ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

Burcu KALE

VERİ MADENCİLİĞİ SINIFLANDIRMA ALGORİTMALARI İLE E-POSTA ÖNEMLİLİĞİNİN BELİRLENMESİ

İSTATİSTİK ANA BİLİMDALI

ADANA-2018

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ SINIFLANDIRMA ALGORİTMALARI İLE E-POSTA ÖNEMLİLİĞİNİN BELİRLENMESİ

Burcu KALE

YÜKSEK LİSANS TEZİ

İSTATİSTİK ANABİLİM DALI

Bu Tez 20/12/2018 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından Oybirliği/Oyçokluğu ile Kabul Edilmiştir.

.....	.....	.....
Prof. Dr. Güzin YÜKSEL	Prof. Dr. Ali İhsan GENÇ	Prof. Dr. Nazif ÇALIŞ
DANIŞMAN	ÜYE	ÜYE

Bu Tez Enstitümüz İstatistik Anabilim Dalında hazırlanmıştır.
Kod No:

Prof. Dr. Mustafa GÖK
Enstitü Müdürü

Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge, şekil ve fotoğrafların kaynak gösterilmeden kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

ÖZ

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ SINIFLANDIRMA ALGORİTMALARI İLE E-POSTA ÖNEMLİLİĞİNİN BELİRLENMESİ

Burcu KALE

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman : Prof. Dr. Güzin YÜKSEL
Yıl 2018, Sayfa: 103
Jüri : Prof. Dr. Güzin YÜKSEL
: Prof. Dr. Ali İhsan GENÇ
: Prof. Dr. Nazif ÇALIŞ

Günümüzde internet ortamında yaptığımız her hareket çeşitli araştırmacılar tarafından takip edilmekte ve kayıt altına alınmaktadır. Veri madenciliği, gizli veri tahminlerini büyük veri tabanlarından çıkaran, veri ambarındaki önemli bilgileri analiz etme potansiyeline sahip güçlü ve yeni bir teknolojidir. Bu çalışmada önemli e-postaları ayırt edici belirgin farkları ortaya çıkarıp bunları öğrenen bir yapay zekâ tasarlanmaktadır. Çalışmada 19 değişkenden oluşan 4.709 birimlik veri üzerinde analizler gerçekleştirilmiştir. Decision Tree, Deep Learning, Gradient Boosted Tree, K-NN, Naive Bayes, Random Forest ve Logistic Regression sınıflandırma algoritmaları seçilmiş ve doğruluk, ağırlıklı ortalama anma, ağırlıklı ortalama duyarlılık ve kök hata kareler ortalaması istatistikleri kullanılarak karşılaştırılmış ve en iyi sınıflandırma algoritması belirlenmiştir. Analizler için Rapid Miner veri madenciliği programı kullanılmıştır.

Anahtar Kelimeler: Büyük Veri, Sınıflandırma Algoritmaları, Önemli E-postalar, Rapid Miner

ABSTRACT

MASTER THESIS

DETERMINATION OF E-MAIL IMPORTANCE WITH DATA MINING CLASSIFICATION ALGORITHMS

Burcu KALE

CUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

Supervisor : Prof. Dr. Güzin YÜKSEL
Yıl 2018, Pages: 103
Jury : Prof. Dr. Güzin YÜKSEL
: Prof. Dr. Ali İhsan GENÇ
: Prof. Dr. Nazif ÇALIŞ

Nowadays, every movement we make on the internet is followed and recorded by various researchers. Data mining is a powerful new technology that takes out hidden data estimates from large databases and has the potential to analyze important information in the data warehouse. In this study, an artificial intelligence is designed to reveal distinctive differences of important e-mails and learn them. In the study, 4.709 data consisting of 19 variables were analyzed. Decision Tree, Deep Learning, Gradient Boosted Tree, K-NN, Naive Bayes, Random Forest and Logistic Regression classification algorithms were selected and compared using accuracy, weighted mean recall, weighted mean precision and root error squares averages statistics and the best classification algorithm was determined. Rapid Miner data mining program was used for analysis.

Key Words: Big Data, Classification Algorithms, Important E-mails, Rapid Miner

GENİŞLETİLMİŞ ÖZET

Günümüzde en büyük kaynak bilgi haline gelmiştir ve artık üretilen ürünün maddesinden daha çok taşıdığı bilgi bir değer ifade etmektedir. Emek, para, kültür, ekonomi, hep bilginin içine gömülmüştür. Günümüzde işletmeler bilgi sistemlerine, yazılımlara ve iletişim araçlarına milyonlarca dolar yatırım yapmaktadırlar.

Büyük veri, yüzyıllardır kuşaktan kuşağa aktarılan bilgi birikimi olarak düşünülebilir. İnternete istenilen her yerden ulaşılabiliyor olmak ve yaşanan teknolojik gelişmeler bilgi paylaşımlarını da arttırmıştır.

Büyük veri olarak tanımlanan veriler, sensör ağlarından, nükleer santrallerden, X-ray ve tarama cihazlarından ve uçak motorlarından makine kaynaklı veriler ve sosyal medyadan tüketici odaklı verileri içerir.

Veri tabanları, kendilerini oluşturmak, güncellemek ve içerdikleri verileri sorgulamak için iş gücü sağlar. İşletmeyi yönetmemize yardımcı olan veri tabanı sistemleri ayrıca kuruluş içinde planlama, tahmin ve organizasyonla ilgili işlevleri gerçekleştirir. Bu işlevler, özellikle mevcut hızlı tempolu dünyada, işletmenin hayatta kalması için çok önemlidir.

E-posta, 1971 yılında bilgisayar mühendisi olan Raymond Samuel Tomlinson tarafından ilk kez denenmiştir. Şubat 2007‘de gmail’in beta sürümü tamamen çıkarılmış ve tüm internet kullanıcılarının kullanabileceği şekilde kurgulanmıştır.

Gmail’in tasarlanma amaçlarından biri e-postanın “önemli” ya da “önemli değil” olarak belirlenebilmesidir. Her satır ayrı bir e-posta olmak üzere her e-postanın önünde iki farklı simge bulunur (Her satır ayrı bir e-posta için). Bu simgelerden soldakini işaretleyip “More” sekmesinden “önemli olarak işaretle” sekmesini işaretleyince e-postamız önemli duruma gelir. Bunu Google, e-postalarınızı bir algoritma çevresinde değerlendirerek otomatik yapar. Ama siz de bu işaretlemeyi değiştirebilirsiniz. Örneğin bir yerden gelen e-postaları gmail size

önemli olarak işaretledi ise bu işareti kaldırırsanız önemli olmadığını belirtmiş oluyorsunuz. Tabi bu işaretlemenin sadece sizin için ve sizin kolay kullanımınız için olduğunu belirtelim. Ayrıca sol tarafta "önemli" sekmesini tıklayarak sadece önemli olarak işaretlenen iletilerinizi de görebilirsiniz.

Bu tezdeki e-posta verileri Louis Dorard'ın 2013 yılında kendi gmail ön tanımlı posta kutusundan alınmıştır. Louis Dorard, inovasyon yönetimi, veri bilimi ve web uygulamaları konusunda uzmanlaşmış bir danışmanlık şirketi olan Codole'nin ortaklarından biridir. Bu çalışmada 4.709 birimlik veri üzerinde analizler gerçekleştirilmiştir. Veri kümesi 18 değişken ve 1 sınıf değişkeninden oluşmaktadır. Bu çalışmanın amacı önemli e-postaları ayırt edici belirgin farkları çıkarıp bunları öğrenen bir yapay zekâ tasarlamaktır. Veri madenciliği, makine öğrenmesi ve veri bilimi tekniklerini kullanarak, bir yapay zekâ yaratılmış, e-postalar seçilen algoritmalara göre analiz edilmiş ve e-postanın önemlilik potansiyeli hakkında bilgi verilmiştir. Çalışmada Rapid Miner yazılımı kullanılmıştır. Rapid Miner yazılımındaki sınıflandırma algoritmaları veriye uygunluğu göz önünde bulundurularak denenmiş, elde edilen sonuçlarda sınıflandırma algoritmalarının nasıl sonuç verdiği gözlemlenmiştir. Kullanılan sınıflandırma algoritmaları; Decision Tree, Deep Learning, Gradient Boosted Tree, K-NN, Naive Bayes, Random Forest ve Logistic regression algoritmalardır. Sınıflandırma algoritmaları karşılaştırılmış ve önemli e-posta olma olasılıkları belirlenmiştir. Algoritmaların performanslarının karşılaştırılması doğru sınıflandırma yüzdesi (Accuracy), anma (Recall), duyarlılık (Precision) ve kök hata kareler ortalaması (RMSE) ölçütlerine göre yapılmıştır.

Bu çalışma benzer veri kümesine sahip araştırmacılara ve e-postanın önemlilik potansiyeli hakkında bilgi edinmek isteyen kuruluşlara örnek teşkil edebilir. Araştırmada veri madenciliği kullanılarak, hangi e-postaların önemlilik arz ettiği belirlenmiştir.

TEŞEKKÜR

Çalışmamın her aşamasında yardımlarını esirgemeyen, beni veri madenciliği konusunda yönlendiren ve bana "Veri Madenciliği Sınıflandırma Algoritmaları İle E-posta Önemliliğinin Belirlenmesi" konulu yüksek lisans tezini veren yapıcı ve yönlendirici fikirleri ile bana daima yol gösteren danışman hocam Sayın Prof. Dr. Güzin YÜKSEL'e sonsuz teşekkürler.

Çalışmalarımın en sıkıntılı zamanlarında dahi yardımlarını ve desteklerini esirgemeyen aileme sonsuz teşekkürler…

İÇİNDEKİLER **SAYFA**

ÖZ I
ABSTRACT II
GENİŞLETİLMİŞ ÖZET III
TEŞEKKÜR V
İÇİNDEKİLER VI
ÇİZELGELER DİZİNİ X
ŞEKİLLER DİZİNİ XII
SİMGELER VE KISALTMALAR XIV
1. GİRİŞ 1
2. ÖNCEKİ ÇALIŞMALAR 5
3. BÜYÜK VERİ 11
3.1. Veri Toplama Yöntemleri 15
3.2. Yapılandırılmış, Yapılandırılmamış ve Yarı Yapılandırılmış Veri 16
3.3. Büyük Verinin 4 V'si 17
3.3.1. Hacim (Volume): 17
3.3.2. Hız (Velocity) 19
3.3.3. Çeşitlilik (Variety) 20
3.3.4. Doğruluk (Veracity) 20
3.4. Büyük Veri Kullanım Alanları 21
3.5. Veri Hacminin Büyümesi 22
4. VERİ TABANI 25
5. VERİ MADENCİLİĞİ 27
5.1. Veri Madenciliği Aşamaları 28
5.2. Veri Madenciliğinin Uygulanabildiği Veriler 30
5.2.1. Veri Ambarı (Data Warehouse) 30
5.2.2. İlişkisel Veri Tabanı (Relational Databases) 32
5.2.3. İşlemsel Veri Tabanı (Transactional Databases) 33

5.2.4. Düz Dosya (Flat File) 34
5.2.5. Diğer Veri Çeşitleri 34
5.3. Makine Öğrenme 34
5.4. Veri Madenciliği Ve İstatistik 36
5.5. Veri Madenciliği Sorunları 37
5.6. Veri Madenciliği Kullanım Alanları 39
5.7. Veri Madenciliğinde Gelişmeler 40
5.8. Veri Madenciliği Yöntemleri 42
5.8.1. Gözetimli Öğrenme 42
5.8.1.1. Veri madenciliğinde sınıflandırma ve tahmin 43
5.8.1.2. Regresyon Analizi (Regression Analysis) 43
5.8.2. Gözetimsiz Öğrenme 45
5.8.2.1. Kümeleme (Clustering) 45
5.8.2.2. Birliktelik Kuralları (Association Rules) 47
5.9. Veri Madenciliği Yazılımları 48
6. MATERYAL VE YÖNTEM 49
6.1. Önemli E-Posta Olma Olasılığı 49
6.2. RAPIDMINER Veri Madenciliği Yazılımı 56
6.3. Sınıflandırma Algoritmaları ve E-mail Önemlilik Analizi 59
6.3.1. En Yakın K Komşuluğu algoritması (K-NN) 59
6.3.2. Bayes Sınıflandırması 61
6.3.2.1. Naive Bayes 62
6.3.3. Karar Ağaçları (Desion Trees) 63
6.3.4. Rastgele Orman (Random Forest) 67
6.3.5. Gradient Boosted Trees 69
4.3.6. Derin Öğrenme (Deep Learning) 70
6.3.7. Lojistik Regresyon (Logistic Regression) 72
7. UYGULAMA Ve ANALİZLER 73
7.1. Veri Kümesi ve Yapısı 73

7.2. Performansların Karşılaştırılması 80
7.2.1. Karar Ağacı Algoritması 81
7.2.2. Rastgele Orman Algoritması 83
7.2.3. Gradient Boosted Algoritması 84
7.2.4. K-NN Algoritması 86
7.2.5. Naive Bayes Algoritması 87
7.2.6. Lojistik Regresyon Algoritması 89
7.2.7. Derin Öğrenme Algoritması 90
7.3. Performansların Karşılaştırılması 92
8. SONUÇLAR VE ÖNERİLER 95
KAYNAKLAR 97
ÖZGEÇMİŞ 103

ÇİZELGELER DİZİNİ **SAYFA**

Çizelge 3.1. Veri Dönüşüm Tablosu (Bayrakçı, 2015) 18
Çizelge 7. 1. Hata Matrisi 80
Çizelge 7.2. Sınıflandırma Algoritmalarının Karşılaştırılması 92

ŞEKİLLER DİZİNİ **SAYFA**

Şekil 6.1. Gmail Kullanıcı Ara Yüzü 49
Şekil 6.2. Compose Kullanıcı Ara Yüzü 50
Şekil 6.3. Gelen E-postanın Bilgilerinin Yer Aldığı Kullanıcı Ara Yüzü 51
Şekil 6.4. Reply Veya Forward Simgelerinin Yer Aldığı Kullanıcı Ara Yüzü 52
Şekil 6.5. Filtreleme yapılan Kullanıcı Ara Yüzü 54
Şekil 6.6. Mailtrack Uzantısının Oluşturduğu Raporun Gösterildiği Pencere 55
Şekil 6.7. Rapid Miner Penceresi 58
Şekil 6.8. K-NN Örneği 60
Şekil 6.9. Örnek Karar Ağacı 64
Şekil 7.1. Rapid Miner Veri Görünümü 77
Şekil 7.2. Rapid Miner Parameters Penceresi Görünümü 79
Şekil 7.3. Rapid Miner Akış Şeması 82
Şekil 7.4. Decisiom Tree Confusion Matrisi 82
Şekil 7.5. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu .. 83
Şekil 7.6. Rapid Miner Akış Şeması 83
Şekil 7.7. Random Forest Confusion Matrisi 84
Şekil 7.8. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu .. 84
Şekil 7.9. Rapid Miner Akış Şeması 85
Şekil 7.10. Gradient Boosted Tree Confusion Matrisi 85
Şekil 7.11. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu 86
Şekil 7.12. Rapid Miner Akış Şeması 86
Şekil 7.13. K-NN Confusion Matrisi 87
Şekil 7.14. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu 87
Şekil 7.15. Rapid Miner Akış Şeması 88
Şekil 7.16. Naive Bayes Confusion Matrisi 88
Şekil 7.17. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu 89
Şekil 7.18. Rapid Miner Akış Şeması 89

Şekil 7.19. Logistic Regression Confusion Matrisi 90
Şekil 7.20. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu 90
Şekil 7.21. Rapid Miner Akış Şeması 91
Şekil 7.22. Deep Learning Confusion Matrisi 91
Şekil 7.23. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu 92
Şekil 7.24. Decision Tree Örnek Seti 93
Şekil 7.25. Gradient Boosted Tree Örnek Seti 94

SİMGELER VE KISALTMALAR

MWC	: Mobile World Congress
IDC	: International Data Corporation
RMSE	: Root Mean Square Error (Kök Hata Kareler Ortalaması)
WEKA	: Waikato Environment for Knowledge Analysis
KNIME	: Konstanz Information Miner
SEER	: Surveillance Epidemiology and End Results
GPU	: Graphic processing unit (Grafik İşlem Birimi)
RO	: Rastgele Orman
NASA	: National Aeronautics and Space Administration
MAE	: Mean Squared Error (Hata Kareler Ortalaması)
CPU	: Central Processing Unit (Merkezi İşlem Birimi)
RDBMS	: Relational Database Management System (İlişkisel Veri Tabanı Yönetim Sistemleri)
SQL	: Structured Query Language (Yapısal Sorgulama Dili)
ER	: Entity-Relationship Model (Varlık-İlişki Modeli)
BLOB	: Binary Large Objects (İkili Büyük Nesneler)
ODBMS	: Object-Oriented Database Management System (Nesne Yönelimli Veritabanı Yönetim Sistemi)
SAP	: System, Application&Products
XML	: extensible Markum Language (Genişletilebilir İşaretleme Dili)
GPS	: Global Positioning System (Küresel Yer Belirleme Sistemi)
NIST	: National Institute of Standars and Technology
CERN	: Conseil Européen pour la Recherche Nucléaire (Avrupa Nükleer Araştırma Merkezi)
LHC	: Large Hadron Collider (Büyük Hadron Çarpıştırıcısı)
KDD	: Knowledge Discovery in Database

SPSS	: Statistical Package for the Social Sciences (Sosyal Bilimler İçin İstatistik Programı)
MATLAB	: Matrix Labaratory
KEEL	: Knowledge Extraction Based on Evolutionary Learning
SPF	: Sender Policy Framework
YALE	: Yet Another Learning Environment
CART	: Classification and Regression Trees
ID3	: Iterative Dichotomiser 3
CAP	: Credit Assignment Path
SPF	: Sender Policy Framework
AGPL	: General Public Lisans

1. GİRİŞ

Günümüzde sadece bilgiye ulaşmak değil, gerekli koşullarda bilgi üretmek de önemli bir konu halini almıştır. Çığ gibi büyüyen sayısal veri ortamları arasından yararlı ve de gerekli olan bilgiye ulaşmayı sağlamak gerçek bir çaba haline gelmiştir. Veri madenciliği bu safhada göze çarpan bir olgudur. Veri madenciliği, gizli veri tahminlerinin büyük veri tabanlarından çıkarılması, veri ambarındaki önemli bilgileri analiz etme potansiyeline sahip güçlü ve yeni bir teknolojidir.

Bir veri ambarı, "İş ne olur?" diye sorar. Bu soruların cevabı, işletmenizin günümüzün enformasyonu için bir gereklilik olan reaktif değil, proaktif olmasını sağlayacaktır. Bugün endüstri eğilimi daha güçlü donanım ve yazılım konfigürasyonuna doğru ilerliyor, şimdi analitik olarak çok fazla bilgiyi işleme yeteneğine sahibiz ki bu altı yıl önce duyulmamıştı bile. Bugün bir işletme, bu yeni teknolojiyi kullanabilmeli veya riski kabul etmelidir. Risk ise bilgilendirilmiş kararlar almak için gerekli olan kritik bilgileri kaçırmaktır. Günümüz dünyasında, teknolojiyi büyük miktarlardaki bilginin üzerinden geçmek için etkin bir şekilde kullanmayan bir işletme, bilgi çağında hayatta kalmayacaktır. Bilgiye erişim ve bilgiyi anlama bir güçtür. Bu güç, rekabet avantajı ve hayatta kalmaya eşittir. Bir birim, bilgi çağında hayatta kalmayı ve gelişmeyi umuyorsa, işletmenin ihtiyaç duyduğu bilgileri depolamak için kendi veri ambarını inşa etmek zorundadır (Rai Technology University, 2018).

Barcelona'da düzenlenen Mobil Dünya Kongresi'nde (Mobile World Congress-MWC) Mobil Ekonomi 2017 başlıklı rapora göre, 2016 sonu itibariyle dünyada mobil abone sayısı 4,8 milyar oldu. 2020 sonundaki beklentisi ise 5,7 milyar aboneye ulaşılacağı yönünde. Bu veriye göreyse 2020'de dünyanın üçte ikisinin mobil kullanıcı olması bekleniyor (Aydın,2017). Küresel olarak, yılda yaklaşık 4.4 ZB elektronik veri üretilmektedir. IDC (International Data Corporation) kurumsal verilerin 2020'ye kadar 40 ZB'ye ulaşacağını iddia ediyor.

Dünya genelinde yaratılan e-posta hesaplarının sayısı 2012 yılında 3,3 milyardan 2016 yılının sonuna kadar 4,3 milyara yükselmiştir. 2012 yılında, gönderilen ve alınan e-posta sayısı günde 89 milyardı. Bu miktar 2016 yılı sonunda 143 milyarın üzerine çıkmıştır. 2016 yılında günde dünya genelinde gönderilen e-posta sayısı 170 milyar civarındadır. Her saniye gönderilen e-posta sayısı 2 milyondur. Son 3 yılda mobil cihazlarda e-posta açma oranı %180 yükselmiş durumdadır (Yaqoob ve ark., 2016).

Verilerin içsel değeri vardır. Ancak bu değer keşfedilene kadar hiçbir faydası yoktur. Bugün, büyük veri sermaye haline geldi. Dünyanın en büyük teknoloji şirketlerinden bazılarını düşünürsek, sundukları değerin büyük bir kısmının, daha fazla verimlilik sağlamak ve yeni ürünler geliştirmek için sürekli olarak analiz ettikleri verilerden geldiğini görürüz.

Son teknolojik gelişmeler, eskisinden daha fazla veriyi depolamayı daha kolay ve daha ucuz hale getirmiştir. Bundan dolayı, daha doğru ve hassas iş kararları verebilirsiniz.

Büyük veride değer bulmak sadece onu analiz etmekle ilgili değildir. Bu, doğru soruları soran, kalıpları tanıyan, bilinçli varsayımlar yapan ve doğru davranışı öngören anlayışlı analistler ve yöneticiler gerektiren bir keşif sürecidir.

Veri madenciliği ziraat alanında, mühendislikte, ekonomide, perakende sektöründe, sağlık sektöründe ve hatta politikada kullanılmaktadır. Veri Madenciliğinin kullanıldığı alanlara örnek olarak, sunulan değeri en üst düzeye çıkarmak için sosyal medyadan, web ziyaretlerinden, arama günlüklerinden ve diğer veri kaynaklarından veri toplamak, kişiselleştirilmiş teklifler sunmaya başlamak ve müşteri kaybını azaltmak için sorunları temkinli olarak ele almaktır. Bir başka örnek, finansal ve planlama konuları ile ilgili kararları iyileştirmek için veri analizlerini kullanmak; trendleri ve müşterilerin istediği yeni ürün ve hizmetleri incelemek ve dinamik fiyatlandırma yapmaktır (ORACLE, 2018).

Bu tezdeki e-posta verileri Louis Dorardın 2013 yılında kendi gmail ön tanımlı posta kutusundan alınmıştır (Bigml,2018). Louis Dorard, inovasyon

yönetimi, veri bilimi ve web uygulamaları konusunda uzmanlaşmış bir danışmanlık şirketi olan Codole'nin ortaklarından biridir. Bu çalışmada 4.709 birimlik veri üzerinde analizler gerçekleştirilmiştir. İçerik özellikleri güvenlik nedeniyle kaldırılmıştır. Veri kümesi 18 değişken ve 1 sınıf değişkeninden oluşmaktadır. Çalışmada Rapid Miner yazılımı kullanılacaktır. Rapid Miner yazılımındaki sınıflandırma algoritmaları veriye uygunluğu göz önünde bulundurularak denenecek, elde edilen sonuçlarda sınıflandırma algoritmalarının nasıl sonuç verdiği gözlemlenecektir. Bu sınıflandırma algoritmaları; Decision Tree, Deep Learning, Gradient Boosted Tree, K-NN, Naive Bayes, Random Forest ve Logistic Regression algoritmalardır. Sınıflandırma algoritmaları karşılaştırılmış ve önemli e-posta olma olasılıkları belirlenmiştir. Algoritmaların performanslarının karşılaştırılması için doğru sınıflandırma yüzdesi (Accuracy), anma (Recall), duyarlılık (Precision) ve kök hata kareler ortalaması (RMSE) ölçütleri kullanılmıştır

Bu çalışmanın amacı önemli e-postaları ayırt edici belirgin farkları çıkarıp bunları öğrenen bir yapay zekâ tasarlamaktır. Bu çalışma benzer veri kümesine sahip araştırmacılara ve e-postanın önemlilik potansiyeli hakkında bilgi edinmek isteyen kuruluşlara örnek teşkil edebilir. Bu araştırmada, en uygun sınıflandırma algoritması veriye uygulanarak, hangi e-postaların önemlilik arz ettiğinin belirlenmiştir. Bu çalışma sayesinde büyük şirketlerde çalışanların iş gücü yükünü hafifletebiliriz. Gmail kullanıcıların önemli e-postaları algılaması için bir kural kümesi oluşturmasına izin verir, ama önemli postaların niteliği zamanla değiştikçe, bu kural kümeleri, kullanıcı tarafından sürekli olarak ayarlanmalı ve yeniden ayarlanmalıdır. Bu zaman alıcı ve genellikle hataya eğilimli, sıkıcı bir süreçtir. Yani kural tabanlı bir yaklaşımın önemli posta filtrelemede sınırlı bir faydası vardır. Bu çalışma sayesinde hangi e-postaların önemli olduğu daha kolay yoldan belirleyebiliriz. Ayrıca reklam içerikli ve okumaya gerek duymadığımız e-postalarda bu yolla daha kolay ortaya çıkar.

2. ÖNCEKİ ÇALIŞMALAR

Krishnan (2013) büyük veri ve veri ambarını incelemiştir.

Yaqoob ve arkadaşları (2016) büyük veri uygulamalarının kökenlerini ve güncel eğilimlerini analiz etmekte, en gelişmiş teknolojiler hakkında kapsamlı bir tartışma sunmaktadır.

Hurwitz ve arkadaşları (2013) bu çalışmada yapılandırılmış ve yapılandırılmamış veriyi büyük veri çevrenize entegre etmeyi öğretmektedir.

Dietrich ve arkadaşları (2015) veri bilimi ve büyük veri analitiği konusunu her yönüyle incelemiştir.

Bayrakçı (2015) sosyal bilimlerdeki akademik çalışmalarda büyük veri kullanımını incelemiştir.

Liew (2013) DIKW hiyerarşisini, verilerin, malumatın, bilginin, bilgeliğin ve bunların karşılıklı ilişkilerini açıklamıştır.

Birincil veriler ve ikincil veriler arasındaki farkları Ajayi (2017)'de ele almıştır.

Gunjal (2003) veri tabanı, fonksiyonları ve türleri, veri yapısı kavramları, veri tabanı dizaynı konularında bilgi vermiştir.

Öğüt (2005) veriden enformasyona ordan bilgiye ve sonuç olarak bilgeliğe ulaşılması sürecini ele almıştır. Veri madenciliğinin gelişim sürecini incelemiştir.

Çığşar (2017) bankacılık ve sigortacılık gibi alanlarda, büyük veri analizlerini, veri madenciliği analizlerini, temerrüde düşme olasılığını hesaplamak için kullanmıştır. Analizler için WEKA açık kaynak kodlu veri madenciliği yazılımı kullanmıştır.

Ayre (2006) veri madenciliğinde bilgi edinme ile metin madenciliği ilişkisini incelemiştir.

Ramzan ve Ahmad (2014) veri madenciliğinin başlangıcından bugüne gelişim aşamasına kadar olan evrimine genel bir bakış sağlamıştır.

Grivas (1999) veri ambarının temel kavram ve mekanizmalarını tanıtmaktadır.

Chao (2011) makine öğrenmesi, temel konseptleri, prensipleri ve etkileri konularında bilgi vermektedir.

Kuonen (2004) araştırmasında, istatistiğin ve veri madenciliğinin ne olduğu, neden ihtiyaç duyulduğu konularını açıklayarak istatistik ve veri madenciliği arasındaki bağlantıyı vurgulamıştır.

Aydın (2017) mobil e-ticaretin Türkiye'de ve Dünya'da yeri, önemi ve kullanıcılarının her geçen gün artmasını değerlendirmiştir.

Alpaydın (2014) gözetimli öğrenme olan sınıflandırma ve regresyon algoritmaları, gözetimsiz öğrenme olan birliktelik kuralı hakkında bilgi vermiştir.

Jantan ve arkadaşları (2011) insan yetenekleri tahmini için veri madenciliği tekniklerini incelemiştir.

Han ve arkadaşları (2012) veri madenciliği, veri hazırlığı, veri ambarı, veri küp teknolojisi, madencilikte birliktelik ve korelasyonda temel kavram ve yöntemler, sınıflandırma, kümeleme analizi gibi konuları ele almıştır.

Mooi ve Sarstedt (2014) regresyon analizi nedir, neye yarar, SPSS'de regresyon analizi nasıl yapılır ve SPSS tarafından üretilen regresyon analizi çıktısı nasıl yorumlanır konularına açıklık getirmiştir.

Pacheco (2015) veri analiz tekniklerini uygulamak için R'yi kullanmayı öğrenmeye çalışan veri analistlerine, istatistikçiler ve veri bilimcilerine bunun yanı sıra denetimsiz öğrenme teknikleri kullanarak veri analizi yapmak isteyen profesyonellere göre konuları incelemiştir.

Avcılar (2014) veri madenciliğinde birliktelik kurallarını açıklamıştır. Ayrıca özel bir giyim ve aksesuar mağazasında birliktelik kuralını kullanarak uygulama yapmıştır.

Jovic ve arkadaşları (2014) günümüzde mevcut olan genel veri madenciliği için en çok kullanılan altı ücretsiz yazılım aracı Rapid Miner, R, Weka, KNIME, Orange ve Scikit-learn'i karşılaştırmış, artılarını ve eksilerini belirtmiştir.

Coşkun ve Baykal (2011) WEKA (Waikato Environment for Knowledge Analysis) programı ile SEER (Surveillance Epidemiology and End Results) veri kaynağı üzerinde çeşitli sınıflandırma metotlarına ait algoritmalar kullanarak modeller oluşturmuşlar, elde edilen modellerin tahminleme başarılarını karşılaştırarak hangi algoritmanın daha başarılı modeller oluşturduğunu incelemişlerdir.

Ceyhan ve Sağıroğlu (2015) çalışmasında adli bilişim ve karşı adli bilişim hakkında bilgi vererek e-postaların adli olarak incelenmesini ve karşı adli bilişim tekniklerini açıklamıştır. E-postaların yapısı ve gönderim aşamalarından bahsedilmiştir.

Şeker (2017) Rapid Miner ile veri madenciliğini anlatmıştır. Veri madenciliği yöntemleri ile twitter üzerinden girişimcilik analizi yapmıştır.

Sahami ve arkadaşları (1998) önemsiz e-postaları filtrelemek için Bayesci yaklaşımı uygulamıştır.

Selvaluxmiy ve arkadaşları (2016) grafik İşlem Birimini (GPU) kullanarak, K-NN sınıflandırma algoritmasını hızlandırarak, daha hızlı bir yürütme süresi göstermesini sağlamıştır.

Dolgun (2014) decision tree, logistic regression, naive bayes ve support vector machines gibi veri madenciliği yöntemleri kullanılarak bankacılık sektöründe doğrudan pazarlama kampanyalarının nasıl yönlendirilebileceği üzerinde durmuştur.

Haciefendioğlu (2012) çalışmasında makine öğrenmesi sınıflandırma yöntemlerinden 3 önemli yöntem olan Support Vector Machines, Artificial Neural Networks ve Decision Tree kullanılarak glokom hastalığının başlangıç safhasında teşhisi için sınıflandırma yapılmış ve birbirleri ile karşılaştırılmıştır. Adı geçen makine öğrenmesi yönetmelerinin performansları X-Validation ile belirlenmiş ve en yüksek sınıflandırma başarısının Support Vector Machines ile elde edileceği görülmüştür.

Ok ve arkadaşları (2011) bu araştırmasında Random Forest sınıflandırma yöntemi kullanılarak, Trabzon ilini kapsayan 4 bantlı IKONOS uydu görüntüsü sınıflandırılmıştır. Random Forest sınıflandırma algoritması ile sınıflandırılan bu görüntü, yaygın olarak kullanılan ECHO Spectral-Spatial, En Çok Benzerlik, Fisher Lineer Diskriminant, En Küçük Mesafe gibi sınıflandırma yöntemleriyle karşılaştırılmıştır. Random Forest ve diğer kullanılan yöntemlerin performanslarının karşılaştırılması sonucunda, %96.80 genel sınıflandırma doğruluğuyla Random Forest sınıflandırma yönteminin en iyi yöntem olduğu gözlenmiştir.

Guo ve arkadaşları (2004) Random Forest'lara dayalı olarak, hata eğilimli modülleri tahmin etmek için yeni bir metodoloji sunmaktadır. NASA veri setlerine dayanan örnek olaylarda Random Forest uygulamışlardır. Önerilen metodolojinin tahmin doğruluğu genel olarak Logistic Regression, Discriminant Analysis ve iki makine öğrenme yazılımı paketindeki algoritmalar, WEKA [32] ve See5 [5] ile elde edilen değerden daha yüksek çıkmıştır.

Krauss ve arkadaşları (2016) Deep Neural Networks, Gradient Boosted Tree ve Random Forest etkisini istatistiksel olarak ve döviz kur farklarından işlem yaparak sağlanan kar içeriği için analiz etmişlerdir.

Korkmaz ve arkadaşları (2012) hem sürekli hem ayrık olan ikili hedef değişken ve bağımsız değişkenler arasındaki ilişkiyi tanımlayabilen Logistic Regression analizini incelediler. Ayrıca hayvancılıkta Logistic Regression'un uygulanabilirliğini araştırdılar.

Baydemir (2014) Logistic Regression modelini kullanarak İnönü Üniversitesi idari personelinin sigara içme olasılıklarını belirleyen etmenler üzerinde durmuş, analizleri SPSS paket programı kullanılarak yapmıştır.

Arabacı (2007) veri madenciliğinde apriori, tahminci apriori ve tertius algoritmalarını WEKA ve YALE programları ile karşılaştırmış ve bir uygulama yapmıştır.

Doğan ve Türkoğlu (2018) bitki yapraklarının sınıflandırılması için kullanılan pek çok yönteme karşı deep learning yoluyla sınıflandırılma yaparak deep learning algoritmalarının başarımlarını ortaya koymuşlardır.

Rende ve arkadaşları (2017) Deep Learning algoritmasının eğitim süreci sonunda ortaya çıkan modellerin birden fazla veri seti ile otomatik olarak testlerinin yapılmasına ve bu testler neticesinde kullanılacak modelin otomatik olarak belirlenmesine yönelik çalışmalar yapmıştır.

Chai ve Draxler (2014) model değerlendirme çalışmalarında düzenli olarak kullanılmakta olan temel kök hata kareler ortalaması (RMSE) hem de hata kareler ortalaması (MAE) da dahil olmak üzere çoğu zaman, model performansını değerlendirmek için bir metrikler kombinasyonunun gerektiğini belirtmişlerdir.

3. BÜYÜK VERİ

Bilgi günümüzde en büyük kaynak haline gelmiştir ve artık üretilen ürünün maddesinden daha çok taşıdığı bilgi bir değer ifade etmektedir. Emek, para, kültür, ekonomi, hep bilginin içine gömülmüştür. Günümüzde işletmeler bilgi sistemlerine, yazılımlara ve iletişim araçlarına milyonlarca dolar yatırım yapmaktadırlar.

Büyük veri; yüzyıllardır kuşaktan kuşağa aktarılan bilgi birikimi olarak düşünülebilir. İnternete istenilen her yerden ulaşabiliyor olmak ve yaşanan teknolojik gelişmeler bilgi paylaşımlarını da arttırmıştır. Bugün kağıt gazete okuyucularının sayısı gitgide azalmakla beraber 78 milyondan fazla insan gazeteleri internet üzerinde takip etmektedir. Her gün 39 milyon kişi çevrimiçi video izlemekte, 66 milyon kişi blog oluşturmaktadır. Sosyal ağ sitesi Facebook ABD'de 2010 yılında aylık 134 milyon, dünya çapında ise 500 milyondan fazla kullanıcı tarafından ziyaret edilmiştir (Ayvaz, 2018).

Büyük Veri olarak tanımlanan veriler, sensör ağlarından, nükleer santrallerden, X-ray ve tarama cihazlarından ve uçak motorlarından elde edilen makine kaynaklı verileri ve sosyal medyadan tüketici odaklı verileri içerir. Bu veriler; operasyonların bilgisayar tarafından gerçekleştirildiği, elektrik sinyalleri şeklinde saklanabilen ve iletilebilen; manyetik, optik veya mekanik kayıt ortamına kaydedilen karakter veya sembollerdir. Bir bilgisayar tarafından işlenen veya saklanan bilgiler; metin belgeleri, resimler, ses klipleri, yazılım programları veya diğer veri türleri biçiminde olabilir. Bilgisayar verileri bilgisayarın CPU'su tarafından işlenir ve bilgisayarın sabit diskindeki dosya ve klasörlerde depolanır (Krishnan,2013).

Büyük verileri analiz etmek, her boyutta ve tüm sektörlerdeki kuruluşlar için en büyük avantajları ve en büyük zorlukları her zaman sunmuştur. İşletmeler; müşterileri, ürünleri ve hizmetleri hakkında bilgi toplamak için pragmatik bir yaklaşım bulmakla uzun zaman uğraştılar. Bir şirket sadece aynı ürünü aynı şekilde

satın alan bir avuç müşteriye sahip olduğunda, işler oldukça basitti. Ancak zamanla, şirketler ve katıldığı pazarlar daha karmaşık hale geldi. Müşterilerle ilgili rekabet avantajı elde etmek için bu şirketler daha fazla ürün grubu eklediler ve ürün teslim yöntemlerini çeşitlendirdiler. Araştırma ve geliştirme (Ar-Ge) kuruluşları, karmaşık modelleri çalıştırmak veya görüntüleri ve diğer bilimsel veri kaynaklarını işlemek için yeterli bilgi işlem gücü elde etmek için çaba sarf etmiştir. Belgeler, müşteri hizmetleri kayıtları ve hatta resimler ve videolar dahil olmak üzere bazı veriler yapılandırılmamışken, diğer veriler geleneksel bir ilişkisel veri tabanında yapılandırılmış ve saklanmıştır. Şirketler ayrıca, sensörler gibi makineler tarafından üretilen yeni veri kaynaklarını da dikkate almak zorundadırlar. Sosyal medyadan elde edilen veriler ve web sitelerinde tıklanma verileri gibi diğer yeni bilgi kaynakları da insan kaynaklı olarak oluşturulmaktadır. Ek olarak, daha yeni, daha güçlü mobil cihazların kullanılabilirliği ve benimsenmesi, küresel ağlara her yerde erişimle birleştiğinde, veriler için yeni kaynakların oluşturulmasını teşvik edecektir.

Organizasyonlar her geçen gün verileri kullanmanın daha fazla yolunu bulmaktadır. Her bir veri yönetimi dalgası, belirli bir tür veri yönetimi problemini çözmek gerekliliğinden doğar. Bu dalgaların veya fazların her biri neden ve sonuçtan dolayı evrimleşmiştir. Yeni bir teknoloji çözümü piyasaya çıktığında, yeni yaklaşımların keşfini gerektirmektedir. Son beş yılda veri yönetimi dalgaları bugün bulunduğumuz noktaya ulaşmıştır: büyük veri çağının başlangıcı. Yani, büyük verilerin anlaşılması için, bu önceki dalgaların temelinin bilinmesi gerekir. Ayrıca, bir dalgadan diğerine geçerken, önceden kullanılan araçları, teknolojileri ve uygulamalarında gereksinimleri önem arz etmektedir.

Hesaplama 1960'ların sonlarında ticari pazara girdiği için, veri hiçbir yapıya dayanmayan düz dosyalarda saklandı. Firmalar, müşteriler hakkında ayrıntılı bir bilgiye ulaşmaya ihtiyaç duyduklarında, insan kaynaklı güç uygulamak zorunda kalıyorlardı. Daha sonra 1970'lerde, ilişkisel veri tabanı; yani organize edilmiş verileri satır ve sütunlar halinde tablolarda saklandığı ve bu tablolar

arasında bağların kurulduğu veri tabanı çeşidi kullanılmaya başlandı. Tablolar üzerinde veriler saklanabilir, eklenebilir, silinebilir ve güncellenebilirdi. İlişkisel veri tabanlarını korumak için ilişkisel veri tabanı yönetim sistemi (RDBMS) olarak bilinen yazılım sistemi gerçekleştirildi. RDBMS, kullanıcı ile veri tabanı arasında bir arabirim oluşturdu ve veri tabanına her türlü erişimi sağladı. Tüm ilişkisel veri tabanı sistemleri dil olarak SQL (Structured Query Language) kullanmaktaydı. SQL veri ekleme, değiştirme, silme ve geri alma için kullanılan bir veri işleme dili olarak tanımlanan özel bir dildi. Bu dil kullanıcılar ve programlama uzmanlarının enformasyon taleplerine karşılık vermek için veri tabanından veri çıkarmalarına izin veren komutları kapsamaktaydı. Böylece RDBMS; şirketlerinin verilerini daha iyi organize etmek isteyen, envanter gibi bilgileri incelemek ve karar verme amacıyla müşteri siparişi bilgilerini karşılaştırmak isteyen iş yöneticilerine yardımcı oldu. Ancak, bu patlayan taleplerden dolayı ortaya bir sorun çıktı: Bu artan veri hacminin depolanması pahalıydı ve buna erişim yavaştı. Verilerin kullanılabilirliğini artırmak için RDBMS den bağımsız modelleme olan Varlık-İlişki (ER) modeli ortaya çıktı. Bu model geliştiriciler, karmaşık programlama olmaksızın modelin en temel öğesi olan varlığın ve aralarındaki ilişkilerin, bağlantıların tasvir edildiği bir modelleme tekniğidir. Bu o zamanlar büyük bir ilerlemeydi ve geliştiricilerin teknolojinin sınırlarını zorlamasıyla, bu günde halen canlı olan İlişkisel veri tabanları pazarı meydana geldi. Organizasyonların yönetmesi gereken veri hacmi kontrol dışı olduğunda, veri ambarı buna bir çözüm sağlamıştı (Hurwitz ve ark., 2013). Böylece 1980'lerin sonunda, karar desteği ve veri ambarı konseptiyle tanışıldı. Eğilimler oluşturabilen, tarihsel analizler gerçekleştirebilen ve tahmine dayalı analitik ve yüksek ölçeklenebilir metrikler sunabilen bu dalga, bir dizi çözüm, şirket ve kendi başına bir endüstri yarattı (Krishnan, 2013).

Veri ambarları 1990'larda ticarileşti ve entegre sistemler ile veri araçları gibi yeni teknolojileri destekleyecek şekilde gelişti. Bazen bu veri ambarları öyle karmaşık ve büyüktü ki; işin gerektirdiği hız ve çevikliği sunmuyorlardı. Şirketler

için şöyle bir problem ortaya çıktı, geleneksel veri yönetimi yaklaşımlarını, yapılandırılmamış veri öğelerinin genişleyen hacmini ele alacak şekilde nasıl geliştirecekleri bilgisinde yetersizlerdi. Şirketler yapılandırılmamış verileri depolamaya başladıkça, üreticiler BLOB'lar (İkili Büyük Nesneler) gibi özellikler eklemeye başladılar. İkili Büyük Nesneler, bir veri tabanı yönetim sisteminde tek bir varlık olarak saklanan ikili veri topluluğudur. BLOB'lar genellikle görüntüler, ses veya diğer multimedya nesneleridir. Özünde, yapılandırılmamış bir veri elemanı, bir bitişik veri parçası olarak ilişkisel bir veri tabanında depolanırdı. Nesne veri tabanı yönetim sistemine (ODBMS) geçince; nesne veri tabanı, BLOB'u adreslenebilir bir parça olarak sakladı, böylece orada ne olduğu görüldü (Hurwitz ve ark., 2013).

ODBMS de kullanılan dil SQL'dir. Bilinen ODBMS Yazılımları; MS SQL Server, Oracle, Sybase, Informix, MySQL, Postrage SQL, MS Access, Advantage, DB/2'dir (Ayvaz, 2018).

1995 yılında, ticari bir internet yaratma etkisi ile "dot-com" dünyasının ortaya çıkışını görüldü ve tüketici dünyasında bireyler arası iletişim kurabilmenin ilk tadı alındı. Bu yeteneğin ortaya çıkmasıyla birlikte, hacim ve veri çeşitliliğinde de önemli bir artış görüldü. Takip eden beş ila yedi yıl içinde, bir kuruluş için iş ortamını hızla değiştiren web ticaret veya e-ticaret tarafından yönlendirilen bir dizi gelişme görüldü. Yeni modeller, tüketiciden doğrudan satışa (web sitesi), tüketiciden tüketiciye pazarlamaya (eBay ve Amazon) ve işten-tüketiciye satışa (Amazon) kadar hızla benimsenen standartlar haline geldi. Tüm bu etkinlikler, veri hacmini her zamankinden daha fazla artırdı. Hacmin yanı sıra, tüketici değerlendirmesi, deneyim üzerine geri bildirim, anketler ve ağızdan ağıza pazarlama gibi ek verilerde ortaya çıkmaya başladı. Bu yeni ve ek veriler, veri işleme ve entegrasyonda karmaşıklıklar meydana getirdi. 1997 ve 2002 yılları arasında, cep telefonları her yerde kullanılmaya başlandı ve insanlar arasında duyguları, görüşleri ve eğilimleri paylaşmak için ses ve metin kullanımı canlı bir trend oldu. Bu sayede iletişim yeteneği ve teknolojiye, yeniliklere son derece

önemli bir katkı sağlayan ürün ve hizmetlere yönelik yakınlık artmış, bu da iş hacmi ve veri yönetiminde veri hacmi, hızı ve çeşitliliği açısından aksamalara yol açmıştır. 2000 yılından sonra, veri tarihinde, arama motorlarının (Google, Yahoo) ortaya çıkışı, müziğin kişiselleştirilmesi (iPod), tablet bilgisayar (iPad), daha büyük mobil çözümler (akıllı telefonlar, 3G ağları, mobil geniş bant, Wi-Fi) ve sosyal medyanın ortaya çıkışı (Facebook, MySpace, Twitter ve Blogger tarafından yönetilen) belirleyici bir an olmuştur. Google, Yahoo, Facebook ve diğer pek çok şirket, veri yönetimi için teknolojik çözümlere yatırım yapmış ve güçlü bir karar destek platformu oluşturmak için farklı derecelerde karmaşıklığa sahip birçok formatta çok sayıda veriyi kısa bir sürede tüketmemize izin vermiştir (Krishnan, 2013).

3.1. Veri Toplama Yöntemleri

Bilgi sunulmadan ve yorumlanmadan önce, veriler toplanır ve analiz edilir. Veri toplama, istatistiksel analizde çok önemli bir rol oynar. Birincil ve ikincil veriler olarak iki farklı yöntem kullanılmaktadır.

Birincil Verilerin Tanımı: Birincil veriler, araştırmacı tarafından ilk defa, özellikle araştırma problemini çözmek amacıyla, doğrudan çabalar ve deneyim yoluyla ortaya çıkmıştır. Ham veri olarak da bilinir. Gerçek zamanlıdır ve orijinaldir. Birincil veri toplama oldukça pahalıdır, çünkü araştırma kurum veya kuruluş tarafından gerçekleştirilir, bu da yatırım ve insan gücü gibi kaynakları gerektirir. Veri toplama, araştırmacının doğrudan kontrolü ve denetimi altındadır. Veriler; gözlemler, fiziksel testler, postalanan anketler, doldurulmuş ve sayım memurları tarafından gönderilen anketler, kişisel görüşmeler, telefon görüşmeleri vb. çeşitli yöntemlerle toplanabilir.

İkincil Verilerin Tanımı: İkincil veriler, hali hazırdaki araştırma problemi ile ilgili olmayan, bir amaç için kullanıcı dışında herhangi bir kişi tarafından toplanan ve kaydedilen ikinci el bilgileri ifade eder. Nüfus sayımları, devlet yayınları, kurumun iç kayıtları, raporlar, kitaplar, dergi makaleleri, web siteleri vb.

çeşitli kaynaklardan toplanan verilerin hali hazırda mevcut halidir. Veriler sadece analiz edilir ve yorumlanır. Geçmiş zamanlı verilerdir. İkincil veriler, kolay erişilebilir olduğu için çeşitli avantajlar sunar, araştırmacıya zaman ve maliyet tasarrufu sağlar. Bununla birlikte bazı dezavantajları vardır; veriler yararlılık, uygunluk ve doğruluk bakımından sorunlu olabilir (Ajayi, 2017).

3.2. Yapılandırılmış, Yapılandırılmamış ve Yarı Yapılandırılmış Veri

Yapılandırılmış veri belirli bir veri modelini baz alarak kolaylıkla depolanabilen, işlenebilen, iletilebilen ve veri formatı kesin ve belirli olan veri çeşididir. Yapılandırılmış veri için Microsoft SQL Server, Oracle, SAP, Access, Excel gibi geleneksel veri tabanı sistemler kullanılmaktadır. Yapılandırılmış veri, format olarak veri tabanlarına uygundur yani satır ve sütun olarak kaydedilebilmektedir. Örneğin telefon numaraları, posta kodları sayılardan oluşmuştur ve yapılandırılmış veriye örneklerdir.

Yapılandırılmamış veri ise geleneksel veri tabanlarına uygun olmayan, tanımlanmış tek bir formatı olmayan verilerdir. NoSQL veya MangoDB gibi veri tabanı sistemleri ise yapılandırılmamış veriler içindir (Bayrakçı, 2015). Yapılandırılmış veri gibi yapılandırılmamış veri de ya makine tarafından üretilir ya da insan tarafından üretilir. Makine tarafından oluşturulmuş yapılandırılmamış verilere örnek olarak; uydu görüntüleri, bilimsel veriler, fotoğraflar ve video, radar veya sonar veriler verilebilir. İnsan tarafından oluşturulmuş yapılandırılmamış verilere ait birkaç örnek ise sosyal medya verileri, mobil veri ve web sitesi içeriği gibidir. İşletmelerin kullanabileceği verilerin yüzde 20'si yapısal veriler ise, yüzde 80'i yapılandırılmamıştır (Hurwitz, 2013).

Yarı yapılandırılmış veriler yapılandırılmış ile yapılandırılmamış veri arasındadır. Yarı yapılandırılmış veriler bir yapıya sahiptir ama bir bütün olarak aynı yapıda değildir yani tabloların formal yapılarına uygun değillerdir. Örneğin, e-postalar; e-posta mesaj içeriğinin ve eklerin yapılandırılmamış verilerine eklenen göndereni, alıcısı, tarihi, zamanı ve diğer sabit alanları içerir. Fotoğraflar veya

diğer grafikler; içerik oluşturucu, tarih, konum gibi anahtar kelimelerle etiketlenebilir ve bu da grafiklerin düzenlenmesini ve konumlandırılmasını sağlar. Yarı yapılandırılmış verileri yönetmek için genellikle XML ve diğer biçimlendirme dilleri kullanılır (Bayrakçı, 2015).

3.3. Büyük Verinin 4 V'si

3.3.1. Hacim (Volume):

2000'lerde dünyada 800.000 petabayt veri depolanmıştır. Bugün ise oluşturulan verilerin tümü dahi analiz edilememektedir. Bununla birlikte oluşturulan verilerin 2020 yılında 35 zettabayta ulaşacağı öngörülmektedir (Çığşar, 2017).

Tech Amerika'nın tahminine göre dünya üzerinde 2015'de her gün 114 milyar e-posta, 24 milyar kısa mesaj gönderilmiş ve 12 milyar telefon çağrısı gerçekleşmiştir. Cisco'ya göre, 2013 yılında 10 milyar nesne Nesnelerin İnternet'ine bağlanmış, çeşitli miktarlarda veri üretilmiştir ve bu sayının 2020 yılı itibariyle 50 milyara çıkması beklenmektedir. 2012 yılında Google'da, her biri kayda alınan günlük 3 milyon arama sorgusu gerçekleşmiştir. Bu sorguların günlük hacmi 24 PB civarındadır. 2011 yılında Facebook'taki aktif kullanıcılar, ayda ortalama 9,3 milyar saatten fazla bir süreyi bu sitede online olarak geçirmişlerdir. Facebook kullanıcısı sayısı 1 milyarı aşmış ve günlük aktif kullanıcı sayısı 650 milyona yaklaşmıştır. Facebook'ta, 2015 yılında günlük ortalama 2,5 milyar içerik paylaşılmış, 2,7 milyar "beğen" düğmesi tıklanmış ve 500 Terabyte'ın üzerinde veri üretilmiştir (Bayrakçı, 2015).

Günümüzdeki bu artış içinde olan veriye ait veri hacim ölçüleri çizelge 3.1. de gösterilmiştir.

Çizelge 3.1. Veri Dönüşüm Tablosu (Bayrakçı, 2015)

Birim	Büyüklük	Anlamı
Bit(b)	0 veya 1	"İkili sayı"nın (binary digit) kısaltılması. Bilgisayarlar verileri ikili kod (0-1) temelinde saklar ve işlerler.
Byte (B)	8 bit	Bir harf veya sayıyı bilgisayar kodunda oluşturmak için yeterli enformasyon. Temel hesaplama birimi.
Kilobyte (KB)	1000 veya 2^{10} byte	Yunancadaki "bin". Bir sayfa yazı 2 KB.
Megabyte (MB)	1000 KB; 2^{20} byte	Yunancadaki "büyük". Shakespeare'in bütün eserleri 5 MB, tipik bir pop şarkı 4 MB.
Gigabyte (GB)	1000 MB; 2^{30} byte	Yunancadaki "dev". İki saatlik bir film 1-2 GB'ye sıkıştırılabilir.
Terabyte (TB)	1000 GB; 2^{40} byte	Yunancadaki "dev". Amerikan Kongre Kütüphanesi'ndeki kitaplar 15 TB.
Petabyte (PB)	1000 TB; 2^{50} byte	Google her saatte 1 PB işliyor.
Exabyte (EB)	1000 PB; 2^{60} byte	Economist dergisinin 10 milyar kopyasına eşdeğer.
Zetabyte (ZB)	1000 EB; 2^{60} byte	2010 yılında var olan tüm bilgi 1,2 ZB olarak öngörülüyor.
Yotabyte(YB)	1000 ZB; 2^{80} byte	Hayal etmesi zor.

Veri tipleri farklı boyutlarda gelir. Örneğin, bir blog metni birkaç kilobayttır, sesli çağrılar veya video dosyaları birkaç megabayttır, sensör verileri, makine kayıtları ve tıklama verileri gigabaytlarda olabilir. Günümüzde endüstriyelden kişisel cihazlara kullandığımız her makine (cihaz) çok fazla veri üretebilir. Makine tarafından üretilen verilere örnek olarak; bir otomotiv fabrikasında montaj hattında bulunan bir robot kolunun, ısıtma ve soğutmayı düzenleyen binaların üstündeki sensörlerin, otomobiller üzerindeki sensörlerin, havalimanlarındaki vücut tarayıcılarının, radyo sinyallerinin, uyduların ve mobil cihazların ürettiği veriler verilebilir.

Ayrıca bugün, tabletler, cep telefonları ve otomobil bilgisayarları; coğrafi bilgi, veri türü, erişim türü, etkinlik süresi ve çok daha fazlası dahil olmak üzere günün her saatinde cihazdan her etkinlik için kayıt oluşturabilirler. Web sayfasının kullanım istatistikleri tıklama akışı verilerinde yakalanır. Bu veri türü, bir

kullanıcının web sayfasında neler yaptığına ilişkin bilgi sağlar böylece davranış analizi, pazarlama ve genel araştırma için oldukça yararlı veriler sağlayabilir. Verilerle ilgili en büyük sorun, eylemler; web sunucuları, yönlendiriciler, proxy sunucular, reklam sunucuları vb. birden çok yerde kaydedilebildiği için hacimdir. Verilerin bir kısmı yapılandırılmış olsa da, verilerin çoğu farklı biçimlere sahiptir ve çoğu zaman ağır hacimlerde gelir (Krishnan, 2013).

Doğru teknolojiyle verinin tamamı olmasa da yapılan işlerde, müşterilerin veya piyasanın daha iyi anlaşılması konusunda yararlı olabilecek kadar veri tanımlanıp analiz edilebilir (Çığşar, 2017).

3.3.2. Hız (Velocity)

Hız; verinin üretilmesini, üretilen verinin yayılımını ve saklanmasını ve anında analiz edilip değerlendirilmesini kapsar (Krishnan, 2013).

Büyük veri, sürekli devam eden bir yapıda gerçek zamanlı ya da gerçeğe çok yakın süreyle üretilir. Hız, süreç içinde her zaman açık olan sistemlerin çalışması ve bu sistemlerde tekrar edilen gözlemlerin zaman veya mekân içerisinde sürekli olarak tekrarlı şekilde devam etmesinden kaynaklanır (Bayrakçı, 2015).

Bugün herhangi bir web sitesinde kullanıcı tarafından üretilen verinin hızı, Büyük veri hızı için en iyi örnektir. Facebook, Yahoo ve Google gibi web sitelerinden her saniyede kullanıcılardan toplanan milyonlarca tıklama vardır ve bunlar büyük hacimli verilerdir. Bu veriler, günün saatine, coğrafyaya, reklam etkinliğine, tıklama davranışına ve gezinme kayıtlarına göre nüfus davranışlarını incelemek için işlenebilir, bölümlere ayrılabilir ve modellenebilir.

Veri hızının bir diğer en önemli örneği, otomobildeki sensörle ilişkili coğrafi ve konum tabanlı istihbarat temeline dayanarak, General Motors tarafından sunulan GPS, lastik-basınç sistemleri, On-Star araç ve yolcu destek hizmetleri gibi çeşitli sensörlerden gelmektedir. Sensör ağlarından üretilen veriler saniyede birkaç gigabayttan saniyede terabayta kadar değişebilir (Krishnan, 2013).

Verilerin anlık olarak izlenmesi, ölçümlenmesi ve analiz edilmesi pazarlama sektörü açısından da hayati önem taşımaktadır. Çünkü rekabetçi ortamda kurumlar veriyi ne kadar hızlı analiz edebilirse, o kadar hızlı pazarlama stratejileri geliştirerek hedef kitlelerini belirleyebilirler. Böylece kişilerin ihtiyaç duydukları bir ürünü/hizmeti anlık olarak belirleyerek kişilere ilgili ürün/hizmetleri gerçek zamanlı olarak sunabilir ve böylelikle pazar paylarını genişletebilirler.

3.3.3. Çeşitlilik (Variety)

Veri çeşitliliği, büyük verinin yapısındaki farklılık ve zenginliğin ölçüsüdür. Veri yapılandırılmış, yapılandırılmamış ve yarı yapılandırılmış, sayı, metin, resim, video, ses ve diğer farklı formatlarda bulunabilir. Örneğin Facebook gönderilerinde fotoğraf, video, diğer web sitelerinden link ve yorumları içerebilen farklı formatlar aynı anda bulunabilmektedir.

Büyük verinin bu çeşitli formatları aynı anda bulundurması analizinde zorluklara sebep olmaktadır. Bu yüzden temelinde mevcut istatistiksel tekniklerin yer aldığı yeni veri analiz teknikleri geliştirilmiştir. NoSQL yapısı kullanılarak dizayn edilen veri tabanlarındaki yapısal olmayan veriler; veri madenciliği yöntemleri, Hadoop ve MapReduce gibi yeni tekniklerle yönetilebilir, işlenebilir ve analiz edilebilir olmuştur.

3.3.4. Doğruluk (Veracity)

Büyük verinin doğruluğu iki anlamı ifade etmektedir. Birincisi, büyük veriyi oluşturan kaynağın güvenilirliğinin yüksek olmasını, ikincisi de verinin hedef kitleye uygun olmasını ifade etmektedir. Doğruluk günümüzde büyük veri açısından en zor temin edilen durumdur. Çünkü büyük verinin çoğunluğu farklı kaynaklardan gelmektedir ve bu kaynakları veriyi toparlayan kurumlar kontrol edememektedir.

Büyük verinin doğruluğunu ve kalitesini etkileyen en önemli unsurlar veri hacmi ve veri miktarıdır. Çünkü veri miktarı arttıkça ve veri kaynakları çeşitlendikçe verini güvenilirliği ve kalitesi azalmaktadır (Bayrakçı, 2015).

3.4. Büyük Veri Kullanım Alanları

Büyük verinin neredeyse dokunmadığı hiçbir alan yoktur denilebilir. İş dünyasından sanata, bilime ve devletlere kadar uzanan bu geniş yelpaze nedeniyle büyük verinin analizine duyulan ilgi giderek artmaktadır (Çığşar, 2017).
Büyük verilerin kullanım alanlarına örnekler, aşağıda listelenmiştir.

- **Devlet Operasyonları:** Ulusal arşiv ve kayıt idaresi, sayım bürosu
- **Ticari:** Bulut tabanlı finans, bulut tabanlı yedekleme, Mendeley, Netflix, internet araştırması, dijital malzemeler, kargo taşımacılığı (UPS'de olduğu gibi)
- **Savunma:** Sensörler, görüntü gözetleme, durum değerlendirmesi
- **Sağlık ve Yaşam Bilimleri:** Tıbbi kayıtlar, grafik ve olasılık analizi, patoloji, biyogörüntüleme, genomik, epidemiyoloji, insan aktivite modelleri, biyoçeşitlilik
- **Derin Öğrenme ve Sosyal Medya:** Kendi kendini süren araçlar, coğrafi konum görüntüler, twitter, kalabalık kaynak kullanımı, ağ bilimi, NIST kıyaslamalı veri kümeleri
- **Araştırma için Ekosistem:** Meta veri, iş birliği, dil çevirisi
- **Astronomi ve Fizik:** Gökyüzü araştırmaları (benzetme ve karşılaştırmalar ile), CERN'de LHC, Japonya'da Belle Accelerator II
- **Toprak, Çevre ve Kutup Bilimi:** Deprem, okyanus, toprak gözlemi, atmosferde radar dağılımı, buz tabakası radar dağılımı, dünya radar haritalama, iklim simülasyonu veri kümeleri, atmosferik türbülans tanımlama, yeraltı biyojeokimyası, AmeriFlux ve FLUXNET gaz sensörleri

- **Enerji:** Akıllı şebeke (NIST Big Data Public Working Group, 2015).

3.5. Veri Hacminin Büyümesi

Sayısı 80 milyonu aşan Türkiye'nin yüzde 58'i internete bağlıdır. Bu nüfusun tam yüzde 90'ının mobil aboneliği bulunmaktadır, ayrıca nüfusumuzun yüzde 45'i mobil üzerinden (akıllı telefon ya da tabletler aracılığıyla) sosyal medya kullanmaktadır. Türkiye'de mobil telefon kullanan yetişkinlerin yüzde 56'sında akıllı telefon vardır. Barcelona'da düzenlenen Mobil Dünya Kongresi'nde (Mobile World Congress-MWC) Mobil Ekonomi 2017 başlıklı rapora göre, 2016 sonu itibariyle dünyada mobil abone sayısı 4,8 milyar oldu. 2020 sonundaki beklentisi ise 5,7 milyar aboneye ulaşılacağı yönündedir. Bu veriye göreyse 2020'de dünyanın üçte ikisinin mobil kullanıcı olması bekleniyor (Aydın, 2017).

Veriler hızlı bir şekilde büyümektedir. Dünyanın veri hacminin yılda %40, 2020 yılında 50 kat artacağı tahmin edilmektedir. Hızlı büyümeden dolayı, 2020 yılında veri üretimi 2009 yılında olduğundan 44 kat daha büyük olacaktır. Verilerdeki günlük artış, ilgili büyüme oranlarını öngörmemizi sağlıyor. Küresel olarak, 2016 yılında yaklaşık 4.4 ZB elektronik veri üretilmiştir. IDC (International Data Corporation) kurumsal verilerin 2020'ye kadar 40 ZB'ye ulaşacağını iddia ediyor. IDC'nin tahminine göre, işten-tüketiciye ve işten-işe internet dünyasında işlemler, 2020 yılına kadar günde 450 milyar dolara ulaşacak. Bu rakam, verilerin Zettabaytlarda ne kadar hızlı arttığını göstermektedir (Yaqoob ve ark., 2016).

Kullanıcıların %50'si, araştırmalarına mobil cihazlarındaki arama motorlarıyla başlıyor. Ayrıca saniyede 40.000 arama sorgusu gerçekleştiriliyor (yalnızca Google'da), bu da günde 3,46 milyon arama ve her yıl 1,2 trilyon aramaya tekabül eder.

Facebook kullanıcıları her dakika yaklaşık 31,25 milyon mesaj göndermekte ve 2,77 milyon video izlemektedir (Anonymous, 2018).

Dünya genelinde yaratılan e-posta hesaplarının sayısı 2012 yılında 3,3 milyardan 2016 yılının sonuna kadar 4,3 milyara yükselmiştir. 2012 yılında, gönderilen ve alınan e-posta sayısı günde 89 milyardı. Bu miktar 2016 yılı sonunda 143 milyarın üzerine çıkmıştır. 2016 yılında, her gün dünya genelinde gönderilen e-posta sayısı 170 milyar civarındadır. Her saniye gönderilen e-posta sayısı 2 milyondur. 2013 – 2016 yılları arasında mobil cihazlarda e-posta açma oranı %180 yükselmiş durumdadır (Yaqoob ve ark., 2016).

4. VERİ TABANI

Temel olarak, veri tabanı sistemi bilgisayar tabanlı bir kayıt tutma sistemidir. Yani genel amacı bilgi / veriyi kaydetmek ve muhafaza etmek olan bir sistemdir.

Veri tabanı teknolojisi, çeşitli durumların bir araya gelmesi sonucu 60'lı yılların sonlarında ortaya çıkmıştır. Çeşitli türlerdeki verileri işlemek için ortaya çıkan teknoloji, 'Veritabanı Yönetim Teknolojisi' olarak adlandırılır ve sonuçta ortaya çıkan yazılımlar, 'Veritabanı Yönetim Sistemi' (DBMS) olarak bilinir. Bir veritabanı yönetim sistemi (DBMS), bir veri tabanında bulunan verilere erişime izni veren bir yazılım sistemidir. DBMS'nin amacı, veri tabanında yer alan bilgilerin tanımlanması, saklanması ve geri alınması için uygun ve etkili bir yöntem sağlamaktır. DBMS, uygulama programlarıyla arabirim kurar, böylece verim tabanında yer alan veriler birden çok uygulama ve kullanıcı tarafından kullanılabilir. Buna ek olarak, DBMS veri tabanının merkezi kontrolünü sağlar, hileli veya yetkisiz kullanıcıların verilere erişmesini engeller ve verilerin gizliliğini sağlar.

'Data' yani ‘Veri’ terimi, bir varlığın niteliğinin değeri olarak tanımlanabilir. Aynı özelliklere sahip varlıkların ilgili veri maddelerinin herhangi bir koleksiyonu bir 'Veri Tabanı' olarak adlandırılabilir. Veriler bilgisayarda; satır, sütun ve tablo halinde düzenlenmiştir ve ilgili bilgileri bulmayı kolaylaştırmak için dizine eklenmiştir. Veri yeni bilgiler eklendikçe güncellenir, genişletilir ve silinir.

İnsanlar günlük yaşamlarında çeşitli veri tabanlarını kullanırlar. Sözlük, telefon rehberi, kütüphane kataloğu, vb. girişlerin alfabetik veya sınıflandırılmış sıraya göre düzenlendiği veri tabanları için örneklerdir. Bilgisayar veri tabanları genellikle satış işlemleri, ürün katalogları ve envanterler ve müşteri profilleri gibi veri kayıtlarının veya dosyalarının bir araya getirilmesini içerir (Gunjal, 2003).

Tüm veri tabanı işlemlerinin; bölünmezlik (atomicity), tutarlılık (consistency), yalıtım (isolation) ve dayanıklılık (durability) anlamına gelen kısaltma ACID'yi kullanan dört özelliği şu şekildedir.

Bölünmezlik: Bir işlemin ya tamamen ya da hiç gerçekleştirilmeyen bölünmez bir birim olmasını sağlayan "hepsi ya da hiçbir şey" özelliğidir.

Tutarlılık: Bir işlem, yeni ve geçerli bir veri durumu oluşturur veya her hangi bir hata oluşursa, işlem başlatılmadan önce tüm verileri kendi durumuna döndürür.

İzolasyon: Süreçte veya henüz taahhüt edilmemiş bir işlem, diğer herhangi bir işlemden yalıtılmış kalmalıdır.

Dayanıklılık: Taahhüt edilen veriler sistem tarafından kaydedilir, böylece bir arıza durumunda ve sistem yeniden başlatıldığında bile veriler kullanılabilir (Rosiji, 2015).

5. VERİ MADENCİLİĞİ

Veri madenciliği ifadesi; büyük miktarda veriden bilgiyi çıkarmaya veya "madenciliğe" atıfta bulunur. Ayrıca bilgi çıkarma, veri analizi, veri arkeolojisi ve veri taramasına benzer.

Birçok kişi veri madenciliğini, popüler olarak kullanılan başka bir terimle, Veri'den Bilgi Bulma veya KDD (Knowledge Discovery in Database) ile eşanlamlı olarak ele alır. Alternatif olarak, bazıları veri madenciliğini sadece bilgi keşif sürecinde önemli bir adım olarak görürler. Bununla birlikte, sektörde, medyada ve veri tabanı araştırma ortamlarında, veri madenciliği terimi, veriden bilgi keşfi teriminden daha popüler hale gelmektedir. Bu nedenle, bu tezde, veri madenciliği terimini kullanmayı tercih ediyoruz (Han ve ark., 2012).

Veri madenciliği, gizli veri tahminlerinin büyük veri tabanlarından çıkarılması, veri ambarındaki önemli bilgileri analiz etme potansiyeline sahip güçlü ve yeni bir teknolojidir. Veri madenciliği, gizli kalıplar için veri tabanlarını araştırmakta, uzmanların kaçırabileceği tahmin edici bilgileri bulmakta, beklentilerinin ötesine geçmektedir (Rai Technology University, 2018).

Veri madenciliği, veri tabanı ve veri ambarı teknolojisi, istatistik, makine öğrenimi, yüksek performanslı bilgi işlem, örüntü tanıma, yapay sinir ağları, veri görselleştirme, bilgi alma, görüntü ve sinyal işleme ve mekânsal veya zamansal gibi çoklu disiplinlerden tekniklerin entegrasyonunu içerir (Han ve ark., 2012).

Bazı küçük işletmeler müşterileriyle ihtiyaçlarını birebir görüşerek, tercihlerini hatırlayarak ve geçmiş etkileşimlerinden gelecekte onlara nasıl daha iyi hizmet edeceğini öğrenerek birebir ilişkiler kurmaktadır. Büyük işletmelerde ise çoğu müşteri şirket çalışanları ile kişisel olarak etkileşimde bulunamadığından bunu gerçekleştiremez. Aslında, birçok şirket müşterileri hakkında ve onlar hakkında hiçbir şey öğrenmeden yüzlerce gigabayt veya terabayt veri toplar. Bu bilgileri öğrenmek için, birçok kaynaktan gelen veriler önce toplanmalı sonra tutarlı ve yararlı bir şekilde organize edilmelidir. Buna veri ambarı denir. Veri

ambarı, işletmeye bir bellek sağlar. Ancak, hafıza zekâsı olmadan çok az kullanışlıdır. Veri madenciliğinin devreye girdiği yer burasıdır. Veri madenciliği, geçmişi anlamaya, kuralları tasarlamaya, yeni fikirlerle gelmeye ve geleceğe dair tahminlerde bulunmayı sağlar. Veriler analiz edilmeli, anlaşılmalı ve eyleme dönüştürülebilir bilgilere çevrilmelidir. Veri madenciliği, veri ambarına zekâ katan araçlar ve teknikler sağlar.

Bir işletme; müşterileri daha iyi tanımak için, veri ambarına zekâ katan birçok veri madenciliği aracı ve tekniğini kullanarak, potansiyel müşterileriyle olan etkileşimlerinin oluşturduğu engin veri dağlarından yararlanabilecektir (Kuonen, 2004).

5.1. Veri Madenciliği Aşamaları

Gösterilen aşamalar ham verilerle başlar ve aşağıdaki aşamaların sonucu olarak elde edilen bilgiyle bitirilir:

1. Problemin Belirlenmesi: Veri madenciliği projesindeki en zor kısım olan bu adımda, ne tür bir analiz yapılması gerektiğinin kesinlikle iyi bir şekilde anlaşılması gerekir. Aksi takdirde tüm proje yanlışlıklar üzerine kurulmuş olacak ve bulunan sonuçlar amacı temsil etmeyecektir. Bu yüzden ilk ve en önemli adım, amacı açıkça belirlemek ve amaca giden süreci geliştirmektir. Amacı tanımlarken, neyi ölçmeye veya öngörmeye çalıştığımıza karar vermek gerekir (Dolgun ve Ersel, 2014).
2. Veri Bulma: Bu aşamada, analist ve son kullanıcı, sorularını cevaplamak için hangi verileri analiz edeceğini belirledikten sonra ihtiyaç duydukları şeylerin var olup olmadığını görmek için elindeki verileri araştırırlar.
3. Veri Seçimi ve Temizleme: Veri seçildikten sonra, temizlenmesi gerekir. Eksik değerler, eksik kayıtların belirlenmesi, el ile doldurulması, her bir eksik değer için bir sabit girilmesi veya bir değer tahmini yapılması gibi tutarlı bir şekilde ele alınmalıdır. Bazı veri kayıtları tam olabilir ama yanlış

(gürültülü) olabilir (Ayre, 2006). Gürültülü veri, veri girişi veya veri toplanması sırasında oluşan sistem dışı hatalardır (Çığşar, 2017). Veri temizleme işlemi, gürültülü verileri düzelterek, verileri eksik değerleri doldurarak, aykırı değerleri belirleyerek veya kaldırarak ve tutarsızlıkları çözerek "temizlemek" için yapılır (Han ve ark., 2012).

4. Veri entegrasyonu: Bu aşamada, çoğu zaman heterojen olan çoklu veri kaynakları ortak bir kaynakta birleştirilebilir.
5. Veri Dönüşümü: Veriler madencilik için uygun bir forma dönüştürülür (Zaiane, 1999).
6. Veri Azaltma: Analiz sürecini yönetilebilir ve düşük maliyetli hale getirmek için verilerin azaltılması gerekecektir. Boyut küçültme, veri sıkıştırması, gerçek veriler yerine model veya örnekler kullanılması, ayrıklaştırma ve kavram hiyerarşisi oluşturma veri azaltma aşamasıdır.
7. Veri Madenciliği: Bu aşamada, veriler için ilginç ilişkiler veya kullanışlı modeller bulmak amacıyla, doğru veri madenciliği tekniğinin seçimi önemlidir. Çoğu zaman, sınıflandırma ve kümeleme algoritmaları ilk önce kullanılır. İlişkilendirme kuralları da uygulanabilir. Bazı kurallar, diğerlerinden daha ilginç desenler verir. İlginçlik genellikle kaliteyi, yeniliği, yararlılığı ve sadeliği birleştiren genel bir model değeri ölçüsü olarak alınmaktadır (Ayre, 2006).
8. Model oluşturma: Veri madenciliği tekniğini belirledikten sonra doğru algoritmanın seçilmesi nispeten daha kolaydır. Her bir veri madenciliği tekniği için geçerli birkaç algoritma mevcuttur. Çoğu zaman örneklem ile denemeler yapmadan en uygun algoritmanın ortaya çıkarılması olası değildir. Algoritmanın doğruluğu verinin yapısı ile ilişkilidir. Doğru yaklaşım değişik algoritmalar kullanarak birkaç model kurmak ve modellerin doğruluğunu karşılaştırmak olabilir (Arabacı, 2007).
9. Sonuç Yorumlama ve Görselleştirme: Çeşitli istatistiksel tekniklere göre en iyi sonuç veren algoritmanın oluşturduğu model seçilerek oluşturulan

modelle yeni durumlar için tahminlerde bulunulur (Çığşar, 2017). Bilgisayar grafikleri ve grafik tasarım araçları, oluşturulan çıktıyı sunmak ve görselleştirmek için kullanılır (Ayre, 2006).

10. Sonuçların uygulamaya geçirilmesi: Bu kullanıcıya keşfedilmiş bilginin sunulduğu ve raporlandığı son aşamadır. Oluşturulan raporlarla gelecekte oluşabilecek benzer problemler için olası çözümler üretilir ve mevcut probleme çözüm önerileri sunulur (Çığşar, 2017). Orijinal sorunun çözülmesine ek olarak, yeni bilgi yeni modellere dahil edilebilir ve tüm bilgi veya veri madenciliği döngüsü tekrar başlayabilir (Ayre, 2006).

5.2. Veri Madenciliğinin Uygulanabildiği Veriler

Veri madenciliğinde, veriler bir hedef uygulama için anlamlı olduğu sürece her türlü veriye uygulanabilir. Veri madenciliği veri ambarları, ilişkisel veri tabanları, işlemsel veri tabanları ve düz dosya gibi veri tabanlarına uygulanabilmektedir. Bunların yanı sıra, çok yönlü formlara ve yapılara ve daha farklı anlamlara sahip başka veri türleri vardır. Veri madenciliği, ortaya çıkabilecek yeni veri türlerini benimsemeye devam edecektir (Han ve ark., 2012).

5.2.1. Veri Ambarı (Data Warehouse)

Tüm işletme bünyesinde güncel işlemler, indeksler ve değişimler hakkında güvenilir ve kesin enformasyonlar gerektiğinde bu enformasyonlara ulaşım kolay olmayacaktır. Çünkü her bir enformasyon farklı veri tabanında bulunmaktadır. Bir veri ambarı işletmenin tümünü ilgilendiren, karar vericilerin ilgi alanlarına yönelik güncel ve geçmiş verileri saklayan büyük veri tabanlarıdır. Veriler, satış ve pazarlama, müşteri hesapları, üretim vb. temel sistemlerden çıkmaktadır ve web sitesi işlemlerini de içerebilir. Veri ambarları, yönetimin analiz yapabilmesi ve karar verebilmesi için enformasyonu farklı veri kaynaklarından çekerek birleştirir ve standartlaştırır.

Veri ambarı sistemi ayrıca bir dizi anlık ve standart sorgu araçları, çözümleyici /analiz araçları ve grafik rapor etme olanakları sağlamaktadır. Pek çok işletme, veri ambarını kullanmak için altyapı olarak intranet kullanmaktadır (Ayvaz, 2018).

Veri ambarı teknolojisi, bilgi işçisini (yürütme, yönetici, analist) karar vermesi için bilgi malzemesi ile destekleyen bir dizi yeni kavram ve araçtan oluşmaktadır. Bir veri ambarı oluşturmanın temel nedeni, organizasyondaki bilginin kalitesini arttırmaktır. En önemli konu, ne zaman olursa olsun şirket genelindeki veri görünümüne erişim sağlanmasıdır. Geleneksel yapısal verilerden, metin veya multimedya gibi yapısal olmayan verilere kadar çeşitli şekillerde bulunan iç ve dış kaynaklardan gelen veriler temizlenir ve tek bir depoya entegre edilir. Bir veri ambarı (DWH), son kullanıcılara bir iş bağlamında anlayabilecekleri ve kullanabilecekleri şekilde sunulan bu verilerin tutarlı bir deposudur (Grivas, 1999).

Veri ambarları operasyonel ve enformasyonel olarak ikiye ayrılır. Operasyonel veri ambarının amacı işletmenin süreçlerinin devamıdır. Veri tipi iş süreçlerindeki durumlardır. Ana kullanıcılar operasyonel kişiler, satış, idari personeldir. Kullanım genişliği dardır, planlıdır, düşük güncelleme gerektirir. Tasarım amacı performans, veriye erişim imkanıdır. Çok sayıda ama kısıtlı veri kaynaklarına erişim sağlar. Operasyonel veri ambarlarına örnek olarak; herhangi bir işletmenin omurgası olan "sipariş girişi", "envanter", "imalat", "bordro" ve "muhasebe" sistemleri verilebilir.

Enformasyonel veri ambarlarının amacı yönetime karar vermede destektir. Veri tipi geçmişe ait rapor ve gelecek tahminleridir. Ana kullanıcılar yöneticiler, iş analistleri, müşterilerdir. Kullanım genişliği komplekstir, hesaplama ve analiz gerektirir. Tasarım amacı basit ve esnek erişim imkanıdır. Hemen her veri kaynağına, seyrek erişim sağlar. "Pazarlama planlaması", "mühendislik planlaması" ve "finansal analiz" gibi işlevler bunları desteklemesi için enformasyonel veri ambarları gerektirir (Şeker, 2015).

Veri ambarı, yönetimin karar verme sürecine destek veren, konuya odaklı, entegre edilmiş, zamana bağlı ve kalıcı bir veri topluluğudur. Bu önemli özellikler şöyledir.

- **Konuya odaklı**: Bir veri ambarı; müşteri, tedarikçi, ürün ve satış gibi ana konular etrafında düzenlenir. Bir veri yazılımı; bir kuruluşun günlük operasyonlarına ve işlemlerine odaklanmak yerine, karar vericiler için verilerin modellenmesi ve analizine odaklanır.
- **Entegre:** Bir veri ambarı genellikle ilişkisel veri tabanları, düz dosyalar ve çevrim içi işlem kayıtları gibi çoklu heterojen kaynakları entegre ederek oluşturulur. Veri kodlaması çoğu zaman tutarsızdır. Örneğin, bir uygulamada cinsiyet, 0 ve 1 iken bir diğerinde "m" ve "f" olarak kodlanabilir. Veriler operasyonel ortamdan veri ambarına taşınırken, tutarlı bir kodlama gerçekleşir. Örneğin; cinsiyet verilerinin tamamı "m" ve "f" ye dönüştürülür.
- **Zamana Bağlı**: Veriler, tarihsel bir perspektiften (ör., Geçmiş 5-10 yıl) bilgi sağlamak için saklanır. Veri ambarındaki her anahtar yapı, örtülü veya açık bir şekilde bir zaman unsuru içerir.
- **Kalıcı:** Veriler, veri ambarına girdikten sonra hiçbir şekilde güncellenmez veya değiştirilmez, ancak yalnızca yüklenir ve erişilir (Han ve ark., 2012).

5.2.2. İlişkisel Veri Tabanı (Relational Databases)

İlişkisel veri tabanı, birbirinden farkı tablolara yerleştirilmiş olan verilerin birbirleri ile belirli alanlara göre ilişkilendirilerek düzenlendiği veri tabanlarıdır. İlişkisel veri tabanlarında veriler tablolarda birbirleri ile ilişkili bir şekilde saklanmaktadır. İlişkisel veri tabanları birden fazla tablodan oluşabilir. Tablolar ise satır ve sütunlardan oluşur. Tablolarda satır ve sütunların kesiştiği noktaya ise "hücre" adı verilir. Tablo içerisindeki satırlar tablonun kayıtlarını oluşturur.

Anahtar alan ise tablonun tanımlayıcısı niteliğindedir. Her sütunun alacağı isim birbirinden farklı olmalıdır. Her satır birbirinden farklı olmalıdır yani birbiri ile tamamen ayni olan iki kayıt kullanılmamalıdır (T.C. Milli Eğitim Bakanlığı, 2012).

İlişkisel veri tabanı için en çok kullanılan sorgu dili SQL'dir (Structured Query Language). SQL tablolarda toplanan bilgilerin işlenmesine, ilişki kurulmasına olanak tanımaktadır. İlişkisel veri tabanları en yaygın ve en zengin bilgi depolarından biridir ve bu nedenle veri madenciliği çalışmalarında önemli bir veri formudur. SQL in veri madenciliğinden ayrılan yönü, veri madenciliği tahmin, karşılaştırma yapabilir ve saptamaları tespit edebilirken SQL sadece veri seçimi ve dönüşü görevlerini yerine getirebilmektedir (Çığşar, 2017).

5.2.3. İşlemsel Veri Tabanı (Transactional Databases)

İşlemsel veri tabanı modelleri günümüzdeki birçok büyük kuruluşun ilk seçimidir. Özellikle analitik için oluşturulmamışlardır, ancak genellikle üretim veri tabanları olarak zaten var olduklarından, fiili analizle ilgili ortamlar haline gelmişlerdir (Anonymous, 2018b). Bir işlemsel veri tabanı, her biri bir zaman niteliği, bir tanımlayıcı ve bir dizi öğe içeren işlemleri temsil eden bir dizi kayıttır (Zaiane, 1999). Bu tür veri tabanlarında bulunan veriler dinamik bir yapıya sahiptirler (sürekli değişmektedirler). Veri bütünlüğünü korurken, tek tek veri satırlarını okuma ve yazma konusunda çok hızlıdırlar (Anonymous, 2018b).

İşlemsel veri tabanı iç içe geçmiş tablolara izin vermediğinden, işlemler genellikle düz dosyalarda saklanır veya biri işlemler için ve diğeri işlem öğeleri için olmak üzere iki normalleştirilmiş işlem tablosunda saklanır. Bu tür veriler üzerinde tipik bir veri madenciliği analizi, birlikte veya sırayla meydana gelen eşyalar arasındaki ilişkilerin incelendiği piyasa sepeti analizi veya birliktelik kuralıdır.

5.2.4. Düz Dosya (Flat File)

Veri madenciliğinde en yaygın olarak kullanılan veri kaynaklarındandır. İkili format veya metin formundaki basit veri dosyalarıdır. Bu veriler bilimsel ölçümler, zaman serileri gibi farklı türlerde olabilmektedir.

5.2.5. Diğer Veri Çeşitleri

Diğer veri türleri birçok uygulamada görülebilir:

- Zamana bağlı veya sıralı veriler (örneğin, tarihsel kayıtlar, borsa verileri, zaman serileri ve biyolojik sıralı verileri).
- Veri akışları (örneğin, sürekli olarak iletilen video izleme ve sensör verileri)
- Mekânsal veriler (örn. haritalar).
- Mühendislik tasarım verileri (örn. binaların tasarımı, sistem bileşenleri veya entegre devreler).
- Hipermetin ve multimedya verileri (metin, resim, video ve ses verileri dahil).
- Grafik ve Ağa bağlı veriler (örneğin, sosyal ve bilgi ağları) Web (İnternet tarafından sağlanan geniş, yaygın olarak dağıtılmış bilgi deposu).

Bu uygulamalar, zengin yapı ve semantik taşıyan örnekleri nasıl işleyeceğimiz ayrıca özel yapılar (örneğin diziler, ağaçlar, grafikler ve ağlar) ve özel semantikler (sipariş, görüntü, ses ve video içeriği ve bağlantı) taşıyan verileri nasıl kullanacağımız gibi yeni zorluklar getirmektedir (Zaiane, 1999).

5.3. Makine Öğrenme

Makine öğrenmesinde, veriler vazgeçilmez bir rol oynar ve öğrenme algoritması, veriden bilgi veya özellikleri keşfetmek ve öğrenmek için kullanılır.

Veri kümesinin kalitesi veya miktarı, öğrenme ve tahmin performansını etkilemektedir.

Gerçek uygulamalarda gözlemlediğimiz şey, bellek eksikliğinden veya diğer kaçınılmaz nedenlerden dolayı evrensel veri kümesinin yalnızca bir alt kümesidir. Bu edinilen veri kümesine eğitim seti (eğitim verileri) denir ve evrensel veri kümesinin özelliklerini öğrenmek için kullanılır. Genel olarak, eğitim setindeki vektörler, bağımsız olarak evrensel veri kümesinden örneklenir. Makine öğrenmesinde, arzu ettiğimiz şey, bu öğrenilen özelliklerin sadece eğitim setini açıklamakla kalmayıp, aynı zamanda, görünmeyen örnekleri veya gelecekteki olayları öngörmek için de kullanılabileceğidir. Öğrenmenin performansını incelemek için, test seti veya test verisi adı verilen başka bir veri kümesi test için ayrılabilir. Örneğin, final sınavlarından önce, öğretmen öğrencilere pratik yapmak için birkaç soru verebilir (eğitim seti) ve öğrencilerin performanslarını yargıladıkları yol, onları başka bir problem seti (test seti) ile incelemektir (Chao, 2011).

Makine öğrenme, örnek verileri veya geçmiş deneyimleri kullanarak bir performans ölçütünü optimize etmek için bilgisayarları programlamaktır (Alpaydın, 2014).

Makine öğrenimi, istatistiğin tekniklerini ve temellerini içerir. Çünkü, verilen verilerden istatistiksel özellikleri öğrenme ve çıkarsama yapılır. Bilgisayar bilimi tekniklerini ve temellerini içerir. Çünkü, optimizasyon, model gösterimi ve performans değerlendirmesi için etkili algoritmalar kullanılmaktadır.

Veri setinin önemine ek olarak, makine öğrenimi genellikle iki kritik faktör, modelleme ve optimizasyondan oluşur. Modelleme, verilen eğitim setinin ayırma sınırı veya olasılık dağılımının nasıl modelleneceği anlamına gelir ve daha sonra seçilen modelin en iyi parametrelerini bulmak için optimizasyon teknikleri kullanılır.

Makine öğrenimi, yapay sinir ağları, örüntü tanıma, bilgi alma, yapay zekâ, veri madenciliği ve işlev yaklaşımı gibi diğer disiplinlerle de ilgilidir. Bu alanlarla

karşılaştırıldığında, makine öğrenimi, makine neden öğrenebilir ve erişilebilir eğitim verilerinin en iyi şekilde kullanılması için nasıl modelleneceğine, optimize edileceğine ve düzenli hale getirileceğine daha fazla odaklanır.

Verilen veri setine göre genellikle gözetimli öğrenme ve gözetimsiz öğrenme olarak iki tip makine öğrenimi vardır. Gözetimli öğrenme için verilen eğitim seti etiketli veri kümesidir. Gözetimli öğrenme, özellik seti ile etiket kümesi arasındaki ilişkileri bulmaya çalışır. Bu öğrenmeden çıkarılan bilgiler genellikle tahmin ve tanıma için kullanılır. Gözetimsiz öğrenmede verilen eğitim seti, etiketlenmemiş veri kümesidir. Gözetimsiz öğrenme kümelenmeyi, olasılık yoğunluk tahminini, özellikler arasındaki ilişkiyi ve boyut küçültmeyi amaçlamaktadır (Chao, 2011).

5.4. Veri Madenciliği Ve İstatistik

Hem veri madenciliği hem de istatistik verilerden öğrenme ile ilgilidir. Verileri bilgi haline getirme amacıyla verilerdeki yapıları keşfetme ve tanımlama ile ilgilidir. Her iki tekniğin amaçları örtüşüyor olsa da farklı yaklaşımları vardır. Bunlar; teorinin rolü, genellenebilirlik, hipotez testi ve güven düzeyidir. İstatistik teori ile ortak yaşama ilişkisi içindedir. İstatistiksel analizler apriori bilgiye dayalı teori ile başlar ve teorinin onaylanması veya ret edilmesi hakkında kanıt arar. İstatistik doğası gereği doğrulayıcı bir süreçtir. Öte yandan veri madenciliği teorinin doğrulanması üzerine odaklanmaz. Bununla birlikte, istatistik ile karşılaştırıldığında veri madenciliği değişkenler arasındaki ilişkiler hakkında daha az varsayımlarla sınırlandırılmıştır. Veri madenciliği tümden gelim, istatistik tüme varım ile ilgilenir (Tüzüntürk, 2010).

İstatistik; veri toplanması için planlamadan ve sonraki veri yönetiminden, veri olarak adlandırılan sayısal verilerden elde edilen sonuçların çıkarılması ve sonuçların sunulması gibi son faaliyetlere kadar her şeyi içerir (Kuonen, 2004).

Veriler, ham bilgi olarak bilinir ve kendi başına bilgi değildir. Verilerden bilgiye kadar olan süreç şu şekildedir: Her şey datanın (veri) oluşması ile başlar.

Data (veri), gelen mesajların anlamlandırılmamış halidir. Örneğin "WifiPassword" dizesini hayal edin. Bu dize sadece veridir. Bir dize olduğunu anlamak ile veri bir anlam kazanır. Ve bir üst basamağa information'a (malumat) ulaşmış olur. Bir üst basamağa baktığımızda knowledge'ı (bilgi) görürüz. Knowledge (bilgi) basamağında information bu kez artık kurumsal/kişisel deneyimlerle, önyargılarla işlenebilir halde ve farklı kaynaklarla birleştirilebilir özelliktedir. Bundan önceki örnekteki dizenin sizin wifi şifresi olduğunu bilmek knowledge (bilgi) olarak tanımlanır. Piramidin en üstünde wisdom (akıl) yer alır. Burası, iyi-kötü, doğru-yanlış, etik veya etik değil değerlendirmelerinin yapılabildiği yerdir. Bu basamağın kendini gerçekleyebilmesi için bir insana, bir ruha kesinlikle ihtiyaç vardır. Aynı örneğe göre, wifi şifresini kullanarak kablosuz ağa erişmek wisdom (akıl) olur (Liew, 2013).

Bu istatistiksel düşünme süreci, belirsizliklerin altında karar verme için istatistiksel modeller oluşturulan verilere dayanmaktadır. İstatistikler, bilgiyi sistematik bir taban üzerine yerleştirme gereğinden doğmuştur. Bu da olasılık yasaları, veri özelliklerinin ve ilişkilerinin ölçülmesi vb. üzerine bir çalışma gerektirmiştir (Kuonen, 2004).

5.5. Veri Madenciliği Sorunları

Veri madenciliği sistemleri, girişinde ham veri sağlamak için veri tabanlarına dayanır ve bu da veri tabanlarındaki sorunların; dinamik, eksik, gürültülü ve büyük olma eğiliminde olmasını sağlar. Diğer problemler, depolanan bilgilerin yeterliliği ve uygunluğunun bir sonucu olarak ortaya çıkar.

Sınırlı bilgi: Sonuçsuz veri problemlere neden olur, çünkü uygulama alanıyla ilgili bilgi için gerekli olan bazı özellikler veride mevcut değilse, belirli bir alan hakkında önemli bir bilginin keşfedilmesi mümkün olmayabilir. Örneğin, eğer hasta veri tabanı kırmızı kan hücresi sayımını içermiyorsa, hasta veri tabanından sıtma teşhisi konamaz.

Gürültü ve eksik değerler: Veri tabanlarının içerdikleri verilerin tamamen doğru olduğu kabul edilmez. Bazı örneklerin yanlış sınıflandırılması gibi hatalar meydana gelebilir. Özellik veya sınıf bilgilerinin değerlerindeki hata, gürültü olarak bilinir. Mümkün olursa, sınıflandırma bilgisinden kaynaklanan gürültü ortadan kaldırılmak istenir, çünkü bu oluşturulan kuralların genel doğruluğunu etkiler.

Eksik veriler, keşif sistemleri tarafından aşağıdaki şekillerde ele alınabilir;

- Eksik değerleri göz ardı etmek
- Uygun gelen kayıtları çıkarmak
- Eksik değerleri bilinen değerlerden çıkarmak
- Eksik verileri ek olarak özellik alanına dahil edilecek özel bir değer olarak ele almak
- Bayes teknikleri kullanılarak eksik değerler üzerinde ortalama yapmak

İstatistiksel yöntemler gürültülü veri problemlerini tedavi edebilir ve farklı gürültü türlerini ayırabilir.

Belirsizlik: Belirsizlik, hatanın ciddiyetini ve verideki gürültü derecesini ifade eder. Veri hassasiyeti, keşif sisteminde önemli bir husustur.

Boyut, Güncellemeler ve Alakasız Alanlar: Veri tabanları, bilgi eklendikçe, değiştirildiklerinde veya kaldırıldıklarında içeriklerinin sürekli değişmesi nedeniyle büyük ve dinamik olma eğilimindedir. Ayrıca, bazı veri değerlerinin zamanla değiştiği ve keşif sisteminin verilerin "güncelliğine" bağlı olduğu için öğrenme sistemi zamana duyarlı olmalıdır. Başka bir konu, veri tabanındaki alanların mevcut keşif odağıyla ilgisi veya ilgisizliğidir. Örneğin, posta kodları, bir ürünün satışı gibi bir öğeye coğrafi bir bağlantı kurmaya çalışan herhangi bir çalışma için önemlidir.

5.6. Veri Madenciliği Kullanım Alanları

Veri madenciliği, bazıları aşağıda listelenen çok çeşitli uygulama alanlarına sahiptir.

Perakende / Pazarlama

- Müşterilerden satın alma modelleri tanımlama
- Müşteri demografik özellikleri arasındaki ilişkileri bulma
- Posta kampanyalarına yanıt vermeyi tahmin etme
- Market sepeti analizi

Bankacılık

- Sahte kredi kartı kullanım örneklerini tespit etme
- 'sadık' müşterileri tanımlama
- Kredi kartı bağlantılarını değiştirecek müşterileri tahmin etme
- Müşteri gruplarının kredi kartı harcamalarını belirleme
- Farklı finansal göstergeler arasında gizli korelasyonlar bulma

Sigorta ve Sağlık

- Talep analizleri yapma yani, hangi tıbbi prosedürlerin birlikte talep edildiğini belirleme
- Hangi müşterilerin yeni hizmetleri satın alacağını tahmin etme
- Riskli müşterilerin davranış kalıplarını belirleme
- Hileli davranışlar

Taşımacılık

- Satış noktaları arasında dağıtım programlarının belirlenmesi
- Yükleme modelini analiz etme

İnternet

- Metin madenciliği uygulamaları geliştirme
- Web pazarlama modeli oluşturma
- Arama motorlarını oluşturma

Tıp

- Ofis ziyaretlerini tahmin etmek için hasta davranışını karakterize etme
- Farklı hastalıklar için başarılı tıbbi tedavileri belirleme (Rai Technology University, 2018).

5.7. Veri Madenciliğinde Gelişmeler

Veri madenciliği, gizli tahmin bilgilerinin büyük veri tabanlarından çıkarılmasıdır. Kuruluşların veri ambarlarındaki en önemli bilgilere odaklanmalarına yardımcı olacak büyük potansiyele sahip güçlü bir teknolojidir. Veri madenciliği ilk kez 1990'larda tanıtıldı. Veri madenciliği süreci bilgisayarlar veri depolamaya başladığı zaman başladı ve işlem gücü, yazılım vb. dahil olmak üzere bilgisayar kapasitesindeki artışla devam etti. Günümüzün teknoloji dünyasında, herkes en iyi kararları vermek için verilerini en iyi şekilde kullanmaya çalışır. 1960'lı yıllarda bilgisayarlarda, kasetlerde ve disklerde veri toplanması ve saklanması başladı. 1980 yılında kullanıcılar; yapılandırılmış sorgu dili kullanarak ilişkisel veri tabanlarında depolanan veriler hakkında analiz yapmaya başladı. 1990 yılında veri ambarı, üretildi. Çok boyutlu veri tabanları ve çevrimiçi analitik işleme, veri ambarının büyümesine katkıda bulunmuştur. Şimdi, ortaya çıkan teknik ise veri madenciliğidir. Evrimin her aşaması incelendiğinde, her adımın önceki adım üzerine inşa edildiği çok açıktır. Verilerin ve veri tabanlarının evrimi çok hızlı bir şekilde gerçekleşmekte ve bu da bu büyük miktardaki verilerden yararlı bilgi elde etme yöntemlerinin bulunmasını sağlamaktadır. Yapay zekâ,

makine öğrenimi, istatistik ve veri tabanları alanları da veri madenciliğinin büyümesine katkıda bulunmuştur.

Televizyon, doğrudan pazarlama ve analiz yazılımlarından sonra daha fazla veri toplanmaya başlanmıştır ve ayrıca bu veri merkezi olarak yönetilebilmekte ve insanların sosyal bağlantılarını içerebilmektedir. Bu arada, belirli konular için adayları destekleyen, ancak resmi kampanya veya partilerle ilgisi olmayan bağımsız savunuculuk grupları, seçmenleri hedeflemek için gelen verileri toplamaktadır. Obama 2012 kampanyasında, veri analizi ile oylama yapmak için deneysel yöntem kullanmıştır. Bunu yaparken, televizyon reklamcılığının uzun süredir siyasette egemenliğini bozmuş ve bireysel seçmenlerin çıkarlarının bilindiği ve ele alındığı bir yerel seçim kampanyası gibi yürütülen ulusal bir kampanya yaratmıştır.

Veri madenciliği teknikleri, verileri daha iyi anlamak, araştırma ve iş amaçlı veriyi kullanmak için sosyal medyaya uygulanmaktadır. Örneğin facebook, twitter, gmail, kullanıcılara kuruluşlar hakkında bilgi sağlamak için sosyal medyayı incelemek ve izlemek için hizmetler sağlar. Verilerden elde edilen bilgiler ile topluluk veya grup saptaması yapılır, bilgi yayılımı incelenir, yayılımın etkisi belirlenir, ayrıca konu algılama ve izleme, bireysel davranış analizi, grup davranış analizi ve tabii ki, işletmeler için pazarlama araştırması yapılır. 2013'ün ilk yarısında twitter, diğer şirketlere tweet verilerini satarak, bir önceki yıla göre%53 artışla 32 milyon dolar kazanmıştır.

Pazarlama amacıyla toplanan veriler diğer kuruluşlara satılır, ancak düzenli olarak devlet kurumları ile paylaşılır. Ayrıca, arama geçmişini ve bir veri tabanında tutulan e-postaları analiz ederek, reklamların arama motorunun kullanıcısına hedeflendiği Google'daki reklamlar gibi doğrudan pazarlama amaçları için de kullanılabilir. Örneğin, dünyanın en popüler arama motoru olan Google, her web araması için tanımlayıcı bilgileri depolar. Kullanılan bir URL adresi ve arama ifadesi, 18 aya kadar bir veri tabanında saklanır. Google ayrıca, kullanıcıların kişisel e-posta yazışmalarında nelerden bahsettiklerini temel alarak hedefli

reklamlar oluşturmak için gmail web posta hizmetinin kullanıcılarının e-postalarının içeriğini tarar. Şirketlerin çoğu, çalışanlarının e-posta trafiğini izlemektedir. Amerika Birleşik Devletleri hükümeti bu veri tabanlarına erişim hakkına sahiptir. FBI, Yurtiçi Güvenlik Dairesi ve diğer istihbarat ajansları, Kızılötesi programlarının bir parçası olarak 34.000'den fazla şirket ile bir bilgilendirme ortaklığı oluşturmuştur (Ramzan ve Ahmad, 2014).

5.8. Veri Madenciliği Yöntemleri

Veri madenciliği modelleri Gözetimli Öğrenme (Supervised Learning) ve Gözetimsiz Öğrenme (Unsupervised Learning) olmak üzere iki ana başlık altında incelenmektedir.

5.8.1. Gözetimli Öğrenme

Bu, eğitim verilerinden bir işlevi çıkarmayı amaçlayan bir dizi süreçle yürütülen bir veri madenciliği yöntemidir. Normal olarak, eğitim verileri bir dizi gözlemden oluşur. Her gözlem, kestiriciler olarak adlandırılan çok sayıda değişkene ve ayrıca etiketler veya sınıflar olarak da bilinen, tahmin etmek istediğimiz bir değişkene sahiptir. Bu etiketler veya sınıflar öğretmenleri temsil eder çünkü modeller bunlardan öğrenir. Modelin yarattığı fonksiyonun nihai amacı, yeni gözlemlere yönelik davranış tahminidir. Gözetimli Öğrenmeye örnek olarak sınıflandırma ve regresyon verilebilir.

Süreci iki aşamaya ayırabiliriz. Modelleme ve Tahmin. Modelleme aşamasında, modeli eğitmek için kullanılacak ham verilerle başlarız. Veri sayısını azaltmak veya dönüştürmek mümkündür. Modeli eğitmeye devam ederiz ve sonunda değerlendirmeyi gerçekleştiririz. Modelin eğitimi, inşası ve geçerliliği, mümkün olan en iyi modele ulaşmayı amaçlayan yinelemeli bir süreçtir. Bu, ayarlama yapmak için bir önceki adıma dönmemiz gerektiği anlamına gelir. İkinci aşama tahmindir. Oluşturduğumuz ve test ettiğimiz modeli kullanarak, yeni veriler için bir tahmin yürütülür ve sonuçlar oluşturulur (Pacheco, 2015).

5.8.1.1. Veri madenciliğinde sınıflandırma ve tahmin

Veri tabanı veya veri ambarı, akıllı karar vermeyi sağlamak için kullanılabilecek gizli bilgiler açısından zengindir. Akıllı karar, insan kararına oldukça benzeyen otomatik karar verme becerisini ifade eder. Veri madenciliğinde sınıflandırma ve tahmin, akıllı karar üretebilecek gözetimli öğrenme yöntemlerindendir. Veri madenciliğinde sınıflandırma ve tahmin, önemli veri sınıflarını tanımlamak veya gelecekteki veri trendlerini tahmin etmek için modelleri çıkarmak için kullanılabilecek iki veri analizidir. Sınıflandırma sürecinde ilk aşama öğrenme sürecidir. Eğitim verileri, sınıflandırma algoritması ile analiz edilecektir. Daha sonra ikinci aşama, sınıflandırma modelinin veya sınıflandırıcının doğruluğunu tahmin etmek için test verilerinin kullanıldığı sınıflandırma işlemidir. Doğruluk kabul edilebilir olarak düşünülüyorsa, yeni verilerin sınıflandırılmasına kurallar uygulanabilir. Veri sınıflandırmasında kullanılan çeşitli teknikler; karar ağacı, Bayes yöntemleri, Bayes ağı, kural tabanlı algoritmalar, yapay sinir ağları, destek vektör makinesi, ilişki kuralı madenciliği, k-en yakın komşu, genetik algoritmalar, kaba kümeler ve bulanık mantıktır.

Son zamanlarda, veri madenciliği yaklaşımını kullanarak İK (İnsan Kaynakları) sorunlarının çözümüne büyük ilgi gösteren bazı araştırmalar vardır. İK'da veri madenciliği tekniği özellikle bir iş için doğru adayları seçmede kullanılmaktadır. Verilen hizmetin süresi, satış primi, sigorta acentelerinin süreklilik endeksleri ve işletmeciler tarafından yapılan yanlış operasyon davranışlarının analizi gibi alanlarda veri madenciliği yaklaşımı kullanılmaktadır (Jantan ve ark., 2011).

5.8.1.2. Regresyon Analizi (Regression Analysis)

En basit haliyle, Regresyon Analizi, araştırmacıların bir bağımsız ve bir bağımlı değişken arasındaki ilişkiyi analiz etmelerine olanak tanır. Bu analizin kullanılmasının temel faydaları şunlardır: 1. Bağımsız değişkenlerin bağımlı değişkenle anlamlı bir ilişkisinin olup olmadığı belirlenir. 2. Bağımlı değişken

üzerindeki farklı bağımsız değişkenlerin etkilerinin göreceli gücü belirlenir. 3. Tahminler yapılır.

Regresyon analizi, pazar araştırmasında en sık kullanılan araçlardan biridir. Pazarlama uygulamalarında, bağımlı değişken genellikle önemsediğimiz sonuçtur (örneğin satışlar), bağımsız değişkenler ise bu sonuçları elde etmek için aldığımız araçlardır (örneğin, fiyatlandırma veya reklam). Regresyon Analizi satışların fiyatlara veya promosyonlara ne ölçüde bağımlı olduğu gibi soruların yanıtlanmasına yardımcı olabilir. Bu analiz ayrıca tahminlerin yapılmasına da yardımcı olabilir. Örneğin, satışlar, fiyatlar ve promosyon faaliyetleri ile ilgili verileri kullanarak bir regresyon modelini tahmin edersek, bu regresyon analizinden elde edilen sonuçlar, fiyatların %5 oranında artması ve tanıtım faaliyetlerinin gerçekleşmesi durumunda satışa ne olacağı konusunda kesin bir cevap verebilir. Böyle kesin cevaplar yöneticilerin sağlam kararlar vermelerine yardımcı olabilir. Ayrıca, %5, %10 ve %15 fiyat artışlarının satış etkilerini hesaplamak gibi çeşitli senaryolar sunarak, yöneticiler pazarlama planlarını değerlendirebilir ve pazarlama stratejileri oluşturabilirler.

Regresyon modelleri genel olarak şöyle gösterilir:

$$y=\alpha+\beta_1 x_1+e$$

y açıklamaya çalıştığınız değişken olan bağımlı değişkeni temsil eder. Bağımsız değişken x_1 ile gösterilir. $\beta_1$bağımsız değişken x_1'in katsayısını gösterir ve ayrıca eğim olarak ifade edilir. Hata, regresyon çizgisi ve gerçek gözlem arasındaki farktır. Birden fazla bağımsız değişken eklersek, bu çoklu regresyon olarak adlandırılır.

Regresyonda bağımlı değişkenin aralık veya oran olarak ölçeklendirilmesi gerekmektedir. Veriler aralık veya oran olarak ölçeklendirilmemişse, alternatif regresyon türlerinin kullanılması gerekir. Bağımlı değişken ikili ise ve sadece iki değeri (ör. Sıfır ve bir) alırsa Lojistik Regresyon kullanılır. Bağımlı değişken,

ikiden fazla seviyeli bir nominal değişkenden oluşuyorsa, Çoklu Lojistik Regresyon kullanılır (Mooi ve Sarstedt, 2014).

5.8.2. Gözetimsiz Öğrenme

Gözetimsiz öğrenim, bir dizi girdi değişkeniyle ilişkili olarak ilişkileri ve kalıpları tanımlamayı amaçlayan bir veri madenciliği yöntemidir. Gözetimli öğrenmeden temel farkı, girdi verilerinin sınıf etiketleri içermemesidir, bu nedenle veri yapılarını kendi ilişkilerine göre tahmin etmek ve bulmaya çalışmak için hiçbir değişken yoktur. Denetimsiz öğrenmede, iki aşamadan söz edebiliriz: Modelleme ve yarar.

Modelleme aşamasında, giriş verileri alınıp, özellik seçimi veya özellik çıkarma teknikleri uygulamak için devam edilir. En uygun değişkenleri tanımlandıktan sonra, eldeki problemi çözmek için gözetimsiz öğrenmenin en iyi yöntemi seçilir. Örneğin, kümelenme veya ilişkilendirme kuralları olabilir. Yöntemi seçtikten sonra, sonuçlardan memnun olana kadar model oluşturmaya ve yinelemeli bir ayarlama işlemi gerçekleştirmeye devam edilir. Gözetimsiz öğrenmede, modelleme aşamasında elde edilen bulgular amacın yerine getirilmesi için yeterli olabilir, bu durumda süreç durur. Artık modele sahip olunur ve onun yeniden kullandığı istediğimiz ikinci aşama vardır. Yeni veriler ile üzerinde çalışmak için ürettiğimiz model kullanıp sonuç alınır (Pacheco, 2015).

5.8.2.1. Kümeleme (Clustering)

Kümeleme analizi veya basit bir şekilde kümelenme, bir dizi veri nesnesini (veya gözlemleri) alt kümelere ayırma işlemidir. Her alt küme, aynı kümedeki nesnelerin birbirine benzediği, ancak diğer kümelerdeki nesnelere benzemediği bir kümedir. Farklı kümeleme yöntemleri aynı veri kümesinde farklı kümelenmeler oluşturabilir. Ayrıştırma tarafından değil, kümeleme algoritması tarafından gerçekleştirilir. Bu nedenle, kümelenme, verilerde önceden bilinmeyen grupların keşfine yol açabilmesi açısından yararlıdır.

Kümeleme analizi; iş zekâsı, görüntü deseni tanıma, Web araması, biyoloji ve güvenlik gibi birçok uygulamada yaygın olarak kullanılmaktadır. İş zekasında, kümeleme, çok sayıda müşteriyi gruplara ayırmak için kullanılabilir; burada bir grup içindeki müşteriler güçlü benzer özellikleri paylaşır. Bu gelişmiş müşteri ilişkileri yönetimi için iş stratejilerinin geliştirilmesini kolaylaştırır.

Kümeleme, Web aramasında da birçok uygulama bulmuştur. Örneğin, bir anahtar kelime araması, çok fazla sayıda web sayfası nedeniyle çok fazla sayıda isabet (yani, arama ile alakalı sayfalar) verebilir. Kümeleme, arama sonuçlarını gruplar halinde düzenlemek ve sonuçları kısa ve kolay erişilebilir bir şekilde sunmak için kullanılabilir.

Veri madenciliği işlevi olarak, kümeleme analizi, verilerin dağılımı hakkında bilgi edinmek, her kümenin özelliklerini gözlemlemek ve daha fazla analiz için belirli kümelere odaklanmak için bağımsız bir araç olarak kullanılabilir. Ayrıca sınıflandırma gibi diğer algoritmalar için bir ön işlem adımı olarak hizmet edebilir.

Kümelenme büyük veri kümelerini benzerliklerine göre gruplara ayırır. Kümeleme, aykırı saptama için de kullanılabilir. Aykırı saptama uygulamaları, kredi kartı sahtekarlığının tespiti ve elektronik ticarette suç faaliyetlerinin izlenmesini içerir.

Literatürde birçok kümeleme algoritması vardır.

Bölümleme yöntemleri: Bir dizi n nesnesi verildiğinde, bir bölümleme yöntemi, her bir grubun en az bir nesne içermesi gerektiği gibi, verileri k gruba ayırır(k<=n). İyi bir bölümlemenin genel ölçütü, aynı kümedeki nesnelerin birbirine “yakın” ya da birbirleriyle ilişkili olması, farklı kümelerdeki nesnelerin “uzak” ya da çok farklı olmasıdır.

Hiyerarşik yöntemler: Hiyerarşik bir yöntem, hiyerarşik ayrışmanın nasıl oluştuğuna bağlı olarak ya kümelenmiş ya da bölücü yöntem olarak sınıflandırılabilir. Aşağıdan yukarıya yaklaşım olarak da adlandırılan kümelenmiş yaklaşımda, tüm gruplar bir sonlandırma koşuluna kadar, nesneleri veya grupları

birbirine yakın olarak birleştirir. Yukarıdan aşağıya yaklaşım olarak da adlandırılan bölücü yaklaşımda, her bir ardışık yinelemede, bir küme, her nesne bir kümede veya bir sona erdirme koşuluna kadar daha küçük kümelere ayrılır. Hiyerarşik kümeleme yöntemleri, uzaklık bazlı veya yoğunluk ve süreklilik temelli olabilir.

Yoğunluğa dayalı yöntemler: Veri uzayındaki düşük yoğunluklu bölgelerle birbirinden ayrılan yoğun veri bölgelerini küme olarak kabul eder.

Izgara tabanlı yöntemler: Veri uzayını sonlu sayıda karelere bölerek kümeleme işlemlerinin tümünün gerçekleştirileceği bir ızgara yapısı elde eden yöntemlere denir (Han ve ark., 2012).

5.8.2.2. Birliktelik Kuralları (Association Rules)

Birliktelik kuralları analizi (aynı zamanda piyasa sepeti analizi olarak da bilinir) perakendeci işlem veri tabanlarından birliktelik veya eş-oluşumları çıkararak müşteri satın alma kalıplarını keşfeden veri madenciliği yöntemidir. Bu birliktelik keşfi, perakendecilerin, müşteriler tarafından hangi ürünlerin sıklıkla satın alındığı hakkında fikir edinerek pazarlama stratejileri geliştirmelerine yardımcı olabilir.

Örneğin, işletmenin elindeki veri kümesinden takım elbise ve kravat satışı arasında bir ilişki olduğu tespit edilebilir çünkü takım elbise satın alan birçok müşteri kravatta da satın almaktadır. Bu durumda {Takım Elbise} → {Kravat} diye bir kural çıkarılabilir.

İşletmenin yöneticileri bu tür ilişkileri belirlediklerinde ilişkili olan ürünleri birbirlerine yakın raflara koyarak ya da çeşitli promosyonlarla ürün pazarlama stratejilerini değiştirerek ürün satışlarının artmasını sağlayabilirler.

Bir ilişkilendirme kuralı, A ve B'nin ayrık öğe kümeleri, yani $A \cap B = \emptyset$ olduğu, form $A \Rightarrow B$'nin bir ima ifadesidir. İlişkilendirme kurallarını bulmak için en yaygın kullanılan tekniklerden biri Apriori algoritmasıdır. Apriori algoritması uygulanırken, standart önlemler ilişkilendirme kurallarını değerlendirmek için kullanılır. Bu kurallar destek ve güven değeridir. $A \Rightarrow B$ gibi birliktelik kuralları için

kural değerlendirmesi için iki kriter birlikte kullanılır. Destek, AUB 'yi içeren işlemlerin yüzdesidir. Form desteği (A⇒B) = P (AUB) şeklindedir, buradaki destek, AUB içeren işlemlerin yüzdesidir. Bu olasılık, P (AUB) olarak alınmıştır. Güven, AUB içeren işlemlerin yüzdesinin A'yı içeren işlemlerin yüzdesine oranıdır. Form güveni (A⇒B) = P (B \ A) = destek (AUB) / destek (A) şeklindedir. Hem minimum destek eşiğini hem de minimum güven eşiğini karşılayan kurallara güçlü denir (Avcılar, 2014).

Birliktelik kuralı e-ticaret, ekonomi, pazarlama, perakende sektörlerinde yaygın olarak kullanılmaktadır. Birliktelik kuralının günümüzde ne kadar sık kullanıldığını gösterecek bir örnek Amazon e-ticaret sitesinin bu analizi kullanarak listelediği ürünlerden müşterilerin satın aldıkları ürünlere benzer olanlarını da listelemeleridir. Böylece hedef müşterilerin beğenecekleri başka ürünleri almaları için cazibe oluşturulmaktadır. Bu verilen örneği şimdilerde neredeyse tüm e-ticaret sitelerinin kullandığı da görülmektedir (Çığşar, 2017).

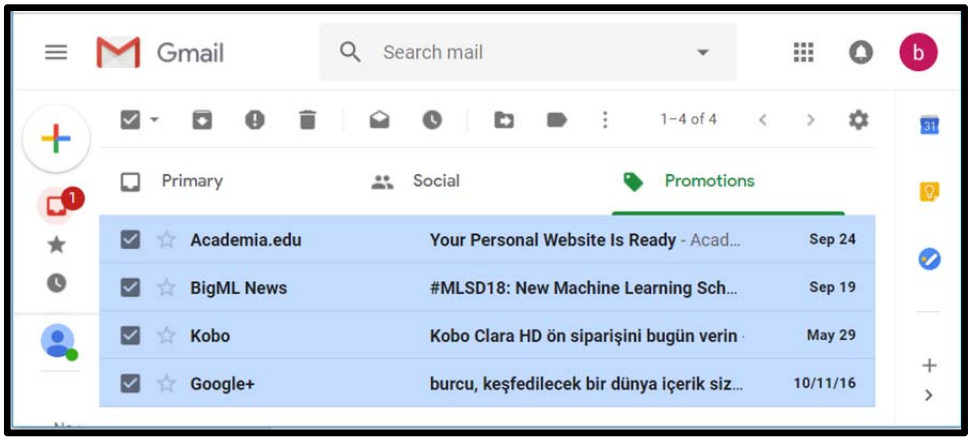
5.9. Veri Madenciliği Yazılımları

Veri madenciliği için ücretsiz ve herkese açık yazılım araçları, son 20 yıldır gelişmektedir. Bu araçların amacı, oldukça karmaşık veri analizi sürecini kolaylaştırmak ve tüm ilgili araştırmacılara ticari veri analiz platformlarına ücretsiz bir alternatif sunmaktır. Ticari yazılım olarak; SPSS Clementine, Excel, SAS, SQL Server, Oracle, MATLAB vb. sayılabilir. Açık kaynak kodlu yazılımlar ise; WEKA, Orange, Rapid Miner, KNIME, R, KEEL (Knowledge Extraction based on Evolutionary Learning) gibi yazılımlardır (Jovic ve ark., 2014).

6. MATERYAL VE YÖNTEM

6.1. Önemli E-Posta Olma Olasılığı

E-posta, 1971 yılında bilgisayar mühendisi olan Raymond Samuel Tomlinson tarafından ilk kez denenmiştir. Hemen yanı başındaki bilgisayara @ işaretini kullanarak posta atmasıyla başlamıştır. Daha sonra 26 Mart 1976 yılında İngiltere' de ARPANET adında internete dönüşebilecek bir bilgisayar ağı çıktı. 1978'de ise bugün e-posta trafiğinin çok büyük bir bölümünü oluşturan ilk reklam gönderimi yapıldı. 1982 yılında "e-mail" yani e-posta kelimesi ilk kez kullanılmaya başlandı. 21. yy 'da internet teknolojilerinin dolayısıyla da e-postanın gelişmesiyle birlikte bu defa güvenlik konusu ön plana çıktı. 2005'te, e-posta kimliklerinin doğrulanmasını sağlayan ilk teknoloji SPF (Sender Policy Framework) geliştirildi. Ardından yani Şubat 2007'de gmailin beta sürümü tamamen çıkarıldı ve tüm internet kullanıcılarının kullanabileceği şekilde kurgulandı (Gümüş, 2018).

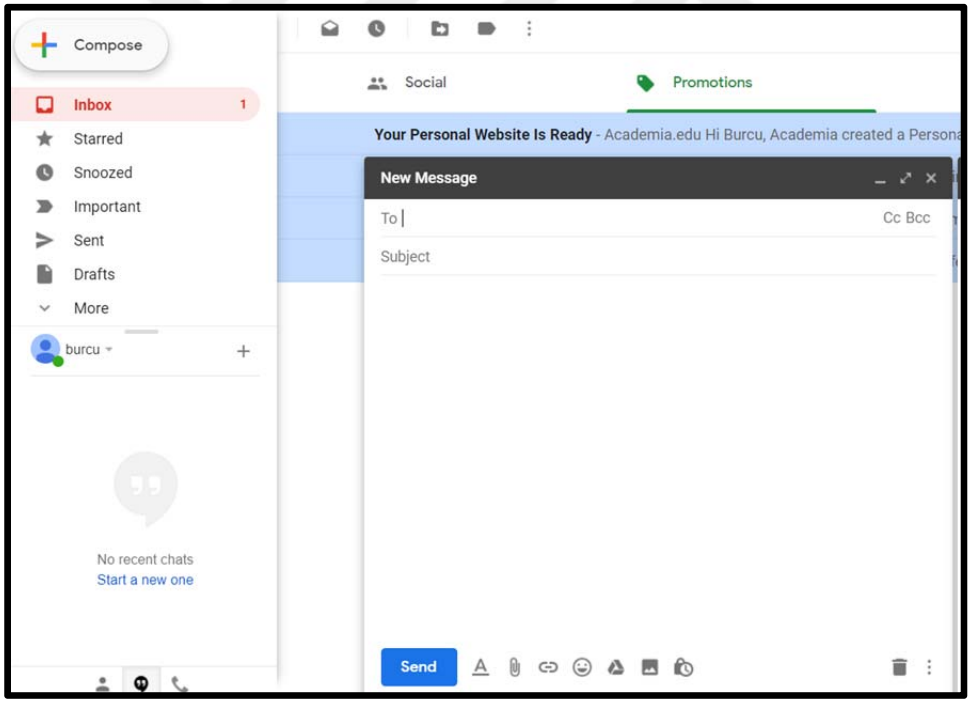


Şekil 6.1. Gmail Kullanıcı Ara Yüzü

Öncelikle Şekil 6.1'deki gmail ara yüzüne bakıldığında ortada, Gelen Kutunuzdaki tüm e-postalar görülür. Gelen e-postaları çeşitli başlıklar altında

kategorize edebilmektedir (Primary, Social, Promotions vs.). Bu ana bölümde soldan sağa e-postanın gönderenin adı, e-postanın konusu ve Gelen Kutunuza ulaştığı zaman bulunmaktadır. Solda gezinme öğeleri bulunur: Gelen Kutusu, Yıldızlı, Önemli, Gönderilmiş Postalar ve Taslaklar. (Telstra Tech Savvy Seniors, 2018).

Gmailin özelliklerinden biride mobil uygulamasının bulunmasıdır ve uygulama 40'tan fazla dilde kullanılabilmektedir. Tüm gelen ve giden e-posta mesajlarındaki eklerde, virüs olup olmadığını otomatik olarak taramaktadır. Gönderilen elektronik postaları, göndereceğiniz kişiyi daha sonra otomatik olarak kaydetmektedir. Yenilikçi Google teknolojisi ile güçlü spam engelleme özelliğine sahiptir (Gümüş, 2018).



Şekil 6.2. Compose Kullanıcı Ara Yüzü

Şekil 6.2'de Compose sekmesi tıklandığında açılan pencerede görünenlerin anlamları şu şekildedir.

To: Mesajın gönderileceği e-posta adresinin yer aldığı bölümdür.

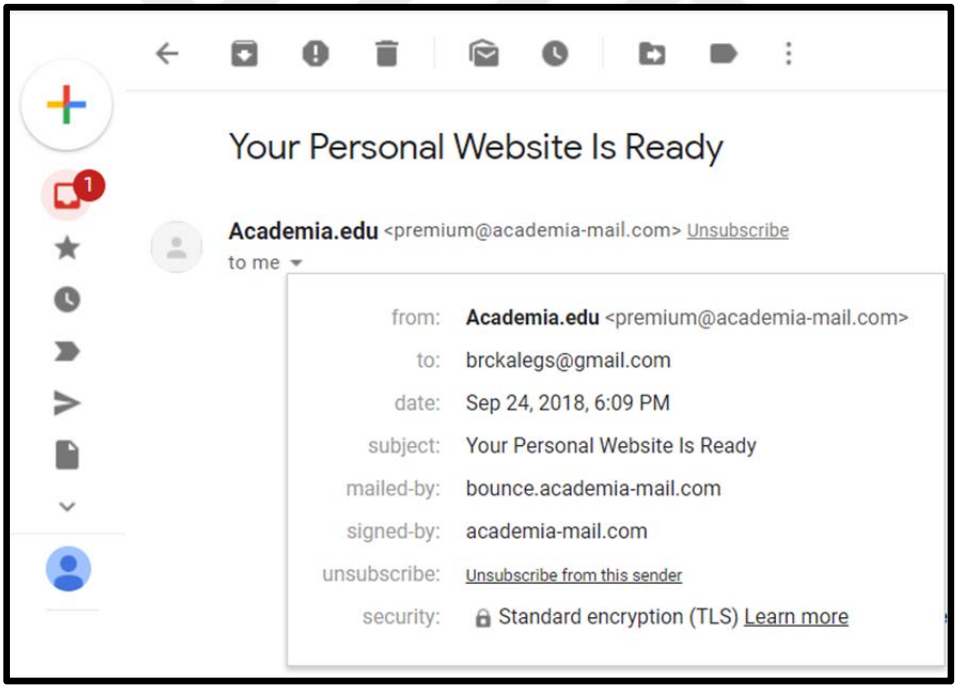
Subject: Mesajın konusu veya başlığının ifade edildiği bölümdür.

Mesaj alanı: Mesajın tam içeriğinin yer aldığı bölümdür.

CC: Mesajın gönderileceği bir başka e-posta adresinin yazılacağı bölümdür.

BCC: Mesajın bir kopyasının gönderileceği ancak diğer alıcılardan gizlenen e-posta adreslerinin yazıldığı bölümdür.

Ek (attachment): Yazıya ek olarak gönderilmek istenen dosyaların yüklendiği bölümdür.



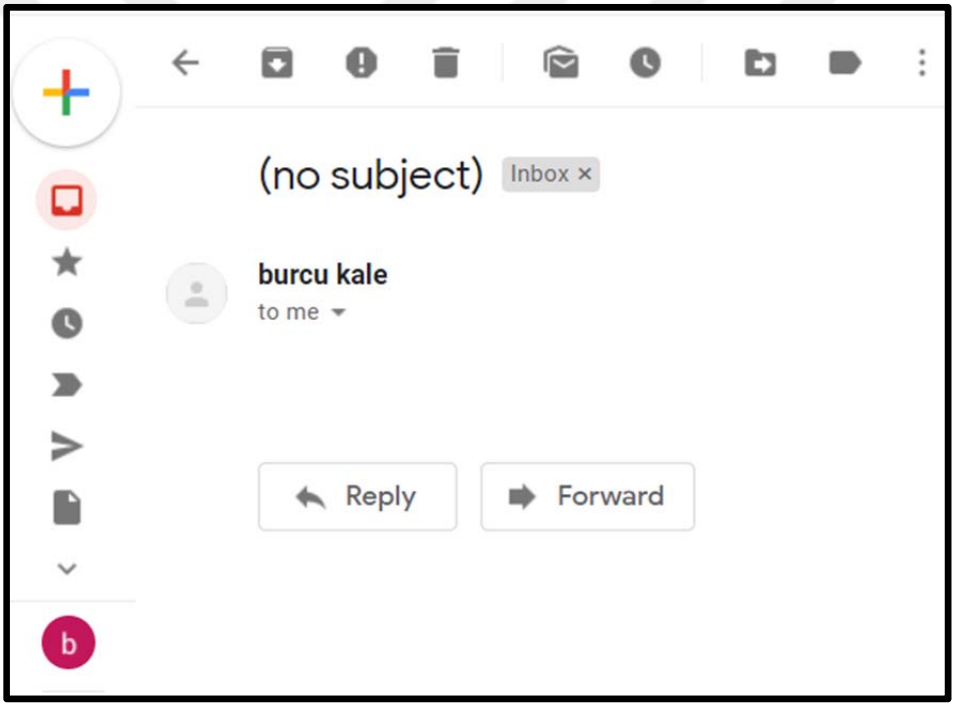
Şekil 6.3. Gelen E-postanın Bilgilerinin Yer Aldığı Kullanıcı Ara Yüzü

Şekil 6.3'de gelen e-postanın bilgilerinin bulunduğu alanda görünenlerin anlamları şu şekildedir.

From: E-postanın kimden geldiğini gösteren başlık alanıdır.

Date: E-postanın geliş zamanını (yıl, ay, gün, saat, dakika, saniye olarak) belirtir.

Subject: Gelen e-postanın konusunun olduğu alandır (Ceyhan ve Sağıroğlu, 2015).



Şekil 6.4. Reply Veya Forward Simgelerinin Yer Aldığı Kullanıcı Ara Yüzü

Şekil 6.4'de reply ve forward simgelerinin anlamları şu şekildedir.

Reply: Bir e-postayı yanıtlamaktır (Reply). Orijinal e-postanın içeriği ana penceredeedir. Yanıtlar, konuşmanın tam geçmişini tutar. Kural olarak, cevabınızı orijinal mektubun üzerindeki alana yazmanız gerekir (böylece alıcının e-postayı

açtığında gördüğü ilk şey olur). Yazdıkça, orijinal e-posta aşağı itilir. Sonra Gönder'e sol tıklanır. Orijinal e-postayı gönderen kişi, cevabınızı hemen almalıdır. Konu satırı, öncekiyle aynı olacaktır, ancak başlangıçta bulunan bir re: bunun bir yanıt olduğunu belirtir.

Forward: Açılan e-postanın altındaki İlet (Forward) sekmesinin tıklanması dışında, yukarıdakilerle aynı adımlar uygulanır. Fakat kime alanındaki alıcı gönderici kişi için doldurulmamıştır. E-postayı yönlendirmek istenilen kişinin adının yazılması gerekir. E-postaya eklemek istediğiniz herhangi bir şey varsa, orijinal içeriğin üzerine yazmanızın beklenmesi gerekir. E-postayı iletilmeye hazır olduğunda Gönder tıklanır. Bu kişi e-postayı aldığında, konu satırı Fwd'ye sahip olacaktır (Telstra Tech Savvy Seniors, 2018).

İnternet'e bağlanan kullanıcı sayısı hızla ilerlerken, elektronik posta (E-posta) hızlı bir şekilde mevcut en hızlı ve en ekonomik iletişim biçimlerinden biri haline gelmektedir.

E-posta son derece ucuz ve kolay gönderilebildiğinden, sadece arkadaşlara ve meslektaşlara mesaj alışverişinde bulunmanın bir aracı olarak değil, aynı zamanda elektronik ticaret yapmak için bir ortam olarak da muazzam halk kitleleri kazanmıştır.

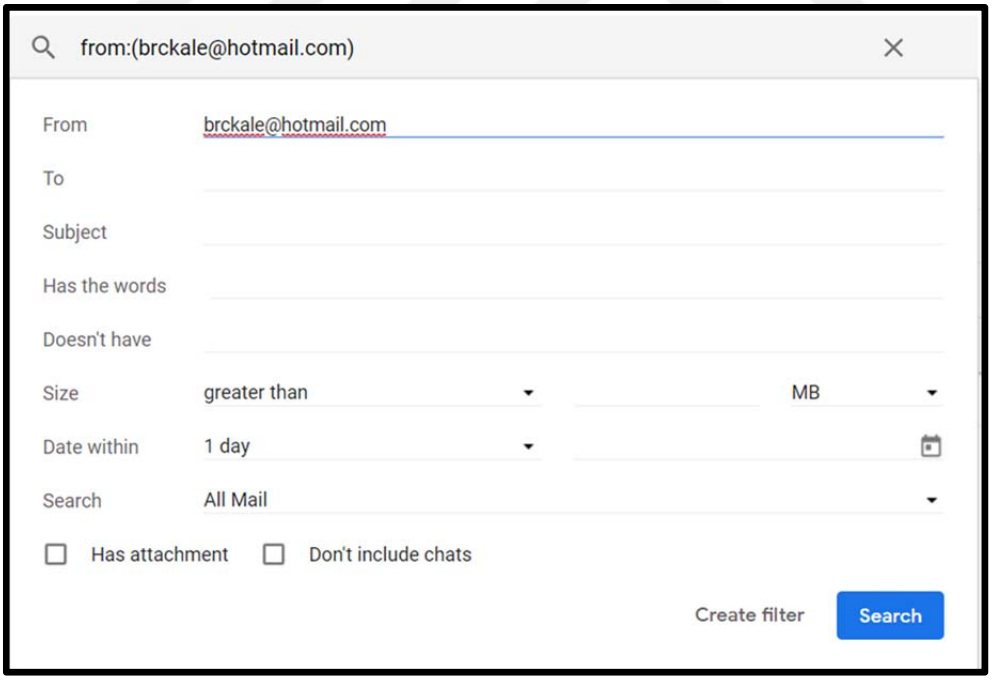
İnternet'teki doğrudan pazarlamacıların çoğalması ve E-posta adresi posta listelerinin kullanılabilirliğinin artmasıyla, önemsiz postaların hacmi (genellikle toplu olarak "spam" olarak anılacaktır) son birkaç yılda muazzam bir şekilde artmıştır (Sahami ve ark., 1998).

Gmailin tasarlanma amaçlarından biri e-postanın önemli ya da önemli değil olarak belirlenebilmesidir. Ayrıca çok önem verilenlere yıldızlama yapılabilir. İşaretleme yapmak istedikleriniz için sarı yıldız veya daha önemlilere başka yıldızlar konulabilir.

Her postanın önünde iki farklı simge bulunur. Bu simgelerden soldakini işaretleyip “More” sekmesinden “önemli olarak işaretle” sekmesini işaretleyince e-postamız önemli duruma gelir. Bunu Google, maillerinizi bir algoritma çevresinde

değerlendirerek otomatik yapar. Ama bu işaretleme de değiştirilebilir. Örneğin bir yerden gelen mailleri, gmail önemli olarak işaretledi ise bu işaret kaldırıldığında önemli olmadığı belirtilmiş olur. Tabi bu işaretleme sadece gönderici için kolay bir kullanım olmaktadır . Ayrıca sol tarafta Önemli sekmesi tıklanarak sadece önemli olarak işaretlenen iletilerde görülebilir (Durmaz, 2013).

Kullanıcıların önemli e-postaları algılaması için bir kural kümesi oluşturmasını gerektiren sistemler, kullanıcılarının sağlam kurallar oluşturabilecek kadar bilgili olduğunu varsayar. Üstelik, önemli postaların niteliği zamanla değiştikçe, bu kural kümeleri, kullanıcı tarafından yeniden ayarlanmalıdır. Bu zaman alıcı ve genellikle hataya eğilimli, sıkıcı bir süreçtir. Yani kural tabanlı bir yaklaşımın önemli posta filtrelemede sınırlı bir faydası vardır. Aşağıda gmail ortamında e-postalarınızı filtrelemek için oluşturulan kuralın aşamaları anlatılmıştır.



Şekil 6.5. Filtreleme yapılan Kullanıcı Ara Yüzü

Şekil 6.5'de filtreleme yapılan alanda görünenlerin anlamları şu şekildedir.

1. Şekil 6.1'de olduğu gibi en üstteki arama kutusunda, aşağı ok simgesi tıklanır.
2. Şekil 6.5'de görüldüğü gibi arama ölçütleri girilir. Aramanın düzgün şekilde çalışıp çalışmadığını kontrol etmek için Arama sekmesi tıklanır ve açılan pencerede görüntülenen e-postalara buradan göz atılabilir.
3. Arama penceresinin alt kısmındaki Filtre Oluştur tıklanır.
4. Filtrenin yapması istenilen işlem belirtilir.
5. Filtre Oluştur tıklanır (gmail yardım, 2018).

Ayrıca Mailtrack uzantısı, gmailde mesajlaşmada faydalı olan bir yoldur. Bu uzantı sayesinde e-postanın gönderildiğinden ve alıcının onu açtığından emin olunur. En önemlisi kaç kere okunduğu da görülür.

Şekil 6.6'da Mailtrack uzantısının oluşturduğu raporun gösterildiği pencere şu şekildedir.



Şekil 6.6. Mailtrack Uzantısının Oluşturduğu Raporun Gösterildiği Pencere (Mailtrack Support Center, 2018).

Bu çalışmanın amacı önemli e-mailleri ayırt edici belirgin farkları çıkarıp bunları öğrenen bir yapay zekâ tasarlamaktır. Yapay zekâ çalışmaları genellikle

insan zekâsı baz alınarak bilgisayarlara algoritmalar geliştirilmesi şeklinde gerçekleşir. Bu çalışmada veri madenciliği, makine öğrenmesi ve veri bilimi tekniklerini kullanarak, bir yapay zekâ yaratılmış ve e-mailler seçilen algoritmalara göre analiz edilmiş ve e-postanın önemlilik potansiyeli hakkında bilgi verilmiştir (Şeker ve Gazioğlu, 2017).

6.2. RAPIDMINER Veri Madenciliği Yazılımı

Bu çalışmada Rapid Miner yazılımı kullanılmaktadır. Rapid Miner; veri hazırlama, makine öğrenimi, derin öğrenme, metin madenciliği ve tahmine dayalı analitik için entegre bir ortam sağlayan bir veri bilimi yazılım platformudur. Araştırma, eğitim, öğretim, hızlı prototip oluşturma ve uygulama geliştirme gibi iş ve ticari uygulamalarda kullanılır ve veri hazırlama, sonuç görselleştirme, model doğrulama ve optimizasyon dahil olmak üzere makine öğrenim sürecinin tüm adımlarını destekler.

Daha önce YALE (Yet Another Learning Environment) olarak bilinen Rapid Miner, 2001 yılında Dortmund Teknik Üniversitesi Yapay Zekâ Birimi'nde Ralf Klinkenberg, Ingo Mierswa ve Simon Fischer tarafından başlatılmıştır. Rapid Miner Java dilinde yazılmıştır. Ayrıca kendi içerisine Java dili ile kod ekleme imkânı sağlamaktadır. Bunun yanında Python, Weka veya R gibi diller/ortamlarla da uyumlu olarak çalışabilmektedir.

1 mantıksal işlemci ve 10.000 veri satırıyla sınırlı olan Rapid Miner Studio Free Edition, AGPL lisansı kapsamında mevcuttur.

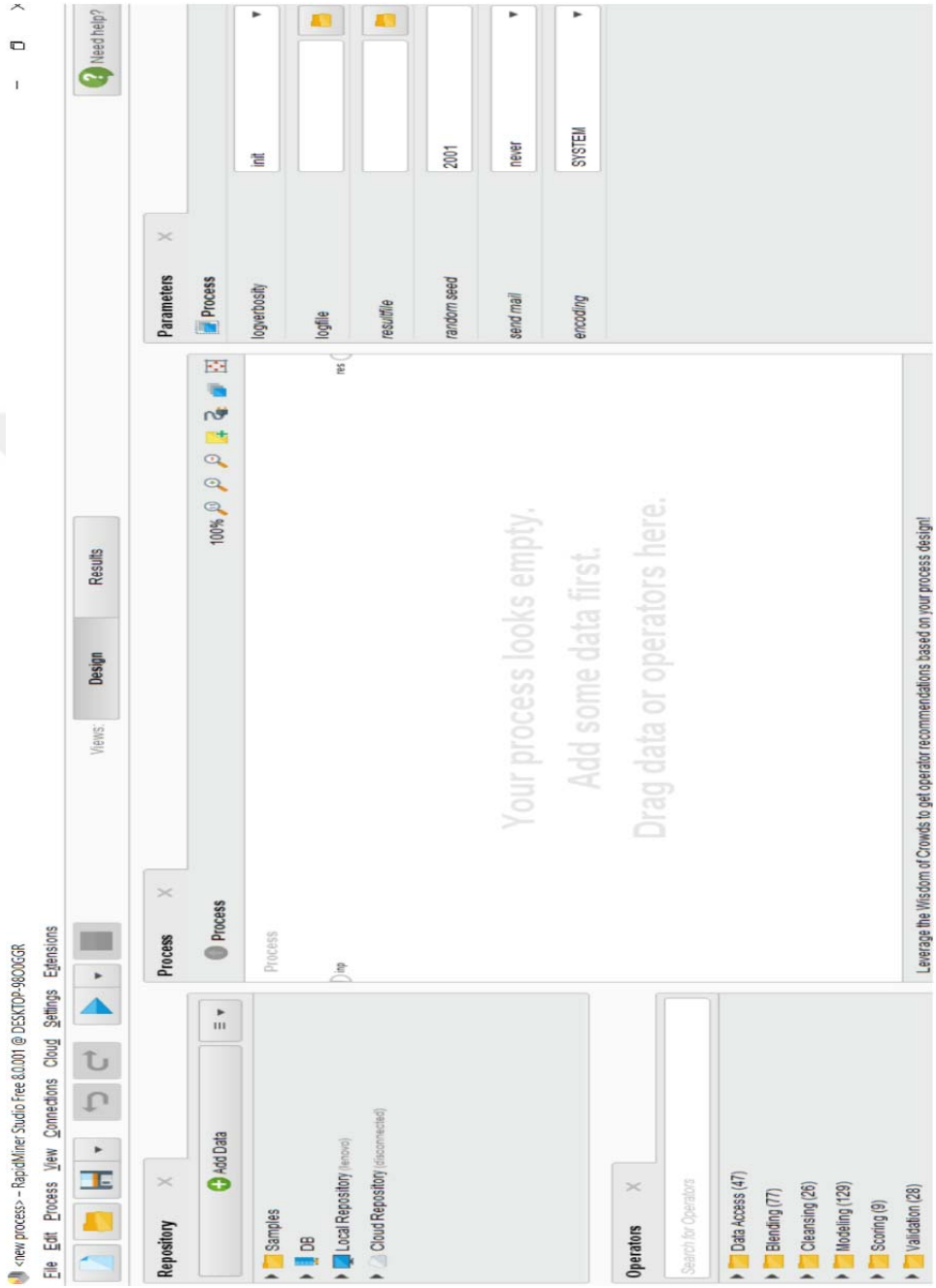
Rapid Miner, sıralamalarda kullanım oranları ve tercih edilebilirlik açısından çoğu ücretli yazılım ile rekabet edebilen bir yazılımdır. Hem yeni başlayan hem de uzman seviyesindeki kullanıcıların ihtiyaçlarına cevap verebilmektedir. Gelişmekte olan büyük veri teknolojisine ayak uydurarak sunucu mimarisinin üzerinde çalışabilmektedir. Radoop olarak isimlendiren hizmeti bulut üzerinden sunmaktadır. Bunun gibi çok sayıda güncel gelişmeye ayak uyduran özelliğinin yanında yüzlerce eklenti seçeneği bulunmaktadır.

Repository penceresi: örnek verileri içerir. Rapid Miner'a dışarıdan veri eklerken bu bölüme eklenir.

Operators penceresi: veri madenciliği ile ilgili işlemleri, operatörleri içerir. Bu operatörlerin parametreleri sağ üst köşede görülmektedir.

Process: bu boş alanda işlemler yapılacaktır (Şeker, 2018).

Şekil 6.7'de bu pencereler görülebilir.



Şekil 6.7. Rapid Miner Penceresi

6.3. Sınıflandırma Algoritmaları ve E-mail Önemlilik Analizi

Rapid Miner yazılımındaki sınıflandırma algoritmaları veriye uygunluğu göz önünde bulundurularak denemeler yapılacaktır, elde edilen verilerle sınıflandırma algoritmalarının nasıl sonuç verdiği gözlemlenecektir. Bu sınıflandırma algoritmaları; Decision Tree, Deep Learning, Gradient Boosted Tree, K-NN, Naive Bayes, Random Forest ve Logistic Regression algoritmalardır. Sınıflandırma algoritmaları karşılaştırılmış ve önemli e-posta olma olasılıkları belirlenmiştir.

6.3.1. En Yakın K Komşuluğu algoritması (K-NN)

K-NN algoritması, basitliği, doğruluğu ve minimum hata oranı sağlaması nedeniyle, makine öğrenimi ve veri madenciliği için en popüler sınıflandırma algoritmalarından biridir.

E-posta spam filtreleme, içerik geri kazanımı, gen ifadesi, protein-protein etkileşimi ve 3 boyutlu yapı tahmini, online alışveriş sitelerinde müşteri segmentasyonu ve kredi başvuru sınıflandırma gibi veri madenciliği uygulamalarında kullanılır.

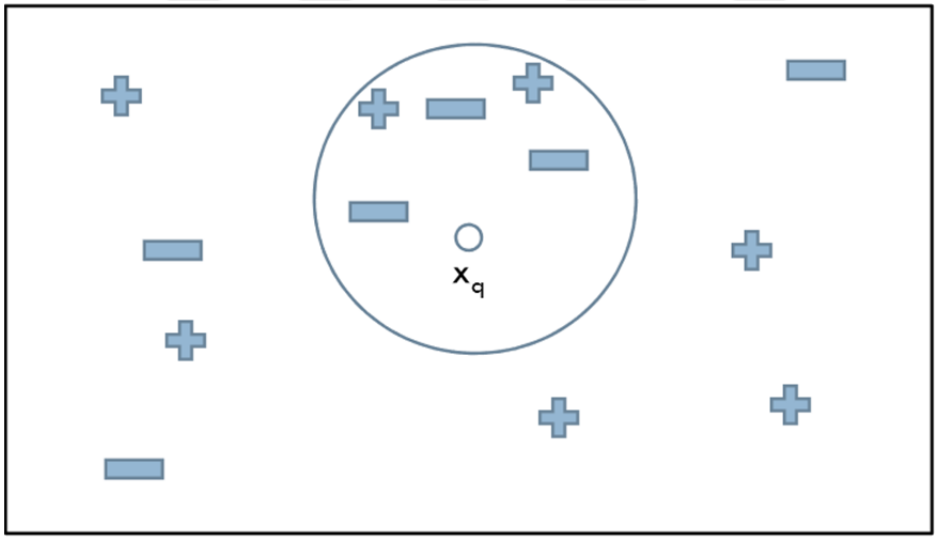
Algoritma yürütme süresi, veri kümesi geniş olduğunda sınıflandırma işlemi için önemli bir engeldir. K-NN algoritmasında yüksek bellek gereksinimi vardır çünkü eğitim verilerinin hemen hemen tümünü saklar. Tahmin aşaması yavaş olabilir.

Gözetimli öğrenme olan k-NN algoritması iki tip veri kümesini; eğitim veri kümesi ve test veri kümesini kullanır. Eğitim veri seti, özellik vektörleri ve önceden tanımlanmış sınıf üyeliğinden oluşan etiketli bir veri kümesidir. Test veri kümesinde, sınıf tanımlanmamıştır ve girdi sadece özellik vektörleriyle birlikte gelir.

Bu algoritmanın dört ana adımı vardır ve bunlar:

1. Öncelikle k değeri belirlenir.
2. İki vektör arasındaki mesafeyi hesaplamanın birçok yolu vardır. Bu ayrıca giriş veri kümesinin türüne de bağlıdır. Eğer bir sayı ise yaygın olarak kullanılan yöntemler Euclidian, Manhattan, Minkowski ve Mahalanobis uzaklıklarıdır. Bunlar arasında Euclidian uzaklık ölçümü yaygın kullanılmaktadır.
3. Veri tabanımızdaki test verisi örneğine minimum mesafedeki k adet girdi seçilir.
4. Son olarak, bu girdilerin en yaygın sınıflandırması bulunur ve bu test verisi, K-NN örneğindeki sınıflandırma olur.

Şekil 6.8'de bir K-NN örneği verilmiştir.



Şekil 6.8. K-NN Örneği

K = 5 ise, bu durumda sorgu örneği Xq en yakın komşularının üçü negatif olarak sınıflandırıldığı için negatif olarak sınıflandırılır (Selvaluxmiy ve ark., 2016).

6.3.2. Bayes Sınıflandırması

Bayes sınıflandırması istatistiksel sınıflandırıcılardır. Bayes teoremi, 18. yüzyılda olasılık ve karar teorisinde çalışma yapan, bir İngiliz papaz olan Thomas Bayes'den alınmıştır. Bayes sınıflandırması, Bayes'in teoremine dayanarak tanımlanmıştır. Sınıflandırma algoritmalarını karşılaştıran çalışmalarda Bayes sınıflandırması, büyük veri tabanlarına uygulandığında yüksek doğruluk ve hız sergilemiştir.

Bayes teoremine göre: X bir veri grubu olsun. X, bir dizi n özniteliğin üzerinden yapılan ölçümlerle tanımlanır. H, veri grubu X belirtilen bir C grubuna aittir hipotezidir. Sınıflandırma problemleri için, P (H \ X)'i, diğer bir deyişle, X veri grubunun C sınıfına ait olma olasılığını ararız. Burada P(H\X) koşullu olasılıktır ve aşağıdaki formül ile gösterilir.

$$P(H\backslash X) = \frac{P(X\backslash H)P(H)}{P(X)}$$

P(H\X): X olayı gerçekleştiği durumda H olayının meydana gelme olasılığıdır.

P(X\H): H olayı gerçekleştiği durumda X olayının meydana gelme olasılığıdır.

P(H) ve P(X): H ve X olaylarının önsel olasılıklarıdır.

Burada önsel olasılık Bayes teoremine öznellik katar. Diğer bir ifadeyle örneğin P(H) henüz elde veri toplanmadan H olayı hakkında sahip olunan bilgidir. Diğer taraftan P(X\H) ardıl olasılıktır çünkü veri toplandıktan sonra, H olayının gerçekleşmiş olduğu durumlarda X olayının gerçekleşme ihtimali hakkında bilgi verir.

6.3.2.1. Naive Bayes

Veri madenciliğinde, Naive Bayes sınıflandırıcıları, verideki tüm değişkenlerin birbirlerinden bağımsız ve eşit derecede önemli olduğu varsayılarak analiz gerçekleştiren, varsayımlarıyla Bayes teoreminin uygulanmasına dayanan basit olasılıksal sınıflandırıcılar ailesidir. Örneğin, bir meyve kırmızı, yuvarlak ve çapı yaklaşık 10 cm ise bir elma olarak kabul edilebilir. Bir Naive Bayes sınıflandırıcı, bu özelliklerin her birinin; renk, yuvarlaklık ve çap özellikleri arasındaki olası korelasyonlardan bağımsız olarak, bu meyvenin bir elma olması olasılığına bağımsız olarak katkıda bulunduğunu göstermektedir.

Naive Bayes sınıflandırıcılar gözetimli öğrenme ortamında çok verimli bir şekilde eğitilebilir. Naive Bayes'ın bir avantajı, sadece sınıflandırma için gerekli parametreleri tahmin etmek için az sayıda eğitim verisi gerektirmesidir (Han ve ark., 2012).

Bayes teoreminin m tane örneğe uygulandığı varsayılsın. S= {$S1$, $S2$, …, Sm} veri kümesi eğitim veri kümesidir. S veri kümesindeki her örnek n boyutlu {$X1$,$X2$,…,Xn} vektörlerle ifade edilsin ve bu veri kümesinde k tane $C1$,$C2$,…,Ck ile gösterilen sınıf değişkeni olsun. Her örnek bu sınıflardan birine ait olur. Bu verilere ek olarak sınıfı bilinmeyen X veri kümesi verildiğinde bu kümenin sınıfları Bayes teorimi geliştirilerek P($Ci\backslash X$) olasılığı ile tahmin edilir (Çığşar, 2017).

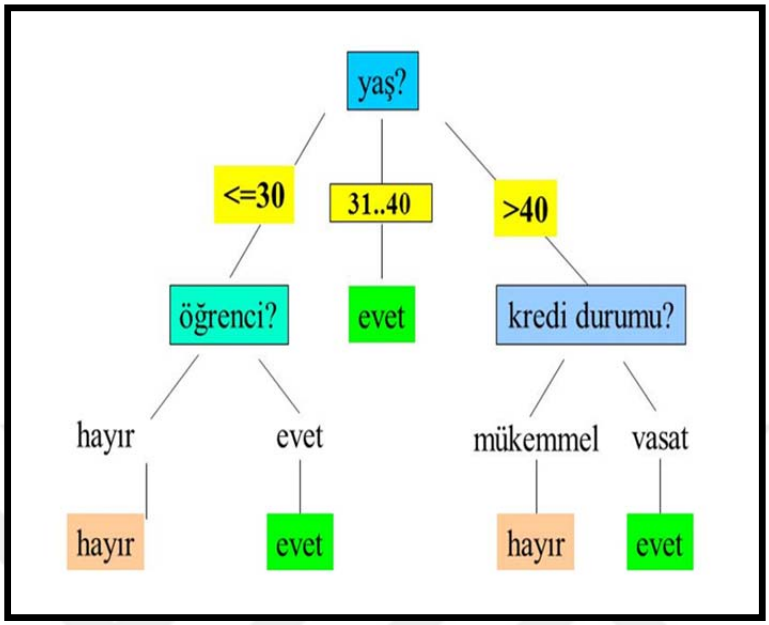
Örnek vermek istersek, bir kişini kadın mı erkek mi olduğuna dair bir veri seti içinde kayıt esnasındaki kişinin uzunluğu, kilosu ve ayak numarası gibi veriler de vardır. Bir insan böyle bir analizi kendi beyninde yaparak tahmin yöntemleri ile kişinin kadın mı erkek mi olduğuna karar verebilir. Ancak veri madenciliği bu kararların tüm kriterlerin etkisi ile verildiği bir yaklaşımdır. Dolayısıyla öğretilen sisteme kişinin uzunluğu 1.80, ağırlığı 95 ve ayak numarası 42 şeklinde bir bilgi verildiğinde eğitilmiş sistem daha önce gerçekleşmiş istatistiklerden faydalanarak kişinin kadın mı erkek mi olduğu ihtimalini hesaplar ve tahminini bildirir (Han ve ark., 2012).

6.3.3. Karar Ağaçları (Desion Trees)

Bir Karar ağaçları ağaç benzeri bir grafik veya modeldir. Ters çevrilmiş bir ağaç gibidir, çünkü kökleri tepede bulunur ve aşağı doğru büyür. Tahmin, test noktaları ve dalları ile bir karar ağacı inşa ederek elde edilebilir. Bir karar ağacının her bir test noktası, belirli bir girdi değişkenini (veya özniteliğini) test etmeyi içerir ve her bir dal, yapılmakta olan kararı temsil eder. Daha fazla dal içermeyen bir düğüme yaprak düğüm denir. Düğümün derinliği, düğümden köke ulaşmak için gereken minimum adım sayısıdır.

Bir karar ağacının girdi değerleri kategorik veya sürekli olabilir. Karar ağaçlarının iki çeşidi vardır: sınıflandırma ağaçları ve regresyon ağaçları. Sınıflandırma ağaçları genellikle kategorik olan, genellikle ikili olan - evet veya hayır, satın alma veya satın almama gibi doğadaki çıktı değişkenlerine uygulanır. Diğer taraftan regresyon ağaçları, bir tüketici malının tahmini fiyatı veya bir aboneliğin satın alınma olasılığı gibi sayısal veya sürekli olan çıktı değişkenlerine uygulanabilir.

Karar ağaçları uygulamada yaygın olarak kullanılmaktadır. Örneğin, hayvanları sınıflandırmak için, sorular (örneğin soğukkanlı veya sıcakkanlı, memeli veya memeli olmayan) cevaplanır. Başka bir örnek, bir doktorun bir hastanın değerlendirmesi sırasında belirtilerin kontrol listesidir. Bir video oyununun yapay zekâ motoru, çeşitli senaryolara yanıt olarak bir karakterin otonom hareketlerini kontrol etmek için genellikle karar ağaçlarını kullanır. Perakendeciler, müşterileri segmentlere ayırmak veya pazarlama ve promosyonlara yanıt oranlarını tahmin etmek için karar ağaçlarını kullanabilirler. Aşağıda Şekil 6.9'da örnek karar ağacı verilmiştir.



Şekil 6.9. Örnek Karar Ağacı (Dietrich ve ark., 2015).

Karar ağaçlarında en önemli sorunlardan birisi herhangi bir kökten itibaren bölünmenin başka bir ifade ile dallanmanın hangi kıstasa göre yapılacağıdır. Karar ağaçlarında sınıflandırma yöntemleri 2 çeşittir. Bunlar, Entropiye Dayalı Algoritmalar ile Sınıflandırma ve Regresyon Ağaçları (CART)'lardır. ID3 Algoritması ve C4.5 Algoritması "Entropiye dayalı algoritmalar" arasındayken Twoing Algoritması ve Gini Algoritması "Sınıflandırma ve regresyon ağaçları" sınıfındadır (Han ve ark., 2012).

Her öznitelik, eğitim örneklerinin sınıflandırmasına karar vermek için istatistiksel test kullanılarak değerlendirilir. Hangi özellik en büyük bilgi kazancını sağlarsa ağacın kökünde yer alacak özellik olarak o seçilir. Bilgi kazancı için entropi adı verilen bir tanımlama kullanılır.

Entropi, olayların olma olasılıklarıyla ilişkili olup belirsizliğin ölçülmesi için kullanılan bir ölçüttür. Entropi bilgi ile ilişkilidir ve belirsizlik arttıkça eldeki veriyi daha iyi tanımlamak için daha fazla bilgi gerekecektir. Entropi 0-1 arası

değerler alır ve 1 değerine yaklaştıkça belirsizliğin arttığını gösterir. ID3, C4.5, CART algoritmaları en iyi ayırıcı özelliğe sahip değişkeni bulmak için entropiden faydalanır (Haciefendioğlu, 2012).

$\{y_1,\ldots,y_m\}$ gibi m farklı değer alan rasgele ayrık değişken D için entropi şu şekilde hesaplanır.

$$Info(D) = -\sum_{i=1}^{m} p_i log_2 (p_i)$$

D'yi v bölümlerine ayırmak için A özelliğini kullandıktan sonra D'yi sınıflandırmak için gerekli bilgi

$$lnfoA(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \times Info(Dj)$$

A özelliğinde dallanma ile ilgili elde edilen bilgi

$$Gain(A) = Info(D) - lnfo_A(D)$$

Sonuç olarak en yüksek bilgi kazancı olan özelliği seçilir.

Kazanma Oranı (Gain Ratio): Bilgi kazancı ölçümü, çok sayıda değeri olan öznitelikler olduğunda iyi sonuç verir.

$$SplitlnfoA(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} \times log_2(\frac{|Dj|}{|D|})$$

$$\text{Gain Ratio(A)} = \frac{Gain(A)}{Splitlnfo(A)}$$

Maksimum kazanç oranına sahip özellik, bölünecek özellik olarak seçilir.

Gini İndeksi: Bir veri kümesi D, n tane sınıftan örnekler içeriyorsa, gini indeksi, gini (D),

$$Gini(D) = 1 - \sum_{j=1}^{n} pj^2$$

olarak tanımlanır. Burada pj D'deki j sınıfının nispi frekansıdır.

Eğer bir veri seti (D) iki alt kümeye (D_1 ve D_2) bölünürse, gini indeksi, gini(D),

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

olarak tanımlanır.

Safsızlıkta azalma,

$$\Delta gini(A) = gini(D) - gini_A(D)$$

Her bir nitelik için hesaplanan Gini değerleri arasından en küçük olanı seçilir ve bölünme bu nitelik üzerinden gerçekleşir. Kalan veri seti üzerinde yukarıdaki bahsedilen adımlar tekrar uygulanır ve diğer bölünme hesaplanır (Şeker, 2015b).

Yinelemeli bölümleme, yalnızca aşağıdaki sonlandırma koşullarından biri doğru olduğunda durur: Örneklerin hepsi aynı sınıfa aitse veya örnekleri bölecek özellik kalmamış ise, kalan özelliklerin değerini taşıyan örnek yoksa veya diğer durdurma kriterleri karşılanır ise (örneğin ağacın maksimum derinliği gibi) bu işlemler son bulur.

Verideki gürültüyü azaltmak için budama yaparız. Bu yöntem düşük öneme sahip özellikleri kullanan dalları kaldırmayı içerir. Böylece ağaçların

karmaşıklığı düşer ve aşırı uyum azalarak tahmin gücü artar. Budama kök veya yapraklardan başlayabilir.

İki çeşit budama yöntemi vardır. Bunlar;

1. Erken budama (Pre-pruning)
   Ağacın büyümesini erken durduran bir yaklaşımdır.
2. Geç budama (Post-pruning) Ağacın tamamlanmasının ardından işlem yapılır. Daha doğru bir çözümdür (Han ve ark., 2012).

Decision Tree'lerde nitelik seçimi hesabı yapılırken çeşitli hesaplamalar kullanılır. Bunların bazıları bilgi kazancı (information gain), kazanç oranı (gain ratio), doğruluk (accuracy), least square, gini indeks fonksiyonlarıdır. Bu tezin uygulamasında Decision Tree algoritması için bilgi kazancı fonksiyonu kullanılmıştır.

6.3.4. Rastgele Orman (Random Forest)

Rastgele orman birden fazla karar ağacı oluşturur ve daha doğru ve istikrarlı bir tahmin elde etmek için onları birleştirir. Her bir ağaç farklı eğitim kümeleriyle eğitilerek sonuçlar kullanır. Her ağaç sınıflandırıcı üretir. Üretilen bu sınıflandırıcılar kendi arasında oylama yapar ve en fazla oyu alan sınıflandırıcıyı algoritma belirler. Seçilen bu sınıflandırıcı yeni veriler verildiğinde, veriyi sınıflandırması için kullanılır (Çığşar, 2017).

Boosting (Shapire, 1996) ve Bagging (Breiman, 1996) ağaçların sınıflandırılmasında toplu öğrenme için çok iyi iki yöntem olarak bilinir. Bagging'de eğitim verisi kullanılarak her bir ağaç inşa edilir. Ardışık gelen ağaçlar bir öncekinden bağımsızdır ve en büyük oyu alan ağaç tahmin için alınır. Boosting'de, ardışık gelen ağaçlar bir öncekine bağımlıdır. Bir önceki öncüller tarafından yanlış tahmin edilmiş noktalar için ekstra ağırlık verilir. 1996'da Berkeley, California Üniversitesi'nde Leo Breiman (CART®'ın babası) tarafından

Bagging tekniği kullanılarak her bir ağacın birbirinden bağımsız olarak eğitim verileri oluşturulmuştur (Zavoral ve ark., 2010).

Rastgele orman, özellikle büyük ölçekli sistemlerde yazılım kalitesi tahmini için iyi bir adaydır, çünkü:

- Mevcut sınıflandırma algoritmaları ile karşılaştırıldığında tutarlı bir şekilde doğru olduğu bildirilmiştir.
- Büyük veri kümelerinde etkili bir şekilde çalışır.
- Eksik verilerin tahmin edilmesi için etkin bir yönteme sahiptir ve verinin büyük bir kısmı eksik olduğunda doğruluğu korur.
- Sınıfta hangi niteliklerin önemli olduğuna dair tahminlerde bulunur.
- Diğer yöntemlere göre gürültü açısından daha sağlamdır (Guo ve ark., 2004).

Rastgele ormanda iki farklı rastlantısallık kaynağı vardır: rastgele eğitim seti ve rastgele nitelik seçimi. Her bir düğümü ayırmak için rastgele bir özellik seçkisinin kullanılması, uygun hata oranları sağlar ve gürültüye göre daha sağlamdır. Çeşitlilik, bir ağacın her düğümündeki özelliklerin rastgele seçilmesi ve en yüksek düzeyde öğrenmeyi sağlayan özellikler kullanılarak elde edilir (Zavoral ve ark., 2010).

Ağaçlar ikili bölümleme kullanılarak yetiştirilir (her ebeveyn düğüm ikiden fazla çocuğa bölünmez) (Guo ve ark., 2004).

Her ağaç, bilgi kaybı nedeniyle daha fazla düğüm oluşturulmadan, budama olmadan mümkün olan en geniş alana kadar büyütülür (Zavoral ve ark., 2010).

Rastgele orman algoritması (hem sınıflandırma hem de regresyon için) aşağıdaki gibidir:

1. Orijinal verilerden yer değiştirmeli olarak N sayıda rastgele eğitim verisi elde edilir.
2. Her bir düğüm için M toplam girdi değişkenlerinden rastgele m ≤ M olacak şekilde değişkenleri seçilir. Bu m değeri orman geliştirme süresince sabittir.
3. Her bir ağaç muhtemel en geniş oranda geliştirilir. Sınıflandırma sırasında, kuralı durdurma ya da budama işlemleri yapılmaz.
4. Ağaçların tahminleri toplanarak yeni veriler tahmin edilir (örn. sınıflandırma için çoğunluk oyu, regresyon için ortalama).
5. m azalınca korelasyon ve güç azalır, m artınca korelasyon ve güç artar (Guo ve ark., 2004).

6.3.5. Gradient Boosted Trees

Herhangi bir makine öğrenme tekniğini kullanarak hedef değişkeni tahmin etmeye çalıştığımızda, gerçek ve tahmin edilen değerlerdeki farkın ana sebepleri gürültü, varyans ve önyargıdır. Topluluk, bu faktörleri azaltmaya yardımcı olur. Toplulukları kullanmamızın nedeni, aynı hedef değişkeni tahmin etmeye çalışan birçok farklı belirleyicinin, tek başına herhangi bir tek tahminciden daha iyi bir iş çıkaracağıdır. Toplama teknikleri daha çok Bagging ve Boosting olarak sınıflandırılmıştır.

Arttırma (boosting), yordayıcıların bağımsız olarak değil, sırayla yapıldığı bir topluluktur. Gradient boosted tree ilk olarak Friedman tarafından geliştirilmiştir.

Bu teknik, sonraki yordayıcıların önceki yordayıcıların hatalarından öğrendiği mantığı kullanır. Bu nedenle, gözlemlerin sonraki modellerde görülme olasılığı eşit değildir ve en yüksek hataya sahip olanlar en çok görünür. Bu yüzden gözlemler hataya göre seçilir (Grover, 2017).

Her bir yükseltici iterasyondan sonra, yanlış sınıflanmış örneklerin ağırlıkları artar ve doğru sınıflandırılmış örneklerin ağırlıkları azalır. Bu nedenle,

her bir başarılı sınıflandırıcı önceki adımlarda sınıflandırılması zor olan örneklere odaklanır (Krauss ve ark., 2016).

Yeni kestiriciler önceki kestiricilerin yaptığı hatalardan ders aldıklarından, gerçek kestirimlere yakın ulaşmak için daha az zaman / yineleme gerektirir. Fakat durma kriterleri dikkatli bir şekilde seçilmelidir yoksa eğitim verileri aşırı yüklenebilir.

4.3.6. Derin Öğrenme (Deep Learning)

Yapay zekâ ve makine öğrenimi, derin öğrenmenin alt yapısını oluşturmaktadır.

Yapay Zekâ, 1950'lerde açığa çıkmış olup, makinaların insanlar kadar kabiliyetli bir şekilde bazı işlemleri yapabilmesidir. Zayıf yapay zekalar sadece programladığınızı yerine getirirken kuvvetli yapay zekalar algoritmik hesaplarda bulunarak programladığınızı iyileştirebilen, hatalardan öğrenebilen sistemlerdir.

Makine öğrenimi, 1980'lerde açığa çıkmış olup veri madenciliğinin kullanılması ile beraber daha popüler hale gelmeye başlamıştır. Sunmuş olduğunuz veriler ve parametreler ile benzetimler yaparak, sizden daha iyi tespitlerde bulunan, programlamadıklarınızı da açığa çıkarabilen, kendi kendini eğitebilen sistemlerdir.

Derin öğrenme, 2010'lu yıllarda kullanılmaya başlanmış, büyük veri denizi ile tek bir katmanda değil, birçok katmanda makine öğreniminde kullanılan hesapları tek bir seferde yapan, makine öğreniminde tanımlamanız gereken parametreleri bile kendisi keşfeden, belki de daha iyi parametreler ile değerlendirmelerde bulunabilen bir sistemdir (Akbulut, 2018).

Günümüzde derin öğrenme algoritmalarının temeli insan beynindeki bir nöronun bilgisayar benzetmesidir.

Derin öğrenme mimarilerinde ağın eğitimi esnasında birkaç farklı öğrenme stratejisi bulunmaktadır. Bunlar öğretmenli öğrenme, öğretmensiz öğrenme ve takviyeli öğrenmedir. Öğretmenli öğrenmede ağın girişinde bulunan girdi verisinin çıkışta ne olduğu bilgisi verilir. Ağın ağırlıklar aynı verinin tekrar tekrar ağa giriş

verisi olarak sokularak ağın ağırlıklarının buna göre düzeltilmesi sağlanır. Böylelikle ağ eğitilerek ağırlıklar belirlenir. Öğretmensiz öğrenmede ise giriş verisinin çıkışta ne olduğu bilgisi verilmez. Ağa giren her veri çıkışta yakın bir kümeye ait bir değer alır. Böylelikle her veri çıkışta bir kümenin üyesi olur. Takviyeli öğrenmede ise ağda girdi verisinin çıkışta ne olduğu bilgisi verilmez. Çıkış sonucunda doğru ya da yanlış olduğu bilgisi verilir. Doğru ise 1 yanlış ise 0 değeri atanır. Buna göre ağın eğitimi gerçekleştirilir (Doğan ve Türkoğlu, 2018).

Derin öğrenme, özellik çıkarma ve dönüşümü için çok katlı lineer olmayan işlem birimlerinin bir kademesini kullanır. Her ardışık katman, önceki katmandan çıktıyı girdi olarak kullanır. Derin öğrenmede, her seviye girdi verilerini biraz daha soyut ve kompozit bir temsile dönüştürmeyi öğrenir. Bir görüntü tanıma uygulamasında, ham girdi bir piksel matrisi olabilir; Birinci temsil tabakası, pikselleri soyutlayabilir ve kenarları kodlayabilir; ikinci tabaka kenar düzenlemelerini oluşturabilir ve kodlayabilir; Üçüncü tabaka bir burun ve gözleri kodlayabilir ve dördüncü katman, görüntünün bir yüz içerdiğini fark edebilir.

Derin öğrenmede "deep", verilerin dönüştürüldüğü katmanların sayısını ifade eder. Daha kesin olarak, derin öğrenme sistemlerinin önemli bir kredi atama yolu CAP (credit assignment path) derinliği vardır. CAP, girişten çıkışa dönüşümler zinciridir. CAP'ler, girdi ve çıktı arasındaki potansiyel olarak nedensel bağlantıları tanımlar. İleriye dönük bir sinir ağı için, CAP'lerin derinliği ağınki kadardır ve gizli katmanlar artı bir tanesidir (çıkış katmanı da parametrelendirilmiştir). CAP derinliği potansiyel olarak sınırsızdır. Çoğu araştırmacı, derin öğrenmenin CAP derinliği> 2'yi içerdiği konusunda hemfikirdir. Ekstra katmanlar, öğrenme özelliklerinde yardımcı olur.

Derin görme ağları, derin inanç ağları ve tekrarlayan sinir ağları gibi deep learning mimarileri; bilgisayar görüşü, konuşma tanıma, doğal dil işleme, ses tanıma, sosyal ağ filtreleme, makine çevirisi, biyoinformatik, ilaç tasarımı ve masa oyunu programları gibi alanlara uygulanmaktadır (Wikipedia, 2018).

Derin öğrenme algoritmalarının makine öğrenmesindeki var olan algoritmalardan ayrılan yönü; çok yüksek miktarda veriye ve karmaşık yapısı ile de bu yüksek veriyi işleyebilecek çok yüksek hesaplama gücü olan donanımlara ihtiyaç duymasıdır. Son yıllarda özellikle görüntü işleme alanında etiketli veri sayısı milyon mertebelerini geçmiştir. Ekran kartı temelli hesaplama gücü alanındaki büyük ilerlemelerin de etkisiyle derin öğrenme algoritmaları çok popüler olmuştur. Yüksek veri ve hesaplama gücü olanaklarına sahip büyük teknoloji firmaları (Google, Facebook, Microsoft, Nvidia vb.) tarafından da kullanılıp ürünlerine entegre edilmektedir (Rende ve ark., 2017).

6.3.7. Lojistik Regresyon (Logistic Regression)

Lojistik regresyon, sınıflandırma sürecinin gerçekleştirilmesine yardımcı olan bir regresyon yöntemidir. Lojistik regresyondaki nokta, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi belirleyen ve en az değişken kullanarak en uygun olanın bilimsel olarak kabul edildiği bir model sunmaktır. Bilinen doğrusal regresyon analizinde bağımlı değişken ve bağımsız değişken(ler) sayısal (ölçümle belirtilen sürekli ya da kesikli sayısal) olarak belirtilir. Bağımlı değişken nitelik olarak belirtilirse, bağımsız değişken ya da değişkenlerle arasındaki ilişki lojistik regresyon yöntemiyle aranır. Lojistik regresyon, bağımlı değişkenler ikili, üçlü ve dörtlü olduğunda da kullanılan bir yöntemdir (Korkmaz ve ark., 2012).

7. UYGULAMA VE ANALİZLER

7.1. Veri Kümesi ve Yapısı

Louis Dorard, inovasyon yönetimi, veri bilimi ve web uygulamaları konusunda uzmanlaşmış bir danışmanlık şirketi olan Codole'nin ortaklarından biridir. Elimizdeki e-posta verileri Louis Dorard'ın 2013 yılında kendi gmail ön tanımlı posta kutusundan alınmıştır (Bigml,2018). İçerik özellikleri güvenlik nedeniyle kaldırılmıştır. Bu çalışmada 4.709 birimlik veri üzerinde analizler gerçekleştirilmiştir. Veri kümesi 18 değişken ve bir sınıf değişkeninden oluşmaktadır.

Sınıf değişkenlerini açıklamadan önce normalizasyondan bahsetmemiz gerekmektedir. Normalizasyonun amacı, veriler arasında farklılığın çok fazla olduğu durumlarda verileri tek bir düzen içerisinde ele almaktır. Farklı ölçekleme sisteminde bulunan veriler birbiri ile karşılaştırılamayabilir. Buradaki amaç, matematiksel fonksiyonlar kullanarak, farklı sistemlerde bulunan verileri, ortak bir sisteme taşımak ve karşılaştırılabilir hale getirmektir. Bizim bu çalışmada kullanılan normalizasyon çeşidi Min-Max Normalisazyonu'dur. Bu yöntemde, bir grup verinin içerisindeki en büyük ve en küçük değerler ele alınır. Diğer bütün veriler, bu değerlere göre normalleştirilir. Buradaki amaç en küçük değeri 0 ve en büyük değeri 1 olacak şekilde normalleştirmek ve diğer bütün verileri bu 0-1 aralığına yaymaktır (Şeker, 2012).

Ayrıca Mailtrack uzantısı, gmailde mesajlaşmada faydalı olan bir yoldur. Bu uzantı sayesinde e-postanızı gönderdiğinizden ve alıcınızın onu açtığından emin olabilirsiniz. Ayrıca kaç kere okunduğunu da göstermektedir.

Sınıf değişkenleri şu şekilde açıklanmıştır.

date: Mesajın geliş veya gidiş zamanını(yıl, ay, gün, saat, dakika, saniye olarak) temsil eder.

me: “to”, “cc”, “bcc” ve “only” özelliklerinden meydana gelir. Özelliğimiz “to” ise bu bize ya mesajın gönderildiği kişinin ismini ve e-posta adresini ya da mesajın geldiği kişinin adresini belirtir. Eğer “cc” (Carbon Copy) ise kime veya alıcı olan kısma maili göndereceğiniz zaman “cc” kısmına mesajın kopyasının gönderildiği tüm kişiler bu e-postayı okuyarak gönderilen diğer adreslerden haberdar olur. Eğer **“**bcc” (Blend Carbon Copy) ise buraya yazılan e-posta adresleri gizli, saklı alıcılar diyebiliriz. “cc” de yazdığınız e-posta adresleri diğer kullanıcılar tarafından görünebilir ama “bcc” yazdığınız e-posta adresleri diğer e-postayı alan kullanıcılar tarafından görünmeyecektir. Özelliğimiz “only” ise bu bize e-postanın gelirken “bcc” olarak gönderildiği ayrıca mesajın geldiği kişinin aynı isimde başka bir hesabından yollandığı ve sonuçta bu gelen mesajın “to”, “cc” ve “bcc” kısımlarının olmadığını belirtir.

attachments: Yazıya ek olarak gönderilmek istenen dosyaların sayısını temsil eder.

length: Karakter olarak mesajın uzunluğunu temsil eder.

conversation: Kişiler arasında yapılan diyalog sayısını temsil eder.

%from: from kısmında hiç adres olmayabilir veya diğer from bölümlerindeki adresler ile aynı olabilir. İşte bu özellik bir adresin tekrar sayısının tüm from kısmındaki adresler içerisindeki yüzde oranını temsil eder. Bu oran normalizasyon sayesinde 0 ve 1 arasına yayılmıştır.

%read: Giden mesaj hiç okunmamış olabilir veya birden fazla okunmuş olabilir. Mailtrack uzantısını gmaile ekleyince gönderdiğimiz mesajın kaç kere okunduğunu öğrenebiliyoruz. Bu özelliğimiz tüm mesajların toplam okunma miktarı içerisinde bir mesajın ne kadar okunduğunun yüzde oranını verir. Bu oran normalizasyon sayesinde 0 ve 1 arasına yayılmıştır.

%starred: Mesajların sol tarafında üç işaret vardır. Bunlardan ortadaki yıldızlama simgesidir. Bu simge tıklama sayısına göre sarı, turuncu, kırmızı, mor, mavi veya yeşil yıldız olabilir. Daha fazla tıklanırsa kırmızı ünlem, turuncu sağ shift, sarı ünlem, yeşil kontrol işareti, mavi bilgi ve mor soru işareti çıkabilir.

Toplam 13 işaret yani 13 kademe vardır. Hiç yıldızlama yapılmamış olabilir. Yıldızlanmış tüm mesajların kademeleri toplanıp bir sonuç elde edilir. Bu %starred özelliği bir mesajın kademesi yani yıldızlanma derecesinin toplam sonuç içindeki yüzde oranını verir. Bu oran normalizasyon sayesinde 0 ve 1 arasına yayılmıştır.

%contact replied: Gönderilen bir mesaja cevap beklenebilir. Cevap gelirse mesajımız reply ile gönderilmiş olabilir. Gönderilen mesaj cevaplanmayabilir. Birden fazla da cevaplanabilir. Gelen tüm reply ile cevaplanmış mesajları teker teker toplayıp bir sonuç elde ederiz. Sonra reply ile gelen cevapları adreslerine göre gruplarız. Her bir gruptaki mesajlar aynı adresten yollanmıştır. %contact replied özelliği bu her reply ile gönderilen mesajın bulunduğu grubun içindeki mesaj sayısının bulduğumuz sonuç içindeki yüzde oranını verir. Bu oran normalizasyon sayesinde 0 ve 1 arasına yayılmıştır.

time contact replied: Cevabın gelene kadar geçen zamanının saniye cinsinden uzunluğunu temsil eder.

length contact replied: Gelen cevabın karakter olarak uzunluğunu temsil eder.

%me replied: Gelen bir mesaja cevap verilebilir. Cevap gönderirsek mesajımızı reply ile göndermiş olabiliriz. Gelen mesajı ya cevaplamayabiliriz ya da birden fazla cevaplayabiliriz. Reply ile giden tüm cevap niteliğindeki mesajları teker teker toplayıp bir sonuç elde ederiz. Sonra reply ile verilen cevapları gönderdiğimiz adreslere göre gruplarız. Her bir gruptaki mesaj aynı adrese cevap niteliğindedir. %me replied özelliği her bir mesaj için bulunduğu grubun içindeki mesaj sayısının bulduğumuz sonuç içindeki yüzde oranını verir. Bu oran normalizasyon sayesinde 0 ve 1 arasına yayılmıştır.

time me reply: Cevabın verilmesine kadar geçen zamanın uzunluğunu temsil eder.

length me reply: Verdiğimiz cevabın karakter olarak uzunluğunu temsil eder.

inaddressbook: Mesajın sahibinin adres defterinde bulunup bulunmamasını temsil eder.(True veya false olarak)

date.day-of-month: Mesajın ayın hangi gününe ait olduğunu temsil eder.

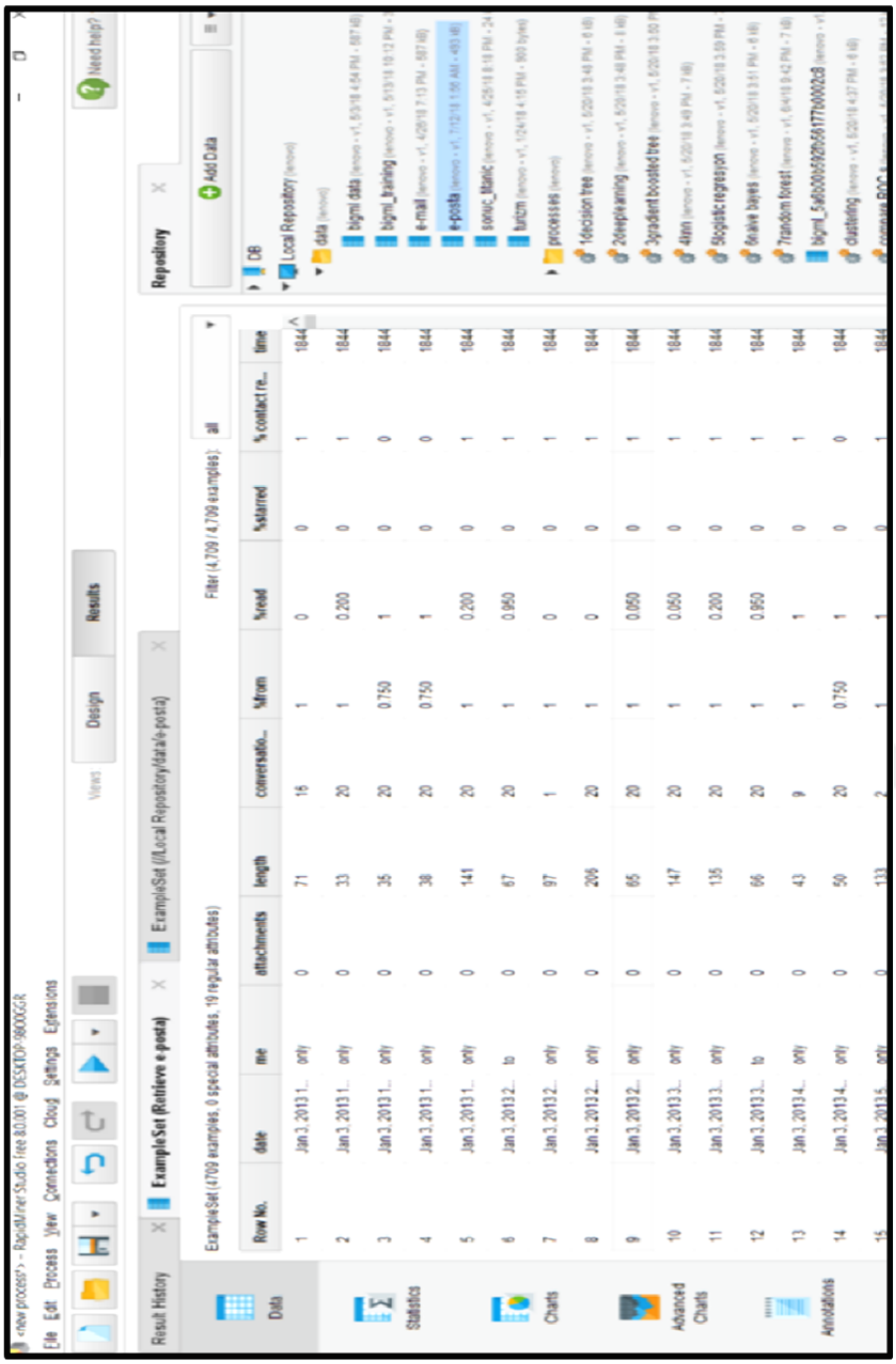
date.day-of-week: Mesajın haftanın hangi gününe ait olduğunu temsil eder.

date.hour: Mesajın günün hangi saatine ait olduğunu temsil eder.

important: Mesajın önemli olup olmadığını belirten bir sınıf değişkenidir.

Çalışmada Rapid Miner yazılımı kullanılmıştır. Rapid Miner yazılımındaki sınıflandırma algoritmaları veriye uygunluğu göz önünde bulundurularak denenmiş, elde edilen sonuçlarda yedi sınıflandırma algoritmasının iyi sonuç verdiği gözlemlenmiştir. Bu sınıflandırma algoritmaları; Decision Tree, K-NN, Naive Bayes, Random Forest, Logistic Regression, Deep Learning, Gradient Boosted Tree'dir. Seçilen yedi sınıflandırma algoritması da kendi içinde doğruluk, anma, duyarlılık ve RMSE kriterlerine göre karşılaştırılmıştır. Karşılaştırma sonucuna göre belirlenen en iyi veri madenciliği sınıflandırma algoritması; benzer özelliklere sahip veri kümelerine uygulanabilir ve e-postanın önemli olma olasılığı bu algoritma kullanılarak hesaplanabilir.

Veri kümesi .csv formatındadır. Rapid Miner programındaki "Repository" bölümüne eklenmiştir. Bunun için "Repository" klasörünün altındaki "add data" seçeneği seçilmiştir. Çıkan ekranda "My computer" ve "database" seçeneklerinden mycomputer seçeneği seçilmiştir. Açılan ekranda verinin bulunduğu klasör ve dosya seçilmiştir. Daha sonra next tuşuna basılmıştır. "select all " tuşuna basılarak kolonların hepsi seçilmiştir. Tekrar next tuşuna basılmıştır. Date kolonunun sağındaki ikonda "Change Type" seçeneğinden "date_time" ı seçtikten sonra yine aynı yolla "in addressbook" ve "important" kolonları için "binomial" seçilmiştir. Next tuşuna basılmıştır. Name kısmından dosyamızın ismi "e-posta" olarak ayarlandıktan sonra finish tuşuna basılmıştır. Verini Rapid Miner programındaki görünümü Şekil 7.1 deki gibidir.



Şekil 7.1. Rapid Miner Veri Görünümü

"Statistics" kısmından "length" kolonunda 13 eksik değer olduğu gözlenmiştir. Eksik veri problemini çözmek için "Replace Missing Value" operatörü "Operators" penceresinden seçilip "process" penceresine eklenip bağlar oluşturmuştur. Parameters penceresinden "Attribute Filter Type" kısmından "single" seçilmiş ve "attribute" kısmından eksik değer içeren "length" seçilmiştir. "default" kısmından "average" seçilmiştir. Bu yöntem ile length kolonunda eksik olan uzunlukları yerine artık 326.633 ortalama değeri yazılmıştır.

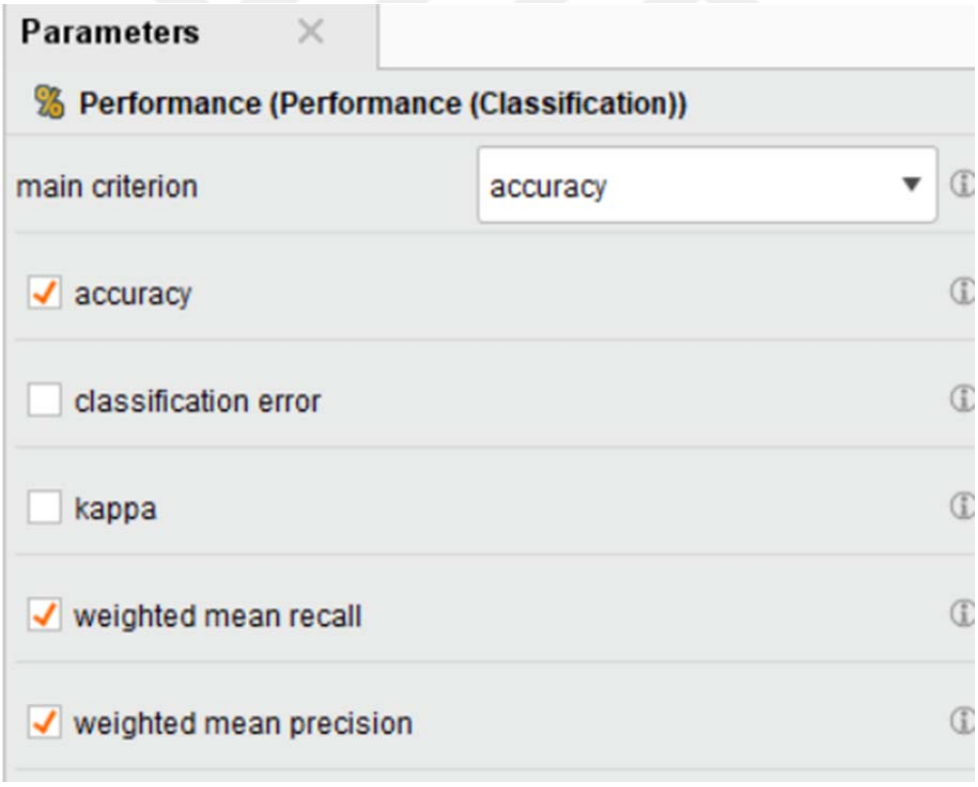
İstatistikte kullanılan "bağımlı değişken" kavramının veri madenciliğindeki karşılığı "hedef kolonudur/özelliğidir". Bu özelliği ayarlamak için "set role" operatörü kullanılmıştır. Bu operatör "Operators" penceresinden seçilip "process" penceresine eklenip bağlar oluşturulmuştur. Parameters penceresinden "attribute name" için important kolonu seçilmiş ve "target role" label olarak belirlenmiştir. Bu işlemle important kolonu hedef kolon olmuştur.

Veri kümesi e-posta %70 ve %30 olacak şekilde 2 parçaya bölünmüştür. Bir parçası "train" eğitim için sınıflandırma algoritmasına gider sonrasında %30'luk unlabeled olarak gelen kısma uygular. Bu işlem için "split data" operatörü "Operators" penceresinden seçilip "process" penceresine eklenip bağlar oluşturulmuştur. Parameters penceresinden "partitions" içeriğinden "Edit Enumarations" tıklanıp "ratio" penceresi açılmıştır. "Add Entry" tıklanıp ratio alanındaki boş çubuğa 0.7 yazılmıştır. Tekrar "Add Entry" tıklanıp açılan 2. Çubuğa 0.3 yazılmıştır. "sampling type" kısmından "stratified" seçilmiştir. Bu seçimle önemli e-postaların %70'ini alır ve önemli olmayanların %70'ini birleştirip veri setinin %70'lik için bölünecek grubuna yazar. Önemli e-postaların %30'unu alır ve önemli olmayanların %30'unu birleştirip veri setinin %30'luk kısmını oluşturur.

Bundan sonraki aşamada Decision Tree, K-NN, Naive Bayes, Random Forest, Logistic Regression, Deep Learning, Gradient Boosted Tree operatörlerinden biri "Operators" penceresinden seçilip "process" penceresine eklenip bağlar oluşturulmuştur. Bu operatörlerden biri ile eğitim yapılmıştır. Bu

operatörlerden biriyle öğrenim yaptıktan sonra “apply model” kullanılır. Apply modelin amacı öğrenilen modeli label’ları belli olmayan (important kolonu olmayan) bir veri kümesi üzerine uygulamaktır. Uygulama sonrası her satır için ayrı ayrı tahminler yapılmıştır.

Son olarak “performance” operatörü seçilip “process” penceresine eklenmiştir. Parameters penceresindeki “main criterion” kısmından “accuracy” seçilmiştir. Diğer seçeneklerden “accuracy”, “weighted mean recall”, “weighted mean precision”, “root mean squared error” ve “use example weights” seçilmiştir. Bu seçim bize tablo olarak accuracy, weight mean recall ve weight mean precision’ı vermiştir. Bundan sonra program çalıştırılmıştır. Şekil 7.2’de Parameters penceresi gösterilmiştir.



Şekil 7.2. Rapid Miner Parameters Penceresi Görünümü

7.2. Performansların Karşılaştırılması

Veri madenciliğinde sınıflandırıcıların karşılaştırılması ve en iyi sınıflandırıcının belirlenmesi çok önemlidir. Algoritmaların performanslarının karşılaştırılması doğru sınıflandırma yüzdesi (accuracy), anma(recall), duyarlılık(precision) ve kök hata kareler ortalaması (RMSE) ölçütlerine göre yapılmıştır.

Doğruluk Oranı (accuracy): Gözlemlerin doğru sınıflandırılma yüzdelerini verir.

Çizelge 7. 1. Hata Matrisi

	Tahmin edilen sınıf		
Gerçek sınıf		**Class=Yes**	**Class=No**
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Rapid Miner sonuç çıktısı:

TP = gerçek pozitif: örneklerinin tahmini pozitif olan gerçekte pozitif olanların sayısı

FP = yanlış pozitif: örneklerinin tahmini pozitif olan gerçekte negatif olanların sayısı

TN = gerçek negatif: örneklerinin tahmini negatif olan gerçekte negatif olanların sayısı

FN = yanlış negatif: örneklerinin tahmini negatif olan gerçekte pozitif olanların sayısı

$$\textbf{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

$$\textbf{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

$$\textbf{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}}$$ (Coşkun ve Baykal, 2011).

Ağırlıklı Ortalama Anma

$$= \frac{'\text{true True}'\text{için anma sınıfı} + '\text{true False}'\text{için anma sınıfı}}{2}$$

Ağırlıklı Ortalama Duyarlılık

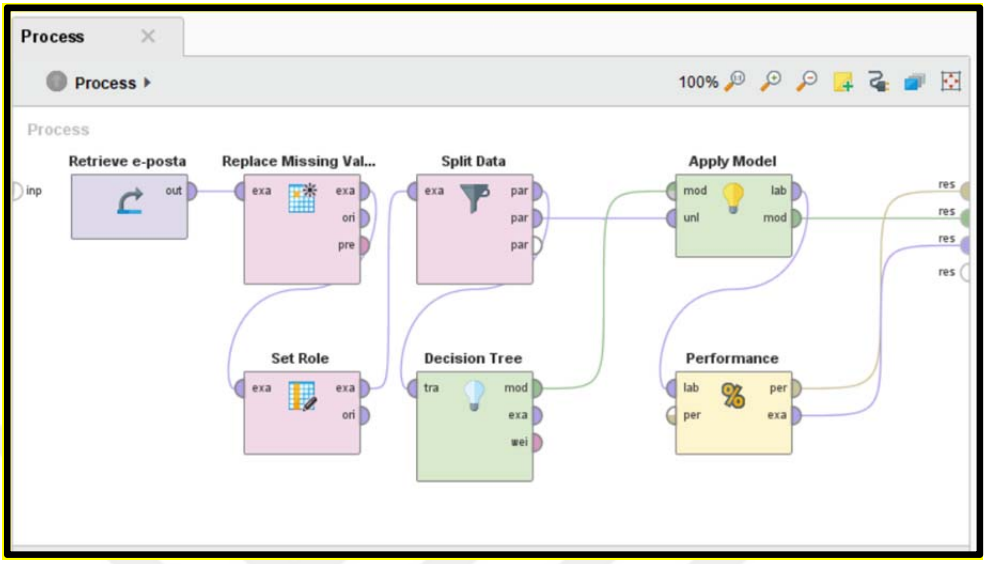
$$= \frac{\text{'pred. True' için duyarlılık sınıfı} + \text{'pred. False' için duyarlılık sınıfı}}{2}$$

(Rapid Miner Studio, 2018).

Kök hata kareler ortalaması (RMSE) = Hatanın ortalama büyüklüğünü ölçen ikinci dereceden bir puanlama kuralıdır. Formül kelimelerle ifade edilirse, tahmin ile karşılık gelen gözlemlenen değerler arasındaki farkın her birinin karesi alınır ve daha sonra örnek üzerinde ortalaması alınır. Son olarak, ortalamanın karekökü alınır. RMSE'nin düşük olması beklenir (Chai ve Draxler, 2014).

7.2.1. Karar Ağacı Algoritması

E-posta veri setimize karar ağacı algoritmasını uygularken Process penceresinin görünümü şekil 7.3'de gösterilmiştir.



Şekil 7.3. Rapid Miner Akış Şeması

Sonuçlar aşağıda şekil 7.4'de gösterilmiştir.

accuracy: 94.90%

	true False	true True	class precision
pred. False	1280	61	95.45%
pred. True	11	60	84.51%
class recall	99.15%	49.59%	

Şekil 7.4. Decisiom Tree Confusion Matrisi

Doğru sınıflandırma yüzdesine bakıldığında oranın %94,90 olduğu görülmektedir. E-posta gerçekten önemsiz olduğunda makine tahmini de 1280 e-posta için bu yöndedir. Makine gerçekte önemli olan 61 e-postanın önemsiz olduğunu tahmin etmiştir. Gerçekte önemsiz olan 11 e-postanın önemli olduğu, 60 e-posta içinde doğru tahminde bulunup önemli olduğu tespitinde bulunmuştur.

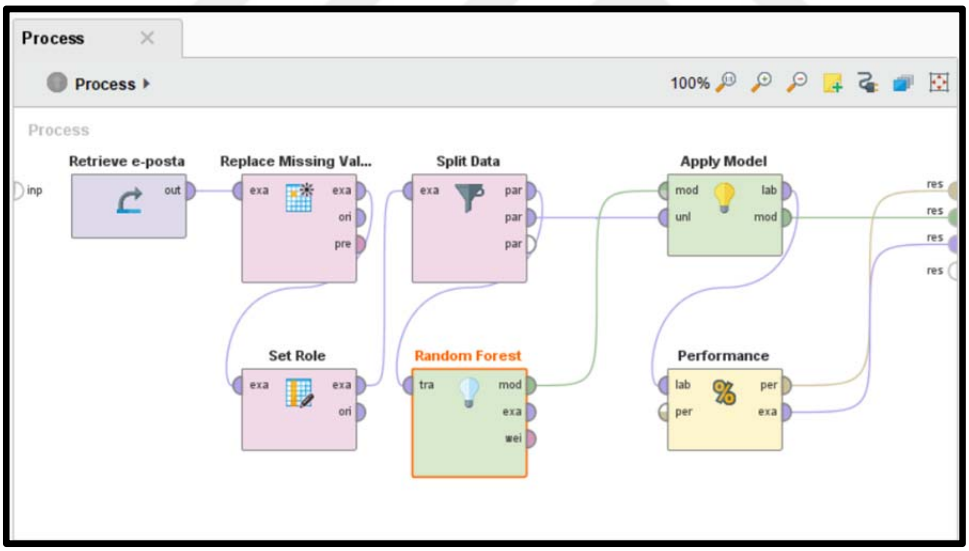
E-posta veri setimize karar ağacı algoritmasını uyguladıktan sonra çıkan diğer sonuçlar şekil 7.5'de gösterilmiştir.

weighted_mean_recall: 74.37%, weights: 1, 1
weighted_mean_precision: 89.98%, weights: 1, 1
root_mean_squared_error: 0.202 +/- 0.000

Şekil 7.5. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu

7.2.2. Rastgele Orman Algoritması

E-posta veri setimize rastgele orman algoritmasını uygularken Process penceresinin görünümü şekil 7.6'da gösterilmiştir.



Şekil 7.6. Rapid Miner Akış Şeması

Sonuçlar aşağıdaki şekil 7.7'de gösterilmiştir.

accuracy: 92.78%

	true False	true True	class precision
pred. False	1284	95	93.11%
pred. True	7	26	78.79%
class recall	99.46%	21.49%	

Şekil 7.7. Random Forest Confusion Matrisi

Doğru sınıflandırma yüzdesine bakıldığında oranın %92,78 olduğu görülmektedir. Makine gerçekte önemsiz olan 7 e-postanın önemli olduğu, 26 e-posta içinde doğru tahminde bulunup önemli tespitinde bulunmuştur. E-posta gerçekten önemsiz olduğunda makine tahmini de 1284 e-posta için bu yöndedir. Ayrıca önemli olan 95 e-posta için de önemsiz olduğunu tahmin etmiştir.

E-posta veri setimize Random Forest algoritmasını uyguladıktan sonra çıkan diğer sonuçlar şekil 7.8'de gösterilmiştir.

weighted_mean_precision: 85.95%, weights: 1, 1

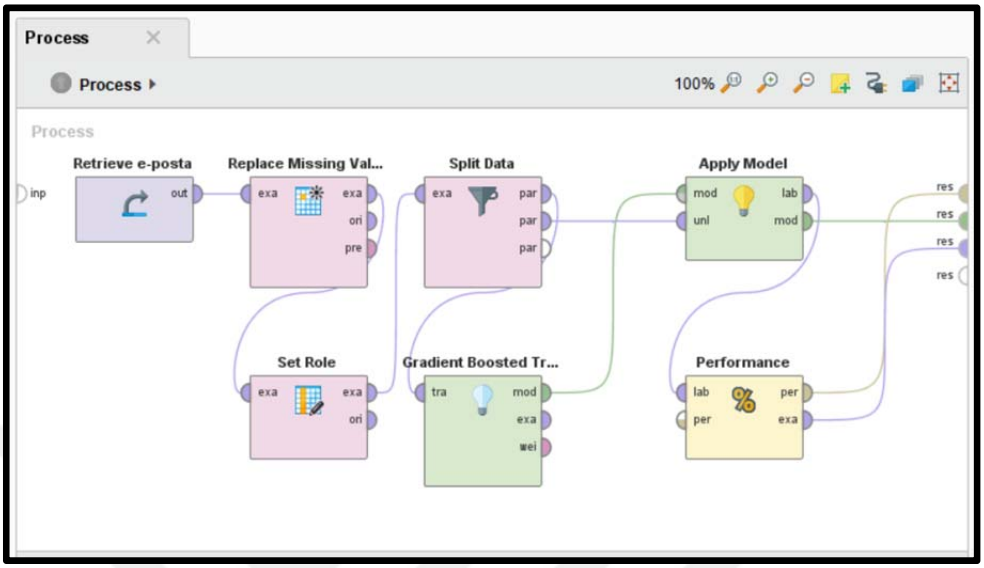
weighted_mean_recall: 60.47%, weights: 1, 1

```
root_mean_squared_error: 0.231 +/- 0.000
```

Şekil 7.8. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu

7.2.3. Gradient Boosted Algoritması

E-posta veri setimize Gradient Boosted Tree algoritmasını uygularken Process penceresinin görünümü şekil 7.9'da gösterilmiştir.



Şekil 7.9. Rapid Miner Akış Şeması

Sonuçlar aşağıdaki şekil 7.10'da gösterilmiştir.

accuracy: 94.97%

	true False	true True	class precision
pred. False	1267	47	96.42%
pred. True	24	74	75.51%
class recall	98.14%	61.16%	

Şekil 7.10. Gradient Boosted Tree Confusion Matrisi

Doğru sınıflandırma yüzdesine bakıldığında oranın %94,97 olduğu görülmektedir. Makine önemli olan 47 e-posta için önemsiz olduğunu tahmin etmiştir. E-posta gerçekten önemsiz olduğunda makine de 1267 e-posta için aynı tahminde bulunmuştur. Makine 74 e-posta içinde doğru tahminde bulunup önemli

tespitinde bulunmuş ve gerçekte önemsiz olan 24 e-postanın için önemli olduğunu belirtmiştir.

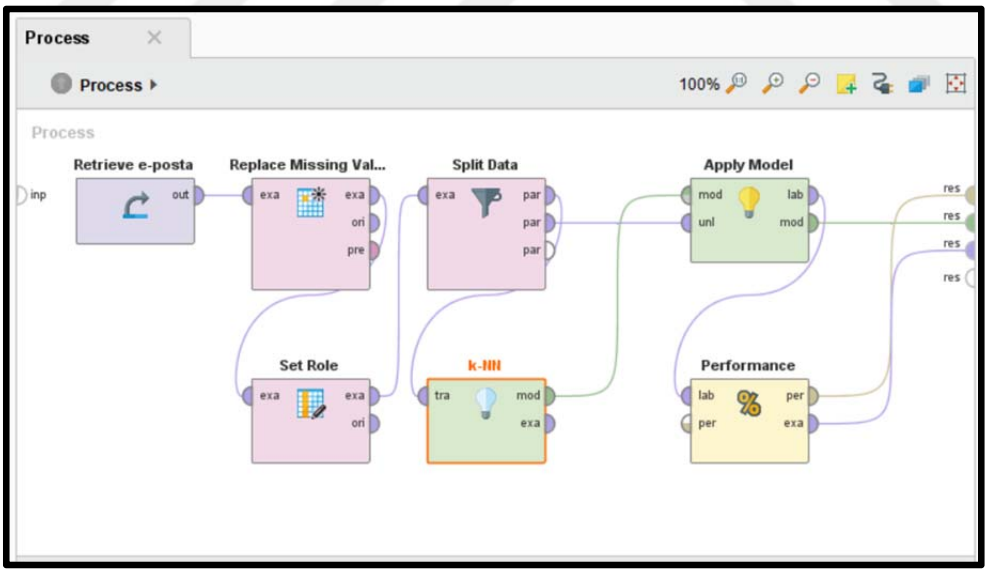
E-posta veri setimize Gradient Boosted Tree algoritmasını uyguladıktan sonra çıkan diğer sonuçlar şekil 7.11'de gösterilmiştir.

weighted_mean_recall: 79.65%, weights: 1, 1
weighted_mean_precision: 85.97%, weights: 1, 1
`root_mean_squared_error: 0.233 +/- 0.000`

Şekil 7.11. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu

7.2.4. K-NN Algoritması

E-posta veri setimize k-NN algoritmasını uygularken Process penceresinin görünümü şekil 7.12'de gösterilmiştir.



Şekil 7.12. Rapid Miner Akış Şeması

Sonuçlar aşağıdaki şekil 7.13'de gösterilmiştir.

accuracy: 87.32%

	true False	true True	class precision
pred. False	1201	89	93.10%
pred. True	90	32	26.23%
class recall	93.03%	26.45%	

Şekil 7.13. K-NN Confusion Matrisi

Doğru sınıflandırma yüzdesine bakıldığında oranın %87,32 olduğu görülmektedir. E-posta gerçekten önemsiz olduğunda makine de 1201 e-posta için bu yönde tahminde bulunmuştur. Bunun yanında önemli olan 89 e-posta için de önemsiz olduğunu tahmin etmiştir. Makine gerçekte önemsiz olan 90 e-postanın önemli olduğunu tahmin etmiştir. 32 e-posta içinde doğru tahminde bulunup önemli tespitinde bulunmuştur.

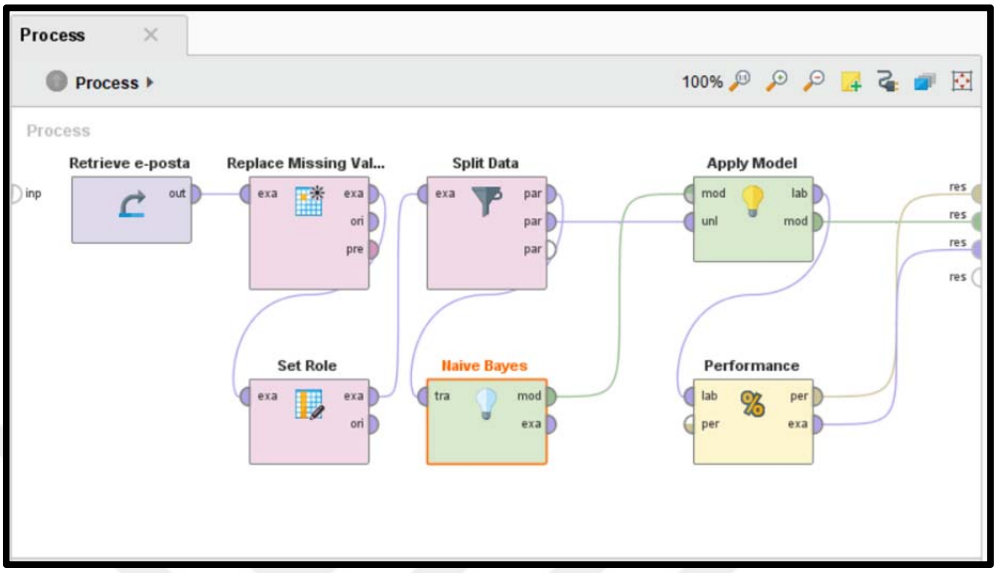
E-posta veri setimize k-NN algoritmasını uyguladıktan sonra çıkan diğer sonuçlar şekil 7.14'de gösterilmiştir.

```
weighted_mean_recall: 59.74%, weights: 1, 1
weighted_mean_precision: 59.67%, weights: 1, 1
root_mean_squared_error: 0.356 +/- 0.000
```

Şekil 7.14. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu

7.2.5. Naive Bayes Algoritması

E-posta veri setimize Naive Bayes algoritmasını uygularken Process penceresinin görünümü şekil 7.15'de gösterilmiştir.



Şekil 7.15. Rapid Miner Akış Şeması

Sonuçlar aşağıdaki şekil 7.16'da gösterilmiştir.

accuracy: 91.78%

	true False	true True	class precision
pred. False	1242	67	94.88%
pred. True	49	54	52.43%
class recall	96.20%	44.63%	

Şekil 7.16. Naive Bayes Confusion Matrisi

Doğru sınıflandırma yüzdesine bakıldığında oranın %91,78 olduğu görülmektedir. Makinenin gerçekte önemsiz olan 49 e-postanın önemli olduğu, 54 e-posta için doğru tahminde bulunup önemli tespitinde bulunduğu görülüyor. E-posta gerçekten önemsiz olduğunda makine tahmini de 1242 e-posta için bu yöndedir. Önemli olan 67 e-posta için de önemsiz olduğunu tahmin etmiştir.

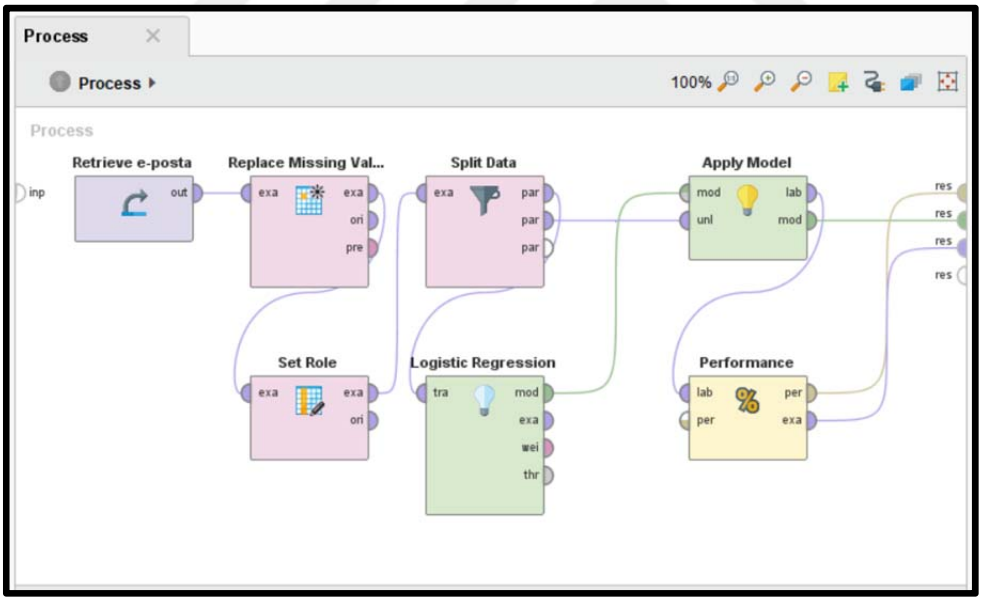
E-posta veri setimize Naive Bayes algoritmasını uyguladıktan sonra çıkan diğer sonuçlar şekil 7.17'de gösterilmiştir.

```
weighted_mean_recall: 70.42%, weights: 1, 1
weighted_mean_precision: 73.65%, weights: 1, 1
root_mean_squared_error: 0.281 +/- 0.000
```

Şekil 7.17. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu

7.2.6. Lojistik Regresyon Algoritması

E-posta veri setimize Lojistik regresyon algoritmasını uygularken Process penceresinin görünümü şekil 7.18'de gösterilmiştir.



Şekil 7.18. Rapid Miner Akış Şeması

Sonuçlar aşağıdaki şekil 7.19'da gösterilmiştir.

accuracy: 94.76%

	true False	true True	class precision
pred. False	1268	51	96.13%
pred. True	23	70	75.27%
class recall	98.22%	57.85%	

Şekil 7.19. Logistic Regression Confusion Matrisi

Doğru sınıflandırma yüzdesine bakıldığında oranın %94,76 olduğu görülmektedir. Makine önemli olan 51 e-posta için önemsiz olduğunu tahmin etmiştir. E-posta gerçekten önemsiz olduğunda makine tahmini de 1268 e-posta için bu yöndedir. Makine gerçekte önemsiz olan 23 e-postanın önemli olduğu, 70 e-posta için ise doğru tahminde bulunup önemli olduğu tespitinde bulunmuştur.

E-posta veri setimize Logistic Regression algoritmasını uyguladıktan sonra çıkan diğer sonuçlar şekil 7.20'de gösterilmiştir.

weighted_mean_recall: 78.03%, weights: 1, 1

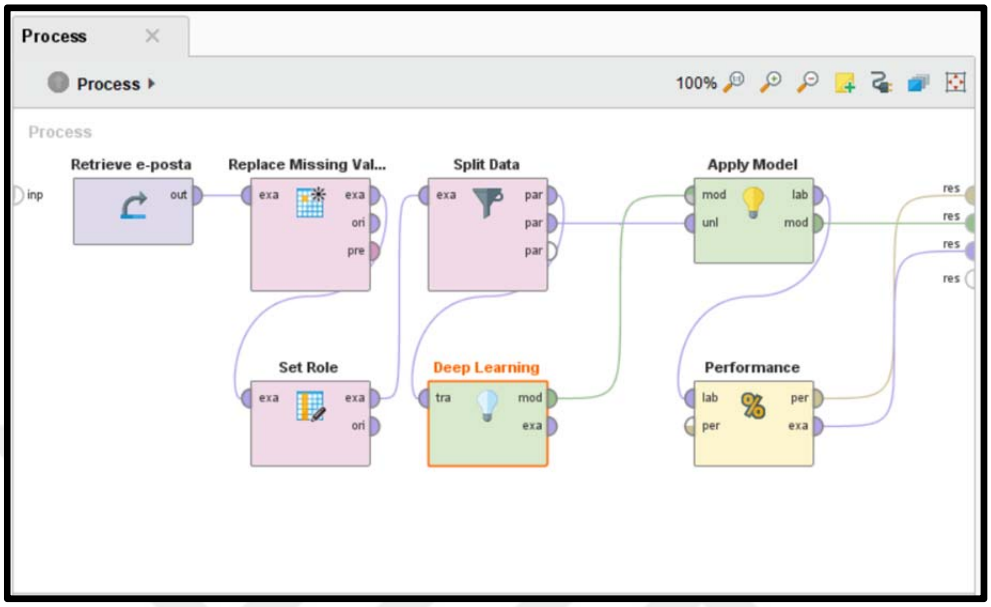
weighted_mean_precision: 85.70%, weights: 1, 1

root_mean_squared_error: 0.211 +/- 0.000

Şekil 7.20. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu

7.2.7. Derin Öğrenme Algoritması

E-posta veri setimize derin öğrenme algoritmasını uygularken Process penceresinin görünümü şekil 7.21'de gösterilmiştir.



Şekil 7.21. Rapid Miner Akış Şeması

Sonuçlar aşağıdaki şekil 7.22'de gösterilmiştir.

accuracy: 94.69%

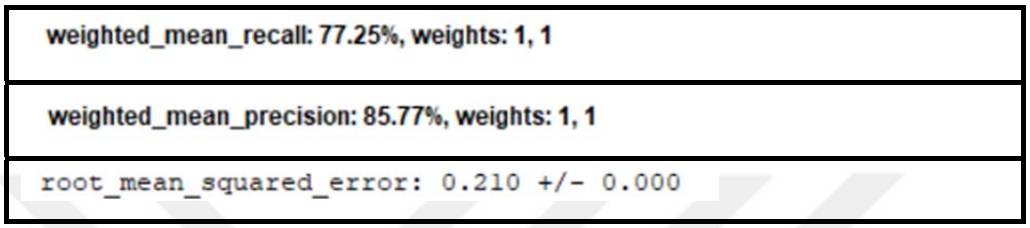
	true False	true True	class precision
pred. False	1269	53	95.99%
pred. True	22	68	75.56%
class recall	98.30%	56.20%	

Şekil 7.22. Deep Learning Confusion Matrisi

Doğru sınıflandırma yüzdesine bakıldığında oranın %94,69 olduğu görülmektedir. E-posta gerçekten önemsiz olduğunda makine tahmini de 1269 e-posta için bu yöndedir. Gerçekte önemsiz olan 22 e-posta için makine tahmini önemli olduğu yönündedir. Önemli olan 53 e-posta için de makine önemsiz

olduğunu tahmin etmiştir. 68 e-posta içinde doğru tahminde bulunup önemli tespitinde bulunduğu görülmektedir.

E-posta veri setimize Deep Learning algoritmasını uyguladıktan sonra çıkan diğer sonuçlar şekil 7.23'de gösterilmiştir.



Şekil 7.23. Weighted Mean Recall, Weighted Mean Precision ve RMSE Tablosu

7.3. Performansların Karşılaştırılması

Çizelge 7.2. Sınıflandırma Algoritmalarının Karşılaştırılması

Özellikler / Algoritmalar	Doğruluk	RMSE	Ağırlıklı Ortalama Anma	Ağırlıklı Ortalama Duyarlılık
Decision Tree	94.90%	**0.202**	74.37%	**89.98%**
Random Forest	92.78%	0.231	65.47%	85.95%
Gradient Boosted Tree	**94.97%**	0.233	**79.65%**	85.97%
k-NN	87.32%	0.356	59.74%	59.67%
Naive Bayes	91.78%	0.281	70.42%	73.65%
Logistic Regresson	94.76%	0.211	78.03%	85.70%
Deep Learning	94.69%	0.210	77.25%	85.77%

Çizelge 7.2'de Rapid Miner veri madenciliği yazılımı kullanılarak yapılan analizde sınıflandırma algoritmaları çeşitli istatistik kriterleri ışığında özetlenmiştir. Çizelgede doğruluk yüzdesine bakıldığında Gradient Boosted Tree algoritmasının

en yüksek doğruluk değerine sahip olduğu görülmektedir. K-NN algoritması ise en düşük doğruluk değerine sahiptir. Kök ortalama hata kareler değerine göre karşılaştırma yapıldığında en düşük değeri alarak bu kritere göre en iyi olduğu görülen algoritma Decision Tree'dir. Bu karşılaştırmaya göre en kötü algoritma ise k-NN'dir. Duyarlılık kriterine bakıldığında en iyi algoritmalar Decision Tree'dir. Anma değerine bakıldığında bu kriterde de Gradient Boosted Tree'nin en iyi algoritma olduğu söylenebilir. Sonuç olarak bunların içinde en iyi algoritmaların Decision Tree ve Gradient Boosted Tree olduğu görülmektedir. Aşağıda şekil 7.24'de Decision Tree ve şekil 7.25'de Gradient Boosted Tree için "important" hedef değişkeni değerleri, tahmini "important" hedef değişkeni değerleri olan örnek veri seti verilmiştir.

Row No.	important	prediction(important)	confidence(False)	confidence(True)	length
1	False	False	0.759	0.241	35
2	False	False	0.759	0.241	38
3	False	False	0.993	0.007	549
4	False	False	0.993	0.007	1481
5	False	False	0.993	0.007	1415
6	False	False	0.993	0.007	335
7	False	False	0.993	0.007	605
8	False	False	0.759	0.241	398
9	True	False	0.520	0.480	327
10	False	False	0.993	0.007	1012
11	False	True	0.114	0.886	156
12	False	False	0.759	0.241	68

Şekil 7.24. Decision Tree Örnek Seti

Row No.	important	prediction(important)	confidence(False)	confidence(True)	length
1	False	False	0.864	0.136	35
2	False	False	0.864	0.136	38
3	False	False	0.938	0.062	549
4	False	False	0.938	0.062	1481
5	False	False	0.938	0.062	1415
6	False	False	0.938	0.062	335
7	False	False	0.938	0.062	605
8	False	False	0.864	0.136	398
9	True	True	0.640	0.360	327
10	False	False	0.938	0.062	1012
11	False	True	0.589	0.411	156
12	False	False	0.864	0.136	68

Şekil 7.25. Gradient Boosted Tree Örnek Seti

8. SONUÇLAR VE ÖNERİLER

Dijital verinin toplanması ve saklanmasındaki gelişmeler, saklanan verilerin üstel bir şekilde büyümesine sebep olmuştur. Diğer yandan hayatın hızla elektronikleşiyor olması, internetin günlük yaşamın ayrılmaz bir parçası haline gelmesi toplanan veri artışını hızlandırmıştır. Veri madenciliği, veri tabanlarının araştırılması, gizli veri tahminlerinin büyük veri tabanlarından çıkarılması, veri ambarındaki önemli bilgileri analiz etme potansiyeline sahip güçlü ve yeni bir teknoloji olmakta ve beklentilerinin ötesine geçmektedir.

Bu çalışmanın amacı önemli e-postaları ayırt edici belirgin farkları çıkarıp bunları öğrenen bir yapay zekâ tasarlamaktır. Veri madenciliği, makine öğrenmesi ve veri bilimi tekniklerini kullanarak, bir yapay zekâ yaratılmış ve e-postalar seçilen algoritmalara göre analiz edilmiş ve e-postanın önemlilik potansiyeli hakkında bilgi verilmiştir. Çalışmada Rapid Miner yazılımı kullanılmıştır. Rapid Miner yazılımındaki sınıflandırma algoritmaları veriye uygunluğu göz önünde bulundurularak denenmiş, elde edilen sonuçlarda sınıflandırma algoritmalarının nasıl sonuç verdiği gözlemlenmiştir. Bu çalışmada 4.709 birimlik veri üzerinde analizler gerçekleştirilmiştir. Veri kümesi 18 değişken ve 1 sınıflandırma değişkeninden oluşmaktadır. Kullanılan sınıflandırma algoritmaları; Decision Tree, Deep Learning, Gradient Boosted Tree, KNN, Naive Bayes, Random Forest ve Logistic Regression algoritmalardır. Sınıflandırma algoritmaları karşılaştırılmış ve önemli e-posta olma olasılıkları belirlenmiştir. Algoritmaların performanslarının karşılaştırılması doğru sınıflandırma yüzdesi (Accuracy), ağırlıklı ortalama anma (Recall), ağırlıklı ortalama duyarlılık (Precision) ve kök hata kareler ortalaması (RMSE) ölçütlerine göre yapılmıştır. Buna göre Decision Tree ve Gradient Boosted Tree sınıflandırma algoritmalarının en iyi performansa sahip algoritma olduğu görülmüştür.

Bu çalışma benzer veri kümesine sahip araştırmacılara ve e-postanın önemlilik potansiyeli hakkında bilgi edinmek isteyen kuruluşlara örnek teşkil

edebilir. Bu araştırmada, en uygun sınıflandırma algoritması veriye uygulanarak, hangi e-postaların önemlilik arz ettiği belirlenmiştir. Bu uygulama sayesinde büyük şirketlerde çalışanların iş gücü yükünü hafifletilebilir. Gmail kullanıcıların önemli e-postaları algılaması için bir kural kümesi oluşturmasına izin verir, ama önemli postaların niteliği zamanla değiştikçe, bu kural kümeleri, kullanıcı tarafından sürekli olarak ayarlanmalı ve yeniden ayarlanmalıdır. Bu zaman alıcı ve genellikle hataya eğilimli, sıkıcı bir süreçtir. Yani kural tabanlı bir yaklaşımın önemli posta filtrelemede sınırlı bir faydası vardır. Bu çalışma sayesinde hangi e-postaların önemli olduğu daha kolay yoldan belirlenebilir. Ayrıca reklam içerikli ve okumaya gerek duyulmayan e-postalarda bu yolla daha kolay ortaya çıkar. Bu tezde olduğu gibi metin madenciliği kullanarak e-postaların önemliliğinin belirlenmesi ileriki çalışmalarda önerilebilir.

KAYNAKLAR

Ajayi, O., V., 2017.Advance Statistical Methods in Education. Phd Thesis, Benue. State University, Nijerya

Akbulut, İ., M., 2018. Derin öğrenme nedir? http://industryolog.com/deeplearning/ (Erişim tarihi: 18 Ekim 2018).

Alpaydın, E., 2014. lntroduction to Machine Learning. The MIT Press, England,517.

Anonymous, 2018a. 21 Big Data Statistics & Predictions on the Future of Big Data http://www.newgenapps.com/big/big-data-statistics-predictions-on-the-future-of-big-data (Erişim tarihi: 18 Ekim 2018).

Anonymous, 2018b. Transactional Databases. http://databases.looker.com/transactional . (Erişim tarihi: 18 Ekim 2018).

Arabacı, G., 2007. Veri Madenciliğinde Appriori, Tahminci Appriori ve Tertius Algoritmalarının WEKA ve YALE Programları ile Karşılaştırılması ve Bir Uygulama. İstanbul Ticaret Üniversitesi, Sosyal Bilimler Enstitüsü İşletme Yüksek Lisans Programı, İstanbul.

Aydın, N., 2017. Mobil E-Ticaret Türkiye'de ve Dünya'da Nereye Gidiyor? Avrasya Ekonomileri Uluslarası Konferansı, İstanbul, 587-592.

Avcılar, M., Y., 2014. Association Rules in Data Mining: An Application on a Clothing and Accesory Speciality Store. Canadian Social Science, 10, 3:75- 83.

Ayre, L., B., 2006. Data Mining: What it is Related to Information Retrieval And Text Mining.

Ayvaz, B., 2018. Değişken ekonomik ve rekabet koşullarında işletmelerdeki enformasyon sistemleri. İstanbul Ticaret Üniversitesi Endüstri Mühendisliği Bölümü, İstanbul.

Baydemir, M., B., 2014. Lojistik Regresyon Analizi Üzerine Bir İnceleme. İnönü Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Ana Bilim DalıMalatya.

Bayrakçı, S., 2015. Sosyal Bilimlerdeki Akademik Çalışmalarda Büyük Veri Kullanımı. Marmara Üniversitesi, İstanbul.Ceyhan, E., B., Sağıroğlu, S., 2015.E-postalarda Adli Bilişim ve Karşı Adli Bilişim Teknikleri. International Conference on Information Security and Cryptology (ISC Turkey), Türkiye, 1-7.

Bigml, 2018. Machine Learning made beautifully simple for everyone. https://bigml.com/. (Erim tarihi 18 Ekim 2018).

Chai, T., Draxler, R., R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? Geoscientific Model Development, 7: 1247-1250.

Chao, W., 2011. Machine Learning Tutorial. Graduate lnstitute of Communication Engineering, National Taiwan University, Çin Cumhuriyeti.

Coşkun, C., Baykal, A., 2011. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. 13. Akedemik Bilişim Konferansı Bildirileri, İnönü Üniversitesi, Malatya, 51-58.

Çığşar, B., 2017. Kredi Risklerinde Veri Madenciliği Sınıflandırma Algoritmaları. Çukurova Üniversitesi, Adana.

Dietrich, D., Heller, B., Yang, B., 2015. Data Science & Big Data Analytics. John Wiley & Sons, U.S.A., 409.

Doğan, F., Türkoğlu, İ., 2018. Derin Öğrenme Algoritmalarının Yaprak Sınıflandırma Başarımlarının Karşılaştırılması. Sakarya University Journal of Computer and Information Sciences, 1: 1-2018.

Dolgun, M., Ö., Ersel, D., 2014. Doğrudan Pazarlama Stratejilerinin Belirlenmesinde Veri Madenciliği Yöntemlerinin Kullanımı. İstatistikçiler Dergisi: İstatistik & Aktüerya, 7: 1-13.

Durmaz, G., 2013.Gmail'de Bilmeniz Gereken 5 Özellik. Weblopedi.net/s=gmailde+bilmemiz+gereken. (Erim tarihi 18 Ekim 2018).

Gmail Yardım, 2018. E-postalarınızı filtrelemek için kural oluşturma. http://support .google.com/mail/answer/6579?hl=tr#. (Erişim tarihi: 18 Ekim 2018).

Grivas, S., G., 1999. Data Warehousing: concepts and Mechanisms, Wirtschaftsinformatik als mitter Zwischen Technic, Ökonomie und Gesselschaft, 61-69.

Grover, P., 2017. Gradient Boosting From Scratch. http://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d. (Erişim tarihi: 18 Ekim 2018).

Gua, L., Ma, Y., Cukic, B., Singh, H., 2004. Robust Prediction of Fault-Proneness by Random Forests. 15th International Symposium on Software Reliability Engineering. Fransa, 417-428.

Gunjal, B., 2003. Database Managements: Concepts and Design. 24. IASLIC, Hindistan, 516-2003.

Gümüş, G., 2018. Dünden Bugüne E-Posta: E-mailin Tarihçesi.http://www.brandingturkiye.com/dünden-bugüne-e-posta-e-mailin- tarihcesi. (Erişim tarihi: 18 Ekim 2018).

Haciefendioğlu, Ş., 2012. Makine Öğrenmesi Yöntemleri ile Glokom Hastalığının Teşhisi. Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques. Morgan Kaufmann, 740.

Hurwitz, J., Nugent, A., Halper, F., 2013. Big Data for Dummies. John Wiley & Sons, New Jersey, 3125.

Jantan, H., Hamdan, A., R., Othman, Z., A., 2011. Data Mining Classification Techniques for Human Talent Forcasting . Faculty of Computer and Mathematical Sciences, Malezya.

Jovic, A., Brkic, K., Bogunovic, N., 2014. An Overview of free software tools for general data mining. 37th lnternational Convention on lnformation and communication Technology, Electronics and Microelectronics (MIPRO), Hırvatistan.

Korkmaz, M., Güney, S., Yiğiter, Ş., Y., 2012. The Importance of Logistic Regression Implementations in The Turkish Livestock Sector and Logistic Regression Implementations/ Fields. Harran Üniversitesi Ziraat Fakültesi Dergisi, 16(2): 25-36.

Krauss, C., Do, X., A., Huck, N., 2016. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S & P 500. European Journal of Operational Research, 259: 689-702.

Krishnan, K., 2013. Data Warehousing in the Age of Big Data. Morgan Kaufmann, 335.

Kuonen, D., 2004.Data Mining and Statistics: What is the connection? The Data Administration Newslatter, 30.

Liew, A., 2013. DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and Their lnterrelationships. Business Management Dynamics, 2, 10: 49-62.

Mailtrack Support Center, 2018. What is Mailtrack?

http://help.mailtrack.io/customer/en/portal/articles_/1385592_ _What_is_ mailrack. (Erişim tarihi: 18 Ekim 2018).

Mooi, E., Sarstedt, M., 2014. A Concise Guide to Market Research. 347.

NIST Big Data Public Working Group. 2015. NIST Big Data lnteroperability Framework: Volume 3, Use Cases and General Requirements. 5-43.

Ok, A., Güngör, O., Akar, Ö., 2011. Rastgele Orman Sınıflandırıcısı ile Arazi Kullanım Alanlarının Belirlenmesi. 3. Uzaktan Algılama ve coğrafi Bilgi sistemleri Sempozyumu, Kocaeli, 142-152.

Oracle, 2018. What is Big Data? http//www.oracle.com/bidata/guide/what-is-big-data.html. (Erişim tarihi: 18 Ekim 2018).

Öğüt, S., 2005. Veri Madenciliği Kavramı ve Gelişim Süreci. Türkiye Bilişim Derneği, 1-12.

Pacheco, E., R., 2015. Unsupervised Learning with R. Packt Publishing Ltd., İngiltere, 192.

Rai Technology University, 2018. Data Warehousing. Hindistan, 1475.

Ramzan, M., Ahmad, M., 2014. Evolution of Data Mining: An Overview. Conference on IT in Business, Industry and Government (CSIBIG), Hindis tan.

Rapid Miner Studio, 2018. http://docs.rapidminer.com/latest/studio/operators/validation/performance/predictive/performance_classification.html (Erişim tarihi: 18 Ekim 2018).

Rende, F., Ş., Karahan, Ş., Bütün, G., 2017. Derin Öğrenme Algoritmalarında Model Testleri: Derin Testler. 10. Ulusal Yazılım Mühendisliği Sempozyumu, Türkiye, 54-59.

Rosiji, A., 2015. Acid Properties, CAP Theorem & Mobile Databases, Colorado Technical University, U.S.A.

Sahami, M., Dumais, S., Heckerman, D., Horvitz, E., 1998. A Bayesian Approachto Filtering Junk E-mail. AAAI workshop, U.S.A.

Selvaluxmiy, S., Ragel, R., Kumara, T., N., 2016. Accelerating K-NN Classification Algorithm Using Graphics Processing Units. IEEE International Conference on Information and Automation for Sustainability (ICIAfS)

Şeker, S., E., 2012. İstatistiksel Normalleştirme.http://bilgisayarkavramlari.Sadievrenseker.com/2012/01/29/istatstiksel- normalleştirme -statistical-normalisation/ (Erişim tarihi: 18 Ekim 2018). ,

2015a. Veri Ambarı. YBS Ansiklopedi, İstanbul, 2,4.,2015b. Karar Ağaçları. https://www.youtube.com/watch?v=z_PCJeDOFOk&t=1212s (Erişim tarihi: 18 Ekim 2018).,

2018. Rapid Miner. Bilgisayar Kavramları Yayınları, Türkiye.

Şeker, S., E., Gazioğlu, K., 2017. Veri Madenciliği Yöntemleri ile Twitter Üzerinden Girişimcilik Analizi. YBS Ansiklopedi, İstanbul, 4, 4.

T.C. Milli Eğitim Bakanlığı, 2012. Veri Tabanı Tasarımı, Ankara.

Telstra Tech Savvy Seniors, 2018. Beginners Guide Introduction To Email-Part1. http://www.telstra.com.au/content/dam/tcom/seniors/pdf/guides/BeginnerGuide_lntroduction_to_Email_Part1_acc.pdf (Erişim tarihi:18 Ekim 2018).

Tüzüntürk, S., 2010. Veri Madenciliği ve İstatistik. Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 19, 1: 65-90.

Yaqoob, l., Hashem, l., A., T., Gani, A., Mokhtar, S., Ahmed, E., 2016. Big Data: From Begining To Future. International Journal of Information Management, 36:1231-1247.

Zaiane, O., R., 1999. Principles of Knowledge Discovery in Data Bases. University of Alberta, Department of Computing Science, Canada.

Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E., 2010. Networked Digital Technologies. Second International Conference. NDT 2010, Çek Cumhuriyeti, 256-262.

Wikipedia, 2018. Deeplearning. http://en.wikipedia.org/wiki/Deep-learning. (Erişim Tarihi: 18.Ekim 2018).

ÖZGEÇMİŞ

10/04/1978 yılında Adana'da doğdu. İlk, orta ve lise öğrenimini Hatay'da tamamladı. 1997 yılında başladığı İstanbul Teknik Üniversitesi, Fen Edebiyat Fakültesi, Matematik Mühendisliği'nden 2004 yılında mezun oldu ve 2016 yılında İstatistik Bölümünde yüksek lisansa başladı.