T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI
YÖNETİM BİLİŞİM SİSTEMLERİ PROGRAMI
YÜKSEK LİSANS TEZİ

YAPAY ZEKA TEKNİKLERİ İLE MÜŞTERİ ŞİKAYETLERİNİN OTOMATİK KATEGORİLERE AYRILMASI

Ahsen EKERYILMAZ

Danışman
Dr. Öğr. Üyesi Kutan KORUYAN

İZMİR - 2021

TEZ ONAY SAYFASI

YEMİN SAYFASI

Yüksek Lisans Tezi olarak sunduğum "Yapay Zeka Teknikleri ile Müşteri Şikayetlerinin Otomatik Kategorilere Ayrılması" adlı çalışmanın, tarafımdan, akademik kurallara ve etik değerlere uygun olarak yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

...../...../2021

Ahsen EKERYILMAZ

ÖZET

Yüksek Lisans Tezi

Yapay Zeka Teknikleri ile Müşteri Şikayetlerinin Otomatik Kategorilere Ayrılması

Ahsen EKERYILMAZ

Dokuz Eylül Üniversitesi
Sosyal Bilimler Enstitüsü
Yönetim Bilişim Sistemleri Anabilim Dalı
Yönetim Bilişim Sistemleri Programı

Çalışmada, Türkiye'nin üç büyük teknoloji firmasına ait olan şikayetler, çevrimiçi şikâyet platformlarından bir tanesi olan şikayetvar.com sitesinden alınmış, şikayetlerin otomatik kategorilere ayrılması ve analizi yapılmıştır. Çalışmanın uygulama bölümünde yöntem olarak, makine öğrenmesi sınıflandırma algoritmalarından olan Lojistik Regresyon, Doğrusal Destek Vektör Makineleri ve Olasılıksal Dereceli Azalma sınıflandırma algoritmaları kullanılarak, eğitim ve test süreci gerçekleştirildikten sonra tahminlemeler yapılmıştır. Çalışma kapsamında kullanılan üç farklı makine öğrenmesi sınıflandırma algoritmalarının sonuçları, doğruluk değerleri temel alınarak karşılaştırıldıktan sonra en iyi sonucu veren algoritma %80'lik bir başarı oranı ile Lojistik Regresyon algoritması olduğu ortaya konulmuştur. Çalışmanın amacı, üç büyük teknoloji firmasına ait şikâyet verileri elde edildikten sonra kullanılan sınıflandırma algoritmalarından en yüksek tahminleme başarısını elde eden algoritma olan Lojistik Regresyon sınıflandırma algoritması ile hangi firmanın en çok hangi kategoride şikâyet aldığını belirlenmesidir.

Uygulama çalışması Python programlama dili kullanılarak yapılmıştır. Çalışma kapsamında elde edilen şikayetlerin dört farklı kategori olan İade/Değişim ve Geri Ödeme, Teslimat/Kargo, Müşteri İlişkileri ve Hizmetleri ve Garanti ve Servis sınıflarından hangisine ait olduğu tahminlenmiştir. Elde edilen bulgular analiz edilerek grafikleri ile birlikte paylaşılmıştır.

Anahtar Kelime: Makine Öğrenmesi, Doğal Dil İşleme, Metin Sınıflandırma, Lojistik Regresyon, Müşteri Şikayetleri

ABSTRACT

Master's Thesis

Automatic Categorization of Customer Complaint with Artificial Intelligence Techniques

Ahsen EKERYILMAZ

Dokuz Eylül University

Graduate School of Social Sciences

Department of Management Information Systems

Management Information Systems Program

The study collected customer complaints directed against Turkey's three big technology companies from an online complaint platform, sikayetvar.com, automatically categorized and analysed. In the application part of the study, estimations were carried out after the training and testing process by using Logistic Regression, Linear Support Vector Machines, and Stochastic Gradient Descent classification algorithms, which are machine learning classification algorithms. The accuracy values comparison of three different machine learning classification algorithms in the study revealed that the most effective algorithm was the Logistic Regression algorithm, with a success rate of 80%. The study aimed to determine the most complained companies and their most complained categories using the Logistic Regression classification algorithm, which achieved the highest estimation success among the classification algorithms utilized in processing the complaint data of the three big technology companies.

The application work was conducted using the Python programming language. The estimations were made on the four different categories that customer complaints were directed to: Return/Exchange and Refund, Delivery/Cargo, Customer Relations and Services, and Warranty and Service. The findings were analysed and shared with their graphics.

Keywords: Machine Learning, Natural Language Processing, Text Classification, Logistic Regression, Customer Complaints

YAPAY ZEKA TEKNİKLERİ İLE MÜŞTERİ ŞİKAYETLERİNİN OTOMATİK KATEGORİLERE AYRILMASI

İÇİNDEKİLER

TEZ ONAY SAYFASI	ii
YEMİN SAYFASI	iii
ÖZET	iv
ABSTRACT	vi
İÇİNDEKİLER	vii
TABLOLAR LİSTESİ	x
ŞEKİLLER LİSTESİ	xi
EKLER LİSTESİ	xii
GİRİŞ	1

BİRİNCİ BÖLÜM
LİTERATÜR TARAMASI

1.1 YORUM VEYA TALEPLERİN SINIFLANDIRILMASI	3
1.2 TAHMİNLEME	4
1.3 E-TİCARETE YÖNELİK METİN SINIFLANDIRMA	5
1.4 İSTENMEYEN MESAJ VEYA YORUMLARIN TESPİTİ	6
1.5 HABER SINIFLANDIRMA	6
1.6 DİĞER METİN SINIFLANDIRMA ÇALIŞMALARI	7

İKİNCİ BÖLÜM
KULLANILAN TEKNOLOJİLER

2.1 MAKİNE ÖĞRENMESİ	9
2.2 DENETİMSİZ ÖĞRENME	13
2.3 YARI DENETİMLİ ÖĞRENME	14

2.4 DENETİMLİ ÖĞRENME 15
2.5 MAKİNE ÖĞRENMESİ ALGORİTMALARI 16
2.6 DOĞAL DİL İŞLEME 18
2.7 METİN SINIFLANDIRMA 21
2.7.1 Metin Sınıflandırmanın Aşamaları 22
2.7.2 Metin Sınıflandırmanın Yaygın Kullanılan Örnekleri 23

ÜÇÜNCÜ BÖLÜM
KULLANILAN ARAÇLAR

3.1 PANDAS 25
3.2 MATPLOTLIB 26
3.3 NUMPY 27
3.4 SCIKIT-LEARN 28
3.5 TURKISHSTEMMER 29

DÖRDÜNCÜ BÖLÜM
MÜŞTERİ YORUMLARININ MAKİNE ÖĞRENMESİ YOLU İLE SINIFLANDIRILMASI UYGULAMASI

4.1 ÇALIŞMANIN AMACI 30
4.2 VERİ ELDESİ 34
4.3 VERİ ÖN İŞLEME 34
4.4 EĞİTİM VE TEST VERİLERİ 35
4.5 VERİNİN SAYISALLAŞTIRILMASI VE SIK GEÇEN KELİMELERİN TESPİTİ 36
4.6 SINIFLANDIRMA ALGORİTMALARI İLE TAHMİNLEME 36
4.6.1 Lojistik Regresyon 36
4.6.2 Olasılıksal Dereceli Azalma 37
4.6.3 Doğrusal Destek Vektör Makineleri 38
4.7 SINIFLANDIRMA ALGORİTMALARININ SONUÇLARININ KARŞILAŞTIRILMASI 40

4.8 ETİKETSİZ VERİLERİN TAHMİNLENMESİ 40

SONUÇ 45
KAYNAKÇA 51
EKLER

TABLOLAR LİSTESİ

Tablo 1: Müşteri Şikayetlerinin Kategori Bazında Örnek Gösterimi s. 35
Tablo 2: A, B ve C Şirketine Ait Tahminlenen Şikayetler ve Oranları s. 42

ŞEKİLLER LİSTESİ

Şekil 1: Yapay Zekâ ve Makine Öğrenmesi İlişkisi s. 9
Şekil 2: Çalışma Akış Şeması s. 33
Şekil 3: Lineer Ayrılabilme Durumunda Optimum Ayırıcı Aşırı Düzlem s. 39
Şekil 4: A Şirketi Tahminlenen Şikâyet Oran Grafiği s. 41
Şekil 5: B Şirketi Tahminlenen Şikâyet Oran Grafiği s. 41
Şekil 6: C Şirketi Tahminlenen Şikâyet Oran Grafiği s. 42
Şekil 7: A Şirketi Aldığı Aylık Şikâyetler ve Şikâyet Kategorileri s. 43
Şekil 8: B Şirketi Aldığı Aylık Şikâyetler ve Şikâyet Kategorileri s. 43
Şekil 9: C Şirketi Aldığı Aylık Şikâyetler ve Şikâyet Kategorileri s. 44
Şekil 10: Koronavirüs Vakası Sonrası A, B ve C Şirketlerine Ait Toplam Şikâyet Sayıları s. 45

EKLER LİSTESİ

EK 1: Kodlar ek s.1

GİRİŞ

Metin sınıflandırma, bir metnin önceden belirlenmiş kategorilerden veya sınıflardan hangisine ait olduğunun belirlenmesidir. Yapay zekâ teknikleri ile metin sınıflandırma ise metnin hangi sınıfa ait olduğunun otomatik olarak bilgisayar tarafından bulunmasıdır. Sınıflandırma yöntemi, ikili veya çoklu olarak ikiye ayrılabilir. İkili sınıflandırma sonucunda sınıfa ait olup olmadığını evet veya hayır cevabına göre değerlendirebiliriz. Tespit etmek istenilen sınıf birden fazla sınıfa sahip ise çoklu sınıflandırmadır.

Şirketler için verinin değerlendirilmesi çok önemlidir. Verilerden analizler ve çıkan sonuçlara göre planlamalar yapmak karlılığı olumlu anlamda etkileyecektir. Gelişen teknoloji ile verileri analiz etmek kolaylaşmış ve önemi artmıştır. İnternetin kullanımının artması ile birlikte veriye ulaşım gittikçe kolaylaşmaktadır. Bu artış, ulaşılan verilerin ve verilerdeki metinlerin kategorilere ayrılması ihtiyacını arttırmıştır.

Metin verilerinin bazen büyük olması otomatik metin sınıflandırma ihtiyacını oluşturmuştur ve bu noktada otomatik sınıflandırma verimlilik ve maliyet konusunda daha iyi çözümler sunmaktadır. Otomatik metin sınıflandırma yapılmasındaki amaç, geçmişe yönelik veriler ile modeli oluşturduktan sonra gelecek verilerin otomatik olarak kurulan model ile sınıflandırılmasıdır. Özellikle veri elde etme oranı yüksek olan şirketler için otomatik sınıflandırma yöntemleri kullanması çok önemlidir.

Gelişen teknoloji ve küreselleşen dünyada insanlar hizmet veya ürün satın almadan önce internetin de kullanımının artması ile birlikte internet üzerinden araştırmalarını kolaylıkla yapabilmektedir. Satın alma eylemi gerçekleştirildikten sonra ise ürün veya hizmet ile ilgili düşüncelerini ve şikayetlerini kolaylık ile paylaşabilmektedir. Bu durum, şirketlere kendilerini geliştirebilmeleri ve hatalarını düzeltebilmeleri için çok büyük faydalar sağlamaktadır. Müşterilerin çevrimiçi mecralarda paylaştığı her yorum şirket için önemli bir veridir. Bu veri topluluğu ile analizler yapıp sonuçlarını değerlendirmek çok önemlidir.

Şikayetler, şirketler açısından şikâyeti oluşturan müşteriye karşı yapılan hatayı telafi edebilme imkânı sunar ve alıcının şirkete karşı olan güvenini artırır. Aksi durumda şikayetleri dikkate almayan bir şirket müşterilerini kaybedecek ve aldığı şikâyet sayısı artacağı için marka değerinin düşmesine sebep olacaktır.

Bundan daha önemlisi, müşteri memnuniyetinin veya müşteri memnuniyetsizliğinin tür ve nedenleri, şirketlerin kendi organizasyon yapılarındaki eksiklik ve aksaklıkları belirleme açısından büyük önem arz eder. Memnuniyetsizliği ifade eden bir şikâyet grubundaki artış, şikâyete konu olan şirket organizasyonu ve satış politikasındaki hataları belirlemek açısından önemli ipuçları ortaya koyabilir. Örneğin, teslimat ve kargo ile ilgili şikayetler, şirketin satış ve dağıtım kanallarının yeniden değerlendirilmesini; müşteri ilişkileri ve hizmetlerindeki şikayetler, şirketin satış elemanları üzerinde yeni eğitim programlarının düzenlemesini gerekli kılabilir.

Bu çalışma kapsamında üç büyük teknoloji firmalarının şikayetleri incelenmiştir. Şikayetler online şikâyet platformları arasında en yaygın olarak kullanılan şikayetvar.com sitesinden alınmıştır. Şikayetler, iade/değişim ve geri ödeme, teslimat/kargo, müşteri ilişkileri ve hizmetleri, garanti ve servis olmak üzere dört kategori olarak incelenmiştir.

Çalışmanın amacı, üç büyük teknoloji firması olan, A firması, B firması ve C firmasının şikâyet verilerini makine öğrenmesi sınıflandırma algoritmaları kullanılarak şikâyet verilerinin dört kategoriden hangisine ait olduğunu tespit edip, firmalar arası kıyaslamalar yapmaktır. Firmaların en çok hangi kategori veya kategorilerde şikâyet aldığını otomatik tespit eden makine öğrenmesi modelleri kurulduktan sonra firmaların şikâyet konuları ve oranlarını kıyaslamak amaçlanmıştır.

Çalışmada makine öğrenmesi sınıflandırma algoritmaları ile modeller kurulmuştur. 2020 yılına ait dört şirketin müşteri şikayetleri ile eğitim ve test süreci gerçekleştirdikten sonra üç farklı sınıflandırma algoritması ile yeni gelen veriler analiz edilmiştir. Teknoloji firmaları arasında sınıflandırma kıyaslamaları yapılması ile birlikte lojistik regresyon, olasılıksal dereceli azalma ve doğrusal destek vektör makineleri algoritmaları arasında da çıkan sonuçlara göre hangisinin tahminleme doğruluğunun daha yüksek olduğunun karşılaştırılması yapılmıştır.

Çalışmada yapılan analiz ve raporlamarın hedeflenen katkısı firmalar için karar vermeye destek sağlamak ve elde ettikleri sonuçlar ile ilgili iyileştirme planları yapmasına katkı sağlamaktır.

BİRİNCİ BÖLÜM
LİTERATÜR TARAMASI

1.1. YORUM VEYA TALEPLERİN SINIFLANDIRILMASI

Yıldız çalışmasında; metin madenciliği çalışmalarında sıkça yapılan anahtar kelime tespiti çalışması yapmıştır. Bu çalışmada veri seti olarak bir üniversiteye ait şikâyet, teşekkür, görüş ve öneri mesajlarının yazıldığı sitenin verilerini kullanmıştır. Veri setinde 3961 mesaj bulunmaktadır. Kullanılan 3961 mesaja veri ön işleme yapılmıştır. Tf-idf ve ki-kare istatistik algoritması ile anahtar kelimeler belirlenmiştir (Yıldız, 2016: 29-50).

Göker ve Tekedere çalışmalarında; FATİH projesinin internet ortamındakini yorumlarını otomatik tespitini ve yorumlarının sınıflandırılmasını yapmışlardır. Yapısal olmayan verinin veri önişleme yöntemi ile yapısal hale getirmişlerdir. 444 yorum bulunduran metin dosyasındaki metinler vektör olarak temsil edilerek sınıflandırma algoritmaları kullanılarak modelin başarısı ölçülmüştür (Göker ve Tekedere, 2017: 291-299).

Aytekin, Sütçü ve Özfidan çalışmalarında; bir işletmenin veri tabanında yer alan müşteri yorumlarını otomatik olarak şikâyet-talep-teşekkür olarak üç sınıfa ayıracak etiketler elde etmiştir. Çalışmada karar ağaçları algoritması modeli kullanılmıştır. Çalışma sonucunda sınıflandırma modellerinin başarısını ölçmek için kesinlik (precision) kriteri kullanılmıştır. Sonuç olarak %60'lık bir başarı elde edilmiştir (Aytekin ve diğerleri, 2018: 783-792).

Arslan, Kaynar ve Şahin çalışmalarında; müşteri talepleri öznitelik çıkarma yöntemleriyle otomatik olarak sınıflandırıldıktan sonra ve taleplerin ilgili birimlere daha doğru iletilmesini sağlayan bir yöntem geliştirmişlerdir. Detay danışmanlık firmasının mevcut yardım masası platformu üzerinden müşteri talepleri toplanmıştır. Bu çalışma BoW ve Doc2Vec öznitelik çıkarma modelleri kullanılmıştır. Veri öznitelik çıkarımından sonra MLP sınıflandırıcısı ile sınıflandırılmıştır. Çalışmanın sonucunda BoW modelinde daha başarılı sonuç elde edildiği tespit edilmiştir (Arslan ve diğerleri, 2019: 1-4).

Hoşgör ve Tosun çalışmalarında; İstanbul'da bir üniversitenin hastanesinin, hasta yakınları tarafından yapılan şikayetlerini incelenmişlerdir. Hastane şikayetlerini sikayetvar.com internet sitesinden elde etmişlerdir. Çalışmada Reader vd. (2014) şikâyet taksonomisi ile klinik, yönetim, ilişkiler olmak üzere üç ana başlık altında şikayetler sınıflandırılmıştır. Microsoft Excel paket programı ile verilerin tanımlayıcı istatistikleri (n, %ortalama) hesaplanmıştır. Çalışma sonucunda en fazla şikâyet alan konu ve buna ilişkin bilgiler bulunmuştur (Hoşgör ve Tosun, 2019: 239-252).

Tekin ve Tunalı çalışmalarında; müşterilerin taleplerini metin sınıflandırma yöntemi ile sınıflandırdıktan sonra taleplerin önem derecesini tahmin eden bir yöntem geliştirmişlerdir. Kullanılan veri seti kurumsal bir şirketin müşteri taleplerinden oluşmaktadır. Çalışmada kullanılan algoritmalar; Naive Bayes, Naive Bayes Multinomial, Sequential Minimal Optimization, Random Forest ve Rotation Forest algoritmalarıdır. Kullanılan algoritmaların başarısını ölçmek için Kesinlik (Precision), Anımsama (Recall), F-Skoru (F-Measure) ve ROC Alanı (Receiver Operating CharacteristicArea) başarı ölçme yöntemleri kullanılmıştır. Çalışma sonucunda algoritma başarıları karşılaştırıldığında Random Forest algoritması genelde çoğu koşulda iyi sonuçlar vermiştir (Tekin ve Tunalı, 2019: 615-620).

Büyükeke, Sökmen ve Gencer çalışmalarında; veri seti olarak Tripadvisor platformundan elde ettikleri müşteri yorumlarını kullanarak yorumlardan duygu analizi yapmışlardır. Lojistik Regresyon, Destek Vektör Makinesi ve Naive Bayes makine öğrenmesi algoritmalarını kullanarak analizlerini gerçekleştirmişlerdir. Yorumları analiz ederken olumlu veya olumsuz olarak iki ayrı kategoride sonuçlar elde etmişlerdir ve sonuç olarak %80 olumlu, %20 olumsuz olmak üzeri çıktı almışlardır (Büyükeke ve diğerleri, 2020: 322-335).

1.2. TAHMİNLEME

Lykourentzou, Giannoukos, Nikolopoulos, Mpardis ve Loumos çalışmalarında; e-öğrenme kurslarına katılan öğrencilerin kursu bırakıp bırakmayacağını tahmin eden bir çalışma gerçekleştirmişlerdir. Çalışmalarında kullandıkları algoritmalar ileri beslemeli sinir ağları, destek vektör makineleri ve olasılıklı topluluk basitleştirilmiş bulanık ARTMAP'tır. Çalışma iki sınıflı bir

sınıflandırma problemidir. Çalışma sonucunda %84'e kadar ulaşan ve genellikle %58-%63'lük sınıflandırma doğruluk oranlarına ulaşılmıştır (Lykourentzou ve diğerleri, 2009: 950-965).

Barboza, Kimura ve Altman çalışmalarında; olaydan bir yıl önce iflası tahmin etmek için makine öğrenmesi ile bir yöntem geliştirmişlerdir. Çalışmada kullanılan teknikler; bagging, boosting, rassal orman, destek vektör makineleri, yapay sinir ağları, lojistik regresyon ve modifiye dağıtık algoritmasıdır. Çalışma sonucunda en iyi sonucu veren modeller Rassal Orman ve Bagging'tir (Barboza, 2017: 405-417).

Bilgin çalışmasında; metin sınıflandırma yöntemi ile yazarı bilinmeyen bir dökümanın yazarını tahmin etmeye çalışmıştır. Diğer metin sınıflandırma yöntemi ile yazar tahmini çalışmalarından farklı olarak, ön işleme aşamasından sonra elde edilen metinleri LZW algoritması ile kayıpsız sıkıştırarak özellik sayısını azaltmak ve bunun başarıda yaptığı değişikliği araştırmaktır. Çalışma, WEKA (Waikato Environment for Knowledge Analysis), açık kaynak kodlu programı ile yapılmıştır. Çalışma sonucunda kullanılan yöntemin sadece yazar tanıma değil tüm metin sınıflandırma çalışmalarında uygulanabileceği ve başarıyı arttırdığı tespit edilmiştir (Bilgin, 2019: 125-136).

1.3. E-TİCARETE YÖNELİK METİN SINIFLANDIRMA

Kaşıkçı ve Gökçen çalışmalarında; metin madenciliği tekniği ile internet sayfalarının e-ticaret sitesi olup olmadığını tespit eden bir yöntem geliştirmişlerdir. Çalışmada k-NN En-Yakın-Komşu ve Naïve Bayes sınıflandırma algoritmaları kullanılmıştır. Çalışma sonucunda Naive Bayes algoritması daha iyi sonuç vermiştir. En iyi sonucu veren Naive Bayes algoritmasına Java programlama dili ile masaüstü uygulaması yazılmıştır (Kaşıkçı ve Gökçen, 2014: 1).

Kazan ve Karakoca çalışmalarında; bir e-ticaret sitesinden aldıkları ürün bilgileri etiketleri ile veri seti elde etmişlerdir. Elde ettikleri veriyi 9 ayrı kategoriye bölmek istemişlerdir ve sonucunda doğru tahmin yapılması için makine öğrenmesi modelleri kullanılmıştır. Kullanılan makine öğrenmesi algoritmaları Random Forest, Karar Ağacı, Multinominal Naive Bayes (Multinominal NB), Lojistik Regresyon, Destek Vektör Makineleri (DVM) ve Yapay Sinir Ağları (YSA) sınıflandırma algoritmalarıdır. Çalışma sonucunda DVM ve MLP sınıflandırma algoritmalarının

diğerlerine göre %97'lik bir oranla daha başarılı sonuç elde edilmiştir (Kazan ve Karakoca, 2019: 18-27).

1.4. İSTENMEYEN MESAJ VEYA YORUMLARIN TESPİTİ

Tretyakov çalışmasında; makine öğrenmesi algoritmaları ile e-posta spam sorununa filtreleme çözümü sunmuştur. Kullandığı sınıflandırıcılar; Naive Bayes, k en yakın komşu (k-NN), yapay sinir ağları, destek vektör makineleridir. Veri seti 481'i spam olan 1099 iletiden oluşur. Çalışmanın sonucunda en iyi sonucu perceptron (algılayıcı) algoritması vermiştir (Tretyakov, 2004: 60-79).

Ahmed, Traore ve Saad çalışmalarında; metin sınıflandırması tekniği ile sahte haberler ve aynı zamanda görüş spam'i olarak da bilinen sahte incelemeler (sahte yorumlar) gibi aldatıcı içerikleri filtreleyen bir çalışma gerçekleştirmişlerdir. Olasılıksal dereceli azalma (SGD), destek vektör makineleri (SVM), lagrangian destek vektör makineleri (LSVM), parçalama algoritması (LR), k-en yakın komşu algoritması (KNN) ve karar ağaçları (DT) algoritmalarını kullanarak bu sınıflandırıcıların uygulamalarını Python doğal dil araç setinden (NLTK) kullanılmıştır. Çalışmanın sonucunda %87'lik bir oranla LSVM algoritması en iyi sonucu vermiştir (Ahmed ve diğerleri, 2017: 1).

Uysal, Günal, Ergin ve Günal çalışmalarında; Basit Bayes ve k-En Yakın Komşu makine öğrenmesi algoritmalarını kullanarak cep telefonuna gelen, istenmeyen SMS mesaj süzgeci yöntemini geliştirmişlerdir. Çalışma sonucunda Android işletim sisteme sahip telefonlarda çalışan mobil bir uygulama geliştirilmiştir (Uysal ve diğerleri, 2020: 1-4).

1.5. HABER SINIFLANDIRMA

Amasyalı ve Yıldırım çalışmalarında; gazetelerin web sayfalarında bulunan haber metinleri 5 farklı kategoride otomatik sınıflandırılmıştır. Naive Bayes ve yapay sinir ağlarının 2 türü olan MLP ve LVQ algoritmaları kullanılmıştır. Çalışmanın sonucunda LVQ ve Naive Bayes metodları en yüksek sonuçları vermiştir (Amasyalı ve Yıldırım, 2004: 224-226).

Chen, Yao ve Yang çalışmalarında; Sina Haber web sitesinden aldıkları metinler ile kısa metin sınıflandırma çalışması yapmışlardır. Çalışmada K-En Yakın Komşu algoritması ve Lineer Diskriminant Analizi (LDA) algoritmaları kullanılmıştır. Eğitim ve test veri seti oranı %80'e %20 olarak ayarlanmıştır (Chen ve diğerleri, 2016: 749-753).

Başkaya ve Aydın çalışmalarında; ekonomi, politika, spor ve sağlık olarak dört farklı kategoriden oluşan haber metinlerini metin madenciliği tekniği ile sınıflandırmışlardır. Toplam 80 tane haber metninden oluşan veri seti hazırlanmıştır, hazırlanan veri setine ön işleme uygulanmıştır, kelime kökü ve kelime seviye n-gram olmak üzere 2 tür öznitelik tekniği kullanmışlardır ve terim ağırlıklandırma yaparak metinleri sınıflandırmışlardır. Çalışma yapılırken WEKA açık kaynak kodlu program kullanılmıştır. Çalışmanın sonucunda kullanılan RO algoritması ile %100'lük bir başarı elde edilmiştir (Başkaya ve Aydın, 2017: 1-5).

Tüfekçi, Uzun, Bektaş ve Sevinç çalışmalarında; bir gazetenin yazarlarına ait köşe yazılarının makine öğrenmesi sınıflandırma algoritmalarından Naive Bayes ve Random Forest ile metin türlerini belirleyen model kurmuşlardır. Çalışma sonucunda Random Forest algoritması ile %95,45 doğruluk oranı ile daha başarılı bir sonuç elde etmişlerdir (Tüfekçi ve diğerleri, 2019: 1-4).

1.6. DİĞER METİN SINIFLANDIRMA ÇALIŞMALARI

Ikonomakis, Kotsiantis ve Tampakas çalışmalarında; dijital formattaki belgelere makine öğrenmesi yöntemi ile metin sınıflandırma yöntemini ele almışlardır. Değerlendirilen algoritmalar; karar ağaçları, Naive-Bayes, kural indüksiyon, sinir ağları, en yakın k komşu ve destek vektör makineleridir (Ikonomakis, 2005: 966-974).

Lilleberg, Zhu ve Zhang çalışmalarında; Word2vec yöntemi ile kelime ve ifadeleri bir vektör gösterimine dönüştürerek, metin madenciliği çalışmasına katkı sağlamışlardır. Çalışma sonucunda tf-idf ve word2vec'in birlikte tf-idf'den daha iyi performans sergileyebileceğini göstermiştir (Lilleberg ve diğerleri, 2015: 136-140).

Xuan, Jiang, Ren, Yan ve Luo çalışmalarında; otomatik hata (bug) triyajını yarı denetimli makine öğrenmesi yöntemi yapmışlardır. Çalışmayı yaparken kullandıkları makine öğrenmesi algoritması Naive Bayes'tir. Çalışma sonucunda Naive Bayes'in

doğruluğu %11 ila %43'tür, bu oranın hata triyajı için tatmin edici bir sonuç olmadığı tespit edilmiştir (Xuan ve diğerleri, 2017: 1-6).

Hartmann, Huppertz, Schamp ve Heitmann çalışmalarında; otomatik metin sınıflandırma yöntemlerini karşılaştırmışlardır. Çalışmada, büyük sosyal medya ağlarından, çeşitli örnekleri ve dilleri kapsayan 41 sosyal medya veri kümesinde bu tür on yaklaşımın (beş sözlük tabanlı, beş makine öğrenme algoritması) performansını karşılaştırmaktadır. Çalışmada ratgele orman (RF) ve Naive Bayes algoritmaları kullanılmıştır. Çalışmanın sonucunda, rassal orman daha iyi sonuç vermiştir (Hartmann, 2019: 20-38).

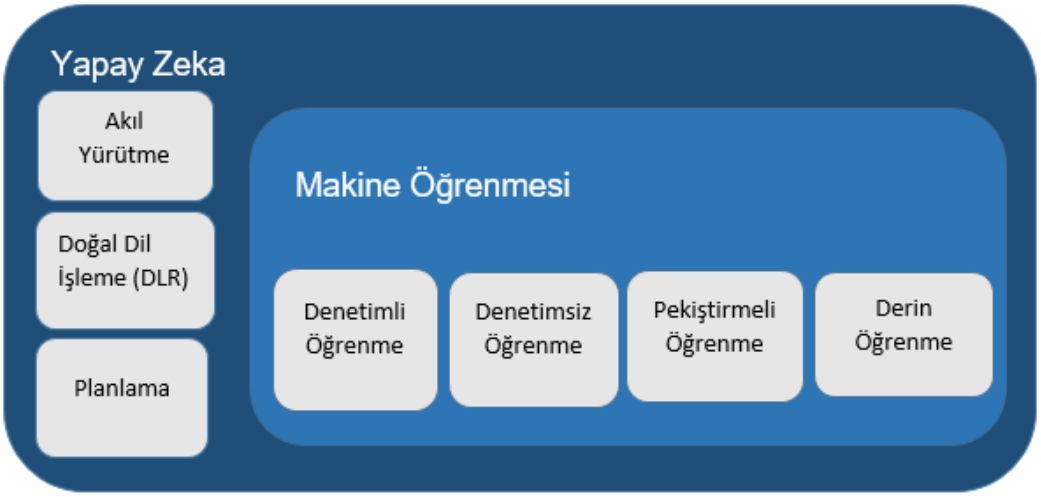
İKİNCİ BÖLÜM
KULLANILAN TEKNOLOJİLER

2.1. MAKİNE ÖĞRENMESİ

Günümüzde verinin artması, verinin artması ile birlikte verinin analiz edilmesi ihtiyacının sonucunda makine öğrenmesi çok büyük önem kazanmıştır. Veri miktarının büyümesi ile veriyi analiz etmenin zorluğu artmaktadır. Küreselleşen ve rekabetin günden güne büyük ölçüde arttığı dünyada veri analizi yapmanın ve bu analizler sonucu rekabette öne geçmenin önemi büyüktür. Makine öğrenmesi insan zihninin yapamayacağı veya çok uzun süreler sonucunda yapacağı işlemleri kolaylıkla yapabilir.

Makine öğrenmesi ile yapay zekanın ilişkisi düşünüldüğünde; makine öğrenmesi yapay zekanın bir alt dalıdır ve bilgisayarlar insan gibi düşünür ve insanın yapabileceği görevleri tek başına yapabilmekle birlikte insan düşüncesini taklit edecek şekilde eğitilebilir.

Şekil 1: Yapay Zekâ ve Makine Öğrenmesinin İlişkisi



Kaynak: (Hurwitz ve Kirsch, 2018: 13)

Makine öğrenimi, bilgisayar bilimi ve mühendisliğinde heyecan verici bir araştırma alanıdır. Makine öğrenmesi, verilen veriden anlamlı kalıpların çıkarılmasını sağlar (Erickson ve diğerleri, 2007: 505).

Makine öğrenmesi, veriden beslenen bir sisteme açık bir şekilde programlanmadan insana benzer bir şekilde otomatik olarak deneyimlerden öğrenebilme yeteneği kazandıran etkili bir yapay zekâ uygulamasıdır. Pek çok alanda başarılı bir şekilde kullanılmaktadır. Ama özellikle belirsizliğin daha çok olduğu özel alanlardaki sorunların çözümünde tercih edilmektedir (Keleş ve diğerleri, 2020: 73).

Makine öğrenmesi sürecinde doğru veri kümesinin seçilmesi çok önemlidir. Bu süreçte büyük veri ile çalışmak şart değildir fakat çalışma sonucunda alacağımız sonuçların doğruluk oranını arttırmada büyük veri önemli bir yere sahiptir. Büyük veri sayesinde veriler sanallaştırılabilir ve ister bulut ister şirket içinde uygun maliyetler ile saklanması mümkündür.

Makine öğrenmesinin asıl amacı insan müdahalesi olmadan otomatik öğrenme yapmaktır. Makine öğrenmesi algoritmaları, gelecekte verilecek olan veriler için tahmin yapmak için eğitilir. Veri miktarı ve yapılan tahmin arasındaki doğruluk oranı birbiri ile doğru orantılı bir şekilde artar veya azalır. Algoritmalar veri ile ilgili birçok örnek gördükten sonra daha önce görmediği olaylar için tahminler yapar ve karar vermede büyük önem taşır.

Makine öğrenmesi, bir probleme ait verileri makine öğrenmesi modelleri ile çözen bilgisayar algoritmasıdır. Mevcut veri seti ve kullanılan algoritma ile oluşturulan model, en yüksek performansı elde etmek için kurulur (Atalay ve Çelik, 2017: 161).

Makine öğrenmesini algoritmalarının kullanımının artması ile fayda sağladığı alanlar; tıp, finans, eğitim gibi birçok alanda olmak üzere kullanılmaya başlanmıştır. Makine öğrenmesinde sıkça yapılan uygulamalar; duygu analizi, spam algılama, borsa tahmini, tıbbi tahminler, zararlı ve güvenli olmayan yazılımların tespiti yaygın bir şekilde üzerinde çalışılan uygulamalardır.

Makine öğrenmesi, bilgisayarı eğitmeye izin veren algoritmalar tasarlamakla ilgilidir. Öğrenme illa bilinci içermez, ancak öğrenme, verilerdeki istatiksel düzenlilikleri veya diğer kalıpları bulma meselesidir. Bu nedenle, birçok makine öğrenimi algoritması, insanın bir öğrenme görevine nasıl yaklaşabileceğine neredeyse hiç benzemeyecektir (Zhang, 2010: 19).

Tipik olarak, makine öğrenimi metodolojileri, bir şeyler öğrenmek amacıyla bir öğrenme sürecini içerir. Makine öğrenimindeki veriler bir dizi örnekten oluşur.

Genelde, bireysel bir örnek, özellikler veya değişkenler olarak da bilinen bir dizi özellik tarafından açıklanır. Bir özellik nominal (numaralandırma), ikili (yani, 0 veya 1), sıralı (örneğin, A + veya B−) veya sayısal (tamsayı, gerçek sayı vb.) olabilir. Makine öğrenmesi, modelinin belirli bir görevdeki performansı, bir performansla ölçülür (Liakos ve diğerleri, 2018: 3).

Makine öğrenmesi, işletmelerde veya bireysel akademik çalışmalarda yaygın olarak kullanılmaktadır. Makine öğrenmesinin kullanılmasının nedenleri aşağıdaki maddelerde verilmektedir:

- İnsan zekasının ve gücünün yapamayacağı veya çok zorlanacağı türden hesaplamalar ve analizleri gerçekleştirmek,
- İnsan iş gücü maliyetlerini ortadan kaldırmak veya azaltmak,
- Daha hızlı ve verimli analizler veya hesaplamalar yapmak,
- Veri toplamak ve bu verileri yararlı bilgilere dönüştürmek,
- Verileri insanların daha kolay anlamasına olanak sağlamak için kullanılır.

Makine öğrenmesinin verimli bir şekilde kullanılabilmesi için aşağıdaki aşamalara sahip olması gerekmektedir.

1. Hangi problemin çözüleceğinin tespiti,
2. Çözülecek problem ile ilgili gerekli olan verinin elde edilmesi,
3. Veri temizleme ve eksik olan verinin tespitinin yapılıp tamamlanması,
4. Makine öğrenmesi çalışması yapılacak olan konu ile uygun programlama dilinin tespit edilmesi,
5. Kullanılacak olan makine öğrenmesi modelinin veya modellerin seçimi,
6. Makine öğrenmesi modelinin veya modellerinin kurulması,
7. Kullanılan modeller ile verinin eğitilmesi,
8. Kullanılan makine öğrenmesi modelinin test edilmesi.

Makine öğrenmesinin birçok avantajı bulunmaktadır. Bu avantajlar aşağıda verilmektedir.

- İçgörüleri ortaya çıkarır ve verilerden bilgi elde eder.
- Kullanıcı deneyimlerini geliştirir.
- Müşterinin geçmiş davranışlarından gelecekte göstereceği davranışları tahmin eder.
- Maliyetleri ve işgücünü arttırarak verimlilik sağlar.

- Dolandırıcılık risklerini azaltır veya ortadan kaldırır.
- Veri madenciliği konusunda geliştirir.
- Makine öğrenmesi uygulamalarının yapıldığı örnekler;
- Spam e-postaları belirleme
- Öneri sistemleri (Netflix, Youtube veya ürün önerisi yapan e-ticaret siteleri)
- Sağlık sistemlerinde hastalık tespiti
- Güvenlik gerektiren konularda sahtekarlık tespiti
- Sürücüsüz araçlar
- Yüz tanıma
- Müşteri yorumlarını sınıflandırma

İnsanlar analizler sırasında veya verilerden gelecek için tahminlerde bulunurken veya sınıflandırma yapmaya çalışırken hata yapmaya eğilimlidir. Bu onların belirli sorunlara çözüm bulmalarını zorlaştırır. Makine öğrenimi genellikle bu sorunlara başarıyla uygulanabilir, sistemlerin verimliliğini ve makinelerin tasarımlarını iyileştirir (Kotsiantis ve diğerleri, 2007: 6).

Makine öğrenmesi, bir modelin istatiksel modellerde olduğu gibi verilere dayalı deneyimlerden otomatik olarak öğrenebildiği bir çalışma dalıdır. Bir dönem boyunca ve daha fazla veriyle, model tahminleri daha iyi hale gelecektir (Dangetti, 2017: 8).

Makine öğrenmesi belirli aşamalardan oluşan bir süreçtir. Bu süreçlerin bütününe makine öğrenmesi terminolojisi denilmektedir. Makine öğrenmesi terminolojisi aşağıdaki gibidir;

- **Gözlemler:** Öğrenmek için olan veri parçaları.
- **Özellikler:** Gözlemi ifade eden veriler.
- **Etiketler:** Gözlemlerin kategorileri.
- **Eğitim verisi:** Algoritmaya öğrenmesi için verilen veri setidir.
- **Test verisi:** Algoritmanın doğruluğunu test eden veri setidir.

Makine öğrenmesi, günümüzde birçok sektörde kullanılmaktadır. Bu sektörlerden bazıları;

- Otomotiv sektörü (sürücüsüz araçlar)
- Fen bilimleri ve tıp (hastalık tanı teşhisi, kimyasal yapı teşhisi)

- Pazarlama (duygu analizi)
- Endüstri (ürünlerin bakım kestirimleri)
- Eğlence (spor performansı değerlendirme)
- Tarım (bitki sağlık durumu)
- Bankacılık ve finans (borsa tahmini)
- Ulaşım (trafik tahmini)

Makine öğrenmesi verilen örneklerdeki gibi birçok alanda çözümler sunmaktadır.

Makine öğrenmesi algoritmaları, denetimli öğrenme, denetimsiz öğrenme ve yarı denetimli öğrenme olmak üzere üç gruba ayrılmaktadır. Denetimli öğrenme; sisteme hem eğitim hem de test verisi yüklendikten sonra veri setinde her verinin etiketlenmesinden dolayı girdi veri seti ile çıktı veri seti arasında ilişki kurulmasıdır (Aydın ve Özkul, 2015:38). Denetimsiz öğrenmede sisteme herhangi bir müdahale yoktur. Sadece girdi verileri sisteme verilir ancak herhangi bir işaretleme yapılmaz. Sistem otomatik olarak ilişki ağını ortaya koymaya çalışır (Alpaydın, 2010:11). Yarı denetimli öğrenme ise hem etiketli hem etiketsiz verilerin birlikte kullanılması ile oluşan bir makine öğrenmesi çeşididir.

Denetimli öğrenme ile yapılacak işlemler;

- Gözlemlenmiş değerlerden yeni çıkarımlar yaparak yeni gözlemin hangi kategoriye ait olduğunu tespit ederek sınıflandırma,
- Değişkenler arasındaki ilişkiyi tahmin ederek regresyon,
- Geçmiş ve şimdiki veriler ile gelecek için tahminleme.
- Denetimsiz öğrenme ile yapılacak işlemler;
- Benzer verileri gruplama yaparak kümeleme,
- Değişken sayısını azaltarak boyut küçültme.

şeklinde sıralanmaktadır.

2.2. DENETİMSİZ ÖĞRENME

Denetimsiz öğrenme, etiketli veriler olmadan, girdi veri setleri üzerinden çıkarımlar yapmak için kullanılan bir makine öğrenme yöntemidir. Denetimsiz

öğrenme, kümeleme, aykırılık tespiti, sinir ağları ve gizli değişken modeller öğrenilmesi gibi yaklaşımları içermektedir (Polat, 2017: 444).

Denetimsiz öğrenmede, çıktı değerinin olmadığı sadece girdi değerlerinden üzerinden işlem yapılarak girdilerin gruplanması, başka bir deyişle kümelenmesi (clustering) sağlanmaktadır (Balaban ve Kartal, 2015: 33).

Denetimsiz öğrenme problemlerinde, veri setinde sınıflara ayrılmamış veriler ile çalışılır. Denetimsiz öğrenme ile veri setindeki etiketlenmemiş veriler arasındaki ilişkileri ortaya çıkartmak amaçlanır. Denetimsiz öğrenmede kümeleme (clustering) ve boyut azaltma (dimensionality reduction) olmak üzere iki farklı görevi yapabilir (Uğuz, 2021: 91).

Denetimsiz öğrenme kullanılan birçok uygulama mevcuttur. Bunlara örnek,

- Sosyal medyada grup veya sayfa önerisi,
- Youtube video önerisi,
- E-ticaret web sayfalarında ürün önerisi

verilebilmektedir.

Denetimsiz öğrenmede kümeleme ve temel bileşenler analizi olmak üzere iki yöntem ile uygulanabilmektedir.

- **Kümeleme:** Denetimsiz öğrenme konusunda kümeleme oldukça önemlidir. Veriler, yakınlık, uzaklık veya benzerliğe göre değerlendirilir. Kümelemeye en yaygın kullanılan kümeleme yöntemi, K-Ortalamalar'dır. K-ortalama kümeleme yöntemi, eğiticisiz bir yöntemdir (Şengöz ve Özdemir, 2016: 88).
- **Temel Bileşenler Analizi (Principal Component Analysis):** anıTma, sınıflandırma, görüntü sıkıştırma alanlarında kullanılır. Bir değişkenler setinin varyanskovaryans yapısını, bu değişkenlerin doğrusal birleşimleri vasıtasıyla açıklar, boyut indirgenmesi ve yorumlanmasını sağlar (Yıldız ve diğeleri, 2010: 248).

2.3. YARI DENETİMLİ ÖĞRENME

Yarı denetimli öğrenme, eğitim verisindeki etiketlenmemiş verilerinde kullanmasıdır. Genel olarak etiketli olmayan veriler etiketli olanlara göre daha

fazladır. Denetimli öğrenme ve denetimsiz öğrenmenin arasındadır. Denetimsiz öğrenmedeki gibi tamamen etiketlenmemiş verilerle çalışmanın yanı sıra az miktarda etiketlenmiş verinin olması sonucun başarısını arttıracaktır.

Yarı denetimli sınıflandırma teknikleri, denetimsiz ve denetimli arasında kalan ve her ikisinden de yararlanan bir tekniktir. Yaygın olarak kullanılır (Satı, 2018: 864).

2.4 DENETİMLİ ÖĞRENME

Denetimli öğrenme Denetimli Makine Öğrenmesi (SML: Supervised Machine Learning) sistemin etiketli veriler kullanılarak eğitilmesi ile öğrenmenin sağlanmasıdır. Sistem eğitilirken veri setinde bulunan her bir örneğe ait giriş ve çıkışlar verilir (Bilgin, 2017: 683).

Denetimli öğrenme uygulamaları regresyon ve sınıflandırma uygulamaları olmak üzere ikiye ayrılır. Elde edilen çıktının veri türü kategorik olursa sınıflandırma ya da nümerik olursa regresyondur. Sınıflandırma, veri setinde modelin öğrenebilmesi için verilen verilerin sonrasında yeni gözlemleri sınıflandırmasıdır. Modelin eğitilip tahminde kullanılması yani modele sunduğumuz veri setinden elde edilen tecrübe ile gelecekteki veriler için tahmin yapılması sınıflandırmadır. Regresyon ise sınıflandırma ile yapısal olarak benzemektedir. Yine aynı şekilde bir veri seti eğitim için kullanılır ve tahmin yapılması beklenir fakat aradaki fark sınıflandırmada çıkan sonuçların kategorik, regresyon için nümerik yani sayısal sonuçlar olmasıdır.

Denetimli makine öğrenmesinde kullanılan sınıflandırma algoritmaları;

- Doğrusal Sınıflandırıcılar
- Destek Vektör Makineleri
- Karar Ağaçları
- Artan Ağaçlar
- Rastgele Ormanlar
- Nöral Ağlar
- En Yakın Komşu

Denetimli makine öğrenmesinde kullanılan regresyon algoritmaları ise

- Linear Regresyon
- Multiple Linear Regresyon

- Polinominal lineer Regresyon
- Karar Ağacı Regresyonu
- Rastgele Orman Regresyonu

olarak araştırmalarda kullanılmaktadır.

Denetimli öğrenme, gerçek dünya uygulamalarında büyük bir başarı oldu. Metin ve Web alanları dahil hemen hemen her alanda kullanılmaktadır. Denetimli öğrenme, makine öğreniminde sınıflandırma veya tümevarımlı öğrenme olarak da adlandırılır. Bu tür bir öğrenme, insanlardan öğrenmeye benzer yeteneğimizi geliştirmek için yeni bilgiler kazanmak için geçmiş deneyimler gerçek dünyadaki görevleri yerine getirmek içindir (Carey ve diğerleri, 2008: 275).

2.5. MAKİNE ÖĞRENMESİ ALGORİTMALARI

Makine öğrenmesi algoritmaları, insanların karmaşık veri kümelerini keşfetmesine, analiz etmesine ve bunlarda anlam bulmasına yardımcı olan kod parçacıklardır. Her algoritma, bir makinenin belirli bir hedefi gerçekleştirmek için izleyebileceği sınırlı ve belirli adım adım ilerleyen yönerge kümesidir (Microsoft Azure, b.t.).

Veri boyutu, kalitesi, verilerden ne elde edilmek istenildiği ve doğruluk gibi faktörler hangi makine öğrenmesi algoritmasının seçileceğinde büyük bir önem taşır. Hangi algoritmanın iyi sonuçlar vereceğini öğrenmek için deneyip her algoritmanın sonuçları karşılaştırılmalıdır.

Bilimsel çalışmalarda en yaygın kullanılan makine öğrenmesi algoritmaları aşağıda verilmektedir:

- **Doğrusal Regresyon Algoritması:** Doğrusal regresyon analizi, bir değişkenin değerini başka bir değişkenin değerine göre tahmin etmek için kullanılır. Tahmin etmek istediğiniz değişken, bağımlı değişken olarak adlandırılır. Diğer değişkenin değerini tahmin etmek için kullandığınız değişken ise bağımsız değişken olarak adlandırılır (IBM, b.t.). Denetimli öğrenme türüne ait bir regresyon algoritmasıdır.
- **Lojistik Regresyon Algoritması:** Lojistik Regresyon Analizinin kullanım amacı, istatistikte kullanılan diğer model yapılandırma teknikleri ile

aynıdır. En az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı ile bağımsız değişkenler arasındaki ilişkiyi tanımlar ve biyolojik olarak kabul edilebilir bir model kurar (Bircan, 2004: 186).

- **Naive Bayes Algoritması:** Naive Bayes algoritması, belirli bir veri kümesindeki değerlerin sıklığını ve kombinasyonlarını sayarak bir olasılık kümesi hesaplayan bir sistemi vardır (Sarıtaş ve Yaşar, 2019: 90).
- **Destek Vektör Makineleri:** Genellikle sınıflandırma problemlerinde kullanılan denetimli öğrenme yöntemidir. İki sınıfa ait verileri birbirinden, en uygun şekilde ayırmak için kullanılır.
- **Karar Ağacı Algoritmaları:** Ağaç tabanlı öğrenme ve sınıflandırma algoritmalarıdır. Karar ağaçları, kurulumu ucuzdur ve kolay yorumlanabilir. Bu sebepler ile yaygın kullanılır (Emel ve Taşkın, 2005: 221-239). Karar ağaçlarının ilk hücrelerine kök denir. Kök hücrelerinin altında düğümler ve karar ağacının en altında yapraklar bulunmaktadır.
- **K-En Yakın Komşu Algoritması:** Makine öğrenim algoritmaları içerisinde en çok bilinen ve kullanılan algoritmalardan biridir. Seçilen özelliğin kendine en yakın olan özellikle arasındaki yakınlığı kullanarak sınıflandırma yapılır (Kılınç ve diğeleri, 2016: 90).
- **Rastgele Orman Algoritmaları:** Karar ağaçları algoritmasını temel alır fakat karar ağaçları algoritmasının aşırı öğrenme (Veri setini ezberleyerek eğitim setindeki durumları ezberler test veri setinde de aynı durumları arar ve yeni veya karmaşık bir veri geldiğinde kötü bir sonuç verir) probleminin önüne geçebilmek için bir ağaç oluşturmak yerine orman oluşturarak ormandaki ağaçları rastgele düzenler. Hem regresyon hem de sınıflandırma için kullanılabilir bir algoritma türüdür.
- **Gradyan Arttırma Algoritması:** Zayıf tahmin modellerini bir araya getirerek karar ağaçlarının oluşturduğu bir modeldir. Regresyon ve sınıflandırma problemleri için kullanılabilen bir makine öğrenmesi algoritmasıdır.
- **K-Ortalamalar Algoritmaları:** Kümeleme, benzerliklerine göre nesneleri gruplamaktır. Kullanımı çok yaygın bir denetimsiz öğrenme algoritmasıdır.

K-ortalamalar algoritmasındaki K değeri küme sayısını belirler. K tane küme oluşturur ve her kümenin merkezi kümedeki değerlerin ortalamasıdır.

2.6. DOĞAL DİL İŞLEME

Bir bilgisayarın insan konuşmasını anlayabilmesi ve insanın anlayabileceği dilde konuşması, insanlarla etkileşimde bulunabilmesi için dilin tüm özelliklerini öğrenmesi gereklidir. Bu nedenle Doğal Dil İşleme alanında bilişimciler dillerin özelliklerini kendi bakış açılarından değerlendirmeye başlamışlardır. Bu bağlamda özellikle şu konularda çalışmalara ağırlık verilmiştir (Adalı, 2020: 4):

- Ses bilimi
- Bürün bilgisi
- Biçim bilimi
- Söz dizimi
- Anlam bilimi
- Söylem bilgisi
- Edim bilimi

Doğal dil, insanların iletişim kurarak hayatta kalabilmeleri için önemli bir gereksinimdir. Konuşma dışında da metinler veya işaretler şeklinde dil örnekleri karşımıza çıkmaktadır. Bu temel gereksinimin insanın hayatını kolaylaştırması adına bilişim teknolojileri ile birleşmesi de doğal dil işlemedir. Doğal dil işleme yapay zeka ve dil biliminin ortak çalışması sonucu ortaya çıkmıştır.

Dil bilimi, Türk Dil Kurumu'nun İnternet'te yayınlanan sözlüğünde "*Dillerin yapısını, gelişmesini, dünyada yayılmasını ve aralarındaki ilişkileri ses, biçim, anlam ve cümle bilgisi bakımından genel veya karşılaştırmalı olarak inceleyen bilim, lisaniyat, lengüistik*." ifadesiyle tanımlanmaktadır (Dil Bilimi, b.t.).

Doğal dil işleme, doğal dilde oluşturulmuş metinler üzerinde otomatik olarak gerçekleştirilen metin ayrıştırma, metin sınıflandırma, bilgi çıkarımı, duygu analizi gibi birçok önemli konuyu bünyesinde barındırmaktadır ve uygulama alanları da gün geçtikçe artmaktadır (Küçük ve Arıcı, 2018: 77).

Doğal dil işleme, yapay zekanın alt dallarındandır. Doğal dil işleme yöntemleri, insanların el ile yapmak zorunda olduğu bazı metin sınıflandırmalarının otomatik olarak yapılabilmesidir ve çok kullanışlı bir yöntemdir (Öztürk ve diğerleri, 2020: 1).

Doğal dil işlemede dört ana teknik kullanılmaktadır; sözdizimi (syntax), semantik (anlamsal), söylem (discourse) ve konuşmadır.

- Sözdizimi, kelimelerin cümle ile dil bilimsel açıdan anlamlı olacak şekilde düzenlenmesidir. Sözdizimi için kullanılan bazı teknikler vardır; Lemmatizasyon (Lemmatization), kelimenin değişik formlarından daha kolay analiz yapmak için tek forma indirgeme tekniği, Morfolojik bölümleme (Morphological segmentation), kelimelerin morfemlerine (ayrı birimlere) ayrılma tekniği, Kelime bölümleme (Word segmentation), uzun metin parçasının kelimelere veya cümlelere bölünme tekniği, Konuşma bölümü etiketleme (Part-of-speech tagging), her sözcüğün konuşmanın bir bölümünü tanımlaması tekniği, Ayrıştırma (Parsing), cümlenin dilbilgisi çözümlemesi tekniği, Cümle kırma (Sentence breaking), cümle sınıflarının uzun bir metine yerleştirilmesi, Kökleşme (stemming), kelimelerinin biçimlerinin değiştirilmesi tekniğidir.
- Semantik (anlamsal), anlam demektir. Metnin ne anlam ifade ettiğini belirtir. Doğal dil işleme açısından zor bir alandır çünkü anlamları analiz etmek zorlayıcıdır. Semantik için kullanılan bazı teknikler vardır; Adlandırılmış varlık tanıma (NER), önceden ayarlanan kategorilere göre yazıyı ayırma yöntemi, Optik karakter tanıma (OCR), bir metni temsil eden karşılığı sistemden karşılık gelen görüntüyle belirleme yöntemi, Doğal dili anlama, bilgisayarın daha rahat anlayabilmesi için metinleri daha açık formlara dönüştürme tekniği, metin içeriğini tanıma, iki metinin içeriklerinin karşılaştırılması örneğin bir metin diğer metnin aynısı veya karşıt anlamlısı mı olduğunun tespiti tekniği, kelimelerde anlam ayrımı yaratma, bir kelimenin diğer anlamları arasından daha anlamlı olanını belirleme tekniğidir.
- Söylem (discourse), çeşitli düzeylerde metinlerden dil yapılarını ortaya çıkarmak için bir Doğal Dil İşleme görevleri paketidir (Joty ve diğerleri, 2019: 1). Söylem analiz konuşma ve metinler aracılığıyla oluşan anlam

ürünleri ile ilgilenen geniş kapsamlı sosyal ve kültürel araştırmalar içinde kullanılan bir araştırma yöntemidir (Çelik ve Ekşi, 2008: 1).

- Konuşma, bir konuşma kaydından metinsel gösterimler çıkarılmasıdır. Konuşma tanıma sürecinin bir alt başlığı ise bir konuşma kaydından metinlerin kelime gruplarına ayrılmasıdır. Diğer bir başlık ise metinden konuşma oluşturmadır. Bu konuda yardıma ihtiyacı olan görme engelliler için yaygın biçimde kullanılmaktadır.

Doğal dil işleme yazılı metin ve ses girdilerine göre ikiye ayrılmaktadır.girdilere göre ikiye ayrılır. Doğal dilin genellikle iki kaynağı bulunmaktadır. Birinin konuşması durumunda ona sesli yanıt verilmesi veya yazılı bir metnin analiz edilmesi olarak düşünülebilir (Şeker, 2015: 15).

Doğal dil işleme, bilgisayarların konuştuğumuz veya yazdıklarımızı anlayıp işimizi kolaylaştırmak veya bize anlamlı geri dönüşler vermek amacıyla kullanılır. Metin veya konuşma gibi yapılandırılmamış doğal dilden oluşan verileri işleyip anlamak için dilbilim teknikleri ve algoritmalardan oluşmuştur.

Doğal dil işleme adımları, kelime yapıların tanımlayıp analiz ederek sözcüksel analiz, kelimelerin arasındaki ilişkiyi ve kelime analinizi yaparak sözdizimsel çözümleme, metindeki kelimelerin tam anlamını sistem üzerinde oluşturup metinin anlamlılığını kontrol ederek anlamsal analiz, cümleler arasındaki anlamların birbiri ile entegrasyonuna bakarak söylem bütünleşmesine ve son olarak oluşan sentezin gerçek bilgilere ihtiyaç duyan dilbilimsel yönlerini üreten pragmatik analizden oluşur (Microsoft Azure, b.t.).

Doğal dil işlemenin aşağıda sayıldığı gibi birçok kullanım alanı mevcuttur:

- Dil çeviri, bir dilden başka bir dile çevirim işlemi örnek olarak Google Translate söylenebilir.
- Ses tanıma, doğal dil işleme tekniği ile ses sinyallerinin anlamlı hallere getirilmesi örnek olarak Siri veya Google Asistan verilebilir.
- Yazım denetimi, bir dilde yazılan bir cümlenin dil bilgisi kuralları açısından kontrol edilmesi.
- Metin özetleme, bir metin veya cümlenin ne anlam ifade ettiğini kısaca belirtmesi.

- Bilgi çıkarılması, bir metinden belli kısımların otomatik tespit edilip alınmasıdır.
- Adlandırılmış varlık tanıma, kişi, yer, tarih veya organizasyon gibi tanımlı olan kategorilerin metinde tespit edip çıkarılma işlemidir.
- Soru cevaplama, makinenin sorulardan anlamlar çıkarıp sorulan soruya uygun cevaplar üretme işlemidir.
- Metin sınıflandırma, bir metin veri setinde geçen kelime veya cümlelerin belirli etiketlere atama işlemidir.

Doğal Dil İşleme, teorik olarak motive edilmiş bir dizi doğal olarak oluşan metinleri analiz etmek ve temsil etmek için hesaplama teknikleri insan benzeri başarmak amacıyla bir veya daha fazla dilbilimsel analiz düzeyinde bir dizi görev veya uygulama için dil işlemedir (Liddy, 2001: 2).

2.7. METİN SINIFLANDIRMA

Bilgisayar ve internet kullanımının artması ile birlikte bilgisayar ortamında gitgide artan bir bilgi ve belge yoğunluğu oluşmuştur. Bilgi ve belge sayısının gitgide artması ile işletmelerin elinde çok sayıda veri oluşmuş ancak verinin değerlendirilmesi başlı başına bir konu haline gelip ortaya veri madenciliği çalışma alanı çıkmıştır. Metin sınıflandırma, veri madenciliğinin alt dalı olan metin madenciliğinin bir çalışma alanıdır. Oluşan verilerin değerlendirilmesi yani verinin bilgiye dönüşme süreci işletmelere kar olarak dönmeye başlamıştır.

Metin sınıflandırma metin madenciliği alt başlıklarından birisidir. Metin madenciliği ile metinler üzerinden bilgi çıkarımı işlemi yapılır. Metin sınıflandırma konusu sınıflandırma methodları ile metinlerden bilgi çıkarımı işlemi yapar. Eğitim verisi kullanarak gelecek veriler için metin sınıflandırma uygulaması gerçekleştirilir.

Metin sınıflandırmanın birçok farklı uygulama alanı vardır. Örneğin kütüphaneye yeni gelen bir kitabın konusunun tespit edilmesi ve uygun konulu kategoriye ait yerin belirlenmesi bir metin sınıflandırma problemidir. Bu işlemi, insanın yapması yerine daha kısa sürede bilgisayarla yapılması durumuna metin sınıflandırma denir (Tantuğ, 2016: 1).

Metin Madenciliği adımları aşağıdaki gibidir:

- **Metin Koleksiyonu Oluşturma:** Ele alınan konularda bilgi erişim sistemlerine dayalı olarak metin koleksiyonu oluşturma sürecidir.
- **Metin Önişleme:** İşaretleme, gövdeleme, sözlük oluşturma ve gereksiz kelimeleri ayıklama, yazım kurallarına uygunluğu tespit etme ve var olan hataları düzeltme gibi metin belgelerinin yapıtaşı olan kelimelerle ilgili işlemleri içeren süreçtir.
- **Veri Madenciliği:** Yapılandırılmış bir formata dönüştürülen metinlerin, geleneksel veri madenciliği teknikleriyle (denetimli veya denetimsiz öğrenim) analizi sürecidir.
- **Değerlendirme ve Yorumlama:** Veri madenciliği yöntemleri ile verileri analizinden elde edilen sonuçların değerlendirilip kullanıcıya uygun ve anlaşılır bir şekilde sunulması işlemidir (Oğuzlar, 2011: 9).

2.7.1 Metin Sınıflandırmanın Aşamaları

Metnin sınıflandırma aşamaları aşağıdaki şekilde sıralanmıştır.

- **Veri Setinin Toplanması:** Sınıflandırma yapılması gereken probleme uygun şekilde veri elde edilmelidir. Veride eksikler var ise bunlar düzeltilmelidir.
- **Veri Ön İşleme Uygulaması:** Veriden doğru ve anlamlı sonuçlar elde edebilmek için metin ön işleme ile daha düzenli ve temiz hale getirilir ve veri uygun bir formatta olmuş olur. Veri ön işlemede uygulanacak işlemler aşağıdaki gibi olur:
  - Noktalama işaretlerinin kaldırılması
  - Metin içindeki gereksiz boşlukların kaldırılması
  - Stopwords olarak adlandırılan gereksiz yani bize bir anlam ifade etmeyen kelimelerin kaldırılması
  - Cümle, kelime, harf veya n-gramlar halinde parçalama (tokenization)
  - Sayısal değerlerin kaldırılması
  - Büyük harfleri küçük harfe dönüştürme
  - Kelime köklerini tespit etme
  - Yazım yanlışlarının düzeltilmesi

- **Özellik Seçimi:** Özellik seçimi, orijinal veri setini temsil edebilecek en iyi altkümenin seçimi işlemidir. Bu işlem, ilgilenilen problem için en faydalı ve en önemli özellikleri seçerek veri kümesindeki özellik sayısını azaltmayı yani veri boyutunu düşürmeyi amaçlar (Budak, 2018: 1).
- **Vektörleştirme:** Metindeki tüm değerlere sınıflandırma işlemi uygulanabilmesi için sayısallaştırılmalıdır. Metindeki verileri bilgisayarın anlayacağı bir hale getirmek gereklidir. Makine öğrenmesi algoritmaları sayısal değerler ile çalışır (Santur, 2020: 3).

2.7.2. Metin Sınıflandırmanın Yaygın Kullanılan Örnekleri

Metin sınıflandırma yaygın olarak kullanılmaktadır. Bu uygulamalarda en sık kullanılanlar aşağıdaki gibidir:

- Spam algılama
- Verilen metnin olumlu ya da olumsuz olduğunu bulan duygu analizi yöntemleri
- Öneri sistemleri
- Dil algılama
- Metinlerin akıcılığının veya zorluğunu tespit etme sistemleri
- Yazar tanıma

Veri madenciliği, bilgi erişimi, web madenciliği, doğal dil işleme ve istatistik metin madenciliğini oluşturan temel alanlardır. Metin sınıflandırma tamamının alt dalında yer alabilir.

Genel olarak, metin sınıflandırma sistemi, uygulanabilecek dört farklı kapsam düzeyi içerir:

- **Belge düzeyi:** Belge düzeyinde, algoritma tam bir belgenin ilgili kategorilerini alır.
- **Paragraf seviyesi:** Paragraf seviyesinde, algoritma tek bir paragrafın ilgili kategorilerini (bir belgenin bir bölümü) elde eder.
- **Cümle düzeyi:** Cümle düzeyinde, tek bir cümlenin ilgili kategorilerini (bir paragrafın bir bölümünü) alır.

- **Alt cümle seviyesi:** Alt cümle seviyesinde, algoritma bir cümle (cümlenin bir kısmı) içindeki ilgili alt ifade kategorilerini elde eder. (Kowsari, 2019: 2)

ÜÇÜNCÜ BÖLÜM
KULLANILAN ARAÇLAR

3.1. PANDAS

Elektronik tablolarda veya veri tabanlarında depolanan veriler gibi tablo verileriyle çalışırken pandas kütüphanesi kullanılır. Pandas, veri keşfetmeye, temizlemeye ve işlemeye yardımcı olur. Pandas'daki bir veri tablosu DataFrame olarak adlandırılır (McKinney, 2021: 3).

Veri ön işleme ve veriyi temizlemek veri analizi konusunda olmazsa olmaz olan ilk basamaklardandır. Veri ön işlemenin uygun ve iyi yapılması veri analizini hızlandıracak ve sonuçların doğruluğunu arttıracaktır. Pandas, veri analizi yapılacak olan veriyi analizi hazır hale getirmeye yarayan önemli bir Python kütüphanesidir.

Pandas ile temel istatistik işlemleri yapılabilir. Pandas'da yapılan temel istatistik işlemleri; minimum, maksimum, mod ve medyan gibi basit istatiksel işlemlerdir. Pandas ile birden çok veri tablosu birleştirilebilir veya veri tabanı içinde sütunlar birleştirilebilir. Pandas ile tarih ve saat ilgilendiren zaman işlemleri kolaylıkla yapılabilir. Pandas ile sadece sayısal veriler ile çalışılmaz aynı zamanda metinsel veriler ile çalışılabilir ve metinsel analizler yapılabilir.

Pandas, 2008 yılında geliştirilmeye başlanmıştır. 2009 yılından beri açık kaynak kodlu olan bir kütüphanedir, 2015 yılında NumFOCUS'un sponsor olduğu bir proje olmuştur.

Pandas, CSV ve metin dosyaları, Microsoft Excel, SQL veri tabanları ve hızlı HDF5 formatındaki verileri okuyup yazabilir, eksik formattaki verileri doldurabilir, veriden grafikler elde edebilir ve matematiksel işlemleri yapabilir.

Pandas, herkes tarafından erişilebilir, kullanıcıların kullanması ve değiştirmesi için ücretsiz, esnek, güçlü, kullanımı kolay ve hızlı bir kütüphanedir. Pandas, pip install pandas şeklinde kurulabilir. Pandas ile çalışmaya başlamak için; import pandas as pd şeklinde içe aktarılır.

Makine öğrenmesi uygulamalarında veri setleri csv, xls, json vb. farklı formatlı dosyalardan elde edilebilir. Farklı formatlı dosya yapıları pandas kütüphanesi aracılığı ile kolaylıkla dataframe yapıları şekline dönüştürülebilirler (Uğuz, 2019: 38).

Pandas, veri işleme aracıdır ve iki temel veri yapısı ile çalışabilir. Çalıştığı veri yapıları; seriler ve DataFrame'lerdir. Seriler iki boyutludur, DataFrame'ler ise iki boyutludur. Seriler; tamsayılar, dizeler, kayan nokta sayıları, Python nesneleri gibi birçok veri türünden birisidir. Serideki her elemanın indis değeri vardır ve ilk elemanın indisi sıfırdır. Pandas ilişkisel veya etiketli veriler ile çalışabilir. Numpy paketi üzerinden geliştirilmiştir ve temel veri yapısı DataFrame'dir. DataFrame'ler satır ve sütunlardan oluşan ve veri analizi işlemlerinde kullanılan yapılardır. Pandas ilişkisel ve etiketli veriler ile çalışan kolay ve hızlı bir Python paketidir.

Pandas veri analizi ve bilimsel hesaplamalar konusunda kullanılacak ise en basit ve en doğru kurulum Anaconda platformunda olmalıdır. Pandas kurulumu; pip install pandas şeklindedir ve import pandas şeklinde içeri aktarılır. Versiyonunu kontrol etmek için ise; pandas.__version__ yazarak kontrol edilebilir.

3.2. MATPLOTLIB

Matplotlib, statik, animasyonlu ve etkileşimli görselleştirmeler oluşturmayı sağlayan bir Python kütüphanesidir. Matplotlib sadece birkaç satır kod ile grafikler oluşturur ve oluşturulan grafiklerde çizgi stilleri, yazı tipi özellikleri özelleştirilebilir (Hunter, 2007: 91).

Makine öğrenmesinde veri analizi yaparken verilerin görselleştirilmesi önemli bir adımdır. Matplotlib verileri görselleştirmeyi sağlar ve 2 veya 3 boyutlu çizimler yapılabilir. Veri görselleştirmeyi, yaptığımız veri analizi sonucunda çıkan karmaşık veriyi düzenleyip daha kolay anlaşılır bir biçime getirmek için yaparız. Soyut bilgilerin somutlaştırılarak ön bilgi alınması için veri görselleştirmesi yapmamız önemlidir.

Matpolotlib kütüphanesinde hazırlanan iki veya üç boyutlu grafiklerin sonuçları PNG, PDF gibi formatlarda kaydedilebilir.

Matplotlib kütüphanesinde yapılacak işlemler; kolaylık sağlaması açısından eşik değeri ile gösterim yapmak mümkündür yani belli bir değerin üstünde bulunan veri kolaylıkla ifade edilebilir, fonksiyonların grafikleri çizilebilir, histogram grafiği, sütun grafiği, eşyükselti grafiği, üç boyutlu yüzey grafiği, üç boyutlu sütun grafiği, üç boyutlu saçılım grafiği, çoklu panel grafikleri, pasta grafiği, kutu grafiği, dal-yaprak

diyagramı, çekirdek yoğunluk tahmini grafiği, çok serili çubuk grafiği, yığın çubuk grafikleri oluşturulabilir.

Veri biliminde verilerin görselleştirilmesi önemli bir yer tutmaktadır. Bu yüzden Python'ın önemli görselleştirme paketlerinden biri olan matplotlib güçlü grafikler çizilmesine olanak tanımaktadır. Bunun dışında seaborn ve plotly paketleride önemli veri görselleştirme paketleri arasındadır.

Veri görselleştirme veri analizi konusunda çok önemli bir yer tutmaktadır. Veri görselleştirmenin iyi olması analizin ve bu analiz ile alınacak kararlarında iyi olması anlamına gelmektedir. Veri görselleştirme yapılar ve ilişkileri daha kolay anlamamızı sağlar.

3.3. NUMPY

Numpy, Python ile bilimsel hesaplamalar yapmak için kullanıllan bir pakettir. Özellikle doğrusal cebir ve fourier dönüşümü uygulamaları ile çok boyutlu dizilerin oluşturulması ve kullanılmasında güçlü fonksiyonlara sahiptir. (Uğuz, 2019: 10).

NumPy'nin ana hedefi homojen çok boyutlu dizidir. Negatif olmayan tamsayılardan oluşan bir demet tarafından indekslenmiş, hepsi aynı türde olan bir elemanlar (genellikle sayılar) tablosudur. NumPy'de boyutlar eksen olarak adlandırılır. NumPy, Python ile sayısal hesaplamayı etkinleştirmeyi amaçlayan açık kaynaklı bir projedir. 2005 yılında Numerical ve Numarray kütüphanelerinin ilk çalışmalarına dayanarak oluşturuldu. NumPy her zaman %100 açık kaynaklı yazılım olacak, herkesin kullanımına açık olacak ve değiştirilmiş BSD lisansının liberal hükümlerine göre piyasaya sürülecektir (Numpy, b.t.).

Numpy kütüphanesinde ndarray dizi sınıfıdır. ndarray.ndim ile dizinin eksen sayısını, ndarray.shape ile dizinin boyutunu, ndarray.size dizideki öğelerin tamamını, ndarray.dtype dizideki öğelerin hangi türde olduğunu, ndarray.itemsize dizideki öğelerin her birinin boyutunun bulunmasını sağlar.

NumPy ndarray tek boyutlu, iki boyutlu veya çok boyutlu olabilir ve hem vektör hem de matrisler ile işlemler yapılabilir. Numpy ile dizi oluşturmak için, np.array() komutu kullanılır.

Numpy bilimsel hesaplamaların kolaylıkla yapılmasını sağlayan bir kütüphanedir. Numpy, np ismi ile import numpy as np şeklinde içeri aktatırılır. Numpy ile numpy dizileri oluşturabiliriz. Numpy dizileri Python listelerine benzer fakat daha hızlı ve daha işlevseldir. Python listesinde farklı veri türleri bulunabilir fakat NumPy dizilerinde tüm öğeler homojendir. NumPy dizilerinde tüm öğelerin homojen olmasının avantajı, matematiksel işlemlerin daha verimli olmasına olanak sağlamasıdır.

Numpy kütüphanesinde yapılabilecek matematiksel işlemler; sum (toplam), max ve min ile maksimum ve minimum değerleri bulunabilir, mod, medyan, varyans ve standart sapma gibi bazı istatiksel hesaplamalar yapılabilir. Numpy basit hesaplamalar dışında, daha karmaşık fonksiyonlar ile işlemler yapabilir.

Numpy kurulumu anaconda için "pip install numpy" şeklindedir ve "import numpy as np" şeklinde içeri aktarılır. Dizi oluşturmak için array komutu kullanılır ve python listesinden veya Tuple'dan dizi oluşturulabilir. Numpy ile diziler yazdırıldığında, iç içe listeler gibi yazdırır.

3.4. SCIKIT-LEARN

Scikit-learn tahmine dayalı veri analizi için basit ve verimli araçlar sunar, herkes tarafından erişilebilir ve çeşitli bağlamlarda yeniden kullanılabilir, NumPy, SciPy ve matplotlib üzerine inşa edilmiştir ve açık kaynak, ticari olarak kullanılabilir. Scikit-learn ile, bir nesnenin hangi kategoriye ait olduğunu belirleyerek sınıflandırma, bir nesne ile ilişkili sürekli değerli bir özniteliği tahmin ederek regresyon, benzer nesnelerin otomatik kümeler halinde gruplayarak kümeleme, Dikkate alınacak rastgele değişkenlerin sayısını azaltarak boyutsal küçülme, Karşılaştırma, doğrulama ve parametreleri ve modelleri seçerek model seçimi ve özellik çıkarma ve normalleştirme ile ön işleme işlemleri yapılabilir (Pedregosa ve diğerleri, 2011: 2826).

Scikit-learn bu artan istatistiksel veri analizi ihtiyacına cevap vermektedir. Python ile entegre edilmiş kolay bir kullanıma ve arayüze sahiptir. Paket çoğunlukla Python'da yazılırken C++ ile SVM'lerin ve uyumlu lisanslarla genelleştirilmiş doğrusal modellerin referans uygulamalarını sağlar (Pedregosa ve diğerleri, 2011: 2826).

Scikit-learn veri bilimi ve makine öğrenmesi konularında yaygın olarak kullanılan bir Python paketidir. Scikit-learn denetimli ve denetimsiz öğrenme uygulamaları yapabilen bir makine öğrenmesi kütüphanesidir. Makine öğrenmesi algoritmalarını kullanarak Scikit-learn ile tahminlemeler yapılabilir.

Scikit-learn kurulumu, pip install -U scikit-learn şeklindedir ve import sklearn şeklinde içeri aktarılır. Versiyonunu öğrenmek için sklearn.__version__ komutu çalıştırılmalıdır.

3.5. TURKISHSTEMMER

Kelimeleri köklerine ayırmak için başarı ile çalışan bir kelime kök ayırıcı kütüphanesidir. Kurulumu, modül klasörünü PythonXX/Lib/site-packages veya uygulama dizininizin içine kopyalayarak yapılır. Türkçedeki isim eklerini iki grupta inceleyebiliriz. İsim ekleri (örn. "doktor-um", "doktorum" anlamına gelir) ve nominal fiil son ekleri (örn. "doktor-dur", 'doktordur' anlamına gelir). İsim fiil eki ile biten kelimeler cümle içinde fiil olarak kullanılabilir. Bu iki genel ek grubunda sınıflandırılan otuzun üzerinde farklı ek vardır. Türkçede ekler belirli sıralama kurallarına göre köke eklenir (Pypi, 2018).

DÖRDÜNCÜ BÖLÜM
MÜŞTERİ YORUMLARININ MAKİNE ÖĞRENMESİ YOLU İLE SINIFLANDIRILMASI UYGULAMASI

4.1. ÇALIŞMANIN AMACI

Küreselleşen dünyanın rekabetçi ortamında firmaların ayakta kalabilmesi veya durumlarını daha iyiye taşıyabilmeleri için müşterilerini memnun etmeleri en önemli etkenlerden birisidir. Bu önemli etki sebebiyle müşterileri memnun etmenin yollarından birisi bir ürün veya hizmet sattıktan sonra müşterinin geri dönüşlerine göre gelecek satışlar için yeni yöntemler veya iyileştirmeler denenmelidir.

Rekabetin ve satın alma davranışlarının artması ile birlikte satış sonrası değerlendirme süreci de büyük önem kazanmıştır. Günümüzde insanlar bir hizmet veya ürün satın almadan önce ürün veya hizmeti değerlendirecek birçok platform bulabiliyorlar. Dijital dünyada tüketicinin bir ürünü satın almadan önce karar vermesini etkileyen ve fayda sağlayan platformlar aynı zamanda satış sonrası yaşadıkları zorlukları da çözüme kavuşturabilmek için şikayetlerini ve yorumlarını dile getirdikleri platformlarda haklarını arıyorlar. Günümüz dijital dünyasında müşterilerin sorunlarını yazdıkları platformların arasında ilk sıralarda "şikayetvar" sitesi gelmektedir.

Dijital bir şikâyet platformu olan "şikayetvar" sitesi tüketicileri ve işletmeleri yaşadıkları memnuniyetsiz durumlar ile bir araya getirirerek hem tüketicilere yaşamış oldukları olumsuz duruma çözüm bulma hem de işletmelere sahip oldukları imaj ve itibarı koruma olanağı sunar (Karapınar ve Öcal, 2017: 104).

Dijital şikâyet platformlarında tüketicilerin sorunlarını anlatması firmalar açısından bir avantaja dönüştürülebilir. Bilinçli müşterilerin haklarını aramak için başvurdukları bu platformdaki yorumları analiz edip işletmelere bu sorunlar ve şikayetler doğrultusunda sonuçlar sunmak mevcut müşterileri memnun etmeye ve gelecekteki müşteri sayısını arttırmaya yardımcı olmada izlenecek iyi bir yöntemdir.

Müşteriler sorun yaşadıkları bu işletmeler tarafından alacakları geri dönüşlere göre işletmeye bağlılıklarını arttırabilir veya tekrar alışveriş yapmamak üzere ilişkilerini tamamen sonlandırabilirler. Bu sebeple şikayetlerini dile getiren

müşterilerin sorunlarını çözmek ve şikâyet sitesi platformlarında şikayetlerin çözüldüğünü göstermek işletmelere kar olarak geri dönecektir. Bir müşteri yaşadığı mağduriyetin sonucuna göre kararlar vermektedir. Günümüzde e-ticaretin de gitgide artması ile bir müşteri için seçenek sayısı oldukça fazladır. Bu sebepten dolayı bir müşteriyi kaybetmek ve rakip firmalara yöneltmemek için yaşanılan mağduriyeti en uygun şekilde çözüme kavuşturup şikâyet siteleri platformlarından faydalanıp uygun analizler ve çözümler bulunmalıdır.

Henüz işletmeyi bırakmamış müşteri, şikâyetinin çözümüne göre hareket ederek, işletmeyi bırakıp bırakamamakla ilgili karar verecektir. Bu durum ise işletme açısından büyük bir fırsattır. Bu fırsatı iyi değerlendirerek, işletme müşterisini memnun etmesi durumunda tekrar kazanabilecektir (Alabay, 2012: 137)

Satılan ürün veya hizmetin kalitesi ve fiyatı gibi birçok etkenin yanında firma ve müşteri açısından da müşteri ilişkileri yönetimi, mevcut müşteri kaybetmeme ve müşteri kazanma konusunda iyi bir etkendir. Müşteri ilişkileri yönetimi, şirketin tüm ilişkilerini yönetmek için kullanılan yöntemdir. İyi bir müşteri ilişkileri yönetimi sistemi kurmak karlılığı arttıracaktır. Müşteri ilişkileri yönetimi için satışlarda yeni stratejiler oluşturulabilir.

Satışın tüm aşamaları değerlendirildiğinde en önemli ve kar getiren süreç müşteri memnuniyeti sürecidir. Gitgide yoğunlaşan rekabet ortamında müşteriyi memnun etmek gitgide zorlaşmaktadır. Memnuniyetten önce artan rekabet ile birlikte pazaryerlerinin de gitgide büyümesi ile ürün çeşitliliği günden güne artarken satış yapmak daha da zorlaşmıştır. Yeni bir müşteri edinmenin maliyeti mevcut müşteriyi korumaktan çok daha fazla olduğundan dolayı mevcut müşterileri kaybetmemek ve marka ile iletişimi koparmamak, daha çok satış yapmak için müşteriyi her zaman memnun tutmak çok önemlidir. Satış öncesinde memnun ettikten sonra satış sonrası da memnun edebilmek için müşteriden gelen her şikâyet büyük önem taşımaktadır. Müşterinin şikayetine sessiz kalmak müşterinin marka tarafından önemsizleştirdiğini hissettirmektir ve bu da büyük oranda müşteri kaybına sebep olmaktadır.

Şikâyet kavramı, işletmelerin müşterileri ile uzun süreli başarılı ilişkiler kurabilmesi için en etkili araçlardandır (Tosun ve Söyük, 2019: 1873).

İşletmeler, satış miktarlarını arttırabilmek için satış süreçlerinin tamamını kapsayan aşamalar ile ilgili veriler elde etmeli ve elde ettikleri verilerden analizler

oluşturmalıdırlar. Veri analizi, işletmelere karlılık ve iyileştirme konularında yol göstermektedir. Veri analizi yapacakları en önemli konulardan birisi de müşteri şikayetleridir. İşletme açısından önemli konular sınıflandırılmalı ve buna göre şikayetler kategorilere ayrılmalıdır. İşletmelerde buna göre en çok şikâyet aldıkları konularda iyileştirmeler yapabilirler.

Müşteri açısından kalite en önemli etkenler arasındadır. İşletmeleri kalite imajlarını belirleyen unsurlar arasında sadece ürün kalitesi değil aynı zamanda müşteri ile kurulan ilişkiler, satış danışmanlığı ve satış sonrası destek gelmektedir. Müşteri şikayetleri, satışın her aşamasıyla ilgili olabilmektedir. Kalite imajını yükseltmek için şikayetler göz ardı edilmeden özenle incelenip müşterinin sorunu çözülmeli ve tekrarlanmaması açısından şikâyet analizleri konusunda çözümler bulunmalıdır.

Bir işletmeyi hem dijital hem gerçek pazar yerlerinde ön plana taşıyacak en önemli unsurların başında olumlu geri dönüşler yer almaktadır. Olumlu geri dönüşler alabilmek için ise olumsuz yorum gelen konuların sık sık analiz edilip en kısa sürede ortadan kaldırılması işletmelere büyük bir avantaj sağlamaktadır. Her müşteri satın aldığı ürün veya hizmetten memnun kalmak ister. Memnun kalmadığı durumlarda ise müşterinin problemleri çözüldüğünde ve mağduriyeti telafi edildiğinde, müşterinin işletmeyi olan bağlılığı artabilir.

Müşteri tatmini, müşterinin beklentilerinin karşılanması ile sağlanır. Bu beklentiler ürünün kullanım alanı, kolaylığı, kalitesi vb. ürün ile ilgili olabileceği gibi, ürünü satan satış elemanları, ürünün dağıtım kanalı ile ilgili de olabilir. "Müşteri ne istiyor, beklentileri nelerdir?" Tüm bu soruların cevaplarını müşteri ilişkileri yönetimi programı kaydetmekte ve bu bilgiler müşteri ilişkilerine yön vermektedir (Ovalı, 2014: 3).

Akın ve Çörek'e (2005: 260) göre müşteri memnuniyet ölçümünde amaçlanan;

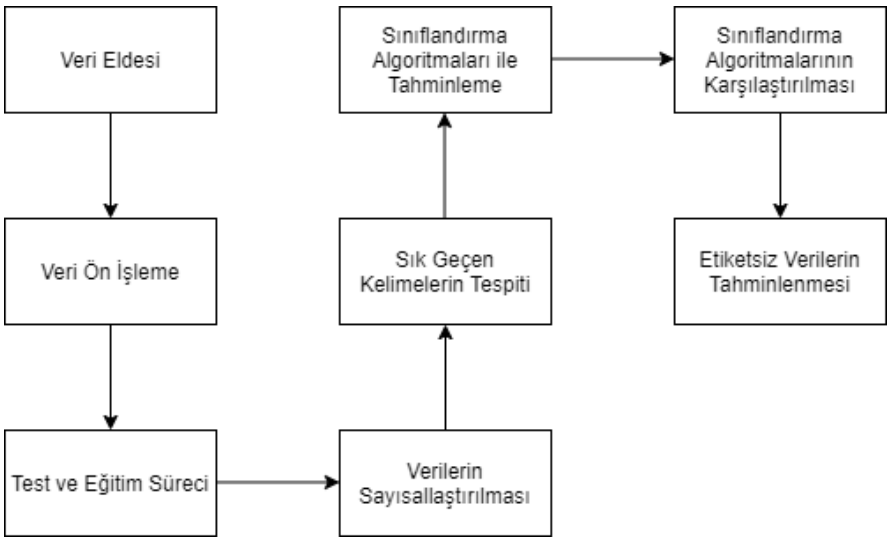
- Müşterilere en iyi şekilde hizmetin sağlanabilmesi için müşterilerin gereksinimlerini öğrenmek,
- Müşterinin ilettiği şikayetler hakkında ürün ve süreçlerde iyileştirme yapmak,
- Müşterinin gözünden rakip performansını anlamak,
- Müşteriye sunulan değerlerin müşteriler gözünden nasıl algılandığı konusunda bilgi edinmek,

- Müşteri önceliklerini belirlemek,
- Müşterilerden yeni ürünlerin yaratılmasında görüşler sağlamaktır.

İşletmeler açısından müşteri şikayetlerini dikkate almak ve buna göre planlamalar yapmak çok önemlidir. Müşteriler ürün veya hizmet satın alımından önce ve sonra satın aldıkları ürün ve hizmetler ile ilgili memnun kalmak ister ve herhangi bir sorun yaşadıklarında şirket tarafından bu sorunların telafi edilmesi çok önemlidir. Bu çalışmada, işletmelere yapılan şikayetleri kategorilere ayrılarak işletmeye en çok hangi kategoride şikâyet geldiğinin analizi yapılmıştır. Müşteriler, internetin de sağladığı kolaylık sayesinde bir ürün veya hizmet satın aldıktan sonra eğer ürün hakkında bir şikâyette bulunmak istediklerinde çoğunlukla çevrimiçi şikâyet sitelerini kullanırlar. İşletmeler de bu şikayetleri dikkate alır ve buna göre iyileştirmeler yaparlar ve bu sayede kazançlarını arttıracaklarını düşünmektedirler.

Çalışmada Türkiye'nin en büyük şikâyet platformlarından biri olan şikayetvar.com sitesinden elde edilen üç rakip teknoloji firmasının (A, B ve C olarak isimlendirilmiştir) en çok hangi konularda şikâyet aldıkları analiz edilmiştir. Şikâyet kategorileri; iade/değişim ve geri ödeme, teslimat/kargo, müşteri ilişkileri ve hizmetleri ve garanti ve servistir olarak sıralanmaktadır. Çalışmanın her adımını gösteren akış şeması Şekil 2'de gösterilmektedir.

Şekil 2: Çalışma Akış Şeması



Kaynak: Yazar Tarafından Derlenmiştir.

4.2. VERİ ELDESİ

Şikayetvar.com adresinden üç büyük teknoloji firmalarının 2020 yılındaki şikayetleri Python programla dili BeautifulSoup kütüphanesi kullanılarak siteden çekilmiştir. BeautifulSoup, HTML ve XML dosyalarından, web sitelerinden veri çekmek için kullanılan bir Python kütüphanesidir. Beautiful Soup, ekran kazıma gibi hızlı geri dönüş projeleri için tasarlanmıştır.

Beautiful Soup kütüphanesi ile alınan veriler, müşteri şikayetleri, müşteri şikayetlerinin tarihleri ve saatleri analiz edilmek üzere şikayetvar.com sitesinden indirilmiştir. Daha sonra veri bütünlüğü açısından şikâyet yorumları arasındaki eksik yorumlar veriden temizlenmiştir.

4.3. VERİ ÖN İŞLEME

Veri ön işleme adımı makine öğrenmesi modelleri kurulmadan önce verinin daha anlaşılabilir ve düzenli hale gelmesini sağlayan bir adımdır. Modelleri kurmadan önce mutlaka yapılması gereken bir işlemdir. Veri kalitesini arttırmaktadır ve veri kalitesi modellerin daha başarılı çalışabilmesi için en önemli etkenlerdendir. Yanlış analizler yapmamak için veri ön işleme adımı mutlaka uygulanmalıdır.

Bu çalışmada veri ön işleme adımları aşağıdaki gibidir;

1. Etkisiz kelimelerin (stop words) kaldırılması: Etkisiz kelimeler, sınıflandırma için herhangi bir anlam ifade etmeyen fakat cümleler içinde sık sık kullandığımız kelimelerden oluşmaktadır. Analizin daha doğru ve anlamlı sonuçlanması için etkisiz kelimelerin kaldırılması gereklidir.
2. Noktalama işaretlerinin kaldırılması: Noktalama işaretleri analiz için doğru ve anlamlı sonuçlar vermeyeceği için kaldırılmıştır.
3. Kelimelerin küçük harflere dönüştürülmesi: Büyük harfli karakterler küçük harfli karakterlere dönüştürülmüştür.
4. Kelimeleri köklerine ayırmak: Türkçe sondan eklemeli bir dildir ve bu sebeple kelimeler bütünüyle değil kökleri ile daha anlamlı olmaktadır. Bu sebeple kelimeleri bütünü ile değil kökleri ile analiz etmek bize daha doğru

sonuçlar verecektir. TurkishStemmer kütüphanesi ile kelimeler köklerine ayrılmıştır ve model kelimeleri kökleri ile değerlendirilmiştir.

4.4. EĞİTİM VE TEST VERİLERİ

Model kurulmadan önce verinin bir kısmı eğitim için ve bir kısmı ise test için kullanılmıştır. Bu veriler önceden belirlenen kategorilere göre etiketlenmiştir. Daha sonra yapılan denemelerde en yüksek başarı oranı %80 eğitim, %20 test olacak şekilde elde edilmiştir. Tablo 1'de belirlenen şikâyet kategorilerine ait örnekler verilmektedir.

Tablo 1: Müşteri Şikayetlerinin Kategori Bazında Örnek Gösterimi

Kategori	**Yorum**
İade / Değişim ve Geri Ödeme	22.12.2019 tarihinde *** üzerinden ***'tan alışveriş yaptım, sistemimdeki adres ve telefon bilgileri yanlış girilmiş, kargo elime ulaşmadı, ***'tan hiçbir şekilde muhatap bulamadım, bu kadar ilgisiz firma olamaz, hala kargomda yok ortada, para iademde yok.
Teslimat / Kargo	Satın aldığım ürünü sipariş 413**2468 kargoya verdiniz ama çalıştığınız *** firması yüzünden bir şikayet de size yazıyorum. Nasıl böyle bir kargo şirketiyle çalışırsınız müşteri mağduriyeti sizi de bağlıyor bu yüzden teslim alamadığım ürün hala sizin sorumluluğunuzda. Ürün kargo merkezinde 1 haftaya yakındır bekliyor.
Müşteri İlişkileri ve Hizmetleri	Cumartesi günü Marmara Park AVM *** mağazasına elektronik alışverişi için girdik. Çalışanlar küme küme sohbet halinde stantlara yaslanmış, keyiflerini bozmuş olsak da "*** kulaklık var mı acaba? " diye sordum. Kollarını göğsünde bağlamış bir arkadaş "bilmiyorum" diye cevap verdi! Koskoca mağazada sadece bir kişi arkadaş bakmaya gitti dedi ancak kimse bize hangi model veya ne arıyorsunuz diye sormadı, diğerlerinin istifini bozmadan muhabbetlerine geri dönmeleri müşteri gözünden inanın hiç hoş bir görüntü değil. Bilmese bile, "ben ilgilenmiyorum ama hemen ilgili arkadaşı yönlendiriyorum" gibi bir cümle duymak isterdik. Çalışanlarınıza sanırım mağaza içi karşılama, hizmet nasıl verilir eğitimi vermiyorsunuz. Başka türlü açıklaması dahi olamaz! *** tarafından iletişim kurulmasını talep ediyorum.
Garanti ve Servis	Daha 3 4 ay önce ***'dan sıfır olarak aldığım bilgisayarım çok nazik kullanmama rağmen 2. Defa bozuldu ürün şarj almıyor. Garantiye gönderdiğimde Bios güncellemesi yapıp göndermişler ancak bir güncelleme 18 İs gününde yapıldı ve en çok ihtiyacım olan egitim süremde çok mağdur oldum. Servisten geldi ve gene bir defa kullanamadan tekrar şarj almıyor. Üründen tiksindim paramın iadesini istiyorum. Teşekkür ederim

Kaynak: Yazar Tarafından Derlenmiştir.

Eğitim verisi için 396 adet müşteri şikâyeti kullanılmıştır. Dört sınıflandırma kategorisi olan, iade/değişim ve geri ödeme, teslimat/kargo, müşteri ilişkileri ve hizmetleri ve garanti ve servis kategorileri için 396 adet müşteri şikâyeti eşit olacak şekilde her birinden 99 adet alınarak eğitilmiştir. Algoritma sonucunda aşırı öğrenme veya yanlış analiz olması istenilen bir sonuç olmayacağı için eğitim verisinde her kategoriden eşit sayıda kullanılmıştır.

4.5. VERİNİN SAYISALLAŞTIRILMASI VE SIK GEÇEN KELİMELERİN TESPİTİ

Veri test ve eğitim olarak ikiye ayrıldıktan sonra makine öğrenmesi modellerinin veriyi anlayabilmesi için veri sayısallaştırılmalıdır. Verinin sayısallaştırılması için birçok yöntem bulunmakta, bu çalışmada sayısallaştırma yöntemlerinden, Tfidf skorlama yöntemini kullanarak veriyi sayısallaştırılmıştır. Tfidf skorlama yöntemi dokümanda sıkça geçen kelimelerin baskın olması ve dokümanı temsil etmesinin önüne geçilmesini sağlamaktadır. Burada n_gram_range olarak unigram ve bigram yani tekli ve ikili kelimeleri göz önüne alınmıştır. max_df ve min_df ise kelime dağarcığı oluştururken kelimelerin metinde geçen sıklığına göre ayarlama yapılmasıdır.

4.6. SINIFLANDIRMA ALGORİTMALARI İLE TAHMİNLEME

Çalışmada makine öğrenmesi sınıflandırma algoritmalarından üç tane model ile çalışılmıştır. Bu modeller; logistik regresyon, olasılıksal dereceli azalma ve doğrusal destek vektör makinalarıdır.

4.6.1. Lojistik Regresyon

Lojistik regresyon, doğrusal sınıflandırma modellerinde kullanılan bir sınıflandırma modelidir. Birden fazla bağımsız değişken bir veri var ise lojistik regresyon istatistiksel yöntemi kullanılır. Tüm regresyon analizleri gibi lojistik regresyon analizi de tahmini analizdir.

Çalışmada sınıflandırma tahminlemesi yapılacağı için lojistik regresyon sınıflandırma modeli aşağıdaki şekilde kurulmuştur.

```
model = OneVsRestClassifier(LogisticRegression(penalty = 'l2', C=1.0))
model.fit(x_train_tfidf, y_train)
```

Lojistik regresyon bir veya çok ön göstergeli değişkene bağlı kategorik bağımlı değişkenlerin sonuçlarını tahmin etmede kullanılan bir tür regresyon analizidir. Çok terimli lojistik regresyon modeli ikiden fazla ayrı sonuca izin veren lojistik regresyonu genelleştiren bir modeldir (Sağbaş ve Ballı, 2016: 379).

Bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi, temel lojistik fonksiyonunu kullanarak olasılıkları tahmin ederek ölçmektedir (Gültepe, 2019: 12).

Doğrusal regresyon analizinin uygulanabilmesi için bazı durumların sağlanması gerekmektedir. Bu durumlardan bir tanesi bağımlı değişkenin sürekli olmasıdır. Ancak, sosyal bilimlerle ilgili uygulamalarda bağımlı değişken genellikle kategorik ve iki düzeylidir. Bağımlı değişkenin iki düzeyli olduğu durumlarda lojistik regresyon analizi kullanılmaktadır (Budak ve Erpolat, 2012: 27).

4.6.2. Olasılıksal Dereceli Azalma

Çalışmada olasılıksal dereceli azalma modeli aşağıdaki şekilde kurulmuştur.

```
model2 = OneVsRestClassifier(SGDClassifier(loss = 'hinge', penalty = 'elasticnet', max_iter = 20))
model2.fit(x_train_tfidf, y_train)
```

Metin sınıflandırma ve doğal dil işleme problemlerinde başarılı sonuçlar veren bir makine öğrenmesi modelidir. Verimlilik ve kod ayarlama anlamından kolay bir algoritmadır.

Olasılıksal dereceli azalmada her adımda veri setindeki her örneği gezer ve sonra veri setini günceller.

Dereceli azalma ile öğrenmenin avantajları aşağıdaki gibidir (Montavon ve diğerleri, 2012: 14);

- Genellikle toplu öğrenmeden çok daha hızlıdır.
- Genellikle daha iyi çözümlerle sonuçlanır.
- Değişiklikleri izlemek için kullanılabilir.

Gradyanın pozitif ile orantılı olarak adımlar atılırsa, bu fonksiyonun yerel maksimuma yaklaştırır; bu fonksiyona gradyan tırmanışı adı verilir. Gradyan inişi, makine öğrenmesi alanında yaygın kullanılan bir yöntemdir, çünkü makine öğrenmesinin amaçlarından biri, eğitim verisi göz önüne alındığında, en yüksek doğruluğu bulmak veya hata oranını en aza indirmektir (Wikipedia, 2021).

4.6.3. Doğrusal Destek Vektör Makineleri

Çalışmada doğrusal destek vektör makinaları modeli aşağıdaki şekilde kurulmuştur.

```
SVC_pipeline = Pipeline([('clf', OneVsRestClassifier(LinearSVC(), n_jobs=1)),])
SVC_pipeline.fit(x_train_tfidf, y_train)
prediction = SVC_pipeline.predict(x_test_tfidf)
```

Sınıflandırma problemlerini çözmek için kullanılan denetimli makine öğrenmesi modellerinden bir tanesidir.

Destek vektör makineleri, uygulaması kolay bir algoritmadır ve birçok problemi çözebilir. Destek vektör makineleri hiç parametre almadan sınıflandırma yapabilirler. Eğitim setlerinde girdiler ve çıktılar eşlenir.

Destek vektör makinelerinin avantajları aşağıdaki gibidir;

- Yüksek boyutlu uzaylarda etkilidirler.
- Boyut sayısı, örneklem sayısından fazla olsa da etkilidirler.
- Karar fonksiyonlarında bazı bazı eğitim noktaları kullanılması sebebiyle bellek verimli şekilde kullanılır.

- Karar fonksiyonu için çok farklı çekirdek fonksiyonları (kernel fonsiyonları) kullanılabilmektedir.

Destek vektör makineleri veri setinin doğrusal olarak ayrılıp ayrılmamasına göre ikiye ayrılmaktadır. Bu çalışmada doğrusal destek vektör makineleri makine öğrenmesi modeli kullanılmıştır.

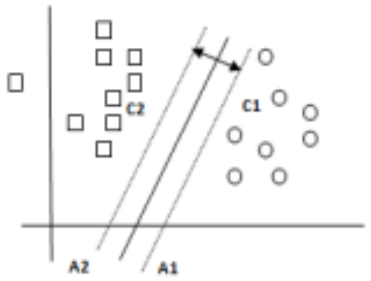
Destek vektör makineleri ile sınıflandırma yönteminde, iki sınıfa ait örnekler doğrusal olarak dağıldığında, bu iki sınıf, eğitim verisi kullanılarak elde edilen karar fonksiyonu ile birbirinden ayrılır. Bu veri setini ikiye ayırmış olan doğrunun ismi, karar doğrusudur.

Doğrusal destek vektör makineleri, verileri ayırabilecek sonsuz doğru içinden marjini en yüksek doğrunun seçilmesidir.

Lineer Ayrılabilme Durumu: Eğitim için kullanılacak N elemandan oluşan verinin $\theta = \{x_i, y_i\}$, i = 1, 2, …, N olduğu varsaydığımızda, $y_i \epsilon \{-1,1\}$ etiket değerleri ve özellikler vektörüdür. Lineer olarak ayrılabilme durumunda, bu iki değerli veriler direkt olarak bir aşırı düzlem ile ayrılabilecektir. Bu aşırı düzleme ayırıcı aşırı düzlem adı verilir (Yakut ve diğerleri, 2014: 143).

Lineer ayrılabilme durumunda optimum ayırıcı aşırı düzlem Şekil 3'te gösterilmektedir.

Şekil 3: Lineer Ayrılabilme Durumunda Optimum Ayırıcı Aşırı Düzlem



Kaynak: Yakut ve diğerleri (2014: 144)

Destek vektör makineleri ile sınıflandırmada genellikle {-1,+1} şeklinde sınıf etiketleri ile gösterilen iki sınıfa ait örneklerin, eğitim verisi ile elde edilen bir karar fonksiyonu yardımıyla birbirinden ayrılması amaçlanır. Bu karar fonksiyonu kullanılarak eğitim verisini en uygun şekilde ayırabilecek hiper-düzlem bulunur (Kavzoğlu ve Çölkesen, 2010: 75).

4.7. SINIFLANDIRMA ALGORİTMALARININ SONUÇLARININ KARŞILAŞTIRILMASI

Müşteri şikayetlerinin makine öğrenmesi sınıflandırma algoritmaları ile sınıflandırılması üç farklı algoritma ile yapılmıştır. Algoritmalar sonuçlarının karşılaştırılabilmesi için başarı kriteri olarak doğruluk oranları kullanılmıştır. Lojistik regresyon, olasılıksal dereceli azalma ve doğrusal destek vektör makineleri algoritmalarının doğruluk oranları aşağıdaki gibidir;

Lojistik regresyon = 0,8

Olasılıksal dereceli azalma = 0,7125

Doğrusal destek vektör makineleri = 0,7625

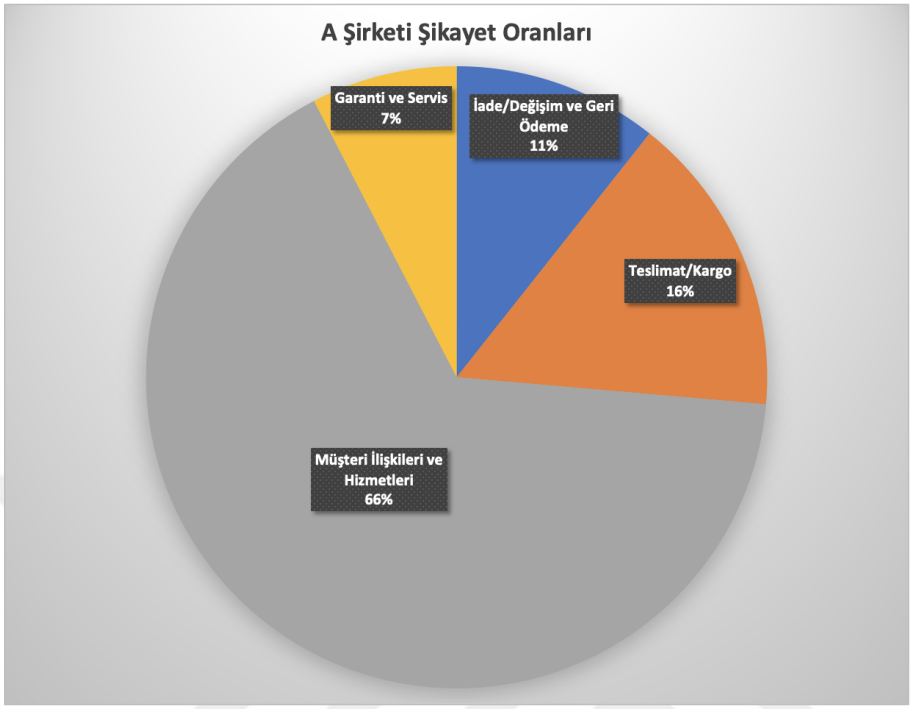
4.8. ETİKETSİZ VERİLERİN TAHMİNLENMESİ

Makine öğrenmesinde eğitim ve test verileri ile model kurulduktan sonra bundan sonra gelecek tüm veriler için tahminlemeye hazır bir halde bir model kurulmuştur. Bu model, eğitim ve test sürecinde en yüksek doğruluk oranını veren lojistik regresyon ile gerçekleştirilmektedir.

A, B ve Ceknoloji firmaları için bundan sonra gelebilecek tüm şikayetler kurulan model altında incelenip şikayetler kategorilere ayrılabilecektir. A, B ve C teknoloji firmaları için incelenmiş olan veri sayıları, A firması için 9638 satır veri, B firması için 7820 satır veri ve Cfirması için 5158 satır veri ve toplamda 22.616 satır olacak şekilde şikâyet incelemesi, üç makine öğrenmesi sınıflandırma algoritması ile analiz edilmiştir.

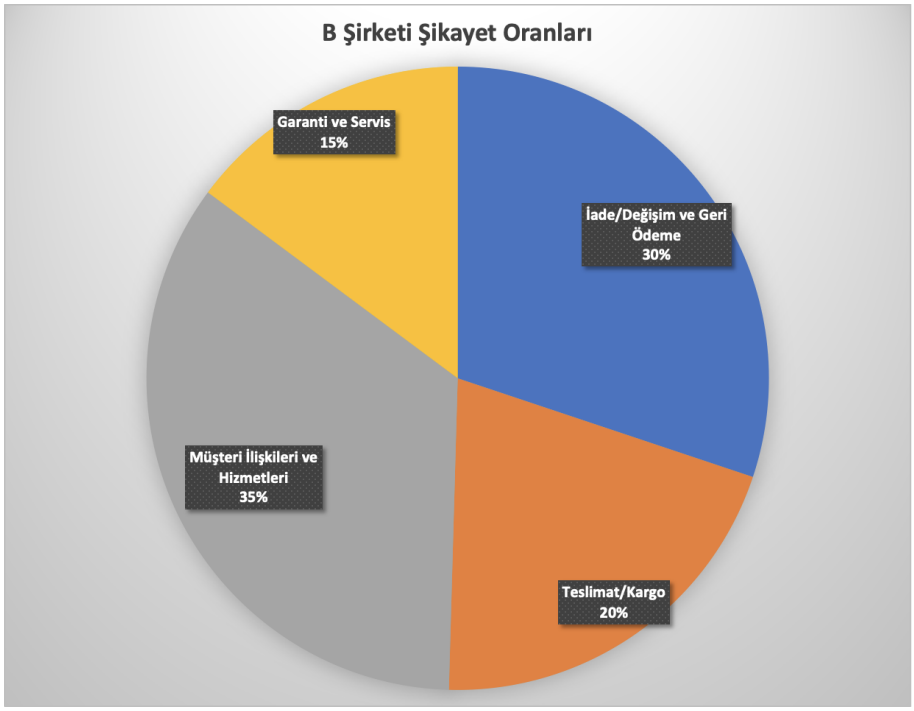
Buna göre, dört farklı kategoriye göre A, B ve Cfirmalarının 2020 yılı içindeki tahminlenen şikayetleri Şekil 4, Şekil 5 ve Şekil 6'da verilmektedir. Tablo 2'de ise bu üç şirkete ait şikayetler toplu halde verilmektedir.

Şekil 4: A Şirketi Tahminlenen Şikâyet Oran Grafiği



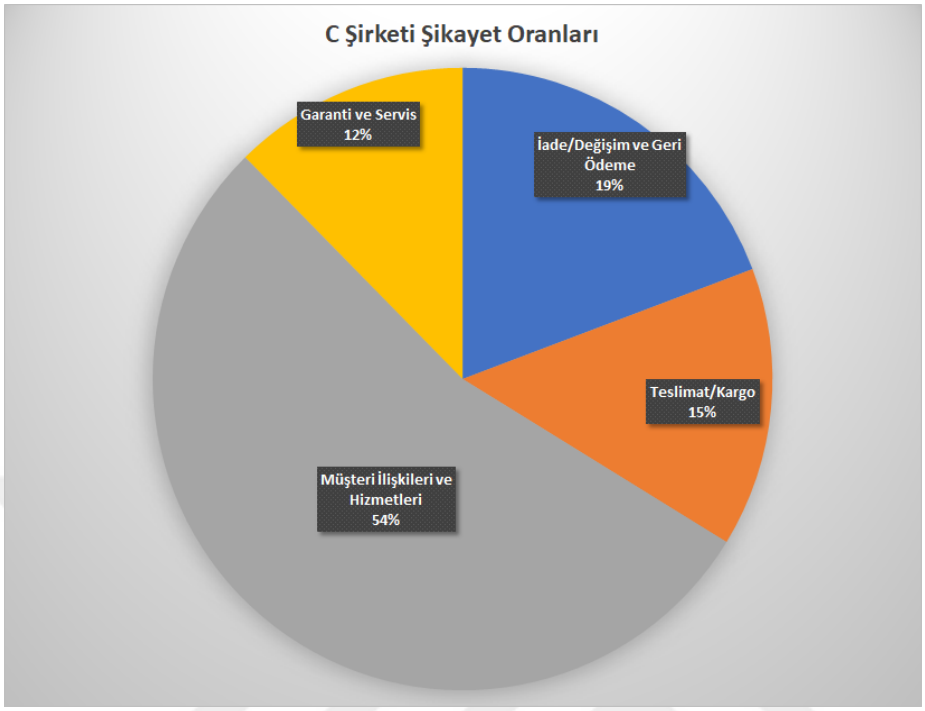
Kaynak: Yazar Tarafından Derlenmiştir.

Şekil 5: B Şirketi Tahminlenen Şikâyet Oran Grafiği



Kaynak: Yazar Tarafından Derlenmiştir.

Şekil 6: C Şirketi Tahminlenen Şikâyet Oran Grafiği



Kaynak: Yazar Tarafından Derlenmiştir.

Tablo 2: A, B ve C Şirketine Ait Tahminlenen Şikayetler ve Oranları

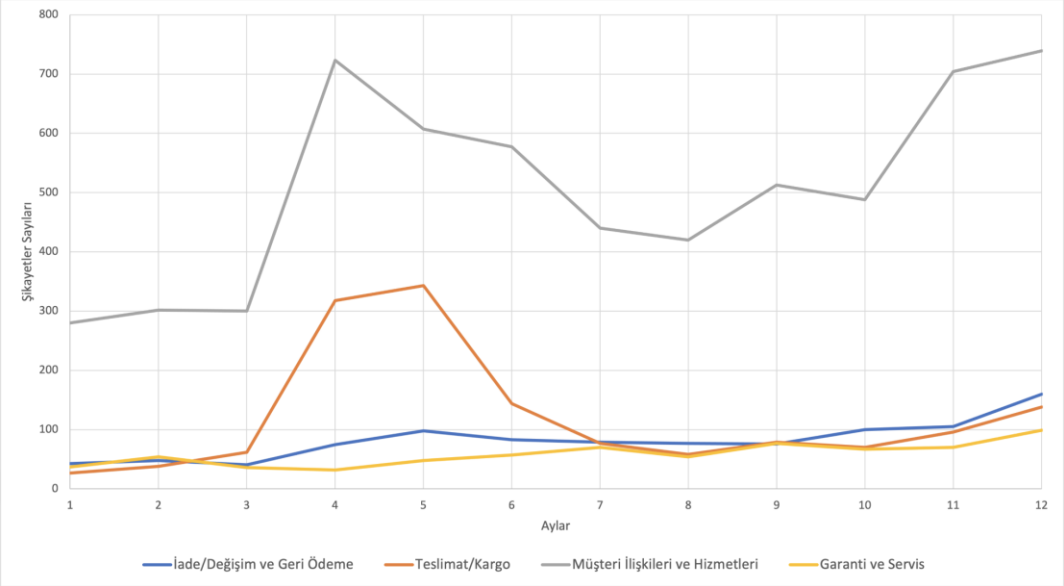
	A (%)	B (%)	C (%)
İade/Değişim ve Geri Ödeme	11	30	19
Teslimat/Kargo	16	20	15
Müşteri İlişkileri ve Hizmetleri	66	35	54
Garanti ve Servis	7	15	12

Kaynak: Yazar Tarafından Derlenmiştir.

Üç şirketin şikâyet oranları incelendiğinde en fazla şikâyet alan kategori müşteri ilişkileri ve hizmetleridir. İkinci en çok şikâyet ise A şirketinde teslimat/kargo iken B ve C şirketlerinde iade/değişim ve geri ödeme kategorisi olarak görülmektedir.

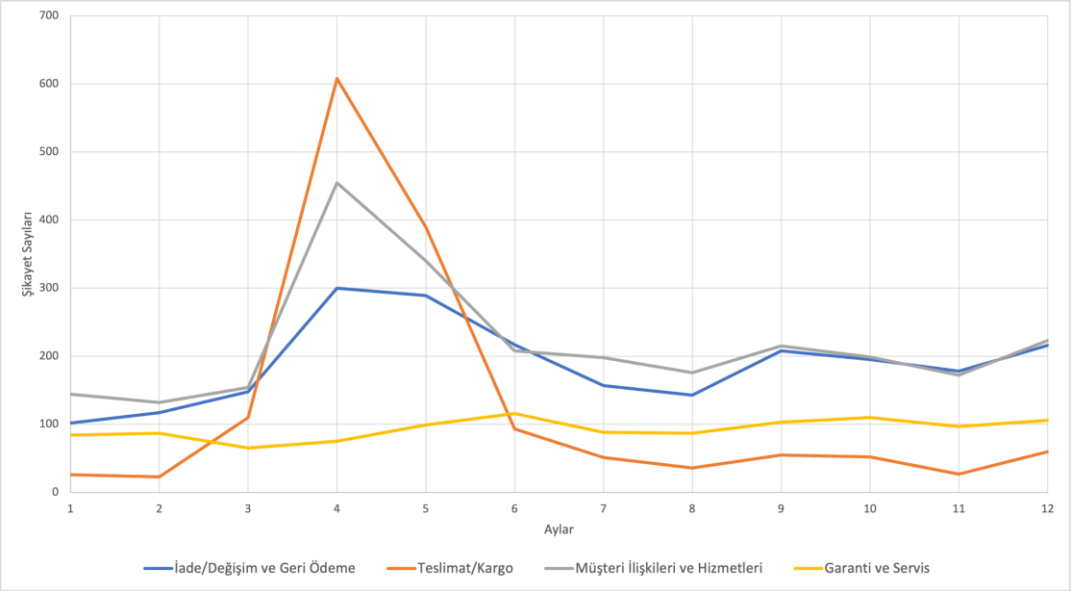
Her üç şirketin aylık olarak aldıkları şikayetlerin grafikleri Şekil 7, Şekil 8 ve Şekil 9'da verilmektedir.

Şekil 7: A Şirketi Aldığı Aylık Şikâyetler ve Şikâyet Kategorileri



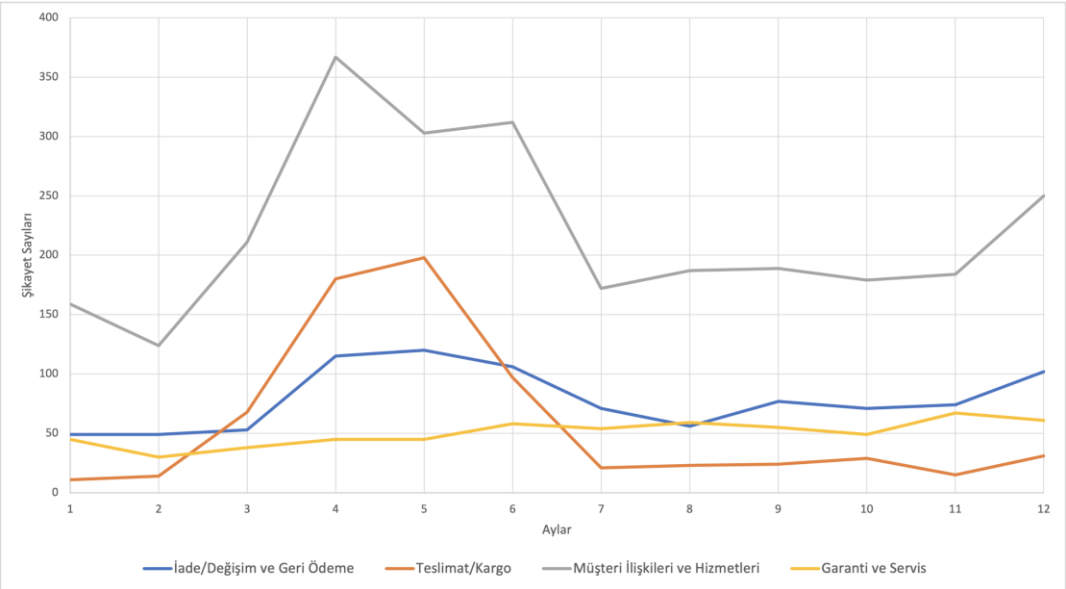
Kaynak: Yazar Tarafından Derlenmiştir.

Şekil 8: B Şirketi Aldığı Aylık Şikâyetler ve Şikâyet Kategorileri



Kaynak: Yazar Tarafından Derlenmiştir.

Şekil 9: C Şirketi Aldığı Aylık Şikâyetler ve Şikâyet Kategorileri



Kaynak: Yazar Tarafından Derlenmiştir.

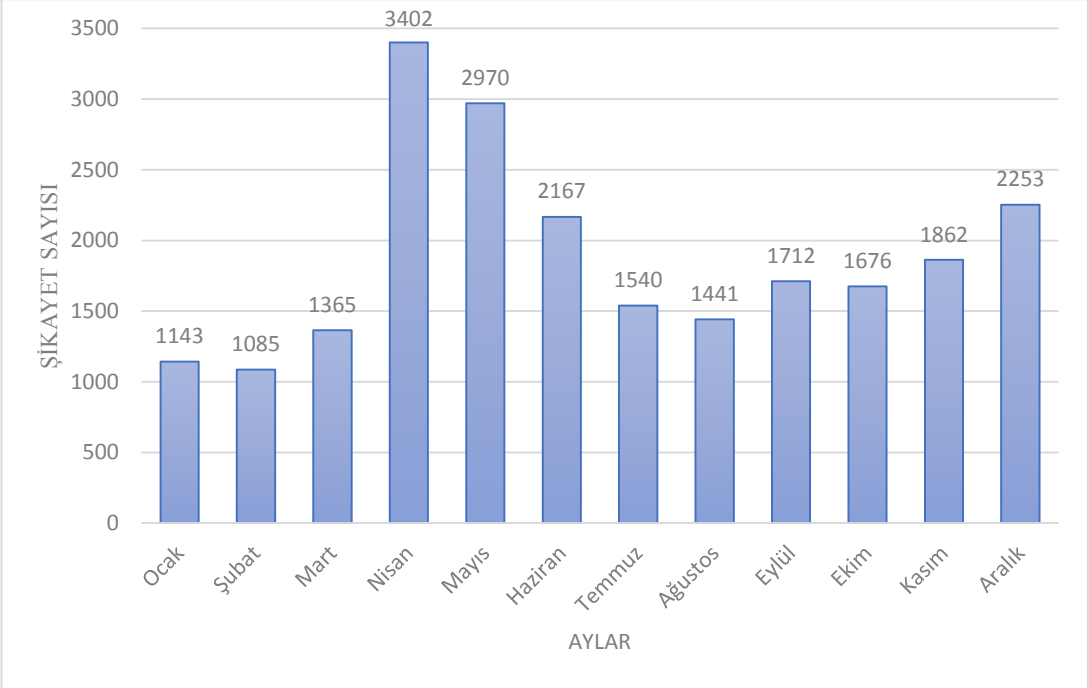
Türkiye'de ilk korona virüs vakasının ortaya çıktığı 11 Mart 2020 tarihinden sonra özellikle "Evde Kal" çağrılarını değerlendiren insanlar alışverişlerinin çoğunu internet üzerinden yapma eğilimi göstermişlerdir (Elektronik Ticaret Bilgi Platformu, 2021). Şekil 7, 8 ve 9'dan da görüldüğü üzere özellikle Mart ayından sonra şikayetlerde gözle görülür bir artış gözlemlenmektedir.

SONUÇ

Yüksek Lisans Tezi, yapay zekâ ve makine öğrenmesi teknikleri kullanılarak müşteri şikayetlerinin sınıflandırılması konusu üzerine yürütülmüştür. Yapılan çalışmada, 2020 yılına ait üç büyük teknoloji firmasına ait müşteri şikayetleri ve bu şikayetlere ait tarih ve saatler şikayetvar.com sitesinden alınmış, daha sonra önceden belirlenen kategorilere göre şikayetler üç farklı tahminleme algoritması ile incelendikten sonra analizleri yapılmıştır.

İlk olarak çalışmada 2020 yılına ait A, B ve C şirketlerine ait toplam şikâyet sayıları incelendiğinde, 2020 yılı Mart ayında pandemi dönemine giriş yapıldığı göz önünde bulundurularak, ilk Korona Virüs vakası görüldükten önceki ve sonraki şikayet verileri analiz edilmiştir. Bu analiz sonucunda üç büyük teknoloji firmasına ait şikâyet verilerinde Mart ayından sonra büyük bir artış gözükmektedir. Analizin sonucu Şekil 10'da gösterildiği gibidir.

Şekil 10: 2020 yılı A, B ve C Şirketlerine Ait Toplam Şikâyet Sayıları



Kaynak: Yazar Tarafından Derlenmiştir.

Şekil 10'da gösterildiği üzere 11 Mart tarihinde ilk vaka görüldükten sonra Nisan ayından itibaren büyük bir oranda artış gözükmektedir. Ocak, Şubat ve Mart

aylarında şikayet sayıları sırası ile 1143, 1085 ve 1365 olarak gitmekte iken Nisan ayında yaklaşık üç katı bir oran ile 3402 olarak görülmektedir. Bu sonuç, şirketlerin pandemi döneminde internet üzerinden yaptıkları satış hacimleri ile ilişkilendirildiğinde doğal karşılanabilmektedir. Ancak unutulmaması gereken bir konu da ilk defa internet üzerinden alışveriş yapan yeni bir müşteri kitlesinin sisteme dahil olmasıdır. Bu nedenle, şirketlerin bu konuları analiz ederken, satış hacmi ve müşteri profillerini de dikkate alması zorunludur.

Çalışmada üç şirkete ait şikayetler önceden belirlenen şikâyet kategorileri temel alınarak, verinin bir kısmı eğitim ve test verisi olarak ayrılarak makine öğrenmesi süreci gerçekleştirilmiştir. Sonuç olarak en iyi başarı oranı %80 ile lojistik regresyon ile elde edilmiştir. Daha sonra geri kalan etiketlenmemiş şikayetlerin hangi kategorilere ait oldukları lojistik regresyon yöntemi ile tahminlenmiştir.

Tahminleme sonucunda A şirketine ait müşteri yorumları incelendiğinde, Şekil 4'te görüldüğü gibi %66'lık bir oran ile Müşteri İlişkileri ve Hizmetleri Kategorisi en fazla şikâyeti alan konu olduğu tespit edilmiştir. %16'lık bir oran ile Teslimat/Kargo, %11'lik bir oran ile İade/Değişim ve Geri Ödeme ve %7'lik bir oran ile Garanti ve Servis kategorisinin şikâyet aldığı tespit edilmiştir. Yapılan analiz sonucunda A şirketinin %66'lık gibi büyük bir oran ile Müşteri İlişkileri ve Hizmetleri kategorisinde yüksek miktarda şikâyet alması sebebi ile bu konuda iyileştirmelere gidebileceği yorumu yapılabilir.

B şirketine ait tahminlenen müşteri yorumları incelendiğinde, Şekil 5'te görüldüğü gibi %35'lik bir oran ile Müşteri İlişkileri ve Hizmetleri Kategorisi en fazla şikâyeti alan konu olduğu tespit edilmiştir. %30'luk bir oran ile İade/Değişim ve Geri Ödeme, %20'lik bir oran ile Teslimat/Kargo ve %15'lik bir oran ile Garanti ve Servis kategorisinin şikâyet aldığı tespit edilmiştir. B şirketine ait analiz sonucu incelendiğinde, ilk sırada yer alan Müşteri İlişkileri ve Hizmetleri Kategorisi ve ikinci sırada yer alan İade/Değişim ve Geri Ödeme kategorisinin birbirine yakın oranlarda şikâyet alması sonucuna göre bu iki kategoride gelen şikayetlerin oranı dikkate alındığında iki kategoride de şirketin düzeltmeler yapmak için çalışmalar yapması gerektiği önerilebilir.

C şirketine ait tahminlenen müşteri yorumları incelendiğinde, Şekil 6'da görüldüğü gibi %54'lük bir oran ile Müşteri İlişkileri ve Hizmetleri Kategorisi en

fazla şikâyeti alan konu olduğu tespit edilmiştir. %19'luk bir oranda İade/Değişim ve Geri Ödeme, %15'lik bir oranda Teslimat/Kargo ve %12'lik bir oranda Garanti ve Servis kategorisinin şikâyet aldığı tespit edilmiştir. C şirketine ait analiz sonucu incelendiğinde, ilk sırada yer alan Müşteri İlişkileri ve Hizmetleri kategorisi büyük bir oranda şikâyet aldığı tespit edilmiştir ve şirketin bu kategoride düzeltmeler yapması gerektiği önerilebilir.

Üç şirketin şikâyet analiz sonuçları incelendiğinde üç şirketin de en fazla yorum alan kategorisi Müşteri İlişkileri ve Hizmetleri kategorisi olduğu tespit edilmiştir. Teknoloji ürünleri ve hizmetleri satın alım işlemi öncesinde ve sonrasında, müşterilerin müşteri ilişkileri konusuna önem veriyor yorumu yapılabilmektedir. Her üç şirket açısından, satış elemanları ve müşteri temsilcilerinin davranışlarının yeniden gözden geçirilmesini gerekli kılabilir. Şikayetlerin niteliği ve yoğunluğu ölçüsünde kısmi veya tüm satış elemanlarını kapsayacak kısa süreli eğitim programları düzenlenebilir.

A şirketinin aylık bazda şikâyet grafiği incelendiğinde (Şekil 7), 2020 Mart ayından sonra Teslimat/Kargo ve Müşteri İlişkileri ve Hizmetleri kategorilerinde büyük oranlarda değişim görülürken, İade/Değişim ve Garanti ve Servis kategorilerinde bir değişim görülmemektedir. Mart ayından sonra oluşan Teslimat/Kargo ve Müşteri İlişkileri ve Hizmetleri kategorilerindeki değişim Haziran ayı itibari ile düşüş göstermektedir. Bu düşüşe sebep ise pandemi sürecindeki normalleşme kararları etkili olmuştur.

B şirketinin aylık bazda şikâyet grafiği incelendiğinde (Şekil 8), 2020 Mart ayından sonra Teslimat/Kargo, Müşteri İlişkileri ve Hizmetleri ve İade/Değişim kategorilerinde büyük oranlarda değişim görülürken Garanti ve Servis kategorilerinde bir değişim görülmemektedir. Mart ayından sonra oluşan Teslimat/Kargo, Müşteri İlişkileri ve Hizmetleri ve İade/Değişim kategorilerindeki değişim Haziran ayı itibari ile düşüş göstermektedir. Bu düşüşe sebep ise pandemi sürecindeki normalleşme kararları etkili olmuştur.

C şirketinin aylık bazda şikâyet grafiği incelendiğinde (Şekil 9), 2020 Mart ayından sonra Teslimat/Kargo, Müşteri İlişkileri ve Hizmetleri ve İade/Değişim kategorilerinde büyük oranlarda değişim görülürken Garanti ve Servis kategorilerinde bir değişim görülmemektedir. Mart ayından sonra oluşan

Teslimat/Kargo, Müşteri İlişkileri ve Hizmetleri ve İade/Değişim kategorilerindeki değişim Haziran ayı itibari ile düşüş göstermektedir. Bu düşüşe sebep ise pandemi sürecindeki normalleşme kararları etkili olmuştur.

Şirketlerin başarılı olmaları ve karlılıklarını artırabilmeleri, satışlarını artırabilmesine ve marka değerlerini korumasına bağlıdır. Her iki durumun sağlıklı olması ve sürekliliği, müşteri memnuniyetine bağlıdır. Şirketlerin mevcut müşteri portföyünü koruyabilmesi ve yeni müşteriler kazanması, şirket imajının korunmasına bağlıdır. Bu nedenle şirket ve müşteri ilişkilerinde olumlu veya olumsuz geri bildirimlerin önemi büyüktür.

Bu yolla şirketler satış politikaları, satış kanalları ve yöntemleri ile müşteri ilişkilerini kontrol altında tutabilmek veya varsa olumsuz geri dönüşleri dikkate alarak aksayan yönleri düzeltme yoluna gidebilir.

Bu çalışma, bu yönüyle şirketlerin karar süreçlerine etki yapabilecek önemli bir faktör olduğunu ortaya koymaktadır.

Yüksek Lisans Tezinde, üç büyük teknoloji şirketine şikâyet verilerinin analizi sunulmuştur ve bu çalışmanın teknoloji firmaları için katkısı şikâyet sonuçlarına göre alınan şikâyet kategorileri ile ilgili iyileştirme plan ve çözümleri yapmaları önerisi getirilmektedir. Şirketler, alacakları kararlar ve iyileştirmeler ile daha fazla müşteri edinebilir ve mevcut müşterilerini kaybetmemek için çalışmalar yapabilirler. Şirketlerin gelişimi ve karlılığı açısından gelişimi için her veriyi değerlendirip verilerden bilgi elde etmeleri konusunda bilişim teknolojilerini kullanmaları, küreselleşen ticaret ortamında rakip firmalardan geride kalmamak için şart olmaktadır. Günümüz küreselleşen rekabet ortamında hem geri kalmamak hem de şirketi geliştirmek için bilişim teknolojilerini kullanarak, geliştirmeler yapılması olmazsa olmaz bir koşul haline gelmiştir. Bununla birlikte, dünya çapında önemi günden güne artan yapay zekâ ve makine öğrenmesi teknolojileri ile analizler yapılmıştır.

Ürün veya hizmetlere, çevrimiçi satış kanallarının da hem yurtiçi hem yurtdışı kanallarda ulaşımın ve satın almanın kolay olması alışveriş alışkanlıklarını arttırmaktadır. 2020 yılı Mart ayı itibari ile evden çıkmadan alışveriş yapma kolaylığı sunmak ve e-ticaret konusunda geride kalmamak firmalar için olmazsa olmaz bir şart haline gelmiştir. Pandemi süreci, insanlarda çevrimiçi alışveriş yapma alışkanlığını

arttırmıştır ve bu durum şirketlere daha fazla satış imkânı ile birlikte daha fazla veriye, daha kolay bir şekilde ulaşma imkânı sunmuştur. Bu konu da rekabeti olağandan daha fazla arttırmaya sebep olduğu için firmaların bir şekilde gelişimlerini daha iyiye taşımalarına olanak sağlayan bilişim teknolojileri çok büyük kolaylıklar ve verimlilik sağlayarak, elde edilebilecek tüm veriyi değerlendirip, şirketlerdeki üst yönetime önemli raporlar sunmada büyük bir kolaylık sağlamıştır. Yapay Zekâ ve Makine Öğrenmesi teknolojileri veriden verimli bir şekilde bulgu elde etmek için gelişen dünyada şirketlere önemli katkılar sağlamaktadır.

Bu çalışmanın, karar vermeye destek sağlayarak, yeni müşteri edinme ve mevcut müşteriyi koruma kapsamında şikâyet oranları ve konuları dikkate alınırsa, karlılık anlamında büyük katkısı olacaktır. Tüketimin artması ve teknolojinin gelişmesi ile birlikte verinin önemi gitgide artmaktadır. Alışveriş alışkanlıklarının ve internetin kullanımının artması, elde edilecek verinin artmasına da olanak sağlamaktadır. Fakat, veriyi elde etmek tek başına anlam ifade etmemektedir. Şirketlerin gelişmesi için elde edilen verinin değerlendirilmesi gerekmektedir. Şirketlerin satış kanallarını geliştirmesi ve müşterilerin internetten ürün ve hizmet ile ilgili tüm bilgilere daha kolay erişebilmesi daha çok veriye ulaşımı kolaylaştırması ile müşteriler artık satın aldıkları ürün veya hizmetler ile ilgili şikayetlerini daha kolay aktarabilmekte ve satın almak istedikleri ürün ve hizmetler ile alakalı daha fazla yoruma ulaşabilmektedir. Şirketler marka bilinirliği ve kalitesini iyi yönde ilerletmek için müşteriler tarafından gelen yorumları dikkate almalılardır. Çalışma kapsamında, artan verinin iyi bir şekilde değerlendirilmesi ve veriyi analiz edip, analiz edilen veriyi şirketlerdeki üst yönetime sunmak amaçlanmıştır. Çalışma, şirketler ile ilgili verileri değerlendirip, şirketlere alınan şikayetlerden sonra şikâyet sayıları, kategorileri, şikâyet zamanları ve rakip firmalar ile karşılaştırmalar yapıp, elde edilen bulgular doğrultusunda sorunun fazla olduğu kategorilerde daha iyi bir yol izlemesi açısından önemlidir.

Yapılan tez kapsamında sadece 2020 yılına ait üç şirket incelenmiştir. Gelecekte yapılacak çalışmalarda daha geniş kapsamlı olarak birkaç yıllık verinin incelenmesi A, B ve C şirketleri için yararlı olacaktır. Bununla birlikte, çalışma kısıtları arasında üç farklı algoritma olan, Lojistik Regresyon, Doğrusal Destek Vektör Makineleri ve Olasılıksal Dereceli Azalma algoritmaları kullanılmıştır.

Gelecekte yapılacak çalışmalarda daha farklı sınıflandırma algoritmaları ile sonuçlar değerlendirilecektir, kategorilerin sayısı arttırılmakla birlikte firmaların ürün kategorileri bazında da analizler yapılacaktır ve şikâyet verilerini sadece şikayetvar.com platformundan değil elde edilebilecek diğer şikâyet platformlarından da devamlı bir şekilde elde edip interaktif bir destek sistemi oluşturulacaktır.

KAYNAKÇA

Adalı, E. (2020). *Türkçe Doğal Dil İşleme.* Ankara: Akçağ

Ahmed, H., Traore, I. ve Saad, S. (2018). Detecting Opinion Spams and Fake News Using Text Classification. *Security and Privacy*. 1(1): e9.

Akın, H. ve Çörek, E. (2005). Müşteri Memnuniyetinde İstatistiksel Yöntemler ve Bir Uygulama. *Öneri Dergisi*. 6(23): 259-266.

Alabay, M.N. (2012). Müşteri Şikayetleri Yönetimi. *Uluslararası Yönetim İktisat ve İşletme Dergisi*. 8(16): 137-157.

Alpaydin, E. (2016). *Machine Learning: The New AI*. Cambridge: MIT Press.

Amasyali, M.F. ve Yildirim, T. (2004). Automatic Text Categorization of News Articles. *Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, 2004* (ss. 224-226), Düzenleyen IEEE Aydın. 30-30 Nisan 2004.

Arslan, H., Kaynar, O. ve Şahin, S. (2019). Classification of Customer Demands by Using Doc2Vec Feaure Extraction Method. *27th Signal Processing and Communications Applications Conference (SIU)* (ss. 1-4). Düzenleyen IEEE Sivas. 24-26 Nisan 2019.

Atalay, M. ve Çelik, E. (2017). Büyük Veri Analizinde Yapay Zekâ ve Makine Öğrenmesi Uygulamaları. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*. 9(22): 155-172.

Aytekin, Ç., Sütçü, C. S. ve Özfidan, U. (2018). Karar Ağacı Algoritması ile Metin Sınıflandırma: Müşteri Yorumları Örneği. *Uluslararası Sosyal Araştırmalar Dergisi*. 11(55): 782-792.

Barboza, F., Kimura, H. ve Altman, E. (2017). Machine Learning Models and Bankruptcy Prediction. *Expert Systems with Applications*. 83: 405-417.

Başkaya, F. ve Aydin, İ. (2017). Haber Metinlerinin Farklı Metin Madenciliği Yöntemleriyle Sınıflandırılması. *IDAP 2017: International Artificial Intelligence and Data Processing Symposium* (ss. 1-5). Düzenleyen IEEE Malatya. 16-17 Eylül 2017.

Bilgin, M. (2017). Gerçek Veri Setlerinde Klasik Makine Öğrenmesi Yöntemlerinin Performans Analizi. *19. Akademik Bilişim Konferansı* (ss. 1-6). Düzenleyen Aksaray Üniversitesi Aksaray. 8-10 Şubat 2017.

Bilgin, M. (2019). Türkçe metinlerin siniflandirma başarisini artirmak için yeni bir yöntem önerisi. *Uludağ University Journal of The Faculty of Engineering*. 24(1): 125-136.

Bircan, H. (2004). Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. *Kocaeli Üniversitesi Sosyal Bilimler Dergisi*. (8): 185-208.

Budak, H. (2018). Özellik seçim yöntemleri ve yeni bir yaklaşım. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*. 22: 21-31.

Budak, H. ve Erpolat, S. (2012). Kredi Riski Tahmininde Yapay Sinir Ağları ve Lojistik Regresyon Analizi Karşılaştırılması. *AJIT-e: Bilişim Teknolojileri Online Dergisi*. 3(9): 23-30.

Büyükeke, A., Sökmen, A. ve Gencer, C. (2020). Metin Madenciliği ve Duygu Analizi Yöntemleri ile Sosyal Medya Verilerinden Rekabetçi Avantaj Elde Etme: Turizm Sektöründe Bir Araştırma. *Journal of Tourism and Gastronomy Studies*. 8(1), 322-335.

Carey, M.J. ve Ceri, S (2008). *Data-Centric Systems and Applications*. Berlin: Springer

Chen, Q., Yao, L. ve Yang, J. (2016). Short text classification based on LDA topic model. *International Conference on Audio, Language and Image Processing (ICALIP)* (ss. 749-753). Düzenleyen IEEE Shanghai. 11-12 Temmuz 2016.

Çakmak Karapınar, D. ve Öcal, D. (2017). Halkla İlişkilerde Konu Yönetimi: Şikayet Var Sitesi Üzerinden Bir İnceleme. *Atatürk İletişim Dergisi*. (13): 103-124.

Çelik, H. ve Ekşi, H. (2008). Söylem analizi. *Marmara Üniversitesi Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi*. 27(27): 99-117.

Dangeti, P. (2017). *Statistics for Machine Learning*. Birmingham: Packt

Dil Bilimi (31 Mayıs 2021). *Öndeyiş*. http://www.dilbilimi.net/anasayfa.htm, (1 Haziran 2021)

Elektronik Ticaret Bilgi Platformu (2021). 2020 Yılı E-Ticaret Verileri Açıklandı. https://www.eticaret.gov.tr/haberler/10040/detay, (20 Nisan 2021).

Erickson, B.J., Korfiatis, P., Akkus, Z. ve Kline, T.L. (2017). Machine learning for medical imaging. *Radiographics*. 37(2): 505-515.

Göker, H. ve Tekedere, H. (2017). FATİH projesine yönelik görüşlerin metin madenciliği yöntemleri ile otomatik değerlendirilmesi. *Bilişim Teknolojileri Dergisi*. 10(3): 291-299.

Gültepe, Y. (2019). Makine Öğrenmesi Algoritmaları ile Hava Kirliliği Tahmini Üzerine Karşılaştırmalı Bir Değerlendirme. *Avrupa Bilim ve Teknoloji Dergisi*. (16): 8-15.

Hartmann, J., Huppertz, J., Schamp, C. ve Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*. 36(1): 20-38.

Hoşgör, H. ve Tosun, N. (2019). İstanbul İlindeki Bir Üniversite Hastanesine Yapılan Şikâyetlerin İncelenmesi. *Cumhuriyet Üniversitesi Edebiyat Fakültesi Sosyal Bilimler Dergisi*. 43(2): 239-252.

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*. 9(03): 90-95.

Hurwitz, J. ve Kirsch, D. (2018). *Machine Learning For Dummies*. Hoboken, NJ: 111 River St.

IBM (b.t.). *Doğrusal Regresyon*. https://www.ibm.com/tr-tr/analytics/learn/linear-regression, (5 Ocak 2021).

Ikonomakis, M., Kotsiantis, S. ve Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*. 4(8): 966-974.

Joty, S., Carenini, G., Ng, R. ve Murray, G. (2019). Discourse analysis and its applications. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts* (ss. 12-17). Düzenleyen Association for Computational Linguistics Florence. 28 Temmuz 2019.

Karapınar, D.Ç. (2017). Halkla İlişkilerde Konu Yönetimi Şikayet Var Sitesi Üzerinden Bir İnceleme. *Atatürk İletişim Dergisi*. 13: 103-124.

Kaşıkçı, T. ve Gökçen, H. (2013). Metin madenciliği ile e-ticaret sitelerinin belirlenmesi. *Bilişim Teknolojileri Dergisi*. 7(1): 25-32.

Kavzoğlu, T. ve Çölkesen, İ. (2010). Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi. *Harita Dergisi*. (76): 73-82.

Kazan, S. ve Karakoca, H. (2019). Makine Öğrenmesi ile Ürün Kategorisi Sınıflandırma. *Sakarya University Journal of Computer and Information Sciences*. 2(1), 18-27.

Keleş, M.B., Keleş, A. ve Keleş, A. (2020). Yapay Zekâ Teknolojisi ile Uçuş Fiyatı Tahmin Modeli Geliştirme. *Turkish Studies*. 15(4): 511-520.

Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M. ve Özçift, A. (2016). KNN algoritması ve R dili ile metin madenciliği kullanılarak bilimsel makale tasnifi. *Marmara Fen Bilimleri Dergisi*. 28(3): 89-94.

Kotsiantis, S.B., Zaharakis, I. ve Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*. 160(1): 3-24.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L. ve Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4): 150.

Küçük, D. ve Arıcı, N. (2018). Doğal Dil İşlemede Derin Öğrenme Uygulamalari Üzerine Bir Literatür Çalişmasi. *Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi*. 2(2): 76-86.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S. ve Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*. 18(8): 2674.

Liddy, E.D. (2001). Natural Language Processing. *Encyclopedia of Library and Information Science, 2nd Ed*. New York: Marcel Decker, Inc.

Lilleberg, J., Zhu, Y. ve Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)* (ss. 136-140), Düzenleyen IEEE Beijing. 6-8 Temmuz 2015.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G. ve Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*. 53(3): 950-965.

McKinney, W. (2020). the Pandas Development Team. *Pandas: Powerful Python Data Analysis Toolkit.* https://pandas. pydata. org/docs/pandas. Pdf, (Erişim tarihi: 10.02.2021).

Microsoft Azure (b.t.). *Azure'da Doğal Dil İşleme Teknolojisini Seçme*. https://docs.microsoft.com/tr-tr/azure/architecture/data-guide/technology-choices/natural-language-processing, (5 Ocak 2021).

Microsoft Azure (b.t.). *Makine Öğrenmesi Algoritmaları*. https://azure.microsoft.com/tr-tr/overview/machine-learning-algorithms/#overview, (3 Ocak 2021).

Montovan, G., Orr, G.B. ve Müller, K.R. (2012). *Neural Networks: Tricks of the Trade.* Berlin: Springer.

Numpy (b.t.). https://numpy.org/about/, (20.01.2021).

Oğuzlar, A. (2011). *Temel Metin Madenciliği*. Bursa: Dora.

Ovalı, E. (2014). Müşteri İlişkileri Yönetimi Programının Bir Parçası Olarak Satış Görüşmelerinde Müşteri Memnuniyetinin Ölçülmesi Üzerine Bir Pilot Araştırma. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*. 2(4): 1-14.

Öztürk, A., Durak, Ü. ve Badıllı, F. (2020). Twitter Verilerinden Doğal Dil İşleme ve Makine Öğrenmesi ile Hastalık Tespiti. *Konya Mühendislik Bilimleri Dergisi*. 8(4): 839-852.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. ve Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research.* 12: 2825-2830.

Polat, S. (2017). Yazılım Hata Kayıtlarının Makine Öğrenmesi Yöntemleriyle Kümelenerek, Hataya Sebep Olan Bileşenlerin Tespit Edilmesi. *Proceedings of the 11th Turkish National Software Engineering Symposium* (ss. 444-453), Düzenleyen Atılım Üniversitesi Antalya. 18-20 Ekim 2017.

Pypi (2018). *TurkishStemmer 1,3* https://pypi.org/project/TurkishStemmer/, (10.01.2021).

Sağbaş, E. ve Ballı, S. (2016). Akıllı Telefon Algılayıcıları ve Makine Öğrenmesi Kullanılarak Ulaşım Türü Tespiti. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*. 22(5): 376-383.

Santur, Y. (2020). Derin Öğrenme ve Aşağı Örnekleme Yaklaşımları Kullanılarak Duygu Sınıflandırma Performansının İyileştirilmesi. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*. *32*(2): 561-570.

Saritas, M.M. ve Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*. 7(2): 88-91.

Seker, S.E. (Aralık 2015). Doğal Dil İşleme (Natural Language Processing). http://ybsansiklopedi.com/wp-content/uploads/2015/09/ybsddi.pdf, (19.02.2021)

Şengöz, N. ve Özdemir, G. (2016). Temel Bileşenler Analizi ve K-Ortalama Kümeleme Yönteminin Birlikte Kullanımı: Bir Örnek Uygulama-Combined Use Of Principal Component Analysis And K-Clustering Method: A Case Study. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*. 8(15): 85-94.

Tantuğ, A.C. (2016). Metin Sınıflandırma. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 5(2): 1-12.

Taşkın, G. (2005). Veri madenciliğinde karar ağaçları ve bir satış analizi uygulaması. *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi*. 6(2): 221-239.

Tekin, M.C. ve Tunalı, V. (2019). Yazılım geliştirme taleplerinin metin madenciliği yöntemleriyle önceliklendirilmesi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*. 25(5): 615-620.

Tokay Argan, M. ve Arıcı, A. (2019). Sağlık İletişiminde E-Şikayet: Hastaların ve Hasta Yakınlarının Sikayetvar.com Paylaşımları Üzerine Bir Araştırma. *Akdeniz Üniversitesi İletişim Fakültesi Dergisi*. (31): 339-355.

Tosun, N. ve Söyük, S. (2019). Şikayet Yönetimi Perspektifinden Müşteri Memnuniyeti ve Davranışsal Niyet Üzerine Kavramsal Bir Model Önerisi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*. (23) Özel Sayı: 1873-1887.

Tretyakov, K. (2004, May). Machine learning techniques in spam filtering. *Data Mining Problem-oriented Seminar MTAT.03.177* (ss. 60-79).

Tüfekci, P. ve Uzun, E. (2013). Author detection by using different term weighting schemes. *2013 21st Signal Processing and Communications Applications Conference (SIU)* (ss. 1-4), Düzenleyen IEEE Haspolat. 24-26 Nisan 2013.

Uğuz, S. (2021). *Makine Öğrenmesi Teorik Yönleri ve Python Uygulamaları İle Bir Yapay Zeka Ekolü*. Ankara: Nobel Yayıncılık.

Uylaş Satı, N. (2018). A collective learning approach for semi-supervised data classification. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*. 24(5): 864-869.

Uysal, A.K., Günal, S., Ergin, S. ve Günal, E.Ş. (2012, April). Detection of SMS spam messages on mobile phones. *2012 20th Signal Processing and Communications Applications Conference (SIU)* (ss. 1-4). Düzenleyen IEEE Muğla. 18-20 Nisan 2012.

Wikipedia (b.t.). *Gradient descent*. https://en.wikipedia.org/wiki/Gradient_descent, (14 Mayıs 2021).

Xuan, J., Jiang, H., Ren, Z., Yan, J. ve Luo, Z. (2017). Automatic bug triage using semi-supervised text classification. *arXiv preprint arXiv:1704.04769*.

Yakut, E., Elmas, B. ve Yavuz, S. (2014). Yapay Sinir Ağları ve Destek Vektör Makineleri Yöntemleriyle Borsa Endeksi Tahmini. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*. 19(1): 139-157.

Yıldız, K., Çamurcu, Y. ve Doğan, B. (2010). Veri madenciliğinde temel bileşenler analizi ve Negatifsiz matris çarpanlarına ayırma tekniklerinin karşılaştırmalı analizi. *Akademik Bilişim'10 - XII. Akademik Bilişim Konferansı Bildirileri* (ss. 207-213). Düzenleyen Muğla Üniversitesi Muğla. 10-12 Şubat 2010.

Yıldız, O. (2017). Metin madenciliğinde anahtar kelime seçimi bir üniversite örneği. *Yönetim Bilişim Sistemleri Dergisi*. *2*(1): 29-50.

EKLER

EK 1: Kodlar

```
pd.set_option('display.max_colwidth', -1)

data = pd.read_csv("bitti//learn_3.csv", delimiter=';')

stop_words=pd.read_csv('https://raw.githubusercontent.com/InJuxSanct/turkish-stopwords/master/src/lib/stopwords/raw-stopwords.txt', sep=" ", header=None)
pat2 = r'\b(?:{})\b'.format('|'.join(list(stop_words[0].str.lower())))
data['Complaint2'] = data['Complaint'].str.lower().str.replace(pat2, '').str.replace('  ', ' ')

stemmer = TurkishStemmer()
data['text_finalst'] = data['text_final'].apply(lambda x : " ".join([stemmer.stem(w) for w in x.split()]))

x_train, x_test, y_train, y_test = train_test_split(data['text_finalst'], data['Category'])
tfidf_vectorizer = TfidfVectorizer(ngram_range=(1,2), max_df=0.9, min_df=5)
words_counts = Counter([word for line in x_train for word in line.split(' ')])
most_common_words = sorted(words_counts.items(), key=lambda x: x[1], reverse=True)[:20]

model = OneVsRestClassifier(LogisticRegression(penalty = 'l2', C=1.0))
model.fit(x_train_tfidf, y_train)
filename = "bitti//vb.csv"
col_list = ["new_id","date","time","comm"]
df = pd.read_csv(filename, usecols=col_list, delimiter=";")
only_comm = open('bitti//vb_only_comm.txt', 'w')

model2 = OneVsRestClassifier(SGDClassifier(loss = 'hinge', penalty = 'elasticnet', max_iter = 20))
model2.fit(x_train_tfidf, y_train)

SVC_pipeline = Pipeline([('clf', OneVsRestClassifier(LinearSVC(), n_jobs=1)),])
SVC_pipeline.fit(x_train_tfidf, y_train)
prediction = SVC_pipeline.predict(x_test_tfidf)
```