T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

MÜŞTERİ KAYIP ANALİZİ PROBLEMİNİN ÇÖZÜMÜNDE ANALİTİK YAKLAŞIMLAR

Fatma Önay KOÇOĞLU

Enformatik Anabilim Dalı

Enformatik Programı

DANIŞMAN
Prof. Dr. Ş. Alp BARAY

II. DANIŞMAN
Yrd. Doç. Dr. Tuncay ÖZCAN

Kasım, 2017

İSTANBUL

Bu çalışma, 24.11.2017 tarihinde aşağıdaki jüri tarafından Enformatik Anabilim Dalı, Enformatik Programında Doktora tezi olarak kabul edilmiştir.

Tez Jürisi

Prof. Dr. Ş. Alp BARAY(Danışman)
İstanbul Üniversitesi
Mühendislik Fakültesi

Prof. Dr. Sevinç GÜLSEÇEN
İstanbul Üniversitesi
Enformatik Bölümü

Prof. Dr. Selim ZAİM
İstanbul Teknik Üniversitesi
İşletme Fakültesi

Prof. Dr. Zuhal TANRIKULU
Boğaziçi Üniversitesi
Uygulamalı Bilimler Yüksekokulu

Doç. Dr. Tarık KÜÇÜKDENİZ
İstanbul Üniversitesi
Mühendislik Fakültesi

20.04.2016 tarihli resmi gazetede yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

Bu tez, İstanbul Üniversitesi Bilimsel Araştırma Projeleri Yürütücü Sekreterliğinin 49762 kodu ve 2221 ID numaralı projesi ile desteklenmiştir.

ÖNSÖZ

Bilimsel çalışmanın nasıl yürütülmesi gerektiği konusunda bilgi ve tecrübesi ile bana yol gösteren, doktora öğrenimimin başından beri sabır ve desteğini eksik etmeyen çok değerli danışmanım Prof. Dr. Ş. Alp BARAY'a sonsuz teşekkürlerimi sunarım.

Tez çalışması süresince her türlü problem karşısındaki sakin duruşu, çözüm odaklılığı ve bilgi birikimi ile desteğini esirgemeyen, kendisinden çok şey öğrendiğim ve daha çok şey öğreneceğime inandığım değerli eş danışmanın Yrd. Doç. Dr. Tuncay ÖZCAN'a çok teşekkür ederim.

Yıllar önce meslek aşkımı, bu yoldaki kendime olan güvenimi ve azmimi fark ederek ışık olup yolumu aydınlatan, kocaman bir aile olduğumuzu bizlere hep hissettiren kıymetli hocam Prof. Dr. Sevinç GÜLSEÇEN'e çok teşekkür ederim.

Güzel insan, değerli hocam Yrd. Doç. Dr. Çiğdem EROL'a sabırla, şefkatle her daim desteğim, yol gösterenim olduğu için özel olarak teşekkür ederim.

Saygıdeğer hocalarım ve sevgili çalışma arkadaşlarım, İ.Ü.Enformatik Bölümü ailesinin tüm üyeleri Yrd. Doç. Dr. Zerrin AYVAZ REİS, Öğr. Gör. Dr. Murat GEZER, Arş. Gör. Dr. Elif KARTAL, Arş. Gör. Dr. Serra ÇELİK, Arş. Gör. Dr. Zeki ÖZEN, Arş. Gör. Dr. Emre AKADAL, İnci AKSAKAL, Nesim YUMLU, Yiğitcan BAKIR ve Mustafa KURT'a desteklerinden ötürü teşekkür ederim.

Annem Muhterem KOÇOĞLU, kardeşim E. Seda KOÇOĞLU, anneannem Ayla ÇIKLA'ya üzerimdeki emeği ve desteği için çok teşekkür ederim.

Övündüğü en büyük eseri evlatları olan, herşeyden önce dürüstlüğü, iyi niyeti ve merhameti, ATA'ya saygıyı, her koşulda sağlam durabilmeyi öğreten, kızı olduğum için her zaman kendimi şanslı hissettiğim canım babam Mustafa KOÇOĞLU'na ayrıca teşekkürü bir borç bilirim.

Can dostum, yol arkadaşım, bugün yanımda olmasını çok istediğim Nazlı KOÇOĞLU, bu yolda yapabileceklerimi ve hayallerimi önce hep seninle paylaştım, seninle de nokta koymak istedim, huzur içinde uyu.

Kasım 2017 Fatma Önay KOÇOĞLU

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

SİMGE VE KISALTMA LİSTESİ

Simgeler **Açıklama**

α : Pozitif Lagrange Çarpanı
A_i : *i*. Girdiye Ait Ağırlık Değeri
b : Eşik Değer
C : Sınıflar Kümesi
Ç : Çıktı Vektörü
d(x,y) : Uzaklık Fonksiyonu
D_i : *i.* Veri Kümesi
ξ : Aylak Değişken
g() : Aktivasyon Fonksiyonu
G : Girdi Vektörü
H : Ayırma Düzlemleri (Destek Vektör Makineleri İçin)
H : Aktivasyon Fonksiyonu Çıktısı (Sinir Ağları İçin)
k : Küme Sayısı
K() : Kernel Fonksiyonu
L() : Lagrange Fonksiyonu
Ñ : Nöron Sayısı
σ : Standart Sapma
μ : Ortalama Değer
p(x) : Olasılık Fonksiyonu
P : Olasılık Değeri
σ : Standart Sapma
sim : Benzerlik Fonksiyonu
Φ : Haritalama Fonksiyonu
T : Sınıf Niteliği
x_i : Veri Setinde Yer Alan *i*. Örnek
x_{min} : Veri Setinde Yer Alan En Küçük Değere Sahip Örnek
x_{maks} : Veri Setinde Yer Alan En Büyük Değere Sahip Örnek
W : Ağırlık Vektörü
w_i : *i*. Ağırlık Vektörü Elemanı
z : Amaç Fonksiyonu

Kısaltmalar	Açıklama
AC	: Associative Classification
AF	: Aktivasyon Fonksiyonu Türü
AUC	: ROC Eğrisi Altındaki Alan
CRISP-DM	: Cross Industry Standard Process for Data Mining
D	: Doğruluk
DN	: Doğru Negatif
DP	: Doğru Pozitif
ED	: Kullanılan Eşik Değeri
ELM	: Extreme Learning Machine, Aşırı Öğrenme Makinesi
F	: F Ölçüsü Değeri
G	: Karışıklık Matrisi Değerler Toplamı
GR	: Kazanım Oranı
Ht	: Hassasiyet Oranı
IEEE	: The Institute of Electrical and Electronics Engineers
IG	: Bilgi Kazancı
Info	: Bilgi Miktarı
K	: Kesinlik Değeri
NGB	: Net Girdi Bilgisi
NS	: Gizli Katmandaki Nöron Sayısı
NTO	: Negatif Tahmin Oranı
Ozg	: Özgünlük Değeri
PGY	: Performans Geçerleme Yöntemi
ROC	: Alıcı İşlem Karakteristiği
SCI	: Science Citation Index
SCI-Exp	: Science Citation Index-Expanded
SEMMA	: Sample, Explore, Modify, Model, Assess
SInfo	: Bölünme Bilgisi
SSCI	: Social Sciences Citation Index
SVM	: Support Vector Machine, Destek Vektör Makinesi
VTBK	: Veri Tabanlarında Bilgi Keşfi
YD	: Yeni Doğruluk Değeri
YN	: Yanlış Negatif
YP	: Yanlış Pozitif

ÖZET

DOKTORA TEZİ

MÜŞTERİ KAYIP ANALİZİ PROBLEMİNİN ÇÖZÜMÜNDE ANALİTİK YAKLAŞIMLAR

Fatma Önay KOÇOĞLU

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Enformatik Anabilim Dalı

Danışman : Prof. Dr. Ş. Alp BARAY

II. Danışman : Yrd. Doç. Dr. Tuncay ÖZCAN

Günümüz dünyası, gelişen bilgi ve iletişim teknolojilerinin etkisi altındadır. Teknolojinin sunduğu olanaklar ile beraber öneminin günden güne artmasnı sonucu bilgi, işletmelere ait sermaye, makine, iş gücü gibi kaynakların arasında önemli bir kaynak haline gelmiştir. Bilginin temel yapı taşının veri olduğu düşünülürse işletmeler her türlü veriyi saklar olmuştur. Tüm bu veri içerisinde müşterilere ait veri ayrı bir önem arz etmektedir. Bir işletmenin varlığını devam ettirebilmesi müşterileri ile arasındaki alışverişin sürekliliğine bağlıdır. Dolayısıyla, işletmeler zorlu rekabet koşulları altında yeni müşteriler kazanmanın yanında, var olan müşterilerini de elde tutmalıdır. Buna göre; müşteri davranışlarının izlenmesi ve bu davranışlardaki değişikliklere göre hareket edilmesi oldukça önemlidir. Bu kapsamda "Müşteri Kayıp Analizi" çalışmaları gerçekleştirilmektedir. Müşteri kayıp analizi çalışmaları ile beraber ayrılma eğilimi gösteren müşteriler ve bu eğilime neden olan durumlar ortaya çıkartılmaya çalışılmaktadır.

Diğer taraftan bilgi ve iletişim teknolojilerinin gelişimi ile beraber daha yüksek işlem gücü ve daha fazla depolama alanına sahip bilgisayarlar daha az maliyet ile üretilebilir hale gelmiştir. Bunun bir sonucu olarak saklanan veri miktarında artış meydana gelmiş, bu verinin analizi için çeşitli yöntemler geliştirilmiştir. Bu yöntemlerden biri "Veri Madenciliği"dir. Veri madenciliği, en genel anlamda, veri yığını içerisinden daha önceden keşfedilmemiş, amaca yönelik olarak işe yarar bilgiyi veya veri içerisindeki ilişkiyi ortaya çıkarma işi ve bu iş için kullanılan yöntemler bütünü olarak tanımlanabilir. Veri madenciliğinin etkin kullanıldığı problemlerden biri de müşteri kayıp analizidir.

Bu tez çalışması kapsamında iki farklı literatür çalışması yapılmıştır. Bu çalışmalardan ilkinde sıkça kullanılan veri madenciliği yöntemlerinin belirlenmesi amaçlanırken, ikincisinde tez problemi için daha önce kullanılmamış yeni yöntemler tespit edilmeye çalışılmıştır. Yeni yöntem önerisi için yapılan literatür incelemesinde araştırmanın yapıldığı dönemde Aşırı Öğrenme Makinelerinin problemin çözümünde kullanılmadığı tespit edilmiştir. Bu doğrultuda, "Aşırı Öğrenme Makineleri"nin bir sınıflandırma problemi olarak müşteri kayıp analizi probleminin çözümündeki etkinliğinin ve belirli koşullar altında en iyi çözümü üreten parametrelerinin araştırılması amaçlanmıştır. Aşırı Öğrenme Makinelerinin ön analizler ile problem çözümü için uygunluğu test edilmiş, olumlu sonuçlar elde edilmiştir. Aşırı Öğrenme Makinelerinin temelde üç farklı parametre değerinin önem teşkil ettiği saptanmıştır. Bunlar gizli katmanda yer alan nöron sayısı, aktivasyon fonksiyonu ve eşik değeridir. Bu parametre değerlerinin değiştirilmesi ile beraber performans değerlendirme ölçülerinde de iyileşme olacağı tezi ile hareket edilmiştir.

Aşırı Öğrenme Makineleri ile elde edilen performans değerlendirme ölçü değerleri, K-En Yakın Komşu Algoritması, Sade Bayes Algoritması ve Destek Vektör Makinelerinin kullanımı ile elde edilen performans değerlendirme ölçü değerleri ile karşılaştırılmıştır. Farklı performans geçerleme yöntemleri ile kurulan modellerde, Aşırı Öğrenme Makineleri %90'ın üzerinde bir doğruluk oranı ile çalışma kapsamında kullanılan diğer yöntemlerden daha iyi performans sergilemiştir. Tez çalışması sonucunda Aşırı Öğrenme Makinelerinin temel sınıflandırma problemi için başarılı olduğu, müşteri kayıp analizi çalışmalarında da etkin bir şekilde kullanılabileceği, belirlenen aralıklarda parametre değerlerinin iyileştirilmesi sağlanırsa, bu parametre değerleri ile beraber sonuçların da iyileşebileği görülmüştür.

Bu tez çalışması, İstanbul Üniversitesi Bilimsel Araştırma Projeleri Yürütücü Sekreterliğinin 49762 kodu ve 2221 ID numaralı projesi ile desteklenmiştir.

Kasım 2017, 169 sayfa.

Anahtar kelimeler: Aşırı öğrenme makineleri, müşteri kayıp analizi, veri madenciliği .

SUMMARY

Ph.D. THESIS

ANALYTICAL APPROACHES FOR SOLVING IN CHURN ANALYSIS PROBLEM

Fatma Önay KOÇOĞLU

İstanbul University

Institute of Graduate Studies in Science and Engineering

Department of Informatics

Supervisor : Prof. Dr. Ş. Alp BARAY

Co-Supervisor : Assist. Prof. Dr. Tuncay ÖZCAN

Today's world is under the influence of developing information and communication technologies. As a result of increasing importance of knowledge and with the opportunities offered by technology, knowledge became an important source for enterprises such as capital, machinery, workforce. When it is considered that data is the basic unit of knowledge, enterprises commence storing all kinds of data. In all kinds of data, customer data has special importance. The ability of an enterprise to maintain its existence depends on the continuity of the relationship with its customers. Therefore, enterprises must keep existing customers as well as acquire new customers under tough competition conditions. So, it is important to monitor customer behavior and act on the changes in these behaviors. Within this scope, "Customer Churn Analysis" studies are carried out. With customer churn analysis studies, it is tried to forecast the customers who tend to leave and reveal the situations that cause this tendency.

On the other hand, with the development of information and communication technologies, computers with higher processing power and more storage capacity can be produced with less cost. As a consequence, the amount of data stored has increased, and various methods have been developed for the analysis of this data. One of these methods is "Data Mining". Data mining can be described in general sense as whole of the methods to discover the relationship between data, which has not been discovered before in a huge data stack. Today, various

algorithms for data mining are being developed and used for solving many problems. One of the problems which data mining is used effectively for is customer churn analysis.

Within the scope of this thesis study, two different literature studies were carried out. In the first of these studies, it was aimed to determine the data mining methods which are frequently used and in the second one tried to find new methods that were not used before for the thesis problem. In the literature review for the new method proposal, it was determined that the Extreme Learning Machine were not used to solve the problem at the time of the research. In this respect, it is aimed to investigate the effectiveness of the "Extreme Learning Machine" problem in solving the customer churn analysis problem as a classification problem and the parameters that produce the best solution under certain conditions. Preliminary analyzes of Extreme Learning Machine have been tested for its suitability for solving the customer churn analysis problem, and positive results have been obtained. It has been determined that three different parameter values are important for Extreme Learning Machines. These are the number of neurons in the hidden layer, the activation function and the threshold value. Acted with the thesis “If these parameter values are changed, the performance evaluation measure will improve”.

The performance evaluation measure values obtained with the Extreme Learning Machine were compared with the performance evaluation measure values obtained with the use of K-Nearest Neighbor Algorithm, Sade Bayes Algorithm and Support Vector Machines. Findings based on the different models established shows that Extreme Learning Machine performed better than the other algorithms used in the study with an accuracy of over 90%. As a result of the thesis study, it is seen that Extreme Learning Machine is successful for the basic classification problem and that it can be used effectively in customer churn analysis studies, if the parameter values are improved in the determined intervals, the results can be improved with these parameters.

This work was supported by Scientific Research Project Coordination Unit of Istanbul University. Project number: 49762.

November 2017, 169 pages.

Keywords: Customer churn analysis, data mining, extreme learning machine.

1. GİRİŞ

Gelişen bilgi ve iletişim teknolojileri hayatımızın hemen hemen her alanını fazlasıyla etkilemektedir. Gerek günlük yaşantımızda, gerek akademik veya sektörel alanda dijitalleşme hızla gerçekleşmekte, bu dönüşüm beraberinde yeni avantaj ve dezavantajları da getirmektedir. Giderek artan veri boyutu veri bilimi ile uğraşan kişilerce dile getirilmektedir. Ancak dijital dönüşüm ile beraber sadece insanların değil akıllı makinelerin de ürettiği veri ile karşı karşıya kalınacaktır.

Teknolojideki gelişim veri miktarındaki artışı tetikleyeceği gibi, artan veri miktarı ile sahip olunan bilgi miktarının artışı arasında pozitif ilişki olacağı beklenmekte, elde edilen bilgi ile beraber teknolojinin daha da gelişmesine olanak sağlanacağı düşünülmektedir. Ancak bilinmektedir ki bilginin elde edilmesi sadece sahip olunan veri yığınlarının boyutunun artması ile mümkün olamamaktadır. Veri yığınlarının işlenmesi, bu devasa yapı içerisinden işe yarar, gizli, daha önceden bilinmeyen bilgi veya ilişkilerin ortaya çıkartılması gerekmektedir.

Veri madenciliği, 1990'lı yıllardan bu yana hızla yaygınlaşan, veriden bilgiye giden yolda en önemli araştırma alanlarından biri olarak karşımıza çıkmaktadır. Veri madenciliği, madencilik kavramı ile benzeşmektedir. Aynı mantıkla yola çıkıldığında veri madenciliği, veri yığını içerisinde gizli kalmış ve değerli olan bilginin ortaya çıkartılması işi, bu iş için gerekli olan yöntemler bütünü şeklinde ifade edilebilir. Veri madenciliği çok çeşitli alanlarda birçok problemin çözümü için kullanılmaktadır. Bu problemler arasında hastalık teşhisi, riskli kredi kartı müşterilerinin belirlenmesi, müşteri alışveriş davranışlarının incelenmesi, bilgisayar güvenlik sistemlerinde kimlik tespiti, web saldırılarının tespiti, görüntü işleme sayılabilir. Veri madenciliği istatistik, makine öğrenmesi, yapay zekâ gibi çeşitli dallardan beslenen ve her ne kadar ortak yönleri olsa da bu alanlardan ayrılan bazı özellikleri ile beraber başlı başına bir araştırma alanıdır.

Diğer taraftan, günümüzde küreselleşen dünya ile beraber çetin bir rekabet ortamı oluşmuş, stratejik karar alma mekanizmalarını ve iş süreçlerini de etkiler olduğundan bilginin önemi işletmeler için de ayrıca önemli hale gelmiştir. Böyle bir ortamda adapte olabilmenin, rekabet avantajı yakalayabilmenin yollarından birisi işletmenin sahip olduğu en önemli kaynaklardan olan veriyi ihtiyaç duyulan bilgiye dönüştürebilmesidir. İşletmeler için en önemli unsurlardan

birisi müşterileridir. İşletmeler yeni müşteriler elde etmenin yanı sıra var olan müşterilerini de kaybetmeme konusunda mücadele vermektedir. Kaybedilen müşteriler işletmeler açısından artan maliyet ve azalan kar oranı ile beraber sorun teşkil etmektedir (Forhad ve diğ., 2014). Bu noktada veri madenciliği ayrılma eğilimindeki müşterilerin belirlenmesi problemi için de etkin çözümler üretmektedir. Kaybedilen müşteriyi geri döndürmeye çalışmak yerine, müşterinin kaybedilmeden önce ayrılma eğiliminde olma durumunun tahmin edilebilmesi konusunda veri madenciliğinden yararlanılmaktadır. Bu sayede müşteri davranışları çözümlenebilmekte ve etkin stratejik kararlar alınabilmektedir. Veri madenciliği yöntemleri ile gerçekleştirilen bu süreç müşteri kayıp analizi şeklinde adlandırılmaktadır. Müşteri kayıp analizinin temel amacı; ayrılma eğilimi gösteren müşterilerin tanımlanması ve mümkün ise de belirlenen bu müşterilerin tutundurma durumunun ortaya konulmasıdır (Mutanen, 2006).

Bu tez çalışması kapsamında, veri madenciliği yöntemlerinin kullanılarak ayrılma eğilimi gösteren müşterilerin belirlenmesi, bu doğrultuda farklı veri madenciliği algoritmalarının performanslarının tespit edilmesi, özel olarak Aşırı Öğrenme Makineleri yönteminin müşteri kayıp analizi probleminin çözümü için kullanılabilirliğinin ve tez çalışmasının malzeme ve yöntem kısmında belirtilen koşullar altında en iyi çözümü üreten parametrelerinin araştırılması ve kullanılan tüm algoritmaların performans kıyaslamasının ortaya konulması amaçlanmıştır. Aşırı Öğrenme Makineleri yöntemi kapsamında kullanılan algoritma daha önce müşteri kayıp analizi probleminin çözümü için kullanılmamış, bu algoritma için parametre araştırmasına gidilmemiştir. Tez çalışmasının sonuçlarının özellikle İşletme Enformatiği alanı için yapılacak benzer çalışmalarda iyileştirme olanakları konusunda öncü bir çalışma olacağı görülmektedir.

Bu tez çalışmasının GENEL KISIMLAR ana başlığı altında veri madenciliği konusunun temel yapı taşları olan veri, enformasyon ve bilgi kavramlarına yer verilmiş; veri madenciliğinin ilk olarak dile getirildiği veri tabanlarında bilgi keşfi süreci açıklanmıştır. Temel kavramlardan sonra veri madenciliği konusuna giriş yapılmış, literatürde sıkça yer verilen tanımlar ele alınarak bu tanımların sentezi yapılmaya çalışılmıştır. Veri madenciliğinin çeşitli uygulama alanları, diğer disiplinler ile olan ilişkisi, veri madenciliği çalışmaları için sürecin nasıl ilerlediği ve yaygın olarak kullanılan kabul görmüş süreçlerin açıklamalarına yer verilmiştir. Veri madenciliğinin kümeleme, birliktelik kuralları vb. işlevleri açıklandıktan sonra tez çalışması problemi için de geçerli olan sınıflandırma işlevi detaylandırılmıştır. Sınıflandırma işlevi için kullanılan çeşitli yöntemler açıklanmıştır. Tez çalışmasının odak noktalarından olan Aşırı

Öğrenme Makinelerinin açıklaması ve algoritma adımlarına ayrıntılı olarak yer verilmiştir. Bölümün son kısmında ise bir sınıflandırma problemi olarak müşteri kayıp analizi problemi için ayrıntılı bir literatür çalışması yapılmış, literatürde yer alan çalışmalar çeşitli kriterlere göre sınıflandırılmıştır.

MALZEME ve YÖNTEM kısmında CRISP-DM sürecine bağlı kalınarak tez çalışması için kullanılan veri setinin detayları, yapılan ön işleme çalışmaları, oluşturulan modellerin ayrıntıları, gerekli kodlama araçları detaylı olarak açıklanmıştır.

BULGULAR bölümünde kurulan modellerden elde edilen sonuçlar öncelikle yöntem bazında ele alınmış, ardından farklı performans değerlendirme ölçülerine göre yöntemlerin karşılaştırmalı performans analizlerine yer verilmiştir.

TARTIŞMA ve SONUÇ bölümünde tez çalışmasının diğer çalışmalardan farkı, bulguların yorumları, elde edilen sonuçların literatüre katkısı, gelecek çalışmalarının neler olabileceği konularında araştırmacıların görüşlerine yer verilmiştir.

2. GENEL KISIMLAR

İnsanoğlu varlığının ilk döneminden günümüze kadar çeşitli değişim ve dönüşüm evrelerini yaşamıştır. Bu süreç insanın ihtiyaçlarına göre çeşitli girdi ve çıktılara sahip olmuştur. Bu girdi ve çıktılara bağlı olarak da toplumlar çeşitli isimler ile anılmıştır. Başlangıçta "Avcı-Toplayıcı" olarak adlandırılan insan yerleşik hayata geçerek toplu yaşama biçimini benimsemiştir. Yerleşik hayat ekme-biçme faaliyetlerinin başlamasına neden olmuş ve buna bağlı olarak da toplum "Tarım Toplumu" olarak tanımlanmıştır. 18. Yüzyıl ortalarında buhar gücü ile çalışan yeni makinaların üretime katkısı sanayi devrimine neden olmuş, bu yeni dönem toplumun "Sanayi Toplumu" olarak adlandırılmasını beraberinde getirmiştir.

Günümüzde ise, gelişen bilgi ve iletişim teknolojileri ile bu teknolojilerin farklı alanlarda kullanımı "bilgi"yi dönemin en önemli olgusu haline getirmiş, insanların günlük hayatlarındaki alışkanlıklar dahi değişiklik göstererek "Bilgi Toplumu"na dönüşüm gerçekleşmiştir. Bu süreç içerisinde işletmeler için fiziksel sermayenin dışında entelektüel sermayenin de değeri artmış, işletmenin sahip olduğu veri, enformasyon ve bilgi ile bunların en doğru şekilde kullanımı önem kazanmıştır. Bu doğrultuda işletmeler, ancak elde bulunan veri ve enformasyonu yararlı bilgiye dönüştürebilirlerse zorlu rekabet koşulları altında avantaj elde edebilmektedirler.

Başka bir açıdan işletmeler için en önemli unsurlardan biri müşterilerdir. Küreselleşme, yeni pazarların oluşması, hızlı değişen ortam koşulları gibi zorlu rekabet koşulları altında işletmelerin odak konularından birisi de "müşteri davranışları"dır (Kisioglu ve Topcu, 2011). Yeni bir müşteri kazanmanın var olan müşteriyi elde tutma maliyetinden yüksek olması da önemli bir faktör olarak göz önünde bulundurulduğunda, işletmeler özellikle var olan müşterilerini kaybetmeme konusunda yeni stratejiler geliştirmeye çalışmaktadırlar. Bu amaçla, müşteri verileri kullanarak ayrılma eğilimi olan müşteriler belirlenmekte, bu müşterilerin elde tutulabilmesi için çeşitli stratejik kararlar verilebilmektedir.

Veriden bilgiye dönüşüm aşamasında "Veri Madenciliği" yöntemleri karşımıza çıkmaktadır. Bu yöntemler işletmelerin stratejik planlama ve karar verme süreçlerinde de sıklıkla kullanılmaktadır. En genel anlamda büyük veri yığını içerisinden anlamlı bilgiye ulaşma işi olarak tanımlanan veri madenciliğinin; ayrılma eğilimli müşterilerin belirlenmesi, bankacılıkta riskli müşterilerin belirlenmesi, dolandırıcılıkların tespiti, olası hastalık durumunun tahmin edilmesi, market sepet analizi gibi farklı alanlarda çeşitli çalışma örnekleri mevcuttur.

Bu bölümde veri, enformasyon, bilgi gibi temel kavramlara, tezin çalışma konusu olan veri madenciliği ve müşteri kayıp analizi hakkında temel bilgilere, sınıflandırma için kullanılan çeşitli algoritmaların açıklamalarına ve tez kapsamında kullanılan Aşırı Öğrenme Makinesinin detaylarına yer verilmiştir.

2.1. VERİ, ENFORMASYON VE BİLGİ

Veri; herhangi bir nesne ya da olayı tanımlayan, bu nesne ya da olaya ait değerleri tutan, işlenmemiş karakterler topluluğu olarak ifade edilmektedir. Enformasyon, yararlılığının arttırılmasına yönelik olarak çeşitli işlemlerden geçen, yararlılığı ve anlamlılığı arttırılmış işlenmiş veri şeklinde tanımlanmaktadır (Ackoff, 1989). Bilgi çeşitli analiz ve sentezler sonucu karar almaya ve tahminler yapmaya yardımcı olan değerler, enformasyon, uzman görüşü ve deneyimlerin bir karışımı; diğer bir deyişle zenginleştirilmiş enformasyondur (Davenport ve Prusak, 1998). Bilgi aynı zamanda veriden başlayıp enformasyonu da kapsayacak şekilde bilgide son bulan yapılanmış bir sürekliliğin içerisinde en değerli içerik biçimidir (Grover ve Davenport, 2001). Bu üç olgu en basit şekilde örneklenecek olursa; bir bölgeye ait günlük yağış miktarının sayısal değeri veriye, aynı bölgeye ait aylık yağış miktarı ortalaması enformasyona ve daha önceden bilgi sahibi olunmayan bu bölgenin kurak, ılıman, yağışlı bir bölge olup olmadığı ve buna bağlı olarak doğal bitki örtüsü hakkında öngörüler ise bilgiye karşılık gelmektedir.

Veri, bilgiye giden yolda temel taş olup, kendisine bir anlam katılmadığı sürece herhangi bir değeri bulunmamaktadır. Bu noktada, verinin öncelikle kullanılabilir bir form olan enformasyona dönüşmesi gerekmektedir. Veriye anlam katılarak enformasyona dönüşüm sürecinde çeşitli yöntemlerin varlığını ifade eden Davenport ve Prusak (1998) bu süreci aşağıdaki adımlar ile özetlemiştir:

Amaca Yönelme: Verinin hangi amaç doğrultusunda toplandığının bilinmesi.

Kategorize Etme: Verinin veya analizin temel bileşenlerinin belirlenmesi.

Hesaplama: Verinin matematiksel veya istatistiksel olarak analiz edilmesi.

Düzeltme: Verideki hataların giderilmesi.

Özetleme: Verinin daha öz bir formda ifade edilmesi.

Veri-enformasyon-bilgi arasında dönüşümün ve güçlü bir bağın varlığı oldukça açıktır. Yukarıdaki çeşitli tanımlamalardan da anlaşılacağı üzere bilgiye erişebilmek için öncelikle verinin varlığı gereklidir. Enformasyon verinin dönüşümü ile elde edilebilirken, bilgi ise ancak veriden çıkarılan enformasyon üzerine inşa edilebilmektedir (Boisot, 1998). Veri, enformasyon ve bilgi arasındaki ilişki Şekil 2.1'de yer alan bilgi hiyerarşisi piramidi ile gösterilmektedir. Piramidin en alt basamağında bulunan veri öncelikle toplanıp organize edilerek anlamlı olan enformasyona, enformasyon ise özetleme, analiz etme ve sentezleme aşamalarından geçerek bilgiye dönüşmektedir. Enformasyon geçmiş ve şimdiki zamanı kapsayan, tanımlayıcı özelliktedir. Bilgi ise, geçmiş ile bugüne dayanarak karar vermeye ve geleceğe yönelik belirli bir düzeyde tahmine olanak sağlayan yapıdadır (Dubin, 1976; Camerer ve Johnson, 1991; Akt:Kock ve diğ., 1997). Bu yapısı nedeniyle piramidin en üst noktasında enformasyondan elde edilen "bilgi" karar alma aşamasında kullanılmaktadır.

Şekil 2.1: Bilgi hiyerarşisi piramidi (Chaffey ve Wood, 2005).

Şekil 2.1'de (Chaffey ve White, 2005) veriden bilgiye dönüşümde her basamakta değer artarken, anlam bakımından en alt seviyede verinin yer aldığı görülmektedir.

2.2. VERİ TABANLARINDA BİLGİ KEŞFİ

Bilgiye sahip olanın bilgi temelli stratejik kararlar verebileceği, bununla beraber potansiyel güce de sahip olabileceği ve karar vericilerin daha çok bilgi ile daha az risk alarak strateji üretebileceği düşüncesi bilginin öneminin her geçen gün artmasına sebep olmuştur. Bu durumda bilgiye erişim isteği ve çabası, gelişen teknoloji ile beraber daha az maliyet ile üretilebilen, depolama alanı daha fazla ve işlem gücü yüksek bilgisayarların da geliştirilmesi ile beraber her türlü verinin saklanması sonucunu doğurmuştur. Bununla beraber verinin

saklanıp gerektiğinde veriye erişim sağlanabileceği yeni teknolojik araçlar ve "veri tabanı", "veri ambarı" gibi kavramlar ortaya çıkmıştır.

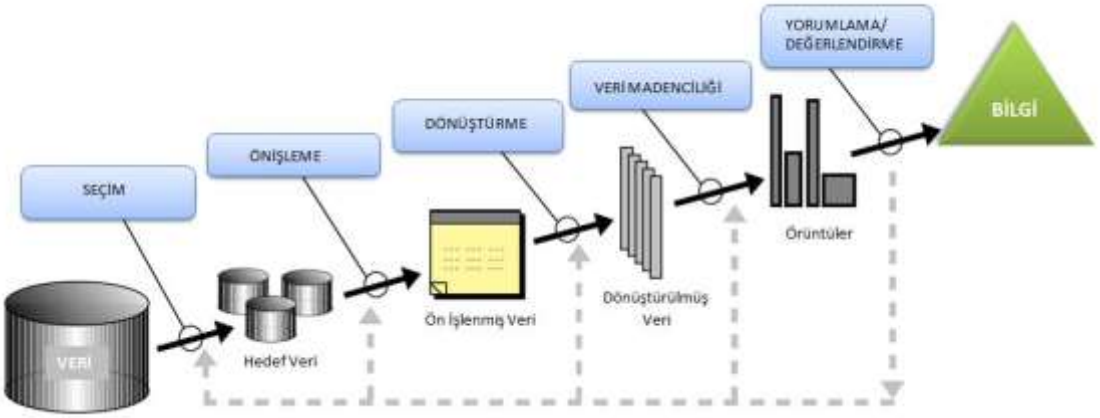
Veri tabanı; karmaşık yapıdaki dosyaların saklanması, kullanıcı erişimi, dosyalar arası ilişkilerin belirlenmesinde yetersizlik gibi sorunlar nedeniyle geleneksel dosya sistemine alternatif olarak çıkan, birbiri ile ilişkili verilerin tutulduğu ortamlar olarak ifade edilebilir. Günümüzde ise daha gelişmiş yapılar olan veri tabanı yönetim sistemleri ve veri ambarları yaygın olarak kullanılmaktadır. Veri tabanı yönetim sistemleri kullanıcı ile etkileşime izin veren bir ara yüze sahip, verilerin düzenli bir şekilde tutulmasına ve yönetilmesine olanak sağlayan yazılımlardır. Veri ambarları da verilerin saklanabildiği ve analiz edilerek sorgulanabildiği ortamlar olup, karar verme sürecine destek sağlayabilme, belirli bir konuya yönelik olabilme, bütünleşik ve zaman boyutluluk gibi özellikleri ile veri tabanı yönetim sistemlerinden ayrılmaktadır (Inmon, 2000). Veri ambarları farklı veri tabanlarından gelen verileri tutabilmekte, müşteri gibi özel bir konuya yönelik ve belirli bir zaman dilimine ait verileri saklayabilmektedir.

Gittikçe artan veri boyutu ile beraber veri tabanlarının sağladığı raporlama ve sorgulama araçları veriyi anlamlı hale getirme konusunda yetersiz kalmıştır. Bunun sonucu olarak yeni hesaplama yöntemleri ve araçlarına ihtiyaç duyulmuştur. Farklı disiplinlerde iş zekâsı, makine öğrenimi, veri madenciliği gibi çeşitli isimlerle anılan, birbirinden bazı özellikler ile farklılaşan ancak temelinde verinin toplanması, düzenlenmesi, modellenmesi ve bilgiye erişimin denenmesi yani verinin analizini sağlayan yöntemler geliştirilmiştir (Akpınar, 2014).

Bu yöntemlerden birisi olan "Veri Tabanlarında Bilgi Keşfi" süreci ilk kez 1989 yılında ortaya çıkmıştır. Bu süreç, optimizasyon, istatistik, veri tabanı yönetimi, örüntü arama, paralel programlama, görselleştirme gibi birçok farklı çalışma alanı ile de etkileşim halindedir (Bradley ve diğ., 1999).

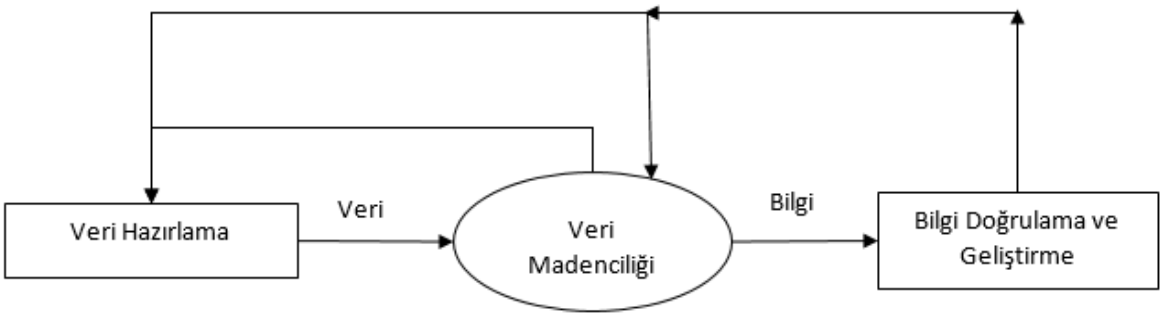
Veri Tabanlarında Bilgi Keşfi (VTBK); veri içerisinde mantıklı, yeni, amaca yönelik olarak faydalı ve anlaşılır örüntüleri ortaya çıkarmak üzere gerçekleştirilen ve gerçekleştirilmesi basit istatistiksel hesaplamalar, sorgulama ve raporlamaya dayanmayan faaliyetler dizisi olarak tanımlanmaktadır (Fayyad ve diğerleri, 1996). Norton (1999) bu sürecin bilginin toplanması, algoritmalar, yeni bilginin araştırılması ve elde edilebilmesi için gerekli mekanizmalar ve alt süreçler etrafında yapılandığını ifade etmektedir.

Fayyad ve diğerleri (1996), VTBK sürecini Şekil 2.2'de yer aldığı üzere beş adımda ifade etmektedir. Bu süreç verinin hazırlanması, örüntülerin aranması ve elde edilen sonuçların değerlendirilip yorumlanması gibi temel işlemlerden meydana gelmektedir.



Şekil 2.2: Veri tabanlarında bilgi keşfi süreci (Fayyad ve diğ., 1996).

VTBK sürecine ait bir başka yapılandırmaya ise Şekil 2.3'te yer verilmiştir. Freitas (2002), Fayyad'a göre bu süreci biraz daha özetleyerek verinin hazırlanma aşamasını ilk adım, veri madenciliği yöntemlerinin uygulanmasını bir sonraki adım ve veri madenciliği sonrası elde edilen bilginin geliştirilmesini ise son adım olarak belirterek üç adımlı bir süreç oluşturmuştur.



Şekil 2.3: Veri tabanlarında bilgi keşfi süreci (Freitas, 2002).

Her iki şekilden de görüleceği üzere "Veri Madenciliği" VTBK süreci altında bir adım olarak gösterilmiştir. VTBK sürecinin bir alt adımı olarak Veri Madenciliği, veri içerisinden örüntülerin elde edilmesi için kullanılan ve karar ağaçları, regresyon ya da kümeleme yöntemlerini içeren algoritmik bir adımı ifade etmektedir. Ancak görülmektedir ki; Veri

Madenciliği bu süreçlerde VTBK'nın çekirdeğini, amaca yönelik olarak en önemli adımını oluşturmaktadır.

Fayyad (1996), veri madenciliğini herhangi bir hipoteze ihtiyaç duyulmadan yeni bilgileri keşfetmek amacıyla yapılan veri analizi şeklinde, VTBK sürecinin bir alt adımı olarak ifade etmektedir. Veri madenciliği her ne kadar VTBK'nin alt süreci olarak ifade edilse de, son dönem çalışmalarda VTBK, veri madenciliği, bilgi çıkarımı, örüntü analizi, veri arkeolojisi, iş zekâsı vb. kavramlar yaygın olarak aynı anlamda kullanılır olmuştur.

2.3. VERİ MADENCİLİĞİ

Bilgi ve iletişim teknolojileri geliştikçe daha çok verinin saklanması olanaklı hale gelmiştir. Bu veri markette yapılan alışverişe ilişkin veriden, bankada gerçekleştirilen işlem verisine; müşteri verisinden, stok verisine; hastalara ilişkin çeşitli sağlık testleri verisinden, bir alana ait görüntü verisine kadar çeşitlilik gösterebilmektedir. Tüm bu veri yığını günümüzde sayısal ortamlara aktarılarak kayıtlı hale gelmektedir. Gelinen son durumda veri kıtlığı yerini veri bolluğuna bırakmış, bilgiye erişim korkusundan ziyade erişim sağlanan bilgi ile mücadele edebilme endişesi dahi oluşmaya başlamıştır. (Usama ve Paul, 1997, Akt: Şentürk, 2006).

1980'lerde var olan istatistiksel yöntemler veri tabanlarında saklanan çok büyük miktardaki verinin analizi ve veri içerisinde lineer olmayan ilişkilerin ortaya çıkartılması konusunda yetersiz kalmaya başlamıştır (Nisbet ve diğ., 2009). Artan veri yığını ile başa çıkabilmek ve bu yığından fayda sağlayabilmek adına birçok yeni istatistiksel ve matematiksel yöntem geliştirilmiştir. Bu yöntemlerden bir tanesi "Veri Madenciliği"dir. Veri madenciliği, 1989'da VTBK sürecinin bir adımı olarak ön plana çıkmış, çeşitli problemlerin çözümü için kullanılmıştır.

Veri madenciliği yöntemlerinin ilk dönemlerinde kullanımı oldukça zorlu, tek bir hedefe yönelik işlem yapılabilen Quinlan'ın C4.5 karar ağacı algoritması, Inselberg'in paralel-koordinat görselleştirme yöntemi gibi temel yöntemler söz konusu iken 1990'larda veri madenciliğine ilgi artmış, 1995'ten itibaren ikinci nesil veri madenciliği sistemleri geliştirilmeye başlanmıştır. Bu sistemler ile beraber farklı analiz türlerinin gerçekleştirilmesi mümkün hale gelmiş, özellikle veri temizleme ve ön işleme adımları uygulanabilir olmuştur (He, 2009). Bilgisayarların işlem gücü göz önüne alındığında, veri madenciliğinin ilk kullanıldığı dönemlerde hesaplama yeteneği açısından soru işaretleri oluşsa da, gelişen

teknoloji ile beraber bu soru işaretleri de aşılmış, azalan işlem süresi ile beraber özellikle var olan yöntemlerin tahmin sonuçlarının doğruluk oranında iyileştirmeler için çalışmalar başlamıştır (Coenen, 2011). Günümüzde ise yapısal ve yapısal olmayan, işlemsel ve akan gibi veri setleri ile analizler gerçekleştirilebilir olmuş, birçok farklı problemin çözümü için çeşitli veri madenciliği yöntemleri geliştirilmiş, bu yöntemlerin uygulanabilirliği için ticari veya açık kaynak kodlu yazılımlar kullanıma sunulmuştur.

2.3.1. Veri Madenciliği Tanımı

Madencilik kavramı bilindiği üzere "*yer altındaki bazı bölgelerde çeşitli iç ve dış doğal etkenlerle oluşan, ekonomik yönden değer taşıyan minerallerin araştırılması, çıkarılması ve işletilmesiyle ilgili teknik ve yöntemlerin bütünü*" şeklinde tanımlanmaktadır. (TDK, 2017). Aynı mantıkla yola çıkıldığında, veri madenciliği de en genel anlamda veri yığını içerisinde gizli kalmış ve değerli olan bilginin ortaya çıkartılması işi, bu iş için gerekli olan yöntemler bütünü şeklinde ifade edilebilir. Veri madenciliğinin amacı veri yığını içerisinde sadece arama yapmak değil; bu veriyi doğru yöntemler ile analiz ederek, amaca yönelik olarak önemli bilgiyi ortaya çıkarmak ve yorumlamaktır. Bu doğrultuda bir arama motoru vasıtasıyla özel bir bilgiyi internette aratmak, hastalık tespiti için hastanın kaydını aramak, kitap listesi içerisinden belirli türden kitapları aratmak, finansal bir rapor içerisindeki görsellerin analiz edilmesi veri madenciliği uygulamaları olarak değerlendirilmemektedir (Gorunescu, 2011).

Literatürde birçok çalışmada şu dört tanıma da sıklıkla yer verilmektedir (Friedman, 1997; Kovalerchuk ve Vityaev, 2000; Mărginean, 2003; Gatnar ve Rozmus, 2005; Korukonda, 2007; Koçoğlu, 2012; Lodhi, 2012):

- Veri içerisinde daha önceden keşfedilmemiş ilişkileri ve örüntüleri ortaya çıkarmak, ayırt etmek üzere veri tabanlarında bilgi keşfi sürecinde kullanılan yöntemler bütünüdür (Ferruzza).
- Daha önceden bilinmeyen, anlaşılır, işlenebilir verileri ayıklamak ve elde edilen bu verileri stratejik kararlar vermek üzere kullanan bir süreçtir (Zekulin).
- Veri içerisindeki avantaj sağlayacak örüntüleri açığa çıkarma sürecidir (John)
- Veri tabanlarında bilinmeyen ve öngörülemeyen bilginin ve örüntülerin ortaya çıkartıldığı karar destek sürecidir (Parsaye).

Bu tanımların dışında literatürde araştırmacılar tarafından çeşitli veri madenciliği tanımı yapılmıştır. Han ve diğerleri (2012), veri madenciliğini çok büyük miktardaki veri içerisinden ilginç örüntü ve bilgileri keşfetme süreci olarak tanımlarken, bu büyük miktardaki verinin betimlenmesi için bilgi madenciliği yerine veri madenciliği ifadesinin tercih edildiğini belirtmiştir. Benzer olarak Hand (2007), veri madenciliği için geniş veri setleri içerisinden ilginç, beklenmeyen ve değerli olan yapıları ortaya çıkarma süreci tanımını yapmıştır. Fayyad ve diğerleri (1996) ise veri madenciliğini VTBK sürecinin bir alt adımı olarak kabul etmekte ve veri içerisinden çeşitli örüntüleri ortaya çıkarmak için işlem süresi ve uygulanabilirliği kabul edilebilir bir oranda olan özel algoritmaların kullanılması işi olarak tanımlamaktadır. Witten ve diğerlerinin (2016) ifadesine göre veri madenciliği veri içerisinde gizlenmiş, anlamlı ve fayda sağlayacak örüntüleri otomatik veya yarı otomatik süreçler ile ortaya çıkarmaktır. Chung ve Gray (2009) veri madenciliğini işletmeler açısından ele almış, organizasyona ait var olan bilgiye hem yeni bilgileri hem de yeni öngörü ve bakış açılarını katabilen bir süreç olarak ifade etmiştir.

2.3.2. Veri Madenciliğinin Uygulama Alanları

Veri madenciliği doğası gereği istatistik, veri tabanı teknolojileri, yapay zekâ, makine öğrenmesi, görselleştirme, paralel ve dağıtık programlama, örüntü tanıma gibi alanlar ile etkileşimli halde olup, çeşitli yöntemler bakımından bu alanlardan yararlanmaktadır (Zhou, 2003).

Veri madenciliğinin her ne kadar birçok farklı disiplin ile ilişkisi olsa da istatistik ve yapay zekâ veri madenciliğini besleyen en önemli iki kaynaktır (Akpınar, 2014). Çünkü veri madenciliği problemlerin çözümü için istatistik ve makine öğrenmesi tabanlı algoritmaları kullanmaktadır. Veri madenciliği klasik istatistiksel uygulamalara benzemek ile beraber kullanılan veri ile ilgili önemli bir ayrım durumu söz konusudur. İstatistik tüm veri setini en iyi şekilde betimleyecek belirli sayıda kaydı içeren özet veri ile çalışırken veri madenciliğinde çok fazla sayıda kayıttan ve nitelik alanından oluşan bir veri seti ile çalışılabilmektedir (Özkan, 2008).

Yine veri madenciliği ve makine öğrenmesi zaman zaman birbiri ile eş olduğu düşünülen iki alandır. Ancak bu iki alan süreç olarak birbirine benzese de temelde bazı farklılıklar göstermektedir. Veri madenciliği için verinin elde edilmesinden bilginin çıkarılmasına kadar geçen tüm adımlar aynı oranda önemlidir. Makine öğrenmesinde ise modelin kurulması ve

modelin öğrenme performansının geliştirilmesi daha fazla önem verilen adımlar olarak karşımıza çıkmaktadır. Bunun yanı sıra, veri madenciliğinde verinin arkasında mutlak surette bir yapı olduğu varsayımı yerine şayet bir örüntü var ise bunu ortaya çıkarmak amaç edinilirken, makine öğrenmesinde veriler arasında bir yapı olduğu var sayılarak model kurulumuna gidildiği söylenebilir (Mannila, 1996). Veri madenciliğinin diğer alanlar ile etkileşimi Şekil 2.4'te gösterilmiştir.

Şekil 2.4: Veri madenciliğinin ilişkili olduğu diğer disiplinler.

Veri madenciliği daha önce de belirtildiği üzere çeşitli problemler için çözümler üretebilmektedir. Bu özelliği ile beraber veri madenciliğinin farklı alanlarda uygulanabildiğini görmek mümkündür. Bu alanlardan bazıları bankacılık, pazarlama, sigorta, sağlık, telekomünikasyon, elektronik ticaret, borsa, spor, taşımacılık, eğitim olarak sayılabilir. Bu alanlarda yapılan uygulamalardan örnekler aşağıda verilmiştir:

- Müşterilerin satın alma davranışlarının belirlenmesi veya benzer özellik gösteren müşteri gruplarının elde edilmesi, buna bağlı olarak kampanyaların düzenlenmesi, mağaza ürün yerleşiminin belirlenmesi, müşteri tutundurma çalışmalarının gerçekleştirilmesi, ayrılma eğiliminde olan müşterilerin belirlenmesi vb.
- Risk yönetimi ve dolandırıcılıkları belirleme çalışmaları kapsamında kredi kartı dolandırıcılıklarının tespiti, riskli kredi gruplarının belirlenerek batık kredi oranlarının

iyileştirilmesi, bilgisayar sistemlerine izinsiz girme girişimlerinin belirlenmesi, internet bankacılığı dolandırıcılıklarının tespiti vb.

- Hastalıkların, hastalık tiplerinin, olası semptomların belirlenmesi.

Veri madenciliğinin araştırılması, uygulanması ve geliştirilmesi gereken hususları Han ve diğerleri (2012) şu beş ana başlık altında listelemiştir:

Veri Madenciliği Metodolojileri: Araştırmacılar her geçen gün yeni veri madenciliği teknikleri geliştirmekte ve uygulamaktadır. Farklı nitelik alanlarının kombinasyonu ile bütünleştirme işleminin yapılması ve elde edilen veri küpleri ile analizlerin gerçekleştirilmesi, farklı disiplinlerdeki teknik ve bilginin veri madenciliği yöntemlerine entegrasyonu, farklı veri nesneleri arasındaki semantik bağların kullanımı, eksik, gürültülü veya belirsizlik içeren verinin veri madenciliği analiz sürecini olumsuz yönde etkilememesi için ön işleme adımlarının geliştirmesi, kullanıcıların öznel ölçütlerine göre elde edilen örüntülerin daha doğru değerlendirilebilmesi için çeşitli yöntemlerin geliştirilmesi gibi konular da ele alınmalıdır.

Kullanıcı Etkileşimi: Kullanıcılar veri madenciliği sürecinin en önemli unsurlarındandır. Dolayısıyla kullanıcıların aktif ve kolay bir şekilde sürecin çeşitli girdilerini değiştirerek sonuçlarda meydana gelen değişiklikleri izleyebileceği, sonuçların görselleştirilerek daha anlamlı hale getirilebileceği sistemler geliştirilerek, sorgulama araçlarının önemli olması nedeni ile veri madenciliği sorgu dillerinin gelişiminin sağlanması gerekmektedir.

Verimlilik ve Ölçeklenebilirlik: Veri madenciliği çalışmalarında algoritmaların verimlilik, ölçeklenebilirlik ve gerçek zamanlı olarak yürütme becerisi gibi performans göstergeleri oldukça önemlidir. Dolayısıyla bu beceriler konusunda ileri düzey algoritmalar geliştirmek önem arz etmektedir. Ayrıca, veri boyutu, dağılımı ve bazı veri madenciliği algoritmalarının hesaplama karmaşıklığı paralel ve dağıtık veri madenciliği algoritmalarının geliştirilmesine de neden olmaktadır.

Veri Tabanı Tiplerinin Çeşitliliği: Çeşitli uygulamalar vasıtasıyla biyolojik diziler, sensör verisi, uzaysal veri, metin verisi, multimedya verisi, web verisi, sosyal ağ verisi gibi çeşitli veri nesneleri elde edilebilmektedir. Bu farklı veri setlerinin de analiz edilebilmesi için çeşitli veri madenciliği yöntem ve araçları geliştirilmektedir. Farklı veri tiplerinin yanı sıra farklı veri kaynakları internet ve çeşitli ağlarla birbirine bağlanmaktadır. Bu nedenle heterojen veri

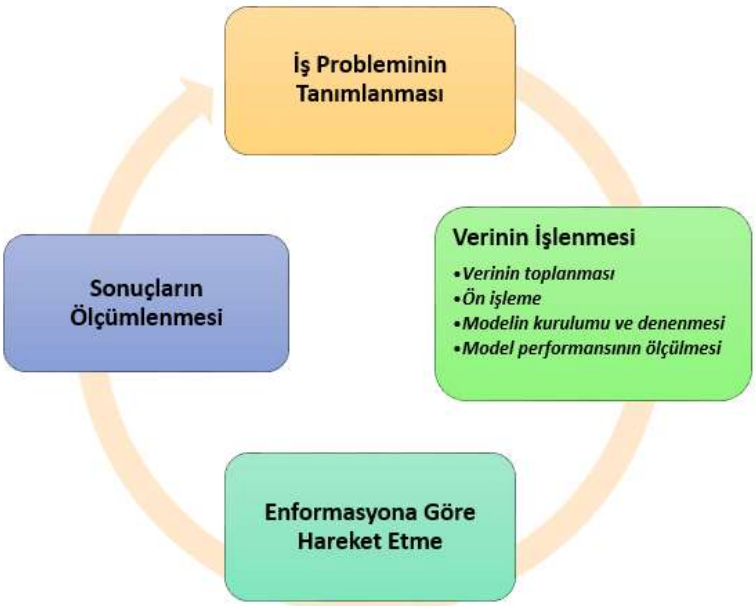
kümeleri oluşmakta bu veri kümelerinin analizi zorlaşmaktadır. Bu alanda yapılan çalışmalar da gelişme göstermektedir.

Veri Madenciliği ve Toplum: Veri madenciliği teknolojilerinin toplumsal boyutta yararları, kötü amaçlı kullanım durumu ve alınması gereken önlemler, kişisel verinin korunması gibi konularda çalışmalar yürütülmektedir.

2.3.3. Veri Madenciliği Süreci

Veri madenciliğinin başarısı her ne kadar iyi araçlar ve yetenekli analistlerin başarısına bağlı olsa da bunların yanı sıra etkin sistematik bir yaklaşım da gereklidir. Ancak böyle bir yaklaşımla veri madenciliği süreci doğru irdelenebilir ve yönetilebilir olacaktır (Wirth ve Hipp, 2000). Bu doğrultuda, veri madenciliği için günümüze kadar araştırmacılar tarafından çeşitli süreç önerileri yapılmıştır.

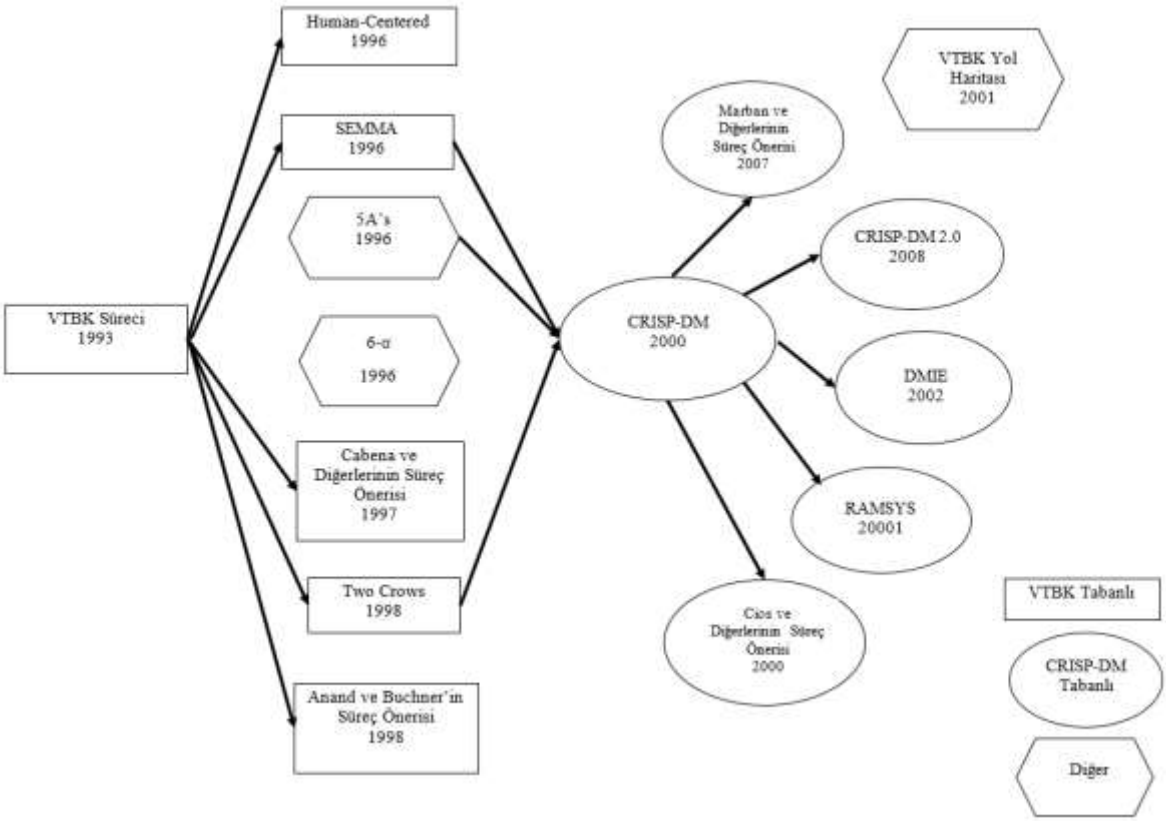
Örneğin; Berry ve Linoff (2000), veri madenciliği sürecini işletme açısından bir iş probleminin ortaya çıkışından, buna ait çözüm üretilinceye kadar geçen bir süreç olarak tasarlamıştır. Bu süreç iş probleminin tanımlanması, verinin işlenmesi, enformasyona göre hareket etme, sonuçların ölçümlenmesi olmak üzere dört adımlı bir döngüdür. Buna göre ilk aşamada iş ihtiyaçlarının neler olduğunun belirlenmesi, alan uzmanlarının görüşlerinin alınması, beyin fırtınası ile çözülmesi istenilen problem ortaya konulmaktadır. İkinci adım, analizlerde hangi verilerin kullanılacağının belirlenmesi ve buna bağlı olarak verinin toplanması, verinin temizlenmesi, verinin doğru forma dönüştürülmesi, gerekli hallerde ek değişkenlerin oluşturulması, modelin kurulumu için kullanılacak veri setinin hazırlanması, model için gerekli algoritmanın seçimi ve denenmesi, model performansının ölçülmesi gibi işlemleri kapsamaktadır. Üçüncü adımda ikinci adımın sonucundaki bulgulara göre strateji ve iş fikirlerinin belirlenmesi, verinin düzeltilmesi, modelin farklı sistemlere entegresi gibi işlemler gerçekleştirilmektedir. Son adımda ise kurulan modelin tahmin çıktıları ile gerçek sonuçlarının karşılaştırılarak performansının değerlendirilmesi yapılmaktadır. Berry ve Linoff (2000)'a ait döngü Şekil 2.5'te verilmiştir.



Şekil 2.5: Veri madenciliği süreci verimli döngü.

Cios (2000) ise VTBK sürecini yapılandırmak suretiyle medikal problemleri ele alarak altı adımlı bir süreç önerisinde bulunmuştur. Bu süreç problemin belirlenmesi, verinin anlaşılması, verinin hazırlanması, veri madenciliği, elde edilen verinin değerlendirilmesi ve elde edilen verinin kullanılması adımlarından meydana gelmektedir.

Mariscal ve diğerleri (2010), veri madenciliği için geliştirilen süreçleri VTBK tabanlı süreçler CRISP-DM tabanlı süreçler ve diğer süreçler olmak üzere 3 ana grupta sınıflandırmıştır. Veri madenciliği süreçlerinin tarihsel gelişimi ve yukarıda belirtildiği üzere sınıflandırılışı Şekil 2.6'da verilmiştir.



Şekil 2.6: Veri madenciliği süreçlerinin sınıflandırılması (Mariscal ve diğ., 2010).

İlk olma özelliği ile VTBK süreci önemini korurken geliştirilen süreçler arasında lüteratürde kendine yer edinmiş ve en yaygın olarak kullanılan süreçler CRISP-DM ve SEMMA'dır (Kurgan ve Musilek, 2006; Azevedo ve Santos, 2008; Mariscal ve diğ., 2010; Shafique ve Qaiser, 2014). Bu süreçlere ait ayrıntılı açıklamalar aşağıda verilmiştir.

2.3.3.1. Veri Tabanlarında Bilgi Keşfi Süreci

Veri madenciliği için ilk süreç Fayyad ve diğerleri (1996) tarafından VTBK ile tanımlanmıştır. Süreç dokuz adımdan meydana gelmektedir:

- Problemin Belirlenmesi: Uygulama alanının anlaşılması ve sürecin amacının belirlenmesi.
- Veri Seçimi: Amaca yönelik olarak verinin veya veri setinden alt veri kümesinin seçilmesi.
- Verinin Temizlenmesi ve Ön İşleme: Gürültülü ve eksik verilerin temizlenmesi veya tamamlanması, aykırı durumların belirlenmesi, gerekli durumlarda atılması.

- Veri Boyutunun Azaltılması ve Dönüştürme: Gerekli görüldüğü taktirde veriden gereksiz görülen nitelik alanlarının atılarak boyutun küçültülmesi, ihtiyaca yönelik olarak nitelik alanlarının dönüştürülmesi.
- Veri Madenciliği İşlevinin Seçilmesi: Veri madenciliğinin işlevleri olan sınıflandırma, regresyon, birliktelik kuralları gibi işlevlerinden birinin seçilmesi.
- Veri Madenciliği Algoritmasının Seçilmesi: Belirlenen amaca ve işleve göre analizde kullanılacak algoritmaların seçimi.
- Veri Madenciliği: Örüntüler, kurallar, kümeler gibi amaca yönelik sonuçların elde edilmesi.
- Yorumlama: Veri madenciliği uygulaması sonrası elde edilen sonuçların yorumlanması.
- Bilgiye Erişim: Elde edilen bilginin farklı kaynaklardan gelen bilgi ile birleştirilmesi, raporlanması, dökümante edilmesi.

Bu adımlar daha önce Şekil 2.2'de belirtildiği üzere seçim, ön işleme, dönüştürme, veri madenciliği, yorumlama/değerlendirme şeklinde beş adımda özetlenebilir.

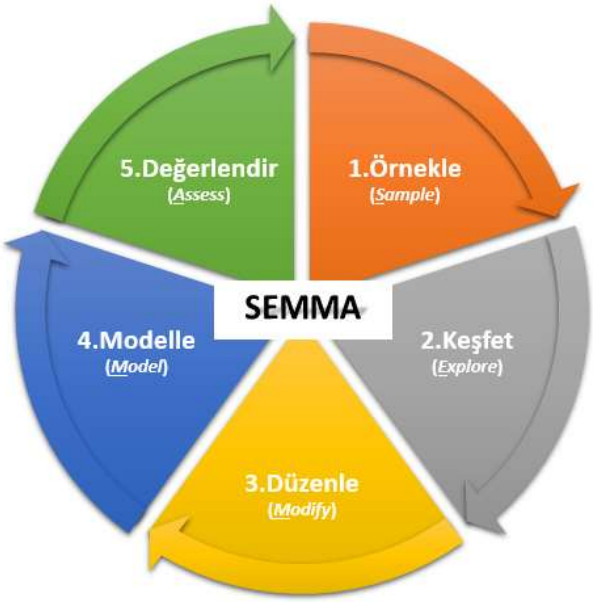
2.3.3.2. SEMMA

SEMMA, SAS Institute tarafından veri madenciliği sürecini daha anlaşılabilir bir hale getirmek ve bir yol haritası çıkarmak üzere, yine SAS Institute tarafından geliştirilen SAS Enterprise Miner veri madenciliği aracı için oluşturulmuş süreçtir. Bu veri madenciliği aracı için oluşturulması sebebiyle bu sistemin sınırları dışında çalışmaması sürecin dezavantajıdır (Rohanizadeh ve Moghadam, 2009). SEMMA, *Sample* (Örnekle), *Explore* (Keşfet), *Modify* (Düzenle), *Model* (Modelle) ve *Assess* (Değerlendir) şeklinde bir açılıma sahip olup toplamda beş adımlı bir süreçtir.

Süreç örnekle adımı ile başlarken, bu adımda verinin örneklenmesi gerçekleştirilir. Örnekleme işleminde bilgi çıkarımına olanak sağlayacak ve modellerin uygulanmasında performansı düşürmeyecek şekilde bir veri örneğinin seçilmesine önem verilir. Keşfetme verinin anlaşılmasına yönelik verinin özetlendiği, aykırı değerlerin belirlendiği adımdır. Düzenle adımında uygulamaya sokulacak nitelik alanları seçilip, gerekli dönüşüm işlemleri yapılarak bir nevi veri ön işleme gerçekleştirilmektedir. Modelleme aşamasında farklı veri madenciliği yöntemleri kullanılarak modeller elde edilmektedir. Değerlendirme olan sona adımda ise elde edilen modellerin performans göstergeleri karşılaştırılarak doğru modelin seçilmesi ve bu

model ile elde edilen sonuçların yorumlaması yapılmaktadır (Azevedo ve Santos, 2008; Akpınar, 2014).

SEMMA süreci Şekil 2.7'de şematize edilmiştir.



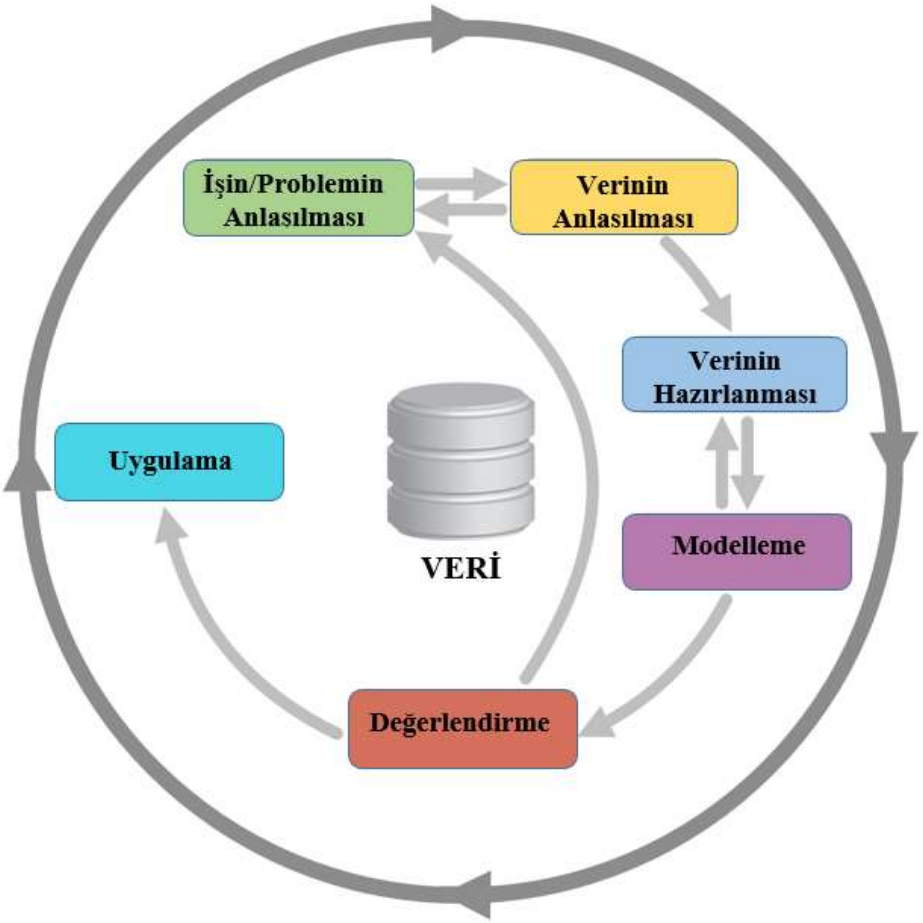
Şekil 2.7: SEMMA süreci.

2.3.3.3. Veri Madenciliği için Çapraz Endüstri Standart Süreci (CRISP-DM)

CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) en çok tercih edilen diğer veri madenciliği süreçlerinden bir tanesi olup ilk sürümü 2000 yılında SPSS, NCR (Teradata), Daimler-Chrysler ve OHRA işbirliği ile geliştirilmiştir.

Yeni veri tiplerinin varlığı, elde edilen sonuçların çeşitli operasyonel sistemler ve mevcut iş süreçleri ile entegrasyonu, geniş ölçekli veri tabanlarının yerinde analizi yerine bu veri tabanlarından çekilecek analitik veri setleri ile işlem yapma, veri madenciliği ve analitik süreçler ile ilgili daha çok sayıda insanın eğitimi gibi konularda iş gereksinimlerinin artması var olan sürümün geliştirilmesi sonucunu doğurmuştur (Mariscal ve diğ., 2010).

CRISP-DM altı adımdan oluşan bir süreçtir. Bu adımlar Şekil 2.8'de gösterilmiştir.



Şekil 2.8: CRISP-DM süreci.

İşin Anlaşılması/Problemin Belirlenmesi: Mevcut durumun değerlendirilmesi, hedeflerin belirlenmesi, proje planının oluşturulması işlemleri gerçekleştirilerek, veri madenciliği ile sonuç üretilmek istenilen problem belirlenir.

Verinin Anlaşılması: Birinci adımda belirlenen probleme göre gereksinin duyulan verinin ne olduğu belirlenmeye çalışılır. Buna bağlı olarak ihtiyaç duyulan veri toplanır. Analizlerde kullanılacak veri tek bir kaynağa bağlı olmaksızın farklı veri kaynaklarından elde edilen veri setlerinin bütünleştirilmesi ile de elde edilebilir. Bu adımda verinin anlaşılabilirliği için istatistiksel özetleme, kümeleme, aykırı değer analizi gibi yöntemlerden yararlanılabilir.

Verinin Hazırlanması: Bu adımda verinin ön işlemesi gerçekleştirilmektedir. Veri ön işleme doğru analiz sonuçlarının elde edilebilir olması için oldukça önem arz eden bir aşamadır. Çünkü analizlere ve probleme uygun veri seti, doğruluğu daha yüksek olan sonuçlar elde edilmesini

sağlayacaktır. Bundandır ki veri madenciliği çalışmalarında en fazla efor sarf edilen aşama ön işleme aşamasıdır (Cabena ve diğ., 1998).

Teknoloji her ne kadar çeşitli fırsatlar sunuyor olsa da bazı problemleri de beraberinde getirebilmektedir. Farklı sistemlerden verinin alınması sırasında oluşabilecek aksaklıklar, tekrar eden, eksik ya da yanlış veri girişleri gibi durumlar nedeniyle veride işe yaramayan ve fazla sayıda, eksik, aykırı, oluşturulacak veri madenciliği modeline uygun olmayan, tutarsız vb. özelliklerde, hatta farklı kaynaklarda kayıtların bulunması mümkündür (Larose, 2005). Bu durumda kullanılacak veriyi herhangi bir ön işleme yapmadan analize almak sonuçların doğruluğuna gölge düşürecektir. Ön işlemeden geçirilen veri setinden bu veri setini en iyi şekilde yansıtabilen örnek veri seti elde edilerek analizlerin doğruluk ve etkinliği arttırılabilirken, ihtiyaca yönelik olan örüntülerin keşfine de olanak sağlanacaktır (Zhang ve diğ., 2003).

Dasu ve Johnson (2003), veri ile çalışmadan önce çeşitli açılardan ön işlemlerin yapılması gerektiğini belirtmişlerdir. Bu işlemi gerekli kılan etmenleri ise verinin heterejon olması, kalitesi, farklı veri tiplerinin varlığı ve verinin ölçeklenebilirliği şeklinde sıralamışlardır. Heterojenlik verinin farklı kaynaklardan elde edilmesi sonucu aynı konuda bile olsa farklılık ve çeşitlilik gösterebileceğini belirtmektedir. Han ve diğerleri (2012), başarılı sonuçlar için kaliteli veri olarak ifade edilen verinin kullanılması gerektiğini ifade etmektedir. Bir veri için kalite kriterleri çeşitli araştırmacılar tarafından ortaya konmuştur. Müller ve Freytag (2005), bu kriterleri doğruluk, bütünlük, tutarlılık, hassasiyet, benzersiz olma şeklinde sıralarken; Wand ve Wang (1996), doğruluk, belirli olma, eksiksiz olma, anlamlılık şeklinde listelenmiştir. Ancak en sık ve ortak olarak ifade edilen kriterler doğruluk, güncellik, tutarlılık, eksiksiz olma, ilgili olma ve uygunluk şeklinde sayılabilir (Wang ve Strong, 1996). Veritabalarındaki veri yapısal olabileceği gibi, ses, metin, görüntü gibi yapısal olmayan veriden de meydana gelebilir. Bu durumda da yine analizlere uygun gerekli dönüşüm işlemleri yapılmalıdır. Çok büyük veri setleri ile çalışıldığında bazı analizler saatler hatta günler boyunca sürebilmektedir. Böyle bir durumda tüm veri seti ile çalışmak yerine veri setini en iyi temsil eden belirli bir ölçeğe göre bir örnek veri kümesi seçilerek gerekli analizler yapılabilmektedir.

Veri ön işleme verinin bütünleştirilmesi, verinin temizlenmesi, verinin dönüştürülmesi ve verinin indirgenmesi olmak üzere dört alt adımdan meydana gelmektedir.

- **Veri Bütünleştirme:** Veri madenciliği için kullanılacak veri setinin tek bir kaynaktan temin edilmesi gerekmemektedir. Ancak farklı kaynaklardan gelen verinin farklılık göstereceği düşünülürse bu veri kümelerinin bir arada birbiri ile uyumlu hale getirilmesi gerekmektedir. Farklı kaynaktaki verinin tek bir fiziki kaynakta bir araya getirilmesi veri konsolidasyonu, farklı kaynaklarda yer alan verinin kopyasının oluşturulması veri yayınımı ve farklı kaynaklardaki verinin sanal olarak bir araya getirilmesi veri federasyonu olmak üzere veri bütünleştirme bu üç yöntem ile gerçekleştirilmektedir (Akpınar, 2014).

- **Verinin Temizlenmesi:** Daha önce de bahsedildiği üzere çeşitli hatalardan ötürü veri gürültülü, eksik veya tutarsız olabilmektedir. Verinin temizlik aşaması verinin bu tarz olumsuz özelliklerinden kurtularak kalitesi arttırılmış bir veri setine dönüştürülmesi amaçlanmaktadır. Verinin eksik olması özniteliklere karşılık gelen değerlerin olmamasını, gürültülü veri kullanıcı giriş hataları veya sistemsel hatalardan kaynaklı olarak meydana gelen hatalı öznitelik değerlerini ifade etmektedir. Örneğin; bir müşteri veri setinde müşteriye ait yaş öznitelik değeri 250 olarak gözüküyorsa bir insanın yaşı 250 olamayacağından bu bir gürültülü veridir. Benzer şekilde müşterinin yaş öznitelik değeri 25 iken doğum tarihi öznitelik değeri 20.05.1982 ise aynı müşteriye ait birbiri ile ilişkili öznitelikler arasında tutarsızlığın olduğu görülmektedir.

Tüm bu problemlerin giderilmesi için çeşitli yöntemlerden yararlanılmaktadır. Eksik verilerin tamamlanması için kullanılan yöntemleri Brown ve Kros (2003), tüm verinin kullanılması, eksik değer içeren kayıt veya özniteliklerin silinmesi, eksik değerlerin yerine değer atama ve model tabanlı yaklaşımlar olmak üzere dört kategori altında toplamıştır. Buna göre eksik veri problemi aşağıdaki yöntemler ile çözülebilmektedir (Grzymala-Busse ve Hu, 2000; Brown ve Kros, 2003; Larose, 2005; Akpınar, 2014):

- Veri seti içerisindeki eksik öznitelik değerine sahip kayıtlar veya öznitelik alanları çıkartılabilir.
- Eksik değerler yerine yeni değerler girilebilir.
    - Veriyi analiz edecek uzman tarafından belirlenen değer girilebilir.
    - Sayısal veri tipleri için o özniteliğe ait tüm değerlerin ortalaması, kategorik veri tipleri için en çok tekrar eden değer girilebilir.
    - Entropi tabanlı çalışan karar ağacı yöntemleri kullanılabilir.
    - Öznitelik değer dağılımı çıkartılarak bu dağılıma göre rastgele değer belirlenebilir.

- Maksimum olasılık tahmini, Monte Carlo yöntemleri gibi model tabanlı yaklaşımlar uygulanabilir.

Bir diğer veri temizleme işlemi aykırı değer analizidir. Aykırı değer, veri seti içerisinde belirli bir ölçüye göre diğer öznitelik değerlerin açıkça farklılık gösteren, diğer öznitelik değerlerine göre beklenenin dışında değerlere sahip değer olarak tanımlanmaktadır (Bakar ve diğ., 2006). Aykırı değer analizi ile aykırılık gösteren bu değerlerin ortaya çıkartılması amaçlanmaktadır. Aykırı değer analizi sınıflandırma problemine benzemek ile beraber analiz edilen veri tabanı kayıtları içerisinde aranılan aykırı değer sayısı oldukça az olması ile sınıflandırma probleminden ayrılmaktadır (Petrovskiy, 2003). Aşırı değer analizi, olasılık ve istatistiksel yöntemler, lineer modeller, yakınlık tabalı modeller gibi çeşitli yöntemler aykırı değer analizi için kullanılmak ile beraber bu modellerin hangisinin seçileceği konusunda veri tipi, verinin büyüklüğü, ilişkili aykırı örneklerin varlığı, yorumlanabilirlik gibi özellikler önem arz etmektedir (Aggarwal, 2015).

Veri tutarsızlığı farklı veri kaynaklarından elde edilen verinin hetereojen yapısı ve kaynaklardan gelen tablolardaki uyuşmazlıklar nedeniyle aynı nesne için birbiri ile uyuşmayan değerlerin ortaya çıkması olarak tanımlandığı gibi farklı kaynaklardan gelen veri bütünleştirildikten sonra veriyi tasvir etme noktasında da tutarsızlıkların olabileceği söylenebilir (Anokhin ve Motro, 2001). Veri seti içerisinde tutarsızlığın ortaya çıkartılması ve çözümü için çeşitli yöntemler önerilmek ile beraber en temel düzeyde ortalama, medyan, mod, standart sapma gibi istatistiksel yöntemler ile tutarsızlıkların saptanması mümkündür (Akpınar, 2014).

- **Verinin dönüştürülmesi:** Veri seti içerisinde yer alan her veri nesnesi için o nesneye ait öznitelik değerinin dönüştürülmesi işlemidir. Veri dönüştürme yöntemleri şu şekilde özetlenebilir (Tan ve diğ., 2006; Gezer, 2016):
  - Gürültülü veriden kurtulmak üzere kutulama, regresyon ve kümeleme tekniklerinin kullanılması.
  - Var olan niteliklerin kulanılarak analizlere uygun yeni öznitelik alanlarının oluşturulması.
  - Öznitelik değerlerinin özetlenerek ifade edilmesi (satış rakamlarının günlük yerine aylık ortalamasının alınması, benzer şekilde hava sıcaklılarının günlük yerine haftalık ortalamasının alınması gibi).

- Veri dönüşümünün uygulanması.
- Sayısal değerlere sahip öznitelik alanlarının bu değerlerini kategorik hale getirmek için ayrıklaştırma işleminin yapılması.

Ayrıklaştırma işlemi, sürekli olan öznitelik değerlerinin kategorik değerlere dönüştürülmesi bir başka ifadeyle veri kaybının en az olacak şekilde sürekli olan öznitelik değerlerinin sonlu komşu aralıklarına dönüştürülme işlemidir (Jin ve diğ., 2007). Bir müşteri veri setinde yer alan yaş nitelik alanında her müşteriye ait ayrı ayrı yaş değeri yerine "genç-orta yaşlı-yaşlı" olacak şekilde üç kategori oluşturularak müşterinin yaşı hangi ategoriye denk geliyorsa o kategori değerini yazmak örnek olarak verilebilir (Koçoğlu, 2012).

Veri setinde yer alan öznitelik alanlarına ait değerlerin aralığı değişiklik gösterebilir. Örneğin; bir öznitelik alanının aldığı değerler 0-100000 olabileceği gibi diğer bir öznitelik alanına ait değerler 0-1 arasında değişebilir. Büyük aralıkta değişen değerlere sahip öznitelikler analiz sonuçlarını, özellikle uzaklık ölçüsü kullanan modellerde kendi yönlerinde etkileyebilirler. Dolayısıyla böyle bir durumda geniş aralıkta değerler alan özniteliğin baskınlığını engellemek üzere veri seti üzerinde düzenlemeye gitmek gerekmektedir. Veri seti içerisinde yer alan öznitelik değerlerinin aynı aralığa denk gelecek şekilde dönüştürülmesi işlemine veri dönüştürme (normalizasyon) denir. En sık kullanılan veri dönüşümü yöntemleri doğrusal veri dönüşümü ve z-skor veri dönüşümüdür.

Doğrusal veri dönüşümü işleminde tüm değerler [0-1] aralığına taşınmaktadır. Veri seti içerisindeki en küçük değer 0'a eşitlenirken en büyük değer 1'e eşitlenmektedir. Veri seti içerisindeki diğer değerlerin [0-1] aralığındaki değerlere dönüştürülmesi için denklem (2.1) uygulanmaktadır.

$$x_{\text{normalDeger}} = \frac{x - x_{min}}{x_{maks} - x_{min}} \tag{2.1}$$

Z-skor veri dönüşümü yönteminde ise ortalama değer (μ) ve standart sapma (σ_x) değeri kullanılarak özniteliklere ait tüm değerler aynı aralık içerisinde yer alacak şekilde dönüştürülür. Bu veri dönüşümü yöntemi için denklem (2.2)'den yararlanılmaktadır.

$$x_{\text{normalDeger}} = \frac{x - \mu}{\sigma_x} \tag{2.2}$$

- **Verinin İndirgenmesi:** Bazı analiz çalışmalarında tüm veri seti yerine veri setini en iyi tanımlayacak şekilde özet veri seti kullanılmaktadır. İndirgeme işlemi de boyutu yüksek olan veri setlerinin amaca yönelik olarak en iyi şekilde temsili için gerçekleştirilen ön işleme adımıdır (Fayyad ve diğ., 1996). Veri indirgeme için gereksiz değerlerin silinip orijinal verinin karakterinin korunduğu özniteliklerin silinmesi, kayıtların silinmesi ve özniteliklere ait değerlerin silinmesi olmak üzere üç farklı yöntem kullanılmaktadır (Kantardzic, 2011).

Modelleme: Benzer problemler için farklı veri madenciliği yöntemleri kullanılarak çeşitli modeller oluşturulabilmektedir (Wirth ve Hipp, 2000). Bu aşamada çeşitli algoritmalar kullanılarak modeller oluşturulmakta, modeller uygulanarak parametreleri en iyi şekilde sonuç vermek üzere optimize edilmeye çalışılmaktadır (Azevedo ve Santos, 2008). Kullanılacak yöntemler amaca yönelik olarak sınıflandırma, kümeleme, birliktelik kuralları işlevlerini yerine getirecek şekilde seçilmektedir. Modellerin ileride performans değerlendirilmesinin yapılabilmesi için ayrıca bu aşamada eğitim ve test verileri de oluşturulmakta, çeşitli performans ölçüleri hesaplanmaktadır.

Değerlendirme: Bu aşamada, bir önceki aşamada oluşturulan farklı modellerin, çeşitli performans geçerleme ve değerlendirme yöntem ve ölçülerine göre değerlendirilmesi yapılmaktadır. Hangi modelin en iyi performansı sergilediği veya amaca hangi modelin uygun olduğu değerlendirilerek ilk adımda belirlenen problemin çözümü için önerilen modele karar verilmektedir. Değerlendirme için geliştirilen çeşitli yöntem ve ölçüler olmak ile beraber bu yöntem ve ölçülerin açıklamalarına 2.5 ve 2.6 başlıklarında yer verilmiştir.

Uygulama: Değerlendirme sonucunda farklı modeller arasında en iyi performansa sahip model seçilerek, problemin çözümü ortaya konmaya çalışılmaktadır. Yani modelin aktif olarak kullanılmaya başladığı evredir.

2.3.4. Veri Madenciliğinin İşlevleri

Veri madenciliği yöntemleri görevlerine göre tanımlayıcı ve tahminleyici olmak üzere iki ana kategori altında sınıflandırılmaktadır. Tanımlayıcı modeller, herhangi bir ön tez olmadan verinin içerisinde gizli olan örüntü ve bilginin aranmasını, bu örüntü ve bilginin ortaya çıkartılmasını, eldeki verinin tanımlanmasını ve veri içerisindeki yapının kullanıcı tarafından anlaşılmasını sağlamaktadır. Tahminleyici yöntemler ise veri ile bir model oluşturup bu modeli kullanarak geleceğe yönelik çıkarım yapılmasını, bir özelliğin diğer özellikler aracılığı ile

tahmin edilmesini sağlamaktadır (Singh ve Chaukan, 2009; Han ve diğerleri, 2012; Kantardzic, 2011; Gorunescu, 2011; Chaudhary, 2015).

Farklı amaçlara yönelik olarak veri madenciliğinin tanımlanan çeşitli işlevleri söz konusudur. Bu işlevler arasında sınıflandırma, regresyon, birliktelik kuralları, kümeleme, aykırı değer analizi, ardışık zaman örüntüleri sayılabilir (Fayyad ve diğ., 1996; Weiss ve Indurkhya, 1998; Shaw ve diğ., 2001; Kantardzic, 2011; Padhy ve diğ., 2012; Chen ve diğ., 2015; Chaudhary, 2015). Bu işlevler göz önünde bulundurulduğunda özellikle sınıflandırma, kümeleme ve birliktelik kurallarının veri madenciliğinin temel işlevleri olduğu söylenebilir.

2.3.4.1. Sınıflandırma

Sınıflandırma yöntemleri sınıf değerleri belirli olan örnekleri içeren veri seti kullanılarak tahminleyici bir modelin oluşturulmasını ve sınıfı belirli olmayan yeni gelecek örneklerin hangi sınıfa ait olduğunun belirlenebilmesini sağlamaktadır. Sınıflandırma yöntemleri bu tez çalışmasında ele alınan problemin çözümünde kullanılan ana işlev olması sebebiyle konu ile ilgili ayrıntılı açıklamalara Bölüm 2.4'te yer verilmiştir.

2.3.4.2. Kümeleme

Kümeleme işlevinde, veri seti içerisindeki farklı gruplaşmalar tespit edilip, benzer özellik gösteren kayıtlar bir araya getirilerek kümeler oluşturulmaktadır (Ahmed, 2004). Kümeleme yöntemleri veri uzayının tanımlanmasına yardım ederek, veri içerisinde gizli kalmış kümelerinin keşfedilmesine olanak sağlamaktadır (Giraud-Carrier ve Povel, 2003). Kümeleme sonucunda heterojen bir yapı gösteren veri setinden, benzer özellikli veri nesnelerinin bir araya getirilmesi ile homojen alt veri kümelerine elde edilmektedir. Kümeleme yöntemleri ile elde edilen kümeler bir sınıf etiketi olarak kabul edilebilmekte bu sayede sınıf etiketi olmayan veri seti için sınıf etiketi oluşturulabilmekte ve daha sonrasında sınıflandırma yöntemleri uygulanabilmektedir. Kümeleme yöntemleri gerçekleştirilirken kümeler arası minimum, küme içi maksimum benzerlik ilkesi kullanılır (Chen ve diğ., 1996). Bu ilke ile beraber aşağıdaki kurallar göz önünde bulundurulur:

- Aynı kümede yer alan veri nesnelerinin benzerliğini maksimize etme,
- Aynı kümede yer alan veri nesnelerinin diğer kümelerde yer alan veri nesnelerine benzerliğini minimize etme,

- Veri nesnelerinin özelliklerine göre veri nesneleri arasındaki benzerliklerin ortaya çıkartılarak benzerliğe dayalı kümeleri elde etme.

Kümeleme yönteminin kullanıldığı alanlardan bir tanesi pazarlamadır. Örneğin; işletmeler ürün ve hizmet sunum stratejilerini belirlemek üzere müşterilerini demografik, sosyal, davranışsal farklılıklar gibi özelliklerine göre gruplama yoluna gitmektedir. Bu doğrultuda da kümeleme yöntemlerinden yararlanılmaktadır (Koçoğlu ve Özcan, 2016).

2.3.4.3. Birliktelik Kuralları

Birliktelik kuralları tahmin yapmaktan ziyade veri içerisindeki ilişkiyi belirlemeye yönelik veri madenciliği yöntemidir. Bu yöntemde veri setinde yer alan veri nesnelerinin eş zamanlı gerçekleşme durumlarından yola çıkılarak kurallar dizisi oluşturulmaktadır. Çeşitli öğelerden oluşan işlemlerin meydana getirdiği bir veri seti söz konusu olsun. *X* ve *Y* bu işlemler içerisinde yer alan öğeler olmak üzere bir birliktelik kuralı *X => Y* şeklinde ifade edilirken bu gösterim, "*X öğesinin yer aldığı öğe grubu söz konusu ise Y öğesinin de bu öğe grubunda bulunma eğilimi vardır*" anlamını taşımaktadır (Agrawal ve diğ., 1993). Örneğin; çocuk bezi satışı ile bira satışının beraber gerçekleşmesi durumunun tanımlanması birliktelik kurallarının sonucudur. Birliktelik kurallarında destek ve güven değerleri önemli büyüklüklerdir. Destek değeri kuralın hangi sıklık ile gerçekleşeceğini belirtirken, güven değeri elde edilen kuralın kabul edilebilirlik durumunu ifade etmektedir. Yani %5 oranındaki bir destek değeri, analiz edilen tüm alışverişlerin %5'inde çocuk bezi ve bira ürünlerinin birlikte satıldığını belirtirken, %72 oranındaki güven değeri çocuk bezi ürününü satın alan müşterilerinin %72'sinin aynı alışverişte bira ürününü de satın aldığını göstermektedir.

Ardışık zamanlı örüntü keşfi birliktelik kuralları ile aynı amacı taşımak ile beraber birbirini takip eden ilişkili durumlar arasında zaman boyutu söz konusudur. A tipinde ameliyat olan bir hastanın bir hafta içerisinde B tipinde bir ağrıya sahip olacağı durumunun tespiti ardışık zaman örüntü keşfinin bir sonucudur.

2.3.4.4. Diğer İşlevler

Regresyon, aslında bir istatistiksel tahmin metodu olup veri madenciliğinde de kullanılmaktadır. Regresyon bağımsız bir dizi değişkenin başka bir değişkeni (bağımlı değişken) modellemek üzere kullanıldığı bir süreçtir (McCarty ve Hastak, 2007).

Sınıflandırmaya temelde oldukça benzemek ile beraber regresyon yöntemleri ile sınıf etiketi yerine sürekli sayısal değerler elde edilmektedir.

Aykırı değer, belirli bir ölçüye göre veri setinde yer alan diğer veri değerleri içerisinde farklılık gösteren değer olarak tanımlanmaktadır. Bu değer bazı durumlarda analiz sonuçlarını etkilememek üzere veri setinden atılırken, bazı durumlarda ise veri setinin anormal davranışını gösteren önemli bir bilgiyi de içerebilmektedir (Aggarwal ve Yu, 2005). Aykırı değer analizi sahtekârlıkların ya da saldırıların tespiti gibi konularda aktif olarak kullanılan bahsi geçen aykırı durumların ortaya çıkarılmasını amaçlayan bir veri madenciliği yöntemidir.

2.4. SINIFLANDIRMA

Veri madenciliğinin en önemli işlevlerinden birisi sınıflandırmadır. Sınıflandırma, veri seti içerisindeki örnekleri daha önce tanımlanmış olan ve sınıf olarak isimlendirilen gruplar ile eşleştirme, bir başka ifade ile en uygun olan sınıfa dâhil etme işlemidir (Dunham, 2003).

Sınıflandırma modelinin oluşturulabilmesi için kullanılacak veri setinde yer alan örneklerin her birinin hangi sınıfa ait olduğunu gösteren ve "sınıf etiketi" olarak adlandırılan nitelik alanı mevcut olmalıdır. Bu veri seti, eğitim veri seti olarak tanımlanmaktadır. Örneğin; *X*, veri setinde yer alan bir örnek (kayıt) olmak üzere $X=(x_1, x_2, ..., x_n)$ *n* boyutlu bir öznitelik vektörü ile ifade edilmektedir. Buradaki her bir boyut, veri tabanında yer alan her bir nitelik alanında *X* örneği için karşılık gelen değeri göstermektedir (Han ve diğ., 2012). Hastane veri tabanından yola çıkılarak hastalara ait değerler ile bu durum örneklenecek olursa, tek bir hastaya ait şeker düzeyi, troid hormonu düzeyi, yaşı, günlük sigara tüketim miktarı, ailedeki kanser hastalık durumu vb. değerler o hastanın öznitelik vektöründeki her bir boyutuna karşılık gelen değerlerdir. *Y* ise *X* örneğine ait sınıf niteliğini (etiketini) göstermek üzere, bahsi geçen hastanın kanser olma durumu (kanser/kanser değil) *Y* sınıf etiketine karşılık gelen değerdir.

Sınıflandırma yöntemi öğrenme ve sınıflandırma olmak üzere iki ana alt süreci kapsamaktadır. Öğrenme adımında eğitim veri setinin analizi gerçekleştirilerek, veri setinde yer alan her bir sınıf için bu sınıflara uygun "sınıflandırıcı" olarak da ifade edilen bir model oluşturulmaktadır. Yani bu adım *X*, *n* boyutlu bir öznitelik vektörü olmak üzere $Y=f(X)$ olacak şekilde bir *f* fonksiyonu belirlenmesi şeklinde değerlendirilebilir (Han ve diğ., 2012). Öğrenme adımını etkileyen faktörleri Balaban ve Kartal (2015), veri seti, analiz sonucunda önem arz eden

nitelikler, öğrenme stratejisi, kullanılacak algoritma ve parametreleri şeklinde dört başlık altında toplamıştır.

Sınıflandırma adımında ise oluşturulan modelin performansını denetlemek üzere ayrılan, örneklerin "sınıf etiketi"nin belirsiz olduğu (ya da hangi sınıfa ait olduğunun bilinmediği) ve test veri seti olarak adlandırılan veri setindeki örneklerin kurulan bu model ile hangi sınıflara ait olduğu tahmin edilmektedir (Chen ve diğ., 1996). Tahmin performansına göre de modelin sınıflandırma için uygunluğuna karar verilmektedir. Sınıflandırma yöntemlerinin veri setindeki örneklerin hangi sınıfa dahil olduğunun tahmini sürecinde kural oluşumundan da yararlanılmaktadır. Karar ağacı algoritmaları ile eğitim veri seti ile kural yapısı oluşturabilmekte, yeni bir durum söz konusu olduğunda bu kurallara uygun olarak verilebilecek kararın ne olması gerektiği tahmin edilebilmektedir.

Sınıflandırma için kullanılan yöntemler arasında En Yakın Komşu Algoritması, Sade Bayes Algoritması, Karar Ağacı Algoritmaları, Destek Vektör Makineleri sayılabilir. Aşağıda bu yöntemlere ait örnek algoritmaların açıklamaları verilmiştir.

2.4.1. K-En Yakın Komşu Algoritması

k-en yakın komşu (*k*-nearest neighbour) algoritması, gözlemler arası uzaklığa bağlı olarak sınıfların belirlenmesine dayalı bir algoritmadır. Temel olarak, yeni bir örneğin belirlenen uzaklık ölçüsüne göre veri seti içerisindeki diğer örneklere uzaklıkların ölçülmesi ve bu yeni örneğe en yakın *k* tane örneğin sınıfına bağlı olarak yeni örneğin sınıfı belirlenmeye çalışılmaktadır. Uzaklık ölçüsü örnekler arası benzerliğin ölçüsü olarak kullanılmakta, birbirine yakın gözlemlerin benzer olduğu, uzak olanların ise benzer olmadığı kabul edilmektedir.

k-en yakın komşu algoritması ile model kurulurken aşağıdaki soruların cevapları önem taşımaktadır (Larose, 2005):

- Kaç tane komşu göz önünde bulundurulacak yani *k* kaç seçilecektir?
- Hangi uzaklık ölçüsü kullanılacaktır?
- Birden fazla sayıdaki gözlemden gelen enformasyon nasıl birleştirilecektir?
- Her örneğin modeldeki etkisi aynı mı olmalıdır? Bazı örneklerin ağırlığı daha fazla mı olmalıdır?

Algoritma için *k* değerinin belirlenmesi önem taşımak ile beraber *k* değerinin büyük seçilmesi birbirine benzemeyen örneklerin bir araya getirilmesine, *k* değerinin küçük seçilmesi ise birbirine benzeyen örneklerin dahi ayrı sınıflara dâhil edilmesine neden olabilmektedir (Silahtaroğlu, 2008).

k-en yakın komşu algoritması için en önemli etkenlerden bir diğeri hangi uzaklık ölçüsünün kullanılacağıdır. Çünkü; algoritmanın başarısının bir ölçüsü olan doğruluk, farklı örnekler arasındaki uzaklıkların hesabı için kullanılan uzaklık ölçülerine de bağlıdır (Weinberger ve Saul, 2009). Veri setinde yer alan örneklerin benzerlik veya farklılığını belirleyebilmek için problemin yapısı ve verinin tipine göre çeşitli uzaklık ölçüleri geliştirilmiştir (Koçoğlu ve Özcan, 2016).

Uzaklık ölçülerinden önce uzaklık fonksiyonu tanımı ve özelliklerine yer vermek gerekmektedir. Bir uzaklık fonksiyonu *x, y* ve *z* koordinat düzleminde noktalar olmak üzere *d(x,y)* şeklinde gösterilir. Denklem (2.3) uzaklığın her zaman pozitif olacağını, koordinat üzerindeki herhangi iki noktanın uzaklığının sıfır olma durumunun ise ancak ve ancak bu iki noktanın aynı olması durumunda geçerli olduğunu, denklem (2.4) simetrikliği, denklem (2.5) ise bir noktanın (*x*) diğer iki noktaya (*y* ve *z*) uzaklıklarının toplamının diğer iki nokta (*y* ve *z*) arasındaki uzaklıktan küçük olamayacağını belirtmektedir.

$$d(x, y) \geq 0,\ d(x, y) = 0 \leftrightarrow x = y \tag{2.3}$$

$$d(x, y) = d(y, x) \tag{2.4}$$

$$d(x, z) \leq d(x, y) + d(y, z) \tag{2.5}$$

Veri setinde yer alan her bir örnek, tüm nitelik alanlarına karşılık bir gözlem değeri almaktadır. Bu nitelik alanı değerleri göz önüne alındığından her bir örnek bir satır olmak üzere tüm gözlemleri ve dolayısıyla tüm veri setini bir matris olarak ifade etmek mümkündür. *n* sayıda değişken ve *p* sayıda gözlemden oluşan veri matrisi *X* olsun. Bu matrise göre birinci gözlemin değerleri ($x_{11},x_{12},x_{13},...,x_{1n}$) ve ikinci gözlemin değerleri ($x_{21},x_{22},x_{23},...,x_{2n}$) şeklinde ifade edilsin. Aşağıda veri seti matrisi *X* ve bu veri setine ait uzaklık matrisi *D* verilmiştir.

$$X=\begin{bmatrix} x_{11} & x_{12} & x_{13} & . & . & x_{1n} \\ x_{21} & x_{22} & x_{23} & . & . & x_{2n} \\ x_{31} & x_{32} & x_{33} & . & . & x_{3n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{p1} & x_{p1} & x_{p1} & . & . & x_{pn} \end{bmatrix} \quad D=\begin{bmatrix} 0 & & & & & \\ d(2,1) & 0 & & & & \\ \mathrm{d}(3,1) & \mathrm{d}(3,2) & 0 & & & \\ . & . & . & . & & \\ . & . & . & . & . & \\ \mathrm{d}(\mathrm{p},1) & \mathrm{d}(\mathrm{p},2) & . & . & . & 0 \end{bmatrix}$$

En çok kullanılan uzaklık ölçülerinden biri *Öklid Uzaklığı*dır. Genelleştirilmiş hali ile Öklid uzaklığı denklem (2.6)'da verilmiştir.

$$d(i,j)=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2} \tag{2.6}$$

Bir diğer uzaklık ölçüsü Minkowski uzaklığıdır. Bu uzaklık ölçüsü denklem (2.7) ile hesaplanmaktadır. Bu denklemde m=2 yazılırsa Öklid, m=1 yazılırsa Manhattan uzaklıkları elde edilir.

$$d(\mathrm{i},\mathrm{j})=\sqrt[m]{\sum_{k=1}^{n}(x_{ik}-x_{jk})^m} \tag{2.7}$$

Belirtilen bu uzaklık ölçüleri veri setinin sadece nümerik değerler alan nitelik alanlarından oluşması durumunda kullanılmaktadır. Eğer nitelik alanları kategorik veya nümerik ve kategorik olacak şekilde karma nitelik alanı değerlerine sahip ise farklı uzaklık ölçüleri kullanılmalıdır. Tablo 2.1'de ikili nitelik değerleri için kullanılan olasılıklar verilmiştir (Sneath ve Sokal, 1963; Gower, 1971; Han ve diğ., 2012).

Tablo 2.1: İkili nitelik değerleri için olasılık tablosu.

		Nesne *j*		
		1	0	Toplam
Nesne *i*	1	*q*	*r*	*q+r*
	0	*s*	*t*	*s+t*
	Toplam	*q+s*	*r+t*	*p*

Simetrik ikili nitelik değerleri için uzaklık ölçüsü (yalın uyum katsayısı- simple matching coefficient)

$$d(i,j)=\frac{r+s}{q+r+s+t} \tag{2.8}$$

Asimetrik ikili nitelik değerleri için uzaklık ölçüsü

$$d(i, j)= \frac{r+s}{q+r+s} \tag{2.9}$$

Asimetrik ikili nitelik değerleri için benzerlik ölçüsü (Jaccard katsayısı)

$$\text{sim}_{\text{jaccard}}(i, j)= \frac{q}{q+r+s} \tag{2.10}$$

k komşu sayısı ve uzaklık ölçüsü belirlendikten sonra elde edilen sonuçlardan yeni örneğin hangi sınıfa dahil edileceği konusunda karar verilmesi gerekmektedir. Bu noktada ya tüm *k* tane komşunun sınıfı yeni örneğin sınıfını belirlemede eşit ya da bazı komşu sınıfların daha etkin olması gerekmektedir. Eğer tüm komşu sınıflar eşit oranda etkili olacaksa belirlenen *k* tane komşu örnek içerisinde en sık tekrarlanan sınıf yeni örneğin sınıfı olacaktır. Bu durumdaki sınıf belirlemeye basit ağırlıksız oylama veya çoğunluk oylaması adı verilmektedir. Eğer *k* tane sınıfın tekrara bakılmaksızın yeni örnek üzerindeki etkisi hesaplamaya dâhil edilmek istenirse bu durumda yeni örneğe olan uzaklıkların tersi cinsinden bir ağırlık değeri elde edilerek yeni örnek en büyük ağırlık değerine sahip sınıfa dâhil edilir. Bu şekildeki sınıf belirlemeye ise ağırlıklı oylama adı verilir (Larose, 2005; Özkan, 2008).

k-en yakın komşu algoritması aşağıdaki şekilde özetlenebilir:

- *k* komşu sayısı belirlenir.
- Problemin cinsine ve veri tipine göre uzaklık ölçüsü belirlenir.
- Yeni örneğin veri seti içerisindeki diğer örneklere uzaklığı (benzerliği) belirlenen uzaklık ölçüsüne göre hesaplanır.
- En yakın *k* tane komşu seçilir.
- Belirlenen oylama yöntemine göre yeni gelen örneğin sınıfı belirlenir.

2.4.2. Sade Bayes Algoritması

Sade Bayes algoritması, Bayes teorimini temel alan istatistik ve olasılık tabanlı bir sınıflandırma yöntemidir. Sade Bayes algoritmasının detaylarına geçmeden önce temel kavramlar ele alınmalıdır.

Sonuçları rastgele meydana gelen bir deneyin, tüm sonuçlarının oluşturduğu örnek uzayın her bir alt kümesi olay olarak ifade edilmektedir. Bir olayın meydana gelme olasılığı başka bir olayın meydana gelme durumuna bağlı ise bu durumda koşullu olasılıktan bahsedilmektedir. Koşullu olasılık denklem (2.11) ile gösterilir. Burada *B* olayının meydana gelme durumu *A* olayının meydana gelmesi koşuluna bağlıdır ve *B* olayının *A* olayına göre koşullu olasılığı şeklinde ifade edilir (*P(A)>0, P(B)>0*) (Taha, 2014).

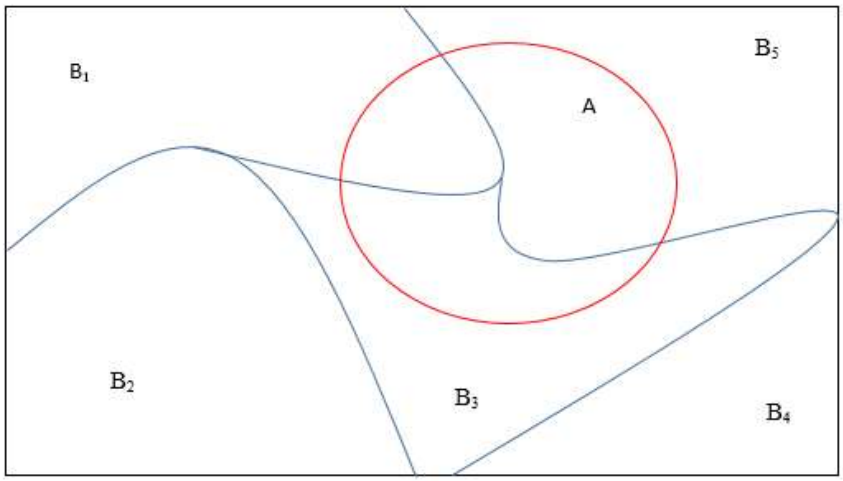
$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{2.11}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{2.12}$$

$$P(A \cap B) = P(A \mid B)\, P(B) \tag{2.13}$$

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)} \tag{2.14}$$

Şekil 2.9'da birbirinden bağımsız B_i olayları ($i=1,2,..,5$) ve *A* olayı görülmektedir. *A* olayı B_i olayları ile ifade edilebilir.



Şekil 2.9: A, B1, B2, B3, B4 ve B5 olayları.

$$A = (B_1 \cap A) \cup (B_2 \cap A) \cup (B_3 \cap A) \cup (B_4 \cap A) \cup (B_5 \cap A) \tag{2.15}$$

$$P(A)= P(B_1 \cap A) + P(B_2 \cap A) + P(B_3 \cap A) + P(B_4 \cap A) + P(B_5 \cap A) \quad (2.16)$$

Denklem (2.16)'daki gösterime Toplam Olasılık Teoremi denir. Denklem (2.14) ve (2.16)'dan yararlanılarak denklem (2.17) elde edilebilir.

$$P(B_1 \mid A) = \frac{P(A \mid B_1)\, P(B_1)}{P(B_1 \cap A)+P(B_2 \cap A)+P(B_3 \cap A)+P(B_4 \cap A)+P(B_5 \cap A)} \quad (2.17)$$

Bu sonuç *n* olay için genelleştirilirse denklem (2.18) elde edilir.

$$P(B_i \mid A) = \frac{P(A \mid B_i)\, P(B_i)}{\sum_{i=1}^{n} P(B_i \cap A)} \quad (2.18)$$

Bayes teoremi A olayının gerçekleşmesi durumuna göre B olayının gerçekleşme olasılığı ve tersine B olayının gerçekleşmesi durumuna göre A olayının gerçekleşme olasılığı arasındaki bağıntıyı ifade etmektedir. Bu bağıntı denklem (2.18)'de görülmektedir.

Tüm bu bağıntılar bir sınıflandırma modeli oluşturmak için kullanılmakta, yeni gelen örneğin veri setinde yer alan sınıflardan herhangi birine girme olasılığı hesaplanmaktadır. Bu yaklaşım Sade Bayes algoritması olarak adlandırılmıştır. $X=\{x_1, x_2, ..., x_n\}$ olmak üzere *n* tane nitelik alanı değerine sahip sınıfı belirlenmek istenen örnek, $C=\{C_1, C_2, ... , C_m\}$ ise veri setinde yer alan sınıflar olsun. *X*'in hangi sınıfa ait olduğu bilinmediğinden herhangi bir sınıfa ait olma olasılığı üzerine hesaplamalar yapılarak *X*'in sınıfı belirlenmeye çalışılmaktadır. $P(C_i \mid X)$ ($i=1,2,..., m$) *X* örneğinin C_i sınıfından olma olasılığını, $P(X \mid C_i)$ ise C_i sınıfındaki bir örneğin *X* olma olasılığını göstermektedir. Bayes teoreminden denklem (2.19) yazılabilir (Murphy, 2006).

$$P(C_i \mid X)= \frac{P(X \mid C_i)\, P(C_i)}{P(X)} \quad (2.19)$$

Sınıflar ne olursa olsun $P(X)$ tüm sınıflar için eşit olarak kabul edilmektedir. Bunun yanı sıra *X* örneğinin tüm nitelik değerlerinin birbirinden bağımsız olduğu kabul edilerek bağıntı denklem (2.20)'deki şekilde sadeleştirilir (Leung, 2007).

$$P(X \mid C_i)=\prod_{k=1}^{n} P(x_k \mid C_i) \quad (2.20)$$

Önemle durulması gereken bir diğer konu kategorik ve nümerik veri tipindeki nitelik değerleri için farklı olasılık fonksiyonlarının kullanıldığıdır. Kategorik nitelik değerleri için frekans tablosundan yararlanılırken nümerik nitelik değerleri için denklem (2.21)'de yer alan olasılık yoğunluk fonksiyonu kullanılmaktadır.

$$f(x_k) = \frac{e^{-\frac{(x_k-\mu_{C_i})^2}{2\sigma_{C_i}^2}}}{\sigma_{C_i}\sqrt{2\pi}} \qquad (2.21)$$

σ: C_i sınıfının belirtilen nitelik alanı için standart sapması

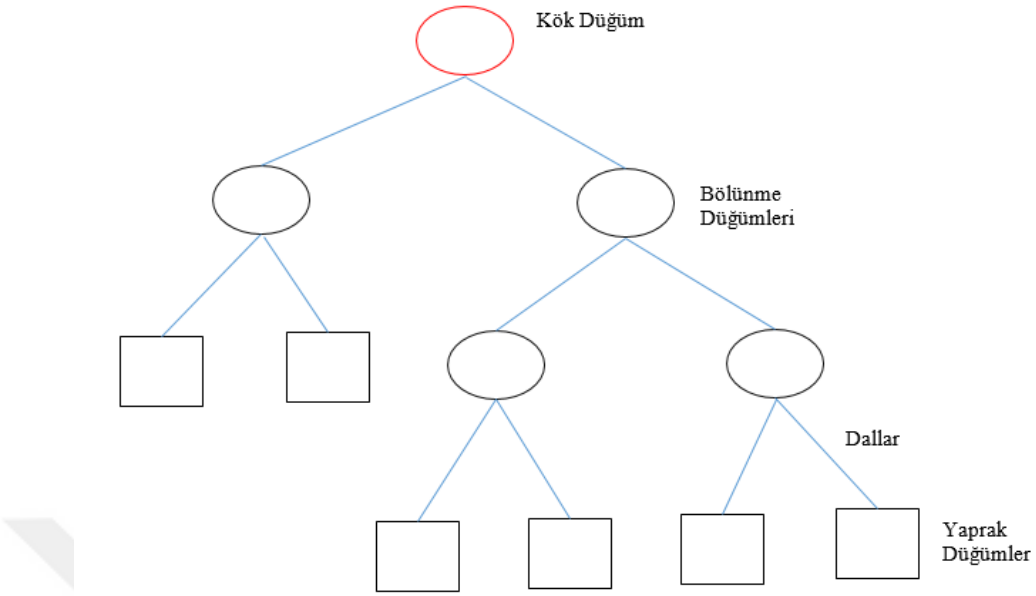
μ: Verilen nitelik alanı için sınıfın ortalaması

Sonuç olarak, denklem (2.19)'da belirtilen pay kısmının hangi veri tipinde nitelik değeri olursa olsun en büyük değer olması aranmaktadır. Buna göre en büyük değer hangi sınıf için elde edildiyse verilen *X* örneğinin o sınıfa ait olduğu söylenebilir.

$$C_{hedef} = \text{maks}\ \{ \prod_{k=1}^{n} P(x_k \mid C_i)\} \quad (i=1,2,\ldots, m) \qquad (2.22)$$

2.4.3. Karar Ağaçları

En çok kullanılan, klasik sınıflandırma yöntemlerinden bir diğeri karar ağacı algoritmalarıdır. Karar ağacı bir veri setini daha küçük alt kümelere yinelemeli olarak ayıran bir sınıflandırma yöntemi olarak tanımlanmaktadır (Friedl ve Brodley, 1997). Karar ağacı algoritmaları hiyerarşik yapıdaki parametrik olmayan yöntemler olup (Kantardzic, 2011), sınıflandırma için gerekli kuralların elde edilmesine olanak sağlamaktadır. Çıkarılan kuralların daha anlaşılabilir ve yorumlanabilir olması, elde edilebilir kesin sonuçlar karar ağacı yöntemlerinin yaygın olarak kullanılmasını sağlamıştır (Mitchell, 1997).



Şekil 2.10: Karar ağacı yapısı.

Hiyerarşik yapı yöntemin isminden de anlaşılacağı üzere gerçek bir ağaç şeklindeki yapı ile uyumlu bir model oluşturmasından kaynaklanmaktadır. Elde edilen modelde kök, dallar ve yaprakların varlığı söz konusu olup yapı kök düğüm ile başlamakta, kök düğüme dallar ile bağlanan alt (bölünme) düğümler ile devam etmekte ve kendisinden sonra başka bir düğüm olmayan yaprak düğümler ile sona ermektedir. Yaprak düğümler her bir sınıf etiketini, kök düğümden yaprak düğüme kadar olan yol bir kuralı temsil etmektedir (Quinlan, 1987). Ancak bu ağaç yapısının gösterimi gerçek bir ağacın tersi şeklindedir. Yani, kök düğüm en üstte yaprak düğümler ise en altta gösterilmektedir. Şekil 2.10'da bir ağaç yapısı görülmektedir.

Karar ağacı algoritmaları ile çalışılırken aşağıdaki durumlara dikkat edilmelidir (Larose, 2005):

- Model oluşturmak için kullanılacak eğitim veri setinde sınıf etiketleri belirli olmalıdır.
- Model kurulurken kullanılacak veri seti, tüm sınıfların iyi bir şekilde tanımlanabilmesi ve kuralların çıkarılabilmesi için yeterli olmalıdır.
- Sınıf etiketleri ayrık olmalı yani belirli bir sınıfa ait olup olmadığını belirtecek şekilde sınırlandırılmış değerlere sahip olmalıdır.

Karar ağaçlarında kök düğümün belirlenerek alt dallanmaların nasıl oluşturulacağı çeşitli hesaplamalara dayanarak yapılmaktadır. Bu yöntemlerden bazıları entropi, gini endeksi, sınıflandırma hatası endeksi, en küçük kareler sapması şeklinde sayılabilir (Akpınar, 2014).

Aşağıda entropi, bilgi kazancı, kazanım oranı ve bölünme bilgisi kavramlarına yer verilmiştir (Quinlan, 1996; Hall ve Holmes, 2003; Raileanu ve Stoffel, 2004, Özkan, 2008).

T sınıf niteliği, *C*= {C_1, C_2,..., C_k}, *T* sınıf niteliğinin *k* farklı sınıfı olsun. $|C_i|$ (*i=1,...,k*), veri kümesinde C_i sınıfına ait örneklerin sayısını, $|T|$ ise veri kümesinde sınıf niteliğine sahip örneklerin sayısını göstersin. *T* sınıf niteliğine ait tüm sınıfların olasılık dağılımı şu şekilde olacaktır.

$$P_T = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \ldots, \frac{|C_k|}{|T|}\right) = (p_1, p_2, \ldots, p_k) \quad (2.23)$$

Burada p_i (*i=1,2,.., k*), *T*'de yer alan C_i sınıfındaki herhangi bir gözlemin olasılığını ifade etmektedir. Buradan hareketle *T* için, *bilgi miktarı (Info)* veya bir diğer ifadeyle *Entropi (E)* denklem (2.24) ile hesaplanmaktadır.

$$Info(T) = E(T) = -\sum_{i=1}^{k} p_i \log_2(p_i) \quad (2.24)$$

Sınıf niteliğine sahip olmayan bir *X* niteliğinin aldığı farklı değerlere göre *T* sınıf niteliği *m* alt kümeye ayrılsın. *X* niteliğindeki değerleri alan örneklerin hangi sınıfa ait olduğunu belirleyebilmek için gerekli bilgi denklem (2.25) ile hesaplanmaktadır.

$$Info_X(T) = \sum_{j=1}^{m} \frac{|T_j|}{|T|} Info(T_j) \quad (2.25)$$

Bilgi kazancına göre dallanma yapılan algoritmalarda en iyi sınıflandırmanın gerçekleştirilebilmesi için denklem (2.26) kullanılarak *Bilgi Kazancı* (*IG*) hesaplanır. En yüksek bilgi kazancı değerine sahip *X* nitelik değerine göre dallanma gerçekleştirilmektedir.

$$IG(X) = Info(T) - Info_X(T) \quad (2.26)$$

Nümerik değerler ile çalışılırken, nümerik veri tipindeki nitelik alanlarının aldığı değerler kategorik değerlere dönüştürülmektedir. Bunun için öncelikle niteliğin aldığı değerler küçükten büyüğe sıralanmakta, sonrasında bu değerlerden herhangi biri seçilerek örneklerin yer aldığı veri seti iki parçaya ayrılmaktadır. Her bir bölünme için bilgi kazancı hesaplanarak en büyük bilgi kazancını sağlayan bölünme gerçekte aranılan eşik değerini belirlemektedir.

Karar ağaçları yapısında gerçek ağaçlarda olduğu gibi bir budama süreci söz konusudur. Bazı karar ağacı algoritmalarında bölünme çok fazla sayıda olup, modelin ezbere davranmasına yani aşırı öğrenmesi (overfitting) durumuna sebep olmaktadır. Dolayısıyla budama süreci ile ağaç yapısının daha sağlıklı bir hale dönüştürülmesi için ağaç yapısının aşağılarında meydana gelen aşırı dallanmalar budanarak daha küçük bir yapı haline getirilmekte, böylece aşırı öğrenme sorununa çözüm sağlanması amaçlanmaktadır. Budama işlemi ya budanan yaprakların bir üstteki düğümlere dâhil edilmesi ile ya da budanan alt ağacın bu ağacı en çok kullanan ağacın yerine alınması ile gerçekleştirilmektedir (Silahtaroğlu, 2008; Akpınar, 2014).

Quinlan (1996), buradan hareketle yukarıda ifade edilen *Bilgi Kazancı* değeri yerine *Kazanım Oranı* (*GR*) değerinin kullanımını öne sürmüştür. Kazanım oranı denklem (2.27) ile hesaplanmaktadır.

$$GR(X) = \frac{IG(X)}{SInfo(X)} \tag{2.27}$$

$$SInfo_X(T) = -\sum_{j=1}^{m} \frac{|T_j|}{|T|} \log_2\left(\frac{|T_j|}{|T|}\right) \tag{2.28}$$

Bölünme bilgisi (*SInfo*), *X* niteliğine ait farklı *m* değeri için *T* kümesi ayrıldığında elde edilen potansiyel bilgiyi göstermektedir. Kazanım oranına göre dallanma yapılan algoritmalarda en iyi sınıflandırma için en yüksek kazanım oranına sahip nitelik değeri seçilerek dallanma gerçekleştirilir.

Karar ağaçları algoritmaları arasında ID3, CART, C4.5 ve C5.0 sayılabilir. Bu algoritmalardan ID3 bilgi kazancına göre C4.5 ve C5.0 ise kazanım oranına göre dallanma gerçekleştirmektedir. C4.5, ID3 algoritmasının eksiklerini gidermek üzere geliştirilmiş bir algoritma olup hem nümerik hem de kategorik değişkenler ile çalışmakta, kayıp değerler bu algoritma için sorun teşkil etmemektedir. C5.0 ise C4.5 algoritmasının ticari bir devamı niteliğinde olup bilgi *www.rulequest.com/* adresinde yer almaktadır (Han ve diğ., 2012). Tso ve Yau (2007), C5.0'ın C4.5'ten üstünlüklerini aşağıdaki şekilde sıralamıştır:

- Yanlış sınıflandırma maliyetleri belirtilmektedir.
- Boosting ve çapraz geçerleme gibi performansı arttıracak yöntemlere olanak sağlamaktadır.

- Kural çıkarımı daha güçlüdür.

C4.5 ve C5.0 algoritmalarının genel adımları özetlenecek olursa:

- Sınıf nitelik alanı için entropi değeri hesaplanır.
- Nümerik nitelik alanları için belirtilen kategorikleştirme işlemi uygulanarak eşik değer belirlenir. Nitelik alanları için bilgi miktarları belirlenir.
- Nitelik alanlarının bilgi kazancı ve bölünme bilgisi değerleri hesaplanır.
- Nitelik alanları için bilgi kazanım oranları hesaplanır.
- Belirlenen değerlere ve daha önce bahsi geçen dallanma kriterlerine göre ağaç yapısı kökten yapraklara doğru oluşturulur.

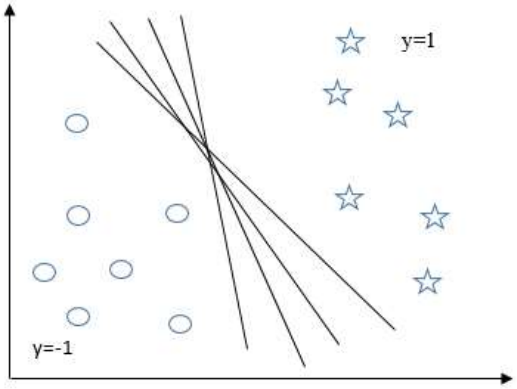
2.4.4. Destek Vektör Makineleri

Sınıflandırma probleminin çözümünün içi geliştirilen bir diğer yöntem Destek Vektör Makineleridir (SVM). Temelleri Vladimir Vapnik ve Alexey Chervonenkis tarafından Vapnik-Chervonenkis Teorisi kapsamında ortaya atılan ve 1992 yılında Berhard Boser, Isabelle Guyon ve Vladimir Vapnik tarafından geliştirilen bu yöntem birçok alanda başarı ile uygulanmaktadır. Diğer yöntemlere göre aşırı öğrenme problemine karşı daha güçlü, doğrusal olmayan sınıflandırmadaki başarı oranının daha yüksek olması ve daha fazla sayıda nitelik alanı ile çalışabilmesi nedeniyle tercih edilen bir yöntemdir.

SVM, genelleştirme kaynaklı hataların üst sınırını minimize eden *Yapısal Risk Minimizasyonu* prensibine göre çalışmaktadır (Shankar, 2002). SVM doğrusal veya doğrusal olmayan sınıflandırma problemleri için kullanılmaktadır. SVM ile doğrusal sınıflandırma probleminde sınıfları en iyi ayıracak şekilde optimum bir hiper düzlem bulunması, bu hiper düzeleme paralel düzlemlerde yer alan ve destek vektörleri olarak isimlendirilen noktalar arasındaki mesafenin maksimum olması amaçlanmaktadır (Han ve diğ., 2012). Doğrusal olmayan sınıflandırma problemlerinde, bahsi geçen amacı gerçekleştirebilmek için veri kümesi doğrusal olmayan haritalama yöntemleri ile doğrusal olarak ayrılabileceği daha yüksek boyutlu bir veri uzayına taşınmaktadır (Cortes ve Vapnik, 1995). Dönüştürme işleminde verilerin izdüşümünü taklit amacıyla çekirdek fonksiyonlardan yararlanılmaktadır (Dixon ve Candade, 2008).

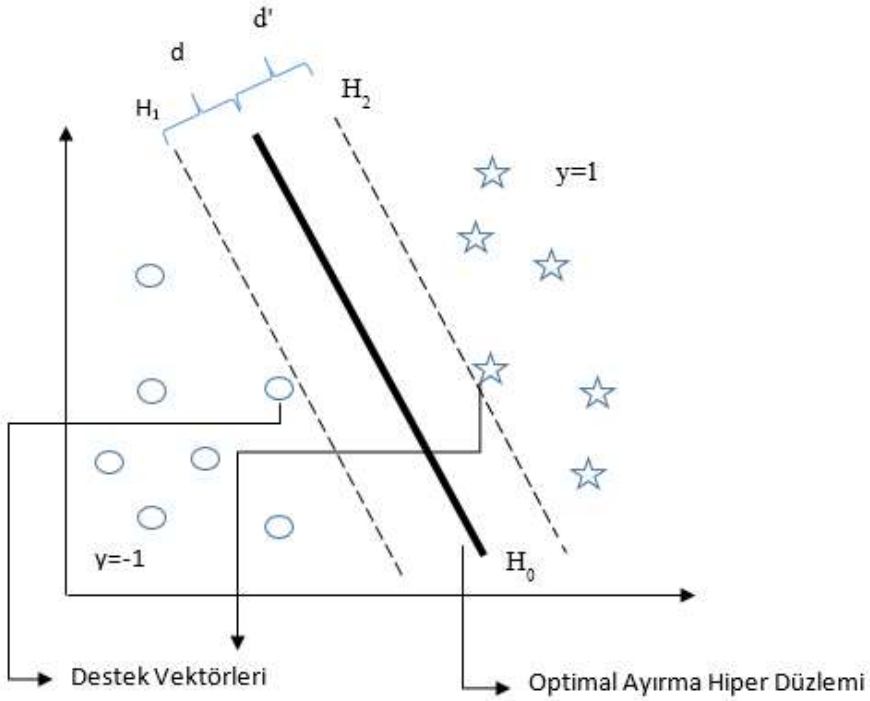
2.4.4.1. Doğrusal Sınıflandırma Problemi

Veri setinin sınıfları arasında doğrusal olarak bir ayırım yapılabilecek şekilde hiper düzlemin mümkün olduğu durumdur. *D, n* elemanlı iki boyutlu veriden oluşan bir veri kümesi olmak üzere (x_1,y_1), (x_2,y_2),.. ,(x_n,y_n) bu kümenin elemanları olsun. $y \in \{-1, +1\}$ olarak kabul edilsin. Şekil 2.11'de veri kümesinin sınıf niteliği *y*'ye göre çeşitli şekillerde ayrılabileceğinin mümkün olduğu görülmektedir.



Şekil 2.11: Doğrusal olarak ayrılabilen veri kümesinde çeşitli ayırma doğruları.

İki boyuttan çok boyutlu uzaya geçildiğinde ayırma doğrusunun yerini hiper düzlemler adı verilen yine çeşitli ayırma düzlemleri alacaktır. SVM ile amaç buradaki en optimal düzlemi belirleyebilmektir. Optimallikten kasıt herhangi bir yeni gözlemin veri kümesine dâhil olması ile bu ayırma hiper düzleminin değişmemesi, bu durumun sağlanabilir olması için de ayırma düzleminin tüm sınıfların sınır düzlemlerine en yakın noktalarına maksimum uzaklıkta olması koşuludur. Bu sınır noktaları destek vektörleri olarak adlandırılmaktadır.



Şekil 2.12: İki boyutlu uzayda ayırma doğrusu ve destek vektörleri.

Şekil 2.12'de iki sınıf arasındaki en büyük uzaklığa sahip olduğu kabul edilen sınır düzlemleri H_1 ve H_2 olmak üzere optimal ayırma düzlemi H_0'dır. $W=\{w_1, w_2,...,w_n\}^T$ ağırlık vektörü ve b sabit olmak üzere belirtilen sınır ve optimal ayırma düzlemleri denklem (2.29), (2.30) ve (2.31)'deki gibi gösterilmektedir (Hearst, 1998).

$$H_0: W^T X + b = \sum_{i=1}^{n} w_i x_i + b = 0 \quad (2.29)$$

$$H_1: W^T X + b = \sum_{i=1}^{n} w_i x_i + b = -1 \quad (2.30)$$

$$H_2: W^T X + b = \sum_{i=1}^{n} w_i x_i + b = 1 \quad (2.31)$$

Şekil 2.12'de belirtilen optimal ayırma hiper düzleminin altında kalan bölge (2.32) ve üzerinde kalan bölge ise (2.33) eşitsizliklerini sağlamalıdır.

$$W^T X + b < 0 \quad (2.32)$$

$$W^T X + b > 0 \quad (2.33)$$

Destek vektörlerinin H_0 olarak belirtilen ayırma hiper düzlemine uzaklığı denklem (2.34) ve (2.35), sınırlarda yer alan farklı sınıf niteliklerine ait destek vektörleri arasındaki uzaklık (2.36) ile verilmektedir.

$$d=\frac{|W^T X-b|}{\|w\|} \tag{2.34}$$

$$d'=\frac{|W^T X+b|}{\|w\|} \tag{2.35}$$

$$d+d'=\frac{2}{\|w\|} \tag{2.36}$$

İki sınıfın destek vektörleri arasındaki uzaklığın büyüklüğünün maksimum olması amaçlandığından (2.36)'da paydanın minimize edilmesi gerekmektedir. Buradan hareketle denklem (2.37) ve (2.38) belirtilen amaç fonksiyonu ve kısıtları altında çözüm bulunmalıdır.

Amaç fonk: $z_{min}=\frac{1}{2}w^T w$ (2.37)

Kısıtlar: $y_i\,(w^T x_i+b)\ \geq 1$ (2.38)

Belirtilen kısıtlar altında amaç fonksiyonunun çözümü için Lagrange fonksiyonuna başvurulmaktadır. Bu doğrultuda düzenlenen denklem (2.39)'da verilmiştir. Denklemde yer alan $\alpha_i \geq 0$ değerleri pozitif Lagrange çarpanları olarak adlandırılmaktadır.

$$L\,(w,\,b,\,\alpha)=\frac{1}{2}w^T w-\sum_{i=1}^{n}\alpha_i[y_i(w_i x_i+b)-1] \tag{2.39}$$

Denklem (2.39)'un çözülebilmesi için Karush-Kuhn-Tucker (KKT) koşullarından yararlanılmaktadır ().

$$\frac{\partial L(w_1,b,\alpha)}{\partial w}\rightarrow w^T=\sum_{i=1}^{n}y_i\alpha_i x_i \tag{2.40}$$

$$\frac{\partial L(w_1,b,\alpha)}{\partial b}\rightarrow\ \sum_{i=1}^{n}y_i\alpha_i=0 \tag{2.41}$$

(2.40) ve (2.41) bağıntıları kullanılarak denklem (2.39) yeniden düzenlenirse (KKT koşulları (2.40)'ın maksimumu için gerek ve yeter koşuldur);

$$L(w,b,\alpha)= \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \left(x_i{}^T x_j\right) - \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \left(x_i{}^T x_j\right) + \sum_{i=1}^{n} \alpha_i \quad (2.42)$$

$$L(w,b,\alpha)= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \left(x_i{}^T x_j\right) \quad (2.43)$$

elde edilir.

(2.43) denkleminin çözümü $\alpha_i \geq 0$ koşulları altında bir optimizasyon problemi şeklinde çözülmektedir. Optimal hiper düzlemin bulunabilmesi için $L(w,b,\alpha)$ duali maksimize edilmelidir. Çözümde yer alacak $\alpha_i \geq 0$ katsayılı x_i örnekleri destek vektörleri olacaktır. Bu optimizasyon probleminin (w^*,b^*,α^*) optimal noktası için α_i^* çözümleri w^* ve b^* parametrelerini belirlemektedir. N destek vektörlerinin sayısını, x_i destek vektörü kümesinin elemanını belirtmek üzere optimal çözüm parametreleri aşağıdaki şekilde ifade edilir (Özkan, 2008; Mammone, 2009).

$$w^*= \sum \alpha_i{}^* y_i x_i \text{ (i=1,...,n)} \quad (2.44)$$

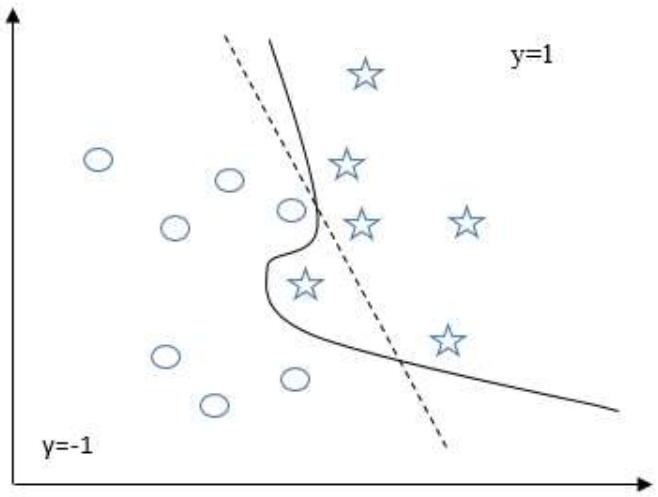
$$b^*= \frac{1}{N} \left(\sum_{s=1}^{N} \left(\frac{1}{y_s} - x_s{}^T w^*\right)\right) \quad (2.45)$$

Buradan hareketle hiper düzleme ait sınıflandırma fonksiyonu (2.46)'da verilmiştir (x_i destek vektörü kümesinin elemanını belirtmek üzere).

$$f(x)= sign \left(\sum \alpha_i{}^* y_i (x_i{}^T x) + b^*\right) \quad (2.46)$$

2.4.4.2. Doğrusal Olmayan Sınıflandırma Problemi

Veri kümesinde yer alan örneklerin sınıflara ayrılması doğrusal bir düzlem ile gerçekleştirilemiyorsa bu şekildeki problemler doğrusal olmayan sınıflandırma problemi olarak adlandırılmaktadır. Bu problemlerin çözümünde negatif olmayan aylak değişkenler ve kernel fonksiyonlarından yararlanılmaktadır. Doğrusal olarak ayrılamayan veri kümesinin örneklerinin iki boyutlu uzaydaki örnek görüntüsü Şekil 2.13'te verilmiştir.



Şekil 2.13: Doğrusal olarak ayrılamayan veri kümesi.

Negatif olmayan aylak değişken ξ_i'nin optimizasyon problemine eklenmesi sonucu denklem (2.38), denklem (2.47)'ye dönüşür. Burada $\xi_i > 1$ değerleri hiper düzlemin doğrusal olarak ayrılmayı önleyen alanda yer alan örneklere ait gözlem değerleridir.

$$y_i\,(w^T x_i + b)\ \geq 1 - \xi_i \quad (i = 1, \ldots, n) ve\ \xi_i \geq 0 \tag{2.47}$$

Lagrange çarpanının alabileceği üst sınır değerini belirten ve ceza parametresi olarak ifade edilen *C* parametresinin dâhil edilmesi ile amaç fonksiyonu denklem (2.48)'deki şeklini alır. Bu *C* parametresi eğitim hatasını minimize etmek ve hiper düzlem ile örnekler arasındaki uzaklığı maksimize etmek arasındaki eşiğin kontrolü için kullanılmakta ve kullanıcı tarafından belirlenmektedir (Çelik, 2017).

$$L(w, \xi) = \frac{1}{2} w^T w + C\left(\sum_{i=1}^{n} \xi_i\right)^k \tag{2.48}$$

Denklem (2.48)'de belirtilen Lagrange fonksiyonu k=1 için (2.49)'a dönüşür (Burges, 1998).

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2} w^T w + C\left(\sum_{i=1}^{n} \xi_i\right) - \sum_{i=1}^{n} \alpha_i \{y_i [w^T x_i + b] - 1 + \xi_i\} - \sum_{i=1}^{n} \beta_i \xi_i \tag{2.49}$$

KKT koşulları sonrası maksimize edilmek istenen dual Lagrange fonksiyonu ve kısıtlar aşağıda verilmiştir.

$$max\ L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j {x_i}^T x_j \quad (2.50)$$

kısıtlar: $\sum_{i=1}^{n} \alpha_i y_i = 0,$

$$C \geq \alpha_i \geq 0$$

Doğrusal olmayan sınıflandırma probleminde kullanılan br diğer yöntem haritalamadır. Daha önce de bahsi geçtiği üzere veri kümesi doğrusal olarak ayrılabileceği daha yüksek boyutlu bir veri uzayına taşınmakta (Cortes ve Vapnik, 1995), bu işlemde çekirdek (kernel) fonksiyonlardan yararlanılmaktadır (Dixon ve Candade, 2008). Dikkat edilmelidir ki sınıflandırma kuralı orijinal veri kümesinin yer aldığı uzayda doğrusal değilken, taşınılan yüksek boyutlu uzayda doğrusaldır (Çelik, 2017). Bir üst boyuta taşıyacak olan haritalama fonksiyonu Φ(x) ile gösterilmektedir (Akpınar, 2014). Kullanılan kernel (çekirdek) fonksiyon haritalama fonksiyonu cinsinden denklem (2.51)'deki şekilde gösterilir.

$$K(x_i,x_j) = \Phi(x_i)^T \Phi(x_j) \quad (2.51)$$

Daha önce bahsi geçen Lagrange ve KKT koşulları da göz önünde bulundurularak elde edilen kernel tabanlı sınıflandırma fonksiyonu şu şekilde gösterilebilir:

$$f(x) = sign\ \left(\sum_{i=1}^{N} y_i \alpha_i K(x, x_i)\right) \quad (2.52)$$

Literatürde kullanılan bazı kernel fonksiyonlarına Tablo 2.2'de yer verilmiştir.

Tablo 2.2: Bazı kernel fonksiyonları.

KERNEL TÜRÜ	**KERNEL FONKSİYONU**	**PARAMETRELER**
Doğrusal	$K(x_i, x_j) = {x_i}^T x_j$	yok
Polinominal	$K(x_i, x_j) = ({x_i}^T x_j + 1)^d$	d
Sigmoid	$K(x_i, x_j) = \tanh(\gamma {x_i}^T x_j + c)$	γ, c

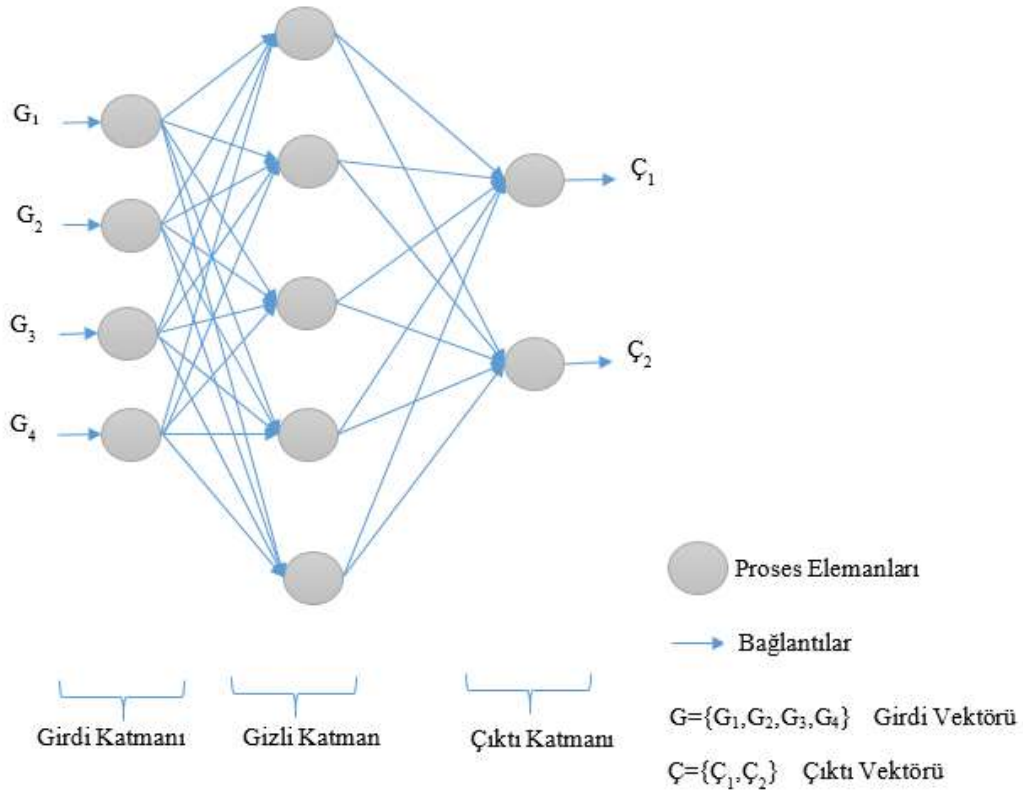
2.4.5. Aşırı Öğrenme Makineleri

Aşırı öğrenme makineleri (Extreme Learning Machine-ELM), ileri beslemeleri yapay sinir ağlarını temel alan, ancak bazı özellikleri ile bu sinir ağlarından ayrılan bir yöntemdir. Dolayısıyla aşırı öğrenme makinelerinin çalışma prensibine geçmeden önce yapay sinir ağları ile ilgili ön bilgiler aşağıda verilmiştir.

2.4.5.1. Yapay Sinir Ağları Temel Kavramlar

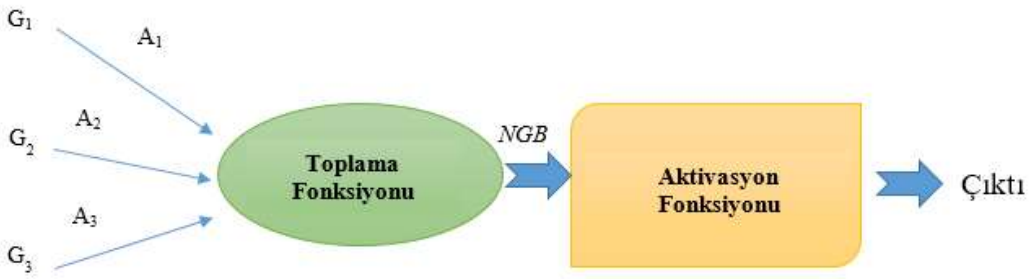
Yapay sinir ağları adından da anlaşılacağı üzere biyolojik sinir ağları temel alınarak modellenmiştir. Yapay sinir ağları insan beyninin fonksiyonlarına benzer şekilde öğrenme, ilişkilendirme, sınıflandırma, genelleme, özellik belirleme, optimizasyon gibi konularda başarı ile uygulanabilen ve öğrenme yolu deneysel bilginin elde edilmesi, depolanması ve kullanılması gibi işlemleri herhangi bir yardım olmaksızın otomatik olarak gerçekleştirmek üzere geliştirilen hücresel bilgisayar sistemleridir (Zurada, 1992; Öztemel, 2012).

Bir yapay sinir ağı girdi katmanı, birden fazla sayıda olabilecek şekilde gizli/ara katmanlar, çıktı katmanı, proses elemanları ve bu elemanlar arası bağlantılardan meydana gelmektedir. Şekil 2.14'te genel bir yapay sinir ağı örneği görülmektedir. Bir yapay sinir ağı göz önüne alındığında girdi setini çıktı setine dönüştüren bir süreç söz konusudur. Bu süreç içerisinde veri setine ait örnekler girdi katmanı ile ağa gönderilmekte, farklı yapay sinir ağı modelleri ile proses elemanları arasında bağlantılar kurulmaktadır. Sonuç olarak elde edilen bağlantılar ile yeni gelen örneklerin hangi sınıflara ait olacağı belirlenmektedir. Girdi katmanında ağa gönderilen girdiler nitelik alanlarına ait değerler, çıktı katmanında üretilen çıktılar ise bu değerlerin sınıf nitelik alanı değerleri olabilir. Girdi katmanında gönderilecek girdiler sayısal değerler olup bir vektör şeklinde olmalıdır.



Şekil 2.14: Yapay sinir ağı örneği.

Bir proses elemanının girdiler, ağırlıklar, toplama fonksiyonu aktivasyon fonksiyonu ve çıktı olmak üzere beş bileşeni mevcuttur. Bu bileşenler Şekil 2.15'te gösterilmiştir.



Şekil 2.15: Proses elemanı bileşenleri.

- *Girdiler*, sinir ağına dışarıdan gönderilen ve veri setindeki örneklere ait değerlerdir. Girdiler aynı zamanda başka proses elemanlarından veya kendisinden de olabilir.
- *Ağırlıklar*, her bir girdi için bu girdinin önemi ve proses elemanı üzerindeki etkisini temsil etmektedir.
- *Toplama fonksiyonu*, proses elemanına gelen net bilginin hesaplanması için kullanılmaktadır. Gönderilen girdiler ve girdiye ait ağırlıklar belirlenen toplama

fonksiyonu vasıtasıyla gerekli işlemlerden geçerek net bilgiye dönüştürülmektedir. En sık kullanılan toplama fonksiyonlarından birisi ağırlıklı toplamdır. Buna göre her bir girdi değeri kendi ağırlığı ile çarpılır ve elde edilen tüm değerler toplanarak net girdi bilgisi (*NGB*) elde edilir. İlgili toplama fonksiyonunun gösterimine denklem (2.53)'te, diğer olası toplama fonksiyonlarına Tablo 2.3'te yer verilmiştir.

Ağırlıklı Toplama Fonksiyonu $=\sum_{i=1}^{n} G_i A_i$ (2.53)

G_i: i. girdi değeri

A_i: i. girdi değerine ait ağırlık değeri

Tablo 2.3: Toplama fonksiyonları.

TOPLAMA FONKSİYONU	FORMÜL
Çarpım	$NGB = \prod_{i=1}^{n} G_i A_i$
Maksimum	$NGB = \text{Max}\{G_i A_i \mid i = 1,2,\dots,n\}$
Minimum	$NGB = \text{Min}\{G_i A_i \mid i = 1,2,\dots,n\}$
Çoğunluk	$NGB = \sum_{i=1}^{n} sgn(G_i A_i)$

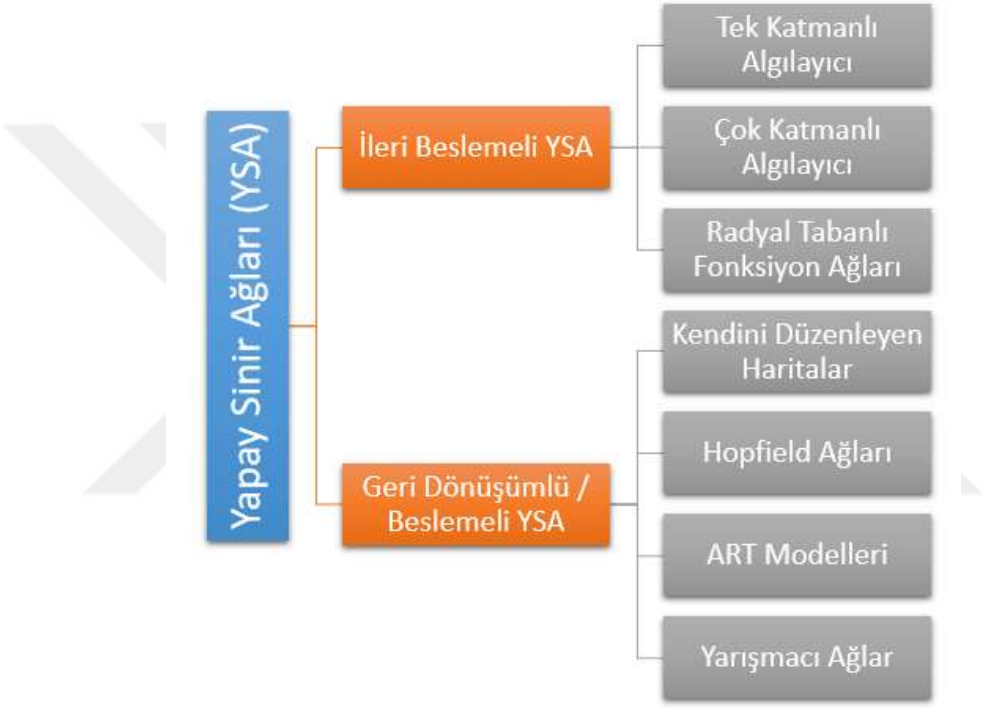
- *Aktivasyon fonksiyonu*, proses elemanına gelen ve toplama fonksiyonu vasıtasıyla hesaplanan net bilginin işlenerek bu net bilgiye karşılık üretilecek çıktı değerinin hesaplanması için kullanılır. Toplama fonksiyonu gibi aktivasyon fonksiyonu için de farklı fonksiyonlar kullanılabilmektedir. Bu fonksiyonlara Tablo 2.4'te yer verilmiştir.

Tablo 2.4: Aktivasyon fonksiyonları.

AKTİVASYON FONKSİYONU	FORMÜL
Lineer	$F(NGB) = NGB$
Sinüs	$F(NGB) = sin(NGB)$
Step	$F(NGB) = \begin{cases} 1, & NGB > eşik\ değer \\ 0, & NGB < eşik\ değer \end{cases}$

- *Çıktı*, aktivasyon fonksiyonu tarafından üretilen değerdir. Çıktı dış dünyaya gönderilebileceği gibi başka bir proses elemanına da gönderilebilir. Her bir proses elemanının tek bir çıktısı bulunmaktadır.

Gardner ve Dorling (1998), yapay sinir ağlarını ileri ve geri beslemeli olmak üzere iki ana kategoriye ayırmış ve bu iki kategoriye göre de diğer sinir ağlarını sınıflandırmıştır. Sınıflandırma Şekil 2.16'da yer almaktadır.



Şekil 2.16: Yapay sinir ağlarının sınıflandırılması.

İleri beslemeli sinir ağlarında girdi tek yönlü olarak girdi, ara ve çıktı katmanlarından geçerek ileri doğru iletilir. Bir katmandan çıkan bilgi diğer katmanda girdiyi oluşturabilir. Geri beslemeli sinir ağlarında ise bir katmandaki çıktı aynı katmandaki başka veya aynı proses elemanının girdisi olabilmektedir.

2.4.5.2. Tek Gizli Katmanlı Aşırı Öğrenme Makineleri

Huang ve diğerleri (2004) tarafından özellikle gradyan tabanlı ileri beslemeli yapay sinir ağı modellerinin fazla sayıda parametre belirleme gerekliliği, yerel minima (hatanın yerel bir noktaya takılması) ve uzun öğrenme süreci gibi problemleri göz önünde bulundurularak geliştirilmiş bir yöntemdir (Cho ve diğ., 2007). Öğrenme sürecinin hızlandırılması amacıyla

girdi ağırlıkları ve eşik değerleri rastgele üretilirken, çıktı ağırlıklarının hesaplanabilmesi için analitik yöntemlerden yararlanılır.

(x_i, y_i), N tane birbirinden farklı ve keyfi seçilmiş örnekler, $x_i = \{x_{i1}, x_{i2},..., x_{iN}\}^T$ girdi ve $y_i=\{y_{i1},y_{i2},...,y_{im}\}^T$ çıktı olmak üzere standart *Tek Gizli Katmanlı İleri Beslemeli Ağlar* (*Single Hidden Layer Feed-Forward Network-SLFN*) için matematiksel model;

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i x_j + b_i) = o_j \qquad (j=1,2,...,N)$$

(2.54)

Ñ: Nöron sayısı, N eğitim veri setindeki örnek sayısı olmak üzere Ñ<N

g(x): Aktivasyon fonksiyonu

$w_i=\{w_{i1},..., w_{in}\}^T$ *i.* girdi ve gizli nöronu ilişkilendiren ağırlık vektörü

$\beta_i=\{\beta_{i1},..., \beta_{in}\}^T$ *i.* gizli nöron ile çıktıyı ilişkilendiren ağırlık vektörü

b_i: *i.* nöron için eşik değer

Normal şartlar altında denklem (2.54) ile belirtilen çıktı sağlanamayabilir. Bu durumda elde edilen çıktı denklem (2.55) ile ifade edilsin.

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i x_j + b_i) = t_j \qquad (j=1,2,...,N) \tag{2.55}$$

Denklem (2.55) genelleştirilir ise denklem (2.56), (2.57) ve (2.58) elde edilir.

$$H\beta = T \tag{2.56}$$

$$H = \begin{bmatrix} g(w_1 x_1 + b_1) & \cdots & g(w_{\tilde{N}} x_N + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1 x_N + b_1) & \cdots & g(w_{\tilde{N}} x_N + b_{\tilde{N}}) \end{bmatrix} \tag{2.57}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix} \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix} \tag{2.58}$$

Aşırı Öğrenme Makineleri ile w_i *i.* girdi ve gizli nöronu ilişkilendiren giriş katmanı ağırlıkları ve b_i gizli katmanda yer alan nöronlar için eşik değerleri rastgele seçilmekte H matrisi ise analitik olarak hesaplanmaktadır (Alçin ve diğ., 2015). Ağ performansını maksimize etmek için hatanın 0 (sıfır) veya minimum olması beklenmektedir (Ertuğrul ve diğ.,2013). Bu durumda;

$$\sum_{j=1}^{N}(o_j - t_j) = 0 \tag{2.59}$$

veya

$$\left\|\sum_{j=1}^{N}(o_j - t_j)^2\right\| \tag{2.60}$$

minimum olmalıdır.

Çıktı katmanındaki *hatayı minimize eden çıktı katmanı ağırlıklarını* $\bar{\beta}$ ile gösterirsek

$$\left\|H\bar{\beta} - T\right\| = min\|H\beta - T\| \tag{2.61}$$

olmalıdır.

Hatayı minimize eden çıktı katmanı ağırlık vektörü:

$$\bar{\beta}=H^{\dagger}T \tag{2.62}$$

Aşırı Öğrenme Makineleri çıktı üretmek üzere öğrenme aşamasında iteratif işlemlerden oluşan bir süreç yerine denklem (2.56) ile verilen doğrusal denklemin çözümünü bulmaya çalışmaktadır (Uçar ve diğ., 2015). Denklemin çözümü için de denklem (2.62) kullanılır. Burada $H^{\dagger}$, H'nin genelleştirilmiş tersi olup Moore–Penrose Sözde Ters yöntemi kullanılmaktadır (Penrose, 1955).

Yukarıdaki bilgiler ışığında *Tek Gizli Katmanlı Yapay Sinir Ağları* için *Aşırı Öğrenme Makinesi Algoritması* adımları aşağıdaki gibi özetlenebilir:

Adım1: Rastgele olacak şekilde w_i ağırlıkları ve b_i eşik değerleri belirlenir.

Adım 2: H gizli katman çıktı matrisi üretilir.

Adım 3: Hatayı minimize eden çıktı katmanı ağırlık vektörü belirtilen esaslara göre hesaplanır.

2.5. MODEL GEÇERLEME YÖNTEMLERİ

Deneysel veri setlerinin bilinmeyen bir olasılıkla dağıldığı varsayıldığında; herhangi bir modelin başarısı veri seti içerisinde yer alan gizli düzenin öğrenilerek, bu düzene uygun bir

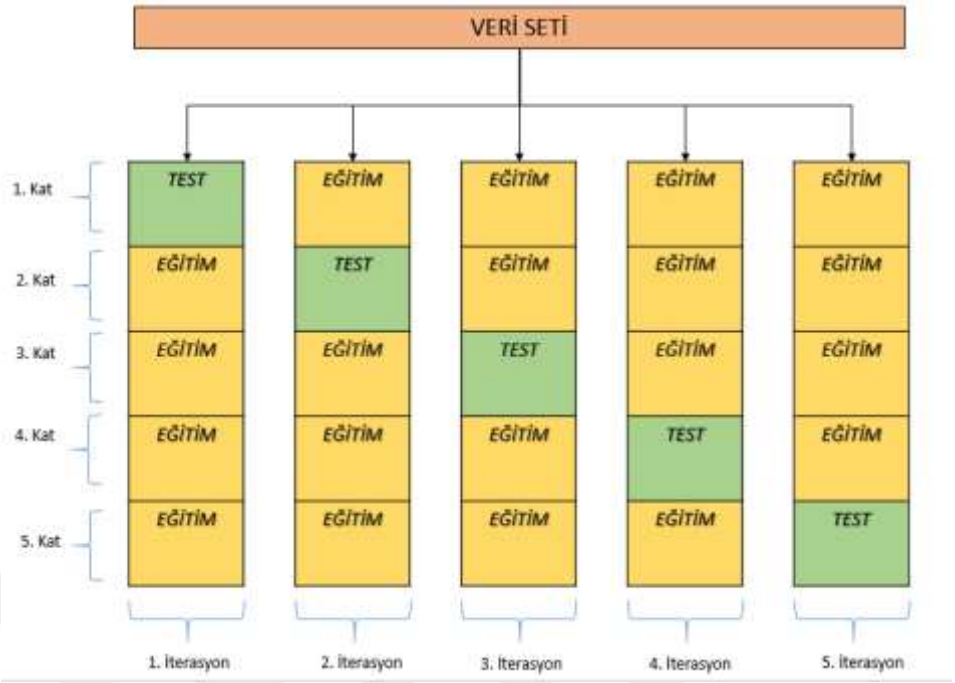
model geliştirilmesi için veri setinde yeterli bilginin olup olmaması ve modelin, sınıfı bilinmeyen yeni gelen örnekler için yüksek doğrulukta sınıf tahminini yapabilmesine bağlıdır. Daha iyi bir modelin oluşturulması da ancak farklı veri grubu örnekleri ile yinelenen bir sürecin geri bildirimlerine göre oluşturulabilmektedir (Kantardzic, 2011).

Daha önce, veri madenciliği süreci modelleme aşamasında eğitim ve test veri setlerinin oluşturulduğu, eğitim veri setinin kullanılarak çeşitli modeller elde edildiği ifade edilmişti. Yukarıdaki açıklama da göz önüne alınırsa eğitim ve test veri setlerinin belirlenmesi özel yöntemler ile mümkündür. Bu amaç doğrultusunda çeşitli yöntemler geliştirilmiştir.

Belirli Oranlarda Bölme (Hold-Out): En bilinen ve kullanılan model performans değerlendirme yöntemleri arasında yer alan belirli oranlarda bölme yöntemi veri setinin belirli oranlar dâhilinde, rastgele olacak şekilde eğitim ve test veri seti olmak üzere iki bağımsız veri kümesine ayrılmasını sağlamaktadır. Bu yöntemde genel olarak veri setinin 2/3 oranı eğitim 1/3 oranı test veri seti olarak ayrılırken eğitim veri seti modelin oluşturulması, test veri seti ise modelin performansının ölçülmesi için kullanılmaktadır (Han ve diğ., 2012). Ayırma işlemi %50-50, %30-70, %20-80 şeklinde de yapılabilmektedir.

Belirli oranlarda bölme yönteminde olduğu gibi eğitim ve test veri setleri ayrılmakta ancak bu ayırma işlemi bir defa değil defalarca tekrarlanarak modelin oluşturulması sağlanmakta ise bu yöntem tekrarlı belirli oranlarda bölme yöntemi olarak adlandırılmaktadır. Burada modelin hatası her ayırma işlemi ile belirlenmekte sonuçta tüm hataların ortalaması alınmaktadır.

Çapraz Geçerleme (Cross Validation): Çapraz geçerleme (doğrulama), veri setini modelin oluşturulmasını sağlayan ve modelin performansı test eden olmak üzere bölümlere ayırmayı sağlayan bir yöntemdir. Çapraz olarak ifade edilmesinin sebebi belirli sayıda tekrarlanan bu geçerleme yönteminde bir tekrarda eğitim veri setinde yer alan örneğin bir sonraki tekrarda test veri setinde yer almasının, keza bir tekrarda test veri setinde yer alan örneğin bir sonraki tekrarda eğitim veri setinde almasının mümkün olmasından kaynaklanmaktadır (Refaeilzadeh ve diğ., 2009). Şekil 2.17’de 5-kat çapraz geçerleme görsel olarak ifade edilmiştir.



Şekil 2.17: 5-kat çapraz geçerleme.

Veri seti mümkün olduğunca eşit sayıda örnek içeren belirli sayıda parçaya bölünmektedir. Veri seti kaç parçaya bölünecek ise bu sayı k olarak ifade edilmekte, yöntem ise k-kat (k-fold) çapraz geçerleme olarak anılmaktadır. Örneğin; 5-kat çapraz geçerleme yönteminde veri seti toplam beş parçaya ayrılacaktır. Buna göre *D* bir veri seti olmak üzere k-kat çapraz geçerlemede D_1, D_2, D_3,..., D_k şeklinde k tane veri kümesi elde edilmektedir. t ϵ {1, 2,..., k} olmak üzere *Dt* kümesi test veri seti olarak ayrılırken geriye kalan (k-1) tane küme eğitim veri seti olarak kullanılmaktadır (Kohavi, 1995). Süreç her bir eğitim kümesi test veri seti olarak kullanılana kadar devam etmektedir (Yadav ve Shukla, 2016). Literatürde en çok 5 ve 10 kat çapraz geçerleme kullanılmaktadır (Kohavi, 1995; Lee ve Lee, 2003; Lin ve diğ., 2011).

Çapraz geçerlemenin özel bir versiyonu birini dışarıda bırak (leave-one-out) çapraz geçerlemedir. Bu geçerleme yönteminde veri setinde yer alan örnek sayısı n olmak üzere, k=n olarak alınmaktadır (Kim, 2009). Ancak çok fazla sayıda örnek içeren bir veri seti düşünüldüğünde her bir örneğin dışarıda bırakılarak tekrar edilmesi işlem süresini bir hayli uzatacaktır.

Bootstrap: Bu yöntem genellikle veri setinde yer alan örnek sayısının az olması durumunda tercih edilmektedir. Efron ve Tibshirani (1994) tarafından öne sürülen yöntemde, n örnekli bir veri setinden iadeli ve rastgele olarak n kez örnek çekme ile eğitim veri seti oluşturulmaktadır.

İadeli çekim söz konusu olduğundan eğitim veri setinde aynı örnek birden fazla kere tekrar edebilmektedir.

2.6. MODEL PERFORMANS DEĞERLENDİRME ÖLÇÜLERİ

Deneysel analizde uygulanan bir modelin performansının ölçülmesi analiz sonuçlarının değerlendirilebilmesi ve daha önce kullanılan model ve yaklaşımlar ile karşılaştırılabilir olması için önem arz etmektedir. Çünkü deneyde kullanılan örneklemin seçimi, deneyin tasarımı araştırmacının seçimine bağlı olup sonuçlar bu seçim ve tasarımlara bağlı olarak değişecektir (Garcia ve diğ., 2009). Veri madenciliği sürecinde de oluşturulan modellerin sonuçlarının performansını değerlendirmek üzere çeşitli performans değerlendirme ölçüleri kullanılmaktadır. En çok kullanılan performans değerlendirme ölçüleri arasında doğruluk (accuracy), duyarlılık (sensitivity/recall), seçicilik/belirleyicilik/özgünlük (specificity), kesinlik (precision), F-ölçüsü (F-score) sayılabilir (Kantardzic, 2011).

Performans değerlendirme ölçüleri modelin uygulanması sonucu her bir sınıf için doğru ve yanlış sınıf tahmin sayılarına göre oluşturulan karışıklık matrisi vasıtası ile hesaplanmaktadır (Sokolova ve diğ., 2006; Subashini ve diğ., 2009). Tablo 2.5'te ikili sınıf niteliğine göre karışıklık matrisi verilmiştir (Akpınar, 2014).

Tablo 2.5: Karışıklık matrisi.

Gerçek Sınıflar / Tahmini Sınıflar	Pozitif	Negatif	Toplam	Değerlendirme Ölçüleri
Pozitif	Doğru Pozitif *(DP)*	Yanlış Pozitif *(YP)*	P^t	Kesinlik (Precisson) $\frac{DP}{DP+YP}$
Negatif	Yanlış Negatif *(YN)*	Doğru Negatif *(DN)*	N^t	Negatif Tahmin Değeri $\frac{DN}{YN+DN}$
Toplam	P	N	$P+N=P^t+N^t=G$	
Değerlendirme Ölçüleri	Hassasiyet (Sensitivity) $\frac{DP}{DP+YN}$	Özgünlük (Specificity) $\frac{DN}{DN+YP}$		Doğruluk (Accuracy) $\frac{DP+DN}{DP+YP+YN+DN}$

P= Pozitif sınıfta yer alan örnek sayısı
N= Negatif sınıfta yer alan örnek sayısı
P^t =Pozitif sınıfta yer aldığı tahmin edilen örnek sayısı
N^t=Negatif sınıfta yer aldığı tahmin edilen örnek sayısı

Veri setinde yer alan her bir örnek için sınıf niteliğinin var olduğu daha önce ifade edilmişti. Eğer bu sınıf niteliği her bir örnek için iki bağımsız değerden birini alıyorsa ikili, ikiden fazla

ve birbirinden bağımsız değerden birini alıyorsa çoklu sınıf niteliği adını almaktadır. Örneğin; hastanın durumu sınıf niteliği "kanser" ve "kanser değil" şeklinde birbirinden bağımsız iki değer içerdiğinden ikili, "Akciğer kanseri", "Lenf kanseri" ve "Mide kanseri" şeklinde ikiden fazla ve bağımsız değer içerdiğinden çoklu sınıf niteliği olarak değerlendirilmektedir. Tez kapsamında ikili sınıf niteliği çalışıldığından buradan sonra yer alan tüm ölçülerin hesaplaması ikili sınıf niteliğine göre anlatılacaktır.

İkili sınıf niteliğine sahip bir veri setinde yer alan her örnek, pozitif ve negatif olacak şekilde iki sınıf etiketi değerinden birine sahiptir. Yani her bir örnek ya pozitif sınıfta ya da negatif sınıfta yer almaktadır. Burada amaca göre bir sınıf niteliği değeri pozitif, diğer sınıf niteliği değeri ise negatif olarak kabul edilmektedir. Pozitif ve negatif sınıf olma durumu, *k* örnekli veri setinde sınıf niteliği *l*, pozitif sınıf niteliği değeri *p* ve negatif sınıf niteliği değeri *n* olmak üzere $l_i \in \{p, n\}$ (i=1, 2, ..., k) şeklinde ifade edilmektedir (Fawcett, 2006).

Eğitim veri setine göre oluşturulan veri madenciliği modeli, uygulama sonrası test veri setinde yer alan her bir örneğin hangi sınıfta yer alması gerektiğine dair sınıf niteliği tahmin değerleri üretecektir. Buna göre, test veri setinde yer alan her bir örnek yine ya pozitif ya da negatif sınıfta yer alacaktır. Örneklerin gerçek sınıf değerleri ve modelin tahmin ettiği sınıf değerleri karşılaştırılarak karışıklık matrisi oluşturulmaktadır. Gerçekte pozitif sınıfa ait olan değer pozitif sınıfa ait olarak tahmin edilmişse "*Doğru Pozitif (DP)*", gerçekte pozitif sınıfa ait olan değer negatif sınıfa ait olarak tahmin edilmişse "*Yanlış Negatif (YN)*", gerçekte negatif sınıfa ait olan değer pozitif sınıfa ait olarak tahmin edilmişse "*Yanlış Pozitif (YP)*" ve son olarak gerçekte negatif sınıfa ait olan değer negatif sınıfa ait olarak tahmin edilmişse "*Doğru Negatif (DN)*" şeklinde ifade edilmektedir.

Doğruluk (Accuracy) (Rand İndeksi): Pozitif ve negatif sınıf fark etmeksizin sınıf nitelik değeri doğru tahmin edilen örnek sayısının veri setinde yer alan tüm örnek sayısına oranı şeklinde ifade edilmektedir.

$$Doğruluk = D = \frac{DP+DN}{G} \tag{2.55}$$

Hata Oranı (Error Rate): Pozitif ve negatif sınıf fark etmeksizin sınıf nitelik değeri yanlış tahmin edilen örnek sayısının veri setinde yer alan tüm örnek sayısına oranı şeklinde ifade edilmektedir. Doğruluk değerinin 1'den çıkarılması ile de elde edilir.

$$\textit{Hata Oranı} = H = \frac{YN+YP}{G} = 1\text{- }\textit{Doğruluk} \tag{2.56}$$

Hassasiyet (Sensitivity/Recall): Modelin gerçekte pozitif sınıf nitelik değerine sahip sınıf içerisinde doğru tahmin oranını veren ölçüdür.

$$\textit{Hassasiyet} = Ht = \frac{DP}{DP+YN} \tag{2.57}$$

Özgünlük (Specificity): Modelin gerçekte negatif sınıf nitelik değerine sahip sınıf içerisinde doğru tahmin oranını veren ölçüdür.

$$\textit{Özgünlük} = Ozg = \frac{DN}{YP+DN} \tag{2.58}$$

Kesinlik (Precision): Modelin pozitif sınıf nitelik değerine sahip olarak tahmin ettiği örnekler içerisinde gerçekte de pozitif sınıf nitelik değerine sahip örneklerin oranıdır.

$$\text{Kesinlik} = K = \frac{DP}{DP+YP} \tag{2.59}$$

Modelin negatif sınıf nitelik değerine sahip olarak tahmin ettiği örnekler içerisinde gerçekte de negatif sınıf nitelik değerine sahip örneklerin oranı ise *Negatif Tahmin Oranı* olarak adlandırılır.

$$\text{Negatif Tahmin Oranı} = NTO = \frac{DN}{YN+DN} \tag{2.60}$$

F-Ölçüsü (F-Score): Rijsbergen (1979) tarafından geliştirilmiş olup kesinlik ve hassasiyet ölçülerine dayalı bir ölçüdür. Kesinlik ve hassasiyet ölçülerinde yanlış negatif sınıf değerlerinin katılımının dengelenmesi sağlaması (Akpınar, 2014) bakımından tek başına kesinlik veya hassasiyet ölçüsüne göre efektif bir yorumlamaya yardımcı olabilir. Geleneksel F-ölçüsü olarak literatürde yer alan ölçüde harmonik ortalama kullanılır.

$$\text{F-ölçüsü: } F = \frac{2*K*Ht}{K+Ht} \tag{2.61}$$

2.7. MÜŞTERİ KAYIP ANALİZİ

Günümüzde yeni pazarların oluşması, işletmeler arası rekabetin gittikçe keskinleşmesi ve zorlaşmasına neden olmuştur. Bu zor koşullar altında işletmelerin odaklandığı konulardan biri de müşteri davranışlarıdır (Kisioglu ve Topcu, 2011). Küreselleşme ile beraber müşteri davranışları da değişim gösterdiğinden böyle bir rekabet ortamında ayakta kalabilmek için işletmeler yeniden yapılanmaya, müşterilerine yönelik stratejilerini gözden geçirerek yeni stratejiler geliştirmeye çalışmıştır (Demir ve Kırdar, 2009). Bu stratejiler ile beraber işletmeler artık müşteri sayısını arttırabilmenin yanı sıra var olan müşterilerini de kaybetmemeyi amaçlamaktadır.

Bu amaç doğrultusunda veri madenciliğinin uygulama alanlarından biri olan Müşteri Kayıp Analizi/Tahmini çalışmaları karşımıza çıkmaktadır. Müşteri kayıp analizi Müşteri İlişkileri Yönetimi altında bir yönetim bilimi problemi olmak ile beraber çözüm için veri madenciliği yaklaşımları benimsenmiştir (Verbeke ve diğ., 2012). Birçok farklı sektörde işletmelere ait veri ambarlarında tutulan müşteriler ile ilgili demografik, fatura, kullanılan servisler vb. alanlardaki veriler üzerinde farklı veri madenciliği yöntemleri uygulanarak müşteri davranışları ile ilgili modeller kurulmaya ve çıkarımlar yapılmaya çalışılmaktadır. Bu veriler içerisinden müşteri memnuniyeti, müşteri değeri ya da hangi müşterinin gelecekte kaybedilme durumunun oluşabileceğinin tahmin edilmesi gibi farklı bilgiler de elde edilebilmektedir (Hadden ve diğ., 2007). Tüm bu bilgiler müşteri yönetiminin daha iyi işleyebilmesi ve performansın arttırılabilmesi dolayısıyla rekabet ortamındaki çetin mücadelede bir adım önde olabilmek için önem arz etmektedir.

2.7.1. Müşteri Kaybı, Sadakati ve Değeri

Müşteriler işletmelerin sunduğu ürün ve hizmetler karşılığında bir bedel ödeyerek satın alan kişilerdir (TDK, 2017). Müşteriler bu ürün veya hizmetleri alma durumlarına göre; mevcut, muhtemel, eski, yeni, hedef müşteri olmak üzere beş grup altında toplanmıştır. Buna göre;

- İşletmenin sunduğu hizmet veya malı sürekli satın alan mevcut,
- İşletmenin satış için görüştüğü ancak henüz müşteri olmamış müşteri muhtemel,
- İşletmenin daha önce müşterisi olan ancak farklı durumlar nedeni ile işletmenin müşterisi olmaktan çıkan müşteri eski,
- İşletmenin ürün veya hizmetini ilk kez alan müşteri yeni,

- İşletmenin belirli ürün veya hizmetlerini satın alması olası müşteriler de hedef müşteri

olarak tanımlanmaktadır (Yıldızel, 2002; Akt: Demir ve Kırdar, 2009).

Birçok müşteri farklı sebeplerle ürün veya hizmet aldığı işletmeden vazgeçebilmekte farklı işletmelere kayabilmektedir. Müşteri kaybı, en genel haliyle mevcut müşterinin ürün veya hizmet almayı kesme durumu olarak ifade edilebileceği gibi, literatürde çeşitli yorumlamalar ile oluşturulan farklı tanımlamalara yer verilmektedir. Kaur ve arkadaşları (2013), belirli bir periyod içerisinde müşterinin işletme ile ilgili ürün veya hizmet alımı işini durdurmaya yönelik eğilimi; Chandar ve arkadaşları (2006), belirli bir zaman dilimi içerisinde müşterinin bağlı olduğu bir şirket ile yaptığı satın alma, kredi kartı kullanımı vb. şekildeki işlerinde duraksama/durma gerçekleşmesi ve durumdaki müşterinin eğilimi; Hung ve arkadaşları (2006) ile Chu ve arkadaşları (2007), müşterinin hizmet veya ürün aldığı bir işletmeden diğerine geçişi; Chen ve arkadaşları (2012) ise, müşterinin bir işletmeden ürün veya hizmet alımını durdurması veya azaltması şeklinde müşteri kaybını tanımlamışlardır.

Lazarov ve Capota (2007), kayıp müşteriyi aktif, dönüşlü ve pasif; müşteri kaybı durumunu ise toplam, gizli ve parçalı olmak üzere üçer alt gruba ayırmıştır. Buna göre;

Aktif, tamamen ürün veya hizmet alımını durduran varsa sözleşmesini feshedip başka bir işletme müşterisi olmaya karar veren; dönüşümlü, ürün veya hizmet alımını durduran varsa sözleşmesini fesheden ancak başka bir işletmeye geçmeyen; pasif ise sözleşme gerektiren hizmet veya ürün alımlarında sözleşmenin sona erdikten sonra devam ettirilmemesi durumundaki kayıp müşterileri temsil etmektedir.

Toplam sözleşmelerin resmen iptal edildiği; gizli, sözleşmenin iptal edilmediği ancak müşterinin hizmet veya ürünü kullanma ya da satın almaya devam etmediği; parçalı ise sözleşmelerin resmen feshedilmediği müşterinin ürün ya da hizmetin bir kısmını kullanması ve başka işletmelerden de aynı anda hizmet ve ürün satın alıyor olma halindeki müşteri kaybı durumlarını ifade etmektedir.

İşletmelerin öncelikli amacı ürün veya hizmetlerini satıp gelir elde ettiği mevcut müşterilerin sayısını arttırmaktır. Müşteri sayısını arttırmak ise pazardaki hedef müşterileri mevcut müşteri haline çevirmek ve bunun yanı sıra mevcuttaki müşterileri de kaybetmemek ile gerçekleştirilebilir. Çünkü kaybedilen her müşteri işletme için yeni kazanılması gereken

fazladan bir müşteri ve dolayısıyla fazladan maliyet anlamına gelmektedir. Bunun yanı sıra Kim ve çalışma arkadaşları (2004) çalışmalarında bir işletmenin müşteri sayısı en üst düzeye ulaştığında yeni bir müşteriye sahip olmanın pazarlama açısından oldukça zor ve maliyetli olabileceğini, dolayısıyla bu noktada müşteri sadakatini arttırarak mevcut müşterileri koruma stratejisinin benimsenmesi gerektiğini belirtmektedirler.

İşletmelerin mevcut müşterileri üzerinde geliştirmeye çalıştıkları stratejiler için önem arz eden ve son dönemde üzerinde durulan kavramlardan birisi de müşteri sadakatidir. Öyle ki, satışları arttırmak ve istikrarlı bir pazar payına sahip olabilmek adına işletmeler müşterilerin sadakatlerini arttırmaya yönelik olarak ödül ve indirimler sundukları sadakat programları düzenlemektedirler (Sharp ve Sharp, 1997; Ou ve diğ., 2011). Müşteri sadakati, pazarda bulunan ve benzer ürün veya hizmetleri sunan rakip işletmeler olmasına rağmen müşterinin mevcut müşterisi bulunduğu işletmeden ürün veya hizmet satın alması, almaya devam etmesi, gelecekte de aynı işletmeyi tercih etme potansiyeli olması ve işletmeye bağlılık hissi duyması gibi tutum ve davranışlar bütünü olarak tanımlanmaktadır (Jones ve Sasser, 1995; Oliver, 1997; Neal, 1999; Chen ve diğ., 2015).

Müşterinin bir işletmeyi tercih etmesi ve işletmeye karşı bağlılığının devamı işletmenin müşteriye sunacağı değer ile sağlanabilmektedir. Dolayısıyla işletmeyi, müşteri kaybının engellenmesi ve sadakatinin sağlanması noktasında başarıya ulaştıracak anahtarlardan biri de müşteri değeridir (Onaran ve diğ., 2013). Müşteri değeri, işletmenin sunmuş olduğu ürün veya hizmeti alan müşteri tarafından, bu ürün veya hizmet hakkında beklenen – alınan arasında beklentinin karşılanıp karşılanmadığının değerlendirilmesidir. (Zeithaml, 1988).

Bir müşteriyi kazanmak, müşteri değeri sağlayarak sadakati arttırmak ve müşteri kaybı müşterinin bir işletme ile olan ilişkileri göz önüne alındığında sistematik bir süreci meydan getirmektedir. Bu süreç Müşteri yaşam döngüsü olarak adlandırılmaktadır.

2.7.2. Müşteri Yaşam Döngüsü

Her işletmenin müşterinin değeri olduğu gibi müşterinin bir de yaşam döngüsü söz konusudur.

Şekil 2.18: Müşteri yaşam döngüsü.

Şekil 2.18'de müşteri yaşam döngüsü müşteri değeri ve sadakati oluşturma geliştirme aşaması şeklinde birleştirilerek şematize edilmiştir.

Müşteri yaşam döngüsü işletmenin müşteri ile iş ilişkilerinin kurulması ile başlayan ve bu ilişkilerin bitmesi ile son bulan farklı aşamalardan meydana gelen bir süreç olup aşağıdaki adımlardan meydana gelmektedir (Wang ve diğ., 2014):

- Potansiyel müşterileri keşfetme ve onlara erişim,
- Hedef müşteriyi mevcut müşteriye çevirmek, müşteriyi elde etmek,
- Müşteri hizmet veya ürün taleplerini sağlayarak müşteri değeri yaratma, ilişkileri geliştirmek,
- Müşterilerin yeni ürün veya hizmetleri kullanmalarını sağlayarak müşteri sadakatini oluşturmak, ilişkilerin sürdürülebilirliğini sağlamak,
- Ayrılması muhtemel müşteriler için uyarı sistemi yaratarak yaşam döngüsü süresini uzatmak,
- Müşteriyi kaybetmek,
- Kaybedilen müşteriyi kazanmak için yöntemler geliştirmek.

2.7.3. Müşteri Kayıp Analizi ve Önemi

Daha önce de belirtildiği gibi yeni müşterilerin kazanımı kadar, belki de çok daha önemlisi var olan müşterilerin kaybedilmemesidir. Kayıp müşteriler maliyetlerin artması ve kar oranının azalması bakımından işletmeler için her zaman sıkıntı doğurmuştur (Forhad ve diğ., 2014).

Müşteri kaybı yaşandıktan sonra standart teknikler ile müşteriyi geri döndürmek oldukça zor olduğundan müşteriyi kaybetmeden önce müşterinin kaybedilip kaybedilmeyeceğine dair bilgileri edinmek için veri madenciliğinin sınıflandırma teknikleri kullanılmaktadır. Bu teknikler tahmin etme ve anlama olmak üzere iki farklı durum için kullanılmakta, buna bağlı olarak müşterinin geleceğe yönelik kaybedilme durumu tahmin edilmekte ve neden kaybedildiği ile ilgili bilgi çıkarılabilmektedir (Richeldi ve Perrucci, 2002).

Olası kayıp müşteriler için gerçekleştirilen bu tahmin etme ve anlama işlemlerinin tümüne "Müşteri Kayıp Analizi" adı verilmektedir. Müşteri kayıp analizinin temel amacı; ayrılma riski taşıyan müşterilerin tanımlanması ve mümkün ise de bu müşterileri elde tutabilme durumlarının ortaya konulmasıdır (Mutanen, 2006).

Müşteri kayıp analizinin önemini ise işletmelerin maliyetler konusundaki durumları arttırmaktadır. Müşteri kazanma maliyetleri bir müşteriyi elde tutma maliyetlerinden çok daha fazla olduğundan müşteri kayıp analizi önemli bir uygulaması haline dönüşmüş, bazı işletmeler müşteri ilişkileri yöntemi altında özel olarak bu analizi gerçekleştiren departmanlar oluşturmuşlardır (Richter ve diğ., 2010).

2.8. LİTERATÜR İNCELEMESİ

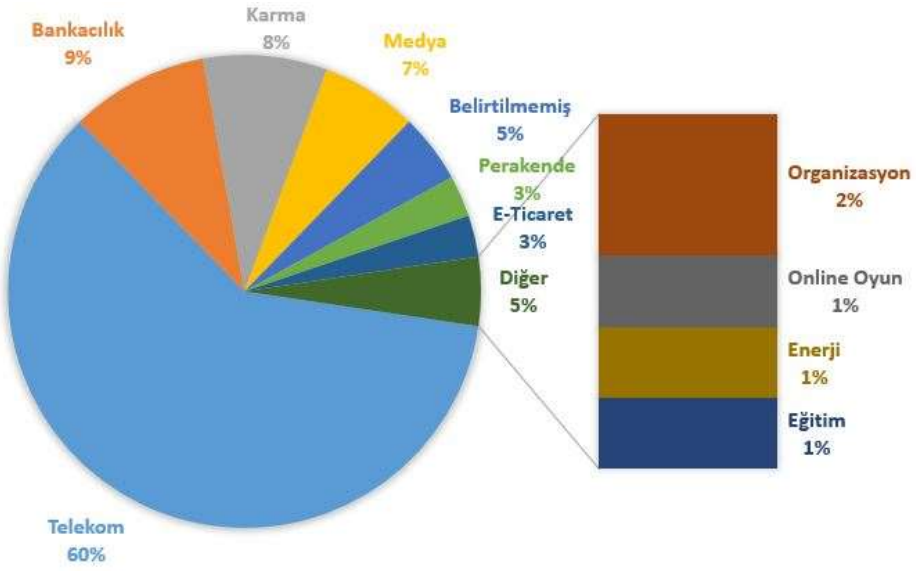
Bu doktora tez çalışması kapsamında müşteri kayıp analizi problemi ile ilgili olarak aşağıdaki sorulara da yanıtlar aranmaya çalışılmıştır:

- Problem hangi sektörde daha fazla gözlemlenmekte ve çözüm yöntemleri daha yoğun olarak uygulanmaktadır?
- Problemin çözümü için kullanılan hangi veri madenciliği algoritma/yöntemleri ön plana çıkmaktadır?
- Konu hakkındaki bilimsel çalışmalar zaman içerisinde nasıl bir eğilim göstermiştir?
- Alanda yapılan çalışmalarda elde edilen sonuçlar ve gelişmeler nelerdir?

- Alanda yapılacak çalışmalarda hangi durumlar için çözümler beklenmektedir?
- Yenilik oluşturabilecek durumlar nelerdir?

Tüm bu soruların yanıtlarının aranması ve tez çalışmasına yön verebilmesi için kapsamlı bir literatür araştırması gerçekleştirilmiştir. Buna göre, 1999-2017 yılları arasında yayınlanmış ve Science Citation Index (SCI), Social Sciences Citation Index (SSCI), Sciences Citation Index Expanded (SCI-Expanded), The Institute of Electrical and Electronics Engineers (IEEE) ve diğer alan indekslerine girmiş uluslararası 106 makale incelenmiştir. Bu incelemede uygulama yapılan sektör, uygulanan yöntemler, performans ölçüm kriterleri, kullanılan veri seti ve araştırmanın genel konusu ile elde edilen genel sonuçlar üzerine odaklanılmıştır. Makalelerin tarandıkları indekslere göre dağılımı 14 SCI, 3 SSCI, 58 SCI-Exp ve 31 Diğer indeksler (IEEE, Alan İndeksi, vd.) şeklindedir. Aşağıda yapılan literatür incelemesine ait sonuçlar paylaşılmıştır (Koçoğlu ve diğ., 2016).

Uygulamaların sektörlere göre dağılımı incelendiğinde en çok telekomünikasyon sektörüne ait veri setlerinin kullanılarak çalışmaların gerçekleştirildiği görüşmüştür. Telekomünikasyon sektörüne ait oran %60 iken diğer sektörlere göre yüzdesel oranların dağılımı Şekil 2.19'da verilmiştir.



Şekil 2.19: Uygulamaların sektörlere göre dağılımı.

Kullanılan veri madenciliği yöntemleri incelendiğinde ise en sık kullanılan algoritmaların Karar Ağacı algoritmaları olduğu görülmektedir. Yüz çalışmanın 59'unda Karar Ağacı algoritmaları kullanılırken sırasıyla 50 çalışmada Lojistik Regresyon, 49 çalışmada Yapay Sinir Ağları ve 29

çalışmada Destek Vektör Makineleri kullanılmıştır. Sade Bayes, Genetik Algoritmalar, K-En Yakın komşu gibi algoritmalar da kullanılan diğer algoritmalar arasındadır. Kullanılan performans değerlendirme ölçüleri arasında ise doğruluk (accuracy) değeri 100 çalışmanın 38'inde öncelikli karşılaştırma ölçüsü olarak kullanılmıştır. Kullanılan diğer ölçüler Kaldıraç (Lift) (35), Alıcı İşletim Karakteristiği (ROC) Eğrisi Altındaki Alan (AUC) (32), Belirleyicilik (Specificity) (16), Duyarlılık (Sensitivity) (15) ve Alıcı İşletim Karakteristiği (ROC) Eğrisi (12)'dir. Veri setleri açısından çalışmalar incelendiğinde ise kullanılan veri setlerinin türü aşağıda listelenmiştir:

- Tek/farklı sektörden veri setlerinin kullanıldığı çalışmalar.
- Tek/birden fazla veri seti kullanılan çalışmalar.
- Daha önce kullanılmamış (orijinal)/herkese açık veri kaynaklarından elde edilen veri setlerinin kullanıldığı çalışmalar.
- Dengeli /dengesiz dağılım gösteren sınıf niteliğine sahip veri setlerinin kullanıldığı çalışmalar.
- Az/çok boyutlu veri setlerinin kullanıldığı çalışmalar.

İncelenen makaleler, sadece ayrılma eğilimli müşterileri tahmin etme amacı dışında diğer çalışma amaçlarına göre altı grupta toplanmıştır. Çalışmalar birden fazla konuyu içerebileceği gibi sadece tek bir amaç doğrultusunda da gerçekleştirilen çalışmalar mevcuttur.

***Hibrit modellerin geliştirilmesi*:** Son dönemde yeni algoritma geliştirmek yerine var olan yöntemlerin performansını arttırmaya yönelik yapılan çalışmaların sayıca fazlalaştığını söylemek mümkündür. Buna göre; çeşitli amaçlar doğrultusunda birden fazla algoritmanın bir arada kullanılması ve bu algoritmaların özellikle avantajlı yönlerinin modele dahil edilmesi ile ortaya çıkan modeller hibrit modeller olarak tanımlanmaktadır. Örneğin; bir algoritma nitelik seçimi, farklı bir algoritma tahmin etme ve başka bir algoritma ise parametre optimizasyonu yapmak üzere bir araya getirilebilmektedir.

***Müşterinin ayrılma durumunu etkileyen faktörlerin belirlenmesi*:** İşletmeler veri tabanlarında müşterileri ile ilgili elde edebildikleri her türlü veriyi saklamaya çalışmaktadırlar. Bu verinin içeriğinde müşterinin adresinden, doğum tarihine, alışveriş sıklığından, alışveriş tutarına, ödeme yönteminden alışveriş tercihlerine kadar kadar farklılık gösterebilmektedir. Ayrılma eğilimi gösteren müşterilerin belirlenmesi çalışmalarında her nitelik alanının eşit oranda etkili

olduğunu söylemek mümkün olmayabilir. Örneğin; doğum tarihi yerine müşterinin ödeme yöntemi daha etkili ve mutlaka göz önüne alınması gereken bir nitelik alanı olabilir. Bu şekilde hem önemli olan nitelik alanlarının belirlenerek müşteri kayıp analizinin bir parçası olan müşterinin ayrılma nedenlerinin yorumlanabilmesi, hem de gereksiz nitelik alanlarının atlarak çalışmaların daha etkin bir şekilde gerçekleştirilerek performansın arttırılması sağlanabilmektedir.

***Dengesiz sınıf dağılımı problemine yönelik çözüm geliştirilmesi*:** Veri madenciliğinin sınıflandırma işlevi için kullanılacak veri setinde sınıf etiketlerini içeren bir nitelik alanının var olması gerekmektedir. Yani veri setinde yer alan her bir örneğin hangi sınıfa ait olduğu bilinmelidir. Sınıf nitelik alanı "var"-"yok", "hasta"-"sağlıklı" veya "iyi"-"orta"-"kötü" gibi ikili veya daha fazla sayıda sınıf içerebilir. Bu sınıfların tüm veri seti içerisinde yüzdesel dağılım oranları farklılık gösterebilir. Aranan/tahmin edilmeye çalışılan sınıf niteliği değeri diğer sınıf nitelik değerlerine göre oldukça düşük bir yüzdesel dağılım oranına sahip ise bu durum dengesiz sınıf dağılımı olarak ifade edilmektedir. Örneğin; müşteri kayıp analizi probleminde aranan/tahmin edilmeye çalışılan sınıf "ayrılma eğilimli" müşterilerdir. Bir işletmenin tüm müşterileri göz önüne alındığında bu eğilimi gösteren müşterilerin diğerlerine oranla oldukça az sayıda olacağı açıktır. Dengesiz sınıf dağılımı için çeşitli örnekleme yöntemleri kullanılarak bu oranın dengelenmesine çalışılır.

***Dinamik müşteri kayıp analizi sistemlerinin geliştirilmesi*:** Veri setleri içerdiği değerler açısından sürekli değişim gösterebilen bir yapıya sahip olabilmektedir. Değişiklik gösteren veri değerleri nedeniyle bu şekildeki veri setleri dinamik veri olarak ifade edilmektedir. Belirli bir döneme ait veri ile bir model oluşturularak veri madenciliği çalışmalarının gerçekleştirilmesi özellikle karar alma seviyesi için her zaman etkin sonuçlar elde edilmesine olanak sağlamaz. Özellikle dinamik veri setleri için modelin takip edilmesi ve değişen veriye göre güncellenmesi gerekebilmektedir. Literatürde de bu özelliğin müşteri kayıp analizi problemi için de etkili olabileceği göz önünde bulundurularak çeşitli dinamik model ve sistem önerileri geliştirilmiştir.

***Ayrılan/ayrılma eğilimi gösteren müşteriler arası sosyal ağ analizinin gerçekleştirilmesi*:** Müşteri Kayıp Analizi sadece tahmin yapma amaçlı değil, müşterilerin ayrılma eğilimlerinin nedenlerinin de belirlenmesini içeren bir yönetim problemidir. Bu doğrultuda araştırmacılar gerek veri madenciliği gerek diğer istatistiksel yöntemler ile beraber ayrılma eğilimi gösteren

müşterilerin özellikle sosyal çevreleri tarafından etki altında kalıp kalmadığı konusunda da çeşitli analizler gerçekleştirmişlerdir.

***Müşteri kayıp analizi çözüm önerilerinin farklı alanlardaki benzer problemlere uygulaması*:** Müşteri Kayıp Analizi probleminin tarafları müşteriler ve işletmeler olarak kabul edilmektedir. Ancak ayrılma eğilimi gösteren her zaman bir müşteri olmayacağını düşünen araştırmacılar buradan hareket ile benzer eğilimi gösterebilecek unsurlar için Müşteri Kayıp Analizi problemine yönelik geliştirilen çözüm önerilerini kullanmışlardır. Bu doğrultuda müşterilerin yerini öğrenciler, çalışanlar; işletmelerin yerini isse okullar veya hizmet sunan değil işveren işletmeler almıştır.

Müşeri Kayıp Analizi problemi için gerçekleştirilen literatür incelemesi sonucu incelenen çalışmalar kronolojik olarak aşağıda verilmiştir.

Masand ve diğerleri (1999), cep telefonu müşterilerinin davranış özelliklerini modelleme temel amacı ile bir prototip geliştirmiştir. Bu çalışma sırasında mobil servis sağlayıcısına ait 200 nitelik alanından oluşan 500000 kayıt içeren bir veri seti kullanmıştır. Karar Ağaçları, K En Yakın Komşu, Genetik Algoritmalar ve Sinir Ağları yöntemlerinden yararlanılmıştır.

Mozer ve diğerleri (2000), kısa zaman aralığında ayrılma eğilimli müşteri olma durumunu tahmin etmek üzere bir model geliştirmiştir. Sunulan modelde, ayrılan müşteri, memnuniyetsizlik, karlılık ve kredi risk durumlarının tahmin edilmesi ele alınarak karar vericilere ortak bir paydada sonuçlar iletilmektedir. Çalışmada bir telekom şirketine ait 46744 kayıt içeren bir veri seti kullanılmıştır. Lojistik Regresyon, Karar Ağaçları ve Sinir Ağları ise model için kullanılan yöntemler arasındadır.

Datta ve diğerleri (2000), aylık periyotlarda ayrılma eğilimli müşteriler için bu duruma sebep olan faktörlerin yeniden belirlenerek gerekli güncellemelerin yapıldığı, Karar ağaçları ve Genetik Algoritmaların nitelik seçimi, Sinir Ağlarının ise tahmin etme için kullanıldığı CHAMP ayrılan müşteri tahmin etme sistemi geliştirilmiştir. Bu çalışma sırasında mobil servis sağlayıcısına ait 200 nitelik alanından oluşan 500000 kayıt içeren bir veri seti kullanmıştır. Geliştirilen prototip dışında kullanılan yöntemler Karar Ağaçları, K-En Yakın Komşu, Genetik Algoritmalar ve Sinir Ağları olmuştur.

Wei ve Chiu (2002), çağrı detayları içerisinde yer alan çağrı desenlerindeki değişimler ve sözleşme verileri kullanarak ayrılma eğilimli müşterilerin tahmin edilmesi amaçlamıştır. Önerilen yöntem ayrıca ayrılma eğilimli olan ve olmayan müşteri sınıfları arasındaki yüksek dengesiz dağılıma yönelik de çözüm sağlamaktadır. Çalışma için Tayvan'da hizmet veren bir telekom şirketine ait dokuz milyonun üzerinde kayıt içeren bir veri seti kullanılmıştır.

Au ve diğerleri (2003), ayrılma eğilimli müşteri desenlerini ortaya koyacak kurallar oluşturarak bir müşterinin yakın zamanda ayrılma durumunun tahmin edilmesini sağlayacak model önerisi sunmayı amaçlamıştır. Çalışma için sekiz farklı veri setinden yararlanılmıştır. Sinir Ağları, Karar Ağaçları ve Evrimsel Öğrenme ile Veri Madenciliği Algoritmaları çalışma kapsamında kullanılan yöntemler arasındadır.

Chiang ve diğerleri (2003), benzer şekilde ayrılma durumuna yatkın müşterilerin kaybedilmeden önce belirlenebilmesini sağlayan bir model önerisi getirmiştir. Sektör olarak bankacılık tercih edilmiş olup, Tayvan'da hizmet veren bir bankaya ait bir milyondan fazla kayıt içeren bir veri seti kullanılmıştır. Hedef Odaklı Sıralı Desen Algoritması, Küresel Odaklı Algoritma, Apriori Algoritmalarından yararlanılmıştır.

Buckinx ve Van den Poel (2005), müşterinin parçalı ayrılma durumunu (zaman içerisinde) tahmin etme amacıyla bir model geliştirmeyi amaçlamıştır. Çalışmada, perakende sektöründe faaliyet gösteren önemli şirketten temin edilmiş 61 nitelik alanı ve 158884 kayıt içeren bir veri seti kullanılmıştır. Lojistik Regresyon, Sinir Ağları ve Rastgele Orman Algoritması kullanılan yöntemler arasındadır.

Hung ve diğerleri (2006), çalışmalarında belirli periyotlarda her bir müşteri için "ayrılma eğilimli olma skoru" ataması sağlayacak şekilde modeller oluşturarak bu modellerin karşılaştırmalı sonuçlarına yer vermiştir. K-Ortalamalar, Sinir Ağları ve Karar Ağaçları çalışma için kullanılan yöntemler arasındadır.

Song ve diğerleri (2006), ayrılma eğilimli müşteri yönetimi alanında statik ve zamana bağlı sürekli veriler ile kullanılabilecek Sinir Ağları tabanlı bir modelin geliştirilmesini amaçlamıştır. China Mobile Communication Company'den temin edilen 12000 kayıt içeren bir veri seti kullanılmıştır.

Ahn ve diğerleri (2006), çalışmalarında işlem ve fatura verilerini kullanarak müşterinin ayrılma durumunu etkileyen bileşenlerin ortaya çıkartılmasını amaçlamıştır. Bu amaçla müşterilerin parçalı ayrılma durumunun (müşterinin bir anda ayrılması değil de belirli bir süre içerisinde işlemlerinde azalma ile beraber) ayrılmayı etkileyen bileşenler ile olan ilişkisi belirlenmeye çalışılmıştır. Güney Kore'de hizmet veren bir telekom şirketine ait 14 nitelik alanı ve 5789 kayıt içeren veri seti kullanılmıştır. Çalışmada yöntem olarak Lojistik Regresyon tercih edilmiştir.

Hadden ve diğerleri (2006), müşteri şikayetleri ile beraber çeşitli niteliklerin ayrılma eğilimli olma durumunu etkileme oranlarını ortaya koymayı ve üç farklı yöntem ile karşılaştırmalı bir çalışma yapmayı amaçlamıştır. Çalışmada 902 müşteriye ait kayıtlar kullanılmış olup, eğitim veri setinde ayrılma eğilimli müşterilerin oranı %50 iken test verisinde bu oran %30 olarak alınmıştır. Karar Ağaçları, Lojistik Regresyon ve Sinir Ağları çalışma için kullanılan yöntemler arasındadır.

Bin ve diğerleri (2007), ayrılma eğilimli olan-olmayan dağılımı, eğitim veri seti seçmek için kullanılan örnekleme yöntemleri, eğitim veri seti alt zaman periyotları olmak üzere 3 farklı faktörün ayrılma eğilimli müşterileri tahmin etme performansı üzerindeki etkisini incelemiştir. Çalışmada uygulama için telekom sektöründen elde edilen veri seti ve Karar Ağaçları yöntemi kullanılmıştır.

Zhang ve diğerleri (2007) tarafından, özellikle Lojistik Regresyona karşı daha iyi performans gösteren, hedef nitelikler ile tahminleyici nitelikler arasında lineer olmayan ilişkilerin varlığında kullanılabilecek hibrit sınıflandırıcı model önerisi getirilmiştir. Çalışmada 5 farklı veri seti ve K-En Yakın Komşu, Karar ağaçları ve Lojistik Regresyon yöntemleri kullanılmıştır.

Burez ve Van den Poel (2007), ayrılma eğilimli müşteri analizi için Rastgele Orman Algoritması tabanlı yeni bir tahmin modeli önerisinde bulunarak, var olan müşterinin elde tutulmasına yönelik farklı uygulamalar sunmuştur. Çalışmada ayrıca Lojistik Regresyon ve Markov zincirlerine yer verilmiştir. Uygulama için bir TV şirketine ait %15 ayrılma eğilimli müşteri oranına sahip veri setinden yararlanılmıştır.

Coussement ve Van den Poel (2008b)'in çalışmalarında, geleneksel veriler dışında e-posta gibi yapısal olmayan verilerin de analizlere katılmasını sağlayarak, tahmin etme performansındaki değişikliği gözlemlemeyi amaçlamıştır. Belçika'da hizmet veren bir yayın organına ait 3161

kayıt içeren veri seti ile uygulamalar gerçekleştirilmiştir. Lojistik Regresyon uygulamada kullanılan yöntemler arasındadır.

Burez ve Van den Poel (2008), statik ve dinamik ayrılan müşteri tahmin modelleri kullanarak finansal sorun yaşayarak ayrılma durumu gösteren müşteriler ile ilgili belirlemeler yapmaya çalışmıştır. Bir TV şirketine ait veri seti kullanılarak dinamik ve statik olmak üzere iki farklı model oluşturulmuştur. Modellerde Rastgele Orman Algoritması kullanılmıştır.

Anil Kumar ve Ravi (2008), bankaların kredi kartı müşteri verileri üzerinde ayrılan müşterileri tahmin etmek üzere çoklu sınıflandırıcıların yer aldığı farklı modeller uygulayarak bu alanda çözüm önerileri sunmayı amaçlamıştır. Uygulama için bir bankaya ait 22 nitelik alanı, 14814 kayıt içeren ve ayrılma eğilimli müşteri oranı %6,76 olan bir veri seti kullanılmıştır. Sinir Ağları, Lojistik Regresyon, Destek Vektör Makineleri, Rastgele Orman Algoritması gibi çeşitli yöntemlerin yanı sıra veri setindeki dengesiz dağılım nedeniyle farklı örnekleme yöntemlerine de başvurulmuştur.

Zhao ve Dang (2008), banka müşteri verisi kullanarak Destek Vektör Makineleri tabanlı bir model önerisi geliştirmiştir. Kullanılan veri seti 2382 müşteriye ait kayıt içermektedir. Çalışmada Destek Vektör Makinelerinin yanı sıra Karar Ağaçları, Lojistik Regresyon, Sinir Ağları, Sade Bayes yöntemleri de kullanılmıştır.

Coussement ve Van den Poel (2008a), ayrılma eğilimli müşterileri tahmin etmek için Destek Vektör Makineleri kullanarak, paremetre optimizasyonu prosedürünün tahminleme başarısı üzerindeki etkisine dikkat çekmiştir. Veri seti olarak Belçika'da hizmet veren bir yayın kuruluşunun müşteri kayıtları kullanılmış olup, Destek Vektör Makineleri, Lojistik Regresyon ve Rastgele Orman Algoritması yöntemlerinden yararlanılmıştır.

Xia ve Jin (2008), yapısal risk minimizasyonu temel alarak farklı kernel fonksiyonları ile Destek Vektör Makineleri algoritmasını kullanmış, algoritmanın performansını farklı algoritmalar ile karşılaştırmıştır. Çalışma için telekom sektörüne ait iki farklı veri seti kullanılmıştır. Sinir Ağları, Karar Ağaçları, Lojistik Regresyon, Sade Bayes gibi yöntemlerden yararlanılmış olup, veri setinin örneklenmesi için de çeşitli yöntemlere başvurulmuştur.

Tsai ve Lu (2009), ilk aşaması verinin küçültülmesi ve gereksiz verinin atılmasını sağlayan toplam iki aşamalı yapay sinir ağları temelli iki farklı hibrit model önerisi getirmiştir. Uygulama

alanı olarak telekom sektörü tercih edilmiştir. Modeller Sinir Ağları yöntemine dayalı çeşitli algoritmalar ile oluşturulmuştur.

Coussement ve Van den Poel (2009), genellikle yapısal veriler üzerine gerçekleştirilen çalışmaların aksine yapısal olmayan verilerde yer alan duygu içerikli ifadelerin de analiz edilerek, ayrılan müşteri tahmin performansının arttırılmasını amaçlamıştır. Belçika'da hizmet veren bir gazetenin 11836 müşterisine ait kayıtları ve 18331 adet e-posta kullanılarak uygulamalar gerçekleştirilmiştir. Veri setindeki ayrılma eğilimli müşteri oranı %18,89 iken e-postalar %35,79 şikayet içeriklidir. Rastgele Orman Algoritması, Destek Vektör Makineleri ve Lojistik Regresyon geliştirilen modeller için kullanılan yöntemler arasındadır.

Wu ve diğerleri (2009), daha önceki yöntemleri de inceleyerek nitelik seçimi için bir yapı oluşturmak amacıyla Makine Öğrenmesi, İstatistik, Metin Kategorileme ve Örüntü Tanıma olmak üzere 4 farklı alan için nitelik seçimi karşılaştırması yapmışlardır. Çalışmada 300 kayıt 206 nitelik alanı içeren bir veri setinden yararlanılmıştır.

Popović ve Bašić (2009), banka müşterilerine ait veriden ayrılma eğilimli müşteri analizi yapabilmek ve var olan tahmin modellerinden daha iyi bir performans gösterecek yöntem elde etmek için bulanık mantık tabanlı yöntem uygulamıştır. Fuzzy C-Ortalamalar, Kanonik Diskriminant Analizi, K-Ortalamalar gibi yöntemlerden yararlanılmıştır. Sektör olarak bankacılık tercih edilmiş, 5000 kayıt ve 73 nitelik alanı içeren bir veri seti kullanılmıştır.

Pendharkar (2009), yapmış olduğu çalışmada genetik algoritma tabanlı iki farklı sinir ağı modeli önermiştir. Uygulama için kullanılan veri seti telekom sektöründen seçilmek ile beraber 195956 kayıt ve 6 nitelik alanına sahiptir.

Xie ve diğerleri (2009), çalışmalarında dengesiz sınıf dağılımının yapılan analiz sonuçlarını olumsuz etkileyeceğini belirtmek ile beraber; maliyet duyarlı algoritmalar ve çeşitli örnekleme yöntemlerini de analizlere dâhil ederek performans açısından daha iyi bir yöntem elde etmişlerdir. Çin'de hizmet veren bir bankaya ait 1524 kayıt ve 15 nitelik alanı içeren bir veri seti kullanılmıştır. Rastgele Orman Algoritması, Yapay Sinir Ağları ve Karar Ağaçları Algoritmasından yararlanılmak ile beraber çeşitli örnekleme yöntemlerine de başvurulmuştur.

Burez ve Van den Poel (2009)'ın çalışması veri setlerindeki dengesiz sınıf dağılımına yönelik olarak çözüm üretmeyi amaçlamaktadır. Çalışmada çeşitli sektörlerden olacak şekilde 6 farklı

veri setinden yararlanılmıştır. Rastgele Orman Algoritması ve Lojistik Regresyon kullanılan yöntemler arasındadır.

Bose ve Chen (2009), danışmanlı ve danışmansız yöntemlerin bir arada kullanıldığı hibrit yöntem önerisinde bulunmuşlardır. Çalışmada telekom sektöründen olmak üzere 3 farklı veri setinden yararlanılmıştır. Sinir Ağları, Karar Ağaçları, K-Ortalamalar başta olmak üzere farklı yöntemler kullanılmıştır.

Wang ve diğerleri (2009), çalışmalarında wireless ağ şirketleri için müşteri kaybı problemini anlama ve çözüm önerisi sunma konusunda destek olacak bir öneri sistemi geliştirmeyi amaçlamıştır. Bu sektörden 60000 kaydın yer aldığı bir veri seti kullanılmıştır. Karar Ağaçları yönteminden yararlanılmıştır.

Coussement ve diğerleri (2010), ayrılma durumunda olan riskli müşterilerin belirlenmesi için Genelleştirilmiş Toplamsal Modeller ve Lojistik Regresyon yöntemlerini kullanarak model önerisinde bulunmuşlardır. Belçika'da hizmet veren bir gazete şirketine ait 134084 kayıt içeren ve %11,95 ayrılma eğilimli müşteri oranına sahip veri seti kullanılmıştır.

Risselada ve diğerleri (2010), gerçekleştirdikleri çalışmada bir tahminleyici modelin tahmin süresinden sonraki zaman periyotlarında tahmin gücündeki değişikliğin ölçülmesi üzerinde durmuştur. Çalışma için bir internet hizmet sağlayıcısı ve sağlık sigorta şirketinden temin edilmek üzere iki farklı veri seti kullanılmıştır. Veri setlerinden çeşitli alt örnekler seçilerek modeller oluşturulmuştur. Karar Ağaçları ve Lojistik Regresyon kullanılan yöntemler arasındadır.

Radosavljevik ve diğerleri (2010), çalışmaları ile kullanılan nitelikler, ayrılan müşteri tanımı ve örneklem grubu üzerinde deneysel çalışmalar yaparak ön ödemeli müşterilerin ayrılma durumlarını tahmin etme performansının arttırılmasına yönelik katkı sağlamaya çalışmıştır. Telekom sektöründe gerçekleştirdikleri uygulamaları için oluşturdukları modellerde Karar Ağaçları ve Lojistik Regresyon yöntemlerinden yararlanmışlardır.

Huang ve diğerleri (2010), hem var olan niteliklerden ayrılan müşteri tahmin oranlarını iyileştirecek yönde yeni girdi nitelikleri oluşturmak, hem de yeni pencere tekniklerinin elde edilmesi ve bu tekniklerin birbiri ile karşılaştırılmasını amaçlamıştır. İrlanda kökenli bir

telekom şirketine ait 47391 kayıt içeren veri seti kullanılmıştır. Geliştirilen modeller için Destek Vektör Makineleri, Sinir Ağları ve Karar Ağaçlarından yararlanılmıştır.

Khan ve diğerleri (2010), literatürde yer alan diğer çalışmalarda da sıkça ve vazgeçilmeden kullanılan demografik, fatura ve kullanım verilerinin önemini araştırmayı yine bu verilere göre çeşitli algoritmalarının performanslarını karşılaştırmayı amaçlamıştır. Bu kapsamda Tahran'da hizmet veren bir telekom firmasına ait 2164261 kayıt ve 36 nitelik alanı içeren bir veri seti kullanılmıştır. Karar Ağaçları, Sinir Ağları, Lojistik Regresyon ve K-Ortalamalar yöntemlerinden yararlanılmıştır.

Huang ve diğerleri (2010), müşteri kayıp analizi problemi için nitelik seçimine yönelik olarak Genetik Algoritma temelli bir yöntem önerisi getirmişlerdir. Çalışma için 18600 kayıt içeren ve telekom sektöründen elde edilen bir veri seti kullanılmıştır. Genetik Algoritmanın yanı sıra Karar Ağaçlarından faydalanılmıştır.

Tsai ve Chen (2010)'in çalışmalarında da amaç nitelik seçimi olmuş, bu seçim işlemi için çeşitli yöntemler kullanılarak elde edilen yeni nitelik kombinasyonları ile analizler gerçekleştirilerek performans sonuçları değerlendirilmiştir. Çalışma için Tayvan'da hizmet veren bir telekom şirketine ait 37882 kayıt içeren bir veri seti kullanılmıştır. Karar Ağaçları ve Sinir Ağlarının yanı sıra Birliktelik Kurallarına da bu çalışmada yer verilmiştir.

Owczarczuk (2010), önceden fatura ödemeli müşteri verileri üzerinden yola çıkılarak çok fazla sayıdaki nitelikler arasından en iyi niteliklerin seçilerek belirlenen yöntemlerin uygulanması ve performanslarının karşılaştırılmasını amaçlamıştır. Polonya'da hizmet veren bir telekom şirketine ait 85274 kayıt ve 1381 nitelik alanına sahip bir veri seti uygulama için kullanılmıştır. Lojistik ve Lineer Regresyon, Karar Ağaçları uygulamada kullanılan yöntemlerdendir.

Huang ve diğerlerinin (2010), başka bir çalışmasında müşteri kayıp analizinde kullanılmak üzere girdi nitelikleri ve sınıf niteliği arasındaki bağımlılığı hesaplama mantığına dayalı yeni bir nitelik seçim yöntemi önerisi getirilmiştir. Bir telekom şirketine ait 147214 kayıt ve 122 nitelik alanına sahip veri setinin kullanıldığı çalışmada Karar Ağaçları, Sade Bayes, Destek Vektör Makineleri gibi yöntemlere yer verilmiştir.

Kraljevic ve Gotovac (2010), genel olarak, hassas ve güvenilir şekilde tahmin etmeye olanak sağlayacak girdi niteliklerini bularak, ayrılma eğilimli müşterileri başarılı bir şekilde tahmin

edecek bir model önerisi hazırlamayı amaçlamıştır. Bosna Hersek'te hizmet veren bir telekom şirketine ait kayıtlar kullanılmıştır. Sinir Ağları, Karar Ağaçları ve Lojistik Regresyon yöntemlerinden yararlanılmıştır.

Tamaddoni Jahromi ve diğerleri (2010), kümeleme ve sınıflandırma yöntemlerini içeren ayrılan müşteri durumunu belirleyen ve tahmin eden iki adımlı bir model geliştirmiştir. Diğer çalışmalarda ağırlıkta olduğu üzere bu çalışmada da telekom sektöründe bir uygulama tercih edilmiştir. İran'da hizmet veren bir telekom şirketine ait 34504 kayıt içeren bir veri seti kullanılmıştır. Karar Ağaçları ve Maliyet Duyarlı Öğrenme Algoritmalarından yararlanılmıştır.

Delen (2010), öğrencilerin okula başladıkları yıldan itibaren ilerleyen yıllarda okulu bırakmaya eğilimlerini ortaya çıkarmayı ve okulu bırakacak öğrencileri tahmin edecek modelleri uygulamaya koymuştur. Delen ayrıca, ayrılma eğiliminin nedenlerini de incelemeye almış bu nedenleri sosyal etkileşim, eğitimden beklentiler, ailelerinin maddi ve eğitim durumları şeklinde sıralamıştır. Delen, Amerika'da eğitim veren bir enstitüye ait 16066 kayıt içeren bir veri seti kullanmıştır. Karar Ağaçları, Destek Vektör Makineleri, Sinir Ağları başta olmak üzere çeşitli örnekleme yöntemlerinden de yararlanmıştır.

Lima ve diğerleri (2011), geriye dönük test çerçevesinin ayrılan müşteri analizinde kullanımını ve tahmin etme performansı üzerindeki etkisini incelemeyi amaçlamıştır. Bu bağlamda 20000 kayıt içeren telekom sektöründen bir veri seti yardımıyla Lojistik Regresyon ve Karar Ağaçları yöntemlerinin kullanıldığı bir uygulama gerçekleştirmiştir.

Fasanghari ve Keramati (2011), ayrılan müşteri olma durumuna neden olan niteliklerin ortaya çıkartılmasını amaçlamıştır. Bu kapsamda İran'da hizmet veren bir telekom şirketine ait 3150 kayıt ve 11 nitelik alanı içeren bir veri seti kullanılarak analizler gerçekleştirilmiştir. Sinir Ağları temelli algoritmalar kullanılmıştır.

De Bock ve Van Den Poel (2011), ayrılan müşterilerin tahmin edilmesi için bir modelleme tekniği olarak kullanılmak üzere rotasyon tabanlı topluluk sınıflandırıcı yöntem önerisinde bulunmayı amaçlamıştır. Banka ve telekomun da içerisinde olduğu çeşitli sektörlerden 4 farklı veri seti kullanılmıştır. Karar Ağaçları ve Rastgele Orman Algoritması başta olmak üzere çeşitli yöntemler analizler için kullanılmıştır.

Abbasimehr ve diğerleri (2011), Uyarlamalı Sinirsel Bulanık Çıkarım Sistemi (ANFIS) tabanlı iki farklı model önerisi getirmiştir. Çalışma kapsamında herkese açık yayınlanmış 5000 kayıt ve 20 nitelik alanı içeren ve %14.13 ayrılma eğilimli müşteri oranına sahip veri seti kullanılmıştır. Model performansının karşılaştırılması için Karar Ağaçları kullanılırken veri setindeki dengesiz sınıf niteliği dağılımına karşı örnekleme yöntemleri de modellere dahil edilmiştir.

Yu ve diğerleri (2011), e-ticaret müşteri davranışlarını belirlemek ve ayrılan müşterileri tahmin etmek üzere, dağınık ve doğrusal olmayan verilerin de kullanılabildiği genişletilmiş Destek Vektör Makinelerini önermiştir. Veri seti olarak Çin'de hizmet veren bir e-ticaret firmasına ait %10 ayrılma eğilimli müşteri oranı ve 150000 kayıt içeren bir veri seti kullanılmıştır. Performans karşılaştırması için Destek Vektör Makineleri, Karar Ağaçları, Sinir Ağları yöntemleri kullanılmıştır.

Kisioglu ve Topcu (2011), müşterilerin ayrılma eğilimlerini tahmin etmek amacıyla Bayes ağları temelli bir yöntem ile beraber ayrılan müşteri karakteristiklerini belirlemeye çalışmıştır. Çalışmada veri seti olarak Türkiye'de hizmet veren bir telekom şirketine ait 2000 kayıt ve 9 nitelik alanı içeren bir veri seti kullanılmıştır.

Keremati ve Ardabili (2011), müşterilerin ayrılma durumunu etkileyen faktörleri ortaya çıkarmayı ve nitelikler arası ilişkilerin belirlenmesini amaçlamıştır. Çalışmada 3150 kayıt içeren telekom sektöründen bir veri seti kullanılmış olup, Lojistik Regresyon yönteminden yararlanılmıştır.

Nie ve diğerleri (2011), müşterinin ayrılma durumundaki en önemli faktörlerin özellikle maliyetin de göz önünde bulundurularak ortaya çıkartılmasını amaçlamıştır. Maliyete dayalı olarak "yanlış sınıflandırma maliyeti" niteliği oluşturulmuş ve tasarlanan modellerde kullanılmıştır. Çalışma için kullanılan Lojistik Regresyon ve Karar Ağaçları algoritmalarının performans karşılaştırmaları da çalışma içerisinde yapılmıştır. Veri seti olarak Çin'de hizmet veren bir bankanın müşterilerine ait 8 milyon kayıt ve 135 nitelik alanı içeren, %8.1 ayrılma eğilimli müşteri oranına sahip bir veri seti kullanılmıştır.

Dierkes ve diğerlerinin (2011) çalışmalarındaki temel amaç sosyal ağdaki etkileşimin ayrılma eğilimine etkisini incelemek olmuş, bu doğrultuda çeşitli algoritmalar ile farklı modeller

oluşturularak modellerin performans karşılaştırmaları gerçekleştirilmiştir. Çalışmada 120000 kayıt ve 32 nitelik alanı içeren telekom sektöründen temin edilen bir veri seti kullanılmıştır.

Yeshwanth ve diğerleri (2011), geliştirdikleri hibrit modelin yanı sıra müşteri üzerinde çevre etkisi incelemesini de analizlere dâhil etmişlerdir. Bu bağlamda Karar Ağaçları ve Genetik Algoritmalardan yararlanılmıştır. Gelişmekte olan bir ülkede hizmet veren telekom şirketine ait 1,2 milyon kayıt içeren veri seti analizlerde kullanılmıştır.

Karahoca ve Karahoca (2011), çalışmalarında nitelik seçimine yönelik algoritmalar kullanmış ve elde edilen sonuçlar ile ayrılma eğilimli müşterileri tahmin etme konusunda performans karşılaştırılması yapmıştır. Veri seti olarak Türkiye'de hizmet veren bir telekom şirketine ait 24900 kayıt içeren veri seti kullanılmıştır.

Keramati ve Ardabili (2011), Bionominal Lojistik Regresyon ile ayrılma eğilimli müşteri olma durumunu etkileyen esas faktörleri belirlemeyi amaçlamıştır. Telekom şirketinin çağrı merkezinden temin edilen 3150 kayıt içeren bir veri seti kullanılmıştır.

Saradhi ve Palshikar (2011), organizasyon şirketinin kendi isteği ile işinden ayrılan ve şirketin işine son verdiği personellere ait veri seti üzerinde bir inceleme gerçekleştirerek, Sade Bayes, Rastgele Orman ve Destek Vektör Makineleri algoritmalarının kullanıldığı üç model geliştirmiştir. Çalışmanın uygulama tarafında bir organizasyonda çalışan 1575 kişiye ait ve 25 nitelik alanı içeren bir veri seti kullanılmıştır.

Verbeke ve diğerleri (2011)'nin gerçekleştirdiği çalışmada geniş bir literatür araştırması ile iki yeni modelin önerisi sunulmuş, bu iki modelin özellikle kural çıkarım tabanlı geleneksel teknikler ile karşılaştırması gerçekleştirilmiştir. Çalışmada telekom sektöründen 5000 kayıt, 21 nitelik alanına ve %14.3 ayrılma eğilimli müşteri oranına sahip herkese açık olarak yayınlanmış bir veri seti kullanılmıştır. Karınca Kolonisi Optimizasyonu, Karar Ağaçları, Lojistik Regresyon, Destek Vektör Makineleri gibi çeşitli yöntemlerden yararlanılmıştır.

Idris ve diğerleri (2012), çalışmalarında çeşitli nitelik azaltma yöntemleri ile işbirliği içerisinde dengesiz sınıf dağılımını işlemek için Parçacık Sürü Optimizasyonu tabanlı bir örnekleme metodunun önemini incelemeyi amaçlamıştır. Çalışmanın uygulama kısmında telekom şirketine ait 50000 kayıt ve 260 nitelik alanına sahip bir veri seti kullanılmıştır. Geliştirilen

modeller ve performans karşılaştırmalarında Rastgele Orman Algoritması ve K-En Yakın Komşu Algoritmalarından yararlanılmıştır.

Verbeke ve diğerleri (2012)'nin çalışması iki ana bölümden oluşmakla beraber ilk bölümde en iyi model ve müşteri grubunu seçen fayda esaslı bir ölçü geliştirilmiştir. İkinci bölümde de istatistiksel ve fayda esaslı geliştirilen ölçüler ile beraber farklı algoritmalar kullanılarak performans karşılaştırması yapılmıştır. Telekom sektöründen olmak üzere 10 farklı veri seti kullanılarak analizler gerçekleştirilmiştir. Karar Ağaçları, Destek Vektör Makineleri, K-En Yakın Komşu Algoritması, Sinir Ağları, İstatistik Tabanlı modeller olmak üzere çok çeşitli yöntemlere çalışmada yer verilmiştir.

Prasad ve Madhavi (2012), banka müşterilerinin satın alma davranışlarını belirlemeye yönelik yöntemler üzerinde çalışılmıştır. Hindistanda hizmet veren bir bankaya ait 1481 kaydın yer aldığı veri setinin kullanıldığı çalışmada Karar Ağaçları yönteminden faydalanılmıştır.

Migueis ve diğerleri (2012)'nin gerçekleştirdiği çalışmada perakende sektöründe (market) bir uygulama yaparak, bu uygulamada parçalı ayrılan müşterilerin tahmin edilebilmesi için modelleme yapmak ve müşterilerin ilk alışveriş deneyimlerini de baz alarak bir çalışma ortaya koymak amaçlanmıştır. Avrupa'da hizmet veren bir marketler zinciri müşterilerine ait 74607 kayıtların yer aldığı ve %44 ayrılma eğilimli müşteri oranına sahip bir veri seti kullanılmıştır. Lojistik Regresyon yönteminden yararlanılmıştır.

De Bock ve Van Den Poel (2012), Genelleştirilmiş Toplamsal Modeller tabanlı hem performans hem de sonuçların yorumlanabilirliği açısından yüksek bir topluluk sınıflandırıcı yöntemi önermiştir. Farklı sektörlerden 6 farklı veri setinin kullanıldığı çalışmada Rastgele Orman Algoritması ve Lojistik Regresyondan da yararlanılmıştır.

Xiao ve diğerleri (2012) tarafından, dengesiz dağılım gösteren veri için topluluk sınıflandırıcılar ve maliyet-duyarlı öğrenme yöntemleri ile dinamik bir ayrılma eğilimli müşteri tahmin yöntemi geliştirilmiştir. İki farklı veri setinin kullanıldığı çalışmada Genetik Algoritmalar, Sinir Ağları ve Rastgele Orman Ağaçları yöntemleri tabanlı çeşitli algoritmalar ele alınmıştır.

Huang ve diğerleri (2012), tarafından gerçekleştirilen çalışmada var olan niteliklerden yeni nitelik seti elde edilmeye çalışılarak, belirlenen 7 veri madenciliği tekniği uygulanmıştır.

Çalışma kapsamında İrlanda'da hizmet telekom şirketine ait 827124 kayıt ve 738 nitelik alanı içeren veri seti kullanılmıştır. Çalışmada kullanılan yöntemler arasında Lojistik Regresyon, Sade Bayes, Sinir Ağları, Destek Vektör Makineleri ve Karar Ağaçları yer almaktadır.

Shim ve diğerleri (2012), çevrimiçi alışveriş mağazasının işlem verisinden elde edilen birliktelik kuralları ve desenler ile Müşteri İlişkileri Yönetimi için strateji önerileri geliştirmeyi amaçlamıştır. Kore'de hizmet veren çevrimiçi alışveriş mağazasına ait 3445 müşterinin 14782 işlem kaydı ve 9 nitelik alanının yer aldığı veri seti çalışmanın analizlerinde kullanılmıştır. Yöntemler arasında Karar Ağaçları, Lojistik Regresyon, Sinir Ağları, Birliktelik Kuralları yer almaktadır.

Ballings ve Van den Poel (2012)'in gerçekleştirdiği çalışmada, müşteri kayıp analizinde iyi bir tahmin yapabilmek için veri setinin hangi süre aralığına ait olmasının yeterli olacağının daha farklı bir ifadeyle zaman penceresi optimizasyonunun tahmin etme performansına etkisinin araştırılması amaçlanmıştır. Bir gazete şirketine ait 129892 kayıt ve yaklaşık %11 ayrılma eğilimli müşeri oranına sahip veri seti seti kullanılara 31-54 arasında değişen sayılarda nitelik alanı ile analizler gerçekleştirilmiştir.

Chen ve Fan (2012), zaman serilerinden, metin verilerine kadar saklanan ve veri tipinde çeşitlilik gösteren, veri tabanlarında bulunan bu çoklu verilerin kullanılabildiği İşbirlikçi Çoklu Kernel Destek Vektör Makinesi (Collaborative Multiple Kernel Support Vector Machine, C-MK-SVM) yaklaşımı önermiştir. Perakende sektöründe 1560 ürüne ve 10281 müşteriye ait 200000 kayıttan meydana gelen veri seti ile gerçekleştirilen uygulamalarda Destek Vektör Makineleri tabanlı algoritmalar kullanılmıştır.

Kim ve Lee (2012), çok boyutlu veri setleri için daha iyi tahmin etme performansına sahip bir yöntem önerisi yapmayı amaçlamıştır. Elektronik ticaret sektörü müşterileri üzerinde bu probleme çözüm arayan araştırmacılar 5000 kayıt 330 nitelik alanı ve %10 ayrılma eğilimli müşteri oranına sahip bir veri seti kullanmışlardır. Çalışmada Sinir Ağları, Lojistik Regresyon, K-Ortalamalar yöntemlerinden yararlanılmıştır.

Zhang ve diğerleri (2012), müşteriler arası etkileşimin önemi üzerinde durmuş ve bu etkileşimi baz alarak bir tahmin modeli önerisi yapmıştır. Çalışmada bir mobil servis sağlayıcıdan temin edilen veri seti ile Karar Ağaçları, Lojistik Regresyon ve Sinir Ağları yöntemleri kullanılmıştır.

Chen ve diğerleri (2012), boylamsal davranış verilerinin dönüşüme uğramadan kullanılabildiği Destek Vektör Makineleri yöntemi temelli bir hibrit yöntem geliştirmişlerdir. Farklı sektörlerden olmak üzere üç veri seti kullanılmıştır. Destek Vektör Makineleri ve Sinir Ağları temelli çeşitli algoritmaların yanı sıra, Karar Ağaçları, Lojistik Regresyon, Ratgele Orman Algoritması gibi yöntemlerden de yararlanılmıştır.

Huang ve Kechadi (2013), tüm veri seti ile tahmin yapmak yerine, sınıfı tahmin edilecek örnek ile benzer karakteristikleri gösteren verinin eğitim seti olarak kullanılması ve buna bağlı tahmin yapılması için danışmanlı ve danışmansız yöntemleri içeren hibrit bir model geliştirmeyi amaçlamıştır. Çalışma kapsamında 104199 kayıt ve 121 nitelik alanının yer aldığı telekom sektöründen elde edilen bir veri seti kullanılmıştır. K-Ortalamalar, Destek Vektör Makineleri, Karar Ağaçları ve Lojistik Regresyon gibi temel yöntemlerin yanı sıra farklı algoritmalardan da yararlanılmıştır.

Verbraken ve diğerleri (2013), son kullanıcıların ana hedefleri ile uyumlu olacak şekilde performans ölçülerini belirlemek amacıyla fayda-maliyet çerçevesini oluşturmuş, bu yapıyı ayrılan müşterilerin tahmin edilmesinde kullanılabilir hale getirmiştir. Hepsi telekom sektöründen elde edilmiş olan 10 farklı veri seti kullanılmıştır. Çalışmada, Karar Ağaçları, Destek Vektör Makineleri, Sinir Ağları, K-En Yakın Komşu gibi yöntemlerin yanı sıra çeşitli kural çıkarımlı ve istatistik tabanlı yöntelere de yer verilmiştir.

Sharma ve Panigrahi (2013), sinir ağları temelli bir yaklaşım önerisinde bulunmuştur. Geliştirilen modelin denemeleri için telekom sektöründen alınan ve herkese açık yayınlanmış bir veri seti kullanılmıştır. Bu veri seti 2427 kayıt ve 20 nitelik içermektedir.

Coussement ve De Bock (2013)'ın gerçekleştirdiği çalışmada, müşteri kayıp analizinin farklı bir alanı olarak bahis siteleri üyeleri/oyuncularının (poker oyuncuları) ayrılma durumlarının analiz edilmesi ele alınmıştır. Çalışmada ayrıca tekil algoritma uygulamaları ile son dönemde popülaritesi artan çoklu sınıflandırıcıların (topluluk sınıflandırıcı) performans karşılaştırılması yapılmaya çalışılmıştır. Çalışmada ayrılma durumunun yanı sıra bu durumda etkili olan niteliklerin analizi de yapılmıştır. Çalışmada kullanılan veri seti 3729 kayıt ve 60 nitelik alanı içermektedir.

Abbasimehr ve Alizadeh (2013), nitelik seçimi çalışmaları için Genetik Algoritma tabanlı bir model önerisi getirmiştir. Bu çalışmalarında telekom sektöründen olacak şekilde iki farklı veri

seti kullanmışlardır. Çalışmada, Genetik Algoritmaların yanı sıra Karar Ağaçları ve çeşitli nitelik seçim algoritmalarına yer verilmiştir.

Phadke ve diğerleri (2013), müşterilerin birbirleri arasındaki çağrıları yani çağrı desenlerini inceleyerek, sosyal ağ analizini de içeren bir model geliştirmiştir. Model ile beraber ayrılan müşterinin sosyal çevresine olan etkisinin dağılımı ve ölçüsü belirlenmeye çalışılmıştır. Araştırmacılar çalışmada telekom sektöründen elde edilen ve 500000 müşteriye ait veri içeren veri seti kullanmışlardır.

Kirui ve diğerleri (2013), çalışmalarında literatürde var olan çalışmaların ayrılan müşterileri tahmin etme performansını arttıracak şekilde yeni nitelik setlerinin oluşturması ana amacıyla hareket etmiştir. Çalışmada, wireless hizmeti veren bir telekom şirketine ait 106405 kayıt, 112 nitielik ve %5,6 ayrılma eğilimli müşteri oranına sahip bir veri seti kullanılmıştır.

Idris ve diğerleri (2013), topluluk sınıflandırıcılar ve nitelik çıkarım yöntemlerinin kullanıldığı bir zeki ayrılan müşteri tahmin sistemi geliştirmeyi amaçlamıştır. Çalışmada telekom sektöründen olmak üzere iki farklı veri seti ile çalışılmıştır. Veri setlerinde ayrılma eğilimli müşteri oranları %7,3 ve %50 şeklindedir. Çalışmada Rastgele Orman algoritması ve Rotasyon Orman Algoritması başta olmak üzere çeşitli yöntemlere yer verilmiştir.

Mohammadi ve diğerleri (2013), çalışmalarında literatürde hiyerarşik modellerin ön plana çıktığını belirterek 4 farklı yöntemin farklı kombinasyonları ile oluşturulmuş 3 hiyerarşik modelin performans karşılaştırmasını yapmıştır. İran'da hizmet veren bir telekom şirketinden temin edilen 3150 müşteriye ait kaydın yer aldığı veri seti kullanılmıştır. Çalışmada kullanılan yöntemlerin Sinir Ağları temelli olduğu görülmektedir.

Abbasimehr ve diğerleri (2013), çalışmalarında tüm müşterilerden ziyade işletme için yüksek değerli müşterilerin ayrılma eğilimi durumunun incelenmesi üzerinde durmuştur. Amerika'da hizmet veren bir telekom şirletine ait 100000 kayıt ve 172 nitelik alanına sahip ayrılma eğilimli müşteri oranı %50 olan bir veri seti üzerinde analizler gerçekleştirilmiştir. Kullanılan yöntemler arasında K-Ortalamalar, Sinir Ağları temelli alforimalar ve Fuzzy temelli algoritmalar olduğu görülmektedir.

Günther ve diğerleri (2014), aylık dilimler olacak şekilde zaman-dinamikli tahmin yapan ve nitelikler arasındaki etkileşimin de kullanıldığı Lojistik Regresyon tabanlı bir model önerisinde

bulunmuşlardır. Çalışma kapsamında Norveç'te sigorta sektöründe hizmet veren bir işletmeye ait 127961 kayıt içeren veri seti kullanılmıştır.

Kim ve diğerleri (2014), çağrı kayıtları ve müşterilerin iletişim ağını dikate alarak ağ analizinin de göz önünde bulundurulduğu bir yöntem önerisi getirmiştir. Telekom sektöründe 89412 kayıt, 36 nitelik ve %9,7 ayrılma eğilimli müşteri oranına sahip veri seti kullanılmıştır. Lojistik Regresyon ve Sinir Ağları yöntemlerinin çalışma kapsamında kullanıldığı gözlemlenmiştir.

Farquad ve diğerleri (2014), müşteri ilişkileri yönetimi amaçlarına göre ayrılan müşterilerin tahmin edilebilmesi için Destek Vektör Makineleri temelli hibrit yöntem geliştirerek kurallar oluşturmayı amaçlamışlardır. Veri seti olarak bir bankaya ait 14814 kayıt ve 22 nitelik alanı içeren, %6,76 ayrılma eğilimli müşteri oranına sahip bir veri seti kullanılmıştır. Destek Vektör Makinelerinin yanı sıra Sade Bayes sınıflandırıcıya ve çeşitli örnekleme yöntemlerine çalışmada yer verilmiştir.

Almana ve diğerleri (2014), telekom sektöründe müşteri kayıp analizi için kullanılan yöntemler için bir literatür taraması ve gelecek çalışmalara yönelik önerilerin sunumu yapmıştır. Çalışma kapsamında incelenen çalışmalarda Sinir Ağları, Lineer ve Lojistik Regresyon, Sade Bayes, K-En Yakın Komşu ve Karar Ağaçları gibi yöntemlerin ön plana çıktığı görülmektedir.

Adwan ve diğerleri (2014), ayrılan müşterilerin tahmin edilebilmesi için Çok Katmanlı Algılayıcı Yapay Sinir Ağları (ÇKA-YSA) topolojilerinin kullanımı ile beraber ÇKA-YSA tabanlı modeller ve bu modellerin karşılaştırmalarını çalışmalarında sunmuştur. Bu çalışmada da telekom sektörü tercih edilmiş, bir telekom şirketine ait 5000 kayıt ve 11 nitelik alanı içeren, ayrılma eğilimli müşteri oranı %7,6 olan bir veri seti kullanılmıştır.

Keramati ve diğerleri de (2014), Karar Ağaçları, Yapay Sinir Ağları, Destek Vektör Makineleri ve K-En Yakın Komşu Algoritması gibi algoritmaları kullanarak bir hibrit yöntem önerisinde bulunmuşlardır. Bu hibrit modelin geliştirilmesinde İran'da hizmet veren bir telekom şirketine ait 3150 kayıt içeren bir veri seti kullanılmıştır.

Gür Ali ve Arıtürk (2014), farklı zaman periyotlarından elde edilen eğitim veri setleri ile dinamik bir model önerisi getirmişlerdir. Kullandıkları yöntemler arasında Karar Ağaçları ve Lojistik Regresyon yer almaktadır. Araştırmacılar sektör olarak bankacılık tercih etmiş, bir bankanın 7204 müşterisine ait kayıtların ve 169 nitelik alanına sahip bir veri seti kullanılmıştır.

Verbeke ve diğerleri (2014), ayrılma eğilimi gösteren müşteriyi tahmin edebilmek için sosyal ağ etkisi ile beraber geniş ölçekli ağlar, zaman bağımlı sınıf etiketi ve çarpık dağınıklık konularını da içeren bir model önerisi geliştirmiştir. Telekom sektöründe gerçekleştirilen çalışmada müşteriler önceden ödeme yaparak hizmet alan ve aldığı hizmetin bedelini sonradan ödeyenler olmak üzere iki grupta incelenmiştir. Kullanılan veri setlerinde ayrılma eğilimli müşteri oranları %0,52 ve %0,57 olmak üzere oldukça düşüktür. Çalışmada, Karar Ağaçları, Lojistik Regresyon, Bayes Ağları, Rastgele Orman Algoritması gibi yöntemlerin yanı sıra sezgisel yöntemler de dahil olmak üzere birçok yöntemden yararlanılmıştır.

Lu ve diğerleri (2014), çalışmalarında hızlandırma (boosting) yönteminin kullanılması ile ilgili bir model önerisi yapmayı amaçlamıştır. Çoğu çalışmada hızlandırma yöntemlerinin kullanım amacının modelin tahmin gücünü arttırmak olmasına karşın bu çalışmada ağırlıklandırma temelli olmak üzere küme oluşturma aşamasında kullanılmıştır. Çalışma kapsamında geniş ölçekte hizmet veren bir telekom şirketinin toplam 7190 müşterisine ait veri kullanılmıştır. Çalışmada Lojistik Regresyon yöntemine de yer verilmiştir.

Vafeiadis ve diğerleri (2015), veri analizi için kullanılan makine öğrenmesi yöntemlerinin müşteri kayıp analizi problemi çözümünde performanlara dayalı karşılaştırmalı bir çalışma yapmayı amaçlamıştır. Özellikle hızlandırma yönteminin performansa etkisi incelenmiştir. Çalışmada telekom sektörüne ait açık erişimde 5000 kayıt ve 19 nitelik alanına sahip bir veri seti kullanılmıştır. Sinir Ağları, Destek Vektör Makineleri, Karar Ağaçları ve Lojistik Regresyon yöntemlerinden faydalanılmıştır.

Hudaib ve diğerleri (2015), her biri kümeleme ve tahmin etme olmak üzere iki aşamadan meydana gelen 3 farklı hibrit model geliştirmiştir. Çalışmada 5000 kayıt, 11 nitelik alanı ve %7,6 ayrılma eğilimli müşteri oranına sahip veri setinin yanı sıra K-Ortalamalar, Karar Ağaçlar ve ve Sinir Ağları yöntemleri kullanılmıştır.

Al-Shboul ve diğerleri (2015), çalışmalarında Hızlı Bulanık C-Ortalamalar (Fast Fuzzy C-Means) ve Genetik Programlama yöntemlerini temel alarak bir hibrit yöntem geliştirmiş, hem Genetik Programlamanın performansının arttırılmasını hem de veri seti içerisinde yer alan aykırı değerlerin ortaya çıkartılabilmesini amaçlamışlardır. Geliştirilen modelin farklı yöntemler ile karşılaştırıldığında bazı performans değerlendirme ölçülerine göre daha iyi

sonuçlar ürettiği görülmüştür. Çalışmada yine telekom sektöründen bir veri seti kullanılmış olup bu veri seti 14573 kayıt ve 11 nitelik alanına sahiptir.

Moeyersoms ve Martens (2015), yüksek kardinaliteli veriler ile çalışmayı, farklı niteliklerin performansa katkısını araştırmayı ve büyük veri ile çalışmayı amaçlamıştır. Bir milyondan fazla müşteriye ait verinin yer aldığı Enerji sektöründen bir veri setinin kullanıldığı çalışmada Karar Ağaçları, Lojistik Regresyon ve Destek Vektör Makineleri yöntemlerine yer verilmiştir.

Xiao ve diğerleri (2015), Transfer Öğrenme Teorisini müşteri kayıp analizi problemi için uygulamış ve nitelik seçimine dayalı bir model önerisi getirmek amaçlamıştır. Çalışmada farklı sektörlerden olmak üzere iki veri seti ile analizler gerçekleştirilmiştir. Topluluk sınıflandırıcı yöntemler ile beraber Sinir Ağları da çalışma kapsamında ele alınmıştır.

Ismail ve diğerleri (2015)'nin çalışmasında, Çok Katmanlı Algılayıcı Yapay Sinir Ağları ile bir telekom şirketine ait veriler üzerinde ayrılan müşteri analizini gerçekleştirmek amaçlanmıştır. Sinir Ağları yöntemine ait çeşitli algoritmalar ile beraber Regresyon yöntemlerine de yer verilmiştir.

Tamaddoni ve diğerleri (2015), iki çevrimiçi satıcıya ait deneysel ve benzetimli veriler kullanarak, paremetrik olan ve olmayan yöntemlerin performanslarını arttırmaya yönelik çalışmalar yapmış, ayrılan müşterilerin tahmin edilmesi için optimal model elde etmeye çalışmıştır. Çalışma kapsamında kullanılan yöntemler arasında Destek Vektör Makineleri ve Lojistik Regresyona yer verilmiştir.

Rodan ve diğerleri (2015), telekom sektöründe ayrılan müşteri analizi için negatif korelasyon öğrenmeyi kullanan Çok Katmanlı Algılayıcı Yapay Sinir Ağları tabanlı bir topluluk sınıflandırıcı yöntemi geliştirmeyi amaçlamıştır. Çalışma kapsamında Ürdün'de hizmet veren bir telekom şirketine ait 5000 kayıt ve 11 nitelik alanına sahip, ayrılma eğilimli müşteri oranı %0,076 olan bir veri seti kullanılmıştır. Sinir Ağları, Karar Ağaçları, K-En Yakın Komşu, Genetik Programlama, Sade Bayes, Destek Vektör Makineleri gibi birçok yöntemden faydalanılmıştır.

Gajowniczek ve diğerleri (2016), Karar Ağaçları yöntemini baz almış, bu yöntemde kullanılan farklı şekillerde hesaplanan entropi değerlerine göre modeller belirleyerek bu modellerin ayrılma eğiliminde olan müşterilerin tahmin performansına etkisini incelemeyi amaçlamıştır.

Herkese açık yayınlanan ve telekom sektöründen elde edilen 71047 kayıt ve 78 nitelik alanına sahip bir veri seti ile gerçekleştirilen uygulamalarda Karar Ağaçları yöntemi ile elde edilen modellerin performansları Lojistik Regresyon, Destek Vektör Makineleri ve Sinir Ağları yöntemleri ile oluşturulan modeller ile karşılaştırılmıştır.

Keremati ve diğerleri (2016), elektronik bankacılık hizmetlerinden yararlanan ve ayrılma eğilimi gösteren müşterilerin karakteristiklerini özellikle demografik veriler ile tanımlamaya çalışmıştır. Modelleme için Karar Ağaçları yönteminden yararlanılan çalışmada bir bankaya ait ve elektronik bankacılık uygulamalarını kullanan 4383 müşteriye ait demografik nitelikleri barındıran veri seti ile beraber, müşterinin ilişkili olduğu süre ve müşteri şikayetleri memnuniyetsizlik seviyesini değerlendirmek üzere kullanılmıştır.

Fathian ve diğerleri (2016), daha önce gerçekleştirilmiş diğer çalışmalara benzer olarak bir hibrit yöntem önerisi getirmiştir. Bu hibrit yöntem iki aşamadan meydana gelmek ile beraber ilk aşamada kümeleme yöntemi ile müşteri kümeleri (sınıfları), ikinci aşamada ise tahmin modeli oluşturulmuştur. Herkese açık olarak yayınlanmış telekom sektöründen bir veri seti kullanılan çalışmada Karar Ağaçları, Sinir Ağları, Destek Vektör Makineleri ve K-En Yakın Komşu Algoritması tahmin modelleri için yararlanılan yöntemler arasındadır. Kullanılan veri seti 40000 kayıt ve 76 nitelik içermektedir.

Dolatabadi ve Keynia (2017), müşteri kayıp analizi problemi ile beraber ayrılma eğilimli çalışan problemi için tahmin modeli geliştirmeyi amaçlamıştır. Bu amaç doğrultusunda aynı işletmeye ait müşteri ve çalışan verisi kullanılan çalışmada Karar Ağaçları, Sade Bayes, Destek Vektör Makineleri ve Sinir Ağları yöntemlerine yer verilmiştir. Toplam 1,5 yıl içerisinde toplanan veri seti çalışanlar için 21 nitelik alanı, 9239 kayıt ve %24,2 ayrılma eğilimli çalışan oranına, müşteriler için ise 15 nitelik alanı, 9239 kayıt ve %15,3 ayrılma eğilimli müşteri oranına sahiptir.

Coussement ve diğerleri (2017), çalışmalarında veri ön işlemenin sınıflandırma modellerinin tahmin başarısındaki etkisini tespit etmeyi ve var olan yöntemlerin farklı veri önişleme yöntemleri ile performansını arttırmayı amaçlamıştır. Bu doğrultuda gerçekleştirilen çalışmada Bayes Ağları, Sade Bayes, Karar Ağaçları, Lojistik Regresyon, Sinir Ağları, Rastgele Orman Algoritması, Destek Vektör Makineleri yöntemlerine yer verilmiştir. Özellikle Lojistik Regresyon yönteminin performansında artış sağlanmıştır. Telekom sektöründen elde edilen veri

seti 30104 müşteriye ait kayıt ve 956 değişken içerirken, ayrılma eğilimli müşteri oranı %4,52'dir.

Lee ve diğerleri (2017), elektronik ortamlarda yer alan haberlerin müşterilerin duygu durumlarını dolayısıyla bir işletmeyi terketme durumunu da etkileyeceği teziyle bir çalışma gerçekleştirmiştir. Çalışma kapsamında Güney Kore'de hizmet veren bir mobil servis sağlayıcı ile ilgili belirli bir süre zarfında elektronik ortamdaki yayınları incelemiş, yine bu süre zarfında ayrılan müşterilerin verisi ile birlikte Lojistik Regresyon, Karar Ağaçları, Sinir Ağları yöntemleri kullanılarak ayrılma eğilimli müşterilerin tahmin edilebilmesi için modeller geliştirilmiştir.

Literatür taraması sonucu incelenen makaleler ile ilgili ayrıntılara Tablo 2.6'da yer verilmiştir.

Tablo 2.6: Literatür incelemesi.

Çalışma	**Sektör**	**Veri Seti**	**Yöntemler**	**Performans Değ. Ö.**
Masand, Datta, Mani, Li (1999)	Telekom	➢ GTE, mobil servis sağlayıcısı ○ 500000 kayıt ○ 200 nitelik	Karar Ağaçları K En Yakın Komşu Genetik Algoritmalar Sinir Ağları	Kaldıraç
Mozer, Wolniewicz, Grimes, Johnson, Kaushansky (2000)	Telekom	➢ Telekom şirketi ○ 46744 müşteri	Lojistik regresyon Karar Ağaçları Sinir Ağları Hızlandırma Algoritmaları	Kaldıraç Eğrisi
Datta, Masand, Mani, Li (2000)	Telekom	➢ GTE, Mobil Servis Sağlayıcı ○ 500000 kayıt ○ 200 nitelik	Sinir Ağları Genetik Algoritmalar Karar Ağaçları K-En Yakın Komşu	Kaldıraç Payoff Metrikleri
Wei, Chiu (2002)	Telekom	➢ Tayvan'da hizmet veren mobil operatör ○ 9100000 kayıt	Çok Sınıflandırıcılı Sınıf Birleştirici Tek Sınıflandırıcılı Sınıf Birleştirici	Başarısızlık Yüzdesi Kaldıraç
Chiang, Wang, Lee, Lin (2003)	Banka	➢ Tayvan'da hizmet veren banka ○ 1 milyondan fazla veri	Hedef Odaklı Sıralı Desen Algoritması Küresel Odaklı Algoritma Apriori Algoritması	İşlem Süresi Oluşturulan Kural Sayısı

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Au, Chan, Yao (2003)	Telekom	➢ Zoo o 101 kayıt ➢ DNA o 3190 kayıt ➢ Credit Card o 690 kayıt ➢ Diabetes o 768 kayıt ➢ Satellite Image o 6435 kayıt ➢ Social o 48843 kayıt ➢ PBX o 3009 kayıt ➢ Malezya'da hizmet veren telefon operatörü o 100000kayıt	Sinir Ağları Karar Ağaçları Evrimsel Öğrenme ile Veri Madenciliği Algoritması (DMEL)	Doğruluk Kaldıraç Eğrisi İşlem Süresi
Buckinx, Van den Poel (2005)	Perakende	➢ Perakende sektöründe faaliyet gösteren önemli şirketlerden birisi o 158884 kayıt o 61 nitelik	Lojistik Regresyon Sinir Ağları Rastgele Orman Algoritması	Doğru Sınıflandırma Yüzdesi ROC Eğrisi Altındaki Alan
Hung, Yen, Wang (2006)	Telekom	➢ Tayvan'da hizmet veren Telekom şirketi o 160000 kayıt	K-Ortalamalar Sinir Ağları Karar Ağaçları T-Test	Kaldıraç Başarı Oranı
Song, Yang, Wu, Wang, Tang (2006)	Telekom	➢ China Mobile Communication Company o 12000 kayıt	Sinir Ağları	Doğruluk İşlem süresi
Ahn, Han, Lee (2006)	Telekom	➢ Güney Kore'de hizmet veren bir telekom Şirketi o 5789 kayıt o 14 nitelik	Lojistik Regresyon	
Hadden, Tiwari, Roy, Ruta (2006)		➢ Belirtilmemiş o 902 müşteri o Eğitim veri setinde churn oranı %50 o Test veri setinde churn oranı %30	Karar Ağaçları Sinir Ağları Lojistik Regresyon	Ayrılma eğiliminde olan/olmayan müşteri oranları Doğruluk
Bin, Peiji, Juan (2007)	Telekom	➢ 4799 müşteri	Karar Ağaçları	Duyarlılık Oranı F-Ölçüsü
Burez, Van den Poel (2007)	TV	➢ Ödemeli-TV Şirketi o %15 churn oranı	Lojistik Regresyon Rastgele Orman Algoritması Markov Zincirleri	Kaldıraç ROC Eğrisi Altındaki Alan Doğru Sınıflandırma Yüzdesi

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Zhang, Qi, Shu, Cao (2007)	Karma	➢ Adult - UCI Machine Learning Repository o 48842 kayıt o 14 nitelik ➢ Credit- UCI Machine Learning Repository o 690 kayıt o 15 nitelik ➢ Telekom (Weiss and Indurkhya, 1995) o 15000 kayıt o 48 nitelik ➢ Wisconsin breast cancer (BCW)- UCI Machine Learning Repository o 699 kayıt o 9 nitelik ➢ Çin'de hizmet veren telekom şirketinin mobil hizmet alan müşteri verileri o 52495 kayıt ➢ 93 nitelik	🔹 K En Yakın Komşu 🔹 Lojistik Regresyon 🔹 Karar Ağaçları 🔹 Sinir Ağları	Doğruluk ROC Eğrisi Altındaki Alan
Coussement, Van den Poel (2008)	Medya	➢ Belçika'da hizmet veren gazete basım şirketi o 3161kayıt	🔹 Lojistik Regresyon 🔹 Salton'un Vektör Uzayları (SMART) 🔹 Örtülü Semantik İndeksleme Algoritması	ROC Eğrisi Altındaki Alan Kaldıraç Alıcı İşletim Karakteristik Eğrisi (ROC)
Burez, Van den Poel (2008)	Medya	➢ Pay-TV şirket veriambarı ➢ Dinamik modelde, o 500000 müşteri ➢ Statik modelde o 100000 kayıt o 171 nitelik	🔹 Rastgele Orman Algoritması 🔹 Sağkalım Analizi	ROC Eğrisi Altındaki Alan
Anıl Kumar, Ravi (2008)	Banka	➢ Latin Amerika kökenli banka - Business Intelligence Cup, Chile University (2004) o 14814 kayıt o 22 nitelik o %6,76 churn oranı	🔹 Sinir Ağları 🔹 Lojistik Regresyon 🔹 Karar Ağaçları 🔹 Rastgele Orman Algoritması 🔹 Destek Vektör Makinaları 🔹 Örnekleme Yöntemleri 🔹 Sentetik Azınlık Örnekleme Algoritması (SMOTE)	Duyarlılık Belirleyicilik Doğruluk Alıcı İşletim Karakteristik Eğrisi (ROC) ROC Eğrisi Altındaki Alan

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Zhao, Dang (2008)	Banka	➢ Banka-CCB o 2382 müşteri o 933 müşteri churn	Destek Vektör Makinaları Lojistik Regresyon Yapay Sinir Ağları Sade-Bayes Sınıflandırıcı Karar Ağaçları	Doğruluk Başarı Oranı Kaldıraç Katsayısı
Coussement, Van den Poel (2008)	Medya	➢ Belçika'da hizmet veren yayın kuruluşu o 90000 kayıt	Destek Vektör Makinaları Lojistik Regresyon Rastgele Orman Algoritması	ROC Eğrisi Altındaki Alan Doğru Sınıflandırma Yüzdesi Kaldıraç
Xia, Jin (2008)	Telekom	➢ UCI Machine Learning Repository o 3333 kayıt ➢ Mobil operatör şirketi o 2340 kayıt	Destek Vektör Makinaları Karar Ağaçları Yapay Sinir Ağları Lojistik Regresyon Sade Bayes Örnekleme Yöntemleri	Doğruluk Yakalama/Başarı Oranı Kaplama Oranı Kaldıraç Katsayısı
Tsai, Lu (2009)	Telekom	➢ Amerika'da hizmet veren Telekom şirketi (Duke University) o 51306 kayıt	Yapay Sinir Ağları T-test	Doğruluk Hata Tipleri (Tip I ve II) Karışıklık Matrisi
Coussement, Van den Poel (2009)	Medya	➢ Belçika'da hizmet veren gazete şirketi o 11836 müşteri o %18,89 churn oranı o 18331 mail o %35,79 şikayet içerikli	Rastgele Orman Algoritması Destek Vektör Makinaları Lojistik Regresyon	Doğru Sınıflandırma Yüzdesi ROC Eğrisi Altındaki Alan Alıcı İşletim Karakteristik Eğrisi (ROC)
Wu, Qi, Wang (2009)	Telekom	➢ Telekom pazarlama verisi o 300 kayıt o 206 nitelik	Temel Bileşenler Analizi Genetik Algoritmalar	Yanlış Sınıflama ASE
Popović, Bašić (2009)	Banka	➢ Bireysel Bankacılık o 5000 kayıt o 73 nitelik	Bulanık C-Ortalamalar Kanonik Diskriminant Analizi K-Ortalamalar Hiyerarşik Kümeleme McQuitty Benzerlik Algoritması Crisp K-Ortalamalar	Doğru Pozitif Sayısı Yanlış Pozitif Sayısı Doğruluk Belirleyicilik
Pendharkar (2009)	Telekom	➢ Telekom şirketi (Teradata Center for Customer Relationship Management at Duke University) o 195956 kayıt o 6 nitelik	Genetik Algoritmalar Sinir Ağları	Doğruluk %10 Kaldıraç Alıcı İşletim Karakteristik Eğrisi (ROC) İşlem Süresi
Xie, Li, Ngai, Ying (2009)	Banka	➢ Çin'de hizmet veren banka o 1524 kayıt o 15 nitelik	Rastgele Orman Algoritması Yapay Sinir Ağları Karar Ağaçları Destek Vektör Makineleri	Kaldıraç Eğrisi

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Burez, Van den Poel (2009)	Karma	➢ Daha önce farklı çalışmalarda kullanılmış 6 farklı veri seti ➢ Banka1 o 117808 kayıt o %6,3 churn oranı o 74 nitelik ➢ Banka2 o 102279 kayıt o %5,99 churn oranı o 38 nitelik ➢ Telekom o 100205 kayıt o %2,98 churn oranı o 73 nitelik ➢ Gazete o 122189 kayıt o %6,3 churn oranı o 33 nitelik ➢ Ödemeli-TV o 143198 kayıt o %13,07 churn oranı o 81 nitelik ➢ Süpermarket o 32371 kayıt o %25,15 churn oranı ➢ 21 nitelik	Rastgele Orman Algoritması Lojistik Regresyon Hızlandırma Algoritmaları Örnekleme Yöntemleri	ROC Eğrisi Altındaki Alan Kaldıraç
Bose, Chen (2009)	Telekom	➢ Teradata Center, Duke University ➢ Veri Seti 1 o 100000 kayıt o 171 nitelik o %50 churn oranı ➢ Veri Seti2 o 50000 kayıt o 171 nitelik o %1,8 churn oranı ➢ Veri Seti 3 o 100000 kayıt o %1,8 churn oranı o 171 nitelik	Karar Ağaçları Hızlandırma Algoritmaları K-Ortalamalar K-Medoid Yapay Sinir Ağları Bulanık C-Ortalamalar	Kaldıraç Dunn indeksi
Wang, Chiang, Hsu, Lin, Lin (2009)	Telekom	➢ Wireless şirketi o 60000 kayıt	Karar Ağaçları	
Coussement, Benoit, Van den Poel (2010)	Medya	➢ Belçika'da hizmet veren gazete şirketi ➢ 134084 kayıt ➢ 18/27 nitelik (iki farklı yöntem için) ➢ %11,95 churn oranı	Genelleştirilmiş Toplamsal Modeller (GAM) Lojistik Regresyon	ROC Eğrisi Altındaki Alan Kaldıraç İndirgenmiş Kazanç/Fayda
Risselada, Verhoef, Bijmolt (2010)	Karma	➢ İnternet Servis Sağlayıcısı o 4 alt örneklem (7063-6967-7146-7001 kayıt) o %50 churn oranı ➢ Sağlık Sigortası Şirketi o 3 alt örneklem (1789-1294-1474 kayıt)	Karar Ağaçları Lojistik Regresyon	Kaldıraç Gini katsayısı

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Radosavljevik, van der Putten, Kyllesbech Larsen (2010)	Telekom		Karar Ağaçları Lojistik Regresyon	Concordance katsayısı ROC Eğrisi Altındaki Alan Kazanç Tablosu
Huang, Kechadi, Buckley, Kiernan, Keogh, Rashid (2010)	Telekom	➢ İrlanda telekom şirketi(Eircom) o 47391 kayıt	Karar Ağaçları Sinir Ağları Destek Vektör Makineleri	Doğruluk
Khan, Jamwal, Sepehri (2010)	Telekom	➢ Tahran'da faaliyet gösteren Sepanta Co. telekom şirketi o 2164261 kayıt o 36 nitelik	Microsoft Karar Ağaçları Microsoft Sinir Ağları Microsoft Lojistik Regresyon K-Ortalamalar	Doğruluk Duyarlılık Belirleyicilik Kesinlik
Huang, Buckley, Kechadi (2010)	Telekom	o 18600 kayıt o 5600 churn kayıt	Genetik Algoritma Karar Ağaçları	Doğruluk
Tsai, Chen (2010)	Telekom	➢ Tayvan'da hizmet veren Telekom şirketi o 37882 kayıt o 22 nitelik	Birliktelik Kuralları Karar Ağaçları Sinir Ağları	Kesinlik Duyarlılık Doğruluk F-Ölçüsü
Owczarczuk (2010)	Telekom	➢ Polonya'da hizmet veren Telekom şirketi o 85274 kayıt o 1381 nitelik	Lojistik Regresyon Karar Ağaçları Lineer Regresyon Fisher Diskriminant Analizi T-test	Kaldıraç Eğrisi
Huang, Huang, Kechadi (2010)	Telekom	➢ Eircom o 147214 kayıt o 122 nitelik	Karar Ağaçları Destek Vektör Makineleri Sade Bayes Lojistik Regresyon Evrimsel Öğrenme ile Veri Madenciliği Algoritması (DMEL)	ROC Eğrisi Altındaki Alan
Kraljevic, Gotovac (2010)	Telekom	➢ Bosna Hersek'te hizmet veren telekom şirketi o 3000kayıt	Sinir Ağları Karar Ağaçları Lojistik Regresyon	Karışıklık Matrisi Doğruluk Hata Oranı
Tamaddoni Jahromi, Sepehri, Teimourpour, Choobdar (2010)	Telekom	➢ İran'da hizmet veren telekom şirketi o 34504 müşteri	Karar ağaçları Maliyet Duyarlı Öğrenme Algoritmaları	Kazanç Ölçüsü Kazanç Tablosu
Delen (2010)	Eğitim	➢ ABD'de egitim veren bir enstitü o 16066 kayıt	Karar Ağaçları Yapay Sinir Ağları Destek Vektör Makineleri Lojistik Regresyon Rastgele Orman Algoritması Hızlandırma Algoritmaları	Karışıklık Matrisi Doğruluk
Lima, Mues, Baesens (2011)	Telekom	➢ Centre for Customer Relationship Management, Duke University o 20000 kayıt	Lojistik Regresyon Karar Ağaçları Geriye Dönük Test Modeli	Duyarlılık Belirleyicilik ROC Eğrisi Altındaki Alan

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Fasanghari, Keramati (2011)	Telekom	➢ İran'da hizmet veren telekom şirketi ○ 3150 kayıt ○ 11 nitelik	Yerel Lineer Model Ağacı Algoritması (LOLIMOT) Yapay Sinir Ağları Lokal Lineer Nöro-Bulanık Model (LLNF)	Ortalama hata kareleri toplamı kökü (RMSE) Ortalama mutlak hata yüzdesi (MAPE)
De Bock, Van Den Poel (2011)	Karma	➢ Banka ○ 20456 kayıt ○ 137 nitelik ○ %5,99 churn oranı ➢ Avrupa Telekomünikasyon Operatörü ○ 35550 kayıt ○ 529 nitelik ○ %2,76 churn oranı ➢ Do-It-Yourself (DIY) Donanım Mağazalar Zinciri ○ 3827 kayıt ○ 15 nitelik ○ %28,14 churn oranı ➢ Mail-order garments ○ 43305kayıt ○ 244 nitelik ➢ %1,76 churn oranı	Rastgele Orman Algoritması Karar Ağaçları Rotasyon Orman Algoritması Temel Parçacık Bileşen Analizi Bağımsız Bileşen Analizi	Doğruluk ROC Eğrisi Altındaki Alan Kaldıraç
Abbasimehr, Setak, Tarokh (2011)		➢ UCI Machine Learning Repository ○ 5000 kayıt ○ 20 nitelik ○ %14,3 churn oranı	Uyarlamalı Sinirsel Bulanık Çıkarım Sistemi (ANFIS) Karar Ağaçları Tekrarlı Budama Algoritması (RIPPER) Örnekleme Yöntemleri	Doğruluk Belirleyicilik Duyarlılık
Yu, Guo, Guo, Huang (2011)	E-Ticaret	➢ Çin'de hizmet veren e-ticaret sitesi ○ 150000 kayıt ○ %10 churn oranı	Yapay Sinir Ağları (ANN) Karar Ağaçları (DT) Destek Vektör Makinaları (SVM) Genişletilmiş Destek Vektör Makinaları (Extended-SVM)	Doğruluk Başarı Oranı Kaldıraç
Kisioglu, Topcu (2011)	Telekom	➢ Türkiye'de hizmet veren telekom şirketi ○ 2000 kayıt ○ 9 nitelik	Sinir Ağları Ki-kare Otomatik Etkileşim Belirleme Algoritması (CHAID) Koralasyon Analizi	
Keremati, Ardabili (2011)	Telekom	➢ Mobil Operatör ○ 3150 kayıt	Lojistik Regresyon Hipotez Testi	
Nie, Rowe, Zhang, Tian, Shi (2011)	Finans	➢ Çin'de hizmet veren bir bankanın veri ambarı (2005) ○ 8 milyon kayıt ○ 135 nitelik ○ %8,1 churn oranı	Lojistik Regresyon Karar Ağaçları	Doğru Sınıflandırma Yüzdesi Alıcı İşletim Karakteristik Eğrisi (ROC) Kaldıraç Gini Katsayısı

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Dierkes, Bichler, Krishnan (2011)	Telekom	○ 120000 kayıt ○ 32 nitelik	Markov Mantık Ağı Lojistik Regresyon	Doğruluk Duyarlılık Belirleyicilik Tanımlama Kesinlik Alıcı İşletim Karakteristik Eğrisi (ROC)
Yeshwanth, Vimal Raj, Saravanan (2011)	Telekom	➢ Gelişmekte olan bir ülkede hizmet veren Telekom şirketi ○ 1,2 milyon kayıt	Genetik Algoritmalar Karar Ağaçları Monte Carlo Yaklaşımı Shapley Değerleri	Doğruluk
Karahoca, Karahoca (2011)	Telekom	➢ Türkiye'de hizmet veren Telekom şirketi ○ 24900 kayıt	X-Ortalamalar Bulanık C-Ortalamalar Uyarlamalı Sinirsel Bulanık Çıkarım Sistemi (ANFIS) Karar Ağaçları	Duyarlılık Belirleyicilik Kesinlik Doğruluk Alıcı İşletim Karakteristik Eğrisi (ROC)
Keramati, Ardabili (2011)	Telekom	➢ Operatör Çağrı Merkezi ○ 3150 kayıt	Lojistik Regresyon Hipotez Testi Ki-Kare	
Saradhi, Palshikar (2011)	Organizsyon	➢ Organizasyon Şirketi ○ 1575 çalışan ○ 25 nitelik	Sade Bayes Rastgele Orman Algoritması Destek Vektör Makinaları	Doğruluk Doğru Pozitif Sayısı Doğru Negatif Sayısı
Verbeke, Martens, Mues, Baesens (2011)	Telekom	➢ Knowledge Discovery and Data Mining (KDD)Library ○ 5000 kayıt ○ 21 nitelik ○ %14,3 churn oranı	Karınca Kolonisi Optimizasyonu Destek Vektör Makinaları Tekrarlı Budama Algoritması (RIPPER) Lojistik Regresyon Karar Ağaçları Ant Miner+ Örnekleme Yöntemleri	Doğru Sınıflandırma Yüzdesi Belirleyicilik Duyarlılık Kural sayısı
Idris, Rizwan, Khan (2012)	Telekom	➢ Orange Telekom Şirketi ○ 50000 kayıt ○ 260 nitelik	Parçacık Sürü Optimizasyonu Rastgele Orman Algoritması K-En Yakın Komşu Temel Bileşenler Analizi Fisher oranı F-skoru	Doğruluk Duyarlılık Belirleyicilik ROC Eğrisi Altındaki Alan
Migueis, Van den Poel, Camanho, Cunha (2012)	Perakende	➢ Avrupa'da hizmet veren marketler zinciri sahibi şirket ○ 74607 müşteri ○ %44 parçalı churn oranı	Lojistik Regresyon	ROC Eğrisi Altındaki Alan Kaldıraç Alıcı İşl. Karakteristik Eğrisi (ROC)
Prasad, Madhavi (2012)	Banka	➢ Hindistan'da hizmet veren banka ○ 1484 kayıt	Karar Ağaçları	Karışıklık Matrisi

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Verbeke, Dejaeger, Martens, Hur, Baesens (2012)	Telekom	➢ Operatör ○ 47761 kayıt ○ 53 nitelik ○ %3,69 churn oranı ➢ Operatör ○ 11317 kayıt ○ 21 nitelik ○ %1,56 churn oranı ➢ Operatör ○ 2904 kayıt ○ 15 nitelik ○ %3,2 churn oranı ➢ Operatör ○ 2180 kayıt ○ 15 nitelik ○ %3,21 churn oranı ➢ Operatör ○ 338874 kayıt ○ 727 nitelik ○ %1,8 churn oranı ➢ Duke ○ 93893 kayıt ○ 197 nitelik ○ %1,78 churn oranı ➢ Duke ○ 38924 kayıt ○ 77 nitelik ○ %1,99 churn oranı ➢ Duke ○ 7788 kayıt ○ 19 nitelik ○ %3,3 churn oranı ➢ UCI Machine L.R. ○ 5000 kayıt ○ 23 nitelik ○ %14,14 churn oranı ➢ KDD Cup 2009 ○ 46933 kayıt ○ 242 nitelik ○ %6,98 churn oranı	Karar Ağaçları Sinir Ağları En Yakın K-Komşu Topluluk Sınıflandırıcılar İstatistiksel metodlar Destek Vektör Makineleri Kural Çıkarım Yöntemleri (RIPPER, PART) Örnekleme Yöntemleri	Kaldıraç Doğru Sınıflandırma Yüzdesi Alıcı İşletim Karakteristik Eğrisi (ROC) ROC Eğrisi Altındaki Alan Gini Katsayısı Maksimum Fayda Kriteri Bonferroni–Dunn testi
Xiao, Xie, He, Jiang (2012)	Karma	➢ German veri seti-UCI Machine Learning Repository ○ 1000 kayıt ○ 20 nitelik ○ %2,33 churn oranı ➢ Sichuan Telekom ○ 3350 kayıt ➢ %12,66 churn oranı	Maliyet Duyarlı Öğrenme Yöntemleri Sentetik Azınlık Örnekleme Algoritması (SMOTE) Rastgele Orman Algoritması Genetik Algoritma Tabanlı Topluluk Sınıflandırıcı (GAE) Sinir Ağları Tabanlı Topluluk Sınıflandırıcı (NNE)	Doğruluk ROC Eğrisi Altındaki Alan
Shim, Choi, Suh (2012)	Online Alışveriş	➢ Kore'de hizmet veren online alışveriş mağazası ○ 9 nitelik ○ 3445 müşteri (14782 işlem)	Karar Ağaçları Yapay Sinir Ağları Lojistik Regresyon Birliktelik Kuralları	Doğruluk Kaldıraç

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
De Bock, Van Den Poel (2012)	Karma	➢ Banka o 20456 kayıt o 137 nitelik o %5,99 churn oranı ➢ Avrupa Telekomünikasyon Operatörü o 35550 kayıt o 529 nitelik o %2,76 churn oranı ➢ Do-It-Yourself (DIY) Donanım Mağazalar Zinciri o 3827 kayıt o 15 nitelik o %28,14 churn oranı ➢ Mail-order garments o 43305kayıt o 244 nitelik o %1,76 churn oranı ➢ Süpermarket Zinciri I o 32371 kayıt o 46 nitelik o %25,15 churn oranı ➢ Süpermarket Zinciri II o 8453 kayıt o 36 nitelik o %47,38 churn oranı	Rastgele Alt uzay Yöntemi (RSM) Rastgele Orman Algoritması Genelleştirilmiş Toplamsal Modeller (GAM) Lojistik Regresyon	Doğruluk ROC Eğrisi Altındaki Alan Kaldıraç Kaldıraç İndeksi
Huang, Kechadi, Buckley (2012)	Telekom	➢ İrlanda'da hizmet veren Telekom şirketi o 827124 kayıt o 738 nitelik	Lojistik Regresyon Lineer Sınıflandırıcı Sade Bayes Karar Ağaçları Sinir Ağları Destek Vektör Makineleri Evrimsel Öğrenme İle Veri Madenciliği Algoritması (DMEL)	ROC Eğrisi Altındaki Alan
Ballings, Van den Poel (2012)	Medya	➢ Gazete şirketi o 129892 kayıt o 31-54 nitelik o Yaklaşık %11 churn oranı	Lojistik Regresyon Ki-Kare Otomatik Etkileşim Belirleme Algoritması (CHAID) Sınıflandırma ve Regresyon Ağaçları (CART)	ROC Eğrisi Altındaki Alan
Chen, Fan (2012)	Perakende	➢ Foodmart 2000 (MS-SQL Server 2000 Database) o 1560 ürün o 10281 müşteri o 200000 kayıt	Destek Vektör Makineleri	Doğru Sınıflandırma Yüzdesi Duyarlılık Belirleyicilik ROC Eğrisi Altındaki Alan İşlem Süresi Maksimum Yarar

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Kim, Lee (2012)	E-Ticaret	➢ E-ticaret o 5000 kayıt o 330 nitelik o %10 churn oranı	Sinir Ağları Lojistik Regresyon K-Ortalamalar K-Medoid	ROC Eğrisi Altındaki Alan Başarı Oranı
Zhang, Zhu, Xu, Wan (2012)	Telekom	➢ Mobil Servis Sağlayıcı	Lojistik regresyon Karar ağaçları Sinir Ağları	Kaldıraç Eğrisi ROC Eğrisi Altındaki Alan Başarı Oranı
Chen, Fan, Sun (2012)	Karma	➢ FoodMart 2000 (MS SQL 2000) o 8842 kayıt ➢ AdventureWorksDW (MS SQL 2005) o 633 kayıt ➢ Telecom (Center for CRM at Duke University) o 3399 kayıt	Destek Vektör Makineleri Yapay Sinir Ağları Karar Ağaçları Rastgele Orman Algoritması Lojistik Regresyon Hızlandırma Algoritmaları	Doğru Sınıflandırma Yüzdesi Duyarlılık Belirleyicilik ROC Eğrisi Altındaki Alan Kaldıraç Maksimum Yarar H Ölçüsü İşlem Süresi
Huang, Kechadi (2013)	Telekom	➢ Telekom veri seti o 104199kayıt o 121 nitelik	K-Ortalamalar Karar Ağaçları Lojistik Regresyon K En Yakın Komşu Destek Vektör Makineleri OneR Kısmi Karar Ağacı (PART)	Doğruluk Doğru Pozitif Sayısı Yanlış Pozitif Sayısı Alıcı İşletim Karakteristik Eğrisi (ROC) ROC Eğrisi Altındaki Alan
Coussement, De Bock (2013)	Online Oyun	➢ bwin Interactive Entertainment (Online Oyun Operatörü) o 3729 kayıt o 60 nitelik alanı	Karar Ağaçları Rastgele Orman Algoritması Genelleştirilmiş Toplamsal Modeller (GAM)	Kaldıraç Kaldıraç İndeksi
Sharma, Panigrahi (2013)	Telekom	➢ UCI Machine Learning Repository o 2427 kayıt o 20 nitelik	Sinir Ağları	Karışıklık matrisi
Abbasimehr, Alizadeh (2013)	Telekom	➢ Amerika'da hizmet veren Telekom verisi (Duke University, Churn Tournament 2003) o 94 nitelik o 15000 kayıt ➢ Telekom şirketi (Duke University- The Center for Customer Relationship Management) o 75 nitelik ➢ 15000kayıt	Genetik Algoritma Sarmal Tabanlı Nitelik Seçimi Karar Ağaçları Korelasyon Tabanlı Nitelik Seçimi Yöntemi (CFS) Ki-Kare Tabanlı Nitelik Seçimi Yöntemi	Doğruluk Duyarlılık Belirleyicilik Geometrik Ortalama Kural Sayısı
Kirui, Hong, Cheruiyot, Kirui (2013)	Telekom	➢ Wireless hizmeti veren telekom şirketi o 106405 kayıt o 112 nitelik o %5,6 churn oranı	Sade Bayes Bayes ağları Karar ağaçları	Duyarlılık Doğru Pozitif Sayısı Yanlış Pozitif Sayısı Alıcı İşletim Karakteristik Eğrisi (ROC)

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Verbraken, Verbeke, Baesens (2013)	Telekom	➢ Operatör ○ 47761 kayıt ○ 53 nitelik ○ %3,69 churn oranı ➢ Operatör ○ 11317 kayıt ○ 21 nitelik ○ %1,56 churn oranı ➢ Operatör ○ 2904 kayıt ○ 15 nitelik ○ %3,2 churn oranı ➢ Operatör ○ 2180 kayıt ○ 15 nitelik ○ %3,21 churn oranı ➢ Operatör ○ 338874 kayıt ○ 727 nitelik ○ %1,8 churn oranı ➢ Duke ○ 93893 kayıt ○ 197 nitelik ○ %1,78 churn oranı ➢ Duke ○ 38924 kayıt ○ 77 nitelik ○ %1,99 churn oranı ➢ Duke ○ 7788 kayıt ○ 19 nitelik ○ %3,3 churn oranı ➢ UCI Machine L.R. ○ 5000 kayıt ○ 23 nitelik ○ %14,14 churn oranı ➢ KDD Cup 2009 ○ 46933 kayıt ○ 242 nitelik ○ %6,98 churn oranı	Karar Ağaçları Topluluk sınıflandırıcı yöntemler Destek Vektör Makineleri Kural Çıkarımlı Yöntemler İstatistik Tabanlı Yöntemler Sinir Ağları K-En Yakın Komşu	ROC Eğrisi Altındaki Alan H Ölçüsü Maksimum Yarar Beklenen Maksimum Yarar
Phadke, Uzunalioglu, Mendiratta, Kushnir, Doran (2013)	Telekom	➢ 500000 müşteri	Sosyal Ağ Analizi Hızlandırma Algoritmaları	Kaldıraç Eğrisi
Idris, Khan, Lee (2013)	Telekom	➢ Orange telekom şirketi ○ 50000 kayıt ○ 260 nitelik ○ %7,3 churn oranı ➢ Cell2Cell- Duke University ○ 40000 kayıt ○ 76 nitelik ○ %50 churn oranı	Rastgele Orman Algoritması Rotasyon Orman Algoritması Hızlandırma Algoritmaları Fisher Oranı Temelli Nitelik Seçimi F-Skoru Tabanlı Nitelik Seçimi Minimum Fazlalık Maksimum Uygunluk Yöntemi Temelli Nitelik Seçimi	ROC Eğrisi Altındaki Alan Duyarlılık Belirleyicilik

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Mohammadi, Tavakkoli-Moghaddam, Mohammadi (2013)	Telekom	➢ İran'da hizmet veren telekom şirketi o 3150 müşteri o 495 churn	Sinir Ağları Bulanık C-Ortalamalar Orantılı Tehlike Modeli (Cox)	Doğruluk Tip 1 Ve 2 Hatalar Yaşam Olasılığı Ortalama hata kareleri toplamı kökü (RMSE) Ortalama Mutlak Sapma (MAD)
Abbasimehr, Setak, Soroor (2013)	Telekom	➢ Amerika'da hizmet veren telekom şirketi(calibration dataset)- Teradata Center, Duke University o 100000 kayıt o 172 nitelik o %50 churn oranı	K-Ortalamalar Uyarlamalı Sinirsel Bulanık Denetim Sistemi (ANFIS) Lokal Lineer Nöro-Bulanık Model (LLNF) Yerel Lineer Model Ağacı Algoritması (LoLiMoT) Sinir Ağları	Doğruluk Belirleyicilik Duyarlılık
Cecilie Günther, Tvete, Aas, Sandnes, Borgan (2014)	Sigorta	➢ Norveç'te hizmet veren sigorta şirketi o 127961 müşteri	Genelleştirilmiş Toplamsal Modeller (GAM) Lojistik Regresyon	Doğru Pozitif Sayısı Yanlış Pozitif Sayısı Alıcı İşletim Karakteristik Eğrisi (ROC)
Kim, Jun, Lee (2014)	Telekom	➢ Telekom şirketi o 89412 müşteri o 36 nitelik o %9,7 churn oranı	Yayılım Aktivasyonu Yöntemi (SPA) Louvain Yöntemi Lojistik Regresyon Sinir Ağları	ROC Eğrisi Altındaki Alan Kaldıraç Başarı Oranı
Farquad, Ravi, Raju (2014)	Finans	➢ Business Intelligence Cup 2004- Latin Amerikan Bankası o 14814 kayıt, o 22 değişken, o %6,76 churn oranı	Destek Vektör Makineleri Sade Bayes Ağaçları Örnekleme Yöntemleri Sentetik Azınlık Örnekleme Algoritması (SMOTE)	Duyarlılık Belirleyicilik Doğruluk
Almana, Aksoy, Alzahrani (2014)	Telekom		Sinir Ağları İstatistiksel Yöntemler Lineer Regresyon Lojistik Regresyon Sade-Bayes K-En Yakın Komşu Karar Ağaçları Tekrarlı Budama Algoritması (RIPPER)	
Adwan, Faris, Jaradat, Harfoushi, Ghatasheh (2014)	Telekom	➢ Umniah-Jordanian Telecommunication Company o 5000 müşteri o 11 nitelik alanı o %7,6 churn oranı	Yapay Sinir Ağları Etki Faktörü Analizi	Karışıklık Matrisi Doğruluk Başarı Oranı Ayrılma Eğilimli Müşteri Oranı
Keramati, Jafari-Marandi, Aliannejadi, Ahmadian, Mozaffari, Abbasi (2014)	Telekom	➢ İran'da hizmet veren Telekom şirketi o 3150 kayıt	Karar Ağaçları Yapay Sinir Ağları K-En Yakın Komşu Destek Vektör Makineleri	Duyarlılık Kesinlik Yanlış Sınıflandırma

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Gür Ali, Arıtürk (2014)	Banka	➢ Banka (Özel Bankacılık Birimi) o 7204 müşteri o 169 nitelik	Lojistik Regresyon Karar Ağaçları Sağkalım Analizi Sentetik Azınlık Örnekleme Algoritması (SMOTE)	ROC Eğrisi Altındaki Alan Kaldıraç
Verbeke, Martens, Baesens (2014)	Telekom	➢ Önceden ödemeli müşteri verileri o 1673724 müşteri o 119 lokal, 58 ağ değişkeni o %0,52 churn oranı ➢ Sonradan ödemeli müşteri verileri o 1226286 müşteri o 119 lokal, 58 ağ değişkeni o %0,57 churn oranı	Karar Ağaçları Rastgele Orman Algoritması Bayes Ağları Lojistik Regresyon Örnekleme Yöntemleri İteratif Sınıflandırma Etiketleme Yöntemleri Benzetilmiş Tavlama Yayılma Aktivasyonu Toplu Çıkarımı Sınıf Dağıtımlı İlişkisel Komşu Sınıflandırıcı (CDRN) Sadece Ağ Bağlantı Tabanlı Sınıflandırıcı (NLB) Yayılma Aktivasyonu İlişkisel Sınıflandırıcı (SPA RC) Ağırlıklı Oylama İlişkisel Komşu Sınıflandırıcı (WVRN)	Kaldıraç
Lu, Lin, Lu, Zhang (2014)	Telekom	➢ Geniş ölçekli müşteri ağına sahip telekom şirketi o 7190 müşteri o 678 müşteri churn	Lojistik Regresyon Ki-Kare Otomatik Etkileşim Belirleme Algoritması (CHAID) Hızlandırma Algoritmaları	Alıcı İşletim Karakteristik Eğrisi (ROC) Kaldıraç ROC Eğrisi Alt. Alan
Vafeiadis, Diamantaras, Sarigiannidis, Chatzisavvas (2015)	Telekom	➢ Churn - UCI Machine Learning Repository o 5000 kayıt o 19 nitelik	Yapay Sinir Ağları Karar Ağaçları Destek Vektör Makineleri Lojistik Regresyon Hızlandırma Algoritmaları	F-ölçüsü Doğruluk Kesinlik Duyarlılık
Hudaib, Dannoun, Harfoushi, Obiedat, Faris (2015)	Telekom	➢ Jordanian Telecommunication Company o 5000 kayıt o 11 nitelik o %7,6 churn oranı	K-Ortalamalar Hiyerarşik Kümeleme Yapay Sinir Ağları Karar Ağaçları	Doğruluk Karışıklık Matrisi
Al-Shboul, Faris, Ghatasheh (2015)	Telekom	➢ Ürdün'de hizmet veren telekom şirketi o 14573 müşteri o 11 nitelik	Sade Bayes Rastgele Orman Algoritması Destek Vektör Makineleri Genetik Programlama Yapay Sinir Ağları En Yakın K-Komşu Bulanık C-Ortalamalar	Doğruluk Ayrılma Eğilimli Müşteri Oranı Yakalama Oranı Kaldıraç Katsayısı

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Moeyersoms, Martens (2015)	Enerji	➢ Belçika'da hizmet veren enerji şirketi o 1000000'dan fazla müşteri	Karar ağaçları Destek Vektör Makineleri Lojistik Regresyon Kukla Kodlama Semantik Gruplama Kanıt Ağırlıkları Danışmanlık Oranı Perlich Oranı	Doğru Pozitif Oranı Kesinlik ROC Eğrisi Altındaki Alan Kaldıraç WOE-Skoru Perlich Oranları
Xiao, Xiao, Huang, Liu, Wang (2015)	Karma	➢ Churn veriseti- UCI Machine Learning Repository o 3333 kayıt o 18 nitelik o %5,9006 churn oranı ➢ China-Churn veri seti- Chongqing, Çin'de hizmet veren banka o 1255 kayıt o 25 nitelik o %8,29 churn oranı	Transfer Öğrenme Yöntemi Çoklu Sınıflandırıcılar Topluluğu (MCE) Sinir Ağları Duyarlılık Analizi	ROC Eğrisi Altındaki Alan Doğruluk
Ismail, Awang, Rahman, Makhtar (2015)	Telekom		Sinir Ağları Çoklu Regresyon Analizi Lojistik Regresyon	Doğruluk Duyarlılık Belirleyicilik
Tamaddoni, Stakhovych, Ewing (2015)	Karma	➢ Online satıcı I o 122489 müşteri ➢ Online satıcı II o 23570 müşteri	Destek Vektör Makineleri Lojistik Regresyon Hızlandırma Algoritmaları	Kümülatif Kaldıraç Kümülatif Fayda
Rodan, Fayyoumi, Faris, Alsakran, Al-Kadi (2015)	Telekom	➢ Ürdün'de hizmet veren telekom şirketi o 5000 müşteri o 11 nitelik o %0,076 churn oranı	Sinir Ağları K-En Yakın Komşu Sade Bayes Random Forest Genetik Programlama Karar Ağaçları Destek Vektör Makineleri Hızlandırma Algoritmaları Maliyet Duyarlı Öğrenme Yöntemleri Sentetik Azınlık Örnekleme Algoritması (SMOTE) Parçacık Sürü Optimizasyonu	Doğruluk Başarı/Yakalama Oranı Doğru Pozitif Sayısı
Keremati, Ghaneei, Mirmohammadi (2016)	Banka	➢ Bir bankanın çevrimiçi hizmetlerini kullanan müşteriler o 4383 müşteri o demografik nitelikler	Karar Ağaçları	Doğruluk F Ölçüsü Kesinlik Özgünlük Hassasiyet Gini İndeksi

Tablo 2.6: Literatür incelemesi (devam).

Çalışma	Sektör	Veri Seti	Yöntemler	Performans Değ. Ö.
Gajowniczek, Orłowski, Ząbkowski (2016)	Telekom	➢ Duke University o 71047 kayıt o 78 nitelik	Lojistik Regresyon Karar Ağaçları Destek Vektör Makineleri Sinir Ağları	AUC Kaldıraç
Fathian, Hoseinpoor, Minaei-Bidgoli (2016)	Telekom	➢ Duke University o 40000 kayıt o 76 nitelik	Karar Ağaçları Sinir Ağları Destek Vektör Makineleri K-En Yakın Komşu Temel Bileşenler Analizi Kümeleme	Doğruluk Hassasiyet Özgünlük AUC
Dolatabadi, Keynia (2017)		➢ Bir işletmeye ait müşteri o 15 nitelik o 9239 kayıt o %15,3 churn ➢ Bir işletmeye ait çalışan ➢ 21 nitelik ➢ 9239 kayıt ➢ %24,2 churn	Karar Ağaçları Sade Bayes Destek Vektör Makineleri Sinir Ağları	Karışıklık Matrisi Doğruluk
Coussement, Lessmann, Verstraeten (2017)	Telekom	➢ Avrupa'da hizmet veren telekom şirketi o 30104 müşteri o 956 nitelik o %4,52 churn	Bayes Ağları Karar Ağaçları Sinir Ağları Sade Bayes Rastgele Orman Algoritması Destek Vektör Makinesi Hızlandırma Algoritmaları	AUC Kadıraç
Lee, Kim, Lee (2017)	Telekom	➢ Güney Kore'de hizmet veren mobil sağlayıcı ➢ 2010-2012 yılları arasında operatörünü değiştiren müşteri ➢ 322 medya organı	Lojistik Regresyon Karar Ağaçları Sinir Ağları	Yanlış Sınıflandırma Oranı Doğruluk

3. MALZEME VE YÖNTEM

Bu bölümde tez çalışmasında kullanılan veri seti ve özelliklerine, gerekli kodlamalar için kullanılan programlama dili ve araçlarına, çalışmaya uygun olabilmesi için veri seti üzerinde yapılan işlemlere ve problem çözümü için uygulanan modellere yer verilmiştir. Tüm bu işlemler için Bölüm 2.3.3.3'te anlatılan CRISP-DM sürecine bağlı kalınmıştır.

3.1. PROBLEM VE KAPSAMIN BELİRLENMESİ

Müşteri kayıp analizi probleminde en önemli amaç ayrılma eğilimli müşterilerin ve ayrılma nedenlerinin belirlenmesidir. Bu doğrultuda veri madenciliği yöntemleri kullanılırken var olan yöntemlere göre çeşitli avantajlar sağlayacak şekilde modeller geliştirilmeye çalışılmaktadır. Bu avantajlar arasında daha yüksek doğruluk oranı ile ayrılma eğilimli müşterileri tahmin etme, daha kısa işlem süresi ile sonuç elde etme, daha az sayıda parametre belirleme gerekliliği, sonuçların anlaşılırlığı, uygulama kolaylığı sayılabilir.

Bahsi geçen avantaj ve dezavantajlar da göz önünde bulundurularak "Müşteri kayıp analizi problemi için kullanılabilecek ve literatürde kendini ispatlamış ya da sektörel alanda da sıklıkla kullanılan veri madenciliği yöntemlerine göre daha iyi performans gösterecek yeni bir yöntem önerisi ne olabilir?" sorusuna cevap aranmaya çalışılmış, bu doğrultuda aşağıdaki şekilde tez çalışma planı oluşturulmuştur.

- Müşteri kayıp analizi problemi çözümü için kullanılan veri madenciliği yöntemlerinin araştırılması.
- Sınıflandırma probleminin çözümü için kullanılabilecek yeni veri madenciliği yöntemlerinin araştırılması.
- Müşteri kayıp analizi probleminin çözümü için yeni bir yöntem önerisinin araştırılması,
- Önerilecek yeni yöntem için performansın arttırılmasına yönelik çalışmaların gerçekleştirilmesi.

Çalışma kapsamında ilk olarak literatür araştırması yapılarak güncel çalışmaların hangi yönde eğilim gösterdiği, alandaki yeniliklerin neler olduğu, elde edilen performans değerlerinin hangi aralıklarda kabul edilebilir olduğu, hangi sektörde müşteri kayıp analizi çalışmalarının daha yoğun gerçekleştirildiği gibi Bölüm 2.8'de de bahsi geçen sorulara yanıt aranmıştır.

Bu araştırma sonucunda öncelikle hangi sektör için uygulamanın gerçekleştirileceğine ve kullanılacak veri setine karar verilmiştir. Özellikle numara taşıma kolaylığının sağlanması, müşterilerin daha sık hizmet alımı yapması, memnuniyet veya memnuniyetsizliğini daha hızlı bir şekilde gösterebilmesi, çok daha kolay ve hızlı bir şekilde hizmet alınan işletmenin değiştirebilmesi vb nedenlerle "telekom" sektöründe müşteri kayıplarının daha net yaşandığını söylemek mümkündür. Müşteri kayıp analizi çalışmaları da bu sektör tarafından özellikle takip edilmekte ve hatta çalışmalar için müşteri ilişkileri bölümü altında özel ekipler oluşturulmaktadır. Tüm bu gelişmeler göz önüne alınarak tez çalışması için telekom sektörüne yönelik bir uygulama gerçekleştirilmesi amaçlanmıştır.

Uygulama alanının belirlenmesinin ardından önerilecek yeni yöntemin performans karşılaştırmaları için literatür araştırması sonucunda en cok bahsi geçen algoritmalardan K-En Yakın Komşu Algoritması, Sade Bayes Algoritması ve Destek Vektör Makineleri seçilerek bu algoritmaların farklı performans geçerleme yöntemlerine göre uygulamaları gerçekleştirilmiş, performans değerlendirme ölçüleri elde edilmiştir.

Önerilecek yeni yöntemin belirlenmesi için yeniden bir literatür araştırmasına gidilmiştir. Bu kapsamda müşteri kayıp analizi probleminin çözümü için daha önce uygulanmamış veri madenciliği yöntemlerinin araştırılmasına yönelik indekslere girmeyen çalışmaların da olma ihtimaline karşı arama, "google scholar" arama motoru kullanılarak gerçekleştirilmiştir. Bu aramada, müşteri kayıp analizi probleminden farklı bir problem için geliştirilmiş, kullanılmış ve veri madenciliği analizlerinde kullanılabilen yöntemler incelenmiştir. Arama esnasında yaklaşık 230 makaleye erişim sağlanmış, yöntemler arasında Aşırı Öğrenme Makineleri (Extreme Learning Machine-ELM) ve İlişkisel Sınıflandırma Yöntemleri (Associative Classification-AC)'nin ön plana çıktığı görülmüştür. Tez çalışması için Aşırı Öğrenme Makineleri'nden "Tek Gizli Katmanlı İleri Beslemeli Sinir Ağları Temelli Aşırı Öğrenme Makinesi Algoritması" seçilmiştir. Tez çalışmasının buradan sonraki bölümlerinde Aşırı Öğrenme Makinesi ile kastedilen "Tek Gizli Katmanlı İleri Beslemeli Sinir Ağları Temelli Aşırı Öğrenme Makinesi Algoritması"dır. Bu algoritmanın seçilmesi ile beraber aşağıdaki sorulara yanıt aramıştır.

- Aşırı Öğrenme Makineleri algoritmalarından «Tek Gizli Katmanlı İleri Beslemeli Sinir Ağları Temelli Aşırı Öğrenme Makinesi Algoritması»nın müşteri kayıp analizi probleminin çözümündeki başarısı nedir?

- «Tek Gizli Katmanlı İleri Beslemeli Sinir Ağları Temelli Aşırı Öğrenme Makinesi Algoritması»nın bu problemin çözümündeki başarının iyileştirilmesine yönelik olarak uygun parametreleri nelerdir?

Tüm modellerin kodlanması aşamasında R programlama dili ve RStudio editörü kullanılmıştır (R Development Core Team, 2008; RStudio Team, 2016). R programlama dili son dönemde birçok çalışmada kullanılan bir dil olmak ile beraber aşağıdaki özellikleri nedeni ile bu çalışmada da kullanılması uygun görülmüştür (Satman, 2010; Matloff, 2011).

- Temel olarak bilimsel hesaplama ve grafikler için geliştirilmiştir.
- Analizlerde kullanılabilecek, kolaylıkla indirilebilen ve kullanıma hazır birçok paket bulunmaktadır.
- Çeşitli grafik olanakları sunmaktadır.
- Farklı paket ve fonksiyonlar geliştirilebilmekte ve entegre edilebilmekte, böylece kolaylıkla genişletilebilmektedir.
- Açık kaynak kodlu bir programlama dilidir.
- Kullanıcı grupları ile iletişime geçerek yaşanılan problemlere çözüm üretilebilmektedir.
- Farklı işletim sistemlerine uyumlu çalışabilmektedir.
- Farklı programlama dilleri ile ilişki kurulabilmektedir.

Tez kapsamında R programlama dili ile beraber kullanılan paketler aşağıda listelenmiştir.

- Veri setini okuma ve yazdırma için - xlsx (Dragulescu, 2014)
- Veri dönüşümü için - clusterSim (Walesiak ve Dudek, 2017)
- Performans geçerleme yöntemlerinden belirli oranlarda bölme için- caret (Kuhn, 2016)
- Performans geçerleme yöntemlerinden çapraz geçerleme için - TunePareto (Müssel ve diğ., 2012)
- K-En Yakın Komşu Algoritması analizleri için - class (Venables ve Ripley, 2002)
- Destek Vektör Makinesi analizleri için - e1071 (Meyer ve diğ., 2015)
- Aşırı Öğrenme Makinesi analizleri için - elmNN (Gosso, 2012)

3.2. VERİNİN ANLAŞILMASI

Tez kapsamında kullanılacak veri setinin sektör olarak telekom sektöründen seçilmesi ve nedenleri bir önceki başlıkta anlatılmıştır. Bu doğrultuda telekom alanında Türkiye'nin önde gelen hizmet sağlayıcıları ile görüşmeler yapılmıştır. Ancak çeşitli güvenlik gerekçeleri ile veri seti talebi geri çevrilmiştir. Bu noktada UCI Machine Learning Repository'de herkese açık olarak yayınlanmış ve literatürde de kullanılmış olan "churn" veri seti analizler için kullanılmaya karar verilmiştir (URL1). Bu durum elde edilen sonuçların literatürdeki çalışmalar ile de karşılaştırılabilir hale gelmesine olanak sağlamıştır.

Kullanılan "churn" veri seti toplamda 20 nitelik alanı ve 5000 kayıt içermektedir. Nitelik alanlarından 5'i kategorik değerler alırken geriye kalan 15 nitelik alanı nümerik değerler almaktadır. Kategorik nitelik alanlarından biri müşterinin ayrılma durumunda olup olmadığını belirten "churn" nitelik alanıdır. Bu nitelik alanı ikili kategorik nitelik olup 0 ve 1 şeklinde değerler almaktadır. 0 değerleri hizmet alımına devam eden müşterileri gösterirken, 1 değerleri ayrılan müşterileri göstermektedir. Veri setinin orijinalinde "state" alanı da bulunmakta olup bu alan analiz çalışmaları dışında tutulmuş bu nedenle toplam 19 nitelik alanı ile çalışılmıştır. Analiz çalışmalarında kullanılan veri setine ait orijinal nitelik alan adları ve veri tipleri Tablo 3.1'de verilmiştir.

Tablo 3.1: Veri seti nitelik adları ve veri tipleri.

Alan Adı	**Veri Tipi**	**Alan Adı**	**Veri Tipi**
account_length	Nümerik	number_csc	Nümerik
area_code	Kategorik	total_day_charge	Nümerik
international_plan	Kategorik	total_eve_minutes	Nümerik
voice_mail_plan	Kategorik	total_eve_calls	Nümerik
number_vmail_message	Nümerik	total_eve_charges	Nümerik
total_day_minutes	Nümerik	total_night_minutes	Nümerik
total_day_calls	Nümerik	total_night_calls	Nümerik
total_night_charge	Nümerik	total_intl_minutes	Nümerik
total_intl_calls	Nümerik	total_intl_charge	Nümerik
		churn	Kategorik

Veri seti incelendiğinde içerik bakımından temel olarak iki özellikte nitelik içermektedir. Bunlar,

- Müşterilerin sahip olduğu/satın aldığı hizmetler ve müşteri hizmetleri bölümü ile olan etkileşimi ile ilgili alanlar (Örneğin sesli mesaj sistemi kullanımı),
- Müşterinin kullanım miktarları ve ödemeleri ile ilgili alanlar (müşterilik süresi, ödeme tutarları vb)

şeklinde sıralanabilir.

Veri setinde ayrılan ve hizmet alımına devam eden müşteri sayıları incelendiğinde 707 ayrılan müşteri ve 4293 hizmet alımına devam eden müşteri olduğu görülmektedir. Yüzdesel oran olarak ayrılan müşteriler tüm veri setindeki kayıtların %14.14'üdür. Bu durum dengesiz sınıf dağılımı açısından incelendiğinde bu veri seti için de dengesiz bir dağılım olduğu söylenebilir. Veri setindeki her bir nitelik alanının aldığı değerlerin kendi içerisinde dağılımı da incelenmiştir. Nümerik değer alan nitelik alanlarının minimum, maksimum, ortalama ve medyan değerlerinden oluşan özete Şekil 3.1'de yer verilmiştir.

```
account_length  number_vmail_messages total_day_minutes total_day_calls total_day_charge total_eve_minutes total_eve_calls
Min.   :  1.0   Min.   : 0.000       Min.   :  0.0     Min.   :  0     Min.   : 0.00    Min.   :  0.0     Min.   :  0.0
1st Qu.: 73.0   1st Qu.: 0.000       1st Qu.:143.7     1st Qu.: 87     1st Qu.:24.43    1st Qu.:166.4     1st Qu.: 87.0
Median :100.0   Median : 0.000       Median :180.1     Median :100     Median :30.62    Median :201.0     Median :100.0
Mean   :100.3   Mean   : 7.755       Mean   :180.3     Mean   :100     Mean   :30.65    Mean   :200.6     Mean   :100.2
3rd Qu.:127.0   3rd Qu.:17.000       3rd Qu.:216.2     3rd Qu.:113     3rd Qu.:36.75    3rd Qu.:234.1     3rd Qu.:114.0
Max.   :243.0   Max.   :52.000       Max.   :351.5     Max.   :165     Max.   :59.76    Max.   :363.7     Max.   :170.0
total_eve_charge total_night_minutes total_night_calls total_night_charge total_intl_minutes total_intl_calls
Min.   : 0.00    Min.   :  0.0       Min.   :  0.00    Min.   : 0.000     Min.   : 0.00      Min.   : 0.000
1st Qu.:14.14    1st Qu.:166.9       1st Qu.: 87.00    1st Qu.: 7.510     1st Qu.: 8.50      1st Qu.: 3.000
Median :17.09    Median :200.4       Median :100.00    Median : 9.020     Median :10.30      Median : 4.000
Mean   :17.05    Mean   :200.4       Mean   : 99.92    Mean   : 9.018     Mean   :10.26      Mean   : 4.435
3rd Qu.:19.90    3rd Qu.:234.7       3rd Qu.:113.00    3rd Qu.:10.560     3rd Qu.:12.00      3rd Qu.: 6.000
Max.   :30.91    Max.   :395.0       Max.   :175.00    Max.   :17.770     Max.   :20.00      Max.   :20.000
total_intl_charge number_customer_service_calls
Min.   :0.000     Min.   :0.00
1st Qu.:2.300     1st Qu.:1.00
Median :2.780     Median :1.00
Mean   :2.771     Mean   :1.57
3rd Qu.:3.240     3rd Qu.:2.00
Max.   :5.400     Max.   :9.00
```

Şekil 3.1: Veri seti özet görünümü.

3.3. VERİNİN HAZIRLANMASI

Veri setinin analizlere uygun hale gelebilmesi için bazı ön işlemlerden geçmesi gerekmektedir. Bu ön işlemler Bölüm 2.3.3.3'te anlatılmıştır. Bu tez çalışmasında kullanılacak veri seti için veri dönüştürme işlemi gerçekleştirilmiştir.

Nitelik alanlarının aldıkları değerlerin aynı aralıkta dağılarak analizler sırasında herhangi birinin sadece aldığı değerin büyük olmasından ötürü baskın gelmemesi adına veri dönüşümü işlemi yapılması uygun görülmüştür. Bu doğrultuda doğrusal veri dönüşümü yöntemi kullanılmıştır. Doğrusal veri dönüşümü işlemi ile beraber tüm nitelik alanlarının aldığı değerler [0, 1] aralığına taşınmıştır.

Doğrusal veri dönüşümünde nitelik alanının sahip olduğu en küçük değer 0'a eşitlenirken en büyük değer ise 1'e eşitlenmektedir. Aradaki tüm değerler denklem (3.1) kullanılarak [0, 1] aralığındaki değerlere dönüştürülmektedir.

$$x_{\text{normalDeger}} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

Destek Vektör Makineleri ve Aşırı Öğrenme Makineleri sadece nümerik değer alan nitelik alanları ile çalışmaktadırlar. Dolayısıyla veri seti içerisinde yer alan kategorik nitelik alanları ya veri setinden çıkartılmak durumunda ya da çeşitli teknikler ile nümerik gösterime dönüştürülmek durumundadır. Bu tez çalışmasında kategorik değişkenlerin dönüşümü tercih edilmiştir. Veri seti içerisinde yer alan kategorik değişkenler «sahte» (dummy) değişkenler yardımıyla ikili değerler ile ifade edilerek matris gösterimine dönüştürülmüştür. Örneğin; "*area_code*" nitelik alanı üç kategoriden meydana gelmektedir. Bu nedenle tek bir "*area_code*" alanı yerine iki yeni nitelik alanı da eklenerek "*area_code1*", "*area_code2*", "*area_code3*" şeklinde toplam 3 nitelik alanı elde edilmiştir. İlk durumda nitelik alanının aldığı 2 değeri "*area_code1*" nitelik alanı için 0, "*area_code2*" nitelik alanı için 1 ve "*area_code3*" nitelik alanı için 0 değerini almıştır.

Bazı durumlarda nitelik alanlarının aldığı değerler içerisinde normalin dışında değerler olabilmektedir. Bu değerler aykırı (uç) değerler olarak ifade edilmektedir. Bu değerler, analiz sonuçlarını kendi yönlerinde etkilememesi adına analizlerin dışında tutulabilmektedir. Bu değerlerin belirlenebilmesi için de aykırı değer analizi yapılmaktadır. Ancak, Şekil 3.1'e göre nitelik alanlarının aldığı minimum ve maksimum değerler incelendiğinde bu değerler içerisinde sonuçları etkileyecek düzeyde aykırılık olmadığı gözlemlenmektedir. Bunun yanı sıra ayrılma eğilimi gösteren müşteriler için nitelik alanlarının aldığı değerler aykırılık içerebilir, buna bağlı olarak aranan müşteriler bu aykırı değere sahip müşteriler olabilir. Dolayısıyla her iki sebep göz önünde bulundurulduğunda ayrıca bir aykırı değer analizi yapılmaması uygun görülmüştür. Veri seti içerisinde eksik veri de yer almamaktadır.

3.4. MODELLEME

Tez çalışması kapsamında gerçekleştirilen analizlerde K-En Yakın Komşu Algoritması, Sade Bayes Algoritması, Destek Vektör Makineleri ve Aşırı Öğrenme Makineleri olmak üzere dört farklı yöntemden yararlanılmıştır. Bu yöntemler ile müşteri kayıp analizi için çeşitli modeller ile probleme yönelik çözüm önerileri oluşturulmuştur. Aşırı Öğrenme Makineleri müşteri kayıp analizi için uluslararası literatürde fazla kullanılmayan hatta tez çalışması kapsamındaki algoritması daha önce kullanılmamış, Türkçe literatürde ise bu problem için kullanılmamış bir yöntem olarak seçilmiş ve farklı parametreler için performansının gözlemlenmesine yönelik çalışmalar gerçekleştirilmiştir. Diğer üç yöntem ise Aşırı Öğrenme Makinelerinin performansını kıyaslamak üzere kullanılmıştır.

Modeller oluşturulurken farklı performans geçerleme yöntemleri kullanılmıştır. Belirli oranlarda bölme yöntemi %70-%30 ve %80-%20 oranları ile çapraz geçerleme yöntemi ise 5 ve 10-kat için gerçekleştirilmiştir. Burada belirtilen oranlar tüm veri setinin eğitim ve test veri setleri şeklinde ayrılabilmesi için kullanılan oranlardır. Çapraz geçerleme yöntemi kullanıldığında performans değerlendirme ölçüleri, tüm katlardan elde edilen ölçülerin ortalaması alınarak elde edilmiştir.

Kullanılan yöntem ve algoritmalar çeşitli parametre değerlerine ihtiyaç duymaktadır. Parametrelerde meydana gelecek değişiklikler analiz sonuçlarını da etkileyebilecektir. Bu doğrultuda her bir model için parametre değerlerinin belirlenmesine ihtiyaç duyulmuştur. K-En Yakın Komşu Algoritması için en önemli parametre değeri *k* olarak ifade edilen komşu sayısıdır. Bu çalışma için *k* değerleri 3, 5 ve 7 olarak seçilmiş, performanslar bu değerler ile elde edilen sonuçlar üzerinden değerlendirilmiştir.

Destek Vektör Makineleri için de bazı parametrelerin belirlenmesine ihtiyaç duyulmuştur. Bu parametrelerden biri hangi çekirdek (kernel) fonskiyonun kullanılması gerektiğidir. Modellerde radial tabanlı, polynominal ve sigmoid kernel fonksiyonlar tercih edilmiştir. R programlama dili ile analizler için kullanılan paketlerde ihtiyaç duyulan diğer parametreler ise *cost, coefo* ve *degree* parametreleri şeklinde karşımıza çıkmaktadır. Paketler vasıtasıyla kullanılan fonksiyonların varsayılan olarak kullandığı parametre değerleri degree parametresi için 3 (polynominal için gerekli), coefo parametresi için 1/veri boyutu (polynominal ve sigmoid için gerekli) ve cost parametresi için 1'dir.

Aşırı Öğrenme Makinesinin ise temelde değişen iki ana parametresi olup bunlar; ara (gizli) katmandaki nöron sayısı ile aktivasyon fonksiyonudur.

Aşırı Öğrenme Makineleri yeni yöntem olarak seçildikten sonra diğer yöntemlerin karşısında performansının ne olacağını genel anlamda görebilmek ve iyileştirilmesine yönelik çalışmaya açık olup olmadığını anlayabilmek adına ön çalışmalar gerçekleştirilmiştir. Bu kapsamda ön analizlerde aktivasyon fonksiyonları olarak sigmoid, sinüs ve radial tabanlı aktivasyon fonksiyonları; gizli nöron sayısı olarak ise 50, 100 ve 200 değerleri seçilmiştir. Yöntem Yapay Sinir Ağları temelli olması nedeniyle ürettiği çıktı değer sayısal sürekli bir değer olup, bu değerin hangi sınıf etiketine (nitelik değerine) karşılık geleceğinin belirlenmesi gerekmektedir. Bu nedenle veri setinin geneline uygulanan doğrusal veri dönüşümü çıktı değerlerine de uygulanarak tüm değerler [0, 1] aralığına taşınmıştır. Ancak bu noktada da bir eşik değer belirlenmesi gerekmekte ve bu eşik değerin altında ve üstünde kalan değer iki farklı sınıfa karşılık gelecek şekilde sınıf etiketlerinin belirlenmesi gerçekleştirilmelidir. Yapılan ön analiz çalışmasında bu eşik değer 0.70 olarak kabul edilmiştir.

Aşırı Öğrenme Makinesi için yapılan ön analiz sonuçlarında eşik değerinin de bir parametre şeklinde değerlendirilip farklı eşik değerlerine göre sonuçların elde edilmesi gerekliliği durumu ortaya çıkmıştır. En iyi parametrelerin aranmasına geçmeden önce yine bir ön çalışma gerçekleştirilerek sonuçların belirli bir eşik değerini işaret edip etmediği belirlenmek istenmiştir. İkinci adımı oluşturan bu analizler için de parametre değerleri

- Gizli nöron sayısı 100, 500 ve 1000
- Aktivasyon fonksiyonları; sigmoid, sin, radial basis, hard-limit, symmetric hard-limit, satlins, tan-sigmoid, triangular basis, positive linear, linear
- Eşik değerleri ise 0.30, 0.40, 0.50, 0.60, 0.70

şeklinde belirlenmiştir. Ancak sadece 0.30 değeri için sonuçların genel olarak hatalı oluştuğu sonucuna varılmıştır.

Son aşamada ise değişiklik öngörülen parametre değerleri ile beraber en iyi parametre kombinasyonlarının belirlenmesi konusunda çalışmalar tamamlanmıştır. Bu analizler için kullanılan parametre değerleri;

- Gizli nöron sayısı [1,1000]'da tamsayılar

- Aktivasyon fonksiyonları sigmoid, sin, radial basis, hard-limit, symmetric hard-limit, satlins, tan-sigmoid, triangular basis, positive linear, linear
- Eşik değerleri 0.40, 0.50, 0.60, 0.70

şeklinde belirlenmiş ve tüm değerler ile analizler gerçekleştirilmiştir.

3.5. PERFORMANS DEĞERLENDİRME

Performans değerlendirme ölçüleri olarak ise doğruluk, hata oranı, duyarlılık, özgünlük, kesinlik, F ölçüsü kullanılmıştır. Performans değerlendirme ölçülerinin hesaplanması için referans değer olarak ayrılan müşteriler alınmıştır. Bunun yanı sıra farklı bir doğruluk oranı hesaplanarak bu oran da göz önünde bulundurulmaya çalışılmıştır. Yeni doğruluk oranı denklem (3.2) ile hesaplanmaktadır.

$$Yeni\_Doğruluk = YD = \frac{DP}{DP+YN+YP} \tag{3.2}$$

Bu denkleme göre ayrılma durumu olmayan ve bu durumu model tarafından doğru tahmin edilen müşteriler doğruluk oranı hesabı dışında tutulmuştur. Bu dışarıda tutma durumu müşteri kayıp analizi ile amacın ayrılan/ayrılma eğilimi gösteren müşterileri tahmin edebilmek olması, zaten ayrılma durumu olmayan bir müşterinin tahmin edilebilmesinin başarı olarak görülerek doğruluğa katılmasına ihtiyaç duyulmadığı fikri ile hesaplanmıştır.

4. BULGULAR

Bu bölümde oluşturulan modellere ait analizlerin belirtilen performans değerlendirme ölçülerine göre sonuçlarına yer verilmiştir. Bulgular modellemede bahsedildiği üzere aşağıdaki sırada listelenmiştir:

- K-En Yakın Komşu Algoritması
- Sade Bayes Algoritması
- Destek Vektör Makinesi
- Aşırı Öğrenme Makinesi Ön Analiz
- Aşırı Öğrenme Makinesi En İyi Parametre Analizi
- Karşılaştırmalı Sonuçlar

4.1. K-EN YAKIN KOMŞU ALGORİTMASINA AİT BULGULAR

K-En Yakın Komşu Algoritması ile elde edilen bulgulara Tablo 4.1'de yer verilmiştir. Performans değerlendirme ölçülerinin aldığı en yüksek değerler performans geçerleme yöntemlerine göre sınıflandırılmıştır.

Tablo 4.1: K-En Yakın Komşu Algoritmasının performans değerlendirme ölçü değerleri.

P. G.Y.		**Doğruluk**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
%70 Eğitim %30 Test	k=3	0,8980	0,1020	0,3750	0,9748	0,6857	0,4848	0,3200
	k=5	**0,9093**	0,0907	0,3542	0,9908	0,8500	**0,5000**	**0,3333**
	k=7	0,9080	0,0920	0,3229	0,9939	0,8857	0,4733	0,3100
%80 Eğitim %20 Test	k=3	0,8970	0,1030	0,3740	0,9758	0,7000	0,4876	0,3224
	k=5	**0,9090**	0,0910	0,3664	0,9908	0,8571	**0,5134**	**0,3453**
	k=7	0,9080	0,0920	0,3282	0,9954	0,9149	0,4831	0,3185
Çapraz Geçerleme (5 Kat)	k=3	**0,8990**	0,1010	0,3214	0,9942	0,9018	0,4735	**0,3291**
	k=5	0,8976	0,1024	0,3345	0,9904	0,8539	**0,4804**	0,3086
	k=7	**0,8990**	0,1010	0,3214	0,9942	0,9018	0,4735	**0,3291**
Çapraz Geçerleme (10 Kat)	k=3	0,8882	0,1118	0,3382	0,9786	0,7201	0,4583	0,3111
	k=5	0,8982	0,1018	0,3341	0,9909	0,8583	**0,4789**	0,3596
	k=7	**0,9004**	0,0996	0,3220	0,9954	0,9150	0,4747	**0,3605**

Tablo 4.1'e göre veri setinin %70 eğitim ve %30 test verisi olarak bölünmesi durumunda en yüksek doğruluk değeri %90,93, en yüksek F ölçüsü %50 ve en yüksek yeni doğruluk değeri olan %33,33 *k*=5 seçildiğinde elde edilmiştir. Performans geçerleme için veri setinin %80 eğitim ve %20 test verisi olarak bölünmesi durumunda ise en yüksek doğruluk değeri %90,90, en yüksek F ölçüsü %51,34 ve en yüksek yeni doğruluk değeri olan %34,53 yine *k*=5 için elde edilmiştir. Performans geçerleme yöntemi olarak çapraz geçerleme kullanıldığında belirli oranlarda bölme yöntemine göre doğruluk ve F ölçüsünde düşme gözlemlenirken yeni doğruluk değerinde artma olduğu görülmektedir. 10 kat çapraz geçerleme sonucu elde edilen en yüksek değerler doğruluk için %90,04, F ölçüsü için %47,89 ve yeni doğruluk değeri için %36,05'tir.

4.2. SADE BAYES ALGORİTMASINA AİT BULGULAR

Sade Bayes Algoritması'na göre elde edilen bulgulara Tablo 4.2'de yer verilmiştir. K-En Yakın Komşu Algoritmasına benzer şekilde performans değerlendirme ölçülerinin aldığı en yüksek değerler performans geçerleme yöntemlerine göre sınıflandırılmıştır.

Tablo 4.2: Sade Bayes algoritmasının performans değerlendirme ölçü değerleri.

Performans G.Y.	**Doğruluk**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
%70 Eğitim %30 Test	**0,8939**	0,1061	0,4481	0,9674	0,6934	**0,5444**	**0,3740**
%80 Eğitim %20 Test	0,8859	0,1141	0,4113	0,9639	0,6517	0,5043	0,3372
Capraz Geçerleme (5 Kat)	0,8872	0,1128	0,4461	0,9602	0,6483	0,5272	0,3584
Capraz Geçerleme (10 Kat)	0,8868	0,1132	0,4494	0,9593	0,6437	0,5262	0,3582

Tablo 4.2'ye göre en yüksek doğruluk değeri %89,39 oranı ile, en yüksek F ölçüsü değeri %54,44 oranı ile ve en yüksek yeni doğruluk oranı %37,40 ile veri setinin %70 eğitim ve %30 test verisi şeklinde bölünmesi sonucu elde edilmiştir.

Hassasiyet değerleri göz önünde bulundurulduğunda gerçekte «ayrılma eğilimli» olan müşteriler içerisinde müşterilerin bu durumunu doğru tahmin etme oranı %44,94 olarak en iyi şekilde çapraz geçerleme (10-kat) performans geçerleme yöntemi ile elde edilmiştir. Kesinlik ölçüsü incelendiğinde ise ayrılma eğilimli olarak tahmin edilen müşterilerin %69,34'ünün

gerçekte de ayrılma eğilimli olduğu tahmin oranı ile en yüksek değer veri setinin %70 eğitim-%30 test verisi şeklinde bölünmesi ile elde edilmiştir.

4.3. DESTEK VEKTÖR MAKİNELERİNE AİT BULGULAR

Destek Vektör Makineleri ile geliştirilen modeller uygulandığında Tablo 4.3'te yer alan analiz sonuçlarına erişilmiştir.

Tablo 4.3: Destek Vektör Makinelerinin performans değerlendirme ölçü değerleri.

P. G.Y.	**Çekirdek Fonk.**	**Doğruluk**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
%70 Eğitim %30 Test	radial	0,9180	0,0820	0,4271	0,9901	0,8632	0,5714	0,4000
	polynomial	**0,9187**	0,0813	0,4583	0,9862	0,8302	**0,5906**	**0,4190**
	sigmoid	0,8113	0,1887	0,0990	0,9159	0,1473	0,1184	0,0629
%80 Eğitim %20 Test	radial	0,9210	0,0790	0,4656	0,9896	0,8714	0,6070	0,4357
	polynomial	**0,9270**	0,0730	0,5191	0,9885	0,8718	**0,6507**	**0,4823**
	sigmoid	0,8140	0,1860	0,0840	0,9241	0,1429	0,1058	0,0558
Capraz Geçerleme 5 Kat	radial	0,9190	0,0810	0,5012	0,9884	0,8765	0,6363	0,4686
	polynomial	**0,9238**	0,0762	0,5388	0,9874	0,8748	**0,6665**	**0,5009**
	sigmoid	0,8112	0,1888	0,1318	0,9232	0,2211	0,1641	0,0896
Capraz Geçerleme 10 Kat	radial	0,9238	0,0762	0,5249	0,9895	0,8923	0,6590	0,4944
	polynomial	**0,9276**	0,0724	0,5493	0,9897	0,8972	**0,6798**	**0,5191**
	sigmoid	0,8058	0,1942	0,1211	0,9188	0,1975	0,1489	0,0808

Tablo 4.3 incelendiğinde performans geçerleme yöntemi olarak tüm veri seti %70 eğitim ve %30 test verisi şeklinde bölündüğünde %91,87 doğruluk oranı, %59,06 F ölçüsü ve %41,90 yeni doğruluk oranı ile en iyi performans, polinominal kernel fonksiyon ve beraberinde kullanılan parametre değerleri ile elde edilmiştir. Performans geçerleme yöntemi olarak çapraz geçerleme (10-kat) tercih edildiğinde ise %92,76 doğruluk, %67,98 F ölçüsü ve %51,91 yeni

doğruluk oranı ile en iyi performansın yine polinominal çekirdek fonksiyon ile elde edildiği görülmektedir. Bu değerler aynı zamanda Destek Vektör Makineleri ile oluşturulan tüm modellerden elde edilen en yüksek performans değerlerdir. Performans geçerleme yöntemi olarak aynı şekilde çapraz geçerleme (10-kat) tercih edildiğinde ve hassasiyet değerleri göz önünde bulundurulduğunda gerçekte «ayrılma eğilimli» olan müşteriler içerisinde bu durumu doğru tahmin etme oranı %54,93 'tür. Bir başka açıdan kesinlik ölçüsü incelendiğinde ise ayrılma eğilimli olarak tahmin edilen müşterilerin %89,72'sinin gerçekte de ayrılma eğilimli olduğu görülmüştür.

Tüm sonuçlar incelendiğinde modeller içerisinde en iyi sonuçların polinominal çekirdek fonksiyon ile elde edildiği görülmektedir.

4.4. AŞIRI ÖĞRENME MAKİNELERİ ÖN ANALİZ BULGULARI

Aşırı Öğrenme Makinesi yöntemi kullanılarak bu yöntemin müşteri kayıp analizi problemi için kullanılır olup olmadığı açısından öngörü oluşturmak için yapılan analiz çalışmaları sonucunda performans geçerleme yöntemi olarak belirli oranlarda bölme kullanıldığında elde edilen bulgular Tablo 4.4'te verilmiştir.

Tablo 4.4: Aşırı öğrenme makineleri (ön analiz) performans değerlendirme ölçü değerleri (belirli oranlarda bölme yöntemi ile).

P.G.Y.	Parametre	Doğruluk	Hata	Hassasiyet	Özgünlük	Kesinlik	F Ölç.	Yeni doğ.
%70 Eğitim %30 Test	sig NS=50	0,8847	0,1153	0,2344	0,9801	0,6338	0,3422	0,2064
	sin NS=50	0,8813	0,1187	0,3646	0,9572	0,5556	**0,4403**	**0,2823**
	radbas NS=50	0,8753	0,1247	0,0469	0,9969	0,6923	0,0878	0,0459
	sig NS=100	0,8760	0,1240	0,0313	1,0000	1,0000	0,0606	0,0313
	sin NS=100	0,8833	0,1167	0,1042	0,9977	0,8696	0,1860	0,1026

Tablo 4.4: Aşırı öğrenme makineleri (ön analiz) performans değerlendirme ölçü değerleri (belirli oranlarda bölme yöntemi ile) (devam).

P.G.Y.	Parametre	Doğruluk	Hata	Hassasiyet	Özgünlük	Kesinlik	F Ölç.	Yeni doğ.
%70 Eğitim %30 Test	radbas NS=100	**0,8900**	0,1100	0,2500	0,9839	0,6957	0,3678	0,2254
	sig NS =200	0,8860	0,1140	0,1250	0,9977	0,8889	0,2192	0,1231
	sin NS =200	0,8873	0,1127	0,1458	0,9962	0,8485	0,2489	0,1421
	radbas NS =200	**0,8900**	0,1100	0,2240	0,9878	0,7288	0,3426	0,2067
%80 Eğitim %20 Test	sig NS =50	0,8850	0,1150	0,2290	0,9839	0,6818	0,3429	0,2069
	sin NS =50	0,8820	0,1180	0,4733	0,9436	0,5586	**0,5124**	**0,3444**
	radbas NS =50	0,8730	0,1270	0,0458	0,9977	0,7500	0,0863	0,0451
	sig NS =100	0,8760	0,1240	0,0534	1,0000	1,0000	0,1014	0,0534
	sin NS =100	0,8790	0,1210	0,0916	0,9977	0,8571	0,1655	0,0902
	radbas NS =100	**0,8880**	0,1120	0,2443	0,9850	0,7111	0,3636	0,2222
	sig NS =200	0,8790	0,1210	0,0763	1,0000	1,0000	0,1418	0,0763
	sin NS =200	0,8810	0,1190	0,1069	0,9977	0,8750	0,1905	0,1053
	radbas NS =200	0,8850	0,1150	0,2214	0,9850	0,6905	0,3353	0,2014

* NS: Gizli Katmandaki Nöron Sayısı

Tablo 4.4 incelendiğinde eğitim ve test veri seti %70-30 oranları ile ayrılarak analizler gerçekleştirildiğinde ve parametre değerleri radyal tabanlı aktivasyon fonksiyonu ile beraber gizli katmandaki nöron sayısı 100 veya 200 olarak seçilirse %89 oranı ile en yüksek doğruluk değerine erişilmektedir. En yüksek doğruluk değeri bu parametreler ile elde edilirken F ölçüsü ve yeni doğruluk değeri için durum değişmektedir. En yüksek F ölçüsü %44,03 ve en yüksek

yeni doğruluk değeri %28,23 olarak sinüs aktivasyon fonksiyonu ve gizli katmandaki nöron sayısı 50 olacak şekilde elde edilmektedir.

Veri setini eğitim ve test veri seti olarak ayırmak için %80-20 oranı tercih edilirse en yüksek doğruluk oranı %88,80 ile radyal tabanlı aktivasyon fonksiyonu ve 100 gizli katmandaki nöron sayısı ile elde edilmektedir. Burada da benzer şekilde F ölçüsü ve yeni doğruluk oranı için en yüksek değerler farklı parametreler ile elde edilmiştir. Buna göre en yüksek değerler olan %51,24 F ölçüsü ve %34,44 yeni doğruluk oranı sinüs aktivasyon fonksiyonu, gizli katmandaki nöron sayısı 50 olacak şekilde elde edilmiştir.

Aşırı Öğrenme Makinesi ön analizlerinde Performans geçerleme yöntemi olarak çapraz geçerleme kullanıldığında elde edilen bulgular ise Tablo 4.5'te verilmiştir.

Tablo 4.5: Aşırı öğrenme makineleri (ön analiz) performans değerlendirme ölçü değerleri (çapraz geçerleme yöntemi ile).

P.G.Y.	**Parametre**	**Doğruluk**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
Çapraz Geçerleme (5 Kat)	AF:sig NS:50	0,8716	0,1284	0,1982	0,9825	0,6476	0,3006	0,1784
	AF:sin NS:50	0,8758	0,1242	0,2507	0,9786	0,6603	**0,3606**	**0,2212**
	AF:radbas NS:50	0,8640	0,1360	0,1688	0,9789	0,4622	NaN	0,1439
	AF:sig NS:100	0,8752	0,1248	0,1426	0,9963	0,8845	0,2405	0,1387
	AF:sin NS:100	**0,8784**	0,1216	0,1799	0,9930	0,8348	0,2884	0,1715
	AF:radbas NS:100	0,8732	0,1268	0,1462	0,9930	0,6512	NaN	0,1376
	AF:sig NS:200	0,8734	0,1266	0,1060	0,9993	0,9716	0,1874	0,1055
	AF:sin, NS:200	0,8700	0,1300	0,0816	0,9998	0,9895	0,1485	0,0814
	AF:radbas NS:200	0,8732	0,1268	0,1349	0,9946	0,8252	0,2174	0,1282
Çapraz Geçerleme (10 Kat)	AF:sig, NS:50	0,8752	0,1248	0,2199	0,9832	0,6799	0,3285	0,1987

Tablo 4.5: Aşırı öğrenme makineleri (ön analiz) performans değerlendirme ölçü değerleri (çapraz geçerleme yöntemi ile) (devam).

P.G.Y.	Parametre	Doğruluk	Hata	Hassasiyet	Özgünlük	Kesinlik	F Ölç.	Yeni doğ.
Çapraz Geçerleme (10 Kat)	AF:sin, NS:50	0,8800	0,1200	0,2941	0,9763	0,6729	**0,4049**	**0,2568**
	AF:radbas, NS:50	0,8662	0,1338	0,2225	0,9718	0,5166	NaN	0,1819
	AF:sig, NS:100	0,8766	0,1234	0,1599	0,9944	0,8616	0,2620	0,1537
	AF:sin, NS:100	**0,8808**	0,1192	0,2224	0,9900	0,8024	0,3367	0,2071
	AF:radbas, NS:100	0,8774	0,1226	0,1828	0,9914	0,7353	NaN	0,1714
	AF:sig, NS:200	0,8806	0,1194	0,1729	0,9962	0,9339	0,2765	0,1669
	AF:sin NS:200	0,8762	0,1238	0,1343	0,9979	0,9484	0,2230	0,1314
	AF:radbas NS:200	0,8806	0,1194	0,2100	0,9916	0,7681	0,3155	0,1980

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu

Tablo 4.5 incelendiğinde 5 kat çapraz geçerleme yöntemi ile analizler gerçekleştirildiğinde ve parametre değerleri sinüs aktivasyon fonksiyonu ile beraber gizli katmandaki nöron sayısı 100 olarak seçilirse %87,84 oranı ile en yüksek doğruluk değerine erişilmektedir. En yüksek doğruluk değeri bu parametreler ile elde edilirken en yüksek F ölçüsü %36,06 ve en yüksek yeni doğruluk değeri olan %22,12 sinüs aktivasyon fonksiyonu ve gizli katmandaki nöron sayısı 50 olacak şekilde elde edilmektedir.

Performans geçerleme yöntemi olarak tercih edilen çapraz geçerleme 10 kat için gerçekleştirilirse en yüksek doğruluk oranı %88,08 ile sinüs aktivasyon fonksiyonu ve 100 gizli katmandaki nöron sayısı ile elde edilmektedir. Burada da benzer şekilde F ölçüsü ve yeni doğruluk oranı en yüksek değerleri farklı parametreler ile elde edilmiştir. Buna göre en yüksek değerler olan %40,49 F ölçüsü ve %25,68 yeni doğruluk oranı sinüs aktivasyon fonksiyonu ve 50 gizli katmandaki nöron sayısı elde edilmiştir.

Her iki tablo da incelendiğinde elde edilen sonuçların diğer algoritmalardan elde edilen performans değerlendirme ölçülerine göre düşük kaldığı görülmektedir.

4.5. AŞIRI ÖĞRENME MAKİNELERİ EN İYİ PARAMETRE BULGULARI

Aşırı Öğrenme Makinesi yönteminin ön analizlerinden sonra en iyi parametre kombinasyonlarının ortaya konması için analizler gerçekleştirilmiştir. Bu analizler sonucunda performans geçerleme yöntemi olarak belirli oranlarda bölme kullanıldığında elde edilen bulgular Tablo 4.6'da verilmiştir.

Tablo 4.6: Aşırı Öğrenme Makineleri performans değerlendirme ölçü değerleri (belirli oranlarda bölme).

P.G.Y.	Parametreler*			Doğruluk	Hata	Has.	Özg.	Kesinlik	F Ölç.	Yeni doğ.
	NS	AF	ED							
%70 Eğitim %30 Test	368	sig	0.5	0,9200	0,0800	0,6458	0,9602	0,7045	0,6739	0,5082
	584	sin	0.5	0,9107	0,0893	0,7292	0,9373	0,6306	0,6763	0,5109
%80 Eğitim %20 Test	584	sin	0.5	0,9310	0,0690	0,6489	0,9735	0,7870	0,7113	0,5519

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.6 incelendiğinde eğitim ve test veri seti %70-30 oranları ile ayrılarak analizler gerçekleştirildiğinde 368 nöron sayısı, sigmoid aktivasyon fonksiyonu ve 0.5 eşik değeri parametre değerleri ile doğruluk performans değerlendirme ölçüsü %92 oranı ile en yüksek değerine erişmektedir. F ölçüsü ve yeni doğruluk değeri için oranlar yakınlık gösterse de en yüksek değerler olan %67,63 F ölçüsü ve %51,09 yeni doğruluk değeri 584 nöron sayısı, sinüs fonksiyonu ve 0.5 eşik değeri ile elde edilmiştir.

Eğitim ve test veri seti %80-20 oranlarına göre ayrılmak istendiğinde tüm performans değerlendirme ölçüleri aynı parametre değerleri için en yüksek değerlere ulaşmaktadır. Buna göre 584 nöron sayısı, sinüs aktivasyon fonksiyonu ve 0.5 eşik değeri olan parametre değerleri ile %93,10 doğruluk oranı, %71,13 F ölçüsü ve %55,19 yeni doğruluk değeri elde edilmektedir.

Modellerin ayrılma eğilimli müşteri olup, sadece bu müşteriler içerisinde doğru tahmin edebilme veya ayrılma eğilimli müşteri olarak tahmin edilenlerin gerçekte de ayrılma eğilimli olması gibi becerileri ölçen hassasiyet ve kesinlik ölçüleri ile belirli oranlarda bölme yöntemi genel olarak göz önünde bulundurulduğunda Tablo 4.6'ya göre %78,70 kesinlik ve %72,92 hassasiyet ölçüleri farklı performans geçerleme yöntemlerine göre olsa da 584 nöron sayısı, sinüs aktivasyon fonksiyonu ve 0.5 eşik değeri olan parametre değerleri ile elde edilmiştir.

Benzer analizler performans geçerleme yöntemi olarak çapraz geçerleme kullanılarak uygulandığında elde edilen bulgular Tablo 4.7'de verilmiştir.

Tablo 4.7: Aşırı Öğrenme Makineleri performans değerlendirme ölçü değerleri (çapraz geçerleme ile).

Performans G.Y.	**Parametreler***			**Doğruluk**	**Hata**	**Has.**	**Özg.**	**Kesinlik**	**F Ölç.**	**Yeni_D**
	NS	**AF**	**ED**							
Capraz Geçerleme (5 Kat)	409	sin	0.40	0,9038	0,0962	0,6677	0,9426	0,6867	**0,6619**	**0,4950**
	527	radbas	0.50	**0,9130**	0,0870	0,5571	0,9713	0,7779	0,6421	0,4747
Capraz Geçerleme (10 Kat)	386	sig	0.50	**0,9138**	0,0862	0,5116	0,9792	0,8290	0,6148	0,4500
	444	sig	0.40	0,9092	0,0908	0,6741	0,9482	0,6872	**0,6736**	**0,5100**

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.7 incelendiğinde 5-kat çapraz geçerleme performans geçerleme yöntemi ile analizler gerçekleştirildiğinde 527 nöron sayısı, radyal temelli aktivasyon fonksiyonu ve 0.5 eşik değeri parametre değerleri ile doğruluk %91,30 oranı ile en yüksek değerine erişmektedir. F ölçüsü ve yeni doğruluk değeri için oranlar yakınlık gösterse de en yüksek değerler olan %66,19 F ölçüsü ve %49,50 yeni doğruluk değeri 409 nöron sayısı, sinüs aktivasyon fonksiyonu ve 0.4 eşik değeri ile elde edilmiştir.

Çapraz geçerleme 10-kat olarak uygulandığında 386 nöron sayısı, sigmoid aktivasyon fonksiyonu ve 0.5 eşik değeri ile en yüksek doğruluk değeri olan %91,38 elde edilirken F ölçüsü %61,48 ve yeni doğruluk değeri ise %45'te kalmaktadır. 10-kat çapraz geçerleme için en yüksek F ölçüsü %67,36 ve en yüksek yeni doğruluk değeri %51 olarak parametre değerleri 444 nöron sayısı, sigmoid aktivasyon fonksiyonu ve 0.4 eşik değeri kabul edildiğinde elde edilmektedir.

Tablo 4.7'ye göre en yüksek doğruluk, F ölçüsü ve yeni doğruluk değerleri göz önünde bulundurulduğunda kesinlik ölçüsü %82,90 ile en yüksek değerine 10-kat çapraz geçelerme ve 386 nöron sayısı, sigmoid aktivasyon fonksiyonu ve 0.5 eşik değeri paremetreleri ile erişmektedir. Hassasiyet ölçüsü ise %67,41 ile en yüksek değerini 444 nöron sayısı, sigmoid aktivasyon fonksiyonu ve 0.4 eşik değeri parametreleri ile elde etmektedir.

Genel olarak bakıldığında 10-kat çapraz geçerleme uygulandığında performans değerlendirme ölçülerine göre daha iyi bir sonuç elde edildiği görülmektedir.

4.6. SINIFLANDIRMA MODELLERİNİN PERFORMANS ANALİZİ

Buraya kadar olan bölümde her bir algoritmanın kendi içerisindeki en iyi performans ölçüleri paylaşılmıştır. Aşırı Öğrenme Makineleri tez çalışması kapsamında müşteri kayıp analizi probleminin çözümü için önerilen yöntem olup, amaçlardan birisi bu yöntemin literatürdeki diğer etkin yöntemler ile performansının kıyaslanabilir düzeyde iyi olduğunun ispatlanabilmesidir.

Bu bağlamda tüm algoritmaların en iyi performans değerleri doğruluk, F ölçüsü ve yeni doğruluk değerleri açısından ve kullanılan performans geçerleme yöntemlerine göre ele alınarak karşılaştırmalı olarak sunulacaktır.

Tablo 4.8'de belirli oranlarda bölme yöntemi %70-30 oranı ile kullanıldığında tüm algoritmalar için elde edilen en iyi doğruluk değerleri verilmiştir. Yöntemler doğruluk değeri en yüksekten en düşüğe doğru olacak şekilde sıralanmıştır.

Tablo 4.8: Doğruluk ölçüsüne göre sınıflandırma modellerinin performans analizi (%70 Eğitim %30 Test).

Yöntem	**Doğruluk**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
Aşırı Öğrenme Makinesi (NS:368, AF: sig, ED:0.5)	0,9200	0,0800	0,6458	0,9602	0,7045	0,6739	0,5082
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9187	0,0813	0,4583	0,9862	0,8302	0,5906	0,4190
K-En Yakın Komşu (k=5)	0,9093	0,0907	0,3542	0,9908	0,8500	0,5000	0,3333
Sade Bayes	0,8939	0,1061	0,4481	0,9674	0,6934	0,5444	0,3740

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.8'e göre doğruluk performans değerleri açısından Aşırı Öğrenme Makinesinin gizli nöron sayısı 368, sigmoid aktivasyon fonksiyonu ve 0.5 eşik değer ile beraber doğruluk değeri %92 olup, en iyi performansa sahiptir. Aşırı Öğrenme Makinesini sırasıyla Destek Vektör Makineleri, K-En Yakın Komşu Algoritması ve Sade Bayes Algoritması takip etmektedir.

Tablo 4.9'da belirli oranlarda bölme yöntemi %80-20 oranı ile kullanıldığında tüm algoritmalar için elde edilen en iyi doğruluk değerleri verilmiştir. Yöntemler doğruluk değeri en yüksekten en düşüğe doğru olacak şekilde sıralanmıştır.

Tablo 4.9: Doğruluk ölçüsüne göre sınıflandırma modellerinin performans analizi (%80 Eğitim %20 Test).

Yöntem	**Doğruluk**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
Aşırı Öğrenme Makinesi (NS:584, AF: sin, ED:0.5)	0,9310	0,0690	0,6489	0,9735	0,7870	0,7113	0,5519
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9270	0,0730	0,5191	0,9885	0,8718	0,6507	0,4823
K-En Yakın Komşu (k=5)	0,9090	0,0910	0,3664	0,9908	0,8571	0,5134	0,3453
Sade Bayes	0,8859	0,1141	0,4113	0,9639	0,6517	0,5043	0,3372

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.9'a göre doğruluk performans değerleri açısından Aşırı Öğrenme Makinesi gizli nöron sayısı 584, sinüs aktivasyon fonksiyonu ve 0.5 eşik değer ile beraber %93,10 doğruluk değerine sahip olup, en iyi performansı göstermiştir. Aşırı Öğrenme Makinesini sırasıyla Destek Vektör Makineleri, K-En Yakın Komşu Algoritması ve Sade Bayes Algoritması takip etmektedir.

Tablo 4.10'da ise 5-kat çapraz geçerleme yöntemi kullanıldığında tüm algoritmalar için elde edilen en iyi doğruluk değerleri verilmiştir. Yöntemler doğruluk değeri en yüksekten en düşüğe doğru olacak şekilde sıralanmıştır.

Tablo 4.10'a göre doğruluk performans değerleri açısından Destek Vektör Makineleri polynominal çekirdek fonksiyon ile beraber %92,38 doğruluk değerine ulaşarak, en iyi performansa sahip olmuştur. Destek Vektör Makinelerini sırasıyla Aşırı Öğrenme Makinesi, K-En Yakın Komşu Algoritması ve Sade Bayes Algoritması takip etmektedir. Görüldüğü üzere performans geçerleme yöntemi değişikliği ile beraber Aşırı Öğrenme Makineleri yaklaşık %1 fark ile yerini Destek Vektör Makinelerine bırakmıştır.

Tablo 4.10: Doğruluk ölçüsüne göre sınıflandırma modellerinin performans analizi (çapraz geçerleme 5-kat ile).

Yöntem	Doğruluk	Hata	Hassasiyet	Özgünlük	Kesinlik	F Ölç.	Yeni doğ.
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9238	0,0762	0,5388	0,9874	0,8748	0,6665	0,5009
Aşırı Öğrenme Makinesi (NS:527, AF:radbas, ED:0.5)	0,9130	0,0870	0,5571	0,9713	0,7779	0,6421	0,4747
K-En Yakın Komşu (k=3)	0,8990	0,1010	0,3214	0,9942	0,9018	0,4735	0,3291
Sade Bayes	0,8872	0,1128	0,4461	0,9602	0,6483	0,5272	0,3584

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.11'de 10-kat çapraz geçerleme yöntemi kullanıldığında tüm algoritmalar için elde edilen en iyi doğruluk değerleri verilmiştir. Yöntemler doğruluk değeri en yüksekten en düşüğe doğru olacak şekilde sıralanmıştır.

Tablo 4.11: Doğruluk ölçüsüne göre sınıflandırma modellerinin performans analizi (çapraz geçerleme 10-kat ile).

Yöntem	Doğruluk	Hata	Hassasiyet	Özgünlük	Kesinlik	F Ölç.	Yeni doğ.
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9276	0,0724	0,5493	0,9897	0,8972	0,6798	0,5191
Aşırı Öğrenme Makinesi (NS:386, AF: sig, ED:0.5)	0,9138	0,0862	0,5116	0,9792	0,8290	0,6148	0,4500
K-En Yakın Komşu (k=7)	0,9004	0,0996	0,3220	0,9954	0,9150	0,4747	0,3605
Sade Bayes	0,8868	0,1132	0,4494	0,9593	0,6437	0,5262	0,3582

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.11'e göre doğruluk performans değerleri açısından Destek Vektör Makineleri polynominal çekirdek fonksiyon ile beraber %92,76 doğruluk değeri ile en iyi performansa sahiptir. Destek Vektör Makinelerini sırasıyla %91,38 doğruluk değeri ile Aşırı Öğrenme Makinesi, %90,04 değeri ile K-En Yakın Komşu Algoritması ve %88,68 değeri ile Sade Bayes Algoritması takip etmektedir. Yine Destek Vektör Makineleri yaklaşık %1 fark ile Aşırı Öğrenme Makinelerine göre daha iyi bir performans göstermiştir.

Tablo 4.12'de belirli oranlarda bölme performans geçerleme yöntemi %70-30 oranı ile kullanıldığında tüm algoritmalar için elde edilen en iyi F ölçüsü değerleri verilmiştir. Yöntemler F ölçüsü değeri en yüksekten en düşüğe doğru olacak şekilde sıralanmıştır.

Tablo 4.12'ye göre F ölçüsü performans değerleri açısından Aşırı Öğrenme Makinesi gizli nöron sayısı 584, sinüs aktivasyon fonksiyonu ve 0.5 eşik değer ile beraber %67,63 F ölçüsü değeri elde edilerek en iyi performansa sahip olmuştur. Aşırı Öğrenme Makinelerini sırasıyla %59,06 değeri ile Destek Vektör Makinesi, %54,44 değeri ile Sade Bayes Algoritması ve %50 değeri ile K-En Yakın Komşu Algoritması takip etmektedir.

Tablo 4.12: F ölçüsüne göre sınıflandırma modellerinin performans analizi (%70 Eğitim %30 Test).

Yöntem	**Doğruluk**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
Aşırı Öğrenme Makinesi (NS:584, AF: sin, ED:0.5)	0,9107	0,0893	0,7292	0,9373	0,6306	0,6763	0,5109
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9187	0,0813	0,4583	0,9862	0,8302	0,5906	0,4190
Sade Bayes	0,8939	0,1061	0,4481	0,9674	0,6934	0,5444	0,3740
K-En Yakın Komşu (k=5)	0,9093	0,0907	0,3542	0,9908	0,8500	0,5000	0,3333

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.13'te belirli oranlarda bölme performans geçerleme yöntemi %80-20 oranı ile kullanıldığında tüm algoritmalar için elde edilen en iyi F ölçüsü değerleri verilmiştir. Yöntemler benzer olarak F ölçüsü değeri en yüksekten en düşüğe doğru olacak şekilde sıralanmıştır.

Tablo 4.13: F ölçüsüne göre sınıflandırma modellerinin performans analizi (%80 Eğitim %20 Test).

Yöntem	**Doğr.**	**Hata**	**Has.**	**Özgünlük**	**Kesinlik**	**F Öl.**	**Yeni doğ.**
Aşırı Öğrenme Makinesi (NS:584, AF: sin, ED:0.5)	0,9310	0,0690	0,6489	0,9735	0,7870	0,7113	0,5519
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9270	0,0730	0,5191	0,9885	0,8718	0,6507	0,4823
K-En Yakın Komşu (k=5)	0,9090	0,0910	0,3664	0,9908	0,8571	0,5134	0,3453
Sade Bayes	0,8859	0,1141	0,4113	0,9639	0,6517	0,5043	0,3372

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.13'e göre F ölçüsü performans değerleri açısından Aşırı Öğrenme Makinesi gizli nöron sayısı 584, sinüs aktivasyon fonksiyonu ve 0.5 eşik değer ile beraber %71,13 F ölçüsü değeri ile en iyi performansa sahiptir. Aşırı Öğrenme Makinelerini sırasıyla %65,07 değeri ile Destek Vektör Makinesi, %51,34 değeri ile K-En Yakın Komşu Algoritması ve %50,43 değeri ile Sade Bayes Algoritması takip etmektedir. Aşırı Öğrenme Makinesi ve Destek Vektör Makinelerine

ait değerler incelendiğinde veri setinin ayrım oranlarındaki değişiklik ile beraber her iki algoritma için de F ölçüsünde kayda değer bir artış söz konusudur.

Tablo 4.14'e gelindiğinde ise F ölçüsü için 5-kat çapraz geçerleme yöntemi yönteminin kullanılması sonucu elde edilen bulgular görülmektedir.

Tablo 4.14: F ölçüsüne göre sınıflandırma modellerinin performans analizi (5-kat çapraz geçerleme ile).

Yöntem	**Doğr.**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9238	0,0762	0,5388	0,9874	0,8748	0,6665	0,5009
Aşırı Öğrenme Makinesi (NS:409, AF: sin, ED:0.4)	0,9038	0,0962	0,6677	0,9426	0,6867	0,6619	0,4950
Sade Bayes	0,8872	0,1128	0,4461	0,9602	0,6483	0,5272	0,3584
K-En Yakın Komşu (k=5)	0,8976	0,1024	0,3345	0,9904	0,8539	0,4804	0,3086

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.14'e göre F ölçüsü performans değerleri, doğruluk ölçüsüne göre benzer bir sıralama göstermektedir. Destek Vektör Makineleri analizlerinde polynominal çekirdek fonksiyonun kullanılması ile beraber %66,65 F ölçüsü değerine ulaşılmıştır. Aşırı Öğrenme Makinesi gizli nöron sayısı 409, sinüs aktivasyon fonksiyonu ve 0.4 eşik değer ile beraber %66,19 F ölçüsü değeri ile ikinci sırada yer almaktadır. Aşırı Öğrenme Makinelerini sırasıyla %52,72 değeri ile Sade Bayes Algoritması ve %48,04 değeri ile K-En Yakın Komşu Algoritması takip etmektedir.

Performans değerlendirmesi için F ölçüsü ve performans geçerleme yöntemi olarak 10-kat çapraz geçerleme göz önüne alındığında, uygulanan modellerin karşılaştırmalı sonuçları Tablo 4.15'te verilmiştir.

Tablo 4.15 incelendiğinde %67,98 değeri ile en yüksek F ölçü değeri Destek Vektör Makineleri ile elde edilmiş olup kullanılan çekirdek fonksiyon polynomial çekirdek fonksiyondur. Destek Vektör Makinelerini %67,36 F ölçü değeri ile Aşırı Öğrenme Makinesi takip etmektedir. Diğer iki yöntemin %52,62 ve %47,89 değerleri ile performans olarak oldukça geride kaldığı görülmektedir.

Tablo 4.15: F ölçüsüne göre sınıflandırma modellerinin performans analizi (10-kat çapraz geçerleme ile).

Yöntem	Doğr.	Hata	Has.	Özgünlük	Kesinlik	F Ölç.	Yeni doğ.
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9276	0,0724	0,5493	0,9897	0,8972	0,6798	0,5191
Aşırı Öğrenme Makinesi (NS:444, AF:sig, ED:0.40)	0,9092	0,0908	0,6741	0,9482	0,6872	0,6736	0,5100
Sade Bayes	0,8868	0,1132	0,4494	0,9593	0,6437	0,5262	0,3582
K-En Yakın Komşu (k=5)	0,8982	0,1018	0,3341	0,9909	0,8583	0,4789	0,3596

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Veri seti %70 eğitim-%30 test olarak bölündüğünde ve performans değerlendirme ölçüsü olarak önerilen yeni doğruluk değerine göre değerlendirme yapıldığında tez çalışması kapsamında uygulanan modellerin karşılaştırmalı sonuçları Tablo 4.16'da verilmiştir.

Tablo 4.16: Yeni doğruluk ölçüsüne göre sınıflandırma modellerinin performans analizi (%70 Eğitim %30 Test).

Yöntem	Doğruluk	Hata	Hassasiyet	Özgünlük	Kesinlik	F Ölç.	Yeni doğ.
Aşırı Öğrenme Makinesi (NS:584, AF:sin, ED:0.5)	0,9107	0,0893	0,7292	0,9373	0,6306	0,6763	0,5109
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9187	0,0813	0,4583	0,9862	0,8302	0,5906	0,4190
Sade Bayes	0,8939	0,1061	0,4481	0,9674	0,6934	0,5444	0,3740
K-En Yakın Komşu (k=5)	0,9093	0,0907	0,3542	0,9908	0,8500	0,5000	0,3333

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Eğitim ve test veri setleri belirtilen oranlarda ayrıldığında Aşırı Öğrenme Makinesi 584 gizli nöron sayısı, sinüs aktivasyon fonksiyonu ve 0.5 eşik değeri parametreleri ile %51,09 oranı ile yeni doğruluk performans değerlendirme ölçüsüne göre en iyi performansı sergilemektedir. Aşırı Öğrenme Makinelerini yaklaşık %10'luk bir fark ile Destek Vektör Makineleri izlemektedir. Sade Bayes ve K-En Yakın Komşu Algoritması ise sırasıyla %37,40 ve %33,33 yeni doğruluk değerleri ile diğer bulgularda olduğu gibi daha düşük bir performans sergilemiştir.

Veri setinin %80'lik kısmı modellerin eğitilmesi %20'lik kısmı test için kullanıldığı taktirde önerilen yeni doğruluk performans değerlendirme değerine göre tez çalışması kapsamında uygulanan modellerin karşılaştırmalı sonuçları Tablo 4.17'da verilmiştir.

Tablo 4.17 incelendiğinde yeni doğruluk değeri açısından en iyi performansın %55,19 oranı ile Aşırı Öğrenme Makineleri ile elde edildiği görülmektedir. Bu durumda Aşırı Öğrenme Makinesi parametre değerleri gizli nöron sayısı 584, aktivasyon fonksiyonu sinüs ve eşik değeri 0.5 şeklindedir. İkinci sırada %48,23 yeni doğruluk değeri ile Destek Vektör Makineleri yer alırken, K-En Yakın Komşu Algoritması Sade Bayes yöntemine göre az da olsa daha iyi bir performans sergilemiştir.

Tablo 4.17: Yeni doğruluk ölçüsüne göre sınıflandırma modellerinin performans analizi (%80 Eğitim %20 Test).

Yöntem	**Doğruluk**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
Aşırı Öğrenme Makinesi (NS:584, AF:sin, ED:0.5)	0,9310	0,0690	0,6489	0,9735	0,7870	0,7113	0,5519
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9270	0,0730	0,5191	0,9885	0,8718	0,6507	0,4823
K-En Yakın Komşu (k=5)	0,9090	0,0910	0,3664	0,9908	0,8571	0,5134	0,3453
Sade Bayes	0,8859	0,1141	0,4113	0,9639	0,6517	0,5043	0,3372

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Performans geçerleme yöntemi olarak 5-kat çapraz geçerleme tercih edildiğinde elde edilen yeni doğruluk performans değerlendirme ölçüsü değerlerine göre modellerin karşılaştırmalı sonuçları Tablo 4.18'de verilmiştir.

Tablo 4.18: Yeni doğruluk ölçüsüne göre sınıflandırma modellerinin performans analizi (5-kat çapraz geçerleme).

Yöntem	**Doğr.**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9238	0,0762	0,5388	0,9874	0,8748	0,6665	0,5009
Aşırı Öğrenme Makinesi (NS:409, AF:sin, ED: 0.40)	0,9038	0,0962	0,6677	0,9426	0,6867	0,6619	0,4950
Sade Bayes	0,8872	0,1128	0,4461	0,9602	0,6483	0,5272	0,3584
K-En Yakın Komşu (k=3)	0,8990	0,1010	0,3214	0,9942	0,9018	0,4735	0,3291

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.18'e göre Destek Vektör Makineleri %50,09 yeni doğruluk değeri oranı ile en iyi performansa sahiptir. Aşırı Öğrenme Makineleri ise %1'den az bir farkla ikinci sırada yer almaktadır. Yeni doğruluk değeri olarak %49,50 değerinin elde edildiği Aşırı Öğrenme Makinesi parametre değerleri 409 gizli nöron sayısı, sinüs aktiasyon fonksiyonu, 0.40 eşik değeri şeklinde sıralanmaktadır.

Son olarak performans geçerleme yöntemi olarak 10-kat çapraz geçerleme tercih edildiğinde elde edilen yeni doğruluk performans değerlendirme ölçüsü değerlerine göre modellerin karşılaştırmalı sonuçları Tablo 4.18'de verilmiştir.

Tablo 4.19: Yeni doğruluk ölçüsüne göre sınıflandırma modellerinin performans analizi (10-kat çapraz geçerleme).

Yöntem	**Doğruluk**	**Hata**	**Hassasiyet**	**Özgünlük**	**Kesinlik**	**F Ölç.**	**Yeni doğ.**
Destek Vektör Makinesi (polynomial çekirdek f.)	0,9276	0,0724	0,5493	0,9897	0,8972	0,6798	0,5191
Aşırı Öğrenme Makinesi (NS:444, AF:sig, ED:0.40)	0,9092	0,0908	0,6741	0,9482	0,6872	0,6736	0,5100
K-En Yakın Komşu (k=7)	0,9004	0,0996	0,3220	0,9954	0,9150	0,4747	0,3605
Sade Bayes	0,8868	0,1132	0,4494	0,9593	0,6437	0,5262	0,3582

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Tablo 4.19 incelendiğinde 5-kat çapraz geçerlemede olduğu gibi 10-kat çapraz geçerlemede de en iyi performansın %51,91 yeni doğruluk değeri ile Destek Vektör Makineleri ile elde edildiği görülmektedir. Aşırı Öğrenme Makineleri yöntemi ve 444 nöron sayısı, sigmoid aktivasyon fonksiyonu ve 0.40 eşik değeri parametreleri ile kurulan model sonucu %51 yeni doğruluk değeri elde edilmiştir. K-En Yakın Komşu Algoritmasında küme sayısı 7 olarak seçildiğinde yeni doğruluk değeri %36,05 olurken, Sade Bayes yöntemi ile elde edilen değer %35,82 şeklindedir.

4.7. GÜVEN ARALIKLARINA İLİŞKİN BULGULAR

Aşırı Öğrenme Makinelerinin en iyi parametre değerlerine ait doğruluk oranları için güven aralıkları binom dağılımı göz önünde bulundurularak hesaplanmıştır. Elde edilen sonuçlar Tablo 4.20'de verilmiştir.

Tablo 4.20: Aşırı öğrenme makinesi için güven aralıkları.

Performans G.Y.	**Parametreler***			**Doğruluk**	**Güven Aralıkları**	
	NS	**AF**	**ED**		**Alt Sınır**	**Üst Sınır**
%70 Eğitim %30 Test	368	sig	0.5	0,92	0,9063	0,9337
	584	sin	0.5	0,9107	0,8963	0,9251
%80 Eğitim %20 Test	584	sin	0.5	0,931	0,9153	0,9467
Capraz Geçerleme (5 Kat)	409	sin	0.4	0,9038	0,8956	0,9120
	527	radbas	0.5	0,913	0,9052	0,9208
Capraz Geçerleme (10 Kat)	386	sig	0.5	0,9138	0,9060	0,9216
	444	sig	0.4	0,9092	0,9012	0,9172

* NS: Gizli Katmandaki Nöron Sayısı, AF: Aktivasyon Fonksiyonu, ED: Eşik Değer

Elde edilen tüm bu bulgulara dayalı yorumlara Bölüm 5'te yer verilmiştir.

5. TARTIŞMA VE SONUÇ

Mirze (2013), işletmeyi toplumun gereksinimi olan mal ve hizmetlerin çeşitli kaynakların bir araya getirilmesi ile beraber üretim ve dağıtımına yönelik faaliyetlerin yürütüldüğü örgüt, kurum, kuruluş olarak tanımlamış, işletmenin en temel faaliyetinin ise alıcı ve satıcılar arasındaki alışveriş olduğunu ifade etmiştir. Buradan hareketle bir işletmenin varlığını devam ettirebilmesinin ürün veya hizmet sunduğu müşteriler arasındaki bu alışverişin devamına bağlı olduğu sonucuna ulaşılabilir.

Diğer taraftan benzer faaliyet gösteren işletmeler arasında bir rekabet ortamın varlığı söz konusudur. Hatta küreselleşmenin bir getirisi olarak bu rekabet ortamı küresel bir alana taşınmış, işetlemeler yerel veya kısıtlı sayıdaki uluslararası rakiplerle değil dünya çapında rakipler ve daha zorlu rakabet koşulları ile karşı karşıya kalmıştır. Rekabet, benzer faaliyet alanında çalışan ve aynı hizmet veya ürünü üreten işletmelerin diğer işletmelere göre müşterilerin talep ve beklentilerini göreceli olarak karşılayabilmesi olarak tanımlanırken, rekabette üstünlük sağlayabilmenin yollarından bazıları müşteri memnuniyetini sağlamak, teknolojik yenilik ve gelişmeler ile birlikte hızlı değişime ayak uydurabilmek olarak belirtilmektedir (Saatçioğlu, 2016).

İşletmelerin iki türlü varlığı söz konusu olup bunlar maddi ve maddi olmayan varlıklar şeklinde ifade edilmektedir. Maddi olmayan varlıklar entelektüel sermaye olarak da adlandırılmaktadır. Entelektüel sermayenin temelinde bilgi yatmaktadır. Dolayısıyla bilginin işletmeler için en önemli kaynaklardan biri olduğunu söylemek mümkündür. Bilgi çağı ile beraber bilgi ve bilginin yönetimi işletmelerin uzun dönemli rekabet üstünlüğü yaratma, daha iyi performans sergileme ve de varlığını sürdürebilmesinde oldukça önemli hale gelmiştir (Büyüközkan, 2010; Çebi ve diğ., 2010).

Günümüz dünyası bilgiye giden yolda yapıtaşı olan verinin hemen her alanda üretildiği ve saklandığı bir ortama dönüşmüştür. Bu durumda gelişen teknoloji ile beraber daha düşük maliyetli, daha büyük kapasiteli ve daha kolay işlem yapılabilen depolama alanlarının geliştirilmesi ve yukarıda da bahsi geçtiği üzere bilginin öneminin keşfedilmesinin olduğu söylenebilir. Ancak çok miktarda kayıt altına alınan veri aynı zamanda çok fazla bilgi anlamına da gelmemektedir. Dolayısıyla amaca yönelik olarak bu veri yığınlarının işlenmesi ve bilginin ortaya konması gerekmektedir. Veriden bilgiye giden yolda çeşitli istatistik ve matematiksel

yöntem geliştirilmiştir. Bu yöntemlerden bir tanesi de veri madenciliğidir. Veir madenciliği daha önce de belirtildiği üzere en genel anlamda veri yığını içerisinde gizli kalmış ve değerli olan, daha önce bilinmeyen, amaca yönelik olarak faydalı bilginin ortaya çıkartılması işi, bu iş için gerekli olan yöntemler bütünü şeklinde tanımlanmaktadır.

Veri madenciliği, işletmeler açısından farklı konularda karşılaşılan problemler için çözüm üretmede etkin araçlardan bir tanesi haline gelmiştir. İşletmelerin varlığı için yeni müşterilerin kazanılmasının yanı sıra var olan müşterilerin de elde tutulması, bu amaca yönelik etkin stratejiler geliştirilmesi gerekmektedir. Dolayısıyla müşteri davranışlarının incelenmesi işletmeler için odak konulardan biri haline gelmiştir. Bu amaçla, müşteri verileri ve veri madenciliği yöntemleri kullanarak ayrılma eğilimi olan müşteriler ve ayrılma nedenleri belirlenmekte, bu müşterilerin elde tutulabilmesi için çeşitli stratejik kararlar verilebilmektedir. Bu çalışmalar bir yönetim problemi olarak ele alınan "Müşteri Kayıp Analizi" kapsamında gerçekleştirilmektedir.

Bu tez çalışması kapsamında, müşteri kayıp analizi problemi için kullanılan veri madenciliği yöntemlerinin çözüm için etkinliliğinin irdelenmesi ve kullanılabilecek yeni yöntem önerisinin geliştirilmesi amaçlanmıştır. Literatürde de belirtilen müşteri kayıplarının daha hızlı ve net bir şekilde gözlemlenebilir olması nedeni ile telekom sektörü uygulama alanı olarak seçilmiştir. Bu doğrultuda, literatürün detaylı incelenmesi, sıklıkla kullanılan farklı veri madenciliği algoritmalarının gösterdikleri performanslarının belirlenen veri seti yardımıyla tespit edilmesi, bu problem için kullanılmamış yöntemlerin araştırılması ve sonuç olarak Aşırı Öğrenme Makineleri yönteminin müşteri kayıp analizi probleminin çözümü için kullanılabilirliğinin ve tez çalışmasının malzeme ve yöntem kısmında belirtilen koşullar altında en iyi çözümü üreten parametrelerinin araştırılması ve kullanılan tüm algoritmaların performans kıyaslamasının ortaya konulması çalışmaları gerçekleştirilmiştir.

Aşırı Öğrenme Makineleri, ileri beslemeleri yapay sinir ağlarını temel alan Huang ve diğerleri (2004) tarafından geliştirilen bir yöntemdir. Yöntem farklı alanlarda kullanılabilir olmak ile beraber özellikle parametre belirleme ve işlem süresi gibi konularda diğer yöntemlere göre üstünlük sağlamaktadır. Bu tez çalışmasında ele alınan müşteri kayıp analizi problemi için Türkçe literatürde Aşırı Öğrenme Makinelerinin kullanıldığı bir çalışma yer almamaktadır. Uluslararası alanda ise yine tez çalışması kapsamında kullanılan Tek Gizli Katmanlı Yapay Sinir Ağları Temelli Aşırı Öğrenme Makinesi algoritması ile yapılan bir çalışmaya

rastlanmamıştır. Bu bakımdan tez çalışması Türkçe literatürde bir ilk, uluslararası alanda ise kullanılan yöntem bakımından farklılık göstermektedir.

Çalışma kapsamında sadece Aşırı Öğrenme Makinesinin probleme uygunluğu test edilmemiş, diğer yöntemler ile rekabet edebilecek düzeyde performansa sahip olup olamayacağı noktasında da araştırmaya gidilmiştir. Bu doğrultuda Bölüm 3'te açıklanan modeller ve kısıtlar dâhilinde, en iyi çözümü üreten parametrelerin aranması çalışması gerçekleştirilmiştir. Bu bakımından da tez çalışması diğer çalışmalardan farklılık göstermektedir.

Uygulanan modellerin performanslarının ölçülmesi ve elde edilen sonuçların diğer yöntem ve sonuçlar ile karşılaştırılabilir olması yapılan deneysel çalışmanın sonuçlarının değerlendirilmesi, geçerliliğinin ispatlanabilmesi adına önem taşımaktadır. Veri madenciliği çalışmalarında birçok performans değerlendirme ölçüsü kullanılabilmektedir. Bunlardan bazıları doğruluk oranı, hata oranı, hassasiyet, kesinlik, özgünlük, F ölçüsü şeklinde sıralanabilir. Tez çalışması kapsamında problemin çıkış noktası baz alınarak, problemin çözümünde önemli olanın ayrılma eğilimli müşterilerin belirlenmesi olduğu göz önünde bulundurulmuş ve Bölüm 2'de verilen karışıklık matrisine göre hesaplanan doğruluk değeri için farklı bir yaklaşım belirlenmiştir. Buna göre, gerçekte ayrılma eğilimi olmayan ve modeller ile de ayrılma eğilimli olmadığı tahmin edilebilen müşteri sayıları hesaplamalardan çıkartılarak yeni bir doğruluk ölçüsü ele alınmıştır. Elde edilen yeni doğruluk ölçüsü sonuçları incelendiğinde bu ölçünün performans kıyaslaması bakımından F ölçüsü ile paralel sonuçlar ürettiği görülmüştür. F ölçüsü, duyarlılık ve hassasiyet ölçülerinden elde edilen bir ölçü olup, doğruluk değerine göre daha geniş bir değerlendirmeye olanak sunmaktadır. Bu bakımdan kullanılan yeni doğruluk ölçüsünün, F ölçüsü ile paralel sonuç üretmiş olması olumlu karşılanmıştır.

Bölüm 4'te verilen bulgular doğruluk, F ölçüsü ve yeni doğruluk ölçüsü bakımından Tablo 5.1'de özetlenmeye çalışılmıştır. Her bir performans değerlendirme ölçüsünün en yüksek değeri tabloda yer almaktadır. Bu değerlerin elde edildiği parametre değerleri değişiklik gösterebilir. Yani Aşırı Öğrenme Makinesi için performans geçerleme yöntemi olarak %70 ve %30 oranı kullanıldığında elde edilen en yüksek doğruluk değeri (%92) ile en yüksek F ölçüsü değeri (%67,63) farklı parametrelerden elde edilmiş olabilir. Tablo 5.1'de sadece performans ölçülerinin farklı parametrelerde de olsa ulaşabildikleri en yüksek değerlerin bir araya getirilmesi amaçlanmıştır.

Tablo 5.1: Sınıflandırma modellerinin performans analizi özeti.

Yöntem	Performans Geçerleme Yöntemi	Doğruluk	F Ölçüsü	Yeni Doğruluk
Aşırı Öğrenme Makineleri	%70eğitim-%30test	0,9200	0,6763	0,5109
	%80eğitim-%20test	0,9310	0,7113	0,5519
	5-kat çapraz geçerleme	0,9130	0,6619	0,4950
	10-kat çapraz geçerleme	0,9138	0,6736	0,5100
Destek Vektör Makineleri	%70eğitim-%30test	0,9187	0,5906	0,4190
	%80eğitim-%20test	0,9270	0,6507	0,4823
	5-kat çapraz geçerleme	0,9238	0,6665	0,5009
	10-kat çapraz geçerleme	0,9276	0,6798	0,5191
K-En Yakın Komşu	%70eğitim-%30test	0,9093	0,5000	0,3333
	%80eğitim-%20test	0,9090	0,5134	0,3453
	5-kat çapraz geçerleme	0,8990	0,4804	0,3291
	10-kat çapraz geçerleme	0,9004	0,4789	0,3605
Sade Bayes	%70eğitim-%30test	0,8939	0,5444	0,3740
	%80eğitim-%20test	0,8859	0,5043	0,3372
	5-kat çapraz geçerleme	0,8872	0,5272	0,3584
	10-kat çapraz geçerleme	0,8868	0,5262	0,3582

*Her bir performans değerlendirme ölçüsünün en yüksek değeri tabloda yer almaktadır. Bu değerlerin elde edildiği parametre değerleri değişiklik gösterebilir.

Sonuçlar incelendiğinde,

- Göz önünde bulundurulan doğruluk, F ölçüsü ve yeni doğruluk performans değerlendirme ölçülerine göre belirli oranlarda bölme performans geçerleme yöntemi kullanıldığında tüm yöntemler içerisinde en iyi performans Aşırı Öğrenme Makinesi ile sağlanmıştır.
- Çapraz gerçeleme yöntemi kullanıldığında ise Aşırı Öğrenme Makinesi yerini Destek Vektör Makinelerine bırakmıştır.

Her ne kadar yerini kaybetmiş olsa da oranlar incelendiğinde bu durumda performans göstergeleri arasındaki fark ancak %1 kadardır. Hâlbuki Aşırı Öğrenme Makinelerinin üstün geldiği durumlarda fark oldukça açık olmaktadır.

Yapılan en iyi paremetrenin belirlenmesi işlemi sonucunda ise Aşırı Öğrenme Makinesi performans değerlendirme ölçülerinde önemli derecede artış görülmüştür. Buna göre Aşırı Öğrenme Makinesi için yaklaşık olarak doğruluk oranında %3-5 arasında, F ölçüsü oranında %20-30 arasında, yeni doğruluk ölçüsünde ise yine %21-27 arasında bir artış kaydedilmiştir. Özellikle F ölçüsü ve yeni doğruluk oranındaki bu artış diğer yöntemler ile yapılan performans karşılaştırmasında önem teşkil etmektedir.

Dengesiz sınıf dağılımı konusu göz önüne alındığında 5000 kayıt içerisinde her ne kadar "ayrılma eğilimli" müşterilerin oranı az da olsa hassasiyet değerleri göz önünde bulundurulduğunda gerçekte "ayrılma eğilimli" olan müşteriler için Aşırı Öğrenme Makinesi ile yapılan tahminlerde doğru tahmin etme oranı %70'lerin üzerini görmüştür. Bir başka açıdan kesinlik ölçüsü incelendiğinde "ayrılma eğilimli" olarak tahmin edilen müşterilerin %80'in üzerinde gerçekte de "ayrılma eğilimli" olduğu görülmüştür. Dolayısıyla sınıf dağılım oranı dengeli hale getirilirse bu oranlarda daha da iyileşme olacağı gözlenebilecektir.

Aşırı Öğrenme Makinesi Algoritması araştırmacılar tarafından özellikle parametrelerde kişinin belirleyiciliğinin mümkün olduğunca azaltılması ve işlem süresinin Sinir Ağları, Destek Vektör Makineleri gibi üst düzey yöntemlere göre daha kısa olması amaçlanmıştır. Bu bağlamda yöntem incelendiğinde modellerin oluşturulması ve kullanılan parametrelerde kişiye özgü müdahalenin diğer yöntemlere göre daha az olması bakımından elde edilen sonuçlar da göz önünde bulundurulduğunda oldukça avantajlı olduğu söylenebilir. Analizler gerçekleştirilirken farklı bilgisayarlar kullanıldığından süre konusunda bir çalışma gerçekleştirilmemiştir. Ancak genel gözlem bakımından analizlerde geçirilen süreler performans geçerleme yöntemi ve nöron sayısına göre değişiklik göstermiştir. Ara katmandaki gizli nöron sayısı arttıkça ve performans geçerleme yöntemi olarak çapraz geçerleme kullanılırsa analiz sürelerinde uzama olduğu gözlenmiştir.

Literatür araştırması ve elde edilen sonuçlar işletmeler açısından incelenecek olursa; stratejik kararlar alma noktasında veri, bilgi ve veri madenciliğinin bunun yanı sıra işletmelerin önemli bir kaynağı olan veri kaynağının etkin bir şekilde kullanılmasının önemini yeniden destekler

niteliktedir. Bir işletmeye ait müşterilerin davranışlarının ayrılma eğilimi noktasında %80-%90 doğruluk oranında belirlenebilir olması, kaybedilme riski taşıyan müşterilere yönelik geliştirmesi gereken çeşitli stratejiler konusunda oldukça önemli bir kaynaktır. Hem veri madenciliği yöntemleri ile gerçekleştirilen analizlerin sonuçları hem de işletmelerin sahip olduğu deneyim kullanılarak çok daha etkin stratejiler belirlenebilir. Bu çalışma göstermektedir ki bilgi işletmeye artı (kar) olarak geri dönebilmektedir. Günümüzde birçok müşteri hizmetleri bölümünde benzer yöntemlerin kullanıldığı, hatta bu analiz çalışmaları için alt ekipler oluşturulduğu bilinmektedir. Ancak birçok işletmede yeniliğe kapalı olarak alışılagelmiş yöntemler kullanılmaktadır. Hâlbuki literatürde her geçen gün veri madenciliği yöntemlerinin daha iyi performanslar ile tahmin yapabilmesi için çalışmalar sürdürülmekte, etkin sonuçlar da elde edilmektedir. İşletmelerin, veri tabanlarında yer alan veri setinin özelliklerini ve literatürde yer alan çalışmaları da göz önünde bulundurup uyguladıkları yöntemler bakımından kendilerini yenilemeleri önerilmektedir.

Her ne kadar telekom ve banka gibi sektörlerde müşteri kayıp analizi çalışmalarının yaygın olduğu görülse de gelişen teknoloji ile beraber özellikle elektronik satış veya üyelik ile işlem yapan farklı sektörlerde de üyelikler üzerinden müşteri davranışları takip edilir olduğundan yapılan çalışmaların ilerleyen dönemde telekom ağırlıklı olmaktan çıkacağı düşünülmektedir. Dolayısıyla gelecek çalışmalarda önerilen yöntem için farklı bir sektör uygulama alanı olarak seçilebilir.

Tez çalışması için veri temin etme konusunda, görüşülen Türkiye'de Telekom sektöründe hizmet veren üç işletmeden olumlu geri dönüş alınamamıştır. Tez çalışması herkese açık bir veri tabanından veri seti elde edilerek gerçekleştirilmiştir. Bu tez çalışmasının orijinalliğini kısıtlayan bir durum olsa dahi elde edilen sonuçların başka araştırmacılar tarafından da denenebilmesine ve benzer veri seti ile yapılan çalışmalar ile kıyaslanabilmesine olanak sağlamaktadır. Veri temini konusunda yaşanan problemde her ne kadar dönemin etkisi de olsa sektörel tarafın sadece proje bazlı değil akademik çalışmaları da destekleyecek şekilde üniversite işbirliğine açık olması beklenmektedir. Orijinal bir veri seti kullanılması durumunda nitelik seçimi analizleri de çalışmaya dahil edilerek çalışma genişletilebilir, sektöre yönelik olarak çözüm önerileri getirilebilir.

Bu tez çalışması kapsamında parametre konusunda özellikle sayısal nitelikler için belirlenen aralıklarda tamsayı kısıtı altında aramalar gerçekleştirilmiştir. Sürekli değerler altında en iyi

parametrenin belirlenebilmesi için bir optimizasyon yöntemi vasıtasıyla analizler gerçekleştirilebir. Bu doğrultuda bir hibrit model önerisi de getirilebilir.

Tez çalışması kapsamında kullanılan veri setinin sınıf niteliğinin dengesiz bir dağılım gösterdiği göz önünde bulundurulursa çeşitli örnekleme yöntemi ile dengesiz dağılım dengelenerek analizler yinelenebilir. Bu noktada özellikle F ölçüsü ve yeni doğruluk oranında artış beklenmektedir. Bu çalışmanın bir ksıtı olarak dengesiz sınıf nitelik ayrımı nedeniyle doğruluk ölçüsü algoritmaların başarısı diğer ölçüler ise algoritmaların karşılaştırılması açısından yararlanılabilir.

Analiz çalışmaları için R programlama dilinden ve kullanılabilir paketlerden yararlanılmıştır. Ancak Aşırı Öğrenme Makineleri için geliştirilen paketler kısıtlı sayıdadır. Dolayısıyla Aşırı Öğrenme Makinesi yönteminin farklı algoritmaları için paketler geliştirilerek bu alanda çalışanlar için literatüre kazandırılabilir.

Sonuç olarak, tez çalışması İşletme Enformatiği alanında yeni bir yaklaşımı da içeren veri madenciliği konusunda hazırlanmıştır. Aşırı Öğrenme Makinesi yöntemi uygulanmış, literatürde etkin olarak bahsi geçen farklı yöntemler ile performansı karşılaştırılarak performansının daha iyi olabileceğine yönelik bir çalışma ortaya konmuştur. Bu tez çalışmasının, çeşitli yönleri ile özgünlüğü ve ileriye yönelik çalışmalar için getirdiği öneriler bakımından literatüre katkısı olacağı düşünülmektedir.

KAYNAKLAR

Abbasimehr, H. Alizadeh, S., 2013, A novel genetic algorithm based method for building accurate and comprehensible churn prediction models, *International Journal of Research in Industrial Engineering*, 2(4), 1.

Abbasimehr, H., Setak, M. Soroor, J., 2013, A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques, *International Journal of Production Research*, 51(4), 1279–1294.

Abbasimehr, H., Setak, M. Tarokh, M.J., 2011, A neuro-fuzzy classifier for customer churn prediction, *International Journal of Computer Applications*, 19(8), 35–41.

Ackoff, R.L., 1989, From data to wisdom, *Journal of Applies Systems Analysis*, 16(1989), 3-9.

Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., Ghatasheh, N., 2014, Predicting customer churn in telecom industry using multilayer preceptron neural networks: modeling and analysis, *Life Science Journal*, 11(3), 75-81.

Aggarwal, C.C., 2015, *Data mining the textbook*, Springer International Publishing, New York, ISBN: 978-3-319-14141-1.

Aggarwal, C., Yu, S., 2005, An effective and efficient algorithm for high-dimensional outlier detection, *The International Journal on Very Large Data Bases*, 14(2), 211-221.

Agrawal, R., Imieliński, T., Swami, A., 1993, Mining association rules between sets of items in large databases, *ACM Sigmod Record*, 22(2), 207-216.

Ahmed, S.R., 2004, Applications of data mining in retail business, *Proceedings International Conference on Information Technology: Coding and Computing ITCC*, 5-7 Nisan 2004 Las Vegas USA, Emerald Group Publishing Limited, 455-459.

Ahn, J.H., Han, S.P., Lee, Y.S., 2006, Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry, *Telecommunications Policy*, 30(10–11), 552–568.

Akpınar, H., 2014, *Data-Veri Analizi, Veri Madenciliği*, Papatya Yayıncılık, İstanbul, ISBN: 978-605-4220-816.

Al-Shboul, B., Faris, H., Ghatasheh, N., 2015, Initializing genetic programming using fuzzy clustering and its application in churn prediction in the telecom industry, *Malaysian Journal of Computer Science*, 28(3).

Alçin, Ö.F., Şengür, A., İnce, M.C., 2015, İleri-geri takip algoritması tabanlı seyrek aşırı öğrenme makinesi, *Journal of the Faculty of Engineering & Architecture of Gazi University*, 30(1), 126-132.

Almana, A.M., Aksoy, M.S., Alzahrani, R., 2014, A survey on data mining techniques in customer churn analysis for telecom industry, *Journal of Engineering Research and Applications*, 4(5), 165-171.

Anil Kumar, D., Ravi, V., 2008, Predicting credit card customer churn in banks using data mining, *International Journal of Data Analysis Techniques and Strategies*, 1(1), 4-28.

Anokhin, P., Motro, A., 2001, Data integration: inconsistency detection and resolution based on source properties, *Proceedings of International Workshop on Foundations of Models for Information Integration FMII*, 16-18 Eylül 2001 Viterbo İtalya.

Au, W.H., Chan, K.C., Yao, X., 2003, A novel evolutionary data mining algorithm with applications to churn prediction, *IEEE Transactions on Evolutionary Computation*, 7(6), 532-545.

Azevedo, A.I.R.L., Santos, M.F., 2008, KDD, SEMMA and CRISP-DM: A parallel overview. http://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf, [Ziyaret Tarihi: 6 Aralık 2017].

Bakar, Z.A., Mohemad, R., Ahmad, A., Deris, M.M., 2006, A comparative study for outlier detection techniques in data mining, *IEEE Conference on Cybernetics and Intelligent Systems*, 7-9 Haziran 2006 *Bangkok Tayland*, IEEE Xplore Digital Library, 1-6.

Balaban, M.E., Kartal, E., 2015, *Veri madenciliği ve makine öğrenmesi temel algoritmaları ve r dili ile uygulamaları*, Çağlayan Kitabevi, İstanbul, ISBN:978-975-436-089-9.

Ballings, M., Van den Poel, D., 2012, Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications*, 39(18), 13517–13522.

Berry, M.J.A., Linoff, G.S., 2000, *Mastering Data Mining*, John Wiley & Sons, New York.

Bin, L., Peiji, S., Juan, L., 2007, Customer churn prediction based on the decision tree in personal handyphone system service, *International Conference on Service Systems and Service Management*, 9-11 Haziran 2007 Chengdu China, IEEE Xplore Digital Library, 1–5.

Boisot, M. H., 1998, *Knowledge assets: securing competitive advantage in the information economy*, OUP Oxford, New York, USA, ISBN: 978-0198296072.

Bose, I., Chen, X., 2009, Hybrid models using unsupervised clustering for prediction of customer churn. *Journal of Organizational Computing and Electronic Commerce*, 19(2), 133–151.

Bradley, P.S., Fayyad, U.M., Mangasarian, O. L., 1999, Mathematical programming for data mining: Formulations and challenges, *INFORMS Journal on Computing*, 11(3), 217-238.

Brown, M.L., Kros, J. F., 2003, Data mining and the impact of missing data, *Industrial Management & Data Systems*, 103(8), 611-621.

Buckinx, W., Van den Poel, D., 2005, Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting, *European Journal Of Operational Research*, 164 (2005), 252–268.

Burez, J., Van den Poel, D., 2007, CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services, *Expert Systems with Applications*, 32(2), 277–288.

Burez, J., Van den Poel, D., 2008, Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department, *Expert Systems with Applications*, 35(1–2), 497–514.

Burez, J., Van den Poel, D., 2009, Handling class imbalance in customer churn prediction, *Expert Systems with Applications*, 36(3), 4626–4636.

Burges, C. J., 1998, A tutorial on support vector machines for pattern recognition, *Data Mining And Knowledge Discovery*, 2(2), 121-167.

Büyüközkan, G., 2010, Entelektüel sermaye ve yönetimi, Bilgi yönetimi ve uygulamaları, İçerisinde: Dinçmen, M. (ed.), Papatya Yayıncılık Eğitim, İstanbul, ISBN:978-605-4220-15-1, 71-98.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A., 1998, *Discovering Data Mining: From Concept To Implementation*. Prentice-Hall, Inc., ISBN: 978-0137439805.

Cebi, F., Aydin, O.F., Gozlu, S., 2010, Benefits of knowledge management in banking, *Journal of Transnational Management*, 15(4), 308-321.

Chaffey, D., White, G., 2005, *Business information management: improving performance using information systems*, Prentice Hall, Harlow, England, ISBN: 9780273711797.

Chandar, M., Laha, A., Krishna, P., 2006, Modeling churn behavior of bank customers using predictive data mining techniques, *National Conference on Soft Computing Techniques For Engineering Applications*, 24-26 Mart 2006 Rourkela India, 24-26.

Chaudhary, P., 2015, Data mining system, functionalities and applications: A radical review, *International Journal of Innovations in Engineering and Technology*, 5(2), 449-455.

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A.V., Rong, X., 2015, Data mining for the ınternet of things: literature review and challenges, *International Journal of Distributed Sensor Networks*, 2015(31047), 1-14.

Chen, M.S., Han, J., Yu, P.S., 1996, Data mining: An overview from a database perspective, *IEEE Transactions on Knowledge and Data Engineering*, 8 (6), 866-883.

Chen, Z.Y., Fan, Z.P., 2012, Distributed customer behavior prediction using multiplex data: A collaborative MK-SVM approach, *Knowledge-Based Systems*, 35, 111–119.

Chen, Z.Y., Fan, Z.P., Sun, M., 2012, A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data, *European Journal of Operational Research*, 223(2), 461–472.

Chiang, D.A., Wang, Y.F., Lee, S.L., Lin, C.J., 2003, Goal-oriented sequential pattern for network banking churn analysis, *Expert Systems with Applications*, 25(3), 293-302.

Cho, J.H., Lee, D.J., Chun, M.G., 2007, Parameter optimization of extreme learning machine using bacterial foraging algorithm, *Journal of Korean Institute of Intelligent Systems*, 17(6), 807-812.

Chu, B.H., Tsai, M.S., Ho, C.S., 2007, Toward a hybrid data mining model for customer retention, *Knowledge-Based Systems*, 20(8), 703-718.

Chung, H.M., Gray, P., 1999, Special section: Data mining, *Journal of Management Information Systems*, 16(1), 11-16.

Cios, K. J., 2000, From the guest editor medical data mining and knowledge discovery, *IEEE Engineering in Medicine and Biology Magazine*, 19(4), 15-16.

Coenen, F., 2011, Data mining: Past, present and future, *The Knowledge Engineering Review*, 26(1), 25-29.

Cortes, C., Vapnik, V., 1995, Support-vector networks, *Machine Learning*, 20(3), 273-297.

Coussement, K., Benoit, D.F., Van den Poel, D., 2010, Improved marketing decision making in a customer churn prediction context using generalized additive models, *Expert Systems with Applications*, 37(3), 2132–2143.

Coussement, K., De Bock, K.W., 2013, Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning, *Journal of Business Research*, 66(9), 1629–1636.

Coussement, K., Lessmann, S., Verstraeten, G., 2017, A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry, *Decision Support Systems*, 95, 27-36.

Coussement, K., Van den Poel, D., 2008a, Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, *Expert Systems with Applications*, 34(1), 313–327.

Coussement, K., Van den Poel, D., 2008b, Integrating the voice of customers through call center emails into a decision support system for churn prediction, *Information & Management*, 45(3), 164–174.

Coussement, K., Van den Poel, D., 2009, Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers, *Expert Systems with Applications*, 36(3), 6127–6134.

Çelik, S., 2017, Web günlük dosyalarının analizi için web kullanım madenciliğinin uygulanması, *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 46(1), 62-75.

Dasu, T., Johnson, T., 2003, *Exploratory data mining and data cleaning*, John Wiley & Sons, Canada, ISBN:0-471-26851-8.

Datta, P., Masand, B., Mani, D.R., Li, B., 2000, Automated cellular modeling and prediction on a large scale, *Artificial Intelligence Review*, 14(6), 485–502.

Davenport, T. H., Prusak, L., 1998, *Working knowledge: How organizations manage what they know*, Harvard Business School Press, Boston.

De Bock, K.W., Van den Poel, D., 2011, An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction, *Expert Systems with Applications*, 38(10), 12293–12301.

De Bock, K.W., Van den Poel, D., 2012, Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models, *Expert Systems with Applications*, 39(8), 6816–6826.

Delen, D., 2010, A comparative analysis of machine learning techniques for student retention management, *Decision Support Systems*, 49(4), 498-506.

Demir, F.O., Kırdar, Y., 2007, Müşteri ilişkileri yönetimi: CRM, *Review of Social, Economic & Business Studies*, 8, 293-308.

Dierkes, T., Bichler, M., Krishnan, R., 2011, Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with Markov logic networks, *Decision Support Systems*, 51(3), 361–371.

Dixon, B., Candade, N., 2008, Multispectral landuse classification using neural networks and support vector machines: One or the other, or both?, *International Journal of Remote Sensing*, 29(4), 1185-1206.

Dolatabadi, S.H., Keynia, F., 2017, Designing of customer and employee churn prediction model based on data mining method and neural predictor, *2nd International Conference on Computer and Communication Systems (ICCCS)*, 11-14 Temmuz 2017 Krakow Poland, IEEE Xplore Digital Library, 74-77.

Dragulescu, A.A., 2014, *xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files*, R package version 0.5.7, https://CRAN.R-project.org/package=xlsx, [Ziyaret Tarihi: 6 Aralık 2017].

Dunham, M.H., 2003, *Data mining ıntroductory and advanced topics*, Pearson Education Inc., USA, ISBN:0-13-088892-3.

Efron, B., Tibshirani, R.J., 1994, *An introduction to the bootstrap*, Chapman & Hall/CRC press, Florida, ISBN: 0-412-04231-2.

Ertuğrul, Ö.F., Tağluk, M.E., Kaya, Y., Tekin, R., 2013, EMG signal classification by extreme learning machine, *21st Signal Processing and Communications Applications Conference (SIU)*, 24-26 Nisan 2013 Kuzey Kıbrıs Türkiye, IEEE Xplore Digital Library, 1-4.

Farquad, M.A.H., Ravi, V., Raju, S.B., 2014, Churn prediction using comprehensible support vector machine: An analytical CRM application, *Applied Soft Computing*, 19, 31–40.

Fasanghari, M., Keramati, A., 2011, Customer churn prediction using local linear model tree for iranian telecommunication companies, *Journal of Industrial Engineering*, 45, 25–37.

Fathian, M., Hoseinpoor, Y., Minaei-Bidgoli, B., 2016, Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods, *Kybernetes*, 45(5), 732-743.

Fawcett, T., 2006, An introduction to ROC analysis, *Pattern Recognition Letters*, 27(8), 861-874.

Fayyad, U.M., 1996, Data mining and knowledge discovery: Making sense out of data, *IEEE Expert: Intelligent Systems and Their Applications*, 11 (5), 20-25.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., 1996, Knowledge discovery and data mining: Towards a unifying framework, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 2-4 Ağustos 1996 Portland Oreon, AAAI Press, ISBN: 978-1-57735-004-0, 82-88.

Forhad, N., Hussain, M.S., Rahman, R.M., 2014, Churn analysis: Predicting churners, *Ninth International Conference on Digital Information Management (ICDIM)*, 29 Eylül-1 Ekim 2014 Phitsanulok Thailand, IEEE Xplore Digital Library, 237-241.

Freitas, A.A., 2002, *Data mining and knowledge discovery with evolutionary algorithms*, Springer, Germany, ISBN: 3-540-43331-7.

Friedl, M.A., Brodley, C.E., 1997, Decision tree classification of land cover from remotely sensed data, *Remote Sensing of Environment*, 61(3), 399-409.

Friedman, J.H., 1997, *Data mining and statistics: What's the connection?*, http://www-stat.stanford.edu/~jhf/ftp/dm-stat.pdf [Ziyaret Tarihi: 15 Kasım 2017].

Gajowniczek, K., Orłowski, A., Ząbkowski, T., 2016, Entropy based trees to support decision making for customer churn management, *Acta Physica Polonica A.*, 129(5), 971-979.

García, S., Fernández, A., Luengo, J., Herrera, F., 2009, A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability, *Soft Computing*, 13(10), 959.

Gardner, M.W., Dorling, S.R., 1998, Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences, *Atmospheric Environment*, 32(14), 2627-2636.

Gatnar, E., Rozmus, D., 2005, *Data mining—the polish experience*, Innovations in Classification, Data Science, and Information Systems, İçerisinde: Baier, D., Wernecke, K.D. (ed.), Springer, Germany, ISBN: 3-540-23221-4, 217-223.

Gezer, M., 2016, *Veri madenciliğinde verinin ön işlenmesi*, R ile veri madenciliği uygulamaları, İçerisinde: Balaban, M.E., Kartal, E. (ed.), Çağlayan Kitabevi, İstanbul, ISBN:978-975-436-093-6.

Giraud-Carrier, C., Povel, O., 2003, Characterising data mining software, *Intelligent Data Analysis*, 7(3), 181-192.

Gorunescu, F., 2011, *Data Mining: Concepts, models and techniques,* Springer, ISBN:978-3-642-19720-8.

Gosso, A., 2012, *elmNN: Implementation of ELM (Extreme Learning Machine) algorithm for SLFN (Single Hidden Layer Feedforward Neural Networks),* R package version 1.0, https://CRAN.R-project.org/package=elmNN, [Ziyaret Tarihi: 15 Kasım 2017].

Gower, J.C., 1971, A general coefficient of similarity and some of its properties, *Biometrics*, 27(4), 857-871.

Grover, T., Davenport, T.H., 2001, General perspectives on knowledge management: Fostering a research agenda, *Journal of Management Information Systems*, 18(1), 5-21.

Grzymala-Busse, J.W., Hu, M., 2000, A comparison of several approaches to missing attribute values in data mining, *International Conference on Rough Sets and Current Trends in Computing*, 16-19 Ekim 2000 Canada, Springer, ISBN: 978-3-540-45554-7, 378-385.

Günther, C.C., Tvete, I.F., Aas, K., Sandnes, G.I., Borgan, Ø., 2014, Modelling and predicting customer churn from an insurance company, *Scandinavian Actuarial Journal*, 2014(1), 58–71.

Gür Ali, Ö., Arıtürk, U., 2014, Dynamic churn prediction framework with more effective use of rare event data: The case of private banking, *Expert Systems with Applications*, 41(17), 7889–7903.

Hadden, J., Tiwari, A., Roy, R., Ruta, D., 2006a, Churn prediction: Does technology matter, *International Journal of Intelligent Technology*, 1(2), 104–110.

Hadden, J., Tiwari, A., Roy, R., Ruta, D., 2006b, Churn prediction using complaints data, *World Academy of Science, Engineering and Technology*, 13(2006), 158-163.

Hadden, J., Tiwari, A., Roy, R., Ruta, D., 2007, Computer assisted customer churn management: State-of-the-art and future trends, *Computers & Operations Research*, 34(10), 2902-2917.

Hall, M.A., Holmes, G., 2003, Benchmarking attribute selection techniques for discrete class data mining, *IEEE Transactions on Knowledge and Data Engineering*, 15(6), 1437-1447.

Han, J., Pei, J., Kamber, M., 2012, *Data mining: Concepts and techniques*, Elsevier, USA, ISBN: 978-0-12-381479-1.

Hand, D. J., 2007, Principles of data mining, *Drug Safety*, 30(7), 621-622.

He, J., 2009, Advances in data mining: History and future, *Third International Symposium on Intelligent Information Technology Application*, 21-22 Kasım 2009 Nanchang China, IEEE, ISBN: 978-0-7695-3859-4, 634-636.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., Scholkopf, B., 1998, Support vector machines, *IEEE Intelligent Systems and Their Applications*, 13(4), 18-28.

Huang, B., Buckley, B., Kechadi, T.M., 2010, Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications, *Expert Systems with Applications*, 37(5), 3638–3646.

Huang, B., Kechadi, M.T., Buckley, B., 2012, Customer churn prediction in telecommunications, *Expert Systems with Applications*, 39(1), 1414–1425.

Huang, B.Q., Kechadi, T.M., Buckley, B., Kiernan, G., Keogh, E., Rashid, T., 2010, A new feature set with new window techniques for customer churn prediction in land-line telecommunications, *Expert Systems with Applications*, 37(5), 3657–3665.

Huang, G.B., Zhu, Q.Y., Siew, C.K., 2004, Extreme learning machine: A new learning scheme of feedforward neural networks, *Proceedings IEEE International Joint Conference on Neural Networks*, 25-29 Temmuz 2004 Budapest Hungary, IEEE, ISBN: 0-7803-8359-1, 985-990.

Huang, Y., Huang, B.Q., Kechadi, M.T., 2010, A new filter feature selection approach for customer churn prediction in telecommunications, *International Conference on Industrial Engineering and Engineering Management (IEEM)*, 7-10 Aralık 2010 Macao China, IEEE Xplore Digital Library, 338-342.

Huang, Y., Kechadi, T., 2013, An effective hybrid learning system for telecommunication churn prediction, *Expert Systems with Applications*, 40(14), 5635–5647.

Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., Faris, H., 2015, Hybrid data mining models for predicting customer churn, *International Journal of Communications, Network and System Sciences*, 8(05), 91.

Hung, S.Y., Yen, D.C., Wang, H.Y., 2006, Applying data mining to telecom churn management, *Expert Systems with Applications*, 31(3), 515–524.

Idris, A., Khan, A., Lee, Y.S. (2013). Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. *Applied Intelligence*, 39(3), 659–672.

Idris, A., Rizwan, M., Khan, A., 2012, Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies, *Computers & Electrical Engineering*, 38(6), 1808–1819.

Inmon, W.H., 2000, What is a data warehouse?, https://www.business.auc.dk/oekostyr/file/What_is_a_Data_Warehouse.pdf, [Ziyaret Tarihi: 9 Mayıs 2012].

Ismail, M.R., Awang, M.K., Rahman, M.N.A., Makhtar, M., 2015, a multi-layer perceptron approach for customer churn prediction, *International Journal of Multimedia and Ubiquitous Engineering*, 10(7), 213–222.

Jin, R., Breitbart, Y., Muoh, C., 2007, Data discretization unification, *Seventh IEEE International Conference on Data Mining (ICDM)*, 28-31 Ekim 2007 Omaha Nebreska, IEEE Conference Publications, 183-192.

Jones, T.O., Sasser, W.E., 1995, Why satisfied customers defect, *Harvard Business Review*, 73(6), 88.

Kantardzic, M., 2011, *Data mining concepts, models, methods and algorithms*, John Wiley & Sons Inc., New Jersey, ISBN: 978-0-470-89045-5.

Kaur, M., Singh, K., Sharma, N., 2013, Data mining as a tool to predict the churn behaviour among ındian bank customers, *International Journal on Recent and Innovation trends in Computing and Communication*, 1(9), 720-725.

Keremati, A., Ardabili, S.M.S., 2011, Churn analysis for an ıranian mobile operatör, *Telecommunications Policy*, 35 (2011), 344–356.

Keramati, A., Ghaneei, H., Mirmohammadi, S.M., 2016, Developing a prediction model for customer churn from electronic banking services using data mining, *Financial Innovation*, 2(1), 10.

Keremati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., Abbas, U., 2014, Improved churn prediction in telecommunication industry using data mining techniques, *Applied Soft Computing*. 24 (2014), 994–1012.

Khan, A.A., Jamwal, S., Sepehri, M.M., 2010, Applying data mining to customer churn prediction in an internet service provider, *International Journal of Computer Applications*, 9(7), 8–14.

Kim, J. H., 2009, Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap, *Computational Statistics & Data Analysis*, 53(11), 3735-3745.

Kim, K., Jun, C.H., Lee, J., 2014, Improved churn prediction in telecommunication industry by analyzing a large network, Expert *Systems with Applications*, 41(15), 6575–6584.

Kim, K., Lee, J., 2012, Sequential manifold learning for efficient churn prediction, *Expert Systems with Applications*, 39(18), 13328–13337.

Kim, M.K., Park, M.C., Jeong, D.H., 2004, The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services, *Telecommunications Policy*, 28(2), 145-159.

Kirui, C., Hong, L., Cheruiyot, W., Kirui, H., 2013, Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining, *IJCSI International Journal of Computer Science Issues*, 10(2), 1694–784.

Kisioglu, P., Topcu, Y. I., 2011, Applying bayesian belief network approach to customer churn analysis: A case study on the telecom industry of Turkey, *Expert Systems with Applications*, 38(6), 7151–7157.

Kock, N.F., McQueen, R.J., Corner, J.L., 1997, The nature of data, information and knowledge exchanges in business processes: Implications for process improvement and organizational learning, *The Learning Organization*, 4(2), 70-80.

Koçoğlu, F.Ö., 2012, *Veri Madenciliğinde Veri Ayrıklaştırma Yöntemlerinin Karşılaştırılması ve Bir Uygulama,* Yüksek Lisans Tezi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü.

Koçoğlu, F.Ö., Özcan, T., 2016, *Veri madenciliğinde kümeleme algoritmaları ile müşteri segmentasyonu*, İçerisinde: Balaban, M.E., Kartal, E. (ed.), Çağlayan Kitabevi, İstanbul, ISBN:978-975-436-093-6.

Koçoğlu, F.Ö., Özcan, T., Baray, Ş.A., 2016, Veri Madenciliğinde Ayrılan Müşteri Analizi Problemi Üzerine Bir Literatür Araştırması, Uluslararası Katılımlı Üretim Araştırmaları Sempozyumu Bildiri Kitabı, 12-14 Ekim 2016 İstanbul Türkıye, pp. 868 – 874.

Kohavi, R., 1995, A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings IJCAI-95*, 20-25 Ağustos 1995 Montreal Quebec, Morgan Kaufmann, Los Altos, CA, 14(2), 1137-1145.

Korukonda, A. R., 2007, Technique without theory or theory from technique? An examination of practical, philosophical, and foundational issues in data mining. *AI & Society*, 21(3), 347-355.

Kovalerchuk, B., Vityaev, E., 2000, *Data mining in finance: Advances in relational and hybrid methods*, Kluwer Academic Publishers, New York, ISBN:0-792-37804-0.

Kraljević, G., Gotovac, S., 2010, Modeling data mining applications for prediction of prepaid churn in telecommunication services, *AUTOMATIKA: Časopis Za Automatiku, Mjerenje, Elektroniku, Računarstvo I Komunikacije*, 51(3), 275–283.

Kuhn, M. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt., 2016, *caret: Classification and Regression Training*. R package version 6.0-73, https://CRAN.R-project.org/package=caret, [Ziyaret Tarihi: 15 Kasım 2017].

Kurgan, L.A., Musilek, P., 2006, A survey of knowledge discovery and data mining process models, *The Knowledge Engineering Review*, 21(1), 1-24.

Larose, D.T., 2005, *Discovering knowledge in data: an introduction to data mining*, John Wiley & Sons, Canada, ISBN: 0-471-66657-2.

Lazarov, V., Capota, M., 2007, *Churn prediction*, Business Analytics Course, TUM. Computer Science, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.7201&rep=rep1&type=pdf., [Ziyaret Tarihi: 6 Aralık 2017].

Lee, E.B., Kim, J., Lee, S.G., 2017, Predicting customer churn in mobile industry using data mining technology, *Industrial Management & Data Systems*, 117(1), 90-109.

Lee, Y., Lee, C.K., 2003, Classification of multiple cancer types by multicategory support vector machines using gene expression data, *Bioinformatics*, 19(9), 1132-1139.

Leung, K.M., 2007, Naive bayesian classifier, https://pdfs.semanticscholar.org/4632/2a2c113db0e28dd1b37aeabd5668adf77de6.pdf, [Ziyaret Tarihi: 6 Aralık 2017].

Lima, E., Mues, C., Baesens, B., 2011, Monitoring and backtesting churn models, *Expert Systems with Applications*, 38(1), 975–982.

Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2011, iDNA-Prot: Identification of DNA binding proteins using random forest with grey model, *PloS One*, 6(9), e24756.

Lodhi, O.S., 2012, Most commonly used techniques in data mining, *International Journal of Advances in Engineering Research*, 4(III).

Lu, N., Lin, H., Lu, J., Zhang, G., 2014, A customer churn prediction model in telecom industry using boosting, *IEEE Transactions on Industrial Informatics*, 10(2), 1659-1665.

Mammone, A., Turchi, M., Cristianini, N., 2009, Support vector machines, *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3), 283-289.

Mannila, H., 1996, Data mining: Machine learning, statistics, and databases, *Proceedings of Eighth International Conference on Scientific and Statistical Database Systems*, 18-20 Haziran1996 Washington USA, IEEE, ISBN: 0-8186-7264-1, 2-9.

Mărginean, F.A., 2003, Computational aspects of data mining, *International Conference on Computational Science and Its Applications*, 18-21 Mayıs 2003 Montreal Canada, Springer, ISBN: 978-3-540-44843-3, 614-622.

Mariscal, G., Marban, O., Fernandez, C., 2010, A survey of data mining and knowledge discovery process models and methodologies, *The Knowledge Engineering Review*, 25(2), 137-166.

Masand, B., Datta, P., Mani, D.R., Li, B., 1999, CHAMP: A prototype for automated cellular churn prediction, *Data Mining and Knowledge Discovery*, 3(2), 219–225.

Matloff, N., 2011, *The art of R programming: A tour of statistical software design*, No Starch Press, ISBN: 978-1-59327-384-2.

McCarty, J.A., Hastak, M., 2007, Segmentation approaches in data-mining: A comparison of RFM, CHAID and logistic regression, *Journal of Business Research*, 60(6), 656-662.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch,F., 2015, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly:E1071), TU Wien*, R package version 1.6-7, https://CRAN.R-project.org/package=e1071, [Ziyaret Tarihi: 15 Kasım 2017].

Mirze, S.K., 2013, *İşletme*, Literatür Yayınları, İstanbul, ISBN:978-975-04-0549-5.

Mitchell, T.M., 1997, *Machine learning*, MIT Press and the McGraw-Hill Companies Inc., Singapore, ISBN:0-07-042807-7.

Moeyersoms, J., Martens, D., 2015, Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector, *Decision Support Systems*, 72, 72–81.

Mohammadi, G., Tavakkoli-Moghaddam, R., Mohammadi, M., 2013, Hierarchical neural regression models for customer churn prediction, *Journal of Engineering*, 2013(543940), 1-9.

Mozer, M.C., Wolniewicz, R., Grimes, D.B., Johnson, E., Kaushansky, H., 2000, Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry, *IEEE Transactions on Neural Networks*, 11(3), 690–696.

Murphy, K. P., 2006, *Naive bayes classifiers*, http://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall06/lectures/naiveBayes.pdf, [Ziyaret Tarihi: 6 Aralık 2017].

Mutanen, T., 2006, Customer churn analysis–a case study, *Journal of Product and Brand Management*, 14(1), 4-13.

Müller, H., Freytag, J.C., 2005, *Problems, methods, and challenges in comprehensive data cleansing*, http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/techreports/2003-hub_ib_164-mueller.pdf, [Ziyaret Tarihi: 6 Aralık 2017].

Müssel, C., Lausser, L., Maucher, M., Kestler, H.A., 2012, Multi-objective parameter selection for classifiers, *Journal of Statistical Software*, 46(5), 1-27.

Neal, W. D., 1999, Satisfaction is nice, but value drives loyalty, *Marketing Research*, 11(1), 20.

Nie, G., Rowe, W., Zhang, L., Tian, Y., Shi, Y., 2011, Credit card churn forecasting by logistic regression and decision tree, *Expert Systems with Applications*, 38(12), 15273–15285.

Nisbet, R., Elder, J., Miner, G., 2009, *Statistical analysis and data mining*, Elsevier, UK, ISBN: 978-0-12-374765-5.

Norton, M. J., 1999, *Knowledge Discovery in Databases*, https://www.ideals.illinois.edu/bitstream/handle/2142/8255/librarytrendsv48i1c_opt.pdf?sequence=1&isAllowed=y, [Ziyaret Tarihi: 6 Aralık 2017].

Oliver, R.L., 1997, *Loyalty and profit: long-term effects of satisfaction*, McGraw-Hill Companies, Inc., New York.

Onaran, B., Bulut, Z.A., Özmen, A., 2013, A study to ınvestigate the effect of customer value on customer satisfaction, brand loyalty and customer relationship management performance, *Business and Economics Research Journal*, 4(2), 37.

Ou, W.M., Shih, C.M., Chen, C.Y., Wang, K.C., 2011, Relationships among customer loyalty programs, service quality, relationship quality and loyalty: An empirical study, *Chinese Management Studies*, 5(2), 194-206.

Owczarczuk, M., 2010, Churn models for prepaid customers in the cellular telecommunication industry using large data marts, *Expert Systems with Applications*, 37(6), 4710–4712.

Özkan, Y., 2008, *Veri madenciliği yöntemleri*. Papatya Yayıncılık, İstanbul, 978- 975-6797-82-2.

Öztemel, E., 2012, *Yapay sinir ağları*, Papatya Yayıncılık Eğitim, İstanbul, ISBN:975-978-6797-39-6.

Padhy, N., Mishra, D., Panigrahi, R., 2012, The survey of data mining applications and feature scope, *International Journal of Computer Science, Engineering and Information Technology*, 2(3), 43-58.

Pendharkar, P. C., 2009, Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services, *Expert Systems with Applications*, 36(3), 6714-6720.

Penrose, R., 1955, A generalized inverse for matrices, *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3), 406-413.

Petrovskiy, M. I., 2003, Outlier detection algorithms in data mining systems, *Programming and Computer Software*, 29(4), 228-237.

Phadke, C., Uzunalioglu, H., Mendiratta, V. B., Kushnir, D., Doran, D., 2013, Prediction of subscriber churn using social network analysis, *Bell Labs Technical Journal*, 17(4), 63–75.

Popović, D., Bašić, B. D., 2009, Churn prediction model in retail banking using fuzzy C-means algorithm, *Informatica*, 33(2), 243-248.

Quinlan, J. R., 1987, Generating production rules from decision trees, *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, 23-28 Ağustos 1987 Milan Italy, Morgan Kaufmann, 304–307.

Quinlan, J. R., 1996, Improved use of continuous attributes in C4.5, *Journal of Artificial Intelligence Research*, 4, 77-90.

R Development Core Team, 2008, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org, [Ziyaret Tarihi: 15 Kasım 2017].

Radosavljevik, D., van der Putten, P., Larsen, K.K., 2010, The ımpact of experimental setup in prepaid churn prediction for mobile telecommunications: What to predict, for whom and does the customer experience matter? *Transactions on Machine Learning and Data Mining*, 3(2), 80–99.

Raileanu, L. E., Stoffel, K., 2004, Theoretical comparison between the gini index and information gain criteria, *Annals of Mathematics and Artificial Intelligence*, 41(1), 77-93.

Refaeilzadeh, P., Tang, L., Liu, H., 2009, *Cross-validation*, Encyclopedia of database systems Springer, US, 532-538.

Richeldi, M., Perrucci, A., 2002, Churn analysis case study, https://sfb876.tu-dortmund.de/PublicPublicationFiles/richeldi_perrucci_2002b.pdf, [Ziyaret Tarihi: 15 Kasım 2017].

Richter, Y., Yom-Tov, E., Slonim, N., 2010, *Predicting customer churn in mobile networks through analysis of social groups*, Proceedings of the 2010 SIAM International Conference on Data Mining, S. Parthasarathy, B. Liu, B. Goethals, J. Pei, C. Kamath (ed.), Philadelphia, Society for Industrial and Applied Mathematics, pp. 732–741.

Rijsbergen, C.J., 1979, *Information Retrieval*, http://www.dcs.gla.ac.uk/Keith/Preface.html, [Ziyaret Tarihi: 6 Aralık 2017].

Risselada, H., Verhoef, P.C., Bijmolt, T.H.A., 2010, Staying power of churn prediction models, *Journal of Interactive Marketing*, 24(3), 198–208.

Rodan, A., Fayyoumi, A., Faris, H., Alsakran, J., Al-Kadi, O., 2015, Negative correlation learning for customer churn prediction: A comparison study, *The Scientific World Journal*, 2015(473283), 1-7.

Rohanizadeh, S.S., Moghadam, M.B., 2009, A proposed data mining methodology and ıts application to ındustrial procedures, *Journal of Industrial Engineering*, 4(1), 37-50.

RStudio Team, 2016, *RStudio: Integrated Development for R*, RStudio, Inc., Boston, URL http://www.rstudio.com/, [Ziyaret Tarihi: 15 Kasım 2017].

Saatçioğlu, Ö, 2016, *Rekabet, strateji ve verimlilik*, Endüsti mühendisliğine giriş, İçerisinde Öztemel, E. (ed.), Papatya Yayıncılık Eğitim, İstanbul, ISBN:978-975-6797-89-1, 63-82.

Saradhi, V.V., Palshikar, G.K., 2011, Employee churn prediction, *Expert Systems with Applications*, 38(3), 1999–2006.

Satman, M.H., 2010, *İstatistik ve Ekonometri Uygulamaları ile R*, Turkmen Kitabevi, İstanbul, ISBN: 978-605-4259-14-4.

Shafique, U., Qaiser, H., 2014, A comparative study of data mining process models (KDD, CRISP-DM and SEMMA), *International Journal of Innovation and Scientific Research*, 12(1), 217-222.

Shankar, A., 2002, *Face detection in images: Neural networks & Support Vector*, Doktora Tezi, Indian Institute of Technology, Kanpur.

Sharma, A., Panigrahi, D.K., 2013, A neural network based approach for predicting customer churn in cellular network services, *International Journal of Computer Applications*, 27(11), 26-31.

Sharp, B., Sharp, A., 1997, Loyalty programs and their impact on repeat-purchase loyalty patterns, *International Journal of Research in Marketing*, 14(5), 473-486.

Shaw, M.J., Subramaniam, C., Tan, G.W., Welge, M.E., 2001, Knowledge management and data mining for marketing, *Decision Support Systems*, 31(1), 127-137.

Silahtaroğlu, G., 2008, *Kavram ve algoritmalarıyla temel veri madenciliği*, Papatya Yayıncılık Eğitim, İstanbul, ISBN:978-975-6797-81-5.

Singh, Y., Chauhan, A.S., 2009, Neural networks in data mining, *Journal of Theoretical & Applied Information Technology*, 5(1), 37-42.

Sneath, A., Sokal, R.R., 1963, *Principles of numerical taxonomy*, W.H. Freeman and Company, San Francisco, ISBN: 0716706970.

Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006, *Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation*, Australian Conference on Artificial Intelligence, İçerisinde: Sattar A., Kang B. (ed.), Springer, ISBN: 978-3-540-49787-5, 1015-1021.

Song, G., Yang, D., Wu, L., Wang, T., Tang, S., 2006, *A mixed process neural network and its application to churn prediction in mobile communications*, Sixth International Conference on Data Mining Workshops, 18-22 Aralık 2006 Hong-Kong China, IEEE, 798–802.

Subashini, T. S., Ramalingam, V., Palanivel, S., 2009, Breast mass classification based on cytological patterns using RBFNN and SVM, *Expert Systems with Applications*, 36(3), 5284-5290.

Şentürk, A., 2006, *Veri madenciliği kavram ve teknikler*, Ekin Yayınevi, Bursa, ISBN:975-8768-24-7.

Taha, H. A., 2014, *Yöneylem araştırması*, Baray, Ş. A., Esnaf, Ş. (çev.), Literatür Yayıncılık, ISBN: 9789758431069.

Tamaddoni Jahromi, A., Sepehri, M.M., Teimourpour, B., Choobdar, S., 2010, Modeling customer churn in a non-contractual setting: the case of telecommunications service providers, *Journal of Strategic Marketing*, 18(7), 587–598.

Tamaddoni, A., Stakhovych, S., Ewing, M., 2015, Comparing churn prediction techniques and assessing their performance a contingent perspective, *Journal of Service Research*, 19(2), 123-141.

Tan, P. N., Steinbach, M., Kumar, V., 2006, *Introduction to data mining*, Pearson Education Inc., Boston, ISBN: 0-321-32136-7

Tsai, C.F., Chen, M.Y., 2010, Variable selection by association rules for customer churn prediction of multimedia on demand, *Expert Systems with Applications*, 37(3), 2006–2015.

Tsai, C.F., Lu, Y.H., 2009, Customer churn prediction by hybrid neural networks, *Expert Systems with Applications*, 36(10), 12547–12553.

Tso, G.K., Yau, K.K., 2007, Predicting electricity energy consumption: A comparison of regression analysis, *Decision Tree And Neural Networks Energy*, 32(9), 1761-1768.

Türk Dil Kurumu (TDK), 2017, http://tdk.gov.tr/, [Ziyaret Tarihi: 6 Aralık 2017].

Uçar, F., Dandıl, B., Ata, F., 2015, Classification of power quality events using extreme learning machine, *23rd Signal Processing and Communications Applications Conference (SIU)*, 16-19 Mayıs 2015 Malatya Türkiye, IEEE, ISBN: 9781467373876, 970-973.

URL, http://www.sgi.com/tech/mlc/db/, [Ziyaret Tarihi: 6 Aralık 2017].

Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K.C., 2015, A comparison of machine learning techniques for customer churn prediction, *Simulation Modelling Practice and Theory*, 55, 1–9.

Venables, W.N., Ripley, B.D., 2002, *Modern Applied Statistics with S. Fourth Edition*, Springer, New York, ISBN: 0-387-95457-0.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B., 2012, New insights into churn prediction in the telecommunication sector: A profit driven data mining approach, *European Journal of Operational Research*, 218(1), 211–229.

Verbeke, W., Martens, D., Baesens, B., 2014, Social network analysis for customer churn prediction, *Applied Soft Computing*, 14, 431–446.

Verbeke, W., Martens, D., Mues, C., Baesens, B., 2011, Building comprehensible customer churn prediction models with advanced rule induction techniques, *Expert Systems with Applications*, 38(3), 2354–2364.

Verbraken, T., Verbeke, W., Baesens, B., 2013, A novel profit maximizing metric for measuring classification performance of customer churn prediction models, *IEEE Transactions on Knowledge and Data Engineering*, 25(5), 961–973.

Walesiak, M., Dudek, A., 2017, *clusterSim: Searching for Optimal Clustering Procedure for a Data Set*, R package version 0.45-2, https://CRAN.R-project.org/package=clusterSim, [Ziyaret Tarihi: 15 Kasım 2017].

Wand, Y., Wang, R.Y., 1996, Anchoring data quality dimensions in ontological foundations, *Communications of the ACM*, 39(11), 86–95.

Wang, Y.F., Chiang, D.A., Hsu, M.H., Lin, C.J., Lin, I.L., 2009, A recommender system to avoid customer churn: A case study, *Expert Systems with Applications*, 36(4), 8071-8075.

Wang, Y., Gao, S., Sheng, X., 2014, Enterprise customer life-cycle value model and applied research, *Journal of Business Administration Research*, 3(2), 68.

Wang, R.Y., Strong, D., 1996, Beyond accuracy: What data quality means to data consumers, *Journal of Management Information Systems*, 12(4), 5-34.

Wei, C.P., Chiu, I.T., 2002, Turning telecommunications call details to churn prediction: a data mining approach, *Expert Systems with Applications*, 23(2), 103–112.

Weinberger, K.Q., Saul, L.K., 2009, Distance metric learning for large margin nearest neighbor classification, *Journal of Machine Learning Research*, 10(Feb), 207-244.

Weiss, S.M., Indurkhya, N., 1998, *Predictive data mining: A practical guide*, Morgan Kaufmann, USA, ISBN: 1-55860-403-0.

Wirth, R., Hipp, J., 2000, CRISP-DM: Towards a standard process model for data mining, *Proceedings of the 4th International Conference On The Practical Applications Of Knowledge Discovery And Data Mining*, 17-31 Ağustos 1998 New York, AAAI Press, 29-39.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016, *Data mining: Practical machine learning tools and techniques*, Morgan Kaufmann, ISBN: 0-12-088407-0.

Wu, Y., Qi, J., Wang, C., 2009, The study on feature selection in customer churn prediction modeling, *International Conference on Systems, Man and Cybernetics*, 11-14 Ekim 2009 San Antonio USA, IEEE, 3205-3210.

Xia, G., Jin, W., 2008, Model of customer churn prediction on support vector machine, *System Engineering – Theory and Practice*, 28(1), 71-77.

Xiao, J., Xiao, Y., Huang, A., Liu, D., Wang, S., 2015, Feature-selection-based dynamic transfer ensemble model for customer churn prediction, *Knowledge and Information Systems*, 43(1), 29–51.

Xiao, J., Xie, L., He, C., Jiang, X., 2012, Dynamic classifier ensemble model for customer classification with imbalanced class distribution, *Expert Systems with Applications*, 39(3), 3668–3675.

Xie, Y., Li, X., Ngai, E. W. T., Ying, W., 2009, Customer churn prediction using improved balanced random forests, *Expert Systems with Applications*, 36(3), 5445–5449.

Yadav, S., Shukla, S., 2016, Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification, *6th International Conference on Advanced Computing (IACC)*, 27-28 Şubat 2016 Bhimavaram India, IEEE, 78-83.

Yu, X., Guo, S., Guo, J., Huang, X., 2011, An extended support vector machine forecasting framework for customer churn in e-commerce, *Expert Systems with Applications*, 38(3), 1425–1430.

Zeithaml, V. A., 1988, Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence, *The Journal of marketing*, 52(3), 2-22.

Zhang, S., Zhang, C., Yang, Q., 2003, Data preparation for data mining, *Applied Artificial Intelligence*, 17(5-6), 375-381.

Zhang, Y., Qi, J., Shu, H., Cao, J., 2007, A hybrid KNN-LR classifier and its application in customer churn prediction, *International Conference on Systems, Man and Cybernetics*, 7-10 Ekim 2007 Montreal Canada, IEEE, 3265–3269.

Zhao, J., Dang, X.H., 2008, Bank customer churn prediction based on support vector machine: taking a commercial bank's VIP customer churn as the example, *4th International Conference on Wireless Communications, Networking and Mobile Computing*, 12-14 Ekim 2008 Dalian China, IEEE, ISBN: 978-1-4244-2108-4, 1–4.

Zhou, Z.H., 2003, Three perspectives of data mining, *Artificial Intelligence*, 143(2003), 139–146.

Zurada, J.M., 1992, *Introduction to artificial neural systems*, West Publishing Company, St. Paul: West, ISBN: 978-0314933911.

EKLER

EK-1 (Veri Setinin %70 Eğitim %30 Test Şeklinde Ayrılması İle Oluşturulan Aşırı Öğrenme Makinesi Modeli En İyi Parametre Arama Kodları)

```
install.packages("XLConnect")
install.packages("xlsx")
install.packages("clusterSim")
install.packages("caret")
install.packages("elmNN")
library(elmNN)
library(caret)
library(clusterSim)
library(XLConnect)
library(xlsx)

#verinin okunması
churnVeri<-read.xlsx("Data1a.xlsx", sheetName = "Sheet1", header = FALSE)
churnVeri<-as.data.frame(churnVeri)
colnames(churnVeri)<-c("account_length", "area_code1", "area_code2",
"area_code3", "international_plan",
                       "voice_mail_plan",   "number_vmail_message",
"total_day_minutes",    "total_day_calls",
                       "total_day_charge",  "total_eve_minutes",
"total_eve_calls",  "total_eve_charges",
                       "total_night_minutes",   "total_night_calls",
"total_night_charge",   "total_intl_minutes",
                       "total_intl_calls",  "total_intl_charge",
"number_csc",   "churn")

#veri donusumu (dogrusal)
churnVeri<-data.Normalization(churnVeri[,c(-2,-3,-4,-5,-6)], type = "n4",
normalization = "column")


#analiz sonuclari icin dosya olusturma
wb <- loadWorkbook("resultFile.xlsx", create = TRUE)
sheet<-createSheet(wb, name = "ResultELM")

#ihtiyac duyulan degisken tanimlamaları
actfunlist <- c("sig", "sin", "radbas", "hardlim", "hardlims", "satlins",
"tansig", "tribas", "poslin", "purelin")
treshold <- matrix(ncol=9)
d<matrix(ncol=8)
d <- matrix(nrow=length(actfunlist), ncol=9)


#performans geçerleme (hold-out %70, %30)
set.seed(1)
trainingCL<-createDataPartition(y=churnVeri$churn, p=.70, list = FALSE)
training<-churnVeri[trainingCL,]
test<-churnVeri[-trainingCL,]
testData<-test[,-21]
testClass<-test[[21]]
trainingData<-training[,-21]
trainingClass<-training[[21]]
```

EK-1 (Veri Setinin %70 Eğitim %30 Test Şeklinde Ayrılması İle Oluşturulan Aşırı Öğrenme Makinesi Modeli En İyi Parametre Arama Kodları)(devam)

```
#k gizli noron sayisi, j aktivasyon fonksiyonları
for(k in 1:1000){
  for(j in 1:10){
    #modelin kurulmasi
    set.seed(1)
    modelELM<-elmtrain(trainingData, trainingClass, nhid = k, actfun =
actfunlist[[j]])
    modelResult<-predict.elmNN(modelELM, testData)
    modelResult<-as.data.frame(modelResult)
    modelResult$V1<-data.Normalization(modelResult$V1, type = "n4",
normalization = "column")
    forecast<-modelResult$V1
    for (i in 1:length(forecast)){
      if(forecast[[i]]>0.70){
        forecast[[i]]<-1
      } else {
        forecast[[i]]<-0
      }
    }

#karisiklik matrisinin olusturulmasi
matris<-table(forecast, trainingClass, dnn = c("tahmini siniflar", "gercek
siniflar"))
tp<-matris[4]
fp<-matris[2]
fn<-matris[3]
tn<-matris[1]
Accuracy<-(tp+tn)/sum(matris)
hata<-1-Accuracy
Recall<-tp/(tp+fn)
specificity<-tn/(fp+tn)
precision<-tp/(tp+fp)
F_olcusu<-(2*precision*Recall)/(precision+Recall)
Accuracy_n<-tp/(tp+fp+fn)

#elde edilen her bir sonucun matrisin satırı olarak atanmasi
layer<-k
actfunction<-actfunlist[[j]]
d[j,] <- c(layer, actfunction, Accuracy, hata, Recall, specificity,
precision, F_olcusu, Accuracy_n)
  }

#elde edilen matrisin daha önce elde edilen matris ile birlestirilmesi
  treshold<-rbind(treshold,d)
  d<-NULL
  d <- matrix(nrow=length(actfunlist), ncol=9)
}

#tum sonuclarin excele yazdirilmasi
write.xlsx(sonucHO80, "C:/Users/Data1/Calisma.xlsx", sheetName
="ref1sonucHO70", append = TRUE)
```

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Fatma Önay KOÇOĞLU
Doğum Yeri	Fatih-İstanbul
Doğum Tarihi	20.07.1984
Uyruğu	☑ T.C. ❒ Diğer:
Telefon	
E-Posta Adresi	fonayk@gmail.com
Web Adresi	http://aves.istanbul.edu.tr/fonayk/

Eğitim Bilgileri	
Lisans	
Üniversite	İstanbul Üniversitesi – Sakarya Üniversitesi
Fakülte	Fen Fakültesi – Mühendislik Fakültesi
Bölümü	Matematik – Endüstri Mühendisliği
Mezuniyet Yılı	2009 - 2017

Yüksek Lisans	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Enformatik Anabilim Dalı
Programı	Enformatik Yüksek Lisans Programı
Mezuniyet Tarihi	26.07.2012

Doktora	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Enformatik Anabilim Dalı
Programı	Enformatik Programı
Mezuniyet Tarihi	24.11.2017

Makale ve Bildiriler

Koçoğlu F.Ö., Emre İ.E., Erol Ç., "Observation of Success Status of Employees in E-Learning Courses in Organizations with Data Mining", International Journal of E-Adoption (IJEA), no.1, pp.38-49, 2017.

Sutaş Bozkurt A.P., Güngör G., Özen Z., Ünlüsoy E.Ö., Uğur O., Sayilgan N.C., et al.,"Older Age Is Related with Higher Pocp in Adult Patient Population in Turkey- Preliminary Report", ANESTHESIA AND ANALGESIA, pp.361-361, 2016.

Koçoğlu F.Ö., Gülseçen S., Erol Ç., Koçoğlu E.S., "Internet-Based learning from the perspective of employees: A study on analysis of awareness, use case and effectiveness", World Journal on Educational Technology, no.1, pp.1-9, 2016.

Koçoğlu F.Ö., Kartal E., Özen Z., Erol Ç., Gülseçen S., "Aspects of Students About Information Technology Courses in Social Science", Procedia - Social and Behavioral Sciences, vol.176, pp.148-154, 2015.

Özen Z., Koçoğlu F.Ö., Beden Ş., "The Examination Of User Habits Through Google Analytic Data Of Academic Education Platforms", International Journal of E-Adoption (IJEA), no.2, pp.31-46, 2014.

Gülseçen S., Kartal Karataş E., Koçoğlu F.Ö., "Can Geogebra Make Easier The Understanding Of Cartesian Co-Ordinates? A Quantitative Study In Turkey", International Journal on New Trends in Education and Their Implications (IJONTE), vol.3, pp.19-29, 2012.

Koçoğlu F.Ö., Özcan T., Baray Ş.A., "Data Mining Methods for Solving Customer Churn Analysis Problem", 4th International Management Information Systems Conference, İSTANBUL, TÜRKIYE, 17-20 Ekim 2017, pp.197-198

Koçoğlu F.Ö., Özcan T., "Importance Of Data Mının In Access To Knowledge As An Intellectual Capıtal And A Study On Churn Analysıs", International Conference of Management Economics and Business, ZONGULDAK, TÜRKIYE, 7-9 Eylül 2017, pp.1-1

Koçoğlu F.Ö., Emre İ.E., Erol Ç., "Observation of Success in e-Learning with Data Mining Methods", 6th International Conference On “Innovations In Learning For The Future”: Next Generation (FL2016) , İSTANBUL, TÜRKIYE, 24-26 Ekim 2016, pp. -

Koçoğlu F.Ö., Özcan T., Baray Ş.A., "Veri Madenciliğinde Ayrılan Müşteri Analizi Problemi Üzerine Bir Literatür Araştırması", Uluslararası Katılımlı Üretim Araştırmaları Sempozyumu "4. Sanayi Devriminde Üretim"", İSTANBUL, TÜRKIYE, 12-14 Ekim 2016, pp. 868 – 874.

Özen Z., Koçoğlu F.Ö., Kartal E., Erol Ç., Gülseçen S., "Awareness of Electronic Waste (E-Waste) in Turkey", 6th International Symposium on Energy From Biomass and Waste Venice 2016, Venedik, ITALYA, 14-17 Kasım 2016.

Sutaş Bozkurt A.P., Güngör G., Özen Z., Koçoğlu F.Ö., Kartal E., Emre İ.E., et al., "Older Age Is Related with Higher Pocp in Adult Patient Population in Turkey-Preliminary Report", 16th World Congress of Anaesthesiologists, Hong Kong, HONGKONG, 28 Ağustos - 2 Eylül 2016, pp.361-361.

Koçoğlu F.Ö., Erol Ç., Gülseçen S., Koçoğlu E.S., "Internet Based Learning from the Perspective of Employees: Study on Analysis of Awareness, Use Case and Effectiveness", 6th World Conference on Learning Teaching and Educational Leadership, Paris, FRANSA, 29-31 Ekim 201.

Erol Ç., Koçoğlu F.Ö., Özen Z., Kartal E., Gülseçen S., "Interpet: Evcil Hayvan Sağlığı Hakkında Bilgiye Erişim Kaynağı Olarak İnternet", Akademik Bilişim 2015, ESKİŞEHİR, TÜRKIYE, 4-6 Şubat 2015.

Koçoğlu F.Ö., Kartal E., Özen Z., Erol Ç., Gülseçen S., "Aspects Of Students About Information Technology Courses In Social Science", International Educational Technology Conference, Chicago, ABD, 3-5 Eylül 2014.

Özen Z., Kartal E., Selçukcan Erol Ç., Koçoğlu F.Ö., "Bilgi Ekonomisi Üzerine Bir Çalışma", Akademik Bilişim 2014, İÇEL, TÜRKIYE, 5-7 Şubat 2014, ss.239-245.

Selçukcan Erol Ç., Koçoğlu F.Ö., Özen Z., Kartal Karataş E., "Kamu Kurumlarında Elektronik İmza Hakkında Karşılaştırmalı Bir Çalışma", 6th International Conference on Information Security and Cryptology, ANKARA, TÜRKİYE, 1-4 Eylül 2013, vol.1, no.1, pp.111-115.

Ecevit Sati Z., Kartal Karataş E., Özen Z., Koçoğlu F.Ö., Selçukcan Erol Ç., "E-Üniversite'ye Yolculuk", Akademik Bilişim 2013, ANTALYA, TÜRKIYE, 1-4 Şubat 2013, cilt.1, no.0, ss.355-360.

Gülseçen S., Kartal E., Koçoğlu F.Ö., "Can Geogebra Make Easier The Understanding Of Cartesian Co-Ordinates? A Quantitative Study In Turkey", 3rd International Conference on New Trends in Education and Their Implications, ANTALYA, TÜRKIYE, 26-28 Nisan 2012.

Özen Z., Koçoğlu F.Ö., Beden Ş., "The Examination Of User Habits Through The Google Analytic Data Of Academic Education Platforms", 4th International Future Learning Conference on Innovations in Learning for the Future: e-Learning, İSTANBUL, TÜRKIYE, 14-16 Kasım 2012.

Sati Z., Özen Z., Koçoğlu F.Ö., Kartal E., Erol Ç., "Yerel Yönetimlerde E-Devlet Uygulamaları: İstanbul İli ve Belediye Yönetimlerinde Kullanılan E-Devlet Hizmetlerinin Değerlendirilmesi", VI. İstanbul Bilişim Kongresi, "Akıllı Yaşam Ve Başarı Örnekleri: Akıllı Şehirler", İSTANBUL, TÜRKIYE, 7-8 Kasım 2012, ss.61-75.

Saraç A.E., Koçoğlu F.Ö., Ayvaz Reis Z., "Web Tabanlı Eğitimde İçerik Tasarımı", XIII. Akademik Bilişim Konferansı, MALATYA, TÜRKİYE, 02-04 Şubat 2011, ss.493-500 (Link)

Koçoğlu F.Ö., Özcan T., "Veri Madenciliğinde Kümeleme Algoritmaları ile Müşteri Segmentasyonu", R ile Veri Madenciliği Uygulamaları, Balaban M.E., Kartal E., Ed., Çağlayan Kitabevi, İstanbul, ss.187-220, 2016

Koçoğlu F.Ö., "Bilgi Yönetimi ve İş Süreçleri Etkileşimi", Bilgi Yönetimi: Bilgi Türeticileri, Büyük Veri, İnovasyon, Kurumsal Zeka, Gülseçen, S., Ed., Papatya Yayıncılık, İstanbul, ss.79-90, 2015