

**GERÇEK LABORATUVAR VERİLERİNİN VERİ
MADENCİLİĞİ TEKNİKLERİ İLE ANALİZİ
YÜKSEK LİSANS TEZİ**

Asiye Betül CİGA

DANIŞMAN
Yrd. Doç. Dr. Ertuğrul ERGÜN

İNTERNET VE BİLİŞİM TEKNOLOJİLERİ
YÖNETİMİ ANABİLİM DALI
Haziran, 2015

AFYON KOCATEPE ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

**GERÇEK LABORATUVAR VERİLERİNİN VERİ MADENCİLİĞİ
TEKNİKLERİ İLE ANALİZİ**

Asiye Betül CİGA

DANIŞMAN

Yrd. Doç. Dr. Ertuğrul ERGÜN

İNTERNET VE BİLİŞİM TEKNOLOJİLERİ YÖNETİMİ

ANABİLİM DALI

Haziran, 2015

TEZ ONAY SAYFASI

Asiye Betül CIGA tarafından hazırlanan “Gerçek Laboratuvar Verilerinin Veri Madenciliği Teknikleri İle Analizi” adlı tez çalışması lisansüstü eğitim ve öğretim yönetmeliğinin ilgili maddeleri uyarınca 12/06/2015 tarihinde aşağıdaki jüri tarafından oy birliği ile Afyon Kocatepe Üniversitesi Fen Bilimleri Enstitüsü **İnternet ve Bilişim Teknolojileri Yönetimi Anabilim Dalı’nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman : Yrd. Doç. Dr. Ertuğrul ERGÜN

Başkan : Prof. Dr. Tülay KÖKEN İmza
Afyon Kocatepe Üniversitesi Tıp Fakültesi

Üye :Yrd. Doç. Dr. Ali Hakan IŞIK İmza
Mehmet Akif Ersoy Üniversitesi Mühendislik Fakültesi

Üye :Yrd. Doç. Dr. Ertuğrul ERGÜN İmza
Afyon Kocatepe Üniversitesi Uzaktan Eğitim MYO

Afyon Kocatepe Üniversitesi

Fen Bilimleri Enstitüsü Yönetim Kurulu’nun

...../...../..... tarih ve

..... sayılı kararıyla onaylanmıştır.

Prof. Dr. İbrahim EROL

Enstitü Müdürü

BİLİMSEL ETİK BİLDİRİM SAYFASI

Afyon Kocatepe Üniversitesi

**Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım
bu tez çalışmada;**

- Tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- Görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- Başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- Atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- Kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- Ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

26/06/2015

Asiye Betül CİGA

ÖZET

Yüksek Lisans Tezi

GERÇEK LABORATUVAR VERİLERİNİN VERİ MADENCİLİĞİ TEKNİKLERİ İLE ANALİZİ

Asiye Betül CİGA

Afyon Kocatepe Üniversitesi

Fen Bilimleri Enstitüsü

İnternet ve Bilişim Teknolojileri Yönetimi Anabilim Dalı

Danışman: Yrd. Doç. Dr. Ertuğrul ERGÜN

Veri madenciliği teknikleri son yıllarda büyük miktarda veri üretilen hemen her alanda kullanılmaktadır. İnsan algısının bazı detayları kaçırma olasılığına karşın gözden kaçan detayların tespit edilip hastalıkların teşhisi, erken teşhis ve tedavi sürecindeki değişiklikler veri madenciliği yaklaşımı ile daha olası hale gelmiştir.

Bu araştırmada; bir kamu hastanesinin altı aylık laboratuvar verileriyle; bilinen ve az sayıdaki tetkikle yapılan çalışmalardan farklı olarak, daha fazla biyokimya tetkik değerleri analizlere eklenerek anemi ve türlerinin teşhisinde etkin rolü olan parametrelerin tespit edilmesi için modeller oluşturulmuştur. Anemi ve sınıflandırmasında istenilen test kalabalığı içerisinde tanıyı gözden kaçırmamaya yardımcı olmak amacıyla sağlık alanında veri madenciliği çalışması gerçekleştirilmiştir. Modellemede WEKA açık kaynak veri madenciliği programı kullanılmış olup, sınıflandırma algoritmaları tek tek uygulanarak sonuçlar karşılaştırılmış ve en yüksek performans değerine sahip sınıflandırma algoritmalarının trees.J48 ile rules.PART algoritmaları olduğu tespit edilmiştir. Kümeleme analizi sonuçlarından FilteredClusterer kümeleme algoritmasının veri seti ile %100 uyumlu olduğu görülmüştür. Tertius birliktelik kuralları uzman görüşü ile doğrulanırken, veri setindeki kısıtlamalardan dolayı FilteredAssociator birliktelik analizlerinin ve karar ağaçlarından elde edilen kuralların pratikte tam olarak doğru ve kullanılabilir olmadığı görülmüştür.

2015, xi + 208 sayfa

Anahtar Kelimeler: Veri madenciliği, Anemi, Anemi Türleri, WEKA

ABSTRACT

M.Sc Thesis

ANALYSIS OF REAL LABORATORY DATA WITH DATA MINING TECHNIQUES

Asiye Betül CIGA

Afyon Kocatepe University

Graduate School of Natural and Applied Sciences

Department of Internet and Management of Information Technologies

Supervisor: Asst.Prof. Ertuğrul ERGÜN

Data mining techniques are widely used in studies where vast amount of data is being produced. Since it is possible that human perception can miss some details about the diagnose process, data mining techniques can aid to make early diagnosis and cure diseases.

In this research, models has been developed in order to determine parameters which has significant roles in diagnosing anemia and anemia types, with laboratory data which consists of 6 months of measurements. A data mining research in the medicine sector has been made in order not to miss the diagnosis in the plentifulness of medical measurements. Open source data mining software (WEKA) has been used in modeling process. Classification algorithms were applied to the data and it has been seen that rules.PART and trees.J48 algorithms had the highest classification success. In clustering analyses it has been seen that FilteredClusterer algorithm has 100% concurrency with data set. Association rules has been produced with Tertius algorithm and they have been verified by a specialist. It has been concluded that FilteredAssociator association rules and rules from decisions trees can't be used practically in diagnose processes.

2015, xi + 208 pages

Key Words: Data mining, anaemia, types of anaemia, WEKA

TEŞEKKÜR

Bu araştırmanın konusu, çalışmaların yönlendirilmesi, sonuçların değerlendirilmesi ve yazımı aşamasında yapmış olduğu büyük katkılarından dolayı tez danışmanım Sayın Yrd. Doç. Dr. Ertuğrul ERGÜN'e, tıbbi verilerin toplanması, düzenlenmesi ve yorumlanması aşamalarında yardımlarından dolayı Sayın Biyokimya Uzmanı Dr. Göksenin ÜNLÜGÜZEL ÜSTÜN'e, manevi destekleriyle büyük bir sabır ve sevgi ile çalışmam boyunca beni yalnız bırakmayan sevgili kızım ve eşime, bana her konuda güvenen ve yüreklendiren sevgili anne ve babama teşekkür ederim.

Asiye Betül CİGA
AFYONKARAHİSAR, 2015

İÇİNDEKİLER DİZİNİ

Sayfa

ÖZET	i
ABSTRACT	ii
TEŞEKKÜR	iii
İÇİNDEKİLER DİZİNİ.....	iv
SİMGELER ve KISALTMALAR DİZİNİ	vi
ŞEKİLLER DİZİNİ	vii
ÇİZELGELER DİZİNİ.....	x
1.GİRİŞ.....	1
2.LİTERATÜR BİLGİLERİ	3
2.1 Temel kavramlar	3
2.1.1 Veri Bilgi ve Enformasyon Kavramları	3
2.1.2 Veri Tabanı Yönetim Sistemleri	4
2.1.3 Veri Tabanlarında Bilgi Keşfi.....	5
2.1.4 OLAP VE OLTP.....	6
2.2 Veri Ambarı.....	7
2.2.1 Veri Ambarı Tanımı.....	7
2.2.2 Veri Pazarı (Data Mart).....	12
2.2.3 Veri Ambarının Amaçları ve Kullanım Nedenleri.....	12
2.2.4 Veri Küpleri	12
2.2.5 Operasyonel Veri Tabanı ile Veri Ambarının Karşılaştırılması	13
2.2.6 Veri Ambarı Oluşturmada Karşılaşılan Zorluklar	15
2.2.7 Veri Ambarı Mimarileri	16
2.3 Veri Madenciliği.....	16
2.3.1 Veri Madenciliği Tarihçesi ve Kullanım Amacı.....	16
2.3.2 Veri Madenciliği Tanımı.....	18
2.3.3 Veri Madenciliği Metodolojisi.....	21
2.3.4 Veri Madenciliği Süreci (CRISP-DM)	23
2.3.5 Veri Madenciliği ve Diğer Disiplinler	41
2.3.6 Veri Madenciliği Yöntemleri	44
2.3.7 Veri Madenciliği Teknikleri.....	55
2.3.8 Veri Madenciliği Uygulama Alanları	69
2.3.9 Veri Madenciliği Uygulamalarında Karşılaşılan Sorunlar.....	86

2.3.10 Veri Madenciliği Tekniklerinin Performans Ölçütleri.....	91
2.3.11 Veri Madenciliği Programları ve WEKA	98
2.4 Tıp ve Veri Madenciliği	111
2.4.1 Veri Madenciliği ve Etik.....	111
2.4.2 Tıpta Veri Madenciliği Gereksinimi ve Tıbbi Verilerin Özellikleri	112
2.4.3 Tıbbi Verilerle Veri Ambarının Oluşturulması.....	114
2.4.4 Tıbbi Verilerle Veri Madenciliği Çalışması.....	115
2.4.5 Metin Madenciliği Uygulamaları.....	116
2.4.6 Veri Madenciliğinin Sağlık ve Tıp Alanında Kullanım Aşamaları ve Uygulamalar.....	118
2.5 Anemi (Kansızlık)	133
2.5.1 Anemi Tanımı	133
2.5.2 Anemi Klinik Belirtileri	135
2.5.3 Anemi Klinik Bulguları	136
2.5.4 Anemi Sınıflama ve Tanısal Yaklaşım	137
2.6 Çalışmanın Amacı, Önemi ve Kapsamı	143
2.6.1 Çalışmanın Amacı.....	143
2.6.2 Çalışmanın Önemi	143
2.6.3 Çalışmanın Kapsamı	144
3. MATERYAL METOT	145
3.1 Verilerin Hazırlanması ve Veri Seti Özellikleri	145
3.2 Modelin Oluşturulması ve Uygulanması.....	154
3.3 Modelin Değerlendirilmesi ve Yorumlanması	155
4. BULGULAR.....	159
5. TARTIŞMA VE SONUÇ	183
6. KAYNAKLAR	191
ÖZGEÇMİŞ.....	207
EKLER	208

SİMGELER ve KISALTMALAR DİZİNİ

Simgeler

µg/dL	Mikro gram/ Desilitre
fL	Femto Litre
g/dL	Gram/Desilitre
mg/dL	Miligram/ desilitre
mm3	Mili metre küp
ng/mL	Nano gram /Mili Litre

Kısaltmalar

ARFF	Attribute-Relation File Format
AUC	Area Under the Curve
BK	Bilgi Keşfi
C&RT	Classification and Regression Trees
CHAID	Chi-Squared Automatic Interaction Detector
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	Comma Separated Values (Text file)
DSY	Doğru olarak Sınıflandırılan Örneklerin Yüzdesi
HCT	Hemotoktit
HGB	Hemoglobin
ID3	Induction of Decision Trees
KDD	Knowledge Discovery in Databases
kNN	k-En Yakın Komşuluk Algoritması
KPSS	Kamu Personeli Seçme Sınavı
MAE	Mean Absolute Error
MCHC	Mean Corpuscular Hemoglobin Concentration
MCV	Mean Coppuscular Volume
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
ÖSS	Öğrenci Seçme Sınavı
ÖSYM	Öğrenci Seçme ve Yerleştirme Merkezi
RDW	Red Cell Distribution width
RMSE	Root Mean Squared Error
ROC	Reciever Operating Characteristic
SPSS	Statistical Packages for the Social Sciences
SQL	Structured Query Language
VM	Veri Madenciliği
VTBK	Veri Tabanı Bilgi Keşfi
VTYS	Veri Tabanı Yönetim Sistemleri
WEKA	Waikato Enviroment for Knowledge Analsis
YSA	Yapay Sinir Ağları

ŞEKİLLER DİZİNİ

	Sayfa
Şekil 2.1 Bir OLAP veri tabanında yapılabilecekler.	7
Şekil 2.2 Veri ambarı temel yapısı (Han and Kamber 2006).	8
Şekil 2.3 Veri ambarı konuya yöneliktir (Özkan 2013).	9
Şekil 2.4 Farklı biçimdeki verilerin veri ambarına taşınırken bütünleştirilmesi (Özkan 2013).	10
Şekil 2.5 Veri ambarında önceki dönemlere ilişkin veriler (Özkan 2013).	10
Şekil 2.6 Veri ambarında veri sadece okunabilir (Özkan 2013).	11
Şekil 2.7 OLAP küpü (Kocabaş 2010).	13
Şekil 2.8 Veri madenciliği çalışmasında kullanılan yöntem bilim (Alpaydın 2000).	21
Şekil 2.9 Veri madenciliği aktiviteleri (Vahaplar ve İnceoğlu 2001).	22
Şekil 2.10 Bilgi keşfi süreci (Bramer 2007).	23
Şekil 2.11 Veri madenciliği için çapraz endüstri standardı süreci (İnt.Kyn.12).	24
Şekil 2.12 CRISP-DM aşamaları ve görevleri (İnt.Kyn.12).	25
Şekil 2.13 Veri indirgeme yöntemleri (Özkan 2013).	36
Şekil 2.14 Veri sıkıştırma (Yıldırım 2008).	37
Şekil 2.15 Nitelik birleştirme (İnt.Kyn.3).	38
Şekil 2.16 VM ve diğer disiplinler arasındaki ilişkiler (Han and Kamber 2006).	42
Şekil 2.17 Yapay sinir hücresi (Esen 2009).	49
Şekil 2.18 Karar ağaçlarının genel yapısı (Çakır 2005).	51
Şekil 2.19 Karar ağacı örneği (Kocabaş 2010).	52

Şekil 2.20 Veri sınıflandırma işlemi (Pala 2013).	53
Şekil 2.21 Veri madenciliği modelleri (Ergün 2008).	55
Şekil 2.22 ROC eğrisi ve keşisen ROC eğrisi (İnt.Kyn.6).	96
Şekil 2.23 Açık kaynak kodlu VM yazılımları karşılaştırması (Kaya ve Özel 2014)..	101
Şekil 2.24 Sınıflandırma algoritmaları açısından yazılımların karşılaştırılması	102
Şekil 2.25 Kümeleme algoritmaları açısından yazılımların karşılaştırılması.....	102
Şekil 2.26 Birliktelik kural algoritmalarının karşılaştırılması (Kaya ve Özel 2014). ..	102
Şekil 2.27 Nitelik seçimi yöntemlerinin karşılaştırılması (Kaya ve Özel 2014).	103
Şekil 2.28 Veri ön işleme algoritmalarının karşılaştırılması (Kaya ve Özel 2014).	103
Şekil 2.29 Örnek bir Arff dosyası görünümü.	105
Şekil 2.30 WEKA grafik ara yüzü.....	106
Şekil 2.31 Explorer modül ekranı.	107
Şekil 2.32 Öznitelik dağılım histogramı.	108
Şekil 2.33 KnowledgeFlow arayüzü.	110
Şekil 2.34 Veri madenciliği döngüsü (Kaur and Wasar 2006).....	119
Şekil 2.35 Normositik aneminin tanısal inceleme şeması (Nalbant ve Karan 2010)..	140
Şekil 2.36 Makrositik aneminin tanısal inceleme şeması (Nalbant ve Karan 2010)....	141
Şekil 2.37 Mikrositik aneminin tanısal inceleme şeması (Nalbant ve Karan 2010). ...	142
Şekil 3.1 Ham veriler.	150
Şekil 3.2 Her bir kayıt için yapılan tetkiklerin ayrıntılı gösterimi.	150
Şekil 3.3 Eksik değerlerin tanımlanması.	151
Şekil 3.4 Verilerin normalizasyonu.....	151

Şekil 3.5 Verilerin kategorik dönüşümü.	152
Şekil 3.6 Veriler normalize edilmeden csv dosyası.	152
Şekil 3.7 Normalize verilerin csv dosyası.	153
Şekil 3.8 Kategorik verilerin csv dosyası.	153
Şekil 3.9 Anemi veri setinin nitelik dağılımı.	155
Şekil 3.10 Verinin sınıflandırma algoritması ile modellenmesi.	156
Şekil 4.1 Verilerin kategorik hale çevrilmesi ile tüm özelliklerin dağılımı.	160
Şekil 4.2 Sınıflandırılmış anemi türlerine göre özelliklerin dağılımı.	161
Şekil 4.3 AnemiTur veri seti için J48 algoritması karar ağacı kuralları.	162
Şekil 4.4 AnemiTur veri seti için J48 algoritmasının classifier output ekran çıktısı. ..	163
Şekil 4.5 AnemiTur veri seti için J48 algoritması karar ağacı.	164
Şekil 4.6 Anemi ve AnemiDegil veri seti için J48 algoritması karar ağacı.	168

ÇİZELGELER DİZİNİ

	Sayfa
Çizelge 2.1 OLTP Sistemler ve Veri Ambarları'nın karşılaştırılması (Çarklı 2010).....	15
Çizelge 2.2 Veri madenciliği gelişim süreci (Şaylan 2013).....	17
Çizelge 2.3 Veri madenciliğinin uygulandığı alanlar (İnt.Kyn.14).....	70
Çizelge 2.4 Hata matrisi (Confusion matris).....	93
Çizelge 2.5 Veri madenciliği için kullanılan araç çeşitleri ve örnekleri	99
Çizelge 3.1 Veri seti referans aralıkları ve boş değerler (Wu and Alan 2006, Üstün 2013)	146
Çizelge 4.1 Anemi ve AnemiDegil veri seti Bayes algoritmaları ve sonuçları.....	169
Çizelge 4.2 Anemi ve AnemiDegil veri seti Functions algoritmaları ve sonuçları.....	170
Çizelge 4.3 Anemi ve AnemiDegil veri seti misc algoritmaları ve sonuçları.....	170
Çizelge 4.4 Anemi ve AnemiDegil veri seti lazy algoritmaları ve sonuçları.....	171
Çizelge 4.5 Anemi ve AnemiDegil veri seti meta algoritmaları ve sonuçları.....	172
Çizelge 4.6 Anemi ve AnemiDegil veri seti için rules algoritmaları ve sonuçları.....	173
Çizelge 4.7 Anemi ve AnemiDegil veri seti için trees algoritmaları ve sonuçları.....	173
Çizelge 4.8 Anemi ve AnemiDegil veri seti için en başarılı algoritmalar	174
Çizelge 4.9 Anemi ve AnemiDegil veri seti DBSCAN kümeleme algoritması.....	174
Çizelge 4.10 Anemi ve AnemiDegil veri seti EM kümeleme algoritması.....	175
Çizelge 4.11 Anemi ve AnemiDegil veri seti FarhestFirst kümeleme algoritması.....	175
Çizelge 4.12 Anemi ve AnemiDegil veri seti FilteredClusterer algoritması	175

Çizelge 4.13 Anemi ve AnemiDegil veri seti MakeDensityBasedClusterer kümeleme algoritması	175
Çizelge 4.14 Anemi ve AnemiDegil veri seti için SimpleKMeans algoritması.....	176
Çizelge 4.15 Tüm Öznitelikler ile SimpleKMeans kümeleme algoritması.....	176
Çizelge 4.16 Anemi Türleri veri seti Bayes sınıflandırma algoritmaları ve sonuçları.	176
Çizelge 4.17 Anemi Türleri veri seti Functions sınıflama algoritmaları ve sonuçları .	177
Çizelge 4.18 Anemi Türleri veri seti Lazy sınıflandırma algoritmaları ve sonuçları...	177
Çizelge 4.19 Anemi Türleri veri seti meta sınıflandırma algoritmaları ve sonuçları...	178
Çizelge 4.20 Anemi Türleri veri seti misc sınıflandırma algoritma ve sonuçları	178
Çizelge 4.21 Anemi Türleri veri seti rules sınıflandırma algoritmaları ve sonuçları...	179
Çizelge 4.22 Anemi Türleri veri seti trees sınıfı algoritmaları ve sonuçları	179
Çizelge 4.23 Anemi Türleri veri seti sınıflandırma sonuçları	180
Çizelge 4.24 Anemi Türleri veri seti EM kümeleme algoritması sonucu	180

1. GİRİŞ

Tarihin sayfalarına baktığımızda güçlü olarak anılan toplumlarda bilginin kullanımının daha fazla olması ünlü düşünür Francis Bacon'ın "bilgi güçtür" sözünü teyit etmektedir. Bugün bilgi, sayfalara ve kitaplara sığmayacak muazzam boyutlara ulaşmıştır. Fakat güç felsefesi değişmemiş ve tahtını bilgi üzerinde bu yüzyılda da korumaya devam etmiştir. Farklı kaynak ve boyutlarda işlenmediği sürece atıl kalan veri yığınları arasında anlamlı ilişkiler kurulması, bilginin bu tahtta veri madenciliği tacını giymesi ile mümkün olabilmektedir. İmkânlar dâhilinde birçok alanda olduğu gibi, tıp alanında da veri madenciliğine kapı aralanmıştır ki, bu alanda verilerin büyüklüğü ve hayati önem taşıması veri madenciliği çalışmalarını aynı zamanda bir zorunluluk haline getirmektedir.

Veri zengini bilgi fakiri birçok kaynak ile özellikle de sağlık veri tabanları, geçmiş kayıtların analiz edilerek, ileriye dönük tahminde bulunabilmeyi kolaylaştırmayı amaçlayan çalışmalar veri madenciliğinde yapılabilmektedir.

Veri madenciliğinin tıp alanındaki kullanımının incelendiği bu tez çalışmasında, uygulama için Amasya iline bağlı bir kamu hastanesinden Temmuz 2014 Ocak 2015 tarihleri arasında farklı servislere, farklı şikâyetlerle başvuran 1 170 260 adet hasta verisi içerisinden, anemi ve benzer semptomlar gösteren kayıtlar ayrıştırılarak, WEKA (Waikato Enviroment for Knowledge Analsis, Waikato Bilgi Analiz Palatformu) açık kaynak veri madenciliği yazılımı ile sınıflandırma, kümeleme ve karar ağaçları algoritmaları analizleri yapılmış, karar kuralları uzman desteği ile incelenerek anemi teşhisinde ve sınıflandırılmasında etkili faktörlere ulaşılmaya çalışılarak, tanıyı gözden kaçırmamaya yardımcı olmak amaçlanmıştır.

Bu amaçla hazırlanan tez şu bölümlerden oluşmaktadır:

İlk olarak veri madenciliğinin yapı taşı olan veri, enformasyon, bilgi, veri tabanı ve veri ambarı kavramları açıklanarak; veri madenciliği, metodolojisi, süreci, disiplinler arası ilişkisi, yöntem ve teknikleri, bu tekniklerin performans ölçütleri ile veri madenciliğinde kullanılan programlar incelenmiştir. Daha sonra; tıpta veri madenciliğinin gereksinimi, veri ambarı oluşturulması, tıbbi verilerle veri madenciliği çalışmaları ve metin

madencilięi uygulamaları ile bu alandaki kullanım aşamaları ile bu alanda literatürde yapılan çalışmalara yer verilerek, hastalık bilgisi ve tedavisi hakkında kısaca açıklamalar yapılmıştır. Son olarak bu bölümde çalışmanın amacı, önemi ve kapsamı açıklanmıştır.

Materyal-metod kısmında ise; çalışmada kullanılacak olan verilerin önışleme aşamaları, modelin kurulması ve değeriendirilmesi hakkında bilgi verilmiştir.

Çalışma sonucunda elde edilen algoritma analizleri kendi arasında kategorize edilerek tablolastırılmış ve çizelge olarak bulgular kısmında ayrıntılı bir şekilde verilmiştir.

Son bölümde elde edilen sonuçların benzer çalışmalar ile benzerlik ve farklılıkları irdelenek, ileride yapılacak çalışmalar için öneriler tartışılmıştır.

2. LİTERATÜR BİLGİLERİ

2.1 Temel kavramlar

Bu bölümde veri madenciliğinin temelini oluşturan veri, enformasyon ve bilgi kavramları açıklanarak veri tabanı yönetim sistemleri, veri tabanlarında bilgi keşfi, OLAP ve OLTP kavramlarına yer verilmiştir.

2.1.1 Veri Bilgi ve Enformasyon Kavramları

Veri madenciliği (VM) kavramı destekleyici bilgi teknolojileri içerisinde incelendiğinden, bilgi ve yönetim sürecinin incelenmesi kavramın daha iyi anlaşılması açısından önemlidir.

Bilişim teknolojilerindeki hızlı değişim ve gelişimle bilginin beşinci üretim faktörü olarak ekonomik sisteme dâhil edilmesi bilginin önemini artırmıştır. Çok boyutlu bir kavram olan bilgi, genellikle veri kavramı ile karıştırılmaktadır (Ergün 2008). İngilizcedeki karşılığı “Knowledge” olan bilgi kavramı, çoğu kez “Information” kavramı yerine kullanılmaktadır. Enformasyon (haber/malumat); “veri” ile “bilgi” arasındaki gri alanda var olan olgu olup, bilgiye ulaşmamıza yardımcı olacak araçlardan biridir. Veri; sadece harf ve sayılar değil, ayrıca onların anlamlarını da içermektedir. Bilgi ise verinin belli bir amaca yönelik olarak işlenmesiyle elde edilir. Veri kendi başına değersiz olup, bir anlam ifade etmezken, veriler sınıflandırma, gruplandırma, analiz edilme ve dönüştürme sürecinden sonra enformasyona dönüştürülür. Bir amaca yönelik olarak bir soruya yanıt vermek için veriden çıkarılan sonuç; bilgi olarak tanımlanabilir. Veri; yapılandırılmamış, sınıflandırılmamış, ilişkilendirilmemiş ve hemen bir anlam verilemeyen enformasyon olguları olarak tanımlanırken; enformasyon işlenmiş, sınıflandırılmış veri olarak karşımıza çıkmaktadır. Bilgi ise açıklayıcı olup enformasyonun deneyim ve çevresel durumlar ile birleştirilmesi sonucunda elde edilmektedir. Han ve Kamber (2001)’e göre bilgi; potansiyel olarak ilginç ve faydalı olan veriler arasındaki ilişki olup, veriyi bilgiye çevirme işlemine veri çözümlemesi (veri analizi) denir. Yani işlenmiş enformasyon bilgi olarak tanımlanmaktadır (Alpaydın 2000, Esen 2009, İnt.Kyn.15).

Verilerin enformasyona çevrilebilmesi için zaman ve maliyet açısından çeşitli süreçler

aşağıda sıralanmıştır (Esen 2009).

- Amaca yöneltme; verilerin hangi amaçlar için toplandığının bilinmesidir.
- Sınıflandırma; verilerin analize uygun bir biçime getirilip gruplandırılmasıdır.
- Hesaplama; veriler matematiksel veya istatistiksel olarak analiz edilebilir.
- Düzeltme; verilerin hatalardan mümkün olduğunca fazla arındırılmasıdır.
- Yoğunlaştırma süreçleri; verilerin kısa ve öz görünecek bir biçimde özetlenmesi sürecidir.

2.1.2 Veri Tabanı Yönetim Sistemleri

Veri tabanı düzenli bilgiler topluluğu olup sistematik erişim imkânı olan, taşınabilir, güncellenebilir, yönetilebilir birbirleri arasında anlamlı ilişkiler bulunabilen bilgi kümesi olarak tanımlanmaktadır. Veri tabanları veriler üzerinde işlem yapmayı oldukça kolaylaştırmıştır (Ergün 2008).

Geleneksel dosya sistemleri karmaşık dosya yapıları, çoklu dosya ve kullanıcı ilişkisi ve erişiminde veriyi saklama ve erişim noktasında oluşan sorunlara cevap veremediğinden veri tabanı yönetim sistemleri (VTYS) kavramı ortaya çıkmıştır. VTYS yaklaşımında veri girişi ve depolanması veriye erişen uygulama programlarından bağımsız olup; birbirleri ile ilişkili veri ve programlar topluluğundan oluşmaktadır. Veri tabanı sistemlerinin bazı yararları bulunmaktadır (Özkan 2013):

- Verinin tekrarlanması önler,
- Verinin tutarlı olmasını sağlar,
- Eşzamanlı erişimlerde tutarsızlıkların oluşmasını önler,
- Verinin güvenliğini sağlar.

VTYS belirli bir veri modeline dayanır. Veri modeli bir veri tabanı yapısının temelini oluşturmaktadır. Veriyi mantıksal düzeyde düzenlemek için kullanılan yapılar ve işlemler topluluğuna veri modeli denir. Sıra düzensel (hiyerarşik), ağ, ilişkisel, nesneye yönelik olmak üzere dört grupta toplanan veri modellerinden en yaygın kullanılan model ilişkisel veri modelidir. İlişkisel modelde varlıklar arasındaki bağıntının, içerdiği değerlere göre sağlanması esasına dayanır. İlişkisel modelde varlıklar arasındaki

karmaşık ilişkiler basite indirgenmek amacıyla veri tabanındaki tüm ilişkiler tablolar biçiminde tanımlanmaktadır (Özkan 2013). Kısaca veri tabanı yapısının oluşturulması, kullanılması, paylaşımı, depolanan verilerin izinsiz erişimlerini ve sistem arızalarını engelleyen, dağıtık veri erişimi veya eş zamanlı kullanım kontrolünü sağlayan yazılımlara VTYS denir (Han and Kamber 2006).

Karar verici için vereceği kararın doğruluğu, deneyimine, yeteneklerine ve bilgi birikimine bağlı olduğu kadar sahip olduğu veri kümesinin yeterliliğine de bağlıdır. Kararın başarısı, verilerin doğru olarak depolanması, ayıklanması, sınıflandırılması, işlenmesi ve yorumlanması oldukça önemlidir. Karar süreçlerinin karmaşıklaşması sonucunda sayısal olarak daha fazla veriye gereksinim duyulması veri depolarının büyüklüğünü elle kontrol edilemeyecek duruma getirmiştir (Ergün 2008).

VTYS teknolojisi büyük boyuttaki verileri depolama, yönetme ve bu veriler üzerinde basit arama işlemleri için büyük kolaylıklar sağlamaktadır. Fakat devasa veri yığınlarını analiz etmek ve anlamak konusunda VTYS yetersiz kalmaktadır. Bunun için veri analiz ve bilgi keşif işlemlerine ihtiyaç duyulmuştur.

2.1.3 Veri Tabanlarında Bilgi Keşfi

Veri tabanı kavramı ve teknolojisi, verilerin belli bir amaca yönelik olarak toplanması, saklanması ve işlenmesini sağlayarak anlamlı bilgi elde edilmesi için alt yapı gereksinimine çözüm olmuştur (Altıntaş 2010).

Geleneksel veri toplama işleminde manuel olarak bilgisayar ortamına yapılan kayıtlar yüksek maliyet ve zaman gerektirdiğinden insanlar kısıtlı veriler ile karar vermek zorundaydı. Artık günümüzde barkod sistemleri, devlet kurumları, uydu, meteoroloji, coğrafya, bankacılık gibi her saniye kayıt yapan görüntüleme sistemlerinden ve hemen her alanda muazzam bir veri yığını ile işlem yapmak zorunda kalan insanoğlu bu verilerden anlam çıkarmak için veri madenciliğine ihtiyaç duymaktadır (Delibaş 2010). Öyle ki bu veriler birçok karar verici konumunda bulunanlar için dikkat edilemeyen veya fark edilemeyen durumlar için öngörücü mahiyette olup farklı boyutları göz önüne alarak en uygun kararı vermeyi kolaylaştırmaktadır.

Otomatikleştirilmiş veri toplama araçları ve gelişmiş veri tabanları; veri tabanlarında,

veri ambarında ve diğer bilgi depolama alanlarında depolanmış durumda olan çok büyük miktarda veri oluşmasına sebep olmaktadır. Dijital medyada kaydedilen ve saklanan verinin boyutu olağanüstü bir şekilde artmış ve giderek artmaktadır. Yılda iki exabite (10^{18}) üzerinde veri üretilmekte, sabit bir fiyattan depolama kapasitesi her dokuz ayda bir ikiye katlanmaktadır. Bu verilerden çok küçük bir kısmı insanlar tarafından incelenip analiz edilebilmektedir. Devasa boyuttaki verilerden anlamlar çıkarmak için “Bilgi Keşfi” bir ihtiyaç durumundadır. Çözüm ise veri ambarı ve VM olarak karşımıza çıkmaktadır (İnt.Kyn.13).

VM ve veri tabanı bilgi keşfi (VTBK) süreci literatürde zaman zaman birbirlerinin yerine kullanılan kavramlar olarak karşımıza çıkmaktadır. VM; VTBK sürecinin bir adımı olduğu halde birçok çalışmada sürecin tümünü anlatmak için kullanılmaktadır. VTBK süreci veri kümesi oluşturma, veri temizleme, birleştirme, seçme, dönüştürme ve VM ile desen değerlendirme ve sunum gibi interaktif ve kullanıcı tarafından karar verilen tekrarlı bir süreçtir.

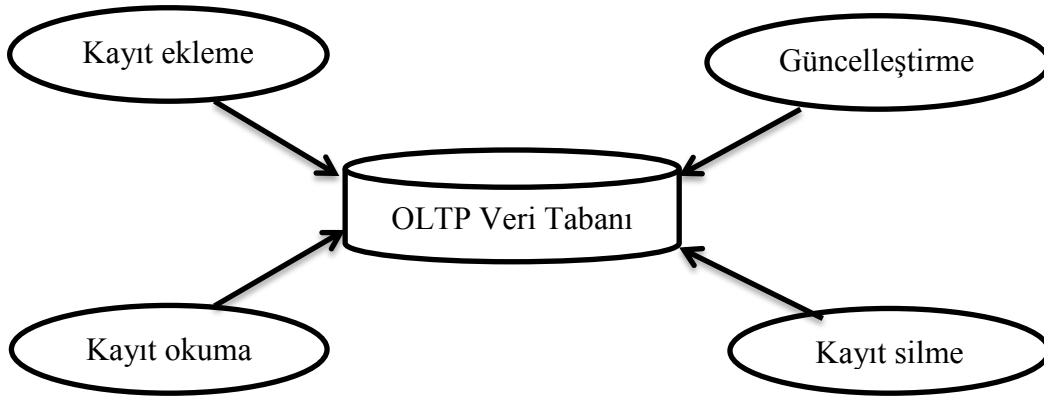
2.1.4 OLAP VE OLTP

Veri ambarında oluşturulan verinin analizi yapılabilir. OLAP (Online Analytical Processing, Çevrimiçi Analitik İşleme); çok boyutlu bir uzay üzerinde tanımlanan ve birçok boyutları olan küpler biçiminde gösterilen çok boyutlu veriler içerisinde derinlemesine farklı boyut analizleri yapılmasını sağlayan sorgu bazlı bir metot olarak tanımlanabilir. Diğer bir tanıma göre; her boyutu veride bir alana karşılık gelen çok boyutlu bir küp olarak veriye bakmayı sağlayan, verinin boyut bazında gruplanması, boyutlar arasındaki korelasyonu incelemesi, sonuçların grafik ve rapor olarak sunabildiği programlara OLAP denir. OLAP çok boyutlu dizi sunumunu kullanırken, ilişkisel veri tabanı veriyi tablolara yerleştirir. Böyle bir veri sunumu ise veri keşif ve analiz işlemlerini daha kolay bir hale getirmektedir (Ergün 2008, Altıntaş 2010, İnt.Kyn.5).

Veri madenciliğinde amaç; bilgi çıkarma sürecinde kullanıcının katkısını olabildiğince azaltmaktır. Çünkü OLAP programlarını kullanırken olası sonuçlar kullanıcının düşünebildiği sorgularla sınırlı iken; veri içerisinde hiç akla gelmeyecek sonuçlar ve ilişkiler veriler arasından çıkabilmektedir. Bu da tam olarak veri madenciliğinin esas

amacına hizmet etmektedir. OLAP uygulamaları sayesinde; büyük miktardaki verinin anlık raporlanması, müşterinin teknik desteğe ihtiyaç duymadan bilgiyi farklı boyutlar ile değerlendirmesi, yeni rapor ihtiyaçlarının hızlı bir şekilde karşılanması, farklı ortamlara XML ile kolay veri aktarımı, Excel ve pivot tabloları ile entegrasyon, kolayca PDF ve Excel'e aktarım mümkün olmaktadır (Yurtay 2008).

Bir kurumun günlük verilerinin işlendiği ortamlara OLTP (Online Transaction processing, Çevrimiçi Hareket İşleme) sistemler adı verilir. Bir OLTP veri tabanının yapabilecekleri işlemleri bir işletme üzerinden örnekleyecek olursak; işletmenin sahip olduğu stok sistemi ile depoya giren ve çıkan ürünler ve ödemelerin izlenmesi işlemleri her gün yapılarak veri tabanına kaydedilmektedir. Çeşitli rapor ve belgeler bu kayıtlara dayalı olarak üretilir. OLTP sistemlerine ilişkin veri tabanlarına veri kaydedilebilir, silinebilir, veriye erişilerek okunup, güncelleştirilip, raporlanabilir (Şekil 2.1) (Özkan 2013).



Şekil 2.1 Bir OLAP veri tabanında yapılabilecekler.

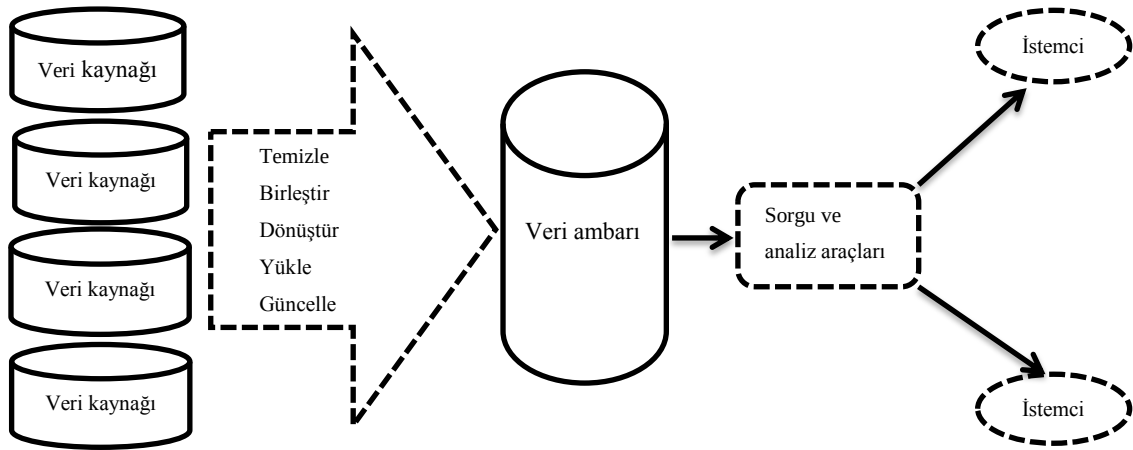
2.2 Veri Ambarı

Bu bölümde veri ambarı tanımı, unsurları, amaçları, kullanım alanları, oluşturulma aşamaları ele alınacaktır.

2.2.1 Veri Ambarı Tanımı

VM çalışmaları için veri ve veri tabanına ihtiyaç vardır. Ancak veri tabanı üzerindeki verilerin düzenlenmesi işlemlerinden sonra VM çalışmaları yapılabilir. Belirli bir

döneme ait, yapılacak çalışmada konuya yönelik ve sadece okunabilen veri tabanlarına veri ambarı denir (Silahtaroglu 2013). Ayrıca veri ambarı karar destek sistemleri için tasarlanan bir ortam olarak da tanımlanabilir. Veri ambarları sahip olunan verinin, eskileri de dâhil olmak üzere karar destek amacıyla kullanılmasını sağlar ki, bunun anlamı; var olduğu halde kullanılmayan geçmişe dönük verilerin de kullanılıp çözümlenebilir hale getirilecek olmasıdır. Veri ambarı genellikle çok boyutlu veri tabanı yapısında modellenir. Her boyut bir özellik ya da özellikler kümesine bağlanır ve her hücre bir veri değerini içerir. Veri ambarının fiziksel yapısı ilişkisel veri deposu veya çok boyutlu veri küpü şeklinde olabilir. Bir veri küpü, verilere çok boyutlu olarak göz atma ve özetlenmiş veriye hızlı erişim ve önişleme imkânı sunmaktadır. Veri ambarları özetlenmiş verinin işlenmesine olanak tanıdıkları ve çok boyutlu görünüme sahip oldukları için çevrim içi analitik işlem (OLAP) için çok uygundur (Han and Kamber 2006, Özkan 2013).



Şekil 2.2 Veri ambarı temel yapısı (Han and Kamber 2006).

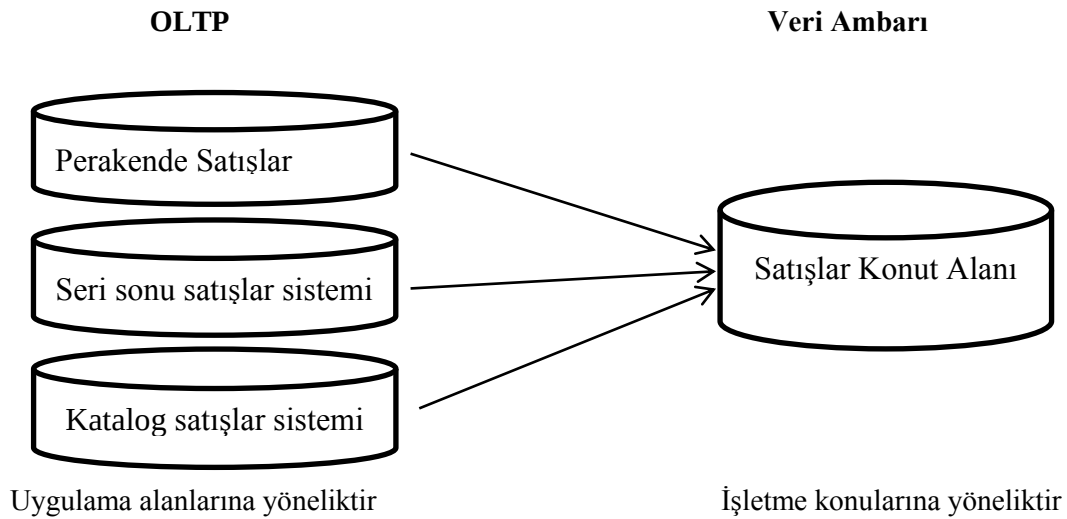
Veri ambarı; veri temizleme veri birleştirme, veri dönüşümü, veri yükleme ve periyodik veri güncelleme işlemlerinden oluşup, veri ambarının şematik gösterimi Şekil 2.2’de görülmektedir. Veri ambarı farklı kaynaklarda tutulan verilerin zaman boyutunda birbirleri ile ilişki kurmasını sağlamak için tutarlı ve doğru verilerin aynı çatı altında toplanmasını sağlayan sistemlerdir (Düzgünoğlu vd. 2006, Han and Kamber 2006).

Veri ambarları kısaca organizasyonel veriye kolay bir şekilde ulaşılmasını sağlayan yapılar olup, veriyi kullanılabilir eğilim, ilişki ve profillerde sınıflandırmazlar, sadece

potansiyel bilgiye sahip veri tabanlarıdır. Veri ambarından veriyi çekebilmek için hangi verinin gerekli olduğu ve bu verinin nerede olduğunu tespit etmek oldukça önemlidir (Akpınar 2000, Akbulut 2006).

Kısaca veri ambarları karar verme sürecinde yöneticilere destek vermek üzere hazırlanmış; konuya yönelik, bütünselik, belirli bir döneme ve zaman dilimine ait olan (zaman boyutu olan), geçici ve uçucu olmayan (sadece okunabilen) veri topluluğudur. Veri ambarının özellikleri konuya yönelik, bütünselik, zaman boyutu, geçici ve uçucu olmayan başlıkları altında incelenebilir (Özkan 2013).

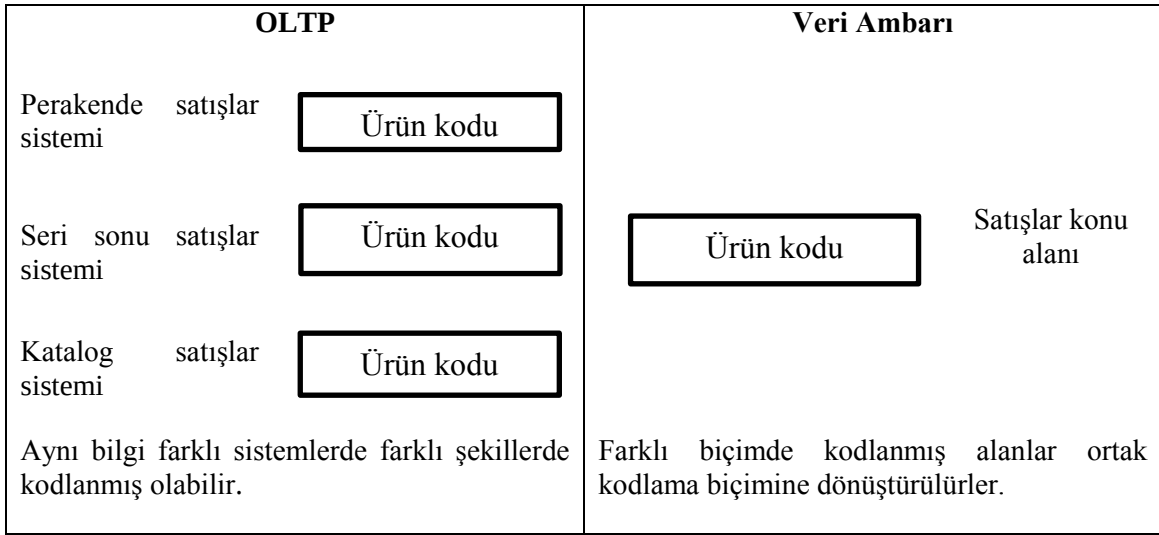
Konuya Yönelik: Aynı olayı ve aynı varlığı ilgilendiren veriler birbirlerine bağlanmışlardır. Veri ambarı işletmelerdeki konular üzerine yoğunlaşırken, klasik işlemsel sistemler; işletmelerdeki süreç, fonksiyon ve işlemler üzerinde durmaktadırlar. Örneğin bir veri ambarı hastane, hasta vs. gibi varlıklarla ya da teşhis, tedavi gibi olaylara yönelik düzenlenmiş olabilir. Veri ambarında yer alan verinin uygulanması ve tasarımı konulara göre düşünülür (Silahtaroglu 2013).



Şekil 2.3 Veri ambarı konuya yöneliktir (Özkan 2013).

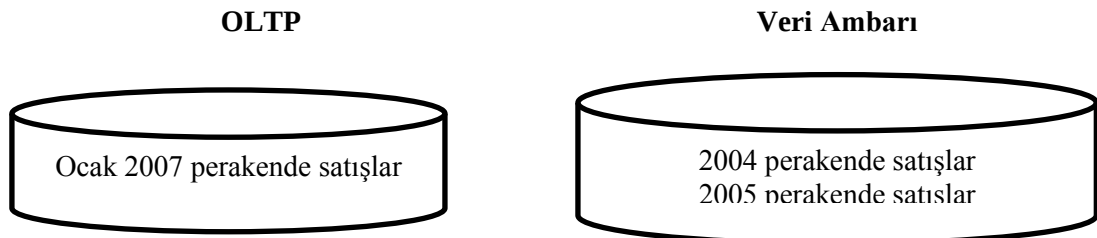
Bütünselik: Bir araya getirilen birden fazla veri tabanı yahut düz sayfalar, internet sayfaları gibi kaynakların veri tabanına aktarılmasıyla veri tabanı bütünsleştirilmiştir. Bütünsleştirme ölçü birimlerinin seçiminde tutarlılık, sayısal değerlerin fiziksel gösteriminde tutarlılık ve verilerin kodlanmasında görüş birliğine varılması şeklinde

farklı biçimlerde olabilmektedir. Örnek vermek gerekirse; cinsiyet ile ilgili bir alan veri tabanı tasarımcısının tercihinine göre bir uygulamada; “E” ve “K” kodları ile diğer bir uygulamada 1 ve 0 değerleri ile bir başka uygulamada ise “Kadın” ve “Erkek” ifadeleri ile kodlanmış olabilir. Veri ambarında kullanılacak kodlama tercihinine göre diğer kullanım kodlarının tümü tercih edilen kodlama türüne dönüştürülmek zorundadır. Şekil 2.4’de bu durum gösterilmektedir. Farklı tarih biçimleri olmasından ötürü tarih alanlarında veri bütünleştirilmesi en sık rastlanan sorunlar arasındadır. Bununla beraber tekrarlanan veriler tek bir alanda toplanarak ve gerekli normalizasyon uygulamaları yapılır.



Şekil 2.4 Farklı biçimdeki verilerin veri ambarına taşınırken bütünleştirilmesi (Özkan 2013).

Zaman Boyutu (Belirli bir döneme ve zaman dilimine ait): İşlemsel sistemlerde veri sadece o anda var olan bir veri ve değerdir.



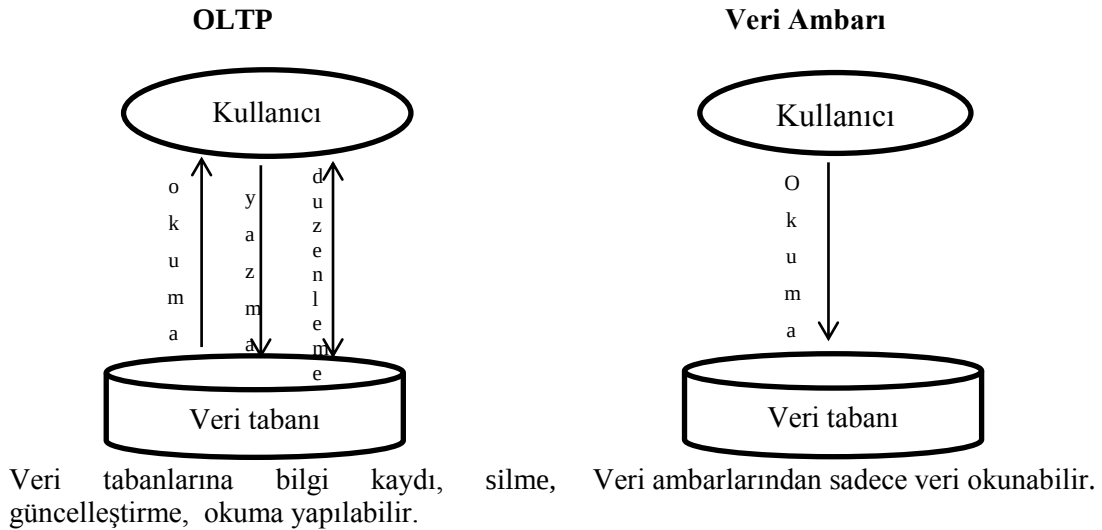
O döneme ilişkin veriler veri tabanında bulunur.

Önceki döneme ait verilerde veri tabanında yer almaktadır.

Şekil 2.5 Veri ambarında önceki dönemlere ilişkin veriler (Özkan 2013).

Veri ambarlarındaki veri ise sadece o andaki değerlerle değil, ayrıca geçmişteki değerlerle de ilgilidir (Şekil 2.5). Veri ambarı verinin zaman içinde aldığı değeri de çözümlmeye katarak, verileri zaman içindeki bir noktayla birleştirilerek değerlendirilir. Her veri doğrudan veya dolaylı olarak zaman değişkeni ile ilişkilendirilerek verilerin zaman içindeki değişimini gösterecek (örneğin son 10 yıl) raporlamaya uygun haldedir. Veri ambarındaki veri; işlemsel sistemlerdeki kayıtların belirli dönemlerdeki anlık görüntüleridir.

Geçici ve Uçucu Olmayan (Sadece okunabilen): Veri ambarları işlemsel veri tabanlarından, düz dosyalardan elde edilmiş, geçmişte belirli bir döneme ait, yeni veri giriş çıkışına uygun olmayan yani değişmeyen veriler topluluğudur. Çünkü veri ambarları veri yönetiminin gereksinimlerine cevap vermek üzere tasarlandığı veri ekleme, güncelleştirme gibi günlük işlemler yapılamaz. Şekil 2.6’da görüldüğü gibi veri ambarları salt okunabilirken; işlemsel veri tabanları karakteristiği gereği verileri değiştirebilir, silebilir ve yeni veri eklemesinde bulunabilir. Veri ambarında verinin yüklenmesi ve ona erişim olmak üzere iki işlem bulunmaktadır. Bundan dolayı veri ambarı ve işlemsel veri tabanlarının veri modelleri birbirlerinden oldukça farklı olacaktır (Silahtaroglu 2013, Özkan 2013).



Şekil 2.6 Veri ambarında veri sadece okunabilir (Özkan 2013).

Hand vd. (2001)'nin belirttiği gibi veri ambarının çeşitli veri tabanlarındaki verileri tek

bir diske atarak oluşturulması her zaman mümkün olan bir durum değildir. Yukarıda belirtildiği gibi toplanan bu özelliklerin uygulamadaki kullanım tutarsızlıklarını ortadan kaldırmak, özelliklerin ve değerlerin anlamlılıklarını öğrenmek gibi birçok bütünleştirme ve dönüştürme işlemlerinin uygulanması gereklidir. İşlemsel veri tabanlarından veri ambarı oluşturmak çok masraflı ve uzun süreli bir çalışmadır (Çataloluk 2012). Bu yüzden VM çalışmalarında her zaman bir veri ambarı olması zorunluluğu bulunmamaktadır.

2.2.2 Veri Pazarı (Data Mart)

Bilgilerin kullanılacağı alanlara göre veri ambarında yer alan bilgiler ayrı alt depolara dağıtılabilirler. Genellikle işletmelerin içindeki departmanların kullanımına göre veri ambarlarında alt depolar (Data-Mart) oluşturulur (Dinçer 2006). Bazı durumlarda veri ambarlarının belirli organizasyonların işlevlerini ve kullanımına göre çeşitli veri ambarları türetilir ve bu tür veri ambarları veri ambarı özeti (Data Mart) olarak adlandırılmaktadır (Çarklı 2010).

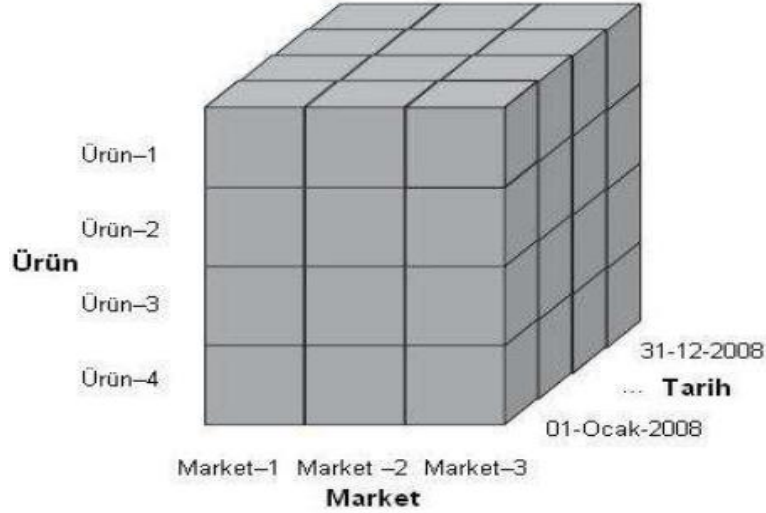
2.2.3 Veri Ambarının Amaçları ve Kullanım Nedenleri

Çok boyutlu verilerin analizi, müşteri ilişkileri yönetimi, istatistiksel analiz ve raporlama gibi VM işlemleri veri ambarları üzerinde karar destek sistemi olarak kullanılırlar (Han and Kamber 2001). Bu işlemlerin tamamına yakını işlevsel veri tabanlarında da gerçekleştirilmekte olmasına karşın, veri ambarı kurma ve kullanmanın temel amacı her iki sistem içinde yüksek performans elde etme isteğidir. Sıralama, arama ve hazır sorguları çalıştırma işlemleri için işlevsel veri tabanları kullanılırken, özel veri organizasyonu, özetleme ve çabuk erişim için veri ambarları optimize edilir. Veri tabanı performansının; veri ambarı kurulmadığı durumlarda ciddi düşüşler yaşaması karar destek işlemlerinin doğruluktan uzaklaşmasını da etkilemektedir (Dinçer 2006).

2.2.4 Veri Küpleri

Veri küpü olarak adlandırılabilir n boyutlu matrisler veri ambarı teknolojilerinin kullandıkları önemli araçlardan bir grubu temsil ederler. Bu matrislerin $n=3$ olduğu kübik üç boyutlu durum kolaylıkla canlandırılabilir. Veri ambarı ve OLAP araçları çok

boyutlu veri modeline dayanır. Çok boyutlu veri modeli, verileri bir küp şeklinde ele alarak, verilerin çok boyutlu olarak yapılandırılmasını ve görüntülenmesini sağlar.



Şekil 2.7 OLAP küpü (Kocabaş 2010).

Klasik kayıt olarak satır ve sütundan oluşan iki boyutlu yapıya, üçüncü bir boyut ilave edildiğinde veri küpü elde edilmiş olur. Veri küplerinde bilgi depolanmasının basit bir gösterimi Şekil 2.7’de görülmektedir (Meriç 2004).

2.2.5 Operasyonel Veri Tabanı ile Veri Ambarının Karşılaştırılması

İşlemsel veri tabanları çevrimiçi verilerin işlenmesi ve bunların sorgulanması özelliğinden ötürü çevrimiçi hareket işleme (OLTP) olarak da isimlendirilir. Diğer taraftan veri ambarı denilince; karar destek sistemleri, karar destek sistemleri uygulamaları ya da karar destek çözümleri adı verilen uygulamaların alt yapısı olarak anlaşılmalıdır. Karar destek sistemleri organizasyonların tarihsel verileri çözümleyip geleceğe yönelik çıkarımlar yapmak için kullanılan sistemlerdir. Bu sistemlerde yer alan bilgiler çeşitli incelemelerden ve araştırmalardan geçirilerek yöneticilerin ileride organizasyonun kârını ya da verimliliğini arttırması, gelecekte izlenecek politikaların belirlenmesi ve benzeri yönetsel kararların alınmasını kolaylaştırıp bu kararların doğru verilmesine yardımcı olurlar (Tantuğ 2002). Bu sistemler veriyi farklı kullanıcıların farklı ihtiyaçlarına göre düzenleyip sunmaktadırlar. Bu tür sistemlerde çevrimiçi analitik işlem süreci (OLAP) olarak adlandırılır. OLTP sistemleri veri ambarları için veri kaynağı oluşturan sistemlerdir. Karar destek sistemleri veriyi genelde

OLTP sistemlerinden elde ederek gerekli depolama ve arşivleme işlemlerini yerine getirmektedir. Bu karar destek sistemleri “Bilgi Tabanlı Sistemler (Knowledge Based)” olarak da bilinirler. Bu sistemlerde veriler, OLTP sistemlerine oranla çok daha büyük boyutlardadır. Dolayısıyla bu verilerin incelenmesi ve bu incelemelerden sonuçlar çıkartılması, sistem kaynaklarını aşırı kullanmakta ve uzun süre almaktadır. Bundan ötürü Karar Destek Sistemlerinde yapılacak incelemelerin ve araştırmaların ve performansını arttırmak için bir takım önlemler alınarak iyileştirmeler yapılmıştır (Tantuğ 2002). OLTP sistemler girdi/çıkış yoğunluklu, veri ambarı ise sorgulama yoğunluklu işlemleri kapsamaktadır. Bu nedenle her iki sistemin amacı ötesinde, yapıları ve işleyiş biçimleri de farklıdır. Genel olarak OLTP ve veri ambarı arasındaki farklılıkları aşağıdaki şekilde sıralamak mümkündür (Han and Kamber 2006, Dinçer 2006, Özkan 2013):

- OLTP sistemleri gündelik işlemleri bilgisayar ortamına aktarmakla görevli iken, karar destek sistemleri verinin analizi amacıyla oluşturulmuşlardır. OLTP uygulamaları kuruluşu çalıştıran, veri ambarı uygulaması ise kuruluşu yol gösteren oluşumlardır.
- OLTP sistemleri gündelik işlemlerin hızlı ve güvenilir bir biçimde yürütülmesi amacıyla optimize edilmişken, veri ambarı ise toplanmış olan veriyi hızlı bir şekilde çözümlemek amacıyla oluşturulur.
- Veri ambarı ve OLTP veri modelleri farklıdır. OLTP sistem varlık-ilişki veri modelini ve uygulama tabanlı veri tabanı tasarımını kullanırken; veri ambarı veri modeli olarak, işletmenin gereksinimlerine bağlı olarak bir “boyutsal model” olan kar tanesi modeli, yıldız şema, veri küpü ve konu tabanlı veri tabanı tasarımını benimsemektedir.
- OLTP uygulamaları girdi/çıkış işlemlerini yoğunlukla yaptıkları için karar destek sistemlerinin en çok ihtiyaç duyduğu sorgulama ve çözümleme işlemlerinde gereksiz kalmaktadırlar. OLTP ortamında çok sayıda tablo üzerinde sorgulama başarımı olumsuz etkilerken veri ambarı üzerinde bu tür sorgular kolaylıkla yapılabilir.
- OLTP sistemlerde yapılan sorgular genelde önceden belirlenmiş sorgular ve raporlar biçiminde iken; veri ambarlarında yapılan sorgular çoğunlukla önceden planlanmamış ve aniden ortaya çıkabilecek türdendir.

- OLTP sistemlerinden farklı olarak veri ambarları tarihsel veriyi de içerirler. Böylece sadece günlük veriye bakarak değil, geçmişteki verilerle geleceğe dönük trend (eğilim) çözümlmeleri yapılabilecektir.
- Veri tabanları okuma/yazma amaçlı, veri ambarı ise sadece okuma amaçlı kullanılırlar.

Çizelge 2.1 OLTP Sistemler ve Veri Ambarları'nın karşılaştırılması (Çarklı 2010).

	Konsantrasyon	Veri Ambarı
İşlev	Veri işleme	Bilgi çıkışı
Çalışma özelliği	Güncelleme	Çoğunlukla okuma
Çalışma şekli	Küçük ve sık işlemler	Kompleks uzun sorgular
Veri hacmi	MegaByte, GigaByte	GigaByte, TeraByte
Veri içeriği	Ham, işlenmemiş	Konsilide edilmiş
Zaman aralığı	Güncel veri	Tarihsel veri
Kullanım amacı	Veriyi işleme	Veriyi analiz etme
Kullanıcı	Son kullanıcı, bilgi işlem	Yönetici, analizci

Ayrıca Çizelge 2.1'de OLTP sistemler ile veri ambarları'nın karşılaştırılması görülmektedir.

2.2.6 Veri Ambarı Oluşturmada Karşılaşılan Zorluklar

- Veri ambarında karşılaşılan en önemli zorluklardan biri verinin kalitesinin kontrolüdür. Verilerin birbirinden farklı heterojen kaynaklardan alınması tutarlılık noktasında veri tabanı yöneticisinin dikkatine bağlıdır. Veri ambarı yöneticisi kaynak veri tabanlarından gelen değişiklikleri ve bu değişikliklerin veri ambarındaki yansımalarını göz önünde bulundurmalıdır.
- Veri ambarı tasarımı yapılırken veri kaynaklarının yahut kaynak farklılıklarının artması veya azalması tasarımın temelini değiştirmeyecek şekilde olmalıdır. Kaynaklardaki bu artma ve azalma veri ambarının yanıt verebileceği şekilde olmalıdır.
- Veri ambarı sağladığı bilgi kaynaklarının yanı sıra statik bir yapıda olup; kaynak veri tabanının gelişmesi ile performansı paraleldir. Ayrıca veri ambarının

yöneticisi veri ambarının karmaşık doğasını iyi bilmek zorundadır ki, veri ambarı yönetimi büyüklüğü ve karmaşıklığı ile orantılı olarak büyük yatırımlar gerektirmektedir.

- Birçok yönetimde olduğu gibi veri ambarı yönetimi de bir ekip işidir. Yönetim ekibinin ve yönetim fonksiyon tasarımının sağlam bir alt yapısı olması, bunun yanında ekip çalışanları; teknik olarak işe hâkim, birlikte grup ruhu ile çalışabilen ve liderlik vasfı olan kişiler arasından seçilmelidir (Doğan 2007).

2.2.7 Veri Ambarı Mimarileri

Kaynaklardan alınan verinin dönüştürülmesi, veri ambarının oluşturulması, kullanıcıların veri ambarına erişiminin sağlanması veri ambarı mimarilerinin genel karakteristik özellikleridir (Özkan 2013). Veriler ile verileri üreten veya kullanan sistemler arasındaki iletişim ve işlemlerin tamamının yer aldığı veri ambarı mimarisinde birbirleriyle ilişkili işletimsel veri tabanı/ dış veri tabanı katmanı, bilgi erişim katmanı, veri erişim katmanı, üst veri katmanı, işlem yönetim katmanı, uygulama ileti gönderme katmanı, veri ambarı katmanı, veri sunma katmanı bölümleri bulunmaktadır (Esen 2009). Bazı kaynaklara göre temel veri ambarı mimarisi, aşamalı alan veri ambarı mimarisi, aşamalı alan ve data mart mimarisi olmak üzere üç çeşit mimari bulunmaktadır. Bazı kaynaklarda veri ambarı mimarileri iki kademeli, üç kademeli, aşağıdan yukarıya kademe olarak yer almaktadır.

2.3 Veri Madenciliği

Bu bölümde veri madenciliğinin tarihi gelişimi, kullanım amacı, tanımı, metodolojisi, süreci, diğer disiplinlerle ilişkisi, uygulama alanları ve karşılaşılan sorunlar, VM yöntemleri, teknikleri ve tekniklerinin performans ölçütleri ile veri madenciliği programları ve WEKA açık kaynak VM programına yer verilmiştir.

2.3.1 Veri Madenciliği Tarihçesi ve Kullanım Amacı

Bilim adamları 1950’li yıllarda mantık ve bilgisayar bilimleri alanlarında VM teknikleri ile yaptıkları çalışmalar neticesinde yapay zekâ ve makine öğrenme alanlarını ortaya çıkarmışlardır. İstatistikçiler 60’lı yıllarda regresyon analizi gibi yeni algoritmalar geliştirmişler. Takip eden 20 yıllık süreçte verilerin sınıflandırılması ve sınıflar

arasındaki ilişkilerin kurulması için veri tabanı kavramı 90'lı yıllarla beraber veri tabanlarında bilgi keşfi, veri ambarları ile emekleme aşamasına gelmişti. Yeni teknolojik gelişmelerle beraber bugün VM çalışmaları büyük bir ivme kazanarak hemen her alanda çalışılabilen multidisipliner bir yapıdır (Şaylan 2013).

Çizelge 2.2 Veri madenciliği gelişim süreci (Şaylan 2013).

Gelişim Adımları	Cevaplanan Problem Kararı	Kullanılan Teknolojiler	Ürün Sağlayıcılar	Karakteristikler
Veri Toplama(1960'lar)	“Benim toplam kararım geçen 5 yıl ne kadardı?”	Bilgisayarlar, Teypler, Diskler	IBM, CDC	Geriye dönük, Statik veri dağıtımı
Veri Erişimi (1980'ler)	“İngiltere’de geçen Nisan ayında birim satışları ne kadardı?”	İlişkisel veri tabanları SQL, ODBC	Oracle, Sybase, Informix, Microsoft	Kayıt düzeyinde geriye dönük, dinamik veri dağıtımı
Veri Ambarlama ve Karar Destek Sistemleri (1999'lar)	“İngiltere’de geçen Nisan ayında birim satışları ne kadardı?”	OLAP, çok boyutlu veri tabanı sistemleri, veri ambarları	Pilot, comshare, Arbor, Cognos, Microstrategy	Çok düzeylerde, geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	“Gelecek ay Boston’daki birim satışlar muhtemelen ne olabilir, niçin?”	İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veri tabanları	Pilot, Lockheed, IBM, SGI, SPSS, SAS, Microsoft cv.	Geleceğe dönük enformasyon dağıtımı

İstatistikte amaç ana kütle hakkında anlamlı bilgiler elde etmek ve yorumlamaktır. VM de aynı şekilde anlamlı bilgiler elde etmek ve bu bilgileri eyleme dönüştürecek kararlar almak için kullanılır. Değişkenler arasındaki ilişkilerden çok, geleceğe yönelik sağlıklı öngörülerin üretilmesi veri madenciliğindeki temel amaçtır. Bu anlamda VM öz bilginin keşfedilmesi anlamında bir “kara kutu” bulma yaklaşımı olarak kabul edilmekte ve bu doğrultuda yalnızca keşifsel veri analizi tekniklerinden değil, sinir ağı tekniklerinden hareketle geçerli öngörüler yapmak, öngörülen değişkenler arasındaki ilişkilerin belirlenmesi mümkün olduğu için aynı zamanda sinir ağı teknikleri de kullanılmaktadır (Altıntaş 2006).

2.3.2 Veri Madenciliği Tanımı

Günümüzde katlanarak değişen ve gelişen bilgisayar ve veri iletişim teknolojilerindeki yenilikler yapılan tüm işlemlerin sayısal ortamda saklanmasına imkân sunmuştur. Geniş, çok hacimli ve dağıtık veri setlerindeki bu sayısal veriler, yeni soru ve sorunlara kapı aralamıştır. Bu veriler ne işe yarayacak yahut nerede kullanılacaktır? Biriken bu veri nasıl bilgiye dönüştürülecek? Veriler üzerinde çözümlemeler yapmak amacıyla birçok matematiksel ve istatistiksel yöntemler kullanılabilir. Özellikle ilişkisel veri tabanları ve veri sayındaki artış bu çözümlemeleri zorlaştırmaktadır. Bu tür çözümlemeler yapabilmek için yeni veri tabanı ve yeni çözümleme yöntemlerine duyulan gereksinim sonucu; verileri yönetmek amacıyla “veri ambarı” ve verileri çözümleyerek “yararlı bilgiye” erişilmesini sağlayan “veri madenciliği” kavramları ortaya çıkmıştır.

VM net olarak tanımlanamayan bir alandır. Literatürde birçok VM tanımına rastlamak mümkündür. Tanımların çoğunluğu tanımlayıcının ilgi alanına ve görüşüne dayanmaktadır (Ergün 2008). VM; MIT (Massachusetts Institute of Technology) tarafından 2001 yılında yayınlanan bildiriye göre dünyayı değiştirecek 10 teknoloji arasında gösterilmiştir (Dinçer 2006).

Veri tabanlarında bilgi madenciliği, veri analizi, örüntü analizi, veri taranması, bilgi çıkarımı ve veri arkeolojisi gibi kavramlar VM konusunda aynı anlamları içeren veya benzer anlamlara gelen farklı tanımları literatürde bulmak mümkündür.

Gartner Group’a göre VM; örüntü tanıma teknolojileri ile birlikte istatistiksel ve matematiksel tekniklerle ayıklanan zengin bir veri deposundan; anlamlı ve yeni ilişkiler, örüntüler ve eğilimler keşfetme sürecidir (Larose 2005).

VM; verilerin içerisindeki desenlerin, değişimlerin, kuralların, düzensizliklerin ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir. VM çalışmalarında sıklıkla vurgulanan geniş veri; tek bir iş istasyonunun belleğine sığmayacak kadar büyük veri kümelerini, yüksek hacimli veri; tek bir iş istasyonundaki yâda bir grup iş istasyonundaki disklere sığmayacak kadar fazla veri anlamında ve dağıtık veri ise farklı coğrafi konumlarda bulunan verileri ifade

etmektedir (Vahaplar ve İnceoğlu 2001).

VM, yüksek miktarda depolanmış verilerden matematiksel ve istatistiksel teknikler ile yeni ve anlamlı desenler ve ilişkiler keşfetme sürecidir (Akpınar 2000). Aynı kavramı Özkan (2013) kurumlarda biriken büyük ölçekli veriler içerisinde kurum için yararlı ve değerli olanların bulup ortaya çıkarılması işlemi olarak tanımlamıştır. Bu bağlamda VM kurumları karar destek sistemleri için önemli bir yere sahiptir. Öyle ki veriler arasındaki gizli ilişkileri ortaya çıkması ile ileriye yönelik öngörülerde bulunmak mümkün olmaktadır.

Berry ve Linoff (2004) ise veri madenciliğini; büyük veri tabanlarında örüntülerin, anormalliklerin ve çeşitli yapıların yarı otomatik bir sistem ile keşfi olarak tanımlamışlardır.

VM büyük ölçekli veriler arasında “değerli olan” bilgiyi elde etme tekniği olup, bu teknik sayesinde veriler arasında var olan ilişkileri ortaya çıkarmak ve gerektiğinde ileriye yönelik kestirimlerde bulunmayı mümkün kılmaktadır. Başka bir ifadeyle VM bir kurumda üretilen tüm verilerin belirli yöntemler kullanarak var olan ya da gelecekte var olabilecek gizli bilgiyi su yüzüne çıkarma süreci olarak değerlendirilir (Silahtaroglu 2013).

Yine bir başka tanımda “VM, otomatik örüntü tanıma sağlar ve geleneksel istatistiki yöntemlerle tespit edilmesi zor olan veri desenleri ortaya çıkarmaya çalışır” ifadesi yer almaktadır (Kaur and Wasan 2006). VM istatistiksel uygulamalara benzerlik göstermektedir. Fakat klasik istatistiksel yöntemler yeterince düzenlenmiş ve genelde özet olan binlerce veriler üzerinde çalışırken; VM çok daha fazla değişkenli, milyonlarca hatta milyarlarca veri ile ilgilenir. Veri sayısının fazlalığı bazı özel analiz algoritmaları ile verinin saklandığı ortamların yeniden düzenlenmesini ve geliştirilmesini gerekli kılmıştır.

Verideki trendleri, ilişkileri ve profilleri belirleyerek veriyi sınıflandıran bir analitik araç ve bilgisayar yazılım paketi olan veri madenciliğinin kümeleme, doğrusal regresyon, bayes ağları, görselleştirme ve ağaç tabanlı modeller gibi özel algoritmaları bulunmaktadır. Veri madenciliğinde uzun yıllar istatistiksel yöntemler kullanılmış olup,

bugün büyük veri kümelerindeki trend ve ilişkileri saptayabilmek için yüksek hızlı bilgisayarlar kullanılmakta, böylece VM, gizli trendleri minimum çaba ve emekle ortaya çıkarmaktadır (Akbulut 2006, Albayrak 2008).

Desenleri bulmak, veri özü hakkında genelleştirme yapmak, eğilimleri tespit etmek yoğun bir arama gerektirmektedir ve VM'de arama alternatifler sınanarak bir çözüme konumlanma işlemidir (Tuğ 2005).

Madencilik ifadesi; toprağın derinliklerinden değerli madenlerin çıkarılması gibi; terabaytlar seviyesindeki veriler arasından değerli olan bilgilerin çıkarılması işleminden ötürü kullanılmaktadır. VM; bir veya daha fazla makine öğrenmesi tekniği ile bir veri tabanı içerisinde bulunan verilerden otomatik olarak bilgi çıkarılması ve analiz edilme işlemidir (Han and Kamber 2006, İnt.Kyn.12).

VM; büyük miktardaki veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanarak aranmasıdır. Yakın geleceğin, geçmişten çok fazla farklı olmayacağını varsayarsak, geçmiş verilerden çıkarılmış kurallar gelecekte de geçerli olacak ve ilerisi için doğru tahminler yapmayı sağlayacaktır (Alpaydın 2000).

VM kendi başına bir çözüm olmayıp, çözüme ulaşmak için verilecek karar sürecini destekleyen ve çözüm için gerekli bilgileri sağlayan bir araç olup, analiste veriler arasındaki ilişkilerin bulunması konusunda yardımcı olmaktadır. Kısaca veri madenciliği (Ergün 2008);

- Verinin faydalı özetinin ortaya çıkarılmasıdır.
- Anlamlı örüntüler ve kurallar bulmak için büyük miktardaki verinin analiz ve keşfidir.
- Verinin analizi ve kaynaklarındaki benzerlik ve örüntüleri bulmak için kullanılan yazılım teknikleri ile ilgilidir.
- Verilerin analizleri ve veriden anlamlı bilgiye, güvenilir karar ve sonuçlara varmak için verinin dönüşümüdür.

VM genellikle büyük veri tabanlarında kullanılmaktadır. Çünkü küçük ölçekli veri tabanlarında veri tabanı uzmanlarının yaptıkları geleneksel sorgu yöntemleri yeterli

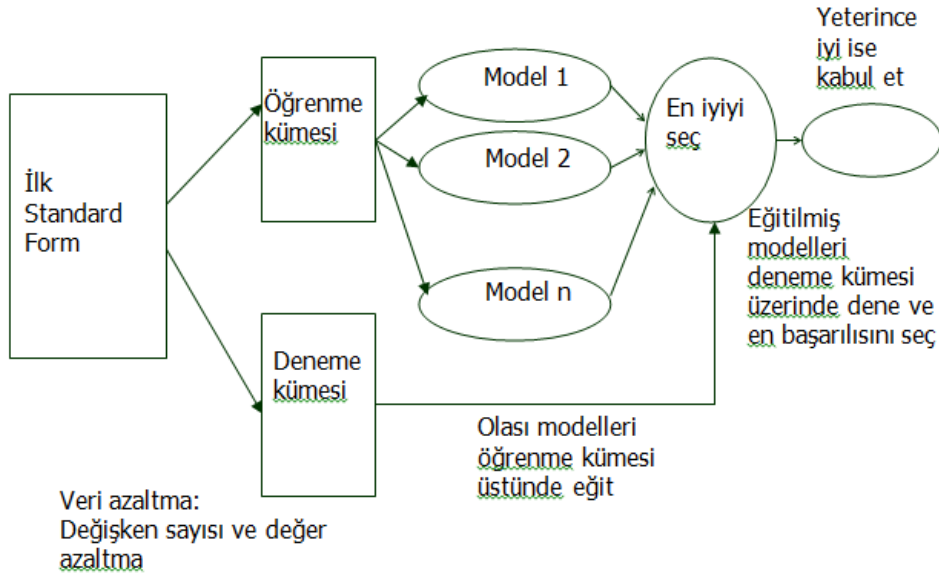
olmaktadır.

Hızla artan veri kayıtları, çok fazla ve farklı türde verinin tek bir ortamda depolanması, geleneksel eski tekniklerin ham verileri işlemede yetersiz kalması, VM bilim insanlarına verileri sınıflandırma ve gruplandırma, hipotez oluşturma, varsayımlar ve tahmin yapabilme konusunda yardımcı olması, rekabet ve gücün işletmeler için büyük öneme sahip olması veri madenciliğinin kullanımını zorunlu kılmaktadır (Hatipoğlu 2013).

2.3.3 Veri Madenciliği Metodolojisi

Büyük boyutlu veri gruplarındaki ilişkiler ağının bulunmasına yönelik yöntemlerden olan VM sınıflandırma, gruplandırma, derecelendirme gibi farklı amaçlar için farklı teknikler kullanılmaktadır (Erdem ve Özdağoğlu 2008).

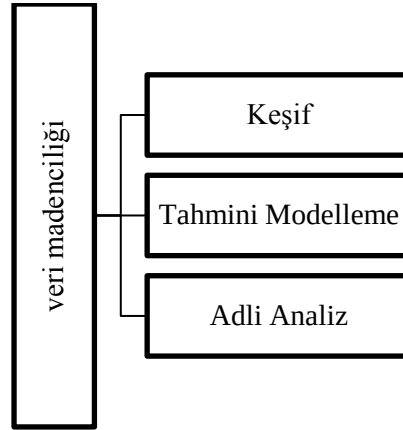
VM çalışması Şekil 2.8’da görüldüğü üzere standart form içindeki veriler öğrenme ve deneme olmak üzere ikiye ayrılır. Veri madenciliğinin disiplinler arası ilişkisi ve hemen her alanda kendine yer edinmiş olması nedeniyle yapılan uygulamalar farklı teknik ve yöntemler içermektedir.



Şekil 2.8 Veri madenciliği çalışmasında kullanılan yöntembilim (Alpaydın 2000).

Bu tekniklerden hangisinin en başarılı sonuç vereceği önceden tahmin edilemeyeceği için öğrenme kümesi üzerinde n değişik teknik ve bu n değişik teknikten, n değişik

model oluşturularak n tane model deneme kümesi üzerinde denenerek tahmin başarısı en yüksek olan model seçilir. Bu seçilen model başarılı değilse tekrar başa dönülerek çalışma tekrarlanır ve başarısız olan örnekler incelenerek; yeni alanlar ekleyerek, mevcut bilgiler değişik şekilde kodlanarak yahut değişik şekilde amaç tanımı yapılarak vb. başarının nasıl artırılabilceği araştırılır (Alpaydın 2000).



Şekil 2.9 Veri madenciliği aktiviteleri (Vahaplar ve İnceoğlu 2001).

Veri madenciliğine işlevleri açısından değerlendirilecek olursa VM aktiviteleri Şekil 2.9’de görüldüğü gibi keşif (discovery), tahmini modelleme ve adli analiz olmak üzere 3 sınıf altında toplanmıştır. Bunlar (Vahaplar ve İnceoğlu 2001);

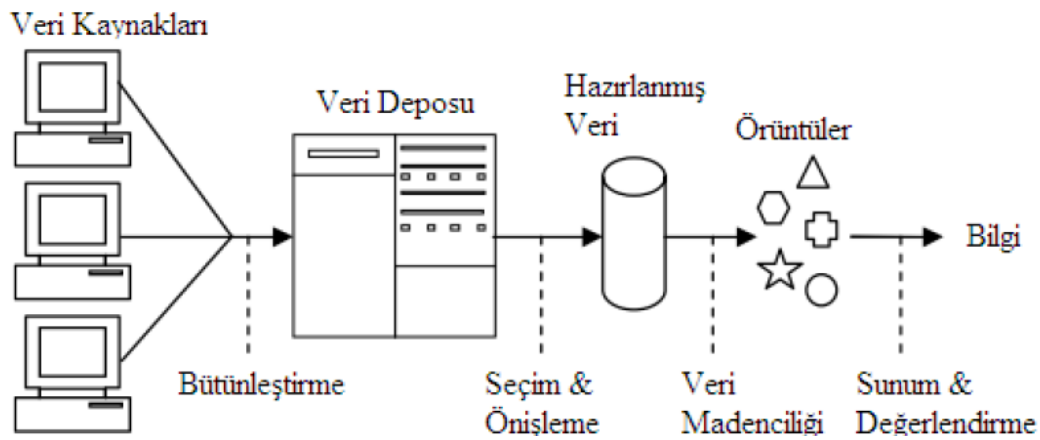
- Keşif, ne olabileceği konusunda önceden belirlenmiş bir hipotez ya da fikir olmadan, veri tabanı içerisinde gizli desenleri arama işlemidir. Geniş veri tabanlarında kullanıcının aklına gelmeyecek birçok gizli desen bulunabilir. Desenlerin zenginliği ve bunlardan çıkarılacak bilginin kalitesi buradaki asıl amaçtır.
- Tahmini modellemede, veri tabanından çıkarılan desenler geleceği tahmin için kullanılarak kullanıcının bazı alan bilgilerini bilmesine dahi kayıt etmesine izin verir. Sistem bu kayıtları, önceki kayıtlara bakarak tahmin yoluyla doldurur. Keşif verideki desenleri bulmaya yönelikken, tahmini modelleme ise bu desenler yeni veri nesnelerini bulmak için uygulanır.
- Adli analiz ise normal olmayan ya da sıra dışı veri elemanlarını bulmak için çıkarılmış desenleri uygulama işlemidir. Sıra dışı olanı bulmak için ilk önce

sıradan kısmı tespit etmek gerekmektedir. Adli analiz keşifte aranılan genel bilginin tersine, sıra dışı olan özel durumları araştırmaktadır.

2.3.4 Veri Madenciliği Süreci (CRISP-DM)

VM çalışmalarının başarılı olabilmesi için sistematik bir işlem sürecinin takip edilmesi gerekmektedir. VM işlem süreci ile ilgili birçok süreç modeli ortaya konulmuştur. Bunlar arasında Çapraz Endüstri Veri Madenciliği Standart Süreci (CRISP-DM Cross Industry Standard Process for Data Mining) projesinin süreç modeli ön plana çıkmaktadır (Şişaneci 2009).

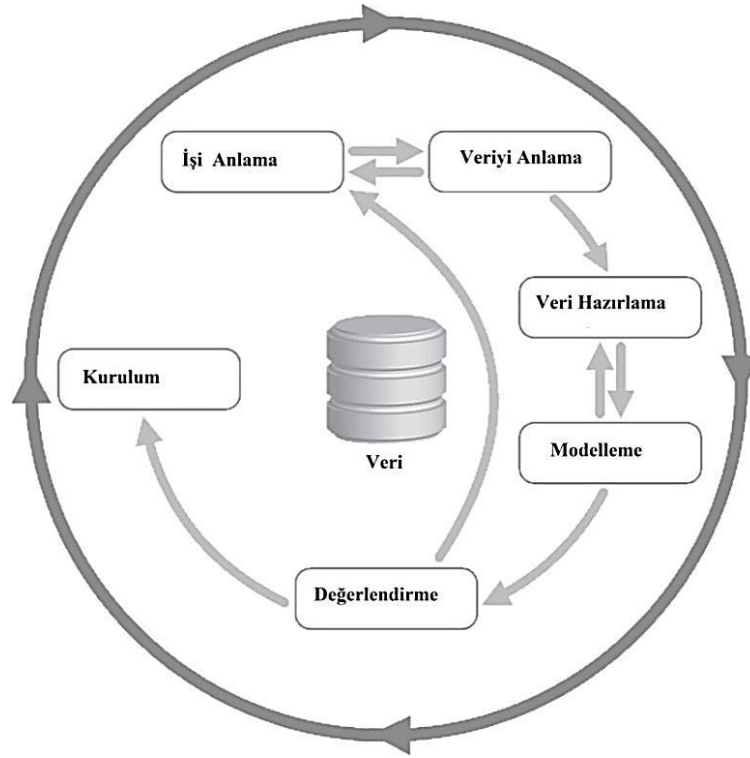
CRISP-DM Şekil 2.11’de görüldüğü gibi VM metodolojisi ve süreç modeli olup, VM projesi yürütme işinde yol haritası görevi görür. Bilgi keşfi için yapılacak veri madenciliğinin temel adımlarını tanımlayan hiyerarşik bir süreç olup, işin anlaşılmasıyla başlar ve gerekli verilerin toplanıp kavranması ile devam eden sürecin bu aşamaları birbirlerine geri bildirimli olarak çalışmakta; böylece VM projesi yaşam döngüsünün gözden geçirilmesini sağlamaktadır. Kavranan veri daha sonra hazırlanma aşamasından geçirilerek modellemede kullanılır. Bu süreçteki iki aşamada birbirini besleyecek ve geliştirecek şekilde geri bildirimli çalışır. Modelleme tamamlandığında değerlendirme aşamasına geçilir ve yapılan modelin istatistiksel olarak ne kadar geçerli olduğu, doğru sonuçlar üretilip üretilmediği ve ihtiyacı karşılayıp karşılayamadığı ölçülerek geçerliliği kanıtlanan modellerde analizler için kullanılabilir hale gelir (Larose 2005, İnt.Kyn.12).



Şekil 2.10 Bilgi keşfi süreci (Bramer 2007).

Bilgi keşfi (BK) ve VM kavramları sıklıkla birbirine karıştırılmaktadır. Veri madenciliğinin yeri BK süreci içerisinde

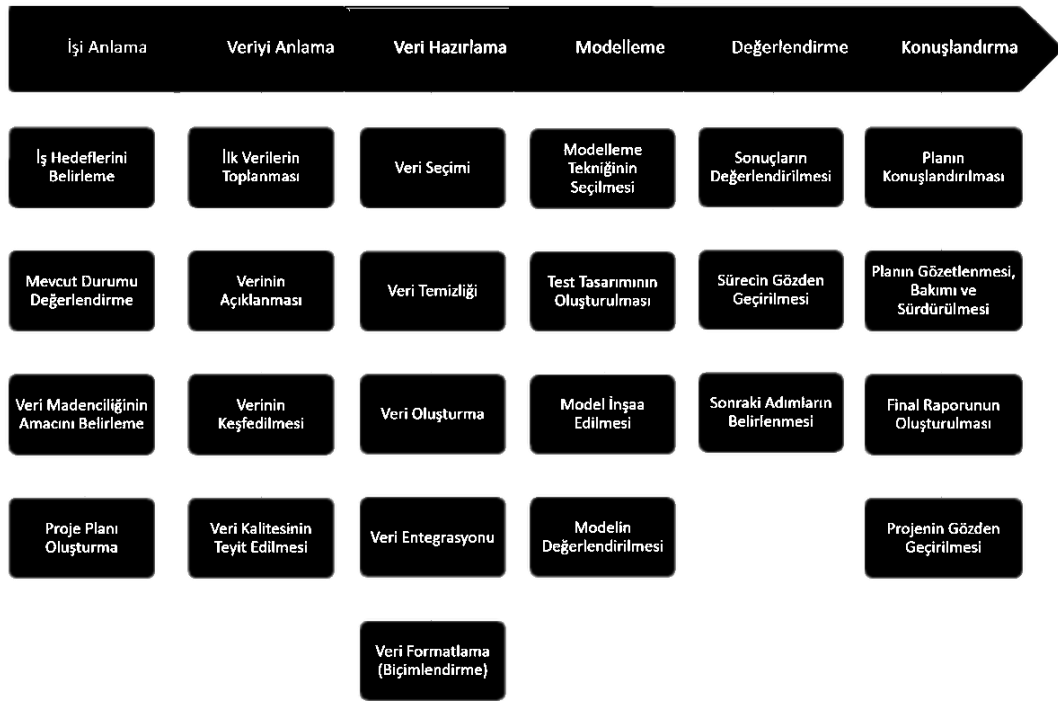
Şekil 2.10’da görülmektedir. VTBK verideki; üstü kapalı, geçerli, özgün, potansiyel olarak kullanışlı ve anlaşılabilir örüntülerin bulunması ve tanımlanması olarak tanımlanan bir süreçtir. VM; VTBK’nin yalnızca bir adımındır (İnt.Kyn.13).



Şekil 2.11 Veri madenciliği için çapraz endüstri standardı süreci (İnt.Kyn.12).

VM süreci problemin tanımlanması, verilerin hazırlanması, modelin kurulması ve değerlendirilmesi, modelin kullanılması ve izlenmesi temel aşamalarından oluşmakta olup; aşamalar arasında elde edilen sonuçlara göre önceki veya sonraki adıma gidilerek problemde iyileştirme veya farklı işlemler yapılabilir (Larose 2005).

CRISP-DM aşamaları ve bu aşamaların görevleri Şekil 2.12’de incelendiğinde tüm görevlerin veriyi hazırlamaya yönelik olduğu görülmektedir.



Şekil 2.12 CRISP-DM aşamaları ve görevleri (İnt.Kyn.12).

VM sürecine daha detaylı bakıldığında, daha karmaşık bir süreç söz konusudur (Altıntaş 2010):

- Problemin analizi ve verilerin anlaşılması
- Veri seçimi
- Veri analizi ve hazırlığı
- Veri azaltma ve dönüştürme
- Önemli özelliklerin seçimi
- Veri kümesi boyutunun küçültülmesi
- Normalleştirme
- Birleştirme
- VM metodunun seçimi
- VM süreci
- Görselleştirme
- Değerlendirme
- Bilgi kullanımı ve sonuçların hedefe uygun olarak değerlendirilmesi

2.3.4.1 Problemin Tanımlanması

VM çalışmalarında başarılı olmanın ilk şartı, üzerinde işlem yapılacak verilerin hangi işletme için hangi amaçla yapılacağının net bir şekilde anlaşılması ve tanımlanmasıdır. İyi bir VM çalışması için ilk ve en temel adım problemin belirlenmesi ve verilerin anlaşılması aşamasında problemin bilinmeyen bağımlılıklara göre değişkenleri belirlenir ve veriler arasındaki ilişkilerden hipotezler oluşturularak sonuçta bir model kurulur. Özetle bu aşamada araştırmanın amaçları ve gerekleri ayrıntılı olarak belirlenerek daha sonra bu amaçlar ve sınırlılıklar doğrultusunda VM probleminin tanımı oluşturulur (Akpınar 2000, Larose 2005, Altıntaş 2010).

VM uygulamasının başarısında en önemli adımlarından biri olan problemin iyi tespit edilmesi için bu adımda bazı önemli kuralları dikkate almak gerekmektedir (Altıntaş 2010);

- Problemin sağlayacak olduğu elle tutulur faydaları açık bir şekilde tanımlanmalı
- Muhtemel sonuç belirlenmeli
- Problem ve veriler derinlemesine anlaşılmalı
- Problem modele dönüştürülmeli
- Varsayımlar belirlenmeli
- Model döngüsel olarak iyileştirilmeli
- Model mümkün olan en basit ve anlaşılır hale getirilmeli
- Modelin kararsızlığı tanımlanmalı (girdilerdeki küçük değişikliklerin çıktılar üzerindeki ciddi değişiklikler oluşturduğu kritik alanlar)
- Modelin belirsizliği tanımlanmalı (veri kümesinin düşük güvenilirlikte çıktılar ürettiği kritik alan ve aralıklar)

Bütün bunlarla beraber bir problemin çözümünde VM kullanmanın uygun olup olmadığının kestirilebilmesi oldukça zor bir iştir. Başlangıçta 4 temel noktada sorulara cevap aranmalıdır (Roiger and Geatz 2003).

1. Problemi açık ve net bir şekilde tanımlayabiliyor muyuz?
2. Potansiyel olarak anlamlı bir veri mevcut mu?
3. Veriler nedensel olup sadece raporlama amaçları için mi uygun yoksa veriler gizli bilgi içeriyor mu?

4. Veriyi işlemenin maliyeti VM projesi sonucunda elde edilen potansiyel bilginin getirisinden daha karlı mı?

Bu soruların cevaplarıyla birlikte bilginin önceden sınıflandırılması probleme ilişkin VM uygulama konusunda karar vermeyi kolaylaştırır (Roiger and Geatz 2003).

- Yüzeysel bilgi; nedensel olup veri tabanlarında kolayca depolanır ve manipüle edilebilir. Veri tabanı sorgulama dillerinden olan SQL (Structured Query Language, yapılandırılmış sorgu dili) veriden yüzeysel bilgi çıkarmak için kullanılan en popüler araçlardandır.
- Çok boyutlu bilgi; nedensel olup çok boyutlu biçimde depolanmıştır. Çevrimiçi analitik işleme araçları (OLAP) çok boyutlu veri üzerinde kullanılan araçlardır.
- Gizli bilgi; SQL gibi sorgulama dilleri kullanılarak kolayca çıkartılamayan verideki örüntüleri veya düzensizliklerin keşfedilmesi ile ortaya çıkan bilgilerdir. VM algoritmaları bu tip örüntüleri kolayca bulabilen araçlardır.
- Derin bilgi; veri tabanında depolanmış olan sadece ne arandığı konusunda verilen yön ile bulunabilen bilgi tipidir. Günümüz VM araçları bu tip bilgiyi bulma yeteneğine sahip değildir.

2.3.4.2 Verilerin Hazırlanması (Önişleme)

Bütün VM araçlarını kullanabilmek ve veri setlerinin bilgi içeriklerinin en iyi şekilde analiz edilmesini sağlamak için veriyi bir takım dönüşüm işlemlerinden geçirmek gerekmektedir. Veri hazırlama işlemi yapıldıktan sonra hatalı tahmin oranlarının hazırlanma işleminden öncesine oranla düşmüş veya aynı oranda kalmış olması gerekir (İnt.Kyn.12).

Modelin kurulumu aşamasında veri hazırlanması süreci doğru yapılandırılmadığı takdirde verilerin sık sık yeniden düzenlenmesi kaçınılmaz olacaktır. Bundan ötürü VM proje çalışmalarında takvimin büyük bir çoğunluğu bu kısımda harcanacaktır. Bu bir kayıp gibi algılsa da, projenin farklı aşamalarında tekrar başa dönme riski göz önüne alındığında kazanç hanesinde yer almaktadır. Verimli modeller üretebilmenin ön koşulu iyi veriye sahip olmaktır. Veri kümesinin zayıf veya var olmayan verilerden hazırlanması sürecin başarısızlığının en büyük nedenidir (Akpınar 2000). Verilerin

hazırlanması yani ön işlem den geçirilmesi aşamaları; verilerin toplanması ve uyumlaştırma, verilerin birleştirilmesi ve temizlenmesi, verilerin seçimi, dönüşümü ve azaltılması şeklindedir.

a) Verilerin Toplanması ve Uyumlaştırma (Değer Biçme)

Tanımlanan problem için gerekli verilerin ve temin edileceği kaynakların belirlenmesi verilerin toplanması aşamasını oluşturmaktadır (Şen 2008). Kullanılacak verilerin farklı kaynaklardan toplanacak olması beraberinde farklı ölçü birimleri, kodlama farklılıkları, farklı zamanlara ait olma gibi veri uyumsuzluklarını da getirecektir. İyi veriler üzerinde iyi modeller kurulabileceği için toplanan verilerin ne ölçüde uyumlu oldukları uyumlaştırma (değer biçme) aşamasında değerlendirilmelidir (Akpınar 2000).

Verilerin nasıl toplanacağı ve nasıl oluşturulacağı sürecinde tasarlanmış ve gözlemsel olmak üzere iki farklı yaklaşım söz konusudur. Tasarlanmış deney; uzman (modelci) kontrolünde yapılmış verilerin oluşturulması sürecidir. Uzman kontrolü olmadan yapılırsa gözlemsel yaklaşım olarak isimlendirilir (Şaylan 2013).

Veri madenciliğinde kullanılacak verilerin aşağıdaki şartları sağlaması gerekmektedir:

- Tüm veri tek bir tabloda yer almalıdır. Bu tabloya data set (veri kümesi) adı verilir.
- Her satır bağımsız bir varlığa (Entity) ait olmalıdır.
- 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 (Berry and Linoff 2004).

Sütunlarda yer alan veri niteliklerinden her birine VM terminolojisinde özellik (attribute) adı verilir. Her özellik genelde taşıdığı niteliğin adının kısaltması ile adlandırılır. Her özellik birer değişken olup veri yapılarına göre (discrete) kesikli ve sürekli (continuous) değişkenler olmak üzere ikiye ayrılırlar (Altıntaş 2010).

Kesikli (Discrete) değişkenler; nominal (kategorik) ve ordinal (sıralı) değişkenler olmak üzere ikiye ayrılır.

Aralarında sıra, ölçü ve uzaklık ölçüsü olmayan veriler için nominal (kategorik) değişkenler kullanılır. Cinsiyet özelliği için erkek veya kadın; hastalık özelliği için hipotroidi, sarılık gibi.

Nominal değişkenlerin gösteriminde en uygun yol kısa adlarla E=Erkek, K=Kadın gibi gösterimlerdir. 1=Erkek, 2=Kadın gibi nominal değişkenlerin rakamlarla gösterimi bazı sayısal algoritmaların uygulanmasında hatalara yol açtığı için tercih edilmez.

Aralarında sıra ölçüsü olan veriler için ordinal (sıralı) Değişkenler kullanılır. Örneğin “x” ürününü satın alma ihtimali için düşük, orta, yüksek seviyeleri. Bu ihtimal ifadeleri arasında sıralı bir ilişki vardır ancak ölçüsel bir ilişki yoktur. Bu tip değişkenlerin rakamla gösterimi uygun değildir.

Bazı ordinal değişkenler ölçü ilişkisi içerirler. Genellikle bir olaya ilişkin frekans sıklığını belirten değişkenlerdir. Örneğin müşterinin aynı ürünü alma sıklığı 0-5, 6-10, 11-15 gibi durumlarda en iyi gösterim; 0-5=1, 6-10=2, 11-15=3 şeklinde rakamsal gösterimdir. Ancak bu gösterim ölçülen orijinal büyüklüğün ordinal yapısını değiştirip, verinin dağınık yapısını azaltarak tahmin modelinin doğruluğunu azalttığı için VM algoritmaları açısından kullanılması tavsiye edilmez.

Sürekli (Continuous) Değişkenler; kullanılması ve yorumlanması en kolay olan değişken tipi olup aralarında sıra, ölçü, uzaklık gibi tüm ilişkilere sahip veriler için kullanılırlar. Özelliğe göre pozitif, negatif ve sıfır değerini alıp ondalık şeklinde yazılabilirler.

Kullanıcı topladığı verilerin hangi değişken şeklinde ifade edileceğini belirler; sürekli değişkenleri eldeki verilerin önemine ve özellik çeşitliliğine göre kategorik verilere dönüştürebilir.

Veri kaynakları hayat çevrimi denilen; tasarım, toplama, değerlendirme depolama ve yönetim, işletme ve uygulama, kullanım adımlarından oluşmaktadır. VM uygulamasında VM kaynaklarını tarihsel değişim olarak sıralama gerekirse;

- Metin dosyaları
- Hesap tablosu dosyası

- Veri tabanları
 - İlişkisel veri tabanları
 - Veri ambarları
 - İşlemsel veri tabanları
 - Gelişmiş veri tabanı sistemleri
 - Nesne ilişkisel veri tabanları
 - Periyodik ya da zaman serili veri tabanları
 - Uzaysal veri tabanları
 - Metinsel veri tabanları
 - Multimedya veri tabanları
 - Akışkan veriler için veri tabanları

VM teknikleri ve yöntemleri veri kaynağına göre farklılık göstermektedir (Han and Kamber 2006). İlişkisel veri tabanları zengin VM çalışmalarında eğilim ve desenleri bularak ileriye dönük öngörülerde bulunan en sık kullanılan veri kaynaklarından (Delibaş 2010).

b) Verilerin Birleştirilmesi ve Temizlenmesi

Farklı kaynaklarda toplanan verilerin getirdiği veri uyumsuzlukları mümkün olduğu ölçüde minimize edilerek veriler tek bir veri tabanında toplanırlar. Ancak basit yöntemlerle ve özensiz olarak yapılan sorun giderme işlemleri, ileriki aşamalarda daha büyük sorunların kaynağı olacağı için bu adım titizlikle gerçekleştirilmelidir (Akpınar 2000). Gerçek hayatta toplanan çelişkili verilerin, eksik ve geniş dağılımlı verilerin üzerinde VM uygulaması yapabilecek bir veri kümesi haline dönüştürülebilmesi için; verilerin temizlenmesi, kaynaştırılması, dönüştürülmesi, kesikli hale getirilmesi, veri hacminin küçültülmesi ve kavramsal hiyerarşi oluşturulması işlemleri verinin hazırlanması kapsamında yapılan işlemlerdir. Verilerin hazırlaması işleminden önce tanımsal veri özetleme işlemi yapılmalıdır. Veri özetleme için kullanılan temel istatistik teknikler (Chakrabarti *et al.* 2008):

- Ortalama, ağırlıklı ortalama, medyan ve mod gibi merkezi eğilim ölçüleri,
- Varyans, standart sapma, açıklık, çeyrek ve ortalama ayrılış gibi verilerin dağılım ölçümü,

- Histogram, matris ve kutu grafiği gibi verilerin görsel kontrolü için grafiksel gösterimler örnek verilebilir.

Verilerin değer biçme aşamasında karşılaşılan sorunlar giderilip, veriler tek bir veri tabanında toplanırlar. Verilerin insan eliyle girişlerinden ötürü; tutarsız, hatalı ve boş olabilecekleri gibi, modelin doğru çalışması için gürültülü verilerin temizlenmesi zorunludur. Eksik değerlerin (missing value) doldurulması, veri gürültülerinin (ortalamadan çok sapan veri değerleri) giderilmesi, ayrık verilerin (outliers) tanımlanması, çelişkili verilerin doğrulanması gibi işlemlerin tamamı verilerin temizlenmesi işlemleridir (Altıntaş 2010).

Kayıp Değerler, Aşırı Değerler, Uyumsuz Veriler

Veri temizlemenin asıl amacı; gürültülü ve ilgisiz verileri veri setinden çıkarmaktır. Eksik, hatalı ve tutarsız verilere gürültü de denilmektedir. Bu verilerin söz konusu sorunlardan temizlenmesi gerekir. Kayıp, yanlış ve gereksiz verilerin ortadan kaldırılması işlemi olan veri temizleme işleminde; kayıp verilerin ortaya çıkaracağı sorunları ortadan kaldırabilmek için bazı teknikler geliştirilmiştir (Han and Kamber 2001, Çakır 2005, Çarklı 2010, Özkan 2013):

- Kayıp verilerin bulunduğu kaydı veri tabanından çıkartmak ya da benzer türdeki kayıtları iptal etmek (eğer kayıp verinin miktarı toplam verinin içinde küçük bir değerse),
- Kayıp değerlerin yerine bir sabit kullanmak. Örneğin doğum tarihi bilgisini vermemiş müşteriler varsa bunlar için DTY (doğum tarihi yok) şeklinde bir veri girişi yapılabilir. Ancak buradan çok farklı bir sonuç çıkıp; doğum tarihini vermemiş olan kişilerin bir ortak özelliği olduğu ve aynı davranışı sergiledikleri, tahmin edilemeyen bir satış fırsatını ortaya çıkardığı durumlar da olabilir (örneğin doğum tarihini yazmayan kişilerin kişisel bakım ürünlerini daha çok satın alması gibi),
- Kayıp verileri elle teker teker doldurmak (kullanılan veri tabanı küçükse ve gerçek hayatta kayıp verilere ulaşmak kolay ve zaman problemi yoksa),
- Kayıp olan verilere tüm verilerin ortalama değerlerinin girilmesi,
- Regresyon yöntemi ve karar ağacı yöntemi kullanılarak diğer değişkenlerin

yardımları ile kayıp olan verilerin tahmin edilmesi.

Eksik değer yerine ikame edilen değer gerçek bir değer olmadığından, verinin yanlılığını özgül özniteliğin aldığı değer yönünde artırır. Böylece veri seti sınıflandırmada en çok eksik değer alan öznitelikleri kural öğrenmede öne çıkaracak şekilde homojenleşecektir (Çakır 2005).

Üzerinde çalışılan veri setinde kayıp değerlerle karşılaşılması durumunda uygulanabilecek yaklaşımlar (Şaylan 2013);

- Bir öznitelik için kayıp değerlerin toplam kayıt sayısına oranı %30'dan büyükse, o öznitelikten vazgeçmek
- Bir öznitelik için kayıp değerlerin toplam kayıt sayısına oranı %30'dan küçükse, o öznitelik için kayıp değerler yerine kaydın ait olduğu sınıf ortalamasını yerleştirmek

VM sonuçlarının daha güvenilir olması için eksik verilerin tahmin edilebilmesi için bazı otomatik metotlar geliştirmişlerdir. Regresyon, bayes formalizmi, karar ağacı ve kümeleme gibi teknikler örnek verilebilir. Temizleme işlemi manuel olarak yapılabilir de büyük miktardaki verilerle yapılan çalışmalarda "Data Tools, Twins, Holmes, dFpower, Integrity, Nodupes" gibi yazılımlar kullanılmaktadır (Altıntaş 2010).

Gürültü ölçülen değişkendeki varyans veya rassal bir hata olup histogram, kümeleme analizi ve regresyon gibi teknikler kullanılabilir. Örneğin ele alınan değişken fiyat gibi sürekli bir değişken ise, gürültülü verilerin düzeltilmesi (smoothing) gerekmektedir. Binning, kümeleme ve regresyon veri düzeltme tekniklerindendir (Oğuzlar 2003).

Uyumsuz veriler ise anormal olan ve gözlemlere uymayan ölçüm ve yazım hatalarından kaynaklanan verilerdir. Modeli olumsuz etkileyecek uyumsuz verilerin tespit edilmesi, silinmesi ve uyumsuz verilere duyarsız modelleme metotları iki temel şekilde geliştirilir (Altıntaş 2010).

Bunlar veri kümesinin çok fazla ve eksik verilerin toplam veriye oranla daha az olduğu durumlarda kullanılan verilerin azaltılması yoludur. Eksik verilerin sıklıkla rastlandığı gerçek hayattan toplanan veriler, bazı VM çalışmalarında olumsuzluklara yol

açmaktadır. Eksik verilerin mutlaka dolu olması ile giderilecek sorunları çözebilmek için; verilerin azaltılması yoluyla eksik veri bulunan tüm özelliklerin silinmesidir.

Diğer bir yol ise; kayıp verilerin ortaya çıkaracağı sorunları ortadan kaldırmak için geliştirilen tekniklerde olduğu gibi; veri kümesinin küçük veya eksik verilerin fazla olduğu durumda bu eksik verilerin doldurulması işlemidir. Mevcut veriler incelenerek uygun bir değerin eksik veri yerine yazılması veri kümesini gürültülü (standarttan çok sapan) bir hale getirebilir. Gürültüyü engellemek için eksik verilerin yerine eksik veri bulunan özelliğin veri ortalaması yazılmasıyla veri kümesi daha homojen bir hale gelecektir.

c) Verilerin Seçimi

Birden fazla veri kümesi içerisinde sorgu yapılmasına uygun veri kümesi oluşturma aşaması olan veri seçiminde; verilerin tüm veri uzayını temsil edebilmesi önemlidir. Çünkü veri tabanlarının üzerinde birden fazla modelin denenmesi maliyet ve zaman gerektirdiği için, verinin bütününe temsil gücüne sahip parçalar üzerinde işlem yapmak bu sorunları ortadan kaldıracaktır (Akpınar 2000, Kaya vd. 2003, Şen 2008). Kurulacak modele bağlı olarak veri seçimi yapılır. Diğer değişkenlerin modeldeki ağırlığını azaltacak anlamlı olmayan değişkenlerin modelden çıkarılması gerekmektedir (Akpınar 2000, Tuğ 2005). Örneğin hasta ID, doktor adı, sıra no, kimlik numarası vb.

Özellik seçimi tümevarıma dayalı öğrenmede budama öncesi yapılan bir işlemdir. Ya da verilen bir ilişkinin içsel tanımını, dışsal tanımın taşıdığı bilgiyi bozmadan onu eldeki niteliklerden daha az sayıdaki niteliklere (yeterli ve gerekli) indirgemektir. Yalnızca arama uzayını küçültmekle kalmayıp, sınıflamanın kalitesini de artırır (Gülbüz 2009).

Veri seçimi aşamasında yapılması gerekenler (Şen 2008):

- Farklı ortamlardaki verilerin mevcut yapılarının incelenerek tablo yapılarının ortaya konulması,
- Gerekli verilerin uygulama için belirlenen veri ortamına aktarılması şeklinde sıralanabilir.

Doğru sonuçlar ve bu sonuçlar için yapılacak doğru yorumlar, yorumların ışığında

alınan doğru kararlar doğru verilerle sağlanabilir (Delibaş 2010).

d) Verilerin Dönüşümü

Veri madenciliğinin bazı çözümlerinde veriyi aynen kullanmak uygun olmayabilir. Değişkenlerin ortalama ve varyansların birbirlerinden önemli ölçüde farklı olmaları, büyük ortalama ve varyansa sahip değişkenlerin diğer değişkenler üzerindeki baskısını artırması ve onların model içindeki rollerini azaltması gibi nedenler çözümlemenin istenen performans ve doğrulukta olmasına engel teşkil eden durumlar arasındadır. Bu nedenle söz konusu değişkenlere bir dönüşüm yöntemi uygulayarak daha küçük aralıklarla veri dönüşümü şeklinde normalleştirme veya standartlaştırma işlemi yapılmalıdır (Özkan 2013).

Veri dönüştürme ile veriler VM için uygun formlara dönüştürme kapsamında; düzeltme, birleştirme, genelleştirme ve normalleştirme gibi değişik işlemlerden biri veya bir kaçını içerebilir (Oğuzlar 2003).

Ayrıca verilerin dönüştürülmesi aşamasında sayısal tabanlı algoritma kullanılacaksa kesikli verilerin sürekli hale getirilmesi, kategorik tabanlı algoritmalar kullanılacaksa sürekli verilerin kesikli hale getirilmesi gerekmektedir (Şaylan 2013).

Sayısal değerler üzerinde standartlaştırma (normalizasyon) işlemi; yani tüm sayısal değerlerin ortak bir aralığa dönüştürülmesi işlemi gerçekleştirilmiş ise VM çalışmasında kullanılan algoritmalarından performans en iyi ölçüde alınacaktır.

Kaynak veri içindeki farklı tip ve biçimdeki veri değerlerinin yapılacak VM çalışması doğrultusunda değiştirilmesi veri dönüşümünün asıl amacıdır (Şen 2008).

i. Ondalık (Decimal) Ölçekleme

Ele alınan değişkenlerin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme yapılır. Değişkenin mutlak değerine bağlı olarak hareket edecek ondalık sayının nokta sayısı için ondalık ölçekleme formülü aşağıda görüldüğü şekildedir (Oğuzlar 2003).

Örneğin 900 maksimum değer ise $n=3$ olacağından 900 sayısı 0,9 şeklinde ondalık ölçekleme ile normalleştirilir.

ii. Max-Min Normalizasyon

Max-Min normalizasyon ve z-score normalizasyon numerik nitelikler üzerinde uygulanan standartlaştırma yöntemleridir. Max-min normalizasyon ile numerik veriler 0 ile 1 arasındaki sayısal değerlere dönüştürülmektedir. Bu yöntem veri içindeki en büyük ve en küçük sayısal değer belirlenerek, diğerlerini buna uygun biçimde dönüştürme esasına dayanmaktadır (Özkan 2013, Silahtaroglu 2013).

Bağıntı şu şekilde ifade edilmektedir:

$$X' = \frac{X - X_{\text{Max}}}{X_{\text{Max}} - X_{\text{Min}}} \quad (2.1)$$

Bağıntıda görüldüğü gibi; Max-min normalizasyon bağlantısında X' normalleştirilmiş değeri, X gözlem değerini, X_{Min} gözlem değerinin en küçüğünü ve X_{Max} ise gözlem büyüğünün en büyüğünü ifade etmektedir (Özkan 2013).

iii. Z-Score Standartlaştırması

Z-score standartlaştırma yönteminde verilerin aritmetik ortalaması ve standart sapması dikkate alınarak yeni değerlere dönüşüm işlemi gerçekleştirilmektedir. Z-score standartlaşması için bağıntı aşağıdaki şekildedir:

$$X' = \frac{X - \bar{X}}{\sigma_X} \quad (2.2)$$

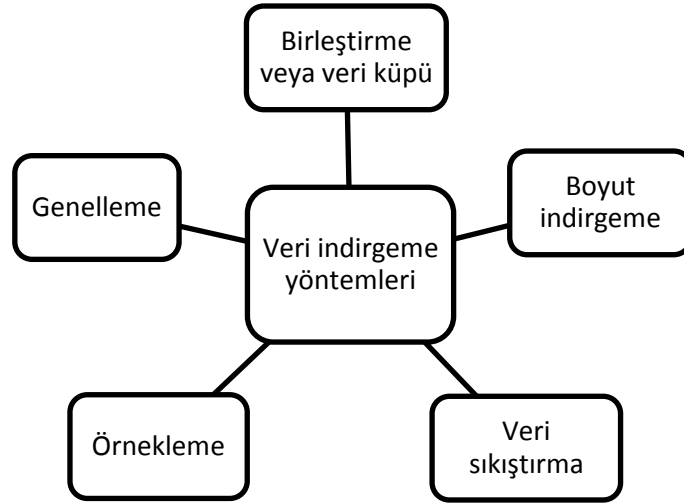
Bağıntıda; X' dönüştürülmüş değeri, X gözlem değerini, $\bar{X}$ gözlem değerinin aritmetik ortalamasını ve σ_X ise gözlem değerinin standart sapmasını ifade etmektedir (Özkan 2013).

Z-Score'da (+) işareti skorun ortalamanın üstünde olduğunu, (-) ise skorun ortalamanın altında olduğunu gösterir. Bütün veri yığınlarındaki birimlerin ortak bir birim aralığına yığılmasını sağlar, böylelikle birlerinden farklı ölçü birimlerinin karşılaştırılmasında kullanılırlar. Z-Scorunda bulunan sonuç standart sapmanın kaç katı kadar aşağısında ya da yukarısında olduğunu gösterir. Z-Scoru standart sapma ile çarpıp, ortalama ile toplarsak ham değerimize ulaşmış oluruz (İnt.Kyn.18).

e) Verilerin Azaltılması (Veri İndirgeme)

VM uygulamalarında çözümleme işlemi uzun zaman alıyorsa ve çözümlemekten elde edilecek sonucun değişmeyeceğine inanılıyorsa, başarıyı artırmak adına veri sayısı ya da değişken sayısı azaltılabilir. Seçilen örnek veri kümesindeki ilgisiz nitelikte ve tekrarlı verilerin çıkarıldığı, aynı bilgi için birden fazla değişkenin olması durumunda çalışma zamanının iyileşmesi amacıyla yalnızca bir değişkenin veriler içerisinde saklanması işlemine veri indirgeme denilmektedir. Verinin boyutu indirgendiğinden VM çalışmalarının daha hızlı olması sağlanır (Şen 2008, Çataloluk 2012, İnt.Kyn.3).

Veri indirgeme Şekil 2.13’de görüldüğü gibi çeşitli biçimlerde yapılabilir.

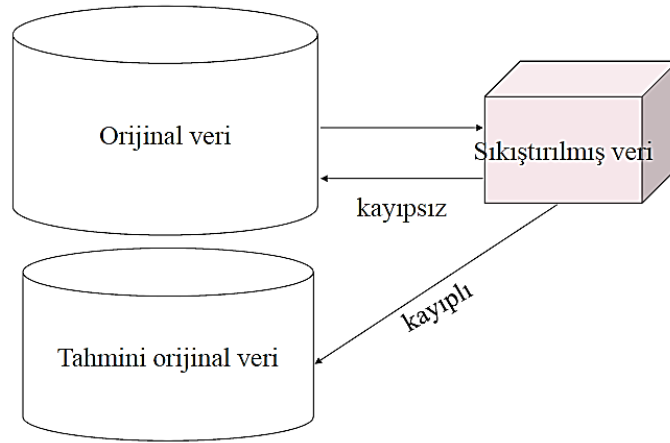


Şekil 2.13 Veri indirgeme yöntemleri (Özkan 2013).

Veriler arasında bir seçme işlemi yapılarak; veri tabanından gereksiz veriler çıkartılarak boyut azaltılması sağlanabilir. Veri kümesinde bazen gereksiz olarak yüzlerce değişken bulunabilir. Örneğin bir ürünün satışına ilişkin olarak düzenlenen bir veri kümesinde tüketiciler gereksiz bir değişken olarak yer alabilir. Bu tür gereksiz değişkenler elde edilecek örüntülerin kalitesini azalttığı gibi VM sürecinin de yavaşlamasına neden olur. Gereksiz değişkenlerin çıkarılması ile ileri veya geri yönlü olarak sezgisel seçimler yapılabilir. İleri yönlü sezgisel seçimde orijinal seçimleri en iyi temsil edebilecek değişkenler seçilerek, her bir değişkenin veya değişkenler grubunun bu kümeye dâhil edilip edilemeyeceği sezgisel olarak belirlenir. Geri yönlü sezgisel seçimde ise;

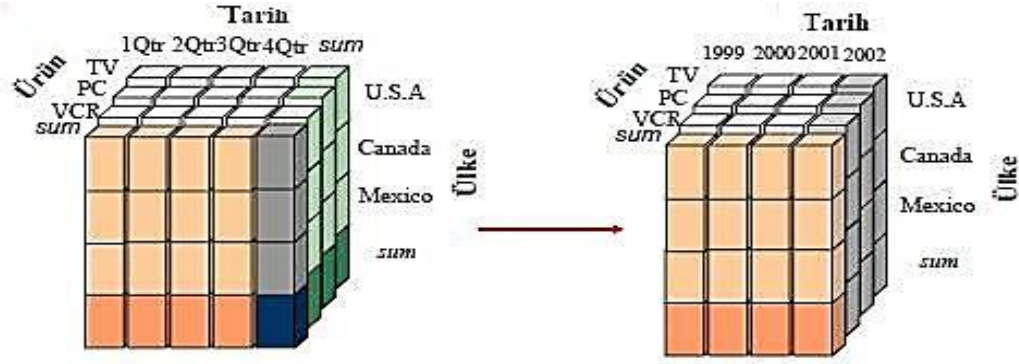
değişkenlerin tüm kümesi ele alınarak gereksiz bulunanlar küme dışına çıkarılıp en iyi değişken kümesi elde edilmeye çalışılır. Karar ağaçları ele alınacak çıktı değişkenini en iyi temsil edecek değişken kümesini verecektir (Oğuzlar 2003).

Veri sıkıştırma daha az saklama ortamı ve veriye ulaşmada daha hızlı erişim sağlamaktadır. Orijinal verileri temsil edebilecek indirgenmiş veya sıkıştırılmış veriler, veri şifreleme veya dönüşümü ile elde edilerek; indirgenmiş veri kümesi orijinal veri kümesini bir bilgi kaybı olmaksızın indirgenmiş biçimde temsil edebilecek algoritmalar mevcuttur. Bir takım sınırlamalar içermelerinden ötürü bu algoritmalar sık kullanılmamakla birlikte; temel bileşenler analizi gibi yöntemler bir bilgi kaybına göz yumarak sıkıştırılmış veri kümesi elde edilmesinde kullanışlıdır (Oğuzlar 2003).



Şekil 2.14 Veri sıkıştırma (Yıldırım 2008).

Örnekleme aşamasında ise; büyük veri topluluğunu temsil edecek daha küçük alt veri kümelerinin oluşturulması sağlanır. Alt küme; yerine koymadan örnekleme (SRSWOR), yerine koyarak örnekleme (SRSWR), demet örnekleme (yerine koymadan veya koyarak) ve katman örnekleme (katman; nitelik değerine göre grup) şeklinde seçilir. Genelleme ise verilerin tek tek değil genel kavramlarla ifade edilmesini sağlar (Özkan 2013, İnt.Kyn.3).



Şekil 2.15 Nitelik birleştirme (İnt.Kyn.3).

Veri indirgeme aşamasında; verilerin çok boyutlu veri kümelerine dönüştürülmesi gerekebilir. Böylece çözümleme sadece belirlenen boyutlara göre Şekil 2.15’de görüldüğü gibi yapılabilir (Özkan 2013, İnt.Kyn.3).

Azaltılmış veri gerçek verinin daha azı olmasına karşın sonuçların aynı olması gereklidir. Verinin çok fazla olduğu durumlarda, bu verilerin algoritmalar tarafından analiz edilmesi uzun zaman almakla birlikte beraberinde hataların oluşmasına da neden olabilir. Veri azaltma işlemi; boyut azaltma, verinin sayısını azaltma ve veri sıkıştırma şeklinde olmaktadır.

Boyut azaltma işleminde, boyutlar artarken, veri uzayda daha seyrek şekilde gözükür. Boyutların fazlalığı; kümeleme ve taşıma bulmada önemli iki nokta arasındaki uzaklık ve yoğunluğu tanımları daha az anlamlı hale gelmektedir. Boyut düşürmenin amaçlarına baktığımızda (İnt.Kyn.4):

- VM algoritmaları tarafından ihtiyaç duyulan bellek ve zaman miktarını azaltmak,
- Boyut fazlalığını problemlerini çözmek,
- Daha kolay görselleştirmeyi sağlamak,
- İlişkisel özellikleri elemeye veya gürültüyü azaltmaya yardımcı olmak.

Bu işlem için kullanılan teknikler (Yıldırım 2008):

- Wavelet dönüşüm (wavelet transforms)

- Temel bileşenler analizi (PCA)
- Niteliklerin seçimi veya oluşturulması

Veri azaltma

- Eğri uydurma
- Histogram, kümeleme, örnekleme
- Veri küpleri

Kesikleştirme, VM algoritmalarının bazıları yalnızca kategorik değerlerle çalıştığından, sürekli verilerin kesikli değerler dönüştürülmesi gerekmektedir. Sürekli verilerin kesikli değer aralıklarına dönüştürülmesi ile elde edilen kategorik değerler orijinal veri değerlerinin yerine kullanılırlar. Kavram hiyerarşileri; verilen sürekli değişken için değişkenlerin ayrıştırılması olarak da tanımlanır. Kavram hiyerarşileri düşük düzeyli kavramların yüksek düzeyli kavramlarla değiştirilmesiyle, verilerin indirgenmesinde kullanılırlar. Örneğin yaş değişkeni 1-15, 16-40, 40+ olacak biçimde daha yüksek kavram düzeyinde ifade edilerek veri indirgemede detay bilgiler kaybolursa dahi genelleştirilmiş veriler daha anlamlı olup yorumlaması daha kolay hale gelmiş olacak ve orijinal verilerden daha düşük hacim kapsamı sağlanmış olacaktır (Oğuzlar 2003).

2.3.4.3 Modelin Kurulması ve Değerlendirilmesi

Yukarıda bahsedilen süreçlerden sonra, VM algoritmalarının uygulanması için hazır hale gelen verilere; probleme yönelik VM modellerinin uygulanması adımıdır.

Model kuruluş çalışmalarının sonucuna bağlı olarak, aynı teknikle farklı parametrelerin kullanıldığı veya araç ve algoritmaların denendiği farklı modeller kurulabilir. Model kuruluş aşamasında hangi tekniğin en uygun olduğuna karar vermek oldukça güçtür. Bu nedenle en uygun modelin bulunabilmesi için mümkün olduğunca çok sayıda modelin kurularak denenmesi gerekmektedir. Veri hazırlama ve model kurma aşamaları en iyi modele ulaşıncaya kadar tekrarlanmalıdır. Fakat modelde kullanılan veri tabanının çok büyük olması durumunda; veri tabanları üzerinde çok sayıda modelin denenmesinin zaman açısından sorun teşkil etmesi nedeniyle; örnekleme yapılarak birkaç model denemek yerine, rastgele örneklenmiş veri tabanı parçası üzerinde birçok modelin denenmesi ve bunlar arasında en güvenilir, en güçlü modelin seçilmesi daha uygun

olabilir. Makine öğrenmesi, yapay zekâ algoritmaları, bulanık mantık gibi birçok yöntem veri madenciliğinde model oluşturmada kullanılabilir (Akpınar 2000, Tuğ 2005, Şen 2008, Arslan 2008).

Model beklenen hedefleri karşılamaya yeterli olduğu zaman süreç bazında daha geniş bir açıdan değerlendirme yapılır. Bu değerlendirmede modelin doğru kurulup kurulmadığı, gelecekte ne tip verilerin kullanılabileceği, modelin geliştirilmesi gibi konular modelin yorumlanmasında değinilmesi gereken durumlardır (Şaylan 2013).

Model kuruluş süreci denetimli (supervised) ve denetimsiz öğrenme (unsupervised learning) modellerinin kullanımına göre farklılık göstermektedir (Akpınar 2000, Dinçer 2006, Arslan 2008).

Bir denetçi tarafından belirlenen bir kritere göre sınıflar ayrılarak, her sınıf için çeşitli örnekler verilerek, her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesine örnekten öğrenme olarak isimlendirilen denetimli öğrenim denir. Süreç sonunda tanımlanan kural cümleleri ile henüz hangi sınıfa ait olduğu bilinmeyen verilerin var olan sınıflardan hangi sınıfa ait olduğu kurulan model tarafından belirlenir. Denetimsiz öğrenme ise başlangıçta verilen ve henüz sınıflandırılmamış bir küme veriyi, gelen verinin özelliklerine ve verinin özellikleri arasındaki benzerliklere göre anlamlı alt kümeler oluşacak şekilde sınıflarının tanımlanmasıdır (Akpınar 2000, Dinçer 2006, Arslan 2008).

Sonuçları değerlendirmeden önce modelleme aşamasındaki modeller kaliteye ve verimliliğe göre değerlendirilir. Modelin ilk aşamasında belirlenen amaca uygunluğuna bakılıp araştırma problemine cevap alınamayan noktalar belirlenip nihayetinde bu VM modelinin kullanılıp kullanılmayacağına karar verilir (Larose 2005).

Bir modelin doğruluğunun test edilmesinde kullanılan performans ölçütleri bölüm 2.3.10’de ayrıntılı olarak ele alınmıştır.

2.3.4.4 Modelin Kullanılması ve İzlenmesi

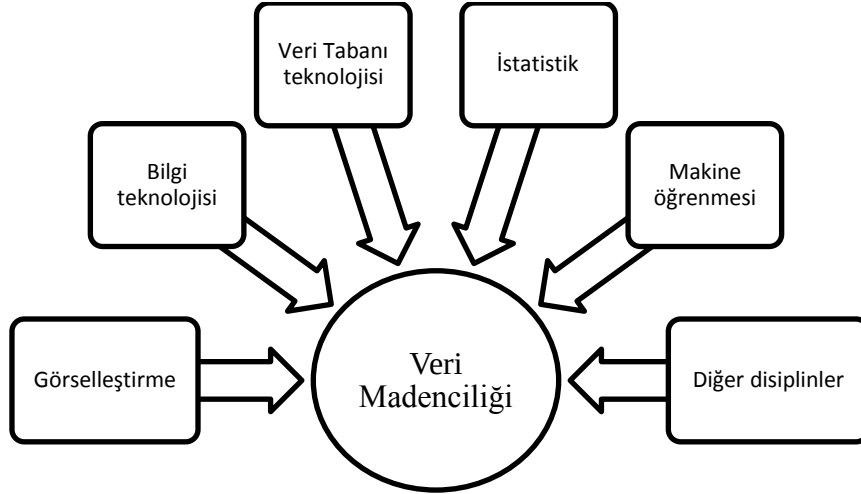
VM ile elde edilen sonuçlar; çeşitli düzenlemelerin ardından kullanıcılar için daha anlaşılır bir hale getirilerek grafikler yardımıyla sunum yapılmalıdır (Akpınar 2000,

Özkan 2013). Değerlendirme aşamasında, keşfedilen bilginin yenilik, yararlılık, geçerlilik ve basitlik ölçütlerine göre değerlendirilmesi gerekmektedir (Sever ve Oğuz 2002). Model doğrudan bir uygulama olabileceği gibi, bir uygulamanın içine de gömülebilir. Zamanla değişen veri ve özellikleri kurulan modelin izlenmesini ve güncellenmesini gerektirebilir (Akpınar 2000).

Evrendeki her nesnenin ayrıntılı bir biçimde tanımlandığı ve nesnelerin alabileceği değerler kümesinin belirli olduğunu varsayımını yaparsak, her nesnenin tanımı kesin ve yeterli olsaydı sınıflandırma işlemi nesnelerin alt kümelerinden faydalanılarak basitçe yapılabilirdi. Kurum ihtiyaçları doğrultusunda toplanan veriler gerçek hayatı yeterince yansıtmayabilir. Örneğin bir hastalığın tanısı için belirlenen kurallar sadece çok yaşlı insanların belirtilerinin bulunduğu bir veri kümesi kullanılarak üretilmiş olsaydı, bu kurallara göre bir çocuğa tanı koymak doğru olmazdı. Bu gibi durumlarda bilgi keşif modeli belirli bir güvenlik derecesinde tahmini kararlar alabilmelidir (Dursun 2005, Gürbüz 2009).

2.3.5 Veri Madenciliği ve Diğer Disiplinler

VM; makine öğrenmesi, örüntü tanıma, veri tabanı teknolojisi, istatistik, bilgi bilimi, yapay zekâ, uzman sistemler, veri görselleştirme alanlarının bir keşişim noktası olarak ortaya çıkmıştır ve multidisipliner olarak gelişimini sürdürmektedir. Bu disiplinlerin hem kendileri arasında, hem de VM arasında kesin sınırlar tanımlamak oldukça güçtür. Örüntü keşif aşamasında; makine öğrenmesi, istatistik ve örüntü tanımlamalarından, bulunan örüntüleri yorumlama aşamasında yapay zekâ teknolojilerinden, eldeki verileri depolama, temizleme, süzme ve sorgulama işlemi aşamasında veri tabanı teknolojilerinden, raporlama ve insan beyni için anlamlı sembollere çevirme aşamasında veri görselleştirmeden faydalanılmaktadır (Hand *et al.* 2001, Dinçer 2006).



Şekil 2.16 VM ve diğer disiplinler arasındaki ilişkiler (Han and Kamber 2006).

Veri madenciliğinin diğer disiplinlerle ilişkisi Şekil 2.16’da görülmektedir. VM uzman sistem yaklaşımı olmamasına karşın, uzman sistemler alanına benzerlik göstermekte ve uzman sistemleri besleyebilecek bir yaklaşımdır. Uzman sistemlerin en önemli sorunu olan kural tabanını oluşturacak olan kuralların elde edilmesi, veri yoğun ortamlarda VM yaklaşımı ile desteklenebilir. Örneğin kanser teşhisini amaçlayan bir uzman sistem geliştirilmesi amaçlanıyorsa, uzman hekimden elde edilecek olan kurallar, kanser veri tabanından elde edilebilecek verilerle çok daha zenginleştirilip, güvenilir uzman sistemlerin oluşturulması mümkün olabilir (Baykasoğlu 2005).

2.3.5.1 Veri Tabanı ile İlişkisi

Tek bir tablo veya düz bir dosya üzerinde çalışan geleneksel VM algoritmaları, gerçek yaşamdaki birden çok tablodan oluşan ve aralarında karmaşık ilişkiler bulunan ilişkisel veri tabanlarında işlem yaparken, bazı verilerin tekrar etmesi yahut kaybolması gibi durumlar ortaya çıkabilmektedir. Bu nedenle çoklu ilişkilere sahip veri tabanları üzerinde VM çalışmaları yapabilmek için tahmin ve tanımlayıcı VM algoritmalarını içeren ilişkisel VM algoritmaları geliştirilmiştir. Bu algoritmalar tahmin ve tanımlarını “Çoklu İlişkisel Karar Ağaçları” ve “Çoklu İlişkisel Kurallar” ile tarif ederler (Yıldırım vd. 2008).

2.3.5.2 İstatistik ile İlişkisi

Büyük veri kümeleri içerisinde gizli kalmış ve ilginç örüntüleri bulmak VM ile mümkündür. VM ile istatistik arasındaki farklılık da tam olarak burada ortaya çıkmaktadır. Veri madenciliğinde bilinmeyen faktörler arasındaki gizli kalmış ve tahmin edilemeyen ilişkiler arasındaki birliktelik, güçlülük söz konusu iken; istatistikte ise bu durum yalnızca bilinen faktörler arasındaki güçlülüğün araştırılmasıdır (Tuğ 2005).

İstatistik daha çok teori odaklıdır ve hipotezin test edilmesine odaklanır. Otomatik öğrenme daha çok heuristic (deneyimsel) ve öğrenme performansını iyileştirmeye odaklanır. VM ve bilgi keşfi teoriyi (istatistik) ve deneyimi (otomatik öğrenme) birleştirir ve bilgi keşfinin bütün süreçlerini içine alır. VM bazen “ikincil veri analizi” olarak adlandırılır. Veri madenciliğinde kullanılan birçok algoritma istatistikçiler ve veri madencileri tarafından ortaklaşa kullanılırlar. Örüntü tespitini hedefleyen veri madenciliğine karşın, istatistik, bulunan bu örüntülerin gerçekliğini belirlemeye çalışır. VM verinin tümünün kullanılması yönüyle de istatistiki verilerden ayrılır (İnt.Kyn.13).

Klasik istatistiksel uygulamalar ve VM arasındaki en temel farklılık, veri kümesinin büyüklüğü olup istatistikçi için büyük veri kümesi birkaç yüz veya bin veriyi içerirken, veri madencisi için bu sayı milyon veya milyar ile ifade edilebilir (Oğuzlar 2003).

2.3.5.3 Makine Öğrenmesi ile İlişkisi

Kuramsal tabanda bilgi destek sistemleri, bilişim teknolojilerinde yazılım ve donanımdaki gelişmelere paralel olarak büyük miktardaki veriler işleme alınıp çıktı olarak elde edilen bilgi karar destek sistemlerinde kullanılabilme imkânı elde etmiştir. İnsan ve makine uzman, veri madenciliğinin birbirini tamamlayan en önemli bileşenleridir. İnsan uzman, problemin tanımlanması, veri tabanının hazırlanması ve hedeflerin belirlenmesinde aktifken; makine uzman, verileri hedeflerle uyuşan örüntüleri tespit etmek amacıyla taramak ve bu örüntülere uygun kuralları öğrenmekten sorumludur. Veri madenciliğinde elektronik ortamda saklanan verilerdeki örüntüler otomatik olarak aranıp, doğrulanmakta böylece tahmin veya sınıflandırma amacıyla kullanılmak üzere kurallar öğrenilmektedir (Çakır 2005).

VM ve makine öğrenmesi yöntemlerinin temel özellikleri aşağıdaki gibi özetlenebilir (Çakır 2005).

- Her iki yöntemde kuramsal olmaktan çok ampirik öğrenme/sınıflandırma problemlerini konu edinir.
- Veride yapısal örüntüleri bulmak ve tanımlamak, veriyi açıklayacak aracı (algoritma) ve ilintili kural kümesini (kümelerini) bulmak ve bilgiyi tahmin veya sınıflandırma amacıyla kullanmak makine öğrenmesinin asıl amacıdır.
- Veri seti örnekleri üzerinde yapılan tahmin/sınıflandırma ya da öğrenilen kurallar her iki yöntemin de çıktısıdır. Elde edilen bilginin tanım alanı bağlamında kuramsallaştırılması önemli bir hedeftir.
- Makine öğrenmesi ve VM yöntemlerinde veriden bilgi çıkarma yarım ya da tam otomatik süreçlerdir.
- Her iki yöntemde karar destek sistemleri açısından, makine uzmanının bilgi işleme kapasitesine dayanılarak, insan uzmanının bu alandaki eksiklikleri giderilmek ve rekabet avantajı sağlamak amaçlanmaktadır.

2.3.6 Veri Madenciliği Yöntemleri

Veri madenciliğinin geniş bir uygulama alanı bulunmaktadır. Farklı disiplinlerle olan ilişkisi de göz önünde bulundurulduğunda; bu geniş yelpazede yapılan çalışmaların araştırma sahalarında farklı teknik ve metotlar kullanılarak yapılması gerekmektedir.

Farklı alanlarda farklı yöntemlerin başarılı olması çeşitli nedenlerden kaynaklanmaktadır. Başarı, kullanılan veri kümesinin büyüklüğüne, öznelilikleri ile sayılarına, görüldüğü alanın özel durumlarına ve kullanılan veri tabanlarına bağlı olarak değişebilmekte ve her alanın kendine özel niteliklerinin olması ve ortak veri standardının olmayışı nedeni ile kontrollü deneyler yapma olanağı bulunmaktadır (Şişaneci 2009).

2.3.6.1 İstatistiksel Yöntemler

VM temel olarak ileri bir istatistik uygulamasıdır. Birçok VM projesinde istatistik tekniklerinin çok fazla kullanıldığı görülmektedir. Veri madenciliğini istatistiğin otomatik veya yarı otomatik olarak bilgisayar tarafından yapılması şeklindeki algının

aksine istatistik yazılım ve yöntemlerini VM olarak adlandırmak doğru olmayacaktır (Koyuncugil 2009).

Veri madenciliğinde kullanılan istatistiksel teknikler; incelenen olay ve çevresindeki iç ve dış faktörler dikkate alınarak problemin yapısına göre bilgilerin incelenmesi ve çözüme yönelik geliştirilen yöntemleri bütünü şeklinde çok değişkenli istatistiksel analiz başlığı altında toplanır. Diskriminat analizi, lojistik regresyon, faktör analizi gibi teknikler çok değişkenli analiz kapsamındadır (Esen 2009).

Karşılaşılan problemlerin genellikle birden fazla faktör etkisinde olması tek değişkenli istatistik tekniklerin sınırlı çözüm aralığı sunmalarından dolayı tercih edilmezler. Gözlemlerin en az hata payı ile ait oldukları gruplara ayırma şeklinde yapılan sınıflandırma problemlerinin çözümünde çok değişkenli problemin tek değişkenli biçime dönüşümü sağlanarak diskriminant analizi kullanılmaktadır (Esen 2009).

Faktör analizi, çok sayıdaki veriler arasındaki ilişkileri inceleyerek verilerin daha anlamlı ve en az kayıpla özet biçiminde sunulmasını sağlayan ve bunların yorumlanmasını kolaylaştıran çok değişkenli istatistiksel bir tekniktir. Sonuçların çok kesin olmaması nedeniyle yorumlanmasının güç olması tekniğin bir dezavantajıdır (Esen 2009).

Lojistik regresyon analizi ise; sınıflama ve atama işlemlerinde kullanılabilen regresyon yöntemi olup normallik, ortak kovaryansa sahip olma şeklindeki çeşitli varsayım bozulmaları olması durumunda diskriminant ve çapraz tablo analizlerine alternatif olarak uygulanan çok değişkenli istatistiksel yöntemlerdendir (Esen 2009).

2.3.6.2 Bellek Tabanlı Yöntemler

İnsanlar kararlarını daha önceden yaşadıkları benzer deneyimlere göre verirler. Örneğin doktorlar bir hastayı incelerken, elde ettikleri bulguları daha önceden benzer hastalığa yakalanmış hastaların tedavi sürecindeki deneyimlerini kullanarak değerlendirmektedir. En önemli özelliği veriyi olduğu gibi kullanabilmek olan bellek tabanlı yöntemler de benzer şekilde deneyimleri kullanarak bilinen kayıtların bulunduğu veri tabanı oluşturup sisteme yeni gelen kaydın komşuları belirlenir ve bu kayıtlar yardımı ile tahminde bulunulur yâda bir sınıflama işlemi gerçekleştirilir. Diğer VM yöntemlerinin aksine

kayıtların şekli (format) yerine; iki kayıt arasındaki uzaklığı belirleyen bir uzaklık fonksiyonu ve komşu kayıtları işleyerek bir sonuç üreten kombinasyon fonksiyonunun varlığı ile ilgilenir (Tantuğ 2002).

Bellek tabanlı yöntemler klinik işlemleri ve sahtekârlık tespiti gibi uygulama alanlarında kullanılmaktadır (Şen 2008).

Bellek tabanlı yöntemlerin güçlü olduğu noktalar aşağıda belirtildiği şekildedir (Tantuğ 2002).

- Rastgele seçilen, hatta birbirleri ile ilgisiz olabilen verilere bile uygulanabilir.
- Kolayca anlaşılabilir sonuçlar üretir.
- Çözümleme alanlarının çok olduğu durumlarda dahi etkili olarak çalışabilirler.
- Eğitim kümesinin oluşturulması basittir.

Bellek tabanlı yöntemlerin zayıf olduğu noktalar ise aşağıda sıralanmıştır.

- Eğitim kümesi için büyük miktarlarda alana ihtiyaç vardır.
- Sınıflama ve tahmin işlemleri için kullanıldığında işlem maliyeti yüksektir.
- Üretilen sonuçlar, seçilen uzaklık fonksiyonuna, kombinasyon fonksiyonuna ve komşu sayısına doğrudan bağlıdır.

2.3.6.3 Genetik Algoritmalar

Genetik algoritmalar evrim sürecini modelleyen problem uzayında çok yönlü ve adaptif bir arama ile en iyi çözümü bulmayı hedefleyen bir makine öğrenmesi tekniğidir. Bu teoriye göre yaşayan canlılar arasında hayatta kalmak için var olan rekabette kabaca güçlü olanlar (başarılı), genetik bilgilerini sonraki nesillere aktarılırken zayıf olanlar elenir. Sınıflama, kümeleme ve örüntü tanıma işlemlerinde veri madenciliğinde kullanılan genetik algoritmalar güçlü bir hesaplayıcı model olup her zaman kesin çözümü garanti edememesine karşın en iyi çözümü elde etmede oldukça başarılıdır (Tuğ 2005, Silahtaroglu 2013, Kaya 2014).

Genetik algoritmalar açıklanabilir sonuçlar üretirler ve değişik tiplerdeki verileri işleme özelliğine sahip olan genetik algoritmalar en iyileme amacı ile kullanılabilirler. Yapay

Sinir Ağları (YSA) ile Genetik algoritmalar ortaklaşa çalışarak başarılı sonuçlar üretmektedirler. Genetik algoritmalar YSA'nın eğitilmesi, bellek tabanlı yöntemlerde kombinasyon fonksiyonunun oluşturulması gibi işlerde de kullanılmışlardır. Genetik algoritmaların kullanımlarında en büyük sıkıntı; karmaşık sorunların genetik kodlamasının çok zor olması ve en iyi sonucun üretildiğine dair bir garanti bulunmamasıdır (Tantuğ 2002).

Genetik algoritmanın en önemli adımı uygun genetik kodlamanın yapılması olup, kodlama işlemi problemin yapısına ve programlayıcının bakış açısına göre değişmektedir. Kodlamada genel kural ikili kodlama olmasına karşın böyle bir zorunluluk olmayıp, onluk sistemde de tasarlanabilmektedir. Popülasyon yapısına sahip olan algoritma üyeleri bireylerden oluşmakta ve her birey aynı zamanda bir kromozomu ifade etmektedir. Her bir kromozom bir çözüm uzayını belirlemekte ve elde edilen en iyi birey çözüm uzayı olarak adlandırılmaktadır (Kaya 2014).

Kromozom veri tabanındaki her bir kayıttan başka bir şey olmayıp üretilecek yeni sonuçlar hakkında bir takım bilgiler içermekte ve bu bilgilerin kullanılabilmesi için kromozomların çözümlenmesi işlemi olan kromozomların kullanılabilir formlar haline dönüştürülmesi gerekir. Kromozomların çözümlenmesi işlemi başlıca ikili, permütasyon, değer çözümü ve ağaç çözümü olmak üzere dört şekilde yapılabilir.

İkili çözümlemede her bir kromozom ikili sayı sisteminde gösterilip elde edilen her bir bit çözümün bir özelliğini oluşturmakta iken permütasyon yönteminde her bir kromozom sayılarla ifade edilip yükleme ve gezgin satıcı problemlerinin çözümünde kullanılmaktadır. Her bir kromozomun değerler dizisi içerdiği değer çözümü yöntemi yüksek değerdeki sayıların olduğu problemlerde kullanılırken, her bir kromozomun nesnelerden oluşan ağaç olduğu ağaç çözümleme yöntemi ise program ve ifade geliştirmede kullanılmaktadır (Silahtaroglu 2013).

Mümkün olduğunca geniş arama uzayını tarama ve global en iyiyi bulmayı amaçlayan çaprazlama ve mutasyon genetik algoritmanın asıl işlevidir. Çaprazlama ve mutasyon çeşitliliği artırmayı sağlamakta ve yararlı bilgiyi elde etmede önem taşımaktadır. Çaprazlama olasılığının yüksek tutulması (0,7 gibi) ve mutasyon oranının düşük tutulması (0,001 gibi) genel bir kabul olmakla birlikte probleme bağlı olarak

değişebilmektedir. Algoritmanın sonlandırılma kriteri; belirli iterasyon sayısına ulaşma veya istenen başarı oranının iterasyon sayısı tamamlanmadan elde edilmesidir. Kromozom sayısı, mutasyon oranı, çaprazlama yöntemi, parametrelerin kodlanması, uygunluk fonksiyonunun belirlenmesi genetik algoritmaların performansını etkileyen faktörler arasındadır (Kaya 2014).

Genetik algoritmalar parametre ve sistem tanımlama, robot uygulamaları, kontrol sistemleri, ses ve görüntü tanıma, yapay zekâ uygulamaları, mühendislik tasarımları, ağ tasarım problemleri, yol bulma problemleri, sosyal ve ekonomik planlama problemleri için diğer en iyileme yöntemlerine kıyasla daha başarılı sonuçlar vermektedir (Siyah 2012).

2.3.6.4 Yapay Sinir Ağları

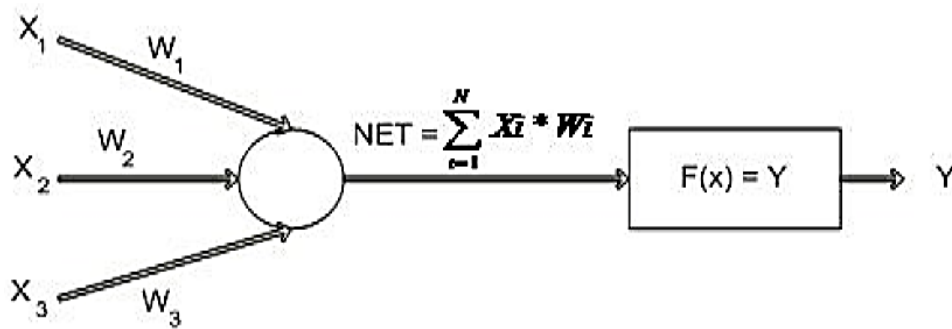
İnsan beyni yaklaşık 10^{11} sayıda nörondan (sinir hücresi) oluşmakta olup çok karmaşık bir yapıya sahip ve bu sinir hücreleri çok sayıda birleşme noktasına (synapse) sahiptir. Bir sinir hücresi diğer bir sinir hücresiyle bu birleşme noktaları sayesinde bağlantı kurar. Yapay sinir ağları (YSA), basit biyolojik sinir sisteminin çalışma şekli simüle edilerek tasarlanan programa yaklaşımdır. Simüle edilen sinir hücreleri nöronlar ve bu nöronların birbirlerine çeşitli şekillerde bağlanarak ağı oluşturulur. Bu ağlar hafıza alma, öğrenme ve veriler arasındaki ilişkiyi ortaya çıkarma kapasitesine sahiptir. YSA normalde bir insanın düşünme ve gözlemlemeye yönelik doğal yeteneklerini gerektiren problemlere yönelik çözümler üretebilmesinin temel sebebi ise insan beyninin sahip olduğu yaşayarak veya deneyerek öğrenme yeteneğidir. İnsan beyninin karmaşık yapısı daha karmaşık olan öğrenme makinelerinin oluşturulmasında model olarak alınmış ve YSA olarak adlandırılmıştır. YSA için farklı yapılar bulunmakta ve bunların her biri verilen işleri yapmak için farklı yol ve öğrenme yöntemleri bulunmaktadır. YSA'nın belirli bir işi yapmak için eğitildiği evreye “şifreleme evresi” sınıflama ya da kestirim yapma evresine ise “şifre çözme evresi” adı verilir (Ergün 2008).

YSA normalde bir insanın yaşadığı, karşılaştığı deneyimlediği ve gözlemlediği benzer ya da farklı olaylardan öğrenmesi, yeni bilgi edinmesi, bu bilgi ve olaylar için genelleme yapması, aralarında ilişki kurması, hatalardan öğrenmesi ve tüm bunlardan yola çıkarak farklı bir durum için karar vermesini sağlamaktadır. YSA'nın insanın

düşünme, gözlemleme, deneme-yanılma hatalardan ders çıkarma şeklinde problemlere çözümler üretebilmesini model almıştır (Pala 2013).

YSA insanın önceki deneyimlerinden bir takım bilgiler çıkartması gibi, kendisine verilen örneklerden bir takım bilgiler çıkartma yeteneğine sahiptir. YSA karmaşık sorunları çözebilir, ayrıca temel uygulamalardan öğrenilebilir sorunu bir sonraki sefer için daha iyi bir çözüm bulacak şekilde değiştirilir. YSA öncelikle sonuçları bilinen belirli bir veri kümesi üzerinde öğrenme algoritmaları çalıştırılarak eğitilir. Bu eğitim neticesinde YSA içerisindeki ağırlıklar belirlenir. Bu ağırlıklar kullanılarak yeni gelen veriler işlenir ve sonuçlar üretilir (Tantuğ 2002).

YSA biyolojik sinir ağlarının yapay bir modeli olduğundan tıpkı biyolojik sinir hücrelerinde olduğu gibi beş bölümden oluşmaktadır. Bu bölümler; girdiler, ağırlıklar, birleştirme fonksiyonu, aktivasyon fonksiyonu ve çıktılarıdır. Basit bir YSA hücresi Şekil 2.17’de görülmektedir (Esen 2009).



Şekil 2.17 Yapay sinir hücresi (Esen 2009).

Girdiler ($x_1, x_2, x_3, \dots$); diğer hücrelerden yada dış ortamlardan hücreye gelen bilgidir.

Ağırlıklar ($w_1, w_2, w_3, \dots$); ilgili girdinin hücre üzerindeki etkisini gösteren katsayılarıdır. Ağırlıkların değerleri negatif, sıfır veya pozitif olabilir.

Birleştirme Fonksiyonu ($f(x)=y$); hücrenin gerçekleştireceği işleve göre seçilen aktivasyon fonksiyonu, girdiye karşılık üretilecek çıktıyı belirlemektedir. YSA hücrelerinin doğrusal olmamasından ötürü, aktivasyon fonksiyonu genellikle doğrusal olmayan bir fonksiyon olarak seçilmektedir.

Çıkış (y); dış dünyaya gönderilebilen veya tekrar ağıın içerisinde kullanılabilen çıktı değeri aktivasyon fonksiyonu tarafından belirlenmektedir.

YSA'larının en olumsuz tarafı ağırlıkların neden ilgili değeri aldığının bilinmemesi ya da çıkan sonucun neden geçerli bir sonuç olduğunu açıklayamamasıdır. YSA; sınıflama, kümeleme ve tahmin analizleri ile kolaylıkla kullanılabilecek güçlü ve genel amaçlı bir teknik olup ekonomiden tıbbi uygulamalara, ses tanımlama ve çözümlemeye robot denetimine, hata algılamadan sorun tanılamaya, değerli müşterilerin belirlenmesi için yapılan kümeleme işlemlerinden kredi kartı sahtekârlıklarının belirlenmesine kadar geniş bir alanda uygulanabilmektedir (Ergün 2008). YSA'nın temel işlevleri arasında tanımlama, sınıflandırma, kümeleme, tahmin etme ve veri birleştirme gibi veri madenciliğinde kullanılan metotlar bulunmaktadır. Plaka ve yüz tanıma sistemleri gibi teknolojiler YSA kullanılarak geliştirilmiştir.

YSA'nın VM açısından kuvvetli yönleri bulunmaktadır (Tantuğ 2002).

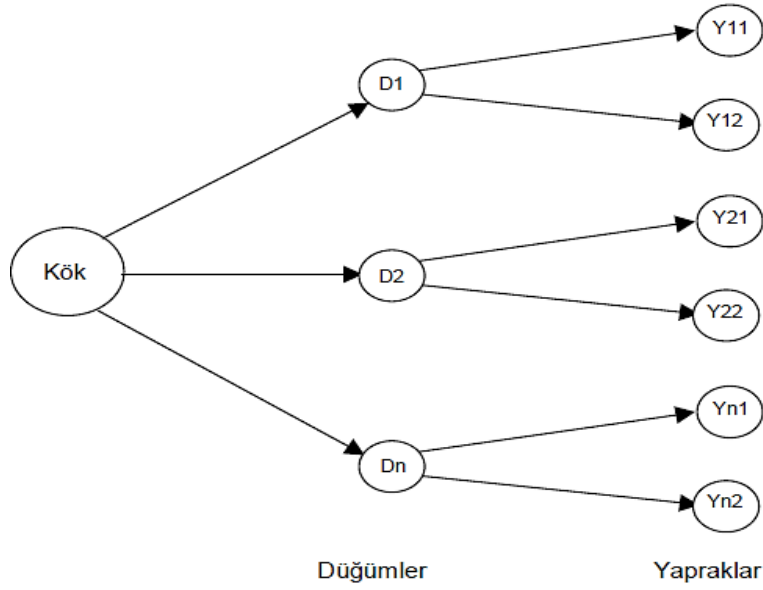
- Çok karmaşık durumlarda dahi iyi sonuçlar üretmektedir.
- Çok geniş spektrumdaki (açıdaki) sorunların çözümünde kullanılabilir
- Hem kategorik hem de sayısal veriler üzerinde işlem yapabilirler.

YSA'nın VM açısından da zayıf yönleri de aşağıda belirtildiği gibidir.

- Ürettikleri sonuçların açıklamasını yapamazlar.
- 0 ile 1 arasında giriş verileri olması zorunludur.
- Varılan sonucun olası en iyi sonuç olduğunun garantisi yoktur.

2.3.6.5 Karar Ağaçları

Karar ağaçları (Decision Tree) verileri analiz etmek için kullanılan sınıflandırma modelleri arasında en sık kullanılan VM yöntemlerindendir.



Şekil 2.18 Karar ağaçlarının genel yapısı (Çakır 2005).

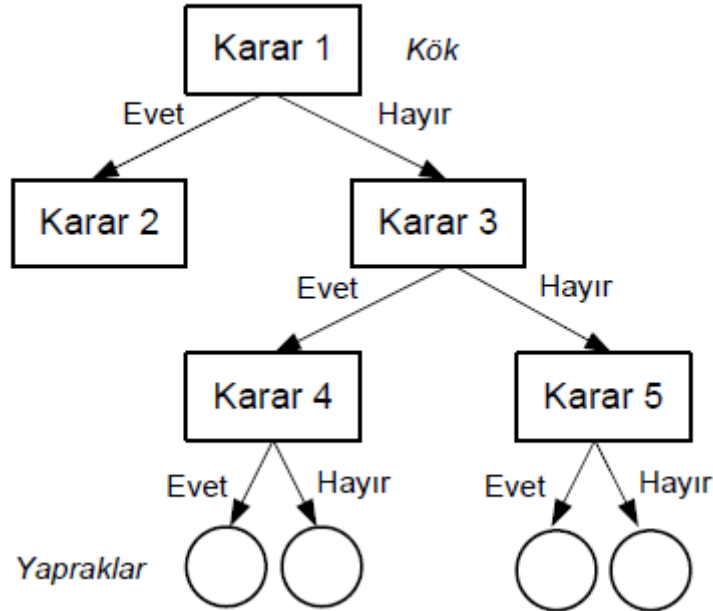
Ağaç yapısındaki bir akış şeması olan karar ağacı (desicion tree); yapılarında her düğüm bir özellik için yapılan testi, her dal bu testin sonucunu ve her yaprak düğüm ise sınıfları ifade etmektedir. Ağacın en tepesindeki düğüm ise kök düğüm olup bu sınıflamada en fazla ağırlığı olan özelliği temsil etmektedir. Karar ağaçlarının oluşturduğu yapılardan öğrenilen kurallar her yaprak üzerinde, kökten o yaprağa kadar olan tüm sınamalar birleştirilerek elde edilir (Çakır 2005, Arslan 2008). Bir karar ağacının genel yapısı Şekil 2.18’de görülmektedir. Karar ağaçları üç bölümden oluşan bir modeldir (Arslan 2008):

1. Tanımda belirtildiği gibi bir karar ağacı
2. Ağacı oluşturacak bir algoritma
3. Ağacı veriye uygulayacak ve mevcut probleme çözüm olacak bir algoritma

Karar ağacına ait yapıyı açıklamak gerekirse sözde kod aşağıdaki şekildedir (Kaya 2014).

1. Adım: Tüm öznitelik durumlarını kontrol et.
2. Adım: Kullanılacak karar ağacı ölçütüne göre her bir öznitelik için bilgi değerine bak.
3. Adım: Bilgi değerine göre öznitelikleri sırala ve en yüksek değerli özniteliği kök değeri olarak belirle.

4. Adım: Belirlenen başarı değeri ve iterasyon sayısına kadar ağaç yapısını 3. Adıma giderek genişlet.
5. Adım: Kriter sağlandıysa algoritmayı sonlandır.

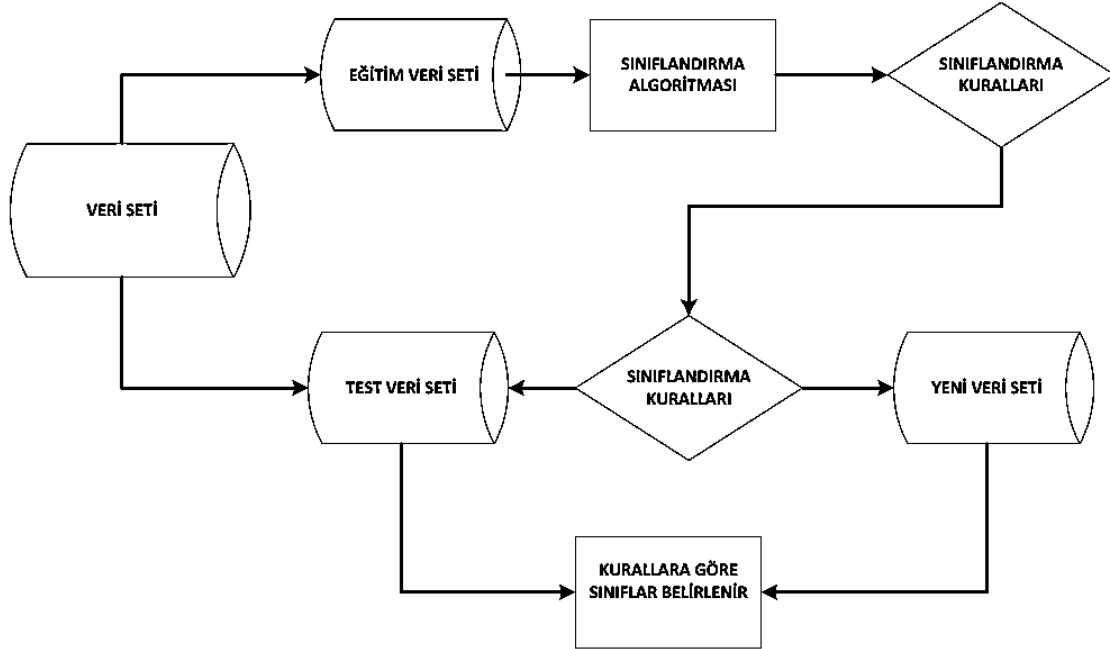


Şekil 2.19 Karar ağacı örneği (Kocabaş 2010).

Karar ağaçlarında her dal kendinden sonra evet ya da hayır şeklinde en az iki dala ayrılarak, her bir dalın belirli bir olasılığı mevcut olup bu sayede son dallardan köke veya istenilen yere ulaşılan kadar olasılıkları hesaplamak mümkündür. Bütün ağacın hata oranı, ayrı ayrı dallar için ağırlıklı hata oranlarının toplamı olarak hesaplanabilir. Hata oranları ve hesaplamaların verimliliği ağacın belirli dalları kesilerek yani daha az yararlı kurallar elenerek geliştirilebilir (Ergün 2008).

Karar ağacı tekniği kullanılarak verinin sınıflandırılması öğrenme ve sınıflama olmak üzere iki basamaklı bir aşamalıdır. Öğrenme basamağında, önceden bilinen bir eğitim verisi sınıflama algoritmaları tarafından model oluşturmak amacı ile analiz edilir ve oluşturulan öğrenme modeli sınıflama kurallarını veya karar ağacını ifade eder. Sınıflama basamağında ise sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek için test verisi kullanılır. Test verisine uygulanan bir modelin doğruluğu, yaptığı doğru sınıflamanın test verisindeki tüm sınıflara oranıdır. Model tarafından

tahmin edilen sınıf ile test örneğinde bilinen sınıf karşılaştırılır. Sonuçta modelin doğruluğu kabul edilebilir düzeyde ise kurallar sonradan gelen verilerin sınıflandırılmasında kullanılır. Bu işlemler Şekil 2.20’de görülmektedir (Pala 2013).



Şekil 2.20 Veri sınıflandırma işlemi (Pala 2013).

Veri seti eğitim ve test veri seti olmak üzere Şekil 2.20’de görüldüğü gibi ikiye ayrılmaktadır. Eğitim verisi üzerinde uygulanan sınıflandırma algoritmalarından sınıflandırma kuralları çıkarılıp bu kurallar test verisi ya da yeni bir veri seti üzerinde uygulanır ve bu veri setleri sınıflandırma kurallarına göre sınıflara ayrılır (Pala 2013).

Tahmin edici ve tanımlayıcı özelliklere sahip karar ağaçları kuruluşlarının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin daha iyi olması gibi nedenlerden ötürü veri madenciliğinde sınıflama modelleri içerisinde sıklıkla kullanılmaktadırlar (Akpınar 2000). Karar ağacı temelli yaygın kullanılan analizlere; belirli bir sınıfın muhtemel üyesi olacak elemanların belirlenmesi, çeşitli vakaların düşük orta yüksek gibi kategorilere ayrılması, gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması, çok miktardaki değişken ve veri kümesinden faydalı olacakların parametrik modelleme kurulması için seçilmesi, sadece belirli alt gruplara özgü olan

ilişkilerin tanımlanması, sürekli değişkenlerin kesikliye dönüştürülmesi ve kategorilerin birleştirilmesi örnek verilebilir (Arslan 2008).

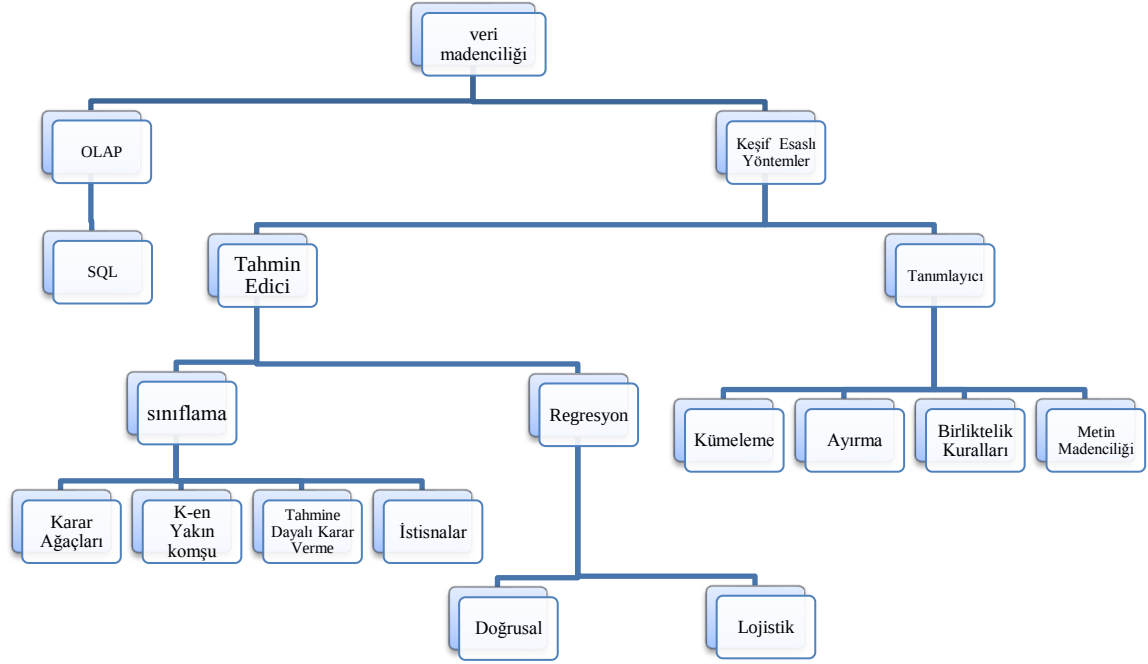
Her hangi bir kuralın ne düzeyde yararlı olduğunu belirlemek için, rastgele seçilen başka bir kuralın veri kuralı doğruluk performansına eşit ya da daha iyi performans sergileyecek bir kural olup olmadığının sorgulanması gerekir. Bir kurala yeni bir terim eklenmesi kuralı “mükemmel” bir duruma getiriyorsa, kuralın eğitim setine özelleşme tehlikesi oluşabilir. Öğrenmenin amacı, algoritmanın yeni örnekleri de başarılı bir şekilde sınıflandırarak genel kurallar öğrenmek olduğu için, mükemmel kurallar öğrenmek çoğu zaman spesifik bir durum için geçerliken bu durumdaki “mükemmel” kurallar şüpheyle karşılanmalıdır. İyi kurallar geliştirilirken mükemmel kuralların öğrenilmesi muhtemel aşırı özelleşmeden dolayı eklenen terimin budanması gerekir. Kural hala iyi vasfını koruyorsa terimin silinmesiyle kuralın performansı düşene kadar silme işlemine devam edilebilir (Çakır 2005).

Karar ağaçlarının kullanım alanlarına bakıldığında; bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi (credit scoring), mektupla yapılan pazarlama uygulamalarına hangi demografik grubun yüksek cevaplama oranına sahip olduğu (Direct Mail), geçmişte işletmeye en faydalı olarak çalışanların özelliklerini kullanarak işe alma süreçlerinin belirlenmesi, üretim verilerinin incelenerek ürün hatalarına yol açan değişkenlerin belirlenmesi, hangi değişkenlerin satışları etkilediğinin belirlenmesi ve tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi gibi birçok örnek verilebilir. 1970’li yılların başlarında karar ağacı temelli ilk algoritma ve yazılım olan AID (Automatic Interaction Detector) sonrasında karar ağacı modellerinin gelişimi CHAID (Chi-Squared Automatic Interaction Detector), C&RT (Classification and Regression Splines), QUEST (Quick, Unibased, Efficient Statistical Tree), C5.0, SLIQ (Supervised Learning in Quest), ve SPRINT (Scalable Parallelizable Induction of Decision Trees) ile devam etmiştir (Akpınar 2000).

En popüler karar ağaçları algoritmaları CHAID, C&RT, ID3, C4.5, C5 algoritmaları olup, WEKA’nın sınıflandırma algoritmalarından olan J48 algoritması da sayısal değerler içeren veri tabanları üzerinde karar ağaçları oluşturulmasına olanak sağlar.

2.3.7 Veri Madenciliği Teknikleri

VM teknikleri eldeki veri türüne ve elde edilen sonuçların kullanım amacına göre temelde tanımlayıcı (descriptive) ve öngörüsöl (predictive-tanımlayıcı) olarak iki kategoride incelenir (Han and Kamber 2001).



Şekil 2.21 Veri madenciliği modelleri (Ergün 2008).

Tanımlayıcı VM; verinin genel karakterini, mevcut durumunu ortaya çıkarmaya yönelikken, öngörüsöl VM; veriler üzerinde geleceğe yönelik tahminler yaparak sonuç çıkarma amaçlı kullanılmaktadır. Bu teknikler kullanılan veri yapılarına ve keşfedebildikleri örüntü biçimlerine göre kategorilere ayrılmaktadır. VM modellerinin şematik gösterimi Şekil 2.21’de görülmektedir.

Literatürde VM teknikleri için farklı gruplandırmalar mevcuttur. En yaygın olan Han ve Kamber (2001)’in yaptığı kategorilerdir:

- Tanımlama ve Ayrımlama (Characterization and Discrimination)
- Birliktelik Analizi ve Ardışık Zamanlı Örüntüler (Association Analysis and

Sequential Patterns)

- Sınıflandırma ve Öngörü (Classification and Prediction)
- Kümeleme Analizi (Cluster Analysis)
- Sıra Dışılık (İstisna) Analizi (Outlier Analysis)
- Evrimsel Analiz (Evolution Analysis)

2.3.7.1 Tanımlama ve Ayrılama

Verilerin ortak özelliklerine göre genelleştirilmiş sınıflara ayrılmasında belirlenen veri kümelerinin diğer veri kümelerinden farklılıklarını yansıtacak şekilde olmalıdır. Bir tümörde evreler belirlenirken yıl bazındaki sürece göre “oluşum evresi” , “sınırlı evre”, “yayılım evresi” ve “yaygın evre” olarak tanımlanır.

Tanımlama (Characterization): Bir veri kümesinin elemanlarının genel özelliklerini özetlemek amaçlı kullanılan tanımlamaya; bir alışveriş merkezindeki “2014 yılı satış oranı %30’un üzerinde artan ürünler” ifadesi örnek olarak verilebilir.

Ayrılama (Discrimination): Bir veri kümesinin diğer bir veri kümesinden farklarını ortaya çıkarma işlemi olan ayırmaya; “2014 yılı satış oranı %20 artan ürünler ve %30 azalan ürünlerin karşılaştırılması” örnek olarak verilebilir. Tanımlama metodundan farkı mukayese yöntemini kullanmasıdır.

Tanımlama ve ayrılama VM yöntemleri birbirlerine çok benzer yöntemler kullanırlar ve elde edilen sonuçlar pasta grafiği, sütun grafiği, eğriler ve çok boyutlu küpler ile ifade edilir.

2.3.7.2 Birliktelik Analizi ve Sıralı Dizi Analizi

Veri tabanı içerisinde yer alan kayıtların birbirleri ile olan ilişkilerinin incelenerek hangi olayların birlikte gerçekleştiklerinin ya da gerçekleşebileceklerinin belirlenmesi durumlarını çözümleyen VM yöntemlerine birliktelik kuralları (association rules) denilmektedir. Birliktelik kuralları aynı işlem içinde çoğunlukla beraber görülen nesneleri içeren kurallar olup birliktelik kuralları yaygın olarak pazarlama alanında kendine yer edinmiştir. Pazar sepet analizleri adı verilen uygulamalar bu tür çözümlerle yardımcıyla müşterilerin alışveriş alışkanlıkları belirlenmeye çalışılır.

Müşterinin herhangi bir ürün aldığındaki sepetine başka hangi ürünleri de koyduğu belirli bir olasılığa göre ortaya konduğunda; birlikte satın alınan ürünler belirlenip mağazalarda raf düzenlemesi müşterilerin kolayca erişebileceği şekle getirilir. Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın alma eğiliminde olduğunun belirlenmesi, müşteriye daha fazla ürün satma yollarından birisidir. Bu analiz sonucunda bulunan ilişkiler katalog tasarımları, çapraz satışlarda ve müşteri davranışlarını anlama konusunda yardımcı olabilir (Han and Kamber 2001, Özkan 2013). Örnek olarak sepet analizi yöntemi raf dizilimlerinin farklılıklarında büyük rol oynar. Birlikte sık alınan ürünler raflarda yakın yerlere dizilip, beraber satılabilecek ürünlerin satış oranını artırabilirler. Yahut birlikte satılan ürünlerden biri rafın başlangıcında diğeri sonunda olmak üzere raf boyunca ilerleyen müşterinin diğer ürünleri alma olasılığını beraberinde getireceği ihtimalini artırmış olur (Han and Kamber 2001).

Birliktelik kuralları büyük veri kümeleri arasındaki ilginç ilişkileri veya korelasyonları bulmak için kullanılır. Birliktelik kuralı, geçmiş verilerin analiz edilerek bu veriler içindeki birliktelik davranışlarının tespiti ile geleceğe dönük çalışmaları destekleyen bir yaklaşımdır. Doksanlı yılların başına kadar satış verilerinde genelde mali açıdan önemli gelir verilerinin depolanmasına karşın doksanlı yıllardan sonra veri toplama uygulamalarındaki gelişmelere paralel olarak ürün ve müşteri tanıma sistemlerinin yaygınlaşmaya başlamasıyla bir satış hareketine ait verilerin satış esnasında toplanması ve elektronik ortama kaydedilmesinde kolaylık sağlamıştır. Toplanan bu veriler yeni pazarlama stratejileri geliştirmede önemli rol oynamaktadırlar (Sever 2002). Birliktelik analizi yalnızca hizmet ve mal birlikte satın alınması için değil aynı zamanda hangi koşulları sağlayan müşterilerin hangi ürünleri alacağı hakkında da çözümler sunmaktadır. Örneğin bir banka kredi kartı kayıtları incelendiğinde 20-29 yaş aralığında değişen müşterilerinden 1500-1700 lira arasında gelirleri değişen müşterilerin bilgisayar satın aldıkları görülmüştür. Bu kural birliktelik analizi ile;

Yaş ($x, "20...29"$) $\wedge$ Gelir ($x, "1500...1700"$) $\rightarrow$ Alır ($x, "Bilgisayar"$) şeklinde gösterilir.

Genellikle satış uygulamalarında kullanılan birliktelik kurallarının kullanım alanı sadece market ve satış alanı olmayıp sağlık, tıp, ekonomi gibi birçok alanda da kullanıldığı örnekleri görmek mümkündür. Örneğin büyük tıbbi veri tabanı ile yapılan

çalıřmalarda semptomlar (belirtiler) ve hastalıklar arasındaki yeni iliřkiler keřfedilip teřhis iřlemlerine yardımcı olunabilir (Tuę 2005).

Birliktelik kurallarında ilginç birliktelik kuralarından ilginç olmayanları ayırt edebilmek için bazı ölçütlerin belirlenmesi gereklidir. Bu ölçütler destek ve güven deęerleridir. Kuralın destek ve güven deęerleri, kuralın ilginçlięini ve ilgililięini ifade eden ölçütlerdir. Destek deęeri keřfedilen kuralların kullanıřlılıęını (yararlılıęını) ve güven deęeri kuralların doęruluęunu (kesinlięini) ifade eder.

Kuralın sol tarafını ifade eden önce (andecent), kuralın saę tarafını ifade eden sonuç (Consequent), destek deęeri, güven deęeri, min_destek olarak gösterilen minimum destek deęeri, min_güven olarak gösterilen minimum güven deęeri, nesne küme, yaygın nesne kümesi ve aday nesne kümesidir (Çarklı 2010).

X ürünü alan bir müřterinin Y ürünü de alma durumu (birliktelik kuralı) $X \rightarrow Y$ ile gösterilir. Destek ölçütü ařağıdaki baęıntıda görölmektedir.

$$\text{Destek}(X \rightarrow Y) = \frac{\text{Sayı}(X, Y)}{n} \quad (2.3)$$

A ve B ürünlerinin birlikte satın alınma olasılıęı güven deęeridir. Güven deęeri ise;

$$\text{Güven}(X \rightarrow Y) = \frac{\text{sayı}(X, Y)}{\text{sayı}(X)} \quad (2.4)$$

İle hesaplanır.

Destek ve güven ölçütlerinin yanı sıra bu deęerleri karşılařtırabilmek için eřik deęerlerine de ihtiyaç duyulmaktadır. Bulunan eřik deęerlerinin, hesaplanan destek ve güven deęerlerinden küçük olması beklenir. Hesaplanan destek ve güven deęerinin büyüklük derecesi birliktelik kurallarının da o kadar güçlü olduęunu ifade eder (Çarklı 2010).

İki önemli ölçüt olan destek (support) ve güven (confidence) sınırları kullanılarak; küçük boyutlu bir veri setinden elde edilebilecek çok sayıdaki birliktelik kuralları sayısı, geçerlilięi olan kaliteli ama az iliřkilerin bulunduęu kural sayısı elde edilebilmektedir. Birliktelik kural madencilięinin amacı; kullanıcı tarafından belirlenen minimum destek

ve güven değerlerini sağlayan kuralların bulunmasıdır. Birliktelik kuralları temel iki aşamadan oluşmaktadır (Şen 2008):

- Sık geçen nesne kümelerinin belirlenmesi: her nesne kümesinin sık geçenler kümesinde yer alabilmesi için minimum destek şartını sağlaması, yani her nesnenin destek değerinin önceden belirlenmiş olan minimum destek değerinden büyük olması gerekmektedir.
- Sık geçen nesne kümelerinden güçlü ilişki kurallarının oluşturulması: elde edilen ilişki kurallarının minimum destek ve minimum güven şartını sağlamaları gerekmektedir.

Birliktelik kuralları eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılırken, ardışık zamanlı örüntüler ise birbiri ile ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır. “Düşük yağlı peynir ve yağsız alan müşteriler %85 ihtimalle diyet süt alırlar” şeklindeki örnekler birliktelik kurallarına, “X ameliyatı yapıldığında 15 gün içinde %45 ihtimalle Y enfeksiyonu oluşacaktır” şeklindeki örnekler ardışık zamanlı örüntülere örnek olarak verilebilir.

Birliktelik kuralları algoritmaların performansları belirleyen öncelikli adımdır. Birliktelik kuralı çıkarmak için en çok kullanılan algoritma apriori algoritmasıdır.

Apriori Algoritması

Bir önceki adımdan alındığı için “Prior” anlamında apriori algoritması boolean ilişki kuralları için geçerli tekrarlayan niteliğe sahip bir VM algoritmasıdır (Han and Kamber 2001).

Verilerin analizinde kullanılacak apriori algoritması için; veri işlemsel veya çizelge veri şeklinde hazırlanır. Çizelge veri yapısında, işlem numarası ve o belgede bulunan ürünlerin sütununda True veya 1, bulunmayan ürünlerin sütununda False veya 0 değeri olacak şekilde düzenlenirken; işlemsel veri yapısında işlem kodu ve o işlemdeki ürünlerin kodları bir satır oluşturmaktadır (Delibaş 2010).

Sıralı dizi analizi (sequential patterns); “x şirketinin hisse fiyatları ile Y şirketinin hisse fiyatları benzer hareket ediyor” şeklinde gözlem sonuçlarının zaman ve mekân

özelliklerine göre sıralanmış olarak gösteren sayı dizileridir. Sayısal sıralı verilerdeki trendleri ve döngüleri anlamak için kullanılır. Bu fonksiyona ilişkin kayıtlar incelenerek zaman içerisinde sıkça rastlanan trendler bulunur. Bu trendler daha sonra veri içindeki ilişkileri tanımlamak için kullanılır. Örneğin bir beyaz eşya perakendecisinin veri tabanından buzdolabı alımını takip eden beyaz eşya alımının bulaşık makinesi olduğunu belirlenmesi, doğal afetler veri tabanından 6 büyüklüğündeki bir depremin 3 gün sonrasında x yanardağının püskürmesi, banka veri tabanından ilk 3 taksitten 2 veya daha fazlasını geç ödeyen müşterilerin %60 olasılıkla kanuni takibe alınmasının belirlenmesi birer sıralı dizi analizidir. Seriler özelliklerine göre; “zaman serileri”, “mekân serileri”, “bölünme serileri” ve “bileşik seriler” olmak üzere dört başlık altında toplanmaktadır (Çarklı 2010).

Zaman serileri (Time Series); gözlem sonuçlarının zamana göre sıralanmış halidir. Zaman serileri analizi; bir değişkenin zamana bağlı olarak değişen değerlerini inceleyerek gelecekte alacağı değerleri tahmin etmek amaçlanmaktadır (Şişaneci 2009). Borsada yer alan hisselerin davranışları sık rastlanan bir örneğidir. Günlere göre hisse değeri, yıllara göre faiz oranları, aylara göre üretim fire oranı gibi örnekler verilebilir. İki malın zaman içindeki satış miktarları gibi tek bir seri dışında birden fazla hareket serisi arasında da bağlantı kurmak mümkündür. Örneğin dondurma satışları ile salep satışları arasında negatif, dondurma ile kola satışları arasında pozitif bir bağıntı beklenebilir. Zaman serisinde yer alan verilerin davranışları döngüler (cycle) ve trend ile tanımlanır. ”Döngü” veride tekrar eden herhangi bir davranışı tanımlamak için kullanılıp sezonsal veya dönemsel olabilir. Sezonsal olanlar tahmin edilebilir zamanlarda gerçekleşirken dönemsel olanlar “n” zaman aralıkları ile kendini tekrar eder. “Trend” serideki verilerin ortalama değerlerinde yaşanan değişimi tanımlamak için kullanılır. Zaman serisi analizinde veri setindeki davranışları belirlemek kadar gelecek değerleri tahmin etme çalışmaları da gerçekleştirilir. Hisse değerlerini, ekonomik değerleri tahmin etme çalışmaları da gerçekleştirilir. Hisse değerlerini, ekonomik değerleri, ürün talebini, hava durumunu tahmin etmek, makine ve cihazların arıza, bakım ve ömür tahminlerinde, otomatik para çekme makinelerinin nakit yönetiminde vb. kullanılır. 1990-2000 yılları arasındaki ekonomik büyüme ve istihdam arasındaki ilişkinin tek bir ülke için incelenmesi zaman serileri analizine örnek olarak verilebilir (Esen 2009).

Mekân serisi; gözlem sonuçlarının mekâna göre sıralanmış şekli olup bölgelere göre satış rakamları, ülkelere göre yaşam süresi vb. örnek olarak verilebilir.

Bölünme serisi (Frekans); gözlem sonuçlarının belirlenen kriterlere göre sıralanmış şekli iken, birleşik seri; gözlem sonuçlarının iki ya da daha fazla özelliğe göre bir arada gösterilmiş şeklidir (Çarklı 2010).

2.3.7.3 Sınıflandırma ve Öngörü

Veri tabanlarındaki gizli örüntüleri ortaya çıkarmakta kullanılan VM yöntemidir. Sınıflandırma işlemi insan düşünce yapısına en uygun VM yöntemi olup aslında insanlar için bir zorunluluktur. İnsanoğlu etrafındaki olayları ve nesneleri daha iyi anlamak ve başkalarına anlatabilmek için; insanları davranışlarına göre, hayvanları türlerine göre, maddeleri elementlere göre, evleri görünüşlerine göre hemen her şeyi sınıflandırma eğilimindedirler (Berry and Linoff 2004, Bramer 2007).

Verilerin içerdiği ortak özelliklere göre ayrıştırılması işlemi sınıflandırma olarak adlandırılır. Sınıflandırma bir öğrenme algoritmasına dayanır. Tüm veriler kullanılarak eğitme işlemi yapılamaz, bunun yerine veri topluluğuna ait bir örnek veri üzerinde gerçekleştirilir. Öğrenmenin amacı ise; bir sınıflandırma modeli oluşturulması yani; hangi sınıfa ait olduğu bilinmeyen bir kayıt için sınıf belirleme sürecidir. Verilerin sınıflandırılması için; verilerin bir kısmı eğitim amacıyla kullanılarak sınıflandırma kuralları oluşturulur ve bu kurallar yardımı ile yeni bir durumda nasıl karar verileceği belirlenir (Özkan 2013). Sınıflandırma günlük yaşamda çok sıklıkla başvurulanan bir işlem olup, nesneler bölünerek ayrıştırılarak karşılıklı özel ya da genel kategorilerden her biri bir sınıf olarak tanımlanabilir. Pek çok pratik karar verme işlemi bir sınıflandırma problemi olarak formüle edilebilir. Örneğin nesneler ve kişiler birçok kategoriden biri olabilir.

Sınıflandırma “tahminleyici” bir modeldir. Sınıflandırmada verinin önceden belirlenen çıktılarına uygun olarak ayrıştırılması ile veri kümesi denetimli olarak öğrenilmesidir ve hedef değişken değerleri kategorik veya ikili (0 ve 1) değerlerinden oluşmaktadır (Şen 2008, Arslan 2008, Şişaneci 2009).

Sınıflandırma yeni bir nesnenin özelliklerinin incelenmesi ve bu nesnenin daha önceden

özellikleri belirlenmiş bir gruba atanmasıdır. Sınıflandırmada istenilen nesneler genel olarak veri tabanında kayıtlar şeklinde olup, sınıflandırma işlemi sonunda sınıf kodu içeren yeni bir sütun tablonun sonuna eklenir. Sınıflandırma işleminde sınıf tanımlarının doğru yapılması ve bu sınıflara örnek olarak verilen bir eğitim veri seti olması önemlidir. Sınıflandırılmamış bir veri girildiğinde bu veriyi sınıflandırılabilecek bir model inşa etmek temel amaçtır (Ergün 2008).

Sınıf kodu içeren sütun yani sınıflandırma işlemi sayısal değilse problem sınıflandırma, sayısal ise problem regresyon problemi olup aralarındaki temel fark; tahmin edilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır. Sınıflandırma modellerinde sınırlı sayıda sınıf olup, kayıtlar bu sınıflardan birine dahil olması beklenirken; regresyon modellerinde ise bir yada daha çok değişkenin başka değişkenler cinsinden tahmin edilmesini sağlayan ilişkiler bulmak amacıyla geliştirilirler (Berry and Linoff 2004, Alpaydın 2000, Ergün 2008).

Resim, örüntü tanıma, dolandırıcılık tespiti, hastalık tanıları, kalite kontrol çalışmaları, bankada kredi taleplerinin değerlendirilmesi ve finansal pazarlamadaki eğilimler konuları gibi daha birçok alan sınıflandırma teknikleri ile çalışmaktadır. Sınıflandırmayı matematiksel olarak tanımlamak gerekirse:

$D=\{t_1, t_2, \dots, t_n\}$ bir veri tabanı ve her bir t_i bir kayıt olmak üzere,

$C=\{C_1, C_2, \dots, C_m\}$ ise m adet sınıftan oluşan sınıflar kümesi olmak üzere,

$F:D \rightarrow C$ ve her bir t_i bir sınıfa dâhil olmalıdır.

Her bir C_i ayrı bir sınıftır ve her bir sınıf kendisine ait kayıtları içerir. Yani;

$C_j=\{t_i \mid f(t_i)=C_j, 1 \leq i \leq n, \text{ ve } t_i \in D\}$.

Sınıflandırma sorgusu yardımıyla bir kaydın nitelikleri belirlenmiş bir sınıfa girmesi amaçlanmaktadır. Sınıflama algoritmaları, öğrenme verileri ile var olan sınıfları ve bu sınıflara hangi özelliklere sahip kayıtların girebileceğini otomatik olarak keşfeder. Karar değişkeni ve örnek ile sınıflama olmak üzere sınıflama algoritmaları iki şekilde kullanılırlar (Şen 2008).

- Karar değişkeni ile sınıflama: karar değişkeni olarak seçilen bir niteliğin aldığı değerlere göre yapılan sınıflandırma işlemi olup; veri tabanındaki kayıtlar karar değişkeni olarak belirlenen niteliğin değerlerine göre sınıflandırılır.
- Örnek ile sınıflama: veri tabanındaki verilerin pozitif ve negatif veriler olarak kümelenmesi şeklindeki sınıflamadır.

Sınıflandırmada kesin sonuçlar vardır. Eldeki bilgi ve sınıf tanımlarına göre kayıtlar sınıflara ayrılırlar, karar vermede ise sürekli değer biçilen çıktılar vardır. Bazı giriş verilerinden bilinmeyen değişkenler hesaplanır. Örneğin reklam kampanyası için bir kredi kartı şirketi müşterilerini tatil yapanlar ve tatil yapmayanlar olarak iki sınıfa ayırabilir. Ya da tüm müşterilere “tatil yapma puanı” şeklinde bir puan tanımlayabilir. Bu puan müşterinin tatil yapma olasılığını 0 ile 1 arasında bir değerle belirtmektedir. Belirli bir değerın üstünde puanı olanlar tatil yapanlar ya da tatil yapma ihtimali bulunanlar, belirli bir değerin altında olanlar ise tatil yapmayanlar ya da tatil yapma ihtimali bulunmayanlar olarak değerlendirilebilir (Berry and Linoff 2004).

Sınıflandırma algoritmalarına örnek olarak bayes gibi istatistiğe dayalı algoritmalar, k-en yakın komşu algoritması gibi mesafeye dayalı sınıflandırma algoritmaları, karar ağaçları, sınıflandırma ve regresyon ağaçları (CART), destek vektör makinesi ve YSA temelli algoritmalar (Silahtaroglu 2013). Sınıflandırma kategorik verilerin tahmininde kullanılır. Sınıflandırmada genellikle karar ağaçları, YSA, Naive Bayes, Diskriminant analizi, kaba kümeler, regresyon analizi ve genetik algoritmalar gibi teknikler kullanılır.

Öngörü işlemi sınıflandırma işlemine çok benzemekle birlikte; öngörü işleminde sınıflandırma, gelecek için tahmin edilen belirli bir davranışa ya da belirli bir değere göre yapılır. Öngörü işleminde yapılan sınıflandırmanın doğru olup olmadığını test etmenin tek yolu “Bekle ve Gör” prensibini uygulamaktır (Han and Kamber 2001). Kayıt altına alınmış geçmişteki verilerin analizleri gelecekte elde edilen benzer durumlar içinde tahmin niteliği taşımaktadır.

Deprem tahmini, enflasyon tahmini, belirli bir bölgeye yağış miktarı, bir ürüne ait bu seneki ürün hasadı, şirketlerin müşteri istek oranlarının tahmini, otellerin yazlık rezervasyon tahminleri, turizm şirketinin müşterilerinin bu yaz hangilerinin yurt dışına

tatil yapmak isteyeceğinin belirlenmesi gibi işlemler öngörü işlemine örnekler verilebilirken, kredi kartı için başvuran kişilerin risk grubunun tespiti, kampanyalara ilgi duyacak müşteri kitlesinin tespiti, çevrimiçi/canlı/görev kritik sistemler üzerinde yapılan işlemlerde yer alan şüpheli ve yasal olmayan işlemlerin tespiti ise sınıflandırma işlemlerine örnek olarak gösterilebilir (Dinçer 2006, Şişaneci 2009).

Sınıflandırma ve öngörü işleminde en çok; karar ağaçları, YSA, K-en yakın komşu, genetik algoritmalar, naive bayesian sınıflama, bellek tabanlı nedenleme ve lojistik regresyon yöntemleri kullanılmaktadır (Akpınar 2000, Dinçer 2006).

2.3.7.4 Kümeleme Analizi

Eldeki verileri kümelere ya da sınıflara ayırarak belirli bir yakınlık kriterine göre, yani veriler arasındaki benzerlik ve farklılıklara göre gruplara ayrılarak yapılan VM modeline kümeleme denir (Dinçer 2006, Haberal 2007, Tapkan vd. 2011). Çok karmaşık ve büyük sorunları çözmekte izlenilen yöntem, genellikle büyük sorunu daha küçük ve tek başına daha rahat çözülebilecek alt sorunlara bölmek ve her bir alt sorunu çözdükten sonra çözümleri birleştirerek sonuca gitmek şeklindedir. Kümeleme işleminde de, heterojen bir yapıda bulunan veri setinde daha homojen alt gruplar oluşturma ya da alt kümelere bölme işlemi gerçekleştirilir. Nesnelerin alt gruplara (dizinler) ayrılması işlemi ile nesneler örneklenen kitle özelliklerini iyi bir şekilde yansıtan etkili bir temsil gücüne sahip olmuş olup sınıflamanın aksine yeniden tanımlanmış sınıflara dayalı değildir (Şen 2008). Veriler sadece taşıdıkları değerlere göre değil, veri tabanındaki diğer verilerin değerlerine göre değerlendirildiğinden kümeleme sonuçları dinamik olup bu özellik kümelemeyi sınıflamadan ayırır (Birtıl 2011).

Kümeleme verideki benzer kayıtların gruplandırılmasını sağlayan bir teknik olup, hangi teknikle yapılırsa yapılsın süreç aynı şekildedir. Her kayıt var olan kümeyle karşılaştırıp kendisine en yakın kümeye atanır ve bu kümeyi tanımlayan değeri değiştirir. En iyi çözüm bulununcaya kadar kayıtlar yeniden atanır ve küme merkezleri ayarlanır (Akbulut 2006).

Kümeleme işleminde matematiksel tanım ise şu şekilde yapılabilir. Veri tabanı

$D=\{X_1,X_2,X_3,\dots,X_n\}$ $n=1,2,\dots,m$ olsun ve her X_n bir kaydı göstere sin. $X=\{x_1,x_2,x_3,\dots,x_i\}$ $i=1,2,\dots,m$ her x_i numara, sınıf, not ortalaması gibi alanlar olsun. Kümeleme işleminde D veri tabanı j adet kümeye bölünür. Kümelere bölme işleminde uzaklık ve benzerlik kavramlarından faydalanılır (Bırtıl 2011).

Uzaklık aşağıdaki formüller ile hesaplanır.

Öklid mesafesi; en çok kullanılan ölçü birimidir. Veri tabanı $D=\{X_1,X_2,X_3,\dots,X_n\}$ $n=1,2,3,\dots,m$ iken X_m ve X_j arasındaki uzaklık;

$$s(X_m-X_j)=\sqrt{\sum_{i=1}^n (X_{mi}-X_{ji})^2} \quad (2.5)$$

Olarak hesaplanır.

Benzerlik kavramı mesafenin tersi olup iki veri arasındaki yakınlığı ifade etmektedir.

$$\text{ben}(X_m,X_j)=\frac{1}{1+\text{mes}(X_m+X_j)} \quad (2.6)$$

Şeklinde gösterilir.

Manhattan uzaklığı boyutlar arasındaki ortalama farka eşittir. $d_{ij}=|X_{i_1}-X_{j_1}|+|X_{i_2}-X_{j_2}|+|X_{i_3}-X_{j_3}|+\dots+|X_{i_p}-X_{j_p}|$ Olarak ifade edilir. Bu ölçüde birimler arasındaki mutlak uzaklık kullanılır.

Uzaklık ve benzerlik ölçütleri yardımı ile yapılan kümeleme analizinde de temel amaç sınıflandırma da olduğu gibi verileri gruplandırmaktır. Sınıflandırma işleminde verilerin sınıfları önceden belirli iken; kümeleme işleminde ise verilerin sınıfları önceden bilinmemektedir. Sınıflandırmada her veri, önceden belirlenmiş bir takım sınıflar üzerinde yapılan eğitim sonundaki ortaya çıkan modele göre önceden belirlenmiş olan bir sınıfa atanmaktadır. Kümeleme işleminde ise örnek sınıflar veya önceden tanımlanmış sınıflar yoktur, verilerin dağılımlarına göre irdelenerek, birbirlerine olan benzerliklerine göre oluşan doğal sınıflandırmalar ile verilerin kümeleme işlemi yapılmaktadır. Sınıflama modelinde verilerin sınıfları belirlenmiştir ve yeni bir veri geldiği zaman hangi sınıfta olacağı tahmin edilmektedir, kümelemede ise sınıfları bulunmayan veriler belirlenen benzerlik yakınlık kriterlerine göre gruplanarak kümelere

ayrılırlar. Kümeleme analizi denetimsiz sınıflama olarak da tanımlanmaktadır. Kümeleme analizi, büyük ölçekli sistemleri daha küçük bileşenlere ayırarak dizayn ve uygulamayı basitleştirmek amacıyla “Böl ve Yönet” metodolojisi üzerine kurulmuştur. Küme üyeleri birbirine çok benzemekte fakat özellikleri birbirinden farklıdır (Akpınar 2000, Şen 2008). Kümeleme yönteminde temel amaç; kümeler arası benzerliğin minimum, küme içi benzerliğin maksimum olmasını sağlamaktır. Bir kümeleme yönteminin kalitesi bu amacı sağlaması ile doğru orantılıdır. Oluşan sınıfların taşıdığı anlamlar çözümlemeyi yapan kişinin değerlendirmesindedir. Kümeleme yöntemleri aşağıdaki gibi kendi aralarında sınıflandırılmaktadır (Han and Kamber 2001, Dinçer 2006, Altıntaş 2006, Şen 2008, Çataloluk 2012, Silahtaroglu 2013).

- Bölümlemeli (Partitioning) Yöntemler
- Hiyerarşik (Hierarchical) Yöntemler
- Yoğunluğa Dayalı (Density-Based) Yöntemler
- Izgara Temelli (Grid-Based) Yöntemler
- Model Tabanlı (Model-Based) Yöntemler

Bölümlemeli yöntemlerde; verileri bölerek n adet nesneden oluşan veri tabanını, giriş parametresi olarak değerlendiren k adet bölüme ayırma temeline dayanır ve her grubu belirlenmiş bir kritere göre değerlendirirler. K-Means algoritması, K-Medoids algoritması, EM algoritması en çok bilinen bölümlendirme algoritmaları arasındadır.

Hiyerarşik kümeleme; nesnelerin yakınlık ilişkisine göre oluşturulan dendogram (ağaç diyagram) adı verilen ağaç yapısı şeklindeki gruplandırma temeline dayanarak birleştirici ve ayrıştırıcı kümeleme olarak iki bölümde incelenir. Analiz etmeden önce nesneler, hiyerarşik bir yapıya çevirmek için değişik bir yapıya göre düzenlenir. Veriyi hiyerarşik bir yapıya çevirmek için Birch, Cure ve Chameleon algoritmaları gibi farklı yöntemler bulunur. Toplayıcı (agglomerative) hiyerarşik algoritması ve bölücü (divisive) hiyerarşik algoritmaları örnek verilebilir.

Veri kümelerinin önceden belirlenmiş bir kritere göre hiyerarşik yapıda ayıran hiyerarşik modellere karşın, yoğunluk tabanlı yöntemler nesnelerin doğal dağılımını bir yoğunluk fonksiyonu aracılığıyla tespit edip bir eşik yoğunluğunu aşan bölgeleri küme olarak adlandırır. Yoğunluktan kasıt, analiz edilen nesnelerin sayısıdır. Yoğunluk bazlı

metotlar DBscan, Optics ve Denclude Algoritmaları örnek olarak verilebilir.

Izgara tabanlı metotlar veri uzayını incelemek için sonlu sayıda kare şeklinde hücrelerden oluşan ızgara yapılar kullanılır. Ana avantajı genelde birbirinden bağımsız sayıda veri nesnelerinde hızlı işlem zamanıdır. Grid temelli yaklaşımın bazı genel özellikleri: STING; grid hücrelerindeki istatistiksel bilgiyi araştırır, Wavecluster; wavelet dönüşüm metodunu kullanan nesneleri kümeler, CLIQUE; yüksek boyutlu veri alanlarını kümelemek için grid ve yoğunluk temelli yaklaşımı temsil eder.

Modele dayalı yöntemlerde ise eldeki verileri bir matematiksel model ile ifade ederek her bir kümenin bir modele uygunluğu varsayılarak gruplamalar yapılır. Bu metotlar genelde olasılık dağılımlarına göre varsayımlar üretir. İstatistiksel yaklaşım ve COBWEB ile Nöral network yaklaşımı olmak üzere iki ana yaklaşımdan oluşmaktadır (Dinçer 2006, Altıntaş 2006).

En yaygın kullanılan kümeleme algoritması “k ortalamalar algoritması (K-Means)” büyük veri tabanlarının kümelenmesinde kullanılan oldukça etkin bir algoritmadır. Yeni bir veri geldiğinde, algoritma tüm veriyi inceleyerek yeni gelene en çok benzeyenlerin bir alt kümesini oluşturur ve çıktıyı tahmin etmek için bunu kullanır.

Pazarlama araştırmalarında, desen tanımlama, resim işleme ve uzaysal harita verilerinin analizinde, genetik yapıların sınıflandırılması ve yeni yapıların keşfinde, uzaysal/düzlemsel (spatial) veri madenciliğinde coğrafi konuma göre yerleşim yerlerine götürülecek mal ve hizmetler için uygun yerler belirlemede, dünya yüzeyinin uydu görüntülerini, astronomik görüntüleri ve medikal görüntüleri analiz etmede, yapay zekâ alanında makine öğrenmesi için ve tıp, ekonomi, iletişim gibi çok farklı alanlarda gözetimsiz öğrenme yöntemi olan kümeleme analizi metodu kullanılmaktadır (Tuğ 2005, Dinçer 2006, Silahtaroglu 2013, Özkan 2013). Ekonomisi aynı olan ülkelerin bulunması, birbirlerine yakın performans gösteren şirketlerin gruplanması, benzer satın alma davranışı gösteren müşteri gruplarının bulunması, benzer semptomları gösteren hastaların tespiti, gen ve protein analizi, benzer içerikli belge gruplarının araştırılması gibi örnek problemler üzerinde kümeleme çalışmaları yapılabilir (Bramer 2007, Çarklı 2010). Küme tanımı kesin ve net olmayıp en iyi kümelemenin verinin doğasına ve arzu edilen sonuçlara bağlıdır (Birtıl 2011).

2.3.7.5 Sıra Dışılık Analizi

Veri modeline uymayan veri setinden, dağılımından ve veri setinin genel davranışından oldukça farklı ve veri setiyle tutarsız olan veri nesneleri “anormal (sıradışı, istisna, outlier)” olarak adlandırılmaktadır ve analizlerde değerlendirilmezler. Anormallik ölçüm ve uygulama hatalarından kaynaklandığı gibi veri değişkeninin doğal bir sonucu da olabilir. Örneğin yönetim biriminde üst düzey bir çalışanın maaşı, diğer çalışanların maaşlarına göre anormal olarak göze çarpabilir (Han and Kamber 2006).

Birçok VM yöntemi istisnaları, gürültü ve aşırı durumlar olarak algılayıp dikkate almazlar. Bundan dolayı VM algoritmalarının birçoğu, anormallik etkisini minimize etmeye veya tamamen ortadan kaldırmaya çalışmaktadır. Bazı durumlarda istisna noktalar diğerlerine göre çok daha fazla bilgi içermektedir. Bu algoritmanın performansını artırıp model üzerinde olumlu bir etkiye sahip olmasına karşın hileli faaliyet gösterebilecek sahtekârlık tespiti durumlarında önemli gizli bilgilerin kaybolmasına yol açabilmektedir. Böylece anormallik tespit ve analiz “anormallik madenciliği” olarak isimlendirilen yeni ve ilginç bir VM çalışması olarak ortaya çıkmıştır. Çok geniş uygulama alanına sahip anormallik madenciliği; kredi kartlarının ve telekomünikasyon hizmetlerinin olağan dışı kullanımının tespiti, tıbbi analizlerde olağan dışı çeşitli tıbbi harcama davranışlarının tespiti, yeni bir hastalığın başlangıcının tespiti, özel pazarlama müşterilerinin aşırı düşük ve yüksek gelirleri, bilgisayar ağlarına yapılan saldırılar gibi farklı alanlarda kullanılabilir. İstisna analizinde istatistik ve yoğunluk tabanlı yöntem olmak üzere iki yöntem söz konusudur (Han and Kamber 2006).

İstatistik tabanlı yöntem: Dağılım analizi ya da standart sapma hesabı gibi istatistik yöntemlerle istisna olabilecek noktaların tespit edilir ve yoğun hesaplama gücü gerektirmesinden ötürü büyük veri yığınlarında performansı sınırlıdır.

Yoğunluk tabanlı yöntem: Uzaklık tabanlı yöntem olarak ta bilinir. Veri analizinde nesneler arasındaki uzaklıkları dikkate alarak istisna tespiti yaparlar. Bu yöntemin temel prensibi; her noktanın çevresindeki komşuları ile olan yakınlığının hesaplanmasında; Öklid uzaklığı kullanılabildiği gibi veri türüne göre farklı yakınlık hesaplama yöntemleri kullanılarak “yeterince komşu olmayan noktalar” tespit etmektir

Han ve Kamber (2006)'e göre anormallik VM veri kümesinde hangi verinin tutarsız kabul edileceği ve belirlenen anormallik için etkin bir yöntemin bulunması olmak üzere iki alt problemten oluşmaktadır. İstisna analizi aynı zamanda bir kümeleme metodudur.

2.3.7.6 Evrimsel Analiz

Zamanla davranışları değişen nesnelerin eğilim (trend) veya düzenliliklerini (regularity) ortaya çıkarmayı amaçlayan evrimsel analiz; zaman serileri, ardışıklık ve periyodiklik örüntüsü bulma, benzerlik analizi gibi metotları kullanarak temel olarak verinin zaman ile ilişkisini ortaya çıkarmayı amaçlamaktadır. Ayrıca tanımlama, ayırlama, birliktelik analizi, sınıflama ve kümeleme metotlarını da kullanmaktadır. Evrimsel analizin kullandığı her bir yöntem evrimsel analiz adı altında değil, kendi başına bağımsız bir teknik olarak kabul edilmektedir (Han and Kamber 2001).

2.3.8 Veri Madenciliği Uygulama Alanları

VM ve teknolojiye yaşanan gelişmelere paralel olarak işletmelerde ve kurumlarda depolanan veri sayısının artmasıyla birlikte her geçen gün farklı uygulamalar literatürde yerini almaktadır.

Tıp ve mühendislik alanlarında diğer alanlara oranla daha az, sosyal bilimlerde daha yoğun uygulamalar mevcuttur (Baykasoğlu 2005, Çınar ve Silahtaroğlu 2012). Geniş bir uygulama alanına sahip VM çalışmaları için sadece büyük miktarda veri tabanına depolanmış verilerin olması yeterlidir. Bundan dolayı pek çok iş alanı ve sektör için kullanılabilir bir disiplindir (Çataloluk 2012).

Hem genel, hem özel amaçlı birçok uygulama VM çalışmalarında geliştirilmektedir (Arslan 2008). 2012 ve 2014 yılı itibariyle veri madenciliğinin sektör bazında kullanımına ilişkin araştırmanın sonucu Çizelge 2.3'de görülmektedir. Sağlık alanında yapılan çalışmaların az olduğu görülmektedir (İnt.Kyn.14).

Çizelge 2.3 Veri madenciliğinin uygulandığı alanlar (İnt.Kyn.14).

Veri Madenciliği Uygulandığı Alanlar	(2012)	(2014)
CRM/Tüketici Analitiği (49)	%28,6	% 22,2
Bankacılık (37)	% 14,3	% 16,7
Sağlık (36)	% 16,3	%16,3
Perakende (30)	% 14,8	%13,6
Sahtecilik Tespiti (30)	% 12,8	%13,6
Bilim (30)	% 11,7	%13,6
Diğer (30)	% 10,2	%13,6
Finans (24)	% 10,2	%10,9
Reklam (23)	% 13,3	%10,4
Petrol/Gaz/Enerji (21)	-	%9,5
e-ticaret (21)	% 5,1	%9,5
İmalat (20)	% 7,10	%9
Telekom/Kablo (20)	% 6,6	% 9
Sosyal Medya/Sosyal Ağlar (19)	% 12,2	%8,6
Sigorta (19)	% 7,7	% 8,6
Kredi Puanlama (18)	% 7,1	%8,1
Eğitim (17)	% 14,3	%7,7
Doğrudan Pazarlama/Bağış Toplama (16)	% 9,7	% 7,2
Tıbbi/ Pharma (16)	% 6,6	% 7,2
Yazılım (16)	% 5,6	%7,2
Biyoteknoloji/Genomik(15)	% 7,7	%6,8
Arama/Web İçerik Madenciliği (14)	% 8,2	%6,3
Devlet/Askeri (14)	% 5,1	% 6,3
Otomotiv (13)	-	% 5,9
İş Gücü Analitiği	-	%5,9
Web Kullanımı/Giriş Çıkarımı (13)	% 6,6	% 5,9
Yatırım/Stoklar (11)	%4,1	% 5,0
Seyahat/Konaklama (7)	%3,1	% 3,2
Mobil Uygulamalar (5)	-	% 2,3
Güvenlik/Anti-terör (5)	%3,6	% 2,3
Oyunlar (4)	-	% 1,8
Eğlence/Müzik/Tv/Fimler (4)	% 4,6	% 1,8
Sosyal Politika/Anket Analizi (4)	% 1,0	%1,8
Önemsiz e-posta/Anti-spam (4)	% 0,5	% 1,8
Sosyoloji/Kar Amacı Gütmeyen (3)	-	%1,4

VM bankacılık, pazarlamacılık, sigortacılık, sağlık gibi farklı alanlarda

uygulanmaktadır. Veri madenciliğinin kullanılmasında sektör farkı öngörülmemekle birlikte, geniş veri ambarlarının meydana getirilmesine imkân sağlayan perakende satış, sigortacılık ve özellikle sağlık gibi alanlarda kullanılması daha yaygın ve doğru bulunmaktadır (Akbulut 2006, Silahtaroglu 2013).

2.3.8.1 Pazarlama ve Perakendecilik

Süpermarketlerde başlangıçta kasalar sadece basit toplama makineleri görevini görürken, günümüzde kasa yerine kullanılan satış terminalleri sayesinde satın alma işlemleri tüm detayları ile saklanabilmektedir. Büyük miktardaki kaydedilen bu veri, VM yöntemleri ile analiz edilerek; veri, seti içerisinde gizli kalmış bağlantılar tespit edilerek işletmelere karar verme sürecinde yol gösterilebilir. Perakende işletmelerinde müşterilerin demografik özellikleri ve alışverişleri ile ilgili veriler müşteri bazında kaydedilmekte, veriler üzerinde analizler yapılarak müşterinin satın alma davranışı, tutum ve eğilimleri hakkında bilgiler üretilebilmektedir. Örneğin işletmeler her müşteriye bir kod numarası verip müşterinin tüm alışveriş hareketlerini izleyerek kayıt etmekte ve bu verileri VM teknikleriyle analiz ederek müşteri karlılığı ve sadakati konusunda yeni bilgilere ve örüntülere ulaşılabilmektedir ve buna göre pazarlama stratejileri geliştirebilmektedir (Alpaydın 2000, Aydoğan 2003, Ergün 2008).

Müşteri satın alma eğiliminin belirlenmesi, servis kalitesinin artırılması, müşteri hatırlama ve memnuniyetinin sağlanması, ürün tüketim oranı, ürünün nakil ve dağıtım düzeninin tasarlanması ve iş maliyetinin düşürülmesi gibi birçok alanda VM çalışmaları yapılmaktadır. Bu alana yönelik yapılan uygulamalar ise;

PILOT: Pazara yönelik çalışmaların yapıldığı uygulama ürünüdür.

IBM: Pos makineler üzerindeki satış bilgilerini çözümlemek ve öneriler sunmak için kullanılır.

SABRE: Müşteri hatırlama yöntemi paketi olup havayolu sanayisi için geliştirdiği veri madencilik uygulamasıdır (Doğan 2007).

Pazarlama alanında yapılan VM çalışmalarına müşteri segmentasyonu, müşterilerin demografik özellikleri arasında bağlantıların kurulması, pazar sepet analizi, çeşitli pazarlama kampanyaları, mevcut müşterilerin elde tutulması için geliştirilecek

pazarlama stratejilerinin oluşturulması, satış tahminleri, çapraz satış analizleri, müşteri değerlendirmesi, çeşitli müşteri analizleri, müşteri ilişkileri yönetimi örnek olarak verilebilir. Satış noktası analizleri, tedarik ve mağaza yerleşim optimizasyonu, alışveriş sepeti analizi ise perakendecilik alanında yapılan çalışmalardandır. Ayrıca e-ticaret noktasında; saldırıların çözümlenmesi, e-CRM uygulamalarının yönetimi ve web sayfalarına yapılan ziyaretlerin çözümlenmesi gibi işlemlerde VM uygulamalarından yararlanılmaktadır (Dinçer 2006, Telcioğlu 2007, Arslan 2008, Özkan 2013, İnt.Kyn.2, İnt.Kyn.13).

Ergün (2008) ürün kategorileri ve ürün sınıfları arasındaki birlikte satış ilişkileri ve ürün sınıflarının birlikte satıldıkları sırada oluşturdukları kümelerin belirlenmesi amacıyla yaptığı çalışmada, bir işletmenin 1 096 125 adet bir yıllık satış fiş verileri VM yöntemleri ile inceleyerek ağ (web) grafikleri, birliktelik kuralları ilişkileri ve kümeleme tablo-grafikleri elde etmiştir. “Birliktelik kuralları analizi sonucunda en büyük kural desteği %18,69 ile içecekler => tatlı ürünler bağıntısında; tatlı ürünler kategorisindeki bir ürünü alan müşterilerin yarıdan fazlası (%52,04) içecekler kategorisindeki bir ürünü de alırken, içecekler kategorisinden ürün alan müşterilerin %36,37’si tatlı ürünler kategorisinden bir ürün almaktadır ve market markası ürün alan müşterilerin yaklaşık %70’i aynı zamanda tatlı ürünler kategorisinden de ürün almaktadır ayrıca kümeleme analizi sonucunda en kuvvetli ilişkiler; makarna- bakliyat (0,403), zeytin-peynir (0,354), çay-şeker (0,337), pasta malzemeleri (un, kabartma tozu vb.), hazır tatlı (puding vb.) (0,308) ürün çiftleri arasındadır” araştırma sonucu elde edilen bulgular arasında en dikkat çekici olanlarıdır.

Şen (2008) bakkal, market ve büfelere toptan tek el ürünleri satan bir firmanın 15 günlük 37 218 adet veri tabanı kaydı kullanılarak yaptığı çalışmada; en sık birlikte geçen tek el ürünlerini bulmak, elde edilen kurallardan stok planlama aşamasının tekrar yapılandırmak ve reyon düzenlemesinin değiştirilerek satışa etkisini artırmayı amaçlamıştır. Firmanın merkez biriminden müşterilere yapılan satış verilerine aşamalı olarak uygulanan apriori algoritması yerel ağ ortamında çalışan PHP tabanlı bir uygulama yazılımı ile gerçekleştirilmiştir. Kullanılan veri boyutunun fazlalığından ötürü algoritma aşamalarında çalıştırılan sorguların sonuç üretimlerinin zaman aldığı gözlemlenmiştir.

Bir süper market zincirinin müşteri işlem verileri kullanılarak yüksek değerli müşteri hedeflerinin optimize edilmesi yapılan bir başka çalışmadır (Bramer 2007).

Çınar ve Silahtaroglu (2013) müşteri memnuniyet ya da memnuniyetsizliğin altında yatan ana ve gizli nedenleri ortaya çıkararak VM yoluyla oluşturulan kuralların tüketicilerin yaş, cinsiyet gibi kişisel özelliklerine bağlı olarak değişimini incelemek için “Pazarlama Araştırması” dersi kapsamında lisans öğrencilerinden 301 kişi ile yaptıkları anket çalışmasında; anketin sonuçları incelenerek elde edilen veriler üzerinde; grafiksel analiz, K-Means ve C5.0 algoritmalar gibi VM kümeleme ve sınıflandırma algoritmalarını kullanmışlardır. İade ve değişim kötü maldaki memnuniyetsizliği memnuniyet anısına dönüştürmektedir, kadınlar kötü mal ile karşılaştıklarında erkeklere oranla daha kolay çözüme ulaşmaktadır gibi sonuçlar analiz sonucu edilen kuralardan bir kaçıdır.

Akbulut (2006) İstanbul yolu üzerinde bulunan bir kozmetik mağazasının sattığı markalardan birinin müşterilerini gruplandırarak her grubun özelliklerine uygun pazarlama stratejileri geliştirmek ve karlı müşterileri elde tutmaya yönelik bir eylem planı hazırlamak amacıyla WEKA yazılımı ile yaptığı çalışmada; sınıflandırma ve kümeleme teknikleri kullanılarak veri modellenmiştir. 100 adet müşteri bilgisi ile yapılan Sınıflama tekniklerinden; NaiveBayes, lojistik regresyon, J.48 karar ağaçları, JRIP ve PART kuralları ile sınır ağları kullanılmış ve doğruluk dereceleri karşılaştırıldığında sınıflandırma için en uygun algoritmanın J.48 karar ağacı olduğu saptanmıştır. J.48 karar ağacı ile müşterilerin ayrılarak rakip şirketi tercih etme (churn) davranışı modellenmiştir. Ayrılma eğiliminde olan müşteri profili belirlendikten sonra bu müşterilere özel kampanyaların düzenlenmesi gerektiği ve pazarlama stratejileri belirlenmesinin mümkün olduğu sonucuna varılmıştır.

2.3.8.2 Bankacılık, Borsa, Sigortacılık ve Finansal Uygulamalar

Bankacılık ve finansal uygulamalardaki verinin tam ve güvenilir olması veri analizi açısından oldukça önemlidir. Kredi kartı sahtekârlığı tahminlemeden, kar analizine; portföy yönetiminden mal fiyat tahminlemeye kadar geniş bir yelpazede çalışmalara yapılmaktadır. Bu alanda yapılan uygulamalar ise;

Netprophet: Gerçek ve tahminlenen stok değerleri şeklinde gösterimle yapılan stok tahmin uygulamasıdır.

Falcon: Bankalarda kullanılan HNC adlı şirketin ürünü olup; sahtekârlık girişimlerini tespit etmeye yönelik uygulamadır (Doğan 2007).

Kredi kartı dolandırıcılık tespiti, müşteri segmentasyonu, kredi taleplerinin değerlendirilmesi, usulsüzlük tespiti, risk analizi, risk yönetimi, yeni poliçe talep edeceklerin tahmininde, farklı finansal göstergeler arasındaki gizli ilişkilerin ortaya konulması bankacılık alanında yapılan çalışmalara örnek olarak verilebilir (Dinçer 2006, Telcioğlu 2007, Arslan 2008, Özkan 2013).

Yeni poliçe talep edecek müşterilerin tahmin edilmesi, sigorta dolandırıcılıklarının tespiti, riskli müşteri tipinin belirlenmesi sigortacılık alanında yapılan başlıca VM uygulamaları olup, hisse senedi fiyat tahmini, genel piyasa analizleri, alım satım stratejilerinin optimizasyonu ise borsa alanında yapılan başlıca VM çalışmalarıdır (İnt.Kyn.2, Telcioğlu 2007, Özkan 2013).

Bir kredi kartı şirketinin müşteri işlemlerinin verilerinden oluşan veri ambarını kullanarak dolandırıcılık tespit edilmesi yapılan bir başka çalışmadır (Bramer 2007).

Çakır (2005) Türkiye Cumhuriyet Merkez Bankası İstatistik Genel Müdürlüğü Reel Sektör Verileri Müdürlüğü veri tabanında 22 000 firmadan imalat sanayinde faaliyet gösteren 4000 firmanın finansal tabloları ile 165 öznitelik ile yapılan firma başarısızlığının dinamiklerini anlamak üzere yapılan ampirik analizlerin ele alındığı çalışmada; mantıksal makine öğrenmesi yöntemleriyle 3715'i WEKA algoritmalarıyla, 184'ü BCFP algoritmasıyla olmak üzere toplam 3899 adet kural tam ve indirgenmiş veri setleriyle öğrenilmiştir.

AXA OYAK sigorta işlemleri SAS veri ambarı ile tespit edilmiş sahtekârlıklarla ilgili kayıplara ilişkin veri kümeleri arasındaki ilişkileri ortaya çıkartarak müşteri verilerinde bölümlenme yaparak sigorta işlemlerinin %5'inin hileli işlemlerden kaynaklandığını ortaya çıkarmışlardır. Turkcell SAS ile başlattığı kredi derecelendirme projesi ile müşterilerin farklı ödeme davranışları hakkında yeni bilgilere sahip olmuş ve doğru ödeme modelleri kurularak müşteri memnuniyetini artırmanın yanı sıra tahsilat

sürecindeki maliyet ve riskleri azaltarak büyük bir avantaj elde etmiştir (İnt.Kyn.11).

Altıntaş (2006) bir bankanın müşterilerini kümelere ayırarak hangi tip müşterilerin kredi ödemesini zamanında yaptığını, ödemeleri geciktirenleri ve hiç ödeme yapmayanları tespit etmek amacıyla yaptığı çalışmada; 12 alan ve 600 kayıttan oluşan banka müşterilerinin bulunduğu müşteri bilgilerinden oluşan veri tabanındaki verileri değerlendirmek için kümeleme yöntemi ve K-Means algoritması kullanılarak bankanın bir sonraki durum için kredi talebinde bulunan müşteriler için model geliştirmiştir.

2.3.8.3 Eğitim

Eğitimi etkileyen faktörlerden en önemlisi öğrencilerin hangi faktörlerden daha fazla etkilendiklerinin belirlenmesidir. Bu faktörlerin belirlenmesi ile etkisinin azaltılması veya ortadan kaldırılması eğitimin kalitesini ve performansını artırabilir. Değerlendirmenin yapılmasına konu olan veriler (aile, okul, çevre vb.) oldukça fazla olduğu için VM sonuçları analiz etmekte oldukça yardımcı olmaktadır (Birtıl 2011).

Öğrenci işlerinde veriler analiz edilerek öğrencilerin başarı ve başarısızlık nedenleri, başarının artırılması için hangi konulara ağırlık verilmesi gerektiği, üniversite giriş puanları ile okul başarısı arasında bir ilişkinin var olup olmadığı vb. soruların cevabı VM çalışmaları ile bulunarak eğitimin kalitesi ve performansı artırılabilir (Baykasoğlu 2005).

2001 yılı Fırat Üniversitesi Teknik Eğitim Fakültesi Elektronik ve Bilgisayar Eğitimi Bölümü'nde okuyan 250 öğrencinin 12 kültür dersinden aldıkları notlar dikkate alınarak nasıl bir dağılım gösterdiği aralarında nasıl ilişkiler olduğu ve bu derslerde ne gibi kuralların bulunduğu analizinin yapıldığı bir çalışmada; ÖSS sayısal puan ile gelen ve yüksek puanlarla bölüme yerleşen öğrencilerin özellikle sayısal derslerinin C aralığında olması, Fizik-II Türkçe-I ve Atatürk İlkeleri ve İnkılap Tarihi-I derslerinden C notu alan öğrencilerin tamamının Matematik-III dersinden de C notu almaları öğrencinin dersleri ile notları arasında bulunan bazı sonuçlardır (Karatabak ve İnce 2004).

Pamukkale Üniversitesi Eğitim Fakültesi İlköğretim Bölümü Sınıf Öğretmenliği Ana Bilim Dalı'na 1999, 2000, 2001 yıllarında kayıt yaptıran 1031 öğrencinin KPSS puan bilgileri, lisan eğitim sürecindeki bazı derslerin geçme notları, genel not ortalamaları ve

öğretim türlerinden oluşan verileri ile öğrencinin üniversiteye giriş yılına göre dört farklı veri kümesi oluşturularak YSA ve regresyon yöntemlerinin başarılarının karşılaştırıldığı çalışmada; YSA modelinin farklı ağ parametreleri ile her veri kümesi açısından daha başarılı sonuçlar ürettiği görülmüştür (Özçınar 2006).

Bırtıl (2011) öğrencilerin başarısızlık nedenlerini belirlemek için Afyonkarahisar ilinde Kız Teknik ve Meslek Lisesi'nde öğrenim gören 555 öğrenciye, “Öğrenci Başarısızlık Nedenleri Anketi” uygulamış ve elde edilen 542 öğrencinin verileri ile VM çalışması gerçekleştirmiştir. Kümeleme yönteminin uygulandığı çalışmada veriler üç kümeye ayrılmıştır. Her kümenin incelenmesi sonucunda, öğrencileri başarısızlığa iten etkenler ve aralarındaki ilişkiler tespit edilmiştir. Anne eğitim düzeyinin düşük olmasının öğrencilerin başarısız olmalarında etkili bir kriter olduğu, tespit edilen sonuçlardandır.

3391 öğrenciye ait ÖSYM tarafından gönderilen ve öğrencinin yerleştiği fakültesi, üniversiteye giriş yaşı, kaçınıcı tercihi, ÖSS puanı, cinsiyet, üniversitede aldığı derslerin ortalaması, kaldığı dersler, anne ve baba eğitim durumu, liseden sonra kaçınıcı yılında üniversiteye yerleştiği verileri ile akademik başarı olarak öğrencilerin not ortalamasının ve kaldığı ders sayısının ele alındığı çalışmada VM teknikleri ile akademik başarıya etki eden nedenler tespit edilmeye çalışılmıştır. Analizler sonucunda; üniversiteye giriş yaşı, üniversiteyi kaçınıcı sıradan tercih ettiği ve bekleme süresi ile not ortalaması ve kaldığı ders sayısı arasında negatif bir ilişki, ÖSS puanı, cinsiyeti ve yerleştiği fakülte ile not ortalaması ve kaldığı ders sayısı arasında pozitif yönlü bir ilişki bulunurken, anne baba öğrenim durumu ile not ortalaması ve kaldığı ders sayısı arasında bir ilişki bulunamamıştır (Taşdemir 2012).

Hatipoğlu (2013) VM modelleri kullanılarak dershanede aldığı eğitim programına göre katılım süresinin ve branş derslerinin öğrencinin üniversiteye yerleşmesindeki katkısının incelendiği çalışmasında; Matematik, Geometri, Türkçe ve Coğrafya derslerinden başarılı olan öğrencilerin lisans bölümlerine yerleşme oranlarının oldukça yüksek olduğu, Matematik ve Geometri derslerinden başarılı olmayıp, Türkçe ve Coğrafya derslerinde başarı gösteren öğrenciler ön lisans bölümlerini tercih etmekte, Matematik, Geometri, Türkçe ve Coğrafya derslerinde başarılı olamayan öğrencilerin ya tercih yapmadığı yada herhangi bir bölümü tercih etmesine rağmen yerleşemediği elde edilen sonuçlardan bazılarıdır.

Veri tabanlarında öğrenciler ve eğitimler ile ilgili birçok veri tutulmaktadır. Tek başına anlamsız olan bu bilgilerden VM teknikleri ile anlamlı sonuçlar alınabilmekte ve böylece eğitim kurumları için önemli bilgilere ulaşılabilir (Özarslan ve Barışçı 2014).

Gürlüer vd. (2007) Muğla Üniversite İİBF öğrencilerinin eğitimdeki profilini ortaya çıkarmak için 1995 ve sonrası 4467 öğrencinin verileri üzerinde yapılan analiz sonucunda; düz liseden orta derecede mezuniyete sahip İstanbul ve İzmir'in de yer aldığı Ege ve Marmara coğrafi bölgelerinden gelen öğrencilerin orta sıralarda yoğunlukta tercih ettikleri, öğrencilerin yarısından azının okulu normal sürede tamamlamasına karşın belirgin bir şekilde bir yıllık uzatma ve okuldan kaydı sildirmenin en çok ilk iki yılda gerçekleştiği tespit edilmiştir.

Kayri (2008) öğrencinin performans göstergelerinin sürekli izlenebilmesi ve ürünler arasında bulunan örüntülerin bilgisayar sistemleri tarafından yapıldığı e-portfolya değerlendirmeleri için veri madenciliğinde kullanılabilecek yöntemler alternatif bir ölçme aracı olarak kullanılmasını önermiştir.

Bozkır vd. (2009) ÖSYM'nin resmi internet sitesinde 2008 yılında 9952 katılımcıya 80 adet sorudan oluşan ÖSS öğrenci anketinden elde edilen veriler üzerinde sınıflandırma ve kümeleme yöntemleri uygulanmış ve öğrencilerin ÖSS sınavında gösterdikleri başarıyı etkileyen faktörler tespit edilmiştir. Araştırma kapsamında orta öğretim başarı puanı tüm öğrencilerin ÖSS'de ortak başarı ölçütü olarak ele alındığında yaş, okul türü, sanat derslerine ilgi, matematik dersi için harcanan ödev zamanı ve fen bilgisi laboratuvar kullanım oranı bu puan türünü etkileyen faktörlerdendir. Ayrıca dershaneye gitme oranının başarı üzerindeki etkisinin beklenenden daha düşük çıktığı araştırmadaki ilginç bulgulardandır.

Koç (2011) Fırat Üniversitesi Teknik Eğitim Fakültesi Bilgisayar Öğretmenliği Bölümü'nde okuyan 145 öğrencinin sosyal ağları kullanma seviyeleri ve amaçları 26 soruluk bir anket formu ile elde edilen veriler ile sosyal ağların öğrencileri ne derecede etkilediği VM teknikleri ile analiz edildiği çalışmada; birliktelik kuralları ile internete bağlandığında ilk olarak ve mutlaka sosyal ağları kontrol eden ve sosyal ağların yararlı bir şekilde kullanıldığını düşünen öğrencilerin tamamı sosyal ağları

kullanarak arkadaşları ve hocalarıyla daha kolay iletişim kurduğu, sosyal ağlara eski ve yeni arkadaşlarını bulmak için kullanan ve sosyal ağlara düşündüğünden daha fazla zaman harcayan öğrencilerin tamamının bir önceki dönemlere ait ders veya derslerinin bulunduğu gibi sonuçlar elde etmiştir.

Alan (2012) Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü Lisansüstü öğrencilerine ait verilerle; öğrencilerin programı, cinsiyeti, Sivas ilinden ya da başka bir ilden olması, kadrosunun araştırma görevlisi olup olmaması ve ders döneminin farklı olmasının notları etkileyip etkilemediği ayrıca en başarılı sınıflandırmayı yapan algoritmanın tespiti için yaptığı çalışmasında; SimpleCART algoritmasının sınıflandırma oranı en yüksek algoritma olduğu, yüksek lisans programına kayıtlı öğrencilere göre doktora programına kayıtlı öğrencilerin başarı notlarının daha yüksek olduğu ancak, cinsiyetin kadronun, ilin ve ders döneminin başarıyı etkilemediği ortaya çıkan sonuçlardır. Karar ağaçları ile öğrenci verilerinin sınıflandırılması yine Alan (2014b) tarafından yapılan bir başka eğitim alanında VM çalışmasıdır. Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Fakültesi'nde 2012 sonu ve 2013 yılı başı itibarıyla okuyan 4106 öğrenciye ait verilerle yapılan analizlerde LANDTree algoritmasının öğrenci verilerini sınıflandırmada en başarılı algoritma olduğu ve 19 değişik sınıf ürettiği tespit edilmiştir.

Şengür ve Tekin (2013) öğrencilerin mezuniyet notlarının VM metotları ile tahmini isimli çalışmalarında; Fırat Üniversitesi Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü öğrencilerinin mezuniyet notlarının tahmin edilmesi için VM yöntemlerinden olan YSA ve karar ağaçları kullanılmış ve sonuçta yapay sinir ağlarının karar ağaçlarına oranla daha iyi tahmin başarımı sağladığı görülmüştür.

Özarslan ve Barışçı (2014) Kırıkkale Üniversitesi öğrencilerine ait 642 adet veri ile “Temel Bilgi Teknolojileri Kullanım” dersi için yüz yüze eğitim sistemi ile uzaktan eğitim yönetiminin öğrencilerin akademik performanslarına etkisinin araştırıldığı çalışmada; J48 algoritması ile yapılan sınıflandırma sonucunda %82,22 ve çok katmanlı algılayıcı ile yapılan sınıflandırma sonucu %80,74 başarı elde edilmiştir. Yapılan bu analizler sonucunda fakültede okuyan ve dersi uzaktan eğitim yöntemi ile alan öğrencilerin başarı notlarının daha yüksek olduğu görülmüştür.

Çetin ve Sezgin (2011); Akdeniz Üniversitesi Eğitim Fakültesi birinci sınıfta okuyan 173 öğrencinin temel bilgisayar kullanımı bilgi düzeylerini tespit etmek amacıyla yaptıkları çalışmada; K-means kümeleme algoritması ile analiz edilen 56 adet çoktan seçmeli testte en düşük seviyedeki ortalama doğru sayısı 10 ve en yüksek seviyede ortalama doğru sayısı 44 olarak tespit edilmiştir. Öğrencilerin en zorlandığı konuların elektronik tablolar yazılımı (%40 doğruluk oranı), internet ve bilgisayar ağları (%56 doğruluk oranı), kelime işleme yazılımı (%56 doğruluk oranı) ve sunu hazırlama yazılımı (%57 doğruluk oranı) olduğu tespit edilmiştir. Buna karşılık bilgisayar temel bilgi düzeylerinin (bilgisayar temel bileşenleri %77, bilgisayar donanımı %74, işletim sistemi %60) yüksek olduğu görülmektedir. Öğrencilerin bilgisayar bilgi düzey yoğunluğu %57 ile orta seviye olduğu tespit edilmiştir. Önceden alınan bilgisayar eğitiminin olumlu bir katkısı olduğu, bilgisayar sahibi olma durumunun testteki doğru cevap sayısı ile doğru orantılı olduğu, cinsiyete göre başarı düzeyleri arasında önemli bir fark olmadığı analiz sonucunda elde edilen diğer sonuçlardır.

Kurt ve Erdem (2012) öğrencilerin akademik başarılarına etki eden faktörlerin belirlenmesine yönelik VM yöntemleri ile yaptığı çalışmada; Gazi Üniversitesi'nde okuyan 545 öğrenciye 38 soruluk likert tipi anket uygulanmıştır. Çalışma sonucunda; başarı ve başarısızlığa etki eden değişkenlerin saptanmasında alışlagelmiş beklentilerin dışında farklı bakış açıları saptamıştır. Yüksek stres düzeyinin başarıya olumlu etkisi olduğu ve öğrenciler için başarı güdüsü sağladığı, mezuniyet sonrası bölümle ilişkili bir işte çalışmama ihtimali ve araştırmacı kişilik başarıyı ciddi şekilde etkilemekte saptanan sonuçlar arasındadır.

Budak (2013) Kocaeli Üniversitesi'nde 2012-2013 öğretim yılı güz döneminde “Temel Bilgi Teknolojileri Kullanımı” dersini alan sosyal ve sağlık bilimleri alanında lisans eğitimi gören 144 öğrenciye bilgisayar okuryazarlığını etkileyen faktörlerin belirlenmesi amacıyla yapılan anket ile elde edilen sonuçlar VM ile analiz edilmiştir. Analiz sonucunda; ailelerin gelir düzeyinin bilgisayar deneyim süresiyle ilişkisinin yüksek, ancak bilgisayar sahipliği ile ilişkisinin düşük olduğu, gelir düzeyleri ile çocuk sayıları arasında kurulan korelasyon da ise ilişkinin ters yönde ve yüksek olduğu, ailelerdeki çocuk sayısının bilgisayar sahipliğini etkileme oranının düşük olduğu görülmüştür.

Taş vd. (2013) Sakarya Üniversitesi Bilgisayar ve Bilişim Fakültesi Bilgisayar Mühendisliği Bölümü'nde öğrenim görmüş/gören öğrencilerin 1999 ve 2012 yılları arasındaki zorunlu donanım ve yazılım stajı kayıtları kullanılarak bölüm öğrencilerinin staj eğilimleri belirlenmeye yönelik çalışmalarının sonucunda 989 tane kural bulunmuş ve bulgular bölüm staj sorumlularına, staj yapan bölüm öğrencileri ile ilgili bilgiler verilmeye çalışılarak staj etkinliğini artırma ve bölüm staj politikasını iyileştirme kararlarını almada destek sağlamıştır. Öğrencilerin staj olarak en çok yazılım ile ilgili firmaları tercih ettiği geri kalanların ise çoğunlukla kendi okudukları üniversitede staj yaptıkları gözlemlenmiştir.

İstanbul ilinde 2011 yılındaki LYS sınavına giren 42 farklı lise türü ve 871 adet okul, başarı performanslarına göre kümelenmiş ve kümelere ayrışmada hangi test türlerinin etkili olduğunun belirlendiği çalışmada; MF puan türünde kimya, TM puan türünde matematik ve TS puan türünde Türk dili ve edebiyatı testlerinin en etkili test olduğu; tüm okul türleri içinde en başarısız olanının Endüstri Meslek Liseleri, en başarılı okul türü olarak Fen Liselerinin olduğu tespit edilmiştir (Bilen vd. 2014).

Ayık vd. (2007) tarafından Atatürk Üniversitesi öğrencilerinin mezun oldukları lise türleri ve lise mezuniyet dereceleri ile kazandıkları fakülteler arasındaki ilişkinin, VM teknikleri kullanılarak incelendiği çalışmada 1976 yılından itibaren mezun olan ve 2007 itibarıyla halen okumakta olan öğrencilerin bilgilerinin Purple Insight MineSet™ 3.2 programı ile analiz edilerek; lise türünün istenilen fakültenin kazanılmasında çok büyük öneminin olduğu ve lise başarısının da aynı derecede önemli olduğu tespit edilmiştir.

2.3.8.4 Yüzey Analizleri ve Coğrafi Bilgi Sistemleri

Bölgelerin coğrafi özelliklerine göre sınıflandırılması, kentlerde yerleşim yerlerinin belirlenmesi, kentlerde suç oranı, zenginlik-yoksulluk, köken belirleme, kentlerdeki otobüs durakları, otomatik para çekme makineleri ve posta kutuları gibi hizmetlerin konumlandırılmasında kullanılır (Dinçer 2006, Arslan 2008).

Yurtay vd. (2014) Sakarya Büyükşehir Belediyesi İtfaiye Dairesi Başkanlığı'ndan alınan 10 550 yangın verisi üzerinde VM tekniklerinden kümeleme ve birliktelik kuralları uygulanarak yapılan çalışmada; hangi ilçede hangi yangın türlerinin çıktığı,

hangi ay ve mevsimlerde hangi sebeplerden dolayı yangınların çıktığı, hangi yangın türlerinin hangi sebeplerden ötürü çıkmakta olduğu ve hangi itfaiye ekiplerinin hangi bölgelerde çalıştıkları, ne kadar çalıştıkları gibi bilgilere ulaşılmıştır.

Dursun (2005) Elazığ iline ait 40 günlük meteorolojik verilerle yapılan çalışmada, VM teknikleri kullanılarak verilerin değerlendirilmesi ve akıllı örüntü tanıma yaklaşımıyla otomatik hava tahmini yapılmasını sağlanmıştır.

2.3.8.5 Bilim, Mühendislik, Uzay Bilimleri ve Teknolojileri

Ampirik veriler üzerinde modeller kurarak bilimsel ve teknolojik problemlerin çözülmesi alanlarına yönelik yapılan bazı VM uygulamalarına örnek olarak; parmak izi ve yüz tanıma teknolojileri ile kimlik tespiti, tümleşik devre çiplerin üretim kusurlarının azaltılması verilebilir (Bramer 2007, Çataloluk 2012).

Fizik, biyoloji ve kimya bilimlerinde de; organik bileşiklerin analizi, toksik risk analizi, yeni ilaç türlerinin keşfi, bitki türlerinin ıslahı ve termik santral optimizasyonu gibi çalışmalar yapılmaktadır. Çeşitli algılayıcılar aracılığıyla tespit edilen görüntülerden yola çıkarak yol, yüz, parmak izi tanıma gibi görüntü tanıma ve robot görüş sistemleri alanında yapılan VM çalışmalarındandır. Kamuoyu yoklamalarının incelenmesi, genel eğilimler ve seçim öngörülerini belirleme çalışmaları sosyal ve davranış bilimlerinde yapılan VM çalışma örneklerindendir. Uzay bilimlerinde; gezegen yüzey şekillerinin ve gezegen yerleşimleri, yeni galaksilerin keşfi, yıldızların konumlarına göre gruplandırılmasında meteoroloji ve atmosfer bilimlerinde ise bölgesel iklim, yağış haritaları oluşturma, hava tahminleri, ozon tabakası deliklerinin tespiti ve çeşitli okyanus hareketlerinin belirlenmesinde kullanılmaktadır (Dinçer 2006, Arslan 2008).

Adak ve Yurtay (2013) VM modellerinden Gini algoritmasını veri seti üzerine uygulayıp karar ağacı oluşturmayı sağlayan Microsoft Visual C# 4.0 ile bir yazılım geliştirmişlerdir. Geliştirilen yazılımın doğru sonuç verdiğini test etmek için; uluslararası kabul görmüş SPSS (Statistical Packages for the Social Sciences, Sosyal Bilimler için İstatistik Paketi) Clementine 12 programının verdiği sonuçlar ile geliştirilen yazılımın sonuçlarının aynı olduğu tespit edilmiştir.

Gürbüz (2009), havacılık sektöründe VM ile yaptığı çalışmada; Federal Havacılık

Yönetimi, kaza veri sisteminden alınmış sivil uçuşlara ait kaza raporları kayıtları üzerinde, kazalardaki ölüm oranını etkileyen parametreler belirlenmeye çalışılmıştır. Sonuç olarak ölümle sonuçlanan kazalar ile ilgili doğrulukları ve güvenilirlikleri test edilmiş kabul edilebilir düzeyde bazı kural ve ilişkiler elde edilmiştir.

2.3.8.6 Haberleşme ve Endüstri

Ses ve görüntünün teknolojinin gelişimi ile daha kolay aktarılabilir olması ayrıca rekabetinde eklenmesi ile telekomünikasyon alanında kaynakları daha iyi kullanarak servis kalitesini artırmak ve sahtekârlık girişimlerini engellemek için VM kullanımı zorunlu hale gelmiştir. Özel sektörde müşteri kaybının önüne geçilmesi amacıyla etkili ve düşük maliyetli kampanyalar düzenlenmesi başta olmak üzere; kamu sektöründe güvenlik sinyal işleme gibi pek çok alanda başarı ile uygulanmaktadır. Bu alanda kullanılan uygulamalar ise:

FraudBuster: Üyelerin iletişim malzeme kullanım örneklerinin modellenmesi ve sahtekârlıktan şüphelenilen bir çağrının tahminlemesi için kullanılan Coloroda şirketinin ürünüdür.

MapInfo: Aynı isimli şirketin ürünü, kablosuz iletişim ağları ve görsel araçlardaki sorunlu bölgeleri tespit edip yeni stratejiler belirlenmesini sağlayan üründür.

Kalite ve iyileştirme analizleri, hisse tespitleri, hataların yoğunluk tahminleri telekomünikasyon alanında, kalite kontrol analizleri, lojistik, üretim süreçlerinin optimizasyonu endüstri alanında yapılan VM çalışma başlıklarıdır (İnt.Kyn.2, Telcioğlu 2007).

Telefon hatlarındaki parazitlenmeden kaynaklanacak veri kayıplarının ve bunula ilgili olarak da konuşmalarda ortaya çıkan gürültünün yok edilmesi, televizyon yöneticilerinin reklam gelirlerini artırmak için pazar payını en üst seviyeye çıkararak televizyon programlarının düzenlenmesine izin verip televizyon programlarının seyirci payının tahmin edilmesi şeklinde çalışmalara literatürde rastlamak mümkündür (Bramer 2007, Silahtaroglu 2013).

Baykasoglu (2005) çimentonun basma dayanımının tahmin edilmesine yönelik yaptığı

çalışmada; çimento üretimi sürecinde her partiden alınan numunelerin 28 gün bekletilerek basma mukavemetin deneysel olarak belirlenmesi endüstri için uzun bir zaman olması nedeniyle basit regresyon modelleri kullanılarak basma mukavemetin daha hızlı belirlenmesi amaçlanmıştır. 19 değişkenden oluşan 104 günlük üretimden alınan örnek verilerden; 79 günlük üretim verisi, eğitim verisi olarak, 25 günlük veri ise test verisi olarak kullanılıp Gen denklem programlama, YSA ve regresyon analizi VM yöntemleri uygulanmış ve performansları karşılaştırılmıştır. Sonuçta gen denklem programlama diğer yöntemlerden daha iyi sonuç vermiştir. Ayrıca çimentonun basma mukavemeti bir çok kimyasal ve fiziksel faktöre bağlı olup faktörler arasındaki ilişkinin kompleks ve doğrusal olmaması regresyon analizinin kötü sonuç vermesi beklenen bir sonuç olarak yorumlanmıştır.

Kocabaş (2011) 7 boyutlu 186 474 veriden bir internet servis sağlayıcı firmanın 2009 yılında ilk ve ikinci altı aylık dönemlerde ADSL hizmetlerini iptal ettiren abone bilgileri kullanılarak, SPSS Clementine programı yardımıyla abonelerin iptal nedenlerinin modellendiği çalışmada; yılın iki dönemi arasındaki farklılaşmanın alternatif servis sağlayıcı firmaların olması, fiyat kırma stratejileri ve ekonomik nedenlerden kaynaklanabileceği fikrine varılmış, incelemelerde iptal nedenlerinin bireysel tercih ve ekonomik nedenlerin baskın olduğu gözlemlenmiştir.

Beton basınç dayanımının belirlenmesi için alternatif bir metot olarak 72 veri için VM kullanılan bir çalışmada; betonun basınç dayanımının karışım tipi, çimento dozajı, numune alma zamanı ve basınç deneyi yapılan numune boyutuna göre VM kullanılarak belirlenmesi, beton basınç dayanımı üzerinde bu parametrelerin etkisinin belirlenmesi amaçlanmıştır. WEKA programında en yüksek kolerasyon katsayısını veren dokuz algoritma için değerlendirme yapılmıştır. En yüksek kolerasyon değeri Kstar algoritmasından ($R=0,9369$) elde edilmişken en düşük kolerasyon katsayısı Gaussian Process algoritmasından elde edilmiştir. Beton basınç dayanımı üzerinde incelenen veriler içinde en etken parametre sıralaması çimento dozajı>numune boyutu>numune alma zamanı>karıştırma metodu olarak elde edilmiş ve çimento dozajı ve numune boyutunun etkisinin diğer iki etkiye göre 2-3 kat daha fazla olduğu gözlemlenmiştir (Özel ve Topsakal 2014).

2.3.8.7 Devlet Uygulamaları

Devlet kurumları vatandaşların ihtiyaçlarına özel hizmet sunmanın önemini kavramışlardır. Kaynakların doğru kullanımını sağlamak ve planlamak kamu yöneticilerinin, güvenlik problemlerinin kamu güvenliği için önceden tahmin edilmesi, olası güvenlik sorunlarının çözümüne dair izleri keşfederek eş zamanlı çözümler üretebilmek, sağlık ödemlerinde ve vergi işlemlerinde ki yolsuzluk, suiistimal ve şüpheli durumların tespit edilmesi, emniyet birimleri için suç işlemeye meyilli profildeki insanların belirlenip eş zamanlı suç engelleme politikalarının oluşturulması VM ile yapılabilecek çalışmalar arasında gösterilebilir. E-devlet uygulamaları ile kamu kuruluşları ziyaretçilerinin sayfalarını kullanım durumları, formlara ulaşım kolaylığı, en iyi kullanılabilen web sayfasının tasarımı, hangi sayfaların hangi sıra üzeri ziyaret edildiği, ziyaretçi davranış eğilimine göre kurumun web sayfasının vatandaşın ihtiyacına optimum cevap verebilmesi için düzenlemeleri VM ile yapmak mümkündür (Arslan 2008).

Alan (2014a) sınıflandırma ve kümeleme yöntemleriyle Sivas Erzincan Kalkınma Projesi (SEKP) kapsamında hibe talebinde bulunan çiftçilere ait 6096 veri girişi kullanılarak yaptığı VM çalışmasında; verilerin BayesNet algoritması ile yedi ve EM algoritmaları ile üç ayrı kümede kümelendiğini saptamıştır. WEKA programı tarafından desteklenen bütün algoritmalar denenmiş ve BayesNet algoritmasının 549 doğru sınıflandırma algoritması ile en başarılı algoritma olduğu görülmüştür.

2.3.8.8 Sağlık ve Tıp Alanında Yapılan Çalışmalar

Sağlık kurumları içinde toplanan belirsiz ve çok büyük boyutlu dağınık verilerin potansiyeli VM analizleri olmadan gerçekleştirilmesi zordur (Kaur and Wasan 2006).

Test sonuçlarının tahmini, ürün geliştirme, gen haritasının analizi ve genetik hastalıkların teşhisi, kanserli hücrelerin tespiti, yeni virüs türlerinin keşfi ve sınıflandırılması, fizyolojik parametrelerin analizi ve değerlendirilmesi, tıbbi teşhis ve tedavi sürecinin belirlenmesi, tecrübe ile elde edilebilecek bilgilerin önceden keşfedilmesi, çok faktörlü hastalıklarda önemli faktörler arasındaki ilişkilerin ortaya çıkarılması gibi çalışmalar sağlık ve ilaç alanında yapılan VM uygulama alanlarıdır

(İnt.Kyn.2, Tuğ 2005, Dinçer 2006, Telcioğlu 2007).

Biomart: Avrupa biyoinformatik enstitüsü ve Cold Spring Harbor Laboratuvarı (CSHL) tarafından geliştirilmiş, ilişkisel veriler için biyoinformatik dünyasında yaygın olarak kullanılan bir VM sistemidir.

Omim: National Center for Biotechnology Information (NCBI) tarafından geliştirilmiş genetik bozukluk hastalıklarının saklandığı veri tabanıdır.

Decipher: İnsan genom projesinin Sanger enstitüsünde yürütülen projede birçok ülke hastanelerinde insan genomundaki değişimlerle hastalıklar arasındaki ilişkinin anlaşılması için oluşturulan bir veri tabanıdır.

Biyolojik veriler üzerinde VM yapacak araştırmacılara kaynak olan farklı veri tabanları da bulunmaktadır. Yüzbinlerce DNA sırasının bulunduğu bir veri tabanı olan GenBank, EMBL, GDB, proteinlerin 3 boyutlu yapılarının saklandığı veri ambarı olan RCSB PDB bunlardan bazılarıdır (İnt.Kyn.7, İnt.Kyn.8, İnt.Kyn.9).

Bir kanser hastasının kemoterapiye cevap verme olasılığının tahmin edilmesi ve böylelikle bakım kalitesini etkilemeden sağlık bakım maliyetlerinin azaltılması, DNA verisi üzerinde VM yöntemleri ile yapılan analizlerde hastalıklara yol açan gen sıralama örneklerinin tespitinin kolaylaştırılması bu alanda yapılan uygulamalardandır (Bramer 2007, Silahtaroğlu 2013).

Aynı tip hastalıklar için muadili ve daha ucuz ilaç tedavilerinin bulunmasına yardımcı olması için, antipsikotik ilaçların kalp kası hastalıkları üzerine etkisi, solunum fonksiyonları testlerinin analizi, genetik bozuklukların tespiti gibi birçok alanda VM tekniklerinden yararlanılmıştır (Roiger and Geatz 2003).

Örneğin bir hastaneye başvuran ve astım teşhisi konulan hastaların demografik bilgileri hastalığa ve seyrine ait klinik bulgular veri tabanlarında tutulmaktadır. Demografik bilgiler bağımsız değişken değerleridir. Bu veriler yardımıyla gelecek dönemlerde hasta özelliklerine göre hastalık risk tahmin değerleri analiz edilebilir (Esen 2009).

MapInfo şirketi, haritalama teknolojisini kullanarak hastaların yoğun olduğu bölgeleri işaretleyerek bu bölgelere daha iyi hizmet sunabilmektedir (İnt.Kyn.11).

Literatür incelendiğinde yapılan çalışmalara ayrıntılı olarak bölüm 2.4.6’de ayrıntılı olarak yer verilmiştir.

2.3.9 Veri Madenciliği Uygulamalarında Karşılaşılan Sorunlar

VM sürecinin gereksinimleri arasında; erişilebilir veri, etkin erişim yöntemleri, açık problem tanımı, etkin algoritmalar, yüksek performanslı uygulama sunucusu sayılabilir (Akbulut 2006).

Veri madenciliğinde sorunlar aşağıdaki başlıklar altında incelenebilir (İnt.Kyn.3).

- Güvenlik ve sosyal haklar
- Kullanıcı arabirimi
- VM yöntemi
- Başarım ve ölçeklenebilirlik
- Veri kaynağı

Gizlilik ve sosyal haklar:

- Kişilere ait verilerin toplanarak kişilerden izinsiz ve habersiz olarak kullanılması
- VM yöntemleri ile bulunan sonuçların izinsiz olarak açıklanması(paylaşılması)
- Gizlilik ve VM politikalarının düzenlenmesi

Kullanıcı arabirimi:

- Görüntüleme; sonucun anlaşılabilir ve yorumlanabilir hale getirilmesi ve bilginin sunulması
- Etkileşim; VM ile elde edilen bilginin kullanılması, VM yöntemine ve sonucuna müdahale etmek

Veri madenciliği yöntemi:

- Farklı seviyede kullanıcı ile etkileşim halinde olabilme
- Farklı tip veriler üzerinde çalışabilme
- Uygulama ortamı bilgisini kullanabilme
- Elde edilen sonucu anlaşılabilir bir şekilde sunabilme

- Gürültü ve eksik veriler işe çalışabilme
- Değişen veya eklenen verileri kolayca kullanabilme
- Örüntü değerlendirme: önemli örüntüleri bulma

Başarım ve ölçeklenebilirlik:

- Kullanılabilirlik ve ölçeklenebilirlik; örnekleme yapabilme, zaman ve yer karmaşıklığı kabul edilebilir
- Paralel ve dağıtık yöntemler; artımlı VM, parçala ve çöz

Veri kaynağı:

VM girdi olarak kullanılacak ham veriyi eksiksiz, dinamik, geniş ve net veri içermeyen veri tabanlarından alması bazı sorunlar oluşturmaktadır (Aydoğan 2003). Küçük veri kümelerinde doğru ve hızlı bir biçimde çalışan sistem çok büyük veri tabanlarında farklı davranabilir. Bir VM sistemi tutarlı veri ve kaliteli veri ile çok iyi bir performans sergilerken, aynı veriye gürültülü veri eklendiğinde sistemin ciddi oranda kötüleşebildiği görülmektedir. Veri setinin yâda kullanılacak örneklemin büyüklüğü; eğitim ve sına örneklemelerinin bulunabilmesi, büyük setlerde istatistiksel anlamlı sonuçların çıkması, daha yoğun arama ve tarama olanağı, boyutun büyümesine bağlı olarak daha kontrollü deneme gibi nedenlerden ötürü önemlidir. Bununla beraber her zaman büyük veri setleriyle sınıflandırma sonucunda daha küçük setlerle elde edilene göre sınıflandırma doğruluğu potansiyelinin daha yüksek olacağı anlamına gelmemektedir. Örneklem ne kadar büyükse sınıflandırma için öğrenilen kuralların doğruluğu ve kitleyi temsil oranı o ölçüde artmaktadır (Çakır 2005).

2.3.9.1 Veri Tabanı Boyutu

Pek çok makine öğrenme algoritması öğrenimini birkaç yüz satırlık örneklem üzerinde çalıştırabilecek şekilde tasarlanmış olup, veri tabanı boyutlarının hızla artmış olması beraberinde bazı örneklem alma aşamasında sorunlar oluşturmaktadır. Öyle ki örneklemin büyük oluşu, gerçekten var olan örüntülerin gösterilmesi açısından bir avantaj olmasına karşın olası örüntü sayısını da artırmaktadır. VM sistemlerinin en önemli sorunlarından biri olan veri tabanının boyutunun çok büyük olması nedeniyle, VM yöntemleri örneklemeyi veri alanının örnekleme alanı olan yatay ve özelliklerin

bulunduđu kolonların azaltılması şeklinde dikey olarak indirgemeli veya sezgisel bir yaklaşımla arama uzayını taramalıdır (Şen 2008).

Yatay indirgeme, nitelik değęerlerinin önceden belirlenmiř genelleme sıra düzenine göre, bir üst nitelik değęeri ile değęiřtirilme iřlemi yapıldıktan sonra aynı olan çoklukların çıkarılması iken, dikey indirgeme, artık niteliklerin indirgenmesi iřlemi olup özellik seçimi yöntemleri ya da nitelik bağımlılık çizelgesi uygulanarak yapılır (Gürbüz 2009).

2.3.9.2 Gürültülü Veri

Büyük veri tabanlarında veri giriři sırasında yapılan kullanıcı hataları veya girilen değęerlerin yanlış ölçülmesi gibi veri giriři veya toplanmasından kaynaklanan sistem dışı hatalara gürültü adı verilir. Gürültü; rastgele hata veya ölçüm değęerlerindeki değęiřim olarak da ifade edilebilir. Her ne kadar günümüzde geliřmiř VTYS'leri bu tip yanlışlıkları en aza indirmede çok başarılı olsa da, problemler hala yařanmaktadır. Bu durum bir VM yönteminin kullanılan veri kümesinde bulunan gürültülü verilere karşı daha az duyarlı olmasını gerektirir ki, eđer veri kümesi gürültülü ise sistem bozuk veriyi tanımalı ve ihmal etmelidir (Gürbüz 2009, Şentürk 2011).

Uygulamalarda toplanan verilerin yetersiz tutarsız ya da gürültülü olabilirler (İnt.Kyn.3).

Hata sebepleri

- Hatalı veri toplama gereçleri
- Veri giriř problemleri
- Veri giriři sırasında kullanıcıların hatalı yorumları
- Veri iletim hataları
- Teknolojik sınırlamalar
- Veri isimlendirmede veya yapısında uyumsuzluk

Sonuçları:

- Çeliřkili (tutarsız) veriler
- Tekrar eden (çift) kayıtlar

- Yetersiz (eksik) veriler

2.3.9.3 Boş (Null) Değerler

Bir veri tabanındaki boş değerlere “null veri” denir. Hiçbir değeri olmayan veri anlamına gelir ki; bir niteliğin boş olması ile boşluk olması aynı anlama gelmez. Bir hücre hiçbir değer içermiyorsa onun değeri NULL’dır. Bir veri tabanında boş değer birincil anahtarda yer almayan herhangi bir niteliğin değerinin olmaması, özellikle ilişkisel veri tabanlarında sıkça karşılaşılan bir durumdur. Bir nitelik değeri eğer boş ise, o nitelik bilinmeyen ve uygulanamaz bir değere sahiptir (Gürbüz 2009, Şentürk 2011).

2.3.9.4 Eksik Veri

Veri kümesindeki bazı değerler bilinmiyor veya yanlışlıkla girilmemiş olabilir. Örneğin kişiler yaşlarını ve kilolarını söylemekten çekinebilirler veya yıllık gelir çocuklar için kullanılmaz. Her veri nesnesi için sabit bir boyut (özellik sayısı) gerektirdiğinden, eksik veriler veri madenciliğindeki birçok yöntem sorun oluşturur (Şen 2008, İnt.Kyn.4).

Verinin; diğer veri kayıtlarıyla tutarsızlığı nedeniyle silinmesi, bazı nitelik değerlerinin hatalı olması dolayısıyla silinmesi, yanlış anlama sonucu kaydedilememesi, verideki değişikliklerin yapılmaması, cihaz arızaları, veri girişi sırasında bazı nitelikleri önemsiz görülmesi gibi nedenlerden ötürü veri için bazı niteliklerin değerleri her zaman bilinmeyebilir (İnt.Kyn.3).

“Eksik veri nasıl tamamlanır?” sorusuna cevap olabilecek çözümler aşağıdaki şekilde sıralanmış olup daha ayrıntılı olarak bölüm 2.3.4.2’de ele alınmıştır (İnt.Kyn.3).

- Eksik nitelik değerleri olan veri kayıtlarını kullanma
- Eksik nitelik değerlerini elle doldurma
- Eksik nitelik değerleri için global bir değişken kullanma
- Eksik nitelik değerlerini o niteliği ortalama değeri ile doldurma
- Aynı sınıfa ait kayıtların nitelik değerlerinin ortalaması ile doldurma
- Olası en fazla olan nitelik değerleriyle doldurma

2.3.9.5 Artık Veri

Verilerin veri kümesi, var olan probleme uygun olmayan veya artık nitelikler içerebilir. Bu durum birçok işlemde karşımıza çıkabilir. Örneğin var olan problemle ilgili veriyi elde etmek için iki ilişki birleştirilirse, elde edilen ilişkide kullanıcının farkında olmadığı artık nitelikler bulunur. Artık nitelikleri eklemek için geliştirilmiş algoritmalar, özellik seçimi olarak adlandırılır. Artık veri oluşumunu engellemek için özellik seçimi yapmak hem arama uzayını küçültür hem de sınıflama işleminin kalitesini artırmaktadır (Dursun 2005, Şen 2008).

2.3.9.6 Dinamik Veri

Sürekli olarak içeriği değişen kurumsal çevrim içi veri tabanları dinamiktir. Bilgi Keşfi metotlarında bu istenmeyen bir durumdur. Sadece okuma yapan ve uzun süre çalışan BK metodu, bir veri tabanı uygulaması olarak mevcut veri tabanı ile birlikte çalıştırıldığında mevcut programın performansını olumsuz olarak etkilemektedir. Bir diğer olumsuzluk ise; veri tabanındaki bilgilerin kalıcı olduğu varsayılp, BK metodu çevrim dışı veriler üzerinde çalıştırıldığında bu değişimin elde edilen örüntülere yansımaları gerekmektedir. Kuralların aynı kalıp kalmadığı ve tutarlılığı sorunları ortaya çıkar (Hulten vd. 2001, Dursun 2005, Gürbüz 2009).

2.3.9.7 Farklı Tipteki Verileri Ele Alma

Gerçek hayatta sadece sembolik veya kategorik veri türleri değil, tamsayı, kesirli sayı, çoklu ortam verisi, coğrafi bilgi içeren veri tipleri gibi farklı tipte veriler üzerinde işlem yapılması gerekebilir. Saklandığı ortama göre basit veya karmaşık veri tipinde bulunabilen veri, ilişkisel veri tabanında olabileceği gibi, nesneye yönelik veri tabanları, çoklu ortam veri tabanları, coğrafi veri tabanları vb. olabilirler. Veri tipinin fazla olması, VM algoritmalarının tüm veri tiplerini ele alabilmesini olanaksızlaştırdığı için veri tipine özgü VM algoritmaları geliştirilmektedir (Gürbüz 2009).

2.3.9.8 Gereğinden Fazla Anlamsız ve Tutarsız Veri

Aynı veri tabanı içerisinde hem yaş hem de doğum tarihi bilgisinin verilmesi gibi bilgisayar çalışma zamanını artıran ve elde edilecek sonuçların güvenilirliğini ve

kalitesini etkileyebilecek veriler gereksiz veri olarak tanımlanır. Mevcut değişkenlerin birleştirilmesi yani veri boyutunun indirgenmesi bir çözüm seçeneği olabilir. Dalga dönüşümü ve temel bileşenler analizi bu amaçla kullanılan en temel yöntemlerdendir (Silahtaroglu 2013).

2.3.10 Veri Madenciliği Tekniklerinin Performans Ölçütleri

Bir modelin performansı yahut farklı modellerin performansları değerlendirilirken modelin ne kadar hızlı sınıflandırdığı veya model oluşturduğundan ziyade modelin tahmin yeteneği ile ölçülür. Model performans değerlendirme ölçütlerinden en yaygın kullanılan ölçüt doğruluk olup kesinlik, anma ve F-ölçümü ölçütleri de kullanılmaktadır. Ayrıca büyük veri kümelerinde değerlendirme tekrarlı-holdout yöntemi, K-Kat çapraz geçişleme ve bootstrap yöntemi kullanılmaktadır. Bir modelin performansı öğrenim algoritmasının gerisinde sınıf dağılımı, hatalı sınıflandırma maliyeti, eğitim ve test setlerinin boyutu gibi faktörlere de bağlı olabilir (İnt.Kyn.6).

Sınıflamadaki performans ile ilgileniyorsa tahminlenen değerler üzerinden, eğer modelin performansı ile ilgileniyorsa gerçek değerler üzerinden doğruluk oranı (accuracy rate) hesaplanır (İnt.Kyn.16).

Basit Doğrulama (Simple Validation); Bir modelin doğruluğu test edilmesinde kullanılan en basit yöntem basit geçerlilik testidir. Bu yöntem büyük veri setleri üzerinde kullanılıp, verilerin %5-%33 arasındaki bir kısmını test verileri olarak ayırır ve kalan kısım üzerinden modelin öğrenimi gerçekleştirildikten sonra veriler üzerinde test işlemi yapılarak modelin doğruluk oranı hesaplanır. Eğer farklı training Set (eğitim seti) ve test verileri kullanılmaz ise modelin geçerliliği tahmin üstü (over estimate) olarak adlandırılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflandırılan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır (doğruluk oranı=1-hata oranı) (Akpınar 2000, Larose 2005, İnt.Kyn.16).

Sınıflandırmanın başarısı, eğitim ve sınav setleri oluşturulabilecek büyüklükteki örneklemelerin bulunması durumunda; elde edilen kavram tanımının sınıf etiketleri bilinen fakat öğrenme algoritmasına sunulmamış bağımsız bir veri seti (test set)

üzerinde denenmesiyle ortaya çıkar. Sınama verisiyle elde edilen başarı oranı, kavramın ne kadar iyi öğrenildiğini gösteren nesnel bir ölçüttür. Örneklemin küçük olması durumunda ise çapraz doğrulama yöntemiyle sınıflandırmanın başarısı sorgulanır (Çakır 2005).

Çapraz Doğrulama (Cross Validation); model kurmak için az sayıda veri varsa, basit doğrulamadaki kadar veri alınmaz. Çapraz doğrulama tüm veriyi kullanmayı sağlayan bir metottur. Veri seti rastgele iki eşit parçaya ayrılıp, ilk veri seti ile model kurulur ve diğer veri setiyle kurulan model test edilip ve doğruluk oranı hesaplanır. Aynı şekilde ikinci veri seti ile model kurulup kurulan model ilk veri seti ile tahmin edilip, doğruluk oranı hesaplanır. Son olarak tüm veriler kullanılarak model oluşturulur. Önceden hesaplanmış olan doğruluk oranlarının ortalaması ile son oluşturulan model karşılaştırılır (İnt.Kyn.16).

n Kere Çapraz Doğrulama (n Fold Cross Validation); bölüm 2.3.10.2’de ayrıntılı ele alınmıştır.

Harici Doğrulama (External Validation); gerçek hayattaki gerçekliktir. Bazı etkiler modele katılmasa da hayattaki değişim modeli etkileyebilmektedir (İnt.Kyn.16).

Önemli bir diğer değerlendirme kriteri de modelin anlaşılabilirliği olup bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, birçok uygulamada ilgili kararın verilme nedenin yorumlanabilmesi daha büyük önem arz etmektedir. Genellikle karar ağaçları ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir (Larose 2005).

2.3.10.1 Karmaşıklık Matrisi (Risk Matrisi)

Karmaşıklık matrisi, sınıflandırma modeli tarafından test verisi üzerinde yapılan gerçek ve tahmini sınıflar hakkında bilgi veren matrisidir. İki sınıfın olduğu bir sınıflandırıcı modelin karmaşıklık matrisi Çizelge 2.4’de görülmektedir.

Çizelge 2.4 Hata matrisi (Confusion matris).

Gerçek	Tahmin	
	Negatif	Pozitif
Negatif	a	b
Pozitif	c	d

Matristeki a değeri gerçekte negatif sınıfının örneği olan doğru tahminlerin sayısını, b değeri gerçekte negatif sınıfının örneği olan yanlış tahminlerin sayısını, c değeri gerçekte pozitif sınıfının örneği olan yanlış tahminlerin sayısını ve d değeri ise gerçekte pozitif sınıfının örneği olan doğru tahminlerin sayısını temsil etmektedir (Çataloluk 2012). Doğru olarak sınıflandırılan örneklerin sayısı köşegenlerdeki sayıların toplamıdır.

2.3.10.2 K-Kere Çapraz Doğrulama

Eldeki mevcut verinin, modeli eğitmek ve modelin geçerliliğini denetlemek amacıyla iki parçaya ayrılarak öğrenme algoritmalarının karşılaştırılması, değerlendirilmesi için kullanılan yöntem çapraz doğrulama denir.

K-kere çapraz doğrulama çapraz doğrulama yönteminin bir çeşidi olup; verilerin k adet eşit parçaya (fold) bölünmesi ile eğitim ve test işlemleri k iterasyon sayısı kadar gerçekleştirilir. Her iterasyonda farklı bir parça (alt küme) test verisi olarak kullanılıp geriye kalan k-1 alt küme ise eğitim verisi olarak kullanılır. Bu işlem her alt küme bir kez test için kullanılıncaya kadar devam eder, sonuçta her bir alt küme için doğruluk gibi performans ölçütü elde edilir. Bu doğruluk değerlerinin ortalaması alınarak değerlendirme yapılır (Çakır 2005, Çataloluk 2012, İnt.Kyn.16).

2.3.10.3 Doğruluk (Accuracy)

En basit değerlendirme ölçütü olup verileri sınıflandırıcı algoritmaları üzerinde kullanılan eğitim algoritmasına göre eğitim sürecinde etiketlenirler. Etiketlendirilen bu verilerin test sürecinde kontrol amacıyla verilerin doğru ve yanlış etiketlenirilmesi durumları kontrol edilerek; veri mevcut durumdan farklı etiketlenirilmiş ise yanlış,

aynı değerlendirilmiş ise doğru olarak adlandırılır. Sınıflandırıcının doğruluğu, veri setiyle ilgili doğru tahminlerin oranı olarak tanımlanmaktadır. Ayrıca modelle verinin ne derece birbirleriyle uyumlu olduklarının bir göstergesi olmakla birlikte sınıflandırma doğruluğunun başarımların kontrolü doğru etiketlenmiş verilerin sayısının; doğru ve yanlış etiketlenmiş verilerin toplam sayısına eşittir (Çakır 2005, Kaya 2014).

Doğruluk performans ölçütünün yanında, duyarlık ve kesinlik ölçümleri de yapılmaktadır. Bu ölçümlerin yapılabilmesi için Çizelge 2.4’de görülen bazı ifade ve terimlerin gibi açıklanması gerekir (sınıflandırıcı model için pozitif ve negatif iki sınıf olduğu var sayılmıştır):

- TP: Doğruluk pozitif sayısı (a)
- TN: Doğruluk negatif sayısı (d)
- FP: Yanlış pozitif sayısı (c)
- FN: Yanlış negatif sayısı (b)

Bu bilgiler ışığında bir sınıflandırıcı modelin tahminin doğruluk değeri aşağıdaki eşitlik ile hesaplanmaktadır.

$$\text{Doğruluk (\%)} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.7)$$

Görüldüğü gibi doğruluk (accuracy) değeri doğru sınıflandırılmış örnek sayısının, toplam örnek sayısına oranı şeklinde ifade edilmektedir. Doğruluğun sınırlayıcı bazı durumları vardır. Örneğin iki sınıf olan bir problemde; Sınıf A gözlem sayısı=9990 ve sınıf B sayısı=10 olduğunu varsayarsak ve model sınıf A’yı tahmin ederse, doğruluk 9990/10000=%99,9 olur. Model sınıf B’yi algılamadığı için doğruluk bu noktada yeterli değildir (İnt.Kyn.6).

Modelin yapmış olduğu tahminin hata değeri ise aşağıdaki denklemde görüldüğü gibidir.

$$\text{Hata (\%)} = \frac{FP+FN}{TP+TN+FP+FN} \quad (2.8)$$

Hata oranı; doğruluk değerinin birimsel tamamlayıcı olup, yanlış sınıflandırılmış örnek sayısının toplam örnek sayısına oranı şeklinde ifade edilmektedir. Modelin yapmış

olduğu tahminin duyarlılık (hassaslık) değeri ise aşağıdaki bağıntı ile hesaplanmaktadır.

$$\text{Duyarlılık(\%)} = \frac{TP}{TP+TN} \quad (2.9)$$

Duyarlılık değeri, testin gerçek pozitif durumlar içinden pozitif olan durumları ayırma yeteneğidir (Adak ve Yurtay 2013).

Modelin yapmış olduğu tahminin kesinlik değeri ise aşağıda gösterilen bağıntı ile hesaplanmaktadır.

$$\text{Kesinlik(\%)} = \frac{TP}{TP+FP} \quad (2.10)$$

Kesinlik (precision) doğru sınıflandırılmış pozitif örnek sayısının, pozitif sınıflandırılmış örnek sayısına oranıdır. Yani gerçekte herhangi bir sınıfta olan kayıtların sınıflandırma algoritması tarafından ne oranda o sınıfa atandığını göstermektedir.

$$\text{Anma} = \frac{TP}{TP+FN} \quad (2.11)$$

Anma (recall) doğru sınıflandırılmış pozitif örnek sayısının, pozitif sınıflandırılmış örneklerin sayısına oranıdır.

$$F_ölçütü = \frac{2 \times \text{anma} \times \text{kesinlik}}{\text{anma} + \text{kesinlik}} \quad (2.12)$$

F-Ölçütü (F-measure) ise; anma ve kesinlik değerlerinin harmonik ortalamasıdır.

2.3.10.4 Kappa İstatistiği

Tahmin doğruluğunun ölçüsü olup aynı nesneyi derecelendiren iki gözlemci arasındaki uyumu test etmek amacıyla kullanılır. Örneğin aynı bireylerin röntgenini değerlendiren iki klinisyen bulgularını karşılaştırılmasında birbirleriyle ne düzeyde uyum içinde olduğu Kappa değeri gösterebilmektedir. Kappa katsayısı 0-1 arasında değer alır ve buna göre;

0,80 -1: neredeyse mükemmel uyuma,

0,61-0,80: önemli derecede uyuşma,

0,41-0,60: ekseriyetle uyuşma,

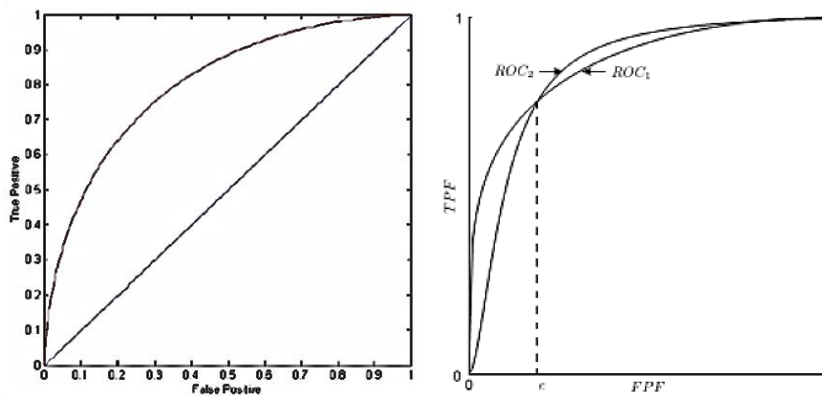
0,21-0,40: orta düzeyde uyuşma,

0,0-0,20: önemsiz uyuşma,

< 0: hiç uyuşma olmamasını tanımlamaktadır (İnt Kyn.19).

2.3.10.5 ROC (Receiver Operating Characteristics) Eğrisi

Alıcı işletim karakteristiği (ROC, Receiver Operating Characteristics) farklı sınıflandırıcı performanslarını test etmek için sıklıkla kullanılan bir yöntemdir. ROC eğrileri ilk olarak İkinci Dünya Savaşında elektrik ve radar mühendisleri tarafından düşman nesneleri örneğin cismin uçak mı, yoksa bir kuş mu olduğunu bulma gibi işlemler için kullanılmasına karşın sonraları psikoloji, tıp, biyometri ve daha sonra ise makine öğrenmesi, VM gibi alanlarında kullanılmaya başlanmıştır. 1950’li yıllarda gürültülü sinyalleri analiz eden sinyal algılama teorisi için geliştirilmiştir. ROC eğrileri tanı testlerinin performanslarını değerlendirmede ve uygun kesim noktası bulmakta kullanılırken ayrıca başka bir uygulaması ise sınıflama analizlerinde performansların karşılaştırılmasıdır. Şekil 2.22’de görüldüğü gibi bir tanı testinin duyarlılığı ile yanlış pozitif oranlarının (1-seçicilik) oluşturduğu grafiğe ROC eğrisi denir.



Şekil 2.22 ROC eğrisi ve keşisen ROC eğrisi (İnt.Kyn.6).

ROC eğrileri y ekseninde doğru pozitif (TP) ve x ekseninde yanlış pozitif (FP) oranına

karşı çizilen bir grafik olup, grafikteki her nokta farklı bir kesim noktası alınarak hesaplanan duyarlılık ve 1-seçicilik değerlerinden oluşur. ROC performans analizi olası bütün kombinasyonlar, yanlış sınıflandırma maliyetleri ve sınıf dağıtımları hakkında bilgi verebilmektedir. ROC üzerindeki her bir nokta sınıflandırıcının oluşturduğu bir modele karşılık gelmektedir (Şişaneci 2009, İnt.Kyn.6, İyisoy 2014).

Grafikte TP ve FP; (0,0) noktasında bütün örnekler negatif; (1,1) noktasında bütün örnekler pozitif olarak sınıflandırılır. (1,0) durumu ise en ideal durum olup, çapraz çizgi rastlantısal tahmindir. ROC grafiği bir sınıflama algoritmasının pozitif durumları doğru olarak belirleme yeteneği ile hatalı olarak sınıflandırılmış negatif durumların sayısı arasındaki ödünleşmeyi incelemek için kullanılan görsel bir araçtır (Yıldırım 2008).

ROC eğrilerinin avantajı sadece seçicilik ve duyarlılık değerlerine bağlı kalınmadan, farklı duyarlılık ve seçicilik değerlerini bir arada değerlendirmeye imkân vermesidir. Böylece maliyet, doğru ve yanlış kararların getirilerinin her seviye için hesaplanabilmesi doğru kesim noktasının bulunmasına yardımcı olur. ROC eğrisi altında kalan tanı testinin performansını gösteren önemli bir ölçüttür. Alanın büyümesi tanı testinde hasta olanla olmayanı daha iyi ayırt ettiğini göstermektedir. Sınıflandırma doğruluğu ROC eğrisi altında kalan alanın büyüklüğüne bağlıdır (İyisoy 2014).

Şekil 2.22’de görüldüğü gibi eşit alanlara sahip iki eğri tanı testlerinde toplam da benzer performans göstermektedir; fakat eğriler aynı olmak zorunda değildir, birbirlerini kesebilirler. Alanlar birbirine yakın olmasına rağmen eğrilerin davranışları farklıdır. Yüksek duyarlılık isteniyorsa 1 numaralı ROC eğrisi, yüksek seçicilik isteniyorsa 2 numaralı ROC eğrisi tercih edilir (İyisoy 2014).

2.3.10.6 Birikimli Kazanç Eğrisi ve Kaldıraç Grafiği

Kaldıraç grafiği bir modelin sağladığı faydanın değerlendirilmesinde kullanılan bir araç olup, eğitim veri kümesine dayalı olarak gerçekleştirilen modelin, test veri kümesinde başarılı olarak yaptığı tahminlerin oranını artan ya da azalan sırada grafiğe yansıtması olarak tanımlanabilir. Kaldıraç grafikleri tahmin modeli olmadan gerçekleştirilen uygulama ile modelin uygulandığı durumda kazancın oranlarını grafiğe yansıtır.

2.3.10.7 Yanlıř Sınıflandırma Maliyeti (Misclassification Cost)

Çalışma sonucunda başarılı olunması halinde elde edilecek kazançlar yahut başarısızlık sonucunda olası mali kayıplar hakkında bilgi verilmelidir.

Hatalı sınıflandırma maliyetlerine baėlı olarak, özgül bir sınıfın doėru sınıflandırılması diėerlerinden daha önemli olabilir. Eėer bu sınıfın dağılımındaki sıklığı daha fazlaysa, standart hata tabanlı algoritmalar özgül sınıfı tahmin etmekte ve toplam maliyeti azaltmaktadır. Hata minimizasyonu tabanlı algoritmalar maliyete duyarlı sınıflandırma amacıyla kullanılabilirler veya ilgi sınıf ya da hedef sınıf örneklerinin kitle ve örneklem içindeki yoğunluğu düşükse, standart maliyet duyarlı yöntemlerin kullanılması amacıyla örneklemdeki sınıf dağılımının deėiştirilmesi gerekmektedir. Sınıf dağılımını deėiştirmekte kullanılan katmanlama yöntemlerinde, sınıf dağılımı sınıflandırma maliyetiyle orantılı olarak deėişmektedir. Hedef sınıfı hatalı tahmin maliyetlerine baėlı olarak düşük yoğunluklu örneklem ve yüksek yoğunluklu örneklem olmak üzere iki tür katmanlama yöntemi kullanılmaktadır. Katmanlama sonucu orijinal veri seti bozulup gizli örüntüler kaybolduėunda elde edilen model ve kavram tanımları gerçeėi yansıtmayacaktır. Sınıflandırma maliyetleri sına, eğitim, hesaplama ve örnek edinme maliyetleri de sınıflandırma maliyetlerindendir (Çakır 2005).

2.3.11 Veri Madenciliėi Programları ve WEKA

Bu bölümde veri madenciliėinde kullanılan programlar ve programların özellikleri kısaca anlatılarak karřılařtırılarak WEKA yazılımını seçim nedenleri üzerinde durulmuřtur.

2.3.11.1 Veri Madenciliėinde Kullanılan Programlar

VM metotlarının uygulanmasını kolaylařtırmak için bu metotları uygulayabilen çeřitli bilgisayar yazılımları geliştirilmiřtir (Altıntař 2010). VM için kullanılan araç çeřitleri ve örnekleri Çizelge 2.5’de görölmektedir (Baloėlu 2006).

Çizelge 2.5 Veri madenciliği için kullanılan araç çeşitleri ve örnekleri.

Kullanılan Araç	Örnekler
Programlama Dilleri	Perl, Python, SQL, XML
Genel Amaçlı	Angoss, Clustran, Cross-Graph, Cross-z, Daisy, Data Distilleries, Database Marksman, DataMind, GVA, IBM Intelligent, Miner, Insightful Miner, Integral Solutions, KXEN, Magnify, MatLab, NeoVista Solutions, Oracle Darwin, Quadstone, SAS, Spotfire, SPSS Clementine, StatPac, Syllogic, WEKA
Biyoenformatik Amaçlı	hinkAnalytics, Thinking Machines, MEME, PIMA, Pratt, PrattWWW, SPEXS

VM işlemlerini yapmak için WEKA, Keel, R, ARTool, RapidMiner, C4.5, Orange, KNIME gibi açık kaynak kodlu programlar ve SPSS, SPSS Clementine, SAS, MATLAB, Angoss, SQL Server ve KXEN gibi ticari programlar kullanılabilir. Kurumsal amaçlarla kullanılmak üzere uygun bir VM yazılımı seçerken dikkat edilmesi gereken hususlar aşağıda belirtildiği gibidir (Danacı vd. 2010, Şaylan 2013):

- Yazılımın uygulayabildiği VM teknikleri
- Yazılımın veri boyutu, kullanıcı sayısı, veri kümesi içindeki özellik sayısı, uygun donanım tipi gibi kapasite kısıtları
- Veri tabanlarına ve dosyalara erişim imkânı
- Oluşturulan modellere ilişkin detaylı açıklama desteği
- Görsel sunum, raporlama formatlarının yetersizliği
- Farklı yazılımlar ile etkileşim uyumu
- Danışma ve eğitim desteği

Kaya ve Özel (2014) açık kaynak kodlu VM yazılımlarının karşılaştırmasını yaptıkları çalışmalarında; altı farklı VM yazılımı detaylı karşılaştırılarak ihtiyaca göre daha etkili bir şekilde kullanılacak olan yazılımlar belirlenmiştir. Çalışma sonucunda içerdiği VM algoritmaları açısından en kapsamlı yazılımların RapidMiner ve WEKA iken en az algoritmaya sahip olan yazılım ise R, makine öğrenmesi paketleri açısından en güçlü yazılım WEKA, metin madenciliği işlemlerini Keel, Orange, Rapidminer ve WEKA kendi başlarına yapabildikleri gözlemlenen sonuçlar arasındadır.

Ayrıca genel olarak; desteklediği dosya formatları, içerdikleri algoritmalar ve makine öğrenmesi paketleri gibi birçok açıdan incelendiğinde en kullanışlı yazılımların WEKA, RapidMiner (Yale) ve Keel olduğu, bu üç yazılım arasından da öğrenim ve kullanım kolaylığı açısından en başarılı yazılım WEKA olarak belirlenmiştir.

Çalışmadaki karşılaştırma sonuçları Şekil 2.23, Şekil 2.24, Şekil 2.25, Şekil 2.26, Şekil 2.27 ve Şekil 2.28’da gösterilmiştir.

Veri madenciliğinde kullanılan bir takım teknik ve algoritmalar sadece bazı türdeki verilerle çalışabilirler. Bazı algoritmalar sadece kategorik değerlerle, bazıları sadece sayısal değerlerle bazıları ise 0-1 değerlerle işlem yapabildikleri için mevcut veri kullanılacak algoritmaya uygun hale getirilmelidir (Tapkan vd. 2011).

	Keel	Knime	Orange	R	RapidMiner (YALE)	WEKA
Görselleştirme	Var	Var (Çok İyi)	Var (6 çeşit)	Var (Çok İyi)	Var (Çok İyi)	Var (5 çeşit)
GUI	İyi	Çok İyi	İyi	Zayıf	Çok İyi	İyi
Grafik	Var	Var	Var	Var	Var	Var
Komut Satırı Arayüzü Bağlantısı	Var	Var	Var	Var (Çok Zayıf)	Var	Var (Çok İyi)
GPL Lisans	Evet (GPLv3)	Evet (GPLv3)	Evet	Evet (GPLv3)	Evet	Evet
Genişletilebilirlik	Evet	Evet	Evet	Evet	Evet	Evet
Esneklik	Evet	Evet	Evet	Evet	Evet	Evet (Az)
Kullanım Kolaylığı	Evet	Evet	Evet	Evet	Evet	Evet (En Kolay)
Öğrenim Kolaylığı	Evet	Evet	Evet	Evet	Evet	Evet (En Kolay)
Hatasız Çalışma	Evet	Evet	Evet (En Az)	Evet	Evet (En Fazla)	Evet (En Az)
Dokümantasyon	Var	Var	Var	Var	Var	Var
Script Yazma	Var	Var	Var (En İyi)	Var	Var	Var
Eklenebilir Paketler	Evet	Evet	Evet	Evet	Evet	Evet
Veri Alma/Verme	Var	Var	Var	Var (Çok Kolay)	Var	Var
Desteklenen Dosya Formatları	.dat, .arff, .csv, .xml, .txt, .prn, .xls, .dif, .html	.arff, .csv	.tab, .basket, .names, .data, .txt, .xls (.arff ve .csv sadece okuyabiliyor)	.r, .txt, .ods, .csv, .xml	.sml, .srff, .stt, .bib, .clm, .cms, .cri, .csv, .dat, .ioc, .log, .matte, .mode, .obf, a bar, one pair, .res, .sim, .thr, .wgt, .wls, .xrff, .arff	.arff, .csv
Veritabanlarıyla Çalışabilme	Var (SQL Veritabanları)	Var (Oracle, MS SQL Server, PostgreSQL, MySQL, Access, ODBC, JDBC)	Var (MySQL)	Var (Informix, Oracle, Sybase, DB2, MS SQL Server, MySQL, PostgreSQL, MS Access, ODBC)	Var (Oracle, MS SQL Server, PostgreSQL, MySQL, JDBC, Sybase, Access, IBM DB2, Ingres, Metin Dosyaları)	Var (JDBC, JDBC aracılığıyla SQL Veritabanları)
Excel Dosyalarıyla Çalışabilme	Evet (import ile)	Hayır	Hayır	Evet	Evet	Evet (Kötü Bağlantı)
Bellek Kullanımı	Limitli	Ayarlanabilir	Limitli	Limitli	Arttırılabilir	Arttırılabilir / Ayarlanabilir
Yazıldığı Dil	Java	Java	Phyton, C++	C, R, C++, Fortran	Java	Java
Kurulum Şartı	Yok	Var	Var	Var	Var	Var
Gerekli Minimum İşletim Sistemi	MS Windows, GNU/ Linux, Mac Os X	MS Windows, GNU/ Linux, Mac Os X	MS Windows, GNU/ Linux, Mac Os X	MS Windows, GNU/Linux, Unix, Mac Os X	MS Windows, GNU/Linux, Mac Os X	MS Windows, GNU/ Linux, Mac Os X

Şekil 2.23 Açık kaynak kodlu VM yazılımları karşılaştırması (Kaya ve Özel 2014).

	Keel	Knime	Orange	R	RapidMiner (YALE)	WEKA
Naive Bayes	Var	Var	Var	Var	Var	Var
Regresyon	Var	Var (Weka)	Var	Var	Var	Var
Destek Vektör Makineleri	Var	Var	Var	Var	Var	Var
KNN	Var	Var	Var	Var (RWeka)	Var	Var
Karar Ağacı	Var	Var	Var	Var	Var	Var
RBF	Var	Var (Weka)	Var	Var	Var	Var
Bagging	Var	Var (Weka)	Var	Var	Var	Var
Lazy Sınıflandırıcılar	Var	Var (Weka)	Var	Var (RWeka)	Var	Var

Şekil 2.24 Sınıflandırma algoritmaları açısından yazılımların karşılaştırılması (Kaya ve Özel 2014).

	Keel	Knime	Orange	R	RapidMiner (YALE)	WEKA
K-Means	Var	Var	Var	Var	Var	Var
EM	Var	Var (Weka)	Yok	Var	Var	Var
SOM	Var	Yok	Var	Var	Var	Var (Eklenti Olarak)
Hiyerarşik Kümeleme	Yok	Var	Var	Var	Var (Modül ile)	Var

Şekil 2.25 Kümeleme algoritmaları açısından yazılımların karşılaştırılması (Kaya ve Özel 2014).

	Keel	Knime	Orange	R	RapidMiner (YALE)	WEKA
Apriori	Var	Var	Var	Var (Paket)	Var (Weka)	Var
FPGrowth	Var	Var	Yok	Var (Paket)	Var	Var

Şekil 2.26 Birliktelik kural algoritmalarının karşılaştırılması (Kaya ve Özel 2014).

	Keel	Knime	Orange	R	RapidMiner (YALE)	WEKA
Kazanç Bilgisi	Var	Var	Var	Var	Var	Var
Kazanç Oranı	Yok	Var	Var	Var	Var	Var
Ki-Kare	Var	Yok	Var	Var	Var	Var
Gini İndeks	Yok	Var	Var	Yok	Var	Var
Genetik Algoritma	Var	Var	Yok	Var (Paket)	Var	Var

Şekil 2.27 Nitelik seçimi yöntemlerinin karşılaştırılması (Kaya ve Özel 2014).

	Keel	Knime	Orange	R	RapidMiner (YALE)	WEKA
Eksik Değer	Var	Var	Var	Var	Var	Var
Kesikleştirme	Var	Var	Var	Var (RWeka)	Var	Var
Veriseti Eğitimi	Var	Var	Var	Var (Kod ile)	Var	Var
Gürültülü Veri Filtreleme	Var	Yok	Yok	Var (Paket)	Var	Var
Normalizasyon	Var	Var	Var	Var (RWeka)	Var	Var
Nominalden İkiliye Dönüştürme	Var	Var	Yok	Yok	Var	Var
Çapraz Doğrulama	Var	Var	Var	Var	Var	Var

Şekil 2.28 Veri ön işleme algoritmalarının karşılaştırılması (Kaya ve Özel 2014).

Tekerek (2011) açık kaynak kodlu VM araçlarını özelliklerine göre karşılaştırdığı çalışmada; YALE, WEKA ve R programlarını açık kaynak kodlu yazılımlar arasında lider olduğunu, kullanım kolaylığı ve yüzlerce özelliği barındırdığından dolayı YALE WEKA'dan üstün olmasına karşın çoğu VM uygulamalarını geliştirmekte WEKA'nın yeterli olduğunu belirtmiştir.

Bu tez çalışmasındaki analizler WEKA 3.6.10 VM platformunda yapılacaktır.

2.3.11.2 WEKA Programı

Yeni Zelanda Waikato Üniversitesi tarafından “WEKA makine öğrenmesi projesi” olarak hazırlanmış ve java tabanlı açık kaynak kodlu genel kamu lisansına sahip VM ve makine öğrenmesi yazılımıdır.

Veri önişleme, sınıflandırma, regresyon, kümeleme, birliktelik kuralları ve görselleştirme araçlarını bünyesinde barındırdığı birçok makine öğrenme algoritması ile VM programları geliştirmek için kullanılan popüler yazılımlardan olan WEKA'nın (Waikato Enviroment for Knowledge Analysis) en güçlü özelliği birçok sınıflama tekniğini içermesi ve uygulamaların komutlar yardımıyla da yapılabiliyor olmasıdır (İnt.Kyn.10). WEKA'nın sürümlerine ve yardımcı araç paketine WEKA'nın sitesinden ücretsiz olarak ulaşılabilir.

1996 yılında ilk resmi sürümü yayınlanan WEKA akademik araştırmalarda, endüstriyel ve eğitim uygulamalarında kullanım yeri edinmiş veri analizi ve tahminleyici modelleme için geliştirilmiş algoritma ve araçların görsel bir birleşimini içerir. Geniş veri önişleme ve modelleme tekniklerini grafiksel kullanıcı ara yüzü sayesinde kullanımının kolay olması ve Java programlama dili ile uygulandığından herhangi bir platformda uygulanabilmesi yani taşınabilir olması WEKA yazılımının temel avantajları arasındadır. Veri önişleme ve VM algoritmaları kapsamında WEKA'nın hiyerarşik yapısı şematize edilebilir (Tapkan vd. 2011).

2.3.11.3 Genel Bilgi

WEKA tüm VM için metotlar içermektedir. Bu metotlar; regresyon, sınıflandırma, kümeleme, birliktelik kuralları oluşturma ve nitelik seçimidir. WEKA'daki bütün algoritmalar basit ilişkisel tablo şeklinde olan ARFF dosya formatını giriş olarak kabul etmektedir. WEKA üzerinde makine öğrenmesi ve istatistik ile ilgili hazır kütüphaneler bulunmaktadır. Uygulamaların komut girilerek de gerçekleştirildiği WEKA yazılımında önişleme (Process), sınıflandırma (Classify), kümeleme (Cluster), birliktelik kuralları (Associate Rules), nitelik seçme (Select Attribute) ve görselleştirme (Visualize) panelleri bulunmaktadır. WEKA, metin tabanlı arff, arff.gz, names, data, csv, c45, libsvm, dat, bsi, xrff, xrff.gz dosya tiplerinin yanısıra , veri tabanlarını ve verilerin olduğu URL adreslerini desteklemektedir (Aydın 2011, Alan 2012) .

ARFF: Makine öğrenmesinde kullanılan değişken tanımlamasına izin veren ASCII metin dosya formatıdır. Veri yığının genel amacını veya ismini belirtmek için @relation, verilerin özelliklerini belirlemek için @attribute ve ham verilerin başlangıcından itibaren @data deyimleri bu dosya formatını oluşturan deyimlerdir.

```

@relation anemivedegil-weka.

@attribute YAS numeric
@attribute CINSIYET numeric
@attribute 'ALT ' numeric
@attribute AST numeric
@attribute BASO numeric
@attribute CRP numeric
@attribute Demir numeric
@attribute DBK numeric
@attribute TIBC numeric
@attribute TSI numeric
@attribute EO numeric
@attribute Ferritin numeric
-
@attribute URE numeric
@attribute VitB12 numeric
@attribute WBC numeric
@attribute ANEMITANI {Anemi,AnemiDegil}

@data

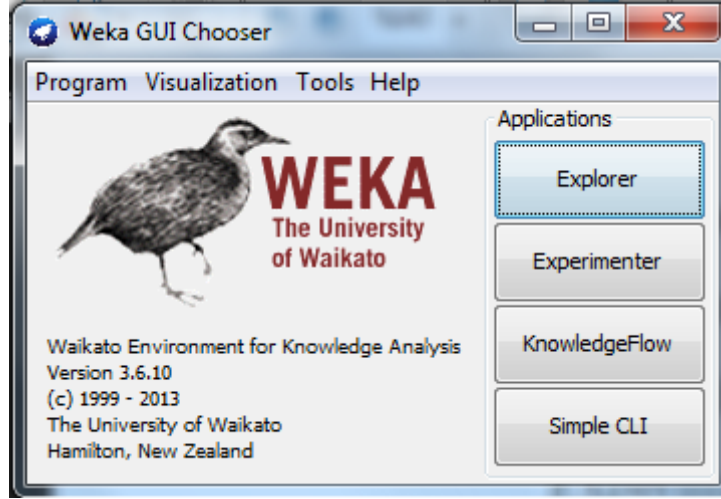
```

Şekil 2.29 Örnek bir Arff dosyası görünümü.

WEKA yazılımında büyük küçük harf duyarlılığı vardır. Şekil 2.29’de örnek bir arff dosyası görülmektedir.

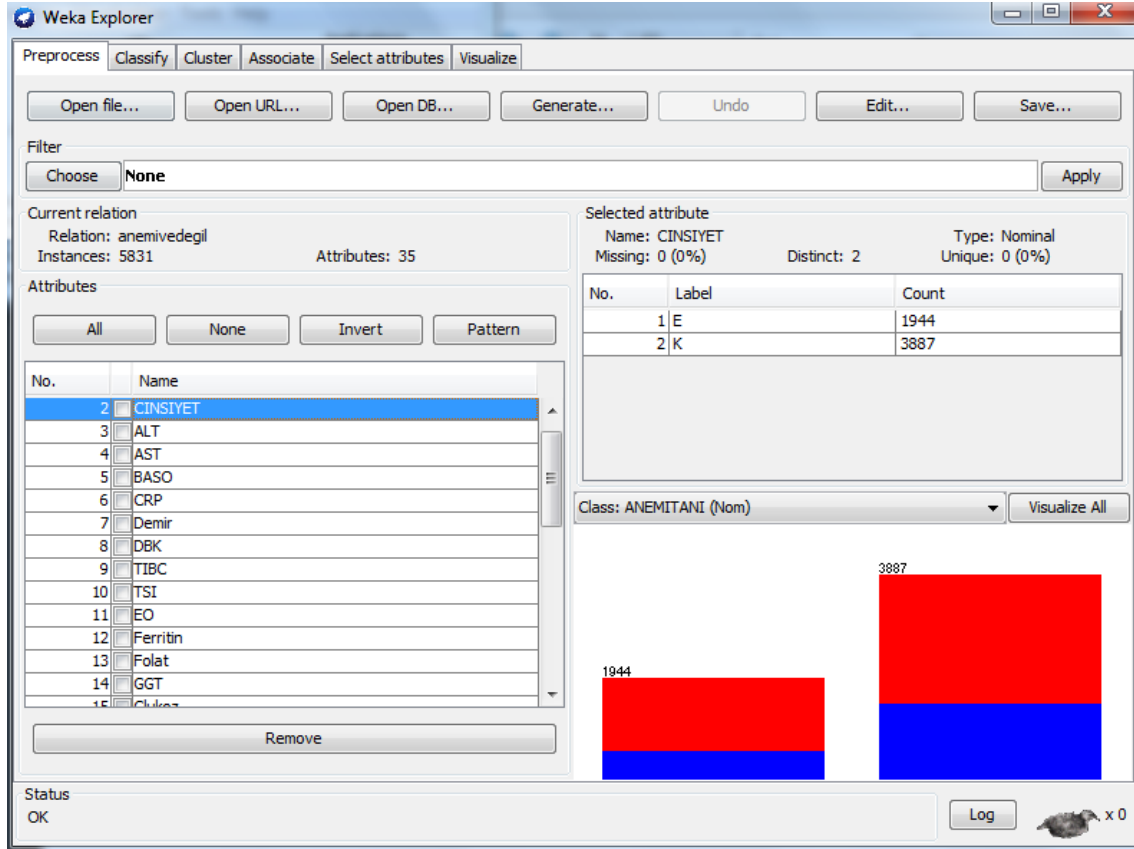
2.3.11.4 Program Modülleri

WEKA üzerinde veri kümesi öğrenme metotlarını uygulayabilir ve veri hakkında daha fazla bilgi sahibi olmak için çıkışı analiz edebilirsiniz. Yeni gelen örnekler üzerinde tahminler yapmak için öğrenilmiş modeller kullanılır. Ayrıca farklı öğrenme algoritmaları uygulamak ve onların yapmış oldukları tahminleri birbirleriyle karşılaştırmak için kullanılır. Öğrenme metotları sınıflandırıcı olarak da isimlendirilmektedir. WEKA içerisinde; Explorer, Exprimenter, KnowledgeFlow ve Simple CLI olmak üzere dört uygulamaya sahip olduğu üzere Şekil 2.30’de görülmektedir (Aydın 2011).



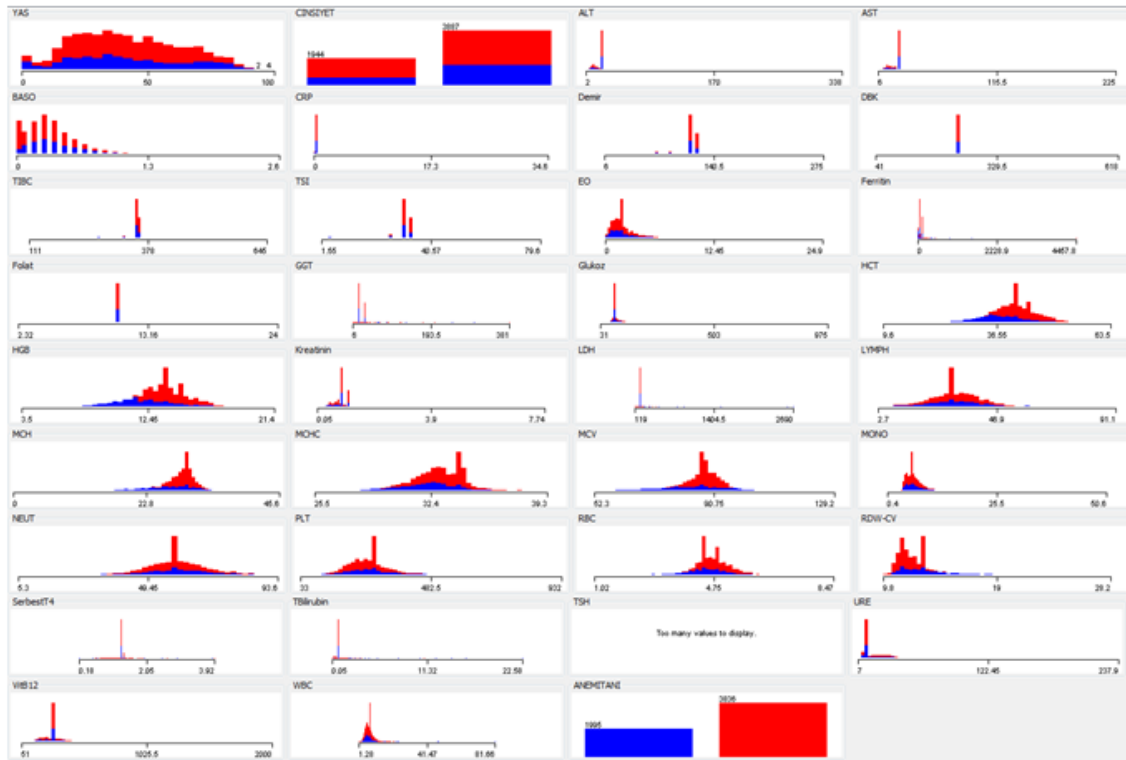
Şekil 2.30 WEKA grafik ara yüzü.

Explorer; WEKA'nın ana grafik arayüzü olan explorer modülü Şekil 2.31'de görüldüğü gibi altı menüden oluşmaktadır. Preprocess menüsü ile eğitim verisi içeren arff, arff.gz, names, data, csv, c45, libsvm, dat, bsi, xrf, xrf.gz dosya türünden biri olarak yüklenir ve veri kümesi ile ilgili ön işlemler yapılır. Classify menüsü ile bir sınıflandırıcı seçilir ve bu sınıflandırıcı eğitim kümesi yardımı ile eğitilir. Cluster menüsü ile birbirine benzer örnekler aynı kümede gruplanır. Associate menüsü ile veri içerisinde birliktelik kuralları elde edilir. Select attributes menüsünde ise veri kümesi içerisindeki ilişkili birçok nitelik seçilir. Visualize menüsünden de öğrenme işlemi sonunda oluşan analizler ve veri kümelerinin analizleri iki boyutlu olarak gösterilir (Aydın 2011).



Şekil 2.31 Explorer modül ekranı.

Temel grafiksel kullanıcı arayüzü olan Explorer ekranında ön işleme (Preprocess), Sınıflandırma (Classify), kümeleme (Cluster), ilişkilendirme (Associate), öz nitelik seçimi (Select Attributes) ve görselleştirme (Visualize) bölümlerinden oluşmaktadır. Veri yükleme işlemi yapıldıktan sonra pasif olan komut ve butonlar, uygun formatta verinin yüklenmesi ile aktif hale gelerek alt kısımda görünen “Status” bölümünde yapılmakta olan ve o zamana kadar yapılmış işlemler hakkında bilgi verir. Sağ alt köşedeki WEKA kuşunu hareket etmesi o anda bir işlemin yapıldığını ve yanındaki sayı kaç adet işlemin yapıldığını göstermektedir (İşler ve Narin 2012). Ayrıca öz niteliklerin dağılımı da görülebilir (Şekil 2.32).



Şekil 2.32 Öznitelik dağılım histogramı.

Preprocess sekmesinde; verinin yüklenmesi ve çeşitli filtreleme işlemleri yapılmaktadır.

Classify sekmesi; çeşitli sınıflandırma algoritmaları veri setine “test option” seçeneklerinden; eğitim setinin tamamı (use training set), ayrı bir veri setine (supplied test set), parçalara ayırarak (cross-validation) ve belli bir oranlarının (percentage split) test amaçlı kullanılması ile uygulanabilir. Sınıflandırıcı ismi, içerdiği örnek sayısı, öznitelik sayısı, doğru ve yanlış şekilde sınıflandırılan örnek sayıları, doğruluk tablosu, ortalama mutlak hata gibi birçok hata gösterilirken, sınıflandırıcı çıkışında (Classifier output) bölümünde çalışma bilgileri, hata matrisi ve sınıflandırma ile ilgili özet bilgiler verilmektedir (İşler ve Narin 2012).

Use Training Set; tüm veri setinin hem modeli geliştirmekte hem de test edilmesinde kullanıldığından sınıflandırma başarısı yanıltıcı olabilir.

Supplied Test Set; modelin geliştirildiği ve test edildiği veri setlerinin farklı olarak yapılan analizlerde kullanılır.

Cross Validation; veri setinde tek bir veri seti dosyası 10 eşit parçaya bölünüp, her defasında 9 parça üzerinde model geliştirip kalan bir parça üzerinde test edilmektedir. Böylelikle analiz 10 kez tekrarlanmış olmaktadır.

Percentage Split; veri setimizin rastgele bölündüğü bu sınıflama seçeneğinde veri setinin bir kısmı (yaklaşık 2/3) modeli geliştirip, bir kısmı üzerinde (yaklaşık 1/3) modeli test etme işleminde %66 (en verimli oran) model geliştirmede %33 modeli test etme işleminde kullanılır.

Cluster sekmesi; kümeleme algoritmaları ile ilgili birçok seçeneğin bulunduğu pencerede seçilen SimpleKMeans, OPTICS, EM, HierarchicalCluster algoritmaları ile sınıflandırmadaki parametre ve “clusterer mode” “test option” seçimleri ile benzerlik göstermektedir. Analizde kullanılmak istenmeyen öznitelikler “Ignore attributes” butonu ile çıkarılarak kalan özniteliklerin başarımı ne kadar etkilediği gözlemlenebilir.

Associate sekmesinde; birliktelik kurallarının uygulanması işlemi için “associater” ile algoritma seçimi yapılır. Seçilen ilişkilendirici seçimi ve ayarları yapılmaktadır. “Associater output” ve “Result list” bölümlerinde algoritma ile ilgili sonuçlar görüntülenmektedir (İşler ve Narin 2012).

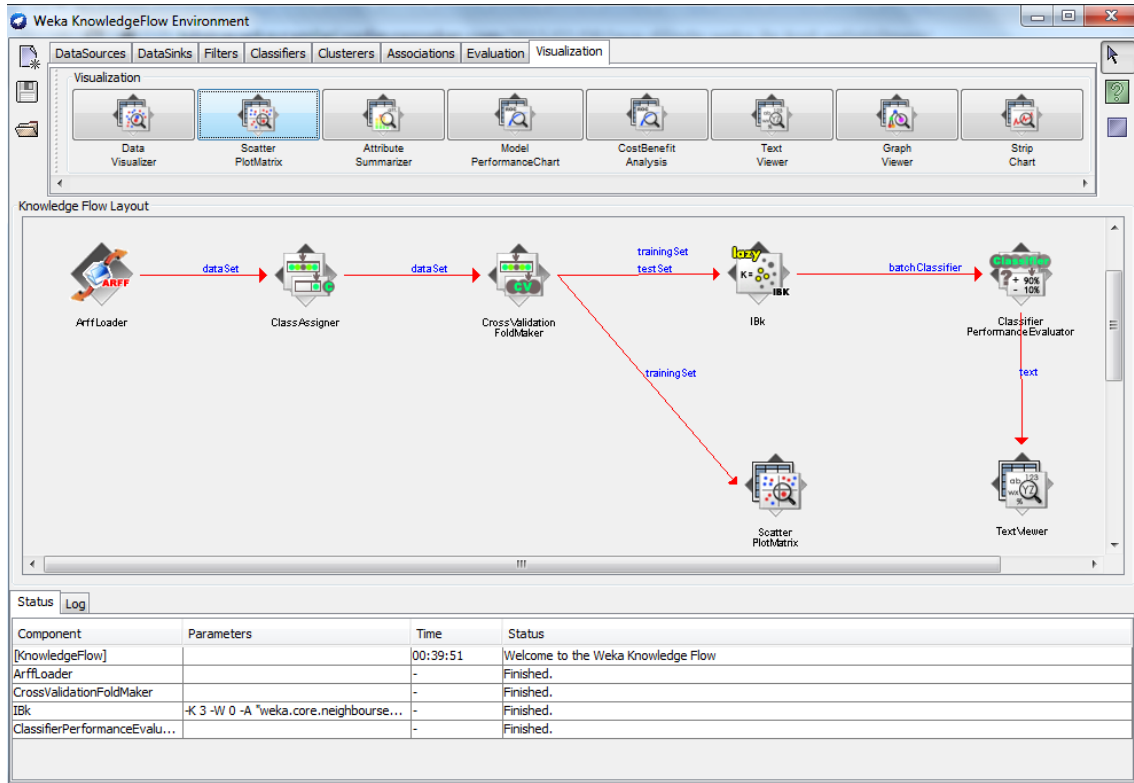
Select attributes sekmesi; öznitelik seçimi yapılabilmek için özniteliklerin her bir alt kümesinin uygunluk derecesinin belirlenmesi için kullanılacak değerlendirici yöntemi için “Attribute Evaluator” ve özniteliklerin içerisinde nasıl arama yapılacağını belirlemek için “Search Method” yöntemlerinin seçilmesi gerekmektedir. “Attribute selection mode” ile eğitim kümesinin tamamı (Use Training Set) ya da çapraz doğrulama (Cross-Validation) yöntemlerinden biri seçilmelidir. Seçilen öznitelikler “Attribute Selection Output” başlığı altında görüntülenmektedir (İşler ve Narin 2012).

Visualize sekmesinde; veri setinin örneklerinin özniteliklere göre 2 boyutlu dağılımının şekillerinin nasıl olduğu gösterilmektedir. İstenilen öznitelik seçimine göre görsel inceleme yapılabilir (İşler ve Narin 2012).

Experimenter; Explorer ve KnowledgeFlow ortamları makine öğrenmesi planlarının verilen bir veri kümesinde ne kadar iyi karar verdiği konusunda kullanıcılara yardım eder. Parametre ayarlamaları, farklı veri kümeleri üzerinde çalışan öğrenme

algoritmalarının sınanması gibi kullanıcıların geniş ölçekte deneyler gerçekleştirebilmelerine olanak sağlar (Aydın 2011). Bir veri setine farklı yöntemler uygulayarak ya da aynı tekniği farklı parametrelerle tekrarlayarak, tek seferde birden fazla deneyin gerçekleşmesini sağlayarak, deneylerin gerçekleştirilmesi ve öğrenme planları arasındaki istatistiksel testleri yapan bir arayüz seçeneğidir.

KnowledgeFlow; Explorer’a alternatif olarak geliştirilmiş bir arayüzdür. Explorer ile yapılabilenler KnowledgeFlow ile araç çubuğu üzerindeki bileşenlerle daha görsel olarak Şekil 2.33’de görüldüğü gibi gerçekleştirilir (Aydın 2011). Explorer ile yapılan işlemlerin sürük-le-bırak şeklinde yapılmasına imkân tanıyan bir ara yüz seçeneğidir.



Şekil 2.33 KnowledgeFlow arayüzü.

Explorer ekranında gerçekleştirilemeyen ekstra bazı özellikler bu ekranda mümkün hale gelmektedir. Explorer’da veriler toplu olarak işlenirken, KnowledgeFlow verileri toplu veya kademeli olarak işleyebilir. Datasource, filters, classifiers, clusterers, evaluation, Visualization sekmelerinin bulunduğu çalışma alanında yapılacak çalışmaya en uygun ara işlemler sürük-le-bırak yöntemiyle Knowledge Flow Layout ekranına bırakılır.

İşlemler tamamlandıktan sonra Arff Loader üzerine sağ tıklanarak çalışma yürütülür (İşler ve Narin 2012).

Simple CLI; WEKA komut satırı ile veri kümeleri üzerinde deneyler gerçekleştirilir. WEKA komutlarının direkt olarak işlenmesini komut ara yüzü ile sağlayan seçenektir.

2.4 Tıp ve Veri Madenciliği

Bu bölüm veri madenciliğinde etik, tıpta veri madenciliğinin gereksinimi, tıbbi verilerin özellikleri, tıbbi verilerle veri ambarının oluşturulması, tıbbi verilerle VM çalışması, metin madenciliği uygulamaları ile veri madenciliğinin sağlık ve tıp alanında kullanım aşamaları ve uygulamaları başlıkları altında incelenmiştir.

2.4.1 Veri Madenciliği ve Etik

Tıbbi veriler ile diğer veriler arasında teoride çok fark olmamasına karşın uygulamada; verilerin biçimlerinin yanı sıra, miktarının çok olması çeşitli ahlaki, hukuki ve tıbbi özel bazı durumlar için önem arz etmektedir.

Tıbbi etik; sağlık kurumlarında her gün binlerce baytlık tıbbi veri üretilmekte, heterojen veri tabanlarında saklanmakta ve sağlık kurumları tarafından belirli ilkeler ve ortak düzenlemeler doğrultusunda sağlık kurumları tarafından yayılmaktadır. Hastaya ait bilginin sahibinin hastalar mı, hekimler mi yoksa sigorta kurumlarını mı olduğu hala tartışılmaktadır.

VM kurumlar için çok önemi faydalar sağlamasına karşın, bilgilerin hukuki ve ahlaki kurallar gözetilerek kullanılmaları gerekmektedir. Problem bireylerden alınan verilerin işlenmesi sonrasında elde edilecek bilginin uygun kullanılıp kullanılmaması noktasında ortaya çıkmaktadır. Öyle ki VM günümüzde güçlü bir rekabet silahı olarak kendine yer edinmekte, aynı zamanda kişilik haklarına karşı da bir tehdit olabilmektedir. Burada, güvenlik; kişilik haklarına saygı, mesleki sırların ifşa edilmemesi ve elde edilen bilgilerin üçüncü taraflara sunulmaması gibi öğelerin göz önünde bulundurulması gerekmektedir (Meriç 2004).

Ülkemizde yaygın bir uygulama olan marketlerin firma kartlarında, kişisel bilgilerin müşterilerden verilen form ile belirli bir promosyondan yararlanmak amacıyla

gönüllülük esasına dayalı olarak birey tarafından verildiği için ahlaki ve hukuki olmayan bir durumdan söz edilemeyeceği gerçeğine karşın, burada sunulan bilgilerin hasta mahremiyeti açısından gizli tutulması, kötüye kullanılmaması ve tanımlayıcı bir etiket olmaması ilişkide bulunulan kuruma aittir. Bilgiyi toplayan kurum, kişisel bilgilerin korunması ve gizliliğinin sağlanması konusunda önlemler almak zorundadır. Aksi durumunda yasalar gerekli şekilde düzenlenerek uygun yaptırımlar gerçekleştirilmelidir (Meriç 2004).

2.4.2 Tıpta Veri Madenciliği Gereksinimi ve Tıbbi Verilerin Özellikleri

Bilgi sistemlerinde ve iletişim teknolojilerindeki hızlı değişim ve gelişimler tıp ve sağlık alanında verinin sayısal ortamda saklanabilmesine ve kolaylıkla erişilebilmesine olanak tanımıştır. Tıptaki değişim ve gelişime paralel olarak insanların yaşam sürelerinin uzaması beraberinde diyabet, astım ve kalp hastalıkları gibi kronik hastalıklarla yaşama zorunluluğu sorunlarını da beraberinde getirmiştir. Sorunların hastalık ayağını oluşturan bu tip kronik hastalıklar hem tıbbi açıdan, hem de hastane kaynak ve maliyetleri açısından doğru bir şekilde yönetilebilmesi için VM yöntemleri kullanılarak gizli ve önemli bilgilerin keşfedilmesi bir ihtiyaç haline gelmektedir (Yıldırım vd. 2008).

Tıp alanında son dönemlerde meydana gelen teknolojik gelişmeler, özellikle biyomedikal elektroniğin ilerlemesi ve verilerin bir arada değerlendirilmesinin öneminin kavranması hem çok fazla sayısal biyomedikal verinin, hem de aynı şekilde çok fazla sayıda kavramsal verinin toplanmasını beraberinde getirmektedir. İnsan algısının bazı detayları kaçırma olasılığına karşın, farkına varılması daha güç detayların da tespit edilip hastalıkların teşhisinin yapılması ve muhtemel erken müdahaleler ya da tedavi değişiklikleri VM yaklaşımı ile daha olası hale gelmiştir. Sadece biyolojik konular değil hastane ve kurum yönetimi, klinik süreçlerin incelenmesinde de VM önemli ölçüde kullanılmaktadır. Bundan dolayı tıpta VM kullanımına biyolojik anlamı dışında hem klinik, hem de yönetsel açıdan bakmak gerekir (Çarklı 2010).

Tıp alanında VM tekniklerinin uygulanması için gerekli altyapı, hastane ve laboratuvarlarda veri tabanı uygulamalarının kullanılmaya başlanmasıyla sağlanmıştır (Şaylan 2013). Sağlık ve tıp alanındaki büyük veri tabanlarından değerli bilgilerin ortaya çıkartılması, hem tıp açısından hem de hizmet kalitesinin artırılması açısından

büyük katkılar sağlaması VM çalışmalarını bu alanda öne çıkarmaktadır.

Tıbbi teşhis sistemlerinde kullanılan standart yazılımlar doktorlar tarafından; veri tabanları üzerindeki sistematik ve tutarlı analizler yapılamaması, verinin dağınık olması, kontrol edilememesi, çok farklı çeşitlerinin kullanılması, örnekleri karşılaştırma, kritik farkları tespit etme gibi nedenlerle eksik görülmüştür. Bu noktada gerekli bilgileri depolayıp işleyen, yöneten, bir uzman bilgi sistemi ile sayısal parametreler ve morfolojilere ilişkin eğrilerin temel analizini sağlayan ayrıca karar için referans yapısı içeren VM devreye girmektedir. Bunun yanı sıra hekime sağladığı kolaylık ve pratisyenlere yardım sebebiyle sağlık alanında kendine yer edinmiştir (Kaya vd. 2003, Albayrak 2008).

Sağlık çalışanlarına zaman ve kaynak tasarrufu sağlamak amacıyla oluşturulacak karar destek mekanizmalarının yüksek başarı oranı ile çalışması insan sağlığı için oldukça önemlidir. Yapılan çalışma sonuçları doğrudan sağlık sektöründe çalışanları ve kamu sağlığını olumlu etkileyip hem uzmanlar için, hem hastane yönetimi için hem de hastaların daha kaliteli hizmet almalarını, ayrıca hastane kaynaklarının verimli kullanılmasını sağlayarak ülke ekonomisine de katkıda bulunacaktır (Tuğ 2005, Şaylan 2013).

Tıbbi alanda kullanılan VM çalışmaları, birçok klinik araştırma gerektiren; hem noninvaziv (hastaya fiziksel bir zarar verme ihtimali olmayan her türlü girişim) olmayan hem de maliyet gerektiren tıbbi araştırmaların yerini ileride bir nebze de olsa doldurarak tıbbi araştırmalar için yeni bakış açıları sağlayacağı düşünülmektedir (Kaya vd. 2003).

Bilişim teknolojilerindeki minimum hata oranı, tıbbi verilerin hayati önem taşıması nedeniyle sıklıkla kullanılmaya başlanmıştır. Bilgisayar uzmanlarına tıbbi veri tabanlarında veri analizi, yönetim biriminde bilgilere hızlı ve etkin bir şekilde ulaşılması, karar destek sistemlerinin oluşturulması büyük kolaylıklar sağlamıştır. Bu hedefler çerçevesinde VM tıp alanında veri içerisinde gizli kalmış değerli ve anlamlı bilgilerin elde edilebilmesinde uygun değer çözüm getirmiştir (Yurtay vd. 2010b).

Tıbbi verilerinde özelliklerinin yapılacak VM çalışmasında göz ardı edilememesi gerekir.

Tıbbi veriler; metin verileri, örüntü verileri ve otomasyon verileri olarak üç gruba ayrılırlar. Raporlar ve tahlil sonuçları metin verilerini oluştururken; filmler, mikroskop, ultrason ve kamera verileri gibi veriler ise örüntü verilerini oluşturmaktadır. Tıbbi verilerin farklılıklarını sıralamak gerekirse (İnt.Kyn.1):

- Veri boyutları çok büyük ve fazla çeşitlilik gösterir (Metin, çeşitli örüntüler)
- Veri dinamikdir
- Eksik veri bulunabilir
- Gürültü içerir
- Geçersiz veri olabilir
- Önemsiz veri olabilir
- Yönetimle ilgili veriler bulunabilir
- Verinin gizlilik özelliği vardır
- Hastane verileri tam olarak disklere yerleştirilmemiş durumdadır.

2.4.3 Tıbbi Verilerle Veri Ambarının Oluşturulması

Elektronik bir devrimin yaşandığı günümüzde; her geçen gün daha fazla insanın tıbbi verileri kayıt altına alınmaktadır. Bununla beraber her tıbbi cihazın veri üreterek bu verilerinde depolanması, doktorun hasta ile görüşmeleri, doktor ve sağlık personelinin gözlem ve girişimleri, laboratuvar verilerinin kayıt altına alınması ile sağlık alanındaki veri miktarı terabayt hatta petabayt'lar seviyesine ulaşmıştır.

Görüldüğü gibi tıp alanında sadece biyolojik anlamda veriler biriktirilmemekte, hastane süreç ve yönetimi ile klinik yaklaşımlar da önemli bir veri yığına sebep olmamaktadır (Çarklı 2010).

Tıbbi veriler üzerinde çalışma yapmak alandaki verilerin iyi tanınması ve yorumlanması uzmanlarla yapılan işbirliği ile mümkündür. Bu nedenle uzman görüşleri ile işlemler arasındaki bağlantının iyi kurulması gerekir. Tıp alanında verilerin belirli bir standartta olmayışı ve var olan standartların uyumsuzlukları nedeniyle farklı kaynaklardan toplanan veriler arasındaki uyumsuzluk göz önüne alınmalıdır ki bu alanda veri ambarı oluşturmak oldukça zor bir işlemdir. Bununla beraber alana özgü tıbbi terimler hem birbirine çok benzemesi hem de oldukça karışık olması veri ambarı oluşumunu negatif

etkileyen faktörler arasındadır (Kaya vd. 2003).

2.4.4 Tıbbi Verilerle Veri Madenciliği Çalışması

1970’lerde hastane bilgi sistemlerinde ilk defa VM teknikleri kullanılmaya başlanmıştır. Bu gelişme sonraki yıllarda yerini uzman sistemlere bırakmasına karşın, tıp alanındaki verilerin hızla değişmesi ve uzmanlar arasındaki görüş farklılıkları nedeniyle çok fazla tanınamamıştır. 1990’lı yıllardan sonra sinir ağları kullanımı maliyet tahmini ve hastaların illerdeki sağlık durumları gibi konular üzerinde araştırma yapmak için kullanılmışlardır (Yıldırım vd. 2008).

Tıp alanında bilginin kullanım şeklindeki değişiklikler, sağlık bakım hizmetleri verenleri etkilemiş, bu hizmetlerde bilgisayar kullanımı, bilginin paylaşımı, ekip yaklaşımı, veri ve bilgi temeli gibi kavramların yaygınlaşması ile bilgisayarlar sağlık bakım ve yönetim hizmetlerinin hemen her alanında sıklıkla kullanılmaya başlanmıştır (Kaya vd. 2003, Albayrak 2008).

Veri tabanlarında bir hastalık ile ilgili hastalara ait demografik bilgiler, biyokimya takipleri, ilaç kullanıp kullanmadıkları gibi birden fazla öznitelik bulunmaktadır. “Bu özniteliklerin hepsi üzerinde çalışma yapılacak hastalığa etki etmekte midir?” sorusunun cevabı için; veri seti üzerinde veri madenciliğinde kullanılan öznitelik seçme algoritmaları kullanılmalıdır (Şaylan 2013). Berry ve Linoff (2004) “insanların deneyimleri ile sonuç çıkarması, geçmişteki uygun örneklerin iyi tanınmasına bağlıdır” derken hastalıklara teşhis koyan doktorun öncelikle benzer vakalara ve mevcut hasta için etkili tedavinin muhtemel benzer hastaların sonuçlarından elde edilen bilgilerle yapılan tedavi olduğu örneğini vermiştir. Klasik hastane bilgi sistemlerinde finansal ve idari konuların ağırlıkta olması teşhis ve tedaviyle ilgili hastaların ayrıntılı bilgilerine ulaşmada oldukça zahmetli bir işlemdir. Öyle ki çoğu zaman hekimin bireysel çabası ile ulaşılabilen hasta teşhis ve tedavi bilgisinde çalışmanın yönlendirilmesi adına alan uzmanından destek alınması gerekebilir. İlgili uzmanın bilgisayar konusunda bilgi sahibi olmaması çalışmayı negatif yönde etkilemektedir (Dinçer 2006).

Tıbbi veri doğası gereği gürültü içeren, tam, kesin ve doğrusal olmayan, bulanık veriler içermektedir. Bu verilerin bilgisayar ortamında oldukça fazla bulunması ve işlenmesinin

zor hale gelmesi, teşhis koyma aşamasında çok fazla hastalık belirtisi ile karşılaşılması, stresli çalışma şartları, hasta ve hasta yakınlarının artan ve uzun ömür beklentisi VM çalışmalarının tıp alanında teşhis, terapi ve prognoz aşamalarında kullanılmasının nedenleri arasında gösterilebilir (Çarklı 2010).

Çalışmalar sadece klinik ve biyolojik olmayıp yönetsel VM çalışmaları da yapılabilmektedir. Sağlık sistemi politikalarının ve yönetsel kararların temeli veri ve veriden elde edilen bilgi olmakla beraber bu politika ve kararların amaçlara uygun ve etkin olabilmesi güncel, doğru ve güvenilir veriye bağlıdır. Sağlık bilgi sistemlerinin amacı çok büyük boyutlardaki sağlık verilerinden faydalı bilgiler çıkarmaktır (Gülten ve Doğan 2008). Bu bilgiler hasta düzeyinde sağlık hizmet kalitesinin artması, kurumun daha iyi yönetilmesi, kaynakların etkin kullanımı ve sağlık politikalarının oluşturulmasında kullanılmaktadır. Sağlık verileri; sağlık kamu ve kuruluşları, sigorta şirketleri gibi birçok kurum ve kuruluş tarafından toplanmaktadır. Hastane yönetimi; hasta yoğunluğu, tedavi sürelerinin önceden tahmini, nöbet ve mesai sürelerinin düzenlenmesi, ilaç, yatak, sarf malzemesi vb. ihtiyaçların önceden belirlenmesi gibi yönetsel ihtiyaçların şekillendirilmesinde önceki tecrübe ve mevcut veri birikimlerine dayanarak yapılacak VM işlemleri yönetsel anlamda tıp alanına önemli katkılar sağlayacaktır (Çarklı 2010).

Bu nedenlerin yanı sıra VM sonucunda tıpta elde edilecekler arasında; tıbbi veri enformatiğindeki fazlaca veri yüklemesini anlamlandırmak ve bunu bir sorun olmaktan çıkarmak, arttırılmış sağlık hizmet kalitesi, düşürülmüş çalışma masrafları ve tıbbi verilerin daha iyi değerlendirilmesi, ağ üzerinden paylaşılan verilerin daha kolay bir şekilde model geliştirilmesini ve açıklama yapma becerisinin artmasını sağlayabilme gibi durumları sıralamak mümkündür (Çarklı 2010).

2.4.5 Metin Madenciliği Uygulamaları

Yapısal olmayan metinlerden VM teknikleri ile dokümanlar arasındaki gizli örüntü ve ilişkileri araştıran BK yapılmasını sağlayan VM yöntemi olan metin madenciliği; doğal dil işleme, bilişsel bilimler ve makine öğrenmesi gibi bilimlerle ortak çalışarak, genellikle birbiri ile ilişkili belgeleri, aynı konuda yazılmış belgeleri ve kavramlar arası ilişkileri bulmak için çok büyük ve anlamsız veri yığınları arasından anlamlı ilişkiler

elde etmekte kullanılırlar (Yıldırım vd. 2008, Arslan 2008).

“Arşivimde (veya internet üzerinde) bu dokümana benzer hangi dokümanlar var?” amaç dokümanlar arasında ayrıca elle bir tasnif gerekmeden benzerlik hesaplayabilmektedir (Text Mining). Bu genelde otomatik olarak çıkarılan anahtar sözcüklerin tekrar sayısı sayesinde yapılır (Alpaydın 2000).

Metin madenciliği yöntemleri (Yıldırım 2008):

- Anahtar kelime ilişkisi
- Otomatik doküman sınıflandırma
- Benzerlik bulma
 - Ortak yazara göre sınıflandırma
 - Ortak konuya göre sınıflandırma
- Bağlantı analizi (varlıklar arası sıradışı ilişkileri analiz etme)
- Sıra analizi (bir sonraki olayı tahmin etme)
- Anormal değerlerin bulunması (aykırı ve anormal değerleri bulma)
- Hypertext analizi
 - Bağlantılı objeler veya dokümanlar arasındaki gizli örüntüleri bulmak

İlaçlar tıbbın önemli araştırmaları arasındadır. İlaçlar piyasaya sürülmeden önce faydalar ve riskler göz önüne alınarak onaydan geçer. İlaçların önceden tanımlanmamış yan etkileri web üzerinden tıbbi yayınlar analiz edilerek metin madenciliği çalışmaları yapılmaktadır.

Sadece ilaç bilgileri değil hastaların tıbbi durumları, tanı, tedavi bilgileri, klinik dokümanlar, uygulanan işlemlere ait faturalar, iş akışını belgeleyen raporlar, tıp alanında bilimsel makaleler, sağlık alanında yapılan araştırma ve yenilikler değerli bilgi kaynakları olup yapısal olmayan metinsel kaynaklardır. Bu yapılar üzerinde BK yapmak için metin madenciliği yöntemleri kullanılmaktadır. Metin madenciliği doğal dil işleme, tıp bilişimi ve istatistik gibi alanlarla ortak çalışan bir araştırma alanıdır. Medline gibi tıbbi alanda bilimsel yayınların saklandığı veri tabanlarında metin yapısındaki veriler analiz edilip BK yapılarak bilgi yönetimi sağlamak temel amaçlardandır. Yapılan çalışmalar neticesinde tedavi, tanı ve bunlarla ilgili yeni kavram ve yaklaşımlar

arasındaki gizli ilişkiler sağlık kurumlarının başarısını artırmanın yanında araştırmacılara da büyük destek sağlamaktadırlar. Avrupa biyoenformatik enstitüsü metin madenciliği araştırma grubu biyomedikal alandaki bilimsel makalelerden tıbbi ontoloji, anlamsal ağlar, doğal dil işleme ve birleştirilmiş tıbbi dil sistemi tekniklerini kullanarak metin madenciliği BK yapan bir araştırma grubudur (Yıldırım vd. 2008).

Amasyalı ve meslektaşları (2012) çeşitli türdeki 6 adet Türkçe sınıflandırma veri kümesi üzerinde 17 adet özellik ile bir metnin hangi özelliklerine bakılarak sınıflandırılabilir ya da sınıflandırma problemi türün; yazarın cinsiyeti, ruh hali, metin konusu, olumlu ifadeler içerip içermemesi nasıl etkiler sorularına cevap olarak incelemişler ve cümle, kelime, ek sayıları, ngramlar, kelime grupları ve saklı anlam indeks özellik gruplarından; ngramların genel olarak diğer temsil yöntemlerinden daha başarılı sonuçlar ürettiği görülmüştür.

İnternette bulunan haber ve forum sayfalarında yer alan metinlerden oluşturulan 2024 metin dosyası içinde 436 707 kelime sözlük Mysql veri tabanında oluşturulan tablolarda saklanarak eğitim kümesi olarak kullanılan çalışmada; İstatistiğe dayalı yöntemler ile makine öğrenmesi yöntemlerinin doğru kelime tercihi üzerindeki başarısı incelenmiştir. Bir metnin tüm parça olarak incelenmesinin, metin hakkında cümle bazlı yöntemle göre daha fazla bilgi verdiği, ayrıca makine öğrenmesi yöntemlerinin istatistiksel bazlı yapılan çalışmalara göre daha iyi sonuçlar sağladığı deneylerle gösterilmiş ve çalışmada Türkçe dili için tanımlanan sözcük anlamı belirleme problemine çözüm getirilmiştir (Okur 2013).

2.4.6 Veri Madenciliğinin Sağlık ve Tıp Alanında Kullanım Aşamaları ve Uygulamalar

VM ile ilgili yapılan çalışmaların tıp ve sağlık ayağını birçok uygulama ve çalışma oluşturmaktadır;

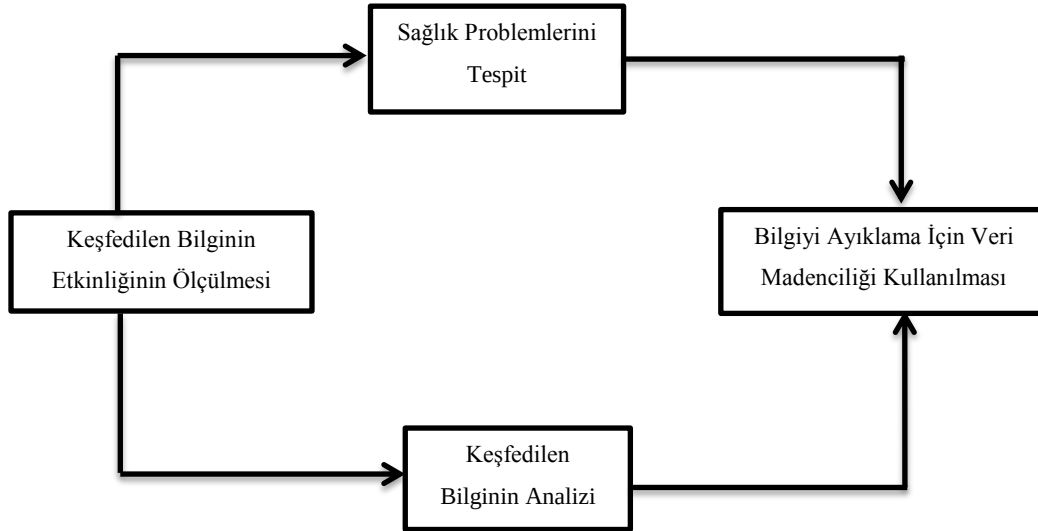
- İlaç ve eczacılık alanında yapılan VM uygulamaları,
- Hastalık tipine göre VM çalışmaları,
- Kullanılan VM tekniklerine göre yapılan çalışmalar,
- Hastane bilgi sistemleri, yönetim ve suiistimal tespiti üzerine yapılan VM

uygulamaları,

- Genetik alanında VM uygulamaları ve çalışmaları,
- Klinik/medikal VM çalışmaları bunlardan bazılarıdır.

Hekim tecrübesi ile elde edilebilecek gizli örüntülerin çok faktörlü hastalıklarda önemli faktör ilişkilerinin önceden tahmini ile hastalıklar ve semptomları arasında gizli kalmış örüntülerin ortaya çıkarılmasının sağlık alanında hayati öneme sahip olması, veri madenciliğini bu alanda birçok çalışma yapmaya teşvik etmiştir.

Sağlık problemlerinin tespiti, bilgiyi ayıklamak için VM uygulaması, keşfedilen bilginin analizi ve keşfedilen bu bilginin etkinliğinin ölçülmesi şeklinde sağlıkta VM döngüsü Şekil 2.34’de görülmektedir.



Şekil 2.34 Veri madenciliği döngüsü (Kaur and Wasar 2006).

Sağlık hizmetlerinde kullanılan VM çalışmaları oldukça fazladır. Rochester kanser merkezi, KnowledgeSEEKER adlı karar ağacı tekniğini araştırmalarında kullanmaktadır. Vysis, ilaç sanayisi için YSA algoritması yardımıyla protein analizini yapmaktadır. “Information Discovery” adlı ürünü California hastanesi VM çalışmaları kullanmaktadır. Hastanede çalışan bir doktor; hiçbir fiziksel test yapmaksızın bu program sayesinde teşhis koyabildiğini açıklamıştır. MapInfo şirketi bölgeye daha iyi hizmet sunmak adına hastaların yoğun bulunduğu bölgeleri haritalama teknolojisini kullanarak işaretlemektedir (Doğan 2007).

Ulusal ve uluslar arası literatür incelendiğinde sağlık alanında sayısız ve değerli bir çok araştırma bulunmaktadır.

Doğan (2007) biyokimya test sonuçlarını kullanarak uzman ve pratisyen doktorlar için; Miyokard infarktüsü, Hiperlipidemi, Demir eksikliği anemisi ve Hipertiroidi-Hipotiroidi hastalıklarının teşhisinde kullanılacak bir karar destek sistemi tasarımı üzerine delphi programlama dilinde bir çalışma yapmıştır. Geliştirilen paket programla belirlenen tip hastalık için hekimlerin tanıda bulundukları biyokimya verilerinden seçilen parametreler temel alınarak lipidler ve enzimlerle yapılan değerlendirmede; örüntü tanıma süreci esas alınmış olup, VM tekniklerinden karar ağaçları yöntemi kullanmıştır. Geliştirilen karar destek sisteminin bu hastalıklar için ulaştığı sonucun doktorun verdiği kararlar % 100 doğrulukla örtüştüğü görülmüştür.

VM yöntemleri ile yapılan bir başka tıbbi karar destek sisteminde; metastatik kolorektal kanser hastalığı ve diyabet hastalığı veri setleri üzerinde sınıflandırma tahmin modeli ve hibrit tahmin olmak üzere iki ayrı model geliştirmiştir. Metastatik kolorektal kanser veri setinde sınıflandırma modelinde en başarılı algoritma destek vektör makineleri iken, hibrit tahmin modelinde ise %100 doğruluk oranı ile en başarılı algoritmanın YSA, buna karşılık diyabet veri seti için çok katmanlı algılayıcı algoritması en yüksek başarı yüzdesi ile en iyi algoritma, karar ağaçları ise başarı yüzdesi en düşük algoritma olarak tespit edilmiştir (Pala 2013).

Altıntaş (2010) oniki ay ve daha uzun süre hemodiyaliz tedavisi görmüş hastaların verileri ile yaptığı çalışmasında sınıflama, kümeleme ve karar ağacı algoritmaları ile elde edilen sonuçlar uzman yardımı ile incelenip risk seviyelerine göre risk faktörlerinin etkileşimlerinden desenlere ulaşılmaya çalışılmıştır.

Şentürk (2011) kanser hastalığının erken tanısına yönelik yaptığı çalışmasında hastane bilgi sisteminden alınan hasta bilgileri kullanılarak, seçilen tahmin modeline göre henüz tanısı koyulmamış hastalara, belirli güvenilirlik oranlarına göre tanı koyulmuştur. Çalışma sonucunda; k-en yakın komşu algoritması doğruluk oranı en yüksek model olurken, modellemede en yanlış sonuçlar karar ağacı yöntemleri ile elde edilmiştir.

Turgut (2012); Süleyman Demirel Üniversitesi Tıp Fakültesi Nöroloji Ana Bilim

Dalı'na demans ayırıcı tanısı için ayrıntılı klinik muayenesi ve nöropsikolojik değerlendirmesi yapılmış 101 hastaya ait verileri kullanarak 15 öznitelik ile WEKA yazılımında VM çalışması yapmıştır. Hekimin yükünün azaltılarak yardımcı olunması ve fikir vermesi amacıyla test ve klinik tanı arasındaki işleyiş, veri-bilgi arasındaki ilişkiye dönüştürülüp VM teknikleri kullanılarak yapılan analiz çalışması sonucunda; sınıflandırma algoritmalarından en başarılı olan meta.LogitBoost algoritması %100 doğruluk değeri ile, web 2.0 ile php dilinde yazılan bir arayüz ile doktorun hastaya uygulamış olduğu testlerin sonuç değerleri girilerek ve demans teşhisi için öngörü üretme işlemi yapılmıştır.

Dermatoloji veri seti üzerinde yapılan bir çalışmada sinirsel bulanık çıkarım sistemi ile yapılan sınıflandırmada %95,5 oranında doğruluk değeri elde edilirken yapay sinir ağı tabanlı CNN yöntemi ile elde edilen sınıflandırmanın %97,77 oranına sahip olduğu görülmüştür (Ubeyli ve Güler 2005, Ubeyli 2009).

Yine dermatoloji alanında; erythemato-squamous hastalıkların teşhisi için tahmin yapabilen, bu hastalıkların patolojik ve klinik özellikleri arasında analiz yapmak amacıyla ve hasta kayıtlarının nitelikleri arasındaki ilişkileri analiz etme imkânı sunabilen k-en yakın komşu (kNN) ve k-means algoritmaları kullanılarak hekimlerin kullanabileceği, yardımcı bir karar verme sistemi tasarlanmıştır. Yapılan çalışmanın sonucunda; 366 kayıttan oluşan dermatoloji veri seti üzerinde normalizasyon ve veri temizleme gibi ön işleme işlemleri sayesinde kalan 358 kayıt için kNN algoritmasının performans yüksek çıktığı görülmüştür (Çataloluk 2012).

Aslandoğan vd. (2004) tarafından göğüs kanseri ve cilt yaralarını sınıflandırmak için kNN, Bayesian, karar ağacı ve Dempster-Shafer teorisi yöntemleri karşılaştırılarak yapılan çalışmada; sırayla elde edilen doğruluk oranları: %42,5, %95 ve %94,9 şeklindedir. Farklı veri tabanlarında Dempster-Shafer birleştirme kuralı teorisinin doğru sınıflandırma yaptığı sonucuna varılmıştır.

Yurtay vd. (2010a) tarafından Sakarya Üniversitesi akademik ve idari personelinin rastgele seçilen 170 kişiye olası migren teşhisine yönelik 24 soruluk anket verilerinden yararlanılarak yapılan çalışmada; sık geçen öğelerin keşfedilmesinde Apriori algoritmasını uygulayabilmek için SQL Server ile veriler ön işlemeden geçirilerek

analiz edilmiş ve .Net platformunda web tabanlı bir yazılım geliştirmiştir. Geliştirilen bu yazılım ile Apriori algoritmasının işleyişi adım adım takip edilerek migren ve migrene bağlı gelişen baş ağrısı rahatsızlıklarına ilişkin dokuz adet birliktelik kuralı elde edilmiştir. Gerçekleştirilen anket uygulaması hastalık teşhisine yönelik olmayıp birliktelik kurallarının bulunmasına yönelik bir çalışma yapılmıştır.

Bir anket çalışması da; dış ameliyatları sonrasında akut ağrı çeken 438 hastanın cevapları uygun formata dönüştürülerek K-Means algoritması ile kümelenecek hastaların sosyal ve psikolojik durumlarının dış ameliyatları sonrası görülen akut ağrı sorununa etkisi araştırılmış, ağrılarının analizi sonucunda bayanlarda ağrıdan dolayı yakınmaların fazla olduğu ve üzüntülü, endişeli, depresyondaki hastalarda ağrı şiddetinin yüksek olduğu görülmüştür (Vickers *et al.* 2006).

Yine Yurtay vd. (2010b) 72 öğrencinin aurasız migrene sahip olup olmadıklarını tespit etmek için 7 adet özellik ve bir adet sınıflama bilgisi ile karar ağacı algoritmalarından olan Gini algoritması ile yaptıkları çalışmada; 17 adet kural belirlenmiş ve bulantı niteliğinin hastalığı belirlemede daha etkin olduğu gözlemlenmiştir.

Ünal vd. (2010) sırt ağrısı hasta verileri üzerinde yaptıkları VM çalışmasında; Selçuk Üniversitesi beyin cerrahisi, Fizik tedavi ve rehabilitasyon kliniklerinde bulunan 350 hasta verisi ve 8 öznitelik ile karar ağaçları, K-Means algoritması ve birliktelik kuralları VM algoritmaları ile incelenmiştir. Çalışmada elde edilen sonuçlarda; yaş, cinsiyet, sırt ağrısı süresi, bacak ağrısı, uyuşma-karıncaşma, gece ağrısı, sabah krampları gibi özelliklerin hastalık teşhisi üzerinde doğrudan etkisi olduğu görülmüştür.

Sağlıkta psikiyatri alanında da çalışmalar yapılmıştır.

K-Means kümeleme yöntemi ile yapılan bir çalışmada; psikiyatri alanında antisosyal kişilik bozukluğu gösteren adli suçluların bilgileri ile yapılan kümeleme analizidir. Analiz sonucunda güçlü suç eğilimi gösteren ve psikopatik kişilikli gruplar olmak üzere iki farklı tipte anti sosyal kişilik bozukluğu tespit edilmiştir. İki suçlu grubun profilleri arasındaki farklılıklar kümelerin analiziyle ortaya çıkarılmıştır (Morana *et al.* 2006).

İntihara teşebbüs eden ve ağır şekilde yaralanan 121 hastadan toplanan bilgilerle K-Means algoritması ile farklı profil klinik grupların görülmesi amacıyla yapılan bir başka

çalışmada; ilaç içerek kendini zehirleyen çoğunluğu kadınlardan oluşan 43 tane intihara çok istekli ve intihar nedeni az olan, orta seviyede intihar isteği olan böcek ilacı içen çoğunluğu erkek olan 53 hasta ve kendine vahşi şekilde zarar veren yüksek düzeyde intihar isteği olan çoğunluğunu erkeklerin oluşturduğu 17 kişi olmak üzere üç küme ortaya çıkmıştı görülmüştür (Rapeli and Botega 2005).

Sol elin işaret parmağının yüzük parmağına oranına (2D:4D) göre bireyin kişilik durumlarının sınıflandırılması ve kişilerin bazı hastalıklara olan yatkınlıklarına ilişkin bilgilerin tespiti amacıyla yapılan çalışmada 108 kişiden alınan sol el parmak görüntüleri ile WEKA yazılımı ile J48, NBtree, LANDTREE sınıflandırma algoritmaları ile VM çalışması yapılmıştır. Çalışma sonucunda en başarılı sınıflandırma sırasıyla %100 oranla LANDTree algoritması, %97,22 oranla NBTree ve %94,44 oranla J48 algoritması olarak gözlemlenmiştir (Daş ve Varol 2013).

Esen (2009) Konya Numune Hastanesi'nden elde edilen çeşitli hastalıklara sahip kişilere ilişkin veriler VM tekniklerinden lojistik regresyon ile analiz edilerek tedavi planlama ve koruyucu tedaviler açısından önemli olan hastalık olasılık tahmini yapılmasının amaçlandığı çalışmasında; erkeklerde astım hastalığı görülme olasılığının kadınlarda görülme olasılığına göre yaklaşık 4,6 kat daha fazla olduğu, mevsimsel olarak hava kalitesinin düştüğü özellikle kış ve sonbahar dönemlerinde astım olasılığının arttığı, 50 yaş ve üstü kişilerin astım hastalığına yakalanma olasılığının bu yaş gurubu altındakilere oranla yaklaşık 6 kat daha fazla olduğu, SSK sağlık güvencesi olan hastaların astım olasılığının olmayanlara göre daha fazla olduğu tespit etmiştir. Kardiyovasküler hastalıklara ilişkin yapılan analizde; kırsalda oturanların kentlerde oturanlara göre kardiyovasküler hastalığa yakalanma oranının fazla olduğu, sonbahar aylarında arttığı ilkbahar aylarında yarı yarıya düştüğü ayrıca kardiyovasküler hastalık olma olasılığının erkeklerde daha fazla olduğu, 50 yaş ve üzerinde olasılığın %50 oranında arttığı tespit edilen diğer sonuçlardır.

Turhan vd. (2011) kadın genital bölge kanser teşhislerinin doğruluğunu ve tutarlılığını test eden birliktelik kuralları ve destek vektör makineleri teknikleri kullanılarak bir model geliştirmişlerdir. Karadeniz Teknik Üniversitesi Farabi Hastanesi laboratuvar verileri ile yapılan çalışmada; 227 228 farklı klinik farklı laboratuvar test işlemleri biyoistatistik ve tıp bilşimi anabilim dalınca hazırlanmış olan DAY isimli dönüşüm

aracı yazılımı aracılığıyla kaydedilmiştir. Çıkarılan birliktelik kurallarına göre; kadın genital bölge kanser tanılarının tutarlılığı destek vektör makineleri ile sınanmıştır.

Güllüoğlu (2011)'nin sağlam kişilerin ileride yakalanma ihtimali olan göğüs, akciğer ve bağırsak kanser tiplerini belirleyerek sağlıklı veya riskli teşhisi koymak amacıyla yaptığı öneri çalışmasında; onkoloji servis verilerinin YSA yöntemi ile sınıflandırma yapılması, böylelikle sağlıkta uzman kişilere sağlık sektöründe veri madenciliğinin kullanımı ile ilgili çalışmanın nasıl olması gerektiği konusunda bilgi sunarak karar verme sürecine yeni bir bakış açısı kazandırmak hedeflenmiştir.

Dinçer ve Duru (2006) tarafından K-Means algoritması kullanarak Borland Delphi'de geliştirilen yazılım ile Kocaeli Tıp Fakültesi Kulak Burun ve Boğaz Bölümü'nden alınan 400 gırtlak kanseri ameliyat veri kaydının analizi amacıyla yapılan çalışmada; K-Means algoritması ile veriler kümelenecek içindeki yoğunlaşmalar belirlenmiş ve grafik aracılığıyla görüntülenmiştir. Yapılan sınamalar sonucu 9 kümeden fazlası için yapılan çalışmadan istenen performansın elde edilemediği görülmüştür.

Ayrıca karar destek sistemleri de sağlık alanında kendine geniş bir yer edinmiştir. Tıbbi teşhislerde bir karar destek mekanizmasının oluşturulması kaliteyi artırırken aynı zamanda maliyeti de düşürmektedir (Şaylan 2013).

Özdemir (2011) yaptığı çalışmada; panik değer aralığındaki test sonuçlarının tespiti, ilgili kişi veya kişilerin uyarılması ve gerekli önlemlerin alınması için ve laboratuvar bilgi yönetim sistemine bütünleşmiş bir parça olarak çalışan klinik karar destek sistemi geliştirmiştir.

Endokrinolojik hastalıklar için oluşturulan veri tabanı ile tiroid hastalığının teşhisini yapan karar destek sistemi tasarlanmıştır (Yılmaz ve Pamuk 2010).

Bir başka çalışmada genetik algoritma, lojistik regresyon ve ki-kare testi yöntemleri kullanılarak asemptomatik karotid arter stenozu risk faktörlerinin anlayıp tanımlayabilmek, inmeyle ilgili komplikasyonlar ve erken tanı çalışmalarına katkı sağlamak amacıyla vakaların tespiti için kural ve ilişkiler VM yöntemleri ile araştırılmış ve karşılaştırılmıştır. Karşılaştırma sonucunda iki yöntem birbirine yakın ve ki-kare testlerinden daha iyi sonuç vererek kronik arter hastalığı veya hipertansiyon varlığı

asemptomatik karotid arter stenozu hastalarının ön tanısında önemli belirteçler olarak ortaya çıkmıştır. Lojistik regresyonun ROC altında kalan değerinin yüksek olmasına karşın kullanım kolaylığı nedeniyle genetik algoritmaların daha iyi bir alternatif olduğu vurgulanmıştır (Bozkurt vd. 2011).

Genetik algoritmalar yönteminin biyomedikal veriler üzerindeki uygulama alanı olarak kan biyokimya parametreleri ile hiperlidemi teşhisinde hekime yardımcı olacak bir karar destek sisteminin oluşturulduğu bir çalışmada; sisteme giriş olarak hiperlidemi hastalığı için temel belirleyicilerden total kolestrol, LDL, Triglisetit, HDL ve LDL testleri kullanılarak yapılan değerlendirme sonuçlarının hekimin verdiği kararlarla tamamen örtüştüğü görülmüştür (Gülten ve Doğan 2008).

TC Sağlık Bakanlığı Kanseri Savaş Daire Başkanlığı'ndan 2001, 2002 ve 2003 yıllarına ait yaklaşık 80 bin kanser hastasına ait verileri ile genetik algoritma kullanılarak anlamlı bilgi kümelerinin bulunmasının amaçlandığı bir çalışmada başlangıç popülasyonu seçimi yapay bağışıklık sistem modeli olan SAND yaklaşımı ile yapılmış ve iki uygunluk fonksiyonu kullanılarak hassas sonuçlar üretmek için bulanık mantıktan yararlanılan çalışmada bulunan veri grupları kanser hastalığı sebeplerine ışık tutmaktadır (Tuğ 2005).

Erdem ve Özdağoglu (2008)'nin Ege Bölgesi'ndeki bir araştırma ve uygulama hastanesinin 214 bin acil servis hasta verisi kullanılarak yaptıkları birliktelik analizi çalışmalarında; acil servis bölümlerinin yeniden yapılandırılması çalışmalarına katkıda bulunmak amacıyla, acil servise bölgesel başvuru nedenleri ve hasta profilleri incelenmiştir. Çalışma sonucunda; kategorik hale getirilen hastaların demografik profilleri, tanı tanımları ve acil serviste kalma sürelerini içeren veriler ile anlamlı 65 kural elde edilmiştir. Elde edilen kuralların genellikle geliş zamanı, yaş grupları, tanı ve cinsiyet odaklı oldukları görülmüştür. Acil servise başvuran hastaların mortalitesine etki eden demografik ve klinik faktörleri ortaya çıkarmaya yönelik yapılan bir başka çalışmada; travma dışı nedenlere 1 aylık 269 hastaya ait veriler genetik algoritma ve lojistik regresyon analizi ile değerlendirilmiştir. Genetik algoritma yönteminin daha esnek ve klinisyen kararlarında daha kullanışlı olduğu söylenmiştir (Özel vd. 2011).

Yıldız vd. (2012) yaptıkları çalışmada; gen ifade verilerinden genetik algoritma ve veri

füzyonu tabanlı gen seçimi gerçekleştirilmiştir. 97 meme kanseri hastasına ait mikro dizi verileri filtreleme ile indirgenerek veri boyutu azaltılmış ve genetik algoritma için arama uzayı oluşturularak iki aşamada analiz edilmiştir. Çalışma sonucunda; on genin meme kanseri hastalarında %94,65 doğruluk oranında sınıflandırma başarısı sağladığı görülmüştür.

Danacı vd. (2010)'nin Xcyt programı ile elde edilen doku örüntülerinin WEKA programıyla meme kanser hücrelerinin tahmin ve teşhisine yönelik yaptığı çalışmada; 569 hastaya ait verilerden, 450 hücre örneği eğitim verisi, 119 hücre test örneği olarak kullanılmış ve C4.5 ağaç yapısı ile yapılan sınıflandırmada % 97,44 doğruluk oranı ile WRADIUS değerinin en yüksek değere sahip olduğu görülmüştür.

Meme kanseri ile tedavi yöntemleri VM kullanılarak belirlenen bir çalışmada Demirel (2008) tarafından yapılmıştır. Ankara Onkoloji Hastanesi'nden beş yıllık, meme kanseri sebebiyle ameliyat geçirmiş 462 hasta ve 15 parametreden oluşan veri seti hormonoterapi, tamoxifen, radyoterapi ve kemoterapi çıkışları için WEKA programı ile sınıflandırma algoritmaları işlenmiş ve doğru tahmin etme performansı yüksek olan algoritmalar ile java Netbeans arabirimi kullanılarak "Tedavi Asistanı" isimli onkoloğa yardımcı bir program geliştirilmiştir. Çalışma sonucunda; çıkış alanlarından hormonoterapi alanı için IB1, tamoxifen ve radyoterapi alanı için MultilayerPerceptron ve kemoterapi alanı için DecisionTable algoritmasının en yüksek doğrulukta tahminde bulunduğu tespit edilmiş, bu algoritmalar modellemede kullanılmış ve yapılan testlerde programın yüksek doğrulukta çalıştığı saptanmıştır.

Uzun vd. (2013) George Washington Üniversitesi'nden alınan down sendromu veri kümesi kullanarak; 8216 gebe üzerinde 31 öznelik ile invaziv (girişimsel) testleri azaltmak down sendromu yüksek riskli sınıfı belirlemek amacıyla yaptıkları çalışmada olasılıksal sınıflandırıcıların en az diğer VM algoritmaları kadar başarımları oranı gösterdiklerini saptamışlardır.

Baloğlu (2006)'nın E.coli bakterisinin DNA'sından alınan 300 adet promoter sırası veri kümesi ile yukarıdan aşağı VM ve genetik algoritma tabanlı hibrit bir çözüm barındıran Java programlama dili kullanılarak geliştirilen Top-Down-GA isimli yazılımla; biyolojik sıralardaki tekrarlı örüntülerin ve potansiyel motiflerin VM yöntemiyle

ıkarılması iřlemi yapılırken mevcut yaklařımlardan zaman aısından daha performanslı alıřan bir yntem geliřtirilmiřtir. Minimum destek deęerinin deęiřmesiyle elde edilen uygunluk deęeri ve performans verileri farklı motif uzunlukları zerinde incelenmiř; srenin rnt uzunluęuna baęlı olarak ta nerilen yntemin artan uzunluklarda da performanslı olduęu grlmřtir.

Kkver vd. (2014) hipertansiyona etki eden faktrlerin VM yntemleriyle inceleyerek hipertansiyon hasta adaylarının hipertansiyon olup olmadıęını tahmin edecek bir teřhis sistemi geliřtirmiřtir. Yapılan alıřmada 150 kiřinin verileri  farklı sınıflama algoritması ile analiz edilmiř ve sınıflandırmada C4.5 ve Naive Bayes'in %91,667 doęruluk yzdesi ile en yksek bařarı elde eden algoritmalar olduęu tespit edilmiřtir. Doęruluk yzdeleri aynı olmasına karřın; gerekte hasta olan bir kiřiye saęlıklı teřhisi koymak tıbben daha riskli bir tanımlama olduęundan C4.5 algoritması Naive Bayes algoritmasından daha bařarılı kabul edilmiřtir.

Hipertansiyon hastalıęı riskini tahmin etmek amacıyla yapılan bir bařka alıřmada; 694 veri zerinde retrospektif bir analiz yapılarak sınıflandırma tekniklerinden olan karar aęaları, istatistiksek algoritmalar ve YSA'nın performansları karřılařtırılmıřtır. Yař, cinsiyet, ailede olup olmaması, sigara kullanımı, lipoprotein, trigliserid, rkasit, total kolesterol ve vcut-kitle indeksi belirleyici deęiřkenleri ile ok katmanlı algılayıcı (MLP) dięer yntemlere gre hipertansiyon teřhisinde daha iyi bir performans gstermiřtir (Tre vd. 2005).

10 769 hastanın kan deęerlerinden oluřan laboratuvar verilerinin kmeleme algoritması ile demir eksiklięi anemisi hastalıęı iin gerekleřtirilen alıřmada; belirleyici faktrlerden olan serum demiri, serum demir baęlama kapasitesi ve ferritin testleri kullanılarak oluřturulan karar aęacı yapısında ıkıř anemi(+) ve anemi (-) řeklinde deęerlendirme yapılmıř, karar destek sistemi sonularının hekim kararları ile %100 rtřtę grlmřtir. 96 hastanın verisi ile RapidMiner 5.3 programı AGglomerative NEsting algoritması ile yapılan alıřma sonucunda hastalar HGB, HCT ve yař deęerlerine gre gruplandırılarak hastalık seviyeleri belirlenmiřtir (Yurtay vd. 2013).

İřler ve Narin (2012)'nin 54 kiřiden elde edilen kalp hızı deęiřkenlięi analiz verileri 59 nitelik ile konjestif kalp yetmezlięi olan hastaların kontrol grubundan ayırt edilmesi iin

WEKA yazılımında k-ortalama kümeleme algoritması başarımının incelendiği çalışmada; k=4 küme değeri için en yüksek başarımın %98,79 olduğu görülmüştür.

Tarimer ve Elmas (2015) hasta verileri üzerinde kalp krizi riskini tespit amacıyla yaptıkları VM çalışmasında, kalp krizi geçiren hastaların kan değerlerinde meydana gelen ortak değişimleri tespit edip kalp krizini tetikleyen faktörler apriori ve gri algoritmaları kullanılarak hasta grubunda test edilmeye çalışılmıştır. Analiz başarımı ölçülen algoritmalarından gri algoritması daha az kural üretmesine karşın apriori algoritması ile aynı başarımı gösterdiği görülmüştür.

Akyol vd. (2012) kronik kalp hastalığı olma olasılığı lojistik regresyon modeli ile 6 aylık kardiyoloji ve diğer servislerde bulunan 169 hastaya ait biyokimya ve hemogram laboratuvar test sonuçlarının analizi ile sınıflandırma işlemi yapılmış ve hastalığın olma durumu incelenmiştir. Çalışma sonucunda bağımlı değişkenken için Kolesterol, Trigliserid, HDLkolesterol, LDLkolesterol ve yaş değişkenlerinin anlamlı olduğu bulunmuştur.

Kalp ritim bozukluğu olan hastaların tedavi süreçlerini desteklemek amaçlı makine öğrenmesine dayalı geliştirilen bir sistemde ise; hastanın bazı metabolik parametreleri alınarak uzak bir sunucudaki veri ambarına aktarılmış sonrasında bu veriler makine öğrenmesi sınıflandırıcısı tarafından eğitim verileri olarak kullanılmıştır. Geliştirilen sistemde makine öğrenme algoritması olarak düşük bias'a sahip nonlineer bir fonksiyon olduğu için k-en yakın komşuluk (kNN) algoritması kullanılmıştır (Aydın 2011).

İlaç çalışmalarına bu alanda yer verildiği görülmektedir.

Delibaş (2010) hastanelerde yazılmış 96 770 reçete bilgisinin 253 farklı endikasyona göre, Tanagra VM programı ve delphi uygulaması geliştirilerek yapılan analiz işleminde vaka sayıları, su analiz raporları ve sıcaklık verileri de dikkate alınarak yapılan çalışmada; yıl bazında, aylar bazında, en sıcak ve en soğuk aylarda, kış mevsimi, ilkbahar mevsimi, yaz mevsimi ve sonbahar mevsimi reçeteli ilaç satışlarında elde edilen birliktelik kurallarında yılın tamamını kapsayan veri kümesinde en yüksek destek değerine sahip olan kuralların, mevsimler veya aylar bazında incelendiğinde yılın tamamından yüksek olmadığı bazı aylarda minimum destek değerinin altında kaldığı

gözlemlenmiştir.

K-Means algoritması yardımıyla ilaçları kategorize ederek ilaçların olumsuz etkilerini araştıran Evans vd. (2005) ilaçların sınıflandırılması yoluyla bazı risk faktörlerinin tüm olumsuz etkilerle ve ilaçların tedavi sınıfları ile tutarlı olmasına karşın bazılarının ait olduğu sınıfa özel faktörleri olduğu izlenmiş ve yüksek risk etkenlerinin hastanın kilo, yaş, cinsiyet gibi karakteristik özellikleri ile ilacın kullanım şekli ve dozajına bağlı olarak izlenmesi gerektiği sonucunu elde etmişlerdir.

Genetik alanında yapılan çalışmalara örnek vermek gerekirse; yüksek boyutlu genetik verilerde önemli genetik faktörlerin bulunması problemi ve sınıflama için gerekli ve destekleyici yöntemlerine sahip boyut indirgeme, genelleştirme yöntemi, sınıflama ve kümeleme analiz özelliklerine sahip Java analiz aracı R yazılımı ile geliştirilmiştir. Geliştirilen “Gene:3E” isimli yazılım ile genetik verilerde sınıflama problemleri için hangi yöntem ya da yöntem grubunun iyi performans gösterdiği konusunda performans ölçütlerini aynı anda vererek kullanıcıların daha bilinçli seçim yapmasına olanak sağlanması açısından büyük ve önemli bir eksikliği gidermektedir. Çalışma da; genetik araştırmalarda çapraz geçerlilik yerine bootstrap yönteminin kullanılmasının farklı genetik faktörlerin bulunmasında daha etkili olduğu görülmüştür. Ayrıca veri analizi oldukça hızlıdır (Coşgun vd. 2011).

Xu vd. (2003) Excavator isimli gen ifade profillerinin kümelenmesi için bir yazılım geliştirmişlerdir. Biyolojik süreçte gen ifade verisinin kümelenmesi, genler arası fonksiyonel ilişkilerin incelenmesi açısından önemli olduğu için; kümeleme gen verilerini minimum örten ağaç gibi göstermek amacıyla graflar oluşturarak önemli verilerde kayıp olmadan çok boyutlu veri seti ağaç yapısına indirgenmiş ve karmaşık gen verisi kümeleme açısından daha kolay işlenebilir hale getirilmiştir.

Şaylan (2013) 21 aylık süreyle tedavi gören böbrek nakli geçirmiş 178 hastaya ait veri seti ile yaptığı çalışmada; WEKA programı ile sınıflama, kümeleme ve karar ağacı algoritmaları ile elde edilen karar kuralları uzman eşliğinde incelenerek kronik arter kalsifikasyon bulunmasında etkili faktörlerin neler olduğu belirlenmiş ve en iyileme yöntem tabanlı öz nitelik seçme algoritması geliştirilerek aynı sonuçların elde edilip edilmediği kontrol edilmiştir.

Yıldırım vd. (2008) tıbbi VM ile Hacettepe Üniversitesinde yapılacak çalışma bilgilendirmesinde; hastane bilgilendirme sisteminde hastane genel bilgi sistemiyle tümleşik çalışan radyoloji bilgi sistemi, hastalar, tetkikler ve iş akışlarına ait verilerden oluşan büyük ilişkisel veri tabanı kaynağının doğru planlanması ve gelecekle ilgili ön görülerin yapılabilmesi için bu veriler ile yapılacak VM analizinin tıbbi çalışmalara katkısı olabileceği düşünülmektedir.

Siyah (2012) Oracle Data Miner ile göğüs kanseri verileri üzerinde yaptığı VM çalışmasında; oluşturduğu model %96,9 doğruluk oranı ile tahmin etme özelliğine sahiptir.

Bir başka çalışmada; büyük hacimli sağlık verileri üzerinde sınıflandırma tabanlı kural, karar ağacı ve yapay sinir ağı gibi sınıflandırma tekniklerini kullanarak diyabetik hastaların tıbbi veri setinde yapılan çalışmada veri madenciliğinin potansiyel kullanımı incelenmiştir. Çalışma sonucunda; yaş ve cinsiyet gibi verilerin karar ağacı yöntemleri ile analizi sonucu hastanın göğüs kanseri olup olmadığı tahmin edilmeye çalışılmıştır (Kaur and Wasan 2006).

Yine diyabet hastalığı ile yapılan tahminlerin doğruluğunu göstermek için kaba kümeler yöntemi kullanılarak akademik ortamda geliştirilen Rosetta isimli bir yazılımla 8 değişkenle analizi yapılan verilerin PIDD (Pima Indian Diabet Database-VM algoritmalarını sınamak için standart hale gelmiş bir veri tabanıdır) üzerinde kaba kümeler ile diyabet tahmin doğruluğu diğer algoritmalara göre daha başarılı bulunmuştur (Breault 2001).

Duru (2005) yaptığı çalışmada; DMAP (Data Mining with Apriori) isimli bir yazılım aracılığı ile Apriori algoritmasını kullanarak diyabet hastalarının sosyal durumlarını ortaya çıkarmıştır. Tavsiye edilen test bilgileri ile hasta bilgilerinin kullanarak teşhis performansını uygun ölçümlenmeyle optimizasyon yapan öğrenme tabanlı bir program geliştirilip kaba kümeler yöntemi kullanılarak yapılan çalışmada; destekli öğrenme ve markov karar süreçleri yöntemleri bir arada kullanılarak, uygun test stratejilerinin elde edilmesi kolaylaştırılmıştır (Fakih and Das 2006).

Sağlık alanındaki problemlerden biri olan kayıp değerli veri setleri ile yapılan bir

çalışmada; Random Forest yönteminin kayıp değerli veri setinde uygulanması durumunda kullandığı algoritmanın, k en yakın komşu ile kayıp değer atama yöntemi karşılaştırılması amaçlanmıştır. İki farklı aşamada gerçekleştirilen çalışmada; ilk aşamada benzetim yolu ile türetilen veri setleri, ikinci aşamada ise Eskişehir Osmangazi Üniversitesi Tıp Fakültesi Hastanesi Endokrinoloji Servis ve Polikliniğine 2011 ve 2013 yılları arasında başvuran tip2 diyabetli kırk yaş üstü hastalarda periferik arter hastalığı varlığı araştırmasına ilişkin 110 hastanın kayıp değerli bir veri seti kullanılarak karşılaştırmalar yapılmıştır. İlk aşamada karşılaştırmalar benzetim yolu ile yapılmıştır. Kayıp değerli veri setleri random forest kayıp değer atama yöntemi ve k=5, 10, 15, 20 olmak üzere kNN ile kayıp değer atama yoluyla ayrı ayrı tanımlanarak doğru sınıflandırma oranları karşılaştırmaları yapılmıştır. Elde edilen sonuçlar sınıfların doğru tahmin edebilme oranlarının ortalamasına göre değerlendirilmiştir (Yılmaz 2014).

Albayrak (2008) Bonn Üniversitesi Epileptomoloji Bölümü'nden aldığı elektroensefolagram (EEG) verileri üzerinde yaptığı çalışmada, epileptik aktivitenin, VERİ MADENCİLİĞİ süreçleri kapsamında sınıflama ve kümeleme algoritmaları kullanılarak belirlenmesi ile elde edilen sonuçlar karşılaştırılarak, verinin türüne göre uygun algoritma seçimi göz önüne alınıp önemi en yüksek doğruluk oranını sağlayan algoritma tespit edilmiştir.

1200 kişiye ait Elektromiyogram (EMG) verileri ile sınıflama ve kümeleme algoritmaları kullanılarak WEKA programında otomatik olarak yapılan veri analizinde en yüksek doğruluk oranını sağlayan MP5 karar ağaçları ile oluşturulan algoritmada nöropati ve miyopati hastalıklarının teşhisinde %95 oranında doğruluk elde edilmiştir (Arslan 2008).

Kalp kapakçığı hastalıklarının belirlenmesinde 15 ve 80 yaş aralığında değişen 83 kadın ve 132 erkek hasta için 95 normal ve 120 normal olmayan vaka sinyalinin 92 sinyal sistemin eğitilmesi için ve 123 sinyal ise test amacıyla kullanılmıştır.

Yapılan çalışma sonucunda normal değerleri tanıma oranı %84 normal olmayanları tanıma oranı ise %96 çıkmıştır (İnt.Kyn.1).

Hile ve yolsuzluk tespitlerinde veri madenciliğinde çalışmalara rastlamak da

mümkündür.

Sağlık uygulamaları ve tedaviler çok maliyet gerektirdiğinden yapılan tetkik ve tedavilerden hile yoluyla çıkar sağlanmaya çalışılabilir. Amerika ve Avrupa’da sağlık sigorta şirketleri bu konu üzerinde çalışmaktadırlar. Hile tespiti için hasta, tetkik ve doktor bilgileri VM yöntemlerinden genellikle kümeleme algoritmaları ile analiz edilir ve ortalama maliyet üzerinde olan tetkik ve tedaviler yahut anormal durumlar şüpheli olarak nitelendirilirler (Yıldırım vd. 2008).

Yapılan çalışmalar sadece biyolojik ve klinik düzeyinde VM algoritmaları ile sınırlı kalmamış ayrıca yönetsel anlamda da birçok çalışmaya konu olmuştur.

“Hastane yönetim bilgi sistemi” veri tabanında bulunan ve karar destek kapsamında bilgi verici konuları içeren verilerin veri ambarı ortamına taşınması çalışmasında; OLAP teknolojileri ile çok boyutlu küpler ile verilerin şekillenmesi sağlanmış; temizlenmiştir. Doğruluğundan emin olunan veriler karar ağacı VM modeli ile yeni anlamlar kazanmış ve veri ambarında tutulan verilerin çok daha hızlı raporlandığı gözlemlenmiştir (Düzgünoğlu vd. 2006).

Şişaneci (2009) sağlıkta performansa dayalı ek ödeme sisteminde VM ile suistimal tespitinin çözülebilmesi için yaptığı çalışmada; üroloji bölümünden alınan veriler veri önileme, tespit edilen modeli oluşturma ve değerlendirme adımlarından oluşan bir hibrit sistem tasarlayarak kural tabanlı, akıllı bir tespit modeli oluşturmuş ve modele göre işlenen veriler suiistimal eğilimli kayıtları uzmanların incelemesine sunmuştur. Geliştirilen suiistimal tespit modeli performansa dayalı ek ödeme sisteminde aynı seansta yapılan girişimsel işlemlerin doğru performans puanı alıp almadığı ve tekrar eden girişten kaynaklanan suiistimal olup olmadığı tespit edilebilmiştir. Ayrıca bu model gelen verilerle sürekli güncellenmektedir.

Sağlık yönetiminde çalışan karar alıcılara sağlık sektöründe VTBK’nin kullanımı ile ilgili örnek oluşturmak ve stratejik karar verme aşamasında o ana değin oluşan yapı hakkında bilgi sağlamak için çalışmalar yapılmıştır. Bu çalışmada ayrıca Gümüşhane Devlet Hastanesi hasta ve hastalık çeşitliliği, poliklinik hasta ilişkileri arasında örüntüler elde edilmeye çalışılmıştır (Özdemir vd. 2010).

2.5 Anemi (Kansızlık)

VM tekniklerinde yaşanan en büyük kısıtlama, üzerinde çalışılan veri setinin özelliklerine hâkim bir uzmanın görüş ve önerilerine mutlak ihtiyaç duyulmasıdır. Sağlık alanında VM ile yapılan çalışmalarda anlamlı kurallar ve ilişkilere; bu alanda uzmanlaşmış bir hekim tecrübesine ve bilgisine ayrıca ortaya çıkan bilginin yorumlanmasıyla ulaşılabilir.

Tetkik tahlil referans aralıkları klinisyen hekimlere hasta değerlendirmesinde yardımcı olan kritik önem sahip parametrelerdir (Özdemir 2011).

Kan, plazma ve kan hücrelerinden meydana gelen bir maddedir. Plazmanın önemli bir kısmı sudan oluşmaktadır. Plazmanın içerisinde suda çözünen proteinler (albümin), mineraller, hormonlar, vitaminler ve antikorlar bulunur. Kan üç ana grup hücre elemanından oluşur (İnt.Kyn.21)

1. Eritrositler (kırmızı kan hücreleri, Alyuvarlar, RBC)
2. Trombositler (kan pulcukları, PLT)
3. Lökosit (beyaz kan hücreleri, WBC); Nötrofil, Monosit, eozofil, bazofil ve lenfositlerden oluşur.

Eritrositler tüm vücuttaki hücrelerin çalışması ve yaşamını sürdürebilmesi için gerekli oksijenin akciğerlerden vücuttaki tüm hücrelere taşınmasını sağlar. Eksikliğinde kansızlık (anemi) meydana gelir. Trombositler kanamayı durduran hücreler olup, yaralanan damar bölgesine giderek kanayan yerde tıkaç oluşturup kanamayı durdururlar. Eritrosit ve trombositlerin aksine lökositler kandan dokulara geçerek mikroplarala savaşır ve enfeksiyonların iyileşmesini sağlarlar. Lökositler nötrofil, monosit, lenfosit, eozinofil ve bazofillerden meydana gelirler. Tüm lökositlerin %50-60'ını oluşturan nötrofil ve monositler mikropları öldürürler. Eozinofil ve bazofiller ise esas olarak alerjik olaylarda rol oynarlar (İnt.Kyn. 21).

2.5.1 Anemi Tanımı

Eritrositlerin yapım ve yıkımı arasındaki dengenin bozulması sonucu ortaya çıkan patolojik klinik tabloya anemi denir. Burada muayyen sayıdaki ve hacimdeki eritrosit ve

hemoglobinin azalması sözkonusudur. Kanın miktarındaki mutlak azalmalar anemi değildir (Akça 2007). Çevre kanında % gram (g/dL) olarak belirtilen hemoglobin konsantrasyonu hastanın yaş ve cinsiyete göre normal değerlerin altında olmasına anemi denir (Ercan 2006). Dünya sağlık örgütü (WHO) hemoglobin değerinin kadınlarda 12-14 gr/dL ve erkeklerde 13-16 gr/dL arasında olmasını normal, bu değerlerin altında olmasını anemi olarak tanımlamaktadır. Bu değerlerin ırklar arasında 1-2 gr/dL civarında farklılık gösterebileceği kabul edilmiştir (Nalbant ve Karan 2010). Anemi bir hastalık değil klinik bir bulgudur. Başlı başına bir hastalık olabileceği gibi birçok hastalığın klinik belirtilerinden biriside olabilir. Genelde anemide Hb konsantrasyonu ile birlikte, eritrosit sayısı ve hematokrit (alyuvarların kandaki hacimsel oranını belirten bir ölçü) değeride azalmış olup bu bir kural değildir. Bazen eritrosit sayısı normal sınırdaki olduğu halde, bu eritrositlerin içerdikleri Hb miktarının normalden az olması nedeniyle yine anemi tanısı konulur. Anemi bütün branşların vizitlerinde en sık rastlanan bulgudur. Ülkeler arasında anemi sıklığında büyük farklılıklar görülmektedir (Nalbant ve Karan 2010).

Demir metabolizmasını ve anemiye anlayabilmek için kırmızı kan hücrelerini (eritrosit) tanımak gerekir. Eritrositler oksijenin taşınmasını sağlayan hücreler olup, kemik iliğinde yapılır ve ortalama 120 gün yaşarlar. Eritrositin içini hemoglobin isimli protein doldurur ve yapısında demir bulunur. Hemoglobin demir yardımıyla oksijenin akciğerlerden alarak dokulara taşınması sağlar. Hemoglobin yapısında bulunan demir, kas dokusunda ve çeşitli enzimlerin yapısında bulunur. Vücudumuz için çok önemli olan demir oniki parmak bağırsağından emilir; oksijenin taşınması, hücre büyüme ve çoğalması gibi birçok olayda önemli rol oynar. Demirin fazlası ilerde kullanılmak üzere vücutta depo edilir, ancak aşırı birikmesi hücrelerin tahribatına neden olup zararlıdır (İnt.Kyn.21).

Normalde erişkinlerde eritrositlerin yapım yeri kemik iliği olup ilikteki pluripotent kök hücrelerden lenfoid sistem hücrelerinin (eritrosit, granülosit, monosit-makrofajlar, eozofiller, bazofiller, trombositler) çıktığı kabul edilir (Ercan 2006).

Dolaşan plazma hacmindeki değişikliklerin Hb konsantrasyonunu etkileyebileceği göz önünde bulundurulmalı, plazma hacminin azaldığı hemokonsantrasyon durumlarında hastada anemi maskelenebilir. Ayrıca plazma hacminde ki artışa bağlı olarak HGB

konsantrasyonunda azalma olmadığı halde, göreceli bir anemi gelişebilir (Ercan 2006).

2.5.2 Anemi Klinik Belirtileri

Hb'nin ana görevi akciğerlerden dokulara oksijen taşımak olup, anemide kanın oksijen taşıma kapasitesi azaldığından dokuların fonksiyonlarını bozarak çeşitli sistemlere ilişkin belirti ve bulguların ortaya çıkmasına yol açan, doku hipoksisi gelişir. Semptomlar öncelikli olarak iskelet kasları, kardiyovasküler sistem ve merkezi sinir sistemi gibi oksijen gereksinimi yüksek dokularda görülür. Anemi belirtilerinin ortaya çıkışı aneminin oluşum hızı, hastanın yaş ve altta yatan hastalık ilişkisidir. Büyük kan kaybı durumlarında ve birden gelişen ağır hemolizlerde semptomlar çok belirgin olmasına karşılık kronik anemi olgularında, düşük değerlere uyum sonucu belirtiler daha hafiftir. Kardiyovasküler belirtiler yaşlılarda, gençlere göre daha erken ve daha şiddetli ortaya çıkar. Anemili hastalarda başlıca yakınmalar aşağıdaki gibidir (Ercan 2006).

- Halsizlik, yorgunluk hissi ve kolay yorulma (psikonörotik hastalardada sıklıkla görülür).
- Efor dispnesi ve hiperpne; ağır anemilerde dinlenme halinde bile hava açlığı vardır.
- Çarpıntı ve taşikardi; orta derecede anemilerde taşikardi efordan sonra gelişir ve ağır anemilerde sürekli.
- Angina pectoris; özellikle yaşlılarda görülür.
- Baş dönmesi ve fenalık hissi; ayağa kalkıldığı zaman beyin hipoksisi arttığı için yakınma artar ve hasta yatarken daha rahattır.
- Ekstremitlerde uyuşma ve iğnelenme.
- Ayaklarda hafif ödem; ağır anemilerde kalp yetersizliği olmadan da görülebilir.
- Menstrüasyon bozuklukları; en önemlisi amenore.
- Lipido azalması, empotans.
- Sindirim sistemi yakınmaları; iştahsızlık, bulantı (Demir eksikliği anemisinde yutma güçlüğü, bazı anemilerde dilde yanma ve tat alamama gibi yakınmalar olabilir.)
- Ateş; az rastlanır bir durum olmakla beraber anemi düzeldiğinde kaybolur.

- Kronik bacak ülseri; konjenital hemolitik anemilerden orak hücreli anemi ve hereditör sferositozda ayak bileklerinin hemen üstünde ülserler ya da geçirilmiş olanların nedbeleri görülür.

2.5.3 Anemi Klinik Bulguları

Aneminin en önemli klinik bulguları arasında deri ve mukozadaki solukluk gelmektedir. Bunu takiben kardiyovasküler sistem ve diğer sistem bulguları izler.

Deri, mukoza ve tırnak yatakları ve konjonktivalarda belirgindir. Deri ve mukozaları sulayan kanda Hb miktarının azalması ve deri damarlarının vazokonstriksiyon yaparak kanın hayati diğer organlara gönderilmesi solukluğun oluşumunun nedenlerindendir. Ayrıca her deri solukluğu anemi olarak nitelendirilemez. Deri solukluğunu epidermisin kalınlığından, derideki kılcıl damarların sayısına kadar birçok faktöre bağlı olabilir (Ercan 2006, Akça 2007).

Kardiyovasküler sistem bulgularında; kan oksijen taşıma kapasitesi azaldığından dolayı organizma, dokulara yeterli oksijeni sağlayabilmek için kardiyovasküler düzenlemelerde kan akımını artırır.

- Taşikardi; kalbin dakikada hacmini artırıcı bir kompensasyon mekanizmasıdır. Anemide yalnızca atım sayısı değil ayrıca atım hacmi de artar.
- Nabız basıncında artma, Karotis pulsasyonları çok belirgindir.
- Dolaşım zamanı kısalabilir.
- Telekardiyografide kalp genişleme bulunabilir. Diğer dokularda olduğu gibi miyokartta da oksijen eksikliği vardır.
- Kalp üfürümleri
- Boyunda iki taraflı, karotisler üzerinde sistolik üfürüm, jugüler venler, üzerimde sürekli bir ses saptanabilir.
- Kesik kesik topallama görülebilir.

Diğer sistem bulguları ise aşağıda belirtildiği gibidir (Ercan 2006).

Nörolojik bulgular; anemili hastalarda baş ağrısı, baş dönmesi, huzursuzluk, sinirlilik ve konsantrasyon yeteneğinde azalma beyin hipoksisine bağlıdır. Pernisiyöz anemide

parestezi, derin duyu kaybı, yürüme güçlüğü ve patolojik refleks gibi bulgular fazladır.

Göz dibi; retina olup aplastik anemi ve akut lösemilerde kanama odakları görülebilir.

Dilde atrofi ve glossit, dudak köşelerinde çatlaklar, demir eksikliği anemileri ile megablastik makrositik anemilerde sık karşılaşılan bulgulardır.

Ağır anemi olgularında hafif olgularında hafif proteinüri ve böbrek işlevlerinde bozulma görülebilir.

2.5.4 Anemi Sınıflama ve Tanısal Yaklaşım

Aneminin genel sınıflandırılması ve başlangıç laboratuvar incelemeleri ile hastaya tanısal yaklaşım sağlamaktır. Anemik hastaya ilk tanısal yaklaşım ayrıntılı öykü, fizik muayene ve az sayıda temel laboratuvar testlerini içerir. Periferik kan yayması anemi tanısında çok faydalı olup klinisyenin bu düşünce sistemiyle düşünüp direkt tarama testlerinin yapılmasını ve gereksiz konsültasyon kaçınılması sağlanmış olur (Akarsu 2009).

Anemilerin tasnifi hakkında bugün için tam bir mutabakata varılamamıştır. Muhtelif sınıflamalar ve tasnifler mevcuttur (Akça 2007).

Anemiler patafizyolojiye göre, eritrosit morfolojisine göre ve klinik bulgulara göre olmak üzere üç şekilde sınıflandırılabilir. Aslında bu sınıflama ölçütlerinin tamamı tanının konulması için önemlidir. Üretime yönelik olarak yapılan sınıflamanın temel mantığı; yetersiz yapım ya da kanama nedeniyle aşırı bir kayıp olmasıdır. Bu temel yaklaşım çerçevesinde anemiler (Nalbant ve Karan 2010);

a)Azaltılmış Yapımlı Anemi; bu anemide kemik iliği üretiminde sorun vardır. Bu fonksiyon kaybına, azaltılmış eritrosit öncüllerine, kemik iliği infiltrasyonuna ya da beslenme faktörlerinde azalmaya bağlı olabilir.

b)Normal Yapımlı Anemi; bu tip anemide kemik iliği düşük eritrosit kitlesine artmış üretimle cevap verir. Bu sınıflamada temel belirleyici retikülosit sayısıdır. (kemik iliğinden dolaşıma yeni çıkan genç eritrositler.)

Pratikte anemilerin ayırıcı tanısında eritrositlerin; ortalama eritrosit hacmi (Mean Corpuscular Volume, MCV) temel alan morfolojik sınıflamadan yararlanılır (Ercan 2006,Nalbant ve Karan 2010). Mikrositik, normositik ya da makrositik olarak yapılabilen bu sınıflama tanının konulmasında önemli bir kolaylık sağlamaktadır. Anemi klinik görünümüne göre de akut (genellikle kanama ya da hemolize bağlı) ya da kronik olarak sınıflandırılabilir.

Bazı anemiler tek bir sınıfa dâhil edilemez; örneğin thalasemia hem yapım bozukluğu hem de yıkımda artma söz konusudur.

A. Kan Kaybı

- a. Akut kanama (akut posthemorajik anemi)
- b. Kronik kanama anemisi (demir eksikliği anemisi)

B. Eritrosit Yapım Bozukluğu

- a. Hb sentezi bozuklukları (hipokromik mikrositik anemiler)
 - i. Demir eksikliği
 - ii. Thalassemia'lar (globin sentezi bozukluğu)
 - iii. Sideroblastik anemiler ve kurşun zehirlenmesi (hem sentezi bozukluğu)
 - iv. Kronik hastalık anemisi (demir kullanımında bozukluk)
- b. DNA sentezi bozuklukları
 - i. B12 vitamin eksikliği
 - ii. Folik asit eksikliği
- c. Kemik iliği bozuklukları
 - i. Aplastik anemi, sadece dizi aplazisi
 - ii. Kemik iliğini en filtre eden hastalıklar; lösemiler, lenfomalar, multipl miyelom, kanser metastazları, miyeloid metaplazili miyeloskleroz, granülmömlü hastalıklar
 - iii. Sekonder anemiler; kronik infeksiyonlar, üremi, habis tümörler, romatizmal hastalıklar, kollajen doku hastalıkları, karaciğer hastalıkları, endokrin bez hastalıkları. Bu grup ta çoğunlukla birden fazla mekanizma rol oynar.

C. Eritrosit Yıkımında Artma (hemolitik anemiler)

- a. Eritrosit içi (intrakorpüsküler) bozukluklar
- b. Eritrosit dışı (ekstrakorpüsküler) bozukluklar

Morfolojik sınıflamada ise; eritrositlerin MCV değerlerine ya da morfolojik görünümüne dayanılarak yapılır. MCV, ortalama eritrosit hemoglobini (Mean corpuscular hemoglobin, MCH) ile de ilişkilidir. MCH eritrosit başına düşen hemoglobin miktarının (Normal, 27-32 pikogram) pikogram cinsinden ifadesidir. Dolayısıyla, MCV ve MCH ile birlikte mikrositer anemide azalırken, makrositer anemide artar. MCH konsantrasyonu (MCHC, Ortalama Eritrosit Hemoglobin Konsantrasyonu) her bir eritrositte ortalama hemoglobin konsantrasyonunu (%32-36) % ile ifade eder. Değişim aralığı çok küçük olduğu için hipokromide bile değişiklik göstermez. MCHC sadece hereditler sferositozda yükselir ve kullanım alanı azdır (Nalbant ve Karan 2010).

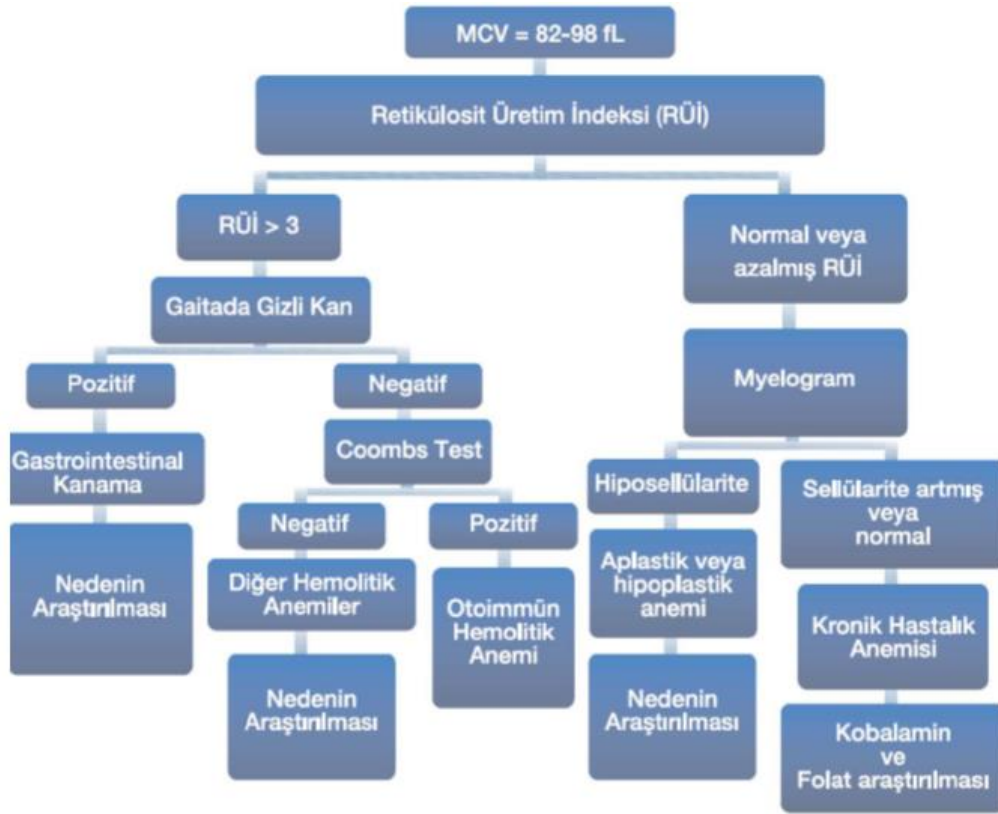
MCV homojen olmayan eritrosit morfolojisi hakkında genel fikir verir. Morfolojinin ne kadar heterojen olduğunun gösterilmesi için eritrositlerin dağılım eğrisi çizilerek buradan bir endeks elde edilir. Buna kırmızı hücre dağılım genişliği (red cell distributen width, RDW) denir. Normal değerleri %10-14 arası olan RDW anizositozun bir ifadesidir. Ayrıca laboratuvar imkânlarının gelişmesine karşın mümkün olan ve nedenin kesin olarak belirlenemediği bütün anemili hastalara mutlak periferik yayma (kanın mikroskop altında incelenmesi) yapılmalıdır (Nalbant ve Karan 2010).

Eğer Hemoglobin miktarı ile hematoktite bulunan hacim eritrosit sayısı ile orantılı olarak azalmışsa normositer anemiden bahsedilir (kanama anemileri, hemolitik anemiler, kemik iliği vazifesi bozukluğu anemileri, aplastik ve myelofitizik anemiler gibi). Eğer eritrosit sayısı hemoglobinden hacimce daha çok azalmış ise hücreler normalden büyük olup buna makrositer anemi denir. Burada karaciğerde hemopoetik maddenin eksikliği vardır. Orak hücreli ve akut hemolitik anemilerde karaciğerde bir bozukluk yoktur. Artan ihtiyacı karşılamak için hemopoetik sistem çokça olgunlaşmamış hücreleri periferik verir. Bu hücreler normal hücrelerden daha büyüktür. Hemoglobin miktarı ve hacmi eritrositlerin sayısına nisbetle daha azalmışsa hücreler normalden küçük olup mikrositer anemiden bahsedilir. B6 vitamini eksikliği ve infleksiyona bağlı anemiler mikrositer anemilerdir (Akça 2007).

A. Normositik Anemiler

MCV normal sınırlar içindedir (80-98 fL). Şekil 2.35’de tanısal inceleme şeması görülmektedir.

1. Akut posthemorajik anemiler
2. Talassemia, paroksimal gece hemoglobinürisi ve kardiyak hemolitik anemi dışında kalan tüm hemolitik anemiler
3. Aplastik anemi
4. Kemik iliğinin infiltrasyonu sonucu oluşan anemiler
5. Sekonder anemiler bu gruba dâhildirler.



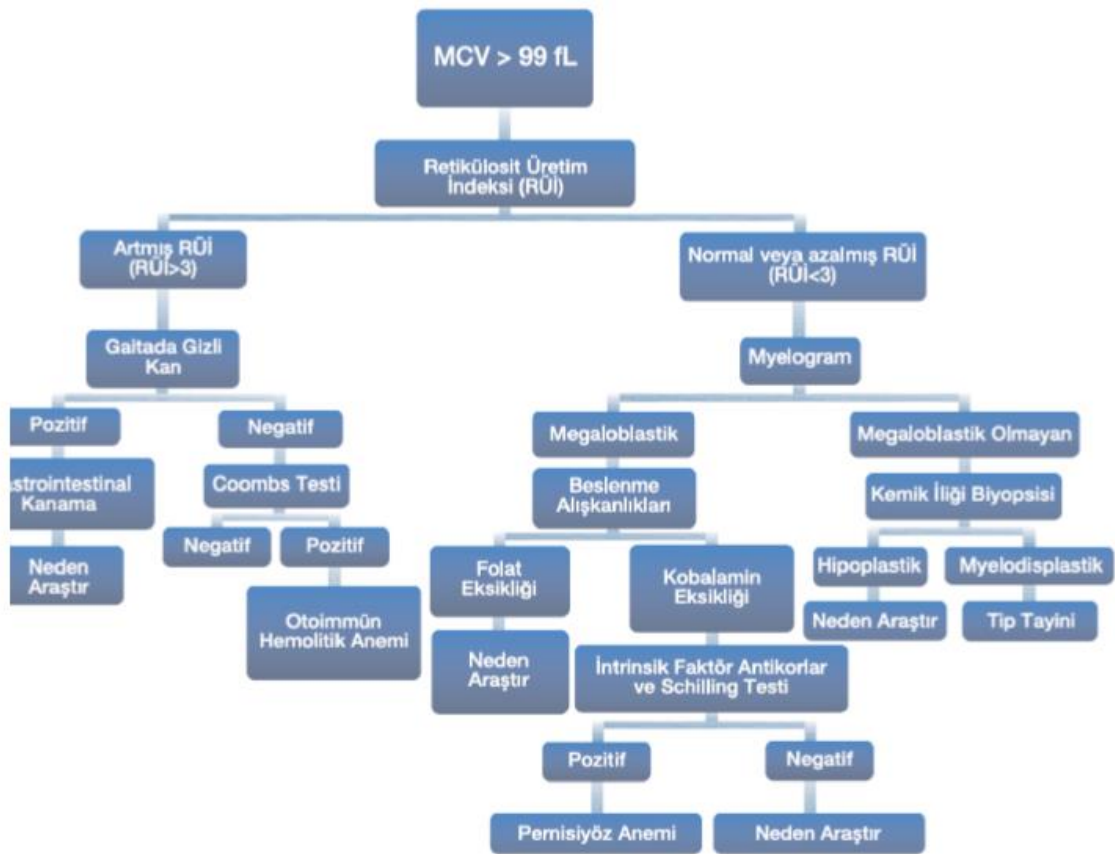
Şekil 2.35 Normositik aneminin tanısal inceleme şeması (Nalbant ve Karan 2010).

B. Makrositik Anemiler

MCV artmıştır (>98 fL). Makrositik anemiler kemik iliğinin megablastik ya da normoblastik olmasına göre ikiye ayrılır. Şekil 2.36’da tanısal inceleme şeması

görülmektedir.

1. Megablastik kemik iliği gösterenler
 - a) B12 vitamini eksikliğine bağlı anemiler
 - b) Folik asid eksikliğine bağlı anemiler
2. Normoblastik kemik iliği gösterenler
 - a) Akut kanama ve akut hemolizden sonra (retikülositoz durumları) Retikülositlerin çapı olgun eritrositlerinkinden büyüktür.
 - b) Akut lösemi, miyelodisplastik sendrom, aplastik anemi, miksödem, skorbüt, karaciğer hastalığı, alkolizm, sitotoksik ilaçları tedaviden sonra, gebelik.

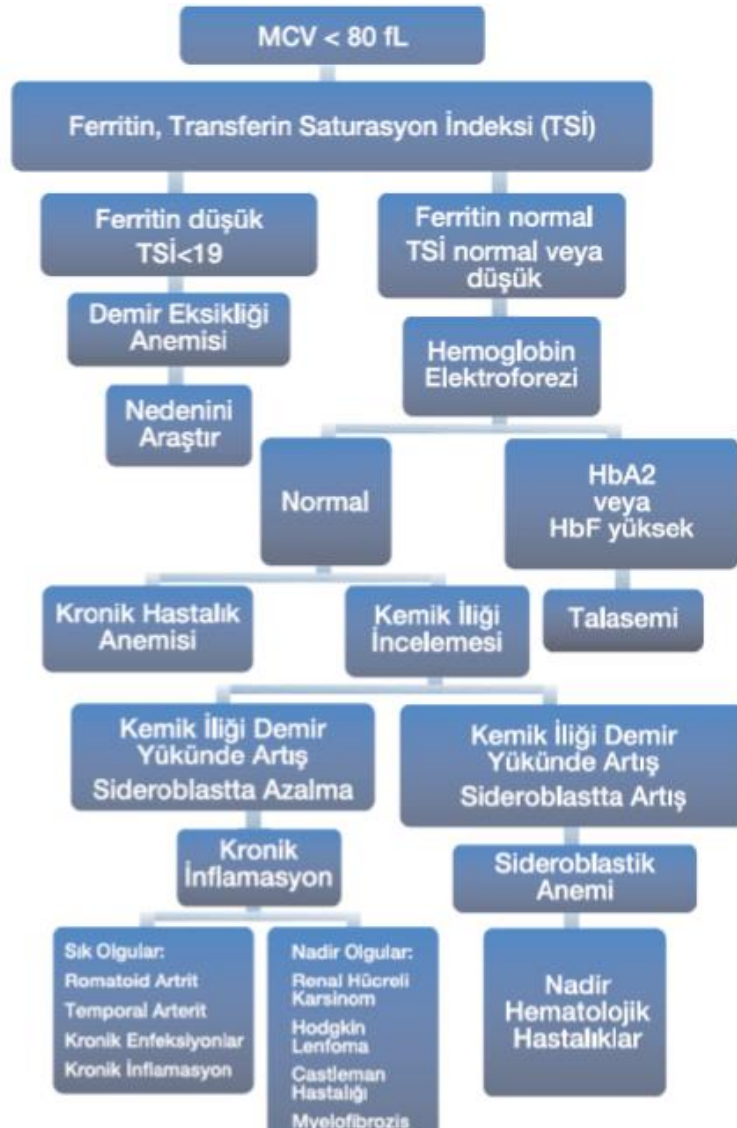


Şekil 2.36 Makrositik aneminin tanısal inceleme şeması (Nalbant ve Karan 2010).

C. Mikrositik Anemiler

MCV düşüktür (<80 fL). Şekil 2.37’de tanısal inceleme şeması görülmektedir. Bu grup içinde başlıca;

1. Demir eksikliği anemileri
2. Talassemia’lar
3. Sideroblastik anemiler ve kurşun zehirlenmesi bulunur
4. Bazen kronik enfeksiyon, habis tümör ve romatoid artrit anemileri mikrositik olabilir.



Şekil 2.37 Mikrositik aneminin tanısal inceleme şeması (Nalbant ve Karan 2010).

Demir eksikliği en sık anemi nedenidir. Demir eksikliği, çeşitli nedenlere bağlı olarak organizmadaki demir miktarının normal değerin altında olmasına denir. Demir eksikliğinde demir depoları normal kan yapımı için yeterli olmadığından gelişen anemi türüne demir eksikliği anemisi denir.

Anemi olmadan mikrositoz talasemi, polisitemi ise demir eksikliği anemisi açısından önemli bulgulardır. RDW talaseminin demir eksikliği anemisinden ayırt edilmesinde önemli bir parametre olup; mikrositoz varlığında RDW normal ise; talasemi, RDW 15’den büyükse demir eksikliği anemisi düşünülmelidir (Nalbant ve Karan 2010).

Anemi türlerinden olan demir eksikliği tanısını koyabilmek için detaylı tetkik gerekmektedir. Klasik olarak bakılan kan tetkikleri ile tam tanı koymak oldukça zordur. Bundan dolayı tam tanı çalışmaları akademik açıdan oldukça önemli olup tanı en az tedavi kadar öneme sahiptir. Bu nedenle uzman hekimler doğru tanı koyma aşamasında çeşitli yapay zekâ tekniklerinden faydalanmaktadırlar (Yılmaz 2012).

2.6 Çalışmanın Amacı, Önemi ve Kapsamı

Bu bölümde çalışmanın amacı, önemi ve kapsamına yer verilmiştir.

2.6.1 Çalışmanın Amacı

Bu çalışmanın amacı, ülkemizde ve dünyada büyük bir halk sağlığı sorunu olan anemi ve aneminin morfolojik sınıflandırılması amacıyla, biyokimya hemogram gibi laboratuvar testlerinin bugün bilinen temel tanımlayıcılar dışında teşhiste etkin rol oynayan diğer faktörler ile VM teknikleriyle tespitinin yapılarak oluşturulan modellerle veriler arasındaki ilişkileri ortaya çıkarmaktır.

2.6.2 Çalışmanın Önemi

VTYS’nin en çok kullanıldığı alanlardan biri olan sağlık sektöründe; hastalara ait, kimlik, demografik, tüm laboratuvar sonuçları, çekilen film ve röntgen görüntüleri vb. verilerin el ile analizinin mümkün olmaması, ayrıca veri tabanlarından geleneksel sorgulama metotları ile sorulması muhtemel soruların dışında bilgiyi filtrelemenin ve verilerin bu haliyle raporunun sunulması, ayrıca çok parametrelili karar vermenin insan faktörü ile zorluğu, veri tabanlarında gizli-önemli örüntülerin, kuralların ortaya

ıkmasını saęlayamaz. İlk olarak akla gelen, belirtilen nedenler dahi bu alanda VM tekniklerinin kullanılması iin yeterli sebepler arasında gsterilebilir.

Hastalıkların erken teřhisi, korunma ve tedavi ařamalarında bilgisayar bilgi sistemlerin kullanımı saęlık personeli aısından iř yknn ve hastane aısından maliyetin azaltılması aısından olduka nem arz etmektedir.

Bu alıřmanın temel hedefi anemiye etki eden parametreleri ve deęerlerini, bu deęerlerin hangi oranda hastalıęa etki ettiklerini tespit edip bu yntemlere dikkat ekmektir. alıřmada VM teknikleri kullanılarak, gerek laboratuvar verileri ile oluřturulan veri setlerinden bilgi keřfi yapılmıř ve bu bilgiler kural setlerine dnřtrlerek uzman bir sistemin alt yapısının oluřturulması, bu alanda alıřma yapacaklara ve saęlık profesyonellerine yardımcı olmak amalanmıřtır. Bu amala alıřmada tm sınıflandırma algoritmaları nce kendi aralarında daha sonra birbirleri ile karřılařtırılmıřtır.

Literatrde anemi ile ilgili birok alıřmaya rastlamak mmkndr. alıřmaların byk bir oęunluęunun temel belirleyiciler ile aneminin bir tr olarak zellikle demir eksiklięi anemisi zerinde yoęunlařtıęı grlmektedir. Bu tez kapsamında yapılan anemi eřitlerinin sınıflandırılmasına (Mikrositer, Normositer, Makrositer) ynelik yapılan bir VM alıřmasına rastlanmamıřtır.

2.6.3 alıřmanın Kapsamı

alıřmanın gerek hasta verilerinden oluřması gerektięi iin; veriler il kamu hastaneler birlięinden alınan izin ile elde edilmiřtir. Alınan veriler, hasta kimlik ve tanımlayıcı bilgileri iermeyen laboratuvar sayısal verileridir. Yapılan resmi yazıřma izin dilekesi Ek 1’de verilmiřtir.

Arařtırmanın kapsamı; Amasya ilinde faaliyet gsteren bir kamu hastanesinin 2014 yılı Temmuz, Aęustos, Eyll, Ekim, Kasım, Aralık ve Ocak aylarına ait laboratuvar verileri ile sınırlıdır. alıřmada anemi ve trleri iin 5831 hasta, 12 znitelik ile analizler belirlenen algoritmalar ile WEKA programında yapılmıřtır.

3. MATERYAL METOT

Bu çalışmada VM uygulaması CRISP-DM referans modeli takip edilerek gerçekleştirilmiştir.

3.1 Verilerin Hazırlanması ve Veri Seti Özellikleri

Verilerin hazırlanması aşaması toplama, birleştirme ve temizleme, seçme ve dönüştürme adımlarından oluşmaktadır. Veriler Amasya iline bağlı bir kamu hastanesinin 2014 yılı Temmuz, Ağustos, Eylül, Ekim, Kasım, Aralık ve Ocak aylarına ait aylık laboratuvar verilerinin, hastane yazılımından Microsoft Office 2010 Excel ortamına aktarılması ile elde edilmiştir. Veriler daha sonra biyokimya uzmanı yardımı ile gerekli birleştirme ve temizleme işlemleri yapıldıktan sonra, WEKA programında çalışması için .csv formatında düzenlenmiştir.

Laboratuvar tetkikleri bir hasta için aynı satırda toplandığında 98 öznitelikten oluşan veri setinden, uzman görüşü ile anemi için teşhis koymada etkinliği olmadığı kuvvetle düşünülen parametrelerin çıkartılması sonucu kalan 34 özneliğin özellikleri referans aralıkları ve boş olan kayıtlara atanan değerler Çizelge 3.1’de görülmektedir.

Referans aralığı; testler için belirtilen referans aralığı sağlıklı bireylerden elde edilen değerlerin %95’ini içeren grubu temsil eden test sonucunu değerlendirmek için genel bir baz oluşturur. Bu değerler çeşitli faktörlere bağlı olarak hastaya göre farklı varyasyonlar gösterebilir. Analiz yönteminin değiştiği durumlarda metoda bağlı olarak referans aralığı da değişebilir (Üstün 2013). VM çalışmalarında birçok algoritma eksik değerlerden olumsuz olarak etkilendiği için referans aralıkları, uzman yardımı ile veri setindeki aynı kayda ait diğer değerler gözönünde bulundurularak sınıf etiketini değiştirmeyecek şekilde tekrar düzenlenerek yazılmıştır.

Çizelge 3.1 Veri seti referans aralıkları ve boş değerler (Wu and Alan 2006, Üstün 2013).

	Parametre	Kısaltması	Referans aralığı		Boş Değerler
1	YAŞ	YAŞ	0-100		Yok
2	CINSİYET	CINSİYET	E:Erkek	K: Kadın	Yok
			0-1 Yaş	13-45 U/L	
			1-60 Yaş	Erkek 10-40 U/L Kadın 7-35 U/L	
3	ALT	ALT	>=60-90 Yaş	Erkek 13-40 U/L Kadın 10-28 U/L	23 U/L
			>=90 Yaş	Erkek 6-30 U/L Kadın 5-24 U/L	
			0-2 Yaş	9-80 U/L	
			>=2-60 Yaş	Erkek 15-40 U/L Kadın 13-35 U/L	
4	AST	AST	>=60-90 Yaş	Erkek 19-48 U/L Kadın 9-36 U/L	25 U/L
			>=90 Yaş	Erkek 11-38 U/L Kadın 18-30 U/L	
5	BASO%	BASO	0-2,5 (%)		0 (%)
6	CRP (Nefelometrik)	CRP	0-0,5 mg/dL		0,25 mg/dL
			Erkek 65-193 µg/dL		<2 Yaş 70
			Kadın 33-170		>=2-12 Yaş 85
7	Demir (serum)	Demir			>=12 Yaş Erkek 120 Kadın 110
	Demir bağlama kapasitesi	DBK	Erkek 110-345 µg/dL		<2 Yaş 200
			Kadın 135-392 µg/dL		>=2 Yaş 240
9	TIBC	TIBC	µg/dL		
10	TSI	TSI	>=19		
11	EO%	EO	0-7 (%)		
			Erkek 22-322 ng/mL		<15 Yaş 73,5
			Kadın 10-291 ng/mL		>=15 Yaş
12	Ferritin	Ferritin			Erkek 135 Kadın 65
13	Folat	Folat	5-16 ng/mL		10,5
14	GGT (Gamma glutamil)	GGT	Erkek 8-61 U/L		Erkek 35
			Kadın 5-36 U/L		Kadın 21

Çizelge 3.1 (Devam) Veri seti referans aralıkları ve boş değerler (Wu and Alan 2006, Üstün 2013).

		0-1 Yaş	80 mg/dL	<1 Yaş	65	
15	Glukoz	Glukoz	>=1-17 Yaş	60-100 mg/dL	>=1-12 Yaş	80
			>=18 Yaş	74-109 mg/dL	>=12 Yaş	90
					<15 Yaş	38,5
16	HCT	HCT	0-14 Yaş	32-45 (%)	>=15 Yaş	
			>=15 Yaş	Erkek 37-45 (%)	Erkek	44
				Kadın 34-48 (%)	Kadın	41
					<12 Yaş	12,73
17	HGB	HGB	0-12 Yaş	11-14,5 g/dL	>=12 Yaş	
			>=13 Yaş	Erkek 12-17,4 g/dL	Erkek	14,7
				Kadın 11,7-16 g/dL	Kadın	13,85
			0-5 Yaş	0,17-0,47 mg/dL	<12 Yaş	0,5
					>12-18 Yaş	0,75
			>=5-11 Yaş	0,32-0,73 mg/dL	>=18 Yaş	
18	Kreatinin	Kreatinin	>=11-15 Yaş	0,53-0,87 mg/dL	Erkek	1,1
			>=15 Yaş	Erkek 0,70-1,2 mg/dL	Kadın	0,85
				Kadın 0,5-1,1 mg/dL		
			0-2 Yaş	180-430 U/L		
19	LDH (Laktik Dehidrogenaz)	LDH	2-15 Yaş	120-300 U/L		210
			>=15 Yaş	Erkek 135-225 U/L		
				Kadın 135-224 U/L		
20	LYMPH%	LYMPH	10-50 (%)			
					<15 Yaş	29
21	MCH	MCH	27-34 pg		>=15 Yaş	
					Erkek	30,5
					Kadın	30
					<12 Yaş	34,5
22	MCHC	MCHC	32-36 g/dL		>=12 Yaş	
					Erkek	34,5
					Kadın	34

Çizelge 3.1 (Devam) Veri seti referans aralıkları ve boş değerler (Wu and Alan 2006, Üstün 2013).

				<2 Yaş 77
				>=2-12 Yaş 81
23 MCV	MCV	0-10 Yaş	70-87 fL	>=12 Yaş
				Erkek 88
		>=11 Yaş	80-99 fL	Kadın 86,5
24 MONO%	MONO	0-12 (%)		6
25 NEUT%	NEUT	37-80 (%)		58
26 PLT	PLT	142-424	10 ³ cells/μL	283
				0-12 Yaş 4,3
27 RBC	RBC	0-12 Yaş	3,5-5,1 x10 ⁶ cells/μL	>=12 Yaş
				Erkek 4,9
		>=13 Yaş	Erkek 4,1-5,7 x10 ⁶ cells/μL Kadın 3,8-5,1 x10 ⁶ cells/μL	Kadın 4,45
28 RDW-CV	RDW-CV	11,7-14,4 (%)		13
29 Serbest T4	SerbestT4	0-11 Yaş	1,07-1,75 ng/dL	<20 Yaş 1,41
		>=12 Yaş	0,7-2 ng/dL	>=20 Yaş 1,35
30 Total Bilirubin	TBilirubin	0,1-1,2 mg/dL		0,75
				<=20 Yaş 3,55
31 TSH	TSH	0,4-4,2 mIU/mL		>20-54 Yaş 2,3
				>=55 Yaş 4,7
				<18 Yaş 1,5
32 ÜRE	URE	10-50 mg/dL		>=18-60 Yaş 13
				>=60-90 Yaş 15,5
				>=90 Yaş 20,5
		0-2 Yaş	160-1300 pg/mL	
33 Vitamin B12	VitB12	>=2-60 Yaş	200-835 pg/mL	289*
		>=60 Yaş	110-770 pg/mL	
34 WBC	WBC	0-15 Yaş	4,5-13 x10 ³ cells/μL	7,75
		>=16 Yaş	4,5-10,5 x10 ³ cells/μL	
35 SONTANI	TANI	Anemi, AnemiDegil/ Makro, Normo, Mikro, Degil		

* Hastanenin bir yıllık vitamin B12 verileri incelendiğinde ortalama değer olarak 289 pg/mL olarak bulunmuştur.

Boyut İndirgeme

Hastane veri tabanından alınan veriler; iki adet Excel çalışma kitabında, her tabloda hiçbir hasta tanımlayıcı bilgisi olmadan salt sayısal veriden oluşan ortalama 30 000 kayıt olacak şekilde, toplam kırk tabloda kodu, ISLEMNO, TUPADI, SONUC,

SONUCTARIH, SERVISADI, TANI, DOGUM TARİH, CINSİYET ve ID alanlarından oluşan toplam 1 170 260 adet veriden oluştuğu Şekil 3.1’de görülmektedir.

Verilerle ilgili bir sorun verilerin VM yazılımında kullanmaya tam olarak hazır olmamasıdır. Laboratuvar sonuçları Microsoft Excel yazılımında satır satır görülmektedir. Ama VM yazılımı açısından analizlerde, tek bir kayda ait verilerin tek bir satırda gözükmesi gerekmektedir. Bu sorunun çözümü için Visual Basic Uygulaması (VBA, Visual Basic Application) kullanılarak bir makro yazılmıştır. Bu işlem sonucunda bir kayda ait bilgiler tek bir satırda toplanmıştır. Yazılan makro sayesinde veri setinde tekrar eden ID değerleri tek kayıt olarak ele alınarak 88 tetkik adı, tanı, ID, KODU, ISLEMNO, SONUCTARIH, YAS, CINSİYET, SERVISADI VE TUPADI olmak üzere toplam 98 öznitelik içerecek şekilde düzenlenmiş olup Şekil 3.2’de görülmektedir. Makrodan ötürü tetkik değerlerinin büyük bir kısmı metin olarak saklanan sayı şeklinde aktarıldığından, referans aralıkları ve boş değerler için formül yazımlarında kullanılmak üzere bu değerler sayıya dönüştürülmüştür. Farklı laboratuvar tetkiklerinden aynı hastaya ait sonuçları tek bir kayıta tutmak için hasta ayırt edici olarak ID isimli hastane tarafından belirlenen değer kullanılmış olup; analizde başlangıç itibari ile ID, KODU, ISLEMNO, ISLEMTARIH, DOGUMTARIH kolonları analiz yapılmadan önce veri setinden çıkarılmıştır. Yapılacak analiz ile ilgili yalnızca ihtiyaç duyulan verileri belirlemek gerektiğinden, bu kapsamda uzman yardımı ile tanısı kansızlık ve kansızlıkla benzer semptom gösteren kayıtlar filtrelenerek yeni bir excel dosyası oluşturulmuştur.

Genel muayeneler diğer, Baş ağrısı, Çarpıntı, Vitamin B12 eksikliği anemisi, tanımlanmamış, Kırgınlık ve yorgunluk, Demir eksikliği anemisi, Anemi, diğer, Baş dönmesi (Vertigo), Vitamin B12 eksikliği anemisi, Genel tıbbi muayene, Demir eksikliği anemisi, tanımlanmamış, Vitamin B12 eksikliği anemisi diğer, Baş ağrısı, diğer sendromları, Anemi tanımlanmamış, Anemiler diğer tanımlanmış, Demir eksikliği, Demir eksikliği anemileri diğer, Vitamin B12 eksikliği anemisi, seçici vitamin B12 malabsorbsiyonuna bağlı proteinüri ile birlikte Baş ağrısı sendromları diğer, tanımlanmış olmak üzere kansızlık ve benzer semptomlara sahip bu tanılar filtrelendiğinde 168 421 adet kayıttan oluşan veri seti elde edilmiştir.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	KODU	İSLEMNO	TETKİKADİ	TUPADI	SONUC	SONUCTARİH	SERVISADİ	TANI		DOĞUMTAR	YAS	CİNSİYET	ID
2	797740	A27701	EO#	HEMOGRAM	0,15	27.01.2015 13:15:48	Kbb Pol. [N	Boynun lokalize şişn	06.04.1955	59		Kadın	12988560684
3	797740	A27701	BASO#	HEMOGRAM	0,07	27.01.2015 13:15:48	Kbb Pol. [N	Boynun lokalize şişn	06.04.1955	59		Kadın	12988560684
4	797740	A27701	RDW-SD	HEMOGRAM	38,1	27.01.2015 13:15:48	Kbb Pol. [N	Boynun lokalize şişn	06.04.1955	59		Kadın	12988560684
5	797740	A27701	PCT	HEMOGRAM	0,240	27.01.2015 13:15:48	Kbb Pol. [N	Boynun lokalize şişn	06.04.1955	59		Kadın	12988560684
6	797740	A27701	P-LCR	HEMOGRAM	23,8	27.01.2015 13:15:48	Kbb Pol. [N	Boynun lokalize şişn	06.04.1955	59		Kadın	12988560684
7	797740	A27701	CRP (Nefelc	BIYOKİMYA	3,5	27.01.2015 13:32:27	Kbb Pol. [N	Boynun lokalize şişn	06.04.1955	59		Kadın	12988560684
8	797620	A27467	VLDL	BIYOKİMYA	52,6	27.01.2015 11:52:05	Kardiyoloji Kronik	iskemik kalp	03.06.1931	83		Kadın	53551533066
9	797656	A27426	Globulin	BIYOKİMYA	2,4	27.01.2015 11:52:29	Fizik Tedav	Hemipleji	28.02.1966	48		Erkek	18889364772
10	797235	A26144	VLDL	BIYOKİMYA	49	27.01.2015 11:53:10	Cildiye Pol	Akne	25.12.1994	20		Kadın	18127389696
11	797741	A19610	Gluko test (	BIYOKİMYA	274	27.01.2015 11:53:57	Kardiyoloji Kalp	yetmezliği	09.09.1947	67		Kadın	22930231214
12	797742	A11826	Gluko test (	BIYOKİMYA	269	27.01.2015 11:53:57	Kardiyoloji Kalp	yetmezliği	07.02.1954	60		Kadın	20323317596
13	797743	A12813	Gluko test (	BIYOKİMYA	191	27.01.2015 11:53:57	Ortopedi S	Muayene ve gözlem	30.09.1939	75		Kadın	15601487788
14	797744	A21390	Gluko test (	BIYOKİMYA	124	27.01.2015 11:53:57	Ortopedi S	Muayene ve gözlem	20.07.1946	68		Kadın	26875099174
15	797745	A27847	WBC	HEMOGRAM	7,59	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
16	797745	A27847	NEUT%	HEMOGRAM	69,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
17	797745	A27847	LYMPH%	HEMOGRAM	20,4	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
18	797745	A27847	MONO%	HEMOGRAM	5,4	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
19	797745	A27847	EO%	HEMOGRAM	3,1	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
20	797745	A27847	BASO%	HEMOGRAM	1,3	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
21	797745	A27847	RBC	HEMOGRAM	4,84	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
22	797745	A27847	HGB	HEMOGRAM	12,9	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
23	797745	A27847	HCT	HEMOGRAM	42,0	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
24	797745	A27847	MCV	HEMOGRAM	86,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
25	797745	A27847	MCH	HEMOGRAM	26,7	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
26	797745	A27847	RDW	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
27	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
28	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
29	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
30	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
31	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
32	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
33	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
34	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
35	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
36	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
37	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
38	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
39	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
40	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
41	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
42	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
43	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
44	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
45	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
46	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
47	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
48	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
49	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
50	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
51	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
52	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
53	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
54	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
55	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
56	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
57	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
58	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
59	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
60	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
61	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
62	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
63	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
64	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
65	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
66	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
67	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
68	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
69	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
70	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
71	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
72	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
73	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
74	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
75	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
76	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
77	797745	A27847	PLT	HEMOGRAM	199	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
78	797745	A27847	MPV	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
79	797745	A27847	PDW	HEMOGRAM	9,5	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
80	797745	A27847	RDW-CV	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
81	797745	A27847	RDW-SD	HEMOGRAM	13,8	27.01.2015 14:04:46	Nöroloji Pç	Tremor, tanımlanma	10.05.1931	83		Kadın	14824483700
82	797745	A27847	PLT	HEMOGRAM									

Şekil 3.1 Ham veriler.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	ID	KODU	İSLEMNO	SONUCTARİH	YAS	CİNSİYET	SERVISADİ	TUPADI	TANI	25-Hid	Albümin	Alfa- fetop	Alkale	ALT	Amilaz	Anti	Anti T	ASO	AST	BASO#	BASO%	Beta-I
5813	24886164864	790506	A9048	12.01.2015 12:	38	Kadın	Kadın Has. Pol. [Ş.Ki	HEMOGRAM	Anemi, diğer	4,9					9	155,3	> 1300,0		13	0,01	0,2	0,0
5814	21481262568	790522	A9072	12.01.2015 12:	21	Kadın	Dahiliye Pol. [S.Sezi	HEMOGRAM	Demir eksikliği anemisi											0,05	0,7	
5815	12728559650	790284	A9171	12.01.2015 09:	19	Erkek	Göğüs Has. Pol. [H.E	HEMOGRAM	Genel muayeneler, diğer											0,03	0,4	
5816	21235292128	790635	A9187	15.01.2015 11:	16	Kadın	Dahiliye Pol. [S.Sezi	HEMOGRAM	Kirgnlık ve yorgunluk											0,02	0,4	
5817	22495670138	790378	A9216	12.01.2015 10:	70	Kadın	Nöroloji Pol. [S.Ayk	HEMOGRAM	Baş dönmesi (Vertigo)											0,01	0,2	
5818	21655256676	790397	A9228	12.01.2015 11:	38	Erkek	Üroloji Pol. [F.Koç	BIYOKİMYA	Genel muayeneler, diğer													
5819	22147255850	790313	A9269	12.01.2015 09:	21	Kadın	Acil Servis	HEMOGRAM	Baş ağrsı											0,03	0,4	
5820	17884399120	790347	A9319	12.01.2015 09:	98	Erkek	Acil Servis	HEMOGRAM	Anemi, diğer					82						93	0,01	0,1
5821	17167410996	790344	A9326	12.01.2015 12:	26	Kadın	Göğüs Has. Pol. [H.E	HEMOGRAM	Genel muayeneler, diğer											0,02	0,5	
5822	19627340844	792576	A9343	16.01.2015 09:	20	Kadın	Kadın Has. Pol. [Ş.Ki	HEMOGRAM	Anemi, diğer	4,2					9					14	0,01	0,1
5823	17317417640	790553	A9347	12.01.2015 13:	50	Kadın	Kardiyoloji Pol.[M.Ç	BIYOKİMYA	Çarpıntı						36					25		
5824	53740086496	790485	A9400	12.01.2015 11:	55	Kadın	Acil Servis	HEMOGRAM	Baş ağrsı					16						15	0,00	0,0
5825	53740086496	790450	A9456	12.01.2015 11:	55	Kadın	Nöroloji Pol. [S.Ayk	HEMOGRAM	Baş ağrsı											0,01	0,1	
5826	19705336396	790456	A9560	12.01.2015 11:	82	Erkek	Genel Cerrahi Pol. [	HEMOGRAM	Demir eksikliği anemisi											0,05	0,8	
5827	25804118662	790499	A9633	12.01.2015 11:	26	Kadın	Nöroloji Pol. [S.Ayk	HEMOGRAM	Baş ağrsı											0,02	0,4	
5828	12361595900	790513	A9687	12.01.2015 11:	46	Kadın	Acil Servis	HEMOGRAM	Genel muayeneler, diğer					24	68					17	0,02	0,5
5829	13375548432	790610	A9779	12.01.2015 14:	75	Kadın	Yoğun Bakım Ser. K:	BIYOKİMYA	Çarpıntı					22						20	0,04	0,3
5830	29434014366	790594	A9817	12.01.2015 14:	69	Kadın	Kardiyoloji Pol.[M.Ç	HEMOGRAM	Çarpıntı											0,03	0,5	
5831	10945619828	790659	A9876	13.01.2015 10:	37	Kadın	Kbb Pol. [N.Akkan]	HEMOGRAM	Baş dönmesi (Vertigo)											0,02	0,3	
5832	23929197754	790625	A9879	13.01.2015 11:	54	Kadın	Dahiliye Pol. [S.Sezi	BIYOKİMYA	Kirgnlık ve yorgunluk													
5833	18103373962	790597	A9880	12.01.2015 14:	65	Erkek	Kardiyoloji Pol. [Y.Y	HEMOGRAM	Genel muayeneler, diğer					15						14	0,02	0,3
5834	12346583372	790600	A9932	12.01.2015 14:	30	Erkek	Göğüs Has. Pol. [H.E	HEMOGRAM	Genel muayeneler, diğer					8						13	0,01	0,1
5835	11569609330	790602	A9936	12.01.2015 14:	32	Erkek	Göğüs Has. Pol. [H.E	HEMOGRAM	Genel muayeneler, diğer					15						14	0,01	0,1
5836	25783135958	790608	A9939	12.01.2015 14:	41	Kadın	Acil Servis	HEMOGRAM	Anemiler diğer, tanımlanmış					60	13					11	0,01	0,2

Şekil 3.2 Her bir kayıt için yapılan tetkiklerin ayrıntılı gösterimi.

Başlangıçta 98 özniteliği olan veriler makine hatası ile okunamayan test sonuçlarının ve tüm kayıtlar için boş değer içeren öznitelikler uzman önerisi ile veri setinden çıkarılması sonucu; veri indirgeme işlemi gerçekleştirilerek 34 Öznitelik ve bir sınıf bilgisinden oluşan 5831 hasta verisi analiz için hazır hale getirilmiştir. Böylelikle veri kalitesinin artması ve VM için harcanacak zamanın azalması sağlanmıştır.

Veri Temizleme, Birleştirme, Dönüştürme ve Seçme

Veri girişindeki insan ve bilgisayar kaynaklı hataların gürültüye neden olmasından ötürü, veri kalitesini artırmak amacıyla ve analizlerdeki yanlışlığın önüne geçebilmek için tutarsız ve hatalı olan veriler uzman görüşü ve onayı alınarak veri setinden silinerek temizlenmiştir. Ayrıca hastalık teşhisinde betimleyici olduğu düşünülen değişkenlerden boş değer içeren kayıtlara, uzman yardımıyla referans aralıkları ve diğer öznel değerleri dikkate alınarak ve sınıf bilgisine bağlı kalınmak suretiyle eksik verilerin manuel olarak tanımlanması Şekil 3.3’de görülmektedir.

BD2		=EĞER(VE(BC2="";A2<12);4,3;EĞER(VE(BC2="";A2>=12;B2="E");4,9;EĞER(VE(BC2="";A2>=12;B2="K");4,45;BC2)))																	
Ad Kutusu	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	
1 LDH (Laktik Dehidrogenaz)	LYMPH%	LYMPH%	MCH	MCH	MCHC	MCHC	MCV	MCV	MONO%	MONO%	NEUT%	NEUT%	PLT	PLT	RBC	RBC	RDW-CV	RDW-CV	
2	210	38,6	38,6	35,3	35,3	34,8	101,6	101,6	10,3	10,3	46,7	46,7	404	404	4,98	4,98	13,6		
3	210	64,3	64,3	20,9	20,9	32,3	32,3	64,5	64,5	5	5	27,8	27,8	295	295	5,22	5,22	21,7	
4	210	60,2	60,2	22,7	22,7	34,7	34,7	65,5	65,5	6,9	6,9	26,6	26,6	344	344	4,98	4,98	14,5	
5	210	61,4	61,4	26,4	26,4	37,2	37,2	70,8	70,8	6,2	6,2	27,7	27,7	406	406	5,08	5,08	11,9	
6	210	69,6	69,6	24,1	24,1	34,3	34,3	70,2	70,2	4,5	4,5	23,3	23,3	440	440	4,28	4,28	15,7	
7	210	57,6	57,6	32,6	32,6	35,1	35,1	93	93	10,6	10,6	27,5	27,5	289	289	4,79	4,79	13,5	
8	210	74,1	74,1	26,4	26,4	34,5	34,5	76,6	76,6	7,6	7,6	16,4	16,4	365	365	4,88	4,88	11,1	
9	210	68	68	27,8	27,8	34,2	34,2	81,2	81,2	8,9	8,9	20,6	20,6	350	350	4,28	4,28	12	
10	210	73,2	73,2	27,7	27,7	33,5	33,5	82,4	82,4	4,7	4,7	19	19	443	443	4,05	4,05	12,2	

Şekil 3.3 Eksik değerlerin tanımlanması.

Veriler bir sonraki aşamada model içerisinde farklı birimler ve değerlerde bulunan özelliklerin sayı değerleri yüksek olanların sayı değeri düşük olanları baskılamaması için veriler, yaş bilgisi de dâhil olmak üzere max-min normalizasyon kurallarına göre normalize edilmiştir (Şekil 3.4).

O5792		=(N5792-N\$5834)/(N\$5835-N\$5834)																	
1	AST	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
	AST		BASO	BASO	CRP	CRP	Demir	Demir	DBK	DBK	TIBC	TIBC	TSI	TSI	EO	EO	Ferritin	Ferritin	Folat
5792	13	0,03196347	0,3	0,115384615	0,25	0,007225	110	0,3866171	240	0,344887348	350	0,447566	31,43	0,38	1,9	0,076305221	65	0,014581184	10,5
5793	27	0,095890411	0,1	0,038461538	1,9	0,054913	120	0,423791822	240	0,344887348	360	0,466292	33,33	0,41	4	0,16064257	135	0,030283997	10,5
5794	25	0,086757991	0	0	0,25	0,007225	120	0,423791822	240	0,344887348	360	0,466292	33,33	0,41	2	0,080321285	135	0,030283997	10,5
5795	25	0,086757991	0,9	0,346153846	0,25	0,007225	110	0,3866171	240	0,344887348	350	0,447566	31,43	0,38	1,7	0,068273092	65	0,014581184	10,5
5796	14	0,03652968	0	0	0,25	0,007225	110	0,3866171	240	0,344887348	350	0,447566	31,43	0,38	2	0,080321285	65	0,014581184	10,5
5797	25	0,086757991	0	0	0,25	0,007225	110	0,3866171	240	0,344887348	350	0,447566	31,43	0,38	2	0,080321285	65	0,014581184	10,5
5798	15	0,04109589	0,1	0,038461538	0,25	0,007225	110	0,3866171	240	0,344887348	350	0,447566	31,43	0,38	1	0,040160643	65	0,014581184	10,5
5799	17	0,050228311	0	0	0,25	0,007225	120	0,423791822	240	0,344887348	360	0,466292	33,33	0,41	1,8	0,072289157	135	0,030283997	10,5
5800	16	0,0456621	0,3	0,115384615	0,25	0,007225	120	0,423791822	240	0,344887348	360	0,466292	33,33	0,41	4,8	0,192771084	135	0,030283997	10,5
5801	19	0,059360731	0,1	0,038461538	0,25	0,007225	110	0,3866171	240	0,344887348	350	0,447566	31,43	0,38	1,9	0,076305221	65	0,014581184	10,5

Şekil 3.4 Verilerin normalizasyonu.

Ayrıca bazı VM algoritmaları kategorik verilerle çalıştığı için veriler referans aralıklarına göre “Normal”, “Low” ve “High” olarak etiketlenmiştir (Şekil 3.5).

3.2 Modelin Oluşturulması ve Uygulanması

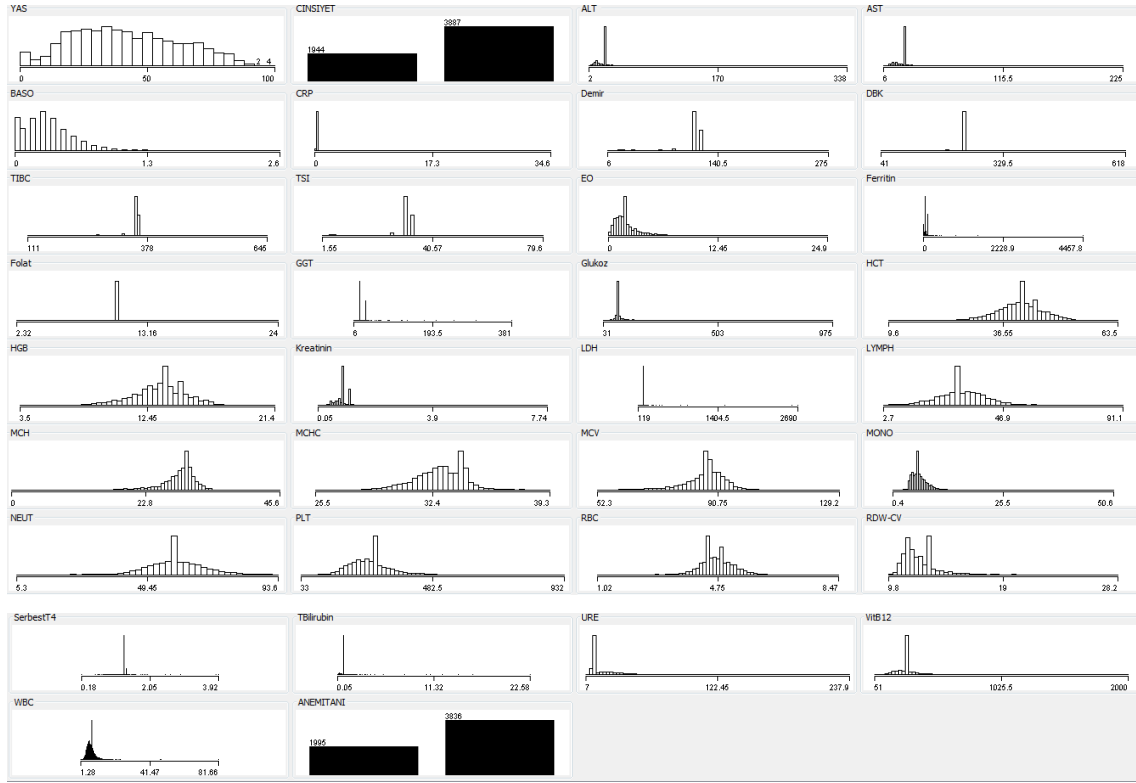
Veri temizleme ve dönüşüm işlemlerinden sonra ki adım modellemedir. VM büyük boyutlardaki verileri işlenmesi ve aralarındaki saklı örüntü ve eğilimlerin keşfedilmesini sağlayan bir süreç olup, bu süreçte farklı görevleri yerine getirmek için farklı algoritmalar kullanılmaktadır. Algoritmalar verileri inceleyerek en uygun modeli seçerler ki algoritmaların amacı da tam olarak budur (Özarslan ve Barışçı 2011). Farklı modeller veri kümeleri üzerinde denenerek en yüksek olan model seçilir.

WEKA veri modelleme algoritmalarından sınıflama algoritmaları problem için uygun formata çevrilmiştir. Teşhis belirlenirken eldeki tetkik değerlerine göre belirlenmesi gerekmektedir. WEKA programında tüm niteliklerin dağılımı Şekil 3.9'da görülmektedir. Bu 34 öznelik tekli veya grupta şeklinde modelden çıkartılarak en iyi başarımın 12 öznelik ile yapılabileceği bulunmuştur. Model oluşturulurken; verilerin belli bir kısmı eğitilmiş (training), eğitilen verilerin oluşturduğu örüntüler kullanılarak geri kalan veriler test edilmiştir. Çalışmada örüntü bulma ve test işlemleri WEKA sınıflandırma algoritmaları ile yapılmıştır. Algoritmaların tüm analizleri öncelikle use training set yöntemi ile yapılmış ve tüm veri setinin hem modeli eğitmek hemde test etmekte kullanıldığı için sonuçlar yanıltıcı olabileceğinden literatürde sıklıkla kullanılan Cross Validation yöntemi ile sınıflandırma analizleri tercih edilmiştir. Parçalara ayırma yöntemi olan cross validation yönteminde; veri seti 10 bölüme ayrılarak; her bölüm bir kez test verisi kalan diğer 9 bölüm ise öğrenme verisi olarak kullanılmıştır. Oluşturulan modeller karşılaştırılırken doğruluk (accuracy) değeri kullanılmıştır.

Verileri işlemede tahmin edilecek alan için doğruluk performansı yüksek olan algoritmalar kategorik ve orijinal veriler için ayrı ayrı elde edilmiştir.

Veriler işlenmeden önce uzman görüşü ile veri işlemede kullanılan giriş alanları olarak; YAS, CINSİYET, Demir, Ferritin, HCT, HGB, MCH, MCHC, MCV, RDW-CV, VitB12 Çıkış değeri olarak Anemi, AnemiDegil ve Makro, Mikro, Normo ve AnemiDegil şeklindedir. Anemi veri setinin nitelik dağılımı Şekil 3.9'da görülmektedir.

Sınıflandırma, kümeleme ve birliktelik analiz algoritmalarının performanslarını test etmek için tüm algoritmalar tek tek denenmiştir.



Şekil 3.9 Anemi veri setinin nitelik dağılımı.

Algoritmaların analizi aşamasında bazı algoritmalar sürekli, bazıları ise kategorik verilerle çalıştığından; bir algoritma; veri setinin orjinal hali, normalize edilmiş ve kategorik hale getirilmiş halleri ile sınıf etiketi Anemi, AnemiDegil veriseti ve AnemiTur veri seti için ayrı ayrı WEKA programında analiz edilerek, karşılaştırmada her veri seti kategorik ve orijinal veri değerleri kendi içinde değerlendirmeye alınmıştır.

Anemi için tasarlanan modeller veri seti üzerinde uygulanarak modellerin tahminleme başarı oranları değerlendirilerek sonuçlar gözlemlenmiştir.

3.3 Modelin Değerlendirilmesi ve Yorumlanması

WEKA’da sınıflandırma da hata hesabı (error calculation), kare ortalamalarının karekökü (RMSE, Root Means Square Error), tip1, tip2 hatalar ve F1- Score gibi bazı hata ölçüm yöntemleri bulunmaktadır (İnt.Kyn.17).

Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem olan basit geçerlilik testi (simple validation test) ve yukarıda bahsi geçen ölçüm yöntemleri

hakkında bölüm 2.3.10’de ayrıntılı bilgi verilmiştir. Örnek bir algoritmanın (trees.J48) Anemi ve AnemiDegil veriseti için veri modellemesi sonuç ekranı Şekil 3.10’da görülmektedir.

Çıkış ekranı örnek alınarak sınıflandırma algoritmasının doğruluk yüzdesi şu şekilde hesaplanmıştır:

8531 kayıt her seferinde 10 eşit parçaya bölünüp 9 parça eğitim ve kalan bir parça test için kullanılacağından toplamda bütün veriler bir kez hem modeli eğitmek, hem de test etmek için kullanılmış olacaktırlar.

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      5582           95.7297 %
Incorrectly Classified Instances    249           4.2703 %
Kappa statistic                    0.8979
Mean absolute error                 0.0744
Root mean squared error            0.199
Relative absolute error            17.2044 %
Root relative squared error        42.8092 %
Total Number of Instances         5831

=== Detailed Accuracy By Class ===

      TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
      0.872     0.003     0.992     0.872     0.928       0.947     Anemi
      0.997     0.128     0.944     0.997     0.97        0.947     AnemiDegil
Weighted Avg.   0.957     0.089     0.959     0.957     0.957       0.947

=== Confusion Matrix ===

      a    b  <-- classified as
1607  236 |    a = Anemi
  13 3975 |    b = AnemiDegil

```

Şekil 3.10 Verinin sınıflandırma algoritması ile modellenmesi.

Şekil 3.10’da sınıflandırıcı çıkış ekranında görülen Confusion matrix’e göre; gerçekte 1841 “Anemi” değerli verinin 1607’si “Anemi”, 236’sı “AnemiDegil” ve gerçekte 3988 “AnemiDegil” değerli verinin 13’ü “Anemi”, 3975’i “AnemiDegil” olarak tahmin edilmiştir. Buna göre toplam 236+13=249 veri yanlış sınıflandırılmış olup diğer 1607+3975=5582 veri doğru sınıflandırılmıştır. Bu bilgiler doğrultusunda bu modelin doğru sınıflandırma yüzdesi şu şekilde hesaplanmıştır:

Doğruluk yüzdesi= $(5582/5831)*100=95,7297\%$ dir.

Correctly Classified Instances değerine bakılarak sınıflandırıcı algoritmanın eğitim verilerini aynı zamanda test verisi olarak ta kullanıp yaptıklarını doğru ve yanlış olarak sınıflandırma oranını gösteren değerdir. Kappa değeri gözlemler arasındaki uyuşmanın büyüklüğünü nicel olarak gösterdiğinden, bir değerine yaklaştıkça uyum oranı artar. Veri setine uygulanan algoritmalar arasında en büyük olan değer kontrol edilmiştir. Algoritma karşılaştırmaları MAE (Mean Absolute Error, Ortalama Mutlak Hata) ve RMSE (Root Mean-Squared Error, Ortalama Karesel Hatanın Karekökü) değerlerine göre yapılmış olup en küçük değer baz alınmalıdır. RMSE her bir veri örneğinin gerçek değerleriyle tahmin edilen değeri arasındaki değer farkının alınması ile elde edilip değer ne kadar az ise hata o kadar azdır. Ayrıca doğruluk değeri (accuracy value) tahmin modeli için karşılaştırma amacıyla kullanılır. Ortalama mutlak hata standart hata olarak seçilip değeri en küçük olan teknik en iyi hata tahmini tekniği olarak seçilmiştir. Şekil 3.10'da sınıflandırıcı çıkış ekranında verilerin detaylandırılmış doğruluk tablosu (Detailed Accuracy By Class) görülmektedir.

TP Rate (Doğru Hassaslık Oranı); a sınıfı için $1607/(1607+236)=0,8719$

b sınıfı için $3975/(3975+13)=0,9967$

FP Rate (Yanlış Hassaslık Oranı); a sınıfı için $13/(13+3975)=0,003$

b sınıfı için $236/(1607+236)=0,128$

Precision (Hassaslık); a sınıfı için $1607/(1607+13)=0,9919$

b sınıfı için $3975/(3975+236)=0,9439$

Recall (anma) a sınıfı için $1607/(1607+236)=0,8719$

b sınıfı için $3975/(3975+13)=0,9967$

F-Score a sınıfı için $2*0,872*0,992/(0,872+0,992)=0,928$

b sınıfı için $2*0,997*0,944/(0,997+0,944)=0,97$

F1 skoru veya F skoru olarak adlandırılan F ölçütü ilgili sınıfın sınıflandırıcı algoritma tarafından ne derece doğru sınıflandırma tahmini yapabileceğimizi gösterir. Bu yüzden değer bir'e yaklaştıkça ilgili sınıfın algoritma tarafından öğrenilmesinin o derece başarılı olduğu anlamına gelir. Modellemeler için kappa, Correctly Classified Instances ve RMSE değerlerine göre en uygun algoritmanın da değerlendirilmesi yapılmıştır. F- değeri 0,5 ve üzeri olan sınıfların modellenebilir oldukları söylenebilir. En yüksek doğrulukla tahmin yapabilen WEKA sınıflandırma algoritmalarının öncelikle doğruluk yüzdelere bakılmış, eşit olan algoritmalarda ise RMSE değeri düşük olan algoritma seçilmiştir.

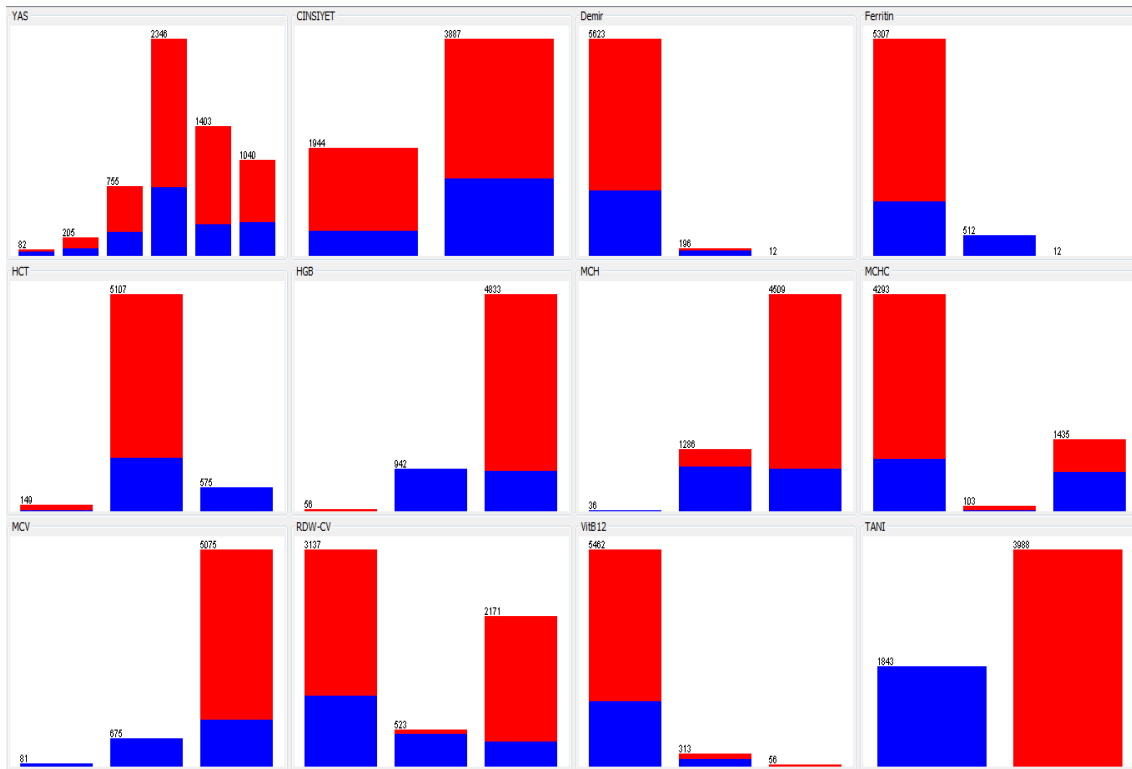
4. BULGULAR

Veri setinde başlangıçta 98 öznitelik ile yapılan analizler sonucunda uzman yardımı ile hastalığın teşhisinde betimleyici özelliğinin çok uzak olduğu düşünülen özellikler teker teker çıkarılarak 34 adet öznitelikten tetkik değeri olmayanlardan yaş ve cinsiyet harici özniteliklerde çıkartılarak kalan 28 adet öznitelik (YAS, CINSIYET, ALT, AST, CRP, Demir, DBK, TIBC, TSI, Ferritin, Folat, GGT, Glukoz, HCT, HGB, Kreatinin, LDH, MCH, MCHC, MCV, PLT, RBC, RDW-CV, Serbest T4, URE, VitB12, WBC) ile analizler yapılmıştır. 28 öznitelik ile analizlerin ortalama başarılarının %90-95 arası değiştiği görülmüştür. 28 özellik ile çalışmadan teker teker ya da grup olarak çıkarılan öznitelikler sonucunda 12 öznitelik (YAS, CINSIYET, Demir, Ferritin, HCT, HGB, MCH, MCHC, MCV, RDW-CV, VitB12,TANI) ile analizler gerçekleştirilmiştir ve doğru sınıflandırılan örnek yüzdesinin arttığı görülmüştür. Analizler için veriler; tetkik değerleri değiştirilmeden, max-min, z-score normalizasyon yapılarak ve kategorik hale getirilerek veri setinden 3 farklı varyasyon elde edilmiştir. Aynı veri setinin 3 farklı şekli için analizler yapılarak, aynı algoritma için en yüksek başarı oranına sahip olan algoritma seçilmiştir. Ayrıca aynı veri setinde ilk olarak sadece anemi olan ve olmayanlar şeklindeki sınıflandırma etiketi uzman yardımı ile aneminin morfolojik sınıflandırmasına (mikrositer, normositer ve makrositer) göre tekrar etiketlendirilmiştir.

Tüm veri grupları WEKA programı altında var olan tüm algoritmalar ile analiz edilmiştir. Modellerin oluşturulması aşamasında Anemi ve AnemiDegil veri seti ile anemi türlerini içeren veri setlerinde (AnemiTur) sınıflandırma algoritmalarından; bayes sınıflandırıcı algoritmaları (AODE, AODEsr, HNB, NaiveBayes, NaiveBayesSimple, NaiveByes.Updateable, WAODE, BayesianLogisticRegression, BayesNet) functions sınıflandırıcı algoritmaları (LibSVM, Logistic, MultilayerPerceptron, RBFNetwork, SimpleLogistic, SMO, SPegasos, VotedPerceptron, Winnow) lazy sınıflandırıcı algoritmaları (IB1, IBk, KStar, LBR, LWL) meta sınıflandırıcı algoritmaları (AdoBoostM1,AttributeSelectedClassifier, Bagging,ClassificationViaClustering, ClassificationViaRegression, Dagging, Decorate, CVParameterSelection, END, FilteredClassifier, Grading, LogitBoost, MultiBoostAB, MultiClassClassifier, MultiScheme, RotationForest, Stacking, nestedSichotomies.ClassBalancedND, nestedDichotomies.DataNearBalancedND, StackingC, nestedDichotomies.ND,

OrdinalClassClassifier, RacedIncrementalLogitBoost, RandomCommitee, RandomSubSpace, Vote, ThresholdSelector), misc sınıflandırma algoritmaları (HyperPipes, VFI), rules sınıflandırma algoritmaları (ConjuntiveRule, DecisionTable, DTNB, JRip, NNge, OneR, PART, Ridor, ZeroR), trees sınıflandırma algoritmaları (ADTree, Id3, BFTree, DesicionStump, FT, J48, j48graft, LADTree, LMT, NBTree, RandomForest, RandomTree, REPTree, SimpleCart) ile analizler yapılmıştır.

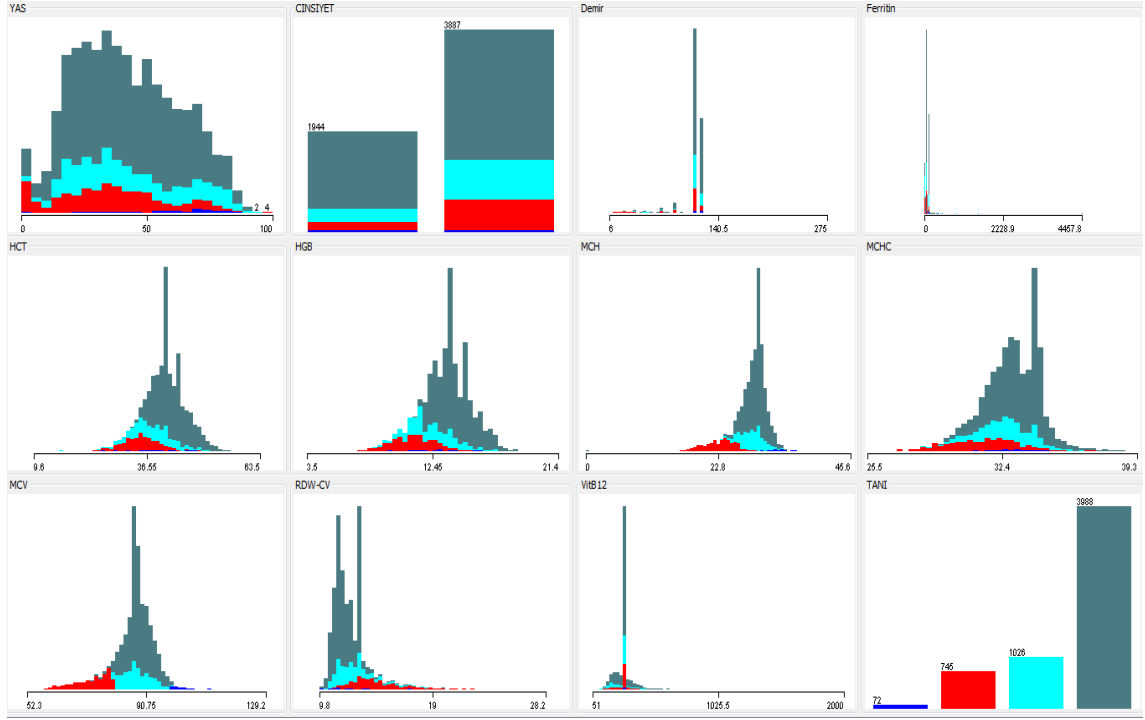
Verilerin özellik dağılımı Anemi ve AnemiDeğil için Şekil 4.1’de ve anemi sınıfları için özellik dağılımı Şekil 4.2’de görülmektedir.



Şekil 4.1 Verilerin kategorik hale çevrilmesi ile tüm özelliklerin dağılımı.

Model oluşturulma aşamasında kümeleme işlemleri için; DBSCAN, EM, FarthestFirst, MakeDensitBasedClusterer, FilteredClusterer ve SimpeKMeans kümeleme analizlerinden faydalanılmıştır.

Birliktelik kurallarından Apriori, FilteredAssociator ve Tertius algoritmaları ile analizler yapılmıştır (Ardıl 2009).



Şekil 4.2 Sınıflandırılmış anemi türlerine göre özelliklerin dağılımı.

AnemiTürleri veri seti için örnek bir model olarak trees.J48 algoritmasını inceleyecek olursak Şekil 4.3, Şekil 4.4 ve Şekil 4.5'te görülmektedir. 5587 adet veri doğru olarak sınıflandırılırken 244 adet veri ise yanlış sınıflandırıldığı Şekil 4.4'te görülmektedir. Karmaşıklık matrisinde ise doğru ve yanlış sınıflandırılan verilerin ayrıntıları verilmektedir. Örneğin bu modelde veri setimizde makrositer aneminin tamamı doğru olarak sınıflandırılmışken, 4 adet gerçekte mikrositer anemi olduğu halde sınıflandırıcı makrositer anemi sınıfında olacak şekilde yanlış sınıflandırmıştır. Aynı şekilde 17 adet veri Anemi olmadığı halde çeşitli anemi gruplarında sınıflandırıldığı görülmektedir.

Şekil 4.5'teki kurallar Şekil 4.3'te görülen kuralların ağaç şeklinde gösterilmiş şeklidir.

J48 pruned tree

```

-----

MCV <= 79.9
|   YAS <= 14
|   |   Ferritin <= 16.9: Mikro (62.0)
|   |   Ferritin > 16.9
|   |   |   HGB <= 10.9: Mikro (32.0)
|   |   |   HGB > 10.9
|   |   |   |   MCV <= 73.3: Mikro (22.0/3.0)
|   |   |   |   MCV > 73.3
|   |   |   |   |   YAS <= 6: AnemiDegil (48.0)
|   |   |   |   |   YAS > 6
|   |   |   |   |   |   HCT <= 37.8: Mikro (5.0)
|   |   |   |   |   |   HCT > 37.8: AnemiDegil (15.0/3.0)
|   |   |   |   YAS > 14: Mikro (624.0)
MCV > 79.9
|   MCV <= 98.9
|   |   HGB <= 11.6: Normo (433.0/2.0)
|   |   HGB > 11.6
|   |   |   Ferritin <= 9.9: Normo (204.0)
|   |   |   Ferritin > 9.9
|   |   |   |   VitB12 <= 199
|   |   |   |   |   YAS <= 29: Normo (91.0)
|   |   |   |   |   YAS > 29
|   |   |   |   |   |   MCHC <= 34
|   |   |   |   |   |   |   CINSIYET = E
|   |   |   |   |   |   |   |   HCT <= 40.5: Normo (10.0/1.0)
|   |   |   |   |   |   |   |   HCT > 40.5: AnemiDegil (66.0/5.0)
|   |   |   |   |   |   |   |   CINSIYET = K: AnemiDegil (118.0/14.0)
|   |   |   |   |   |   |   MCHC > 34: AnemiDegil (42.0)
|   |   |   |   |   VitB12 > 199
|   |   |   |   |   |   Ferritin <= 16.9
|   |   |   |   |   |   |   CINSIYET = E: Normo (21.0)
|   |   |   |   |   |   |   CINSIYET = K
|   |   |   |   |   |   |   |   YAS <= 14: Normo (12.0)
|   |   |   |   |   |   |   |   YAS > 14: AnemiDegil (105.0/2.0)
|   |   |   |   |   |   |   Ferritin > 16.9
|   |   |   |   |   |   |   |   YAS <= 30
|   |   |   |   |   |   |   |   YAS <= 8
|   |   |   |   |   |   |   |   |   MCV <= 85.9: AnemiDegil (41.0)
|   |   |   |   |   |   |   |   |   MCV > 85.9: Normo (8.0)
|   |   |   |   |   |   |   |   |   YAS > 8: AnemiDegil (1111.0/7.0)
|   |   |   |   |   |   |   |   YAS > 30
|   |   |   |   |   |   |   |   |   HGB <= 12.73
|   |   |   |   |   |   |   |   |   |   CINSIYET = E
|   |   |   |   |   |   |   |   |   |   |   HGB <= 12.5: Normo (34.0)
|   |   |   |   |   |   |   |   |   |   |   HGB > 12.5
|   |   |   |   |   |   |   |   |   |   |   |   YAS <= 67: Normo (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   YAS > 67: AnemiDegil (10.0/1.0)
|   |   |   |   |   |   |   |   |   |   CINSIYET = K: AnemiDegil (360.0/25.0)
|   |   |   |   |   |   |   |   |   HGB > 12.73: AnemiDegil (2280.0/158.0)
|   MCV > 98.9: Makro (72.0)

Number of Leaves :      26

Size of the tree :      51

```

Şekil 4.3 AnemiTur veri seti için J48 algoritması karar ağacı kuralları.

```

=== Stratified cross-validation ===
=== Summary ===

```

```

Correctly Classified Instances      5587           95.8155 %
Incorrectly Classified Instances    244           4.1845 %
Kappa statistic                     0.9103
Mean absolute error                 0.0371
Root mean squared error            0.1395
Relative absolute error             15.2792 %
Root relative squared error        40.0804 %
Total Number of Instances         5831

```

```

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	0.986	1	0.993	1	Makro
	0.995	0.002	0.989	0.995	0.992	1	Mikro
	0.783	0.002	0.99	0.783	0.874	0.933	Normo
	0.996	0.123	0.946	0.996	0.97	0.953	AnemiDegil
Weighted Avg.	0.958	0.085	0.96	0.958	0.956	0.956	

```

=== Confusion Matrix ===

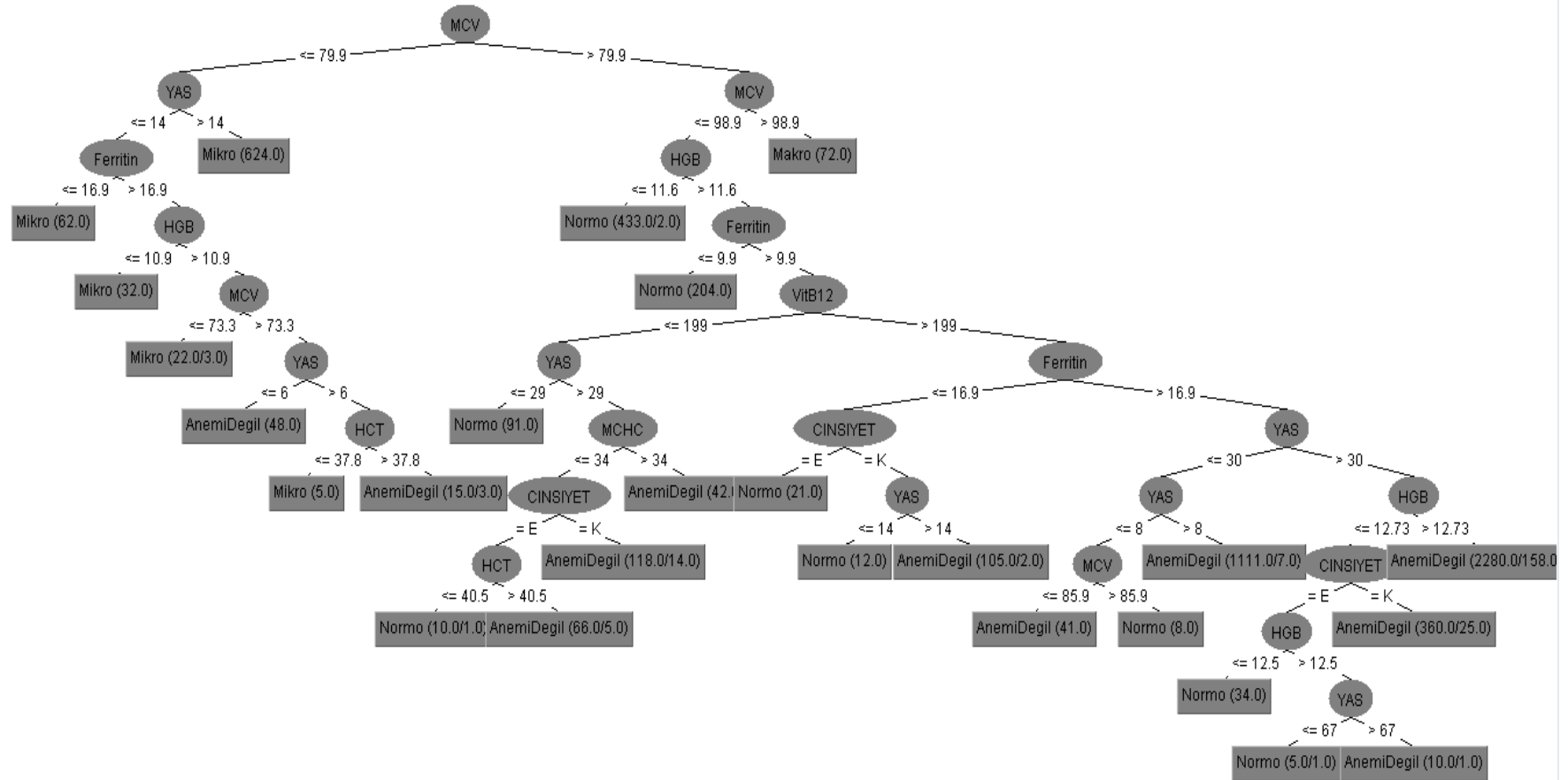
```

```

  a   b   c   d  <-- classified as
72   0   0   0 |   a = Makro
 0 741   0   4 |   b = Mikro
 0   0 803 223 |   c = Normo
 1   8   8 3971 |   d = AnemiDegil

```

Şekil 4.4 AnemiTur veri seti için J48 algoritmasının classifier output ekran çıktısı.



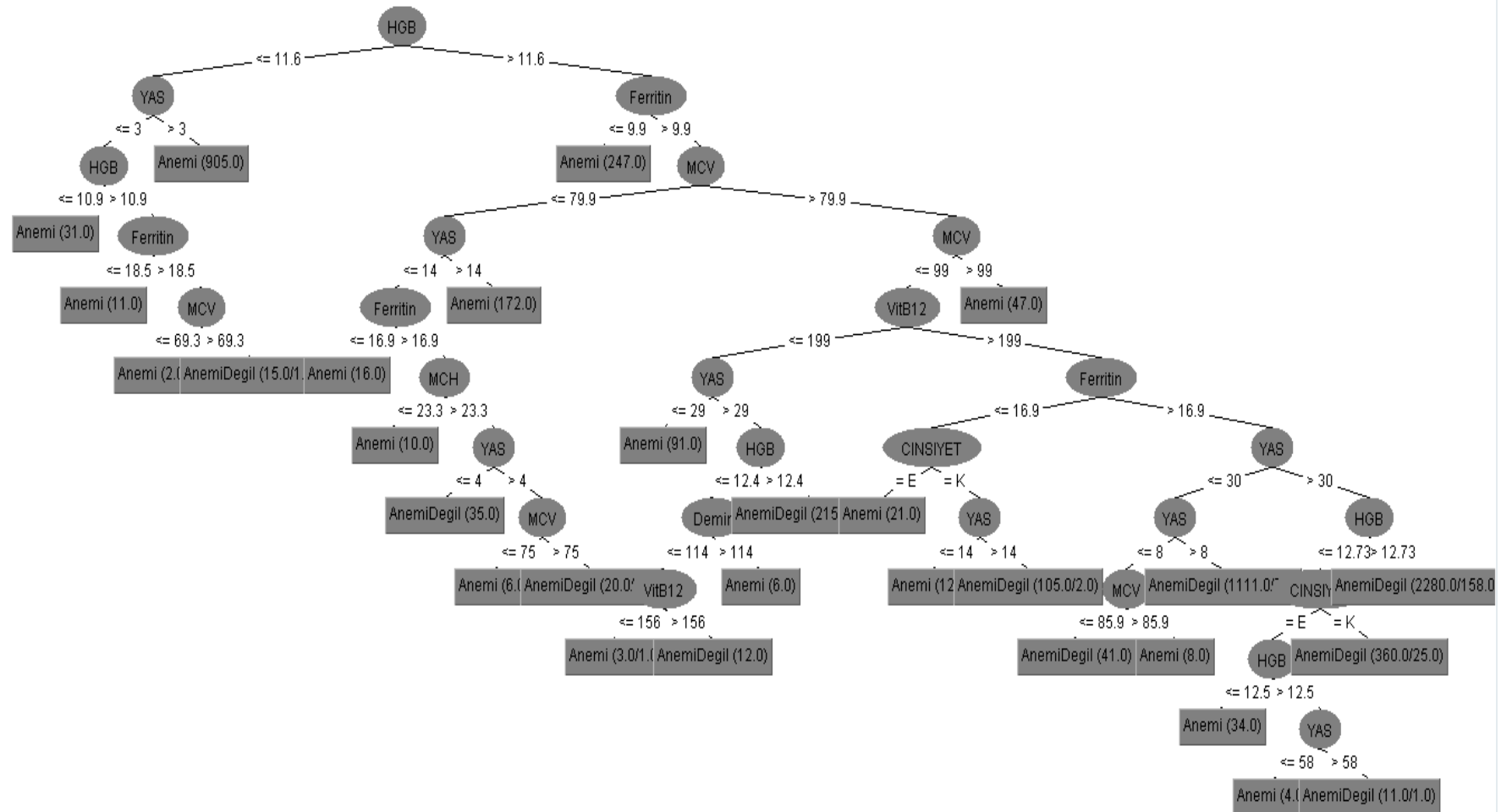
Şekil 4.5 AnemiTur veri seti için J48 algoritması karar ağacı.

Trees.J48 algoritması ile analiz sonucunda elde edilen kurallar şu şekildedir:

1. **Eğer** MCV 79,9 değerinden küçük ya da eşitse ve yaşı 14'e eşit ve küçükse ve ferritin değeri 16,9'a eşit ve az **ise Mikrositer Anemi**'dir.
2. **Eğer** MCV 79,9 değerinden küçük ya da eşitse ve yaşı 14'e eşit ve küçükse ve ferritin değeri 16,9'dan büyük ve hemoglobini 10,9'a eşit ve küçük **ise Mikrositer Anemi**'dir.
3. **Eğer** MCV 79,9 değerinden küçük ya da eşitse ve yaşı 14'den büyük **ise Mikrositer Anemi**'dir.
4. **Eğer** MCV 79,9 değerinden küçük ya da eşitse ve yaşı 14'e eşit ve küçükse ve ferritin değeri 16,9'dan büyük ve hemoglobini 10,9'dan büyük ve MCV'si 73,3'e eşit ve küçük **ise Mikrositer Anemi**'dir.
5. **Eğer** MCV 79,9 değerinden küçük ya da eşitse ve yaşı 14'e eşit ve küçükse ve ferritin değeri 16,9'dan büyük ve hemoglobini 10,9'dan büyük ve MCV'si 73,3'den büyük ve yaşı 6'ya eşit ve küçükse **AnemiDegil**'dir.
6. **Eğer** MCV 79,9 değerinden küçük ya da eşitse ve yaşı 14'e eşit ve küçükse ve ferritin değeri 16,9'dan büyük ve hemoglobini 10,9'dan büyük ve MCV'si 73,3'den büyük ve yaşı 6'dan büyükse ve hematokriti 37,8'den küçük **ise Mikrositer Anemi**'dir.
7. **Eğer** MCV 79,9 değerinden küçük ya da eşitse ve yaşı 14'e eşit ve küçükse ve ferritin değeri 16,9'dan büyük ve hemoglobini 10,9'dan büyük ve MCV'si 73,3'den büyük ve yaşı 6'dan büyükse ve hematokriti 37,8'den büyük **ise AnemiDegil**'dir.
8. **Eğer** MCV değeri 79,9'dan ve 98,9'dan büyük **ise Makrositer Anemi**'dir.
Makrositer anemi için MCV'nin 99'dan büyük olması gereklidir.
9. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit ve küçük olan ve Hemoglobini 11,6'ya eşit ve küçük **ise Normositer Anemi**'dir.
10. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit ve küçük olan ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'a eşit ve küçük **ise Normositer Anemi**'dir.
11. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit ve küçük olan ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'a eşit ve küçükse ve yaşı 29'a eşit ve küçük **ise Normositer Anemi**'dir.
12. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit ve küçük olan ve Hemoglobini

- 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'a eşit ve küçükse ve yaşı 29'dan büyük ve MCHC değeri 34'den büyük **ise AnemiDegil**'dir.
13. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit ve küçük olan ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'a eşit ve küçükse ve yaşı 29'dan büyük ve MCHC değeri 34'e eşit ve küçük ve Cinsiyeti Kadın **ise AnemiDegil**'dir.
14. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit ve küçük olan ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'a eşit ve küçükse ve yaşı 29'dan büyük ve MCHC değeri 34'e eşit ve küçük ve Cinsiyeti Erkek **ise ve Hematokrit değeri 40,5 değerine eşit ya da küçük ise Normositer Anemi**'dir.
15. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit ve küçük olan ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'a eşit ve küçükse ve yaşı 29'dan büyük ve MCHC değeri 34'e eşit ve küçük ve Cinsiyeti Erkek **ise ve Hematokrit değeri 40,5 değerinden büyük ise AnemiDegil**'dir.
16. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'a eşit küçük ve cinsiyeti erkek **ise Normositer Anemi**'dir.
17. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'a eşit küçük ve cinsiyeti kadın ve yaşı 14' eşit küçük **ise Normositer Anemi**'dir.
18. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'a eşit küçük ve cinsiyeti kadın ve yaşı 14'den büyük **ise AnemiDegil**'dir.
19. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'dan büyük ve yaş 30'a eşit küçük yaş 8'den büyük **ise AnemiDegil**'dir.

20. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'dan büyük ve yaş 30'a eşit küçük yaş 8'e eşit küçük ve MCV değeri 85,9'dan büyük **ise Normositer Anemi'dir.**
21. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'dan büyük ve yaş 30'a eşit küçük yaş 8'e eşit küçük ve MCV değeri 85,9'a eşit küçük **ise AnemiDegil'dir.**
22. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'dan büyük ve yaş 30'dan büyük ve hemoglobin değeri 12,73'den büyük **ise AnemiDegil'dir.**
23. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'dan büyük ve yaş 30'dan büyük ve hemoglobin değeri 12,73'e eşit küçük ve cinsiyet kadın **ise AnemiDegil'dir.**
24. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'dan büyük ve yaş 30'dan büyük ve hemoglobin değeri 12,73'e eşit küçük ve cinsiyet erkek **ise ve hemoglobin değeri 12,5'e eşit küçük ise Normositer Anemi'dir.**
25. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'dan büyük ve yaş 30'dan büyük ve hemoglobin değeri 12,73'e eşit küçük ve cinsiyet erkek **ise ve hemoglobin değeri 12,5'den büyük ve yaş 67'ye eşit küçük ise Normositer Anemi'dir.**
26. **Eğer** MCV değeri 79,9'dan büyük ve 98,9'a eşit küçük ve Hemoglobini 11,6'dan büyük ve ferritini 9,9'dan büyük ve Vitamin B12 değeri 199'dan büyük ve ferritini 16,9'dan büyük ve yaş 30'dan büyük ve hemoglobin değeri 12,73'e eşit küçük ve cinsiyet erkek **ise ve hemoglobin değeri 12,5'den büyük ve yaş 67'den büyük ise AnemiDegil'dir.**



Şekil 4.6 Anemi ve AnemiDegil veri seti için J48 algoritması karar ağacı.

Şekil 4.5 için yazılan karar ağacı kurallarına benzer şekilde, Şekil 4.6 karar ağacı kuralları da yazılabilir.

Uygulamada en iyi model elde edilinceye kadar WEKA’da tüm algoritma analizleri farklı varyasyonları ile başlangıçta 98 parametre sınıflandırmada etkin rol oynadığı düşünülen 34 parametreye uzman yardımı ile indirgenerek, sonrasında parametreler bazen tekli bazen grup halinde çıkartılarak kalan 12 parametre ile en başarılı algoritma bulunana kadar denenmiştir.

Uygulamada elde edilen sonuçlar sınıflandırma çıktısı (Classifier Output) ekranında görülen şekliyle sınıflandırma algoritmaları için; doğru olarak sınıflandırılan örnek yüzdesi, Kappa İstatistiği, Ortalama mutlak hata, ortalama hata karekökü, TP oranı, FP oranı ve F-ölçütü verilmiştir.

WEKA içerisinde bulunan makine öğrenme algoritmaları bayes, functions, lazy, meta, misc, trees olmak üzere altı temel sınıfta analizler gerçekleştirilmiştir.

Anemi ve AnemiDegil veri seti kategorik veriler (üst satır) ve orjinal tetkik değerleri (alt satır) ile gösterilmiştir. Çizelge 4.1’den Çizelge 4.7’ye kadar yapılan algoritma analizlerinde verilerin tetkik değerlerinin sürekli ve kategorik hale çevrilmesi ile yapılan Anemi ve AnemiDegil şeklindeki sınıflama algoritmaları kategorize edilerek değerlendirildiği görülmektedir.

Çizelge 4.1 Anemi ve AnemiDegil veri seti Bayes algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
AODE	92,91	0,8288	0,1081	0,2489	0,929	0,136	0,927
HNB	93,43	0,8417	0,123	0,2388	0,934	0,127	0,934
NaiveBayes	91,06	0,787	0,1088	0,2702	0,911	0,147	0,909
WAODE	93,02	0,8313	0,1112	0,247	0,93	0,134	0,928
BayesianLogisticRegression	73,14	0,2216	0,2686	0,5182	0,731	0,555	0,671
BayesNet	88,82	0,7368	0,1286	0,3123	0,888	0,164	0,887
NaiveBayes	84,38	0,6221	0,1633	0,3676	0,844	0,249	0,84
NaiveBayesSimple	84,31	0,6199	0,1667	0,3684	0,843	0,251	0,839
NaiveBayesUpdate	84,38	0,6221	0,1663	0,3676	0,844	0,249	0,84

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Bayes sınıflandırma algoritmaları içerisinde doğru olarak sınıflandırılan örnek yüzdesi %93,43 ile en başarılı olan algoritma kategorik veriler ile yapılan HNB algoritması olduğu Çizelge 4.1’de görülmektedir. Bayes sınıflayıcı algoritmaları bayes teoremine göre istatistiksel kestirim yaparlar. Bir örneğin sınıf üyelik olasılığını kestirir. Burada

amaç; birden fazla özellik taşıyan verilerden bir eğitim oluşturmak ve eğitim sonunda gelen yeni veriyi doğru bir şekilde sınıflandırmaktır.

Çizelge 4.2 Anemi ve AnemiDegil veri seti Functions algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
LibSVM	93,67	0,847	0,0633	0,2516	0,937	0,127	0,935
Logistic	93,57	0,8438	0,1073	0,2322	0,936	0,132	0,934
Multilayer Perceptron	93,84	0,852	0,891	0,2318	0,938	0,119	0,937
RBFNetwork	92,23	0,8134	0,1301	0,2567	0,922	0,139	0,921
SimpleLogistic	93,53	0,843	0,1086	0,2324	0,935	0,133	0,933
SMO /SPegasos	93,48	0,8413	0,652	0,2553	0,935	0,136	0,933
VotedPerceptron	93,62	0,8451	0,0638	0,2526	0,936	0,131	0,934
Winnow	85,58	0,6677	0,1442	0,3798	0,856	0,185	0,856
LibSVM	74,57	0,2592	0,2543	0,5043	0,746	0,54	0,686
Logistic	85,30	0,6387	0,2384	0,3405	0,853	0,252	0,848
Multilayer Perceptron	89,20	0,7341	0,154	0,2967	0,892	0,202	0,892
RBFNetwork	85,01	0,6396	0,239	0,3465	0,85	0,235	0,847
SimpleLogistic	85,42	0,6418	0,2414	0,3411	0,854	0,25	0,849
SMO /SPegasos	84,84	0,6205	0,1516	0,3894	0,848	0,275	0,841
VotedPerceptron	73,37	0,1693	0,2763	0,5255	0,724	0,593	0,643

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Functions sınıflandırma algoritmaları içerisinde % 93,84 ile en başarılı algoritmanın MultilayerPerceptron algoritması olduğu Çizelge 4.2’de görülmektedir. WEKA’da function sınıflandırma algoritmaları arasında YSA çok katmanlı algılayıcılar (Multilayer Perceptron); örnekleri sınıflandırmak için geri yayılım kullanan sınıflandırıcı (backpropagation) kullanan sınıflandırıcı içerir. Ağ görüntülenip eğitim zamanı süresince modifiye edilebilir. Ağdaki düğümler sigmoidtir (Bir şeyin haricinde; sınıf sayısalken çıkış düğümleri eşiksiz lineer birimler olur) (Ardıl 2009).

Çizelge 4.3 Anemi ve AnemiDegil veri seti misc algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
HyperPipes	31,47	0,0009	0,4924	0,4931	0,315	0,313	0,152
VFI	85,43	0,615	0,4306	0,448	0,854	0,314	0,84

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Misc sınıflandırma algoritmaları arasında en başarılı algoritmanın %58,43 ile VFI algoritması olduğu Çizelge 4.3’te görülmektedir.

VFI (Voting Feature Interval, Öznitelik Aralıklarını Oylama); her öznitelik sınıflandırmaya sınıflar arasında gerçek değerli oylar dağıtarak katılır. Yüksek oyu alan sınıf kestirimi yapılan sınıf olarak seçilir. Aralıklar her bir sınıf etrafında her bir

öznitelik için yapılandırılır ve sınıf ayarları her bir öznitelik üzerindeki her bir aralık için kaydedilir. Sınıflandırma voting ile yapılır (Demiröz ve Güven 1997).

Çizelge 4.4 Anemi ve AnemiDegil veri seti lazy algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
IB1	89,67	0,7599	0,1032	0,3213	0,897	0,141	0,896
IBk	93,03	0,8322	0,1042	0,2421	0,93	0,131	0,929
KStar	90,34	0,7571	0,1576	0,2747	0,903	0,207	0,898
LBR	94,12	0,8576	0,1054	0,2295	0,941	0,121	0,94
LWL	87,48	0,6762	0,2098	0,3212	0,875	0,27	0,865
IB1	85,58	0,5637	0,1842	0,4292	0,816	0,256	0,813
IBK	86,35	0,6598	0,1914	0,336	0,863	0,251	0,857
KStar	85,25	0,6378	0,174	0,3471	0,853	0,252	0,847
LWL	84,45	0,5873	0,2507	0,3555	0,844	0,332	0,828

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Lazy sınıflandırma algoritmaları içerisinde LBR algoritmasının %94,12 ile en başarılı sınıflandırıcı algoritması olduğu Çizelge 4.4’de görülmektedir. Naive bayes sınıflandırıcı öğrenme için basit ve etkili bir yaklaşım sunar. Bu metodun naive bayes algoritmasının geliştirmeye çalışan tekniklere göre temel avantajı kavramsal ve basit hesaplanabilir olmasıdır (Ardıl 2009).

Çizelge 4.5 Anemi ve AnemiDegil veri seti meta algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
AdoBoostM1	92,88	0,8253	0,1828	0,2819	0,929	0,151	0,926
AttributeSelectedClassifier	93,40	0,8388	0,1203	0,2456	0,934	0,139	0,932
Bagging	94,29	0,8618	0,0995	0,2237	0,943	0,119	0,941
Dagging	93,41	0,8394	0,0677	0,2438	0,934	0,138	0,932
FilteredClassifier	94,25	0,8609	0,1015	0,2266	0,943	0,12	0,941
LogitBoost	93,41	0,8393	0,1261	0,2368	0,934	0,138	0,932
MultiClassClassifier	93,57	0,8438	0,1073	0,2322	0,936	0,132	0,934
RacedIncrementalLogitBoos	93,45	0,8402	0,1442	0,2485	0,934	0,138	0,932
RotationForest	94,27	0,8614	0,1075	0,2263	0,943	0,12	0,941
ThresholdSelecector	93,64	0,8474	0,1134	0,2329	0,936	0,119	0,935
AdoBoostM1	87,65	0,6823	0,1938	0,3061	0,877	0,263	0,867
AttributeSelectedClassifiers	92,83	0,8264	0,1135	0,2504	0,928	0,139	0,926
Bagging	95,35	0,8889	0,826	0,2029	0,954	0,094	0,953
ClassificationViaRegression	93,16	0,8342	0,1287	0,2479	0,932	0,136	0,93
CVParameterSelection	68,39	0	0,4324	0,4649	0,684	0,684	0,556
Dagging	84,65	0,6108	0,1574	0,3725	0,847	0,289	0,837
Decorate	94,96	0,8797	0,1237	0,2276	0,95	0,097	0,949
END	95,78	0,8993	0,0732	0,1988	0,958	0,087	0,957
FilteredClassifier	93,11	0,8342	0,1106	0,2448	0,931	0,129	0,93
Grading	68,39	0	0,3161	0,5622	0,684	0,684	0,556
LogitBoost	91,84	0,8015	0,1568	0,2649	0,918	0,156	0,916
MultiBoostAB	84,43	0,5872	0,1556	0,3925	0,844	0,332	0,828
MultiClassClassifier	85,30	0,6387	0,2384	0,3405	0,853	0,252	0,848
MultiScheme	68,39	0	0,4324	0,4649	0,684	0,684	0,556
OrdinalClassClassifier	95,78	0,8993	0,0732	0,1988	0,958	0,087	0,957
RacedIncrementalLogitBoost	91,07	0,7841	0,1862	0,2844	0,911	0,16	0,908
RandomCommitee	94,62	0,8729	0,0893	0,2207	0,946	0,091	0,946
RandomSubSpace	93,65	0,8467	0,173	0,2529	0,937	0,126	0,935
RotationForest	93,16	0,836	0,1397	0,2453	0,932	0,125	0,93
Stacking/ StackingC/ Vote	68,39	0	0,4324	0,4649	0,684	0,684	0,556
ThresholdSelector	84,93	0,6407	0,2465	0,3437	0,849	0,228	0,847

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Meta sınıflandırma algoritmaları içerisinde en başarılı sınıflandırma algoritmasının %98,78 ile OrdinalClassClassifier ve END algoritmaları olduğu Çizelge 4.5’de görülmektedir.

OrdinalClassClassifier algoritmasında Sınıf değerlerinin çoğunlukla düzensiz olduğu varsayımı ile makine öğrenmesi metotları tanımlanmıştır. Sınıf değeri doğal sıralama E_g gösterilir. Örneğin derecelendirmeyi öğrenmek için sırasal (ordinal) sınıflandırmaya standart yaklaşım, sınıf değerini sayısal niceliğe dönüştürür ve dönüştürülen veriye regresyon öğrenici uygulayarak, son işleme adımında çıktıyı ayırık sınıf değerlerine çevirir (Ardıl 2009).

Çizelge 4.6 Anemi ve AnemiDegil veri seti için rules algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
DecisionTable	93,35	0,8877	0,1233	0,2461	0,933	0,139	0,931
DTNB	94,29	0,8619	0,1004	0,2284	0,943	0,118	0,941
NNge	91,01	0,7898	0,0899	0,2998	0,91	0,129	0,91
ConjunctiveRule	84,62	0,5887	0,2553	0,355	0,846	0,335	0,829
DecisionTable	92,59	0,8222	0,1433	0,2537	0,926	0,13	0,925
DTNB	91,63	0,7982	0,1157	0,2733	0,916	0,147	0,915
JRip	95,71	0,8969	0,0796	0,2011	0,957	0,091	0,956
NNge	84,42	0,6401	0,1558	0,3947	0,844	0,2	0,845
OneR	84,87	0,6001	0,1513	0,389	0,849	0,321	0,834
PART	95,97	0,9033	0,0757	0,1963	0,96	0,085	0,959
Ridor	95,56	0,8931	0,0444	0,2107	0,956	0,094	0,955
ZeroR	68,63	0	0,432	0,464	0,686	0,686	0,559

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Rules sınıflandırma algoritmaları içerisinde en başarılı algoritmanın %95,97 sınıflandırma doğruluk oranı PART algoritması olduğu Çizelge 4.6’da görülmektedir.

Çizelge 4.7 Anemi ve AnemiDegil veri seti için trees algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
Id3	93,28	0,8448	0,0975	0,2352	0,935	0,118	0,934
ADTree	94,31	0,8623	0,1976	0,2568	0,943	0,118	0,942
FT	94,25	0,861	0,0663	0,2311	0,943	0,119	0,941
J48	94,25	0,8609	0,1015	0,2266	0,943	0,12	0,841
RandomTree	93,31	0,8397	0,0999	0,2397	0,933	0,123	0,933
ADTree	94,20	0,8584	0,2046	0,2606	0,942	0,125	0,942
BFTree	95,92	0,9022	0,0716	0,1935	0,959	0,084	0,958
DecisionStump	84,62	0,5887	0,2548	0,3549	0,846	0,335	0,829
FT	94,20	0,8622	0,0627	0,2303	0,942	0,099	0,941
J48	95,8	0,8999	0,074	0,1996	0,958	0,085	0,957
j48graft	95,41	0,8902	0,0769	0,2095	0,954	0,09	0,953
LADTree	94,40	0,8661	0,1197	0,2208	0,944	0,103	0,943
LMT	95,21	0,8852	0,0811	0,2108	0,952	0,094	0,951
NBTree	94,40	0,866	0,098	0,2287	0,944	0,104	0,943
RandomForest	94,91	0,8785	0,1023	0,222	0,949	0,094	0,948
RandomTree	90,57	0,7807	0,0975	0,3077	0,906	0,126	0,906
REPTree	95,36	0,8885	0,0881	0,2077	0,954	0,095	0,953
SimpleCart	95,92	0,9017	0,0762	0,1947	0,959	0,089	0,958

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Trees sınıflandırma algoritmalarından en başarılı algoritma kategorik verilerle yapılan %95,92 doğruluk oranı, 0,9022 kappa değeri, 0,0716 ortalama mutlak hata, 0,1935 ortalama karesel mutlak hata oranı ile BFTree algoritması olduğu Çizelge 4.7’de görülmektedir. BFTree algoritmasında; ilk önce kök düğüme yerleştirilmek için bir öznelik seçilir ve dallanmalar yapılır. Sonra kök düğüme uzanan her bir dal için bir tane olmak üzere eğitim örnekleri alt kümelere bölünür (Ardıl 2009).

Çizelge 4.8 Anemi ve AnemiDegil veri seti için en başarılı algoritmalar.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
Bayes.HNB	93,43	0,8417	0,123	0,2388	0,934	0,127	0,934
Functions.MultilayerPerceptro	93,84	0,852	0,891	0,2318	0,938	0,119	0,937
Lazy.LBR	94,12	0,8576	0,1054	0,2295	0,941	0,121	0,94
Meta.END	95,78	0,8993	0,0732	0,1988	0,958	0,087	0,957
Meta.OrdinalClassClassifier	95,78	0,8993	0,0732	0,1988	0,958	0,087	0,957
Misc.VFI	85,43	0,615	0,4306	0,448	0,854	0,314	0,84
Rules.PART	95,97	0,9033	0,0757	0,1963	0,96	0,085	0,959
Trees.BFTree	95,92	0,9022	0,0716	0,1935	0,959	0,084	0,958

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Yapılan sınıflandırma algoritmalarından her sınıflandırma algoritmasını kendi içinde değerlendirilerek Anemi ve AnemiDegil veri seti için en başarılı algoritmanın %95,97 doğru sınıflandırma oranı ile rules.PART algoritması olduğu Çizelge 4.8’de görülmektedir.

Anemi ve AnemiDegil sınıfı ile yapılan kümeleme algoritmaları analizlerinden elde edilen sonuçlar ise; Çizelge 4.9’dan Çizelge 4.15’e kadar görülmektedir.

Çizelge 4.9 Anemi ve AnemiDegil veri seti DBSCAN kümeleme algoritması.

Test seti üzerinde model ve değerlendirme		
Küme numarası	Örnek sayısı	Yüzde (%)
0	455	8
1	1388	24
2	2498	43
3	2498	26

Normalde Anemi ve AnemiDegil şeklinde iki sınıf etiketine sahip veri kümesi için DBSCAN kümeleme algoritması veri setini özelliklerine göre 4 kümede toplamış ve bir örnek sınıflandıramamıştır.

Çizelge 4.10 Anemi ve AnemiDegil veri seti EM kümeleme algoritması.

Test veri seti üzerinde model ve değerlendirme		
Küme numarası	Sınıflandırılmış örnek sayısı	Yüzde (%)
0	176	3
1	338	6
2	400	7
3	1409	24
4	194	3
5	685	12
6	2242	38
7	387	7

EM algoritmasının ise veri setini 8 kümeye ayırdığı Çizelge 4.10’da görülmektedir.

Çizelge 4.11 Anemi ve AnemiDegil veri seti FarhestFirst kümeleme algoritması.

Küme ağırlık merkezleri	
Küme 0	23,0 K 110,0 65,0 43,2 13,7 28, 7 31,7 90,3 12,8 289,0 AnemiDegil
Küme 1	77,0 E 120,0 135,0 19,3 6,8 45,6 35,2 129,2 20,8 77,0 Anemi

Test veri seti üzerinde model ve değerlendirme		
Küme numarası	Sınıflandırılan küme sayısı	Yüzde (%)
0	5367	92
1	464	8

Veriler 5367 adet veri anemi değil ve 464 adet veri anemi olarak sınıflandırılmıştır.

Çizelge 4.12 Anemi ve AnemiDegil veri seti FilteredClusterer algoritması.

Test veri seti üzerinde model ve değerlendirme		
Küme numarası	Sınıflandırılmış örnek sayısı	Yüzde (%)
Küme 0	1843	32
Küme 1	3988	68

FilteredClusterer Kümeleme algoritması verileri %32’lik bir grup ve %68’lik olmak üzere ilgili özellik değerlerine göre Anemi ve AnemiDegil şeklinde kümelendiği Çizelge 4.12’de görülmektedir.

Çizelge 4.13 Anemi ve AnemiDegil veri seti MakeDensityBasedClusterer kümeleme algoritması.

Test verisi üzerinde model ve değerlendirme		
Küme numarası	Sınıflandırılmış örnek sayısı	Yüzde (%)
Küme 0	1884	32
Küme 1	3947	68

FilteredClusterer ve MakeDensityBasedClusterer kümeleme algoritmaları birbirlerine çok yakın şekilde verileri iki kümeye ayırmışlardır.

Çizelge 4.14 Anemi ve AnemiDegil veri seti için SimpleKMeans algoritması.

Test verisi üzerinde model ve değerlendirme		
Küme numarası	Sınıflandırılmış örnek sayısı	Yüzde (%)
Küme 0	1843	32
Küme 1	3988	68

SimpleKMeans algoritması verileri iki kümeye ayırdığı Çizelge 4.14’de görülmektedir.

Çizelge 4.15 Tüm Öznitelikler ile SimpleKMeans kümeleme algoritması.

Test veri seti üzerinde model ve değerlendirme		
Küme numarası	Sınıflandırılan örnekler	Yüzde (%)
0	1843	32
1	3988	62

Anemi türleri veri seti Anemi ve AnemiDegil veri seti için kullanılan sınıflandırma algoritmalarının tümü ile analiz edilmiştir. Analiz sonucunda elde edilen sonuçlar kategorik veriler (üst satırlar) ve orjinal tetkik değerleri (alt satırlar) olacak şekilde Çizelge 4.16’dan Çizelge 4.22’ye kadar görülmektedir.

Çizelge 4.16 Anemi Türleri veri seti Bayes sınıflandırma algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
AODE	92,83	0,8426	0,062	0,1758	0,928	0,136	0,923
AODEsr	92,93	0,8449	0,0611	0,1751	0,929	0,135	0,925
BayesNet	91,05	0,8047	0,0679	0,1939	0,91	0,14	0,904
HNB	93,26	0,8541	0,0628	0,1694	0,933	0,122	0,929
NaiveBayesSimple	91,05	0,8046	0,0682	0,1942	0,91	0,14	0,904
WAODE	92,76	0,8418	0,0617	0,1762	0,928	0,135	0,923
BayesNet	87,29	0,7369	0,077	0,2271	0,873	0,153	0,872
NaiveBayes	81,84	0,6179	0,1088	0,2744	0,818	0,21	0,811
NaiveBayesSimple	81,73	0,616	0,109	0,2749	0,817	0,21	0,81
NaiveBayesUpdate	81,84	0,6179	0,1088	0,2744	0,818	0,21	0,811

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Bayes sınıflandırma algoritmaları içerisinde en başarılı algoritmanın %93,26 başarı oranı ile HNB algoritması, doğru sınıflandırılan örnek yüzdesinin en az NaiveBayesSimple algoritması olduğu Çizelge 4.16’da görülmektedir.

Çizelge 4.17 Anemi Türleri veri seti Functions sınıflama algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
LibSVM	93,12	0,8494	0,0344	0,1854	0,931	0,133	0,931
Logistic	93,35	0,8541	0,546	0,1672	0,993	0,132	0,929
MultilayerPerceptron	92,93	0,846	0,0465	0,1778	0,929	0,124	0,925
RBFNetwork	91,80	0,8208	0,0674	0,1861	0,918	0,14	0,913
SimpleLogistic	93,40	0,8551	0,0552	0,1658	0,934	0,133	0,93
SMO	93,38	0,8544	0,2557	0,3205	0,934	0,135	0,929
LibSVM	72,49	0,1953	0,1375	0,3709	0,725	0,587	0,646
Logistic	86,97	0,7065	0,1089	0,2298	0,87	0,226	0,856
MultilayerPerceptron	89,93	0,7767	0,0735	0,2038	0,899	0,182	0,892
RBFNetwork	83,23	0,637	0,1271	0,2532	0,832	0,228	0,822
SimpleLogistic	87,07	0,707	0,1118	0,2305	0,871	0,232	0,856
SMO	82,88	0,5824	0,2654	0,3356	0,829	0,335	0,793

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Functions sınıflandırma algoritmaları içerisinde en başarılı algoritma %93,40 doğruluk oranı ile SimpleLogistic algoritması iken en başarısız algoritmanın LibSVM algoritması olduğu Çizelge 4.17’de görülmektedir. Simple Logistic; temele öğrenenler olarak basit regresyon modelleri ile LogitBoost, lojistik modelleri ayarlamak için kullanılarak, yenilemelerinin optimal sayısı çapraz sağlama ile yapılır (Ardıl 2009).

Çizelge 4.18 Anemi Türleri veri seti Lazy sınıflandırma algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
IB1	88,84	0,7675	0,0558	0,2363	0,888	0,134	0,888
IBk	92,40	0,835	0,0557	0,1774	0,924	0,129	0,92
KStar	89,57	0,7594	0,0863	0,2019	0,896	0,215	0,884
LBR	93,33	0,8541	0,0611	0,1738	0,933	0,126	0,929
LWL	80,47	0,5007	0,1451	0,274	0,805	0,394	0,73
IB1	79,54	0,566	0,1023	0,3198	0,795	0,245	0,792
IBk	79,90	0,5715	0,1022	0,3166	0,799	0,245	0,795
KStar	83,81	0,6379	0,0964	0,2565	0,838	0,246	0,828
LWL	80,08	0,4935	0,1552	0,2792	0,801	0,409	0,718

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Lazy sınıflandırma algoritmalarından en başarılı sınıflandırma algoritmasının %93,33 başarı oranı ile LBR algoritması olduğu Çizelge 4.18’de görülmektedir. Lazy sınıflandırmada; lokal olarak ağırlıklandırılmış öğrenme diğer algoritmalarla göre tercih sebebidir.

Çizelge 4.19 Anemi Türleri veri seti meta sınıflandırma algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
AttributeSelectedClassifier	92,75	0,8402	0,0662	0,1822	0,927	0,137	0,923
Dagging	93,12	0,8485	0,2568	0,3181	0,931	0,138	0,926
FilteredClassifier	93,98	0,8685	0,0529	0,1629	0,94	0,121	0,936
LogitBoost	93,05	0,8467	0,0625	0,1698	0,931	0,139	0,926
MultiClassClassifier	93,31	0,8525	0,3116	0,3611	0,933	0,138	0,928
RacedIncrementalLogitBoost	92,73	0,8396	0,0736	0,1802	0,927	0,14	0,922
RotationForest	94,01	0,8694	0,0571	0,1638	0,94	0,12	0,937
Vote	68,39	0	0,2425	0,3481	0,684	0,684	0,556
AdaBoostM1	80,08	0,4935	0,2461	0,314	0,801	0,409	0,718
AttributeSelectedClassifier	88,85	0,7486	0,094	0,2184	0,889	0,212	0,874
Bagging	95,35	0,8995	0,0412	0,1439	0,953	0,094	0,951
ClassificationViaRegression	94,25	0,8748	0,0602	0,1638	0,943	0,117	0,939
CVParameterSelection	68,39	0	0,2425	0,3481	0,684	0,684	0,556
Dagging	80,52	0,5044	0,2689	0,3384	0,805	0,393	0,753
Decorate	94,39	0,8809	0,0603	0,1619	0,944	0,091	0,942
END	95,78	0,9095	0,0374	0,1393	0,958	0,086	0,956
FilteredClassifier	93,38	0,8589	0,0556	0,1707	0,934	0,127	0,93
Grading	68,39	0	0,158	0,3975	0,684	0,684	0,556
LogitBoost	92,52	0,8362	0,0634	0,1764	0,925	0,143	0,92
MultiBoostAB	80,09	0,4935	0,103	0,3134	0,801	0,409	0,718
MultiClassClassifier	86,47	0,687	0,325	0,3774	0,865	0,254	0,844
MultiScheme	68,39	0	0,2425	0,3481	0,684	0,684	0,556
ClassBlancedND	95,61	0,9058	0,0374	0,143	0,956	0,087	0,954
DataNearBalancedND/ND	95,71	0,9079	0,037	0,1412	0,957	0,087	0,955
OrdinalClassClassifier	95,76	0,9092	0,0369	0,1404	0,958	0,086	0,956
RacedIncrementalLogitBoost	91,99	0,8247	0,0872	0,1938	0,92	0,148	0,914
RandomCommite	94,55	0,8844	0,047	0,1579	0,945	0,089	0,944
RandomSubSpace	93,50	0,8579	0,0977	0,1855	0,935	0,129	0,931
RotationForest	92,81	0,8448	0,0735	0,1762	0,928	0,124	0,925

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Meta sınıflandırma algoritmaları içerisinde en başarılı algoritmanın %95,78 doğruluk oranı ile END algoritması iken doğru sınıflandırılan örnek yüzdesi Vote algoritmasının en az olduğu Çizelge 4.19’da görülmektedir.

END algoritmasında ise çok sınıflı veri kümelerini iki sınıflı sınıflandırıcılarla Nested (İçiçe) ve Dischotomies (ikiye ayrılma) toplulukları yapılandırılır (Ardıl 2009).

Çizelge 4.20 Anemi Türleri veri seti misc sınıflandırma algoritma ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
HyperPipes	31,95	0,1785	0,3637	0,4204	0,319	0,148	0,207
VFI	84,41	0,6687	0,329	0,3905	0,844	0,205	0,834
HyperPipes	13,98	0,0926	0,3726	0,4303	0,14	0,04	0,072
VFI	85,41	0,6931	0,3169	0,3785	0,854	0,156	0,842

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Misc sınıflandırma algoritmaları içerisinde en başarılı algoritma %85,41 doğruluk oranı

ile VFI algoritmasıdır (Çizelge 4.20). HyperPipes; her bir kategori için o kategorinin tüm noktalarını içeren HyperPipes yapılandırılarak; test örnekleri “en çok örnek içerir” kategorisine göre sınıflandırılır. Test durumlarında sayısı sınıf veya kayıp değerleri ele alınmaz (Ardıl 2009).

Çizelge 4.21 Anemi Türleri veri seti rules sınıflandırma algoritmaları ve sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
DecisionTable	93,29	0,8527	0,0706	0,1749	0,933	0,132	0,929
DTNB	93,48	0,8572	0,067	0,1764	0,935	0,128	0,931
NNge	90,36	0,7997	0,0482	0,2195	0,904	0,118	0,903
OneR	81,08	0,5147	0,0946	0,3075	0,811	0,405	0,734
ConjunctiveRule	80,08	0,4935	0,1598	0,2825	0,801	0,409	0,718
DecisionTable	91,84	0,8228	0,0911	0,195	0,918	0,14	0,913
DTNB	95,55	0,7964	0,0782	0,2024	0,906	0,152	0,901
JRip	95,40	0,9011	0,0411	0,147	0,954	0,092	0,952
NNge	78,11	0,5933	0,1094	0,3308	0,781	0,155	0,794
OneR	80,64	0,5288	0,0968	0,3111	0,806	0,372	0,742
PART	95,28	0,8987	0,0387	0,1484	0,953	0,092	0,951
Ridor	95,15	0,8953	0,0243	0,1558	0,951	0,097	0,949
ZeroR	68,39	0	0,2425	0,3481	0,684	0,684	0,556

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Rules sınıflandırma algoritmaları arasında en başarılı algoritmanın %95,55 değeri ile DTNB algoritması ve doğru sınıflandırma yüzdesi en az olan algoritmanın ZeroR algoritması olduğu Çizelge 4.21’de görülmektedir.

Çizelge 4.22 Anemi Türleri veri seti trees sınıfı algoritmaları ve sonuçları

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
Id3	93,00	0,8562	0,0496	0,1686	0,933	0,114	0,931
RandomTree	92,95	0,8484	0,0517	0,1741	0,93	0,117	0,927
BFTree	95,10	0,8951	0,0397	0,1508	0,951	0,091	0,949
DecisionStump	80,08	0,4935	0,1596	0,2825	0,801	0,409	0,718
FT	95,58	0,8837	0,0291	0,1606	0,946	0,1	0,944
J48	95,82	0,9103	0,0371	0,1395	0,958	0,085	0,956
J48graft	95,73	0,9084	0,0367	0,1414	0,957	0,086	0,955
LADTree	94,13	0,8723	0,0577	0,1638	0,941	0,118	0,938
LMT	95,40	0,9011	0,0389	0,145	0,954	0,093	0,952
NBTree	94,36	0,8783	0,0503	0,1607	0,944	0,103	0,941
RandomForest	94,79	0,8888	0,0497	0,1585	0,948	0,092	0,946
RandomTree	90,70	0,8068	0,0478	0,2151	0,907	0,114	0,907
REPTree	95,37	0,9004	0,0407	0,1464	0,954	0,093	0,952
SimpleCart	95,40	0,9012	0,0402	0,1461	0,954	0,092	0,952

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü

Trees sınıflandırma algoritmaları içerisinde en başarılı olan algoritma %95,82 başarı oranı ile J48 algoritması iken, doğru sınıflandırılan örnek yüzdesinin en az olduğu algoritmanın DecisionStump algoritması olduğu Çizelge 4.22’de görülmektedir.

J48 algoritması; ağacı yapılandırmak için kullanılır, eksik öznitelik gibi çeşitli problemler ele alır. Karar ağacı algoritmaları durumlar veya örnekler kümesi ile başlayıp yeni durumları sınıflandırmak için kullanılan ağaç veri yapısı oluşturulur. Her durum sayısal veya sembolik değer alabilen öznitelikler kümesi ile tanımlanır. Her bir eğitim durumu sınıfın adıyla ifade edilen etikettir (Ardıl 2009).

Anemi türleri veri seti için yapılan sınıflandırma algoritmaları sonuçlarında ise en iyi sınıflandırma yapan algoritma %95,82 doğruluk oranı ile karar ağaçları algoritmalarından trees.J48 olduğu Çizelge 4.23’de görülmektedir.

Çizelge 4.23 Anemi Türleri veri seti sınıflandırma sonuçları.

Algoritmalar	DSY	Kappa	MAE	RMSE	TP	FP	F-ölç.
Bayes.AODEsr	92,93	0,8449	0,0611	0,1751	0,929	0,135	0,25
Functions.SimpleLogistic	93,40	0,8551	0,0552	0,1658	0,934	0,133	0,93
Lazy.LBR	93,33	0,8541	0,0611	0,1738	0,933	0,126	0,929
Meta.END	95,78	0,9095	0,0374	0,1393	0,958	0,086	0,956
Misc.VFI	85,41	0,6931	0,3169	0,3785	0,854	0,156	0,842
Rules.DTNB	95,55	0,7964	0,0782	0,2024	0,906	0,152	0,901
Trees.J48	95,82	0,9103	0,0371	0,1395	0,958	0,085	0,956

Kappa: kappa istatistiği, TP: TP oran, FP: FP oran, F-ölç: F ölçütü.

Anemi türleri veri setinde yapılan kümeleme analizine DBSCAN kümeleme algoritmasında veri seti 8 adet kümeye ayrılmıştır (Çizelge 4.23).

Çizelge 4.24 Anemi Türleri veri seti EM kümeleme algoritması sonucu

Test veri seti üzerinde model ve değerlendirme		
Küme numarası	Sınıflandırılmış örnek sayısı	Yüzde (%)
0	2932	50
1	1572	27
2	342	6
3	985	17

EM Kümeleme algoritması ile veri kümesi 4 kümeye ayrılmıştır. FarthestFirst Algoritması ile yapılan analiz sonucunda ise veriler 5586 (%96) ve 245 (%4) olacak şekilde 2 kümeye ayırdığı Çizelge 4.24’de görülmektedir.

FilteredClusterer Kümeleme analizi ile veri seti 1944 (%33) ve 3887 (%67) olacak şekilde 2 kümeye ayrılmıştır. Aynı şekilde MakeDensityBasedCluster ve SimpleKMeans kümeleme analizlerinde veriler iki küme olacak şekilde %33 ve %67 olarak kümelemiştir.

Yapılan analizlerde elde edilen birliktelik kuralları için; Apriori, FilteredAssociater, Tertius, Predicitive Apriori algoritmaları analiz sonuçları kurallar şeklinde verimiştir.

Apriori algoritması ile elde edilen en iyi kurallar aşağıdaki şekildedir.

1. **Eğer** Ferritin ve MCV değerleri normal **ise** %99 olasılıkla demirde normaldir.
2. **Eğer** Hemoglobin normal **ise** Demir %98 olasılıkla normaldir.
3. **Eğer** Ferritin ve vitamin B12 normal **ise** demir %98 olasılıkla normaldir.
4. **Eğer** Ferritin normal **ise** %98 olasılıkla demirde normaldir.
5. **Eğer** MCV normal **ise** %98 olasılıkla demirde normaldir.
6. **Eğer** Hematokrit normal **ise** demirde %98 olasılıkla normaldir.
7. **Eğer** Hemoglobin değeri normal **ise** hematokrit değeri %97 olasılıkla normaldir.
8. **Eğer** Vitamin B12 değeri normal **ise** %97 olasılıkla demirde normaldir.
9. **Eğer** Demir ve ferritin normal **ise** %94 olasılıkla vitamin B12 normaldir.
10. **Eğer** Ferritin normal **ise** vitaminB12 %94 olasılıkla normaldir.
11. **Eğer** Demir ve MCV normal **ise** %94 olasılıkla demirde normaldir.
12. **Eğer** Demir normal **ise** vitamin B12 %94 olasılıkla normaldir.
13. **Eğer** MCV normal **ise** %94 olasılıkla vitamin B12’de normaldir.
14. **Eğer** Hematokrit normal **ise** %94 olasılıkla vitamin B12 normaldir.
15. **Eğer** MCV Normal **ise** %93 olasılıkla ferritinde normaldir.

FilteredAssociator birliktelik analiz sonuçları apriori algoritmasının ilk on kuralı ile aynı olduğu görülmüştür.

Tertius birliktelik algoritma birliktelik kuralları ise aşağıda sıralanmıştır.

1. **Eğer** Ferritin, hemoglobin ve MCV normal **ise** TANI AnemiDegil’dir.
2. **Eğer** TANI Anemi **ise** Ferritin, hemoglobin veya MCV düşüktür.
3. **Eğer** TANI Anemi **ise** Ferritin, hemoglobin veya MCH düşüktür.
4. **Eğer** Ferritin, Hemoglobin ve MCH normal **ise** TANI AnemiDegil’dir.
5. **Eğer** TANI Anemi **ise** Ferritin, hemoglobin veya vitamin B12 düşüktür.
6. **Eğer** TANI Anemi **ise** Ferritin veya hemoglobin düşük veya MCV yüksektir.
7. **Eğer** TANI Anemi **ise** Ferritin veya Hemoglobin Düşük veya RDW-CV yüksektir.

8. **Eğer** Ferritin, Hemoglobin ve vitamin B12 normal **ise** TANI AnemiDegil'dir.
9. **Eğer** Ferritin, hematokrit ve MCV normal **ise** TANI AnemiDegil'dir.
10. **Eğer** TANI Anemi **ise** Ferritin, hematokrit veya MCV Düşüktür.

5. TARTIŞMA VE SONUÇ

Anemi; halk arasında bilinen adıyla “kansızlık” eritrosit (kırmızı kan hücresi/alyuvar) sayısının azalması veya eritrosit içindeki hemoglobin miktarının azalması veya her ikisinin birlikte olması sonucu oluşan ülkemizde ve dünyada önemli bir halk sağlığı sorunudur. Kan değerlerinin yer aldığı laboratuvar incelemesi ile HGB, HCT, RBC, MCV, RDW gibi değerlere bakılarak anemi tespiti yapılır. Sonraki aşama MCV kullanılarak yapılan eritrosit morfolojisinden de yararlanılarak ayırıcı tanıya gidilir. Tedaviye başlamadan önce altta yatan durumun tespiti ve bu durumun ortadan kaldırılması önemlidir. Koruyucu önlemlerin alınması, geniş ölçekli çalışmaların yapılması gereklidir (İnt.Kyn.20).

Çalışmada Amasya ilinde faaliyet gösteren bir kamu hastanesinden hastaların kişisel mahremiyetleri gözetilerek alınan 6 aylık hasta verileri kullanılmış olup, ön tanısı anemi ve anemi ile benzerlik gösteren tanılara sahip 0-100 yaş aralığında 1944 erkek ve 3887 kadın olmak üzere toplam 5831 hasta kaydı üzerinde anemi teşhisinde etkin rol oynadığı düşünülen parametrelerle WEKA açık kaynak VM yazılımında analizler gerçekleştirilmiştir.

Yapılan sınıflandırma algoritmaları; Anemi ve AnemiDegil veri setinin sürekli ve kategorik verilerle yapılan analizleri sonucunda, en başarılı algoritmanın %95,97 başarı oranı ile kural tabanlı algoritmalarından rules.PART algoritması olduğu görülmüştür.

Anemi türleri veri seti (AnemTur) için sürekli ve kategorik varyasyonlar ile yapılan analiz sonuçlarında en başarılı algoritmanın ise %95,82 doğruluk oranı ile karar ağaçlarında trees.J48 algoritması olduğu görülmüştür.

Anemi ve AnemiDegil veri seti için kök düğüm hemoglobinin 11,6 değerinden küçük eşit ya da büyük olması iken, anemi türleri için kök düğüm MCV değerinin 79,9 değerinden küçük eşit ya da büyük olması şeklinde tanımlanmaktadır. Bu tanımların teşhiste öncelikle bakılan parametrelerle uyduğu görülmektedir.

Anemi ve AnemiDegil etiketlerine sahip veri kümesinde kümeleme algoritmalarından elde edilen bulgular incelendiğinde; normalde Anemi ve AnemiDegil şeklinde iki sınıf etiketine sahip veri kümesi için DBSCAN kümeleme algoritması veri setini

özelliklerine göre 4 kümede toplamıştır. Bu durum; aneminin alt sınıfları olan makrositer, normositer, mikrositer ve Degil şeklinde sınıflandırılmış olabileceği düşündürmektedir.

FarthestFirst Kümeleme Algoritmasında veriler 5367 adet veri anemi değil olarak ve 464 adet veri anemi olarak sınıflandırıldığı için bizim veri setimizle uyuşmadığı görülmüştür. Buna karşın FilteredClusterer Kümeleme algoritması veri setimizin orijinal hali ile %100 uyumlu şekilde 1843 adet anemi ve 3988 adet AnemiDegil olmak üzere iki kümeye ayırmıştır. FilteredClusterer ve MakeDensityBasedClusterer kümeleme algoritmaları birbirlerine çok yakın şekilde verileri iki kümeye ayırmışlardır. MakeDensityBasedClusterer Kümeleme Algoritması veri setini gerçeğe çok yakın bir şekilde kümelere ayırmıştır. SimpleKMeans Kümeleme Algoritması ise veri setimizi 1843 adet Anemi ve 3988 adet AnemiDegil şeklinde sınıf etiketleri ile %100 uyumlu olarak iki kümeye ayırmıştır. Hiçbir parametre çıkarılmadan yapılan SimpleKMeans Kümeleme Algoritması analizi sonucu yine aynı oranda çıktığı görülmüştür.

Anemi türleri veri seti için yapılan kümeleme algoritmalarının analiz sonuçlarına baktığımızda ise; tıpkı Anemi ve AnemiDegil veri setinde olduğu gibi, FilteredClusterer Kümeleme analizi ile veri seti 1944 (%33) ve 3887 (%67) olacak şekilde 2 kümeye ayrılmıştır. Aynı şekilde MakeDensityBasedCluster ve SimpleKMeans kümeleme analizlerinde veriler iki küme olacak şekilde %33 ve %67 olarak kümelemiştir. FarthestFirst Algoritması ile yapılan analiz sonucunda ise veriler 5586 (%96) ve 245 (%4) olacak şekilde 2 kümeye ayrılmıştır. Buna karşın EM Kümeleme algoritması ile veri kümesi 4 kümeye ayrılmıştır. Bu orijinal veri setimizle daha çok örtüşmektedir.

Yapılan birliktelik analizi algoritma sonuçları incelendiğinde tespit edilen bulgulara ilişkin elde edilen kuralların uzman eşliğinde değerlendirmesi aşağıdaki şekilde yapılmıştır.

1. Ferritin ve MCV normalse demire normal diyebiliriz.
2. Hemoglobin normalken demir düşük te olabilir. Önce demir deposu ferritin azalır sonra serum demiri azalır ve en sonra hemoglobin etkilenir. Ve düşer dolayısıyla hemoglobin normalken serum demiri tedavi sürecine ve bazı hastalıklara bağlı olarak düşük, normal veya yüksek olabilir.

3. Ferritin normalken vitamin B12'den bağımsız olarak serum demiri normal bulunur.
4. Ferritin normal ise demire normal diyebiliriz.
5. MCV normalken normositer anamide demir düşük veya normal olabilir.
6. Hemotokrit normalken, demir aneminin başlangıcına, türüne ve tedavi gibi değişkenlere bağlı olarak normal, düşük veya yüksek olabilir.
7. Hemoglobin değeri normalken hematokrit değeride normaldir (Ağır dehidratasyon veya sıvı yükleme tedavileri haricinde).
8. Bu kural doğru değildir, çünkü demir vitamin B12'den bağımsız olarak yüksek normal veya düşük olabilir. Vitamin B12 normalken demir bağımsızdır. Sağlıklı bir bireyde ikisinin normal olması beklenir. Düşüklük veya yükseklikleri birbirinden bağımsızdır.
9. Elde edilen bu kural, kural 8'deki gibi demir ve vitamin B12 değerlerinin düşüklük veya yükseklikleri birbirinde bağımsızdır.
10. Sağlıklı bir bireyde ferritin ve vitamin B12 değerlerinin ikisinin normal olması beklenir. Düşüklük veya yükseklikleri birbirinden bağımsızdır.
11. Demir, MCV normalken ferritin düşük veya normal olabilir. Depo demiri olarak ferritin anemi başlangıcında ilk düşen parametredir.
12. Sağlıklı bir bireyde demir ve vitamin B12 değerlerinin ikisinin normal olması beklenir. Düşüklük veya yükseklikleri birbirinden bağımsızdır.
13. Bu kural için doğru diyemeyiz çünkü MCV yüksekken büyük olasılıkla B12 düşük bulunur (Makrositer Anemi). Şu durumda hariç tedavi sürecinde olan ve B12 replasman tedavisi alan makrositer anemilerde MCV düşüşü görülmeden önce B12 vitamin yüksekliği bulunabilir. MCV normalken B12 normal olması yüksek olasılıktır. Şu durumda hariç miks tip anemilerde hem B12 hemde demir eksikliği varken MCV normal ama B12 düşük olabilmektedir.
14. Bu kural doğru kabul edilebilir. Sağlıklı bir bireyde Hematokrit ve vitamin B12 değerlerinin ikisinin normal olması beklenir. Hematokrit normalken B12'de normalse birey sağlıklıdır. HCT normalken B12 düşük ise B12 eksikliği var, henüz anemi tablosu yoktur (başlangıç). HCT normalken B12 yüksek ise replasman tedavisi almış ve anemi tablosu çözülmüş hasta, HCT düşüken B12'de düşükse makrositer anemi veya miks tip anemidir. HCT düşüken B12

normal veya yüksek ise mikrositer veya normositer anemi durumlarında görülebilmektedir. Dolayısı ile tek başına HCT değerlerinden B12 çıkarımı MCV ve HGB gibi ek parametreler görülmeden yapılmaması daha uygundur.

- 15. MCV Normal ise kesin olarak ferritinde normaldir diyemeyiz** çünkü MCV normalken ferritin düşük veya normal olabilir. Depo demiri olarak ferritin anemi başlangıcında ilk düşen parametredir. Örneğin MCV normal (Normositer anemi) ferritin düşük veya normal olabilir. Veya anemi tablosu henüz başlamamış demir eksikliği durumlarında MCV normalken ferritin düşük olabilmektedir.

FilteredAssociator birliktelik analiz sonuçları apriori ile benzer olduğunda yorumları apriori ile paraleldir. Tertius birliktelik algoritmalarının tümü uzman görüşü ile doğrulanmıştır.

Trees.J48 algoritması ile elde edilen sonuçlar uzman hekim eşliğinde incelenerek karar ağacından çıkan kurallar şu şekilde değerlendirilmiştir diğer kurallar ise olduğu şekli ile doğru kabul edilmiştir.

6. ve 7. Kural için; hematokriti anemi tanısında tek başına kullanamadığımız bu çıkarımların kesin olarak doğru ya da yanlış olduğunu söyleyemeyiz.

10. Kural için; Ferritin düşük ve hemoglobin normal iken demir eksikliği (mikrositer) anemi başlangıcında olup henüz hemoglobin ve MCV'yi etkileyen bir anemi tablosu içinde olmayıp demir eksikliği tanısı koymak daha doğru bir yaklaşımdır.

11. Kural için; Vitamin B12 değerinin 200'den küçük olması B12 eksikliğinin başladığını fakat henüz kan tablosuna yansımadığını görürüz. Bu tip hastalar vitamin B12 eksikliğinde görülecek olan makrositer anemi başlangıcında olup henüz MCV ve HGB'nin etkilenmediği ilk evrede olup şu durumda ancak Anemi Değil denilebilir.

12. Kural için; MCHC'den bağımsız olarak bu tip hastalar B12 vitamini eksikliği olan ama henüz anemi tablosu olmayan bir hastadır. Yani Anemi Değil'dir.

13. Kural için; MCHC'den ve cinsiyetten bağımsız olarak bu tip hastalarda B12 eksikliği olan ama henüz anemi tablosu olmayan bir hasta olup Anemi Değil'dir.

14. Kural için; MCHC'den, cinsiyetten ve HCT'den bağımsız olarak, vitamin B12 eksikliğinin başladığı fakat henüz kan tablosuna yansımadağı için makrositer aneminin başlangıç aşamasında olan hastada henüz MCV ve HGB etkilenmemiş yani ilk evre de olup hasta henüz anemi tablosunda olmayan bir hastadır. Yani Anemi Değil'dir.

15. Kural için; yaş, cinsiyet, MCHC ve HCT sonucu olmadan bu hasta anemi değildir, sadece vitamin B12 eksikliği olan bir hastadır. Ama henüz anemi değildir.

16. Kural için; Ferritin değeri düşük ve hemoglobin normal sınırdan ise demir eksikliği (Mikrositer) anemi başlangıcında olup henüz HGB ve MCV'yi etkileyen bir anemi tablosu içinde olmayıp demir eksikliği tanısı koymak daha doğru olacaktır.

17. Kural için; bu kural için anemi değil tanısını yapabiliriz.

20. Kural için; tedavi görmüş olan makrositer anemi denilebilir.

Bu tez kapsamında yapılan çalışmada; anemi tanısı ve benzer semptomlara sahip hastalıklarda dâhil edilerek hastanedeki mevcut tüm servislerden elde edilen 5831 veri ile çalışılmıştır. Bunlardan 1843 adedi anemi ve 3988 adedi değil şeklinde olup, anemi teşhisi için başlangıçta 98 olan parametre sayısı uzman görüşü ile 34 adet parametre olarak belirlenmiş ve analiz süresince parametreler teker teker veya gruplar halinde çıkartılarak en iyi sonucu veren 12 adet parametre kullanılarak analizler yapılmıştır. Ve en çok başarı elde edilen algoritmalar modellemede kullanılmıştır.

Literatürdeki birçok çalışma incelendiğinde; analizlerde teşhiste betimleyici olan tetkikler dışında diğer tetkikler dâhil edilmediği, belli yaş grubu ve kayıt sayısı ile sınırlı ve aneminin yalnızca demir eksikliğine bağlı bir hastalık olarak ele alındığı görülmüştür.

Doğan (2007) tarafından, teşhiste bilinen parametrelerden sadece dâhiliye servisine başvuran hastaların verileri ile demir eksikliği anemisi teşhisinde; serum demiri, serum bağlama kapasitesi ve ferritin enzimlerini kullanarak bir çalışma gerçekleştirmiştir. Yapılan çalışma tez kapsamındaki çalışma ile giriş parametrelerinin sayısının fazla olması böylece her yaş grubundaki bireye hitap edebilmesi ve anemi seviyesine göre gruplandırması amaçları ile paralellik göstermektedir.

Allahverdi vd. (2011); hemoglobin ve hemotokrit testinin anemi seviyesini belirlemede

sıklıkla kullanıldığını ve bunun çocuklarda önemli bir sağlık sorunu olduğunu belirleyerek, Hb ve hematokriti giriş parametresi, çıkış parametresi olarak ise; anemi seviyesi kullanılarak %90-95,8 arası başarı oranı elde etmişlerdir. Çalışmada aneminin sadece bir türü ele alınıp, çocuklara yönelik olması çalışmanın sahasını anemiye ait tüm hastaları ve çeşitlerinden ziyade daha spesifik bir alanda yapılmıştır. Yapılan tez çalışmasında ise pratikte kullanılabilir olmamasına karşın mevcut veri setinde yapılan analizlerde %95,97-%95,82 başarı elde edilmiştir.

Demir eksikliği için yapılan bir bulanık uzman sistem tasarımında mamdani çıkarım yöntemi ile aneminin bir çeşidi olan demir eksikliği anemisi için bilinen giriş parametreleriyle yalnızca demir eksikliği seviyesi belirlenmiştir (Yılmaz 2012). Bu kapsamda yapılan tez çalışmasında ise anemi ve diğer türleri için bilinen parametreler dışında diğer parametrelerinde etkisi araştırılmıştır.

Yurtay vd. (2013) kansızlık tanısına ilişkin yaptıkları çalışmada kümeleme analizi sonucunda yaş, HGB ve HCT değerlerine göre hastalar gruplandırılmış ve hastalık seviyeleri belirlenmiştir. 36-45 yaşları arasındaki kadın hastalarda HCT, HGB azalan değerlerine bağlı olarak hastalık seviyelerinin arttığı, farklı yaş gruplarına aynı küme içerisinde hamilelik, geçirilen rahatsızlıklar, beslenme gibi etkenlerden kaynaklı rastlandığı tespit edilen diğer sonuçlardır. Bu çalışmada da aynı şekilde aneminin bir türü ele alınarak belirli yaş grubu incelenmiştir. Yapılan tez kapsamında ise yaş daha geniş bir yelpazede olup sadece demir eksikliğinde belirleyici olan parametreler dışında daha fazla parametrenin anemi ve türlerine etkisi araştırılmıştır.

Doğan ve Türkoğlu (2008) karar ağaçları kullanarak demir eksikliği anemisi üzerinde değerlendirmede bulunmuşlardır. Kök düğümü olarak serum demirinin 37'den küçük olup olmaması yer almaktadır. Çalışma sonucunda karar destek sistemi ile doktorun verdiği kararların tamamen örtüştüğü görülmüştür. Aynı şekilde bu çalışmada temel belirleyici olan parametreler kullanılarak serum demiri kök düğüm olarak bulunurken bu tez kapsamında J48 algoritması ile Anemi ve AnemiDegil veri seti için kök düğümde Hemoglobinin 11,6'dan büyük olup olmaması yer almaktadır. Yine tez kapsamında yapılan karar ağacı algoritmasının uzman tarafından pratikte çokta uygulanabilir olmadığı vurgulanmıştır.

Yine Ercan (2006)'da diyabetik nefropatinin ilerleme hızının RBC, hematokrit, hemoglobin, MCV ve MCH değerlerinin başlangıç dönemindeki diyabetik nefropatiyle ilişkisini araştırdığı çalışmada; belirgin bir ilişki olduğu saptanamamıştır. Bu çalışma kapsamında 12 adet öznitelik dışında kalan özniteliklerin teşhiste belirgin bir ilişkisi olmadığı saptanmıştır.

Bu çalışmada sadece açık kaynak kodlu yazılımlardan WEKA programının sonuçlarına bağlı kalınmayıp diğer VM programları ile (Rapid Miner, SPSS Clementine vb.) aynı veri setleri üzerinde yapılan analizlerin karşılaştırılması yapılarak farklı çıkan sonuçlar için analizler iyileştirilerek çeşitli tıbbi verilerin anemi değerlendirilmesinde sınıflama, kümeleme ve birliktelik algoritmalarını katkısı gözlenebilir. Ayrıca tıp alanında geliştirilecek ve VM algoritmalarını içeren çalışmalar için bilgilendirici bir kaynak olabilir ve sağlık profesyonellerine yardımcı tıbbi karar destek sistemleri oluşturulmak suretiyle geliştirilmesi önerilir.

Hasta verilerinin sadece altı aylık bir zaman dilimine ait olması, hastalığın başlangıcı, türü ve hastanın tedavi görmesi gibi durumlara bağlı olarak bazı parametrelerdeki değişiklikler yanıltıcı olmakta ve eldeki veri seti ile pratikte teşhisi tam olarak ortaya koyan parametreler net olarak ortaya çıkmamaktadır. Veri setindeki kısıtlamalardan ötürü algoritma analizlerinin tam olarak kullanılabilir olmadığı görülmüştür.

Sınıflandırma algoritmalarında yanlış sınıflandırılan kayıtlar silinerek kümeleme analizi şeklinde; kümeleme artı sınıflandırma gibi birçok hibrit çözüm veri seti üzerinde denenerek sonuçların başarısı değerlendirilebilir. Yapılan analizler bir karar destek sistem yazılımı ile desteklenerek sağlık profesyonellerine sahada aktif kullanımda yardımcı olabilir.

Özellikle tıp alanında bu tip uygulamalar hastalara doğru teşhis koyabilmek için yüksek maliyet gerektiren testlerin ve kısıtlı laboratuvar imkânına sahip sağlık kuruluşların risk taşıyan yöntemler ve mesleğe yeni başlayan sağlık profesyonellerine hastalık sonucuna götürebilecek daha fazla ihtimali bir arada değerlendirerek hızlı ve doğru tanı koymalarına yardımcı olmak amacıyla internet tabanlı bir karar destek sistemi oluşturulabilir.

İleriki dönemlerde; çalışmada kullanılan veri miktarı artırılarak ve laboratuvar verilerinin yanı sıra fizik muayenedeki tanısal yaklaşımlar ile varsa tarayıcı incelemeler de dâhil edilerek algoritmaların daha kapsamlı örüntülerle sonuçların iyileştirilmesi sağlanarak daha geniş çaplı analizler yapılabilir.

6. KAYNAKLAR

- Adak, M., F., Yurtay, N. (2013). Gini Algoritmasını Kullanarak Karar Ağacı Oluşturmayı Sağlayan Bir Yazılım Geliştirilmesi. *Bilişim Teknolojileri Dergisi*, **6(3)**: 1-6.
- Akarsu, S. (2009). Anemilerin sınıflandırılması. *Türk Pediatri Arşivi Araştırma Dergisi*, **44**: 3-13.
- Akbulut, S. (2006). Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu. Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.
- Akça, E. (2007). Gebelerdeki Aneminin Türkiye’deki Yöresel Dağılımı. Uzmanlık Tezi, TC Sağlık Bakanlığı Haseki Eğitim ve Araştırma Hastanesi, Aile Hekimliği Koordinatörlüğü, İstanbul.
- Akgöbek, Ö., Kaya, S. (2011). Veri Madenciliği Teknikleri İle Veri Kümelerinden Bilgi Keşfi: Medikal Veri Madenciliği Uygulaması. *E-Journal of New World Sciences Academy Engineering Sciences*, **1A0145**, **6(1)**: 237-245.
- Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, **29(1)**: 1-22.
- Akyol K., Şen B., Çalık E. (2012). Biyokimya ve Hemogram Laboratuvar Test Sonuçlarının Lojistik Regresyon Yöntemiyle Analizi. AB2012 Akademik Bilişim Konferansları, Uşak, 1-3 Şubat, 1-14.
- Alan, M.A. (2012). Veri Madenciliği ve Lisansüstü Öğrenci Verileri Üzerine Bir Uygulama. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, **33**: 165-174.
- Alan, M.A. (2014a). Sivas Erzincan Kalkınma Projesi (SEKP) Verilerinin Veri Madenciliği ile Sınıflandırılması ve Kümelenmesi. *Manas Sosyal Araştırmalar Dergisi*, **3(10)**: 129-144.
- Alan, M.A. (2014b). Karar Ağaçlarıyla Öğrenci Verilerinin Sınıflandırılması. Atatürk

Üniversitesi İktisadi ve İdari Bilimler Dergisi, **28(4)**: 101-112.

Albayrak, M. (2008). EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci ile Tespiti. Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya.

Allahverdi, N., Tunalı, A., Işık, H., Kahramanlı, H. (2011). A Takagi- Sugeno Type Neuro-Fuzzy Network for Determining Child Anemia. *Expert Systems with Applications*, **38(6)**: 7415-7418.

Alpaydın, E. (2000). Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri. Bilişim 2000 Eğitim Semineri.

Altıntaş, T. (2006). Veri Madenciliği Metotlarından olan Kümeleme Algoritmalarının Uygulamalı Etkinlik Analizi. Yüksek Lisans Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Kocaeli.

Altıntaş, Y.Y. (2010). Veri Madenciliğinin Tıpta Kullanımı ve Bir Uygulama: Hemodiyaliz Hastaları İçin Risk Seviyelerine Göre Risk Faktörlerinin Etkileşimlerinin İncelenmesi. Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Amasyalı, M. F., Balcı, S., Varlı, E. S., Mete, E. (2012). Türkçe Metinlerin Sınıflandırılmasında Metin Yöntemlerinin Performans Karşılaştırılması. *Elektrik, Elektronik, Bilgisayar, Biyomedikal Mühendisliği Dergisi*, **2(4)**: 95-104.

Ardıl, E. (2009). Esnek Hesaplama Yaklaşımı ile Yazılım Hata Kestirimi. Yüksek Lisans Tezi, Trakya Üniversitesi, Fen Bilimleri Enstitüsü, Edirne.

Arslan, E. (2008). EMG İşaretlerinin İncelenmesi ve Veri Madenciliği Uygulaması. Yüksek Lisans Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya.

Aslandoğan, A. Y., Mahajani, G. A., Taylor, S. (2004). Evidence Combination in Medical Data Minig. *International Conference on Information Technology: Coding and Computing (ITCC'04)*, **2**: 465-469.

- Aydın, F. (2011). Kalp Ritim Bozukluğu Olan Hastaların Tedavi Süreçlerini Desteklemek Amaçlı Makine Öğrenmesine Dayalı Bir Sistemin Geliştirilmesi. Yüksek Lisans Tezi, Trakya Üniversitesi, Edirne.
- Aydoğan, F. (2003). E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı Ve Gerçekleştirimi. Yüksek Lisans Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.
- Ayık, Y.Z., Özdemir, A., Yavuz, U. (2007). Lise Türü ve Mezuniyet Başarısının, Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi, *Sosyal Bilimler Enstitüsü Dergisi*, **10(2)**: 441-454.
- Baloğlu, U.B. (2006). DNA Sırasındaki Tekrarlı Örüntülerin ve Potansiyel Motiflerin Veri Madenciliği Yöntemiyle Çıkarılması. Yüksek Lisans Tezi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elazığ.
- Baykal, A. (2006). Veri Madenciliği Uygulama Alanları. Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi, **7**: 95-107.
- Baykal, A. (2015). Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Bir Veri Seti Üzerinden Karşılaştırılması. Akademik Bilişim 2014.
- Baykasoğlu, A. (2005). Veri Madenciliği ve Çimento Sektöründe Bir Uygulama. Akademik Bilişim Konferansı 2005, Gaziantep, 2-4 Şubat, 1-14.
- Berry, M.J.A., Linoff, G.S. (2004). Data Mining Techniques For Marketing, Sales and Customer Relationship Management 2nd Ed., John Wiley & Sons, 586.
- Bırtıl, F.S. (2011). Kız Meslek Lisesi Öğrencilerinin Akademik Başarısızlık Nedenlerinin Veri Madenciliği ile Analizi. Yüksek Lisans Tezi, Afyon Kocatepe Üniversitesi, Afyonkarahisar.
- Bilen, Ö., Hotaman, D., Aşkın, Ö.E., Büyüklü, A.H. (2014). LYS Başarılarına Göre Okul Performanslarının Eğitsel Veri Madenciliği Teknikleriyle İncelenmesi: 2011 İstanbul Örneği. *Eğitim ve Bilim*, **39(172)**: 78-94.

- Bozkır, A.S., Sezer, E., Gök, B. (2009). Öğrenci Seçme Sınavında (ÖSS) Öğrenci Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti. 5.Uluslararası İleri Teknolojiler Sempozyumu (IATS'09), Karabük, 13-15 Mayıs, 1-7.
- Bozkurt, S., Bilge, U., Durmaz, S., Gülkesen, K.H., Yılmaz, S. (2011). Asemptomatik Karotid Arter Stenozunun Veri Madenciliği Yöntemleriyle İncelenmesi. 8.Ulusal Tıp Bilişimi Kongresi, Antalya, 17-20 Kasım, 123.
- Bramer, M. (2007). Principle of Data Mining. Springer-Verlag, London, UK.
- Breault, J.L. (2001). Data Mining Diabetic Databases: Are Rough Sets a Useful Addition? *Computing Science and Statistic*, **1(34)**: 1-10.
- Chakrabarti, S., Witten, I. *et al.*, (2008). Data Mining Know It All, Morgan Kaufmann, 2, 14, 39, 40, 41, 108.
- Coşkun, E., Kaya, A., Keçeli, A.S., Aksarı, Y., Karaağaoğlu, E. (2011). Gene 3E: Genetik Veri Madenciliği için Yen, Analiz Aracı. 8.Ulusal Tıp Kongresi, Antalya, 17-20 Kasım, 42-48.
- Çakır, M. (2005). Firma Başarısızlığının Dinamiklerinin Belirlenmesinde Makine Öğrenmesi Teknikleri: Ampirik Uygulamalar ve Karşılaştırmalı Analiz. Uzmanlık Yeterlilik Tezi, Türkiye Cumhuriyet Merkez Bankası İstatistik Genel Müdürlüğü, Ankara.
- Çarklı, B. (2010). Sağlık Sektöründe Apriori Algoritması ile Bir Veri Madenciliği Uygulaması. Yüksek Lisans Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya.
- Çataloluk, H. (2012). Gerçek Tıbbi Veriler Üzerinde Veri Madenciliği Yöntemlerini Kullanarak Hastalık Teşhisi. Yüksek Lisans Tezi, Bilecik Üniversitesi, Fen Bilimleri Enstitüsü, Bilecik.
- Çetin, H., Sezgin, E. (2011). Eğitim Fakültesi Öğrencilerinin Enformatik Bilgi Düzeylerinin K-Means Algoritması ile Gruplandırılması. 5th International

Computer & Instructional Technologies Symposium, Elazığ, 22-24 Eylül, 1-6.

Çınar, A., Silahtaroglu, G. (2012). Veri Madenciliği Teknikleri ile Müşteri Memnuniyetine Etki Eden Gizli Nedenleri Keşfi. *Marmara Üniversitesi İ.İ.B. Dergisi*, **33(2)**: 309-330.

Çoban Budak, E. (2013). Üniversite Öğrencileri için Bilgisayar Okuryazarlığını Etkileyen Faktörlerin Etkisinin Veri Madenciliği ile Analizi. *Online Academic Journal of Information Technology*, **4(11)**: 57-70.

Danacı, M., Çelik, M., Akkaya, A. E. (2010). Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhisi. Akıllı Sistemlerde Yenikler ve Uygulama Sempozyumu, Kayseri, 21-24 Haziran, 9-12.

Daş, B., Varol, A. (2013). 2D:4D Sayısal Parmak Oranına Göre Bireylerin Kişilik Durumlarının Sınıflandırılması. 1st International Symposium on Digital Forensics and Security (ISDFS'13), Elazığ, 20-21 Mayıs, 44-48.

Demirel, B. (2008). Meme Kanseri Tedavi Yöntemlerinin Veri Madenciliği ile Belirlenmesi. Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Isparta.

Demiröz, G., Güvenir, H.A. (1997). Classification by Voting Feature Intervals, ECML '97 Proceeding of the 9th European Conference on Machine Learning, Prague, Czech Republic, 23-25 April 85-92.

Dinçer, E. (2006). Veri madenciliğinde K-Means Algoritması ve Tıp Alanında Uygulanması. Yüksek Lisans Tezi, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Kocaeli.

Dinçer, E., Duru, N. (2006). Gırtlak Kanseri Ameliyat Verilerinin K-Means Yöntemiyle Analizi. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, **2(2)**: 1-10.

Doğan, Ş. (2007). Veri Madenciliği Kullanılarak Biyokimya Verilerinden Hastalık Teşhisi. Yüksek Lisans Tezi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elazığ.

- Dursun, Ö.O. (2005). Meteorolojik Verilerin Veri Madenciliği ile Değerlendirilmesi. Yüksek Lisans Tezi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü. Elazığ.
- Duru, N. (2005). An Application of Apriori Algortihm on a Diabetic Database. *Knowledge-Based Intelligent Information and Engineering System. Lecture Notes in Computing*, **3681**: 398-404.
- Düzgünoğlu, S., Yazıcı, A., Yarımağan, Ü. (2006). Tıp Bilişiminde Veri Ambarı ve Veri Madenciliği Uygulaması. 3.Ulusal Tıp Bilişimi Kongresi/Medikal Information, Antalya, 16-19 Kasım, 110-115.
- Ercan, V. (2006). Tip 2 Diabetes Mellitus'lu Hastalarda Eritrosit Sayısı, Hemoglobin, Ortalama Eritrosit Hacmi, Ortalama Eritrosit Hemoglobin Değerlerinin Başlangıç Halinde (İncipient) Diyabetik Nefropati ile İlişkisi. Uzmanlık Tezi, T.C. Sağlık Bakanlığı Taksim Eğitim ve Araştırma Hastanesi Aile Hekimliği, İstanbul.
- Erdem, S., Özdağoğlu, G. (2008). Ege Bölgesi'nde Bir Araştırma ve Uygulama Hastanesinin Acil Hasta Verilerinin Veri Madenciliği İle Analizi. *Anadolu Üniversitesi Bilim Ve Teknoloji Dergisi*, **9(2)**: 261-270.
- Ergün, E. (2008). Ürün Kategorileri Arasındaki Satış İlişkisinin Birliktelik Kuralları ve Kümeleme Analizi ile Belirlenmesi ve Perakende Sektöründe Bir Uygulama. Doktora Tezi, Afyon Kocatepe Üniversitesi, Sosyal Bilimler Enstitüsü, Afyonkarahisar.
- Esen, M.F. (2009). Veri Tabanlarından Bilgi Keşfi: Veri Madenciliği ve Bir Sağlık Uygulaması. Yüksek Lisans Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.
- Evans, R.S., Lloyd J.F., Stoddard, G.J., Nebeker J.R., Samore, M.H. (2005). Risk Faktors for Adverse Drug Events. *The Annals of Pharmacotherapy*, **39(7-8)**: 1161-1168.
- Fakih, S., Das, T. (2006). A Methodology for Learning Efficient Approached to Medical Diagnosis. *IEEE Transactions on Information Technology in*

Biomedicine, **10(2)**: 220-228.

Güllüoğlu, S.S. (2011). Tıp ve Sağlık Hizmetlerinde Veri Madenciliği Çalışmaları: Kanser Teşhisine Yönelik Bir Ön Çalışma. *Online Academic Journal of Information Technology*, **2(5)**: 1-7.

Gülten, A., Doğan, Ş. (2008). Genetik Algoritmalar Yönteminin Biyomedikal Verileri Üzerindeki Uygulamaları. *Doğu Anadolu Bölgesi Araştırma Ve Uygulama Merkezi Dergisi*, **1(7)**: 12-16.

Gürbüz, F. (2009). Havacılık Sektöründe Veri Madenciliği ile Farklı Sınıflandırma Tekniklerinin Karşılaştırılmalı Olarak Uygulanması. Doktora Tezi, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Kayseri.

Gürüler, H., Karahasan, M., İstanbullu, A. (2007). Üniversite Öğrencilerinin Profilini Belirleme: Muğla Üniversitesi Veri Tabanları Üzerinde Bir Durum Çalışması. *Muğla Üniversitesi Sosyal Bilimler Enstitüsü Dergisi (İLKE)*, **18**: 1-11.

Güven, A., Bozkurt, Ö.Ö., Kalıpsız, O. (2007). Veri Madenciliğinin Geleceği, Akademik Bilişim Konferansı 2007, Dumlupınar Üniversitesi, Kütahya,

Haberal, İ. (2007). Veri Madenciliği Algoritmaları Kullanılarak Web Günlük Erişimlerinin Analizi. Yüksek Lisans Tezi, Başkent Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Han, J., Kamber, M. (2001). Data Mining: Concepts and Techniques. Morgan Kaufmann, Publishers Inc, USA, 550.

Han, J., Kamber, M. (2006). Data Mining: Concepts and Techniques. Second Edition. San Francisco: morgan kaufman, 743.

Hatipoğlu, B. (2013). Dershane Eğitiminin Üniversiteye Yerleşmedeki Etkisinin Veri Madenciliği ile İrdelenmesi. Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi, İstanbul.

He, Y., Hui, S.C. (2009). Exploring Ant-Based Algorithms For Gene Expression Data

- Analysis, *Artificial Intelligence in Medicine*, **47(2)**: 105-119.
- Hulten, G., Spencer, L., Domingos, P. (2001). Mining Time-Changing Data Streams. *7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Fransisco, CA: ACM Pres*, 97-106.
- İşler, Y., Narin, A. (2012). WEKA Yazılımında K-Ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi. *Süleyman Demirel Üniversitesi Teknik Bilimler Dergisi*, **2(4)**: 21-29.
- İyisoy, M.S. (2014). Tanı Test Ölçütlerinde ROC Eğrisi ve Sınıflama Analizlerinin Karşılaştırılmasında Kullanımı. Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.
- Karatabak, M., İnce, M.C. (2004). Apriori Algoritması ile Öğrenci Başarısı Analizi. Elektrik Elektronik Bilgisayar Mühendisliği Sempozyumu, Bursa, 8-12 Aralık, 348-352.
- Kaur, H., Wasan, S. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. *Journal of Computer Science*, **2(2)**: 194-200.
- Kaya, E., Bulun, M., Arslan, A. (2003). Tıpta Veri Ambarı Oluşturma ve Veri Madenciliği Uygulamaları. Akademik Bilişim 2003, Çukurova Üniversitesi, Adana.
- Kaya, M. (2014). Gen İfade Verilerinde Öznetelik Seçimi ve Sınıflandırma. Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.
- Kaya, M., Özel, S.A. (2014). Açık Kaynak Veri Madenciliği Yazılımlarının Karşılaştırılması. Akademik Bilişim 2014, 5-7 Şubat, Mersin Üniversitesi, Mersin.
- Kayri, M. (2008). Elektronik Portfolyo Değerlendirmeleri için Veri Madenciliği Yaklaşımı. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, **5(1)**: 98-110.
- Kocabaş, F.M. (2010). Veri Madenciliği Süreci ve Gerçek Bir Veri Seti Üzerinde

Uygulanması. Yüksek Lisans Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü,

Koç, M. (2011). Sosyal Ağların Öğrenciler Üzerindeki Etkisinin Veri Madenciliği Kullanılarak İncelenmesi. 5th International Computer & Instructional Sysposium, Elazığ, 22-24 Eylül, 1-6.

Kökver, Y., Barışcı, N., Çiftçi, A., Ekmekçi, Y. (2014). Hipertansiyona Etki Eden Faktörlerin Veri Madenciliği Yöntemleriyle İncelenmesi. *NWSA-Engineering Sciences*, **9(2)**: 15-25.

Kurt, Ç., Erdem, O.A. (2012). Öğrenci Başarısını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle İncelenmesi. *Politeknik Dergisi*, **15(2)**: 111-116.

Larose, D.T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. Wiley Interscience A. John Wiley & Sons Inc., Canada. 214.

Meriç, O.A. (2004). Veri Madenciliği Aracı Olarak Genetik Algoritmalar ile Yapay Sinir Ağları ve Genetik Algoritma-Yapay Sinir Ağı Melez Modelinin Müşteri Değerlendirmesinde Kullanılması. Yüksek Lisans Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Morana, H. C.P., Camara, F.P., Arboleda-Florez, J. (2006). Cluster Analysis of a Forensic with Antisocial Personality Disorder Regarding PCL-R Scores: Differentiation of Two Patterns of Criminal Profiles. *Forensic Science International*, **64(2)**: 98-10.

Nalbant, S., Karan, M.A. (2010). İç Hastalıkları Uzmanının Anemiye Yaklaşımı. *İç Hastalıkları Dergisi*, **17(1)**: 7-15.

Oğuzlar, A. (2003). Veri Önişleme. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, **21**: 67-76.

Okur, B.Ç. (2013). İngiliz Alfabesi Kullanılarak Yazılmış Türkçe Metinlerin Türk Alfabesine Göre Yeniden Oluşturulması. Yüksek Lisans Tezi, Gebze İleri Teknoloji Enstitüsü Mühendislik ve Fen Bilimleri Enstitüsü, Kocaeli.

- Özarslan, S., Barışçı, N. (2014). Öğrenci Performansının Veri Madenciliği ile Belirlenmesi. 2 International Symposium on Innovative Technologies in Engineering and Science, Karabük, 18-20 Haziran, 1958-1965.
- Özçınar, H. (2006). KPSS Sonuçlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi. Yüksek Lisans Tezi, Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Denizli.
- Özdemir, A., Aslay, F.Y., Çam, H. (2010). Veri Tabanında Bilgi Keşfi Süreci: Gümüşhane Devlet Hastanesi Uygulaması. *Selçuk Üniversitesi İktisadi Ve İdari Bilimler Fakültesi, Sosyal ve Ekonomik Araştırmalar Dergisi*, **14**: 347-367.
- Özdemir, C. (2011). Panik Değerli Tıbbi Tahlil Sonuçlarının Analizi İçin Klinik Karar Destek Sisteminin Geliştirilmesi. 8.Ulusal Tıp Bilişimi Kongresi, Antalya, 17-20 Kasım, 35-41.
- Özel, C., Topsakal, A. (2014). Veri Madenciliği Kullanılarak Beton Basınç Dayanımının Belirlenmesi. *Cumhuriyet Üniversitesi Fen Fakültesi Fen Bilimleri Dergisi*, **35**: 43-57.
- Özel, D., Bilge, U., Tahtacı, R. (2011). Acil Servislerde Mortalite Tahmini İçin Veri Madenciliği Tekniklerinin Kullanılması. 8.Ulusal Tıp Bilişimi Kongresi, Antalya, 17-20 Kasım, 124-125.
- Özkan, Y. (2013). Veri Madenciliği Yöntemleri, Papatya Yayıncılık, 2.Basım, İstanbul, Türkiye.
- Pala, T. (2013). Tıbbi Karar Destek Sistemlerinin Veri Madenciliği Yöntemleriyle Gerçekleştirilmesi. Yüksek Lisans Tezi, Marmara Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.
- Rapeli, C.B., Botega, N.J. (2005). Clinical Profiles of Serious Suicide Attempters Consecutively Addmitted to a University-Based Hospital. *Revista Brasileira de Psiquiatria*, **27(4)**: 285-289.
- Roiger, R.J. , Geatz, M.W. (2003). Data Mining: A Tutorial-Based Primer, Addison-

Wesley, Pearson Education, Inc.

Sever, H., Oğuz, B. (2002). Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım: Kısım I: Eşleştirme Sorguları ve Algoritmalar. *ÜNAK, Bilgi Dünyası*, **3(2)**:173-204.

Silahtaroglu, G. (2013). Veri Madenciliği Kavram Ve Algoritmaları, Papatya Yayıncılık, 2. Basım, İstanbul, Türkiye.

Siyah, B. (2012). Oracle Data Miner ile Göğüs Kanseri Verileri Üzerine Bir Veri Madenciliği Uygulaması. Lisans Tezi, Mustafa Kemal Üniversitesi, Fen Bilimleri Enstitüsü, Hatay.

Şaylan, Ç.A. (2013). Böbrek Nakli Geçirmiş Hastalarda Akıllı Yöntem Tabanlı Yeni Öznitelik Seçme Algoritması Geliştirilmesi. Yüksek Lisans Tezi, Kadir Has Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Şen, F. (2008). Veri Madenciliği İle Birliktelik Kurallarının Bulunması. Yüksek Lisans Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya.

Şengür, D., Tekin, A. (2013). Öğrenci Mezuniyet Notlarının Veri Madenciliği Metotları ile Tahmini. *Bilişim Teknolojileri Dergisi*, **6(3)**: 7-16.

Şentürk, Z.K. (2011). Veri Madenciliği ile Kansere Teşhisi. Yüksek Lisans Tezi, Düzce Üniversitesi, Fen Bilimleri Enstitüsü, Düzce.

Şişaneci, İ. (2009). Sağlık Sisteminde Veri Madenciliği ile Suistimal Tespiti. Yüksek Lisans Tezi, Gebze İleri Teknoloji Enstitüsü, Mühendislik ve Fen Bilimleri Enstitüsü, Kocaeli.

Tantuğ, A.C. (2002). Veri Madenciliği ve Demetleme. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Tapkan, P. Özbakır, L., Baykasoğlu, A. (2011). WEKA ile Veri Madenciliği Süreci ve Örnek Uygulama. Endüstri Mühendisliği Yazılımları Kongresi, İzmir, 30 Eylül-01/02 Ekim, 247-262.

- Tarimer, İ., Elmas, F. (2015). Determination of Heart Attack Risk on Patients Data by Data Mining Applications. *Mugla Journal of Science and Technology*, **1(1)**: 18-23.
- Taş, M., Adak, M.A., Yurtay, N. (2013). Öğrencilerin Staj Verileri Üzerine Uygulan Apriori Algoritması ile Birliktelik Kurallarının Çıkarılması ve Staj Eğiliminin Belirlenmesi. 1st International Symposium On Innovative Technologies in Engineering and Science. Sakarya, 7-9 Haziran, 1086-1095.
- Taşdemir, M. (2012). Veri Madenciliği (Öğrenci Başarısına Etki Eden Faktörlerin Regresyon Analizi ile Tespiti). Yüksek Lisans Tezi, Dicle Üniversitesi, Sosyal Bilimler Enstitüsü, Diyarbakır.
- Tekerek, A. (2011). Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği. Akademik Bilişim 2011, 2-4 Şubat, İnönü Üniversitesi, Malatya.
- Telcioğlu, M.B. (2007). Veri Madenciliğinde Genetik Programlama Temelli Yeni Bir Sınıflandırma Yaklaşımı ve Uygulaması. Yüksek Lisans Tezi, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Kayseri.
- Tosun, T. (2006). Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.
- Tuğ, E. (2005). Genetik Algoritmalar ile Tıbbi Veri Madenciliği. Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.
- Ture, M., Kurt, I., Kurum, A.T., and Ozdamar, K. (2005). Comparing Classification Techniques for Predicting Essential Hypertension. *Expert Systems with Applications*, **29**: 583-588.
- Turgut, H. (2012). Veri Madenciliği Süreci Kullanılarak Alzheimer Hastalığı Teşhisine Yönelik Bir Uygulama. Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Isparta.
- Turhan, K., Engin, Y.Z., Kurt, B. (2011). Laboratuvar Test İstemleri Üzerinde Veri

Madenciliği: Destek Vektör Makineleri ile Kadın Genital Bölge Kanseri Tanı Kontrolü. 8. Ulusal Tıp Bilişimi Kongresi, Antalya, 17-20 Kasım, 29-34.

Ubeyli, E.D. (2009). Combined Neural Networks for Diagnosis of Erythematous Squamous Diseases. *Expert With Application*, **36(3)**: 5107-5112.

Ubeyli, E.D., Guler, I. (2005). Automatic Detection of Erythematous Squamous Diseases Using Adaptive Neuro-Fuzzy Inference Systems. *Computer in Biology and Medicine*, **35**: 421-433.

Uzun, Ö., Kaya, H., Gürgen, F., Varol, F.G. (2013). Olasılıksal Sınıflandırıcılar ile Doğum Öncesinde Troizmi 21 Risk Hesaplaması, IEEE 21. Sinyal İşleme ve İletişim Uygulamaları Konferansı, Girne, 24-26 Nisan, 1-4.

Ünal, Y., Köklü, M., Taşkiran, U. (2010). Application of Data Mining Techniques on Backpain Patient Data. International Science and Technology Conference, Famagusta, 27-29 Ekim, 127-132.

Üstün, G.Ü. (2013). T.C. Sağlık Bakanlığı Merzifon Kara Mustafa Paşa Devlet Hastanesi, Biyokimya Laboratuvarı, Laboratuvar Test Rehberi, Amasya, 1-44.

Vahaplar, A., Inceoglu, M.M. (2001). Veri Madenciliği ve e-Ticaret. Türkiye’de İnternet Konferansları 7, İstanbul, 01-03 Kasım, 1-8.

Vickers, E.R., Boocock, H., Harris, R.D., Bradshaw J., Cooper, M., Vickers, P., Cannon, P. (2006). Analysis of the Acute Postoperative Pain Experience Following Oral Surgery. *Australian Dental Journal*, **51(1)**: 69-77.

Wu, Alan, H.B. (2006). Tietz Clinical Guide to Laboratory Tests. Fourth Edition. Elsevier Science Health Science div, 1952.

Xu, D., Olman, V., Wang, L., Xu, Y. (2003). A Computer Program for Efficiently Mining Gene Expression Data. *Nucleic Acids Research*, **31(19)**: 5582-5589.

Yıldırım, P. (2012). Okan Üniversitesi Elektronik Ders Notları.

Yıldırım, P., Uludağ, M., Görür, A. (2008). Hastane Bilgi Sistemlerinde Veri

Madenciliği. Akademik Bilişim, Çanakkale On sekiz Mart Üniversitesi, Çanakkale, 30 Ocak- 01 Şubat, 429-434.

Yıldız, O., Tez, M., Bilge, H.Ş., Akcayol, M.A., Güler, İ. (2012). Meme Kanseri Sınıflandırması İçin Veri Füzyonu Ve Genetik Algoritma Tabanlı Gen Seçimi. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, **27(3)**: 659-998.

Yılmaz, A. (2012). Demir Eksikliği Anemisi Teşhisi için Bir Bulanık Uzman Sistem Tasarımı. Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.

Yılmaz, H. (2014). Random Forest Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama. Yüksek Lisans Tezi, Eskişehir Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, Eskişehir.

Yılmaz, Z., Pamuk, N. (2010). Uzman Sistem Kullanarak Tiroit Teşhisi. Bilimde Modern Yöntemler Sempozyumu Bildiriler Kitabı, Diyarbakır, 14-16 Ekim, 1-13.

Yurtay, N., Ateş, Z., Yurtay, Y., Çelik, U. (2010b). Tıbbi Veri Madenciliği Üzerine Bir Uygulama. International Science and Technology Conference, Famagusta, 27-29 Ekim, 787-792.

Yurtay, N., Çarklı, B., Yurtay, Y., Çit, G. (2010a). Apriori Algoritması Kullanılarak Bir İstatistik Çalışması. International Science and Technology Conference, Famagusta, 27-29 Ekim, 133-138.

Yurtay, N. (2008) Sakarya Üniversitesi Yayınlanmamış Ders Notları.

Yurtay, Y., Salman, Y., Salman, M.E., Gençali, F. (2013) . Kansızlık Tanısına İlişkin Bir Veri Madenciliği Uygulaması. 1ST International Symposium on Innovative Technologies in Engineering and Science. Sakarya, 7-9 Haziran, 896-900.

Yurtay, Y., Yurtay, N., Çelebi, N., Bacınoğlu, N.Z., Ak, G. (2014). Sakarya İline Ait Yangın Verilerinin Veri Madenciliği Yöntemleriyle Değerlendirilmesi. 2 International Symposium on Innovative Technologies in Engineering and Science, Karabük, 18-20 Haziran, 930-939.

İnternet kaynakları

- 1- <http://www.slideserve.com/earl/yapay-zeka-teknikleriyle-t-bbi-verilerin-lenmesi-ver-madenc-l> 23.09.2014
- 2- <http://danismend.com/kategori/altkategori/isletmelerde-tutulan-musteri-verilerinin-anlamli-hale-getirilmesi-ve-etkin-kullanilmasi/> 13.10.2014
- 3- <http://ninova.itu.edu.tr/tr/dersler/bilisim-enstitusu/195/bbl-606/ekkaynaklar?g8396> 20.03.2014
- 4- http://htakci.sucati.org/datamining/ikinci_ders.pdf 19.03.2014
- 5- http://htakci.sucati.org/datamining/ucuncu_ders.pdf 19.03.2014
- 6- http://htakci.sucati.org/datamining/dorduncu_ders.pdf 19.03.2014
- 7- <http://www.ncbi.nlm.nih.gov/> 22.04.2014
- 8- <http://www.embl.de/> 22.04.2014
- 9- <http://www.rcsb.org/pdb/home/home.do> 22.04.2014
- 10- <http://www.cs.waikato.ac.nz/ml/WEKA/> 23.04.2014
- 11- <https://volkansoner.wordpress.com/2011/09/15/veri-madenciligi-uygulamasina-ornekler/> 05.05.2014
- 12- <http://mertricks.com/2014/11/29/veri-madenciligi-3-veri-hazirlama-bolum-1/> 07.05.2014
- 13- <http://mertricks.com/2014/11/12/1-veri-madenciligine-giris-101/> 07.05.2014
- 14- <http://www.kdnuggets.com/polls/2014/industries-applied-analytics-data-mining-data-science.html> 04.01.2015
- 15- <https://aydinerceis.wordpress.com/2011/06/19/bilgi-ve-enformasyon-uzerine-bir-deneme/> 10.02.2015
- 16- <http://www.stat.gen.tr/index.php?istek=sinif&dersid=ist01&konuid=ver01&max=1> 22.02.2015
- 17- <http://bilgisayarkavramlari.sadievrenseker.com/2013/03/31/siniflandirma-classification/> 03.03.2015
- 18- www.slideshare.net/mgocmen37/z-skorzscore 10.01.2015
- 19- http://tr.wikipedia.org/wiki/Fleiss'in_kappa_katsay%C4%B1s%C4%B1 10.01.2015

- 20- http://www.centralhospital.com/TibbiBolumler/ic-hastaliklari--dahiliye-/hastaliklar_ve_tedavi_yontemleri/kansizlik--anemi--b12-ve-folik-asit-eksikligi/ 13.01.2015
- 21- <http://www.hematolojika.com/content2.asp?m1=1&m2=8> 05.11.2014

ÖZGEÇMİŞ

Adı Soyadı : Asiye Betül CİGA

Doğum Yeri ve Tarihi : Çorum 1984

Yabancı Dili : İngilizce

İletişim (Telefon/e-posta) : 0 (506) 456 76 38 - betulciga@hotmail.com

Eğitim Durumu (Kurum ve Yıl)

Lise : Çorum Merkez İmam Hatip Lisesi, Çorum (2001)

Lisans : Selçuk Üniversitesi Teknik Eğitim Fakültesi Elektronik
ve Bilgisayar Eğitimi, Konya (2007)

Yüksek Lisans : Afyon Kocatepe Üniversitesi Fen Bilimleri Enstitüsü
İnternet ve Bilişim Teknolojileri Yönetimi, Afyon
(2015)

Çalıştığı Kurumlar ve Yıl : Havza Mesleki ve Teknik Anadolu Lisesi, Samsun
(2007-2009)

Merzifon Mesleki ve Teknik Anadolu Lisesi, Merzifon
(2009-...)

EKLER

Ek 1. Amasya Kamu Hastaneler Birlięi Genel Sekreterlięi'nden alınan izin yazısı



T.C.
SAęLIK BAKANLIęI
TÜRKİYE KAMU HASTANELERİ KURUMU
Amasya İli Kamu Hastaneleri Birlięi Genel Sekreterlięi
Merzifon Kara Mustafa Paşa İlçe Devlet Hastanesi

Sayı : 44269710-903.02/434
Konu : Akademik Bilimsel Çalışmalar

5.102/2015

AMASYA KAMU HASTANELERİ BİRLİęİ GENEL SEKRETERLİęİNE
AMASYA

İlgi: 24.11.2014 tarih ve 10455 sayılı yazınız.

İlgi sayılı yazınız ile Afyon Kocatepe Üniversitesi Fen Bilimleri Enformatik Ana Bilim Dahil İnternet ve Bilişim Teknolojilere Yönetme bölümünde yüksek lisans öğrencisi Asiye Betül CİGA'ya hastanemize ait son altı aylık Vitamin B12 eksikliği anemisi, talasemi, hepatit hastalıklarına ait mikrobiyoloji, biyokimya, hemogram, hormon, koagülasyon, eliza ve patoloji laboratuvarı sayısal verilerinin hasta kimlik bilgileri ve tanımlayıcı bilgileri olmadan verilmesinde sakınca yoktur.

Bilgilerinize arz ederim.

Uzm.Dr. Hasan ERGİN
Hastane Yöneticisi 108728