AĞIRLIKLI BAYES SINIFLANIRICIDA AĞIRLIKLARIN OPTİMİZASYONU

Gamzepelin AKSOY

Yüksek Lisans Tezi

Yazılım Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Murat KARABATAK

AĞUSTOS-2018

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

AĞIRLIKLI BAYES SINIFLANDIRICIDA AĞIRLIKLARIN OPTİMİZASYONU

YÜKSEK LİSANS TEZİ

Gamzepelin AKSOY

(151137104)

Anabilim Dalı: Yazılım Mühendisliği

Tez Danışmanı: Doç. Dr. Murat KARABATAK

Tezin Enstitüye Verildiği Tarih: 14.08.2018

AĞUSTOS – 2018

T.C.

FIRAT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

AĞIRLIKLI BAYES SINIFLANDIRICIDA AĞIRLIKLARIN OPTİMİZASYONU

YÜKSEK LİSANS TEZİ

Gamzepelin AKSOY

(151137104)

Anabilim Dalı: Yazılım Mühendisliği

Tezin Enstitüye Verildiği Tarih : 14 Ağustos 2018
Tezin Savunulduğu Tarih : 14 Eylül 2018

Tez Danışmanı : Doç. Dr. Murat KARABATAK (F.Ü.)
Diğer Jüri Üyeleri : Doç. Dr. Bilal ALATAŞ (F.Ü.)
Dr. Öğr. Üyesi Özal YILDIRIM (M.Ü.)

AĞUSTOS-2018

ÖNSÖZ

Yüksek lisans çalışmalarım sürecinde maddi ve manevi destek sağlayan kurum ve kişilere teşekkürlerimi sunmayı bir borç bilirim.

Yüksek lisans eğitimim boyunca her türlü sorumu yılmadan cevaplayan ve bilgi ve tecrübelerini paylaşıp bana yol gösteren danışman hocam Doç. Dr. Murat KARABATAK'a teşekkür ederim.

Hayatımın her anında maddi ve manevi desteğini benden esirgemeyen, sabır ile eğitim hayatımda bana yol gösteren canım kardeşim Gonca Pervin AKSOY'a ve aileme sonsuz teşekkürler.

Gamzepelin AKSOY

ELAZIĞ – 2018

İÇİNDEKİLER

Sayfa No

ÖZET

Teknolojinin hızla gelişmesi ve bununla birlikte artan veri miktarı, veri analizini güçleştirmektedir. Birçok işlemin elektronik ortamda kaydedilmesi, saklanabilmesi ve istenildiği zaman veriye erişilebilmesi önem kazanmaktadır. Veri işlenmediği sürece anlam ifade etmemektedir. Verilerin anlamlı bir bütün haline getirilebilmesi için veri madenciliğinden faydalanılmaktadır. Veri madenciliği büyük ölçekli veriler arasında bilgileri ayrıştırarak yararlı bilgilere ulaşılması ve bu verilerden gelecekle ilgili tahminde bulunabilmek için bağıntıların bilgisayar programı kullanılarak aranması işlemidir.

Bu tez kapsamında, veri madenciliği sınıflandırma yöntemleri ve bu yöntemler arasında yer alan Bayes sınıflandırma algoritması incelenmiştir. Ayrıca Bayes sınıflandırma algoritmasının geliştirilmiş bir modeli olan Ağırlıklı Bayes algoritması da tez kapsamında incelenerek, bu yöntemde kullanılan ağırlıkların optimize edilmesi amaçlanmıştır. Bu amaçla öncelikle ağırlıkların hızlı bir şekilde bulunabilmesi için Hızlandırılmış Ağırlıklı Sade Bayes (Fasted Weighted Naive Bayesian- FW-NB) yöntemi önerilmiş ve daha sonra ağırlıkların optimizasyonu için Genetik Algoritma kullanılarak (Genetik Algoritma Tabanlı Ağırlıklı Sade Bayes- GAW-NB) ağırlıklar optimize edilmiştir.

Tez kapsamında kullanılan yöntemler 5 farklı veri setine uygulanmış ve elde edilen sonuçlar karşılaştırmalı olarak değerlendirilmiştir. Elde edilen sonuçlardan FW-NB algoritması ile ağırlıkların daha hızlı bulunduğu ve GAW-NB algoritması ile performans değerinin W-NB algoritmasına göre daha yüksek olduğu sonucuna ulaşılmıştır.

Anahtar Kelimeler: Veri Madenciliği, Sınıflandırma Algoritmaları, Bayes Algoritması, Ağırlıklı Sade Bayes Algoritması

SUMMARY

Optimization of the weights of Weighted Naïve Bayesian Classifier

The rapid development of the technology, along with the increasing amount of data, makes data analysis inconvenient. Today, it is important that many processes can be recorded, stored and accessed in an electronic environment. As long as the data are not processed, it does not make any sense. Data mining is used to make the data meaningful. Data mining is the process of retrieving useful information from large-scale data by separating the information among the large-scale data and retrieving the data by using a software to make predictions about the future.

In this thesis, methods of data mining and Bayes classification algorithm are examined. The Weighted Bayes algorithm, which is an improved model of the Bayes classification algorithm, is also examined in the thesis and it is aimed to optimize the weights used in this method. For this purpose, we propose Fasted Weighted Naive Bayes (FW-NB) method to find weights quickly and then weights are optimized by using Genetic Algorithm (Genetic Algorithm Based Weighted Naive Bayes-GAW-NB) for the optimization of weights.

The methods used in this thesis are applied on 5 different data sets and the results are evaluated comparatively. The results show that FW-NB algorithm is faster than W-NB algorithm and performance of GAW-NB algorithm is higher than W-NB.

Key Words: Data Mining, Classification Algorithms, Bayesian Algorithms, Weighted Naïve Bayesian Algorithms

ŞEKİLLER LİSTESİ

Sayfa No

TABLOLAR LİSTESİ

Sayfa No

KISALTMALAR LİSTESİ

RTMS **:** Uzaktan Mikrodalga Trafik Sensörü (Remote Traffic Microwave Sensor)

SVM **:** Destek Vektör Makinesi (Support Vector Machine)

KNN **:** k-En Yakın Komşu (K Nearest Neighborhood)

YSA **:** Yapay Sinir Ağları

CPB **:** Koşullu Tercihler Tabanı (Conditional Preferences Base)

MLP **:** Çok Katmanlı Algılayıcı (Multi-Layer Perceptron)

HRA **:** Sezgisel Azaltma Yaklaşımı (Heuristic Reduction Algorithm)

DDTRS **:** Hastalık Teşhis ve Tedavi Öneri Sistemi (Disease Diagnosis and Treatment Recommendation System)

SPM **:** Ardışık Desen Madenciliği (Sequential Pattern Mining)

FPM **:** Esnek Desen Madenciliği (Flexible Pattern Mining)

DLPCA **:** Dağıtılmış Yük Dengeleme Ana Bileşen Analizi (Distributed Load Balancing Principal Component Analysis)

DDM **:** Dağıtılmış Veri Madenciliği (Distribute Data Mining)

KE **:** Bilgi Çıkarımı (Knowledge Extraction)

KDD **:** Veri Tabanlarından Bilgi Keşfi (Knowledge Discovery in Databases)

NB **:** Sade Bayes (Naive Bayesian)

W-NB **:** Ağırlıklı Sade Bayes (Weighted Naive Bayesian)

FW-NB **:** Hızlandırılmış Ağırlıklı Sade Bayes (Fasted Weighted Naive Bayesian)

GAW-NB **:** Genetik Algoritma Tabanlı Ağırlıklı Sade Bayes (Genetic Algorithm Based Weighted Naive Bayes)

1. GİRİŞ

Teknolojik gelişmelerin sonucunda, veri miktarında yaşanan artış beraberinde bu verinin faydalı ve kullanılabilir hale getirilmesini zorunlu kılmıştır. Elektronik ortamlarda kaydedilen verilerin saklanması ve istenildiği zaman erişilebilmesi önem kazanmıştır. Verilerin kullanılabilir hale getirilebilmesi için veri madenciliğinden faydalanılmaktadır.

Verileri bilgiye dönüştürmenin geleneksel yöntemi, analizlerin manuel olarak yapılmasına ve yorumlamaya dayanır. Bu şekilde yapılan analizler; yavaş, pahalı ve özneldir. Veri boyutlarında yaşanan artış ile manuel veri analizi yapmak pratikliğini kaybetmiştir ve yerini bilgisayarlar ile yapılan analizler almaya başlamıştır. Bilgisayarlar, insanların elde edip yorumlayabileceğinden çok daha fazla verinin toplanmasını sağlayabileceği gibi muazzam miktardaki verilerden anlamlı kalıpların ve yapıların ortaya çıkarılmasını sağlamaktadır. Bilgisayarların veri işleme süreçlerinde kullanılmasıyla birlikte, veri madenciliği kavramı ortaya çıkmıştır.

Veri madenciliği, yeni teknikler ile veri sahibine kullanışlı ve anlaşılabilir bir şekilde verileri özetleme imkânı sunmaktadır. Ayrıca tahmin edilemeyen ilişkileri bularak, gözlemlenen veri setlerinin analizini yapabilmeyi sağlamaktadır. Veri madenciliği veri analiz sürecini ölçen ve sorunun tanımı ile başlayıp veri kaynaklarının incelenmesi ile devam eden, değerlendirme ve sunum aşamaları ile sonlanan bir yaklaşımdır. Başlı başına bir çözüm olmamakla birlikte, bir çözüm bulmak için karar verme sürecini destekleyen ve gerekli bilgiye ulaşmayı sağlayan bir araçtır. Araştırmacılara veri içindeki kalıpları ve ilişkileri bulmada yardımcı olmaktadır. Bu sebeple veri madenciliği birçok alanda çeşitli araştırmalarda kullanılmıştır.

1.1. Literatür Araştırması

Literatürde veri madenciliği teknikleri kullanılarak birçok farklı uygulama gerçekleştirilmiştir. Bu çalışmaların çoğunun amacı, elde edilen verilerin daha kullanışlı bilgi paketlerine dönüştürülmesini sağlamaktır. Bunun için genel olarak istatistiksel yöntemler, farklı ilişkilendirme kuralları ve sınıflandırma algoritmaları kullanılmıştır. Çalışmaların bazıları aşağıda tartışılmıştır.

Zengin vd. [1], bir eğitim çalışmasından elde ettikleri verileri, işlemeye uygun veri madenciliği tekniklerini kullanarak analiz eden örnek bir çalışma sunmuşlardır. Bu amaçla eğitim bilimlerinde kullanılan "Bilgisayar Öz yeterlilik Ölçeğini" çalışma gruplarına uygulamışlardır. Verilere tanımlayıcı istatistikler (t testi ve varyans analizi), veri madenciliği tekniklerinden karar ağacı ve kümeleme tekniklerini uygulayarak verileri analiz etmişlerdir. İstatistiksel analizleri gerçekleştirmek için Microsoft SQL Server 2008 ve Delphi 2009 programlama dilinde yazılmış bir program kullanarak hesaplamaları yapmışlardır. Microsoft SQL Server 2008'i, veri madenciliği tekniklerinden kümelemenin yapılması için doğrudan kullanmışlardır. Araştırmadan elde edilen bulgulara göre, veri madenciliği teknikleri ve istatistiksel tekniklerin kullanılması sonucu elde edilen ortak sonuçlar şu şekilde belirtilmiştir; "Bilgisayar terimleri ve kavramlarında yetkin olduklarını düşünenler bilgisayar kullanımı konusunda özel yeteneklere sahip olduklarına inanmaktadırlar." "Bilgisayarları kullanırken özel bir yeteneğe sahip olduklarını düşünenler, bilgisayarları vücudunun bir parçası olarak görmektedirler" ve "Altı yıldan uzun bir süredir bilgisayar kullanan öğrenciler bilgisayar kullanma konusunda özel yetenekleri olduğuna inanmaktadırlar."

Yıldırım ve Çataltepe [2], İstanbul Büyükşehir Belediyesi'nin sitesinden almış oldukları Uzaktan Trafik Mikrodalga Sensörü (Remote Traffic Microwave Sensor- RTMS) cihazlarından elde edilen hız değerlerini kullanarak, ileri yönlü trafik hızı tahminine yönelik bir çalışma yapmışlardır. Bu tahminlerde bulunabilmek için bir sensöre yakın sensörlerinde hız bilgileri alınmıştır. Ayrıca yüksek bağıntıya sahip olan sensörlerlerinde hız bilgileri kullanılmıştır. Bu amaçla k-En Yakın Komşu (K Nearest Neighborhood-KNN) ve Destek Vektör Makinesini (Support Vector Machine-SVM) kullanmışlardır. Çalışma sonucunda, SVM kullanılarak elde edilen sonuçların KNN metoduna göre daha iyi sonuçlar verdiği gözlemlenmiştir. Yakın ve yüksek bağıntılı sensörlerden alınan verilere göre yapılan tahminlerin daha iyi sonuçlar verdiğini elde etmişlerdir.

Coşkun ve Baykal [3], çeşitli kanser gruplarının yer aldığı bir veri kümesini kullanarak veri madenciliği sınıflandırma algoritmalarının performansını kıyaslamışlardır. Çalışmada, veri madenciliği tekniklerinin birçoğunun yer aldığı java programlama dili ile yazılmış olan WEKA paket programını kullanmışlardır. Ayrıca sınıflandırma algoritmalarının kıyaslanmasında önemli parametreler olan veri önişleme, özniteliklerin seçimi, test veri kümesinin belirlenmesi ve modelin başarımını belirleyen doğruluk, kesinlik,

duyarlılık ve F-Ölçütü gibi özelliklerine değinmişlerdir. J48, Sade Bayes (NB), Lojistik Regresyon ve KStar algoritmalarının kullanılmıştır.

En yüksek doğruluk ve F-Ölçütü değerlerinin J48 algoritmasında, en yüksek kesinlik değerinin Lojistik Regresyon algoritmasında ve en yüksek duyarlılık değerinin KStar algoritmasında elde edildiğinin belirtmişlerdir. Ayrıca, kesinlik ölçütünün doğruluk ölçütünden bağımsız olarak değerlendirilmesi söz konusu olmadığı için her iki değerin ortalaması olan F-Ölçütünden elde edilen değere bakmak gerektiğini vurgulamışlardır. Sonuçlar incelendiği zaman ise doğruluk ve F-Ölçütü sonucu elde edilen değerlere göre, her ikisinde de J48, KStar, Lojistik regresyon ve Sade Bayes (NB) algoritmalarının en yüksek değerden en düşük değere doğru bir sıralama yapmışlardır.

Öztürk [4], veri madenciliği tekniklerini lojistik alanına uygulanması üzerine bir çalışma yapmıştır. Bu çalışmadaki veriler, kara lojistiği için düşünülerek kurgusal olarak hazırlanmış olup Karar Ağaçları kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Veri setinde 9 niteliğe ve 1000 örneğe yer verilmiştir. Çalışmada Knime yazılımı kullanılmıştır. Karar Ağacı sonucunda elde edilen verilere göre fiyatlandırma işleminde kayda değer değişiklikler elde edilebileceğini belirtmiştir. Lojistik sektöründe veri madenciliği tekniklerinin kullanılmasının fiyatlandırmayı iyileştireceği çalışmanın sonuçları arasında yer almaktadır. Verilerden öğrenme yolu ile elde ettiği Karar Ağacından %95.333 başarım elde etmiştir.

Uzun vd. [5] yaptıkları çalışmada veri madenciliği tekniklerini kullanarak down sendromunun doğumdan önce teşhisinde kullanılabilecek bir yöntem geliştirmişlerdir. George Washington Üniversitesinden bir veri kümesini (Her biri 31 özelliğe sahip 8216 gebe) alarak sınıflandırma algoritmalarından Sade Bayes ve Karar Ağacı, Çok Katmanlı Algılama Sistemi, Destek Vektör Makinesi, k-En Yakın Komşu algoritmalarını uygulamışlardır. Algoritmaların verimini arttırmak amacıyla, özellik seçimi ve çıkarımı işlemlerini uygulamışlardır. Olasılıksal sınıflandırma yöntemlerinin Down sendromu vakalarının erken teşhisinde etkili bir yöntem olduğunu ve özellik seçimi, özellik çıkarımı gibi uygulamalar ile boyut azaltmanın başarımı arttırdığını belirtmişlerdir.

Eşiyok çalışmasında [6], bir kümeleme algoritması olan DBSCAN ve bir sınıflandırma algoritması olan KNN kullanarak tıbbi verilerin kümelenmesi amaçlayan bir sistem önermiştir.

Mamografi verileri DBSCAN algoritması ile optimal sayıda kümeye ayrılmıştır ve hiçbir kümeye dahil edilmeyen veriler KNN sınıflandırma algoritması yardımı ile uygun kümelere yerleştirilmiştir. Bir noktanın yoğunluğu, EPS değerine bağlıdır. Parametre değerleri incelendiğinde, DBSCAN algoritmasına göre en uygun değerlerin EPS=40 ve MinPts=8 olduğu gözlemlenmiştir. Sonuç olarak kullanılan veri tabanındaki optimal sonuç veren parametrelerin EPS ve Min Pts olduğunu belirlemiştir.

Atak [7] çalışmasında, Apriori algoritmasını, Karar Ağaçlarını ve sınıflandırma yöntemlerinden Sade Bayes algoritmasını kullanarak bir kurumun gerçek ağ verileri üzerinde elde edebileceği verileri incelemiştir. Bu çalışmada, haftanın 3 gününü seçerek, belirlediği günlerdeki ağ verilerine göre çalışanların en çok hangi arama motorunu kullandığı, hangi sitelere giriş yaptığı ve sosyal medyanın gün içerisinde hangi zaman dilimlerinde kullanıldığı gibi bilgiler ortaya çıkarılmıştır. Böylece, çalışanların aktif olarak hangi süreler içinde çalıştığı bilgisine de erişilmiştir. C# programlama dilinin yanı sıra RapidMiner uygulaması kullanılarak hazırlanan Sade Bayes algoritmasından elde edilen sonuçların, Karar Ağacına göre daha yüksek başarım gösterdiğini vurgulamıştır.

Olgun ve Özdemir [8], Shewhart istatistiksel süreç kontrol grafiklerinden elde edilen ham veriler, ortalama, çarpıklık, basıklık katsayısı, standart sapma, pearson korelasyon katsayısı gibi istatistiksel verileri, veri setini hazırlamak için kullanmışlardır. Anormal değişimlerin örüntülerini tanımlamak için Bayes ve Yapay Sinir Ağları (YSA) kullanan araştırmacılar, istatistiksel özelliklerin kullanıldığı YSA ve Bayes sınıflandırıcıların performansının, ham verilerin kullanılmasıyla elde edilen sonuçlara göre performansının daha yüksek olduğunu belirlemişlerdir.

Hasanlı [9], veri madenciliği tekniklerinden biri olan kümeleme tekniğini matematiksel olarak incelemiştir. Kümeleme problemi için yeni bir algoritma önermiştir. Çalışmasında bankacılık ile alakalı bir veri kümesine, literatürde yer alan algoritmaları ve yeni oluşturduğu algoritmayı uygulamıştır. Veri seti, bankaların kredi kartı verdiği müşterilerin bilgilerinden hazırlanmış olup, sayısal verileri ve sayısal olmayan verileri incelemek için K-means ve K-mode algoritmaları kullanılmıştır. Önerilen yeni algoritma C# programlama dili kullanılarak hazırlanmıştır. Veri setinde yer alan veriler 20 kümeye ayrılmıştır.

Çalışma neticesinde ise en çok hangi insanların kredi kartı kullandığı, ödeme bilgisinin düzenli yapılıp yapılmadığı ve evlilik durumunun kredi kartı kullanımı üzerinde etkisi olup olmadığı gibi bilgilere ulaşılmıştır.

Gökgöz [10] çalışmasında, veri madenciliği tekniklerinden faydalanarak tıbbi cihazlar için ideal bakım sürecini oluşturmayı amaçlamıştır. Veriler bakım onarımı yapılan cihazlar için oluşturulan formlar göz önüne alınarak hazırlanmıştır. Bu çalışma kapsamında veri madenciliğinin Apriori Birliktelik Kuralından faydalanmıştır. Yapmış olduğu çalışma sonucunda Apriori sonucu elde edilen 353 Birliktelik Kuralı içinden destek değeri büyükten küçüğe doğru sıralanmış ilk 10 kuralı açıklamıştır.

Özçakır ve Çamurcu [11], bir pastanenin satış verilerini göz önüne alarak, veri seçme ile uygulama veri tabanına verileri aktarmışlardır. Aktarılan verilere, veri önişleme ve veri indirgeme süreçleri uygulamışlardır. Veri madenciliğine uygun olan bir veri seti elde ettikten sonra Birliktelik Kurallarından Apriori algoritmasını kullanmışlardır. Çalışma sonucunda farklı zaman dilimlerinde birlikte satın alınan ürünler olduğunu gözlemlemişlerdir. Uygulamayı internet web ortamında kullanılabilecek hale getirerek maliyet açısından kazanım sağlamışlardır.

Solmaz vd. [12] tiroit hastalığını teşhis etmek amacıyla kullanılabilecek bir Karar Destek Sistemi için Sade Bayes sınıflandırıcıyı kullanmışlardır. Kan değerleri baz alınarak hazırlanan iki veri setine uyguladıkları sınıflandırıcı ile %97.02 ve %95.04 doğruluk elde etmişlerdir. Bu sonuçlar neticesinde kan değeri temel alınan tiroit tanımlama sistemi için Sade Bayes algoritmasının kullanışlı olacağını belirtmişlerdir.

Köklü [13] yapmış olduğu çalışmada, veri tabanında yer alan özniteliklerin ayrık ve gerçel veri tipleri olmak üzere ikiye ayırıldığını belirtmiştir. Veriler evet- hayır, var-yok gibi net bilgiler içerebildiği gibi belirli aralık değerlerine sahipte olabilir. Bu veriler yorumlanırken belirli aralık değerleri düşük, orta ve yüksek olacak şekilde saptanabilir. Yapılan bu belirleme işlemi ise kişiden kişiye farklılıklar gösterebilir. Köklü' nün çalışmasında ise etkili bir sınıflandırma kural çıkarımı yöntemi elde etmek amaçlanmıştır. Böylece hata payının düştüğü gözlemlenmiştir. Yapmış olduğu yöntem ile çok sınıflı problemlerde kural çıkarımı yapılabileceğini belirtmiştir. Uygunluk değerinin optimizasyonu için CLONALG algoritmasından faydalanmıştır. Verilere optimizasyon uygulanarak gerçel değerlerin aralıkları uygun bir şekilde belirlemiştir.

Gerçel değerlerin dışında veri setinde yer alan ayrık değerleri ise ikili kodlamıştır. Bu şekilde gerçel ve ayrık değerler farklı biçimlerde kodlamıştır. Yapılan uygulamaya ise "Aralık Keşfi" ismini vermiştir. Aralık Keşfi işlemini 8 farklı veri setine uygulamış ve başarım değerinin arttığını belirtmiştir.

Acun [14], bir yazılım firmasında kullanılan yazılımların sistem hatalarını incelemiştir. Geliştirdiği yazılım ile günlük olarak hata verilerini çekerek bu verilere Apriori algoritmasını uygulamıştır. Elde ettiği Birliktelik Kurallarını Graphviz kullanarak görselleştirmiş ve kritik hataları e-mail yoluyla ilgili kişilere iletmiştir. Bu görsellerden hataların en çok hangi tarayıcıda gerçekleştiği, kim tarafından gerçekleştirildiği gibi kurum çalışanlarından kaynaklanan hata verilerinin elde edildiğini gözlemlemiştir.

Altay [15] Atatürk Üniversitesi Diş Hekimliği Fakültesinden edindiği verileri, mekânsal ve zamansal veri madenciliğinde kullanmıştır. Hastalar ile ilgili olarak bölge ve poliklinik bazlı sınıflandırma yapılmış ve hastaların hastalık grupları belirlenmiştir. Veri setinde yer alan aykırı veriler de elde edilmiştir. Bu verileri geliştirmiş olduğu PETEK-O sisteminde kullanmıştır. Bu sistem ile hastaların tedavi sürecinde izlemiş olduğu yolun veri madenciliği yöntemi ile incelenmesini sağlamıştır.

Yangın [16] çalışmasında, eğitim yayıncılığı yapan bir firmanın 2010-2015 yılları içinde dağıtım yapmış olduğu illere göre satış verilerini kapsayan bir veri setini kullanmıştır. Uzman görüşlerini alarak satışları etkileyen 7 önemli etkeni (Tüfe, Üfe, Satış Hacmi, Hane Eğitim Giderleri, Dolar Kuru, Müşteri Sayısı) belirlemiştir. Bağımsız değişkenlerin satış üzerindeki etkileri, Yapay Sinir Ağları yöntemi kullanılarak bulunmuştur. Bazı illerde aile içi gelirlerin yüksekliği ve öğrenci sayısının fazlalığı nedeni ile satışların daha yüksek olduğu gözlemlenmiştir. YSA tekniği ile elde edilen tahmini veriler ile gerçek veri değerlerini karşılaştırmış (Tahmini ve gerçek değerler arasındaki sapma: 3.87) ve YSA'nın daha başarılı olduğu belirtmiştir.

Freitas vd [17], trafik gürültüsünü girdi parametresi olarak alıp, veri madenciliği tekniklerini uygulamışlardır. Bu çalışmada, her bir faktörün önemini değerlendirmek için, asfalt, doku, kaplama sıkıntıları ve araç hız türüne dayalı olarak, lastik-asfalt gürültüsünün akustik ve psikoakustik göstergelerini tahmin etmek için modeller geliştirmiş ve kullanılmışlardır. Veri madenciliği, özellikle Yapay Sinir Ağları ve Destek Vektör Makineleri kullanılarak hem akustik hem de psikoakustik gürültü göstergelerinin iyi tahmin

kapasitesine sahip modellerini elde etmişlerdir. Geliştirilen modellerin uygulanabilirliğini 3 tip asfaltın kullanılması ile sınırlandırmışlardır. Diğer bir sınırlayıcı faktörün ise yol yapımında kullanılan malzemenin ülkeden ülkeye farklılık göstermesi olduğunu belirtmişlerdir. YSA ve SVM modelleri ses yüksekliğini tahmin etmekte benzer bir davranışa sahiptir. Ancak SVM'nin keskinliği tahmin etmekte daha iyi olduğunu söylemişlerdir. İki modelde de yüzey pürüzlülüğü tahmin edilememiştir. Gürültüdeki en etkili faktörlerin hız ve yol yüzey bozukluğu olduğu belirtilmiştir.

Sene vd. [18] tıbbi karar verme süreci için koşullu tercihler, kanıt teorisi ve veri madenciliğinin yüksek fayda sağlayan modellerinden yararlanmışlardır. Bu amaç ile karar vericiye yardımcı olmak için Koşullu Tercihler Tabanı (Conditional Preferences Base-CPB) kullanmışlardır. Daha sonra kanıt teorisinin desteklediği belirsizliği bütünleyen Dempster-Shafer ontolojisi inşa etmişlerdir. Uçuş içi elektronik sağlık kayıt verilerini kullanmışlardır. Bu kayıtlar tıbbi olayların yönetilmesini sağlayan öğeleri içermektedir. Belirsizlik toleransını yönetmek için İnanç Füzyon Algoritması geliştirmişlerdir. Çalışmada belirsizlik yönetimini, CPB ve Klinik Akıl Yürütme Modellerine entegre etmeyi ve veri madenciliği kümeleme yöntemini kullanarak olay yönetiminin belirleyicileri arasındaki ilişkileri kullanmayı amaçlamışlardır. Uçuş içi tıbbi olayların ne olduğu ve belirsizlik sorunları altında nasıl bir karar olduğu irdelenerek veri madenciliğinin ilk aşamasını oluşturmuşlardır. Daha sonra bu belirsizliğin yönetilmesi için uçakta karar verme süreçlerini desteklemek amacıyla kanıtsal ve ontolojik bir mantık önermişlerdir.

Liu vd [19] gerçek dünya problemleri için pertürbasyon temelli gizliliği koruyan veri madenciliği uygulanabilirliğini incelemişlerdir. Bu yöntemde, gizliliğe duyarlı verilere rastgele gürültü eklendiğini belirtmişlerdir. Daha sonra yeniden yapılandırma işlemi yapılmış ve bu yeni yapılandırılmış dağıtımı veri madenciliğinde kullanmışlardır. Çalışmada bireylerin kendi mahremiyet seviyelerini seçmelerini sağlayan, bireysel olarak uyarlanabilen bir model sunmuşlardır. Yeni modeli hem gerçek dünya veri setleri hem de yapay veri setlerinde denemişlerdir ve sonucunda basit ama etkili, verimli bir teknik olduğunu belirtmişlerdir. Gerçek dünya veri setlerine uygulama yapıldığı zaman, yeni yapılanmanın sorun olabileceği için yeniden yapılandırma adımını atlayan ve veri madenciliği sonuçlarını doğrudan hesaplayan yöntemler önermişlerdir. Gelir verilerinden rastgele seçilen farklı boyuttaki veri kümelerine, Karar Ağacı, Sade Bayes, Yapay Sinir Ağı ve Çok Katmanlı Algılayıcı uygulamışlardır.

Orijinal veri seti ve gürültü değeri eklenen veri setlerinden elde edilen sonuçları inceleyerek veri modelciliği başarımını azaltmadan kullanıcılara daha fazla gizlilik sunabileceklerini belirtmişlerdir.

Compieta vd [20], mekânsal ve zamansal veri madenciliği keşfi ve görselleştirmesi üzerine çalışmışlardır. Bu veri setlerinin oldukça büyük ve analizinin zor olduğunu belirten araştırmacılar, büyük boyuttaki mekânsal-zamansal veri kümeleriyle ilgilenebilmek, veri madenciliği sürecini etkin bir şekilde destekleyebilmek ve veri setinin mekânsal-zamansal boyutlarını ele alıp sonuçlarını görselleştirip yorumlamak için bir veri madenciliği sistemi önermişlerdir. Bu sistemin Apriori algoritmasının uyarlanmış bir versiyonu olduğunu söylemişlerdir. Veri seti olarak Hurricane Isabel (2003'te Atlantik kasırga sezonunun kategori 5 kasırgası) ile ilgili veriler Hava Araştırması ve Tahmin modeli tarafından üretilmiş ve araştırmacılar tarafından kullanılmıştır. Sonuçları yorumlayabilmek amacıyla görsel tekniklere odaklanmışlardır. İki bağımsız görselleştirme aracı geliştirmişlerdir. Bu araçlardan ilki Google Earth uygulamasıyla, belirli coğrafi bölgeyle ve ilgili özelliklerle ilişkilendirmeyi sağlamışlardır. Bir diğerinin ise veri madenciliği aşamasında çıkarılan kurallar ile verileri kıyaslamaya olanak sağlayan karmaşık bir etkileşim sunan Java 3D uygulaması olduğunu söylemişlerdir. Böylece veri kümesindeki yapı ve ilişkileri anlamayı kolaylaştırmışlardır.

Song vd [21] intihar oranı oldukça yüksek olan Kore'de, ergen gençler arasındaki intihar ile ilişkili kelimeleri içeren sosyal ağ sitelerinde yer alan web tabanlı belgeleri veri madenciliği yöntemleri kullanarak analiz etmişlerdir. 1 Ocak 2011-31 Aralık 2012 tarihleri arasında Güney Kore'deki 163 sosyal medyada 2 yıl boyunca 2,35 milyardan fazla mesaj ve 99,693 intiharla ilgili belge metin madenciliği ve görüş madenciliği kullanılarak incelenmiştir. Çalışmanın sonucunda; sınıf baskısı, zayıflık, zorbalık mağdurları, hastalığa ilişkin kaygılar, maddi zorluklar ve depresyon ile ilgili kaygıların intihar riskini arttıran en etkili yordayıcılar olduğunu belirtmişlerdir. Ergenlerin siber alanda bu duygularını diğer kullanıcılara yansıttıkları gözlemlenmiştir. 2011 yılında online ifadelerin toplamda %23.1, 2012 yılında ise %22.2 intihar riskini gösterdiği sonucuna ulaşmışlardır.

Medvedev vd. [22] verilerin karmaşıklığı sebebi ile veri madenciliği yöntemlerinin bulut teknolojileri kullanılarak uygulanması gerektiğini belirtmişlerdir. Bu amaç ile Bulut

sisteminden esinlenerek DAMIS olarak isimlendirdikleri yeni bir web tabanlı çözüm önermişlerdir.

DAMIS mimarisini, kolay erişilebilirlik, kullanılabilirlik, ölçeklenebilirlik ve çözümün taşınabilirliğini sağlamak açısından faydalı olacağını düşünerek tasarlamışlardır. Geniş bir uygulama yelpazesine sahip bu çözümün bilgi keşfi aşamasında verilerden derinlemesine bilgi almayı sağlayacağını belirtmişlerdir. Veri seti olarak UCI'den Meme Kanseri veri setini (9 özellik 699 veri) kullanmışlardır. DAMIS'te öncelikle veri dosyasını sisteme yüklemişlerdir ve eksik verilerin olduğu kayıtlar temizlenmiştir. Böylece veri önişleme aşamasını gerçekleştirmişlerdir. Daha sonrasında verileri sınıflandırmak için Rastgele Karar Ormanı (RDF)'i ve Çok Katmanlı Algılayıcıyı (Multi-Layer Perceptron-MLP)'yi ve karmaşık verileri anlamlı şekilde sunmak için, farklılığa dayalı görselleştirme yöntemi SMACOF'u kullanmışlardır. Araştırmacılar DAMIS'in diğer veri madenciliği araçları ile rekabet edebilecek seviyede olduğu ve veri sınıflandırma, kümeleme ve boyut azaltma sorunlarını çözmede kullanışlı olacağı sonucuna ulaşmışlardır.

Jing vd [23] dinamik veri madenciliği için artırımlı bir öznitelik azaltma yöntemi geliştirmişlerdir. Öznitelik azaltmanın veri madenciliği için önemli bir önişlem adımı olduğunu belirten araştırmacılar, birçok gerçek verinin zamanla dinamik olarak değişebileceğini, bu nedenle de karar verme sistemlerinde nesnelerin ve katkılarının da dinamik olarak değişeceğini ifade etmişlerdir. Klasik nitelik azaltma yöntemlerinin, dinamik karar sistemleri ile başa çıkmada yetersizliği sebebi ile karar sistemlerinin nesneleri ve nitelikleri dinamik olarak değiştiğinde, bilgi işlem düşüşünün artımsal mekanizmalarını sunmuşlardır. Daha sonra öznitelikler ve nesneler eşzamanlı olarak arttığında bilgi düşüşünü güncellemek için yöntemler geliştirmişlerdir. Son olarak, önerilen artımlı nitelik azaltma yöntemlerini onaylamak için bir dizi deney gerçekleştirmişlerdir. Matrise dayalı bir azaltma yaklaşımı (IAMRCD), Matris dışı metot kullanarak hesaplama redüksiyonunun artımlı indirgeme yöntemi (IARCD) ve Sezgisel Azaltma Yaklaşımı (Heuristic Reduction Algorithm-HRA) modellerini açıklamışlardır. IARCD, bazı özniteliklerin ve nesnelerin aynı anda değiştirilmesini sağlayan karar sistemleri için verimli bir veri önişleme aracı olarak kullanılabilir sonucuna ulaşmışlardır.

Cayci vd [24] her yerde bulunan bir veri madenciliği algoritmasının yürütülmesini otomatik olarak yapılandırma problemi üzerine çalışmışlardır. Algoritma, şartların sıkılıkla

değiştiği ve bilgi işlem kaynaklarının çoğunlukla ciddi şekilde sınırlı olduğu bir ortamda yürütüldüğü için çözümlerinde kaynak ve koşullar dikkate alınacak bir şekilde yapılandırma kararları üretmişlerdir.

Veri madenciliği algoritmasının yürütme davranışını geçmiş uygulamaları inceleyerek analiz etmişlerdir. Böylece kaynakların, koşulların ve veri madenciliği kalitesindeki parametre ayarlarının etkilerini incelemişlerdir. Sınıflandırma modelinin, algoritmanın yürütülme davranışını tahmin etmede uygun olacağını düşünerek, Karar Ağacı sınıflandırıcısını seçmişlerdir. Problemin zorluklarına çözüm bulmak için, makine öğrenmesine dayanan bir yaklaşım önermişlerdir. Bu önerme ile farklı durumlarda veri madenciliği algoritmasının davranışını, algoritmanın konfigürasyonu için kullanılacak şekilde modellemişlerdir. Davranış modeli bileşenlerini ve sınıf etiketi dönüşümlerini resmi olarak tanımlanmışlardır ve önerdikleri yaklaşımı deneysel olarak da doğrulamışlardır. Bu yaklaşımda, veri madenciliği kalitesi davranış modelinin bir parçasıdır. Yapılandırma kalitesinin hedeflerine ulaşıp ulaşmadığını ve hedeflere ulaşmadığı durumlarda yeni bir veri madenciliği modeli oluşturarak değişen koşullara uyum sağlamasının mümkün olduğunu belirtmişlerdir.

Chen vd. [25] büyük veri madenciliği ve bulut bilişim temelli bir hastalık tanı ve tedavi öneri sistemiyle ilgili çalışmışlardır. Farklı belirtilerdeki çeşitli semptomlara göre bir hastalık için uyumlu tedavi şemalarını sağlamanın önemine değinmişlerdir. Araştırmacılar, birçok sınıflandırma yönteminin hastalığın doğru şekilde sınıflandırmasında yetersiz olabileceğini ve farklı hastanelerin ve deneyimsiz doktorların hızlı bir şekilde tanı koymakta zorlanacağını belirtmişlerdir. Bu sebeple Hastalık Teşhis ve Tedavi Öneri Sistemi (Disease Diagnosis and Treatment Recommendation System -DDTRS)'yi önermişlerdir. İlk olarak, hastalık semptomlarını doğru tanımlamak ve hastalık belirti kümelemesi için Yoğunluk Kümeleme Analizi (DPCA) algoritması önermişlerdir. İkinci aşamada hastalık teşhis ve hastalık tedavi kuralları ile ilgili ilişkileri analiz etmek için Apriori algoritmasını kullanmışlardır. Bu önerimi yüksek performans ve düşük gecikme olacak şekilde gerçekleştirebilmek için Apache Spark bulut platformunu kullanarak DDTRS için paralel bir çözüm uygulamışlardır. Yaptıkları kapsamlı deneysel çalışmalar sonucunda önerilen DDTRS'nin hastalık semptom kümelemesini etkili bir şekilde gerçekleştirdiğini ve hastalık teşhislerini etkin bir şekilde ortaya koyduğunu gözlemlemişlerdir.

Sanmiquel vd. [26] veri madenciliği tekniklerini kullanarak İspanya'daki madencilik kazalarının incelenmişlerdir. İspanyol madencilik sektöründe 2003-2012 yılları arasında meydana gelen yaklaşık 70.000 iş kazası ve ölüm raporlarından oluşan bir veri tabanını kullanarak bu kazaların ana nedenlerini analiz etmişlerdir.

Bayes Sınıflandırma Algoritmaları, Karar Ağaçları ve Olasılık Tabloları gibi istatistiksel veri madenciliği tekniklerini kullanmışlardır. Bu analizleri WEKA yazılımını kullanılarak gerçekleştirmişlerdir ve belirli kurallara dayalı davranışlar elde etmişlerdir. Bu kurallardan yaralanma ve ölümleri azaltmak için uygun ödeme politikalarının geliştirilmesine yardımcı olacak sonuçlar elde etmişlerdir. Ayrıca çalışmada önleyici nedenler, yer, boyut, fiziksel aktivite, önleyici organizasyonlar, deneyim ve yaş gibi faktörlerin kazaların oluşumunda etkili olduğunu belirtmişlerdir.

Yuan ve Chen [27] çalışmalarında, hükümetler ve işletmeler için iyi bir bilgi güvenliği değerlendirmesi metodu bulmak ve bilgi güvenliği denetim departmanlarının etkili kararlar vermelerine yardımcı olmak amacıyla bir yöntem önermişlerdir. Veri madenciliği teknolojisine dayanan verilere dayalı bilgi güvenliği konusunda bir korelasyon analizi uygulamışlardır. İş Birliği Analizi Modeli kullanarak e- Devlet bilgi güvenliği için bilgi güvenliği risk değerlendirmesi yapmışlardır. K-means algoritması, Karar Ağaçları ve Yapay Sinir Ağları teknikleri uygulamışlardır. Korelasyon Analizi modelini, büyük veri kümelerinden anlamlı bağlantıyı gizlemek için adapte etmişlerdir.

Naganathan vd [28], tersine mühendislik için veri madenciliği teknikleri ile enerji tüketimi ve talep modellerinin öğrenilmesi üzerine çalışmışlardır. Geleneksel tekniklerin yetersizlikleri vurgulanmış ve tersine mühendislik için önerilen veri madenciliği algoritmalarıyla, ekipman, fiziksel sistemler ve binaların enerji kullanım modellerini birleştirerek enerji talebinin tahminini geliştirebilecek bir öneri sunmuşlardır. Bunun için öncelikle büyük karmaşık enerji verilerinden veri örneklerini analiz edip, daha sonra tersine mühendislik için bir dizi hesaplama ve etkin veri madenciliği algoritması sunmuşlardır. Araştırmanın ilk aşamasında enerji arz-talep özelliklerine katkıda bulunan değişkenleri tanımlamışlardır. Değişenleri algoritmanın amaçlanan çıktıları için test etmişlerdir. İnsan davranışlarının tüm değişkenlerinin; elektrik, ısıtma, soğutma, ısı endeksi ve iklim değişiklikleri, sıcaklık ve nem gibi doğal faktörleri içeren diğer ana değişkenlerle ilgili

olduğunu belirten araştırmacılar bu verilerin analizi için Regresyon analizi, Yapay Sinir Ağları ve Bayes Teorisi'ni kullanmışlardır.

Amos vd. [29] üretim sistemlerinin yenilikçi tasarımı ve analizi için çok amaçlı optimizasyon ve veri madenciliği tekniklerini kullanmışlardır. Araştırmacılar, çeşitli tasarım hedeflerinin tek bir matematiksel fonksiyonda doğrusal olarak birleştirildiği üretim sistemleri tasarımı için sıradan optimizasyon yaklaşımlarından farklı olarak, çoklu tasarım alternatifleri üretebilen ve performanslarını verimli bir bölüme ayıran çok amaçlı optimizasyonun, tasarımcının daha eksiksiz bir resme sahip olmasını sağlayabileceğini ifade etmişlerdir. Üretilen çok sayıda optimal tasarım alternatifi nedeniyle, tasarım değişkenleri ve hedefler arasındaki ilişkiler hakkında bilgi elde edebilmek için veri madenciliği algoritmalarına artan bir veri kümesi oluşturmuşlardır. Entegre optimizasyon ve veri madenciliği yaklaşımı için ayrı üretim sistemlerinin tasarımının oluşturulduğu belirli zorlukları ele alan araştırmacılar, gerçek dünyadaki üretim hattı tasarım örneğiyle gösterilen bu zorlukları karşılamak için geliştirilmiş yeni bir etkileşimli veri madenciliği algoritması önermişlerdir. Ardışık Desen Madenciliğinin (Sequential Pattern Mining-SPM) genişletilmiş bir versiyonu olan Esnek Desen Madenciliği (Flexible Pattern Mining-FPM) adlı yeni bir algoritma sunmuşlardır.

Govada ve Sahay [30] 2020'de 60PB arşivlenmiş verinin gökbilimcilere açık olacağını belirtmişlerdir. Ancak bu verileri analiz etmenin, karmaşık coğrafi olarak dağıtılmış arşivlerden verileri merkezi bir siteye indirmek ve daha sora yerel sistemlerde analiz etmek gerekeceği için zor olacağını vurgulamışlardır. Bu sebeple dağıtılmış veri madenciliği katmanının VO'ya eklenmesi, bilginin astronomlar tarafından ham verilerin yerine indirgenebileceği ve astronomların ya indirilen bilgiden önceki veriyi yeniden oluşturabileceği ya da daha ileri analiz için bilgiyi doğrudan kullanabileceği mantığından yola çıkmışlardır. İletim maliyetini en aza indirmek için mevcut düğümler arasında hesaplamayı en uygun şekilde dağıtmak amacıyla Dağıtılmış Yük Dengeleme Ana Bileşen Analizini (Distributed Load Balancing Principal Component Analysis-DLPCA) kullanmışlardır. Farklı gözlemevlerinde saklanan astronomik verileri indirmek için DLPCA kullanan etkin ve ölçeklenebilir Dağıtılmış Veri Madenciliğini (Distribute Data Mining-DDM) önermişlerdir.

Karabatak [31] yapmış olduğu çalışmada meme kanserinin kadınlarda en sık görülen kanser türü olduğunu ve bu yüzden otomatik kanser tespit sistemlerinin talep edildiğini belirtmiştir. Bu amaçla çalışmasında yeni bir Sade Bayes sınıflandırıcı (Ağırlıklı Sade Bayes) önermiş ve meme kanseri tespiti için kullanmıştır. Deneylerde UCI'den aldığı meme kanseri veri setini kullanarak, 5 katlı çapraz geçerlilik testi uygulamıştır. Ayrıca duyarlılık, özgünlük ve doğruluk gibi performans değerlendirme teknikleriyle çalışmıştır. Duyarlılık %99.,11, özgünlük %98.25 ve doğruluk %98.54 olarak bulmuştur.

Literatürde yer alan diğer veri madenciliği algoritmalarının (C4.5, RIAC, LDA, SVM, NEFCLASS, SFC, ME, NB) bu veri seti üzerindeki başarım değerlerini de hesaplamıştır. Sonuç olarak Ağırlıklı Sade Bayes (Weighted Naive Bayesian- W-NB)'in, doğruluk değerinin daha yüksek olduğu sonucuna ulaşmıştır.

Bu tez çalışmada veri madenciliği sınıflandırma tekniklerinden olan Sade Bayes ve Ağırlıklı Sade Bayes algoritmaları incelenmiştir. Sade Bayes algoritmasındaki veri setinin özniteliklerinin sonuca olan etkisinin aynı olması sorununu çözen W-NB algoritmasının ağırlıklarının optimize edilmesi için iki yöntem önerilmiştir. Bunlardan ilkinde en yüksek doğruluk değeri veren ağırlık değerinin sabit tutulması ve diğer ağırlık değerlerinin de sabitlenene kadar aranması işlemi yapılmıştır. Bu yöntem Hızlandırılmış Ağırlıklı Sade Bayes (FW-NB) olarak adlandırılmıştır. İkinci yöntem olarak ağırlık değerlerinin bulunmasında Genetik Algoritmalar kullanılarak ağırlıkların optimize edilmesi sağlanmıştır. NB, W-NB ve önerilen iki yöntemden elde edilen sonuçlar tartışılmıştır.

Toplam altı bölümden oluşan bu tezin ikinci bölümde veri madenciliği ve tekniklerinden bahsedilmiştir. Üçüncü bölümde Sınıflandırma algoritmalarından olan NB, W-NB ve FW-NB yöntemleri hakkında bilgiler verilmiştir. Dördüncü bölümde beş farklı veri seti belirlenmiş ve bu veri setlerine NB, W-NB, FW-NB algoritmaları uygulanarak hassasiyet, özgünlük ve doğruluk değerleri kıyaslanmıştır. Beşinci bölümde Genetik algoritmalar hakkında kısa bir bilgi verilerek, ağırlıkların nasıl optimize edildiği açıklanmıştır. Ayrıca W-NB algoritması ve bu yöntemden elde edilen değerlerin performansları incelenmiştir. Altıncı bölümde ise tezde yer alan çalışmalar değerlendirilmiş ve yorumlamıştır.

2. VERİ MADENCİLİĞİ

Disiplinlerarası bir konu olarak veri madenciliği, birçok farklı şekilde tanımlanmıştır. Bu tanımlardan bazıları şu şekildedir;

Birçok kişi veri madenciliğini, "veriden bilgi keşfi" ile eş anlamlı olarak kullanırken, diğerleri veri madenciliğini, yalnızca bilgi keşif sürecindeki veri temizleme, veri entegrasyonu, veri seçimi, veri dönüşümü, model değerlendirmesi ve bilgi sunumu gibi önemli bir parametre olarak görmektedir [32].

Veri madenciliği, veri setlerinden yararlı bilgilerin çıkarılması, veri içindeki gizli örüntülerin tespit edilmesi ve değişkenler arasındaki ilişkilerin belirlenmesi işlemidir. Ancak veri madenciliği sadece bilgisayar teknolojilerine bağlı değildir. Aynı zamanda insanların, veri toplama, temizleme, model oluşturma, test etme ve uygulama gibi aşamalardaki yorumları da oldukça önemlidir [33].

Dünya genelinde veri tabanlarında depolanan veri miktarının her 20 ayda bir yaklaşık olarak iki katına çıkması sonucu önem kazanan veri madenciliği, verideki kalıpları keşfetme ve verileri analiz ederek problemleri çözme süreci olarak tanımlanmaktadır [34]. Veri madenciliği, örüntü tanıma, istatistik, matematik, makine öğrenmesi gibi farklı alanlardaki tekniklerin bir araya gelmesi sonucu oluşan, veri tabanlarından bilgi çıkarma işlemidir [35].

Veri madenciliği ortaya çıkışından bugüne kadar birçok farklı şekilde adlandırılmıştır. Bunlardan bazıları Veri Tabanı Bilgi Keşfi (Knowledge Discovery in Databases-KDD); büyük veri tabanlarında yararlı bilgilerin yarı otomatik olarak keşfedilmesi sürecidir [36]. Bilgi Çıkarımı (Knowledge Extraction-KE); genellikle hacmi, çeşitliliği, karmaşıklığı nedeniyle uzmanlar tarafından işlenemeyen ve anlaşılamayan büyük veri setlerinin bir çözümüdür [37]. Veri ve Örüntü Analizi; veri analizi için prosedürler, bu prosedürlerin sonuçlarını yorumlama teknikleri, veri analizi yapmak için bu verilerin daha kolay ve hassas erişimini sağlayan bir bütündür [39]. Veri Tarama (Data Dredging); istatistiksel olarak anlamlı bir şekilde sunulabilecek veriyi ortaya çıkarmaktır [40].

Veri Yakalaması (Data Fishing), veri tarama olarak da bilinen bilgi keşfi sağlayan veriye dayalı bir analiz ve sunumdur [41]. Bunun yanı sıra literatürde, veri madenciliğinin veri arkeolojisi, bilgi üretimi ve bilgi hasadı olarak adlandırıldığı da göze çarpmaktadır [7].

2.1. Veri Madenciliğinin Tarihsel Gelişimi

Veri madenciliği bir terim olarak 1980'lerden itibaren kullanılmaya başlamasına rağmen bu alana yön veren çalışmalar çok daha eskiye dayanmaktadır. Veri madenciliğinin tarihsel gelişim sürecini etkileyen olaylar ve çalışmalar şu şekilde sıralanabilir [42]:

1763 yılında Thomas Bayes veri madenciliği sınıflandırma yöntemlerinden olan Bayes Teoremini oluşturmuştur. Bu teorem ile olasılıklara dayalı karmaşık gerçeklerin anlaşılmasını sağlamıştır.

1805 yılında Adrien-Marie Legendre ve Carl Friedrich Gauss istatistiksel Regresyon Analizini gerçekleştirmişlerdir. Regresyon Analizi yapmalarının amacı değişkenler arasındaki ilişkileri tahmin etmektir.

1936 yılında Alan Turing günümüzdeki en etkili teknoloji olan bilgisayar biliminin yaratıcısı olarak ilk çalışmalarını gerçekleştirmiştir. Sayısal sayılar üzerine yaptığı çalışmalar ile evrensel makine fikrini tanıttı. Böylece günümüz bilgisayarlarının temelini oluşturdu.

1943 yılında Warren McCulloch ve Walter Pits "Sinirlerin aktivitesi için fikirlerin mantıksal hesabı" adlı çalışmaları ile sinir ağlarının kavramsal modelini oluşturmuşlardır. Böylece bir ağdaki nöron fikrini ortaya atmışlardır. Her bir nöronun girişleri almak, girişleri işlemek ve çıktı üretmek gibi 3 olayı gerçekleştirebileceğini belirtmişlerdir.

1965 yılında Lawrence J Fogel, evrimsel hesaplama ve insan faktörleri analizinde öncülük eden Decision Science Inc. adlı bir şirket kurmuştur. Gerçek problemleri çözmek için özel olarak evrimsel hesaplama yöntemlerini uygulayan ilk şirket olmuştur.

1970'lerde veri tabanı yönetim sistemleri ile büyük miktardaki verileri depolamak ve sorgulamak mümkün hale gelmiştir. 1975'te John Henry Holland "Doğal ve Yapay Sistemlerde Adaptasyon" adlı genetik algoritmalar üzerine bir kitap yayınlamıştır.

Bu, çalışma alanını başlatan, veri madenciliği teorik temellerini sunan ve uygulamaları araştıran bir kitaptır.

1980'lerde HNC'deki konu uzmanlarının karmaşık algoritmalardaki, ilişkilerin ne anlama geldiği hakkında fikir yürütmesine olanak sağlayan, verilerden "bilgi" üretebileceğini belirterek "Veri Madenciliği" terimini kullanmıştır. 1989 yılında Gregory Piatestsky Shapiro tarafından Veri Tabanlarında Bilgi Keşfi (KDD) kavramı ortaya atılmıştır [38]. 1980'lerin sonunda sınıflandırma, kümeleme, aşırı uç analizi, birleşme, ilişkilendirme ve korelasyon, eğilim ve sapma analizi gibi kavramlar analizler için kullanılmaya başlandı [32].

1990'larda kurumlarda veri madenciliği kavramı kullanılmaya başlanmıştır. 1992'de Bernhard E. Boser, Isabelle M. Guyon ve Vladimir N. Vapnik verilerin analiz eden ve sınıflandırma, regresyon analizi için kullanılan kalıpları tanıyan denetimli bir öğrenme yaklaşımı sağlayan Destek Vektör Makinelerini geliştirmeye yönelik bir öneri sunmuşlardır. 1993 yılında ise Agwaral, Immielinski ve Swami Birliktelik Kuralı kavramını ortaya çıkardılar ve birtakım değişiklikler ile bir algoritma (AIM) önerisi sundular [35].

2001 yılından itibaren 'veri bilimi' terimi bağımsız bir disiplin olarak ortaya çıktı. 2017 yılında keşfedilen en etkili tekniklerden biri olarak Derin Öğrenme Tekniği göze çarpmaktadır. Derin öğrenme, çoklu soyut katmanlardaki verilerin temsilini öğrenmek için çoklu işlem katmanlarından oluşan bir hesaplama modelidir [43].

2.2. Veri Madenciliği Kullanım Alanları

Veri madenciliği birçok alanda bilgi çıkarım sürecinde kullanılmaktadır. Bunlardan bazıları şu şekilde sıralanabilir [7];

Bankacılık sektöründe;

- ✓ Kredi kartı talebi bulunan müşterilerin değerlendirilmesi sürecinde,
- ✓ Bireylerin yapmış oldukları harcamalara göre kampanyaların belirlenmesinde,
- ✓ Kredi kartı sahtekarlıklarının belirlenmesinde,
- ✓ Risk yönetiminde

Pazarlama sektöründe;

- ✓ Müşterilerin hangi ürün ya da hizmeti almaya eğimli olduğunun belirlenmesi sürecinde,
- ✓ Müşterinin satın alma potansiyelinin belirlenmesinde,
- ✓ Ürünlerin satış tahmininde,
- ✓ Sektördeki müşteri-çalışan ilişkilerinin denetlenmesinde,
- ✓ Mevcut olan müşterilerin sayısının arttırılmasında kullanılmaktadır.

E-Ticaret sektöründe;

- ✓ Elektronik ticaret sitelerine yapılan saldırıların belirlenmesinde,
- ✓ Dolandırıcılıkların tespitinde,
- ✓ Satış verilerinin incelenmesinde,
- ✓ En çok satılan ürünlerin belirlenmesinde tercih edilmektedir.

Tıp ve genetik biliminde;

- ✓ Genetik hastalıkların bulunmasında,
- ✓ Birçok hastalığın tespit edilmesinde,
- ✓ Tanısı konulmuş hastalıkların tedavi sürecinin belirlenmesinde,
- ✓ Virüslerin sınıflandırılmasında,
- ✓ Gen haritalarının analizinde kullanılmaktadır.

Coğrafi bilgi sistemlerinde;

- ✓ Bölgelerin sınıflandırılmasında,
- ✓ Kentlerde sunulan hizmetlerin belirlenmesinde,
- ✓ Suç oranı analizlerinde,
- ✓ Yerleşim yerlerinin düzgün şekilde konumlanmasında kullanılmaktadır.

Uzay bilimlerinde;

- ✓ Astrolojik verilerin incelenmesinde,
- ✓ Galaksilerin keşfedilmesinde,
- ✓ Yıldız gibi gök cisimlerinin bulundukları yerlere göre gruplandırılmasında tercih edilmektedir.

Sosyal bilimlerde;

- ✓ Kamuoyu görüşlerini incelemede,

- ✓ Genel yönelimlerin belirlenmesinde,
- ✓ Davranışsal ve psikolojik süreçlerin incelenmesinde kullanılmaktadır.

Eğitim bilimlerinde;

- ✓ Öğrenci davranışlarının analiz edilmesinde,
- ✓ Sınav verilerinin incelenmesinde,
- ✓ Öğrencilerin başarım oranlarının test edilmesinde,
- ✓ Başarısızlık sürecindeki sebeplerin analizinde kullanılmaktadır.

Bilim ve Mühendislik alanlarında;

- ✓ Enerji arz-talep dengelerinin belirlenmesinde,
- ✓ Cep telefonu iletişim protokolü olan (GSM) şebekelerinin performansında,
- ✓ Bir baraj veya binanın yapımında yaşam alanlarının verimli kullanılmasının ve en uygun inşaatın yapılmasının sağlanmasında [44],
- ✓ Endüstri mühendisliği uygulamalarında ve daha birçok mühendislik alanında veri analiz sürecinde tercih edilmektedir.

2.3. Veri Madenciliği Bilgi İşlem Süreci

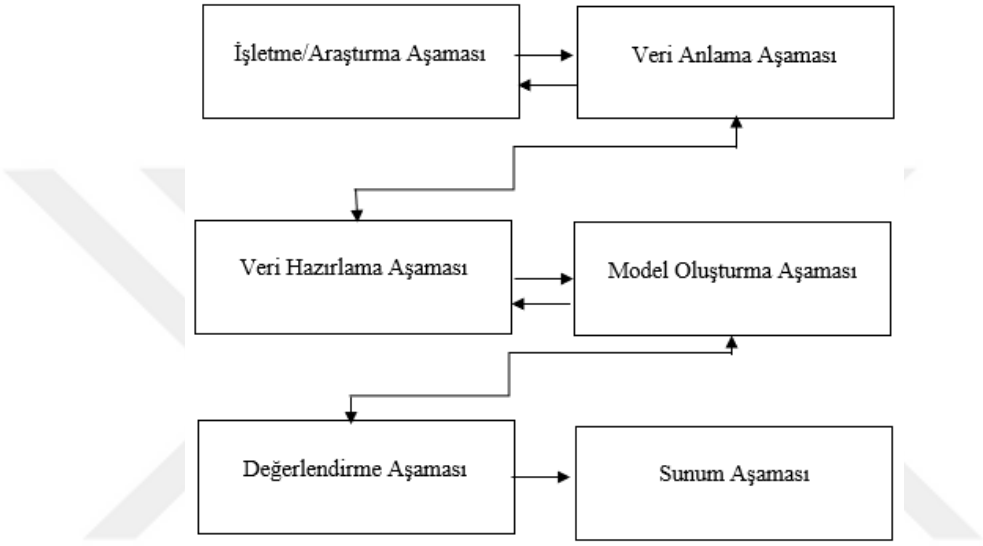
Veri Madenciliği için Sektörler Arası Standart Süreç (CRISP-DM) Daimler-Chrysler, SPSS ve NCR'yi temsil eden analistler tarafından 1996 yılında geliştirilmiştir. Böylece işletme ve araştırma birimlerinin genel sorun çözme stratejilerine veri madenciliği tekniklerinin yerleştirilebilmesi için tescilli olmayan ve serbestçe kullanılabilen bir standart süreç sunulmuştur. CRIPS-DM'ye göre 6 aşamadan meydana gelen bir yaşam döngüsü vardır [45].

CRISP-DM'de belirtilen aşamalar şu şekildedir;

- ✓ İşletme/araştırma aşaması,
- ✓ Veri anlama aşaması,
- ✓ Veri hazırlama aşaması,
- ✓ Model oluşturma aşaması,
- ✓ Değerlendirme aşaması,
- ✓ Sunum aşaması.

Bu döngüde sonraki aşama genellikle önceki aşama ile ilişkili sonuçlara bağlıdır. Şekil 2.1'deki gibi fazlar arasındaki en önemli bağımlılıklar oklar ile gösterilmiştir. Modelin davranışına ve özelliklerine bağlı olarak, model değerlendirme aşamasına geçmeden önce daha fazla geliştirme için veri hazırlama aşamasına dönmek gerekmektedir.

Analistleri değerlendirme aşamasında karşılaşılan bir sorunu düzeltebilmek için önceki aşamalardan herhangi birine geri gönderebilir [45].



Şekil 2.1. Veri madenciliği bilgi işlem süreci

2.3.1. İşletme/ Araştırma Aşaması

- ✓ Proje hedeflerinin ve gereksinimlerinin bir bütün olarak ortaya konulduğu aşamadır.
- ✓ Hedefler ve kararlar, veri madenciliği problemine uygun hale dönüştürülür.
- ✓ Son olarak, hedeflere ulaşmak için bir ön strateji hazırlanır.

2.3.2. Veri Anlama Aşaması

- ✓ İncelenen problem için gerekli olan veriler toplanır.
- ✓ Verileri tanımak için keşifsel veri analizi kullanılır.
- ✓ Verilerin kalitesi değerlendirilir.
- ✓ Son olarak işleme geçirilebilir alt kümeler belirlenir.

2.3.3. Veri Hazırlama Aşaması

- ✓ Problem için kullanılacak olan nihai küme hazırlanır.
- ✓ Analize uygun vaka ve değişkenler belirlenir.
- ✓ Gerekirse belirli değişkenler üzerinde dönüşümler gerçekleştirilir.
- ✓ Ham veriler, modelleme araçlarına hazır olacak şekilde temizlenir.

2.3.4. Model Oluşturma Aşaması

- ✓ Uygun modelleme teknikleri seçilir ve uygulanır.
- ✓ Sonucu optimize edebilmek için model ayarlanır.
- ✓ Aynı veri madenciliği problemi için birden fazla yöntem kullanılabilir.
- ✓ Bu aşamada kullanılan veri madenciliği tekniğinin gereksinimleri doğrultusunda veri setini uygun hale getirebilmek için veri hazırlama aşamasına geri dönülebilir.

2.3.5. Değerlendirme Aşaması

- ✓ Kullanılacak olan modeller kalite ve etkinlik açısından değerlendirilmelidir.
- ✓ Modelin işletme/araştırma aşamasındaki hedeflere ulaşıp ulaşmadığı belirlenir.
- ✓ Son olarak veri madenciliği sonuçlarının kullanımı ile ilgili bir karara varılır.

2.3.6. Sunum Aşaması

- ✓ Veri madenciliğinde bilgi keşfi sürecinin son aşamasını oluşturmaktadır.
- ✓ Model oluşturma aşamasının tamamlanması, oluşturulan modelden faydalanılmadığı sürece projenin tamamlandığını göstermez.
- ✓ Rapor oluşturma, başka bir bölüm ile paralel veri madenciliği uygulamaları gibi işlemler yapılmalıdır.
- ✓ Gelecek adımların planlaması gerekmektedir [45].

2.4. Veri Madenciliği Modelleri

Bu bölümde veri madenciliğinin hedeflerini ele almak için kullanılan veri madenciliği modellerine ve bu modelleri içeren algoritmalara genel bir bakış açısı sunulmuştur.

Bilgi keşif sistemleri, sistemin amaçlanan kullanımına göre tanımlanır. Hedef doğrulama ve keşif olmak üzere iki şekilde ayırt edilebilir. Doğrulamada, sistem kullanıcının hipotezini doğrulamak ile sınırlı iken keşif ile sistem nesnelerin gelecekteki davranışlarını tahmin etmeye yarayan bağımsız yeni kalıplar bularak, insan tarafından anlaşılır bir şekilde kullanıcıya sunar [46].

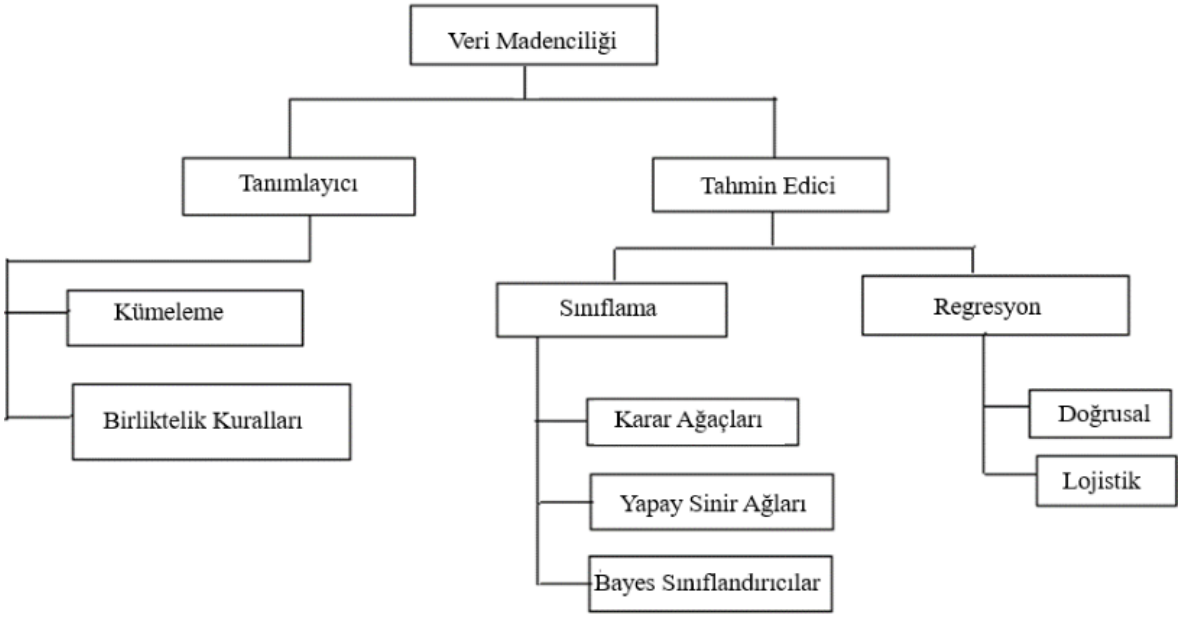
Veri madenciliği gözlemlenen verilerin modellere uydurulması veya modellerin belirlenmesini içerir. Modeller mantıksal ve istatistiksel olacak şekilde ikiye ayrılırlar. İstatistiksel yaklaşım modelde deterministtik olmayan etkilere izin verirken, mantıksal model determinist tiktir. Çoğu veri madenciliği metodu, makine öğrenimi, örüntü tanıma ve istatistikten denenmiş ve test edilmiş tekniklere dayanmaktadır [46].

Veri madenciliğinde kullanılan modeller tahmin edici ve tanımlayıcı olmak üzere ikiye ayrılmaktadır. Tahmin edici modellerin mantığı, sonucu bilinen verilerden yola çıkılarak sonuçları bilinmeyen verilerin elde edilmesine dayanırken, tanımlayıcı modellerin mantığı, var olan verilerdeki faydalı örüntülerin tanımlanmasına dayanır [47].

Veri madenciliği modellerini 3 başlık altında toplamak mümkündür. Bunlar;

1. Sınıflama ve Regresyon Modelleri
2. Kümeleme Modeli
3. Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Şekil 2.2'de veri madenciliği modelleri şematik olarak gösterilmiştir [48].



Şekil 2.2. Veri madenciliği modelleri [48].

2.4.1. Sınıflandırma

Sınıflama problemlerinde yer alan hedef değişkenler iki ya da daha fazla etiket değerine sahiptirler. Bu onları diğer problemlerden ayıran temel özelliktir. Sınıflandırma işleminin yapılabilmesi için önceden belirli sınıflara dahil olan bir veri kümesinin olması gerekmektedir.

Amaç veri kümesine dahil edilmemiş öğeler eklendiği zaman var olan sınıflardan uygun olanına yerleştirilmesidir. Diğer bir deyişle sınıflandırmada amaç, daha önce var olan sınıflandırma etiketleri kullanılarak bir model oluşturmak ve bu oluşturulan modelin yeni kayıtların sınıflandırılmasında kullanılmasını sağlamaktır [49].

Sınıflandırma modeli, çeşitli tahmin değişkenleri ve özellikleri bilinen hedef değişkene ait birçok kaydın incelenmesi ile elde edilir. Model oluşturmak için öncelikle veri setinde yer alan verilerden rastgele seçilen kayıtlar ile alt kümeler oluşturulur. Oluşturulan bu alt kümelerin tamamına eğitim seti denilmektedir. Bu aşama ile denetimli öğrenme işlemi gerçekleştirilir. Gerçekleştirilen bu işlem ile elde edilen model sınıflandırma yöntemlerinde kullanılır. Oluşturulan model sonucu henüz belli olmayan verilerin sınıfların

belirlenmesinde bir tahmin süreci oluşturur. Modelin başarım oranının tespit edilebilmesi için veri setinden rastgele seçilen bir alt veri kümesine uygulanır.

Alt veri kümesine test seti denilmektedir. Eğer modelin başarım oranı yeterliyse, elde edilen modelin mevcut problemin çözümünde etkili olduğu söylenebilir.

Veri madenciliği modellerinden sınıflandırma modelini temel alan birçok algoritma bulunmaktadır. Bunlardan başlıcaları;

- ✓ Karar Ağaçları,
- ✓ Bayes Sınıflandırma,
- ✓ Yapay Sinir Ağları,
- ✓ Genetik Algoritmalar,
- ✓ Diskriminant Analizi,
- ✓ Kaba Küme Yaklaşımı,
- ✓ Destek Vektör Makineleridir [49].

2.4.1.1. Karar Ağaçları

Karar ağacı, çoklu değişkenlere dayalı sınıflandırma sistemlerinin kurulması ve hedef değişkenler için tahmin algoritmalarının geliştirilmesi amacıyla kullanılmaktadır. İlk olarak 1960'larda tanıtılan karar ağaçları veri madenciliği için en etkili yöntemlerdendir. Ayrıca çeşitli disiplinlerde sıklıkla kullanılmaktadır [50].

Karar ağaçları ve kuralları basit bir temsili forma sahiptir. Bu da çıkarılan modelin kullanıcının anlaması için nispeten kolay olmasını sağlar. Ancak belirli bir ağaç veya kural temsilinde yapılacak kısıtlama, modelin işlevsel biçimini önemli ölçüde sınırlayabilir.

Çoklu regresyon modellerinde bulunan sınırlamaların giderilebilmesi amacı ile Karar Ağacı analiz teknikleri geliştirilmiştir. Örnekler sınırlı sayıda sınıfa ayrılır. Her bir nitelik farklı bir düğüm tarafından temsil edilir. Bir Karar ağacındaki düğümler, öznitelik adlarıyla etiketlenir ve yapraklar farklı sınıflarla etiketlenir. Nesneler ise, bir nesnede özniteliklerin değerlerine karşılık gelen kenarları alarak, ağaçtan aşağı doğru bir yol izleyerek sınıflandırılır. Verilerin tanımlanması, sınıflandırılması ve genelleştirilmesi için Karar Ağaçlarının otomatik olarak yapılandırılması ve kullanılması ile ilgili çalışmalar, matematik, istatistik ve mühendislik başta olmak üzere çeşitli disiplinlerde vardır [51].

Kök düğümde yer alan eleman en yüksek karar düğümüdür. En alt yapıyı ise yapraklar oluşturur. Yapraklar ile kök arasında kalan yapılara ise dal denilmektedir.

Her düğümün iki ya da daha fazla dalı olabilir. Bu dallar ikili ise ikili ağaç, ikiden fazla ise çok yollu ağaç denilmektedir. Her dal bir başka yaprak düğüm ile sonlandırılır. Kategorik olan değişkenleri sınıflandırmak için sınıflandırma Karar ağaçları kullanılır iken, sürekli sayısal değişkenler için Regresyon Ağaçları kullanılır [52].

Karar ağaçlarında en önemli sorunlardan biri dallanmanın neye göre yapılacağının belirlenmesi işlemidir. Bu amaçla 3 farklı başlık altında algoritmalar geliştirilmiştir. Bu başlıklar şu şekilde sıralanabilir [53];

- ✓ Entropiye dayalı algoritmalar (ID3, C4.5, SPRINT ve SLIQ Algoritmaları)
- ✓ Sınıflandırma ve regresyon ağaçları (Twoing ve Gini Algoritması)
- ✓ Bellek tabanlı sınıflandırma algoritmaları (k-En Yakın Komşu algoritması)

1970'lerde Morgan ve Sonquist tarafından, Karar ağacı modelini temel alan ilk algoritma Otomatik Etkileşim Dedektörü (AID) tanıtılmıştır. Yöntem en yüksek tahmin değerinin bulunabilmesi için tüm değişkenler arasındaki bütün ilişkileri incelemektedir. Değişkenler arasındaki anlamlı, anlamsız ilişkilerin algoritma tarafından ayırt edilememesi diğer algoritma ve yazılımların gelişmesine sebebiyet vermiştir [48]. ID3 algoritması, 1986 yılında Quinlan tarafından geliştirilmiştir. Bir sistemdeki belirsizliğin ölçüsü olan entropiye dayalı algoritmalar arasında yer almaktadır.

Belirsizlik olmayan durumlar için entropi düşüktür. Bu algoritma örnekler arasındaki ayırıcı değişkenin bulunmasını sağlar. ID3 algoritması kategorik veriler için kullanılmaktadır. Ağaç tamamlanana kadar düğümlerdeki tüm verilerin sıraya dizilmesini sağladığı için büyük veri setlerinde kullanışlı değildir. C4.5 algoritması ise ID3'ten farklı olarak kategorik ve sürekli değerdeki veriler için kullanılır. Her iki algoritmada graf teorisindeki düğümleri dolaşma yöntemlerinden önce derinlik ilkesine göre çalışırlar. Kayıp verileri kullanmaması, kötümser budama adı verilen yöntem ile hata oranının azaltılmış olması ID3'e göre üstünlüğüdür. Ancak büyük veri setleri için tüm verileri ID3 algoritması gibi sıraya dizmesi sebebi ile C4.5 algoritmasının kullanımı da uygun değildir [4].

SPRINT algoritması, disk tabanlı sınıflama yapılabilmesi için tasarlanan, paralel sınıflandırma algoritmalarındandır. Kategorik ve sürekli değerler bu algoritmada kullanılabilir. Büyük veri setlerinde kullanımı uygundur.

Algoritmada düğümlerdeki veriler ilk aşamada sürekli değişkenlere göre sıralanır ve kök düğüm ile ilişkili hale getirilir. Ağaçta bölünmeler ile yeni ürünler oluşturulur. Her bölünmede kayıtların sırası korunduğu için oldukça hızlı bir algoritmadır. SLIQ algoritması da ilk aşamada sıralama işlemi yapmaktadır. Bu algoritma graf teorisindeki düğümleri dolaşma yöntemlerinden önce genişlik prensibine göre çalışmaktadır. Eğitim setinin tamamını kullanarak ağacın oluşturulmasını sağlar. Bu sebeple kayıtlı veriye sürekli ihtiyaç duyar. Bu da SLIQ algoritmasının sınıflandırabileceği veri miktarını azaltır [54].

Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Tree-CART) ilk olarak Breiman vd. tarafından tanıtılmıştır. CART, sürekli ve kategorik yordayıcısı ve hedef değişkenlerini işleyebilen bir ikili Karar Ağacı algoritmasıdır. Ayrıca, Karar Ağacı modeli sonuçların net olmasını sağlar. Karar Ağacı modeli sonuçları, tahmin veya sınıflandırma için önemli faktörlerin önemi hakkında açık bilgi vermektedir [55].

CART algoritması yinelemeli olarak çalışır. Her alt kümedeki kayıtları önceki alt kümelere göre daha homojen hale getirmek için verileri iki alt kümeye ayırır. Ayrılma işleminin amacı alt grupların/dalların homojenliğini maksimize edebilmektir. Daha sonra iki alt grup, homojenlik kriteri ya da başka zamana dayalı durma kriterleri karşılanana kadar tekrar bölünür. Aynı tahmin değişkeni Karar Ağacını büyütme sürecinde birkaç kez kullanılabilir [55].

Sınıflandırma ağaçları oluşturulurken verileri bölme işleminde kullanılan değişkeni seçmek için çeşitli kriterler önerilmiştir. Bunlar Gini ve Twoing kriterleridir. Bu iki kriter, sınıflandırma ağaçlarında yayın olarak kullanılan safsızlık ölçümleri olarak bilinirler. Twoing algoritması ayırma işlemini her defasında eğitim setini iki parçaya ayırarak gerçekleştirirken, Gini algoritması nitelik değerlerini iki parçaya ayırarak uygulanır. Her ayrılma işleminde en küçük Gini değeri kullanılır. Sınıflandırma ağaçlarında, ağacın her düğümünde kullanılan değişken Karar Ağacının performansını etkiler. Bu sebep ile ayrılma değişkeninin belirlenmesi önemli bir kriterdir. Değişkenin seçilmesinden sonra, en iyi bölünmeyi veren değer seçilir. Sınıflandırmanın amacı minimum hata ile doğru sınıflara ayırmaktır [55].

Bellek tabanlı sınıflandırma algoritmalarından olan KNN algoritmasında eğitim verisi hazırlanır. Sınıflandırma yapılmamış bir yeni kayıt eklendiği zaman, eğitim verisindeki benzer kayıtlar ile kıyaslanır. Bu işlemin yapılabilmesi için en kısa mesafeye sahip olan k adet gözlem seçilir.

Kullanılan *k* değeri rastgele belirlenmektedir. *k* değerinin büyüklüğü sınıflandırılacak olan verinin benzerlikleri az olan veya benzerlikleri fazla olan sınıfa dahil edilmesinde önemli rol oynar. Yeni kaydedilen veri ile sınıflandırılmış veriler arasındaki uzaklığın belirlenmesinde Öklid formülü kullanılır.

2.4.1.2. Bayes Sınıflandırıcı

Bayes sınıflandırıcı, 1763 yılında Thomas Bayes tarafından tanıtılan Bayes teoremine dayanmaktadır. Bayes sınıflandırıcıları istatistiksel sınıflandırma teknikleri arasında yer alan hızı ve hesaplamadaki performansı sebebi ile araştırmacılar tarafından sıklıkla tercih edilen bir algoritmadır. Sınıflandırılacak olayları birbirinden bağımsız olarak ele alan bu teorem verilerin hangi sınıfa ait olduğunu öngörmektedir.

Sınıflandırmadan önce bir başlangıç zamanı gerektirmez ve yapılacak olan tüm sınıflandırmalar için veri kümelerinin tamamını işler. Kolay uygulanabilirliği, çoğu durum için iyi sonuçlar elde edilmesi ve yüksek performansı Bayes teoreminin avantajları arasında sayılabilir. Ancak uygulamada değişkenler birbirine bağımlı olduğu için değişkenler arası ilişkiyi modellemede sorun yaşanmaktadır [8].

Bu teorem koşullu olasılıklar ile rastgele değişenlerin marjinal olasılıkları arasında bir ilişki olduğunu göstermektedir. P(A), A'nın ilk olasılığı, P(B) ise B'nin olasılığıdır. P(A\B) ise şarta bağlı olasılığı göstermektedir. Bayes teoremi matematiksel olarak Denklem 2.1'de ifade edilmiştir.

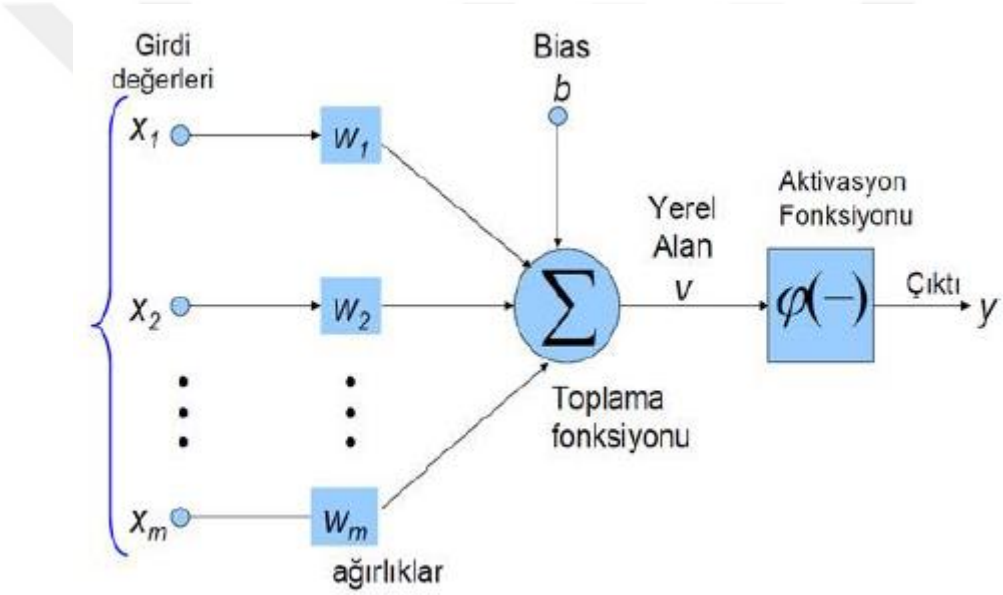
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (\mathbf{2.1})$$

Bayes sınıflandırıcılardan olan NB algoritması ve W-NB algoritmasından 3. Bölümde bahsedilmiştir.

2.4.1.3. Yapay Sinir Ağları

Sınıflandırma teknikleri arasında olan Yapay Sinir Ağları ilk olarak 1943 yılında insan beyninin sinir sistemi model alınarak oluşturulmuştur. Bu sistem öğrenilen bilgiler kullanılarak, yeni bilgiler üretilmesini sağlayan programlardır.

Paralel dağıtılmış, bağlantılı ve nuromorfik ağlar olarak da adlandırılmaktadır Şekil 2.3'te basitçe bir Yapay Sinir Ağı Modeli gösterilmiştir [56]. YSA, verilen girdi değerlerine karşılık bir çıktı üretmektedir.



Şekil 2.3. YSA yapısı [56].

YSA'da Şekil 2.3'te görüldüğü gibi, her bir girdi değeri sayısal bir ağırlığa sahiptir. Ağırlıklar girişlerin sinir üzerindeki etkisini belirleyen katsayılardır ve bir elemanın diğer eleman üzerindeki etkisini gösterir. Ağırlıkların negatif olması etkinin ters yönde olduğunu ifade ederken, sıfır olması hiçbir etkinin olmadığını göstermektedir. Ancak ağırlık değerinin büyük ya da küçük olması girişin önemli ya da önemsiz olması anlamına gelmemektedir.

Ağ topolojisinin seçimi YSA'nın genellemesi ile doğrudan ilişkilidir. Ağ için kullanılan mimarinin problemi öğrenmek için uygun büyüklükte olması gerekmektedir. Uygun mimariden küçük olan bir ağda iyi bir öğrenme gerçekleşemezken, daha büyük bir ağda ise aşırı öğrenme gerçekleşir ve genelleme işlemi zayıf kalır. Ağın yapısı belirlenirken

iki yaklaşım kullanılmaktadır. Eğer ağ yapısı başlangıçta küçük seçilip öğrenme aşamasında büyüyorsa büyüyen/yapıcı yaklaşım, büyük seçilip öğrenme aşamasında küçülüyorsa budama/yıkıcı yaklaşım kullanılmaktadır [57].

YSA tek katmanlı ve çok katmanlı algılayıcılar olmak üzere ikiye ayrılır. Tek katmanlı algılayıcılar, yalnızca giriş ve çıkış katmanından oluşur.

Katmanlardaki nöron sayısı birden fazla olabilir. Tek katmanlı algılayıcılardaki temel modeller, Basit Algılayıcı Modeli ve Adaline Modelidir. Çok katmanlı algılayıcılarda ise birden fazla ara katmanı bulunan bir giriş katmanı ve bir çıkış katmanı bulunur. Çok katmanlı algılayıcılar, geri öğrenme algoritmasını kullanmaktadır [57].

2.4.1.4. Genetik Algoritmalar

1975 yılında John Henry Holand tarafından tanıtılan Genetik algoritmalar, doğal seçim prensibine dayanmaktadır. Bir olasılıksal sınıflandırma algoritması olan genetik algoritmalar parametre kümesinin kodlanmış halini kullanmaktadır. Çözüm için çözüm uzayının belirli bir kısmını taramaktadır. Böylece kısa sürede etkili bir çözüm sunmaktadır. Genetik algoritmaların aşamaları şu şekilde açıklanabilir [9];

- ✓ Arama uzayında yer alan tüm olası çözümler ikili bit dizisi olarak kodlanır.
- ✓ Rastgele bir çözüm kümesi belirlenir.
- ✓ Her bir dizi için uygunluk değeri bulunur.
- ✓ Bir grup dizi olasılık değerine göre rastgele seçilip çoğaltma işlemi uygulanır.
- ✓ Yeni eklenen nesnelerin uygunluk değerleri belirlenir, genetik operatörlerinden çaprazlama ve mutasyon işlemleri uygulanır. Böylece yeni bir popülasyon elde edilir.
- ✓ Önceden belirlenen kuşak sayısına ulaşana kadar tüm işlemler tekrarlanır.
- ✓ İterasyon sayısı, belirlenen kuşak sayısına ulaştığında işlem tamamlanarak amaç fonksiyonuna en uygun olan dizi seçilir.

Genetik algoritmaların ilk adımı için, tüm olası çözümler aynı boyuta sahip olacak şekilde ikili bit dizisi olarak kodlanır. Her bir dizi çözüm uzayındaki rastgele bir noktayı temsil eder. Böylece genetik algoritmalarda kullanılacak şekilde bilgiler tanımlanmış olur. Muhtemel çözümlere göre bir çözüm grubu yani popülasyon oluşturulur. Bu grupta çözümlerin kodlarına kromozom denilmektedir. İlk popülasyonu oluşturmak için rastgele

sayı üreticisi kullanılır. Popülasyonda yer alan her üye için uygun bir değer belirlenir. Genellikle bu değer o noktadaki amaç fonksiyonunun değeridir. Uygunluk değerinin yüksek olması, çözümün sonraki kuşaklarda temsil edilme oranının yüksek olmasını sağlar. Sonraki aşamada amaç fonksiyonuna göre, diziler kopyalanarak iyi olan kalıtsal özellikleri gelecek kuşaklara aktarılır. Mevcut gen potansiyelini araştırmak için çaprazlama, mevcut kromozomlardan yeni kromozomlar üretmek için ise mutasyon işlemi gerçekleştirilir.

Bu işlemlerin yapılmasından sonra yeni kuşak oluşturulmuş olur ve bunlar bir sonraki kuşağın ebeveynleridir. Yapılan işlem hedeflenen uygunluk kriterine ya da maksimum iterasyon sayısına ulaşılıncaya kadar devam eder [58].

2.4.1.5. Diskriminant Analizi

Diskriminant analizi çok değişkenli bir analizdir. Sınıflandırma işlemi yapılırken, sınıflar önceden biliniyor ise diskriminant analiz kullanılır. Bu analizin amacı grupları birbirinden ayıran özellikleri belirlemektir. Aynı zamanda gruplarda yer alan değişkenlerin bu farklılık üzerindeki etkileri ve gruplar arasındaki ayrımı maksimize eden tahmin değişkenlerinin belirlenmesi gerekmektedir. Son olarak da yeni gelen bireylerin sınıfların belirlenmesi işlemi gerçekleştirilebilir. Kısaca diskriminant analizi, diskriminant fonksiyonlarını belirleyip, bu fonksiyonlar ile gruplar arası ayrıma sebep olan değişkenleri ve hangi gruptan geldiği bilinmeyen bir verinin hangi gruba dahil edileceğini belirler [59].

Diskriminant analizinin gerçekleştirilmesi sürecindeki aşamalar şu şekildedir [60];

- ✓ Apriori ile grup üyelikleri belirlenir.
- ✓ Değişkenler için gruplar arasındaki farklar belirlenir. Bu amaç için MANOVA testi kullanılır. Gruplar arasında anlamlı bir fark olması durumunda analize devam edilirken, kayda değer bir fark olmaması durumunda grupların aynı olduğu düşünülerek analiz sonlandırılır.
- ✓ Bir önbilgi veya istatistiksel yöntem uygulanarak değişkenler seçilir.
- ✓ Değişkenler arasında çoklu korelasyon olup olmadığı incelenir. Bunun için grup içi korelasyon matrisi oluşturulur. Bu matriste korelasyonun mutlak değerinin %75'in üzerinde olması durumunda bazı değişkenler elimine edilir ve böylece değişken grup belirlenmiş olur.

- ✓ Matrisin özdeğerleri ve özvektörleri tanımlanır. Özvektörler, diskriminant fonksiyonları için gerekli olan ağırlık değerlerinin elde edilmesinde kullanılır. Özdeğerler ise diskriminant fonksiyonlarının önemlerinin tespit edilmesinde kullanılırlar. Fonksiyonların herhangi biri önemli ise, işlem başarılı kabul edilir.
- ✓ Standartlaştırılmamış diskriminant fonksiyonu kullanılarak, her bir birey için fonksiyon değerleri elde edilir ve bu değerler sonraki aşamalarda sınıflandırma işlemi için kullanılır.
- ✓ Grup üyeliklerinin belirlenmesi için öncelikli olasılıklar belirlenir. Bu olasılıklar ve diskriminant değerleri kullanılarak bir frekans olasılığı elde edilir. Olasılığı veren grup, bireyin ait olduğu grup olarak belirlenir ve sınıflandırma işlemi tamamlanır.
- ✓ Tüm bireylerin sınıflandırma işlemi tamamlandıktan sonra doğru sınıflandırma yüzdesi bulunarak diskriminant fonksiyonunun başarısı belirlenmiş olur.

2.4.1.6. Kaba Küme Yaklaşımı

1982 yılında Pawlak tarafından önerilen teori, eksik bilgi ile başa çıkılabilmesini sağlamaktadır. Kaba Küme yaklaşımında sadece veri içerisindeki bilgi kullanılmaktadır. Her nesne ile bazı bilgilerin bağdaşlaştırılması varsayımına dayanmaktadır. Varsayıma göre bir kümenin oluşturulabilmesi için öncelikle bazı bilgilere gereksinim vardır. Aynı hastalığa sahip bireyler nesne olarak kabul edilir. Hastalığın belirtileri hastalar hakkındaki bilgiyi oluşturur ve nesneler aynı bilgiler ile bağdaştırılırsa, bu nesnelerin aynı olduğu ve ayırt edilemeyeceği bilgisine ulaşılır. Ayırt edilememe ilişkisi kaba küme yaklaşımının temelini oluşturur. Tüm kaba kümelerin sınıflandırılamayan elemanları bulunmaktadır. Ancak kesin kümelerde bu durum söz konusu değildir [61].

Kaba küme yaklaşımı verilerdeki gizli bilgilerin bulunmasında etkili bir yöntemdir. Veriler hakkında ek bilgilere ihtiyaç duymaz. Asıl verideki bilgi ile aynı bilgiyi sağlayacak verinin minimal altkümesi bulunur. Verinin önemli olması değerlendirme işlemini kolaylaştırır. Paralel ve dağıtık algoritma yapıları için uygundur [61].

2.4.1.7. Destek Vektör Makineleri

1995'te Vapnik'in ortaya çıkardığı Destek Vektör Makineleri hem sınıflandırma hem de regresyon problemlerine uygulanabilen bir yöntemdir. İstatistiksel öğrenme, bir dizi

eğitim setinden bir fonksiyonun tahmin edilmesi anlamına gelmektedir. Bunun için bir öğrenme makinesi, belirli bir fonksiyon setinden bir fonksiyonu seçmeli ve bu tahmini fonksiyonun gerçek fonksiyondan farklı olduğu amprik riski en aza indirmelidir. Risk eğitim setine ve seçilen fonksiyonun karmaşıklığına bağlıdır.

Bir öğrenme makinesi en iyi fonksiyonu bulmalıdır. Ancak riske bağlı olmak kolay hesaplanamayan bir işlemdir ve çözümün kalitesini analiz etmekte faydalı değildir [62].

Destek Vektör Makinesi verileri iki kategoriye ayırarak n boyutlu hiperdüzlem oluşturmaktadır. Sigmoid bir kernel fonksiyonu kullanmaktadır. İki katmanlı bir YSA'ya sahiptir.

Eğitim setindeki tüm örneklerin bağımsız ve benzer olarak dağıtılmış olduğu varsayımına dayanan DVM, yapısal risk minimizasyonu prensibi ile çalışır. DVM'de temel amaç eğitim verilerini mümkün olduğunca gerçeğe uygun bir şekilde yansıtan doğrusal ayırıcı fonksiyonun bulunmasıdır. DVM'de lineer olarak ayrılabilme ve ayrılamama gibi iki durum söz konusudur. Lineer olarak ayrılma durumunda iki sınıf doğrusal bir düzlem ile ayrılır ve bu düzlem iki sınıfa da eş uzaklıktadır. Lineer ayrılamamada ise çekirdek fonksiyonu kullanılarak ayırma işlemi gerçekleştirilir. Bu amaç için kullanılan çekirdek fonksiyonları, doğrusal. polinomiyal, sigmoid ve radyal tabanlı fonksiyondur [63]

2.4.2. Regresyon Analizi

Regresyon analizi bir bağımlı değişken ve bir ya da daha fazla tahmin edici (bağımsız) değişken arasındaki bağıntıların niteliğini incelemek için kullanılan analiz yöntemidir. Sınıflandırma yönteminde bağımlı değişken kategorik iken regresyon analizinde bağımlı değişken sayısal bir değerdir. Kullanılan bağımsız değişken sayısı bir tane ise basit regresyon, birden fazla ise çoklu regresyon analizi denilmektedir.

Bağımsız değişkenlerden her birinin bağımlı değişken üzerindeki etkisinin hesaplanması olan regresyon, tahmin edilen değerden yola çıkarak bağımlı değişkenin değerin tahmin edilmesi sürecidir [52].

Veri madenciliği modellerinden regresyon modelini temel alan başlıca yöntemler;

- ✓ Basit Doğrusal Regresyon

- ✓ Çoklu Regresyon (Standart Çoklu Regresyon, Hiyerarşik Çoklu Regresyon, İstatistiksel ve Kademeli Regresyon)

2.4.2.1. Basit Regresyon Analizi

Biri bağımlı diğeri bağımsız iki değişken arasındaki ilişkiyi incelemek için kullanılan regresyon analizi yöntemidir. Basitçe gösterimi Denklem 2.2'de verilmiştir [52].

$$Y_i = b_0 + b_i X_i + e_i \quad \textbf{(2.2)}$$

b_0: Doğrunun y eksenini kestiği nokta

b_1: Regresyon katsayısı

e_i: Hata değeri

2.4.2.2. Çoklu Regresyon Analizi

Çoklu regresyon bir bağımlı ve birden fazla bağımsız değişken arasındaki ilişkiyi incelemek için kullanılan regresyon analizi yöntemidir. Bu yöntemde bağımsız değişkenlerin bağımlı değişken üzerindeki anlamlı bir etkiye sahip olup olmadığı belirlenmektedir [52]. Çoklu regresyon modeli Denklem 2.3'te verilmiştir.

$$Y_i = b_0 + b_i X_{1i} + b_i X_{2i} + \cdots + b_n X_{ni} + e_i \quad \textbf{(2.3)}$$

b_0: Doğrunun y eksenini kestiği nokta

b_1,b_2,b_n: Regresyon katsayıları

e_i: Hata değeri. En küçük kareler yöntemi ile hesaplanır.

X_{1i},X_{2i},….X_{ni} değerlerine göre Y_i tahmin edilir.

Bağımsız değişkenlerin tümünün aynı anda Denkleme katılması ile yapılan analize standart çoklu regresyon analizi denilir. Böylece bağımsız değişkenlerden her birinin bağımlı değişken üzerindeki anlamlı olan etkileri değerlendirilir.

Bağımsız değişkenlerin hangi sıra ile Denkleme dahil edileceğine araştırmacının karar verdiği regresyon modeline hiyerarşik çoklu regresyon denilmektedir. Araştırmacı değişkenlerin sıralanmasına mantıksal veya teorik değerlendirmelere göre karar verir. Yani daha önce kayıtsız kaldığı varsayılan değişkenlere ya da daha fazla teorik öneme sahip değişkenlere sıralamada öncelik verilebilir. Değişken girişlerinin sırasının sadece istatistiksel kriterlere dayandığı yöntem istatistiksel ve kademeli regresyondur. Oldukça tartışmalı olan bu modelde hangi değişkenlerin sürece dahil edildiği ve çıkarıldığı sadece belirli bir örneklemden hesaplanan istatistiğe dayanır ve bu istatistiklerdeki ufak farklılıklar bile bir değişkenin önemi ile ilgili büyük bir etkiye sahip olabilir. İstatistiksel regresyonun ileri seçim, geriye doğru silme, kademeli regresyon olmak üzere üç türü vardır. İleri doğru seçimde Denklem ilk önce boştur [64].

İstatistiksel kriterleri sağlayan değişkenler Denkleme birer birer eklenir. Geriye doğru silmede, Denklem tüm değişkenlerin eklenmesi ile başlar. Ancak regresyona önemli bir katkısı bulunmayan değişkenler birer birer silinir. Kademeli regresyon yönteminde, istatistiksel ölçütleri karşıladığı sürece Denkleme her seferinde bir değişken eklenir. Ancak regresyona önemli bir katkılarının olmadığı herhangi bir aşamada değişkenler silinebilirler [64].

2.4.3. Kümeleme Modeli

Sınıflandırma gözetimli bir öğrenme yaklaşımı iken kümeleme, gözetimsiz bir öğrenme yaklaşımıdır. Kümeleme genellikle uzaklık ölçütlerine dayanan benzerlik ve farklılık ölçütlerini kullanarak kümeler arasındaki gözlemlerin farklılıklarını belirler. Gözlemlere ait grafikler de bu amaç için kullanılabilir. İki değişkenin olduğu durumlarda açılım grafikleri, değişken sayısı ikiden fazla olduğu durumlarda temel bileşen analizi kullanılabilir [9].

Kümeleme işleminde, verilerin bölündükleri gruplara göre kendi aralarında benzer başka gruplar ile farklı özellikler gösterir. Bu işlemin ne kadar iyi yapıldığı ise o kümeleme işleminin ne kadar başarılı olduğunu ifade etmektedir.

Kümeleme analizi, modelin benzerliğe dayalı olarak kümeler halinde düzenlenmesidir. Verilerin daha az sayıda küme tarafından temsil edilmesi, ince ayrıntıların kaybolmasına sebep olurken, sadeleştirmeyi sağlar [65].

Küme analizi, nesnelerin hem sınıfların sayısı hem de bileşimlerinin belirlenmesi gerektiğinde, anlamlı sınıflara ayrıştırılmasıdır. Kümeleme işleminde kümelerin sayısını belirlemek için çoğaltma ve çapraz doğrulama prosedürleri kullanılabilir [66].

Kümeleme analizinde uygulanan aşamalar şu şekildedir;

- ✓ Nesneler arasındaki benzerliğin ölçütünün belirlenmesi,
- ✓ Belirlenen benzerliklere göre nesnelerin kümelere ayırılması,
- ✓ Kümelerin doğru oluşturulup oluşturulmadığının incelenmesi,
- ✓ Belirlenen kümelerin doğru uygunluğunun istatistiksel yöntemler kullanılarak ortaya konulmasıdır.

Kümeleme analizi için hiyerarşik kümeleme ve hiyerarşik olmayan kümeleme olmak üzere iki yöntem bulunmaktadır. Bu yöntemlerden Hiyerarşik Kümeleme yönteminde, bir ana küme belirlenip, sonra benzerliklerine göre alt kümelere ayrılırsa Bölücü Hiyerarşik Kümeleme yöntemi, her bir nesne bir küme gibi düşünülüp özelliklerine göre birleştirilirse Birleştirici Hiyerarşik Kümeleme yöntemi denilir [9].

Hiyerarşik olmayan kümeleme yöntemi oluşturulacak olan küme sayısı belli ise karmaşıklığı az olması sebebi ile kullanılabilir. Kuramsal dayanakları güçlü olan bu yöntemdeki temel sorun uygun küme sayısının belirlenmesi için kullanılan güvenilir bir yöntemin olmamasıdır [67].

Hiyerarşik olmayan kümelemede farklı birçok yöntem kullanılmaktadır. Bunlardan bazıları ayırma tipli, yoğunluk tabanlı, ızgara tabanlı kümelemedir. En çok bilinen ve sıklıkla kullanılan algoritması ise k-Means algoritmasıdır. Bir veri setinde yer alan n tane veri kümesi oluşturulup, bu kümeleri kümelerin ortalamalarına göre ayırmaktadır. Kısaca küme içi benzerlikleri ve diğer kümeler ile farklılıkları fazla olacak şekilde kümeleme işlemini gerçekleştirir.

2.4.4. Birliktelik Kuralı

İlk olarak 1993 yılında Agrawal, Imielinski ve Swami tarafından tanıtılan Birliktelik Kuralları verilerin analiz edilerek veriler içindeki örüntülerin belirlenmesi amacıyla oluşturulmuştur. Bu örüntülerin belirlenmesi için veriler arasındaki birliktelik davranışları tespit edilmektedir [68].

Bağıntı eşleme ve ilişkisel kurallar olarak da bilinen Birliktelik Kuralları, büyük veri kümelerindeki farklı veriler arasında yüksek sıklıkla birlikte görülen özelliklere ait ilişkisel kuralların çıkarılması işlemidir. Birliktelik Kurallarının en iyi bilinen örneği sepet analizidir. Müşterilerin birlikte satın aldığı ürünler incelenerek, ürünler arasındaki pozitif korelasyonları bulmaktır. Böylece müşterilerin ürünleri satın alırken ne tür alışkanlıkları olduğu belirlenebilir. Bu sektörde yer alan marketlere birlikte alınma sıklığı çok olan ürünleri yakın raflara koyarak satışları arttırmaları anlamında katkı sağlamaktadır [69].

Veri madenciliğinin en iyi örnekleri arasında yer alan birliktelik kuralları veriler arasındaki potansiyel ilişkilerin keşfedilmesini sağlar. Oluşturulan kuralların karakterize edilebilmeleri için destek ve güvenilirlik düzeylerine bakılır. Destek ve güvenirlik seviyesi yüksek olan bir kural güçlü bir kural olarak tanımlanabilir [10].

Birliktelik Kuralları X→Y şeklinde gösterilmektedir. İfadeye göre, X'in olduğu bir durumda Y'nin de olma olasılığının yüksek olduğunu belirtir. İfadede X olarak gösterilen bölüme öncül denilmektedir ve birden fazla öncül bulunabilir. Y olarak gösterilen kısma ardıl ya da sonuç denilmektedir. Kuralın önem seviyesini belirlemek için destek ve güvenirlilik değerinin hesaplanması gerekir. Güvenilirlik X'i içeren olayların Y'yi de içerme yüzdesidir [70]. Destek ise X→Y = (X ve Y nin bulunduğu satır sayısı) /toplam satır sayısıdır [10].

Birliktelik Kuralları için geliştirilen AIS, SETM, RARM, Apriori, Apriori-TID, Apriori-Hybrid, OCD, Partitioning, Sampling, DIC, CARMA, Fp-Growth gibi birçok algoritma bulunmaktadır. Ancak en iyi bilineni Apriori algoritmasıdır.

3. SADE BAYES SINIFLANDIRICI

Bayes teoremini temel alan Sade Bayes Sınıflandırıcı, olasılık tabanlı bir yaklaşımdır. Verinin öğrenilmesi mantığına dayanan NB'de modelin öğrenmesi amacı ile eğitim verileri için her bir çıktının kaç kez tekrarlandığına bakılarak öncelikli olasılık değerleri bulunur. Bağımsız değişkenlerin bağımlı değişkenlere bölünmesinin kombinasyonu hesaplanarak olayın oluşma sıklığı belirlenir.

Veri setleri içerisindeki verilerin birden çok özniteliği vardır. NB sınıflandırıcı hem kategorik hem de sayısal öznitelikleri olan verilerin sınıflandırılması işleminde kullanılmaktadır. Ayrıca sınıf etiketi denilen bir öznitelik mevcuttur. Böylece yeni eklenen ve sınıf etiketi olmayan kayıtların, sınıf etiketlerinin tahmin edilmesinde kullanılacak bir model oluşturmayı hedefler. Bağımsız değişkenlerin, bağımlı değişkenler üzerindeki etkileri kullanılarak, yeni bir verinin sınıflandırılma işlemi gerçekleştirilir [7].

NB sınıflandırıcının çalışma süresinin kısa olması, gerçek ve ayrık verileri işleme yeteneği, veri akışını iyi idare etmesi avantajları arasında sayılabilir. Ancak ilgisiz özelliklere karşı duyarlı olmaması yani bütün özniteliklerin eşit sayılması dezavantajıdır. Çünkü yapılan tahminlerin yetersiz kalmasına sebep olmaktadır.

D bir veri tabanı eğitim seti olsun. Her veri tabanını temsil eden bir n-D nitelik vektörü bulunur. X, *n* tane bağımsız nitelik (x_1,x_2,x_3,….x_n) içerir. *m* tane sınıfı (C_1,C_2,C_3,….C_m) olduğunu varsayalım. Denklem 3.1 ile maksimum P(Ci/X) türetilmektedir.

$$P(Ci\backslash X) = \frac{P(X\backslash Ci)P(Ci)}{P(X)} \qquad (\mathbf{3.1})$$

P(X) tüm sınıflar için eşit değere sahip olduğundan maksimize edilmelidir. Bu işlem Denklem 3.2'de gösterilmiştir.

$$P(Ci\backslash X) = P(X\backslash Ci)P(Ci) \qquad (\mathbf{3.2})$$

Özniteliklerin arasında bağımlılık ilişkisi bulunmamaktadır. X_i değerlerinin birbirinden bağımsız olduğu kabul edilerek Denklem 3.3'te yer alan bağıntı kullanılır.

$$P(X\backslash Ci) = \prod_{k=1}^{n} P(Xk\backslash Ci) \qquad (\mathbf{3.3})$$

Sınıf üyeliği bilinmeyen X için Denklem 3.1'de gösterilen bağıntıda payda değerleri birbirine eşit olduğu için yalnızca pay değerleri kıyaslanır. Bu değerlerden en büyük olanı belirlenerek X'in ait olduğu sınıfa karar verilebilir.

$$\arg\max_{Ci}\{P(X\backslash Ci)P(Ci)\} \qquad (\mathbf{3.4})$$

Sonrasal olasıkları kullanılan Denklem 3.4'teki ifade, en büyük sonrasal sınıflandırma yöntemi (MAP) olarak bilinmektedir. Belirtilen Denklemlerden yola çıkılarak NB sınıflandırıcı için Denklem 3.5'te yer alan bağıntı kullanılabilir.

3.1. Ağırlıklı Sade Bayes Sınıflandırıcı

Ağırlıklı Sade Bayes sınıflandırıcı, NB sınıflandırıcının geliştirilmiş bir versiyonudur. NB sınıflandırıcılardaki ilgisiz özniteliklere karşı duyarlı olmaması, bütün özniteliklerin eşit kabul edilmesi sorununun çözümü için geliştirilmiştir. Temel amacı birkaç unsurun etkisini aynı kümede yer alan diğer öğelere göre arttırmaktır.

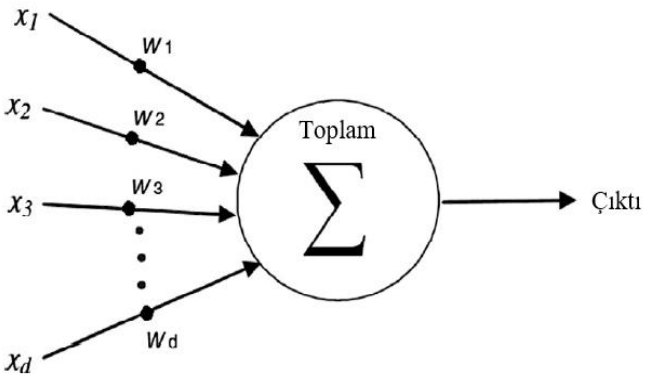
Ağırlık fonksiyonu, toplam, çarpım ve ortalama gerçekleştirmek için kullanılan bir matematiksel sistem olarak kabul edilmektedir. Bir ağırlık fonksiyonunun ana amacı, her bir öğenin sonuçlar üzerindeki etkisini değiştirmek veya arttırmaktır. Ağırlık fonksiyonu olan $\omega: A \rightarrow \mathbb{R}^+$, A üzerinde pozitif bir fonksiyon olarak tanımlanır. Eğer ağırlık fonksiyonu ω(a):=1 ise buna tüm unsurlar eşit ağırlıkta ya da ağırlıksız denir. Eğer f: A $\rightarrow\mathbb{R}$ fonksiyonu gerçek değerli bir fonksiyon ise, A'nın üzerindeki f'nin ağırlıksız toplamı Denklem 3.5'teki gibi tanımlanır [31].

$$\sum_{a\in A} f(a); \qquad (\mathbf{3.5})$$

Fakat ağırlık işlevi $\omega: A \rightarrow \mathbb{R}^+$ verildiğinde ağırlıklı toplam Denklem 3.6'daki gibi olur;

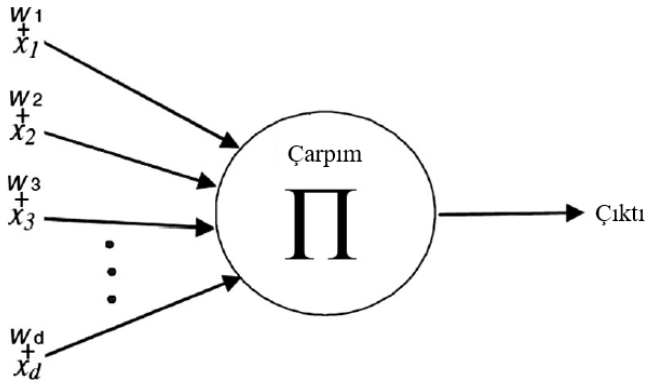
$$\sum_{a \in A} f(a)\omega(a). \qquad (\mathbf{3.6})$$

Şekil 3.1'de YSA yapısında bulunan ağırlıklandırma işlemi gösterilmiştir. Giriş elemanları (x_1, x_2, x_3,……x_d) ağırlık değerleri ile çarpılıp toplam fonksiyonundan geçirilerek çıkış elde edilir.



Şekil 3.1. Toplam fonksiyonu için ağırlıklandırma işlemi

Ağırlıklı Sade Bayes sınıflandırıcıda kullanılan ağırlıklandırma işlemi Şekil 3.2'de gösterilmiştir. Her bir ağırlık, karşılık gelen girişler ile çarpılmak yerine toplanmaktadır.



Şekil 3.2. Çarpım fonksiyonu için ağırlıklandırma işlemi

W-NB sınıflandırıcıda ağırlıkların eşit olması gerektiği zaman w(a)= 0 olur. Bu durum Denklem 3.7'de gösterilmiştir.

$$\prod_{a\in A}\big(f(a)+w(a)\big) \qquad (\mathbf{3.7})$$

Örnek bir veri tabanı eğitim setinde, veri tabanını temsil eden n adet nitelik vektörü bulunur. X, n tane bağımsız nitelik (x_1,x_2,x_3,….x_n) içerir. i tane sınıfı (C_1,C_2,C_3,….C_i) olduğunu varsayıldığında, Denklem 3.8'de ağırlıklandırma işlemi gerçekleştirilmiş P(Ci/X) türetilmektedir.

$$\psi(\mathrm{k}) = W_k + \mathrm{P}(\mathrm{X_k}\backslash\mathrm{C_i}) \qquad (\mathbf{3.8})$$

$$P_w(X\backslash C_i) = \prod_{k=1}^{n}\psi\,(k) \qquad (\mathbf{3.9})$$

Sınıf üyeliği bilinmeyen X için Denklem 3.10'da gösterilen bağıntı ile X'in ait olduğu sınıfa karar verilebilir.

$$\arg\max_{\mathrm{C_i}}\{\mathrm{P_w}(\mathrm{X}/\mathrm{C_i})\mathrm{P}(C_i)\} \qquad (\mathbf{3.10})$$

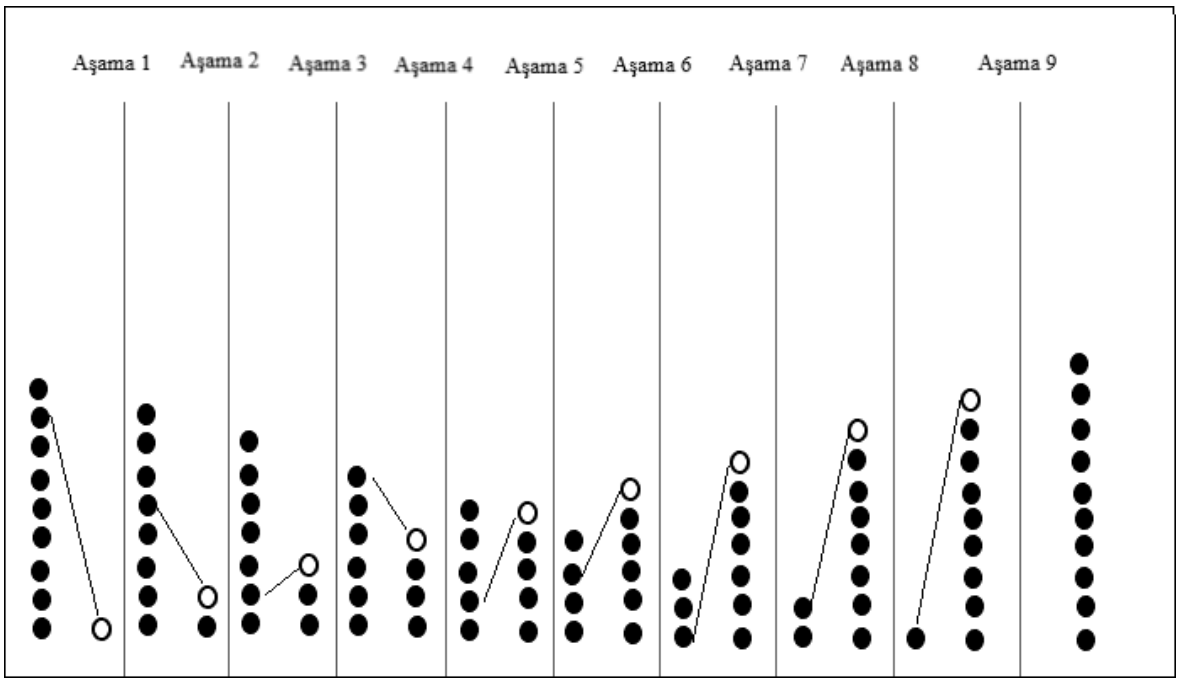
3.2. Önerilen Yöntem

Ağırlıklı Sade Bayes sınıflandırıcıların başarımı, Sade Bayes sınıflandırıcılara göre daha yüksektir. Ancak W-NB'de ağırlıklandırma işlemi yapılması maliyeti yükseltmektedir. Ağırlıkların belirlenebilmesi için literatürde yapılan çalışmada ızgara arama yapısı kullanılarak iç içe döngüler oluşturulmuştur [31]. Oluşturulan bu döngüler algoritmanın hesaplama yapmak için kullandığı süreyi uzatmaktadır. Algoritmanın karmaşıklığının giderilmesi performansını etkileyen bir parametredir. Bu tez çalışmasında, özniteliklerin etkili olanlarını ön plana çıkarmayı sağlayan ağırlıkların, farklı bir algoritma ile elde edilmesi sağlanacaktır. Hazırlanan algoritmanın çalışma süresini kısaltması ve NB sınıflandırıcıya göre başarımının yüksek olması hedeflenmektedir.

Önerilen yöntemde, n adet özniteliğe sahip bir veri seti için ağırlık değerleri (w_1, w_2, w_3, ….w_n) şeklinde oluşturulur. Tüm ağırlık değerleri başlangıçta sıfır olarak belirlenmiştir.

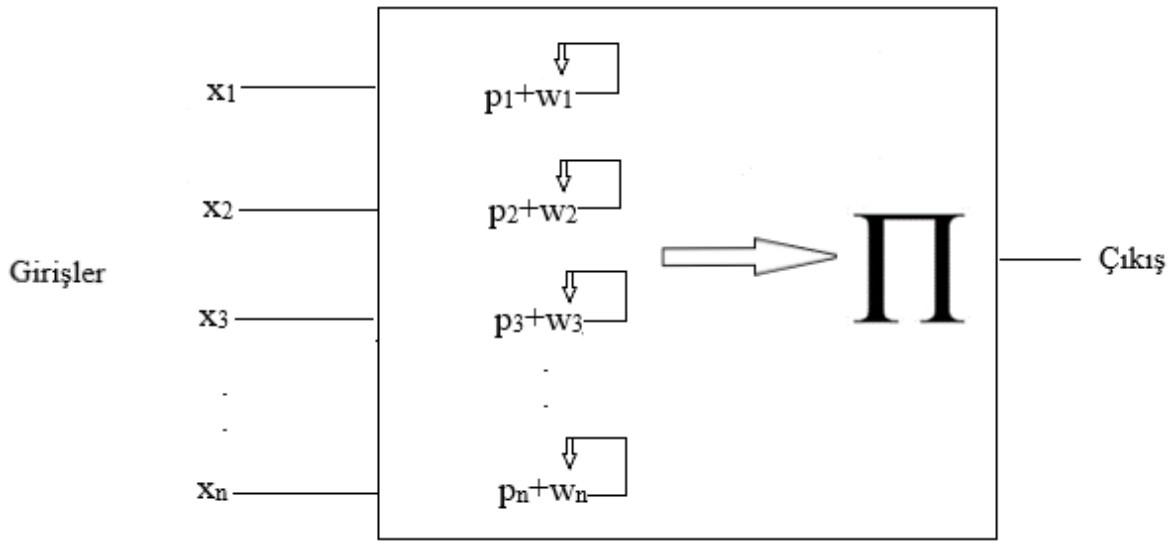
Ağırlık değerlerine ayrı ayrı 0 ve 1 arasında, her adımda x kadar arttırılacak şekilde değerler verilmiştir. Ağırlık değerleri, Bayes sınıflandırıcılardaki olasılık değerleri 0-1 aralığında olduğu için bu aralıkta seçilmiştir.

Kullanılan artım miktarı çalışmanın başında kullanıcı tarafından belirlenmektedir. Bir ağırlık değerine bu şekilde değerler verilirken, ilk etapta diğer ağırlıklar sıfırdır. n tane ağırlık için bu işlem gerçekleştirildikten sonra elde edilen başarım oranı en yüksek olan değer sabit alınarak n-1 defa daha aynı işlem tekrarlanmıştır. Bu şekilde döngü tüm değerler sabit kabul edilene kadar tekrar eder. En son adımda hem her öznitelik için ağırlık değerleri belirlenir hem de en yüksek başarım değeri elde edilir.



Şekil 3.3. Önerilen yöntemde kullanılan ağırlıklandırma işleminin şematik gösterimi

Şekil 3.3'te gösterilen aşamalar 10 özniteliğe sahip olan bir veri setine göre hazırlanmıştır. İlk aşamada ağırlıkların tamamı sıfır değerine sahiptir. Her aşamada yöntemin başarım değerini yükselten ağırlık değeri sabit kabul edilip, analiz tekrarlanmıştır. Analizdeki iterasyon sayısı öznitelik sayısı kadardır. Şekil 3.4'te yöntemin blok yapısı verilmiştir.



Şekil 3. 4. Önerilen yöntemin blok yapısı

Şekil 3.4'te gösterilen blok diyagram incelendiğinde Bayes teoreminden farklı olarak, her bir olasılık değerine ağırlıkların eklendiği görülmektedir. Eklenen bu ağırlıklar kendi içinde belirlenen sayısal değerler arasında değişmektedir. Her döngüde değişen bu ağırlık değerleri her defasında olasılık değerlerine eklenerek yöntemin başarımı hesaplanmaktadır. Bu yöntem ile W-NB algoritmasında hesaplama yapmak için geçen süreyi kısaltarak, arama uzayını genişletebiliriz.

W-NB sınıflandırma algoritmasının uygulama aşamaları şu şekilde sıralanabilir;

- ✓ Veri seti modele tanıtılır.
- ✓ Veri setinin eğitim ve test için ayrılan kısımları belirtilir.
- ✓ Veri setinin çıkışında elde edilmesi gereken değerlere göre, Bayes teoreminin temelini oluşturan olasılık hesaplama işlemi gerçekleştirilir.
- ✓ Ağırlıklar (w1,w2,w3,….wn) ızgara arama (grid search) yöntemi kullanılarak (w1,w2,w3,….wn $\in$ R , $0 \leq$ w1,w2,w3,….wn ≤ 1) belirlenir.
- ✓ Her öznitelik için bulunan olasılık değerlerine ağırlıklar eklenir.
- ✓ Bulunan sonuçlar Bayes sınıflandırma kuralları çerçevesinde değerlendirilir.
- ✓ Eğitilen sistemin kontrolü için test verileri incelenir.
- ✓ Test verilerine göre doğru sonuçlar ile hatalı sonuçlar elde edilir.
- ✓ Sistemin başarım oranı bulunur.

3.2.1. Önerilen Yöntemin Algoritmik Analizi

Literatürde yer alan W-NB yönteminde uygun ağırlık parametrelerinin belirlenmesinde ızgara arama yöntemi kullanılmıştır [31]. Bu yöntemde ağırlıkların alabileceği alt değer, üst değer ve uygun bir aralık değeri belirlenmiştir. Ağırlık değerleri alt ve üst sınır değerleri içerisinde belirlenen aralık değerleri kadar arttırılır ve her bir ağırlık değeri elde edilen olasılıklar ile toplanarak doğruluk değerleri hesaplanır. Bulunan doğruluk değerlerinin bir önceki değerden daha yüksek olduğu durumda bu değerler yeni doğruluk değeri olarak kabul edilerek ağırlıklar belirlenir. Her bir veri setinin *m* adet özniteliğinin olduğu ve her niteliğe ait ağırlık değerinin *n* adımdan oluştuğu varsayıldığında, algoritmanın karmaşıklığı $O(n^m)$ olarak bulunur.

Önerilen yöntemde ise ağırlıkların aranmasında, öznitelik sayısı kadar ağırlık değeri ($w_1,w_2,w_3,…..w_m$) belirlenmiştir. Ağırlıkların ilk değerleri ilk aşamada sıfır olarak ($w_1,w_2,w_3,…..w_m=0$) kabul edilmiştir. Her bir ağırlık için ayrı ayrı sıfırdan, belirlenen üst sınıra kadar arama işlemi gerçekleştirilmiştir. Verilen bu değerlere göre ağırlıkların her birinin ayrı ayrı doğruluk değeri üzerindeki etkisi gözlemlenmiştir. Doğruluk değerinin en yüksek olduğu değerdeki ağırlık değeri sabit kabul edilir. Bu işlem en yüksek doğruluk değeri bulunana ve tüm ağırlıklar belirlenene kadar her defasında tekrarlanır. Bu şekilde yapılan arama işleminde algoritmanın karmaşıklığı $O(nm^2)$ olarak bulunur. Böylece W-NB'in ağırlıklarının belirlenmesi daha hızlı hale gelmiş olacaktır.

4. UYGULAMA

Bu tez kapsamında 5 farklı veri setine literatürde yer alan NB ve W-NB sınıflandırma algoritmaları ve önerilen yöntem ile elde edilen Ağırlıklı Sade Bayes algoritması uygulanmıştır. Bu bölümde veri setleri ve analiz sonucu elde edilen verilere yer verilmiştir.

4.1. Veri Setleri

Uygulama için kullanılan veri setleri UCI makine öğrenme ve akıllı sistemler merkezinden alınmıştır [71]. Eksik verilerin olduğu kayıtlar temizlenerek deneyler gerçekleştirilmiştir. Bu veri setleri şu şekilde sıralanabilir;

1. Tic-Tac-Toe son oyun veri seti,
2. Ameliyat sonrası hasta veri seti,
3. Niteliksel iflas veri seti,
4. Mamografik kitle veri seti,
5. Meme kanseri veri seti'dir.

4.1.1. Tic- Tac-Toe Veri Seti

Tic-tac-toe oyunu bilinen en eski zekâ oyunlarından biri olup 3'e 3'lük bir tahtada *x* ve o koyularak, yatay düşey ya da çapraz 3'lü (*x* o *x*) yapılmaya çalışılarak oyun oynanmaktadır. Bu veri tabanı *x*'in ilk oynandığı varsayıldığında olası tüm tahta konfigürasyonlarının kodunu kodlar. *X* için kazanma hedefine sahip olan veri tabanı *x*'in 8 de bir satır oluşturmak için 8 olası yoldan birine sahip olması durumunda geçerlidir. Veri tabanında 9 özellik 960 örnek yer almaktadır. Tic-tac-toe veri tabanı özellik bilgisi Tablo 4.1'de verilmiştir. Veri setinde *x: x* oyuncusu, y: y oyuncusu, b: boş alanı göstermektedir.

Tablo 4.1. Tic-tac-toe veri tabanı özellik bilgisi

Özellik Bilgisi	
Sol üst kare	*x*,o,b
Üst-orta kare	*x*,o,b
Sağ üst kare	*x*,o,b

Sol orta kare	*x*,o,b
Orta ortadaki kare	*x*,o,b
Sağ orta kare	*x*,o,b
Sol alt kare	*x*,o,b
Orta alt kare	*x*,o,b
Sağ alt kare	*x*,o,b
Sınıf	Pozitif, negatif

4.1.2. Ameliyat Sonrası Hastanın Durumu Veri Seti

Ameliyat sonrası, hastanın durumunu veri tabanında sınıflandırma görevi, postoperatif kurtarma alanındaki hastaların ameliyat sonrası nereye gönderileceğini belirlemektir. Postoperatif, ağır cerrahi travma ile başlayan bir akut ağrı olup hipotermi cerrahiden sonra oluşan önemli sorunlardan biridir. Bu veri setinde 8 özellik bilgisi ve 90 örnek yer almaktadır. Veri tabanının özellik bilgisi Tablo 4.2'de gösterilmiştir.

Tablo 4.2. Ameliyat sonrası hastanın durumu veri tabanı özellik bilgisi

Özellik Bilgisi	
Hastanın iç sıcaklığı	Yüksek>37, orta ≥36 ve ≥37, düşük <36
% olarak oksijen satürasyonu	Mükemmel≥98, iyi≥90 ve <98 orta <36.5 ve ≤35 kötü<80
Son kan basıncı ölçümü	Yüksek>130/90, orta ≤ 130/90 ve ≥90/70, düşük <90/70
Hastanın yüzey sıcaklığının stabilitesi	Stabil, orta stabilite, karasız
Hastanın iç sıcaklığının stabilitesi	Stabil, orta stabilite, karasız
Hastanın kan basıncı dengesi	Stabil, orta stabilite, karasız
Hastanın algılanan konforu	0-20
Karar	A hasta genel katına gönderilir. I hasta yoğun bakıma gönderilir.

4.1.3. Niteliksel İflas Veri Seti

Niteliksel iflas veri tabanı, riskler üzerinden iflas durumunun belirlenmesi için hazırlanmış bir veri kümesidir. Bu veri setinde kullanılan özellik bilgileri şu alt başlıkları içermektedir [72].

- ✓ Endüstriyel risk: hükümet politikaları ve uluslararası anlaşmalar, rekabet derecesi, piyasa arzının istikrarı, makroekonomik faktör değişiklikleri, yurtiçi ve yurtdışı rekabet gücü, ürün yaşam döngüsü,
- ✓ Yönetimsel risk: yönetim yeteneği ve yeterliliği, yönetim istikrarı, yönetim/mal sahibi, insan kaynakları yönetimi, büyüme süreci/iş performansı, kısa ve uzun vadeli iş planlaması, başarı ve fizibilite,
- ✓ Mali esneklik: doğrudan finansman, dolaylı finansman, diğer finansmanlar,
- ✓ Güvenilirlik: kredi geçmişi, bilgi güvenilirliği, finansal enstitülerle olan ilişkiler,
- ✓ Rekabetçilik: Pazar konumu, temel kapasitelerin seviyesi, farklılaştırılmış stratejiler,
- ✓ Çalışma riski: satın alma istikrarı ve çeşitliliği, işlem kararlılığı, üretim verimliliği, ürün ve hizmet talebi, satış çeşitliliği, satış fiyatı, satış ağı etkinliği.

Veri setinde toplam 7 özellik 250 örnek yer almaktadır. Tablo 4.3'te veri tabanının özellik bilgisi verilmiştir.

Tablo 4.3. Niteliksel iflas veri tabanı özellik bilgisi

Özellik Bilgisi	
Endüstriyel risk	Pozitif, ortalama, negatif
Yönetimsel risk	Pozitif, ortalama, negatif
Mali esneklik	Pozitif, ortalama, negatif
Güvenilirlik	Pozitif, ortalama, negatif
Rekabetçilik	Pozitif, ortalama, negatif
Çalışma riski	Pozitif, ortalama, negatif
Karar	İflas, İflas değil

4.1.4. Mamografik Kitle Veri Seti

Mamografik kitle veri tabanı, göğüste oluşan lezyonun iyi veya kötü huylu olduğunu tahmin etmek için oluşturulmuştur. Temel amaç gereksiz yapılan göğüs biyopsisi sayısını azaltmaktır. Bu veri tabanında, 400 iyi huylu, 425 kötü huylu olarak belirlenen veri yer

almaktadır. Veri seti, 6 özellik 825 örnekten oluşmaktadır. Tablo 4.4'te mamografik kitle veri setinin özellik bilgisi yer almaktadır.

Tablo 4.4. Mamografik kitle veri tabanı özellik bilgisi

Özellik Bilgisi	
BI-RADS değerlendirmesi	1,2,3,4,5
Yaş	Hasta yaşı (tam sayı)
Şekil	Yuvarlak=1, oval=2, lobüler=3, düzensiz=4
Kütle marjı	Sınırlanmış=1, mikrolobulize=2, belirsiz=3 tanımlanmamış=4, spiküle=5
Yoğunluk	Yüksek=1, orta=2, düşük=3, yağ içeren=4
Şiddet	İyi huylu, kötü huylu

4.1.5. Meme Kanseri Veri Seti

Meme kanseri veri seti, meme kanseri olan hastaların belirli özelliklerine bakarak ışınlama tedavisi almaları gerekip gerekmediğine karar vermek amacıyla düzenlenmiştir. Veri seti 9 özellik bilgisi ve 680 örnekten oluşmaktadır. Meme kanseri özellik bilgisi Tablo 4.5'te verilmiştir.

Tablo 4.5. Meme kanseri veri tabanı özellik bilgisi

Özellik Bilgisi	
Sınıf	Tekrarlama olmayan olaylar
Yaş	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
Menopoz	40 yaş üstü,40 yaş, erken menopoz
Tümör boyutu	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
İnv. Düğümleri	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
Düğüm kapakları	Evet, hayır
Göğüs	Sol, sağ
Göğüs yuvası	Sola, sola aşağıya, sağ-yukarı, sağ düşük, orta
Işınlama	Evet, hayır

4.2. Deneysel Bulgular

Yapılan çalışmada, 5 farklı veri setinde NB ve W-NB algoritmaları ve FW-NB kullanılarak performans karşılaştırması yapılmıştır. Bu amaç doğrultusunda, her iki algoritmaya ait kodlar MATLAB programı ile hazırlanmıştır. 5 katlı çapraz geçerleme testi kullanılarak gerçekleştirilen sınıflandırma işleminde, her katmanda elde edilen sonuçların aritmetik ortalaması alınarak başarım değeri hesaplanmıştır. Ayrıca veri setlerinin özgünlük ve hassasiyet değerleri de belirlenmiştir. Her bir veri seti üzerinde ayrı ayrı gerçekleştirilen sınıflandırma işlemi sonucunda elde edilen değerler Tablo 4.6 ve Tablo 4.7'de verilmiştir.

Tablo 4.6. Sade Bayes, Ağırlıklı Sade Bayes ve Önerilen yöntem ile elde edilen başarım değerleri (%)

Tic- Tac-Toe Veri Seti			
Veri seti adı	**NB**	**W-NB**	**FW-NB**
1. Test verisi	69.79	77.60	75.52
2. Test verisi	75.00	80.20	78.12
3. Test verisi	72.39	77.60	77.60
4. Test verisi	68.75	75.00	71.87
5. Test verisi	67.70	73.95	72.39
Ortalama Başarım Değeri	**70.72**	**76.87**	**75.10**
Ameliyat Sonrası Hastanın Durumu Veri Seti			
Veri seti adı	**NB**	**W-NB**	**FW-NB**
1. Test verisi	44.44	83.33	77.77
2. Test verisi	55.55	55.55	55.55
3. Test verisi	66.66	77.77	72.22
4. Test verisi	66.66	72.22	72.22
5. Test verisi	72.22	77.77	77.77
Ortalama Başarım Değeri	**61.11**	**73.32**	**71.10**

Niteliksel İflas Veri Seti			
Veri seti adı	**NB**	**W-NB**	**FW-NB**
1. Test verisi	98.0	100	100
2. Test verisi	100	100	100
3. Test verisi	100	100	100
4. Test verisi	100	100	100
5. Test verisi	100	100	100
Ortalama Başarım Değeri	**99.6**	**100**	**100**
Mamografik Kitle Veri Seti			
Veri seti adı	**NB**	**W-NB**	**FW-NB**
1. Test verisi	84.24	86.06	84.84
2. Test verisi	83.03	86.06	85.45
3. Test verisi	76.96	82.42	82.42
4. Test verisi	84.24	89.09	88.48
5. Test verisi	86.66	88.48	89.09
Ortalama Başarım Değeri	**83.02**	**86.42**	**86.06**
Meme Kanseri Veri Seti			
Veri seti adı	**NB**	**W-NB**	**FW-NB**
1. Test verisi	97.05	99.26	97.79
2. Test verisi	93.38	98.52	98.52
3. Test verisi	94.11	96.32	96.32
4. Test verisi	97.05	98.52	97.79
5. Test verisi	97.79	98.52	97.79
Ortalama Başarım Değeri	**95.88**	**98.23**	**97.79**

Tablo 4.6 incelendiğinde, deneysel çalışma yapılan 5 farklı veri seti üzerinde de Ağırlıklı Bayes Sınıflandırıcının, Sade Bayes Sınıflandırıcıya göre başarımının daha yüksek olduğunu görülmektedir. Bu durumu deneylerin tüm katmanlarında da gözlemlenmektedir. Bu sonuca göre W-NB algoritmasının NB algoritmasına göre daha yüksek başarımla kullanılabileceği ortaya çıkmaktadır. Ancak, W-NB algoritmasında uygun ağırlıklarının bulunması işlemi gerekliliğinden dolayı bu algoritmanın çalışma zamanını NB Algoritmasına göre daha fazladır. FW-NB ile ağırlıkların belirlenmesi işlemi sağlandığında ise doğruluk değerlerinin NB sınıflandırıcıdan daha iyi olduğu ve W-NB sınıflandırıcının sonuçlarına oldukça yakın olduğu görülmektedir. Tablo 4.8'de NB, W-NB, FW-NB yöntemlerinin veri setine uygulanması sonucu hesaplanan hassasiyet ve özgünlük değerleri verilmiştir. Bu değerlerin hesaplanmasında Denklem 4.1 ve Denklem 4.2 kullanılmıştır.

$$\text{Hassasiyet} = \frac{\text{TP}}{\text{TP+FN}} \tag{4.1}$$

$$\text{Özgünlük} = \frac{\text{TN}}{\text{TN+FP}} \tag{4.2}$$

Tablo 4.7. NB, W-NB ve FW-NB ile elde edilen hassasiyet ve özgünlük değerleri

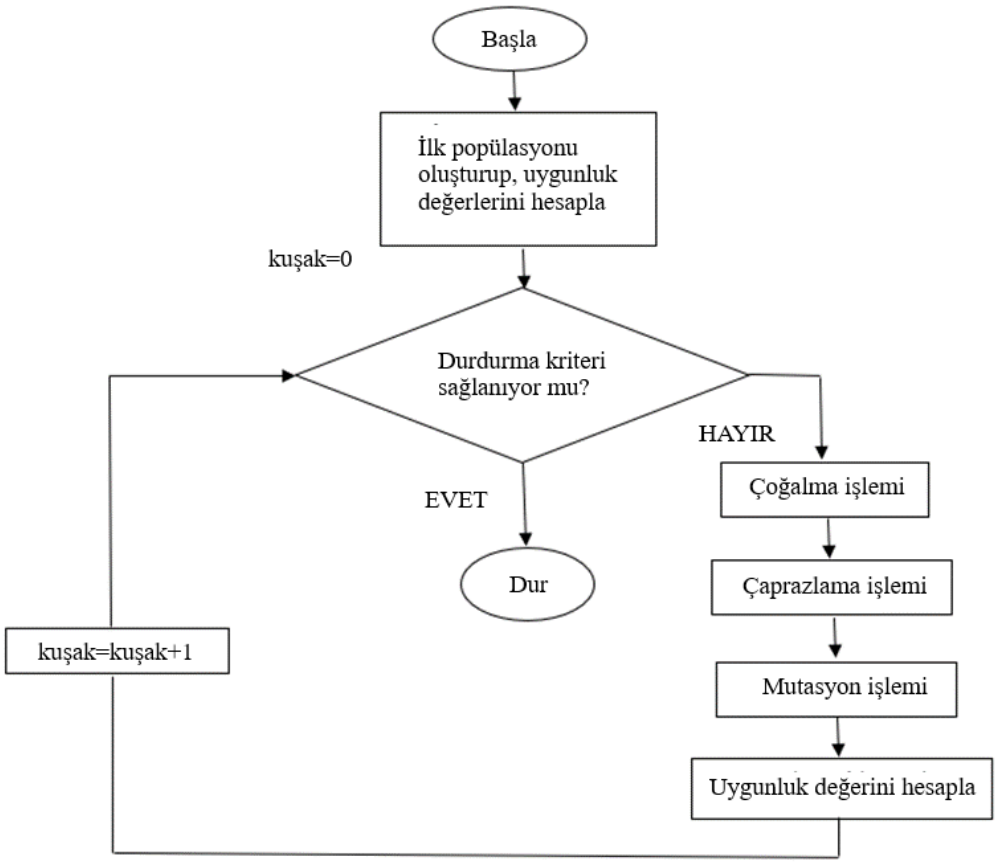
	NB		**W-NB**		**FW-NB**	
Veri Seti	Hassasiyet (%)	Özgünlük (%)	Hassasiyet (%)	Özgünlük (%)	Hassasiyet (%)	Özgünlük (%)
Tic-Tac-Toe Veri Seti	73.79	61.27	74.87	87.83	75.36	74.11
Ameliyat Sonrası Hastanın Durumu Veri Seti	68.83	15.38	70.90	75.00	71.59	50.00
Niteliksel İflas Veri Seti	99.30	100.0	100.0	100.0	100.0	100.0
Mamografik Kitle Veri Seti	80.66	85.53	88.91	84.39	85.01	87.08
Meme Kanseri Veri Seti	98.81	91.11	99.5	95.96	99.07	95.56

5. AĞIRLIKLI NAİVE BAYES'TE AĞIRLIKLARIN OPTİMİZASYONU

Optimizasyon, kelime anlamı olarak iyileştirmek, en iyi hale getirmek anlamına gelmektedir. Belirli bir fonksiyonu maksimize ya da minimize etmek amacı ile kullanılmaktadır. W-NB algoritmasında kullanılan ağırlıkların optimize edilmesi ve doğruluk değerinin maksimum değerinin bulunması için Genetik Algoritma kullanılmıştır.

5.1. Genetik Algoritma ile Ağırlıkların Optimizasyonu

Genetik algoritmalar, doğal genetik olaylar sonucu ortaya çıkan arama ve optimizasyon algoritmalarıdır. Darwin' in çevre şartlarına uyum sağlayan en iyilerin hayatta kalması mantığına dayanmaktadır. Dizilerden oluşan bir popülasyonda çoğalma, çaprazlama ve mutasyon adımlarının uygulanması adımlarını içerir. Şekil 5.1'de genetik algoritma akış diyagramı gösterilmiştir [73].



Şekil 5.1. Genetik algoritma akış diyagramı [73].

Ağırlıklar veri setinde yer alan öznitelik sayısına göre belirlenmektedir ve ağırlıklar genetik algoritmadaki karar değişkenlerini oluşturmaktadır. Karar değişkenleri 0-1 aralığındadır. Doğruluk değeri hedef fonksiyonu oluşturmaktadır. Doğruluk değerinin maksimize edilebilmesi için öncelikle rastgele belirlenen bir popülasyon oluşturulmuştur. Bu popülasyondaki değerler 8 bitlik olacak şekilde ikili kodlama işlemi gerçekleştirilmiştir. Değerlerin uygunluk değerleri doğruluk değerlerine göre belirlenmiştir. Uygunluk değerinin yüksek olması bir sonraki kuşakta yaşama şansını arttırmasına dayanan Rulet Tekerleği yöntemi kullanılarak mevcut kuşaktan yeni bir popülasyon üretilmiştir. Böylece bir sonraki popülasyonda uygunluk değeri yüksek bireylerin yaşaması, uygunluk değeri düşük olan bireylerin ise daha az yer alması sağlanmıştır. Yeni popülasyonda %60 oranında çaprazlama işlemi gerçekleştirilmiştir. Çaprazlama için rastgele seçilen 8 bitlik bireylere tek noktalı çaprazlama ile 4 bitlik değiştirme işlemi uygulanmıştır. Böylece ebeveynlerden uyum değeri yüksek bireyler elde edilmiştir. İkili kodlama ile kodlanmış dizideki, herhangi bir değer seçilerek değeri 1 ise 0, 0 ise 1 olacak şekilde değiştirilerek mutasyon işlemi gerçekleştirilmiştir. Algoritmada, yeni popülasyonun uygunluk değerleri 2000 iterasyon süresince artmadığı durum bitim şartı olarak belirlenmiştir. Yapılan işlemler için kullanılan kontrol parametreleri De Jong'a göre belirlenmiştir. Bu parametreler Tablo 5.1'de verilmiştir [74].

Tablo 5.1. Genetik Algoritma için kullanılan kontrol parametre değerleri [74].

Kontrol Parametreleri	**De Jong**
Popülasyon Büyüklüğü	50
Çaprazlama Oranı	0.60
Mutasyon Oranı	0.001

5.2. Uygulama Sonuçları

Hazırlanan algoritma 5 farklı veri setine uygulanmıştır. Dördüncü bölümünde elde edilen sonuçlar değerlendirildiğinde en yüksek performansın elde edildiği algoritma W-NB olduğu için bu bölümde karşılaştırma W-NB ile yapılmıştır. Genetik Algoritma kullanılarak ağırlıklandırma işlemi gerçekleştirilen Ağırlıklı Sade Bayes yönteminin (GAW-NB) doğruluk sonuçları ve ızgara arama yöntemi kullanılarak ağırlıkları belirlenen Ağırlıklı Sade Bayes yönteminin sonuçları Tablo 5.2'de gösterilmiştir.

Tablo 5.2. W-NB ve GAW-NB ile elde edilen doğruluk değerleri (%)

Tic- Tac-Toe Veri Seti		
Veri seti adı	**W-NB**	**GAW-NB**
1. Test verisi	77.60	79.16
2. Test verisi	80.20	81.77
3. Test verisi	77.60	77.60
4. Test verisi	75.00	75.52
5. Test verisi	73.95	74.47
Ortalama Başarım Değeri	**76.87**	**77.70**
Ameliyat Sonrası Hastanın Durumu Veri Seti		
Veri seti adı	**W-NB**	**GAW-NB**
1. Test verisi	83.33	83.33
2. Test verisi	55.55	55.55
3. Test verisi	77.77	72.22
4. Test verisi	72.22	72.22
5. Test verisi	77.77	77.77
Ortalama Başarım Değeri	**73.32**	**72.21**
Niteliksel İflas Veri Seti		
Veri seti adı	**W-NB**	**GAW-NB**
1. Test verisi	100	100
2. Test verisi	100	100
3. Test verisi	100	100
4. Test verisi	100	100
5. Test verisi	100	100
Ortalama Başarım Değeri	**100**	**100**

Mamografik Kitle Veri Seti		
Veri seti adı	**W-NB**	**GAW-NB**
1. Test verisi	86.06	86.66
2. Test verisi	86.06	86.06
3. Test verisi	82.42	83.63
4. Test verisi	89.09	89.09
5. Test verisi	88.48	88.48
Ortalama Başarım Değeri	**86.42**	**86.78**
Meme Kanseri Veri Seti		
Veri seti adı	**W-NB**	**GAW-NB**
1. Test verisi	99.26	99.26
2. Test verisi	98.52	98.52
3. Test verisi	96.32	97.05
4. Test verisi	98.52	98.52
5. Test verisi	98.52	98.52
Ortalama Başarım Değeri	**98.23**	**98.37**

Tablo 5.2 incelendiğinde, deneysel çalışma yapılan 5 farklı veri seti üzerinde de GAW-NB'nin W-NB algoritmasının performansları karşılaştırılmıştır. Ameliyat sonrası hastanın durumu veri setinin 3. test verisi dışında tüm veri setlerinde GAW-NB algoritmasının doğruluk değerinin daha yüksek olduğu gözlemlenmiştir. Bu durumu deneylerin tüm katmanlarında da görmek mümkündür. Ayrıca yapılan çalışmada GAW-NB algoritmasının çalışma zamanının, W-NB algoritmasına göre daha kısa olduğu ve arama uzayının daha geniş olabileceği sonucuna ulaşılmıştır.

6. SONUÇLAR ve TARTIŞMA

Bu tez çalışmasında, veri madenciliği sınıflandırma yöntemlerinden olan Ağırlıklı Sade Bayes algoritmasının ağırlıklarının hızlı bir şekilde elde edilmesi konusu üzerinde durulmuştur. Bu doğrultuda, ağırlıkların hızlı bir şekilde bulunması için yeni bir yaklaşım önerilmiş ve 5 farklı veri setine üzerinde deneyler yapılmıştır. Bu deneyler, literatürde yaygın olarak kullanılan Tic-Tac-Toe, Mamografik Kitle, Meme Kanseri, Niteliksel İflas ve Ameliyat Sonrası Hastanın Durumu veri setleri üzerinde gerçekleştirilmiştir. Elde edilen sonuçlara göre önerilen FW-NB algoritmasının çok hızlı çalıştığı ve W-NB algoritmasına yakın başarım değerleri elde ettiği görülmüştür. Ayrıca, FW-NB ile ağırlıkların belirlenmesi ile elde edilen başarımın değerinin, NB sınıflandırıcısına göre ise çok daha iyi olduğu da gözlenmektedir. Modelde her ağırlığın hesaplanması için çözüm uzayının arttırıldığı düşünüldüğü zaman W-NB algoritmasının çalışma süresinin, Sade Bayes ve FW-NB algoritmasına göre çok daha fazla zaman aldığı gözlemlenmiştir.

Ayrıca bu tezde, W-NB algoritmasının hem performansını arttırabilmek hem de ağırlıkların hesaplanma süresini kısaltabilmek için optimizasyon tekniklerinden biri olan Genetik Algoritmalar kullanılmıştır. Yapılan deneyler sonucunda GAW-NB'in test edilen veri setlerinin çoğunda başarımının daha yüksek olduğu görülmüştür. Kullanılan veri setleri içerisinde sadece ameliyat sonrası hastanın durumu veri setinin 3. test verisi için W-NB ile elde edilen başarım değeri GAW-NB ile elde edilen başarım değerlerine göre daha yüksek olmuştur. Önerilen GAW-NB yöntemi ile algoritmanın arama uzayının genişletebileceği ve algoritmanın çalışma zamanının kısaldığı gözlemlenmiştir.

W-NB algoritmasında, çalışma süresinin kısalması ve yüksek başarım elde edilmesi için farklı optimizasyon tekniklerinin de kullanılması mümkündür. Ayrıca literatürde bulunan diğer optimizasyon algoritmaları ile elde edilen sonuçlar NB, W-NB, FW-NB ve GAW-NB ile karşılaştırılarak değerlendirilebilir. Hız ve performans arasındaki dengeyi sağlayacak en iyi yöntem tespit edilebilir.

KAYNAKLAR

[1] **Zengin, K., Esgi, N., Erginer, E., Aksoy, M.E.**, 2011. A sample study on applying data mining research techniques in educational science: developing a more meaning of data, Procedia Social and Behavioral Sciences, **15,** 4028-4032.

[2] **Yıldırım, Ü. ve Çataltepe, Z.**, Örüntü tanıma ve öznitelik seçme yöntemleri kullanarak kısa zaman sonraki yol trafik hız öngörüsü, İstanbul Teknik Üniversitesi, İstanbul, 21-27.

[3] **Coşkun, C. Ve Baykal, A.**, 2011. Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması, XIII. Akademik Bilişim Konferansı, Malatya, Türkiye, 2-4 Şubat.

[4] **Öztürk, Ü.**, 2014. Lojistikte fiyatlandırmayı iyileştirme amaçlı olarak veri madenciliği teknikleri ile bir öneri, Yüksek Lisans Tezi, Beykent Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[5] **Uzun, Ö., Kaya, H., Gürgen, F. ve Varol, F.G.**, 2013. Olasılıksal sınıflandırıcılar ile doğum öncesinde trizomi 21 risk hesaplaması, İstanbul.

[6] **Eşiyok, T.**, 2011. Data mining supported hospital information system solutions, Graduate Thesis, Dokuz Eylül University Graduate School of Natural and Applied Sciences, İzmir.

[7] **Atak, F.**, 2014. Gerçek ağ verisi üzerinde veri madenciliği uygulamalarının karşılaştırılması, Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara.

[8] **Olgun, M.O. ve Özdemir, G.**, 2012. İstatiksel özellik temelli bayes sınıflandırıcı kullanılarak kontrol grafiklerinde örüntü tanıma, Journal of the Faculty of Engineering and Architecture of Gazi University, **Vol 27, No 2,** 303-311.

[9] **Hasanlı, H.**, 2014. Çok boyutlu veritabanlarında veri madenciliği yöntemleri kullanılarak bilgi keşfi, Yüksek Lisans Tezi, Ege Üniversitesi Fen Bilimleri Enstitüsü, İzmir.

[10] **Gökgöz, K.B.**, 2015. Veri madenciliği ile tıbbi cihaz bakım karar modeli, Yüksek Lisans Tezi, Başkent Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

[11] **Özçakır, F.C. ve Çamurcu, A.Y.**, 2007. Birliktelik Kuralı yöntemi için bir veri madenciliği yazılımı tasarımı ve uygulaması, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, İstanbul, **12,** 21-37.

[12] **Solmaz, R., Günay, M., Alkan A.**, 2014. Fomksiyonel Tiroit hastalığı tanısında Naive Bayes Sınıflandırıcının kullanılması, Akademik Bilişim Konferansı, Mersin, 891-897.

[13] **Köklü, M.**, 2014. Sınıflandırma problemlerinde kural çıkarımı için yeni bir yöntem geliştirilmesi ve uygulamaları, Doktora Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya.

[14] **Acun, G.**, 2014. Yazılım hata loğları kullanılarak veri madenciliği uygulaması gerçekleştirilmesi, Yüksek Lisans Tezi, Maltepe Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[15] **Altay, Ş.Y.**, 2014. Sağlık sektöründe aykırı yörüngelerin algılanması ve yorumlanması için mekânsal-zamansal veri madenciliği kullanımı, Yüksek Lisans Tezi, Atatürk Üniversitesi Fen Bilimleri Enstitüsü, Erzurum.

[16] **Yangın, A.**, 2017. Yapay sinir ağı teknikleri kullanarak eğitim yayıncılığı sektöründe veri madenciliği, Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[17] **Freitasa, E.F., Martins, F.F., Oliveira, A., Segundo, I.R. and Torres, H.**, 2018. Traffic noise and pavement distresses: Modelling and assessment of input parameters influence through data mining techniques, Applied Acoustics, **138,** 147-155.

[18] **Sene, A., Kamsu-Foguem, B. and Rumeau, P.**, 2018. Data mining for decision support with uncertainty on the airplane, Data & Knowledge Engineering.

[19] **Liu, L., Kantarcioglu, M. and Thuraisingham, B.**, 2008. The applicability of the perturbation based privacy preserving data mining for real-world data, Data &Knowledge Engineering, **65,** 5-21.

[20] **Compieta, P., Di Martino, S., Bertolotto, M., Ferucci, F. and Kechadi, T.**, 2007. Exploratory spatio-temporal data mining and visualization, Journal of Visual Languages & Computing, **18,** 255-279.

[21] **Song, J., Song, T. M., Seo, D. and Jin, J.H.**, 2016. Data mining of web-based documents on social networking sites that ıncluded suicide-related words among korean adolescents, Journal of Adoslescent Healt, **59,** 668-673.

[22] **Medvedev, V., Kurasova, O., Bernatavicien, J., Treigys, P., Marcinkevicius, V. and Dzemyda**, 2017. Anew web-based solution for modelling data mining processes, Simulation Modelling Practice and Theory, **76,** 34-46.

[23] **Jing, Y., Li, T., Fujita, H., Wang, B. and Cheng, N.,** 2018. An incremental attribute reduction method for dynamic data mining, Information Sciences, **465,** 202-218.

[24] **Cayci, A., Menasalvas, E., Saygin, Y. and Eibe, S.,** 2013. Self-configuring data mining for ubiquitous computing, Information Sciences, **246,** 83-99.

[25] **Chen, J., Li, K., Rong, H., Bilal, K., Yang, N. and Li, K.,** 2018. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing, Information Sciences, **435,** 124-149.

[26] **Sanmiquel, L., Rossell, J.M. and Vintro, C.,** 2015. Study of Spanish mining accidents using data mining techniques, Safety Science, **75,** 49-55.

[27] **Yuan, T. and Chen, P.,** 2012. Data Mining Applications in E-Government Information Security, International Workshop on Information and Electronics Engineering, Harbin, Heilongjiang,10-11 March, 235-240 (in China).

[28] **Naganathan, H., Chong, W.K. and Ye, N.,** 2015. Learning Energy Consumption and Demand Models through Data Mining for Reverse Engineering, Procedia Engineering, **118,** 1319-1324.

[29] **Ng, A.H.C., Bandaru, S. and Frantzen, M.,** 2016. Innovative Design and Analysis of Production Systems by Multi-objective Optimization and Data Mining, Procedia CIRP, **50,** 665-671.

[30] **Govada, A. and Sahay, S.K.,** 2016. A communication efficient and scalable distributed data mining fort he astronomical data, Astronomy and Computing, **16,** 166-173.

[31] **Karabatak, M.,** 2015. A new classifier for breast cancer detection based on Naive Bayesian. Measurement, **72**, 32-36.

[32] **Han, J., Kamber, M. and Pei, J.,** 2012. Data mining concepts and techniques, Morgan Kaufmann Publishers, Burlington, Massachusetts.

[33] **Seyrek, İ.H. ve Ata, H.A.,** 2010. Veri Zarflama analizi ve veri madenciliği ile mevduat bankalarında etkinlik ölçümü, BDDK Bankacılık ve Finansal Piyasalar, **2,** 67-84.

[34] **Witten, l.H., Frank, E., Hall, M.A. and Pal, C.J.,** 2000. Data mining practical machine learning tools and techniques, Morgan Kaufmann Publishers, Burlington, Massachusetts.

[35] **Doğan, O.,** 2015. Bir E-Ticaret sitesi kullanıcı hesaplarında şifre yapılarının birliktelik kuralları ile incelenmesi, İnternet Uygulamaları ve Yönetimi, **6(2),** 50-61.

[36] **Hoche, S., Geist, I., Castillo L. P. ve Schulz, N.,** 2006. CIGR handbook of agricultural engineering volume vı ınformation technology, CIGR-The International Commission of Agricultural Engineering, Michigan.

[37] **Johnson, C.,** 2007. Visualization and knowledge discovery: Report from the DOE/ASCR workshop on visual analysis data exploration at extreme scale, Argonne National Laboratory, Salt Lake City, Utah.

[38] **Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P.,** 1996. Knowledge discovery and data mining: Towards a unifying framework, **96,** 82-88.

[39] **Ball, G.H. and Hall, D.J.,** 1965. Iso data, a novel method of data analysis and pattern classification, Stanford Research Institute Menlo Park Technical Report, California.

[40]https://ipfs.io/ipfs/QmXoypizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco/wiki/Data_dredging.html, 8 Ağustos 2018.

[41] **Bender, R. and Lange, S.,** 2001. Adjusting for multiple testing-when and how?, Journal of Clinical Epidemiology, **54,** 343-349.

[42] Li R., History of data mining. https://hackerbits.com/data/history-of-data-mining, 10 Ağustos 2018.

[43] **LeCun Y., Bengio Y. and Hinton G.,** 2015. Deep learning. İnternational Journal of Science, **521**, 436-444

[44] **Durap, A. ve Doğan,Y.,** 2015. İnşaat mühendisliğinde bilişim kavramı ve veri madenciliği algoritmaları ile bir uzman sisteminin oluşturulması, İnşaat Mühendisliğinde Veri Madenciliği.

[45] **Larose, D.T. and Larose, C.D.,** 2005. Discovering knowledge in data, A John Wiley & Sons, Inc., Publication, Hoboken, New Jersey.

[46] **Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P.,** 1996. From Data Mining to Knowledge Discovery in Databases, 37-54.

[47] **Akpınar, H.,** 2000. Veri tabanlarında bilgi keşfi ve veri madenciliği, İ.Ü. İşletme Fakültesi Dergisi, İstanbul, **29,** 1-22.

[48] **Dolgun, M.Ö.,** 2006. Büyük alışveriş merkezleri için veri madenciliği uygulamaları, Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

[49] **Çakır, Ö.,** 2008. Veri Madenciliğinde sınıflandırma yöntemlerinin karşılaştırılması "Bankacılık müşteri veri tabanı üzerinde bir uygulama", Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

[50] **Song, Y. and Lu, Y.,** 2015. Decision tree methods: applications for classification and prediction, Shanghai Archives of Psychiatry, **27(2),** 130-135.

[51] **Sörensen, K. and Janssens, G.K.,** 2003. Data mining with genetic algorithms on binary trees, European Journal of Operational Research, **151,** 253-264.

[52] **Özçınar, H.,** 2006. KPSS sonuçlarının veri madenciliği yöntemleriyle tahmin edilmesi, Papatya Yayıncılık, İstanbul.

[53] **Özkan, Y.,** 2013. Veri madenciliği yöntemleri, Allyn & Bacon, Inc. Needham Heighs, MA, USA.

[54] **Gemici, B.,** 2012. Veri madenciliği ve bir uygulaması, Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü, İzmir.

[55] **Kayri, M. and Kayri, İ.,** 2015. The Comparison of Gini and Twoing Algorithms in Terms of Predictive Ability and Misclassification Cost in Data Mining: An Empirical study, International Journal of Computer Trends and Technology, **Vol 27, No 1,** 21-30.

[56] **Keskenler, M.F. ve Keskenler, E.F.,** 2017. Geçmişten Günümüze Yapay Sinir Ağları ve Tarihçesi, Takvim-i Vekayi, **Cilt 5, No 2,** 8-18.

[57] **Arı, A. ve Berberler, M.E.,** 2017. Yapay Sinir Ağları ile Tahmin ve Sınıflandırma Problemlerinin Çözümü İçin Arayüz Tasarımı, Acta In Pologica, **1(2),** 55-73.

[58] **Emel, G.G. ve Taşkın, Ç.,** 2002. Genetik Algoritmalar ve Uygulama Alanları, Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, **Cilt XXI, Sayı 1,** 129-152.

[59] **Sağlam, İ., Yıldız, N. ve Şimşek, A.,** THY örneği verilerin diskriminant analizi ile açıklanması, https://slideplayer.biz.tr/slide/3240250/.

[60] **Tektaş, N.,** 2014. Classification Performance Comparison of Discriminant Analysis and Logistic Regression Analysis, Turkish Studies, **Vol 9/2,** 1517-1527.

[61] **Ağırgün, B.,** 2009. Bulanık kaba küme yöntemi ile nitelik indirgemede yeni bir algoritma, Doktora Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

[62] **Shmilovici, A.,** 2010. Support Vector Machies, Data Mining and Knowledge Discovery Handbook, 231-247, Springer, Heidelberg.

[63] **Yakut, E., Elmas, B. ve Yavuz, S.,** 2014. Yapay sinir Ağları ve destek Vektör Makineleri Yöntemiyle borsa Endeksi Tahmini, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, **C.19, S.1,** 139-157.

[64] **Tabachnick, B.C. and Fidell, L.S.,** 2006. Using Mutivariate Statistics, Allyn & Bacon, Inc. Needham Heighs, MA, USA.

[65] **Abbas, O.M.A.,** 2008. Comparisons Between Data Mining Clustering Algorithms, IAJIT, **Vol. 5, No 3,** 320-325.

[66] **Fonseca, J.R.S.,** 2013. Clustering in the Field of Social Sciences: That's Your Choice, International Journal of Social Research Methodology, **16(5),** 403-42.

[67] **Yılmaz, Ö. ve Temurlenk, M.S.,** 2005. Türkiye'deki istatistik bölgelerin kişi başına düşen gelir açısından hiyerarşik ve hiyerarşik olmayan kümeleme analizi ile değerlendirilmesi: 1965-2001, Atatürk Üniversitesi İktisadi ve İdari Bilimler dergisi, **Cilt 19, Sayı 2,** 75-92.

[68] **Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I.,** 1996. Fast Discovery of Association, American Association for Artificial Intelligence Menlo Park, CA.

[69] **Argüde, Y. ve Erşahin, B.,** 2008. Veri Madenciliği, ARGE Danışmanlık Yayınları, İstanbul.

[70] **Özçalıcı, M.,** 2017. Veri Madenciliğinde Birliktelik Kuralları ve İkinci El Otomobil Piyasası Üzerine Bir Uygulama, Ordu Üniversitesi Sosyal Bilimler Araştırmaları Dergisi, **7(1),** 45-58.

[71] UCI machine learning repository, https://archive.ics.uci.edu/ml/datasets.html , 16 Şubat 2018.

[72] **Han, M. K.,** The discovery of expert decision rules from qualitative bankruptcy data using genetic algorithms, Expert System with Application, **25**, 637-646.

[73] **Taşkın, Ç. ve Emel, G.G.,** 2009. Sayısal yöntemlerde genetik algoritmalar, Star Matbaacılık, Bursa.

ÖZGEÇMİŞ

Gamzepelin AKSOY, 28 Kasım 1988 yılında Niğde'de doğmuştur. Ortaöğrenimini Elazığ Merkez Anadolu Lisesinde tamamlayarak, 2007 yılında Fırat Üniversitesi Teknik Eğitim Fakültesi Elektronik Bilgisayar Eğitimi bölümü Bilgisayar Öğretmenliği programına başlamıştır ve 2011 yılında mezun olmuştur. 2015 yılında Fırat Üniversitesi Mekatronik Mühendisliği Anabilim Dalı Mekanik Sistemler programında yüksek lisansını tamamlamıştır. 2014 yılında Mühendislik Tamamlama Programı ile Teknoloji Fakültesi Yazılım Mühendisliğine yerleştirilmiş ve 2015 yılında mezun olmuştur. 2015 yılında Fırat Üniversitesi Teknoloji Fakültesi Yazılım Mühendisliği Anabilim Dalındaki yüksek lisansına başlamıştır. Fırat Üniversitesi Teknoloji Fakültesi Yazılım Mühendisliği bölümünde ÖYP araştırma görevlisi olarak görev yapmaktadır.