin yapısal olarak gelişimine paralel olarak bu veri yapılarını kullanan uygulamaların karmaşık veri yapılarını en etkili şekilde depolayabilmesi, en uygun metotlarla işlemesi, farklı şekillerde gösterim imkanları sunabilmesi gerekmektedir. Bu ihtiyaçlara cevap vermek üzere gelişmiş veri tabanı sistemleri geliştirilmiştir.

Veri yapısına bağlı olarak özelleşen gelişmiş veri tabanı sistemleri altı ana başlıkta kısaca özetlenmiştir :

(I) Nesne İlişkisel veri tabanları (Object-relational databases)

Amaca yönelik veri tabanlarının temel yapısı model alınarak her veri birimi bir nesne kabul edilir. Bu nesneler bireysel olarak çalışanlar, müşteriler veya cisimler olabilir. Nesneler her birini tanımlayan bir değişkenler kümesine, diğer nesnelerle veya veri tabanı sistemi ile iletişim kurabilmesini sağlayan bir mesaj kümesine, her mesajı uygulamayı sağlayan kodu içeren bir metotlar kümesine sahiptir. Farklı özellikte nesneler nesne sınıfı oluşturacak şekilde hiyerarşik olarak gruplanabilir. Böylece gelişmiş bir veri ilişkilendirme, depolama ve paylaşım veri tabanı ortaya çıkmaktadır.

(II) Periyodik ya da Zaman Serili veri tabanları

Zamana bağlı olan verilerin depolandığı veri tabanlarıdır. Borsa, stok kontrolü ve doğal olayların gözlem verilerinin (sıcaklık, rüzgar vb.) toplanmasında kullanılır.

(III) Uzaysal (Spatial) veri tabanları

Uzay boyutu içeren verilerin depolanması için geliştirilmiştir. Grafik haritalar, çok büyük boyutlu ya da bilgisayar destekli tasarım verileri, tıbbi görüntüleme teknikleriyle elde edilen veriler, uydu fotoğrafları vb. verilerin toplanması, depolanması ve işlenmesinde bu tip gelişmiş veri tabanları kullanılır.

(IV) Metinsel veri tabanları

Kelimelerin nesne olarak kabul edildiği sistemlerdir. Kelime olarak tanımlanan metin uzun cümleler, paragraflar, ürün spesifikasyonları, hata ya da virüs raporları, uyarı mesajları, özet raporlar, notlar ya da farklı tip dokümanlar olabilir. Metinsel veri tabanları World Wide Web gibi yüksek nitelikte düzensiz olabileceği gibi email mesajları ya da pek çok HTML/XML web sayfası gibi

yarı düzenli olabilir. İnternet arama motorları bu tip veri tabanları üzerinde çalışan bir uygulamadır.

(V) Multimedya veri tabanları

Multimedya teknolojisi ile internetin gelişimiyle büyük bir ihtiyaç haline gelen multimedya verilerinin (görüntü, ses ve video) depolanması için geliştirilmiştir.

(VI) Akışkan veriler için veri tabanları

Dinamik olarak bir ortamda ya da ekranda akan ve değişen bir veri türü olan akışkan veriler (stream data) için geliştirilmiştir. Bu tür veriler dev ya da muhtemelen sonsuz hacimde, dinamik olarak değişken, sabit bir düzende içeri ya da dışarı doğru akan, tek ya da küçük miktarda taramaya elverişli ve hızlı cevap zamanı isteyen kendine özgü özelliklere sahiptir. Bilimsel veriler, mühendislik verileri, dinamik çevrelerden elde edilen zaman serisi şeklinde veriler bu yapıdaki verilerdir. Ağ trafiği, güç kaynağı, borsa, telekomünikasyon, video akımı ve hava-çevre görüntüleme verileri bu tip verilere örnektir [5].

Veriler düzenlenmesinde kullanılan ve bu bölümde anlatılan veri kaynakları olan metin dosyaları, hesap tablosu dosyaları ve veri tabanlarında kullanılan yazılımlar Çizelge 2.3'te özetlenmiştir.

Veri kalitesi

Veri tabanları ve veri ambarları sayesinde verilere toplu halde ulaşmak kolaylaşsa da erişilen verilerin kalitesi, veri madenciliği algoritmalarının hemen uygulanmasına engel olmaktadır.

Veri kalitesi, işin (problemin) görünümü ve buradan hareketle kurulan model üzerinde son derece etkilidir. Ayrıca çıkacak sonuçları olduğu kadar kullanıcıların alacağı kararları da etkilemektedir.

Verilerin kalitesinin ölçümü için veri madenciliği sürecinin erken adımlarında (ikinci veya üçüncü adım) bir analiz gerçekleştirilmesi gereklidir. Bu analiz veri miktarının yüksekliği nedeniyle elle yapılamaz.

Çizelge 2.3 Veri yapılarına göre yazılımlar [8]

Kategori	Yazılım	Özellikler
Metin Dosyaları	Notepad	Basit doküman oluşturmada kullanılır. Dosyalar farklı karakter kümelerinden oluşan Unicode, ANSI ve UTF-8 formatlarında kaydedilebilir.
	UltraEdit-32	Sütun modu, DOS formatından UNIX dosya tiplerine dönüşüm yapılabilen, HTML destekli bir metin editörüdür.
Spread sheet	Microsoft Excel	Spreadsheet pazarının %90'ını ele geçiren Microsoft Office paket programı
	Lotus 1-2-3	IBM DB2 ve Oracle gibi veri tabanlarında kullanılan Excel öncesi spreadsheet paket programı
	Quattro Pro	Borland International tarafından geliştirilen, önce Novell ardından Corel firmalarına satılan Lotus 1-2-3'ün gelişmiş versiyonudur. Excel sonrasında yeni versiyonları çıkmıştır.
Veri Tabanı	Microsoft SQL	Web standartlarında veri tabanı programlama özelliklerine sahip ilk nesil bir sorgu dilidir.
	Oracle 9i	Client/Server tipi veri tabanı yönetim sistemi XML desteği yanında gelişmiş OLAP desteği ilaveli Windows ve Linux işletim sistemleri ile uyumlu hale getirilmiş bir yazılım
	IBM DB2	e-iş için tasarlanmış ilk çok platformlu multimedya veri tabanı
OLAP	Microsoft OLAP	Microsoft SQL Server tarafından sağlanan bir hizmet yazılımı
	Oracle Discoverer	Oracle tarafından geliştirilen OLAP yazılımı
Veri Ambarı	SAP	Excel ve Web sunucularını raporlama aracı olarak kullanabilen pek çok hazır analiz modeli içeren, hızlı sorgu kabiliyetli coğrafi bilgi sistemi uyumlu veri ambarı yazılımı
	SAS	Gelişmiş veri çıkarım, değişim ve yükleme işlemlerine sahip, veri kalitesini ölçebilen bir veri ambarlama yazılımı

Veri kalitesini analize yardımcı bazı göstergeler aşağıda verilmiştir :

- 1) Veri doğru olmalıdır. Analiz yapan veri niteliğinin adının doğruluğunu, verilen kodun doğru aralık için verildiğini, veri değerinin doğru ve tam olduğunu vb.kontrol etmelidir.
- 2) Veri mutlaka ilgili veri tipinde depolanmış olmalıdır. Analiz yapan sayısal bir değer karakter formunda olup olmadığını, tam sayıların reel sayı şeklinde olup olmadığını vb. kontrol etmelidir.
- 3) Veri güvenilir olmalıdır. Farklı kullanıcılar arası anlaşmazlık nedeniyle güncellemeler kaybedilmemeli, veriler Veri Tabanı Sistemi içinde değil ise sağlam yedekleme ve kurtarma prosedürleri uygulanmalıdır.
- 4) Veri tutarlı olmalıdır. Veriler farklı kaynaklardan toplandıktan sonra veri kümesinde tutarsızlık olmamalı, tek bir form ve başlıkta olmalıdır.
- 5) Veri gereksiz olmamalıdır. Gereksiz veri en azlanmalı, tekrar eden veriler kontrol edilip çıkarılmalıdır.
- 6) Veri güncel olmalıdır. Verinin zaman bileşeninden, verileri kontrol ederek (kapalı olarak) ya da verilerin toplanma şekline bakarak (açık olarak) haberdar olunmalıdır.
- 7) Veri çok iyi anlaşılmış olmalıdır. İsimler anlaşılmanın bir gösteresidir. Kullanıcı verinin modelle olan ilgisini iyi bilmelidir.
- 8) Veri kümesi tam olmalıdır. Gerçekte rastlanan eksik (kayıp) veri miktarı minimize edilmelidir. Eksik (kayıp) veri, modelin kalitesini düşürse de bazı gelişmiş teknikler kayıp verilere rağmen geçerli veri madenciliği uygulaması yapabilmektedir [7].

2.2.3. Verilerin hazırlanması

Verilerin hazırlanması, veri madenciliği sürecinde zamanın %60 ila %90'ını alsa da sürecin başarılı olmasında %75 ila %90 arası bir katkıya sahiptir. Veri kümesinin zayıf veya var olmayan verilerden hazırlanması sürecin başarısızlığında %100 sorumludur [8].

Veri madenciliği sürecinin bu önemli adımı ya veri hazırlama konusunda beceri sahibi olunarak ya da bu beceri iyi belirlenmiş bir prensipler kümesini takip suretiyle kazanılarak gerçekleştirilebilir.

Verilerin hazırlaması adımı veri madenciliği sürecinin diğer adımlarından tamamen bağımsız değildir. Sürecin her adımında yapılan tüm işlemler birlikte yeni ve daha gelişmiş bir veri kümesi elde etmeye yardımcı olur [7]. Verilerin hazırlanması, gerçek hayattan toplanan eksik, geniş dağılımlı, çelişkili verilerin veri madenciliği uygulaması yapılabilecek bir veri kümesi haline getirilmesi için yapılan verilerin temizlenmesi, verilerin kaynaştırılması ve dönüşümü, veri hacminin küçültülmesi, verilerin kesikli hale getirilmesi ve kavramsal hiyerarşi oluşturulması işlemlerinden oluşur.

Bu işlemlere geçmeden önce verilerin hazırlanması için analitik bir temel olan Tanımsal Veri Özetleme işlemi gerçekleştirilir.

Veri Özetleme için kullanılan temel istatistiksel tekniklere;

- Merkezi eğilim ölçüleri olan Ortalama, Ağırlıklı Ortalama, Medyan, Mod,
- Verilerin dağılımını ölçen Açıklık, Çeyrek Ayrılış, Ortalama Ayrılış, Varyans ve Standart Sapma,
- Verilerin görsel kontrolü için grafiksel gösterimler olan Histogram, Kutu Grafiği, Saçılım Grafiği ve Matrisi örnek verilebilir [6].

Verilerin temizlenmesi

Verilerin temizlenmesi, eksik değerlerin (missing values) doldurulması, uyumsuz verilerin (outliers) tanımlanması, veri gürültülerinin (ortalamadan çok sapan veri değerleri) giderilmesi, çelişkili verilerin doğrulanması gibi işlemlerden oluşur.

Uyumsuz veriler (Outliers) gözlemlere uymayan anormal veri değerleridir. Genelde ölçüm ve yazım hatalarından kaynaklanır. Bu tip değerler kurulacak modeli olumsuz etkiler. Uyumsuz verilerle başa çıkmak için verilerin temizlenmesi adımının bir parçası olarak uyumsuz verilerin tespiti ve silinmesi ve uyumsuz verilere duyarsız güçlü modelleme metotları geliştirilmesi olmak üzere iki temel yol vardır:

Gerçek hayattan toplanan verilerde sıklıkla eksik verilere (missing values) rastlanır. Bazı veri madenciliği teknikleri eksik değerlerden etkilenmeden çalışabilirken, bazıları tüm veri alanlarının mutlaka dolu olmasını ister. Bu durumda eksik verilerin tamamlanması gerekliliği ortaya çıkar.

Bu sorunu çözmenin en kolay yolu, verilerin azaltılması yoluna giderek eksik veriler bulunan tüm özelliklerin silinmesidir. Veri kümesinin çok büyük olduğu ve eksik verilerin toplam verilerin düşük bir oranında olduğu durumlarda bu yol tercih edilir.

Veri kümesinin görece küçük ya da eksik verilerin yüksek bir oranda olduğu durumda bu eksik verilerin doldurulması gerekmektedir. Elde bulunan veriler incelenerek uygun bir değer eksik veri yerine yazılabilir ancak bu durum veri kümesini gürültülü (standarttan çok sapan) bir duruma getirebilir. Gürültüyü engellemek için eksik veriler bulunan özelliğin veri ortalaması bu verilerin yerine yazılabilir. Bu durumda da veri kümesi olduğundan daha homojen bir hale gelecektir.

Veri madenciliği ile elde edilecek sonuçların daha güvenilir olmasını sağlamak amacıyla bazı otomatik metotlar geliştirilmiştir. Bu metotlar Regresyon, Bayes formalizmi, kümeleme ve karar ağacı gibi teknikler yardımıyla nitelikler arasında ilişki kurup eldeki veriler yardımıyla eksik verilerin değerini tahmin ederler.

Verilerin temizlenmesi işlemi manuel olarak rastlantısal kontrollerle yapılabilirse de yüksek miktarda veri gerektiren veri madenciliği sürecinde bu işlemin en iyi şekilde yapılabilmesini sağlayan çeşitli yazılımlar geliştirilmiştir. Çizelge 2.4'te veri temizleme işlemi yapmak için kullanılabilecek yazılımlar belirtilmiştir.

Çizelge 2.4 Ticari veri temizleme yazılımları ve üretici şirketler [10].

Yazılım Adı	Firma	Web Adresi
Centrus Merge/Purge	Qualitative Marketing Software	http://www.qmsoft.com
Data Tools Twins	Data Tools	http://www.datatools.com.au
DataCleanser DataBlade	Electronic Digital Documents	http://www.informix.com
DataSet V	INTERCON	http://www.ds-dataset.com
DeDuce	The Computing Group	
DeDupe	International Software Publishing	
dfPower	DataFlux Corporation	http://www.dataflux.com
DoubleTake	Peoplesmith	http://www.peoplesmith.com
ETI Data Cleanse	Evolutionary Technologies Intern	http://www.evtach.com
Holmes	Kimoce	http://www.kimoce.com
i.d.Centric	firstLogic	http://www.firstlogic.com
Integrity	Vality	http://www.vality.com
matchIT	helpIT Systems Limited	http://www.helpit.co.uk
matchMaker	Info Tech Ltd.	http://www.infotech.ie
NADIS Merge/Purge Plus	Group 1 Software	http://www.g1.com
NoDuples	Quess Inc.	http://www.quess.com/nodupes.html
PureIntegrate	Carleton	http://www.carleton.com/products/View/index.htm
PureName PureAddress	Carleton	http://www.carleton.com/products/View/index.htm
QuickAddress Batch	QAS Systems	http://207.158.205.110/
reUnion and MasterMerge	PitneyBowes	http://www.pitneysoft.com
SSA-Name/Data Clustering Engine	Search Software America	http://www.searchsoftware.co.uk
Trillium Software System	Trillium Software	http://www.trilliumsoft.com
TwinFinder	Omikron	http://www.deduplication.com/index.html
Ultra Address Management	The Computing Group	

Verilerin kaynaştırılması ve dönüştürülmesi

Verilerin kaynaştırılması, farklı kaynaklardan toplanan verilerin uyumlu bir forma (tablo) sokulması için yapılan Metadata, korelasyon analizi, veri çatışması tespiti ve semantik uyumsuzluğun giderilmesi gibi işlemlerden oluşur.

Verilerin dönüştürülmesi ise verilerin veri madencilik uygulamasında daha iyi sonuç verecek forma sokulması işlemidir. İki farklı değerin ayrı ayrı verilerinin kullanılması yerine farkının alınması ya da oransal bir değere çevrilmesi gibi.

Dönüştürme işleminde dönüşüm sonucu elde edilen değer, model için bir önem ifade etmelidir. Böylece ayrı bir özellik olarak tanımlanabilir. Örnek olarak tıbbi veri kümelerinde sıklıkla rastlanan hasta ağırlık ve boy değerleri kullanılarak veri dönüştürme yoluyla Vücut Kitle İndeksi denen bir orana çevrilir. Bu oran yardımıyla hastaların kilolu olup olmadığının tespiti daha iyi yapılabilmektedir [7].

Veri hacminin azaltılması

Veri hacminin azaltılmasında temel amaç, veri küp bileşimi, özellik alt küme seçimi, boyutsal azaltım, kesikli hale getirme gibi işlemler yardımıyla en az veri kaybı ile veri madenciliği algoritmasının işlem zamanının kısılmasını sağlamaktır.

Verilerin kesikli hale getirilmesi ve kavramsal hiyerarşi oluşturma

Verilerin kesikli hale getirilmesi sürekli değişkenler (sayısal veriler) üzerinde uygulanır. Bunun için Birleştirme, Histogram analizi, Entropi temelli kesiklileştirme, KiKare Analizi, Kümeleme analizi ve Sezgisel bölümlendirme gibi teknikler kullanılır.

Verilerin dönüştürülmesi kararı, uygulanacak algoritmanın tipine bağlıdır. Sayısal tabanlı algoritmalar kullanılacaksa kesikli verilerin sürekli hale getirilmesi; kategorik tabanlı algoritmalar kullanılacaksa sürekli verilerin kesikli hale getirilmesi gereklidir [8].

Bazı kesikli veriler, uygulanan veri madenciliği tekniği içinde farklı ağırlıklara sahip olabilir. Bir özelliğe ait verileri $[0,1]$ aralığında değerler alırken bir diğerinin verilerinin $[-100,100]$ arasında değerler alması durumunda olduğu gibi. Bu durum elde edilecek sonuçları olumsuz etkileyecektir. Ölçekleme yardımıyla bu sürekli değişkenlerin kesikli hale getirilmesi ile uygulamada

ağırlıklarının daha dengeli olması ve elde edilecek sonuçların daha doğru olması sağlanır [7].

Kategorik veriler için kavramsal hiyerarşi, ancak özelliklerin belirli değerleri hiyerarşi ifade ediyorsa oluşturulabilir.

2.2.4. Modelin kurulması ve değerlendirilmesi

Tanımlanan problem için en uygun modelin bulunabilmesi, mümkün olan en çok sayıda model kurularak bunların denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele ulaşıncaya kadar yinelenen bir süreçtir.

Modelin kurulması aşamasında veri madencisinin vereceği kararlar, veri madenciliği çalışmasının başarıya ulaşmasında kritik öneme sahiptir. Model kurulumu aşamasına gelinceye kadar yeterli ve doğru verilerin toplanması, verilerin hazırlanması gibi işlemlerde veri madencisinin kararlarının doğrudan sürece etki ettiği unutulmamalıdır [1].

Model kuruluş süreci, denetimli ve denetimsiz öğrenmenin kullanıldığı modellere göre farklılık göstermektedir.

Örnekten öğrenme olarak da isimlendirilen denetimli öğrenmede, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir.

Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklerle uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir.

Denetimsiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır.

Denetimli öğrenmede seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenilmesi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenilmesi, öğrenim kümesi kullanılarak gerçekleştirildikten sonra test kümesi ile modelin doğruluk derecesi belirlenir.

Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik testidir. Bu yöntemde tipik olarak verilerin % 5 ile % 33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır.

(Doğruluk Oranı = 1 - Hata Oranı)

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

Bir kaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin n gruba ayrıldığı n katlı çapraz geçerlilik testi tercih edilebilir. Verilerin örneğin 10 gruba ayrıldığı bu yöntemde, ilk aşamada birinci grup test, diğer gruplar öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Elde edilen on hata oranının ortalaması kurulan modelin tahmini hata oranı olacaktır.

Bootstrapping, küçük veri kümeleri için modelin hata düzeyinin tahmininde kullanılan bir başka tekniktir. Çapraz geçerlilikte olduğu gibi model bütün veri kümesi üzerine kurulur. Ardından en az 200 bazen 1000'in üzerinde olmak üzere çok fazla sayıda öğrenim kümesi tekrarlı örneklemelerle veri kümesinden oluşturularak hata oranı hesaplanır.

Model kuruluşu çalışmalarının sonucuna bağlı olarak aynı teknikle farklı parametrelerin kullanıldığı veya başka algoritma ve araçların denendiği değişik modeller kurulabilir. Bu nedenle farklı modeller kurarak, doğruluk derecelerine göre en uygun modeli bulmak üzere sayısız deneme yapılmasında yarar bulunmaktadır.

Özellikle sınıflama problemleri için kurulan modellerin doğruluk derecelerinin değerlendirilmesinde basit ancak faydalı bir araç olan risk matrisi kullanılmaktadır. Aşağıda bir örneği görülen bu matriste sütunlarda fiili, satırlarda ise tahmini sınıflama değerleri yer almaktadır. Örneğin fiilen B sınıfına ait olması gereken 46 elemanın, kurulan model tarafından 2'sinin A, 38'inin B, 6'sının ise C olarak sınıflandırıldığı Çizelge 2.5'teki matrisde kolayca görülebilmektedir.

Çizelge 2.5 Risk Matrisi

	Fiili		
Tahmini	A Sınıfı	B sınıfı	C Sınıfı
A Sınıfı	45	2	3
B sınıfı	10	38	2
C Sınıfı	4	6	40
Toplam	59	46	45

Önemli diğer bir değerlendirme kriteri, modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, bir

çok kuruluş uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir.

Kaldıraç oranı ve grafiği, bir modelin sağladığı faydanın değerlendirilmesinde kullanılan önemli bir yardımcıdır. Örneğin kredi kartını muhtemelen iade edecek müşterilerin belirlenmesi amacını taşıyan bir uygulamada, kullanılan modelin belirlediği 100 kişinin 35'i gerçekten bir süre sonra kredi kartını iade ediyorsa ve tesadüfi olarak seçilen 100 müşterinin aynı zaman diliminde sadece 5'i kredi kartını iade ediyorsa kaldıraç oranı 7 olarak bulunacaktır.

Kurulan modelin değerinin belirlenmesinde kullanılan diğer bir ölçü, model tarafından önerilen uygulamadan elde edilecek kazancın bu uygulamanın gerçekleştirilmesi için katlanılacak maliyete bölünmesi ile edilecek olan yatırımın geri dönüş oranıdır.

Kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmamasıdır. Örneğin modelin kurulması sırasında varsayılan enflasyon oranının zaman içerisinde değişmesi, bireyin satın alma davranışını belirgin olarak etkileyecektir [11].

Kurulan modelin değerlendirme aşamasında, uygun model ya da modeller kurulduktan sonra, elde edilen veri madenciliği sonuçlarının araştırma probleminin amaçlarını gerçekleştirip gerçekleştirmediği değerlendirilir. Bu aşama sonuçların değerlendirilmesi, veri madenciliği sürecinin gözden geçirilmesi ve sonraki adımların ne olacağı hususlarını içermektedir. Bu aşamanın sonunda veri madenciliği sonuçlarının kullanımı üzerinde karara varılmaktadır [12].

2.2.5. Modelin yorumlanması ve elde edilen sonuçların sunumu

Modelin, elde edilmesi beklenen hedeflerini karşılamaya yeterli olduğu kararını aldıktan sonra, süreç bazlı daha geniş bir perspektiften değerlendirme yapılır. Bu değerlendirme, modelin doğru kurulup kurulmadığı, sadece eldeki verilerden mi yararlanıldığı, gelecekte kullanılabilecek farklı verilerin neler olabileceği, modelin genişletilmesi ya da güncellenmesi gerekliliği gibi konuları içerir.

Elde edilen sonuçların sunumunda veri görselleştirme tekniklerinden yararlanılır. Veri görselleştirme kavramı ve teknikleri Bölüm 3.6. Diğer veri madenciliği metotları başlığı altında incelenecektir.

2.3. Veri Madenciliği Sistemleri

2.3.1. Tipik bir veri madenciliği sisteminin yapısı

Veri madenciliği sürecini sistematik olarak uygulayan tipik bir Veri Madenciliği Sistemi, aşağıda belirtilen altı temel bileşenden oluşur :

Veri tabanı, Veri ambarı, Web Veri deposu veya başka bir veri bankası : Verilerin toplu olarak ilişkili veya dağınık olarak tutulduğu ortamlar olup bir veya daha çok veri tabanları, veri ambarları veya veri bankalarından oluşur. Bu kaynaklardaki veriler üzerinde veri temizleme ve veri bütünleştirme teknikleri uygulanabilir.

Veri tabanı veya Veri ambarı sunucusu : Bu sunucular kullanıcının veri madenciliği uygulaması için kullanmak istediği verileri ilgili veri tabanı ya da veri ambarından çekip getirme görevi yapar.

Bilgi tabanı : Araştırmayı yönlendirmek, model kurmak ve Veri madenciliği uygulaması sonucu elde edilecek desenlerin ilginçliğini değerlendirecek

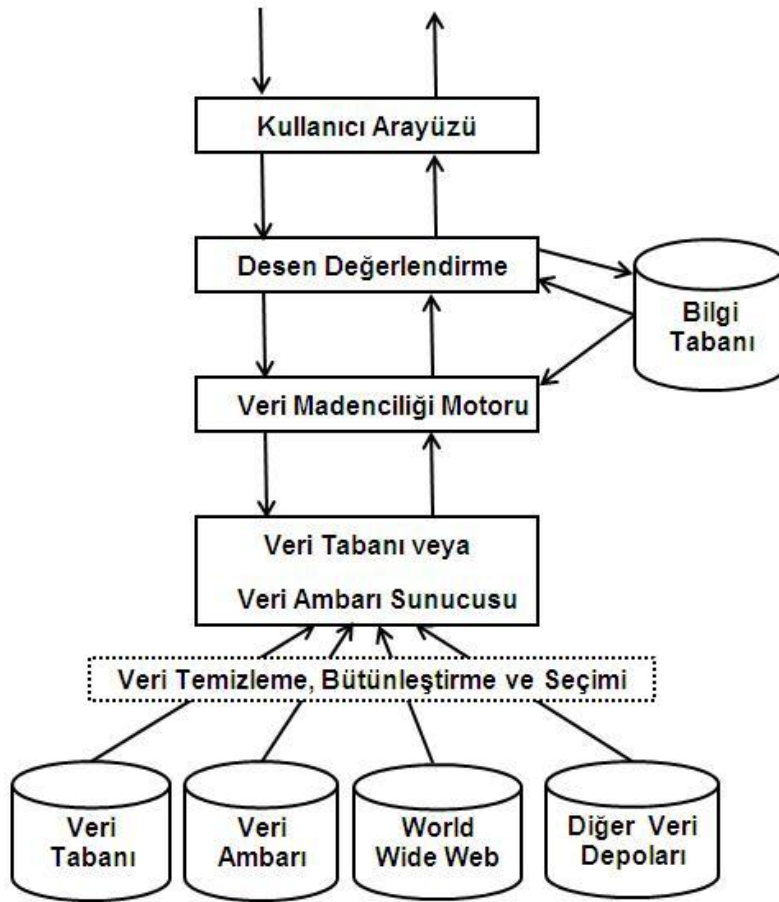
bilgileri içerir. Bilgi tabanı içinde yer alan bilgilere örnek olarak özelliklerin organize edilmesi ve farklı seviyelerde değer atanmasında kullanılan bilgiler, sonuçların beklenmezliği ve ilginçliğini belirleyecek kullanıcı varsayım ve inançları, ilginçlik kısıtları, farklı kaynaklardan toparlanan bilgiler (meta data) verilebilir.

Veri madenciliği motoru: Veri madenciliği sisteminin temelini oluşturur ve gerekli fonksiyonel modüller kümesini ihtiva eder. Bu modüller tanımlama, ilişki ve korelasyon analizi, sınıflama, tahmin, kümeleme analizi, uyumsuz verilerin (outliers) analizi ve değerlendirme analizi görevlerini yerine getirir.

Desen değerlendirme modülü: Bu bileşen ilginçlik ölçüsü kullanır ve desenin ilginçliğinin araştırılmasında veri madenciliği modülleri ile etkileşim kurar. İlginçlik eşiği belirleyerek uygulama sonuçlarını filtreleyebilir. Kullanılan verinin yapısına göre bu modül Veri madenciliği modülüne de bağlanabilir.

Kullanıcı arayüzü: Bu modül veri madenciliği sistemi ile kullanıcı arasında iletişimi sağlar. Kullanıcıların madencilik uygulaması için veri seçimi, metot seçimi, parametre belirleme vb. işlemleri yapmasını, çıkan sonuçların farklı formatlarda görüntülenmesini sağlar.

Şekil 2.9'da Tipik bir Veri Madenciliği Sisteminin yapısı verilmiştir.



Şekil 2.9 Tipik bir veri madenciliği sisteminin yapısı [5].

2.3.2. Veri madenciliği sistemlerinin sınıflandırması

Veri Madenciliği Sistemi, veri madenciliği süreci içinde dikkate alınan veri kaynağı, veri tipi vb. özellikler kendi içinde değiştikçe otomatik olarak değişir ve değişen niteliğin türü ile adlandırılır. Örneğin ilişkisel veri tabanı kullanılarak uygulanacak veri madenciliği sistemi, ilişkisel Veri Madenciliği Sistemi diye adlandırılır.

Buna göre sınıflandırma kriterleri ve veri madenciliği sistemi tipleri aşağıda özetlenmiştir :

a) Veri tabanlarının tipine göre,

- İlişkisel (Relational) Tabanlı Veri Madenciliği Sistemi,
- İşlemsel (Transactional) Tabanlı Veri Madenciliği Sistemi,
- Nesne-İlişkisel Tabanlı Veri Madenciliği Sistemi,
- Veri Ambarı Madenciliği Sistemi

b) Kullanılan özel veri tipine göre,

- Uzaysal (spatial) Veri bazlı Veri Madenciliği Sistemi,
- Zaman serileri bazlı Veri Madenciliği Sistemi,
- Akım veri (stream data) bazlı Veri Madenciliği Sistemi
- Text (Metin) Madenciliği Sistemi
- Multimedia Veri Madenciliği Sistemi
- Web Madenciliği Sistemi

c) Madencilik uygulaması yapılan bilginin türüne göre sınıflandırılan Veri Madenciliği sistemleri Özelleştirme (Characterization), Ayrımlandırma (Discrimination), Ortaklaştırma (Association) and Korelasyon Analizi, Sınıflandırma (Classification), Tahmin (Prediction), Kümeleme (Clustering), Sapma Analizi (Outlier analysis) ve Gelişim Analizi (Evolution analysis) gibi çeşitli Veri Madenciliği fonksiyonlarından birine odaklıdır.

Kapsamlı bir veri madenciliği sistemi genellikle bu veri madenciliği fonksiyonlarından birden çoğunu ayrı ayrı ya da entegre olarak sağlar.

Yine bu sınıflandırma içinde düzenliliği araştıran (sıklıkla ortaya çıkan desenleri) ya da düzensizliği araştıran (istisna verileri ve sapmaları) veri madenciliği sistemleri şeklinde de bir ayrıma gidilebilir.

d) Veri Madenciliği uygulaması yapılan bilginin soyutlama seviyesine göre,

bilginin soyutlama seviyesi, ilkel seviyede veri (işlenmemiş durumda), genelleştirilmiş bilgi (yüksek soyutlama seviyesinde) veya çok seviyeli bilgi (çeşitli soyutlama seviyelerini içeren) şeklinde üç grupta toplanır.

Kapsamlı bir veri madenciliği sistemi çok seviyeli bilgi (çeşitli soyutlama seviyelerini içeren) içinden bilgi keşfi yapabilmelidir.

e) Uygulanan veri analiz metoduna göre veri tabanı ya da veri ambarı tabanlı sistemler, makine öğrenimli sistemler, İstatistiksel sistemler, görselleştirici sistemler, desen tanımlayıcı sistemler, Sinirsel ağlar şeklinde sınıflandırılabilir.

Yine bu sınıflandırma içinde Kullanıcı etkileşimi derecesine göre Otonom Sistemler, İnteraktif Keşifçi Sistemler, Sorgu tabanlı sistemler şeklinde bir ayırım yapılabilir.

Karmaşık bir Veri Madenciliği sistemi, çoklu veri analiz metotlarını kullanabilen ya da farklı metotlardan yaklaşımları entegre bir şekilde uygulayabilen bir seviyede olmalıdır.

f) Kullanılabilen uygulama alanlarına göre Finans, Telekomünikasyon, DNA, hisse senedi, email vb. odaklı veri madenciliği sistemleri geliştirilmiştir [5].

2.4. Veri Madenciliğinde Kullanılan Metotlar

Veri Madenciliği uygulaması hazır hale getirilen veri kümesi dosyası üzerinde kurulan modele ve verilerin yapısına uygun olarak seçilen çeşitli veri madenciliği metotları kullanılarak gerçekleştirilir. Veri madenciliğinde kullanılan metotlar 3. Bölüm’de detaylı olarak incelenmiştir.

2.5. Veri Madenciliğinin Kullanım Alanları

Veri Madenciliği verinin yoğun olarak üretildiği hemen her ortamda uygulama kullanım alanı bulabilir ve bulmuştur. Bazı uygulama alanları şu şekilde özetlenebilir [13].

Pazarlama alanında Veri madenciliği uygulamaları,

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,
- Pazar Araştırması : Hedef pazar , müşteriler arası benzerliklerin saptanması, çapraz pazar incelemesi,
- Pazar sepeti analizi (Market Basket Analysis),
- Müşteri ilişkileri yönetimi (Customer Relationship Management),
- Müşteri değerlendirme (Customer Value Analysis),
- Satış tahmini (Sales Forecasting) yapmak amacıyla kullanılır.

Bankacılık alanında Veri madenciliği uygulamaları,

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Müşteri kredi risk araştırmaları,
- Kredi kartı dolandırıcılıklarının tespiti,
- Hisse senetlerinin değer değişimi ile ilgili tahminler yapılması için kullanılır.

Sigortacılık alanında Veri madenciliği uygulamaları,

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri örüntülerinin belirlenmesi amacıyla kullanılır.

Tıp-Eczacılık alanındaki Veri madenciliği uygulamaları 3. Bölümde geniş şekilde ele alınmıştır.

Bu uygulama alanları dışında Taşımacılık ve Ulaşım, Turizm ve Otelcilik, Telekomünikasyon ve Kamu (Belediye hizmetleri, e-devlet vb.) alanlarında da Veri madenciliği uygulamalarının faydalı olabileceği söylenebilir [14].

2.6. Veri Madenciliği Yazılımları

Veri madenciliği metotlarının uygulanmasını kolaylaştırmak amacıyla bu metotları uygulayabilen çeşitli bilgisayar yazılımları geliştirilmiştir. Bu yazılımlardan belli başlılarının Genel özellikleri Çizelge 2.6'da, veri madenciliği ile ilgili teknik özellikleri Çizelge 2.7'de belirtilmiştir [15].

Yazılımların genel özelliklerini gösteren çizelgede, ürünün adı, üreticisi, yapısı, yasal statüsü, demo içeriği, mimari yapısı, boyutu, veri modeli, desteklediği özellik tipleri ve kullanılabildiği işletim sistemi gibi özellikler yer almaktadır.

Yazılımların veri madenciliği özelliklerini gösteren çizelgede ise, bilgi keşif görevleri, keşif metodolojisi ve kullanıcı etkileşimi özellikleri yer almaktadır.

Kurumsal amaçlarla kullanmak üzere uygun bir Veri Madenciliği Yazılımı seçmek için dikkat edilmesi gereken hususlar aşağıda verilmiştir [9]:

- Yazılımın uygulayabildiği veri madenciliği teknikleri
- Yazılımın kapasite kısıtları (Veri Boyutu, kullanıcı sayısı, veri kümesi içinde özellik sayısı, uygun donanım tipi)
- Veri tabanlarına ve dosyalara erişim imkanı
- Çok boyutlu kullanıcı arayüzü desteği ve kullanım kolaylığı
- Oluşturulan modellere ilişkin detaylı açıklama desteği
- Görsel sunum, raporlama formatlarının yeterliliği
- Farklı yazılımlar ile etkileşim uyumu
- Danışma ve eğitim desteği

Çizelge 2.6 Veri madenciliği yazılımlarının genel özellikleri

Ürün Adı	Üreticisi	Yasal Statü	De mo İçe	Mimarî Yapı	Veri Boyutu	Veri Modeli	Algoritma	Co	Ca	S	MF	Unix	Maz	Dos	W3.a	W9a	WNT	OS/2
Alice	Issoft	C	D	S	L	I	X	X	X	X				X		X		
AutoClass	Nasa	PD	D	S.P	L	I	X	X	X	X				X				
Bayesian Knowl. Disc.	Open U.	F	D	S	L	I	X	X	X	X								
BrainMaker	Cal. Sc Software	C	N	S	S	I	X	X	X	X				X		X		
Brule	Univ. Of Washington	C	N	S	S	I	X	X	X	X				X		X		
BusinessMiner	Business Objects	C	N	S	M	R	X	X	X	X				X		X		
Claudian	K.U. Leuven	C	D	S	S	R	X	X	X	X								
Clementine	Integral Solutions Ltd.	C	N	S	M	R	X	X	X	X								
CN2	Univ. Of Texas	PD	D	S	S	I	X	X	X	X								
CrossGraphs	Bellmont Research	C	N	S	M	R	X	X	X	X				X		X		
Cubist	RuleQuest	C	D	S	M	I	X	X	X	X				X		X		
C5.0	RuleQuest	C	D	S	M	I	X	X	X	X				X		X		
Darwin	Thinking Machines	C	N	C/S, P, S	L	R	X	X	X	X								
Data surveyor	Data Destilleries	C	R	C/S, P	L	R	X	X	X	X				X		X		
DB Miner	SFU	C	N	C/S	L	R	X	X	X	X								
Delta Miner	Biseritz	C	D	C/S, S	L	R	X	X	X	X								
Decision Series	NeoVisda	C	N	C/S, P	L	R	X	X	X	X								
DI Driver	Dimensional Insight	C	D	C/S, S	M	R	X	X	X	X				X		X		
Ecobweb	Tei Aviv Univ.	PD	D	S	S	I	X	X	X	X								
IDIS	Information Discovery	C	N	C/S, P, S	L	R	X	X	X	X				X		X		
IND	Nasa	C	N	S	S	I	X	X	X	X								
Intelligent Miner	IBM	C	R	C/S, P, S	L	R	X	X	X	X								
KATE-Tools	AcknoSoft	C	N	C/S, S	M	R	X	X	X	X				X		X		
Kepler	GMD	C	R	S	M	R	X	X	X	X								
Knowledge SEEKER	Angosa	C	D	C/S, S	L	I	X	X	X	X				X		X		
MineSet	Silicon Graphics	C	N	C/S, S	L	R	X	X	X	X								
MLC++	Silicon Graphics	F	D	S	L	R	X	X	X	X								
Model	GMD	F	D	S	S	R	X	X	X	X								
ModelQuest	AbTech	C	N	S	S	R	X	X	X	X								
MSN	Microsoft	F	D	S	S	R	X	X	X	X								
MVSP	KCS	S	D	S	M	I	X	X	X	X				X		X		
PolyAnalyst	Megaputer	C	D	C/S, S	M	R	X	X	X	X								
PVE	IBM	C	N	S	M	I	X	X	X	X								
Q-Yield	Qualition	C	D	S	S	R	X	X	X	X								
Ripper	AT&T	C	D	S	L	I	X	X	X	X								
Rosetta	NTNU	C	D	S	S	R	X	X	X	X								
Rough Enough	Troll Data	F	D	S	S	R	X	X	X	X				X		X		
Scenario	Cognos	C	R	S	M	R	X	X	X	X								
Sipma-W	Univ. Of Lyon	S	D	S	M	I	X	X	X	X								
SuperQuery	AZMY	C	D	S	M	R	X	X	X	X								
ToolDog	UFES	F	D	S	S	I	X	X	X	X								
Weka	Univ. Of Waikato	F	D	S	M	R	X	X	X	X				X				
WizRule	Wizsoft	C	D	S	M	I	X	X	X	X								

Yasal Statü : PD = Engine Açık (Public Domain) F = Öcretsiz (Freeware) S = Sınırlı Öcretsiz Kullanım (Shareware) C = Öcretli (Commercial)

Demo İçeriği : D = İnternette Demosu mevcut R = İstek üzerine Demosu mevcut U = Unknown (Bilinmeyen)

Mimarî Yapısı : Yazılımın kullandığı bilgisayar yapıları S = Standalone (PC) C/S = Client server (Network) pp = Paralel İşlemcili

Nitelik Tipi : İşlenebilen nitelik tipleri Co = Sürekli (Continuous) Ca = Kesik nitelik (Categorical) S = Sembolik

Veri Boyutu : İşlenebilen maksimum veri sayısı S = Küçük (10 bin kayıta kadar) M = Orta boyutlu (10 bin ile 1 milyon kayıta kadar) B = Büyük (1 milyon kayıttan fazla)

Veri Modeli : Analiz edilecek veriler için veri modeli R = İlişkisel (Relational) O = Nesne tabanlı (Object Oriented) N = Model yok (Yazılım sadece tek tablo alır)

Çizelge 2.7 Veri madenciliği yazılımlarının teknik özellikleri

Ürün Adı	Üreticisi	Bilgi Keşif Görevleri					Keşif Metodolojisi										Kullanıcı Etk.				
		Pre.	P	Regr	Clu.	Clu.	A	Vis.	EDA	NN	GA	FS	RS	St.	DT	RI	BN	CSR	A	G	H
Alice	Isolt	X			X	X		X											X	X	
AutoClass	Nasa				X	X													X	X	
Bayesian Knowl. Disc.	Open U.				X	X													X	X	
BrainMaker	Cal Sc Software		X	X	X	X											X		X	X	
Brute	Univ. Of Washington				X	X													X	X	
BusinessMiner	Business Objects		X		X	X													X	X	
Claudian	K.U. Leuven				X	X													X	X	
Clemence	Integral Solutions Ltd	X	X	X	X	X		X	X	X									X	X	
CN2	Univ. Of Texas				X	X		X	X										X	X	
CrossGraphs	Beimont Research	X			X	X		X	X										X	X	
Cubist	RuleQuest		X	X	X	X													X	X	
C5.0	RuleQuest				X	X													X	X	
Darwin	Thinking Machines		X	X	X	X		X	X										X	X	
Data surveyor	Data Destilleries	X	X	X	X	X		X	X	X									X	X	
DB Miner	SFU	X	X	X	X	X		X	X	X									X	X	
Delta Miner	Bissantz	X	X	X	X	X		X	X	X									X	X	
Decision Series	NeoVista	X	X	X	X	X		X	X	X									X	X	
DI Diver	Dimensional Insight	X	X	X	X	X		X	X	X									X	X	
Ecobweb	Tel Aviv Univ.		X		X	X		X	X										X	X	
IDIS	Information Discovery		X	X	X	X		X											X	X	
IND	Nasa		X	X	X	X													X	X	
Intelligent Miner	IBM	X	X	X	X	X													X	X	
KATE-Tools	AcknoSoft	X	X	X	X	X		X	X	X									X	X	
Kepler	GMD	X	X	X	X	X		X	X	X									X	X	
Knowledge SEEKER	Angoss	X	X	X	X	X		X	X	X									X	X	
MineSet	Silicon Graphics	X	X	X	X	X		X	X	X									X	X	
MLC++	Silicon Graphics	X	X	X	X	X		X	X	X									X	X	
Modal	GMD	X	X	X	X	X		X	X	X									X	X	
ModelQuest	AbTech	X	X	X	X	X		X	X										X	X	
MSBN	Microsoft	X	X	X	X	X		X	X	X									X	X	
MVSP	KCS	X	X	X	X	X		X	X										X	X	
PolyAnalyst	Megaputer	X	X	X	X	X		X	X	X									X	X	
PVE	IBM		X	X	X	X		X	X	X									X	X	
Q-Yield	Quadrillion		X	X	X	X		X	X	X									X	X	
Ripper	AT&T	X	X	X	X	X		X	X	X									X	X	
Rosetta	NTNU	X	X	X	X	X		X	X	X									X	X	
Rough Enough	Troll Data	X	X	X	X	X		X	X	X									X	X	
Scenario	Cognos	X	X	X	X	X		X	X	X									X	X	
Sipina-W	Univ. Of Lyon	X	X	X	X	X		X	X	X									X	X	
SuperQuery	AZMY	X	X	X	X	X		X	X	X									X	X	
ToolDiag	UPES	X	X	X	X	X		X	X	X									X	X	
Weka	Univ. Of Waikato	X	X	X	X	X		X	X	X									X	X	
WizRule	Wissoft				X	X		X											X	X	

Bilgi Keşif Görevleri : Pre = Veri Hazırlama, P = Tahmin, Regr = Regresyon, Clu = Sınıflama, Clu. = Kümeleme, A = Birliklik Kuralları,

Vis = Model Gökşelleşme, EDA = Açıklayıcı Veri Analizi

Keşif Metodolojisi : NN = Sinirsel Ağlar, GA = Genetik Algoritmalar, FS = Bulanık Küme, RS = Pürüzlü Küme, St. = İstatistiksel Metotlar, DT = Karar Ağacı

Kullanıcı Etkileşimi : RI = Kural Çıkarımı, BN = Bayes Ağları, CBR = Durum Bazlı Muhakeme

A = Otonom, G = İnsan rehberli Keşif, H = Yüksek derecede İnteraktif

Bazı ticari veri madenciliği yazılımlarının güçlü ve zayıf nitelikleri Çizelge 2.8'de özetlenmiştir.

Çizelge 2.8 Bazı veri madenciliği yazılımlarının güçlü ve zayıf nitelikleri [16]

ÜRÜN	GÜÇLÜ NİTELİĞİ	ZAYIF NİTELİĞİ
Clementine	Görsel Arayüzü, Algoritma genişliği	Ölçeklenirlik
Darwin	Etkili hizmet sağlayıcı, sezgisel arayüz opsiyonu	Sınırlı görselleştirme
DataCruncher	Kullanım kolaylığı	Tek algoritma içermesi
Enterprise Miner	Algoritma derinliği, görsel arayüzü	Kullanımı zor, kompleks yapısı
GainSmarts	Veri dönüştürme, algoritma seçeneği genişliği	Sınırlı görselleştirme
Intelligent Miner	Algoritma genişliği, grafiksel ağaç/küme çıktısı	Algoritma seçeneği az
MineSet	Veri Görselleştirme	Algoritma seçeneği az, model ihracına izin vermez
Model1	Kullanım kolaylığı, otomatik model keşfi	Dikey bir araç olması
ModelQuest	Algoritma genişliği	Sezgisel olmayan arayüz seçenekleri
PRW	Kapsamlı algoritmalar, otomatik model seçimi	Sınırlı görselleştirme
CART	Ağaç opsiyonları derinliği	Dosya açılması zorluğu, sınırlı görselleştirme
Scenario	Kullanım kolaylığı	Dar analiz yolu
NeuroShell	Çoklu Sinirsel ağ mimarisi	Alışılmışın dışında arayüz, sadece sinirsel ağ içermesi
OLPARS	Çoklu istatistiksel algoritmalar, sınıf tabanlı görselleştirme	Eski arayüz, Dosya açılması zorluğu
See-5	Ağaç opsiyonları derinliği	Sınırlı görselleştirme, az veri seçeneği
S-Plus	Algoritma derinliği, görselleştirme, programlanabilir	Sınırlı çıkarım metodu, dik öğrenme eğrisi
WizWhy	Kullanım ve model kavrama kolaylığı	Sınırlı görselleştirme

Tez kapsamında yapılan uygulama çalışmasında ESTARD Data Miner Version 2.1.0.1 ve WEKA Version 3.7.0 Veri Madenciliği yazılımları kullanılmıştır.

2.7. Veri Madenciliğinin Geleceği

Veri madenciliği konusunda bugüne kadar yapılan çalışmaları dikkate alarak mevcut çalışmaların yapıldığı araştırma alanlarını ve gelecekte ortaya çıkabilecek araştırma trendlerini içeren Veri madenciliği haritası Şekil 2.10'da gösterilmiştir [17].

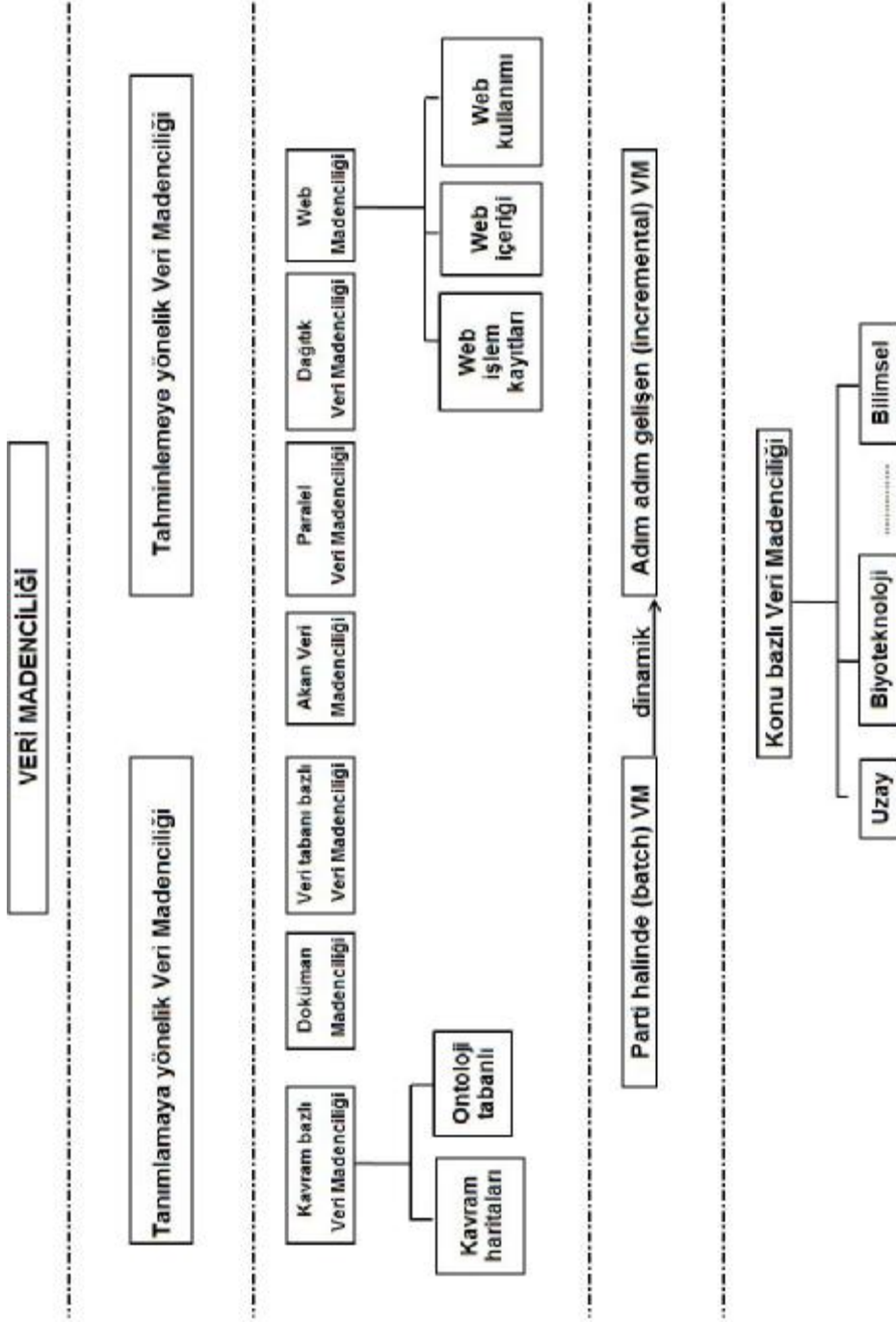
Veri madenciliğinin geleceği, 2.7.1. ve 2.7.2. başlıkları altında incelenecektir.

2.7.1. Potansiyel veri madenciliği uygulama alanları

Gelecek yıllarda Veri madenciliği uygulama potansiyeli olan konular aşağıda ana başlıklar altında özetlenmiştir [18,19].

- *Ülke Güvenliği ve İstihbarat :*

- * İstihbarat verilerinin analizi,
- * Güvenlik riski olan bölgelerin belirlenmesi ,
- * Kritik altyapıyı tehdit eden siber saldırı izlerini (pattern) tespit etmek ve bu saldırıları engellemek
- * Bina ve tesis Giriş-çıkışlarından elde edilecek verilerin (giriş kodu, parmak izi, retina izi vb.) analizi,
- * El yazısı okuma ve kimlik tespiti,
- * Şehirlere ve tesislere kurulan güvenlik kamerası sistemlerinden görüntü analizi, suç mahalleri modellemesi, riskli mekanların tespiti vb.,
- * Havalimanı, gümrük vb. Giriş noktalarında kişi,eşya kontrol verileri üzerinde veri analizi, (SARS, domuz gribi gibi salgın hastalıkların ülkeye girişte engellenmesi için havalimanlarına kurulan vücut ısı dedektörlerindeki verilerin analizi gibi.)



Şekil 2.10 Veri madenciliği haritası [17]

-Enerji Sektörü

- * Enerji nakil hatlarında arızaların hangi bölgelerde yoğunlaştığının ve ne tip arızalar ile karşılaşıldığının tespiti,
- * Enerji nakil hattı arızaların önlenmesi için nelerin yapılması gerektiğinin, hangi bölgelere ve merkezlere dikkat edilmesi gerektiğinin önceden tespit edilmesi,
- * Enerji müşteri kaydını analiz ederek satış, faturalandırma ve müşteri hizmetlerini iyileştirilmesi,
- * Müşteri promosyon sistemi gibi uygulamaların modellenmesi,
- * Elektrik enerjisi üretim sistemlerinde güç değişimlerinin tahmini. Bu konuda Kusiak ve Zheng [20], rüzgar türbinlerinin zaman aralıklarına bağlı güç değişimlerini tahmin etmek için veri madenciliği tekniği de kullanan bir çok değişkenli zaman serileri modeli geliştirmiştir. Model üzerinde beş farklı veri madenciliği algoritması ayrı ayrı kullanılmış ve elde edilen sonuçlar karşılaştırılmıştır.

-Kamu Sektörü

- * E-devlet uygulamalarının yaygınlaşması ile çeşitli veri tabanlarında kayıt altına alınan her türlü verinin analizi,
- * Kamu hizmet planlarının oluşturulmasında kamu hizmeti alan müşteri verilerinin analizi.

Bu alanlar dışında Veri Madenciliği uygulamaları gelecekte genetik, biyoloji (oşinografi vb. Alt dalları), kriptoloji, bilgi sistemleri güvenliği, uzay bilimleri, haberleşme teknolojisi gibi alanlarında daha fazla kullanılacaktır.

2.7.2. Veri madenciliğinden doğan yeni kavramlar

Veri Madenciliği alanı zamanla gelişerek literatüre farklı kavramlar katmaktadır. Gelecekte üzerinde daha fazla çalışma yapılacak Veri Madenciliği temelli iki yeni kavram *Veri füzyonu* ve *Veri Çiftçiliği*'dir.

Veri füzyonu, karar vericiler için farklı veri tiplerinin ortak bir çevrede birleştirilmesi işlemidir.

Veri füzyonu algoritmaları, savunma sektöründe hedef takibi, hedef kimlik tespiti amacıyla istihbarat, keşif ve gözetleme operasyonlarında kullanılmaktadır. Veri madenciliği ve veri füzyonu birbirini tamamlayan prosesler olmasına rağmen, araştırmacılar bu iki alanda birbirinden bağımsız olarak, herhangi bir ilişkiye girmeden çalışmaktadırlar [21].

Veri Füzyonu konusunda yapılan az sayıdaki çalışmadan birinde Waltz [22] görüntü verisi, yersel veri, video görüntüleri, istatistiksel veri setleri, anahtar kelimeler içeren metin verisi içinde gizlenmiş hedeflerin tespiti ve sınıflandırması işlemlerinin iyileştirilmesi için veri madenciliği ve veri füzyonu tekniklerinin birleştirilmesi yollarını araştırmıştır.

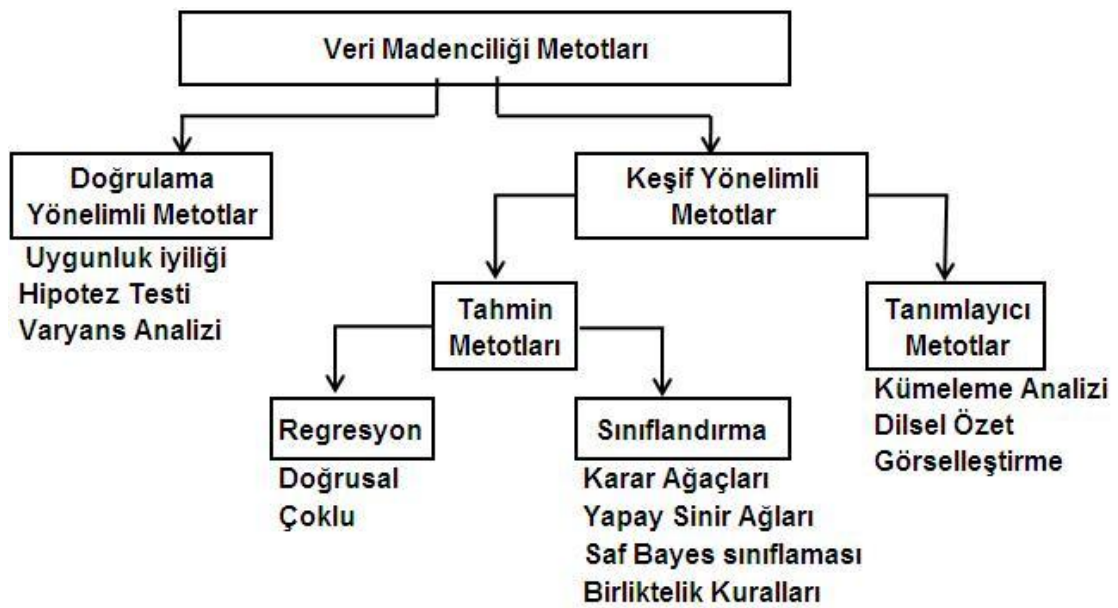
Bir diğer çalışmada Aygün ve Adalı [21], dört temel sınıflandırma algoritması farklı kombinasyonlarda birleştirilerek on farklı veri takımı üzerinde yapılan deneylerde ortalama olarak mevcut basit ve bileşik sınıflandırma algoritmalarından %5-15 arasında daha iyi sonuçlar alınmıştır.

Veri Çiftçiliği ise en az veri toplama maliyeti ile en iyi performans ölçüsü (tahmin gerçekliği, gruplama performansı vb.) değeri veren veri kümesini oluşturmak için kullanılan metotlardan oluşan bir veri madenciliği kavramıdır. Literatürde çok yeni olan bu kavram hakkında Kusiak'ın iki çalışması mevcuttur [23].

3. VERİ MADENCİLİĞİ METOTLARI

3.1. Veri Madenciliği Metotlarının Sınıflandırılması

Veri Madenciliği metotları farklı amaçlar için kullanılabilmeleri nedeniyle çeşitli gruplar sınıflandırılmaktadır. Şekil 3.1’de bu sınıflandırma genel olarak verilmiştir.



Şekil 3.1 Veri madenciliği metotları

Veri Madenciliği metotları temel olarak iki ana gruba ayrılırlar.

Keşif yönelimli Metotlar (Discovery-oriented Methods)

Verilerin içindeki desenleri doğrudan tanımlayan metotlardır. Keşif yönelimli modellerin çoğu (nicel olanlar) yeterli sayıda çalışma örneklerinden genelleme yoluyla model oluşturur, yani tümevarımsal öğrenme tabanlıdır. Tümevarımsal yaklaşımın temel varsayımı çalıştırılan modelin gelecekte ortaya çıkacak örnekler için uygulanabilir olmasıdır.

Keşif yönelimli modeller kendi içlerinde Tanımlama ve Tahmin metotları diye ikiye ayrılır.

Tanımlayıcı metotlar, verilerin içinde tespit ettikleri desenlerin diğer veri parçaları ile ilişkisini anlamaya odaklıdır.

Tahmin metotları ise yeni ve görülmemiş örnekler elde eder ve bu örneklerle dayanarak bir veya daha fazla değişken değerini tahmin etmeye yarayan davranışsal bir model oluşturmayı amaçlar. Ayrıca keşfedilen bilgileri anlaşılır ve işlenebilir kılacak desenler de geliştirir.

Doğrulama yönelimli Metotlar (Verification-oriented Methods)

Doğrulama yönelimli metotlar bir dış kaynak (bir uzman vb.) tarafından ileri sürülen hipotezin değerlendirmesi ile ilgilidir. Geleneksel istatistiğin en yaygın kullanılan metotları olan uygunluk iyiliği testi, hipotez testi ve varyans analizini içerir. Ancak bu metotlar veri madenciliği ile keşif yönelimli metotlara göre daha az ilgilidir çünkü veri madenciliği var olan bir hipotezin test edilmesinden çok yeni bir hipotez keşfetmeye çalışır. Veri Madenciliği Metotları başlıklı 4. bölümde temel olarak Keşif Yönelimli Metotları incelenecektir.

Literatürde kullanılan bir başka terminoloji ise *Denetimli öğrenme* (*Supervised / Directed learning*) veya *Denetimsiz öğrenme* (*Unsupervised / Undirected learning*) olarak iki sınıfa ayırmaktadır.

Denetimli öğrenim metotları, girdi özellikleri (input attributes) ya da bağımsız değişkenler ile çıktı özellikleri (output attributes) ya da bağımlı değişkenler arasındaki ilişkiyi keşfetmeye çalışır. Keşfedilen ilişki bir model olarak temsil edilir. Modeller genellikle veri kümesi içinde gizli olan ve girdi özellikleri bilindiğinde hedef özellikleri tahmin etmek için kullanılabilecek desenlerin anlaşılmasında kullanılırlar. Denetimli öğrenme metotları pazarlama, finans,

üretim gibi alanlarda kullanılabilir [10]. Denetimli metotlar sınıfına Regresyon modeller, Birliktelik Kuralları, Karar Ağaçları (Decision trees), Geri beslemeli Yapay Sinir Ağları (Neural networks) gibi metotlar girmektedir.

Denetimsiz öğrenim metotları, yüksek boyutlu girdi uzayında örneklerin dağılımını modeller. Elde edilmesi istenen sonuç için özel bir tanımlama yapılmamış veya belirsizlik sözkonusudur [24].

Denetimli öğrenim metotlarından farkı, veri kümesi içinde çıktı (bağımlı) özellikleri olmadan modelleme yapmasıdır. Model yardımıyla veri kümesi içinde oluşturulan desenlerin veri kümesi dışındaki çıktı durumlarına ne kadar uyduğu belirlenir [10].

Denetimsiz metotlar sınıfına En yakın k-komşuluk (k-Nearest-Neighbor), K-ortalamalar kümeleme (K-means clustering), Aşamalı Kümeleme (Hierarchical clustering), Kendi kendini düzenleyen haritalar (Self organized maps; Kohonen Networks) gibi metotlar girmektedir.

Önemli veri madenciliği metotlarının detayları aşağıda verilmiştir.

3.2. Regresyon Tabanlı Metotlar

Regresyon tabanlı metotlar, temeli Regresyon tekniğine dayalı geliştirilen metotlardır. İstatistiksel bir metot olan Regresyon, eldeki veriler içinde belirlenen iki değişken kümesi (Bağımsız Değişken ya da Tahmin Edici ile Bağımlı Değişken) arasındaki ilişkiyi belirlemek için kullanılır [25].

Regresyon tabanlı metotlar iki ana başlık altında incelenecektir.

3.2.1. Doğrusal regresyon

Basit doğrusal regresyon, iki sürekli değişken (tahmin edilmeye çalışılan bağımlı değişken ve bağımsız değişken) arasındaki ilişkiyi tanımlamayı amaçlayan bir tekniktir. Teknik verileri kullanarak bir doğru denklemi oluşturmayı hedefler. Bu doğru oluşturulurken tüm veri noktalarından tahmin edilen eğriye olan uzaklığın karelerinin minimize edilmesi ile doğrunun optimize edilmesi sağlanır. Doğru elde edildikten sonra iki değişken arasındaki ilişkinin gücü R-kare (R-Square) değeri ile tanımlanır. R-kare verinin değişiminin ne ölçüde oluşturulan model (çizilen doğru) ile açıklanabildiğini gösterir [11].

Basit Doğrusal Regresyon Modeli :

n adet gözlem için, bağımlı değişken değeri y_i ile bağımsız (tahmin edici) değişken değeri x_i arasındaki doğrusal ilişkinin

$$\hat{y} = b_0 + b_1 x \quad (3.1)$$

şeklinde Regresyon doğrusu gösterimine *Regresyon Eşitliği ya da Tahmin Edilen Regresyon Eşitliği* denir.

Bu eşitlikte;

– $\hat{y}$ Bağımlı değişkenin yaklaşık değeri ,

– b_0 regresyon doğrusunun y ile kesişimi

– b_1 Regresyon doğrusunun eğimi,

– b_0 ve b_1 , Regresyon katsayılarıdır.

$$b_1 = \frac{\sum x_i y_i - [(\sum x_i) (\sum y_i)] / n}{\sum x_i^2 - (\sum x_i)^2 / n} \quad (3.2)$$

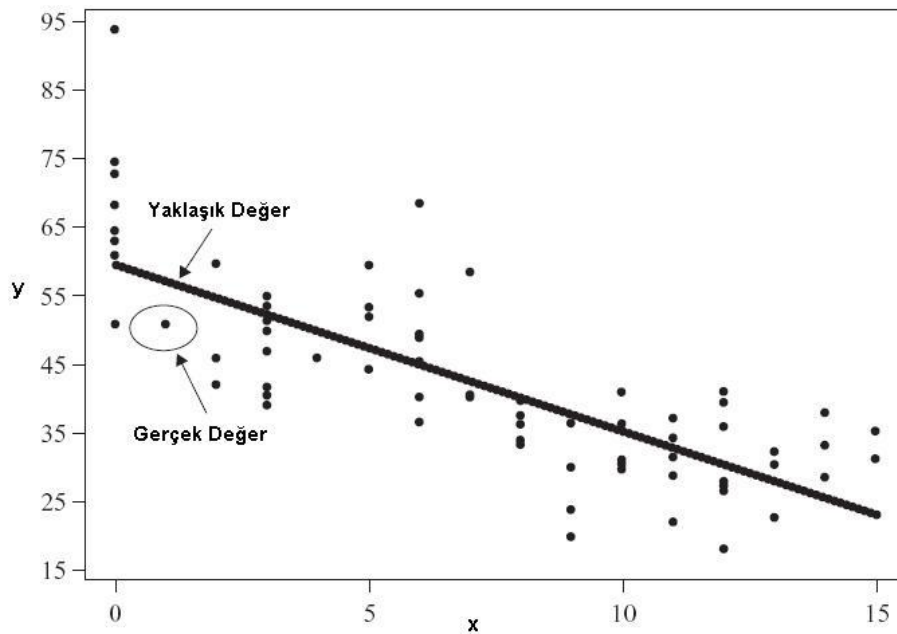
$$b_0 = \bar{y} - b_1 \bar{x} \quad (3.3)$$

$\bar{x}$ bağımsız değişkenin ortalama değeri, $\bar{y}$ bağımlı değişkenin ortalama değeri, n ise (i=1'den n'e kadar) toplam gözlem sayısıdır.

Tahmin yapmak için gerekli regresyon doğrusunun ne kadar güvenilir olduğunu belirlemek için doğrusal yaklaşımın gerçek gözlem değerlerine ne kadar uyduğunu ölçen Belirlilik Katsayısı adı verilen r-kare istatistiği kullanılır.

$$r^2 = y - \hat{y} \quad (3.4)$$

Şekil 3.2'de bir Doğrusal Regresyon modelinin grafiksel gösterimi yer almaktadır.



Şekil 3.2 Doğrusal Regresyon modelinin grafiksel gösterimi

Doğrusal Regresyon Modeli :

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (3.5)$$

β_0 ve β_1 y kesişimi ve eğim için model parametrelerini, ε ise hata terimini temsil eder.

Hata Terimi ile ilgili Varsayımlar

1. Hata terimi ortalaması ve beklenen değeri 0 olan bir rassal değişkendir.

$$E(\varepsilon) = 0 \quad (3.6)$$

2. Hata teriminin σ_2 ile gösterilen varyansı x değerinden bağımsız olarak sabittir.
3. Hata teriminin değerleri bağımsızdır.
4. Hata terimi ε Normal dağılımlı bir rassal değişkendir. Hata terimi değerleri ε_i bağımsız normal rassal değişkendirler. Ortalaması ve varyansı 0'dır.

Regresyon Modelinde $\beta_1 = 0$ değeri alırsa regresyon eşitliği $y = \beta_0 + \varepsilon$ olur, bu da x ile y arasında doğrusal bir ilişki olmadığı anlamına gelir. β_1 0'dan farklı bir değer alırsa bu durum x ile y arasında bir çeşit doğrusal ilişkinin varlığı anlamına gelir [26].

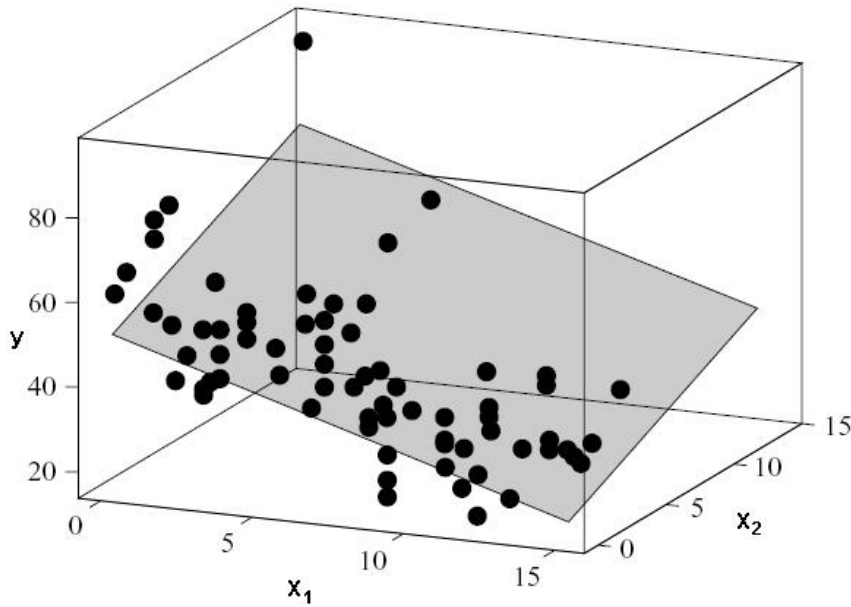
İki değişken arasındaki ilişkiyi bulmak, ilişki varsa bu ilişkinin gücünü belirlemek, değişkenler arasındaki ilişkinin türünü belirlemek, ileriye dönük değerleri tahmin etmek gibi konularda kullanılır. Regresyon analizi, araştırma, matematik, finans, ekonomi, tıp gibi bilim alanlarında yoğun olarak kullanılmaktadır [11].

3.2.2. Çoklu regresyon

Çoklu Regresyon'da bir Bağımlı Değişken kümesine karşılık birden fazla Bağımsız Değişken kümesi söz konusudur. Çoklu Regresyon'dan Doğrusal Regresyon modelinden daha büyük veri kümeler için daha karmaşık ilişkilerin tespitinde yararlanılır [25].

Çoklu Regresyon modeli, doğrusal modeldeki tek boyutlu doğru yerine bir düzlem ya da hiper düzlem gibi iki boyutlu bir doğrusal yüzey kullanarak sürekli bağımlı değişken ile bağımsız değişken kümes arasındaki ilişkiyi tahmin etmeye çalışır.

Şekil 3.3'te bir Çoklu Regresyon modelinin grafiksel gösterimi yer almaktadır.



Şekil 3.3 Çoklu Regresyon modelinin grafiksel gösterimi

Çoklu Regresyon Modeli:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 \quad (3.8)$$

m değişkenli durumda

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \varepsilon \quad (3.9)$$

$\beta_0, \beta_1, \dots, \beta_m$ model parametrelerini, ε ise hata terimini temsil eder.

Hata Terimi ile ilgili Varsayımlar

1. Hata terimi ortalaması ve beklenen değeri 0 olan bir rassal değişkendir.

$$E(\varepsilon) = 0 \quad (3.10)$$

2. Hata teriminin σ_2 ile gösterilen varyansı $x_1, x_2, \dots, x_m$ değerlerinden bağımsız olarak sabittir.

3. Hata teriminin değerleri bağımsızdır.

4. Hata terimi ε Normal dağılımlı bir rassal değişkendir. Hata terimi değerleri ε_i bağımsız normal rassal değişkendirler. Ortalaması ve varyansı 0'dır [25].

Veri madencisi, çoklu regresyon modelinde hangi değişkenlerin yer alması gerektiğine karar vermek durumundadır. Bu karar verilirken aşağıdaki değişken seçim metotlarından biri kullanılabilir :

- (1) İleriye Doğru Seçim (Forward selection),
- (2) Geriye Doğru Ayıklama (Backward Elimination),
- (3) Adım Adım Seçim (Stepwise selection),
- (4) En iyi Alt kümeler (Best subsets).

Çoklu regresyon analizi, veri kümesini bir bütün olarak homojen kabul etmektedir. Bu nedenle, araştırmalarda bu yöntemlerle tahminlenen parametrelerin güvenilirliği ve evrene genellenebilirliği tartışılır olmuştur [26].

Bağımsız değişken(ler) ile bağımlı değişken arasındaki ilişkiyi inceleyen regresyon analizi (basit ve çoklu), bir takım varsayımların (doğrusallık, normallik, homojenlik, toplanabilirlik gibi) yerine getirilmesinden sonra uygulanabilen bir test istatistiğidir. Varsayımların yerine gelmemesi durumunda bilinen bir takım dönüştürme (transformation) işlemlerine tabi tutulacak olan veri seti, uygun hale getirilmeye çalışılmaktadır. Bu da veri kümesindeki orijinal değerlerin ya logaritmik dönüşümleri, ya da karekök gibi dönüşüm yöntemleri ile yapılabilmektedir [28].

Bu bölümde incelenen Doğrusal ve Çoklu Regresyon modelleri dışında Veri Madenciliği'nde parametrik (Lojistik, Üstel ve Weibull ve yarı parametrik (Cox) regresyon modelleri de kullanılmaktadır.

3.3. Sınıflandırma Metotları

En yaygın uygulanan veri madenciliği metotlarından biri olan Sınıflandırma metotları, sınıfı tanımlanmış mevcut verilerden yararlanarak sınıfı belli olmayan verilerin sınıfını tahmin etmek için kullanılır. Sınıflama iki adımdan oluşur. Birinci adımda tahmin için kullanılacak bir model oluşturulmaktadır.

İkinci adımda, oluşturulan bu model sınıfı belli olmayan veriler üzerinde uygulanarak bu verilerin sınıfları tahmin edilmektedir. Sınıflandırma metotları dört grup altında incelenecektir:

- Karar Ağaçları (Decision Trees)
- Yapay Sinir Ağları (Artificial Neural Networks),
- Saf Bayes Sınıflaması (Naive – Bayes Classification),
- Diğer Sınıflandırma Metotları (Genetik Algoritmalar, K- En Yakın Komsu (K–Nearest Neighbour), SEE5 vb.)

Belirtilen metotlardan Karar ağacı/karar kuralları oluşturma algoritmaları tez kapsamında kullanıldığından, bu gruba giren sınıflandırma algoritmaları detaylı bir şekilde anlatılmıştır. Diğer sınıflandırma metotları bölüm içinde ilgili başlıklar altında kısaca anlatılmıştır.

3.3.1. Karar ağaçları ve karar ağacı algoritmaları

Karar Ağaçları, verileri belli özellik değerlerine göre sınıflandırmaya yarar. Bunun için algoritmaya girdi olarak çeşitli veri özellikleri belirlenir. Çıktı olarak da belli bir veri özelliği seçilerek algoritmanın bu çıktı özelliği değerlerine ulaşmak için hangi girdi özelliklerinin bir arada olması gerektiğini ağaç veri yapıları şeklinde keşfetmesi sağlanır [29].

Karar ağacı algoritmalarını bir probleme uygulayabilmek için aşağıdaki koşulların sağlanması gerekir:

- Olayların özelliklerle ifade edilebilmesi gerekir. Nesnelerin belli sayıda özellik değerleriyle ifade edilebilmesi gerekir. Örneğin; soğuk, sıcak vb. özelliklerin sürekli ve ya kesikli olması fark etmez.

- Sınıfları belirleyebilmek için gereken ayırıcı özelliklerin olması gerekir.

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

•**Karar düğümü:** Veriye uygulanacak test tanımlanır. Her düğüm bir özellikteki testi gösterir. Test sonucunda ağacın dalları oluşur. Dalları oluştururken veri kaybı yaşanmaması için verilerin tümünü kapsayacak sayıda farklı dal oluşturulmalıdır.

•**Dal:** Testin sonucunu gösterir. Elde edilen her dal ile tanımlanacak sınıfın belirlenmesi amaçlanır. Ancak dalın sonucunda sınıflandırma tamamlanamıyorsa tekrar bir karar düğümü oluşur. Karar düğümünden elde edilen dalların sonucunda sınıflandırmanın tamamlanıp tamamlanmadığı tekrar kontrol edilerek devam edilir.

•**Yaprak:** Dalın sonucunda bir sınıflandırma elde edilebiliyorsa yaprak elde edilmiş olur. Yaprak, verileri kullanarak elde edilmek istenen sınıflandırmanın sınıflarından birini tanımlar [11].

Karar ağacının kurulması için kullanılacak girdi olarak bir dizi kayıt verilirse bu kayıtlardan her biri aynı yapıda olan birtakım özellik/değer çiftlerinden oluşur. Bu özelliklerden biri kaydın hedefini belirtir. Amaç, hedef-olmayan özellikler kullanılarak hedef özellik değerini doğru kestiren bir karar ağacı belirlemektir. Hedef özellik çoğunlukla sadece {evet, hayır}, veya {başarılı, başarısız} gibi ikili değerler alır.

Örnek [33]: Golf oynamak isteyen kişiler için hava koşulları kayıtlarıyla ilgilenilsin. Hedef özellik olan golf oynama durumu iki değer alabilsin. Bunlar oynanabilir ve oynanamaz olsun. Çizelge 3.1’de bu örnek için verilmiş özellik ve bunların alabileceği olası değerler verilmiştir.

Çizelge 3.1 Golf örneği için hedef-olmayan özellikler

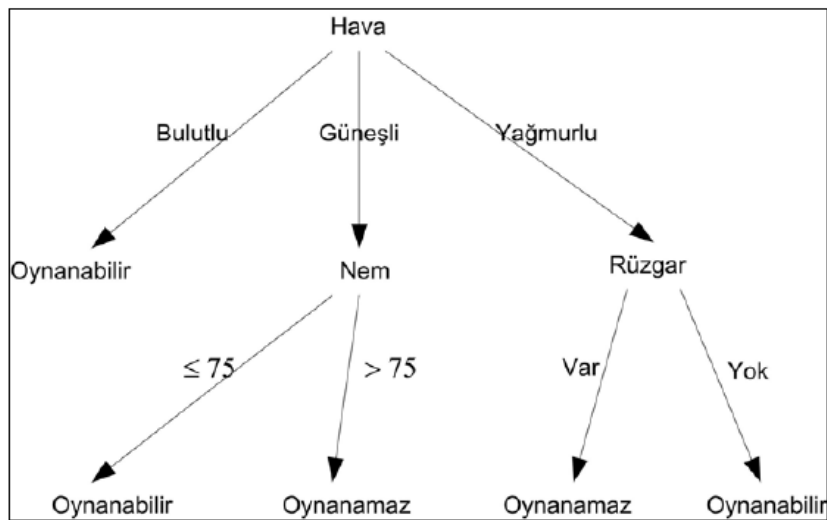
NİTELİK	OLASI DEĞERLER
Hava durumu	Güneşli, Bulutlu , Yağmurlu
Sıcaklık	Sürekli
Nem	Sürekli
Rüzgar	Var , Yok

Çizelge 3.2’de bu özellikler için kullanılan öğrenme verisi verilmiştir.

Çizelge 3.2 Golf örneği için öğrenme verisi [33].

Hava durumu	Sıcaklık (°F)	Nem(%)	Rüzgar	Sonuç
Güneşli	75	70	Var	Evet
Güneşli	80	90	Var	Hayır
Güneşli	85	85	Yok	Hayır
Güneşli	72	95	Yok	Hayır
Güneşli	69	70	Yok	Evet
Kapalı	72	90	Var	Evet
Kapalı	83	78	Yok	Evet
Kapalı	64	65	Var	Evet
Kapalı	81	75	Yok	Evet
Yağmurlu	71	80	Var	Hayır

Çizelge 3.1 ve 3.2’deki veriler kullanıldığında ortaya Şekil 3.4’teki karar ağacı elde edilir.



Şekil 3.4 Golf örneği için Karar ağacı örneği [32]

Bir karar ağacı ne bildiğimizi (örn., öğrenme verisi) özetlediği için değil, yeni durumların sınıflamasını doğru yaptığı için önemlidir.

“Karar ağacı tekniğini kullanarak verinin sınıflanması üç aşamadan oluşur.

•**Öğrenme:** Önceden sonuçları bilinen verilerden (eğitim verisi) model oluşturulur.

•**Sınıflama:** Yeni bir test verisi kümesi modele uygulanır, bu şekilde karar ağacının doğruluğu belirlenir. Test verisine uygulanan bir modelin doğruluğu, yaptığı doğru sınıflamanın test verisindeki tüm sınıflara oranıdır. Her test örneğinde bilinen sınıf, model tarafından tahmin edilen sınıf ile karşılaştırılır.

•**Uygulama:** Eğer doğruluk kabul edilebilir oranda ise, karar ağacı yeni verilerin sınıflanması amacıyla kullanılır [11].

Test verisini en iyi şekilde dallara ayıran özellik tespit edilerek, bu özellik karar ağacının oluşturulmasında daha önce seçilir. Böylece daha iyi bir karar ağacı oluşturulur. En iyi dallara ayıran özelliğin tespitinde çeşitli ölçütler geliştirilmiştir.

Bu ölçütlerin bazı örnekleri şunlardır:

$p(i | t)$, i sınıfına ait verilerin, verilen bir t düğümündeki bölünmesini gösterebilir. c sınıf sayısıdır ve entropi hesaplamasında $0 \log_2 0 = 0$ şeklinde düşünülmüştür.

$$\text{Entropi}(t) = - \sum_{i=0}^{c-1} p(i | t) \log_2 p(i | t), \quad (3.11)$$

Entropi bir düğümün ne kadar bilgi verici olduğunu ölçmede kullanılır. Bu “İyi” ile ne kastedildiğini belirtir. Bu kavram Claude Shannon tarafından ilk kez Bilgi Teorisi içinde tanımlanmıştır.

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i | t)]^2, \quad (\text{Breiman et al., 1984}) \quad (3.12)$$

$$\text{Sınıflandırma hatası}(t) = 1 - \max_i [p(i | t)], \quad (3.13)$$

Karar ağacı oluşturulduktan sonra, kökten yaprağa doğru inilerek kurallar belirlenir. Ancak çok sayıda dal ve düğümden oluşan karar ağaçlarında kuralların belirlenmesi zorlaşır. Karar ağacı modelini daha okunabilir hale getirmek için kuralların yazımında IF-THEN (Eğer-ise-O Zaman) biçiminde şartlı ifadeler kullanılır. IF (Eğer-ise) kısmı dalın sonuna giden yol üzerindeki tüm testleri içerirken THEN (O Zaman) kısmı son sınıflamayı gösterir. IF-THEN yapısındaki kurallara *Karar Kuralları (Decision Rules)* denir.

Örneğin Şekil’de oluşan basit karar ağacından yola çıkılarak;

Eğer Hava=Bulutlu ise O Zaman Golf Oynama= Oynanabilir

Eğer Hava=Güneşli ve Nem>75 ise O Zaman Golf Oynama= Oynanamaz

Eğer Hava=Güneşli ve Nem<=75 ise O Zaman Golf Oynama= Oynanabilir

Eğer Hava=Yağmurlu ve Rüzgar=Yok ise O Zaman Golf Oynama= Oynanabilir

Eğer Hava=Yağmurlu ve Rüzgar=Var ise O Zaman Golf Oynama= Oynanamaz karar kurallarını yazılabilir.

Bu kurallar uygulama konusunda uzman bir karar verici tarafından incelenerek sonucun anlamlı olup olmadığı anlaşılır.

Karar ağaçları, model kurulumu ve sonuçlarının yorumlanmasının kolay olması, veritabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir [30].

Karar ağaçlarının güçlü yönleri aşağıdaki gibi özetlenebilir:

- * Karar ağaçları anlaşılabilir kurallar üretirler.
- * Karar ağaçları aşırı hesaplama gerektirmeden sınıflandırma yaparlar.
- * Karar ağaçları hem sürekli ve hem de kesikli değişkenler için uygundur.
- * Karar ağaçları sınıflandırma ve kestirim için hangi alanların en önemli olduğunu açık biçimde gösterir.

Karar ağacı temelli analizlerin yaygın olarak kullanıldığı alanlar şunlardır [4]:

- Belirli bir sınıfın olası üyesi olacak elemanların belirlenmesi,
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması,
- Parametrik modellerin kurulmasında kullanılmak üzere çok sayıdaki değişkenden en önemlilerinin seçilmesi,
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması,
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması,
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikli değişkenlere dönüştürülmesi.

Karar Ağacı Algoritmaları ID3, C4.5 ve C5, CART, CHAID ve Diğer Karar Ağacı Algoritmaları olmak üzere beş ana başlık altında incelenecektir.

En sık kullanıma sahip karar ağacı modelleri ID3 ve daha gelişmiş modeli C4.5, Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees-CART) (Breiman vd., 1984) ve Otomatik Ki-kare Etkileşim Keşfedicisi (Dedektörü)-OKEK(D)(Chi-Square Automatic Interaction Detector-CHAID) (Ribic ve Miller, 1998) 'dir [31].

ID.3 algoritması:

ID3 (Induction of Decision Trees) Algoritması, 1986'da Quinlan tarafından geliştirilmiş bir karar ağacı algoritmasıdır.

Karar ağacında, her bir düğüm hedef-olmayan bir özelliğe, düğümler arasındaki her ok (*arc*) ise özelliğin olası bir değerine karşılık gelir. Ağacın bir yaprağı, bu yapraktan köke kadar ki yolda tanımlanan kayıtlar için hedef özelliklerin beklenen değerini belirler.

Ağaçta her bir düğüm, kökten başlayarak yol üzerinde henüz dikkate alınmamış özellikler arasından en büyük kazançlısı seçilerek belirlenir. Bu düzenlemenin amacı küçük karar ağaçları oluşturmaktır. Böylece bir kayıtın hangi sınıfa girdiği sadece bir kaç sorudan sonra belirlenebilir.

ID3 Algoritması, test edilecek en kazançlı niteliği seçmek için Kazanç (Gain) adlı bir ölçüt kullanır.

KAZANÇ (Gain): Bir T elemanını tanımak için gereken bilgi ile X özelliği de kullanılarak T elemanını tanımak için gereken bilgi arasındaki farktır. Bu aynı zamanda, X özelliğine dayalı bilgi kazancı olarak adlandırılır ve aşağıdaki gibi ifade edilir.

$$\text{Kazanç}(X, T) = \text{Bilgi}(T) - \text{Bilgi}(X, T) \quad (3.14)$$

Örnek : Bir özellik için bilgi kazancının hesaplanması [32]

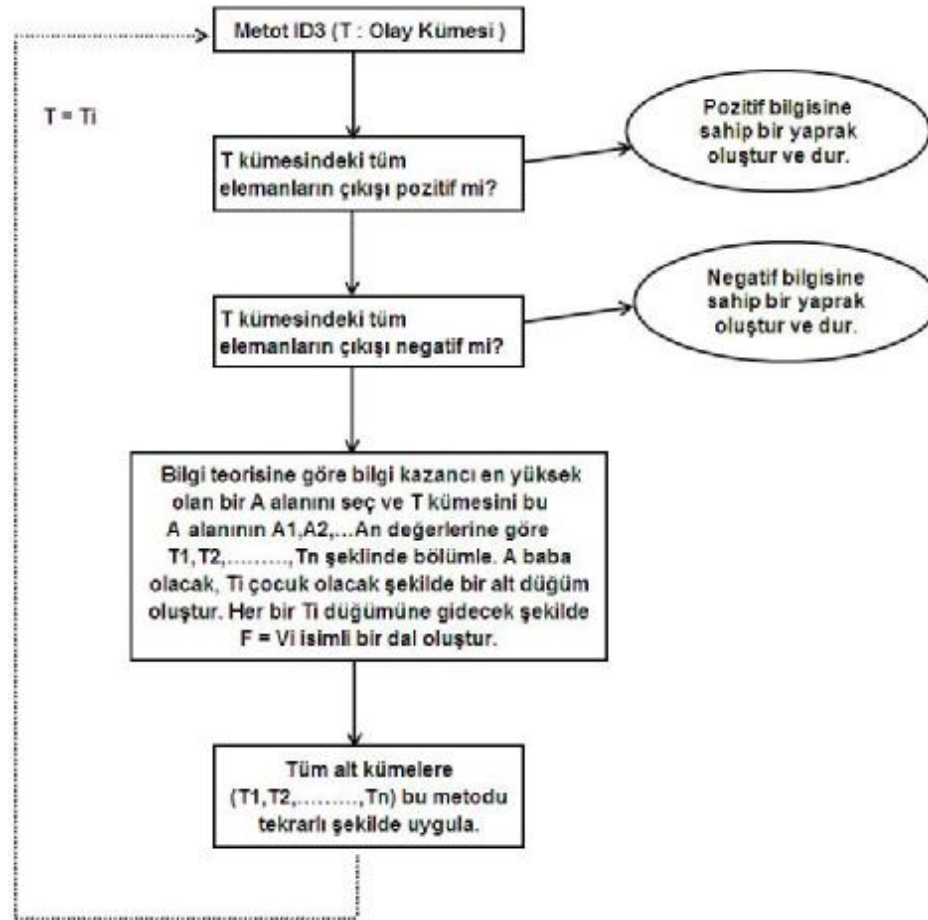
$\{A_1, A_2, \dots, A_n\}$ değerlerine sahip A özelliği ağacın bölünmesi için kullandığında, T kümesi de $\{T_1, T_2, \dots, T_n\}$ şeklinde bölünecektir. Bu bölümlenmede T kümesindeki A özelliğinin A_i olduğu bölgelere T_i kümesi densin. Bu kümedeki pozitif olayların sayısı p_i , negatif olayların sayısı n_i olarak gösterildiğinde T_i alt ağacı için beklenen bilgi gereksinimi $I(p_i, n_i)$ olur. T ağacı için beklenen bilgi kazancı tüm T_i ağaçlarının beklenen bilgi kazançlarının ağırlıklı ortalamalarının toplamı olur ve

$$E(A) = \sum_{i=1}^n \frac{p_i + n_i}{p + n} I(p_i, n_i) \quad (3.15)$$

şeklinde hesaplanır. Dolayısı ile A özelliği üzerinden sağlanan bilgi kazancı aşağıdaki eşitlikle ifade edilir.

$$\text{Bilgi kazancı (A)} = I(p,n) - E(A) \quad (3.16)$$

Şekil 3.5'te ID3 algoritması adımları verilmiştir.



Şekil 3.5 ID3 Algoritması [31].

C4.5 algoritması, ID3 algoritmasında sayısal değerlere sahip bir özellik için sonuç alınamaması nedeniyle ID3 algoritmasını da geliştiren Quinlan tarafından 1993 yılında geliştirilmiştir. ID3 algoritmasından farkları aşağıda özetlenmiştir:

- Sayısal değerler ile çalışılırken bir eşik değeri belirlenir. Bu eşik değeri bulunurken özelliğin değerleri sıralanır ve $[v_i, v_{i+1}]$ aralığının orta noktası

alınır. Bu değer t eşik değeri olarak belirlenir ve özellik değeri bu t eşik değerinden büyük veya küçük eşit olmak üzere ikiye ayrılır.

-C4.5 kayıp verilere sahip örneklerde bir düzeltme faktörü kullanır.

- F faktörü kullanılarak kazanç ölçütü düzeltilir.

F faktörü Veri tabanında değeri bilinen niteliğe sahip örneklerin sayısının Veri tabanındaki tüm örneklerin sayısına bölümüyle elde edilir.

J4.8 algoritması, C4.5 algoritmasının WEKA grubu tarafından [www.cs.waikato.ac.nz/ml/weka/, 2002] Java'da kodlanan 8. versiyonudur (*Revision* 8) [32]. Uygulama çalışmasının karar ağacı / kuralı oluşturma aşamasında WEKA yazılımında yer alan J4.8 algoritması ile beraber RandomTree ve Part algoritmaları kullanılmıştır.

Sınıflandırma ve regresyon ağacı (Classification and Regression Trees-CART) metodu, analiz edilen verilerin özeline göre farklı isimlerle anılmaktadır. Bağımlı değişkenin kategorik olması durumunda Sınıflandırma Ağacı (Classification Tree; CT), sürekli değişken olması durumunda da Regresyon Ağacı (Regression Tree; RT) kullanılmaktadır. 1984 yılında Breiman ve arkadaşları tarafından geliştirilmiştir [3].

Sınıflandırma ve regresyon ağacı metodu, Çoklu regresyon analizindeki gibi veri setini bir bütün olarak homojen kabul etmez. Veri setindeki heterojenliği de göz önünde bulundurarak, homojen alt sınıflar için bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi modeller ve bu ilişkiyi ağaç yapısında diyagram (alt ağaçlar; node) biçimindeki göstererek diğer regresyon tabanlı metotlar içinde avantajlı bir konum elde eder [34].

Sınıflama ve regresyon ağacı yöntemi, sadece bağımlı değişken ile bağımsız değişken arasındaki ilişkinin yapısını araştırmakla kalmaz; aynı zamanda bağımsız değişkenlerin birbirleri ile olan etkileşimlerini de ortaya koymaya çalışmaktadır. Yöntemin kullandığı algoritma, benzerlik gösteren (similarity) değişkenlerin aynı ağaç düğümünde toplanmasına dayalı olup, bütün

oluşturduğu alt dalları (sub-branches) bağımlı değişken olan kök düğüme bağlamayla son bulmaktadır.

Sınıflama ve regresyon ağacı yöntemine ait bu algoritma temelde 3 unsurdan meydana gelmektedir. Bunlar, “ağacın oluşturulması”, “budama (pruning)” ve “en uygun ağaç yapısının seçimi” şeklindedir.

Maksimum düzeyde homojen alt sınıflar (ağaçlar) oluşturma ilkesine dayalı çalışan sınıflama ve regresyon ağacı algoritması, “ağacın oluşturulması” kısmında olabilecek en fazla sayıda alt ağaçları belirlemektedir. Algoritmanın ikinci kısmı olan “Budama” kısmında alt ağaçlar arasında bağımlı değişkenle önemli ilişkiler gösteren ağaçları seçmek gerekmektedir. Böylece budamadan sonra “en uygun ağaç yapısının seçimi” ile regresyon ağacı elde edilebilmektedir [12].

Sınıflandırma ve regresyon ağacı yöntemi daha çok tıp, endüstri ve mühendislik bilimlerinde yaygın bir şekilde kullanılmaktadır [33].

CHAID (Kikare Otomatik Etkileşim Bulucu) metodu, 1980’de Kaas tarafından geliştirilmiştir [38]. Bir bağımlı değişken ile bir dizi bağımsız değişken arasındaki ilişkilerin belirlenmesine yönelik çalışmalarda kullanılan açıklayıcı bir metottur. Bir başka ifade ile bağımlı değişkenin farklı biçimlerde tahminini sağlayan bağımsız değişkenlerden yararlanarak temel yapıların nasıl olduğunu gösteren bir sınıflandırma aracıdır. En uygun bölümleri seçmek için kullanılan entropi veya gini metrikleri yerine ki-kare testi kullanılmaktadır.

CHAID algoritması aşağıda verilmiştir [36]:

Adım 1. Her bir tahmin edici değişken X için, X ’in, Y hedef değişkenini dikkate alan en az öneme sahip (En büyük p değerine sahip) kategori çiftini bul. Yöntem, Y ’nin ölçüm düzeyine bağlı olarak p değerlerini hesaplayacaktır.

a. Eğer Y sürekli ise F testini kullan.

- b. Eğer Y isimsel ise X'in kategorileri satırlarda ve Y'nin kategorileri sütunlarda olacak biçimde iki yönlü çapraz tablo düzenle. Pearson ki-kare testini veya olabilirlik oranı testini kullan.
- c. Eğer Y sıralı ise bir Y birliktelik modeli (Clogg ve Eliaisin, 1987; Goodman, 1979; Magidson, 1992) uydur. Olabilirlik oranı testini kullan.

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

- a. Eğer p değeri *abirleş* 'den büyük ise bu çifti bir tek kategori altında birleştir. X'in yeni kategori kümesi için süreci Adım 1'den başlat.
- b. Eğer p değeri *abirleş* 'den küçük ise Adım 3'e git.

Adım 3. X'in ve Y'nin kategori kümesi için uygun Bonferroni düzeltmesini kullanarak, düzeltilmiş p değerini hesapla.

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

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

CHAID ile diğer yöntemler arasındaki en önemli farklardan biri, ağaç türetimidir. ID3, C4.5 ve CART ikili ağaçlar türetirken, CHAID ikili olmayan çoklu ağaçlar türetir. CHAID sürekli ve kategorik tüm değişken tipleriyle çalışabilmektedir. Bununla beraber, sürekli tahmin edici değişkenler otomatik olarak analizin amacına uygun olarak kategorize edilmektedir. CHAID, Ki-Kare metriği vasıtasıyla, ilişki düzeyine göre farklılık rastlanan grupları ayrı ayrı sınıflamaktadır. Dolayısıyla, ağacın yaprakları, ikili değil, verideki farklı yapı sayısı kadar dallanmaktadır [36].

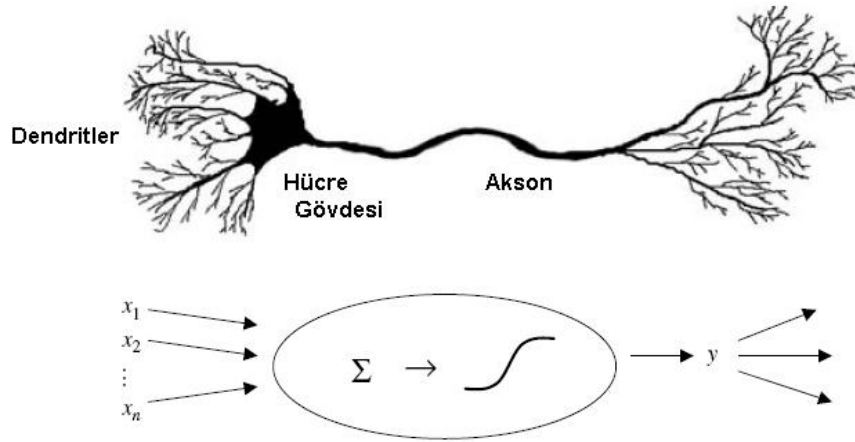
Karar Ağacı Algoritmaları başlığı altında detaylı incelenen belli başlı algoritmalar dışında pek çok karar ağacı algoritması mevcuttur. Bu algoritmalara örnek olarak MARS (Multivariate Adaptive Regression Splines), QUEST (Quick, Unbiased, Efficient Statistical Tree), SLIQ (Supervised Learning in Quest), SPRINT (Scalable Parallelizable Induction of Decision Trees), S-Plus, FACT, OC1, LMDT, CAL5, T1 karar ağacı algoritmaları verilebilir [13].

3.3.2. Yapay sinir ağları (Artificial Neural Networks)

Yapay sinir ağları; insan beyninden esinlenerek, öğrenme sürecinin matematiksel olarak modellenmesi sonucu ortaya çıkan bilgisayar sistemleridir. Yapay sinir ağları üzerinde çalışmalar ilk olarak beyni oluşturan biyolojik üniteler olan nöronların modellenmesi ve bilgisayar sistemlerinde uygulanması ile başlamış, bilgisayar sistemlerinin gelişimine paralel olarak pek çok alanda kullanılır hale gelmiştir.

İlk yapay sinir ağı modeli, 1943'te bir sinir hekimi olan Warren McCulloch ile bir matematikçi olan Walter Pitts tarafından gerçekleştirilmiştir. McCulloch ve Pitts, insan beyninin hesaplama yeteneğinden esinlenerek, elektrik devreleriyle basit bir sinir ağı modellemişlerdir.

Şekil 3.6'da gerçek bir nöron (insan sinir hücresi) ve yapay sinir hücresi modeli görülmektedir.



Şekil 3.6 Gerçek ve yapay nöron modeli [37]

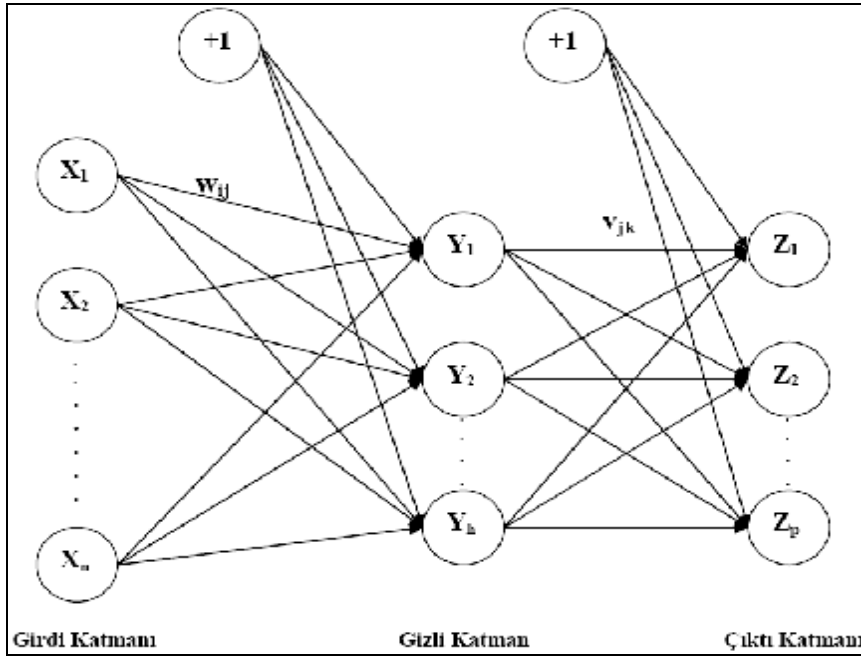
Üç çeşit Yapay sinir ağı tipi vardır :

- a-) *İleri beslemeli Yapay Sinir Ağları* : Her bir katmandaki hücreler sadece bir önceki katmanın hücrelerince beslenir.
- b-) *Basamak bağlantılı Yapay Sinir Ağları* : Hücreler sadece önceki katmanlardaki hücrelerce beslenir.
- c-) *Geri beslemeli Yapay Sinir Ağları* : En az bir hücre sonraki katmanlardaki hücrelerce de beslenir.

Veri madenciliğinde Yapay Sinir Ağları, geçmiş verilerden yararlanarak karar kurallarının tahmininde kullanılır.

Öngörü amaçlı YSA tiplerinden en yaygın kullanılanı geri beslemeli öğrenmeye (backpropagation) sahip *çok katmanlı algılayıcı*dır. Çok katmanlı algılayıcı (ÇKA, Multi Layer Perceptron), dışarıdan verileri alan girdi katmanından, ağın çıktılarını dışarıya veren çıktı katmanından ve bu ikisi arasında genellikle bir ve bazen de daha fazla gizli (ara) katmandan meydana gelmektedir [38].

Şekil 3.7'de tipik bir çok katmanlı algılayıcının yapısı görülmektedir.

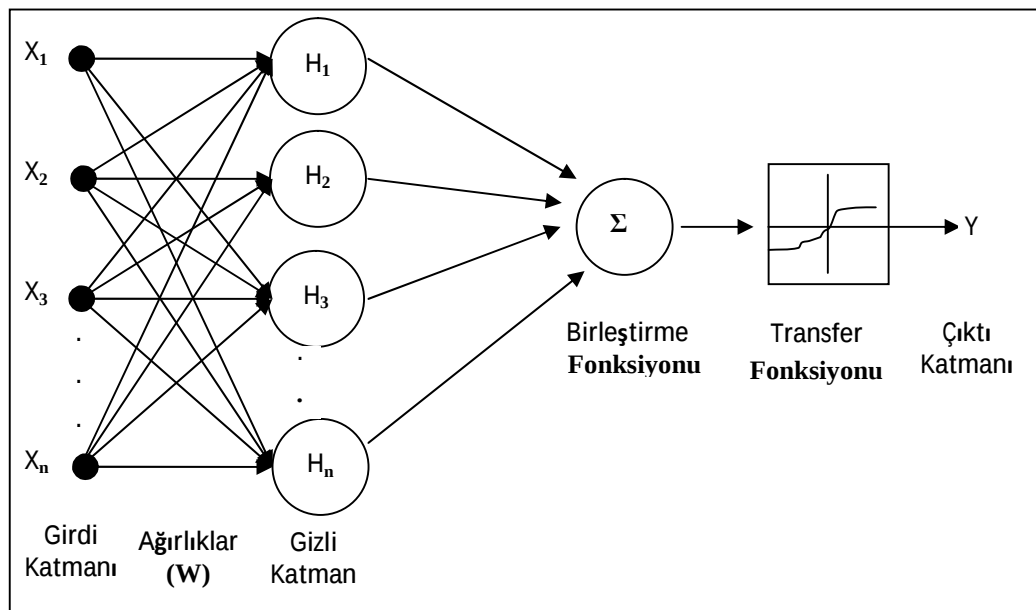


Şekil 3.7 Tipik bir Çok Katmanlı Algılayıcının yapısı [39]

Şekil 3.7’de, $(x_i; i=1, \dots, n)$ girdi katmanı, gizli katman ve çıktı katmanındaki işlem elemanları olan sırasıyla $x_i; i=1, \dots, n$; $z_i; i=1, \dots, n$ ve $y_i; i=1, \dots, n$ görülmektedir. Ayrıca w_{ij} ; i . girdi işlem elemanı ile j . gizli işlem elemanı olan arasındaki bağlantının ağırlığını ve v_{jk} ; j . gizli işlem elemanı ile k . çıktı işlem elemanı arasındaki bağlantının ağırlığını göstermektedir. Eşik değerler +1 olarak gösterilmiştir [40].

Temel olarak bir ÇKA’nın yapısında girdiler arasında bağlantı kuran işlem elemanları, bağlantı ağırlıklarının belirlenmesini sağlayan öğrenme algoritmaları ve doğrusal olmamayı sağlayan transfer fonksiyonu bulunur. Her bir işlem elemanı, diğer işlem elemanlarıyla bağlantılı olup bu bağlantıların kuvvetini gösteren bir ağırlık değerine sahiptir. Bu durum bilginin tüm bu bağlantılar aracılığıyla işlem elemanları ve katmanlar arasında ağırlı çıkışına kadar iletilmesini ve ağırlı dağıtılmış bir hafızaya sahip olmasını sağlamaktadır. İşlem elemanlarının sahip olduğu ağırlık değerleri, gizli katmandaki ve çıktı katmanındaki işlem elemanlarının net girdisinin hesaplanmasında kullanılır [40].

Bir Yapay Sinir Ağından istenen, örnek veri setindeki yapıyı öğrenerek, modelden elde edilmesi hedeflenen genelleştirmelere ulaşmasıdır. Bunu yapabilmesi için ağ, ilgili olayın örnekleri ile eğitilerek genelleme yapılabilecek yeteneğe kavuşturulur. Bu genelleme ile benzer olaylara karşılık gelen çıktı setleri belirlenir. YSA'nın öğrenmesi, işlem elemanlarının sahip olduğu ağırlıkların, seçilen eğitim algoritmalarıyla değiştirilmesi ile yapılmaktadır. Ağın ürettiği sonuçlar ise ağa girilen bilgilerin kendi ağırlıkları ile çarpımlarının toplanması sonucu elde edilen net girdinin, bir transfer fonksiyonu ile işlenmesi ile çıktı katmanından alınmaktadır. Bu işlemler Şekil 3.8'de gösterilmiştir.



Şekil 3.8 Bir ÇKA modelinin Girdi-Çıktı hesaplama süreci [41].

Şekil 3.8'de bir ÇKA'nın gizli katmanında meydana gelen işlemler gösterilmektedir. Girdi katmanında yer alan girdilere ait veriler, gizli katmandaki işlem elemanlarıyla olan bağlantı ağırlıklarıyla çarpıldıktan sonra gizli işlem elemanlarına giriş yapmakta ve burada işlem elemanına gelen net girdiyi hesaplamak amacıyla birleştirme işlevinden geçirilmektedir. Daha sonra, net girdiler bir transfer fonksiyonu yardımıyla işlenerek işlem elemanının girdiye karşılık üreteceği çıktı belirlenmektedir. Gizli katmandaki

transfer fonksiyonundan geçen çıktılar, gizli katmandaki işlem elemanlarından elde edilen net girdiler ise transfer fonksiyonu yardımı ile net çıktılara dönüştürülmekte ve ağıın üretmiş olduğu çıktılar elde edilmektedir [41].

Yapay sinir ağları tabanlı algoritmalara LVQ (Learning Vector Quantization) ve RBF (Radial Basis Function) algoritmaları örnek verilebilir.

Yapay Sinir Ağları kullanmanın getirdiği çeşitli yararlar ve olanaklar aşağıda özetlenmiştir [7]:

- *Doğrusal Olmayan Yapısı* sayesinde doğrusal olmayan gerçek hayattan mekanizmalardan bilgi çıkarımında tercih edilir.
- *Örneklerden Öğrenme Özelliği* ile ara katman ağırlıkları deneme ya da öğrenme veri seti yardımıyla değişebilir ve böylece model eldeki verilere tam olarak uygun hale gelir.
- *Uyarlanabilirliği*: YSA içindeki ağırlıklar model çevresindeki değişimlere göre değiştirilebilir.
- *Güvenilirlik Değeri Özelliği* sayesinde bir YSA, verilecek kararın güvenilirlik değerini de belirtecek şekilde tasarlanabilir.
- *Hata Toleransı Özelliği* ile çalışma performansı nöronlar arasında bağlantısızlık ya da eksik / dağınık veri olması gibi zor işlem durumlarında dahi azalmaz.
- *Analiz ve Tasarımın Uygunluğu* : YSA'ları temel bir bilgi işlem metodu olarak aynı kural, notasyon ve metodolojik adımlarla uygulanır.

Yapay Sinir Ağlarının en önemli dezavantajı öğrenme işlem süresinin uzun ve çıkan sonucun ifade edilmesinin / tanımlanmasının güç olmasıdır [11].

3.3.3. Saf bayes sınıflaması (Naive Bayes Classification)

Eldeki verilerin belirlenmiş olan sınıflara ait olma olasılıklarını öngörür. İstatistikteki Bayes teoremine dayanır. Bu teorem; belirsizlik taşıyan herhangi bir durumun modelinin oluşturularak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler doğrultusunda belli sonuçlar elde edilmesine olanak sağlar. Belirsizlik taşıyan durumlarda karar verme konusunda çok kullanışlıdır.

Bayes yöntemi koşullu olasılık durumları ile ilgilidir. Her hangi bir koşullu olasılık durumu aşağıda belirtilen eşitlik şeklinde tanımlanır :

$$P(X=x | Y=y) = R \quad (3.18)$$

Bu ifade; “Eğer $Y = y$ doğru ise, $X = x$ olma olasılığı R ’dir” anlamına gelmektedir. X ve Y ‘nin alabileceği değerlerin her kombinasyonu için koşullu olasılıkları belirleyen tabloya koşullu olasılık dağılımı adı verilir ve $P(X|Y)$ ile ifade edilir.

Bayes Kuralı şu şekilde tanımlanır.

$$P(X|Y) = P(Y|X) \times P(X) / P(Y) \quad (3.19)$$

Bu ifade; Y ’nin gerçekleşmesi halinde X ’in gerçekleşme ihtimalinin ne olduğunu belirtmektedir. Bu değeri bulabilmek için “ X ’in gerçekleştiği durumlarda Y ’nin gerçekleşme ihtimali” ile X ’in gerçekleşme ihtimalini çarpmak ve bunu Y ’nin gerçekleşme ihtimaline bölmek gereklidir.

Belirsiz durumlarda tahmin ve sınıflandırma yapmak için kullanılır. En önemli dezavantajı değişkenler arası ilişkinin modellenmemesi ve değişkenlerin birbirinden tamamen bağımsız olduğu varsayımıdır [25].

3.3.4. Diğer sınıflama metotları

Diğer sınıflama metotları başlığı altında k-en yakın komşu algoritmasından ve Genetik Algoritmalarından kısaca bahsedilecektir:

K-en yakın komşu (KNN) algoritması, yeni bir örnek geldiğinde var olan öğrenme verisi üzerinde sınıflandırma yapmak amacıyla kullanılan bir eğitilmiş öğrenme algoritmasıdır.

Algoritma, yeni bir örnek geldiğinde, onun en yakın K komşusuna bakarak örneğin sınıfına karar verir. Örneğin, a ve b isimli iki niteliğimiz olduğunu ve bu özelliklerin Çizelge 3.3'te yer alan değerlere sahip olduğunu varsayalım.

Çizelge 3.3 a ve b özelliklerinin değerleri

a	b	Sınıf
8	12	Evet
6	10	Evet
7	21	Hayır
4	9	Hayır

a niteliğine ait 12, b niteliğine ait 10 değerine sahip yeni bir örnek için aşağıdaki K-en yakın komşu (KNN) algoritması adımları gerçekleştirilmelidir [42]:

1. K değerini seç.
2. Tüm öğrenme örnekleri ile olan uzaklığını hesapla. (Bunun için uzaklık ölçütlerinden biri kullanılabilir. En çok kullanılan ölçüt Euclid Uzaklık ölçütüdür.)

$$d(p, q) = \sqrt{\sum_i (p_i - q_i)^2} \quad (3.20)$$

3. Minimum uzaklığa göre sırala.
4. Ait oldukları sınıf değerlerini bul
5. Değeri baskın olan sınıfı seç.

Genetik algoritmalar, evrimsel sürecin simülasyonunu bilgisayarda gerçekleştirerek olası çözümleri rastgele arama tekniğidir. İlk defa Michigan Üniversitesi'nde John Holland ve çalışma arkadaşları tarafından geliştirilmiştir.

Optimizasyon problemlerinde karışık değişkenlerin (sürekli ve kesikli) olduğu veya araştırma alanında süreksizlikler mevcut olduğu durumlarda iyi bir çözüm oluşturmaktadır [43].

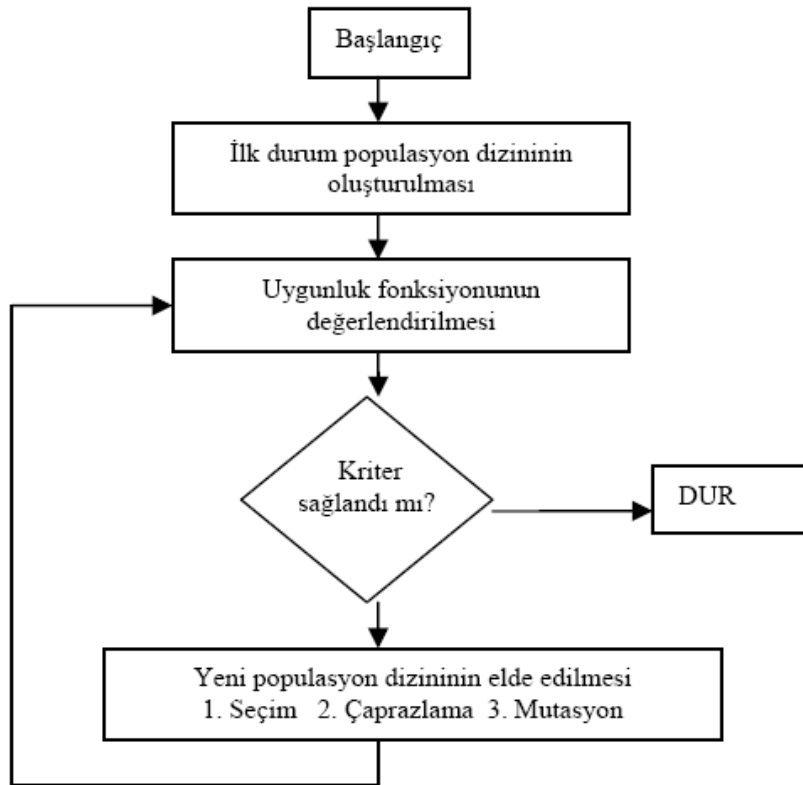
Genetik algoritmalar, çözüm topluluğuna adım adım genetik operatörler uygulayarak ve uygun topluluktan arama yoluyla yeni nesiller üreterek en iyi çözümlere ulaşılmasını sağlar. Genetik algoritmaların çoğu üç ana operatörden oluşur :

i. Seçim (Selection): Seçim, uygunluk değerini temel alarak, popülasyondan uygunluk değeri düşük olan bireylerin elenmesi ve yerlerine uygunluk değerleri yüksek bireylerin kopyalarının konmasıdır. Uygunluk değeri; hangi bireyin sonraki topluluğa taşınacağını belirler.

ii. Çaprazlama (Crossover): Çaprazlama, iyi uygunluk değerine sahip iki bireyin iyi özelliklerinin rassal olarak birleştirilerek daha iyi sonuçlar elde edilmesidir. Genetik algoritmalarından sınıflandırma algoritması olarak faydalanılmasını bu özelliği sağlar.

iii. Mutasyon (Mutation): Genetik algoritmalarda çeşitlendirmeye gitmede mutasyondan faydalanılır. Mutasyonda genlerden biri rassal olarak değiştirilir.

Genetik algoritmanın durması için sağlanması gereken kriter; genellikle belli bir uygunluk değerinin yakalanması veya belirlenen sayıda döngünün sağlanması olarak verilmektedir. Bir genetik algoritma süreci genel olarak aşağıdaki Şekil 3.9'daki gibidir:



Şekil 3.9 Genetik algoritma süreci [43]

3.4. Kümeleme Analizi

Kümeleme, veriyi birbirlerine benzeyen elemanlardan oluşan sınıflara (kümelere) ayırarak, heterojen bir veri grubundan, homojen alt veri grupları elde edilmesi işlemidir.

Kümelemenin temel hedefleri arasında; geniş veri yığınları için tanımlayıcı veriler belirleyerek, işlenecek veri hacmini daraltmak, veri yığınlarındaki doğal kümeleri ortaya çıkarak aynı kümede olması gereken verileri belirlemek, belirlenmiş kümelerin dışında kalan istisna durumları tanımlamak sayılabilir.

Kümeleme algoritmaları; küme içinde benzerliğin en büyüklenmesi (küme içi uzaklıkların en azlanması), kümeler arası benzerliğin en azlanması (kümeler arası uzaklıkların en büyüklenmesi) kavramına dayanır.

Kümeleme fonksiyonu ile sınıflandırma fonksiyonu arasındaki en önemli fark, kümelemenin önceden tanımlanmış girdilere dayanmıyor olmasıdır. Sınıflandırma fonksiyonunda tanımlı girdiler ve bunların geçmişte aldıkları değerler temel modeli oluştururken, kümeleme fonksiyonunda önceden tanımlanmış girdiler ve örnekler yoktur. Veriler kendi içlerindeki benzerliklere göre gruplanırlar. Benzerliği tanımlayacak boyutlar ve özellikler modeli kuran tarafından öngörülür [11].

Kümeleme analizlerinde, sonuç kümelemelerinin değerlendirilmesi kümeleme modeli geliştirme işleminin ayrılmaz bir parçasıdır. Bir veri kümesinde, veriler küme oluşturacak benzerlikte olmasa bile kümeleme algoritmaları bu veri seti içerisinde istenilen sayıda küme bulacaktır. Bu nedenle kümeleme algoritmalarının sonuçlarının değerlendirilmesine yönelik çeşitli sayısal küme doğrulama (cluster validity) yöntemleri geliştirilmiştir. Küme doğrulama yöntemi kullanımı sayesinde küme kalitesi ve uygun küme sayısı belirlenerek kümeleme işlemi başarıyla tamamlanabilir [44].

Kümeleme algoritması tarafından üretilen sonuçların değerlendirilmesine yönelik içsel (internal), dışsal (external) ve görelî (relative) olmak üzere, 3 farklı kritere göre hesaplanmış, çeşitli küme doğrulama (cluster validity) yöntemleri geliştirilmiştir.

Bunlara örnek olarak Silhouette, Davies-Bouldin, Dunn, Calinski-Harabasz, Krzanowski ve Lai ve Hartigan küme doğrulama (cluster validity) endeksleri sayılabilir [45].

Kümeleme yöntemleri genel olarak beş farklı şekilde sınıflandırılabilir:

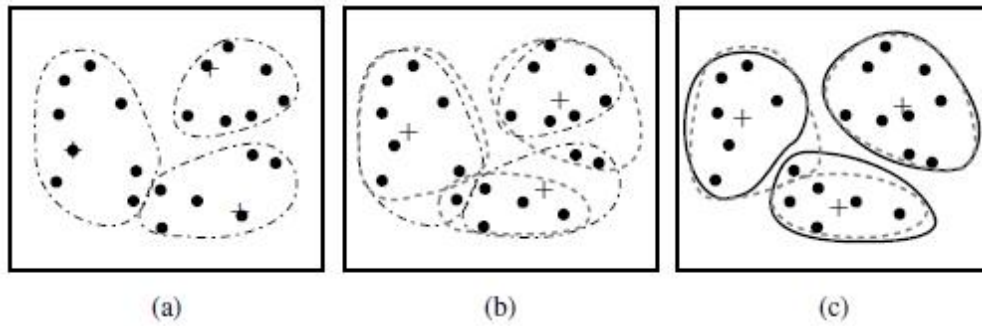
(1) Bölümlemeli (Partitioning) Kümeleme Metotları : Bölümlemeli kümeleme metotlarında, n veri tabanındaki nesne sayısı ve k oluşturulacak küme sayısı olarak kabul edilir. Bölümleme algoritması n adet nesneyi, k adet kümeye böler ($k \leq n$). Her bir bölüm bir kümeyi gösterir. Kümeler tarafsız bölme ölçütü olarak nitelendirilen bir ölçüte uygun oluşturulduğu için aynı kümedeki nesneler birbirlerine benzerken, farklı kümedeki nesneler birbirinden farklıdır. K-Means (K-Ortalama) Kümeleme metodu, k-medoid (K-orta nokta) Kümeleme metodu, Ortogonal Kümeleme (O-Cluster) algoritması, PAM algoritması, CLARA algoritması, CLARANS algoritması geliştirilen bölümlemeli kümeleme metotlarından birkaçıdır. K-means (K-Ortalama) Kümeleme metodu tez kapsamında kullanıldığından, aşağıda detaylı bir şekilde anlatılmıştır.

K-Means (K-Ortalama) kümeleme metodu :

K-Means algoritması, veritabanındaki n tane nesnenin k adet kümeye bölünmesini sağlar. Kümeleme sonucu küme içi (intra-cluster) elamanlar arasındaki benzerlikler çok iken, kümeler arası (inter-cluster) elamanları arasındaki benzerlikler çok düşüktür. Kümeleme sürecinde türüne özgü olarak hata kareleri ölçütü (square-error criterion) toplamı kullanılır.

K-means yöntemi, ilk önce n adet nesneden rastgele k adet nesne seçer ve bu nesnelerin her biri, bir kümenin merkezini veya orta noktasını temsil eder. Geriye kalan nesnelerden her biri kendisine en yakın olan küme merkezine göre kümelere dağılırlar. Yani bir nesne hangi kümenin merkezine daha yakın ise o kümeye yerleşir. Ardından her küme için ortalama hesaplanır ve hesaplanan bu değer o kümenin yeni merkezi olur. Bu işlem tüm nesneler kümelere yerleşinceye kadar devam eder.

Bir nesne grubunun, Şekil 3.10'da görüldüğü gibi uzayda konumlanmış olduğu varsayılın. Kullanıcının bu nesneleri üç kümeye ayırmak istediği varsayılırsa, $k=3$ olur. Şekil 3.10 (a)'da ilk önce rastgele üç nesne, üç kümenin merkezi olarak seçilmiş ve + şeklinde gösterilmiştir. Diğer nesneler de bu merkezlere olan yakınlıklarına göre üç kümeye ayrılmışlardır. Bu ayrıma göre her üç kümenin nesnelerinin yeni ortalaması alınmış ve bu değer kümelerin yeni merkezleri olmuştur. Bu yeni merkezler Şekil 3.10 (b)'de + işareti bulunan noktalarla gösterilmektedir. Algoritma farklı kümelere eşit ya da benzer yakınlıktaki nesneleri bir kümeye atar ve açıkta bir nesne kalmayınca durur. Her nesnenin içinde bulunduğu kümenin merkezine en yakın durumda bulunduğu durumda k-means yöntemi ile kümelere bölünme işlemi sona erer. Son durum Şekil 3.10 (c)'de görülmektedir [5].



Şekil 3.10 k-means yöntemiyle kümeleme örneği [5]

K-means algoritmasının en büyük problemi başlangıçta merkez noktayı belirlemek için seçilen nesnelerin küme içindeki seçiciliğiyle ilgilidir. Eğer kötü bir seçim yapılırsa nesnelerin kümelenmesindeki değişiklik çok sık olur ve farklı sonuçlar doğurabilir [10].

K-means yöntemi, sadece kümenin ortalaması tanımlanabildiği durumlarda kullanılabilir, kullanıcıların k değerini, yani oluşacak küme sayısını belirtme gerekliliği bir dezavantaj olarak görülebilir. Esas önemli olan dezavantaj ise dışarıda kalanlar (outliers) olarak adlandırılan nesnelere karşı olan duyarlılıktır. Değeri çok büyük olan bir nesne, dâhil olacağı kümenin

ortalamasını ve merkez noktasını büyük bir derecede değiştirebilir. Bu değişiklik kümenin hassasiyetini bozabilir. Bu sorunu gidermek için kümedeki nesnelerin ortalamasını almak yerine, kümede ortaya en yakın noktada konumlanmış olan nesne anlamındaki medoid kullanılabilir [46].

k-means kümeleme metodu doğrusal hesaplama kolaylığı nedeniyle doğrusal olmayan yapıdaki hiyerarşik kümeleme metotlarından daha çok kullanılmaktadır. Kolay uygulanabilirliği, hesaplama zorluğunun düşük olması, kolay yorumlanabilir sonuçlar vermesi, farklı verilere uyarlanabilmesi k-means algoritmasının avantajlarıdır [10].

(2) Hiyerarşik Kümeleme Metotları : Veri kümelerini ya da nesneleri önceden belirlenmiş bir kritere göre hiyerarşik olarak ayıran metotlardır. Bu sınıfa giren kümeleme algoritmalarına SLINK Algoritması ve Tek Bağlantı Tekniği, CURE algoritması, CHAMELEON algoritması, BIRCH algoritması örnek verilebilir.

(3) Yoğunluk Tabanlı (Density-based) Kümeleme Metotları : Nesnelerin yoğunluğuna göre kümeleme oluşturan metotlardır. DBSCAN (Density-Based Spatial Clustering of Application with Noise) algoritması, OPTICS (Ordering Points To Identify The Clustering Structure) algoritması, DENCLUE (Density Based Clustering) algoritması yoğunluk tabanlı metotlar arasında sayılabilir.

(4) Model Tabanlı (Model-based) Kümeleme Metotları (Model-base): Her kümenin bir modele uyduğu varsayılarak bu modellere uyan verilerin gruplandırıldığı metotlardır. Kendini Düzenleyen Haritalar (Kohonen Ağları, SOM-Self Organising Maps), COBWEB, CLASSIT algoritması model tabanlı metotlardandır.

(5) Izgara Tabanlı (Grid-based) Kümeleme Metotları : STING (Statistical Information Grid Approach) algoritması, Dalga Kümeleme (Wave Cluster), CLIQUE (Clustering High-Dimensional Space) algoritması ızgara tabanlı kümeleme metotlarından birkaçıdır [5].

Bazı kümeleme algoritması uygulamaları aşağıda verilmiştir:

- Marketlerde farklı müşteri gruplarının keşfedilmesi ve bu grupların alışveriş
- örüntülerinin ortaya konması,
- Biyolojide bitki ve hayvan sınıflandırmaları ve işlevlerine göre benzer genlerin sınıflandırılması,
- Şehir planlanmasında evlerin tiplerine, değerlerine ve coğrafik konumlarına göre gruplara ayrılması
- Web üzerinde bilgi keşfi için dokümanların sınıflanması

3.5. Birliktelik Kuralları

Birliktelik kuralları, işlemlerden oluşan ve her bir işlemin de elemanlarının birlikteliğinden oluştuğu düşünülen bir veri tabanında, bütün birliktelikleri tarayarak, sık tekrarlanan birliktelikleri veritabanından ortaya çıkarma teknikleridir.

Birliktelik kuralı çıkarmada kullanılmak üzere birçok algoritma geliştirilmiştir. Bu algoritmalarda birliktelik kuralı çıkarım problemi Minimum Destek (support) ve güven (confidence) ölçütlerini sağlayan iki adımda çözülmektedir.

Birinci adım, kullanıcı tarafından belirlenmiş minimum destek ölçütünü sağlayan ürün kümelerinin bulunmasıdır. Minimumu Destek (s) birliktelik içeren işlemlerin sayısının veritabanındaki tüm işlemlerin sayısına oranıdır. Minimum destek ölçütünü sağlayan işlem kümelerine “sık geçen eleman kümesi” adı verilmektedir. Verilen örnekleme N adet eleman var ise, potansiyel olarak 2^N adet sık geçen eleman kümesi olabilir. Bu adımda üstel arama uzayını etkili biçimde tarayarak sık geçen eleman kümelerini bulan etkili yöntemler kullanılmalıdır.

İkinci adım ise sık geçen eleman kümeleri kullanılarak minimum güven ölçütünü sağlayan birliktelik kurallarının bulunmasıdır. Bu adımdaki işlem oldukça basittir ve şöyle yapılmaktadır. Sık geçen her bir eleman kümesi için, boş olmayan I 'in tüm alt kümeleri üretilir. I 'in boş olmayan alt kümeleri a ile gösterilsin. Her a kümesi için $a \Rightarrow (I - a)$ durumu, I kümesinin destek ölçütünün a kümesinin destek ölçütüne oranı minimum güven ölçütünü sağlıyorsa $a \Rightarrow (I - a)$ birliktelik kuralı olarak üretilir. Minimum destek ölçütüne göre üretilen çözüm uzayında, minimum güven ölçütüne göre taranarak bulunan birliktelikler kullanıcının ilgilendiği ve potansiyel olarak önemli bilgi içeren birlikteliklerdir.

Birliktelik Kuralları Eğer-İse ifadeleri ile aşağıdaki formatta yazılır:

Eğer <bazı şartlar sağlanırsa> *İse* <bazı özelliklerin değerlerini tahmin et>

Örnek: Eğer Ürün=Şapka İse Kar=Medium. Güven %80 Destekleyen 25 satır

Birliktelik sorgusu algoritmalarının performansı birinci adım belirler. Bu adımda sık geçen öge kümeleri belirlendikten sonra birliktelik kurallarının bulunması adımları karmaşık değildir [47].

Birliktelik kuralına ilişkin olarak geliştirilen bazı algoritmalar arasında AIS (Agrawal vd., 1993), SETM (Houtsma ve Swami, 1995), Apriori (Agrawal ve Srikant, 1994), Partition (Savasere vd., 1995), RARM - Rapid Association Rule Mining (Das vd., 2001), CHARM (Zaki ve Hsiao, 2002) sayılabilir. Bu algoritmalar içerisinde, ilk olanı AIS, en bilineni ise Apriori algoritması (Agrawal ve Srikant, 1995) dir [48].

Birliktelik kuralı çıkarım algoritmaları içerisinde en fazla bilineni Apriori Algoritması'dır. Bu algoritmada sık geçen eleman kümelerini bulmak için birçok kez veritabanını taramak gerekir.

İlk taramada bir elemanlı minimum destek metriğini sağlayan sık geçen eleman kümeleri bulunur. İzleyen taramalarda bir önceki taramada bulunan sık geçen eleman kümeleri aday kümeler adı verilen yeni potansiyel sık geçen eleman kümelerini üretmek için kullanılır. Aday kümelerin destek değerleri tarama sırasında hesaplanır ve aday kümelerinden minimum destek metriğini sağlayan kümeler o geçişte üretilen sık geçen eleman kümeleri olur. Sık geçen eleman kümeleri bir sonraki geçiş için aday küme olurlar. Bu süreç yeni bir sık geçen eleman kümesi bulunmayana kadar devam eder [49].

Bu algoritmada temel yaklaşım, eğer k-eleman kümesi minimum destek metriğini sağlıyorsa bu kümenin alt kümeleri de minimum destek metriğini sağlar şeklindedir.

3.6. Diğer Veri Madenciliği Metotları

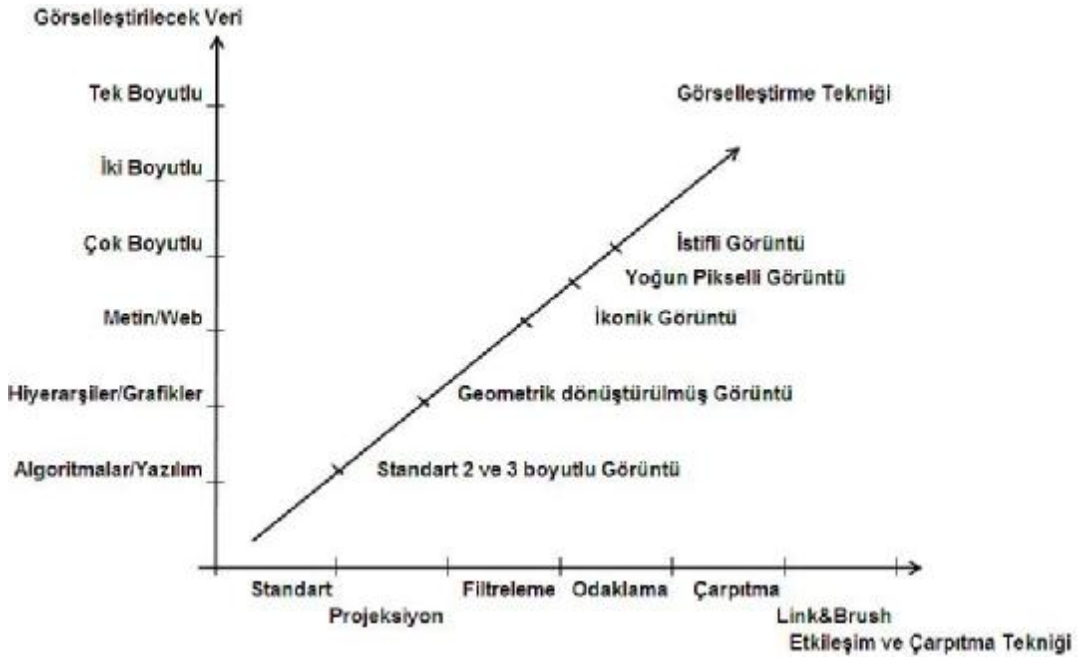
Veri Görselleştirme:

Verileri grafiklere dönüştürerek analiz sürecine sunan bir Veri Madenciliği tekniğidir. Veri analizi, insan yorumlaması olmadan eksik kalacaktır. Veri Görselleştirme kavramı veri analizinden insanın görme duyusundan yararlanılmasını sağlamaktadır. Böylece verilerin analizci ile etkileşimi sağlanır.

Kullanılacak Veri Görselleştirme metotları görselleştirilecek veri tiplerine bağlı olarak değişir. Görselleştirilecek veri tipleri temel olarak aşağıdaki gibi sınıflandırılır.

- Tek boyutlu veriler (Zaman serileri gibi)
- İki boyutlu veriler (Coğrafi haritalar gibi)
- Çok boyutlu veriler (İlişkisel tablolar gibi)
- Metin ve hiper metin (Haber makaleleri ve Web dokümanları gibi)
- Hiyerarşi ve grafikler (telefon konuşmaları ve Web dokümanları gibi)
- Algoritmalar ve yazılım (debugging işlemleri gibi)

Veri tiplerine göre kullanılan Görselleştirme teknikleri Şekil 3.11'de gösterilmiştir.



Şekil 3.11 Veri tiplerine göre kullanılan görselleştirme teknikleri [50]

Veri Görselleştirme ile ilgili uygulamaları kolayca yapmak için pekçok yazılım geliştirilmiştir. Bu yazılımlara Daisy, Datascope 3.0, JWAVE, NETMAP, VisualMine, IRIS, Graf-FX örnek verilebilir.

Diğer Veri Madenciliği yöntemlerinin başlıcaları olarak ;

- Temel Bileşenler Analizi,
- Diskriminant Analizi,
- Faktör Analizi,
- Bulanık Mantığa dayalı yöntemler,
- Bayes Ağları, (Bayesian Networks)
- Kaba Küme (Rough set) teorisine dayalı yöntemler sıralanabilir.

Bu yöntemlerin dışında birden fazla tekniği içine alan hibrid yöntemler ve zaman verilerine dayalı yöntemlerden de veri madenciliği yöntemi olarak

faaydalanılmaktadır [51]. Özet olarak, bilgi keşfine yarayan her yöntem veri madenciliği yöntemi olarak kullanılabilir [52].

3.7. Veri Madenciliği Metotlarının Kullanımı

Verilerin yapısına ve elde edilmesi amaçlanan hedeflere bağlı olarak hangi Veri Madenciliği tekniğinin kullanılabileceği Çizelge 3.4.'te verilmiştir.

Çizelge 3.4 Veri madenciliği teknikleri ve kullanılabileceği alanlar

Amaç	Girdi Değişkenler (Tahmin Ediciler)	Çıktı Değişkenler (Sonuçlar)	Veri Madenciliği Tekniği	Örnek (SPSS, 2000)
Popülasyonu en iyi ayıran girdi değişkenlerin doğrusal kombinasyonunu bulmak	Sürekli	Kesikli	Diskriminant Analizi	Sahtekarlık olayları tahmini Müşterilerin kalma ya da gitme durumu tahmini Yeni ürün veya teklife hangi müşterinin yanıt vereceği Çeşitli tıbbi prosedürlerin sonuç tahmini Sigorta poliçesi yenileme tahmini
Belli bir kategoride olma olasılığının tahmini	Sürekli	Kesikli	Lojistik ve Çoklu Regresyon	Sahtekarlık tahmini Bir müşterinin hangi ürünü alacağı tahmini Bir ürünün başarısız olmaya yakınlık tahmini
Girdi değişkenlerinin doğrusal kombinasyonu olan çıktı	Sürekli	Sürekli	Doğrusal Regresyon	Yeni müşteriden beklenen gelir tahmini Mağaza için satış geliri tahmini Bir hizmet numarasını arayan müşterilerin bekleme tahmini Hasta tipi ve tıbbi duruma göre hastanede kalma süresi tahmini
Deneyler ve aynı örneğin tekrarlanan ölçümü	Çoğu Girdi Kesikli olmalı	Sürekli	Varyans Analizi (ANOVA)	Hangi çevresel faktörlerin olası kanser nedeni olduğu tahmini
Düzenli aralıklarla kaydedilen olayların gelecekte ortaya çıkma tahmini	Sürekli	Sürekli	Zaman Serileri Analizi	Eski satış kayıtlarından gelecek satış verilerinin tahmini
En yakın komşuların değerine göre çıktı tahmini	Sürekli, Kesikli ve Metin	Sürekli veya Kesikli	Hafıza tabanlı Muhakeme (MBR)	Tıbbi sonuçların tahmini
Veriyi alt gruplara (dallara) ayırarak tahmin etme	Sürekli veya Kesikli	Sürekli veya Kesikli	Karar Ağaçları	Ayrılacak müşterilerin tahmini Sahtekarlık olayları tahmini
Kompleks doğrusal olmayan çevrelerde çıktı tahmini	Sürekli veya Kesikli	Sürekli veya Kesikli	Sınırsız Ağlar	Beklenen gelir tahmini Kredi riski tahmini
Veriyi ikiden fazla alt gruplara (dallara) ayırarak tahmin etme	Sürekli, Kesikli ve ya Sıra sayısı	Kesikli	CHAID	Clası satış için gerekli demografik faktörlerin tahmini Üretimde ürün hasarlarına yol açan faktörlerin tahmini
Veriyi ikiden fazla alt gruplara (dallara) ayırarak tahmin etme	Sürekli	Kesikli	C5.0	Kredi alanı hangi müşterilerin az riskli olduğu tahmini Bir ülkenin yatırım riski ile ilgili faktörlerin tahmini
Benzerlikleri Küçük bir girdi karakteristik kümesi olan büyük olay grupları bulmak	Sürekli veya Kesikli	Çıktı değişken yok	k ortalamalar Kümeleme Analizi	Pazarlamada müşteri bölümlenme Benzer sigorta amaçlı gruplar
Büyük Küme üyelikleri oluşturmak			Kohonen Sınırsız Ağları	Müşterileri demografik verilere ve alım desenlerine göre kümeleme
Küçük Küme birliktelikleri oluşturmak ve bir çok kategori arasında desen aramak	Mantıksal	Çıktı değişken yok	Pazar Sepet Analizi veya Apriori Algoritması	Birlikte satın alınması: olası ürünleri belirleme Öğrencilerin birlikte alması olası dersleri belirleme

Farklı tipteki veri madenciliği tekniklerini nitelikleri bakımından karşılaştırması Çizelge 3.5'te gösterilmiştir.

Çizelge 3.5 Farklı tipteki veri madenciliği tekniklerini nitelikleri bakımından karşılaştırması [53]

	Yapay Sinir Ağları	Genetik Algoritmalar	İstatistiksel Teknikler	Kural Çıkarımı ve Karar Ağaçları	Veri Görselleştirme
Yapılandırma Kolaylığı	Düşük	Çok Düşük	Yüksek	Çok Yüksek	Orta
Esneklik	Yüksek	Orta	Orta	Düşük	Düşük
Otonomi	Yüksek	Yüksek	Düşük	Düşük	Çok Yüksek
Hesaplama Karmaşıklığı	Çok Yüksek	Çok Yüksek	Orta	Düşük	Çok Yüksek
Açıklayabilirlik	Çok Düşük	Yüksek	Orta	Çok Yüksek	Çok Yüksek

4. VERİ MADENCİLİĞİNİN TIP ALANINDA KULLANIMI

İnsanların sağlık sorunlarını çözmeye çalışan Tıp bilimi, insan türünün bireysel çeşitliliği nedeniyle genel geçerliliği kanıtlanmış çözümleri ancak bir seviyeye kadar oluşturabilmektedir. Bireysel farklılıkların dikkate alınması gerekliliği yanında insan vücudunun zaman ile sürekliliğini koruyan, hiç durmadan çalışan dinamik bir sistem olması Tıp biliminin ancak olasılıklar ile konuşabilmesine izin vermektedir. Bu durumda insanların sağlık sorunlarına çözüm getirme noktasında kişisel bilgilerin ayrıntılı ve güncel olarak kayıt altında tutulması gerekmektedir.

Veri tabanı teknolojisinin gelişimi ile Tıp biliminin ihtiyaç duyduğu hasta bilgilerinin kayıt altına alınması ve takibi, sağlık sorunlarına çözüm bulma noktasında tek başına yeterli değildir. Elde tutulan verilerden çözüm bulma sürecine katkı sağlayacak anlamlı bilgilere ulaşılmasını sağlayan veri madenciliğinin Tıp alanında uygulanması insanlık için çok önemli bir gelişmedir.

Yaklaşık 30 yıllık geçmişi ile oldukça genç bir disiplin olan Veri Madenciliği, telekomünikasyon ve pazarlama gibi alanlarda başarısını kanıtladıktan sonra çok daha önemli bir alan olan tıp alanında da uygulanmaya başlanmıştır.

Veri madenciliği tekniklerini uygulamak için gerekli altyapı, veri tabanı uygulamalarının hastane ve laboratuvarlarda kullanılmaya başlanması ile sağlanmıştır.

Böylece bilgisayarlar hasta bakım hizmetlerinin destekleme, sağlık bakım hizmetlerinin kalitesinin değerlendirilmesi gibi doğrudan sağlık bakım hizmetlerinin sunulmasında kullanılmasının yanı sıra, karar verme, yönetim, planlama ve tıbbi araştırmalar gibi yönetsel ve akademik fonksiyonların yerine getirilmesinde daha fazla kullanılmaya başlanmıştır.

Basit veri tabanları hasta bilgileri (ad,soyad, adres, kan grubu vb.) gibi temel verileri saklarken, daha gelişmiş veri depo sistemleri tıbbî personele hastanın sağlık durumuna ilişkin bir çok veriyi kaydetme imkanı verir. Günümüzde hastaların hastane yerleşiminden finansal verilerine, ilaçlarından ziyaret planına kadar pek çok veri de bu sistemlerce kayıt altına alınmaktadır [55].

4.1. Tıbbi Verilerin Yapısı

Veri Madenciliğinin tıp alanında uygulanması, tıbbi verilerin yapısından kaynaklanan çeşitli zorluklar nedeniyle diğer alanlardan daha sonra olmuştur. Tıbbi verilerin kendine has yapısından kaynaklanan zorluklar maddeler halinde aşağıda incelenmiştir :

- Tıp alanındaki veriler genellikle farklı kaynaklardan toplanmaktadır. Örneğin hastanın laboratuvar ile ilgili verileri ile hastanın teşhis bilgileri farklı kaynaklarda ve farklı şekillerde tutulmaktadır. Bu nedenle bu verilerin anlamlı bir bütün haline getirilmeleri için uzman desteğine ihtiyaç vardır [54].
- Tıbbi veriler metinden tıbbi görüntüye (röntgen filmi vb.), EEG, EKG gibi sinyallerden diğer görüntüleme yöntemlerine kadar pek çok farklı formatta tutulmaktadır [54]. Farklı formattaki tıbbi verilerin bir arada kullanılması gerektiği durumlarda anlamlı bir model oluşturmak zorlaşmaktadır.
- Tıbbi veriler homojen değildir, çok farklı numerik ya da kategorik değerler alabilir [55]. Bu nedenle veri kümesi oluşturulurken her bir verinin değerine ayrıca dikkat edilmesi gereklidir. Tıbbi veriler kişiye özel olduğundan veri değişim aralığının çok dışında veri değerlerine de rastlanabilmektedir. Tıbbi Veri kümelerinin bu heterojen yapısı Veri Madenciliği algoritmalarını zorlamaktadır.
- Tıbbi veriler özellikle zaman bazlı tutulan veriler ve görüntü verileri çok büyük boyutlara ulaşabilmektedir [55]. Bu durum bu verilerin depolanması için büyük boyutlu veri depoları ve verileri işlemek için büyük bir işlemci gücü ve yeni teknikler gerektirmektedir.

- Tıbbi veriler tek başlarına anlam taşımazlar [55]. Mutlaka bir uzman tarafından yorumlanmaları gereklidir. Bu nedenle tıp alanında yapılacak veri analizi çalışmalarında insan faktörü çok büyük öneme sahiptir.
- Tıbbi veriler kesin değildir, her zaman içinde hata ihtimali barındırır. Hata ölçümden kaynaklanabileceği gibi verinin kişiye özel olmasından da kaynaklanabilir. Bu nedenle tıbbi verilerin bir Veri Madenciliği çalışmasından kullanılmadan önce yeterli güvenilirlikte kesinlik taşıması sağlanmalıdır.
- Tıbbi veriler fiziksel verilere göre daha fazla metinsel karakter taşır. Tıbbi verilerin metin yapısından numerik yapıya dönüştürülmesi süreci (kategorilere ayırma vb. tekniklerle) Veri Madenciliği tekniklerini uygulamayı zorlaştırmaktadır.
- Tıbbi veriler insanlardan elde edildiği için yasal ve gizlilik kurallarına uygun olarak sağlanmalı ve kullanılmalıdır.
- Tıp alanında belirli bir standardın olmayışı ve var olan standartlar arasında tam bir uyumun olmaması, bu alanda bir veri ambarının oluşturulmasını oldukça zorlaştırmaktadır. Çizelge 4.1’de görüldüğü üzere çeşitli standartlar arasında farklı kodlama sistemleri mevcut olmaktadır. Bu nedenle veri ambarı oluştururken farklı kaynaklardan toplanan veriler arasındaki standart uyumunun sağlanması gereklidir [3].

Çizelge 4.1 RCC (Read Clinical Classification) ve ICD-9-CM ((International Classification of Diseases 9th Revision) standartları arasındaki ilişki [3].

Seviye	Terim	RCC	ICD-9-CM
1	Infectious/parasitic diseases	A	001-139
2	Viral disease with exanthema	A5	050-057
3	Rubella	A56	56
4	Rubella+neurological complications	A560	560
5	Rubella + encephalomyelitis	A5601	056.01

4.2. Tıp Alanında Veri Madenciliği Uygulamaları

Tıp alanında veri madenciliği uygulamaları yedi başlık altında özetlenmiştir :

- Tıp Literatürü üzerinde Metin Madenciliği Uygulamaları
- Hastane Bilgi Sistemi üzerinde Veri Madenciliği çalışmaları
- Genetik alanında Veri Madenciliği çalışmaları
- Eczacılık alanında Veri Madenciliği çalışmaları
- Veri Madenciliği Tekniklerine göre Veri Madenciliği çalışmaları
- Hastalık Çeşitlerine göre yapılan Veri Madenciliği çalışmaları
- Hemodiyaliz ile ilgili Veri Madenciliği çalışmaları

4.2.1 Tıp literatürü üzerinde metin madenciliği uygulamaları

Tıp alanında yayınlanan bilimsel makaleler, sağlık alanında yapılan araştırmalar ve yenilikler için değerli bilgi kaynaklarıdır ve metinsel yapılarda saklanırlar. Bu yapılar üzerinde bilgi keşfi yapmak için Medline gibi tıbbi alanda yapılmış bilimsel yayınların saklandığı büyük veritabanları üzerinde çalışabilen metin madenciliği yöntemleri geliştirilmektedir. Aşağıda tıp literatürü üzerinde yapılan veri madenciliği uygulamalarına bazı örnekler verilmiştir.

Prether et al. [56], obstetrik (kadın doğum ile ilgili) veri tabanını veri madenciliği teknikleri yardımıyla analiz ederek erken doğuma sebep olan potansiyel faktörleri değerlendirmeye çalışmıştır.

Houston et al. [57], Arizona Üniversitesi Yapay Zeka Laboratuvarı'nda geliştirdikleri çeşitli veri madenciliği algoritma ve tekniklerini Amerikan Ulusal Kanseri Enstitüsü (NCI) ve farklı kanser veri tabanları üzerinde internet ortamında kullanabilme imkanları üzerinde duran bir çalışma yapmıştır.

Mullins et al. [58], tıp alanındaki yayınlar üzerinde araştırma yapmak için HealthMiner adı verilen bir veri madenciliği yaklaşımı geliştirmiştir.

HealthMiner uygulaması ile 667 bin hasta ve sağlıklı birey verilerinden oluşan veri kümesi üzerinde CliniMiner, Tahmin Edici Analiz ve Desen Keşfi adlı üç denetimsiz öğrenme metodu uygulanmıştır.

Kuo ve Shih [59], birliktelik kuralları çıkarmada çok boyutlu kısıtlar problemini çözmek amacıyla yeni bir teknik önermiştir. Bu teknik geniş bir veri tabanı olan Ulusal Sağlık Sigortası Araştırma Veritabanından Karınca Koloni sistemi adlı meta-sezgisel teknik yardımıyla verimli şekilde birliktelik kuralları elde etmeyi sağlamıştır.

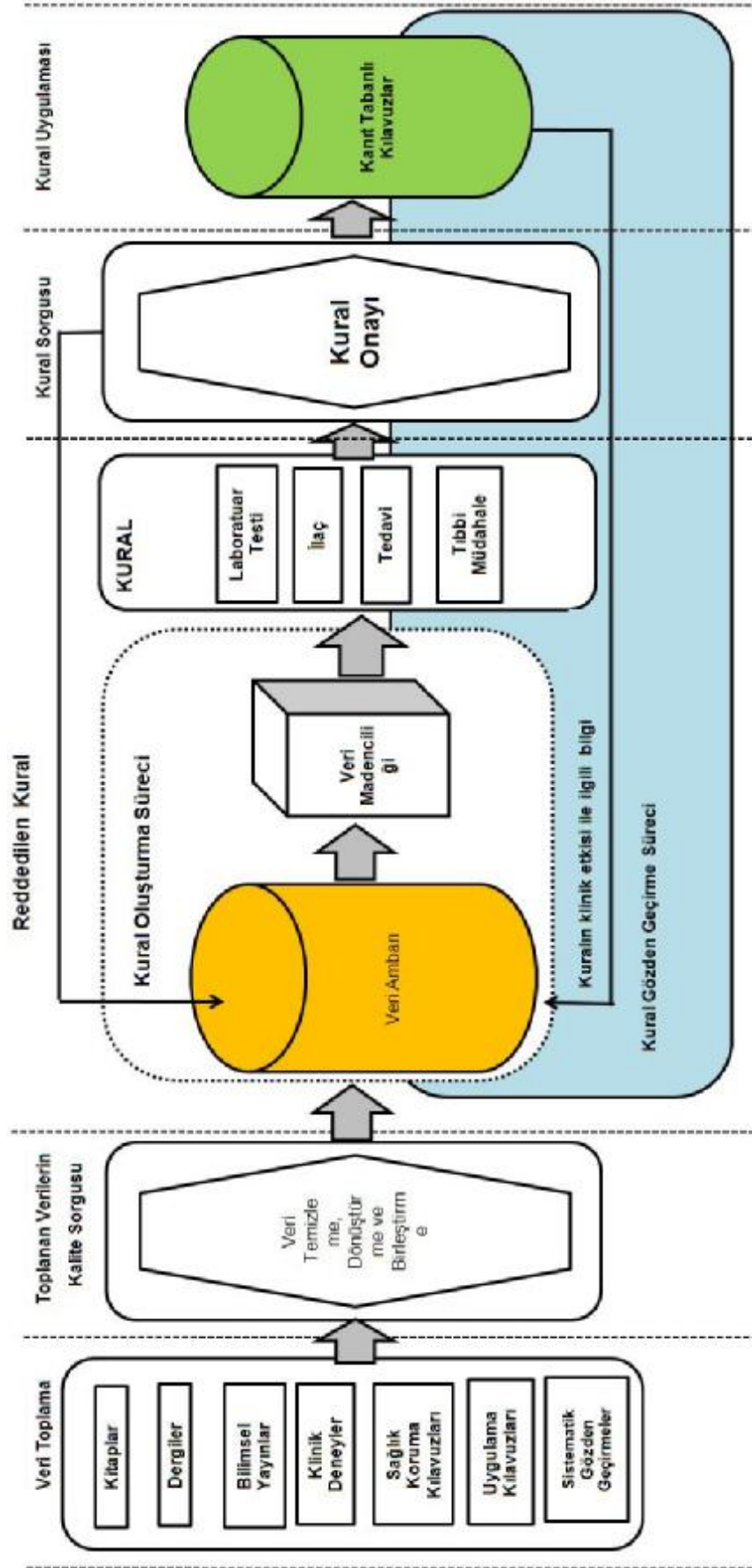
Avrupa Biyoenformatik Enstitüsü (European Bioinformatics Institute) Metin Madenciliği Araştırma Grubu, biyomedikal alandaki bilimsel makalelerden bilgi keşfi yapmak için Tıbbi Ontoloji (Medical Ontology), Anlamsal Ağlar (Semantic Networks), Doğal Dil İşleme (Natural Language Processing) ve Birleştirilmiş Tıbbi Dil Sistemi (Unified Medical Language System-UMLS) gibi teknikler kullanarak metin madenciliği çalışmaları yapmaktadır. Yine bu araştırma grubu tarafından geliştirilen EBIMed yazılım sisteminde girilen anahtar kelime için, Medline veritabanındaki özetler bulabilmekte, biyoinformatik kaynaklardan seçilen terminolojik cümleler bulunarak genel bir tabloda görüntülenmektedir [60].

Tıp bilimi kendi içinde pek çok uzmanlık alanına ayrılmıştır. Her uzmanlık alanı kendi konusu olan vücut bölgesi, organ ya da hastalık ile ilgili Hastalık belirtileri (semptomlar), test ve tedavi metotları, kullanılan ilaçlar, hasta bakımı kuralları vb. çeşitli tıbbi çalışmalar yapmaktadır. Bu çalışmalarda bulunan yenilikler ve belirlenen yeni tıbbi kurallar gerekli sorgu süreçlerini geçtikten sonra o alanın uzmanlarının bir araya geldiği bilim kurulları tarafından Kanıt Tabanlı Kılavuzlar şeklinde yayınlanmaktadır. Nefroloji alanında sürekli güncellenen KDOQIG (Kidney Disease Outcomes Quality Initiative Guideline) kanıt tabanlı kılavuzlara bir örnektir.

Metin madenciliđi, Kural tabanlı Kılavuzların Oluřturulması sürecinde tıbbi uzmanlık alanlarında yeniliklerin bulunmasında yardımcı olarak kullanılmaktadır.

řekil 4.1'de Kural tabanlı tıbbi kılavuzların oluřturulması süreci özet olarak verilmiřtir.

Sađlık Bilimlerinde (Tıp, Biyoloji, Genetik vb.) kullanılan Metin madenciliđi uygulama araçları Çizelge 4.2'de özetlenmiřtir.



Şekil 4.1 Tıbbi Kılavuzların Kural tabanlı oluşturulması süreci [61]

Çizelge 4.2 Sağlık Bilimlerinde (Tıp, Biyoloji, Genetik vb.) kullanılan Metin Madenciliği uygulama araçları [66].

Kaynak Adı	Tanımı ve URL Adresi
PubMed	Tıbbi Makale Arama Servisi http://www.pubmed.org
NCBI E-Utilities	Web sorgu arayüzü dışında Entrez verilerine erişim http://www.ncbi.nlm.nih.gov/entrez/query/static/eutilsJhelp.html
E-BioSci	Yaşam bilimleri ile ilgili Tam metin ve bilgilere erişim için Avrupa Platformu http://www.e-biosci.org
iHOP	Protein fonksiyonları ve etkileşimi ile ilgili biyomedikal literatür erişimi http://www.ihop-net.org/
EBI Literature DBs	Avrupa Patent Ofisi'nin biyoteknoloji ile ilgili patent uygulamaları http://www.ebi.ac.uk/Databases/literature.html
GoPubMed	PubMed sorgularını Gen Ontolojisine göre sınıflar http://www.gopubmed.org
MedMiner	Literatürden Tıbbi içerikli makaleleri filtreler http://discover.nci.nih.gov/textmining
EBIMed	Medline veri tabanından Bilgi Çıkarımı yapan Web uygulamaları http://www.ebi.ac.uk/Rebholz-srv/ebimed
XplorMed	PubMed sorgularını MeSh Ontolojisine göre sınıflar http://www.ogic.ca/projects/xplormed
Textpresso	C. elegans. Biyoloji literatüründen bilgi çıkarımı http://www.textpresso.org
BioIE	PubMed sorgularından kural bazlı bilgi içerikli cümleler keşfi http://umber.sbs.man.ac.uk/dbbrowser/bioie/
JournalMine	Biyomedikal literatürde nesne ilişkisel sorgulama http://textmine.cu-genome.org/gridsphere/gridsphere
iProLINK	Protein bilgilendirme ve etiketleme http://pir.georgetown.edu/lprolink
Arrowsmith	İki PubMed makale kümesi arasında bağlantı tanımlama aracı http://arrowsmith.psych.uic.edu
BITOLA	Biyomedikal kavramlar arasında potansiyel yeni ilişkiler amaçlı bir araç http://www.mf.uni-lj.si/bitola
HCAD	İnsan kromozom bozuklukları, genler ve hastalıklar arası ilişki bilgi içeren bir veri tabanı http://www.ihop-net.org/UniPub/HCAD/

4.2.2. Hastane bilgi sistemi üzerinde veri madenciliği çalışmaları

Bu kısımda literatürde yer alan Hastane Bilgi Sistemi üzerindeki veriler kullanılarak yapılan Veri Madenciliği çalışmalarından bazı örnekler verilmiştir.

*Tedavi Planı Oluşturma

Johnson et al. [63], Tip II Diyabet hastalarında gelecekte ortaya çıkabilecek olumsuz durumların (kalp krizi riski vb.) tahmin edilmesi için hastanın sağlık durum bilgileri ve belirlenen tedavilere verdiği yanıtlardan oluşan verileri kullanarak yapay sinir ağları ile bir veri madenciliği modeli kurulmuşlardır.

Amerika'da hastane ve sağlık merkezlerine Diyabet hastaları için tedavi kalitesini artıran ve maliyetleri düşüren gelişmiş sağlık planları tasarlama hizmeti veren American Healthways, yüksek riskli hastaların tespit edilmesinde, kısa dönemde ortaya çıkabilecek sağlık problemlerinin tahmininde ve bu durumları önleyici tedavi planlarının geliştirilmesinde veri madenciliği teknikleri yardımıyla tahmin edici modelleme teknolojisi kullanmaktadır [64].

Razali ve Ali [65], Malezya'da farklı sağlık merkezlerinden toplanan hasta verileri üzerinde C.5 karar ağacı algoritması uygulayarak hastalarda en sık rastlanan akut üst solunum yolları enfeksiyonu için bir tedavi planı geliştirme modeli oluşturmuştur.

*Hastane Cerrahi İhtiyaç Tespiti

Persson ve Lavesson [66], hasta verileri ile cerrahi operasyon verilerinden hareketle hastanenin cerrahi ihtiyaçlarını (cerrahi müdahale olasılığı, ameliyat süresi, iyileşme süresi vb.) belirlemekte kullanılacak bir otomatik tahmin modelinin veri madenciliği teknikleri yardımıyla geliştirilme olanaklarını incelemiştir.

*Hastane Enfeksiyon Kontrol Sistemi

Brosette et al. [67] çalışmasında aynı zamanda bir veri madenciliği uygulaması olan Alabama Üniversitesi Enfeksiyon Gözetim Sistemi'ni incelemiştir. Bu gözetim sistemi sayesinde laboratuvar bilgi yönetim sisteminden elde edilen enfeksiyon kontrol verileri içinde Birliktelik kuralları yardımıyla yeni ve ilginç desenler tespit edilmekte ve her ay tespit edilen desenler yayınlanarak enfeksiyon uzmanları tarafından incelenmektedir.

*Hastane Performans Değerlendirme Sistemi

Amerika'da çeşitli organizasyonlar, hastane ve sağlık merkezlerini verdikleri sağlık hizmetleri ve elde ettikleri sonuçlar bakımından değerlendirerek sıralamaya tabi tutmaktadır. Sıralamanın oluşmasında veri madenciliği teknikleri kullanılarak oluşturulan risk faktörleri raporlaması etkili olmaktadır. Bu raporlamada hasta bilgileri, hastalık tipleri ve risk faktörleri verileri üzerinde kümeleme ve birliktelik analizi teknikleri kullanılmaktadır [68].

*Uyumsuz (Outlier) Veri tespiti ve Analizi

Sağlık alanında tutulan veri tabanlarında hatalı görünen ya da normalden aşırı sapmalar gösteren verilerin tespit edilmesi ve bu verilerin bilgi keşfi için analiz edilmesi, bir veri madenciliği çalışması gerektirmektedir. Bu alanda Kumar et al. [69], Wisconsin Göğüs Kanseri Merkezi Veri Bankası'ndan sağladıkları veriler üzerinde TANAGRA yazılımını kullanarak uyumsuz verilerin tespiti ederek bilgi keşfi için bu verilerin analizini gerçekleştirmiştir.

*Tıbbi Hataların Analizi

Tsumoto ve Hirano [70], çalışmasında tıbbi müdahale hatalarının nedenlerine ilişkin analiz yapmak amacıyla tıbbi hata verilerden oluşan bir veri kümesi üzerine karar ağacı algoritması kullanmıştır. Sonuçta hemşirelerin çalışma

saatlerinin tıbbi müdahale hatalarının ortaya çıkışında etkisi olduğu, ayrıca hemşirelere yardımcı olmayan hastaların hemşirelerin iş konsantrasyonunu azaltmaları ile tıbbi müdahale hatalarına yol açtığı tespit edilmiştir.

Kimura et al. [71], çalışmasında ilaçlar ve tıbbi cihazlarla ilgili tıbbi unutma durumlarının nedenlerinin tespiti için ilgili veriler üzerinde veri ve metin madenciliği teknikleri (Pazar sepet analizi ve C5.0 karar ağacı algoritması) kullanmıştır.

*Hastanın Hastanede Yatış Süresinin Tahmini

Hong [72] tez çalışmasında multi-mini-destek birliktelik kuralları ve zaman serisi algoritması yardımıyla ve hemodiyaliz hastalarına ait biyokimyasal test verilerini analiz etmiş ve hastaneye yatma oranının tahmininde çeşitli desenler bulmuştur.

Yeh ve Wu [73], veri madenciliği teknikleri (karar ağacı ve birliktelik kuralları analizi) ile zaman soyutlama işlemini birleştirerek hastaneye yatma oranının tahmini için hemodiyaliz hastalarının biyokimyasal verilerini analiz eden bir karar destek sistemi geliştirmiştir.

4.3.3. Genetik alanında veri madenciliği çalışmaları

İnsan Genomu Projesi ile birlikte Genetik, geleceğin bilimi olarak ilan edilmiş ve insanlığın sorunlarının kökeninin belirlenmesi ve bu sorunların çözümü ve tekrarının engellenmesi için büyük umutlar beslenmesine yol açmıştır.

Genler şifrelerinin çözülmesi ile birer veri tabanı olarak tıp biliminin kullanımına sunulmuştur. Bu veri tabanlarında veri madenciliği uygulamalarının kullanımı insanlığa çok büyük faydalar sağlayacaktır.

Bu alanda yapılan alıřmalara İnsan Genom Projesi'nin Avrupa ayağı olan Sanger Enstitüsü tarafından yürütölmekte olan DECIPHER projesi ile National Center for Biotechnology Information (NCBI) tarafından geliştirilen OMIM (Online Mendelian Inheritance in Man) projesi örnek verilebilir.

Birok ölkeden hastanelerin bilgilerini paylařtığı DECIPHER projesi ile insan genomunda kromozom düzeyindeki deęiřimler ile hastalıklar arasındaki iliřkinin anlařılması için kapsamlı bir veritabanı oluşturulması amaçlanmıřtır.

OMIM (Online Mendelian Inheritance in Man) projesi ise genetik bozukluklarla ilgili bilinen hastalıkların saklandığı bir veritabanıdır. Bu veritabanı metin bilgiler, resimler ve referans bilgilerinden oluřmuřtur ve ayrıca Medline veritabanına da baęlantısı vardır [74].

Veri Madencilięi teknikleri veri kaynağı bakımından ok büyük ve karmařık veri bankalarına sahip Genetik ve Protein Biyolojisi alanlarında sıklıkla kullanılmaktadır. Bu alanlarda yapılan veri madencilięi alıřmalarına iliřkin örnekler ařağıda özetlenmiřtir:

Tessier et al. [75] proteinler ile ilgili yaptığı arařtırmada sistein adlı aminoasitin disölfid baę durumunun güvenilir olup olmadığının tahmini için kurallar ıkarımı yapan kullanıřlı bir veri madencilięi teknięi geliřtirmiřtir.

Saito et al. [76] geliřtirdikleri özelleřtirilmiř bir veri madencilięi metodu aracılığıyla mutasyona uğramıř hücre grupları arasındaki farkları açıka göstermeyi bařarmıřtır.

Georgii et al. [77] gen mikrodizileri tanım bilgilerini analiz etmek için veri madencilięi teknikleri yardımıyla bir araç geliřtirmiřtir.

Perez et al. [78] erişime açık bir web hizmet sağlayıcıda genlerin insan genomu içinde ilişkili olduğu hastalığa bağlı olarak sıralanmasını sağlayan bir metot geliştirmiştir.

Korn et al. [79] genleri Kanser Genom Anatomi Projesi verilerini kullanarak, ilişkili olduğu önceden belirlenmiş fenotiplerden ve/veya moleküler yollardan tanımlayan alışılmadık bir veri madenciliği prosedürü geliştirmiştir.

.

4.3.4. Eczacılık alanında veri madenciliği çalışmaları

Eczacılık alanında yapılan çeşitli veri madenciliği çalışmaları ilgili başlıklar altında özetlenmiştir.

*Yeni İlaçlar Geliştirme

Yeni ilaçların geliştirilmesi sürecinde sınıflama, kümeleme ve sinir ağları gibi veri madenciliği tekniklerinden benzer etkinlik gösteren bileşenlerin bulunmasında ve ilgili genlerde aktivite tespit çalışmalarında kullanılmaktadır.

*İlaç Dozlarının Tespiti

Başarıyla geliştirilen ilaçlar, istenen etkileri hangi dozda hangi ölçüde gösterdiklerinin tespit edilmesi amacıyla insan ve hayvan dokuları üzerinde test edilirler. Bu test aşamasında sınıflama ve Sinir ağları tekniklerinden yararlanılmaktadır. Sinir ağlarından ayrıca insan denekler üzerinde yapılan ilaç testlerinde hastanın gelişimini izlemek için de yararlanılmaktadır [80].

Tedavinin hastaya faydalı olup olmadığının tespiti, ilaçtan faydalanabilecek hastaların özelliklerinin tahmini gibi işlemler de çeşitli veri madenciliği teknikleri yardımıyla yapılabilmektedir [80].

*İlaçların Yan Etkilerinin Araştırılması

İlaçların yan etkilerinin tespiti, zorunluluğu yasal düzenlemelerle belirlenmiş önemli bir konudur. Bu alanda yapılan çeşitli veri madenciliği çalışmaları mevcuttur. İlaç kullanan hastalar ait zamana bağlı takip verileri, Bayes sinir ağları tekniği yardımıyla incelenerek ilaç-durum birlikteliği analizleri gerçekleştirilmektedir [81].

Tsumato [82], antibiyotikler için allerji durumu ve steroidlerin yan etkileri ile ilgili veriler içeren iki veri kümesi üzerinde Kaba küme tekniğini model alan ve negatif kural çıkarımında bulunabilen PRIMEROSE-REX2 algoritmasını kullanarak ilginç desenler elde etmeye çalışmıştır.

4.3.5. Tıpta alanında kullanılan veri madenciliği tekniklerine göre çalışmalar

Veri Madenciliği tekniklerinin hemen hemen tamamı tıp alanındaki çalışmalarda kullanılabilmektedir. Tıp alanında kullanılan çeşitli veri madenciliği teknikleri için yapılan literatür taramasında bulunan örneklerin özetleri, ilgili tekniğin adı altında yer almaktadır.

*Yapay Sinir Ağları

Zhou et al. [83], çalışmasında vücuttan alınan doku örneklerinde akciğer kanserine yakalanan hücrelerinin tespiti için çoklu sinir ağı yapısı kullanan bir otomatik hastalık teşhis prosedürü önermiştir. İki aşamadan oluşan bu prosedür ilk aşamada hangi hücrelerin normal hangi hücrelerin kanserli olduğunu tahmin etmekte, ikinci aşamada ise kanserli hücrelerin beş farklı tip akciğer kanseri arasından hangisine ait olduğu tespit etmeye çalışmaktadır.

Cheney et al. [84] çalışmasında 10 yıl boyunca yaralanma vakaları için kaydedilen hasta bilgileri, yaralanma ile ilgili detaylı klinik veriler ile tedavi

verilerden oluşan travma veri tabanını Yapay Sinir Ağları tekniği yardımıyla modellemiş ve modelden elde edilen sonuçları aynı veri tabanı üzerinde lojistik regresyon tekniği kullanılarak elde edilen sonuçlarla karşılaştırmıştır.

Comak et al. [85], çalışmasında kalp atışlarından elde edilen Doppler sinyallerini Yapay Sinir Ağları yardımıyla normal ve anormal olarak sınıflandırarak kardiyologlara yardımcı olabilen bir karar destek sistemi geliştirilme olanaklarını ele almıştır.

*Karar Ağacı

Vlahou et al. [86] çalışmasında yumurtalık kanserinin teşhisinde yardımcı olmak üzere karar ağacı tekniği ile sınıflandırma yapan bir model önermiştir.

Hunt et al. [87], bayanlarda idrar kaçırmının teşhisinde karar ağacı tekniği ile Bayes analizi uygulamasını karşılaştıran bir çalışma yapmıştır.

*Kümeleme Metotları

Sleeman et al. [88], diyaliz hastalarının tedavi verilerinden oluşturduğu veri kümesi üzerinde makine öğrenimli veri madenciliği teknikleri (hiyerarşik kümeleme, Bayes ağı vb.) kullanarak düşük tansiyona yatkınlık ve kandaki kimyasal dengesizlik eğilimlerine dair bilgiler elde etmeye çalışmıştır.

*Makine Öğrenimli Teknikler

Demsar et al. [89], ciddi yaralanma vakalarında cerrahların tıbbi müdahaleye uygun cevap verebilecek hastaları seçmesinde yardımcı bir model olarak hasta özelliklerinden oluşan veri kümesinin makine öğrenimli tekniklerle analizi üzerinde çalışmıştır.

Mena ve Gonzalez [90], kalp-dolaşım sistemi hastalıklarının teşhisinde kullanılan yeni bir makine öğrenimli kural çıkarım algoritması önermiştir. Makine öğrenimi için Kalp-dolaşım hastalıkları ile ilgili bir veri kümesi kullanmış, daha sonra Kalp-dolaşım hastalıkları ile ilgili bir veri tabanı kümesi ile üç farklı kamusal veri tabanı bir araya getirilerek orijinal bir veri kümesi oluşturulmuştur. Bu veri kümesi üzerinde farklı sınıflandırma algoritmaları (Saf Bayes, C4.5, k en yakın komşuluk) kullanılmıştır.

*Saf Bayes Sınıflaması

Matthew Wiggins et al. [91] çalışmasında EKG (elektrokardiyogram) sinyalleri üzerinden hastaların yaşlarına göre sınıflandırılması için bir Bayes Ağ sınıflandırıcısı geliştirmiştir.

Ranjit et al. [92], Saf Bayes sınıflandırıcısının tıbbi veri kümeleri üzerinde sınıflandırma doğruluğunu arttırmak amacıyla farklı kesikli hale getirme tekniklerinin karşılaştıran bir çalışma yapmıştır.

Yine Ranjit et al. [93], Saf Bayes sınıflandırıcısının tıbbi veri kümeleri üzerinde sınıflandırma doğruluğunu arttıran yeni bir özellik seçim lagoritması önermiştir.

*Kaba Küme Metodu (Rough Set)

Stefanowski ve Slowinski [94], çalışmasında çoklu yaralanma durumundaki hastalara ait veriler üzerinde kaba küme tekniğine kural çıkarım ve Ki kare testi entegre ederek geliştirdikleri karma bir metot kullanarak Kaba küme tekniğinden daha iyi sonuçlar elde etmiştir.

Stepaniuk [95], Diabetes Mellitus hastası çocuklara ait verilerin yer aldığı bir veri kümesi üzerinde kaba küme tekniği kullanarak bir veri madenciliği çalışması yapmıştır.

Tsumato [82], ilaç yan etkileri ile ilgili iki veri kümesi üzerinde Kaba küme tekniğini model alan bir kural çıkarım algoritması kullanarak ilginç desenler elde etmeye çalışmıştır.

Breault [96] çalışmasında bir diyabet veri tabanında yer alan sekiz farklı değişken üzerinde kaba küme tekniği kullanarak yaptığı veri madenciliği uygulaması ile kişilerin diyabet olma durumunu tahmin etmede kullanılabilecek bir araç bulma imkanını araştırmıştır.

Kusiak et al. [97], Hipoplastik sol kalp sendromu ile ilgili 73 ayrı veri özelliği belirleyerek topladığı veri kümesi üzerinde kaba küme teorisi temelli bir makine öğrenme algoritması kullanarak karar kuralları oluşturmaya çalışmıştır.

*Bulanık Küme Metotları

Cosic et al. [98], tomografi resimlerinin etiketlenmesinde denetimsiz öğrenimli bulanık sınıflandırma algoritmaları ve kural tabanlı sistemlerin kullanım imkanlarını araştırmıştır.

Duch et al. [99], çalışmasında sinirsel ağ tekniği ile global minimizasyon teknikleri kullanarak geliştirdikleri bulanık mantıksal kural çıkarım algortmasını göğüs kanser, kalp hastalığı ve hipotiroid veri tabanları üzerinde uygulamıştır.

*Genetik Algoritmalar

Brameier ve Banzhaf [100], farklı tıbbi teşhis problemlerinde doğrusal genetik programlama metodunun kullanım imkanlarını araştıran bir çalışma yapmıştır.

Novobilski et al. [101], akciğer embolisine yol açan ve tesbit edilmesi çok zor olan akciğerlerdeki kan pıhtılarının tespitinde genetik algoritma tabanlı bir veri madenciliği aracı kullanarak olasılık tabanlı bir çıkarım modeli önermiştir.

*Birliktelik Kuralları

Richard et al. [102], diyabet hastaları ile ilgili verilerin yer aldığı bir veri tabanı üzerinde birliktelik kuralları tekniği uygulayarak hastaneye ilk başvuru sırasındaki veriler ile erken ölüm arasında ilginç desenler elde etmeye çalışmıştır.

Gupta et al. [103], sınıflandırma kuralları kullanarak tıbbi teşhis verileri analizi sonucunda kalp hastalığı teşhisi için yapılması gereken test sayısını belirlemede yardımcı olacak bir birliktelik kuralı modeli önermiştir.

Doğan ve Türkoğlu [104], biyokimyasal kan parametreleri analizi ile Hiperlipidemi teşhisinde yardımcı olan birliktelik kuralları tabanlı bir model önermiştir.

Niimi ve Tazaki [105], Menenjit veri tabanından kural çıkarımı için Apriori birliktelik kuralı algoritması ile genetik programlama metodunu birleştiren bir model önermiştir.

*Graf tabanlı Metotlar

You et al. [106], çalışmasında metabolik yolların tüm özelliklerini graf tabanlı gösterimine yer vererek bu gösterim üzerinde graf tabanlı ilişkisel öğrenme algoritmaları kullanarak metabolik yolları birbirinden ayıran farkların tespit eden, ayrıca ortak özelliklerini belirleyen bir model ortaya koymuştur.

*Sınıflandırma Metotları

Tang et al. [107], kalp hastalığı veri kümesi üzerinde çeşitli karar ağacı sınıflandırma algoritmaları (ID3, C4.5, CART, CHAID) kullanarak karşılaştırmalı bir çalışma yapmıştır.

Huang et al. [108], kronik hastalıkların tahmin ve teşhisinde yardımcı olacak veri madenciliği teknikleri ile durum tabanlı muhakemeyi bir araya getiren bir karar destek sistemi önermiştir. Sistemin amaçları arasında sağlık tarama verilerinin veri madenciliği teknikleri (sınıflandırma ve diğer) ile analizi sonucunda elde edilecek kuralların kronik hastalık tahmininde kullanımı, yorumlanan verilerden kronik hastalıkların teşhisinde yararlanılması, durum tabanlı muhakeme ile kronik hastalıkların teşhis ve tedavisine yardımcı olunması yer almaktadır.

Podgórec et al. [109] çalışmasında sınıflandırma kuralları çıkarımı için önerdiği evrimsel karar ağacı çıkarımı tekniği ile otomatik programlama tekniklerini kullanılarak oluşturduğu AREX metodunu pediatrik kardiyoloji veri kümesi üzerinde uygulamış ve elde ettiği sonuçları incelemiştir.

*Tıbbi Görüntü Analizi

Xie et al. [110] çalışmasında zatürre ve SARS (Ciddi Akut Solunum Hastalığı) gibi solunum yolu hastalıklarına ait hasta akciğer filmlerinin çeşitli veri madenciliği algoritmaları (sınıflandırma - C4.5, CART; sinirsel ağ ve birliktelik kuralları) ile analiz ederek, SARS teşhisinde yararlanılacak Bilgisayar destekli SARS Tespiti (Computer Aided Detection –SARS) adlı bir karar destek sistemi önermiştir.

Oliveiraa et al. [111], göğüs radyografi filmlerinin k en yakın komşuluk sınıflandırma algoritması ile inceleyebilen ve zatürre hastalığının teşhisinde yardımcı olacak Pneumo-CAD adlı bir karar destek sistemi önermiştir.

Lovis et al. [112], hastaların tıbbi kayıtlarından oluşan heterojen bir veri kümesinden otomatik enfeksiyon desenleri ve antibiyotik reçeteleri keşfeden bir veri madenciliği modeli önermiştir.

Spyros et al. [113], akciğer radyografi filmleri üzerinde bir veri madenciliği tekniği (sınıflandırma) uygulayarak akciğerde enfeksiyon noktalarının tespitinde yardımcı olacak bir model önermiştir.

*Tıbbi Ses Analizi

Cox et al. [114] çalışmasında sayısal, kaetgorik ve metinsel veriler içeren heterojen bir ses kayıt veri kümesini kendi kendini düzenleyen haritalar ve istatistiksel ki kare testi ile analiz eden bir veri madenciliği aracı önermiştir.

*Zaman Serileri Analizi

Sacchi et al. [115], hemodiyaliz seansları sırasında farklı klinik değişkenlerin zamana bağlı kayıtlarından oluşan bir veri kümesini kural çıkarım algoritması kullanılarak analiz etmiştir.

*Veri Görselleştirme Metotları

Chittaro et al. [116] çalışmasında zamana bağlı verilerden oluşan tıbbi veri tabanları üzerinde görsel veri madenciliği yapan ve böylece hemodiyaliz tedavisinin yönetimine yardımcı olan bir yaklaşım geliştirmiştir.

*Çeşitli Metotlar Karşılaştıran Çalışmaları

Bir sağlık problemi ile ilgili verilerin incelenmesinde hangi veri madenciliği metodunun daha iyi sonuç verdiği konusu araştıran çalışmalar da yapılmıştır.

Sandholm et al. [117] atlarda sindirim sistemi koliği için ölüm sebebi ve ölüm riski tahmininde kullanılabilen altı veri madenciliği algoritmasını (Karar Ağacı, Sinir Ağları, Doğrusal ve Lojistik Regresyon, en yakın komşu algoritması) karşılaştırmıştır.

Aftarczuk [54] tez çalışmasında tıbbi karar destek sistemlerinde uygulanan veri madenciliği metotlarından üç tanesinin (Saf Bayes, C4.5 sınıflandırma ve Çok Boyutlu algılama algoritması) beş farklı tıbbi veri kümesi (Kalp hastalığı, Hepatit, Göğüs kanseri, Dermatoloji ve Diyabet) üzerindeki performanslarını karşılaştırmıştır.

Matyja [118] tez çalışmasında göğüs kanserine ilişkin elde ettiği veri kümesi üzerinde altı farklı veri madenciliği tekniği (Karar ağacı, birliktelik kuralları, kümeleme, Saf Bayes, Yapay Sinir Ağları ve Lojistik Regresyon) uygulamış ve sonuçları değerlendirmiştir.

Daud ve Corne [119] çalışmasında tıbbi veri madenciliğinde kullanılan altı farklı kural çıkarım ve beş farklı karar ağacı algoritmasının on farklı tıbbi veri kümesi üzerinde nominal sınıf tahmini performanslarını karşılaştırmıştır.

4.3.6. Hastalık tiplerine göre yapılan veri madenciliği çalışmaları

Farklı hastalıklar için literatürde pek çok veri madenciliği çalışması mevcuttur. Bu kısımda belli başlı hastalıklar için yapılmış bazı Veri Madenciliği çalışmalarından ilgili başlık altında örnekler verilmiştir :

***Lösemi**

Kumarkolla ve Sekhar [120], hastalık modeli geliştirmede veri madenciliği kullanma imkanını araştırdığı çalışmalarında Akut lösemi hastalığını gen verilerinden hareketle makine öğrenme teknikleri yardımıyla modellenmesini ele almışlardır.

*Tüberküloz

Sebban et al. [121], veri madenciliği metotlarının Tüberküloz hastalığının Moleküler Epidemiolojisinde kullanım imkanını deneysel kısıtları eleyen ve bilinmeyen desenlerle ilgili uzman bilgilerinden yararlanan bir teknik geliştirerek göstermiştir.

*Tip II Diyabet

Wren ve Garner [122], Tip II Diyabet hastalığını veri madenciliği yardımıyla incelemiş ve vücut içinde epigenetik (genetik olmayan irsi) değişimlerin Tip II diyabetin ortaya çıkmasında etkisi olduğunu belirtmişlerdir.

*Astım

Schmidt et al. [123], tıp biliminin veri madenciliği kullanarak teknoloji ile yapabileceği işbirliği imkanlarından bahsettiği çalışmasında geniş bir astım hastalığı veri kümesi üzerinde birliktelik kuralları, bölümlleme (segmentation) ve kümeleme algoritmaları kullanımını gerçekleştirmiştir.

*Akciğer Kanseri

Zhou et al. [83] bölüm 4.3.5. 'te özetlenen çalışmasında akciğer kanserini ele almıştır.

*Kalp Hastalığı

Le Duff et al. [124] çalışmasında kalp durması geçiren hastaların hastaneden taburcu olduktan sonra kurtulma oranının etkileyen en önemli faktörün tespiti ve hasta profillerinden kurtulma olasılığının çıkarımını sağlayan bir model geliştirmek için Bayes ağları tekniğini kullanmıştır.

Kusiak et al. [97], bölüm 4.3.5. 'te özetlenen çalışmasında kalp hastalığı üzerinde çalışmıştır.

Palaniappan ve Awang [125], çeşitli veri madenciliği teknikleri (Karar ağaçları, Saf Bayes ve Sinir ağları) kullanarak geleneksel karar destek sistemlerinin cevaplayamadığı "what if" (olursa) sorgularına cevap verebilen Akıllı Kalp Hastalığı Tahmin Sistemi (Intelligent Heart Disease Prediction System- IHDPDS) adlı prototip bir sistem geliştirmiştir.

*Göğüs Kanseri

Antoine et al. [126] çalışmasında dijital meme filmlerinden tümör tespit edilmesini sağlamak amacıyla sinirsel ağlar ve birliktelik kuralları algoritmaları gibi çeşitli veri madenciliği tekniklerinin kullanım imkanlarını incelemiştir.

Bellaachia ve Güven [127], 151,886 kayıttan oluşan işlenmemiş bir göğüs kanseri veri kümesi üzerinde Saf Bayes, geri yayımlı sinir ağı ve C4.5 karar ağacı algoritmaları kullanarak göğüs kanserinden kurtulabilirliği tahmin etmeye çalışan bir model geliştirmiştir.

Abe et al. [128], bütünleşik bir tıbbi veri madenciliği uygulama aracı olarak Siber Entegre Tıbbi Altyapı (Cyber Integrated Medical Infrastructure-CIMI)) adlı bir çerçeve sistem önermiştir. Bu sistem, biyoloji, mühendislik, kültürel bilimler ve sosyoloji gibi farklı alanlardan yararlanarak yapılacak tıbbi çalışmalar için farklı veri tipleri üzerinde entegre veri madenciliği uygulama imkanı sunmaktadır.

Karabatak ve İnce [129] çalışmasında göğüs kanseri için otomatik teşhis sistemi geliştirmek amacıyla Birliktelik kuralları ve Sinirsel Ağ algoritmaları kullanmıştır. Wisconsin göğüs kanseri veri kümesi üzerinde önce birliktelik kuralları yardımıyla özellik seçimi yaparak veri kümesi boyutunu küçülmüş, ardından Sinirsel Ağ algoritması yardımıyla akıllı sınıflama işlemi yapmıştır.

%95,6'lık bir sınıflama oranı yakalayan bu uygulamanın başka hastalıklar için de kullanılabileceği belirtilmiştir.

4.3.7. Hemodiyaliz ile ilgili yapılan veri madenciliği çalışmaları

Bellazzi et al. [130], hemodiyaliz hizmeti kalitesinin değerlendirilmesinde zamana bağlı veri madenciliği yaklaşımı ile birliktelik kuralı algortimasını başarılı bir şekilde bir araya getiren yeni bir model önermiştir.

Chittaro et al. [116] çalışmasında hemodiyaliz hastalarının hemodiyaliz seansı boyunca zamana bağlı ölçülen verilerini çubuk grafik metodu ile 3 boyutlu olarak görselleştirerek verilerin daha iyi analiz edilmesini sağlayan İnteraktif Paralel Çubuk Grafikler (*Interactive Parallel Bar Charts-IPBS*) bir sistem önermiştir.

Yeh ve Wu [73], hemodiyaliz hastalarının biyokimyasal verileri üzerinde zamana bağlı veri madenciliği uygulaması yaparak hastaların klinikte yatma sürelerini tahmin eden bir karar destek sistemi önermiştir.

Sai et al. [131], böbrek verilerini çok katmanlı algılayıcı ve geri beslemeli bir sinirsel ağ algoritması kullanılarak analiz etmiş ve böbrek hastalarının tespit edilmesinde kullanılabilecek kurallar elde etmeye çalışmıştır.

Kusiak et al. [132], hemodiyaliz hastalarına ait bir veri kümesi üzerinde makine öğrenimli Kaba küme ve karar ağacı algoritmaları uygulayarak hastaların hastalıktan kurtulma süresini tahmin etmeye çalışan bir model geliştirmiştir.

Kusiak et al. [133], çalışmasında hemodiyaliz hastalarının hastalıktan kurtulma sürelerinin tahmininde tedavi sürecinde kontrol altında tutulan 50'yi aşkın parametrenin toplu etkisini incelemek amacıyla dört farklı diyaliz merkezinden elde ettiği 188 hastaya ait veri kümesi üzerinde sırasıyla Kaba küme algoritması ve karar ağacı modeli kullanmıştır. Elde edilen karar

kuralları yardımıyla hastalıktan kurtulma süreleri 3 yıldan az ya da çok hastaların parametre etkileşimlerini incelemiştir. Ayrı ayrı uygulanan Kaba küme algoritması ve karar ağacı modeli sonuçlarını karşılaştırmıştır.

Bries [134] tez çalışmasında demografik, klinik ve vizite verilerinden oluşan bir veri kümesi üzerinde Kümeleme ve sınıflandırma metotları kullanarak diyaliz hastalarının kan hücreleri ile ilgili çeşitli modeller geliştirmiştir.

Sleeman et al. [88], hemodiyaliz seanslarında hemodiyaliz cihazları tarafından toplanan ve hastanın süreçteki kan basıncı, kan değerleri ve diğer fizyolojik durumunu belirten veriler üzerinde hiyerarşik kümeleme ve Bayes Ağları gibi makine öğrenimli teknikler kullanarak analiz etmiş ve sonuçların tekil hastaların klinik durumları ile ilgisini araştırmıştır.

Tez kapsamında yapılan uygulama çalışmasında Kusiak et al. [133] temel alınmıştır.

5. BİR UYGULAMA: HEMODİYALİZ HASTALARI İÇİN ÖLÜM RISKİ SEVİYELERİNE GÖRE RISK FAKTÖRLERİNİN ETKİLEŞİMLERİNİN İNCELENMESİ

Uygulama çalışması, hemodiyaliz hastalarına ait demografik ve klinik veriler ile aylık test verilerinden oluşan bir veri kümesi üzerinde her hasta için belirlenen risk skoru dikkate alınarak uygulanan veri madenciliği teknikleri yardımıyla risk faktörlerinin etkileşimlerinin (risk deseninin) incelenmesini içermektedir.

Uygulamanın anlatılacağı bölüm başlıklar halinde aşağıda özetlenmiştir:

- Problemin belirlenmesi ve verilerin anlaşılması
- Hastalık bilgisi (Son Aşama Kronik Böbrek hastalığı ve Hemodiyaliz Tedavisi)
- Verilerin toplanması,
- Verilerin hazırlanması
- Modelin kurulması
- Uygulama ve Elde Edilen Sonuçlar

5.1. Problemin Belirlenmesi ve Verilerin Anlaşılması

Veri madenciliği tekniklerinin tıp alanında kullanım konularından biri de kronik hastalıklardır. Kronik hastalıklar, hastanın hastalığa yakalandığı andan hayatının sonuna kadar devam eden, sürekli tıbbi gözetim ve tedavi altında bulunmayı gerektiren, en önemlisi hayati tehlike içeren hastalıklardır. Kanser, kalp hastalığı, kronik böbrek yetmezliği, AIDS gibi kronik hastalıklardan her yıl yüzbinlerce insan hayatını kaybetmektedir. Bu tehlikeli hastalıklar üzerinde yapılacak bilimsel çalışmalardan elde edilecek yeni bilgilerle yeni tedavi metotları geliştirilmesi bu hastalıklara yakalanmış pek çok insan için hayati önem taşımaktadır. Bu kronik hastalıklardan biri de kronik böbrek yetmezliğidir.

Böbreklerin kan filtrasyon hızının doğuştan, kalıtsal hastalıklar, kardiyovasküler hastalıklar, bağışıklık sistemi hastalıkları, idrar yollarını tıkayıcı diğer hastalıklar ya da metabolik hastalıklar nedeniyle geri dönüşümsüz bir şekilde azalmasına kronik böbrek yetmezliği adı verilir. Böbreklerin normal süzme kapasitesinin %10-%15'ini kullanabildiği son aşama kronik böbrek hastalığı durumunda böbrek fonksiyonlarının bir şekilde dışarıdan yerine getirilmesi gerekir. Bu amaçla diyaliz tedavisi veya böbrek nakli uygulanır.

Türkiye`de 2008 yılı sonu itibariyle 400'ü özel 837 diyaliz merkezindeki 13879 diyaliz cihazından diyaliz hizmeti alan hasta sayısı yaklaşık 45 bindir [135] . Hasta sayısının her yıl ortalama 5 bin artmakta olduğu düşünüldüğünde 2010 yılında diyaliz tedavisi gören hasta sayısının 55 bini geçmesi beklenmektedir. Yıllık yapılan böbrek nakil sayısının 600'lerle sınırlı olması, kronik böbrek yetmezliğinin daha korkulan bir hastalık olmasına yol açmaktadır. Kronik böbrek yetmezliği, hayati tehlikesi yanında ekonomik olarak da ülkeye çok ciddi bir yük getirmektedir. 2008 yılı için Türkiye`de diyalize giren hastaların devlete yıllık maliyeti 600 milyon dolara ulaşmıştır.

Bu konuda giderek artan hasta sayısı ve yurdumuzda açılan özel tıp merkezleri sayısının artışı ile birlikte Nefroloji alanında daha fazla bilimsel araştırma yapma ihtiyacı doğmaktadır.

Son yıllarda bilgisayar teknolojisinin sağlık sektöründe çok daha yaygın kullanımı ile birlikte Nefroloji alanında yapılan bilimsel çalışmalarda istatistiksel tekniklerin kullanımı, yerini Veri Madenciliği gibi daha kompleks tekniklerin kullanımına bırakmaktadır. Uygulama çalışmasında kullanmak üzere problem belirlemek amacıyla yapılan literatür taramasında, hemodiyaliz hastalarına ait az sayıda veri üzerinde tekil risk faktörleri ile ilgili istatistiksel teknikler kullanan pek çok çalışmaya rastlanmakla birlikte hemodiyaliz hastaları ile ilgili risk faktörlerinin birbirleri ile etkileşimi üzerinde

veri madenciliği teknikleri kullanan yalnız bir çalışmanın [133] mevcut olduğu dikkati çekmiştir.

Bu nedenle uygulama çalışmasında Hemodiyaliz hastalarına ait bir veri kümesi üzerinde veri madenciliği teknikleri kullanılarak risk faktörlerinin birbirleri ile etkileşimi (risk desenleri) incelenecektir.

5.2. Hastalık Bilgisi (Son Aşama Kronik Böbrek Hastalığı ve Hemodiyaliz Tedavisi)

Son Aşama Kronik Böbrek Hastalığı ve Hemodiyaliz Tedavisi konusunda bilgiler aşağıda özetlenmiştir.

Böbreklerimiz karın boşluğunun arka kısmında omurganın her iki tarafına yerleşik fasulyeye benzer organlardır. Her biri yaklaşık 10-12 cm uzunluğunda ve 150 gr ağırlığındadır. İki böbreğimiz olsa da, sağlıklı yaşam sürmek için tek sağlam böbrek yeterlidir.

Böbreklerimiz ana atardamardan (aort) çıkan iki büyük arter (renal arter) ile beslenir. Her bir böbreğimizde yaklaşık 1 milyon adet nefron denilen ünite vardır.

Nefronlar süzme işlevinin gerçekleştiği glomerül denilen yumakçık ve bunlara bitişik tüplerden oluşur. Tüplerde glomerüllerden süzülen vücut için gerekli bazı maddeler geri emilirken, atılması gereken maddeler idrara salgılanır. Temizlenmiş olan kan toplardamar (renal ven) aracılığıyla dolaşıma katılır. Tüplerden geçen idrar böbrek havuzuna (pelvis), buradan da üreterleri geçerek idrar torbasına (mesane) gelir ve işeme ile dışarı atılır. İdrar miktarı sıvı alma alışkanlığına bağlı olarak değişiklik gösterirse de, günde yaklaşık 1-2 litre kadardır.

Böbreklerimiz beyin, kalp, karaciğer ve akciğer ile birlikte yaşamsal öneme sahip organlarımızdan birisidir. Böbreklerin organizma içindeki görevleri;

- Metabolizma sonucu ortaya çıkan üre, kreatinin gibi zararlı yıkım ürünlerinin vücuttan atılması,
- Yaşamsal öneme sahip sodyum, potasyum gibi tuzların dengesinin ayarlanması,
- Kan basıncının düzenlenmesi,
- Kemik iliğinde kan yapımını sağlayan eritropoetin hormonunun yapımı,
- Kemik ve kasların normal yapısı ve işlevi için gerekli olan D vitamininin aktive edilmesi olarak özetlenebilir.

Böbrek yetmezliği, böbreğin tüm işlevlerinin ani veya yavaş bir şekilde kaybedilmesidir. Akut ve kronik (süreğen) olmak üzere iki tip böbrek yetmezliği vardır.

Akut Böbrek Yetmezliği, böbrek işlevlerinin birkaç saat veya hafta gibi kısa bir sürede aniden bozulmasıdır. Böbreğin özellikle zararlı yıkım ürünlerini atma ve sıvı-tuz dengesini düzenleme işlevleri bozulur. Böbrek işlevi, nedene bağlı olarak birkaç hafta veya ay süreyle bozuk devam edebilir ve bu sürede bazen geçici bir diyaliz uygulanması gerekebilir. Tedavisi yol açan nedene göre değişiklik gösterir. Çoğunlukla geri dönüşlü bir durumdur, ancak bazı hastalarda kalıcı böbrek işlev bozukluğu gelişebilir.

NKF-KDOQI (The National Kidney Foundation - Kidney Disease Outcomes Quality Initiative)'nin tanımına göre Kronik Böbrek Yetmezliği, böbrek hasarı veya glomerüler filtrasyon hızının (GFR) 3 veya daha fazla ay için 60 ml/dak veya 1.73m^2 altında olması olarak tanımlanmaktadır [136]. Buna göre Kronik böbrek hasarı evreleri aşağıdaki Çizelge 5.1'de özetlenmiştir.

Kronik Böbrek Yetmezliği, böbreğin tüm işlevlerinin kalıcı ve ilerleyici bir şekilde bozulmasıdır. Geri dönüşlü bir durum olmamakla birlikte hipertansiyonun etkin bir şekilde kontrolü, diyet uygulanması, yeterli sıvı alınması, bazı ilaçların kullanımından kaçınılması ve düzenli kontrol gibi önlemlerle böbrek işlev bozukluğunun ilerlemesini önemli ölçüde yavaşlatmak olasıdır. Ülkemizde ve birçok ülkede kronik böbrek yetmezliğinin en sık görülen iki nedeni şeker hastalığı ve hipertansiyondur. Bu hastalıklarda tuz kısıtlanması, düzenli egzersiz yapılması ve kilo verilmesi gibi yaşam tarzı değişikliklerinin uygulanması, kan şekerinin ve kan basıncının etkin bir şekilde kontrol edilmesi ve düzenli izlem gibi önlemlerle böbrek yetmezliği gelişimini önlemek mümkündür.

Çizelge 5.1 Kronik böbrek hasarı evreleri [137]

Evre	Tanım	Glomerular Filtrasyon Hızı (GFR)
Yüksek Riskli	Böbrek hastalığı için Risk faktörleri (Diyabet, yüksek tansiyon, kalıtım, yüksek yaş, vb.)	GFR > 90
1	Böbrek hasarı (idrarda protein bulunması) ve normal GFR	GFR > 90
2	Böbrek hasarı ve GFR'de hafif düşüş	60 < GFR < 89
3	GFR 'de orta derecede düşüş	30 < GFR < 59
4	GFR 'de ciddi düşüş	15 < GFR < 29
5	Böbrek Yetmezliği (diyaliz veya böbrek nakli gerekli)	GFR < 15

Kronik Böbrek Yetmezliğine çeşitli nefritler, taş hastalığı, idrar yollarındaki yapısal bozukluklar, enfeksiyonlar, özellikle polikistik hastalık olmak üzere böbreğin kalıtsal hastalıkları, bağışıklık sistemi hastalıkları, doğuştan gelen böbrek yapı bozuklukları gibi durumlar yol açmaktadır [138].

Böbrek işlevinin ileri derecede bozulduğu son evrede (süzme işlevinin % 85'ten fazlasının yitilmesi) yaşamın devam edebilmesi için diyaliz veya böbrek nakli tedavilerinin uygulanması gerekir.

Diyaliz yöntemleri, böbreğin sadece birikime uğrayan atık maddelerin atılması ve vücutta sıvı ve tuz dengesinin sağlanması işlevlerini yerine getirirken, başarılı bir böbrek naklinden sonra nakli yapılan böbrek organizmanın ihtiyacı olan tüm işlevleri yerine getirir. Bu nedenle, böbrek yetmezlikli hastalarda seçkin tedavi yöntemi böbrek naklidir. Diyalize aday veya diyaliz tedavisi gören tüm hastaların böbrek nakli için değerlendirilmesi gerekir.

Böbrek nakli sağlıklı canlı akrabalarından veya kadavradan (beyin ölümü gerçekleşmiş kişi) alınan böbreklerle yapılabilir. Canlı akrabalarından yapılan nakillerin başarı şansı daha yüksektir. Kan ve doku grubu uyumu olmadan da böbrek nakli yapılabilmeyle beraber, nakledilen böbreğin uzun süreli işlevi bakımından alıcı ile verici arasında kan ve doku grubu uyumu olması önemlidir.

Böbrek nakli sayısının sınırlı olması ve doku uyuşması, nakil sırası bekleme gibi nedenlerle zorluğu, hastaları diyaliz tedavisi alma durumunda bırakmaktadır. Hastanın durumuna bağlı olarak iki çeşit diyaliz tedavisi vardır.

Hemodiyaliz, vücut dışına alınan kanın makine aracılığıyla yapay bir filtreden (diyalizör) geçirilerek birikime uğrayan zararlı maddelerden ve fazla sıvıdan arındırılması ve sonra vücuda geri verilmesi işlemidir.

Periton Diyalizi, önceden yerleştirilen bir kateter aracılığıyla belirli aralıklarla yaklaşık 2 litrelik diyaliz sıvısının karın boşluğuna doldurulması, bekletilmesi ve boşaltılması şeklinde uygulanan bir diyaliz yöntemidir. Kanda birikime uğrayan atık maddeler ve sıvı fazlası periton zarını geçerek karın

boşluğundaki diyaliz sıvısına geçer ve boşaltılan diyaliz sıvısı ile de vücuttan uzaklaştırılır. Periton diyalizi tedavisi için, olası diyaliz zamanından yaklaşık 3 hafta önce karın boşluğuna ameliyatla silikondan yapılmış yumuşak bir kateterin yerleştirilmesi gerekir. Periton diyalizi bazı tip barsak hastalıkları, karın içinde yapışıklılar, fıtık ve ciddi psikiyatrik bozukluğu olmayan tüm hastalara uygulanabilir. Kalp ve damar hastalıkları, damar giriş yeri sorunu, kanama eğilimi, şiddetli kansızlık, bazı karaciğer rahatsızlıkları, şeker hastalığına bağlı göz sorunları olan hastalarda, daha serbest sıvı ve beslenme, daha aktif yaşam isteyen, iğne girişlerinden korkan, çalışan ve okuyan hastalarda periton diyalizinin özellikle tercih edilmesi gerekir.

Sürekli ayaktan veya aletli olmak üzere iki tip periton diyalizi yöntemi vardır. Periton diyalizi yöntemi hemodiyalize göre daha az tercih edilen bir yöntemdir.

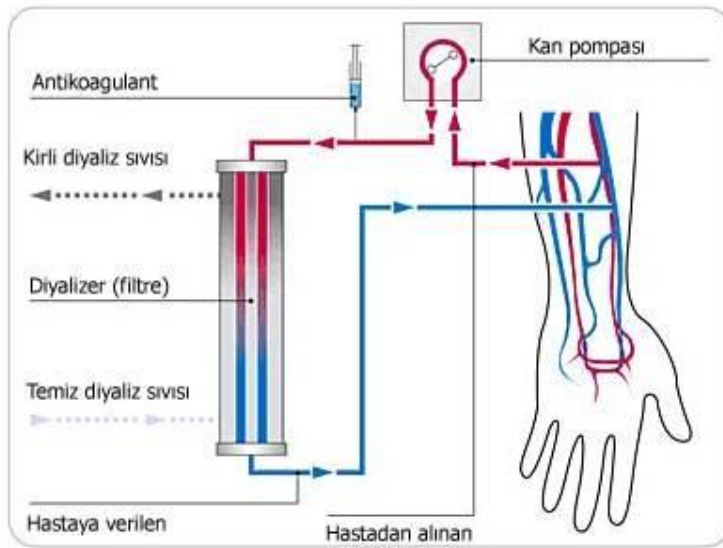
Hemodiyaliz

Diyaliz basit olarak yarı geçirgen bir zarla ayrılmış, bir tarafında temizlenmesi gereken kanın, diğer tarafında ise diyaliz sıvısının bulunduğu iki odacıklı bir sistemdir. Uygulama sırasında vücutta eksik olan bazı tuzlar da diyaliz sıvısından vücuda geçer. Hemodiyaliz uygulaması için kolda atardamar ile toplardamarın birleştirilmesi (fistül) ameliyatının yapılması gerekir. Fistül ameliyatından en az 3-4 hafta geçtikten sonra hemodiyaliz tedavisine başlanabilir. Acilen diyalize alınması gereken hastanın fistülü yok veya yeterince gelişmemişse, boyun veya kasıktaki orta genişlikteki toplardamarlardan birine çift kanallı kateter yerleştirilerek de hemodiyaliz uygulanabilir. Hemodiyaliz tedavisi hastanın klinik seyrine göre hemodiyaliz merkezlerinde haftada nadir hastalarda 1-2 gün, standart 3 gün uygulanır ve her tedavi seansı yaklaşık 4 saat sürer.

Hemodiyaliz sırasında her defasında fistül olarak isimlendirilen damara biri vücuttan kanı temizlenmek üzere filtreye götüren, diğeri temizlenmiş kanın

vücuda dönmesini sağlayan olmak üzere iki iğne yerleştirilir. Tedavi sırasında vücut dışına alınan kanın pıhtılaşmasını önlemek için kan sulandırıcı (antikoagülasyon) ilaçların uygulanması gerekir. Haftalık seans sayısı, seansın süresi, diyaliz sırasında kullanılacak filtre, diyaliz sıvısı hastanın bireysel özelliklerine göre tedavi ekibi tarafından ayarlanır.

Şekil 5.1’de hemodiyaliz süreci şematik olarak anlatılmıştır.



Şekil 5.1 Hemodiyaliz Süreci

Hemodiyalizin Olumlu Yönleri

- Zararlı maddeler ve fazla sıvı vücuttan hızlı ve etkin bir şekilde uzaklaştırılır.
- Kullanılabilme süresi periton diyalizinden daha uzundur.
- Daha yakın bir doktor ve hemşire kontrolü sağlar.
- Tedavi sırasında diğer hastalarla tanışılarak dostluk kurulması hastaya psikolojik olarak avantaj sağlar.

Hemodiyaliz Olumsuz Yönleri

- Hemodiyaliz aralıklı uygulanan diyaliz yöntemi olduğundan tedavi seansları arasında sıvı, tuz ve bazı besinlerin sınırlı tüketilmesi gerekir.
- Tedavinin uygulanması için sıklıkla haftada 3 gün bir hemodiyaliz merkezine gidilmesi iş, okul ve sosyal yaşantıyı olumsuz yönde etkileyebilir.
- Tedavi sırasında kan basıncı düşmeleri ve kas krampları görülebilir.
- Kanın pıhtılaşmasını önlemek için uygulanan ilaçlar kanamalara neden olabilir.
- Hemodiyaliz uygulanması için oluşturulmuş olan fistülde zamanla sorunlar ortaya çıkabilir ve yeni fistül ameliyatlarının yapılması gerekebilir.

Ülkemizde diyaliz tedavisi gören hastalar Sosyal Güvenlik Kurumu güvencesinde kamu veya özel sektöre ait yetkili sağlık kuruluşlarında bu tedaviyi görebilmektedirler. Yasal anlamda diyaliz tedavisi ile ilgili Sağlık Bakanlığı tarafından çıkarılan Diyaliz Merkezleri Yönetmeliği yürürlüktedir. Bu yönetmelik gereği özel diyaliz merkezlerinde hastalara ait gerekli resmi, tıbbi kayıtlar tutulmaktadır.

Çalışmada özel bir diyaliz merkezinin hemodiyaliz tedavisi gören hastalarına ait demografik ve tıbbi verilerden, yetkililerin bilgi ve izni dahilinde tıbbi etik kuralları ve yasalar gereği gizliliği korunacak şekilde yararlanılmıştır.

5.3. Verilerin Toplanması

Bu çalışmada verileri üzerinde çalışılacak hastalar, Özel Can Diyaliz Merkezi'nde Hemodiyaliz tedavisi görmüş / görmekte olan hastalardır.

Özel Can Diyaliz Merkezi 29 Şubat 1996'da Ankara'da faaliyete geçmiş, halen Yahya Kemal Caddesi Bergama Sokak No: 10 Yenimahalle / ANKARA adresinde bulunan ve hemodiyaliz merkezi olarak projelendirilip inşa edilmiş, Türkiye'de bir başka örneği olmayan, 3500 m² açık, 3000 m² kapalı alana sahip 5 katlı yeni binasında,

- özel odalarda (VIP) hemodiyaliz tedavisi sunan tek merkez olma özelliği,
- binası içinde yer alan her tür hematolojik ve biyokimyasal tetkik işleminin yapılabilirdiği Özel Ankalab Laboratuvarları,
- 6 Hekim, 8 Hemşire, Psikolog, Diyetisyen, Sosyal Hizmet Uzmanı, Teknisyen, diğer yardımcı sağlık personeli ve idari personel dahil olmak üzere toplam 45 kişiden oluşan eğitimli ve güleryüzlü personeli,
- Hastaların Ankara'nın her semtinden ve 4 ilçesinden diyaliz merkezine ulaşımını sağlayan 6 adet servis aracından oluşan filosu ile sağlık hizmeti vermeye devam etmektedir.

Merkezde tedavi sırasında kullanılan tüm malzemeler, dünya tıbbının kabul ettiği standartlar esas alınarak özenle seçilmekte ve bir kullanım sonunla kesinlikle atılmaktadır.

Özel Can Diyaliz Merkezi binasına ait resimler Resim 5.1 ve 5.2'de görülmektedir.



Resim 5.1 Özel Can Diyaliz Merkezi Binası - Giriş



Resim 5.2 Özel Can Diyaliz Merkezi Binası - Genel görünüm

Hemodiyaliz işlemi için 20 adet Fresenius 4008B model Hemodiyaliz makinesi kullanılmaktadır. Resim 5.3'te bu makinenin resmi verilmiştir.



Resim 5.3 Fresenius 4008B model Hemodiyaliz makinesi

Verilerin eldesi ve özellikleri

Verilerin toplanması, Özel Can Diyaliz Merkezi arşivinde yaklaşık 1 ay süren bir çalışma (verilerin uygun tablo formatına manuel olarak girilmesi) sonucunda ile tamamlanmıştır. Verilerin toplanması çalışmasında Özel Can Diyaliz Merkezi arşivini gösteren bir resim Ek-1'de yer alan Resim 1.1'de sunulmuştur.

Her hasta için tutulan

- Hasta Bilgi formundan hastalara ait demografik veriler, etiyoloji ve komorbid durumlarına ilişkin veriler,
- Aylık Hekim İzlem formlarından hastalara ait rezidiv idrar, EPO tedavisi bilgileri, diyet durumu vb. veriler,
- Seans İzlem formlarından diyaliz seansı ve tedavi verileri,
- Rutin Tetkik İzlem formlarından periyodik yapılan tıbbi testlere ait biyokimyasal veriler elde edilmiştir.

Verilerin toplandığı formlara ait örnekler, Ek-2'de yer alan Resim 2.1-2.3'te Hasta Bilgi Formu, Resim 2.4'te Aylık Hekim İzlem Formu, Resim 2.5'te Seans İzlem Formu, Resim 2.6'da Rutin Tetkik İzlem Formu yer alacak şekilde gösterilmiştir.

Hemodiyaliz tedavisi standart etkinliğine yaklaşık 3 aylık bir başlangıç devresi sonunda ulaşmaktadır. Kullanılacak verilerin güvenilir olması için uzmanlar tarafından hastaların 12 aylık tedavi sürecine ilişkin verilerin ortalamasının alınması uygun görülmüştür.

Yapılan çalışmada Özel Can Diyaliz Merkezi'nden 12 ay ve daha uzun süredir tedavi görmekte/görmüş olan hastaların bu sürece ait verileri dikkate alınmıştır.

Diyaliz merkezi arşivinde yapılan çalışmada 12 ay ve uzun süre tedavi gören/görmüş 170 hastanın kayıtlarına ulaşılmıştır. Bu hastalardan bir kısmı halen tedavi görmekte iken bir kısmı başka bir diyaliz merkezine Nakil, periton diyalizine geçiş, iyileşme, transplantasyon olma, ya da vefat etme gibi durumlar nedeniyle tedavisi sonlanmış durumdadır. 3 hastanın ayrılış nedeni ile ilgili kayıt bulunamamıştır. Çalışma kapsamına giren hastaların halen tedavi görme durumlarına göre dağılımı bilgisi Çizelge 5.2’de gösterilmiştir.

Çizelge 5.2 Çalışma kapsamındaki hastaların halen tedavi görme durumlarına göre dağılımı

Durum	Sayısı	Yüzdesi
Tedavide	54	31,8
EX	68	40,0
Nakil	34	20,0
Transp.	8	4,7
Periton	2	1,2
İyileşti	1	0,6
Belirtilmemiş	3	1,8
Toplam	170	100

Periyodik tıbbi testlere ait verilerde 12 aylık ortalamalarının temel alınması ile bu verilerin doğruluğu ve güvenilirliği artmış, böylece hatalı ya da ortalamadan sapan değerlerin sayısının en aza indirgenmesi amaçlanmıştır.

Benzer bir durum diyaliz seansı verileri için geçerlidir. Hemodiyaliz hastaları ile ilgili Kusiak [133] tarafından yapılan veri madenciliği çalışmasında veri sapmalarını azaltmak için 15 diyaliz seansına ait verilerin ortalaması alınmıştır. Bu çalışmada da veri kümesine her hasta için verilerin derlendiği tarihten geriye doğru girdiği son 15 diyaliz seansına ait verilerin ortalamaları kullanılmıştır.

Veri kümesinin düzenlenmemiş ilk hali, Özel Can Diyaliz Merkezi'nde 12 aydan fazla bir süredir hemodiyaliz tedavisi görmüş toplam 170 hastanın 76 farklı parametre verilerinden oluşmaktadır. Veri kümesi ilk boyutu 170x76'dır. Veri kümesini oluşturan veriler iki ana başlık altında incelenebilir:

1)Hastaların Demografik Verileri : Demografik veriler kısmını oluşturan veri sütunu başlıklar ve içerikleri şunlardır :

Hasta ID : Her bir hasta için 1 adet olmak üzere 1'den 170'e kadar verilen tanımlama numarasıdır.

Cinsiyet : Hastanın cinsiyet durumunu içerir. Kadın ve Erkek olarak iki tiptir. Veri kümesi gösteriminde Kadın için 0, Erkek için 1 rakamı kullanılmıştır.

Yaş : Hastanın yaş durumunu içerir. Birimi yıl'dır.

Tedavi görmekte olan hastalar için yaş hesabı, veri toplama çalışmasının tamamlandığı 26.05.2009 tarihi baz alınarak "Yaş = 26.05.2009 – Doğum Tarihi" formülü kullanılarak yapılmıştır.

Tedavi görme durumu dışında herhangi bir durumdaki hastalar için yaş hesabı, "Yaş= Ayrılış Tarihi – Doğum Tarihi" formülü kullanılarak yapılmıştır. Hastalara ilişkin doğum tarihleri, diyaliz merkezine müracaat tarihleri ve varsa ölüm/ayrılış tarihleri ve nedenleri ile hesaplanan yaşlarına ilişkin bilgiler Ek-3'te yer alan Çizelge 3.1'de gösterilmiştir.

Kan Grubu : Hastanın kan grubunu bilgisini içerir. 4 tip kan grubu ve kan grubu bilinmeyen durumlar için toplam 5 farklı durum belirtecinden oluşur.

Boy : Hastanın boy ölçüsünü içerir. Birimi cm'dir.

Ağırlık : Hastanın ağırlık ölçüsünü içerir. Birimi kg'dır. Ampute (bacağı/bacakları kesik) hastalar için mevcut ağırlığı 1,15 ile çarpılarak hesaplanır [139].

VKI (Vücut Kitle İndeksi) : Hastanın ağırlığı boyunun karesine bölünerek elde edilir. Birimi kg/m^2 'dir.

Tansiyon : Hastanın Büyük (sistolik) ve Küçük (diastolik) tansiyon değerlerini içerir. Birimi mmHg'dir.

Sigara Kullanımı : Hastanın sigara kullanım durumunu belirtir.

Alkol Kullanımı : Hastanın alkol kullanım durumunu belirtir.

2) Hastalara ait Klinik Veriler (Teşhis ve Tedavi Süreci): Klinik veriler üç alt başlıktan oluşur :

a) **Tıbbi Dosya Verileri** : Her hasta için hastanın tıbbi durumuna dair klinik tanı ve verilerden oluşur.

Hastalık Etiyolojisi : Hastanın Kronik Böbrek Yetmezliği hastalığına yakalanmasına yol açan ana nedeni teşkil eden hastalıktır.

Komorbid Durumlar : Hastanın bünyesinde Kronik Böbrek Yetmezliği haricinde mevcut olan diğer kronik hastalık ve rahatsızlıklardır. Sayılarına bağlı olarak ana veri kümesinde beş sütunluk bir alan ayrılmıştır.

Beslenme Risk İndeksi (NRI) : Hastanın Beslenme İndeksi değerini içerir.

“NRI = $(1.489 \times \text{serum albümin, g/L}) + [41.7 \times (\text{şimdiki ağırlık} / \text{İdeal ağırlık})]$ ” formülü ile ana tablodaki veriler kullanılarak hesaplanmıştır [140]. Birimi yoktur.

Transplantasyon durumu : Hastanın böbrek nakline Uygun olup olmadığı ile ilgili veri olup “Uygun” ve “Uygun Değil” şeklinde iki yanıt içerir. 65 yaş üstü

ve kronik kalp hastalığı gibi etkenlerin Transplantasyon yapılmasını engellediği bilinmektedir.

Rezidiv İdrar Miktarı : Böbrek süzme fonksiyonun aylık kontrolü için kullanılan bir parametredir. Normal bir insan günde 1,5-2 lt idrar üretirken, kronik böbrek yetmezliği hastası bir kişi böbreği süzme fonksiyonunu büyük ölçüde kaybettiğinden çok azdır. Hastalar için 12 aylık Rezidiv idrar miktarı ortalamaları kullanılmıştır.

b) Tedavi Verileri : Tedavi verileri üç alt başlıktan oluşur :

(i)Diyet Durumu Verileri : Hastaya standart ya da diyabetik diyet uygulanması durumunu belirtir.

(ii)Anemi Tedavisi Verileri : Hastanın anemi tedavisi görüp görmediği ve görüyorsa kullandığı ilaç tipi ve dozaj verilerini içerir. Bu veriler :

EPO (Eritropoetin hormonu) İlaç tipi : Piyasada anemi tedavisinde Epoetin ihtiyacını karşılayan Eprex, Aranesp ve Neorecormon adlı üç ilaçtan hangisini kullandığını ya da ilaç kullanmadığını belirten verilerdir. Anemi tedavisi sürecinde 12 aylık ilaç dozajları dikkate alındığında birden çok ilaç çeşidi kullanma durumu ortaya çıkan hastalar için, süreç içinde diğerlerine göre daha çok kullanılan ilaç tipi dikkate alınmıştır.

EPO İlaç Dozajı : Anemi tedavisi gören hastalar için 12 aylık ilaç dozajı ortalaması alınmıştır. 12 aylık süreç içinde birden fazla ilaç tipi kullanan hastalar için ortak bir dozaj hesaplaması yapmak gerekmektedir. Üç farklı ilacın etken maddeleri farklı olsa da dozaj anlamında birbirlerine çevrilmeleri mümkündür. Bu konuda kullanılacak formül “1mcg darbepoetin = 200 ünite epoetin alfa /epoetin beta” olacaktır [141].

(iii) Diyaliz Tedavisi Verileri :

Alınan Toplam Seans Sayısı : Hastanın tedavi gördüğü toplam seans sayısıdır.

Tedavi Yaşı : Aylık Hekim İzlem formlarında yer alan ve hastanın ay olarak tedavi gördüğü süreyi belirten veridir.

Diyalizde Yaşam Süresi : Ana tabloda hastanın birimi ay olan tedavi yaşının 12'ye bölünmesi ile elde edilen hastanın yıl olarak tedavi gördüğü süredir.

Hemodiyaliz Seansı Verileri : Hastanın gördüğü en son 15 seansta ölçümü yapılan verilerden oluşur. Bu veriler :

**Haftalık Seans Süresi:* Haftada görülen seans sayısı ve süresi çarpımıdır.

**Kan Akım Hızı :* Kanın Diyaliz makinasından geçişi boyunca akış hızıdır.

**Antikoagülasyon Dozu :* Kanın damar dışında iken pıhtılaşmasını engellemek için antikoagülasyon ilacı kullanılır. 15 seanslık süreçte bazı hastalar birbirinden farklı dozajları olan Heparin, Fluxum (Parnaparin), Clexane (Enoxoparin) ve Fragmin (Dalteparin) olmak üzere dört farklı tip antikoagülasyon ilacından bazılarını birlikte kullanmıştır. Bu durumda dozaj hesaplaması, uzman görüşü olarak alınan 5000 Heparin = 5000 Fragmin = 0,4 Clexane = 0,4 Fluxum eşitliğini dikkate alarak Heparin'in biriminde olacak şekilde yapılmıştır.

**Damar Yolu Tipi :* Hastayı diyaliz makinasına bağlamak için kullanılan kateterler/fistüller, hastanın durumuna bağlı olarak cerrahi operasyonla yerleştirildikleri uygun damar yolunun yerine göre adlandırılır. Veri kümesinde 8 farklı damar yolu tipi mevcuttur. Veri kümesinde nominal hale getirilirken 0'dan 7'ye kadar artan sayılarla sembolize edilen bu damar yolları sembol sayı sıralarına göre;

LAVF (Left ArterioVenous Fistule-Sol Arteriovenöz Fistül)

RAVF (Right ArterioVenous Fistule-Sağ Arteriovenöz Fistül)

LSVC (Left Subclavian Venous Catheter – Sol subklavian venöz Kateter),

RSVC (Right Subclavian Venous Catheter – Sağ subklavian venöz Kateter),

LIJVC (Left IntraJuguler Venous Catheter - Sol IntraJuguler venöz Kateter),

RIJVC (Right IntraJuguler Venous Catheter - Sağ IntraJuguler venöz Kateter)

LFVC (Left Femoral Venous Catheter – Sol Femoral venöz Kateter),

RFVC (Right Femoral Venous Catheter – Sağ Femoral venöz Kateter)

Dikkate alınan 15 seanslık süreçte bazı hastalar için farklı damar yolu tipleri kullanılmakla birlikte veri kümesinde diğerlerine göre daha fazla kullanılan damar yolu yer almıştır.

**Pre Diyaliz Ağırlığı* : Seans öncesinde ölçülen kg cinsinden hasta ağırlığıdır.

**Pre Diyaliz Sistol* : Seans öncesinde ölçülen hastanın sistolik kan basıncıdır.(büyük tansiyon)

**Pre Diyaliz Diyastol* : Seans öncesinde ölçülen hastanın diyastolik kan basıncıdır.(küçük tansiyon)

**Post Diyaliz Ağırlığı* : Seans sonrasında ölçülen kg cinsinden hasta ağırlığıdır. Literatürde kuru ağırlık olarak da geçer ve hesaplamalarda temel olarak kullanılır.

**Post Diyaliz Sistol* : Seans sonrasında ölçülen hastanın sistolik kan basıncıdır.(büyük tansiyon)

**Post Diyaliz Diyastol* : Seans sonrasında ölçülen hastanın diyastolik kan basıncıdır.(küçük tansiyon)

**Kilo Farkı* : Seans sonrası ölçülen hasta ağırlığı ile seans öncesi yapılan ağırlık ölçümünün farkıdır. Hekimler tarafından her seans için 2-2,5 kg arasında tutulmak istenir. Ana tabloda 15 seanslık ortalama seans sonrası ve öncesi ağırlıklar arası fark olarak hesaplanmıştır.

c) *Hastalara ait Periyodik Test Verileri* : Periyodik test verileri test yapılış periyoduna göre üç alt başlıktan oluşur. Bu kısımdaki veri sütunlarının başlıkları ve içeriği şunlardır :

(I) *Aylık Test Verileri* : Ölçümü aylık yapılan test verilerinin başlıkları ve birimleri şunlardır : Hb (Hemoglobin), Htc (Hematokrit), MCV (Mean Corpuscular Volume), MCH (Mean Corpuscular Hemoglobin), MCHC (Mean Corpuscular Hemoglobin Concentration), Lökosit, Platelet, Na (Sodyum), K (Potasyum)(Giriş), K (Potasyum)(Çıkış), K (Potasyum) Farkı, Glisemi, Üre-BUN Giriş, Üre-BUN Çıkış, BUN Ratio, Kreatinin (Giriş), Kreatinin (Çıkış), URR (Urea Reduction Ratio), Kt/V (Üre Temizleme Oranı), Ca (Kalsiyum), P (Fosfor), Ca x P (Kalsiyum-Fosfor Çarpımı), ALT (Alanin Transaminaz), Toplam Protein, Albümin.

(II) *3 Aylık Test Verileri* : Ölçümü 3 ayda yapılan test verilerinin başlıkları ve birimleri şunlardır : CRP (C Reaktif Protein), Ürik Asit, PTH (Paratiroid hormone), Alkalen Fosfataz, Ferritin, Fe (Demir), Fe B.K. (Demir Bağlama Kapasitesi), TSI (Transferrin Satürasyon İndeksi), Venöz Bikarbonat.

(III) *6 Aylık Test Verileri* : Ölçümü 6 ayda bir yapılan test verilerinin başlıkları ve birimleri şunlardır : T. Kolesterol (Total Kolesterol), HDL (High Density Lipoprotein), LDL (Low Density Lipoprotein), Triglicerid.

Ana tablo olarak da anılan ana veri kümesi ve bu veri kümesini oluşturan seans, tedavi ve tıbbi testlere ilişkin verilerin hesaplandığı çizelgeler Ek-3'te yer alan Çizelge 3.2, 3.3, 3.4 ve 3.5'te kısmi olarak verilmiştir. Anılan çizelgelerin tamamı tezin ekinde yer alan CD içinde elektronik ortamda mevcuttur.

5.4. Verilerin Hazırlanması

Elde edilen ana veri kümesi üzerinde bu bölümde anlatılan işlemler yapılarak en anlamlı özelliklerden oluşan eksiksiz ve işlenebilir bir veri kümesi elde edilmeye çalışılmıştır.

5.4.1. Eksik verilerin tamamlanması

Diyaliz hastalarının yapılması Sağlık Bakanlığı tarafından çıkarılan Diyaliz Merkezleri Yönetmeliği'nde zorunlu olarak belirtilen periyodik laboratuvar testlerinin kapsam değişikliği nedeniyle kayıtlarda bazı hastalar için bazı özelliklerin ölçümü yer almamaktadır.

Bu özelliklere ait boş veri oranı bilgisi, boş hücelere doldurulan veriler ve boş hücrelerin ID leri Çizelge 5.3'te belirtilmiştir.

Çizelge 5.3 Eksik veri içeren özelliklere ait boş veri oranı bilgisi, boş hücelere doldurulan veriler ve boş hücrelerin ID leri

Eksik Verisi Olan Özellik	Eksik Verilerin		Yerine Girilen Değer
	Oranı	Yüzdesi	
CRP	51/170	30	4,72
PTH	6/170	3,5	234,13
T. Kolesterol	1/170	0,6	177,7
LDL	5/170	31,1	TND formülü ile hesaplandı.
	48/170		81,7
HDL	52/170	30,6	39,89
Trigliserid	1/170	0,6	227,71
Ferritin	4/170	2,3	504,47
Venöz Bikarbonat	66/170	38,8	Değer girilmedi. Özellik veri kümesinden çıkarıldı.

HDL özelliğindeki 4 adet boş hücre (ID No : 92,105,116,127) Türk Nefroloji Derneği web sayfasında belirtilen Friedewald Formülü

“ $LDL = \text{Total Kolesterol} - (\text{Trigliserid} / 5) - HDL$ ”Trigliserid düzeyi <400 mg/dL olan ve LDL, T. Kolesterol değerlerinin bilindiği durumlar için kullanılarak yapılan hesaplama ile doldurulmuştur [142].

Venöz Bikarbonat özelliği sonradan Diyaliz Merkezleri Yönetmeliği'nde yapılan bir değişiklik sonrası ölçümü zorunlu kapsama alındığından fazla sayıda boş hücre içerdiğinden (66/170,%38,8) veri kümesinden çıkarılmıştır.

Her hasta için veri toplama süreci öncesindeki son 15 hemodiyaliz seansının verileri, veri kümesinde ortalamaları kullanılmak üzere toplanırken bazı hastalar için 15'ten daha az hemodiyaliz seansı verisine ulaşılabilmektedir. Ancak verilerine ulaşılamayan seans sayısı toplam seans sayısının binde 9'u kadardır ve seans verilerinin ortalamalarını alındığı düşünüldüğünde ihmal edilebilecek bir orandır.

Bu hastaların ID'leri ve verilerine ulaşılamayan hemodiyaliz seansı sayısı aşağıdaki Çizelge 5.4'te belirtilmiştir.

Çizelge 5.4 Eksik hemodiyaliz seansı verileri

Hasta ID	Verisine Ulaşılamayan Seans Sayısı
17	1
59	1
118	3
126	1
127	1
148	12
155	1
159	2
Toplam	8 hasta
Verilerine tam ulaşılan	162 hasta
	2430
	0,9
	Verilerine Ulaşılabilen Seans Sayısı
	Eksik Veri Yüzdesi

Her hasta için veri toplama süreci öncesindeki son 12 ayın tıbbi test verileri, veri kümesinde ortalamaları kullanılmak üzere toplanırken sadece 147 ID nolu hasta için bu veriler 12 aydan daha az bir süre için tıbbi test verisine ulaşılabilmiştir. Hastanın tedavi süreci 7 ay sürse de tedavisinin Ex (Ölü) olması nedeniyle sonlanması ve yaşının fazlalığı dikkate alınarak veri kümesine dahil edilmiştir. Bu durum tıbbi test verilerinin ortalamalarını alındığı düşünüldüğünde ihmal edilebilecek bir durumdur. Hastalar için ID'lere göre 12 aylık tıbbi test verilerinin kullanıldığı tarihleri gösteren Ek-3'teki Çizelge 3.6'da yer almaktadır.

5.4.2. Veri kümesine eklenen özellikler

Hastalara ait demografik verilerden doğum tarihi verileri diyaliz merkezine müracaat tarihlerinden çıkarılarak her hasta için tedavi başlangıç yaşı hesaplanmış ve ayrı bir özellik olarak veri kümesine dahil edilmiştir.

Literatür taramasında “Üre-BUN (G) / Üre-BUN (Ç)” formülü ile hesaplanan BUN Ratio adlı yeni bir parametreye rastlanmış [143] ve eldeki Üre-BUN (G) ve Üre-BUN (Ç) özellikleri kullanılarak bu oran hesaplanıp veri kümesine yeni bir özellik olarak ilave edilmiştir.

Vücut Kitle İndeksi diyaliz hastaları için post-diyaliz ağırlığı (kuru ağırlık) kullanılarak “VKİ = post-diyaliz kuru ağırlığı [kg] / Boyun karesi [m²]” formülü ile hesaplanmıştır [144,145].

Beslenme Risk indeksi (NRI) serum albümin değeri ve kilo kaybının oransal olarak kullanıldığı diyaliz hastaları için önemli bir parametredir.

“NRI= (1.489Xserum albümin, g/L) + [41.7X(şimdiki ağırlık/İdeal ağırlık)]” formülü ile hesaplanmıştır. Bu formülde belirtilen İdeal Ağırlığın hesaplanmasında Lorentz Formülü erkekler için $WLo = Boy-100-[(Boy-150)/4]$ ve

kadınlar için $WLo = Boy - 100 - [(Boy - 150) / 2.5]$ olarak kullanılmıştır [146].

EPO Direnci, hasta vücudunun alınan eritropoetine karşı gösterdiği olumsuz reaksiyon olup diyaliz hastalarının gördüğü anemi tedavisinin değerlendirmesinde önemli bir parametredir. EPO Direnci için “Doz verimliliği / Ortalama Hemoglobin miktarı” formülü kullanılmıştır. Doz Verimliliği ise “EPO dozajı / Post-diyaliz kuru ağırlığı” formülü ile hesaplanır [147]. Veri kümesinde mevcut verilerle yukarıdaki formüller kullanılarak Doz Verimliliği ve EPO direncinin hesaplanmış ve veri kümesine yeni birer özellik olarak ilave edilmiştir.

5.4.3. Dönüştürülen özellikler

Hemodiyaliz hastaları için risk değerlendirmesinde aylık ölçülen test verilerinden olan Ca (Kalsiyum) ve P (Fosfor) mineral değerleri çarpım şeklinde dikkate alınmaktadır. Bu nedenle bu iki özellik veri kümesinde çarpım şeklinde CaXP adı altında bir özellik şeklinde gösterilmiştir.

Transferrin Satürasyon Oranı Fe (Demir) seviyesinin Fe B.K.(Demir Bağlama Kapasitesi) seviyesine bölünmesi ile elde edildiğinden Fe ve Fe B.K. özellikleri veri kümesinden çıkarılmıştır.

Seans Sayısı ve Seans Sıklığı özellikleri çok sayıda aynı sayısal değer içerdiğinden, bu iki özellik değerlerinin çarpımı ile hesaplanan Haftalık Seans Süresi adlı özellik altında tek bir özellik olarak gösterilmiştir.

EPO tedavisi verilerinde ilk aşamada 12 aylık ortalama EPO dozajı ortalama olarak Aranesp ilacı kullanan hastalar için mikrogram/hafta biriminde, ortalama olarak Neorecormon veya Eprex kullanan hastalar için ünite/hafta veya biriminde olarak belirlenmiştir. Ardından aylık ortalama dozaj miktarı Neorecormon veya Eprex kullanan hastalar için ünite/hafta x 4 işlemi ile, Aranesp kullanan hastalar için ise dozaj eşitleme için (mikrogram/hafta x 4) x

200 işlemi yapılmıştır [141]. Sonuçta elde edilen yıllık ortalama dozaj miktarı bir özellik olarak veri kümesinde yer almıştır.

Veri hazırlama işlemleri sonucunda elde edilen 170 x 54 boyutundaki nihai veri kümesinde yer alan özellikler ve bunlara ait bilgiler Çizelge 5.5'te yer almaktadır.

Çizelge 5.5 Veri özellikleri ve bunlara ilişkin bilgiler

Özellikler	Kısaltma	İçerik - Formül	Birimi
Hasta ID	ID	Nümerik (1-170 arası rakam)	-
Cinsiyet	SEX	Nominal (0=Erkek, 1= Kadın)	-
Yaş	AGE	Nümerik	yıl
Tedavi Başlangıç Yaşı	SAGE	Nümerik	yıl
Ağırlık	HEI	Nümerik	kg
Boy	WEI	Nümerik	cm
Vücut Kitle İndeksi	VKI	Nümerik (VKI = Diyaliz sonrası kuru ağırlık [kg] / Boy karesi [m ²])	kg/m ²
Başı. Tansiyon (Sistol)	SS_AKB	Nümerik	mmHg
Başı. Tansiyon (Diastol)	SD_AKB	Nümerik	mmHg
Beslenme Risk İndeksi	NRI	Nümerik (NRI=(1,489Xserum albümin, g/L) + [41.7X(gımdiki ağırlık/ideal ağırlık), İdeal Ağırlık = H-100-[(H-50)/4] erkekler için, H-100-[(H-150)/2.5] kadınlar için)	-
Toplam Seans Sayısı	SES	Nümerik	Adet
Tedavi Süresi	SEAG	Nümerik	Ay
Diyalizde Hayat Süresi	DHS	Nümerik	yıl
Haftalık Seans Süresi	HSS	Nümerik	saat
Damar yolu tipi	DYT	Nominal (0=LAVF, 1=RAVF, 2=L SVC, 3=RSVC, 4=LIJVC, 5=RIJVC, 6=LFVC, 7=RFVC)	-
Antikoagülasyon Dozu	AKG	Nümerik	(ünite/seans)
Pre-Diyaliz Ağırlık	Pre-W	Nümerik	kg
Pre-Diyaliz Systol	Pre-S	Nümerik	mmHg
Pre-Diyaliz Diastol	Pre-D	Nümerik	mmHg
Post-Diyaliz Ağırlık	Ps-W	Nümerik	kg
Post-Diyaliz Systol	Ps_S	Nümerik	mmHg
Post-Diyaliz Diastol	Ps-D	Nümerik	mmHg
Seans Başı-Sonu Kilo Farkı	ISW	Nümerik	kg
Kan Akım Hızı	KAH	Nümerik	Qb
Eritropoetin İlaç Tipi	EIT	Nominal (0=Eprex 1=NeoRecormon 2=Aranesp 3= kullanılmıyor)	-
Eritropoetin Dozajı	EPOD	Nümerik	(ünite/ay)
EPO Dozaj Verimliliği	EPOU	Nümerik (EPO Dozajı (ünite/ay) / Diyaliz sonrası kuru ağırlık [kg])	(ünite/ay) /kg
Eritropoetin Direnci	EPOR	Nümerik (EPO Dozajı/ortalama Hb)	-
Rezidiv İdrar	RZI	Nümerik	ml
Htc	Htc	Nümerik	%
CRP	CRP	Nümerik	mg/l
Na	Na	Nümerik	mEq/dl
K (G)	K_I	Nümerik	mEq/dl
K (Ç)	K_O	Nümerik	mEq/dl
K Farkı	K_F	Nümerik	mEq/dl
Glisemi	GLI	Nümerik	mg/dl
Üre-BUN (G)	BUN_I	Nümerik	mg/dl
Üre-BUN (Ç)	BUN_O	Nümerik	mg/dl
BUN Oranı (Ratio)	BUN_R	Nümerik	-
Kreatinin (G)	KRE_I	Nümerik	mg/dl
Kreatinin (Ç)	KRE_O	Nümerik	mg/dl
Ürik Asit	URA	Nümerik	mg/dl
URR (Üre Azalma Oranı)	URR	Nümerik (URR= [(Üre-BUN (G) - Üre-BUN (Ç)) / Üre-BUN (G)]*100)	-
Kt/V (Üre Temizleme Oranı)	Kt/V	Nümerik (Kt/V = (URR*0,023)-0,284)	-
Ca X P	CaXP	Nümerik	-
PTH	PTH	Nümerik	pg/ml
T.Kolesterol	TKoll	Nümerik	mg/dl
HDL	HDL	Nümerik	mg/dl
LDL	LDL	Nümerik	mg/dl
Trigliserid	TGL	Nümerik	mg/dl
Total Protein	TPRO	Nümerik	g/dl
Albumin	ALB	Nümerik	g/dl
Ferritin	FERT	Nümerik	mg/dl
TSI	TSI	Nümerik (TSI= (Fe / Fe B.K.)^100)	-
Risk Skoru	RSCORE	Nümerik	-

5.5. Modelin Kurulması

Uygulama sürecinde takip edilecek adımlar Şekil 5.2'de özetlenmiştir.



Şekil 5.2 Uygulama süreci adımları

Uygulama sürecinin ilk adımında anahtar özellik olarak diyaliz hastaları için ölüm riskini belirten bir skor seçilmiş, ölüm riski sınıflarına göre veri kümesinde yer alan özelliklerin etkileşimlerinin incelenmesi düşünülmüştür.

Uzman nefrolog görüşü doğrultusunda veri kümesinde ölüm riski ile doğrudan ilgili özelliklerin yer alması gereği anlaşılmış ve bunu sağlamak için hazırlanan ilk veri kümesinde yer alan Kan Grubu, Sigara Kullanımı, Alkol Kullanımı, Diyet Uygulaması, Transplantasyon Durumu, Hb (Hemoglobin), MCV, MCH, MCHC, Lökosit, Platelet, Alkalen Fosfataz, ALT (Alanin

ransaminaz), Fe (Demir), Fe B.K. (Demir Bağlama Kapasitesi) özellikleri veri kümesinden çıkarılmıştır.

Ölüm riski ile ilgisi olmayan özelliklerin ana veri kümesinden çıkarılması sonrasında uygulamada kullanılacak veri kümesi kısmi olarak Ek-3'te yer alan Çizelge 3.7'da gösterilmiştir.

Anahtar özelliğin ölüm riski ile ilgili olması risk sınıflarının aralıklarının ve sayısının belirlenmesini gerektirmektedir.

Sınıf aralıklarının belirlenmesi için veri kümesinde yer alan veri değişkenliği yüksek özelliklerin anahtar özellik ile etkileşimi dikkate alınacaktır.

Keyfi belirlenecek farklı sınıf sayıları için veri kümesi üzerinde yapılacak kümeleme algoritması uygulaması sonucu elde edilecek kümeleme performansı, mevcut veriler için en anlamlı risk sınıfı sayısını belirlemede kullanılacaktır.

Anahtar özelliğin sınıf sayısının belirlendikten sonra elde edilen yeni veri kümesinde anahtar özelliğin tahmininde en etkili özelliklerin tespit edilmesi işlemi gerçekleştirilecektir.

Ardından aynı veri kümesi üzerinde çeşitli karar ağacı algoritmaları çalıştırılacak ve en iyi performansa sahip algoritma tarafından oluşturulan karar kuralları değerlendirmeye alınacaktır.

Özetle belirlenen risk sınıflarına göre uygulanacak kümeleme algoritması yardımıyla en yüksek uygunluk oranında kümelenecek veri kümesinden karar kuralları elde edilmeye çalışılacaktır. Uygulama çalışmasının sonucu olan bu karar kuralları, veri kümesini oluşturan ve diyaliz hastaları için ölüm riski faktörlerinden olan özelliklerin etkileşimleri açısından incelenecektir.

5.5.1. Metot seçimi

Uygulama çalışmasında kullanılacak algoritmaların seçimi yapılırken algoritmanın açık kaynak koduna sahip olması yani üzerinde değişiklik yapılabilmesi ve buna imkan tanıyan yasal sorunların olmaması ve sık kullanılan ve performansı bilinen bir algoritma olması kriterleri dikkate alınarak;

- o Kümeleme analizi için alanında en sık kullanılan ve uygun çözümler üreten K-means algoritması,
- o Anahtar özelliklerle ilgili tahminlerde en iyi sonuç veren özelliklerin tespit edilmesi amacıyla Weka'da yer alan özellik değerlendirici olarak CfsSubsetEval algoritması ile Best first algoritması,
- o Karar ağacı oluşturmak için bu konuda en gelişmiş algoritma olan ve etkin çözüm veren J4.8 ile bu algoritmanın karar ağacı oluşturma performansının kıyaslamasını yapmak için RandomTree algoritması ve PART algoritması kullanılmıştır.

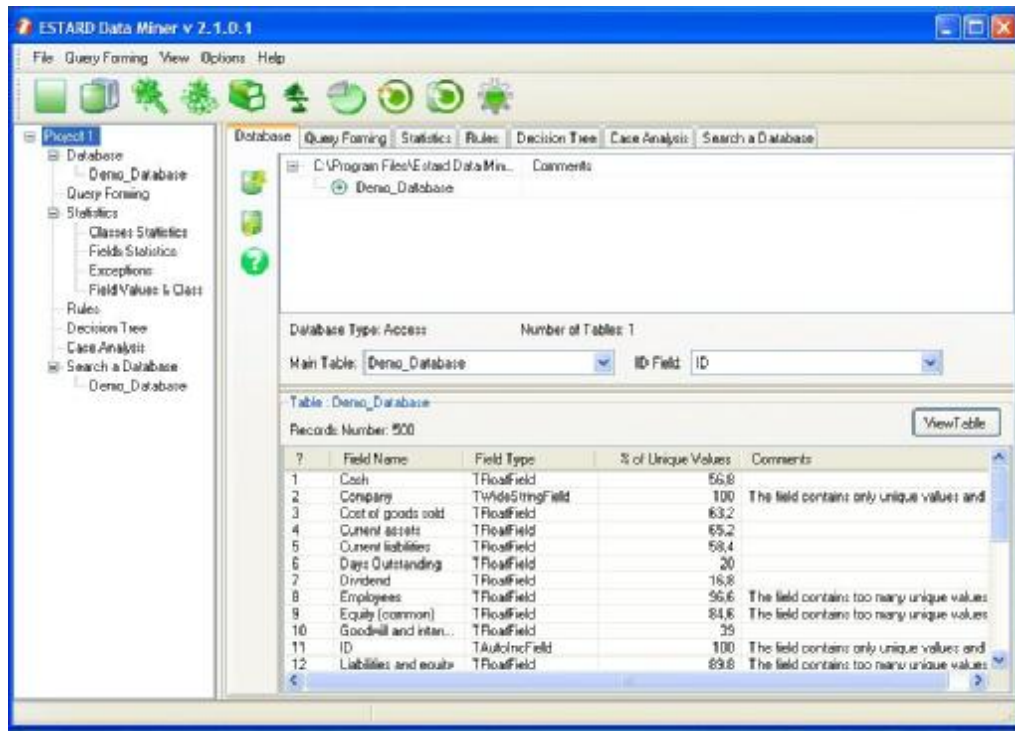
5.5.2. Yazılım seçimi

Uygulamada yazılım olarak anahtar özellik için sınıf aralıklarının belirlenmesi adımı ESTARD Data Miner Version 2.1.0.1, kümeleme ve karar ağacı algoritmaları kullanımı ile anahtar özelliklerle ilgili tahminlerde en iyi sonuç veren özelliklerin belirlenmesi adımlarında WEKA Version 3.7.0 Veri Madenciliği yazılımı kullanılmıştır.

ESTARD Data Miner Version 2.1.0.1

Estard Data Miner karar kuralları oluşturma konusunda gelişmiş performansa sahip bir yazılımdır. Tahmin edici analiz adlı karar modeli oluşturan bir iş zekası metodu kullanır. Böylece sadece istatistik teknikler kullanımı sunan pek çok benzer yazılımdan daha iyi karar kuralları oluşturur. Şekil 5.3'te ESTARD Data Miner Version 2.1.0.1 ekran görüntüsü görülmektedir.

ESTARD Data Miner yazılımının 17 Haziran 2009'da internetten ücretsiz kullanıma açılan 2.1.0.1 Demo versiyonu bazı kullanım kısıtları içermektedir. Özellikle karar kuralı oluşturma algoritmalarında üretilen karar kuralı sayısı 15 ile sınırlıdır. Bu durum 15'ten fazla sayıda karar kuralı üretilen uygulamalarda sorun oluşturmaktadır. Bu nedenle çalışma kapsamında karar ağacı oluşturma aşamasında ESTARD Data Miner yazılımı kullanılmayacaktır. Bu yazılım yerine açık kodlu, kullanım kısıtı olmayan Weka yazılımının en son sürümü Weka Versiyon 3.7.0 tercih edilecektir.



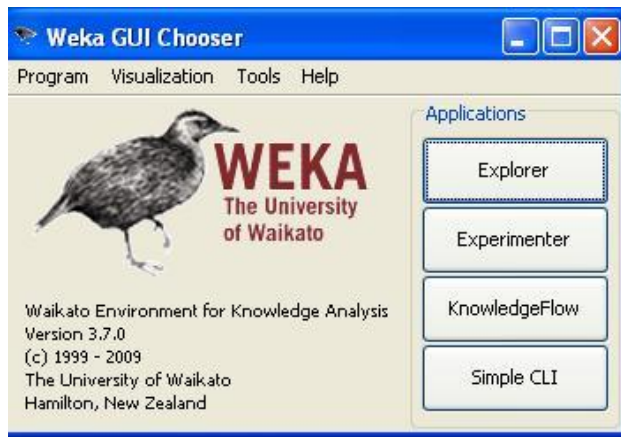
Şekil 5.3 ESTARD Data Miner ekran görünümü

ESTARD Data Miner Version 2.1.0.1, veri kümesi yazılıma yüklendikten sonra veri kümesinde yer alan özelliklerin farklı veri içerme yüzdesini hesaplamakta ve veri değişkenliği düşük veri özelliklerini seçenek dışı tutarak sınıflandırma ve karar kuralları oluşturma algoritmalarında bu verilerin kullanımını engellemektedir. Yazılımın bu özelliğinden yararlanarak RSCORE kısaltmalı risk skoru özelliği ile seçilebilen yüksek veri değişkenlikli özellikleri

arasındaki etkileşim dikkate alınarak 2, 3, 4 ve 5 farklı risk sınıfı için verilerin sınıf bazlı dağılım yüzdeleri ve sınıf aralıkları belirlenebilecektir.

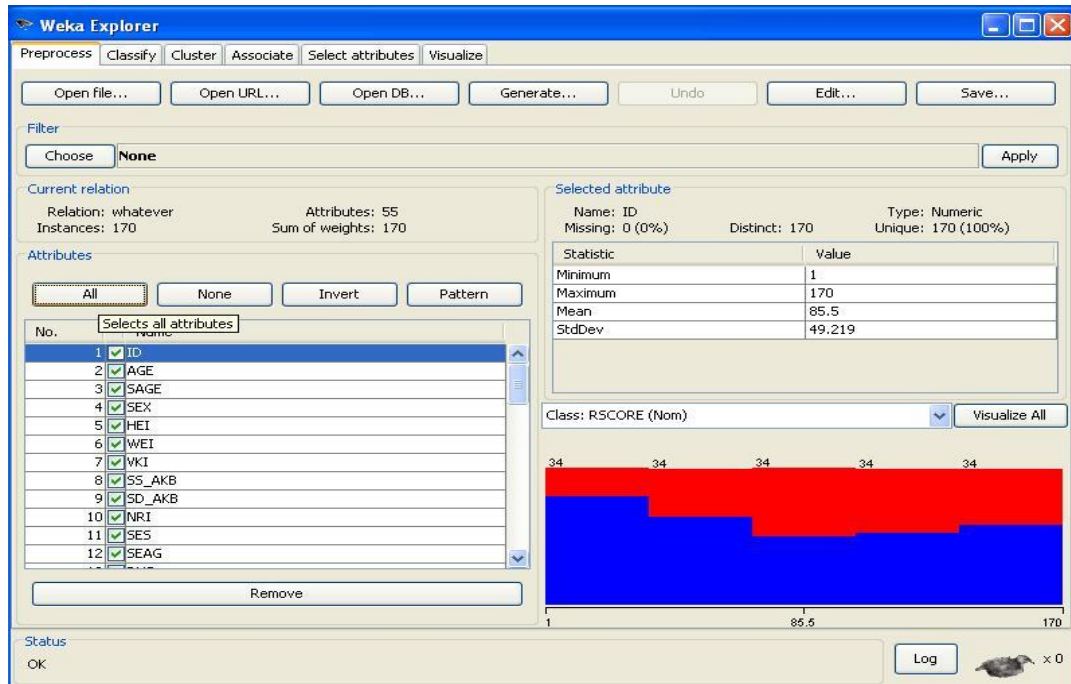
WEKA Version 3.7.0

Waikato üniversitesinde açık kaynak kodlu olarak JAVA dili üzerinde geliştirilmiştir ve GPL lisansı ile dağıtılmaktadır. Yazılımın adı Waikato Environment for Knowledge Analysis kelimelerinin baş harflerinden oluşmaktadır. Şekil 5.4 ve Şekil 5.5'te WEKA Version 3.7.0 ekran görüntüleri verilmiştir.



Şekil 5.4 WEKA Kullanıcı Ara yüzü

WEKA verileri basit bir dosyadan okur ve veriler üzerindeki stokastik değişkenlerin sayısal veya nominal değerler olduğunu kabul eder. Uygulama çalışması için hazırladığımız veri kümesinde yer alan bazı özellikler nominal olarak kodlandıktan sonra kullanılabilir durumdadır. WEKA yazılımına nominal olarak yüklenecek bu özelliklere ilişkin etiketlemeler Çizelge 5.6'da verilmiştir.



Şekil 5.5 WEKA Explorer bölümünün ekran görünümü

WEKA verileri basit bir dosyadan okur ve veriler üzerindeki stokastik değişkenlerin sayısal veya nominal değerler olduğunu kabul eder. Uygulama çalışması için hazırladığımız veri kümesinde yer alan bazı özellikler nominal olarak kodlandıktan sonra kullanılabilecek durumdadır. WEKA yazılımına nominal olarak yüklenecek bu özelliklere ilişkin etiketlemeler Çizelge 5.6'da verilmiştir.

Çizelge 5.6 Nominal hale getirilen özelliklere ilişkin etiketleme

Özellikler	
Etiketleme Öncesi	Etiketleme Sonrası
Cinsiyet (Erkek-E, Kadın-K)	Cinsiyet (1,0)
Damar Yolu Tipi (LAVF,RAVF, LSVC,RSVC,LIJVC,RIJVC, LFVC, RFVC)	Damar Yolu Tipi (0, 1,2,3,4,5,6,7)
EPO ilaç Tipi (Eprex, Neorecormon, Aranesp,Kullanılmıyor)	EPO ilaç Tipi (0,1,2,3)

WEKA yazılımı verilerin bir dosya verisi şeklinde olması durumunda veritabanı (database) üzerinden de veri çekebilir. Dosya formatı olarak .arff

(Attribute Relation File Format) formatını kullanır. Bu nedenle hazırlanan veri kümelerinin WEKA yazılımına yüklenmeden önce .arff dosya formatına dönüştürülmesi gereklidir.

Çalışmada kullanılan veri kümesi ilk olarak Microsoft 2007 Excel yazılımında hazırlanmış, WEKA yazılımına yüklemek için önce Microsoft 2007 Excel yazılımında Farklı Kaydet seçeneği altında yer alan .csv dosya formatında kaydedilmiş, ardından .csv formatından .arff formatına Ljubljana Üniversitesi'nden Marko Tkalčič tarafından bu amaçla hazırlanan internet sayfası kullanılarak [148] dönüşümü sağlanmıştır.

Arff formatı @relation, @attribute ve @data deyimleri aracılığıyla dosyanın yapısını belirler. @relation veri yığının genel amacını ya da ismini belirtir. @attribute verideki veri tabanında sütunlara karşılık gelen özellik isimlerini belirtmek için kullanılır. @data ham verilerin başladığı satıra işaret eder. Veri kümesinin WEKA yazılımına yüklenirken .arff dosyası formatındaki görünümü temsili olarak Şekil 5.6'da gösterilmiştir.

WEKA üzerinde makine öğrenmesi ve istatistik ile ilgili pek çok kütüphane hazır olarak gelmektedir. Örneğin veri ön işleme (data preprocessing), regresyon, sınıflandırma (classification), kümeleme (clustering), özellik seçimi veya özellik çıkarımı (feature extraction) bunlardan bazılarıdır. Ayrıca bu işlemler sonucunda çıkan neticelerinde görsel olarak gösterilmesini sağlayan görüntüleme (visualization) araçları bulunmaktadır.

```

@relation whatever

@attribute ID numeric
@attribute AGE numeric
@attribute SAGE numeric
@attribute SEX {1,0}
@attribute HEI numeric
@attribute WEI numeric
@attribute VKI numeric
@attribute SS_AKB numeric
@attribute SD_AKB numeric
@attribute NRI numeric
@attribute SES numeric
@attribute SEAG numeric
@attribute DHS numeric
@attribute HSS numeric
@attribute DYT {0,2,1,5,7,3,4,6}

.....

@attribute FERT numeric
@attribute TSI numeric
@attribute RSCORE {0,1}

@data
1,72,66,1,160,45.6,20.9,130,70,44.5,229,71,5.9,12,0,3500,55.1,145,69....
2,28,26,1,182,71.2,21.3,140,80,46,224,20,1.7,12,0,5933,72.8,131,82....
3,46,39,1,173,64.5,22.4,180,100,47.6,884,111,9.3,12,0,5000,69.7,130,81....

.....

169,78,73,0,160,64,24.9,70,50,53.2,763,59,4.9,12,0,2500,64.3,83,60....
170,71,67,0,158,57.5,23.3,150,110,50.7,722,53,4.4,12,3,3500,59.9,124,73.....

```

Şekil 5.6 Veri kümesinin .arff dosya formatında temsili görünümü

Bu çalışmada açık kodlu oluşu, çok miktarda ve çeşitli algoritmanın kullanımına imkan tanınması, kullanım kısıtı olmaması gibi nedenlerle WEKA yazılımının kullanımı tercih edilmiştir. Yazılımdan kümeleme analizi, anahtar özellikle ilgili tahminlerde en iyi sonuç veren özelliklerin tespiti ve karar ağacı / karar kuralları oluşturma amaçlı çeşitli algoritmalar kullanılmıştır.

5.6. Uygulama ve Elde Edilen Sonuçlar

5.6.1. Risk skorunun hesaplanması

Ölüm riskini gösteren bir skor hesaplamada kullanılabilecek bir risk skorum sistemi bulmak amacıyla yapılan literatür taraması sonucunda Nefrolojide APACHE II (Acute Physiology and Chronic Health Evaluation), APACHE III, Liano, PINDEX, SAPS 2, SAPS 3, SOFA, MODS vb. Risk skorum sistemlerinin kullanıldığı görülmüştür [149].

Ancak bu sistemlerden APACHE II ve APACHE III gibi bazı skorlar sadece yoğun bakım ünitesinde yatan diyaliz hastaları için kullanılırken Liano, PINDEX gibi bazı skorların ise hazırlanan veri kümesinde yer almayan bazı parametreleri (ciğerlerde belli bir sürede temizlenen karbondioksit miktarı gibi) kullandığı anlaşılmıştır. Bu durumda uzman görüşü doğrultusunda Nefrolojide kullanılan tıbbi skorum sistemlerinin mevcut veri kümesinin ölüm riski skorumasında kullanılamayacağı sonucuna varılmıştır.

Devam eden araştırmalar neticesinde sağlık sigortacılığı alanında hemodiyaliz hastaları için özel olarak kullanılan ve mevcut veri kümesi ile hesaplama yapmaya uygun bir risk skorum sisteminin olduğu görülmüştür.

Veri madenciliği uygulaması gerçekleştireceğimiz diyaliz hastalarına ait veri kümesinin risk değerlendirmesinde CMS HCC (the Centers for Medicare and Medicaid Services -- Hierarchical Coexisting Conditions) modelinin Health Economics Research Inc. tarafından diyaliz hastaları için özel olarak geliştirilmiş bir versiyonu olan CMS HCC ESRD Risk Değerlendirme Modeli kullanılacaktır [150].

Amerika'da 1997'de Kongre tarafından alınan bütçe kararları gereği, sağlık sigortası hizmeti veren kuruluşlar 2000 yılından itibaren kişinin ödeme periyodu öncesi 6-18 ay boyunca aldığı sağlık hizmetlerini 10 grup altında

ücretlendiren bir değerlendirme sistemi olan PIP-DCGs (Principal Inpatient Diagnostic Cost Groups) kullanmaya başladılar. Bu sistemin yerini Ocak 2004'te kişinin demografik ve klinik durumuna göre risk skorlamasına tabi tutan CMS HCC (the Centers for Medicare and Medicaid Services -- Hierarchical Coexisting Conditions) Modeli almıştır [151].

CMS HCC Modeli'nde risk skorlamada kullanılacak katsayılar, risk değerlendirme uzmanları tarafından kişiye ait demografik (yaş,cinsiyet) ve klinik verileri (sağlık sorunlarının) sağlık masraflarındaki arttırıcı etkisi dolayısıyla ölüm riski düzeyi dikkate alınarak hesaplanmıştır. Bu katsayılar dönemsel olarak gözden geçirilmekte ve yıllık olarak masraflardaki değişime bağlı olarak gerektiğinde güncellenmektedir.

CMS HCC Modeli risk değerlendirmede klinik sınıflama sistemi olarak Uluslararası Hastalık Sınıflandırması'nın 9. Revizyonu'nda (International Classification of Diseases, 9th Revision, ICD-9) yer alan Klinik Durum kodlarını kullanmaktadır. Modelin klinik kodların daha detaylı alt başlıklar içeren hali olan ICD-10 sınıflama sistemine geçiş çalışmaları devam etmekte ve ICD-10 sınıflamasının risk kodlarının hesaplanması ile birlikte 2015 veya 2016'da geçişin tamamlanması planlanmaktadır [151].

CMS HCC Modeli'nin farklı hastalıklar için o hastalığa özel bir model olarak geliştirilmesi fikrinden hareketle 2005 yılında CMS HCC ESRD Modeli geliştirilmiştir. Tez kapsamında kullanılacak CMS HCC ESRD Modeli'nde klinik durum kodları listesi Kronik Böbrek Yetmezliği hastalığı ile ilgisi olan ve sağlık maliyetlerini en yüksek düzeyde artıran dolayısıyla yüksek düzeyde ölüm riski içeren hastalıklardan oluşturulmuştur [152].

Hemodiyaliz hastaları için özel olarak kullanılan Risk skorlama sisteminde tipik bir hemodiyaliz hastası için katsayı 0'dır. Bu da modeldeki hastalık katsayılarının, bu hastalıkların diyaliz hastası bir kişide görülmesi durumuna göre hesaplandığı anlamına gelmektedir.

Tıbbi risk skorumlama sistemlerinde herhangi bir hasta için tekil parametrelerden yararlanarak etkileşimli bir hesaplama sözkonusu iken CMS HCC ESRD modeli diyaliz hastaları arasında komorbid durumlar açısından bir kıyaslama yapmaya imkan vermektedir. CMS HCC ESRD modelinin bir diğer önemli özelliği de komorbid durumlar arasındaki ikili hatta üçlü etkileşimleri de risk skoruna yansıtabilmesidir. CMS HCC Diyaliz Modeli'nde 2007 yılı için belirlenen risk skoru katsayı listesi Demografik ve Klinik katsayılar olmak üzere Çizelge 5.7 ve 5.8'de verilmiştir.

Çizelge 5.7 CMS HCC ESRD Modeli Demografik Katsayılar

Yaş Grupları	Katsayı	
	Kadın	Erkek
0-34 Yaş	0,699	0,614
35-44 Yaş	0,699	0,650
45-54 Yaş	0,715	0,675
55-59 Yaş	0,746	0,699
60-64 Yaş	0,749	0,722
65-69 Yaş	0,813	0,776
70-74 Yaş	0,813	0,776
75-79 Yaş	0,831	0,790
80-84 Yaş	0,850	0,790
85 ve Üstü Yaş	0,872	0,826

Çizelge 5.8 CMS HCC ESRD Modeli Klinik Katsayılar [153]

HCC Kodu	Hastalık Tanımı	Katsayı
HCC1	HIV / AIDS	0,235
HCC2	Septisemi / Şok	0,073
HCC5	Fırsatçı Enfeksiyonlar	0,051
HCC7	Metastatik Kanser ve Akut Lösemi	0,189
HCC8	Akciğer, Üst Sindirim alanı ve Diğer Ciddi Kanserler	0,189
HCC9	Lenf, Baş ve Boyun, Beyin ve Diğer Ciddi Kanserler	0,160
HCC10	Göğüs, Prostat, Kalın Bağırsak ve Diğer Kanserler ve Tümörler	0,058
HCC15	Böbrek veya Çevresel Dolaşım Bozukluklu Diyabet	0,080
HCC16	Nörolojik veya Diğer Bozukluklu Diyabet	0,080
HCC17	Akute Komplikasyonlu Diyabet	0,080
HCC18	Oftalmolojik veya Belirsiz Bozukluklu Diyabet	0,080
HCC19	Komplikasyonsuz Diyabet	0,079
HCC21	Protein-Kalori Eksikliği	0,050
HCC25	Son Aşama Karaciğer Hastalığı	0,259
HCC26	Karaciğer Sirozu	0,095
HCC27	Kronik Hepatit	0,051
HCC31	Bağırsak Tıkanıklıkları / Delikleri	0,057
HCC32	Pankreas Kanseri	0,084
HCC33	Ateşli Bowel Hastalığı	0,088
HCC37	Kemik/Eklemler/ Kas Enfeksiyonları / Nekroz	0,115
HCC38	Hastalık	0,077
HCC44	Ciddi Hematolojik Hastalıklar	0,000
HCC45	Bağışıklık Sistemi Hastalıkları	0,113
HCC51	İlaç / Alkol Psikozu	0,000
HCC52	İlaç / Alkol Bağımlılığı	0,000
HCC54	Şizofreni	0,179
HCC55	Ana Depresif, Bipolar ve Paranoid Hastalıklar	0,123
HCC67	Boyundan Aşağısı Felç (Quadripleji), Diğer Kapsamlı Felçler	0,229
HCC68	Omurilik Hasarları (Parapleji)	0,229
HCC69	Omurilik Hastalıkları / Yaralanmaları	0,148
HCC70	Kas Bozukluğu (MS)	0,000
HCC71	Çoklu Sinir Rahatsızlıkları	0,056
HCC72	Multiple Skleroz	0,087
HCC73	Parkinson ve Huntington Hastalıkları	0,038
HCC74	Nöbetli Bozukluklar ve Sarsıntılar	0,094
HCC75	Koma, Beyin Sarsıntısı, Beyinde Oksijensizlik Hasarı	0,201
HCC77	Solunuma Bağlılık / Trakeostomi Durumu	0,349
HCC78	Solunum Yetmezliği	0,156
HCC79	Kalp-Solunum Bozukluğu ve Şok	0,088
HCC80	Konjestif Kalp Yetmezliği	0,086
HCC81	Akut Miyokard Enfarktüsü	0,107
HCC82	Değişken Anjin ve Diğer Akut İskemik Kalp Hastalığı	0,107
HCC83	Anjina Pektoris / Eski Miyokard Enfarktüsü	0,027

Çizelge 5.8 (Devam) CMS HCC ESRD Modeli Klinik Katsayılar [153]

HCC Kodu	Hastalık Tanımı	Katsayı
HCC92	Belirli Kalp Aritmisi	0,061
HCC95	Beyin Felci	0,058
HCC96	İskemik veya Belirsiz Kalp Krizi	0,058
HCC100	İnme (Hemipleji) / Geçici Felç (Hemiparezi)	0,088
HCC101	Beyin Felci ve Diğer Felçli Sendromlar	0,040
HCC104	Damar Hastalığı Komplikasyonlu	0,169
HCC105	Damar Hastalığı	0,059
HCC107	Kistik Fibroz	0,078
HCC108	KOAH (Kronik Obstrüktif Pulmonar Hastalık)	0,078
HCC111	Aspirasyon ve Belirli Bakteriyel Zatürre	0,123
HCC112	Pnömonok Zatürresi, Amfizem, Akciğer Apsesi	0,051
HCC119	Artımlı Diyabetik Retinopati ve Vitreus Hemoraji	0,000
HCC130	Diyaliz Durumu	0,000
HCC131	Böbrek Yetmezliği	0,000
HCC132	Nefrit	0,000
HCC148	Yatak Yarası (Decubitus Cilt Ülseri)	0,182
HCC149	Kronik Cilt Ülseri, Decubitus harici	0,110
HCC150	Kapsamlı Üçüncü Derece Yanıklar	0,088
HCC154	Ciddi Kafa Travması	0,201
HCC155	Önemli Kafa Travması	0,022
HCC157	Omurilik hasarsız Omur kırıkları	0,035
HCC158	Bel Kırığı / Disk Kayması	0,054
HCC161	Travmatik Amputasyon	0,073
HCC164	Tıbbi Bakım Gerektiren Büyük Travma ve Komplikasyonlar	0,000
HCC174	Büyük Organ Nakli Durumu	0,199
HCC176	Sindirim veya Boşaltım için Kullanılan Vücut Açıklığı	0,062
HCC177	Amputasyon Durumu, Düşük Kol / Amputasyon Komplikasyonlar	0,073
Hastalık Etkileşimleri		
INT1	DM_CHF (Diabet + Kronik Kalp Yetmezliği)	0.020
INT2	DM_CVD (Diabet + Kronik Damar Hastalığı)	0.051
INT3	CHF_COPD (Kronik Kalp Yetmezliği + KOAH)	0.000
INT4	COPD_CVD_CAD (KOAH + Kalp-Damar Hastalığı + Koroner Damar Hastalığı)	0.000

Risk skoru, hastaların yaş, cinsiyet gibi demografik verileri ile kronik böbrek yetmezliği etiyolojisi dışındaki comorbid hastalıkları ve etkileşimli comorbid durumlar için belirlenen katsayılar kullanılarak dört farklı bileşenin toplamından oluşur.

CMS HCC ESRD modeli örnek risk skoru hesabı aşağıda gösterilmiştir [153]:

Erkek, 81 yaşında: Demografik skor	-0.790
HCC 18 - Göz kusuruna yol açmış Diyabet	-0.080
HCC 80 - Konjestif Kalp Yetmezliği	-0.086
HCC 108 – Kronik Obstrüktif Pulmonar Hastalık	-0.078
Toplam Risk Skoru =	1.034

Veri kümesindeki 170 hastanın hastalık durumları ve bu durumlar için CMS HCC ESRD modeline göre kullanılacak katsayıları Çizelge 5.9'da gösterilmiştir.

Veri kümesindeki 170 hasta için bu yöntemle yapılan Risk Skoru hesaplaması ve sonuçları Ek-4'teki Çizelge 4.1'de yer almaktadır.

Risk skoru hesaplamada kullanılan hastalara ait hastalık etiyolojisi ve komorbid durum bilgileri, sınıflandırma zorlukları nedeniyle hesaplamaların ardından veri kümesinden çıkarılmıştır.

Çizelge 5.9 Veri kümesindeki hastaların hastalık durumlarına göre kullanılan CMS HCC ESRD modeli katsayıları

Hastalık Adı	HCC Kodu	Hastalık Tanımı	Katsayı
Polikistik Böbrek Hastalığı	HCC131	Böbrek Yetmezliği	0,000
Renal Celsea Kanseri			
Hipertansif Nefropati			
Poststreptokoksik GN			
Nefrolitiazis			
Diffüz crescent			
Obst. Nefropati			
Amiloidozis			
Granülo parotit			
Amiloidoz nedeni FMF (Ailevi Akdeniz Ateşi)			
Renal Amiloidozis			
Ürolitiazis			
Membranepiloriferative GN			
Bilateral Amarozis			
Berger Hastalığı			
Kronik Kalp Yetmezliği	HCC80	Konjestif Kalp Yetmezliği	0,086
Bypass			
ASKH (aterosklerotik kalp hastalığı)			
PANCARD			
Hiperlipidemi	HCC44	Ciddi Hematolojik Hastalıklar	0,000
Hiperproteinemi			
Hiperkolesterol			
Hipotiroid			
Anemi			
Burger Hastalığı	HCC104	Damar Hastalığı Komplikeşyonlu	0,169
Wegener Granül.			
TAH+BSO (total abdominal histerektomi + bilateral salpingooferektomi Rahim,her iki yumurtalık ve tüplerin alınması)	HCC10	Göğüs, Prostat, Kalın Bağırsak ve Diğer Kanseler ve Tümörler	0,058
BPH (Benign Prostat Hiperplazi)			
Prostat			
Osteoporoz	HCC38	Hastalık	0,077
Hipertansiyon			
Klenfositik Lösemi	HCC9	Lenf, Kafa ve Boyun, Beyin ve Diğer Ciddi Kanseler	0,160
Diabetik Nefropati	HCC15	Böbrek veya Çevresel Dolaşım Bozukluklu Diyabet	0,080
Diabetik Nöropati	HCC16	Nörolojik veya Diğer Bozukluklu Diyabet	0,080
Diabet Tip II	HCC19	Komplikasyonsuz Diyabet	0,079
Diabetes Mellitus			
PKH	HCC25	Son Aşama Karaciğer Hastalığı	0,259
Gastroduodenit	HCC31	Bağırsak Tıkanıklıkları / Delikleri	0,057
Antifosfolipid Send.	HCC45	Bağışıklık Sistemi Hastalıkları	0,113
SJögren			
Astım Bronşit	HCC78	Solunum Yetmezliği	0,156
AC Enfarktüs	HCC81	Akut Miyokard Enfarktüsü	0,107
Irregular Tansiyon	HCC92	Belirli Kalp Aritmisi	0,061
İskemik	HCC96	İskemik veya Belirsiz Kalp Krizi	0,058
KAH	HCC108	KOAH (Kronik Obstrüktif Pulmonar Hastalık)	0,078
Diabetik Retinopati	HCC119	Artımlı Diabetik Retinopati ve Vitreus Hemoraji	0,000
Nefrit	HCC132	Nefrit	0,000
Ampute	HCC177	Amputasyon Durumu, Düşük Kol / Amputasyon Komplikeşyonları	0,073

5.6.2. Anahtar özellik için sınıf sayısının tespiti

CMS HCC ESRD modelinde hesaplanan risk skoru, hastanın dahil olduğu sigorta modeli için belirlenen yıllık ortalama sağlık masrafı miktarı ile çarpılarak o hasta için yapılacak sağlık sigortasının değerini tespit etmede kullanılmaktadır. Bu nedenle model içinde risk skorları için ayrı bir sınıflama yapılmamıştır.

Veri kümesi için temel değerlendirme kriteri olarak CMS HCC ESRD modeline göre hesaplanacak Risk skoru kullanılacağından bu özelliğin kendi içinde en iyi gruplama performansı veren sınıflamasını tespit etmek gerekmektedir.

Bu sınıf sayısını belirlemek üzere numerik risk skoru içeren veri kümesi ESTARD Data Miner Version 2.1.0.1 yazılımına yüklenmiş ve tüm özellikler veri değişkenliği oranları Şekil 5.7 ve 5.8’de gösterildiği şekilde belirlenmiştir.

Veri değişkenlik oranı yüksek veriler birbirinden farklı değerler içerdiğinden sınıflama işlemi performansının düşmesine neden olurlar. ESTARD yazılımı sınıflama performansını artırmak için veri değişkenliği diğerlerine göre daha az oranda 24 özelliğin seçimine izin vermiştir.

ESTARD Data Miner v 2.1.0.1

File Query Forming View Options Help

Database Query Forming Statistics Rules Decision Tree Case Analysis Search a Database

Table: vigilius
Records Number: 170

?	Field Name	Field Type	% of Unique Values	Comments
1	AGE	TFloatField	33,5294117647059	
2	AKG	TFloatField	18,2352941176471	
3	ALB	TFloatField	76,4705882352941	The field contains too many unique values and is not recommended for use.
4	BUN_I	TFloatField	95,2941176470588	The field contains too many unique values and is not recommended for use.
5	BUN_Q	TFloatField	90	The field contains too many unique values and is not recommended for use.
6	BUN_R	TFloatField	99,4117647058823	The field contains too many unique values and is not recommended for use.
7	CaKp	TFloatField	100	The field contains only unique values and can only be used as an ID field.
8	CRP	TFloatField	68,8235294117647	
9	DHS	TFloatField	51,7647058823529	
10	DYT	TFloatField	5,29411764705882	
11	EIT	TWideStringField	2,35294117647058	
12	EPOD	TFloatField	79,4117647058823	The field contains too many unique values and is not recommended for use.
13	EPOIR	TFloatField	95,2941176470588	The field contains too many unique values and is not recommended for use.
14	EPOIU	TFloatField	95,2941176470588	The field contains too many unique values and is not recommended for use.
15	FEPT	TFloatField	97,6470588235294	The field contains too many unique values and is not recommended for use.
16	GLI	TFloatField	92,9411764705882	The field contains too many unique values and is not recommended for use.
17	HDL	TFloatField	35,2941176470588	
18	HEI	TFloatField	13,4117647058824	
19	HSS	TFloatField	5,29411764705882	
20	Htc	TFloatField	96,8235294117647	The field contains too many unique values and is not recommended for use.
21	ID	TFloatField	100	The field contains only unique values and can only be used as an ID field.
22	ISW	TFloatField	95,2941176470588	The field contains too many unique values and is not recommended for use.
23	K_F	TFloatField	90	The field contains too many unique values and is not recommended for use.
24	K_I	TFloatField	81,1764705882353	The field contains too many unique values and is not recommended for use.
25	K_O	TFloatField	82,3529411764706	The field contains too many unique values and is not recommended for use.
26	KAH	TFloatField	12,9411764705882	
27	KRE_I	TFloatField	94,1176470588235	The field contains too many unique values and is not recommended for use.

Şekil 5.7 ESTARD Data Miner veri değişkenliği oranları ekran görüntüsü I

ESTARD Data Miner v 2.1.0.1

File Query Forming View Options Help

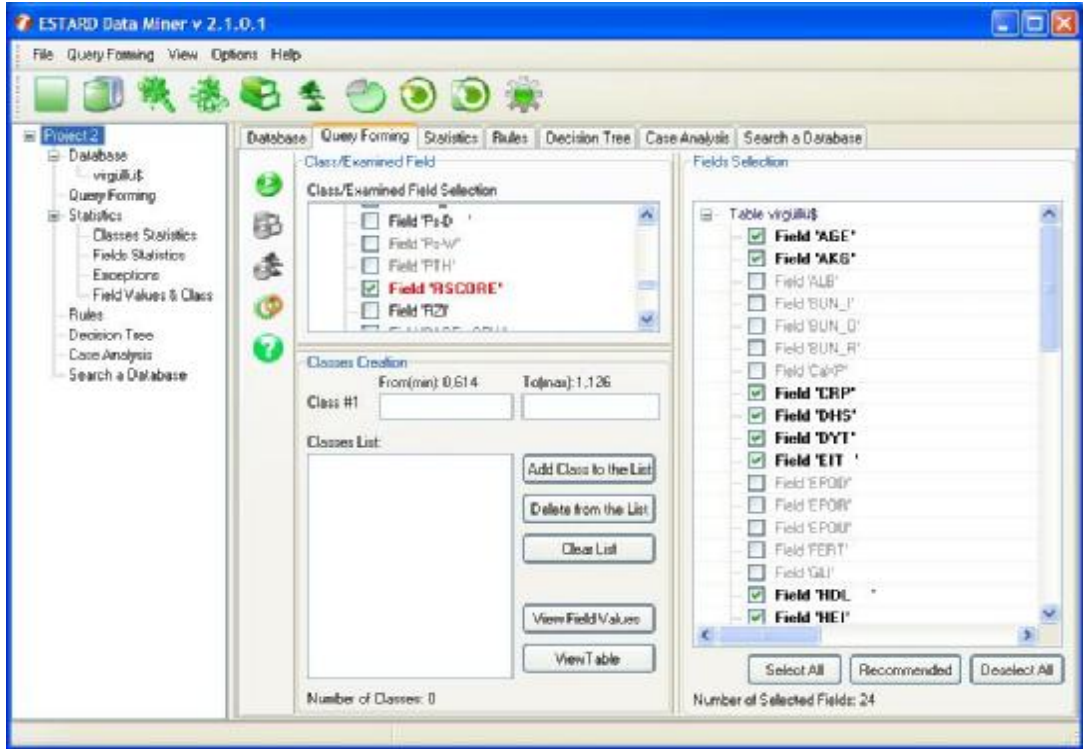
Database Query Forming Statistics Rules Decision Tree Case Analysis Search a Database

Table: vigilius
Records Number: 170

?	Field Name	Field Type	% of Unique Values	Comments
28	KRE_O	TFloatField	92,3529411764706	The field contains too many unique values and is not recommended for use.
29	KV	TFloatField	100	The field contains only unique values and can only be used as an ID field.
30	LDL	TFloatField	68,8235294117647	
31	Na	TFloatField	62,9411764705882	
32	NRH	TFloatField	100	The field contains only unique values and can only be used as an ID field.
33	Pse-O	TFloatField	31,7647058823529	
34	Pse-S	TFloatField	95,2941176470588	
35	Pse-W	TFloatField	100	The field contains only unique values and can only be used as an ID field.
36	Pse_5	TFloatField	53,5294117647059	
37	Pse-D	TFloatField	35,2941176470588	
38	Pse-W	TFloatField	100	The field contains only unique values and can only be used as an ID field.
39	PTH	TFloatField	97,0588235294118	The field contains too many unique values and is not recommended for use.
40	RSCORIE	TFloatField	46,4705882352941	
41	RZI	TFloatField	57,0588235294118	
42	SAGE SEX	TFloatField	34,1176470588235	
43	SD_AKB	TFloatField	4,70588235294118	
44	SEAG	TFloatField	51,7647058823529	
45	SES	TFloatField	93,5294117647059	The field contains too many unique values and is not recommended for use.
46	SEX	TFloatField	1,17647058823529	
47	SS_AKB	TFloatField	7,64705882352941	
48	TGL	TFloatField	92,3529411764706	The field contains too many unique values and is not recommended for use.
49	TKol	TFloatField	23,5294117647059	The field contains too many unique values and is not recommended for use.
50	TPRO	TFloatField	77,6470588235294	The field contains too many unique values and is not recommended for use.
51	TSI	TFloatField	100	The field contains only unique values and can only be used as an ID field.
52	URA	TFloatField	28,2352941176471	The field contains too many unique values and is not recommended for use.
53	URR	TFloatField	100	The field contains only unique values and can only be used as an ID field.
54	VKI	TFloatField	100	The field contains only unique values and can only be used as an ID field.
55	WEI	TFloatField	60	

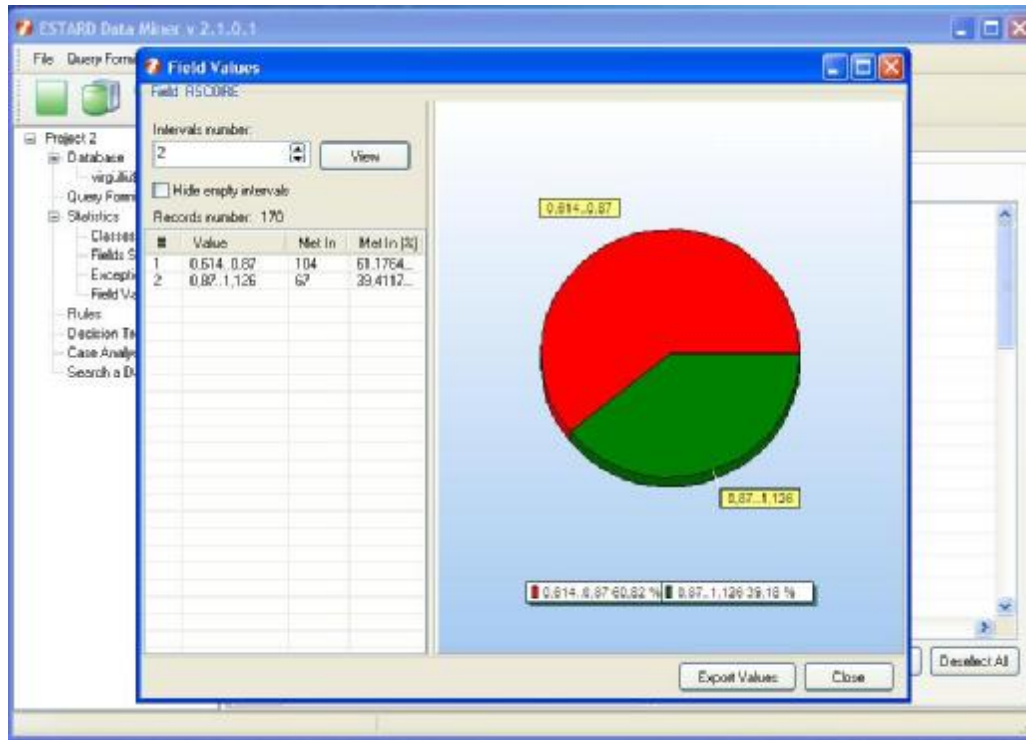
Şekil 5.8 ESTARD Data Miner veri değişkenliği oranları ekran görüntüsü II

Anahtar özellik olan risk skoru, yazılımın sınıflama işlemi için seçimine izin verdiği 24 adet özelliikle etkileşimine göre sırasıyla 2, 3, 4 ve 5 farklı sınıfına (risk seviyeleri) ayrılmıştır. Seçime ilişkin ekran görüntüsü Şekil 5.9'da verilmiştir.

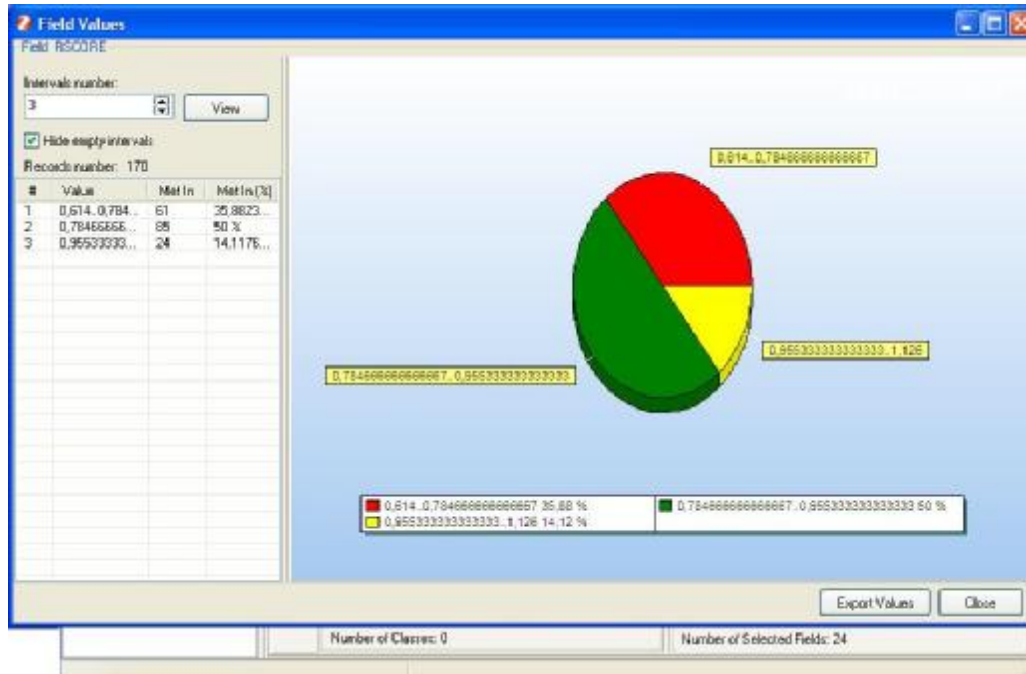


Şekil 5.9 ESTARD Data Miner sınıflama için özellik seçimi ekran görüntüsü

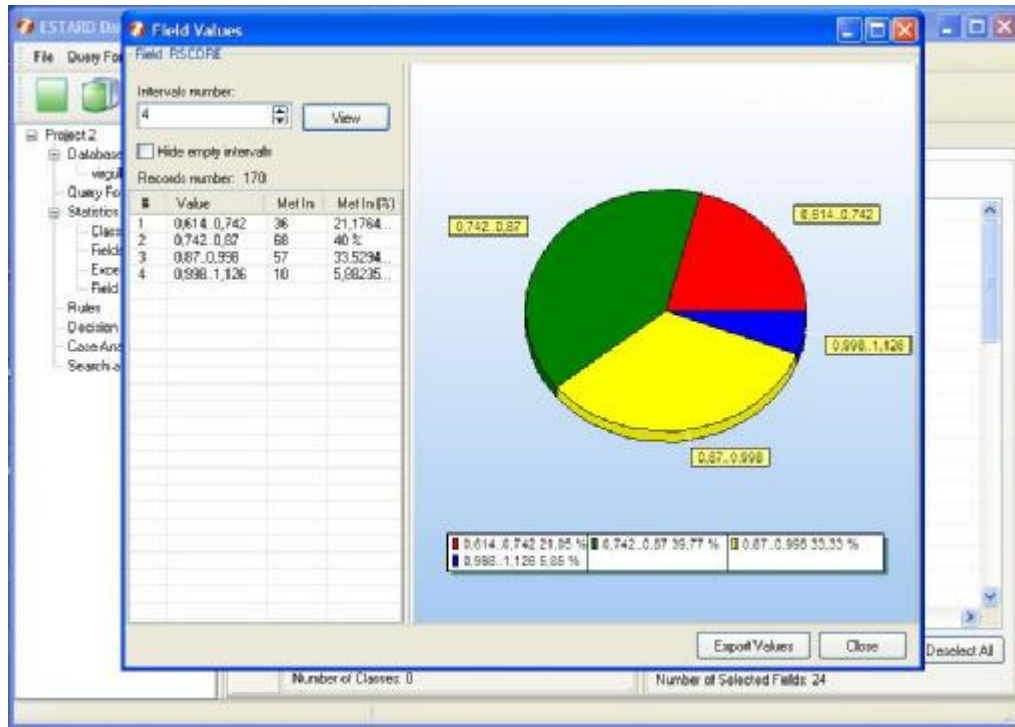
Risk skoru özelliğinin farklı sınıf sayıları (risk seviyeleri) için ESTARD Data Miner Version 2.1.0.1 yazılımı ile elde edilen dağılımlarını gösteren ekran görüntüleri Şekil 5.10-5.13'te yer almaktadır.



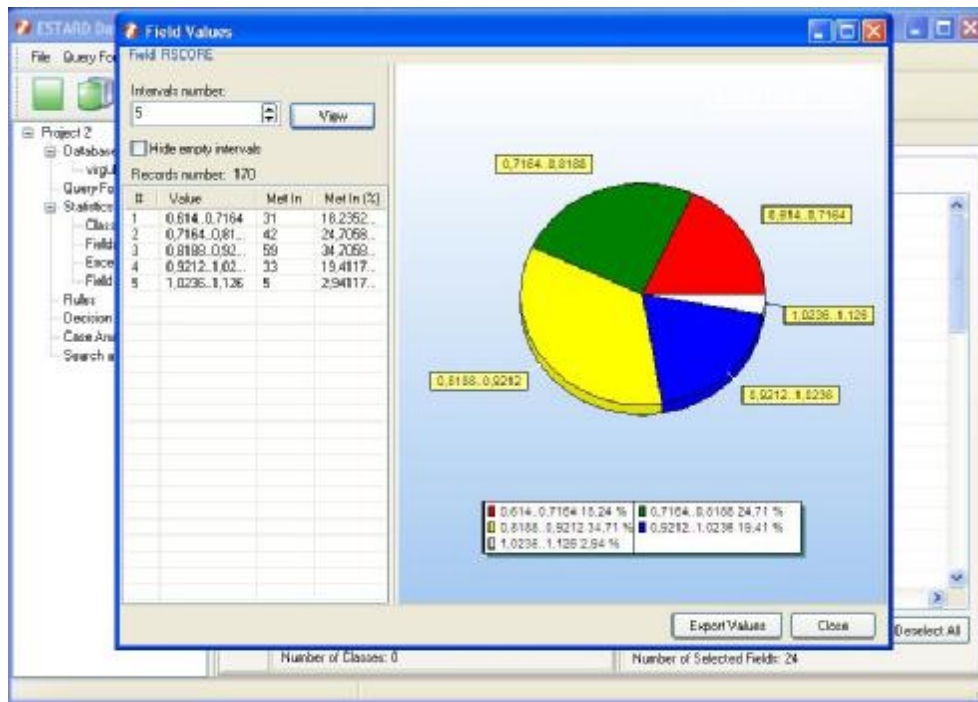
Şekil 5.10 ESTARD Data Miner RSCORE iki sınıf seviyesi için veri dağılımı



Şekil 5.11 ESTARD Data Miner RSCORE üç sınıf seviyesi için veri dağılımı



Şekil 5.12 ESTARD Data Miner RSCORE dört sınıf seviyesi için veri dağılım



Şekil 5.13 ESTARD Data Miner RSCORE beş sınıf seviyesi için veri dağılım

Her sınıf sayısı için ilgili risk seviyesine giren kayıtların nominal etiketlemesi ve sınıflara göre dağılım oranları Çizelge 5.10'da gösterilmiştir.

Çizelge 5.10 Her sınıf sayısı için ilgili risk seviyesine giren kayıtların nominal etiketlemesi ve sınıflara göre dağılım oranları

Risk Skoru	Sınıf Nominal Etiketi	Sınıf Aralığı	Sınıfa Giren Veri Sayısı	Yüzdesi
2 Sınıflı	0	0.614 - 0.870	104	61.17
	1	0.870 - 1.126	67	39.4
3 Sınıflı	0	0.614 - 0.786	61	35.8
	1	0.786 - 0.955	85	50.0
	2	0.955 - 1.126	24	14.2
4 Sınıflı	0	0.614 - 0.742	36	21.1
	1	0.742-0.870	68	40.0
	2	0.870 - 0.998	57	33.5
	3	0.998 - 1.126	10	5.4
5 Sınıflı	0	0.614 - 0.716	31	19.0
	1	0.716 - 0.810	42	25.0
	2	0.810 - 0.920	59	34.0
	3	0.920 - 1.020	33	19.1
	4	1.020 - 1.126	5	2.9

Buna göre iki sınıflı yapıda Risk skoru 0,624 ile 0,8 arasında kalan hastalar Risk Seviyesi 0 (RS0) grubuna, Risk skoru 0,8 ile 1,8 arasında kalan hastalar Risk Seviyesi I (RSI) grubuna girmektedir. Ölüm riski risk skoru arttıkça arttığından Risk Seviyesi 0 Risk Seviyesi I'e göre daha az risk taşımaktadır. Aynı risk seviyesi mantığı 3, 4 ve 5 sınıflı yapılarda da geçerlidir.

K-means kümeleme algoritması nominal bir anahtar özellik ile kullanılabildiğinden Risk skoru özelliği, Risk skoru sınıfı sayısı 2, 3, 4 ve 5 için ayrı ayrı nominal hale getirilmiş ve her biri veri kümesine ilave edilerek dört farklı veri kümesi elde edilmiştir. Risk skoru özelliğinin farklı sınıf sayıları için nominal hale getirilmiş hali Ek-4'teki Çizelge 4.2'de yer almaktadır.

Sınıf sayısı 2, 3, 4 ve 5 olarak belirlenen farklı risk skoru sınıflamalarından hangisinin karar kuralları oluşturacak karar ağacı algoritması ile çalıştırılacak veri kümesinde Risk sınıfı olarak yer alacağını belirlemek için kümeleme performanslarının karşılaştırılacaktır. Kümeleme performansı, sınıf sayısı sırasıyla 2, 3, 4, ve 5 olan Risk skoru özelliği içeren ayrı veri kümeleri üzerinde k-means kümeleme algoritmasının sırasıyla 2, 3, 4 ve 5 olarak belirlenmiş küme sayısı için çalıştırılmasıyla ölçülmüştür..

Dört farklı sınıf sayısı içeren veri kümeleri üzerinde WEKA yazılımında k-means algoritması çalıştırılmıştır. Elde edilen çıktı raporlarına örnek olarak RSCORE 2 Sınıf için 2 Küme oluşturacak K-means uygulaması Çıktı Raporu- I Ek-5'te sunulmuştur.

Çıktı raporlarına göre her sınıf için elde edilen kümeleme performans sonuçları aşağıdaki Çizelge 5.11'de özetlenmiştir.

Çizelge 5.11 Her sınıf için elde edilen kümeleme performans sonuçları

	Kümeleme Algoritması			
RSCORE özelliği için Sınıf Sayısı	kmeans			Sınıf için Ortalama Hata Oranı
	Küme Sayısı	Hatalı Gruplanan Birim		
		Sayısı	Oranı	
2	2	57	33,5	46,15
	3	64	37,6	
	4	94	55,3	
	5	99	58,2	
3	2	92	54,1	60,1
	3	95	55,8	
	4	109	64,1	
	5	113	66,4	
4	2	89	52,3	56,875
	3	89	52,3	
	4	102	60	
	5	107	62,9	
5	2	115	67,6	66,55
	3	111	65,2	
	4	111	65,2	
	5	116	68,2	

Çizelge 5.11.'e göre en iyi kümeleme performansı 2 Risk seviyesi için 2 kümeye atama yapılması işleminde elde edilmiştir. Bu sonuca göre bundan sonra yapılacak işlemlerde 2 farklı seviye için nominal hale getirilmiş Risk skoru içeren veri kümesi kullanılacaktır.

5.6.3. Anahtar özellik tahmininde etkili özelliklerin tespiti

İki risk sınıflı risk skoru özelliğine sahip veri kümesi üzerinde WEKA yazılımında özellik değerlendirici olarak CfsSubsetEval algoritması ile Best first algoritması uygulanması sonucu elde edilen Çıktı Raporu-II Ek 5'te gösterilmiştir.

Çıktı raporunda anahtar özelliklerle ilgili tahminlerde en iyi sonuç veren özellikler olarak hasta yaşı (AGE), tedaviye başlama yaşı (SAGE), cinsiyet (SEX), boy (HEI), Post Diyaliz Diyastolik Kan Basıncı (Ps-D), C Reaktif Protein (CRP), Potasyum (G) (K_I), Glisemi (GLI), Üre_BUN (G) (BUN_I), Paratiroidhormon (PTH) on farklı özellik tespit edilmiştir. Bu özelliklerden normalde Post Diyaliz Diyastolik Kan Basıncı ölüm riski ile ilgisiz, Boy ise Vücut Kitle İndeksi'ne katılması ile dolaylı ilgilidir. Diğer sekiz özellik ölüm riski ile ilgili tahmin edici olarak normalde de etkilidir.

Karar ağaçlarının oluşturulması adımıyla elde edilecek karar ağaçları ve kuralları incelenirken bu tahmin edici özellikler dikkate alınacaktır.

5.6.4. Karar ağaçlarının oluşturulması

Karar ağacı oluşturma sürecinde kullanılan iki risk sınıflı Risk skoru içeren veri kümesinde yer alan 55 adet özelliğe ait istatistik bilgileri Çizelge 5.12'de verilmiştir.

İki Risk seviye sınıflı Risk skoru özelliği içeren veri kümesi üzerinde J4.8, RandomTree ve PART karar ağacı algoritmaları uygulanmıştır. J4.8 ve

RandomTree algoritması çıktı olarak karar ağacı yapısı oluştururken PART algoritması çıktı olarak karar kuralları oluşturmaktadır.

WEKA yazılımı üzerinde üç algoritma kullanarak yapılan karar ağacı/karar kuralları oluşturma işlemlerinin çıktı raporları Çıktı Raporu-III, IV ve V Ek-5'te sunulmuştur.

İşlemler sonucunda elde edilen karar ağacı yanlış sınıflama oranları aşağıdaki Çizelge 5.13.'te özetlenmiştir.

Çizelge 5.12 Veri kümesinde yer alan özelliklere ilişkin istatistikler

Veri Niteliği	Kısaltma	Birimi	ESTARD	WEKA			
			Değişkenlik Oranı (%)	Min	Max	Ortalama	Standart Sapma
Hasta ID	ID	-	100,0	-	-	-	-
Cinsiyet	SEX	Nominal	1,2	0:81, 1:89			
Yaş	AGE	yıl	33,5	21	86	59,41	15,99
Tedavi Başlangıç Yaşı	SAGE	yıl	34,1	20	82	55,11	15,98
Ağırlık	HEI	kg	19,4	145	186	163,74	8,62
Boy	WEI	cm	60,0	40	99,7	65,38	12,01
Vücut Kitle İndeksi	VKI	kg/m2	100,0	13,9	39,3	24,06	4,39
Başl. Tansiyon (Systol)	SS_AKB	mmHg	7,7	70	200	136,36	22,39
Başl. Tansiyon (Diastol)	SD_AKB	mmHg	4,1	60	120	78,82	12,11
Beslenme Risk İndeksi	NRI	-	100,0	30,4	80,1	51,06	8,53
Toplam Seans Sayısı	SES	Adet	93,5	86	2066	566,15	410,68
Tedavi Süresi	SEAG	ay	61,8	11	207	58,86	42,48
Diyalizde Hayat Süresi	DHS	yıl	61,8	0,9	17,3	4,91	3,54
Haftalık Seans Süresi	HSS	saat	5,3	6	13,5	11,67	1,12
Damar yolu tipi	DYT	-	5,3	0:94, 1:27, 2:10, 3:16, 4:3, 5:17, 6:1, 7:3			
Antikoagülasyon Dozu	AKG	(ünite/seans)	18,2	0	10000	4406,35	1340,22
Pre-Diyaliz Ağırlık	Pre-W	kg	100,0	37,9	106,5	66,29	12,45
Pre-Diyaliz Systol	Pre-S	mmHg	66,3	82	186	130,68	18,78
Pre-Diyaliz Diastol	Pre-D	mmHg	31,8	52	103	76,80	8,49
Post-Diyaliz Ağırlık	Ps-W	kg	100,0	37,8	102,2	64,41	12,15
Post-Diyaliz Systol	Ps_S	mmHg	63,5	77	172	121,08	19,90
Post-Diyaliz Diastol	Ps-D	mmHg	36,3	60	93	72,84	8,66
Seans Başı-Sonu Kilo Farkı	ISW	kg	96,3	-1,4	4,2	1,88	0,89
Kan Akım Hızı	KAH	Qb	12,9	180	300	244,96	22,76
Eritropoetin İlaç Tipi	EIT	-	2,4	0:102, 1:39, 2:20, 3:9			
Eritropoetin Dozajı	EPOD	(ünite/ay)	79,4	0	56000	20936,21	11881,59
EPO Dozaj Verimliliği	EPOU	(ünite/ay)/kg	95,3	0	1057,7	336,19	197,83
Eritropoetin Direnci	EPOR	-	95,3	0	108,6	32,70	21,33
Rezidiv İdrar	RZI	ml	67,1	0	4227	332,71	490,37
Htc	Htc	%	98,8	25,92	44,69	32,44	3,50
CRP	CRP	mg/l	68,8	0,08	64	4,72	9,22
Na	Na	mEq/dl	62,9	127,08	147,33	136,82	3,41
K (G)	K_I	mEq/dl	81,2	3,8	6,55	5,20	0,49
K (Ç)	K_O	mEq/dl	82,4	2,84	4,73	3,92	0,36
K Farkı	K_F	mEq/dl	90,0	0,46	2,32	1,27	0,41
Glisemi	GLI	mg/dl	93,0	73,83	278	121,72	49,58
Üre-BUN (G)	BUN_I	mg/dl	95,3	65,17	213,67	140,90	28,49
Üre-BUN (Ç)	BUN_O	mg/dl	90,0	18,42	91,67	51,57	11,34
BUN Ratio	BUN_R	-	99,4	1,94	4,15	2,77	0,39
Kreatinin (G)	KRE_I	mg/dl	94,1	2,36	16,96	8,28	2,25
Kreatinin (Ç)	KRE_O	mg/dl	92,4	1,27	6,73	3,84	1,04
Ürik Asit	URA	mg/dl	78,2	3,93	9,4	6,33	0,91
URR (Üre Azalma Oranı)	URR	-	100,0	49,22	76,04	63,33	4,70
Kt/V (Üre Temizleme Oranı)	Kt/V	-	100,0	0,86	1,46	1,17	0,11
Ca X P	CaXP	-	100,0	24,72	69,17	44,66	8,62
iPTH	PTH	pg/ml	97,1	12	1433,36	234,13	240,61
T.Kolesterol	TKoll	mg/dl	73,5	104,5	310	177,72	36,01
HDL	HDL	mg/dl	35,3	24	123,3	40,42	9,61
LDL	LDL	mg/dl	68,8	8	180,5	81,68	24,63
Trigliserid	TGL	mg/dl	92,4	53,5	982	227,71	114,02
Total Protein	TPRO	g/dl	77,7	5,41	7,88	6,80	0,40
Albumin	ALB	g/dl	76,5	2,84	8,54	3,94	0,50
Ferritin	FERT	mg/dl	97,6	40,2	2014,8	506,42	318,64
TSI	TSI	-	100,0	10,23	58,21	26,93	9,27
Risk Skoru	RSCORE	-	46,5	0:104, 1:66			

Değerlendirmeye alınacak J4.8. ve PART algoritmalarının oluşturduğu karar ağacı ve karar kuralları aşağıda sıralanmıştır :

- *J4.8 Algoritmasının Ürettiği Karar Ağacının Karar Kuralı olarak Listesi*

- (1) EĞER AGE $\leq$ 58 ve GLI $\leq$ 155 İSE O ZAMAN RS0
- (2) EĞER AGE $\leq$ 58 ve GLI $>$ 155 ve Htc $\leq$ 31.16 İSE O ZAMAN RS1
- (3) EĞER AGE $\leq$ 58 ve GLI $>$ 155 ve Htc $>$ 31.16 İSE O ZAMAN RS0
- (4) EĞER AGE $>$ 58 ve SEX=1 ve GLI $\leq$ 136.92 ve WEI $\leq$ 68.5 ve DYT= 0 ve HDL $\leq$ 37.5 İSE O ZAMAN RS1
- (5) EĞER AGE $>$ 58 ve SEX=1 ve GLI $\leq$ 136.92 ve WEI $\leq$ 68.5 ve DYT = 0 ve HDL $>$ 37.5 İSE O ZAMAN RS0
- (6) EĞER AGE $>$ 58 ve SEX=1 ve GLI $\leq$ 136.92 ve WEI $\leq$ 68.5 ve DYT=2 İSE O ZAMAN RS0
- (7) EĞER AGE $>$ 58 ve SEX=1 ve GLI $\leq$ 136.92 ve WEI $\leq$ 68.5 ve DYT=1 ve ID $\leq$ 59 İSE O ZAMAN RS0
- (8) EĞER AGE $>$ 58 ve SEX=1 ve GL $\leq$ 136.92 ve WEI $\leq$ 68.5 ve DYT=1 ve ID $>$ 59 İSE O ZAMAN RS1
- (9) EĞER AGE $>$ 58 ve SEX=1 ve GLI $\leq$ 136.92 ve WEI $\leq$ 68.5 ve DYT=5 İSE O ZAMAN RS0
- (10) EĞER AGE $>$ 58 ve SEX=1 ve GLI $\leq$ 136.92 ve WEI $\leq$ 68.5 ve DYT=3 İSE O ZAMAN RS1
- (11) EĞER AGE $>$ 58 ve SEX=1 ve GLI $\leq$ 136.92 ve WEI $>$ 68.5 İSE O ZAMAN RS0
- (12) EĞER AGE $>$ 58 ve SEX=1 ve GLI $>$ 136.92 ve K_F $\leq$ 1.66 İSE O ZAMAN RS1
- (13) EĞER AGE $>$ 58 ve SEX=1 ve GLI $>$ 136.92 ve K_F $>$ 1.66 İSE O ZAMAN RS0
- (14) EĞER AGE $>$ 58 ve SEX=0 ve GLI $>$ 136.92 ve SEAG $\leq$ 91 İSE O ZAMAN RS1

(15) EĞER AGE>58 ve SEX=0 ve GLI>136.92 ve SEAG>91 İSE O ZAMAN RS0

- *PART Algoritmasının Ürettiği Karar Kurallarının Listesi*

(1) EĞER AGE ≤ 58 VE GLI ≤ 155 İSE O ZAMAN RS0

(2) EĞER SEX = 0 VE SEAG ≤ 91 VE SEAG > 23 İSE O ZAMAN RS1

(3) EĞER K_I ≤ 4.83 VE CRP > 0.58 İSE O ZAMAN RS1

(4) EĞER K_I ≤ 5.98 VE Pre-S ≤ 142 VE AKG ≤ 4423 VE SES > 112 İSE O ZAMAN RS0

(5) EĞER DYT = 0 VE TSI > 25.91 İSE O ZAMAN RS0

(6) EĞER EIT = 0 VE SAGESEX ≤ 66 VE TSI ≤ 19.44 İSE O ZAMAN RS0

(7) EĞER URA > 6.45 İSE O ZAMAN RS1

(8) EĞER Pre-S ≤ 142 VE TKoll > 158 İSE O ZAMAN RS1

6. SONUÇ VE ÖNERİLER

6.1. Bulguların Değerlendirilmesi

J4.8. ve PART algoritmalarının oluşturduğu karar kuralları, veri kümesinde yer alan 55 risk faktörünün (ID hariç) birbirleri ile etkileşimlerinin hangi Risk Skoru sınıfında hangi aralıklar için olduğunu belirtmektedir.

Karar kuralları Uzman eşliğinde incelendiğinde tespit edilen bulgulara ilişkin değerlendirmeler aşağıda maddeler halinde özetlenmiştir:

a) J4.8. ve PART algoritmalarının oluşturduğu ilk karar kuralı birbirinin aynı ve doğrudur. Glisemi değeri 155 mg/dl ve altı olan 58 yaş ve altı hastalar için Risk Seviyesi 0'dır. Glisemi değeri 155 mg/dl üstü olan 58 yaş üstü hastalar için risk seviyesi yüksektir.

Açlık kan şekeri (Glisemi) normalde 70-110 mg/dl arasındadır. 70 altı değerlerde Hipoglisemi, 110 üstü değerlerde Hiperglisemi söz konusudur. Glisemi değeri 155 mg/dl üstü durumda risk seviyesi yüksektir. Sonuç olarak yaş ve kan şekeri düzeyi yüksekliği bir arada riski yükseltmiştir.

b) J4.8 algoritmasının oluşturduğu 2. ve 3. karar kurallarına göre Glisemi değeri 155 mg/dl üstü olan 58 yaş ve altı hastalarda Hematokrit seviyesi %31,16 ve altı olduğunda risk seviyesi 0 dan 1'e çıkmakta yani artmaktadır.

Hemodiyaliz hastalarında hematokrit değeri %33-36 arasında olmalıdır. Kurala göre 58 yaş ve altı hastalar için kan şekeri yüksek ve hematokrit düzeyi normalin altında durumda risk artmaktadır.

Kan şekeri düzeyi yüksekliği ve Hematokrit düzeyi düşüklüğü bir arada olduğunda riski yükseltmektedir.

c) J4.8 algoritmasının oluşturduğu 4. ve 5. karar kurallarına göre 58 yaş üstünde, 68,5 kg ve altı ağırlıktaki, hemodiyaliz tedavisi için damar yolu erişimi olarak LAVF (Sol Arteriovenöz Fistül) tipi damar yolu kullanılan Erkek hastalarda glisemi düzeyi 136,92 mg/dl ve altı için HDL-Kolesterol (Yüksek Yoğunluklu Lipoprotein) düzeyi 37,5 mg/dl ve altına düştüğünde risk seviyesi 0'dan 1'e çıkmakta yani risk artmaktadır.

Glisemi düzeyi 136,92 mg/dl ve altı olduğu durum, normal sınır değeri olan 70-110 aralığı dikkate alındığında normale yaklaşan bir değerdir. Glisemi düzeyi HDL ve LAVF ile birlikte düşünüldüğünde de risk arttırıcı bir durum bulunmamaktadır.

Diyaliz hastalarında düşük beden ağırlığı riski azaltmaktadır. Ancak bu kuraldaki ağırlık faktörünü tek başına yorumlamak uygun değildir. Kuralda ağırlık özelliği yanında Vücut kitle indeksi veya Beslenme Risk İndeksi de yer alsaydı birlikte değerlendirilebilirdi.

20 yaşın üzerindeki kişilerde kan HDL-kolesterol düzeylerinin 40 mg/dl'nin üzerinde olması istenilen değerdir. HDL-kolesterol düzeyinin 37,5 mg/dl ve altı olması riskin artması demektir. Bu karar kurallarında HDL düzeyi doğrudan risk artışına neden olurken yaş, cinsiyet ile HDL düzeyi özellikleri anlamlı bir desen oluşturmuştur.

d) PART algoritmasının oluşturduğu 2. karar kuralına göre 23 ay ile 91 ay arasında hemodiyaliz yaşı olan kadın hastaların risk seviyesinin 1 olduğu görülmektedir.

Kronik Böbrek Yetmezlikli bir hasta için renal replasman tedavi seçenekleri arasında yer alan Hemodiyaliz Tedavisi görme süresi "ay" olarak takip edilmektedir. Tedavi süresi uzadıkça riskin arttığı bilinmektedir. Kuralda kadın hastalar için tedavi süresinin riske etkisine dair bir desen bulunmuştur. Buna göre yaklaşık 2 yıldan daha fazla bir süre (7,5 yıla kadar) diyalize giren kadın

hastalar yüksek risk grubuna girmektedir. Cinsiyet ve tedavi süresi faktörlerinin etkileşimi ile risk artışı gözlenmiştir.

e) PART algoritmasının oluşturduğu 3. karar kuralına göre Potasyum Giriş (hemodiyaliz seansı öncesi) düzeyi 4,83 mEq/L ve altında, C reaktif protein düzeyi 0,58 mg/l üstü hastalar 1 nolu risk seviyesine girmektedir.

Hemodiyaliz Tedavisi görmekte olan hastaların Kan Potasyum (K) düzeyleri bir hemodiyaliz tedavi seansı öncesi (Potasyum Giriş) ve aynı seans bitiminde (Potasyum Çıkış) olmak üzere rutin olarak ayda bir kez “Giriş” ve “Çıkış” olarak ölçülmektedir. Potasyum Giriş ve Çıkış düzeyleri için 3,5 mEq/L altında ve 6,5 mEq/L üzerinde değerlerde risk artmaktadır [154]. Kuralın belirttiği 4,83 mEq/L ve altı değerlerden 3,5 mEq/L altı seviyesi için risk yüksektir.

C reaktif protein düzeyi ile ilgili yapılan çalışmalarda 10 mg/l üstü durumlarda riskin 3,5 kat arttığı tespit edilmiştir [155]. Kuralın belirttiği 0,58 mg/l üstü değerler 10 mg/l üstü seviyesi için riski yükseltmektedir. Özetle Potasyum Giriş düzeyi ve C reaktif protein düzeyi etkileşimi ile risk artmıştır.

f) PART algoritmasının oluşturduğu 5. karar kuralına göre LAVF (Sol Arteriovenöz Fistül) tipi damar yolu kullanımı ve TSİ (Transferrin Saturasyon İndeksi) düzeyi %25,91 üstü olan durumlar 0 risk seviyesindedir.

Eritropoetin direncinin en önemli nedeni demir eksikliği olduğundan, diyaliz hastalarında TSİ düzeyi en az %20 olmalıdır [156]. Kuralda belirtilen %25,91 üstü durumlar için risk daha azdır.

g) PART algoritmasının oluşturduğu 8. karar kuralına göre ürik asit düzeyi 6,45 mg/dl üstü durumlar yüksek risk seviyesine girmektedir.

Kanda ürik asit düzeyinin üst sınırı kadınlarda 6-7 mg/dl, erkeklerde 7-8 mg/dl dır [157]. Kuralın belirttiği gibi 6,45 mg/dl üstü ürik asit seviyesi riski artırmaktadır.

Elde edilen karar kuralları incelendiğinde, temel özellik olarak seçilen risk skorunu en çok etkilediği tespit edilen 10 özellik arasından hasta yaşı (AGE), cinsiyet (SEX), C Reaktif Protein (CRP), Potasyum (G) (K_I), Glisemi (GLI) olmak üzere 5'inin karar kuralları içinde yer aldığı görülmüştür.

Yaş ve kan şekeri düzeyi, Glisemi ve Hematokrit düzeyi , Yaş-cinsiyet- HDL düzeyi, cinsiyet ve tedavi süresi, Potasyum Giriş düzeyi ve C reaktif protein düzeyi etkileşimleri riski yükselten risk faktörü etkileşimleri olmuştur.

Karar kurallarına göre C reaktif protein düzeyi ile Ürik asit düzeyi yüksekliği tek başına riski artıran faktörlerdir.

Damar yolu tipleri hastanın tıbbi uygunluk durumuna göre kullanılmakta olup, risk düzeyine etkisinin bilinmemesi nedeniyle araştırılması gereken bir parametredir. Bu nedenle değerlendirmeye alınan karar kurallarında yer alan damar yolu tipi özelliği hakkında bir yorum yapılmamıştır.

Gelecekte bu konuda yapılacak çalışmalarda, kaba küme teorisine dayalı metotlar, bulanık kümeleme algoritmaları gibi farklı veri madenciliği teknikleri kullanılması yanında risk skorunun tıbbi risk skorum sistemi kullanılarak elde edilmesi durumunda, klinik anlamda daha geçerli bilgiler elde edileceği düşünülmektedir. Bununla birlikte Türkiye'de hemodiyaliz hastaları ile ilgili ilk veri madenciliği uygulama çalışmalarından biri olduğunu düşündüğümüz bu tez kapsamında elde edilen sonuçlar tatmin edici düzeydedir.

6.2. Öneriler

Veri Madenciliğinin tıbbi kullanımı, daha önce birçok klinik araştırma gerektiren, hem ekonomik hem de insan sağlığı açısından sakıncaları olan tıbbi araştırmaların yerini kısmen de olsa doldurarak tıbbi araştırmalar için yeni bir ufuk sağlayacaktır.

Tez kapsamında yapılan uygulama çalışmasında, uygun verilerin elde edilmesi süreci, ülkemiz özelinde tıp alanındaki pek çok verinin bütünlük veri tabanlarında tutulmaması nedeniyle uzun ve zorlu geçmiştir.

Türkiye’de hemodiyaliz hizmeti veren sağlık kuruluşlarının kamu ve özel sektör olarak ikiye ayrılması yanında hizmetin yerinde verilmesi gerekliliğinden kaynaklı dağınıklık, hemodiyaliz hastalarına ait verilerin bugüne kadar bütünlük bir veri tabanında toplanmasına imkan vermemiştir.

Ancak aynı durumda olan Amerika Birleşik Devletleri ve diğer gelişmiş ülkeler bu konuda bütünlük bir veri tabanı kurmayı başarmışlardır. Bu sayede çok çeşitli veri kümeleri derleyerek daha fazla veri madenciliği çalışması yapılmasını ve kıymetli bilgiler elde edilmesini sağlamışlardır. Amerika’da 1988’de kurulan ve tüm hemodiyaliz merkezlerinden güncel veriler toplayan Amerikan Böbrek Veri Tabanı (US Renal Database) [158] bu konuda örnek alınabilecek bir veri tabanı uygulamasıdır.

Veri madenciliğinin sağladığı imkanlardan hayati öneme sahip tıp alanında en iyi şekilde yararlanmak için ulusal bir böbrek veri tabanı kurulması gereklidir. Bunu başarmak amacıyla öncelikle kamu veya özel sektöre ait hemodiyaliz merkezlerinin kendi veri tabanları oluşturulmalı, elde mevcut verilerin bir araya getirilerek muhafazası sağlanmalıdır. Ardından düzenleyici otorite olan kamu önderliğinde ulusal bir böbrek veri tabanı oluşturulması için bir araya gelinmeli ve Türkiye Böbrek Veri Tabanı en kısa sürede faaliyete geçirilmelidir.

Ulusal veri tabanının kurulması ve farklı disiplinler (Tıp, Endüstri Mühendisliği, Bilgisayar Mühendisliği, İstatistik vb.) arasında yapılacak ortak çalışmalar, gelecekte Nefroloji alanında yapılacak veri madenciliği çalışmalarının sayısının artmasını, hemodiyaliz hastalarını etkileyen risk faktörlerinin etkileşimi hakkında daha fazla bilgi elde edilmesini ve bu bilgilerden yararlanarak yeni tedavilerin geliştirilmesini sağlayacaktır.

Hızlı bir gelişim trendi gösteren Veri madenciliği, beraberinde getirdiği Veri Çiftçiliği ve Veri Füzyonu gibi kavramlar ve yeni geliştirilen teknikler yardımıyla tıbbın her alanında daha fazla uygulanmaya devam edecektir. Böylece hastalara daha kaliteli bir sağlık hizmeti sunulmasında etkin bir rol oynayacaktır.

KAYNAKLAR

1. Olson, D. L., Heidelberg, D. D., " Advanced Data Mining Techniques", **Springer**, 3 (2008)
2. Vahaplar, A., Dr. İnceoğlu, M.M., " Veri Madenciliği ve Elektronik Ticaret", **VII. Türkiye’de İnternet Konferansı**, 2-6 (2001)
3. Kaya, E., Bulun, M., Arslan, A., " Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları ", **AKADEMİK BİLİŞİM 2003, Çukurova Üni-Üniversitesi**, Adana, (2003)
4. Prof.Dr. Akpınar, H., " Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", **İ.Ü. İşletme Fakültesi Dergisi**, 29 (1):1-22 (2000)
5. Han, J., Kamber, M., " Data Mining Concepts and Techniques 2nd Ed.", Editor : Jim Grey, The Morgan Kaufmann Series in Data Management Systems, **Morgan Kaufmann** 2,8,12,14,15,29,30,398-403, (2006)
6. Chakrabarti, S., Witten, I. et al., " Data Mining Know It All ", **Morgan Kaufmann**, 2,14,39,40,41,108 (2008)
7. Kantardzic, M., "Chapter 9: Artificial Neural Networks Chapter 1-1.4", Data Mining Concepts, Models, Methods and Algorithms, **John Wiley & Sons**, (2003)
8. Nong, Ye, " Part II Management of Data Mining Chapter 1 Data Collection, Preparation, Quality and Visualization ", The Handbook of Data Mining, **Lawrence Erlbaum Associates**, 368,391,395,396,406 (2003)
9. Berry, M.J.A., Linoff, G.S., "Data Mining Techniques For Marketing, Sales, and Customer Relationship Management 2nd Ed.", **John Wiley & Sons**, 586 (2004)
10. Maimon, R., " Data Mining & Knowledge Discovery Handbook", **Springer**, 334 (2005)
11. Dr. Argüden, Y., Erşahin, B., " Veri Madenciliği : Veriden Bilgiye, Masraftan Değere", **ARGE Danışmanlık**, 48-63, (2008)
12. Albayrak, S., Koltan Yılmaz, Ş., " Veri Madenciliği Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama", **Süleyman Demirel Üniv. İktisadi ve İdari Bil. Fak. Dergisi**, 14 (1): 31-52 (2009)
13. Sumathi, S., Sivananda, S.N., " Introduction to Data Mining and its Applications", **Springer**, 178, 629 (2006)

14. Baykasoğlu, A., "Veri Madenciliği ve Çimento Sektöründe Bir Uygulama", **Akademik Bilişim 2005**, Gaziantep, 171:14, (2005)
15. Goebel, M., Gruenwald, L., " A survey of data mining and knowledge discovery software tools", **SIGKDD Explorations**, 1:20-33 (1999)
16. Elder, John F., Abbott, D. W., " A Comparison of Leading Data Mining Tools", **4th Annual Conference on Knowledge Discovery & Data Mining**, New York, USA, 234 (1998)
17. Güven, A. , Bozkurt, Ö. Ö., Kalıpsız, O., "Veri Madenciliğinin Geleceği", **Akademik Bilişim Konferansı 2007**, Dumlupınar Üniversitesi, Kütahya, (2007)
18. Hastie, T., " Modern Trends in Data Mining", **Stanford University Presentation**, August (2009)
19. Kriegel, H. P., Borgwardt, K. M., Kröger, P., Pryakhin, A., Schubert, M., Zimek, A., "Future trends in data mining", **Data Min Knowl Disc**, 15:87–97 (2007)
20. Zheng, H., Kusiak, A., " Prediction of Wind Farm Power Ramp Rates: A Data Mining Approach ", **Journal of Solar Energy Engineering**, 131/(031011) :1-8 (2009)
21. Akgün, H., Adalı, E., " Dempster - Shafer algoritmasının kullanımı ile sınıflandırma algoritmalarının birleştirilmesi ", **itüdergisi mühendislik**, 5 (4) : (2006)
22. Waltz, E., " Information understanding : Integrating data fusion and data mining processes", **IEEE 1999 International Symposium on Circuits and Systems**, Monterey, CA, USA, IEEE Report 0-7803-4455-3/98 (1999)
23. Kusiak, A., " Data Farming : A Primer ", **International Journal of Operations Research**, 2 (2): 48-57 (2005)
24. Hastie, T., Tibshirani, R., Friedman, J., " The Elements of Statistical Learning; Data Mining, Inference and Prediction", **Springer Series in Statistics**, New York, USA, 533 (2001)
25. Wang, J. (Editor), " Encyclopedia of Data Warehousing and Mining", **Information Science Reference**, 49, 140 (2006)
26. Larose, D.T., " Chapter 2 Regression Modelling, Chapter 4 ", Data Mining Methods and Models, **Wiley-IEEE Press**, 35 (2007)

27. Chen, Z., Kou, L., " A Note on the Estimation of the Multinomial Logit Model with Random Effects", **The. American Statistician**, 55 (2):89-95 (2000)
28. Efe, E., Bek, Y., Şahin, M., "SPSS'te Çözümleri ile İstatistik Yöntemler II", Kahramanmaraş Sütçü İmam Üniversitesi Rektörlüğü Yayın No : 7 **K.S.Ü. Basımevi**, Kahramanmaraş (2000)
29. Aydoğan, F., " E-Ticarete Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirimi", Yayınlanmamış Yüksek Lisans Tezi, **Hacettepe Üniversitesi Fen Bilimleri Enstitüsü**, Ankara, (2003)
30. Vahaplar, A., " Bir Coğrafi Veri Madenciliği Uygulaması", Yayınlanmamış Yüksek Lisans Tezi, **Ege Üniversitesi Fen Bilimler Enstitüsü**, İzmir, (2003)
31. Koyuncuğil, A. S., Özgülbaş, N., " İMKB'de İşlem Gören KOBİ'lerin Güçlü ve Zayıf Yönleri : Bir CHAID Karar Ağacı Uygulaması ", **Dokuz Eylül Üniversitesi İİBF Dergisi**, 23 (1): 1-22 (2008)
32. Yıldırım, S., " Tümevarım Öğrenme Tekniklerinden C4.5'in İncelenmesi", Yüksek Lisans Tezi, **İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü** , İstanbul, (2003)
33. Chang, L. Y., Wang, H. W., " Analysis of traffic injury : An application of non - parametric classification tree techniques ", **Accident Analysis Prevention**, 38:1019-1027 (2006)
34. D'Alisa, S., Miscio, G., Baudo, S., Simone, A., Tesio L., Mauro, A., "Depression is the main determinant of quality of life in multiple sclerosis: A classification-regression (CART) study" **Disability and Rehabilitation**, 28 (5):307-314 (2006)
35. Kayri, M., Boysan M., " Bilişsel Yatkınlık ile Depresyon Düzeyleri İlişkisinin Sınıflandırma ve Regresyon Ağacı Analizi ile İncelenmesi", **H. Ü. Eğitim Fakültesi Dergisi**, 34:168-177 (2008)
36. Özgülbaş, N., Koyuncuğil, A. S., " İMKB'de İşlem Gören KOBİ'lerin Veri Madenciliği Karar Ağaçları Algoritmalarından CHAID ile Profillendirilmesi ve Küreselleşme Sürecinde Güçlü ve Zayıf Yönlerinin Belirlenmesi", **10. Ulusal Finans Sempozyumu**, İzmir, (2006)
37. Larose, D. T., " Discovering Knowledge in Data An Introduction to Data Mining", **John Wiley & Sons**, Hoboken, New Jersey, 129 (2004)

38. MacKay, D. J. C., " Information Theory, Inference and Learning Algorithms", **Cambridge University Press**, UK, (2003)
39. Smith, K. A., " Neural Networks for Business : An Introduction, Neural Networks in Business: Techniques and Applications", Ed.: Smith Kate A., Gupta J. N.D. Gupta, **Idea Group Publishing**, USA, (2002)
40. Akel, V., Bayramoğlu, M. F., " Kriz Dönemlerinde Yapay Sinir Ağları İle Finansal Öngöründe Bulunma: İMKB 100 Endeksi Örneği", **International Symposium on International Capital Flows and Emerging Markets**, Bandırma, Balıkesir, (2008)
41. Öztemel, E., " Yapay Sinir Ağları ", **Papatya Yayınları**, İstanbul, 49-50 (2003)
42. Yıldız, T., Yıldırım, S., Altılar, T., " İstenmeyen İletilerin Paralelleştirilmiş Knn Algoritması ile Tespiti ", **Akademik Bilişim 2008 Çanakkale Onsekiz Mart Üniversitesi**, Çanakkale, (2008)
43. Er, H., Çetin, M.K., Çetin, E.İ., " Finanasta Evrimsel Algoritmik Yaklaşımlar : Genetik Algoritma Uygulamaları ", **Akdeniz İ.İ.B.F. Dergisi**, (10):73-94 (2005)
44. Toledo, M. D. G., "A Comparison In Cluster Validation Techniques ", M.S. Thesis, **Mathematics Department, University Of Puerto Rico**, (2005)
45. Vatansever, M., Büyüklü, A.H., "Using Visual Data Mining Techniques in Clustering Analysis and An Application", **Sigma Journal of Engineering and Natural Sciences**, Sigma 27:83-104 (2009)
46. Özekes S., " Veri Madenciliği Modelleri ve Uygulama Alanları", **İstanbul Ticaret Üniversitesi Dergisi**, 3:65-82 (2003)
47. Karabatak, M., İnce, M. C., " Apriori Algoritması ile Öğrenci Başarısı Analizi", **Elektrik Elektronik Bilgisayar Mühendisliği Sempozyumu, ELECO 2004**, Bursa, Elektronik-Bilgisayar 348-352, (2004)
48. Özçakır, F.C., Çamurcu, A. Y., " Birliktelik Kuralı Yöntemi için Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması ", **İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi**, 6 (12):21-37 (2007)
49. Alataş, B., " Nicel Birliktelik Kurallarının Keşfinde Bulanık Mantık ve Genetik Algoritma Yaklaşımı ", Yüksek Lisans Tezi, **Fırat Üniv. Fen Bilimleri Enstitüsü**, Elazığ, (2003)

50. Keim, D. A., " Information Visualization and Visual Data Mining", **IEEE Trans. on Visualisation and Computer Graphics**, 8 (1):1-8 (2002)
51. Kovalerchuk, B., Vityaev E., " Data Mining in Finance: Advances in Relational and Hybrid Methods ", **Kluwer Academic Publishers**, USA, (2002)
52. Becerra-Fernandez, et al., " Knowledge Management ", **Prentice Hall Additional Material**, Chapter 12 Discovering New Knowledge, Dekai Wu, 36 (2007)
53. Alkan, A., " Finansal Uygulamalarda Veri Madenciligi ", **TBD İstanbul Bilisim Kongresi**, İnforA Advanced Analytic Solutions Sunumu, Haziran (2007)
54. Aftarczuk, K., " Evaluation of Data Mining Algorithms implementing in Medical DSS", Masters Thesis, **School of Engineering at Blekinge Institute of Technology**, Ronneby, Sweden (2007)
55. Cios, K.J., Moore G.W., "Uniqueness of Medical Data Mining", **Artificial Intelligence in Medicine journal**, 26(1-2):1-24 (2002)
56. Prather, J.C., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L., Hammond, W.E., " Medical Data Mining : Knowledge Discovery in a Clinical Data Warehouse", **AMIA Annu Fall Symp.** 1997:101-5 (1996)
57. Houston,A.L., Chen,et al., "Medical data mining on the Internet: Research on a Cancer Information System", **Intelligence Review**, 13:437-466 (1999)
58. Mullins, I.M., Siadat, M.S., Lyman, J., Scully, K., Garrett, C. T., Miller, W.G., Muller, R., Robson, B., Apte, C., Weiss, S., Rigoutsos, I., Platt, D., Cohen, S., Knaus, W. A., " Data mining and clinical data repositories: Insights from a 667,000 patient data set", **Computers in Biology and Medicine**, 36 (12):1351-1377 (2006)
59. Kuo, R.J., Shih, C.W., " Association rule mining through the ant colony System for National Health Insurance Research Database in Taiwan", **Computers and Maths with Applications**, 54:1303–1318 (2006)
60. Yıldırım, P., Uludağ, M., Görür, A., " Hastane Bilgi Sistemlerinde Veri Madenciliği ", **Akademik Bilişim 2007, Çanakkale Onsekiz Mart Üniversitesi**, Çanakkale, (2007)
61. Stolba, N., Tjoa, A. M., " The relevance of data warehousing and data mining in the field of evidence - based medicine to support healthcare decision making", Computer Science, **Enformatika**, 12-17 (2006)

62. Dubitzky, W., Granzow, M., Berrar, D., " Fundamentals of Data Mining in Genomics and Proteomics", **Springer**, 257 (2007)
63. Johnson, P.E., Adomavicius, G., McCabe, R.M., Elidrisi, M., Meyer, G., O'Connor, P.J., Ramsey G.W., "Improving Chronic Disease Care Using predictive Modeling and Data Mining", **3rd INFORMS Workshop on Data Mining and Health Informatics**, Washington D.C., USA, (2008)
64. Ridinger, M., " American Healthways uses SAS to improve patient care". **DM Review**, 12:139 (2002)
65. Razali, A.M., Ali, S., " Generating Treatment Plan in Medicine: A Data Mining Approach", **Am J of Applied Sciences**, 6 (2):345-351 (2009)
66. Persson, M., Lavesson, N., "Identification of Surgery Indicators by Mining Hospital Data: A Preliminary Study", **20. DEXA Workshops 2009: Linz, Austria, First International Workshop on Database Technology for Data Management in Life Sciences and Medicine DBLM 2009**, 323-327 (2009)
67. Brosette, S. E., Spragre, A. P., Jones, W. T., Moser, S. A., "A data mining system for infection control surveillance", **Methods Inf Med**, 39:303-310 (2000)
68. Obenshain, M.K., "Application Of Data Mining Techniques To Healthcare Data", **Infection Control and Hospital Epidemiology**, 25 (8): 690–695 (2004)
69. Kumar, V., Kumar, D., Singh, R.K., "Outlier mining in medical databases : An Application of data mining in health care management to detect abnormal values presented in medical databases ", **Int. J. of Comp. Sci. & Network Sec.**, 8 (8):272–277 (2008)
70. Tsumoto, S., Hirano, S., "Data Mining as Complex Medical Engineering", **Journal of Japanese Society for Artificial Intelligence**, 22 (2):201-207 (2007)
71. Kimura, M., Watabe, S., Hayasaka, T., Tatsuno, K., Takahashi, Y., Aoto, T., Ohkura, M., Tsuchiya, F., " Application of Data Mining and Text Mining to the Analysis of Medical near Miss Cases ", Data Mining in Medical and Biological Research, Edited by: Eugenia G. Giannopoulou, **In-tech**, Croatia, (2008)
72. Hong, G. Q., " Multiple-minimum-support association rule mining for hospitalization prediction of hemodialysis patients", M.S. thesis, **Dept. of Information Engineering, National Tung Hua Univ.**, Taiwan, (2004)

73. Yeh, J. , Wu, T. , " A Decision Support System for Predicting Hospitalization of Hemodialysis Patients ", **World Academy of Science, Engineering and Technology**, 53:1128-1138 (2009)
74. Papageorgiou, E., Kotsioni, I., Linos, A., "Data Mining A New Technique In Medical Research", **HORMONES**, 4 (4):189-191 (2005)
75. Tessier D., Bardiaux B., Larre C., Popineau Y., "Data mining techniques to study the disulfide– bonding state in proteins: signal peptide is a strong descriptor", **Bioinformatics**, 20:2509-2512 (2004)
76. Saito, T., Sese, J. , Nakatani, Y., et al, " Data Mining tools for the Saccharomyces cerevisiae morphological database ", **Nucleic Acids Res**, 33: (Web Server Issue): W753-W757 (2005)
77. Georgii E., Richter L., Ruckert U., Kramer S., "Analyzing microarray data using quantitative association rules", **Bioinformatics**, 21 (2):123-129 (2005)
78. Perez-Iratxeta C., Wjst M., Bork P., Andrade M.A., "G2D: a tool for mining genes associated with disease", **BMC Genet**, 6:45 (2005)
79. Korn, R., Rohrig, S., Schulze-Kremer, S., Brinkmann, U., " Common denominator procedure : a novel approach to gene – expression data mining for identification of phenotype - specific genes", **Bioinformatics**, 21:2766-2772 (2005)
80. Ranjan, J., " Application of Data Mining Techniques in Pharmaceutical Industry ", **Jour. of Theoretical and Applied Information Technology**, 3 (4):61-67 (2007)
81. Hauben, M., Madigan, D., Gerrits, C., Walsch, L., Van Puijenbroek, E., "The role of Data mining in Pharmacovigilance", **Expert Opinion Drug Safety**, 4 (5):929-948 (2005)
82. Tsumato, S., "Discovery of Knowledge about Drug Side Effects in Clinical databases based on Rough Set Model", **AAAI Tech.Report SS-99- 01** (1999)
83. Zhou, Z. H., Jiang, Y., Yang, Y. B., Chen, S. F., " Lung cancer cell Identification based on artificial neural network ensembles", **Artificial Intelligence in Medicine**, 24 (1):25–36 (2002)
84. Chesney, T., Penny, K., Oakley, P., Davies, S., Chesney, D., Maffulli, N., Templeton, J., "Data Mining Medical Information: Should Artificial Neural Networks Be Used to Analyse Trauma Audit Data", **Int. Journal of Healthcare Information Systems and Informatics**, 1 (2):51-64 (2006)

85. Comak, E., Arslan, A., Turkoglu, I., "A decision support system based on support vector machines for diagnosis of the heart valve diseases", **Computer in Biology and Medicine**, 37 (1):21-27, (2007)
86. Vlahou, A., Schorge, J. O., Gregory, B. W., Coleman, R. L., "Diagnosis of ovarian cancer using decision tree classification of mass spectral data" **Journal of Biomedicine and Biotechnology**, 5:308-314 (2003)
87. Hunt Kinsky, B., Venkatesh, S., Petros, P., " Bayesian networks and Decision trees in the diagnosis of female urinary incontinence", Engineering in Medicine and Biology Society, **Proceedings of the 22nd Annual International Conference of the IEEE**, 1:551-554 (2000)
88. Sleeman, D., Fluck, N., Gyftodimos, E., Moss, L., Christie, G., "An Intelligent Aide for Interpreting a Patient's Dialysis Data Set", **Blood Pressure Prediction & Clustering**, 89:169-179 (2007)
89. Demsar, J., Zupan, B., Aoki, N., Wall, M.J., Granchi, T.H., Beck, J.R., "Feature mining and predictive model construction from severe trauma patient's data," **Int. J. of Medical Informatics**, 63:41-50 (2001)
90. Mena, L., Gonzalez, J.A., " Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic", **Proclamation of the 19th Int. FLAIRS Conf. (FLAIRS-2006)**, Melbourne Beach, Florida, USA, 574-579 (2006)
91. Wiggins, M., Saad, A., Litt, B., Vachtsevanos, G., "Genetic Algorithm-Evolved. Bayesian Network Classifier for Medical Applications", **Tenth World Soft Computing. Conf. and Trans. (WSC'10 2005)**, (2005)
92. Abraham, R., Simha, J. B., Iyengar, S. S., " A comparative analysis of discretization methods for Medical Data mining with Naive Bayesian classifier", **9th Int. Conf. on Information Technology (ICIT'06)**, icit, 235-236, (2006)
93. Abraham, R., Simha, J.B., Iyengar, S.S., "Medical Datamining with a New Algorithm for Feature Selection and Naïve Bayesian Classifier", **10th Int. Conf. on Information Technology (ICIT 2007)**, icit, 44-49, (2007)
94. Stefanowski, J., Slowinski, K., " Rough set theory and rule induction techniques for discovery of attribute dependencies in medical information systems", **Bulletin of the Polish Academy of Sciences**, 46 (2):247-263 (1998)
95. Stepaniuk, J., " Rough Set Based Data Mining in Diabetes Mellitus Data Table", **EUFIT '98, September 7-10, 1998**, Intelligent techniques and soft computing, Aachen, Germany, Verlag Mainz: 980-984, (1998)

96. Internet : Berault, J., "Data Mining Diabetic Databases Are Rough Sets a Useful Addition?", <http://www.galaxy.gmu.edu/interface/I01/I2001Proceedings/JBreault/JBreault-Paper.pdf> (2001)
97. Kusiak, A., Caldarone, C., Kelleher, M., Lamb, F., Persoon, T., Burns, A., "Hypoplastic left heart syndrome knowledge discovery with a data mining approach", **Computers in Biology and Medicine**, 36:21-40 (2006)
98. Cosic D., Loncaric, S., "Rule-based labeling of CT head image", Lecture Notes in AI, **Springer-Verlag**, Berlin, Germany, 1211:453-456, (1999)
99. Duch, W., Adamczak, R., Grabczewski, K., Zal, G., Hayashi, Y., "Fuzzy and crisp logical rule extraction methods in application to medical data", Computational Intelligence and Applications, **Springer-Verlag**, Berlin, Germany, 23:593–616 (2000)
100. Brameier, M., Banzhaf, W., "A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining", **IEEE Trans. on Evolutionary Computation**, 5 (1):17-26 (2001)
101. Novobilski, A., Kline, J., Fesmire, F., " Using a Genetic Algorithm to Identify Predictive Bayesian Models in Medical Informatics", **Int. Conf. on Information Technology (ITCC 2004)**, Las Vegas, Nevada, (2004)
102. Richards, G., Rayward-Smith, V.J., Sönksen, P.H., Carey, S., Weng C., "Data mining for indicators of early mortality in a database of clinical records", **Artificial Intelligence in Medicine**, 22 (3):215–231 (2000)
103. Gupta, A., Kumar, N., Bhatnagar, V., " Analysis of Medical data using data mining and formal concept analysis", **World Academy of Science Engineering and Technology**, 11:61-65 (2005)
104. Dogan, S., Turkoglu, I., " Diagnosing Hyperlipidemia Using Association Rules", **Math. and Comp. Applications**, 13 (3):193-202 (2008)
105. Niimi, A., Tazaki, E., " Combined Learning Method of Apriori Algorithm and Genetic Programming for Medical Database", **Japan Journal of Medical Informatics**, 20:584-585 (2000)
106. You, C., Holder, L., Cook, D., "Application of graph-based data mining to metabolic pathways", **ICDM Workshop on IEEE Int. Conf. on Data Mining in Bioinformatics**, (2006)
107. Tang, T., Zheng, G., Huang, Y., Shu, G., Wang, P., " A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis", **IEMS**, 4 (1):102-108 (2005)

108. Huang, M. J., Chen, M.Y., Lee, S.C., "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis", **Expert Systems with Applications**, 32:856-867 (2007)
109. Podgorelec, V. , Kokol, P. , Stiglic, M. M. , Hericko, M. , Rozman, L. , "Knowledge discovery with classification rules in a cardiovascular data set ", **Comp. Meth. and Prog. in Biomedicine**, 80:S39-S49 (2005)
110. Xie, X. , Li, X. , Wan, S. , Gong, Y. , " Mining -Ray images of SARS patients", Editors: G.J. Williams , S.J. Simoff, Data Mining, **LNAI 3755**, 282-294 (2006)
111. Oliveiraa, L.L.G., Almeida e Silvaa, S., Vilela Ribeirob, L.H., Coelhoc, C.J.,de Oliveiraa, R.M., Andrade, A.L.S.S., "Computer-aided diagnosis in Chest Radiography for detection of childhood Pneumonia", **Int. J. of Medical Informatics**, 77 (8): 555-564 (2007)
112. Lovis, C., Colaert, C., Stroetmann, V.N., " Debug IT for Patient Safety – Improving the treatment with Antibiotics through Multimedia Data Mining of Heterogeneous Clinical Data", **Stud Health Technol. Inf.**, 136:641-646 (2008)
113. Tsevas, S., Iakovidis, D., Papamichalis, G., " Mining patterns of lung infections in chest radiographs", **Art. Intelligence Appl. and Innov. III**, 296:205-213 (2009)
114. Cox, S., Oakes, M., Wermter, S., Hawthorne, M., "AudioMine: Medical Data mining in heterogeneous Audiology records", **World Academy of Science, Engineering and Technology**, 1:5-8 (2005)
115. Sacchi, L., Larizza, C., Combi, C., Bellazzi, R., "Data mining with temporal abstractions: learning rules from time series", **Journal of Data Mining and Knowledge Discovery**, 15 (2):217-247 (2007)
116. Chittaro, L., Combi, C., Trapasso, G., "Data Mining on Temporal Data: A Visual Approach and its Clinical Application to Hemodialysis", **Journal of Visual Languages & Computing**, 14 (6):591-620 (2003)
117. Sandholm, T., Brodley, C., Vidovic, A., Sandholm, M., " Comparison of regression methods, symbolic induction methods and NN in morbidity diagnosis and mortality prediction in equine gastrointestinal colic" **AAAI Spring symposium series**, AI in medicine: Appl.of current technologies, Stanford University, CA, USA, 154-159 (1996)
118. Matyja, D., " Applications of Data Mining Algorithms to Analysis of Medical Data", Masters Thesis, **School of Engineering Blekinge Inst. of Technology**, Ronneby, Sweden (2007)

119. Daud, M.N.R., Corne, D.W., "Human readable rule induction in medical data mining: A survey of existing algorithms", **WSEAS European Computing Conference**, Athens, Greece, (2007)
120. Kumarkolla, P., Sekhar, R., " Data mining : understanding data and disease modelling Guntur Region ", http://www.csre.iitb.ac.in/~csre/conf/wp-content/uploads/fullpapers/PS1/PS1_2.pdf, 2001
121. Sebban, M., Mokrousov, I., Rastogi, N., Sola, C., "A data mining approach to spacer oligonucleotide typing of Mycobacterium tuberculosis", **Bioinformatics**, 18 (2):235-243 (2002)
122. Wren, J., Garner, H., " Data mining analysis suggests an epigenetic pathogenesis for Type 2 Diabetes ", **Journal of Biomedicine and Biotechnology**, 2:104–112 (2005)
123. Schmidt, S., Vuillermine, P., Jenner, B., Ren, Y., Li, G., Phoebe, Y. C., " Mining Medical Data: Bridging the Knowledge Divide", **eResearch Australasia Conference 2008**, Melbourne, Australia, (2008)
124. Le Duff, F., Muntean, C., Cuggia, M., Mabo, P., " Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", **MEDINFO**, 1256-1259, (2004)
125. Palaniappan, S., Awang, R., " Intelligent Heart Disease Prediction System Using Data Mining Techniques", **Int. J. of Computer Science and Network Security**, 8 (8):108 - 115 (2008)
126. Antonie, M., Zaiane, O., Coman, A., " Application of Data Mining Techniques for Medical Image Classification", **2nd Int. Workshop on Multimedia Data Mining**, San Fransisco, USA, 94-101 (2001)
127. Bellaachia, A., Guven, E., " Predicting Breast Cancer Survivability using Data Mining Techniques ", **9th Workshop on Mining Scientific and Engineering Data sets in conjunction with the 6th SIAM Int. Conf. on Data Mining (SDM 2006)**, Bethesda, MD, USA, (2006)
128. Abe, A., Hagita, N., Furutani, M., Furutani, Y, Matsuoka, R., "Possibility of Integrated Data Mining of Clinical Data", **Data Science Journal**, 6:S104-S115 (2007)
129. Karabatak, M., Ince, M.C., " An expert system for detection of breast cancer based on association rules and neural network", **Expert Systems with Applications**, Corrected Proof In Press, (2008)

130. Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R., Cetta, S., " Intelligent data analysis techniques for quality assessment of hemodialysis services", **Proc. of the Workshop on Intelligent Data Analysis and Pharmacology**, London, UK, (2001)
131. Sai, M.S.S., Thrimurthy, P., Purushothaman, S., " Implementation of back-propagation algorithm for renal data mining", **Int. J. of Computer Science and Security**, 2 (2):35-48 (2009)
132. Shah, S., Kusiak, A., Dixon, B., " Data Mining in Predicting Survival of Kidney Dialysis Patients - Invariant object approach", in Proceedings of **Photonics West - Bios 2003, Bass, L.S. et al. (Eds.), Lasers in Surgery : Advanced Characterization, Therapeutics and Systems XIII, Vol. 4949, SPIE**, Bellingham, WA, USA, 1-8 (2003)
133. Kusiak A., et al., " Predicting Survival Time for Kidney Dialysis Patients A Data Mining Approach", **Computers in Biology and Medicine**, 35:311-327 (2005)
134. Bries, M.F., " Modeling of Hemodialysis Patient Hemoglobin : A Data Mining Exploration ", Masters Thesis, **Dept. of Industrial Engineering, University of Iowa**, USA (2007)
135. İnternet : DIADER Ekim 2009 Raporu, Türkiye’de Hemodiyaliz Hizmetlerinde Özel Diyaliz Merkezlerinin Rolü, Kalite Maliyet İlişkisi, Diyaliz Maliyet ve Geri Ödemeleri, Kamu Maliyetinin Düşürülmesi, <http://www.diader.org.tr/dosya/DiyalizRaporEkim2009.pdf> (2009)
136. Dr. Kurubacak, N., " Hemodiyalize Giren Diabetes Mellituslu Olgularda Troponin I, Yüksek Duyarlıklı C-Reaktif Protein (HS_CRP) ve Leptin düzeylerinin Karşılaştırılması", Uzmanlık Tezi, **T.C. Sağlık Bak. Taksim Eğitim ve Araştırma Hastanesi Dahiliye Kliniği**, İstanbul (2007)
137. İnternet : National Kidney Foundation, <http://www.kidney.org/kidneydisease/ckd/knownGFR.cfm> (2009)
138. İnternet : National Kidney Foundation, <http://www.kidney.org/kidneydisease/ckd/index.cfm> (2009)
139. İnternet : Implementation of Changes in End Stage Renal Disease (ESRD) Payment for Calendar Year 2006, <http://www.cmh.hhs.gov/MLN/MattersArticles/downloads/MM4196.pdf> (2006)
140. Sungurtekin, H., Gürses, E., Hancı, V., Sungurtekin, U., " Hospitalize Hastalarda Malnütrisyonun Nütrisyonel Risk İndeksi ile Saptanması", **Türk Anest Rean Der Dergisi** 31:368-372 (2003)

141. İnternet : " Darbepoetin alfa in oncology patients", The United Kingdom Medicines Information Pharmacists' Group and The National Prescribing Centre Report, Pub. Date: February 2002, Monograph Number: 3/02/02, www.ukmi.nhs.uk/NewMaterial/Secure/darbepoetin.pdf (2009)
142. İnternet : Türk Nefroloji Derneği Nefrolojide Kullanılan Formüller, <http://www.tsn.org.tr/index.php?file=formuller/index> (2009)
143. Goldwasser, P., Mittman, N., Antignani, A., Burrell, D., Michel, M.A., Collier, J., Avram, M., "Predictors of mortality in hemodialysis patients", **J. Am. Soc. Nephrol.** 3:1613-1622 (1993)
144. Salahudeen, A.K., "Obesity and survival on dialysis", **Am J Kidney Dis** 41:925-932 (2003)
145. Kalantar-Zadeh, K., Abbott, K.C., Salahudeen, A.K., Kilpatrick, R.D., Horwich, T.B., "Survival advantages of obesity in dialysis patients", **Am J Clin Nutr** 81:543-554 (2005)
146. Bouillanne, O., Morineau, G., Dupont, C., Coulombel, I., Vincent, J., Nicolis, I., Benazeth, S., Cynober, L., Aussel, C., "Geriatric Nutritional Risk Index : a new index for evaluating at-risk elderly medical patients", **Am J Clin Nutr** 82:777-783 (2005)
147. Kalantar-Zadeh, K., McAllister, C.J., Lehn, R.S., Lee, G.H., Nissenson, A. R., Kopple, J. D., " Effect of malnutrition - inflammation complex syndrome on EPO hyporesponsiveness in maintenance hemodialysis patients", **Am J Kidney Dis** 42:761-773 (2003)
148. İnternet : Online website for csv to arff conversion, <http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php> (2009)
149. Tsaib, C., Linb, Y., Wub, V., Chub, T., Chenb, Y., Huc, F., Wub, K., Koa, W., " SAPS 3 at dialysis commencement is predictive of hospital Mortality in patients supported by extracorporeal membrane oxygenation and acute dialysis", **Eur. J. Cardiot. Surg** 34:1158-1164 (2008)
150. Levy, J. M., Robst, J., Ingber, M. J., " Risk-Adjustment System for the Medicare Capitated ESRD Program", **J Health Care Financing Review** 27 (4):53-69 (2006)
151. Buntin, M.B., Garber, A. M., McClellan, M., Newhouse, J.P., "The Costs of Decedents in the Medicare Program : Implications for Payments to Medicare Choice Plans", **Health Serv Res.** 39(1):111-130 (2004)

152. Pope, G.C., Kautter, J., Ellis, R.P., Ash, A.S., Ayanian, J.Z., Iezzoni, L.I., Ingber, M. J., Levy, J. M., Robst, J., " Risk Adjustment of Medicare Capitation Payments Using the CMS - HCC Model", **J. Health Care Financing Review** 27 (4):119-141 (2004)
153. All Medicare Advantage Organizations and Other Interested Parties Note, Published Date: 02/04/2007, **Centers for Medicare & Medicaid Services (CMS)**, USA (2007)
154. Factor, K.F., " Potassium Management in Pediatric Peritoneal Dialysis Patients : Can a Diet with Increased Potassium Maintain a Normal Serum Potassium without a Potassium Supplement? ", **Advances in Peritoneal Dialysis** 23:167-169 (2007)
155. Korevaar, J. C., v. Manen, J. G., Dekker, F. W., de Waart, D. R., Boeschoten, E. W., Krediet, R. T., " Effect of an Increase in CRP level during a HD session on mortality", **J Am Soc Nep.** 15:2916-2922 (2004)
156. Evrenkaya, T. R., Atasoyu, E. M., Ünver, S., Gültepe, M., Narin, Y., Tülbek, M.Y., "Hemodiyaliz Yeterliliği ile Komorbid Faktörler Arasındaki İlişki", **Türk Nefroloji Diyaliz ve Transp. Dergisi** 11(1):44 (2002)
157. Obermayr, R.P., Temml C., Gutjahr,G., Knechtelsdorfer, M., Oberbauer, R., Klauser-Braun R., "Elevated Uric Acid Increases the Risk for Kidney Disease", **J Am Soc Nep.** 19:2407–2413 (2008)
158. İnternet: US Renal Database, <http://www.usrds.org/> (2009)

EKLER

EK-1 Özel Can Diyaliz Merkezi arşiv görünümü



Resim 1.1 Verilerin toplanması çalışması sırasında Özel Can Diyaliz Merkezi arşivinden bir görünüm

EK-2 Veri toplamada kullanılan formlar

ÖZEL CAN HEMODİYALİZ MERKEZİ HASTA BİLGİ FORMU - I

07-06-2004
MÜRACAAT TARİHİ

07-06-2004 / 07 / Resadine
DOĞUM TARİHİ/CİNSİYET/MEMLEKET

0 Rh(+)
KAN GRUBU

SSK
SOSYAL GÜVENLİK SİSTEMİ

hal Emekli
MESLEK

GATA
HEMODİYALİZ RAPORU VEREN KURUM

MÜRACAATINDA EPİKRİZİ VAR MI?

Hds RAPOR TARİH/NO

EV ADRESİ/EV TELEFONU

İŞ ADRESİ/İŞ TELEFONU

BBY NEDENİ PRİMER HASTALIK: Hipertansiyon Nefropati

CO-MORBİD DURUMLAR:

1. Hipertansiyon 2. 3. 4. 5. 6. 7. 8.

ALERJİK HASTALIK-İLAÇ ÖYKÜSÜ

6
AŞI ÖYKÜSÜ

Monogril 20 1x1 (Ak), Lasix 1x2, Amlodipin 10mg 1x1 (Sg)
MÜRACAATI SIRASINDA KULLANDIĞI İLAÇLAR

Alfa D3 1x1 4mloliz 1x1/2 (Ak)

Evlü / 1949 φ
MEDENİ DURUMU/AKRABA EVLİLİĞİ

1949 / Evli
EŞİNİN YAŞI/MESLEĞİ/SOSYAL GÜVENCE

1949 / Hipertansiyon
EŞİNİN ÖĞRENİM VE SAĞLIK DURUMU

3
EVDEKİ FERT SAYISI

Okul / 0.14
ÖĞRENİM DÜZEYİ/SOSYOEKONOMİK DURUM

5 / Sağ Sağlıklı
ÇOCUK SAYISI/SAĞLIKLARI

+ /
EV SAHİBİ/KİRA/LOJMAN

φ / φ
SİGARA/ALKOL

Resim 2.1 Hasta Bilgi Formu Örneği (I)

EK-2 (Devam) Veri toplamada kullanılan formlar

ÖZEL CAN HEMODİYALİZ MERKEZİ HASTA BİLGİ FORMU - II

HASTANIN MÜRACAAT SIRASINDAKİ YAKINMALARI:

İştahsızlık	Bulantı	Hazımsızlık hissi	Dışme (teberrünnat)
Halsizlik	Kusma	Şişkinlik	Yastık sarımsı
Bitkinlik	Kabızlık	Pirüzis	Çarpıntı (teberrünnat)
Yorgunluk	İshal	Karın ağrısı	Geğir ağrısı
Bacaklarda kesilme hissi	Kanlı gata	Epigastriumda kızınma, yanma hissi	Yan ağrısı
Kilo kaybı	Kanlı kusma	Erken doyma hissi	Öksürük
	Burun kanaması	Tad alma hissinde değişiklik	Balgam (kanlı/kamçı)
Ateş	Baş ağrısı	Kaşınma	Dysuri/lik idrar
Üşüme	Baş dönmesi	Cilt kuruluğu	Noktür
Türme	Ortostatik değişiklik	Terleme hissi	Hematur
Artralji	Unutkanlık	Saç dökülmesi	İnce idrar
Myalji			Çatalı idrar
Ellerde uyuşma/karınalanma		Menstrual düzensizlik	Görme problemi
Bacaklarda ayaklarda uyuşma/karınalanma		Vajinal kanama	İşitme problemi
Bacaklarda huzursuzluk hissi		Libido kaybı/empotans	Koku doyma
Bacaklarda şişlik			

HİKAYESİ (Böbrek hastası olarak ne kadar süredir, hangi merkez ve hekimce izleniyor?, Hangi RRT şeklinde ne zaman tedaviye başlanmış?, Damar yolu nerede, ne zaman açılmış?)

2-3 ay kadar önce ağız ve burundan kanama ile hastaya basırcan hasta (SSK Dişhekimliği) yapılan tetkiklerinde hipertansiyon, diyabet ve üreterik deşar ların yetersiz olduğu tespit edilmiş ve GATA Nefroloji'de tabibe alınmış. KBY teşhisi konulmuş.

09.06.2024 ile özel Can Dk Mkt

HASTANIN MÜRACAATI SIRASINDA BESLENMESİ İLE İLGİLİ YAKLAŞIMI VE BUNUN BELİRLEYİCİLERİ

HASTANIN MERKEZİMİZE MÜRACAAT NEDENİ:

SSK Sıra revkli	Nefrolog duyumu	Hemşire duyumu
Evinin yakınlığı	Nefrolog/doktor tavsiyesi	İş sahsını tanıyor olması
Personeli tanıyor olması	Seans hekimlerinin methi	Diğer

Resim 2.2 Hasta Bilgi Formu Örneği (II)

EK-2 (Devam) Veri toplamada kullanılan formlar

ÖZEL CAN HEMODİYALİZ MERKEZİ HASTA BİLGİ FORMU - III

HASTANIN FİZİK DEĞERLENDİRMESİ (OBJEKTİF VE POZİTİF BULGULAR)

Boy: 175cm Ağırlık: 78.800kg Yaş: 77kg Kuru ağırlık (EDW):

MAC: TSE: Kuru ağırlık (EDW):

Vücut sıcaklığı: 36.4°C Vücut sıcaklığı: 36.4°C Keith-Wegener:

Görüşme:
 Sağ PAUF (+)
 Sağ nefisaltı ve bağışık akciğer mevcut
 Solunum muayenesi N. olarak değerlendirildi.

İmza: _____

Uzman değerlendirme:

İmza: _____

KLİNİK VE LABORATUAR PROBLEMLER

01	2003 HSG → Sağ Nefisaltı	09
02	30. Grade I Renal parankimal Hast.	10
03	Minimal BPH	11
04		12
05		13
06		14
07		15
08		16

Resim 2.3 Hasta Bilgi Formu Örneği (III)

EK-2 (Devam) Veri toplamada kullanılan formlar

ÖZEL CAN DİYALİZ MERKEZİ HEMODİYALİZ HASTASI AYLIK HEKİM İZLEM FORMU

Hasta İsmi: Hasta Yaşı: 67 Dx Yaşı: 29 ay Tarih: 10.09

Pozitif fizik muayene bulguları: AKB: 100/40 mm/Hg Vuru: /dk

Damaryolu: Sol SCV katman katman (5/04)

Post dx ağırlığı: 51 Kg

Boy: 180 cm Rezidiv idrar miktarı: 0 ml/gün.

Klinik problem ve tanıları:

1. Kolesterolüleri (1998)
2. tx nefrotelimi (5/07)
3. Ctx böbrekte malignite
4. Surrenal yetmezliği
5. G.U.T
6. Gastrointestinal
7. Atrial fibrilasyon
8. Katman sol (5/09) sağ (6/09)
9.
10.

Laboratuvar problem sonuçları:

1.
2.
3.
4.
5.
6.
7.
8.
9.

Hepatit Aşı Takip: HBs Ab: (-)

1. Doz 2. Doz 3. Doz

Önerilen Diyaliz Tedavisi:

Dx sıklığı: 3 seans/haft. Qb: 250 ml/d Diyasol: Ren-Med Bic.

Dx süresi: 4 saat/seans Qd: 500 ml/dk Konduktivite/Na: 140

Diyalizör: 1... polysulfon Diyalizat ısı: 36°C Antikoagülasyon: 2500 + 50000

Önerilen Diyet:

Su ve tuz kontrollü ☐ Normal Diyet ☐ Diyabetik Diyet ☐

Önerilen İlaç Tedavisi:

1.
2.
3.
4. Antifosfat ca 3x1
5. Lenoxon 1x1
6. Deltacortril 5x1
7. Methilum 1x1
8. Lenoxon 30x1
9. Desigrel 50 2x1
10. Cipram 1x1 + 1/2
11. Zorax 1x1
12. Dekun 1x1 (6)
13. Renal 1x1 (6)
14. Lenoxon 25x1
15. Avil ampil (28/10) ✓ (6)
16.
17.
18. **CAN DİYALİZ MERKEZİ**
19. **Prof. Dr. Mustafa K. ALI**
20. **İç Hastalıkları Uzmanı**
21. **Özellikler: Hemodiyaliz**
22. **Sorumlu: 0532-3437**
23. **Diploma No: 0532-3437**

Resim 2.4 Aylık Hekim İzlem Formu Örneği

EK-2 (Devam) Veri toplamada kullanılan formlar

ÖZEL CAN HEMODİYALİZ MERKEZİ HEMADİYALİZ SEANSI HEMŞİRE İZLEM FORMU

HASTA İSMİ :
Dx ŞİKLİĞİ : Haftada 3 Seans

TARİH : 05.10.2009
Dx BAŞLAMA / BİTİŞ SAATİ : 16⁰⁰ - 18⁰⁰

SEANS NO : 351
MAKİNE NO : 5
Dx SÜRESİ : 4 Saat

DIYALİZÖR : F-5
DIYALİZAT : FRESINIUS
DAMAR YOLU (Vascular access): RAUF
DAMAR YOLU HAZIRLIĞI : B+SS
ANTİKOAGÜLASYON : 5000 + 1500 G Neq.

KAN AKIM HIZI (QB): 250
DIYALİZAT AKIM HIZI (Qd): 250
DIYALİZAT ISI: 36
KONDÜKTİVİTE (Nk): 142
UF ŞEKLİ:

PREDİYALİZ DEĞERLENDİRMESİ

PreDx Ağırlığı: 76.7 / 2-2
AKB: 120/80
Nabız: 76
İsi: 36
Yakınması:
Mental Durum:
Pozisyonu
Dispne
Solunum
Göğüs Ağrısı:
Öden:
Damar Yolu: RAUF
Kaşıntı:

POSTDİYALİZ DEĞERLENDİRMESİ

PreDx Ağırlığı: 76
AKB: 120/80
Nabız: 76
İsi: 36
Yakınması:
Mental Durum:
Pozisyonu
Dispne
Solunum
Göğüs Ağrısı:
Öden:
Damar Yolu: RAUF
Kaşıntı:

Saat: 16⁰⁰ 17⁰⁰ 18⁰⁰ 19⁰⁰

AKB: 120/80

Nabız: 76

Qb

Vp

Cdtv

230

220

210

200

190

180

170

160

150

140

130

120

110

100

90

80

70

60

Proseski bitki ve sağlıdır.

20ml Arter (2.1)

250cc % 0.9 Nacl

Sorumlu Hemşire: Kaya

ile çıkarıldı.

İmza

Resim 2.5 Seans İzlem Formu Örneği

EK-2 (Devam) Veri toplamada kullanılan formlar

CAN DİYALİZ MERKEZİ RUTİN TETKİK İZLEM FORMU

Adı Soyadı: _____

	1/08	2/08	3/08	4/08	5/08	6/08	7/08	8/08	9/08	10/08	11/08	12/08
Hb	10,1	10,9	9,6	9,7	11,2	10,5	9,8	9,0	9,5	11,1	15,1	13
Htc	30,7	32,7	27,9	28,1	32,9	31,9	29,6	30,4	28,4	30,6	27,1	32,7
MCV	91,3	97,1	93,9	92,7	86,2	84,3	84,8	97,7	93,6	92,5	95	89,6
MCH	29,9	32,5	31,5	29,5	28,5	22,0	25,4	27,8	30,4	29,6	30,5	30,1
MCHC	32,8	33,4	33,5	31,8	22,0	22	30,0	31,7	33,3	33,6	32,1	33,6
Lökosit	9,8	10,5	8,9	7,8	8,6	8,2	13,7	8,9	7,6	8,4	8,8	7,1
Platelet	223,0	355,0	289,0	140,0	277	255	340	255	200	261	245	219
CRP			5,7			13,9			0,4			
Na	133,0	139,0	139,0	135,0	133	133	138	135	134	135	136,3	143
K (G)	4,0	4,3	8,7	4,7	5,5	5,7	6,1	5,8	5,9	5,7	7,13	3,6
K (C)	2,9	3,2	3,9	3,3	3,4	3,2	4,0	3,7	4,0	4,1		3
HbA1c												
Glisemi	71,0	70,0	70,0	77,0	74	76	81	80	72	93	104	73
Ure-BUN (G)	103,0	151,0	225,0	132,0	139	109	11	143	135	107	220	176
Ure-BUN (C)	35,0	57,0	84,0	56,0	56	46	21	52	54	79		77
Kreatinin (G)	6,2	10,2	9,8	5,9	6,5	5,8	5,6	5,1	4,9	5,0	5,5	5,5
Kreatinin (C)	2,6	4,8	4,0	3,1	3,1	2,9	2,9	2,6	2,5	2,7		3,1
Unik Asit			6,5			4,8			5,3			
URR	66,0	62,3	62,7	57,6	59,7	57,8	55,9	63,6	60	59,9		58,3
Kt / V	1,2	1,1	1,2	1,0	1,1	1,0	1,0	1,2	1,1	1,1		1,0
Ca	9,9	8,8	8,8	8,8	8,6	9,0	9,9	8,0	9,0	8,1		8,2
P	3,7	5,1	6,0	6,1	6,0	5,8	6,2	5,6	5,6	6,2		6,3
Ca X P	36,6	44,9	52,8	52,5	51,6	52,2	49,0	44,8	51	53,3		51,7
PTH			94,3			16,9			132,9			
Alk Phos.			57,0			266			74		256	
ALT	15,0	18,0	14,0	19,0	13	11	24	21	25	27	14	20
T Kolesterol						140						
HDL						32						
LDL						80						
Trigliserid						140						
Tot Prot.	7,4	7,4	8,9	8,6	8,0	8,5	7,5	7,2	7,7	7,6		7,1
Albumin	3,7	3,5	3,6	3,6	3,2	3,2	3,0	3,3	3,3	3,9		3,6
Ferritin			278,3			374			1034			
Fe			49,3			104			74,8			
Fe B K.			233,0			502			358			
TSH			21,2			20,7			20,9			
Venöz Bic.			20,6			27,2			22,8			
CLN: 307 (7.25) α: 21 (21.90)												
HEPATİT MARKERLERİ:												
HbsAg			3/08			6/08			9/08			
Hs Hbs			(-)			(-)			(-)			
Hs HCV			(-)			(-)			(-)			
Hs HIV			(+)			(+)			(+)			

Resim 2.6 Rutin Tetkik İzlem Formu Örneği

EK-3 Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.1 Hasta Yaşı ve Durumu çizelgesi

Hasta ID	Müracaat Tarihi	Doğum Tarihi	Ölüm/Ayrılış Tarihi	Yaş	Durumu
1	17.05.2003	01.03.1937	15.05.2009	72	EX
2	15.08.2007	25.08.1981	-	28	Tedavide
3	20.05.2002	01.04.1963	-	46	Tedavide
4	24.05.1999	18.08.1965	-	44	Tedavide
5	17.07.2007	01.01.1942	-	67	Tedavide
6	26.06.2002	21.05.1947	-	62	Tedavide
7	07.12.2007	15.02.1942	-	67	Tedavide
8	14.12.2007	01.01.1926	-	83	Tedavide
9	05.11.2007	10.09.1953	-	56	Tedavide
10	14.12.2005	01.07.1923	-	86	Tedavide
11	15.09.2006	12.05.1932	-	77	Tedavide
12	14.07.2004	05.11.1947	-	62	Tedavide
13	14.09.2006	20.03.1958	-	51	Tedavide
14	06.12.2004	01.01.1944	-	65	Tedavide
15	07.06.2004	01.07.1942	-	67	Tedavide
16	19.08.2002	01.01.1952	-	57	Tedavide
17	17.11.2006	01.02.1953	-	56	Tedavide
18	02.11.1996	02.06.1966	-	43	Tedavide
19	30.07.2003	01.01.1942	-	67	Tedavide
20	20.01.2007	01.01.1949	-	60	Tedavide
21	16.06.2003	01.01.1950	-	59	Tedavide
22	05.11.2007	04.06.1926	-	83	Tedavide
23	16.06.2006	01.01.1952	-	57	Tedavide
24	17.01.2007	01.01.1963	-	46	Tedavide
25	09.05.1998	01.01.1947	-	62	Tedavide
26	09.05.2008	22.02.1944	-	65	Tedavide
27	26.06.2002	01.01.1971	-	38	Tedavide
28	05.04.2008	15.01.1940	-	69	Tedavide
29	24.08.1999	01.01.1941	-	68	Tedavide
30	31.07.2007	20.08.1970	-	39	Tedavide
31	28.05.2007	17.04.1944	-	65	Tedavide
32	09.12.2003	17.12.1939	-	70	Tedavide
33	24.10.2007	01.01.1949	-	60	Tedavide
34	14.01.1999	21.04.1962	-	47	Tedavide
35	07.09.2006	01.01.1981	-	28	Tedavide
36	18.01.2007	14.09.1946	-	63	Tedavide
37	16.11.2002	01.01.1932	-	77	Tedavide
38	07.12.2007	26.03.1945	-	64	Tedavide
39	16.11.2006	01.07.1933	-	76	Tedavide
40	24.04.2007	07.02.1939	-	70	Tedavide
41	05.05.2007	01.01.1930	-	79	Tedavide
42	21.02.2008	01.07.1930	-	79	Tedavide
43	07.11.2006	01.01.1930	-	79	Tedavide
44	18.04.2008	15.07.1972	-	37	Tedavide
45	11.02.1998	06.07.1963	-	46	Tedavide
46	30.01.2002	04.05.1945	-	64	Tedavide
47	31.03.2003	10.08.1965	-	44	Tedavide
48	31.07.2006	07.05.1970	-	39	Tedavide
49	07.04.2004	01.01.1950	-	59	Tedavide
50	22.03.1996	23.11.1973	-	36	Tedavide
51	17.08.2005	26.07.1949	-	60	Tedavide
52	24.03.2008	01.01.1988	-	21	Tedavide
53	09.12.2006	01.01.1977	02.01.2009	32	Nakil
54	21.12.2007	01.12.1978	24.11.2008	30	Nakil

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.1 (Devam)Hasta Yaşı ve Durumu çizelgesi

Hasta ID	Müracaat Tarihi	Doğum Tarihi	Ölüm/Ayrılış Tarihi	Yaş	Durumu
55	11.11.2005	14.05.1964	19.09.2008	44	Nakil
56	29.01.1999	09.03.1946	06.12.2008	62	EX
57	18.01.2007	01.01.1967	07.03.2008	41	Transp.
58	02.02.2007	01.01.1969	16.02.2008	39	Transp.
59	24.03.2008	22.01.1960	02.02.2009	49	Nakil
60	23.03.2002	01.01.1938	06.06.2007	69	EX
61	08.07.1999	01.01.1946	28.11.2007	61	Nakil
62	30.09.2002	01.01.1932	02.02.2008	76	EX
63	19.06.2000	01.01.1931	15.09.2007	76	EX
64	22.02.2001	01.01.1933	02.04.2007	74	EX
65	07.10.1996	10.10.1944	18.07.2007	63	EX
66	16.12.2000	01.01.1935	04.04.2008	73	EX
67	04.12.2004	07.12.1930	03.04.2007	77	EX
68	22.08.2003	01.07.1938	25.04.2008	70	EX
69	28.12.1999	01.01.1942	01.07.2008	66	EX
70	12.11.2002	01.01.1932	01.06.2007	75	EX
71	01.12.2006	01.01.1940	21.04.2008	68	EX
72	23.08.2004	01.01.1934	05.05.2008	74	EX
73	21.06.2004	01.01.1946	07.06.2007	61	EX
74	29.06.2002	01.01.1936	01.04.2007	71	EX
75	05.12.2001	01.01.1935	06.10.2006	71	EX
76	28.06.2005	18.01.1940	28.07.2007	67	EX
77	01.10.2000	01.01.1944	27.02.2007	63	EX
78	20.04.1999	01.01.1951	10.03.2007	56	Nakil
79	30.01.2003	01.01.1970	25.08.2007	37	Nakil
80	14.01.2002	01.01.1945	28.10.2004	59	Nakil
81	12.10.2002	01.01.1935	02.10.2007	72	Nakil
82	01.02.2001	01.01.1973	01.05.2008	35	Nakil
83	13.12.2002	01.01.1962	18.05.2007	45	Transp.
84	30.11.2002	01.01.1968	04.07.2008	40	Transp.
85	28.07.2003	01.01.1926	20.03.2007	81	İyileşti
86	03.06.2006	01.01.1980	28.06.2008	28	Transp.
87	22.07.2002	13.05.1949	15.02.2006	57	Transp.
88	15.02.2000	01.01.1953	17.11.2007	54	Nakil
89	11.04.2005	01.01.1927	13.10.2007	80	EX
90	17.12.2003	01.01.1983	19.07.2005	22	Nakil
91	01.01.2000	01.01.1928	01.09.2004	76	EX
92	31.01.2000	01.01.1947	24.09.2005	58	Nakil
93	15.12.2005	01.01.1947	05.02.2007	60	Belirtilmemiş
94	03.11.1999	01.01.1975	08.05.2007	32	Nakil
95	05.04.2003	01.01.1936	31.03.2007	71	Nakil
96	23.03.2005	01.01.1974	01.05.2007	33	EX
97	29.08.2003	01.01.1938	20.03.2007	69	EX
98	13.11.1999	25.06.1949	24.08.2007	58	EX
99	01.04.2003	01.01.1960	30.08.2007	47	EX
100	22.12.2005	01.08.1950	05.03.2007	57	EX
101	29.03.2005	01.01.1960	08.03.2007	47	EX
102	23.11.2000	01.01.1934	01.12.2006	72	EX
103	17.01.1999	01.01.1927	21.04.2007	80	EX
104	25.10.2002	11.01.1943	10.05.2006	63	EX
105	25.12.1998	01.01.1940	20.05.2006	66	EX
106	18.04.2003	16.06.1955	-	54	Tedavide
107	02.11.1998	01.07.1938	-	71	Tedavide
108	16.03.2001	01.01.1949	08.07.2005	56	Nakil
109	18.05.2000	01.01.1933	04.03.2005	72	Nakil
110	17.01.2002	01.01.1929	06.04.2005	76	EX
111	29.01.2004	10.09.1984	22.03.2006	22	Nakil

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.1 (Devam) Hasta Yaşı ve Durumu çizelgesi

Hasta ID	Müracaat Tarihi	Doğum Tarihi	Ölüm/Ayrılış Tarihi	Yaş	Durumu
112	01.03.1997	02.02.1929	15.06.2003	74	EX
113	23.06.2004	01.01.1965	02.09.2005	40	Nakil
114	06.11.2007	20.09.1929	-	80	Tedavide
115	15.11.2002	01.01.1927	07.02.2004	77	Nakil
116	15.09.2003	01.01.1952	15.01.2005	53	Nakil
117	22.12.2001	27.03.1974	01.09.2003	29	Nakil
118	20.06.1998	01.01.1942	29.01.2003	61	EX
119	01.04.2003	01.01.1933	27.02.2004	71	EX
120	21.12.2002	06.04.1954	14.04.2004	50	EX
121	18.04.2000	01.01.1935	16.08.2003	68	Nakil
122	26.12.2003	01.01.1974	04.01.2005	31	Nakil
123	25.10.2001	01.01.1934	09.10.2004	70	Nakil
124	04.08.2000	01.01.1919	04.09.2004	85	EX
125	11.03.2003	22.05.1949	02.10.2005	56	EX
126	07.05.1998	01.01.1968	01.02.2005	37	EX
127	15.02.2001	01.01.1958	08.03.2004	46	Nakil
128	15.02.1999	01.01.1930	10.07.2004	74	EX
129	24.05.2002	01.01.1942	07.11.2005	63	EX
130	04.12.2001	01.01.1980	10.09.2003	23	Nakil
131	17.06.2005	01.01.1930	20.05.2006	76	Periton
132	08.06.1998	01.01.1942	22.06.2006	64	EX
133	17.03.2005	01.01.1940	16.09.2006	66	EX
134	27.03.1999	01.01.1945	10.09.2005	60	EX
135	07.12.2001	01.01.1935	30.09.2003	68	Nakil
136	10.05.2002	16.05.1941	08.12.2003	62	EX
137	03.09.1998	01.01.1941	11.11.2004	63	EX
138	29.05.2001	01.01.1935	31.07.2004	69	Nakil
139	05.02.2002	01.01.1931	02.04.2003	72	Belirtilmemiş
140	06.12.1997	01.01.1920	03.05.2005	85	EX
141	06.08.1998	01.01.1937	26.04.2005	68	EX
142	15.12.1998	01.01.1931	04.08.2003	72	EX
143	01.06.1999	01.01.1926	31.01.2007	81	EX
144	25.03.1996	01.01.1964	21.09.2004	40	Periton
145	02.04.1996	01.01.1949	21.06.2006	57	EX
146	25.04.2002	01.01.1940	22.10.2003	63	EX
147	18.11.2002	01.01.1954	01.06.2003	49	EX
148	08.06.2001	01.01.1930	03.01.2003	73	EX
149	09.03.1996	01.01.1946	06.04.2004	58	EX
150	16.03.2002	01.01.1945	09.10.2005	60	EX
151	02.11.1999	01.01.1948	01.01.2005	57	EX
152	16.03.2004	01.01.1947	18.06.2005	58	Nakil
153	01.05.2003	15.02.1956	06.04.2004	48	Nakil
154	22.06.2006	01.06.1982	31.01.2008	26	EX
155	20.03.2007	01.01.1925	05.02.2008	83	EX
156	23.06.1998	01.01.1917	10.10.2003	86	EX
157	31.12.2003	01.01.1940	11.10.2006	66	EX
158	30.01.2000	01.01.1955	26.03.2004	49	Transp.
159	15.03.2001	01.01.1941	31.08.2006	65	EX
160	27.10.2001	01.01.1972	24.09.2002	30	Nakil
161	01.05.2003	08.08.1983	19.03.2004	21	Transp.
162	20.11.1999	01.01.1959	18.06.2003	44	Nakil
163	16.05.1998	01.01.1948	06.03.2004	56	EX
164	20.11.2000	01.01.1938	15.12.2004	66	EX
165	28.01.2002	01.01.1932	26.05.2003	71	Belirtilmemiş
166	05.02.2000	01.01.1920	01.06.2005	85	EX
167	03.01.1999	01.01.1934	10.08.2003	69	EX
168	01.11.2002	30.04.1950	09.10.2004	54	Nakil
169	12.07.1999	01.01.1926	12.06.2004	78	EX
170	13.04.1999	01.01.1932	18.12.2003	71	EX

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.2 EPO Tedavisi ve Rezidiv idrar verileri hesaplama çizelgesi

Hasta ID	1				2				3				4				5			
Ay	EPO Tedavisi		Rezidiv İdrar Miktarı		EPO Tedavisi		Rezidiv İdrar Miktarı		EPO Tedavisi		Rezidiv İdrar Miktarı		EPO Tedavisi		Rezidiv İdrar Miktarı		EPO Tedavisi		Rezidiv İdrar Miktarı	
	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)
1	E	1000	100	9000	E	300	E	2000	E	4000	E	20	E	4000	E	20	E	4000	E	20
2	E	1000	100	4500	E	300	E	2000	E	4000	E	25	E	4000	E	25	E	4000	E	25
3	E	1000	100	4000	E	300	E	4000	E	4000	E	40	E	4000	E	40	E	4000	E	40
4	E	1000	100	4000	E	4000	E	4000	E	4000	E	20	E	4000	E	20	E	4000	E	20
5	E	1000	100	4000	E	4000	E	3000	E	4500	E	20	E	4500	E	20	E	4500	E	20
6	E	2000		3000	E	300	E		E	4000	E	20	E	4000	E	20	E	4000	E	20
7	E	4000		6000	E	300	E		E	4000	E	40	E	4000	E	40	E	4000	E	40
8	E	4000		8000	E		E		E	4000	E	40	E	4000	E	40	E	4000	E	40
9	E	4000		5000	E	300	E		E	4000	E	40	E	4000	E	40	E	4000	E	40
10	E	4000		5000	E	300	E		E	4000	E		E	4000	E		E	4000	E	
11	E	3000		5000	E		E		E	4000	E		E	4000	E		E	4000	E	
12	E	1500	100	4000	E		E		E	6000	E		E	6000	E		E	6000	E	
Ortalama	E	2292	100	5136	E	300	E	3000	E	4208	E	0	E	4208	E	0	E	4208	E	0

Hasta ID	51				52				53				54				55			
Ay	EPO Tedavisi		Rezidiv İdrar Miktarı		EPO Tedavisi		Rezidiv İdrar Miktarı		EPO Tedavisi		Rezidiv İdrar Miktarı		EPO Tedavisi		Rezidiv İdrar Miktarı		EPO Tedavisi		Rezidiv İdrar Miktarı	
	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)	İlaç Tipi (E/NR/AR)	Dozaj (ünite/hafıza)
1	E	8000		45	E	1500	E	2000	E	15000	E	1500	NR	15000	NR	1500	E	12000	E	12000
2	E	8000		45	E	1250	E	9000	E	15000	NR	1500	NR	15000	NR	1500	E	12000	E	12000
3				45	E	45	E	9000	E	5000	NR	450	NR	5000	NR	450	E	12000	E	12000
4	E	8000		45	E	45	E	9000	E	5000	NR	450	NR	5000	NR	450	E	12000	E	12000
5	E	8000		30	AR	30	1100	9000	E	1000	NR	500	NR	1000	NR	500	E	8000	E	8000
6	E	8000		30	AR	30	750	9000	NR	5000	NR	500	NR	5000	NR	500	E	8000	E	8000
7	NR	5000		50	AR	50	750	6000	E	12000	NR	200	E	12000	NR	200	NR	7500	NR	7500
8	E	4000		50	AR	50	500	6000	E	5000	NR	150	E	5000	NR	150	NR	6000	NR	6000
9	E	8000		40	AR	40	500	8000	E				NR				NR	6000	NR	6000
10	E	8000		40	AR	40	1250	8000	E								E	6000	E	6000
11	E	8000		1000	E			6000	E								E	12000	E	12000
12	E	9000			E				E								E		E	
Ortalama	E	7364	0	43	AR	955,6	0	6000	E	7556	NR	756,3	NR	7556	NR	756,3	E	9125	E	9125

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.3 Diyaliz seans verileri hesaplama çizelgesi

Hasta ID	TEDAVİ VERİLERİ (Diyaliz Seansları Ortalaması)										
	Seans No (Son 15)	Damar Yolu Tipi	Antikoagülasyon	Pre Diyaliz			Post Diyaliz			Kan Akım Hızı (QB)	Seans Süresi (saat)
				Ağırlık (kg)	AKB (mmHg)	Ağırlık (kg)	AKB (mmHg)				
1	1	LAVF	3500	55,30	170	70	53,20	120	70	250	4
	2	LAVF	3500	55,50	180	70	53,70	140	60	250	4
	3	LAVF	3500	55,10	140	80	53,30	100	70	250	4
	4	LAVF	3500	55,80	120	70	53,50	130	60	250	4
	5	LAVF	3500	54,90	140	70	53,25	140	70	250	4
	6	LAVF	3500	54,90	140	70	53,25	140	70	250	4
	7	LAVF	3500	54,90	150	60	53,20	140	80	250	4
	8	LAVF	3500	55,40	150	70	53,60	110	70	250	4
	9	LAVF	3500	55,00	150	60	53,60	120	70	250	4
	10	LAVF	3500	54,60	140	60	53,70	130	70	200	4
	11	LAVF	3500	55,50	130	80	53,40	120	70	200	4
	12	LAVF	3500	55,00	140	70	53,00	130	80	250	4
	13	LAVF	3500	54,20	140	60	53,35	120	60	250	4
	14	LAVF	3500	56,00	140	60	53,50	110	60	250	4
	15	LAVF	3500	54,50	150	80	53,30	130	70	250	4
	Ortalama	LAVF	3500	55,1	145,3	68,7	53,4	125,3	68,7	243,3	4
38	1	LAVF	2500	74,50	140	70	72,00	120	60	250	3,5
	2	LAVF	2500	73,50	130	70	71,50	120	60	250	3,5
	3	LAVF	2500	74,40	140	80	72,10	140	60	250	3,5
	4	LAVF	2500	74,70	140	70	72,40	140	70	250	3,5
	5	LAVF	2500	73,60	130	80	72,15	130	80	250	3,5
	6	LAVF	2500	73,30	150	80	72,00	140	80	250	3,5
	7	LAVF	2500	74,20	120	80	72,00	120	80	250	3,5
	8	LAVF	2500	74,20	120	70	71,70	150	60	250	3,5
	9	LAVF	2500	72,90	150	80	71,50	150	80	250	3,5
	10	LAVF	2500	72,50	140	70	71,40	100	60	250	3,5
	11	LAVF	2500	73,60	130	70	71,70	140	80	250	3,5
	12	LAVF	2500	73,70	140	80	71,60	100	60	250	3,5
	13	LAVF	2500	74,70	140	60	72,60	120	70	250	3,5
	14	LAVF	2500	73,50	170	80	72,00	140	60	250	3,5
	15	LAVF	2500	74,20	150	80	71,50	110	60	250	3,5
	Ortalama	LAVF	2500,0	73,8	139,3	74,7	71,9	128,0	68,0	250,0	3,5
64	1	RSVC	3500	67,00	130	70	65,20	130	70	200	3,5
	2	RSVC	3500	67,00	140	80	65,20	110	60	250	4
	3	RSVC	3500	65,90	120	70	64,50	110	70	250	4
	4	RSVC	3500	66,50	150	90	65,50	110	60	250	3,5
	5	RSVC	3500	68,00	140	80	66,30	12	60	250	3,5
	6	RSVC	3500	68,50	180	90	66,00	90	60	250	4
	7	RSVC	3500	66,30	160	90	64,50	100	60	250	3,5
	8	RSVC	3500	65,60	100	80	64,50	110	70	250	3,5
	9	RSVC	3500	66,90	120	70	65,50	90	60	200	3,5
	10	RSVC	3500	66,80	120	70	64,90	100	70	200	3,5
	11	RSVC	3500	66,40	90	60	65,50	80	50	200	3,5
	12	RSVC	3500	67,50	120	80	65,00	80	60	200	3,5
	13	RSVC	3500	67,00	110	60	65,70	80	60	200	3,5
	14	RSVC	3500	66,60	110	60	64,50	80	50	200	3,5
	15	RSVC	3500	67,20	100	60	64,50	80	60	200	3,5
	Ortalama	RSVC	3500	66,9	126,0	74,0	65,2	90,8	61,3	223,3	3,6

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.4 Ana veri çizelgesi formatı

Hasta ID	Yaş	Başlangıç Yaşı	Cinsiyet (E/K)	Kan Grubu	Boy (cm)	Kilo (kg)	Vücut Kütlesi (kg/m ²)	AKB (mmHg)	İdeal Kilo (kg)	Nüfuziyon İndeksi (NRI)	Sigara Kullanımı	Alkol Kullanımı	Primer Hastalık	Co-Morbid durumlar			Diyaliz				Haftalık Seans Süresi			
														HTN	KKY	D.M.	Yaş Sayısı (ay)	Diyaliz Seans Süresi (g)	Seans Süresi					
1	72	66	E	0	Rh(+)	160	45,6	20,86	130	70	57,5	44,5	Brakmış	yok	Polikistik BH				229	71	5,92	3	4	12
2	29	26	E	B	Rh(-)	182	71,2	21,27	140	80	74	46,0	Yok	yok	biyopsi teyisiz PGN				224	20	1,67	3	4	12
3	47	39	E	A	Rh(+)	173	64,5	22,39	180	100	67,3	47,6	Var	yok	biyopsi teyisiz PGN	var			884	111	9,25	3	4	12
4	45	34	K	0	Rh(+)	165	49	17,77	130	90	59	40,6	Yok	yok	biyopsi teyisiz PGN	var			1341	119	9,92	3	4	12
5	68	65	E	A	Rh(+)	180	67,7	21,87	140	90	72,5	46,0	Yok	yok	Renal Celsea Kanser				371	22	1,83	3	4	12
6	63	55	E	0	Rh(+)	165	68,5	23,44	120	80	61,3	49,3	Yok	yok	Diyabetik Nefropati	var	var		1013	188	15,67	3	4	12
7	68	65	E	0	Rh(+)	170	57	23,04	140	80	65	48,7	?	?	PKB				200	18	1,33	3	4	12
8	64	61	K	A	Rh(+)	150	44,5	19,62	120	70	50	43,6	Yok	yok	Hipertansif Nefropati	var		kanama eğ.	122	16	1,33	3	4	12
9	57	54	E	B	Rh(+)	167	62	24,43	180	80	62,8	51,4	?	?	Diyabetik Nefropati		var	Tip II görme yok	179	71	5,92	3	4	12
10	87	82	E	B	Rh(+)	185	62	16,21	130	70	76,3	36,2	Var	var	Hipertansif Nefropati	var			490	39	3,25	3	4	12
11	78	74	K	B	Rh(+)	155	48,8	19,84	120	80	53	42,6	Yok	yok	KKY				445	31	2,58	3	4	12
12	63	57	K	B	Rh(+)	160	71,9	27,90	180	80	56	59,2	Yok	yok	Diyabetik Nefropati		var	osteoporoz	726	56	4,67	3	4	12
13	52	48	E	A	Rh(+)	182	76	22,95	120	80	74	49,0	Yok	yok	biyopsi teyisiz PGN				378	40	3,33	3	4	12
14	66	60	E	A	Rh(+)	170	76	26,22	150	80	65	58,4	Yok	yok	Diyabetik Nefropati		var		643	52	4,33	3	4	12
15	68	62	E	0	Rh(+)	165	70,2	24,17	130	80	61,3	51,0	Yok	yok	Hipertansif Nefropati	var			698	58	4,83	3	4	12
16	58	50	K	A	Rh(+)	160	49	20,15	130	80	56	44,1	Yok	yok	KKY				467	82	6,83	3	4	12
17	57	53	K	AB	Rh(+)	160	66,6	27,50	110	60	56	58,5	var	yok	biyopsi teyisiz PGN				467	29	2,42	3	4	12
18	44	30	K	0	Rh(-)	159	59,5	24,99	110	70	55,4	53,2	Yok	yok	biyopsi teyisiz PGN				1324	164	13,67	3	4	12
19	68	61	E	0	Rh(+)	170	69	22,86	110	70	65	48,5	Var	yok	iskemik				882	69	5,75	3	4	12
20	61	58	K	0	Rh(+)	155	40	19,31	120	70	53	42,2	Yok	yok	Polikistik BH				334	207	17,25	3	4	12
21	60	53	E	B	Rh(+)	170	63	23,58	170	100	65	49,6	?	?	iskemik	Burger H			861	69	5,75	3	4	12
22	84	81	E	0	Rh(+)	160	58	23,13	200	90	57,5	48,6	?	?	Hipertansif Nefropati	var			214	17	1,42	3	4	12

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.4 (Devam) Ana veri çizelgesi formatı

Hasta ID	Damar Yolu Tipi	Damar Yolu Tipi	Antikoagülasyon	Pre Divaliz		Post Divaliz		Kan Akım Hız (Q/B)	İnter session Weight (kg)	Diyet		Transp. Durumu (U/UD)	EPO Tedavisi				Rezidiv İdar Miktar (12 aylık ort.)	Hb (Aylık)
				Ağırlık (kg)	AKB (mmHg)	Ağırlık (kg)	AKB (mmHg)			Stand.	Diyab.		İlaç Tipi (E/N/R/A/R)	Dozaj (12 aylık ort.)	Dozaj (12 aylık ort.)	(Ünitede) (kg)		
1	LAVF	0	3500	55,1	145,3	68,7	125,3	243,3	1,7	Var	yok	UD	E	2292	9167	172	15,433	11,13
2	LAVF	0	5933	72,8	130,7	82,0	130,0	258,0	2,4	Var	yok	U	E	5136	20545	252	25,943	11,24
3	LAVF	0	5000	69,7	130,0	81,3	118,7	250,0	2,7	Var	yok	UD	E	3000	12000	179	13,475	13,29
4	LAVF	0	3500	49,5	138,7	88,0	152,0	250,0	1,1	Var	yok	U	E	4208	16833	348	31,46	11,06
5	LSVC	2	5000	72,1	102,7	69,3	93,3	250,0	1,3	Var	yok	UD	AR	29	23556	332	26,869	11,52
6	LAVF	0	3000	66,6	130,0	77,3	110,7	203,3	2,8	Var	var	UD	AR	40	32000	501	50,267	9,98
7	RAVF	1	3500	68,3	108,7	70,0	62,7	250,0	1,7	Var	yok	UD	E	3813	15250	229	21	10,92
8	LAVF	0	4500	46,5	135,3	80,7	128,7	250,0	1,9	Var	yok	UD	E	3625	14500	325	27,271	11,93
9	LAVF	0	3500	70,5	163,3	78,7	128,7	250,0	2,3	Var	var	U	E	1600	6400	94	8,014	11,73
10	LSVC	2	5000	56,2	123,3	64,0	123,3	250,0	0,7	Var	yok	UD	E	5682	22727	410	37,718	10,86
11	RLVC	9	3500	49,0	107,3	69,3	108,0	250,0	1,3	Var	yok	UD	E	5167	20667	433	40,6	10,67
12	RFVC	7	5000	73,3	117,5	74,0	118,7	250,0	1,8	Var	var	UD	E	9333	37333	523	51,3	10,18
13	LAVF	0	4667	77,6	126,7	81,3	120,0	271,7	1,6	Var	yok	U	NR	4333	17333	228	19,5	11,70
14	RAVF	1	5000	84,3	130,7	77,3	124,7	250,0	2,7	Var	var	U	E	4208	16833	206	18,1	11,38
15	LAVF	0	3500	67,3	128,0	76,0	128,0	250,0	1,5	Var	yok	UD	E	2000	8000	122	10,5	11,54
16	LAVF	0	2000	53,2	132,0	72,7	132,0	250,0	1,6	Var	yok	?	AR	40	32000	620	69,7	8,90
17	LAVF	0	3500	71,2	114,3	67,9	105,0	250,0	0,8	Var	yok	?	E	4167	16667	237	21,1	11,21
18	LAVF	0	7000	65,1	103,3	62,0	89,3	250,0	2,0	Var	yok	?	E	2250	9000	142	10,3	13,83
19	RAVF	1	5000	68,5	136,7	76,0	118,7	250,0	2,4	Var	yok	U	E	3571	14286	216	19,3	11,18
20	RLVC	9	3500	47,6	113,3	70,7	110,7	250,0	1,2	Var	yok	UD	E	5333	21333	480	42,8	10,73
21	LAVF	0	5000	70,1	134,0	82,0	102,0	250,0	2,0	Var	yok	U	E	1667	6667	98	8,1	12,09
22	LAVF	0	3500	59,9	156,0	80,7	136,0	250,0	0,7	Var	yok	UD	E	3167	12667	214	18,5	11,59

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.4 (Devam) Ana veri çizelgesi formatı

Hasta ID	Htc (Aylık)	MCV	MCH	MCHC	Lökosit	Platelet	CRP (3 ayda bir)	Na	K(G)	K(Ç)	K Farkı	Glisemi	Üre-BUN (G)	Üre-BUN (Ç)	BUN Ratio	Kreatinin(G)	Kreatinin(Ç)	Unik Asit (3 ayda bir)	URR	kV	Ca
1	33,23	99,03	33,35	33,93	4,20	155,17	0,25	137,75	5,55	3,73	1,82	123,33	143,67	49,06	2,93	9,67	3,97	5,90	66,03	1,23	10,23
2	32,74	87,77	30,01	34,53	9,40	175,75	0,46	136,92	5,01	3,70	1,31	99,33	190,33	69,00	2,76	10,77	4,85	6,68	64,04	1,19	9,05
3	39,54	93,12	31,39	34,17	7,19	162,17	0,17	136,67	5,83	3,93	1,89	83,92	167,06	60,17	2,76	11,48	5,43	6,56	64,15	1,19	8,75
4	32,29	92,31	31,42	34,58	7,49	169,92	0,13	136,83	5,10	3,45	1,65	97,08	157,50	37,92	4,15	8,39	2,86	7,95	76,04	1,48	9,42
5	35,29	90,70	29,52	32,75	9,14	255,25	5,12	136,67	5,23	3,49	1,73	77,75	159,67	63,17	2,53	6,29	3,12	5,80	60,53	1,11	8,72
6	28,74	91,08	31,72	35,39	5,46	93,00	2,19	134,92	4,43	2,99	1,44	108,00	188,08	75,33	2,50	9,49	4,35	9,40	60,09	1,10	9,00
7	32,08	90,26	30,74	34,35	5,98	156,92	0,19	136,00	5,28	3,63	1,65	110,92	165,50	56,42	2,93	11,15	4,57	6,60	65,80	1,23	8,82
8	35,21	86,72	29,29	34,23	16,14	231,25	1,35	137,41	4,50	3,16	1,34	120,33	140,92	37,17	3,79	5,50	2,00	6,30	74,00	1,40	9,21
9	32,83	98,80	35,39	35,89	7,80	102,83	1,91	137,17	4,63	3,33	1,30	122,25	148,83	54,33	2,74	9,72	4,28	5,85	66,70	1,20	8,66
10	32,70	88,71	29,65	33,93	6,95	136,00	0,60	136,58	4,76	3,49	1,29	102,58	127,92	52,83	2,42	6,68	3,17	5,76	58,62	1,06	8,92
11	32,14	85,70	28,66	33,63	7,17	185,83	1,30	135,25	4,83	3,46	1,37	96,17	136,58	36,33	3,81	6,09	2,22	5,63	73,80	1,41	8,66
12	31,59	85,15	27,13	32,03	6,63	223,08	0,50	135,17	5,36	3,92	1,47	215,67	129,08	51,77	2,49	6,81	3,41	5,55	60,35	1,10	9,31
13	33,91	92,03	31,91	34,68	5,47	174,25	0,10	139,50	5,11	3,44	1,67	78,67	129,91	51,58	2,52	7,53	3,40	6,85	60,16	1,10	9,35
14	33,51	85,73	29,11	34,26	8,18	184,00	0,21	134,83	5,77	3,98	1,78	258,50	155,25	59,75	2,60	6,94	3,38	6,85	61,64	1,13	9,13
15	32,91	86,55	29,99	34,82	5,55	200,58	0,27	136,83	5,22	3,63	1,58	97,25	176,83	71,00	2,49	6,18	2,99	5,60	60,48	1,11	8,71
16	26,14	90,98	30,82	34,31	3,76	98,50	0,39	135,83	4,88	3,40	1,48	119,08	158,75	49,25	3,22	6,04	2,43	5,90	68,87	1,30	10,29
17	32,03	86,85	30,37	35,28	6,23	206,75	0,39	137,83	5,05	3,42	1,63	99,17	132,14	49,35	2,66	6,42	3,54	5,05	62,06	1,14	9,44
18	40,80	91,16	30,73	34,31	8,01	277,33	0,40	134,25	5,73	3,87	1,86	93,00	200,05	67,58	2,96	9,25	4,11	7,23	66,15	1,24	9,16
19	34,63	89,14	22,29	32,38	10,60	259,75	0,54	140,17	5,35	3,69	1,66	82,17	185,67	72,50	2,56	11,21	5,28	6,50	60,67	1,11	9,38
20	31,91	93,99	31,63	34,14	5,04	185,92	0,17	139,83	5,13	3,53	1,60	80,42	126,58	36,56	3,46	7,43	2,90	5,58	70,88	1,35	8,96
21	35,85	85,09	28,73	33,83	6,17	220,17	1,19	135,25	5,31	3,80	1,51	78,92	179,92	62,67	2,87	11,08	5,05	7,43	65,40	1,22	8,73
22	34,99	89,42	29,63	33,37	5,52	218,75	0,30	137,58	5,78	4,16	1,62	87,25	207,08	76,67	2,70	7,43	3,52	6,68	62,52	1,15	8,81

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.4 (Devam) Ana veri çizelgesi formatı

Hasta ID	P	Ca X P	PTH (3 ayda bir)	Alk. Phos.(3 ayda bir)	ALT	T Koiles t (6 ayda bir)	HDL (6 ayda bir)	LDL (6 ayda bir)	Trigliser id (6 ayda bir)	TotProt	Albumin	Ferritin (3 ayda bir)	Fe (3 ayda bir)	FeB.K. (3 ayda bir)	Tsl (3 ayda bir)	Venöz Bic. (3 ayda bir)
1	5,04	60,72	46,91	47,00	13,00	182,00	39,00	83,90	295,50	6,63	3,87	596,85	53,13	255,00	20,86	25,80
2	4,69	42,32	32,58	47,75	15,50	173,00	36,50	69,90	333,00	6,72	4,22	286,08	47,43	273,00	18,28	24,98
3	5,33	46,53	40,07	63,25	32,33	168,00	39,50	75,80	263,50	6,98	4,06	604,43	99,60	267,00	34,91	22,58
4	4,39	41,22	136,10	58,50	14,50	178,50	42,50	82,90	265,50	6,89	4,28	374,58	53,43	288,50	18,71	25,90
5	5,74	49,81	74,46	112,00	19,08	158,00	35,00	71,20	259,00	7,88	3,53	236,95	32,20	348,50	10,23	23,33
6	5,52	49,41	191,25	62,88	20,75	143,00	32,00	50,10	304,50	7,11	3,96	2014,80	109,30	243,00	46,30	19,73
7	3,44	30,36	112,48	60,98	23,25	176,50	35,00	75,30	331,00	6,75	3,99	429,03	62,95	273,00	23,04	25,43
8	4,63	42,43	43,04	60,50	12,58	167,00	47,00	68,90	255,50	7,22	4,29	162,90	48,50	311,25	16,04	21,90
9	5,27	45,18	18,53	82,50	19,00	164,50	39,00	70,90	273,00	6,47	4,13	1105,25	97,83	286,25	34,53	22,78
10	4,12	36,65	155,01	60,50	14,25	185,00	50,50	85,80	243,50	6,43	3,90	119,46	48,20	284,00	16,48	23,55
11	4,28	37,18	111,43	136,75	10,58	185,00	37,00	95,60	262,00	6,72	3,45	1192,52	38,10	272,75	15,05	26,90
12	3,83	35,56	25,34	61,25	12,50	176,50	42,50	84,60	247,00	7,21	4,03	232,53	38,03	286,75	13,49	26,33
13	3,60	33,82	205,33	48,75	26,00	163,50	35,50	66,70	306,50	6,88	4,14	438,80	66,75	276,75	24,03	25,68
14	4,51	41,03	209,98	60,75	17,08	151,00	34,00	50,80	331,00	6,75	4,08	513,98	51,18	276,75	18,54	23,35
15	3,53	30,76	86,57	51,75	17,25	167,00	39,50	77,10	252,00	6,89	4,18	375,70	44,13	280,00	16,42	25,88
16	5,53	57,10	193,00	55,50	12,25	161,00	44,50	72,10	222,00	7,53	3,78	608,30	51,78	275,75	18,74	22,73
17	5,32	50,08	46,26	52,75	11,58	192,00	43,00	92,60	282,00	6,88	4,09	353,60	64,80	302,00	21,78	25,45
18	5,48	50,23	147,69	83,25	15,00	165,50	35,50	76,40	268,00	6,96	3,81	118,65	29,70	296,75	10,32	23,13
19	5,30	49,65	381,20	91,25	18,67	163,50	40,00	73,00	252,50	7,12	4,12	599,88	35,15	240,50	15,34	22,95
20	4,21	37,82	50,17	50,00	12,08	173,00	41,50	85,30	231,00	7,16	3,84	812,55	72,33	303,50	25,25	23,45
21	5,18	45,08	196,68	100,50	13,08	186,50	34,50	94,40	288,00	6,73	3,98	858,35	67,08	259,75	25,79	22,63
22	5,73	50,52	114,10	56,38	14,17	161,00	39,00	92,80	246,00	7,07	3,83	453,65	61,63	227,25	27,54	24,40

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.5 Periyodik test değerleri hesap çizelgesi

CAN DİYALİZ		HASTA ID : 82													
Ölçüm Periyodu (ay)	Ölçülen Değer	Birimi	1	2	3	4	5	6	7	8	9	10	11	12	Ort.
1	Hb	g/dl	12,3	14,3	12	13,8	13,9	13,6	13,4	13,6	12,1	14,1	13,2	12,2	13,2
1	Htc	%	36,8	44	34,7	42,1	43,6	44,1	41,6	45,1	32,6	43,9	41,6	37,9	40,7
1	MCV	fL	85,4	85,5	92,2	93,3	94,4	95,7	97,6	101	103	104	102	98,1	96,0
1	MCH	pg	28,5	29,1	29,2	30,6	30	29,4	31,5	30,3	31,5	33,4	32,3	33,3	30,8
1	MCHC	g/dL	33,4	32,5	31,7	30,6	32	30,7	32,3	30,2	30,6	32	31,8	32,1	31,7
1	Lökosit	UL	5,2	7,3	6,2	6,9	6,9	7,5	6,8	6,3	6,4	6,7	9,2	8,9	7,0
1	Trombosit	UL	221	207	203	213	193	189	175	206	176	188	231	198	200,0
3	CRP	mg/l			0,3			0,7			0,23			0,2	0,4
1	Na	mEq/dl	141	135	135	133	140	140	136	135	136	133	133	133	135,8
1	K (G)	mEq/dl	4,6	5	5	5,3	5,7	4,7	4,7	5	5,1	5,5	5,9	5,6	5,2
1	K (Ç)	mEq/dl	4,1	3,1	3,5	3,6	4,3	3,5	3,6	3,6	3,7	3,9	3,8	3,8	3,7
3	HbA1c	%													
1	Glisemi	mg/dl	117	70	91	98	82	77	78	74	72	80	74	91	83,7
1	Üre-BUN (G)	mg/dl	96,5	132	143	151	143	96	102	101	98	142	144	144	124,4
1	Üre-BUN (Ç)	mg/dl	22	42	52	61	64	31	32,7	36	33	52	50	45	43,4
1	Kreatinin (G)	mg/dl	8	10,1	10,5	11,9	13,7	10,6	12,6	12,4	12,4	9,6	10,1	10,6	11,0
1	Kreatinin (Ç)	mg/dl	2,6	4,4	5,1	6,1	7,6	4,8	6	5,9	6,1	4,6	4,8	4,4	5,2
3	Ürik Asit	mg/dl			7,6			9,4			5,8			5,5	7,1
1	URR		77,2	68,2	63,6	59,6	55,2	67,7	67,9	64,4	66,3	63,4	65,3	68,8	65,6
1	Kt / V		1,5	1,3	1,2	1,1	1,0	1,3	1,3	1,2	1,2	1,2	1,2	1,3	1,2
1	Ca	mg/l	9,5	8,3	9,5	9,7	9,9	10,2	9,6	9	9,6	9,2	9,4	9,8	9,5
1	P	mg/l	4,8	5,3	6	5,9	6,4	5	5	5,5	5,3	4	4,1	4,2	5,1
1	Ca X P		45,6	44,0	57,0	57,2	63,4	51,0	48,0	49,5	50,9	36,8	38,5	41,2	48,6
6	PTH	pg/ml						1900						967	1433,4
6	Alk. Fos.	U/L						544						310	427,0
1	ALT	U/L	65	29	51	108	54	37	23	25	29	78	43	47	49,1
6	T.Kolest.	mg/dl						112						101	106,5
6	HDL	mg/dl						20						32	26,0
6	LDL	mg/dl						31						49	40,0
6	Trigliserid	mg/dl						305						100	202,5
1	Tot.Prot.	g/dl	6,6	6,5	6,8	7,5	7,5	6,8	7	7,2	7	6,7	6,5	6,8	6,9
1	Albumin	g/dl	3,7	3,7	3,8	4,1	4,1	3,8	3,8	3,9	3,9	3,6	3,8	3,9	3,8
3	Ferritin	mg/dl			387			138			58,6			94,5	169,4
3	Fe	ug/dl			51,4			35,5			52,5			88,3	56,9
3	Fe B.K.	ug/dl			360			377			403			267	351,8
3	TSI	%			14,3			9,42			13			33,1	17,4
3	Venöz Bikar.				14,3			9,4			21			23,7	17,1

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.6 Kullanılan 12 aylık test verilerinin periyotları

HASTA ID	Test Verileri Kullanılan Ay ve Yıllar		HASTA ID	Test Verileri Kullanılan Ay ve Yıllar	
1	1-12 2008		49	1-12 2008	
2	1-12 2008		50	1-12 2008	
3	1-12 2008		51	1-12 2008	
4	1-12 2008		52	3-12 2008	
5	1-12 2008		53	1-12 2008	
6	1-12 2008		54	12 2007	1-11 2008
7	1-12 2008		55	9-12 2007	1-9 2008 (-6)
8	1-12 2008		56	11-12 2007	1-10 2008
9	1-12 2008		57	3-12 2007	1-2 2008
10	1-12 2008		58	3-12 2007	1-2 2008
11	1-12 2008		59	3-12 2007	1 2008
12	1-12 2008		60	6-12 2006	1-5 2007
13	1-12 2008		61	12 2006	1-11 2007
14	1-12 2008		62	1-12 2007	
15	1-12 2008		63	9-12 2006	1-8 2007
16	1-12 2008		64	4-12 2006	1-3 2007
17	1-12 2008		65	8-12 2006	1-7 2007
18	1-12 2008		66	1-12 2007	
19	1-12 2008		67	4-12 2006	1-3 2007
20	1-12 2008		68	4-12 2007	1-3 2008
21	1-12 2008		69	6-12 2007	1-6 2008
22	1-12 2008		70	7-12 2006	1-5 2007
23	1-12 2008		71	2-12 2007 (-6,7,8)	1-4 2008
24	1-12 2008		72	6-12 2007	1-5 2008
25	1-12 2008		73	5-12 2006 (-6)	1-5 2007
26	5-12 2008	1-4 2009	74	4-12 2006	1-3 2007
27	1-12 2008		75	9-12 2005	1-8 2006
28	4-12 2008	1-3 2009	76	8-12 2006	1-7 2007
29	1-12 2008		77	1-12 2006	
30	1-12 2008		78	2-12 2006 (-7,8)	1-3 2007
31	1-12 2008		79	9-12 2006	1-8 2007
32	1-12 2008		80	11-12 2003	1-10 2004
33	1-12 2008		81	9-12 2006	1-8 2007
34	1-12 2008		82	1-12 2007	
35	1-12 2008		83	5-12 2006	1-5 2007
36	1-12 2008		84	8-12 2007	1-7 2008
37	1-12 2008		85	3-12 2006	1-2 2007
38	1-12 2008		86	1-12 2007	
39	1-12 2008		87	3-12 2005	1-2 2006
40	1-12 2008		88	9-12 2006	1-8 2007
41	1-12 2008		89	10-12 2006	1-9 2007
42	2-12 2008	1 2009	90	5-12 2007	1-5 2008
43	1-12 2008		91	1-12 2003	
44	4-12 2008	1-4 2009	92	10-12 2004	1-9 2005
45	1-12 2008		93	5-12 2006	1-4 2007
46	1-12 2008		94	6-12 2006	1-5 2007
47	1-12 2008		95	4-12 2006	1-3 2007
48	1-12 2008		96	5-12 2006	1-4 2007

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.6 (Devam) Kullanılan 12 aylık test verilerinin periyotları

HASTA ID	Test Verileri Kullanılan Ay ve Yıllar		HASTA ID	Test Verileri Kullanılan Ay ve Yıllar	
97	4-12 2006	1-3 2007	134	8-12 2004	1-7 2005
98	8-12 2006	1-7 2007	135	10-12 2002	1-9 2003
99	9-12 2006	1-8 2007	136	1-12 2003	
100	3-12 2006 (-8)	1-3 2007	137	7-10 2003	4-11 2004
101	4-12 2006	1-3 2007	138	1-12 2003	
102	9-12 2005	1-8 2006	139	3-12 2002	1,3 2003
103	5-12 2006	1-4 2007	140	1-12 2003 (-4,12)	10-11 2004
104	4-12 2005	1-3 2006	141	5-12 2004	1-4 2005
105	1-12 2005		142	7-12 2002	1-6 2003
106	1-12 2008		143	1-12 2006	
107	1-12 2008		144	10-12 2003	1-9 2004
108	1-12 2003		145	7-12 2005	1-6 2006
109	1-7 2004	1-5 2005	146	9-12 2002	1-8 2003
110	5-12 2004	1-4 2005	147	12 2002	1-6 2003
111	1-12 2005		148	1-12 2002	
112	4-12 2002 (-9)	2-5 2003	149	1-12 2003 (-6)	1 2004
113	10-12 2004	1-9 2005	150	10-12 2004	1-9 2005
114	1-12 2001		151	10-12 2004	1-10 2005 (-7)
115	1-12 2002 (-4,7,11)		152	7-12 2004	1-6 2005
116	1-12 2004		153	5-12 2003	1-4 2004
117	9-12 2002	1-8 2003	154	10-12 2006	3-11 2007
118	12 2002, 1 2003		155	4-12 2007	1-3 2008
119	3-12 2003	1-2 2004	156	1-8 2003 (-4)	12 2002
120	6-12 2003	1-6 2004	157	10-12 2005	1-10 2006 (-2)
121	9-12 2002	1-8 2003	158	1-12 2003	
122	1-12 2004		159	8-12 2005	1-8 2006 (-2)
123	11-12 2003	1-10 2004	160	7-12 2002	1-6 2003
124	8-12 2003	1-8 2004	161	4-12 2003	1-3 2004
125	10-12 2004	1-9 2005	162	6-12 2002	1-5 2003
126	2-12 2004	1 2005	163	1-12 2003	
127	4-12 2002	1-3 2003	164	9-12 2003	1-7 2004
128	8-12 2003	1-7 2004	165	1-12 2002	
129	9-12 2004	1-10 2005 (-6,8)	166	1-12 2004	
130	1-12 2002		167	9-12 2002	1-7 2003
131	6-12 2005	1-6 2006 (-5)	168	11-12 2003	1-10 2004
132	7-12 2005	1-6 2006	169	7-12 2003	1-6 2004
133	9-12 2005	1-8 2006	170	1-12 2002	

Not : Parantez içinde - işareti yanında yer alan numara ilgili periyotta verisine ulaşamayan ayı simgeler. Örneğin 1-12 2003 periyodu yanında yer alan (-4,12) ifadesi 2003 yılı için 12 aylık periyotta 4. ve 12. ayların verilerinin kullanılmadığını, diğer 10 ayın kullanıldığını göstermektedir.

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.7 Uygulamada kullanılacak veri kümesinin kısmi gösterimi

ID	AGE	SAGE	SEX	HEI	WEI	VKG	SS_AKB	SD_AKB	NRI	SFS	SEAG	DHS	HSS	DYT	AKG	Pre-W	Pre-S	Pre-D	Ps-W
1	72	66	1	160	45,6	20,9	130	70	44,5	229	71	5,9	12	0	3500	55,1	145	69	53,4
2	28	26	1	162	71,2	21,3	140	80	46,0	224	20	1,7	12	0	5933	72,8	131	82	70,4
3	46	39	1	173	64,5	22,4	180	100	47,6	884	111	9,3	12	0	5000	69,7	130	81	67,0
4	44	34	0	165	49	17,8	130	90	40,6	1341	119	9,9	12	0	3500	49,5	137	86	48,4
5	67	65	1	180	67,7	21,9	140	90	46,0	371	22	1,8	12	2	5000	72,1	103	69	70,9
6	62	55	1	165	68,5	23,4	120	80	49,3	1013	188	15,7	12	0	3000	66,6	130	77	63,8
7	67	65	1	170	57	23,0	140	80	48,7	200	16	1,3	12	1	3500	68,3	109	70	66,6
8	83	81	0	150	44,5	19,8	120	70	43,6	122	16	1,3	12	0	4500	46,5	135	81	44,6
9	56	54	1	167	62	24,4	180	80	51,4	179	71	5,9	12	0	3500	70,5	163	79	68,1
10	86	82	1	185	62	16,2	130	70	36,2	490	39	3,3	12	2	5000	56,2	123	64	55,5
11	77	74	0	155	48,8	19,8	120	80	42,6	445	31	2,6	12	5	3500	49,0	107	69	47,7
12	62	57	0	160	71,9	27,9	180	80	59,2	726	56	4,7	12	7	5000	73,3	117	74	71,4
13	51	48	1	182	76	22,9	120	80	49,0	378	40	3,3	12	0	4867	77,6	127	81	76,0
14	65	60	1	170	78	28,2	150	80	58,4	643	52	4,3	12	1	5000	84,3	131	77	81,6
15	67	62	1	165	70,2	24,2	130	80	51,0	698	58	4,8	12	0	3500	67,3	128	76	65,8
16	57	50	0	160	49	20,2	130	80	44,1	457	82	6,8	12	0	2000	53,2	132	73	51,6
17	56	53	0	160	66,6	27,5	110	60	58,5	467	29	2,4	12	0	3500	71,2	114	68	70,4
18	43	30	0	159	59,5	25,0	110	70	53,2	1324	164	13,7	12	0	7000	65,1	103	62	63,2
19	67	61	1	170	69	22,9	110	70	48,5	882	69	5,8	12	1	5000	68,5	137	76	66,1
20	60	58	0	155	40	19,3	120	70	42,2	334	207	17,3	12	5	3500	47,6	113	71	46,4
21	59	53	1	170	63	23,6	170	100	49,6	861	69	5,8	12	0	5000	70,1	134	82	68,2
22	83	81	1	160	58	23,1	200	90	48,6	214	17	1,4	12	0	3500	59,9	156	81	59,2
23	57	54	1	175	60	20,2	150	90	43,4	432	34	2,8	12	5	5000	63,8	143	86	61,9
24	46	44	1	163	61	23,7	150	70	49,5	292	26	2,2	12	5	3500	64,4	132	81	63,0
25	62	51	1	172	59,5	18,6	110	80	39,5	1531	164	13,7	12	0	5000	57,5	123	73	55,1
26	65	64	1	160	84,7	33,0	130	70	67,6	137	12	1,0	12	0	5000	87,1	132	85	84,5
27	38	31	1	165	69,3	23,6	140	80	49,6	1048	81	6,8	12	0	3500	67,2	151	91	64,4
28	69	68	1	165	73,5	26,8	160	100	55,6	147	12	1,0	12	0	3500	73,9	142	83	72,9

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.7 (Devam) Uygulamada kullanılacak veri kümesinin kısmi gösterimi

ID	Ps_S	Ps_D	ISW	KAH	EIT	EPOD	EPO_U	EPOR	RZI	Hic	CRP	Na	K_J	K_O	K_F	GLI	BUN_I	BUN_O
1	125	89	1,7	243	0	9167	172	15,4	100	33,23	0,25	137,75	5,55	3,73	1,82	123,33	143,67	49,08
2	130	85	2,4	258	0	20545	292	25,9	300	32,74	0,46	136,92	5,01	3,70	1,31	99,33	190,33	69,00
3	119	75	2,7	250	0	12000	179	13,5	0	39,54	0,17	138,67	5,83	3,93	1,89	83,92	167,08	60,17
4	152	93	1,1	250	0	16833	348	31,5	0	32,29	0,13	136,83	5,10	3,45	1,65	97,08	157,50	37,92
5	93	81	1,3	250	2	23558	332	28,9	0	35,29	5,12	136,67	5,23	3,49	1,73	77,75	159,67	63,17
6	111	68	2,8	203	2	32000	501	50,3	0	28,74	2,19	134,92	4,43	2,99	1,44	108,00	188,08	75,33
7	96	63	1,7	250	0	15250	229	21,0	250	32,08	0,19	136,00	5,28	3,63	1,65	110,92	185,50	56,42
8	129	73	1,9	250	0	14500	325	27,3	2000	35,21	1,35	137,41	4,50	3,16	1,34	120,33	140,92	37,17
9	129	66	2,3	250	0	6400	94	8,0	160	32,83	1,91	137,17	4,63	3,33	1,30	122,25	148,83	54,33
10	123	72	0,7	250	0	22727	410	37,7	75	32,70	0,60	138,58	4,78	3,49	1,29	102,58	127,92	52,83
11	108	68	1,3	250	0	20667	433	40,6	300	32,14	1,30	135,25	4,83	3,46	1,37	96,17	138,58	36,33
12	119	73	1,8	250	0	37333	523	51,3	100	31,59	0,50	135,17	5,38	3,92	1,47	215,67	129,08	51,77
13	120	79	1,6	272	1	17333	228	19,5	1375	33,91	0,10	139,50	5,11	3,44	1,67	78,67	129,91	51,58
14	125	72	2,7	250	0	16833	206	18,1	415	33,51	0,21	134,83	5,77	3,98	1,78	258,50	155,25	59,75
15	128	75	1,5	250	0	8000	122	10,5	1020	32,91	0,27	136,83	5,22	3,63	1,58	97,25	176,83	71,00
16	132	77	1,6	250	2	32000	620	69,7	100	26,14	0,39	135,83	4,88	3,40	1,48	119,08	158,75	49,25
17	105	65	0,8	250	0	16667	237	21,1	427	32,03	0,39	137,83	5,05	3,42	1,63	99,17	132,14	49,35
18	89	58	2,0	250	0	9000	142	10,3	0	40,80	0,40	134,25	5,73	3,87	1,86	93,00	200,05	67,58
19	119	67	2,4	250	0	14286	216	19,3	821	34,63	0,54	140,17	5,35	3,69	1,66	82,17	185,67	72,50
20	111	71	1,2	250	0	21333	460	42,8	0	31,91	0,17	139,83	5,13	3,53	1,60	80,42	126,58	36,58
21	102	66	2,0	250	0	6867	98	8,1	200	35,85	1,19	135,25	5,31	3,80	1,51	78,92	179,92	62,67
22	136	75	0,7	250	0	12667	214	18,5	608	34,99	0,30	137,58	5,78	4,16	1,62	87,25	207,08	76,67
23	113	77	1,9	250	0	17167	278	24,7	575	33,08	0,45	140,17	4,93	3,83	1,10	171,25	137,25	58,58
24	116	71	1,4	250	0	7273	115	9,9	350	34,09	0,32	142,25	4,40	3,33	1,07	140,33	127,42	48,75
25	99	66	2,4	250	0	4000	73	5,1	0	43,27	0,35	138,50	5,52	4,14	1,38	82,42	183,83	66,75
26	117	74	2,5	250	0	36800	435	41,0	800	31,80	0,64	136,33	5,47	3,92	1,55	82,67	144,25	62,86
27	145	91	2,9	250	0	16833	261	23,4	0	32,52	0,86	136,75	5,67	3,73	1,94	79,50	186,92	71,56
28	126	73	1,0	250	1	20909	287	27,7	308	31,16	0,27	138,33	4,93	3,73	1,20	125,17	135,17	50,25

EK-3 (Devam) Veri kümesini oluşturmada kullanılan çizelgeler

Çizelge 3.7 (Devam) Uygulamada kullanılacak veri kümesinin kısmi gösterimi

ID	BUN_R	KRE_I	KRE_O	URA	URR	KIV	CaXP	PTH	TKoll	HDL	LDL	TGL	TPRO	ALB	FERT	TSI	RSORE
1	2,93	9,67	3,97	5,90	66,03	1,23	60,72	46,91	182,00	39,00	83,90	295,50	6,63	3,87	596,85	20,86	0,776
2	2,76	10,77	4,85	6,68	64,04	1,19	42,32	32,58	173,00	36,50	69,90	333,00	6,72	4,22	286,08	18,28	0,614
3	2,78	11,48	5,43	6,58	64,15	1,19	46,53	40,07	168,00	39,50	75,80	263,50	6,98	4,06	604,43	34,91	0,752
4	4,15	8,39	2,86	7,95	76,04	1,46	41,22	136,10	178,50	42,50	82,90	265,50	6,69	4,28	374,58	18,71	0,862
5	2,53	6,29	3,12	5,80	60,53	1,11	49,81	74,46	158,00	35,00	71,20	259,00	7,88	3,53	236,95	10,23	0,776
6	2,50	9,49	4,35	9,40	60,09	1,10	49,41	191,25	143,00	32,00	50,10	304,50	7,11	3,96	2014,80	46,30	0,879
7	2,93	11,15	4,57	6,60	65,80	1,23	30,36	112,48	176,50	35,00	75,30	331,00	6,75	3,99	429,03	23,04	0,776
8	3,79	5,50	2,00	6,30	74,00	1,40	42,43	43,04	167,00	47,00	68,90	255,50	7,22	4,29	162,90	16,04	0,927
9	2,74	9,72	4,28	5,85	66,70	1,20	45,18	18,53	164,50	39,00	70,90	273,00	6,47	4,13	1105,25	34,63	0,885
10	2,42	6,68	3,17	5,78	58,62	1,06	36,65	155,01	185,00	50,50	85,80	243,50	6,43	3,90	119,46	16,48	0,903
11	3,81	6,09	2,22	5,63	73,60	1,41	37,18	111,43	185,00	37,00	95,60	262,00	6,72	3,45	1192,52	15,05	0,917
12	2,49	6,81	3,41	5,55	60,35	1,10	35,56	25,34	176,50	42,50	84,60	247,00	7,21	4,03	232,53	13,49	0,906
13	2,52	7,53	3,40	6,85	60,16	1,10	33,82	205,33	163,50	35,50	66,70	306,50	6,88	4,14	438,80	24,03	0,675
14	2,60	6,94	3,38	6,85	61,64	1,13	41,03	209,98	151,00	34,00	50,80	331,00	6,75	4,08	513,98	18,54	0,856
15	2,49	6,18	2,99	5,60	60,48	1,11	30,76	86,57	167,00	39,50	77,10	252,00	6,69	4,18	375,70	16,42	0,853
16	3,22	6,04	2,43	5,90	68,87	1,30	57,10	193,00	161,00	44,50	72,10	222,00	7,53	3,78	608,30	18,74	0,832
17	2,68	8,42	3,54	5,05	62,06	1,14	50,08	46,26	192,00	43,00	92,60	282,00	6,88	4,09	353,60	21,78	0,746
18	2,96	9,25	4,11	7,23	66,15	1,24	50,23	147,69	165,50	35,50	76,40	268,00	6,96	3,81	118,65	10,32	0,699
19	2,56	11,21	5,28	6,50	60,67	1,11	49,65	361,20	163,50	40,00	73,00	252,50	7,12	4,12	599,88	15,34	0,834
20	3,46	7,43	2,90	5,58	70,88	1,35	37,82	50,17	173,00	41,50	85,30	231,00	7,18	3,84	812,55	25,25	0,749
21	2,87	11,08	5,05	7,43	65,40	1,22	45,08	196,68	186,50	34,50	94,40	288,00	6,73	3,98	858,35	25,79	0,926
22	2,70	7,43	3,52	6,68	62,52	1,15	50,52	114,10	181,00	39,00	92,80	246,00	7,07	3,83	453,65	27,54	0,867
23	2,43	9,01	4,67	6,13	58,46	1,06	40,15	97,88	132,00	32,00	47,30	263,50	7,08	3,98	506,68	33,88	0,856
24	2,61	7,43	3,50	6,35	60,71	1,11	41,14	617,95	181,50	35,00	93,00	267,50	6,47	3,73	154,03	25,91	0,755
25	2,75	9,81	4,86	7,05	63,90	1,19	55,10	325,55	150,00	42,00	63,40	223,00	5,93	3,36	209,55	18,18	0,722
26	2,29	7,31	3,83	6,53	56,15	1,01	52,64	137,08	169,00	38,00	77,80	266,00	7,06	4,24	399,23	26,46	0,853
27	2,61	10,62	4,75	7,15	61,71	1,14	45,84	204,93	155,50	39,50	68,20	239,00	6,45	3,88	794,84	25,05	0,727
28	2,69	4,53	6,40	7,05	62,78	1,16	38,61	88,92	181,00	34,00	76,70	351,50	6,80	4,01	265,98	19,75	0,856

EK-4 CMS HSS ESRD Modeli risk skoru hesaplama çizelgeleri

Çizelge 4.1 CMS HCC ESRD modeline göre veri kümesindeki hastaların Risk Skoru hesaplama çizelgesi

(Risk Skoru = Yaş-Cinsiyet Katsayısı + Primer Hastalık Katsayısı + Co Morbid Durum Katsayıları + Etkileşim Katsayısı)

HASTA ID	Yaş	Cinsiyet (E/K)	Primer Hastalık	Co-Morbid durumlar				Etkileşim	Risk Skoru
				1	2	3	4		
1	72	E	Polikistik BH 0						0,776
2	28	E	Poststreptokoksik GN 0						0,614
3	46	E	Poststreptokoksik GN 0	Hipertansiyon 0,077					0,752
4	44	K	Poststreptokoksik GN 0	Hipertansiyon 0,077	Kronik Kalp Yet. 0,086				0,862
5	67	E	Renal Çelsea Kanseri 0						0,776
6	62	E	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				0,879
7	67	E	Polikistik BH 0						0,776
8	83	K	Hipertansif Nefropati 0	Hipertansiyon 0,077	Anemi 0				0,927
9	56	E	Diabetik Nefropati 0,08	Dişabet TIP II 0	Kronik Kalp Yet. 0,086	Görme yok 0		Dişabet-KİKY 0,020	0,885
10	86	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,903
11	77	K	Kronik Kalp Yet. 0,086						0,917
12	62	K	Diabetik Nefropati 0,08	Dişabet 0	Osteoporoz 0,077				0,906
13	51	E	Poststreptokoksik GN 0						0,675
14	65	E	Diabetik Nefropati 0,08	Dişabet 0					0,856
15	67	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,853
16	57	K	Kronik Kalp Yet. 0,086						0,832
17	56	K	Poststreptokoksik GN 0						0,746
18	43	K	Poststreptokoksik GN 0						0,699
19	67	E	İskemik 0,058						0,834
20	60	K	Polikistik BH 0						0,749
21	59	E	İskemik 0,058	Burger H 0,169					0,926
22	83	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,867
23	57	E	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				0,856
24	46	E	Diabetik Nefropati 0,08	Dişabet 0					0,755
25	62	E	Polikistik BH 0						0,722
26	65	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,853
27	38	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,727
28	69	E	Diabetik Nefropati 0,08	Dişabet 0					0,856
29	68	K	Poststreptokoksik GN 0						0,813
30	39	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,776
31	65	E	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				0,933
32	70	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,853
33	60	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,799
34	47	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,792
35	28	K	Polikistik BH 0						0,699

EK-4 (Devam) CMS HSS ESRD Modeli risk skoru hesaplama çizelgeleri

Çizelge 4.1 (Devam) CMS HCC ESRD modeline göre veri kümesindeki hastaların Risk Skoru hesaplama çizelgesi

(Risk Skoru = Yaş-Cinsiyet Katsayısı + Primer Hastalık Katsayısı + Co Morbid Durum Katsayıları + Etkileşim Katsayısı)

HASTA ID	Yaş	Cinsiyet (E/K)	Primer Hastalık	Co-Morbid durumlar				Etkileşim	Risk Skoru
				1	2	3	4		
36	63	K	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				0,906
37	77	K	Diabetik Nefropati 0,08	Kronik Kalp Yet. 0,086	Dişabet 0			Dişabet-İKKY 0,020	1,017
38	64	K	Diabetik Nefropati 0,08	Dişabet 0					0,829
39	76	K	Hipertansif Nefropati 0	Hipertansiyon 0,077	Kronik Kalp Yet. 0,086				0,994
40	70	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,853
41	79	E	Diabetik Nefropati 0,08	Dişabet 0					0,870
42	79	E	iskemik 0,058						0,848
43	79	E	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				0,947
44	37	K	Poststreptokoksik GN 0						0,699
45	46	K	Nefrit 0						0,715
46	64	K	Poststreptokoksik GN 0						0,749
47	44	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,776
48	39	K	Poststreptokoksik GN 0						0,699
49	59	E	Nefrolitozis 0						0,699
50	36	E	Diffüz crescent 0						0,650
51	60	E	Poststreptokoksik GN 0						0,722
52	21	E	Poststreptokoksik GN 0						0,614
53	32	K	Poststreptokoksik GN 0						0,699
54	30	E	Poststreptokoksik GN 0						0,614
55	44	E	Poststreptokoksik GN 0						0,650
56	62	K	Wegener Granül. 0,169	Hipertansiyon 0,077					0,995
57	41	K	Nefrit 0	Hipertansiyon 0,077					0,776
58	39	E	Poststreptokoksik GN 0						0,650
59	49	E	Polikistik BH 0						0,675
60	69	E	Diabetik Nefropati 0,08	Bypass, ASKH 0,086	Dişabet 0			Dişabet-İKKY 0,020	0,962
61	61	E	Diabetik Nefropati 0,08	Dişabet 0					0,802
62	76	E	Obst. Nefropati 0						0,790
63	76	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,908
64	74	K	Diabetik Nefropati 0,08	Dişabet 0					0,893
65	63	K	Hipertansif Nefropati 0						0,749
66	73	E	Diabetik Nefropati 0,08	Kronik Kalp Yet. 0,086	Dişabet 0			Dişabet-İKKY 0,020	0,962
67	77	E	KLL 0,16						0,950
68	70	K	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				0,970
69	66	E	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet Tip II 0				0,933
70	75	K	Hipertansif Nefropati 0	TAH+BSO 0,058	AC Enfarktüs 0,107				0,996

EK-4 (Devam) CMS HSS ESRD Modeli risk skoru hesaplama çizelgeleri

Çizelge 4.1 (Devam) CMS HCC ESRD modeline göre veri kümesindeki hastaların Risk Skoru hesaplama çizelgesi

(Risk Skoru = Yaş-Cinsiyet Katsayısı + Primer Hastalık Katsayısı + Co Morbid Durum Katsayıları + Etkileşim Katsayısı)

HASTA ID	Yaş	Cinsiyet (E/K)	Primer Hastalık	Co-Morbid durumlar				Etkileşim	Risk Skoru
				1	2	3	4		
71	68	K	Hipertansif Nefropati 0	Hipertansiyon 0,077	Dişabet 0,079				0,969
72	74	E	Antifosfolipid Send. 0,113						0,889
73	61	E	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	KAH 0,078	Dişabet 0	BPH 0,058		1,015
74	71	K	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	Kronik Kalp Yet. 0,086	Dişabet 0	Görme yok 0	Dişabet-KKY 0,020	1,076
75	71	K	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	Astım bronşit 0,156	Dişabet 0			1,126
76	67	K	Poststreptokoksik GN 0	PANCARD 0,086					0,899
77	63	K	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				0,906
78	56	E	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	Hiperglisemi 0	Dişabet 0			0,856
79	37	E	Hipertansif Nefropati 0	Hipertansiyon 0,077	Dişabet 0,08	Görme yok 0			0,807
80	59	K	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				0,903
81	72	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,853
82	35	K	Poststreptokoksik GN 0	Hipoproteinemi 0	Görme yok 0				0,699
83	45	K	Manit. Glom. 0						0,715
84	40	K	Renal Amiloidosis 0	Gastroduodenit 0,057					0,756
85	81	K	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				1,007
86	28	E	Poststreptokoksik GN 0						0,614
87	57	E	Polikistik BH 0						0,699
88	54	E	Renal Amiloidosis 0						0,675
89	80	K	Kronik Kalp Yet. 0,086						0,936
90	22	K	Poststreptokoksik GN 0	Hipertansiyon 0,077	Kronik Böbrek Yet. 0	Granulo parotit 0			0,776
91	76	K	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0	Görme Yok 0			0,988
92	58	E	Poststreptokoksik GN 0						0,699
93	60	E	Hipertansif Nefropati 0	Hipertansiyon 0,077	KAH 0,078				0,877
94	32	E	Poststreptokoksik GN 0						0,614
95	71	E	Hipertansif Nefropati 0	Hipertansiyon 0,077	Kronik Kalp Yet. 0,086	KAH 0,078	Hiperkolesterol 0		1,017
96	33	E	Poststreptokoksik GN 0						0,614
97	69	E	Kronik Böbrek Yet. 0						0,776
98	56	E	Dişabetik Nefropati 0,08	Dişabet 0					0,779
99	47	E	Renal Amiloidosis 0	FMF 0,113					0,788
100	57	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,823
101	47	K	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				0,872
102	72	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,890
103	80	E	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet TIP II 0				0,947
104	63	E	Dişabetik Nefropati 0,08	Dişabet 0	Ampute 0,073				0,875
105	66	E	Dişabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0				0,933

EK-4 (Devam) CMS HSS ESRD Modeli risk skoru hesaplama çizelgeleri

Çizelge 4.1 (Devam) CMS HCC ESRD modeline göre veri kümesindeki hastaların Risk Skoru hesaplama çizelgesi

(Risk Skoru = Yaş-Cinsiyet Katsayısı + Primer Hastalık Katsayısı + Co Morbid Durum Katsayıları + Etkileşim Katsayısı)

HASTA ID	Yaş	Cinsiyet (E/K)	Primer Hastalık	Co-Morbid durumlar				Etkileşim	Risk Skoru
				1	2	3	4		
106	54	E	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Diyabet 0				0,832
107	71	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,853
108	56	K	Polikistik BH 0	Hipertansiyon 0,077					0,823
109	72	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,890
110	76	K	Diabetik Nefropati 0,08	Hipertansiyon 0,077	KAH 0,078	Diyabet 0			1,066
111	22	K	Nefrotik Sendrom 0						0,699
112	74	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,890
113	40	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,776
114	80	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,867
115	77	E	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Diyabet 0	Retinopati 0	Prostat 0,058		1,005
116	53	K	Diabetik Nefropati 0,08	Diyabet 0					0,795
117	29	E	Diabetik Nefropati 0,08	Diyabet 0					0,694
118	61	E	Diabetik Nefropati 0,08	Diyabet 0					0,802
119	71	K	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Diyabet 0				0,970
120	50	K	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Diyabet 0				0,872
121	66	E	Kronik Böbrek Yet. 0						0,776
122	31	E	Renal Amila. 0	Ag. Anemi 0					0,614
123	70	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,853
124	85	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,949
125	56	K	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Diyabet 0				0,903
126	37	E	Poststreptokoksik GN 0	İneg. Tansiyon 0,061					0,711
127	46	E	Poststreptokoksik GN 0						0,675
128	74	K	Ürolitiazis 0	Hipertansiyon 0,077					0,890
129	63	K	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Görme az 0	Diyabet 0			0,906
130	23	K	Membranep. GN 0						0,699
131	76	E	Hipertansif Nefropati 0	Hipertansiyon 0,077	Hipotiroid 0	Hiperlipidemi 0			0,867
132	64	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,826
133	66	K	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Hiperlipidemi 0	Diyabet 0			0,970
134	60	K	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Diyabet 0				0,906
135	66	K	Kronik Böbrek Yet. 0						0,813
136	62	K	Obst. Nefropati 0	Hipertansiyon 0,077	Diyabet 0,08				0,906
137	63	K	Sjögren 0,113	Hipertansiyon 0,077					0,939
138	69	E	Polikistik BH 0	Hipertansiyon 0,077					0,853
139	72	E	PKH 0,259						1,035
140	85	E	Hipertansif Nefropati 0	Hipertansiyon 0,077	Kronik Kalp Yet. 0,086				0,989

EK-4 (Devam) CMS HSS ESRD Modeli risk skoru hesaplama çizelgeleri

Çizelge 4.1 (Devam) CMS HCC ESRD modeline göre veri kümesindeki hastaların Risk Skoru hesaplama çizelgesi

(Risk Skoru = Yaş-Cinsiyet Katsayısı + Primer Hastalık Katsayısı + Co Morbid Durum Katsayıları + Etkileşim Katsayısı)

HASTA ID	Yaş	Cinsiyet (E/K)	Primer Hastalık	Co-Morbid durumlar				Etkileşim	Risk Skoru
				1	2	3	4		
141	68	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,853
142	72	K	Diabetik Nefropati 0,813	Hipertansiyon 0,077	Osteoporoz 0,077	Dişabet 0	Diş Ret.+Nör. 0		1,047
143	81	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,927
144	40	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,727
145	57	E	Nefroditosis 0	Dişabet 0,079					0,778
146	63	E	Kronik Böbrek Yet. 0						0,722
147	49	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,752
148	73	K	Kronik Böbrek Yet. 0	Hipertansiyon 0,077	Dişabet 0,079				0,969
149	58	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,776
150	60	E	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Bilateral Amarozis 0	Dişabet 0			0,879
151	57	E	Poststreptokoksik GN 0	Hipertansiyon 0,077					0,776
152	58	K	Poststreptokoksik GN 0	Hipertansiyon 0,077					0,823
153	48	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,752
154	28	K	Renal Amiloidosis 0	FMF 0,113					0,812
155	83	K	Berger H 0	Hipertansiyon 0,077	Kronik Kalp Yet. 0,086				1,013
156	86	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,949
157	68	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,853
158	49	E	Diabetik Nefropati 0,08	Dişabet 0					0,756
159	65	K	Diabetik Nefropati 0,08	Dişabet 0					0,893
160	30	K	Poststreptokoksik GN 0						0,699
161	21	K	Post streptokok GN 0						0,699
162	44	K	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,776
163	56	K	Diabetik Nefropati 0,08	Hipertansiyon 0,077	Dişabet 0	Retinopati 0			0,903
164	66	K	Hipertansif Nefropati 0	Hipertansiyon 0,077	Dişabet 0,079				0,969
165	71	E	Hipertansif Nefropati 0	Hipertansiyon 0,077					0,853
166	85	K	Diabetik Nefropati 0,08	Dişabet 0					0,952
167	69	E	Diabetik Nefropati 0,08	Dişabet 0					0,856
168	54	E	Hipertansif Nefropati 0	Hipertansiyon 0,077	Dişabet 0,079				0,831
169	78	K	Hipertansif Nefropati 0	Hipertansiyon 0,077	Kronik Kalp Yet. 0,086				0,994
170	71	K	Kronik Böbrek Yet. 0						0,813

EK-4 (Devam) CMS HSS ESRD Modeli risk skoru hesaplama çizelgeleri

Çizelge 4.2 Risk skorunun farklı sınıf sayıları için nominal hale getirilmiş hali

ID	Risk Skoru	Sınıf Sayısı			
		2	3	4	5
1	0.776	0	0	1	1
2	0.614	0	0	0	0
3	0.752	0	0	1	1
4	0.862	0	1	1	2
5	0.776	0	0	1	1
6	0.879	1	1	2	2
7	0.776	0	0	1	1
8	0.927	1	1	2	3
9	0.885	0	1	2	2
10	0.903	1	1	2	2
11	0.917	1	1	2	2
12	0.906	1	1	2	2
13	0.675	0	0	0	0
14	0.856	0	1	1	2
15	0.853	0	1	1	2
16	0.832	0	1	1	2
17	0.746	0	0	1	1
18	0.699	0	0	0	0
19	0.834	0	1	1	2
20	0.749	0	0	1	1
21	0.926	1	1	2	3
22	0.867	0	1	1	2
23	0.856	0	1	1	2
24	0.755	0	0	1	1
25	0.722	0	0	0	1
26	0.853	0	1	1	2
27	0.727	0	0	0	1
28	0.856	0	1	1	2
29	0.813	0	1	1	2
30	0.776	0	0	1	1
31	0.933	1	1	2	3
32	0.853	0	1	1	2
33	0.799	0	1	1	1
34	0.792	0	1	1	1
35	0.699	0	0	0	0
36	0.906	1	1	2	2
37	1.017	1	2	3	3
38	0.829	0	1	1	2
39	0.994	1	2	2	3
40	0.853	0	1	1	2
41	0.870	1	1	1	2
42	0.848	0	1	1	2
43	0.947	1	1	2	3
44	0.699	0	0	0	0
45	0.715	0	0	0	0
46	0.749	0	0	1	1
47	0.776	0	0	1	1
48	0.699	0	0	0	0
49	0.699	0	0	0	0
50	0.650	0	0	0	0
51	0.722	0	0	0	1
52	0.614	0	0	0	0
53	0.699	0	0	0	0
54	0.614	0	0	0	0
55	0.650	0	0	0	0
56	0.995	1	2	2	3
57	0.776	0	0	1	1
58	0.650	0	0	0	0
59	0.675	0	0	0	0
60	0.962	1	2	2	3
61	0.802	0	1	1	1
62	0.790	0	1	1	1
63	0.908	1	1	2	2
64	0.893	1	1	2	2
65	0.749	0	0	1	1
66	0.962	1	2	2	3
67	0.950	1	1	2	3
68	0.970	1	2	2	3
69	0.933	1	1	2	3
70	0.996	1	2	2	3
71	0.969	1	2	2	3
72	0.889	1	1	2	2
73	0.879	1	1	2	2
74	1.076	1	2	3	4
75	1.126	1	2	3	4
76	0.899	1	1	2	2
77	0.906	1	1	2	2
78	0.856	0	1	1	2
79	0.807	0	1	1	2
80	0.903	1	1	2	2
81	0.853	0	1	1	2
82	0.699	0	0	0	0
83	0.715	0	0	0	0
84	0.756	0	0	1	1
85	1.007	1	2	3	3
86	0.614	0	0	0	0

EK-4 (Devam) CMS HSS ESRD Modeli risk skoru hesaplama çizelgeleri

Çizelge 4.2 (Devam) Risk skorunun farklı sınıf sayıları için nominal hale getirilmiş hali

ID	Risk Skoru	Sınıf Sayısı			
		2	3	4	5
87	0.699	0	0	0	0
88	0.675	0	0	0	0
89	0.936	1	1	2	3
90	0.776	0	0	1	1
91	0.988	1	2	2	3
92	0.699	0	0	0	0
93	0.877	1	1	2	2
94	0.614	0	0	0	0
95	1.017	1	2	3	3
96	0.614	0	0	0	0
97	0.776	0	0	1	1
98	0.779	0	0	1	1
99	0.788	0	1	1	1
100	0.823	0	1	1	2
101	0.872	1	1	2	2
102	0.890	1	1	2	2
103	0.947	1	1	2	3
104	0.875	1	1	2	2
105	0.933	1	1	2	3
106	0.832	0	1	1	2
107	0.853	0	1	1	2
108	0.823	0	1	1	2
109	0.890	1	1	2	2
110	1.066	1	2	3	4
111	0.699	0	0	0	0
112	0.890	1	1	2	2
113	0.776	0	0	1	1
114	0.867	0	1	2	2
115	1.005	1	2	3	3
116	0.795	0	1	1	1
117	0.694	0	0	0	0
118	0.802	0	1	1	1
119	0.970	1	2	2	3
120	0.872	1	1	2	2
121	0.776	0	0	1	1
122	0.614	0	0	0	0
123	0.853	0	1	1	2
124	0.949	1	1	2	3
125	0.903	1	1	2	2
126	0.711	0	0	0	0
127	0.675	0	0	0	0
128	0.890	1	1	2	2

ID	Risk Skoru	Sınıf Sayısı			
		2	3	4	5
129	0.906	1	1	2	2
130	0.699	0	0	0	0
131	0.867	0	1	2	2
132	0.826	0	1	2	2
133	0.970	1	2	2	3
134	0.906	1	1	2	2
135	0.813	0	1	1	2
136	0.906	1	1	2	2
137	0.939	1	1	2	3
138	0.853	0	1	1	2
139	1.035	1	2	3	4
140	0.989	1	2	2	3
141	0.853	0	1	1	2
142	1.047	1	2	3	4
143	0.927	1	1	2	3
144	0.727	0	0	0	1
145	0.778	0	0	1	1
146	0.722	0	0	0	1
147	0.752	0	0	1	1
148	0.969	1	2	2	3
149	0.776	0	0	1	1
150	0.876	1	1	2	2
151	0.776	0	0	1	1
152	0.823	0	1	1	2
153	0.752	0	0	1	1
154	0.812	0	1	1	2
155	1.013	1	2	3	3
156	0.949	1	1	2	3
157	0.853	0	1	1	2
158	0.755	0	0	1	1
159	0.893	1	1	2	2
160	0.699	0	0	0	0
161	0.699	0	0	0	0
162	0.776	0	0	1	1
163	0.903	1	1	2	2
164	0.969	1	2	2	3
165	0.853	0	1	1	2
166	0.952	1	1	2	3
167	0.856	0	1	1	2
168	0.831	0	1	1	2
169	0.994	1	2	2	3
170	0.813	0	1	1	2

EK-5 Uygulama çıktı raporları

Çıktı Raporu I RSCORE 2 Sınıf için 2 Küme oluşturacak K-means uygulaması (1/2)

=== Run information for 2 Classes for RSCORE===

Scheme: weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance
-R first-last" -I 500 -S 10

Relation: whatever-weka.filters.AllFilter

Instances: 170

Attributes: 55

ID, AGE, SAGE, SEX, HEI, WEI, VKI, SS_AKB, SD_AKB, NRI, SES, SEAG, DHS, HSS, DYT, AKG, Pre-W, Pre-S, Pre-D, Ps-W, Ps_S, Ps-D, ISW, KAH, EIT, EPOD, EPOU, EPOR, RZI, Htc, CRP, Na, K_I, K_O, K_F, GLI, BUN_I, BUN_O, BUN_R, KRE_I, KRE_O, URA, URR, KtV, CaXP, PTH, TKoll, HDL, LDL, TGL, TPRO, ALB, FERT, TSI

Ignored: RSCORE

Test mode: Classes to clusters evaluation on training data (for 2 Clusters)

=== Model and evaluation on training set ===

kMeans

Number of iterations: 5

Within cluster sum of squared errors: 431.25707845064153

Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Cluster#		
	Full Data (170)	0 (89)	1 (81)
=====			
ID	85.5	79.5506	92.037
AGE	59.4118	59.3933	59.4321
SAGE	55.1118	55.0787	55.1481
SEX	1	1	0
HEI	163.7412	168.9663	158
WEI	65.3824	66.5921	64.0531
VKI	24.0624	22.9449	25.2901
SS_AKB	136.3529	137.9775	134.5679
SD_AKB	78.8235	79.4382	78.1481
NRI	51.0494	48.4955	53.8556
SES	566.1529	547.7079	586.4198
SEAG	58.8471	59.1685	58.4938
DHS	4.91	4.936	4.8815
HSS	11.6682	11.736	11.5938
DYT	0	0	0
AKG	4406.3471	4457.7865	4349.8272
Pre-W	66.2906	67.4955	64.9667
Pre-S	130.6765	130.8539	130.4815
Pre-D	76.8	76.5955	77.0247
Ps-W	64.4071	65.5011	63.2049
Ps_S	121.0824	120.9551	121.2222
Ps-D	72.8353	72.764	72.9136

EK-5 (Devam) Uygulama çıktı raporları

Çıktı Raporu I RSCORE 2 Sınıf için 2 Küme oluşturacak K-means uygulaması (2/2)

Cluster centroids:

Attribute	Cluster#		
	Full Data (170)	0 (89)	1 (81)
=====			
ISW	1.8806	1.991	1.7593
KAH	244.9588	250.809	238.5309
EIT	0	0	0
EPOD	20936.2059	19144.3483	22905.037
EPOU	336.1929	300.3135	375.616
EPOR	32.7024	28.727	37.0704
RZI	332.7059	357.8989	305.0247
Htc	32.436	33.2198	31.5748
CRP	4.719	4.9803	4.4319
Na	136.8174	136.9974	136.6195
K_I	5.1958	5.1926	5.1994
K_O	3.9226	3.9181	3.9275
K_F	1.2736	1.2746	1.2725
GLI	121.7224	115.0002	129.1085
BUN_I	140.8966	146.0566	135.2269
BUN_O	51.5708	55.1003	47.6926
BUN_R	2.765	2.6707	2.8686
KRE_I	8.2795	8.6433	7.8798
KRE_O	3.8388	4.0967	3.5554
URA	6.3265	6.4265	6.2167
URR	63.3246	62.3496	64.3959
KtV	1.1721	1.1494	1.1969
CaXP	44.6532	44.6717	44.6328
PTH	234.1285	249.7388	216.9764
TKoll	177.7241	175.1011	180.6062
HDL	40.4159	40.1798	40.6753
LDL	81.6829	82.0809	81.2457
TGL	227.7071	223.0989	232.7704
TPRO	6.8005	6.7553	6.8501
ALB	3.9376	3.9679	3.9043
FERT	506.4196	483.5843	531.5104
TSI	26.9316	27.0912	26.7563
Clustered Instances			
0	89 (52%)	1	81 (48%)

Class attribute: RSCORE

Classes to Clusters:

0 1 <-- assigned to cluster

68 36 | 0 21 45 | 1

Cluster 0 <-- 0 Cluster 1 <-- 1

Incorrectly clustered instances : 57.0 33.5294 %

EK-5 (Devam) Uygulama çıktı raporları

Çıktı Raporu II RSCORE 2 Sınıf için Özellik Seçimi uygulaması (1/1)

=== Run information for 2 Classes RSCORE===

Evaluator: weka.attributeSelection.CfsSubsetEval

Search: weka.attributeSelection.LinearForwardSelection -D 0 -N 5 -I -K 50 -T 0 -Z false

Relation: whatever

Instances: 170

Attributes: 55

ID, AGE, SAGESEX, SEX, HEI, WEI, VKI, SS_AKB, SD_AKB, NRI, SES, SEAG, DHS, HSS, DYT, AKG, Pre-W, Pre-S, Pre-D, Ps-W, Ps_S, Ps-D, ISW, KAH, EIT, EPOD, EPOU, EPOR, RZI, Htc, CRP, Na, K_I, K_O, K_F, GLI, BUN_I, BUN_O, BUN_R, KRE_I, KRE_O, URA, URR, KtV, CaXP, PTH, TKoll, HDL, LDL, TGL, TPRO, ALB, FERT, TSI, RSCORE

Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:

Linear Forward Selection.

Start set: no attributes

Forward selection method: forward selection

Stale search after 5 node expansions

Linear Forward Selection Type: fixed-set

Number of top-ranked attributes that are used: 50

Total number of subsets evaluated: 713

Merit of best subset found: 0.35

Attribute Subset Evaluator (supervised, Class (nominal): 55 RSCORE):

CFS Subset Evaluator

Including locally predictive attributes

Selected attributes: 2,3,4,5,22,31,33,36,37,46 : 10

AGE

SAGESEX

SEX

HEI

Ps-D

CRP

K_I

GLI

BUN_I

PTH

EK-5 (Devam) Uygulama çıktı raporları

Çıktı Raporu III RSCORE 2 Sınıf için J4.8 Karar ağacı algoritması uygulaması (1/2)

=== Run information for 2 Classes of RSCORE===J48

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: whatever-weka.filters.AllFilter

Instances: 170

Attributes: 55

ID, AGE, SAGESEX, SEX, HEI, WEI, VKI, SS_AKB, SD_AKB, NRI, SES, SEAG, DHS, HSS, DYT, AKG, Pre-W, Pre-S, Pre-D, Ps-W, Ps_S, Ps-D, ISW, KAH, EIT, EPOD, EPOU, EPOR, RZI, Htc, CRP, Na, K_I, K_O, K_F, GLI, BUN_I, BUN_O, BUN_R, KRE_I, KRE_O, URA, URR, KtV, CaXP, PTH, TKoll, HDL, LDL, TGL, TPRO, ALB, FERT, TSI, RSCORE

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

Note that J48 (implementation of C4.5 algorithm) does not require discretization of numeric attributes, in contrast to the ID3 algorithm from which C4.5 has evolved.

AGE <= 58

| GLI <= 155: 0 (60.0)

| GLI > 155

| | Htc <= 31.16: 1 (4.0)

| | Htc > 31.16: 0 (4.0)

AGE > 58

| SEX = 1

| | GLI <= 136.92

| | | WEI <= 68.5

| | | | DYT = 0

| | | | HDL <= 37.5: 1 (4.0)

| | | | HDL > 37.5: 0 (8.0/1.0)

| | | | DYT = 2: 0 (6.0/1.0)

| | | | DYT = 1

| | | | ID <= 59: 0 (2.0)

| | | | ID > 59: 1 (2.0)

| | | | DYT = 5: 0 (3.0/1.0)

| | | | DYT = 7: 0 (0.0)

| | | | DYT = 3: 1 (2.0)

| | | | DYT = 4: 0 (0.0)

| | | | DYT = 8: 0 (0.0)

| | | | DYT = 6: 0 (0.0)

| | | WEI > 68.5: 0 (14.0)

| | GLI > 136.92

| | | K_F <= 1.66: 1 (10.0)

| | | K_F > 1.66: 0 (2.0)

| SEX = 0

| | SEAG <= 91: 1 (43.0/3.0)

| | SEAG > 91: 0 (6.0/1.0)

EK-5 (Devam) Uygulama çıktı raporları

Çıktı Raporu III RSCORE 2 Sınıf için J4.8 Karar ağacı algoritması uygulaması (2/2)

Number of Leaves : 19
 Size of the tree : 30
 Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	138	81.1765 %
Incorrectly Classified Instances	32	18.8235 %
Kappa statistic		0.6059
Mean absolute error		0.2115
Root mean squared error		0.4064
Relative absolute error		44.4869 %
Root relative squared error		83.3544 %
Total Number of Instances	170	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.837	0.227	0.853	0.837	0.845	0.825	0
0.773	0.163	0.75	0.773	0.761	0.825	1
0.812	0.202	0.813	0.812	0.812	0.825	Wei. Avg.

=== Confusion Matrix ===

a b <-- classified as
 87 17 | a = 0
 15 51 | b = 1

EK-5 (Devam) Uygulama çıktı raporları

Çıktı Raporu IV RSCORE 2 Sınıf için RandomTree Karar ağacı algoritması uygulaması (1/4)

=== Run information for 2 Classes of RSCORE===RandomTree

Scheme: weka.classifiers.trees.RandomTree -K 1 -M 1.0 -S 1

Relation: whatever-weka.filters.AllFilter

Instances: 170

Attributes: 55

ID, AGE, SAGESEX, SEX, HEI, WEI, VKI, SS_AKB, SD_AKB, NRI, SES, SEAG, DHS, HSS, DYT, AKG, Pre-W, Pre-S, Pre-D, Ps-W, Ps_S, Ps-D, ISW, KAH, EIT, EPOD, EPOU, EPOR, RZI, Htc, CRP, Na, K_I, K_O, K_F, GLI, BUN_I, BUN_O, BUN_R, KRE_I, KRE_O, URA, URR, KtV, CaXP, PTH, TKoll, HDL, LDL, TGL, TPRO, ALB, FERT, TSI

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

RandomTree

ISW < 2.85

| SES < 294.5

| | HSS < 9.25

| | | NRI < 46.45 : 0 (1/0)

| | | NRI >= 46.45

| | | | TGL < 320.25

| | | | | BUN_R < 2.59 : 0 (1/0)

| | | | | BUN_R >= 2.59 : 1 (3/0)

| | | | TGL >= 320.25 : 0 (1/0)

| | HSS >= 9.25

| | | EPOR < 49.5

| | | | ID < 91.5

| | | | | Na < 137.36 : 0 (13/0)

| | | | | Na >= 137.36

| | | | | URR < 62.95 : 0 (4/0)

| | | | | URR >= 62.95

| | | | | FERT < 357.39 : 1 (2/0)

| | | | | FERT >= 357.39 : 0 (2/0)

| | | | ID >= 91.5

| | | | | EPOD < 13470

| | | | | TKoll < 223.5

| | | | | | HEI < 162.5 : 0 (1/0)

| | | | | | HEI >= 162.5 : 1 (1/0)

| | | | | TKoll >= 223.5 : 1 (2/0)

| | | | EPOD >= 13470

| | | | | TKoll < 213.75 : 0 (8/0)

| | | | | TKoll >= 213.75

| | | | | | Ps-D < 73.5 : 1 (1/0)

| | | | | | Ps-D >= 73.5 : 0 (1/0)

| | | EPOR >= 49.5

| | | | EPOR < 57 : 1 (2/0)

| | | | EPOR >= 57

| | | | | TSI < 18.47 : 0 (1/0)

EK-5 (Devam) Uygulama çıktı raporları

Çıktı Raporu IV RSCORE 2 Sınıf için RandomTree Karar ağacı algoritması uygulaması (2/4)

```

| | | | | TSI >= 18.47
| | | | |   EIT = 0
| | | | |   | WEI < 66.3 : 1 (2/0)
| | | | |   | WEI >= 66.3 : 0 (1/0)
| | | | |   EIT = 2 : 0 (0/0)
| | | | |   EIT = 1 : 0 (1/0)
| | | | |   EIT = 3 : 0 (0/0)
| | | | | SES >= 294.5
| | | | |   URR < 60.47
| | | | |     KAH < 262.5
| | | | |       ISW < 2.05
| | | | |         K_F < 1.15
| | | | |           SEAG < 54.5
| | | | |             EPOD < 44033.5
| | | | |               BUN_I < 129.92 : 1 (2/0)
| | | | |               BUN_I >= 129.92
| | | | |                 HEI < 165 : 1 (1/0)
| | | | |                 HEI >= 165 : 0 (1/0)
| | | | |               EPOD >= 44033.5 : 0 (1/0)
| | | | |             SEAG >= 54.5 : 1 (4/0)
| | | | |           K_F >= 1.15 : 1 (7/0)
| | | | |         ISW >= 2.05 : 1 (7/0)
| | | | |       KAH >= 262.5
| | | | |         HEI < 175.5 : 0 (1/0)
| | | | |         HEI >= 175.5
| | | | |           ALB < 4.01 : 1 (1/0)
| | | | |           ALB >= 4.01 : 0 (1/0)
| | | | |   URR >= 60.47
| | | | |     Ps-W < 42.1 : 0 (3/0)
| | | | |     Ps-W >= 42.1
| | | | |       VKI < 18.3
| | | | |         URR < 61.26 : 1 (1/0)
| | | | |         URR >= 61.26 : 0 (8/0)
| | | | |       VKI >= 18.3
| | | | |         HSS < 11.45
| | | | |           NRI < 53.2 : 1 (3/0)
| | | | |           NRI >= 53.2
| | | | |             AGE < 75 : 1 (1/0)
| | | | |             AGE >= 75 : 0 (1/0)
| | | | |         HSS >= 11.45
| | | | |           TGL < 118.75 : 0 (4/0)
| | | | |           TGL >= 118.75
| | | | |             TPRO < 7.74
| | | | |             LDL < 125.25
| | | | |             EIT = 0
| | | | |             VKI < 23.85

```


EK-5 (Devam) Uygulama çıktı raporları

Çıktı Raporu IV RSCORE 2 Sınıf için RandomTree Karar ağacı algoritması uygulaması (4/4)

```

| | | | | | | | | | DYT = 6 : 1 (1/0)
| | | | | | | | | | EIT = 3
| | | | | | | | | | Ps-D < 76
| | | | | | | | | | SAGESEX < 47.5 : 0 (1/0)
| | | | | | | | | | SAGESEX >= 47.5 : 1 (2/0)
| | | | | | | | | | Ps-D >= 76 : 0 (1/0)
| | | | | | | | | | LDL >= 125.25 : 0 (3/0)
| | | | | | | | | | TPRO >= 7.74 : 0 (2/0)
ISW >= 2.85
| HSS < 10
| | K_I < 5.47 : 1 (1/0)
| | K_I >= 5.47 : 0 (1/0)
| HSS >= 10
| | Pre-D < 86.5 : 0 (15/0)
| | Pre-D >= 86.5
| | | CRP < 2.8 : 0 (3/0)
| | | CRP >= 2.8
| | | K_I < 5.29 : 0 (2/0)
| | | K_I >= 5.29 : 1 (1/0)

```

Size of the tree : 140

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	93	54.7059 %
Incorrectly Classified Instances	77	45.2941 %
Kappa statistic		0.0595
Mean absolute error		0.4516
Root mean squared error		0.6712
Relative absolute error		94.9679 %
Root relative squared error		137.6754 %
Total Number of Instances	170	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.606	0.545	0.636	0.606	0.621	0.535	0
0.455	0.394	0.423	0.455	0.438	0.535	1
0.547	0.487	0.553	0.547	0.55	0.535	Wei. Avg.

=== Confusion Matrix ===

a b <-- classified as

63 41 | a = 0

36 30 | b = 1

EK-5 (Devam) Uygulama çıktı raporları

Çıktı Raporu V RSCORE 2 Sınıf için PART Karar kuralları algoritması uygulaması (1/2)

=== Run information for 2 classes===PART Rules

Scheme: weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1

Relation: whatever

Instances: 170

Attributes: 55

ID, AGE, SAGESEX, SEX, HEI, WEI, VKI, SS_AKB, SD_AKB, NRI, SES, SEAG, DHS, HSS, DYT, AKG, Pre-W, Pre-S, Pre-D, Ps-W, Ps_S, Ps-D, ISW, KAH, EIT, EPOD, EPOU, EPOR, RZI, Htc, CRP, Na, K_I, K_O, K_F, GLI, BUN_I, BUN_O, BUN_R, KRE_I, KRE_O, URA, URR, KtV, CaXP, PTH, TKoll, HDL, LDL, TGL, TPRO, ALB, FERT, TSI, RSCORE

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list

AGE <= 58 AND

GLI <= 155: 0 (60.0)

SEX = 0 AND

SEAG <= 91 AND

SEAG > 23: 1 (38.0/1.0)

K_I <= 4.83 AND

CRP > 0.58: 1 (12.0)

K_I <= 5.98 AND

Pre-S <= 142 AND

AKG <= 4423 AND

SES > 112: 0 (15.0)

DYT = 0 AND

TSI > 25.91: 0 (11.0)

EIT = 0 AND

SAGESEX <= 66 AND

TSI <= 19.44: 0 (7.0)

URA > 6.45: 1 (9.0)

Pre-S <= 142 AND

TKoll > 158: 1 (8.0)

: 0 (10.0)

Number of Rules : 9

Time taken to build model: 0.3 seconds

EK-5 (Devam) Uygulama çıktı raporları

Çıktı Raporu V RSCORE 2 Sınıf için PART Karar kuralları algoritması uygulaması
(2/2)

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	133	78.2353 %
Incorrectly Classified Instances	37	21.7647 %
Kappa statistic	0.5354	
Mean absolute error	0.2249	
Root mean squared error	0.4532	
Relative absolute error	47.3009 %	
Root relative squared error	92.9562 %	
Total Number of Instances	170	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.846	0.318	0.807	0.846	0.826	0.791	0
	0.682	0.154	0.738	0.682	0.709	0.791	1
Weighted Avg.	0.782	0.254	0.78	0.782	0.781	0.791	

=== Confusion Matrix ===

a b <-- classified as

88 16 | a = 0

21 45 | b = 1

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : ALTINTAŞ, Yunus Yesari
 Uyruğu : T.C.
 Doğum tarihi ve yeri : 24.07.1980 Ankara
 Medeni hali : Bekar
 Telefon : 0 (312) 220 17 79
 Faks : -
 e-mail : yy_altintas@yahoo.com

Eğitim

Derece Tarihi	Eğitim Birimi	Mezuniyet
Lisans	Gazi Üniversitesi / Endüstri Müh. Bölümü	2002
Lise	Özel Muradiye Fen Lisesi	1998

İş Deneyimi

Yıl	Yer	Görev
2009-	Sanayi ve Ticaret Bakanlığı	AB Uzm. Yrd.
2003-2006	Termikel Madeni San. Tic. A.Ş.	KYS Sorumlusu.

Yabancı Dil

İngilizce

Yayınlar

-

Hobiler

Basketbol, Felsefe, Dilbilim