DÜZCE ÜNİVERSİTESİ | LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

TÜRKÇE SPAM MAİLLERİN DUYGU ANALİZİ VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE ANALİZİ

YUNUS EMRE PALAVAR

YÜKSEK LİSANS TEZİ
SİBER GÜVENLİK ANABİLİM DALI

DANIŞMAN
DR. ÖĞR. ÜYESİ AHMET ALBAYRAK

DÜZCE, 2024

T.C.
DÜZCE ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

TÜRKÇE SPAM MAİLLERİN DUYGU ANALİZİ VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE ANALİZİ

Yunus Emre PALAVAR tarafından hazırlanan tez çalışması aşağıdaki jüri tarafından Düzce Üniversitesi Lisansüstü Eğitim Enstitüsü Siber Güvenlik Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Dr. Öğr. Üyesi Ahmet ALBAYRAK

Düzce Üniversitesis

Jüri Üyeleri

Dr. Öğr. Üyesi Ahmet ALBAYRAK
Düzce Üniversitesi _____

Dr. Öğr. Üyesi Enver KÜÇÜKKÜLAHLI
Düzce Üniversitesi _____

Dr. Öğr. Üyesi Yasin ORTAKCI
Karabük Üniversitesi _____

Tez Savunma Tarihi: 04/06/2024

BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanmasından yazımına kadar bütün aşamalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları da kaynaklar listesine aldığımı, yine bu tezin çalışılması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

4 Haziran 2024
YUNUS EMRE PALAVAR

TEŞEKKÜR

Yüksek lisans öğrenimimde ve bu tezin hazırlanmasında gösterdiği her türlü destek ve yardımdan dolayı çok değerli hocam Dr. Öğr. Üyesi Ahmet ALBAYRAK'a en içten dileklerimle teşekkür ederim.

Bu çalışma boyunca yardımlarını ve desteklerini esirgemeyen sevgili aileme ve çalışma arkadaşlarıma sonsuz teşekkürlerimi sunarım.

4 Haziran 2024 **YUNUS EMRE PALAVAR**

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

ÖZET

TÜRKÇE SPAM MAİLLERİN DUYGU ANALİZİ VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE ANALİZİ

Yunus Emre PALAVAR

Düzce Üniversitesi
Lisansüstü Eğitim Enstitüsü, Siber Güvenlik Anabilim Dalı
Yüksek Lisans Tezi
Danışman: Dr. Öğr. Üyesi Ahmet ALBAYRAK

Haziran 2024, 42 sayfa

Çevrimiçi platformların kullanımının artmasıyla birlikte metin verilerinin hacmi artmakta ve bu verilere erişim kolaylaşmaktadır. Bu durum metin sınıflandırma alanında yapılan çalışmaların sayısının artmasına neden olmuştur. Özellikle spam tespiti ve duygu analizi gibi alanlarda metin sınıflandırma teknikleri büyük önem taşımaktadır. Literatürde İngilizce metinler üzerine yapılan çalışmaların sayısı oldukça fazla olmasına karşın Türkçe veriler üzerine yapılan çalışmalar oldukça kısıtlıdır. Bu çalışmanın amacı, Türkçe maillerin duygu analizi ve makine öğrenmesi teknikleri ile morfolojik analizini gerçekleştirmek ve modellerin spam ve normal mailleri tespit etmedeki başarısını karşılaştırmaktır. Bu amaçla literatürde yer alan iki Türkçe mail veri seti kullanılmıştır. Bu veri setleri spam ve normal olarak etiketlenmiş maillerden oluşmaktadır. Çalışma kapsamında bu iki veri setinden bir veri seti elde edilmiştir. Bu veri kümesine üç işlem uygulanmıştır ve bu uygulanan işlemler sonucu üç adet veri seti elde edilmiştir. İlk veri seti, verilere temel veri ön işleme adımları uygulanarak oluşturulmuştur. Bu adımda sırasıyla veriler küçük harflere dönüştürülmüştür. Daha sonra web sitesi adları "website" ve mail adresleri "email" olarak yeniden adlandırılmıştır. Buna ek olarak ilk veri seti olarak noktalama işaretleri ve sayısal ifadelerin kaldırılması elde edilmiştir. İkinci veri seti, ilk veri setinden Türkçe kökenli olmayan kelimeler ve dört harften kısa sözcüklerle oluşurulmuştur. Üçüncü veri seti ise ikinci veri seti ile birinci veri setinin kesişim kümesinden elde edilmiştir. Bu çalışma ile literatürde yer alan çalışmalarda etkisi göz ardı edilen Türkçe veri setleri içerisindeki Türkçe kökenli olmayan kelimelerin de sonuçlar üzerindeki etkisi gözlemlenmiştir. Bu veri kümeleri K-means ve Isolation Forest yöntemleri ile kümelenmiş ve bu yöntemlerin performansı değerlendirilmiştir. Ayrıca bu veri kümeleri üzerinde duygu analizi yapılarak spam ve normal maillerin duygu durumları gözlemlenmiştir. Son olarak veriler Naive Bayes, Random Forest, Logistic Regression ve Support Vector Machine sınıflandırma algoritmaları ile sınıflandırılmış ve yöntemlerin sonuçları doğruluk, kesinlik, geri çağırma ve f1-skor kriterleri ile değerlendirilmiştir. Çalışma sonucunda en yüksek başarım puanlarına ilk veri seti ile ulaşılmıştır. Naive Bayes ve destek vektör makinesi 0,92 doğruluk değeri ile en başarılı sonucu verirken Lojistik Regresyon ile 0,90 ve Random Forest ile 0,89 doğruluk değerleri elde edilmiştir. K-means ve Isolation Forest, orijinal veri kümesindeki etiketlere kıyasla verileri etiketlemede yetersiz kalmıştır. Yapılan işlemler sonucunda maillerin kategorik morfolojisi çıkarılmıştır.

Anahtar Sözcükler: Spam Mail, Duygu Analizi, Makine Öğrenmesi, Morfolojik Analiz

ABSTRACT

ANALYSIS OF TURKISH SPAM MAILS WITH SENTIMENT ANALYSIS AND MACHINE LEARNING METHODS

Yunus Emre PALAVAR
Düzce University
Graduate School, Department of Cyber Security
Master Thesis
Supervisor: Asst. Prf. Ahmet ALBAYRAK

June 2024, 42 pages

The increasing use of online platforms has led to a growth in the volume of text data, and access to these data has become easier. This has resulted in a rise in the number of studies conducted in the field of text classification. Text classification techniques are particularly important in areas such as spam detection and sentiment analysis. While there is a significant number of studies on English texts in the literature, studies on Turkish data are quite limited. The aim of this study is to perform sentiment analysis and morphological analysis of Turkish emails using machine learning techniques, and to compare the success of the models in detecting spam and normal emails. For this purpose, two Turkish email datasets from the literature, which are labeled as spam and normal, were used. A single dataset was obtained from these two datasets. Three processing steps were applied to this dataset, resulting in three datasets. The first dataset was created by applying basic data preprocessing steps to the data, such as converting to lowercase, renaming website names to "website" and email addresses to "email", and removing punctuation marks and numerical expressions. The second dataset was created from the first dataset by removing non-Turkish words and words shorter than four characters. The third dataset was obtained from the intersection of the second and first datasets. This study observed the impact of non-Turkish words in Turkish datasets, which has been overlooked in previous studies. These data sets were clustered using K-means and Isolation Forest methods, and the performance of these methods was evaluated. Additionally, sentiment analysis was performed on these data sets to observe the sentiment states of spam and normal emails. Finally, the data was classified using Naive Bayes, Random Forest, Logistic Regression, and Support Vector Machine classification algorithms, and the results of the methods were evaluated using accuracy, precision, recall, and F1-score criteria. As a result of the study, the highest performance scores were achieved with the first dataset. Naive Bayes and Support Vector Machines achieved the most successful results with an accuracy of 0.92, while Logistic Regression achieved 0.90 and Random Forest achieved 0.89 accuracy. K-means and Isolation Forest were insufficient in labeling the data compared to the original dataset labels. As a result of the performed operations, the categorical morphology of the emails was extracted.

Keywords: Spam Mail, Sentiment Analysis, Machine Learning, Morphological Analysis

1. GİRİŞ

E-mail ya da diğer değişle mail, internet üzerinden gönderilen dijital mektuplardır. E-mail, geleneksel postaya nispeten daha ucuz, daha pratik ve daha hızlı olduğundan günlük hayatımızda geleneksel postanın yerini almaktadır. E-mail üzerinden her türlü özel ve resmi yazışmalar yapılmaktadır. Günümüzde bu yazışmaların yanında istenmeyen e-maillere maruz kalınmaktadır. 2023 yılı itibariyle 4.2 milyar mail kullanıcısı varken, günlük olarak alınan ve gönderilen mail sayısı ise 333 milyar olmuştur. Bu gönderilen maillerin %56.5 spam olarak gruplanmaktadır [1]. Bu istenmeyen e-mailler ise spam mail olarak adlandırılmaktadır [3]. Bu spam maillerin kullanıcılar üzerinde olumsuz etkileri mevcuttur. Bunlardan bazıları aldatmak, dolandırmak, sanal zorbalık ya da sanal hırsızlık vb. olarak sıralanabilir [5],[7]. Kullanıcıların bu durumlardan korunması amacıyla mail sağlayıcıları bu spam maillerin engellemesine yönelik çalışmalar gerçekleştirmektedir [3]. Literatürde yer alan çalışmalarda yabancı kaynaklı mailler üzerinde yapılan çalışmalar yaygınken Türkçe kaynaklı mailler üzerinde yapılan çalışmalar daha kısıtlıdır [2].

Bu çalışma kapsamında spam ve normal maillerde oluşan veri setleri üç farklı adımda incelenmiştir. İlk adımda klasik veri önişleme işlemleri yapılmıştır. İkinci adımda ilk adıma ek olarak verilerden Türkçe kökenli olmayan kelimeler, bağlaçlar ve dört kelimeden kısa cümleler çıkarılmıştır. Üçüncü adımda ise ilk adımda elde edilen veri seti ile ikinci adımda elde edilen veri setinin kesişim kümesinden yeni bir veri seti elde edilmiştir. Üç adım sonucunda elde edilen üç veri seti K-means, Isolation Forest, duygu analizi, Naive Bayes (NB), Random Forest (RF), Lojistik Regresyon (LR) ve destek vektör makinesi (DVM) yöntemleri ile analiz edilmiştir. Bu çalışma, literatürde yer alan Türkçe veri seti kullanılan çalışmalardan farklı olarak, Türkçe veri setlerinin içerisindeki yer alan Türkçe kökenli olmayan kelimelerde analiz edilmiştir. K-means, Isolation Forest, duygu analizi, NB, RF, LR ve DVM yöntemleri ile hem geleneksel veri ön işleme adımları uygulanan veriler değerlendirilmiş hem de geleneksel yöntemlere ek olarak Türkçe kökenli olmayan kelimelerin çıkarılması ile elde edilen veri setleri de değerlendirilmiştir.

2. LİTERATÜR İNCELEMESİ

Türkçe e-maillerde spam içeren e-mailleri tespiti için Lojistik Regresyon, Naive Bayes, Rastgele Orman, Yapay Sinir Ağları makine öğrenme yöntemleri ve ALBERT, BERT, DistilBERT, ELECTRA, dil modelleri analiz edilmiştir [2]. Makine öğrenmesi tekniklerinden yapay sinir ağları %90.15 doğruluk değeri elde etmiştir. Çalışma sonucunda en başarılı dil modelleri %94.08 doğruluk değeri ile ELECTRA ve BERT olmuştur [2].

Eryılmaz ve Kılıç (2020) yaptıkları çalışmada, spam mailler filtrelenmesi için literatürde kullanılan yöntemleri, yapay zekâ tabanlı olan ve yapay zekâ tabanlı olmayan olarak sınıflandırarak incelemişlerdir. Çalışma sonucunda eskiden yapay zekâ tabanlı olmayan sistemler etkili iken günümüzde yapay zekâ tabanlı sistemler daha yaygın olduğu gözlemlenmiştir. Aynı zamanda Türkçe spam mail tespiti uygulama geliştiricilerinin en büyük sıkıntısının, Türkçe maillerden oluşan veri setlerinin azlığı olduğunu belirtmişlerdir [3].

Özdemir, Ataş & Özer (2013) yapay bağışıklık algoritmaları ile spam maillerin tespit edilmesi üzerine çalışmışlardır. AIRS1 (Artificial Immune Recognition System 1), AIRS2 (Artificial Immune Recognition System 2), AIRS2PARALLEL (Parallel Artificial Immune Recognition System 2), CLONALG (Clonal Selection Algorithm) ve CSCA (Crow Search Algorithm) algoritmaları bu çerçevede incelenmiştir. Bu algoritmalar arasında en yüksek sınıflandırma başarısına ulaşan algoritma %86 sınıflandırma başarısı ile CSCA olmuştur [4].

Karim, Azam, Shanmugam, Kannoorpatti & Alazab (2019) yaptıkları çalışmada spam mail tespitinde Yapay Zeka ve Makine Öğrenmesi yöntemlerine odaklanmış bir literatür araştırmasını açıklamışlardır. Çalışmanın sonucunda DVM ve NB algoritmalarının yüksek talep gördüğü ve tek algoritmalı anti-spam uygulamalarının yaygın olduğu gözlemlenmiştir [30].

Duygu Analizi'ni kullanarak spam e-mail tespiti öneren bir çalışmada 5,572 mesajdan oluşan veri setini Bag of words, Hashing ve Long short-term memory algoritması (LSTM) ile test etmişlerdir. LSTM ile test edilen veri setinden %98 doğruluk elde edilmiştir [6].

NB, K-NN (K- Nearest Neighbor), yapay sinir ağı (YSA), DVM, yapay bağışıklık sistemi ve Rough set sınıflandırma algoritmaları açıklamışlar ve SpamAssassin spam korpusu üzerinde performanslarının karşılaştırmasını yapmışlardır. Altı yöntem arasından en yüksek doğruluk değerine NB ve Rough set sahiptir. Çalışmanın sonucunda hibrit sistemlerin şuan da başarılı bir anti-spam filtresi oluşturmanın en etkili yolu olduğu ortaya konulmuştur [35].

[36] yapmış oldukları çalışmada NB, SVM, K-NN, Decision Tree (DT), Bagging, RF ve Boosting ve Adaboost makine öğrenmesi algoritmaları ile Kaggle üzerinden elde ettikleri veri seti üzerinde sınıflandırma işlemlerini gerçekleştirmişlerdir. Elde edilen bu verilerin, kullanılan makine öğrenmesi yöntemleri ile spam mı normal mı olduğu tespit edilmeye çalışılmıştır. Çalışma sonucunda Multinomial Naïve Bayes en iyi sonucu vermiştir [36].

Yaseen (2021) makine öğrenmesi yöntemlerinden NB ve K-NN ile derin öğrenme yöntemlerinden bert-base-cased, BiLSTM (çift yönlü Uzun Kısa Süreli Bellek) performansları karşılaştırmıştır. Çalışma da kullanılan bert-base-cased, BERT (Transformatörlerden Çift Yönlü Kodlayıcı Gösterimleri) doğal dil işleme modelinin bir türüdür. Çalışma kapsamında Biri modeli eğitmek için, diğeri ise modelin görünmeyen verilere karşı kalıcılığını ve sağlamlığını test etmek için iki açık kaynaklı veri seti kullanılmıştır. Çalışma sonucunda %98,66 doğruluk ve %98,66 F1 puanıyla en iyi model bert-base-cased olmuştur [37].

Sharma ve Bhardwaj (2018) spam mail tespiti için Naive Bayes ve J48 (karar ağacı) olmak üzere iki makine öğrenmesi algoritmasını uygulayarak makine öğrenimi tabanlı bir hibrit torbalama yaklaşımı önermişlerdir. Çalışma kapsamında veri seti farklı veri setlerine bölünerek her algoritmaya girdi olarak verilmiştir. Toplam üç deney yapılarak elde edilen sonuçlar recall, precision, f-measure, accuracy, yanlış pozitif oranı, gerçek negatif oranı ve yanlış negatif oranı açısından karşılaştırılmıştır. İki deneyde NB ve J48 algoritmaları kullanılmıştır. Üçüncü deneyde, hibrit torba yaklaşım kullanılarak uygulanan önerilen SMD (spam mesaj tespit) sistemidir. NB ve J48 algoritmasının elde ettiği doğruluk sırasıyla %83,5 ve %91,5'tir. Önerilen sistem olan SMD ise %87,5 genel doğruluk yalnızca J48 algoritması üzerinde gerçekleştirildiğinde deneysel sonuçların daha iyi oldu saptanmıştır [38].

[39] yapmış oldukları çalışmada mobil cihaz iletişiminde spam ve normal mesajların sınıflandırılması için LR, K-NN ve DT makine öğrenmesi yöntemleri kullanılmıştır.

Çalışma kapsamında kullanılan veri seti kaggle elde edilmiştir. Veri seti iki etiketten oluşmaktadır, mesajın v1 spam ya da normal olduğu belirtirken, v2 mesaj metnini içerir. Veri setinde 4900 spam olmayan ve 672 spam mesaj olmak üzere toplamda 5572 adet veri vardır. Çalışma sonucunda önerilen yöntem olan Lojistik Regresyon %99 doğruluk oranına ulaşmıştır [39].

Siddique vd. (2021) içeriği Urdu dilinde yazılmış spam maillerin tespiti üzerine çalışmışlardır. Spam mail tespiti için makine öğrenmesi yöntemlerinden olan NB, Convolutional Neural Network (CNN), DVM, LSTM kullanmışlardır. Çalışma kapsamında kullanılan veri seti 5000 mailden oluşmaktadır ve kaggle'dan elde edilmiştir. Çalışma sonucunda DVM algoritması %97,5 doğruluk, NB %98 doğruluk, CNN %96,2 doğruluk ve LSTM %98,4 doğruluk değerlerine ulaşılmıştır. Urduca spam mailleri tespit etmek için %98,4'lük doğrulukla LSTM daha güçlü bir yöntem olduğu sonucuna ulaşılmıştır [40].

[41] yapmış oldukları çalışmada Short Message Service (SMS) spam tespiti alanında önceki araştırma çalışmalarında toplanan farklı veri setleri üzerinde farklı sınıflandırma tekniklerini karşılaştırmışlardır. Çalışma kapsamında kullanılan veri seti ücretsiz araştırma kaynaklarından elde edilmiştir ve iki kümede toplanmıştır. İlk veri kümesi Singapur Ulusal Üniversitesi SMS Corpus (3375 Ham SMS), Grumbletext Web Sitesi (425 Spam SMS), Caroline Tag'in Doktora Tezleri (450 Ham SMS) ve SMS Spam Corpus v.0.1 Big (1002 Ham SMS ve 322 Spam SMS) bu veri koleksiyonlarının bir araya getirilmesi ile oluşturulmuştur. İkinci veri kümesi 1000 spam ve 1000 normal SMS'ten oluşur. Elde edilen veri seti 8 farklı sınıflandırma algoritması ile test edilmiştir ve sonuçlar doğruluk, kesinlik, duyarlılık, AR (doğruluk oranı) ve pozitif oranı parametreleri ile değerlendirilmiştir. Bu sınıflandırma algoritmaları sırasıyla DVM, NB, DT, LR, RF, AdaBoost, Artifical Neural Network (ANN) ve CNN'dir. İlk veri kümesi için DVM %98,57 doğruluk ve 0,9846 AR, NB %98,48 doğruluk ve 0,9892 AR, DT %96,05 doğruluk ve 0,8853 AR, LR %98,65 doğruluk ve 0,9891 AR, RF %98,65 doğruluk ve 0,9883 AR, AdaBoost %97,85 doğruluk ve 0,9645, ANN %98,39 doğruluk ve 0,9667 AR, CNN %99,10 doğruluk ve 0,9926 AR değerlerine ulaşmıştır. İkinci veri kümesi için ise DVM %96,25 doğruluk ve 0,9689 AR, NB %96,75 doğruluk ve 0,9904, DT %91,25 doğruluk ve 0,8354 AR, RF %94,25 doğruluk, ve 0,9868, LR %96,25 doğruluk ve 0,9878 AR, AdaBoost %92,50 doğruluk ve 0,9294 AR, ANN %98 doğruluk ve 0,9918 AR, CNN %98,25 doğruluk ve 0,9994 AR değerlerine ulaşmıştır. Sonuç olarak iki veri kümesi için

en yüksek doğruluk ve AR değerlerine CNN ulaşmıştır. Ancak çalışma kapsamında AR değerinin risk modellerinin ayırt edici gücünü belirlemede daha değerli olduğu belirtilmiştir. [41].

Nandhini ve K.S: (2020) bir mailin spam mail mı yoksa normal mail mı olduğunun tespiti için makine öğrenmesi yöntemlerini kullanmışlardır. Çalışma kapsamında Lojistik Regresyon (LR), Naive Bayes (NB), Decision Tree (DT), K-nearest neighbor (K-NN) ve destek vektör makinesi (DVM) makine öğrenmesi algoritmaları kullanılmıştır. WEKA aracı veri setini eğitmek ve test etmek için kullanılmıştır. Çalışma kapsamında University of California (UCI) Machine Learning Repository Spambase veri seti kullanılmıştır. Çalışma sonucunda LR: %93,1319 doğruluk, 0,931 kesinlik, 0,931 duyarlılık, 0,931 F1 puanı, DT: %99,9348 doğruluk, 0,999 kesinlik, 0,999 duyarlılık, 0,999 F1 puanı, NB: %79,5262 doğruluk, 0,845 kesinlik, 0,795 duyarlılık, 0,797 F1 puanı, KNN: %99,9348 doğruluk, 0,999 kesinlik, 0,999 duyarlılık, 0,999 F1 puanı, DVM: %90,7629 doğruluk, 0,117 kesinlik, 0,908 duyarlılık, 0,907 F1 puanı değerlerine ulaşmıştır. Çalışma sonucunda Decision Tree ve KNN aynı sonucu vermiş ve tüm performans metrikleri açısından diğer sınıflandırma algoritmalarına göre başarılı olmuştur ancak KNN aynı sonucu vermesine rağmen modeli oluşturmak DT modelini oluşturmaktan fazla zaman almıştır [42].

Kontsewaya, Antonov & Artamonov (2021) spam tespit etmek için bir sınıflandırıcı kullanarak spam miktarını azaltmayı amaçlamışlardır. Çalışma kapsamında K-NN, NB, DVM, DT, LR ve RF makine öğrenmesi algoritmaları kullanılmıştır. Çalışma kapsamında 4360 spam ve 1368 normal e-mail olmak üzere toplam 5728 e-mailden oluşan hazır veri seti kullanılmıştır. Çalışma sonucunda K-NN: 0,90 doğruluk, 0,91 kesinlik, 0,63 duyarlılık, 0,74 F1 puanı, Receiver Operating Characteristic (ROC) 0,95, NB: 0,99 doğruluk, 0,97 kesinlik, 0,99 duyarlılık, 0,98 F1 puanı, ROC 1,00, DT: 0,94 doğruluk, 0,82 kesinlik, 0,96 duyarlılık, 0,88 F1 puanı, ROC 0,95, DVM: 0,98 doğruluk, 0,98 kesinlik, 0,95 duyarlılık, 0,96 F1 puanı, ROC 1,00, RF: 0,84 doğruluk, 1,00 kesinlik, 0,28 duyarlılık, 0,42 F1 puanı, ROC 0,99, LB: 0,99 doğruluk, 0,98 kesinlik, 0,96 duyarlılık, 0,97 F1 puanı, ROC 0,95 değerlerine ulaşmıştır. En başarılı modeller LR ve NB olmuştur [43].

[44] NB, DVM, RF, DT ve MLP kullanılarak yedi farklı mail veri kümesinde özellik çıkarma ve ön işlemenin yanı sıra makine öğrenimi modellerini uygulayarak kapsamlı bir çalışma yapmışlardır. Sınıflandırıcıların performansını optimize etmek için Particle

Swarm Optimization (Parçacık Sürü Optimizasyonu) ve Genetic Algorithm (Genetik Algoritma) gibi biyolojik temelli algoritmalar kullanılmıştır. Çalışma kapsamında önerilen modeller yaklaşık 50000 mail ile test edilmiştir. WEKA kullanılarak 14 algoritma bu veriler ile test edilmiştir. Çalışma sonucunda en başarılı yöntem Multinomial Naive Bayes (MNB) olmuştur [44].

Trivedi (2016) düşük yanlış pozitif oranıyla iyi doğruluk sağlayan verimli ve hassas bir sınıflandırma modeli oluşturarak normal mailler ile spam mailleri birbirinden ayırmayı amaçlamışlardır. Çalışma kapsamında Enron mail veri kümesinin bilgilendirici özelliklerini araştırmak için Greedy Stepwise arama yöntemi kullanılmıştır. Çalışma kapsamında karşılaştırma farklı makine öğrenimi sınıflandırıcıları (Bayesian, NB, DVM, J48, Adaboost ile Bayesian, Adaboost ile NB gibi) arasında yapılmıştır. Çalışma sonuçlarının değerlendirilmesi için F1 skoru ve FP Oranı (Yanlış Pozitif Oranı) parametrelerinden yaralanılmıştır. Çalışma sonucunda DVM'nin bu çalışmanın en başarılı sınıflandırıcı olduğu saptanmıştır [45].

[46] maildeki spam mesajları sınıflandırmak için farklı öğrenme algoritmaları kullanılarak karşılaştırmalı bir analiz yapmışlardır. Çalışma kapsamında kullanılan veri seti, iki ay boyunca çeşitli e-maillerden elde edilmiştir. Spam maillerin 57 özelliği tanımlanmış veri seti içerisinde kullanılmıştır. Bu özelliklerden birkaçı: alıcı ve gönderici adresi, spam türü, spamın alındığı kuruluştur. Makine öğrenimi tekniklerine ilişkin veri seti UCI Machine Learning Repository elde edilmiştir. UCI'den toplanan spam veri seti 2788 normal mail ve 1813 spam mail olmak üzere toplamda 4601 mailden oluşmaktadır. Çalışma kapsamında WEKA aracı kullanılarak J48, MLP ve NB makine öğrenmesi algoritmaları ile sınıflandırma işlemi yapılmıştır. Çalışma sonucunda MLP en yüksek performansı sergileyen algoritma olmuştur. Çalışmanın neticesinde performansı artırmak ve daha iyi sonuç üretebilmek için NB Filtered Bayesian Learning algoritması kullanılmıştır [46].

Aski ve Sourati (2016) MLP kullanarak geçerli maillerden gelen istenmeyen mailleri düşük hata oranları ve yüksek verimlilikle filtrelemek için üç makine öğrenmesi algoritmasını açıklamışlardır. Çalışma kapsamında C4.5 DT, NB ve MLP kullanılmıştır. Çalışma sonucunda NB %98,6 doğruluk, J48 %96,6 doğruluk, MLP %99,3 doğruluk sonuçlarına ulaşmıştır [47].

Bassiouni, Ali & El-Dahshan (2018) spam mailleri sınıflandırmışlardır. Hangi

sınıflandırıcının daha iyi sonuç verdiği değerlendirmek için veri setine on alternatif sınıflandırıcı karşılaştırılmıştır. Doğruluğu sağlamak için 10 kat çapraz doğrulama kullanılmıştır. Çalışma kapsamında UCI ve Enron veri setleri kullanılmıştır. Çalışma kapsamında kullanılan sınıflandırıcılar: RF, ANN, LR, DVM, RT, Decision Table, KNN, Bayes Net, NB ve Radial Basis Function (RBF). Çalışma sonucunda RF 95,4, ANN 92,4, LR 92,4, DVM 91,8, RT 91,5, KNN 90,7, Decision Table 90,3, Bayes Net 89,8, NB 89,8 ve RBF 82,6 doğruluk sonuçlarına ulaşılmıştır. En yüksek performans gösteren sınıflandırıcı RF olmuştur [48].

Charanarur vd. (2023) Windows 10 platformundaki Internet Explorer, Firefox ve Chrome geçmiş verilerini, önbellek verilerini, oturum geri yüklemelerini, flash yapıtlarını ve süper çerezleri incelemişlerdir. Çalışmada önerilen sistemde spam maillerin daha verimli bir şekilde tanımlanması ve sınıflandırılması için iki farklı filtreleme modeli kullanılmıştır. İlk yaklaşım, kullanıcının güvenilirliğini mail adresine göre değerlendirmek için tasarlanmış bir mekanizma olan Opinion Rank (Görüş Sıralaması) olarak bilinen bir kavramın kullanılmasını içerir. İkinci yaklaşım ise ilk yaklaşımın üzerine Latent Dirichlet Allocation eklenmesi ile oluşturulmuştur. Çalışma kapsamında KN, NB, ETC, RF, SVC AdaBoost, xgb, LR, GBDT, BgC ve DT olmak üzere 11 makine öğrenmesi algoritması kullanılmış ve sonuçlar karşılaştırılmıştır. Çalışma sonucunda KN ve NB algoritmaları, en yüksek doğruluk ve hassasiyet sonuçları ulaşmıştır [49].

Panigrahi (2012) maillerin içeriğini dikkate alarak, mevcut bir veri küme üzerini öğrenerek bir mailin spam olup olmadığı tahmin edebilecek bir sınıflandırma modeli geliştirmiştir. Çalışma kapsamında kullanılan veri seti UCI Machine Learning Repository elde edilmiştir. Çalışma kapsamında ANN, DVM, C 5.0 DT, Bayes Net, CRT, NB, Lazy-IBK, Lazy-LWL, Lazy-Kstar, J48 ve RF makine öğrenmesi yöntemleri kullanılmış ve sonuçlar karşılaştırılmıştır. Çalışma sonucunda en yüksek doğruluk değerine sahip makine öğrenmesi yöntemi ANN olmuştur [50].

Yu ve Xu (2008) NB, Neural Network (NN), DVM ve Relevance Vector Machine (RVM) makine öğrenmesi algoritmalarını kullanarak, SpamAssassin ve Babletext karşılaştırmalı spam filtreleme derlemleri üzerinde ampirik bir çalışma yapmışlardır. Bu makine öğrenmesi algoritmalarının performanslarını test etmek için eğitim seti boyutu ve çıkarılan özellik boyutu değiştirilerek iki deney yapılmıştır. Deneysel sonuçlar, NN sınıflandırıcının eğitim seti boyutuna daha duyarlı olduğunu ve spam reddetme aracı olarak tek başına kullanılmasının uygun olmadığının kanaatine varılmıştır. Genel olarak,

DVM ve RVM sınıflandırıcılarının performansları veri kümelerinden ve özellik boyutlarından daha az etkilenmektedir ve NB sınıflandırıcıya göre açıkça daha üstündür. DVM ile karşılaştırıldığında, RVM'nin önemli ölçüde daha düşük RV oranı ve çok daha hızlı test süresiyle benzer sınıflandırma sağladığı gösterilmiştir. Ancak RVM'nin öğrenme prosedürü normalde DVM'den çok daha yavaştır. Çalışma sonucunda tüm bu sebeplerden dolayı RVM sınıflandırmasının, düşük karmaşıklık gerektiren uygulamalar açısından DVM sınıflandırmasına daha uygun olduğu sonucuna varılmıştır [51].

Rayan (2022) RF ve J48 makine öğrenmesi yöntemlerini birleştirerek yeni bir makine öğrenmesi tabanlı bir hybrid bagged yöntemi önermiştir. Çalışma kapsamında deneyler için, iki sınıflandırma yönteminin her birinde 500 mail içeren 1000 e-postodan oluşan bir veri seti kullanılmıştır. Çalışma kapsamında üç test yapılmıştır. RF ve J48 makine öğrenmesi yöntemleri için iki test ve RF ve J48 birlikte kullanıldığı hybrid bagged için üçüncü test yapılmıştır. RF ile yapılan testlerde doğruluk, kesinlik, duyarlılık ve F-skoru puanları sırasıyla: 84, 85, 82 ve 90 olmuştur. J48 ile yapılan testlerde doğruluk, kesinlik, duyarlılık ve F-skoru puanları sırasıyla: 92, 94, 90 ve 85 olmuştur. Hybrid bagged ile yapılan testlerde doğruluk, kesinlik, duyarlılık ve F-skoru puanları sırasıyla: 88, 90, 86 ve 88 olmuştur. Çalışma sonucunda J48, önerilen sistem olan hybrid bagged yönteminden daha yüksek performans sergilemiştir [52].

Srinivasan vd. (2021), doğal dil işleme (NLP) bağlamında derin öğrenme mimarilerine dayalı olarak spam mailleri tespit etmek için yeni bir metodoloji önermiştir. Çalışma kapsamında üç farklı veri seti kullanılmıştır. Bunlar: Lingspam, PU, Spam Assassins ve Enron veri setleridir. Bu çalışmada, her yöntemin metni temsil etmenin kendine özgü bir yolu olduğundan, çalışmadaki soruna en uygun metin temsil tekniğini incelemek için farklı teknikler uygulanmıştır. Bu yöntemler TDM, TF-IDF, SVD ile TDM, NMF ile TDM, SVD ile TF-IDF, NMF ile TF-IDF, Keras gömme, FastText, NBOW ve kelime gömmedir. Çalışma kapsamında özellik temsilinin yüksek boyutluluğu nedeniyle, DNN, Recurrent Neural Network (RNN), CNN, LSTM ve CNN-LSTM gibi birçok derin öğrenme modeli kullanılarak optimum sayıda özellik çıkartılmıştır. Ayrıca, çalışmadaki yaklaşımı karşılaştırmak için, LR, K-NN, GaussianNB, DT, RF, AdaBoost ve DVM dahil olmak üzere çeşitli klasik makine öğrenme algoritmaları kullanılarak ve bunlarla birleştirilen farklı metin temsil yöntemleri önerilen yaklaşım ile karşılaştırılmıştır. Çalışma kapsamında önerilen model, kara listeye alma ve makine öğrenimi sınıflandırıcılarına dayanan mevcut spam tespit yaklaşımlarından daha iyi performans

göstermiştir [53].

Hossain, Uddin & Halder (2021) mailleri spam ve spam olmayan olarak sınıflandıran bir model önermişlerdir. Çalışma kapsamında DBSCAN ve Isolation Forest, aykırı değerleri tanımlamak için kullanılmıştır. Çalışma kapsamında etkili özellikleri seçmek için Heatmap, Recursive Feature Elimination ve Chi-Square özellik seçme teknikleri kullanılmıştır. Önerilen model, karşılaştırmalı bir analiz oluşturmak için hem makine öğrenmesi algoritmalarına hem de derin öğrenmesi algoritmalarına uygulanmıştır. Makine öğrenmesi uygulamasında topluluk yöntemini tanıtmak için KNN, MNB, RF, Gradient Boosting (GB) kullanılmıştır. Derin öğrenme uygulaması için Gradient Descent (GD), RNN, ANN kullanılmıştır. Birden fazla sınıflandırıcının çıktısını birleştirmek için bir topluluk yöntemi oluşturulmuştur. Topluluk yöntemleri, tek bir sınıflandırıcıya kıyasla daha iyi tahmin doğruluğu üretilmesini sağlamıştır. Çalışma kapsamında önerilen model, UCI makine öğrenimi havuzundan toplanan bir mail spam temel veri setine dayalı olarak makine öğrenimi uygulaması için %100 doğruluk, AUC (Area Under the Curve) =100, MSE (Mean Squared Error) hatası = 0 ve RMSE (Root Mean Squared Error) hatası = 0 ve derin öğrenme uygulaması için %99 doğruluk, kayıp değeri = 0,0165 elde etmiştir. Çalışma sonucunda makine öğrenimi algoritmalarının derin öğrenme algoritmalarından daha iyi performans gösterdiğini gözlemlenmiştir [54].

Lai (2007) NB, K-NN ve DVM algoritmaları karşılaştırmalı olarak analiz etmiştir. Çalışma kapsamında iki adet veri koleksiyonu kullanılmıştır. İlk veri koleksiyonu 11291 spam ve 5552'si normal olarak etiketlenmiş 16843 mailden oluşmaktadır ve Babletext ve SpamAssassin üzerinden elde edilmiştir. İkinci veri koleksiyonu EM Canada web sitesinden gelen 24038 spam postayı ve SpamAssassin web sitesinden gelen aynı sayıda normal postayı içermektedir ve spam oranı %81,24'tür. Çalışma sonucunda NB ve DVM K-NN göre daha yüksek bir performans göstermiştir [55].

[56] yapmış oldukları çalışmada WEKA kullanarak spam mailleri filtrelemek için farklı sınıflandırma teknikleri kullanmışlardır. Çalışma kapsamında Clustering, J48, Naive Bayes (NB), destek vektör makinesi (DVM) ve ID3 makine öğrenmesi teknikleri kullanılmıştır. Çalışma kapsamında kullanılan veri seti laboratuvar üyelerinden toplanan 450'si spam ve 570 normal olmak üzere toplam 1020 mailden elde edilmiştir. Çalışma sonucunda, çalışmada kullanılan tüm tekniklerden DVM ve ID3 hariç Naive Bayes daha hızlı sonuç ve daha yüksek doğruluk oranına sahip olmuştur. DVM ve ID3, Naive Bayes'ten daha iyi doğruluk sağlamasına rağmen modeli oluşturmak daha fazla zaman

almıştır [56].

[57] yapmış oldukları çalışmada Harris Hawks Optimizer (HHO) algoritmasını K-NN algoritması ile entegre eden yeni bir spam sınıflandırma tekniği üzerinde çalışmışlardır. Çalışma kapsamında önerilen modelin performansını değerlendirmek için UCI makine öğrenimi havuzundan elde edilen Spambase veri seti kullanılmıştır. Veri seti, her biri 57 öznitelikten oluşan 4601 örnek içermektedir. Çalışma kapsamında önerilen binary HHO algoritmasının performansını değerlendirmek için Teaching-Learning-based Optimization, binary Dragonfly, Seagull optimizasyon algoritması, Equilibrium Optimizer ve Marine Predators algoritması ile karşılaştırılmıştır. Çalışma sonucunda, önerilen yöntemin sınıflandırma doğruluğunun beş farklı optimize edici ile karşılaştırıldığında, en yüksek doğruluğa sahip olmuştur [57].

3. YAPAY ZEKA VE MAKİNE ÖĞRENMESİ

Bu bölümde çalışma kapsamında kullanılan K-means, Isolation Forest, Sentiment Analysis, Naive Bayes, Random Forest, Lojistik Regresyon, destek vektör makinesi teknikleri, değerlerdirme metrikleri ve TF-IDF (Term Frequency-Inverse Document Frequency) açıklanmıştır.

3.1. K-MEANS

K-means, 1967 senesinde MacQueen tarafından önerilmiştir [15]. Arama verimliliğini artırmak için sezgisel bilgiyi kullanır ve aramanın daha objektif olmasını sağlar. Temel fikri, küme sayısı K'nin atanmasıdır. İlk olarak, başlangıçta bir bölüm oluşturularak, ardından küme merkezini sürekli olarak hareket ettirerek bölümü iyileştirmek için iterasyon yöntemi kullanılır [16]. Aslında, bu arama yöntemi kullanılarak en iyi çözüm gerekli değildir. Ancak sezgisel bilgiyi kullanarak, her kümenin merkezini belirtmek için ortalama değeri kullanarak hesaplama karmaşıklığını azaltıp arama verimliliğini artırır. Bu, belirli bir verimlilik kısıtlaması altında büyük verilerin en iyi çözümünü elde etmeyi mümkün kılar [17]. K-means algoritmasının formülsel ifadesi için denklem (3.1)'de verilen denklemlerden yararlanılır [33].

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{ci} (||x_i - v_j||)^2 \tag{3.1}$$

Denklem (3.1) için, $||x_i - v_j||$, x ve y arasındaki Öklid mesafesi 'c_i', i^{th} kümesindeki veri noktalarının sayısı, c ise küme merkezlerinin sayısıdır.

K-means kümelemenin algoritmik adımları:

$\{x_1, x_2, x_3, \dots, x_n\}$ kümesi veri noktalarının, $V = \{v_1, v_2, v_3, \dots, v_c\}$ ise merkez noktalarının kümesi olsun.

1. Rastgele 'c' küme merkezlerini seç.
2. Her veri noktası ile küme merkezleri arasındaki mesafeleri hesaplayın.

3. Her veri noktasını, en yakın olduğu küme merkezine atayın, yani küme merkezi ile diğer küme merkezleri arasındaki mesafeden daha küçük olanı seçin.
4. Yeni küme merkezlerini denklem (3.2) kullanarak yeniden hesaplayın:

$$v_i = (1/c_i)\sum_{j=1}^{ci} x_i \tag{3.2}$$

5. Her veri noktası ile yeni küme merkezleri arasındaki mesafeleri yeniden hesaplayın.
6. Eğer hiçbir veri noktası ataması değişmediyse algoritmayı sonlandırın, aksi takdirde 3. adımdan başlayarak tekrarlayın.

3.2. ISOLATION FOREST

Isolation forest 2008 yılında Fei Tony Liu ve Zhi-Hua Zhou tarafından geliştirilmiştir. Isolation forest anomalilerin tespiti için kullanılmaktadır [18]. Denklem (3.3)'te Isolation Forest bir x örneğinin anomali puanı hesaplanmasında kullanılan denklem verilmiştir [34].

$$s(x,n) = 2^{\frac{-E(h(x))}{c(n)}} \tag{3.3}$$

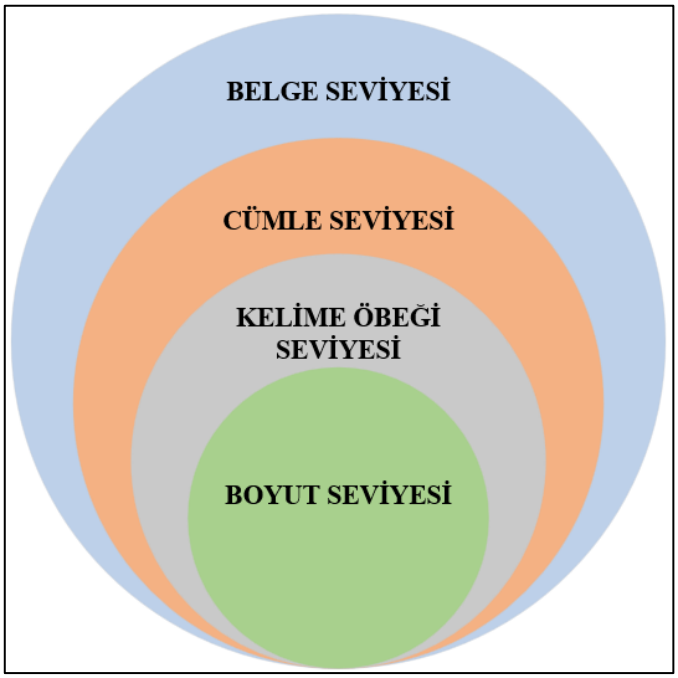
Denklem (3.3) için x örneğimiz, n harici düğümlerin sayısını temsil eder. h(x), x veri noktasının yol uzunluğunu ifade ederken, E(h(x)) ise bu yol uzunluğunun beklenen veya ortalama değerini gösterir. c(n), bir ikili arama ağacında başarısız aramanın ortalama yol uzunluğunu ifade eder ve denklem (3.4)'teki gibi ifade edilir [18],[34].

$$c(n) = 2H(n-1) - (2(n-1)/n) \tag{3.4}$$

3.3. DUYGU ANALİZİ

Duygu analizi (sentiment analysis), doğal dil işleme, hesaplamalı dilbilim ve biyometri tekniklerini kullanarak öznel tercihleri ve duygusal durumları sistematik olarak anlamak, belirlemek, ölçmek ve incelemek amacını taşır [10]-[12]. Genel olarak bir metnin, yazarının bir konuya, bir bağlama ya da bir belgeye yönelik tutumunu ve duygusal yaklaşımını belirlemeyi hedefler [13],[14].

Duygu analizi çeşitli seviyelerde incelenmektedir: Belge seviyesi, cümle seviyesi, kelime öbeği seviyesi ve boyut seviyesi [58]. Şekil 3.1'de duygu analizi seviyeleri yer almaktadır.



Şekil 3.1 Duygu analizi seviyeleri [58]

3.3.1. Belge Seviyesi Duygu Analizi

Belge düzeyinde duygu analizi tüm belge üzerinde gerçekleştirilir ve belgenin tamamına tek kutupluluk verilir. Bu tür duygu analizi çok fazla kullanılmaz. Bir kitabın bölümlerini veya sayfalarını olumlu, olumsuz veya tarafsız olarak sınıflandırmak için kullanılabilir [58]. Bu seviyede belgeyi sınıflandırmak için hem denetimli hem de denetimsiz öğrenme yaklaşımlarından yararlanılabilir [59].

3.3.2. Cümle Seviyesi Duygu Analizi

Bu analiz düzeyinde her cümle analiz edilir ve karşılık gelen kutupluluk bulunur. Bu, bir belgenin geniş bir yelpazeye yayılan ve kendisiyle ilişkilendirilen duyguların karışımını içerdiği durumlarda oldukça faydalıdır [60].

3.3.3. Kelime Öbeği Seviyesi Duygu Analizi

Duygu analizi, fikir kelimelerinin kelime öbeği düzeyinde çıkarıldığı ve sınıflandırmanın yapılacağı yerlerde de gerçekleştirilebilir. Her bir ifade birden fazla yön veya tek bir yön içerebilir. Bu, birden fazla satırdan oluşan faydalı ürün incelemeleri olabilir; burada, bir

ifadede tek bir yönün ifade edildiği görülmektedir [61]. Belge düzeyinde analiz, tüm belgeyi olumlu ya da olumsuz olarak öznel bir şekilde kategorize etmeye odaklanırken, bir belge hem olumlu hem de olumsuz ifadeler içerdiğinden, kelime düzeyinde analiz daha faydalıdır [58].

3.3.4. Boyut Seviyesi Duygu Analizi

Duygu analizi boyut seviyesinde gerçekleştirilir. Her cümle birden fazla boyut içerebilir; bu nedenle, boyut düzeyinde duyarlılık analizi yapılır [58]. Cümlede kullanılan tüm yönlere öncelikli olarak dikkat edilir ve tüm yönlere polarite atandıktan sonra tüm cümle için toplam bir duyarlılık hesaplanır [62], [63].

3.4. NAIVE BAYES

Naive Bayes (NB), sınıflandırma için en iyi bilinen veri madenciliği algoritmalarından biridir [19]. Naive Bayes belirli bir sınıfa ait tüm niteliklerin birbirinden bağımsız olduğu varsayımına dayanarak yeni bir örneğin belirli bir sınıfa ait olma olasılığını ortaya çıkarır [20]. Bu varsayım, eğitim verilerinden çok değişkenli olasılıkları tahmin etme ihtiyacından kaynaklanmaktadır. Uygulamada çoğu nitelik değeri kombinasyonu ya eğitim verilerinde mevcut değildir ya da yeterli sayıda mevcut değildir. Bu nedenle ilgili çok değişkenli olasılıkların doğrudan tahminleri güvenilmez olacaktır. Naive Bayes, koşullu bağımsızlığı varsayarak bu çıkmazı aşmaktadır [22]. Katı bağımsızlık varsayımına rağmen, Naive Bayes birçok gerçek dünya uygulamasında gerçekten yetkin bir sınıflandırıcıdır [21], [22].

Bayes teoremi, $P(c)$, $P(x)$ ve $P(c \mid x)$'den sonsal olasılığı, $P(c \mid x)$, hesaplamanın bir yolunu sağlar. Naive Bayes yöntemi, bir tahmincinin (x) değerinin belirli bir küme (c) üzerindeki etkisinin diğer tahmincilerin değerlerinden bağımsız olduğunu varsayar. Bu varsayıma sınıf koşullu bağımsızlık adı verilir [64]-[66]. Denklem (3.5)'te Naive Bayes denklemi yer almaktadır.

$$P(c \mid x) = P(x \mid c)P(c)/P(x) \tag{3.5}$$

- $P(c \mid x)$, tahmin edici(öznitelik) göz önüne alındığında sınıfın (hedef) sonsal olasılığıdır.
- $P(c)$, sınıfın önceki olasılığıdır.

- $P(x \mid c)$, sınıf göz önüne alındığında tahmin edicinin olasılığı olan olabilirliktir.
- $P(x)$, tahmin edicinin önceki olasılığıdır [66].

3.5. RANDOM FOREST

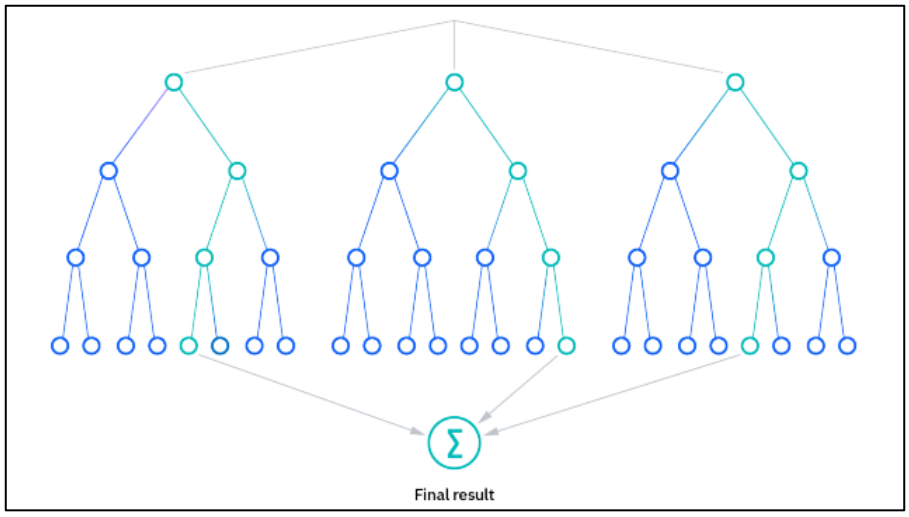
Random Forest (RF), 2001 senesinde Leo Breiman tarafından geliştirilmiştir. Random Forest regresyon ve sınıflandırma problemlerinde kullanılır [23]. Random Forest, Bagging ve Random Subspace yöntemlerin birleştirilmesiyle elde elde edilmiştir [67]. Bu algoritmanın amacı, çoklu karar ağaçları oluşturarak sınıflandırma sürecindeki sınıflandırma değerini iyileştirmektir. Random Forest algoritması, birbirine bağımlı olmayan bir şekilde çalışan birden çok karar ağacının birleştirilerek aralarından en yüksek puana sahip olanın seçilmesi işlemine dayanır [23].

Breiman'ın orijinal ormanlarında, tek bir ağacın her bir düğümü bir hiper dikdörtgen hücre ile ilişkilendirilir. Ağacın kökü kendisidir ve ağaç inşasının her adımında bir düğüm (ya da eşdeğer olarak karşılık gelen hücre) iki parçaya bölünür. Terminal düğümler (veya yapraklar) birlikte ele alındığında, $\mathcal{X}$ 'in bir bölümünü oluşturur [23].

Algoritma, M farklı (rastgele) ağacı aşağıdaki şekilde büyüterek çalışır. Her bir ağacın oluşturulmasından önce, gözlemler orijinal veri setinden değiştirilerek (veya değiştirilmeden) rastgele çekilir [23]. Bunlar ve yalnızca bunlar (olası tekrarlarla birlikte) ağaç oluşturmada dikkate alınır. Daha sonra, her bir ağacın her bir hücresinde, p orijinal arasından rastgele seçilen yönler üzerinden CART kriterini maksimize ederek bir bölme işlemi gerçekleştirilir [23]. Son olarak, her hücre düğüm boyutundan daha az nokta içerdiğinde bireysel ağaçların inşası durdurulur. Herhangi bir sorgu noktası için $x \in \mathcal{X}$, her regresyon ağacı kendisine tekabül eden X_i'nin x hücresine düştüğü (a_n noktaları arasında olan) Y_i'nin ortalamasını tahmin eder [23].

Özetle Rastgele Forest algoritması bir dizi karar ağacından oluşur ve topluluktaki her ağaç, önyükleme örneği adı verilen, değiştirilen bir eğitim setinden alınan bir veri örneğinden oluşur. Bu eğitim örneğinin üçte biri, daha sonra geri döneceğimiz torba dışı (oob) örnek olarak bilinen test verileri olarak ayrılır [68]. Daha sonra özellik torbalama yoluyla başka bir rastgelelik örneği enjekte edilir, böylece veri kümesine daha fazla çeşitlilik eklenir ve karar ağaçları arasındaki korelasyon azaltılır. Sorunun türüne bağlı olarak tahminin belirlenmesi farklılık gösterecektir. Bir regresyon görevi için bireysel karar ağaçlarının ortalaması alınacak ve bir sınıflandırma görevi için çoğunluk oyu (yani

en sık görülen kategorik değişken) tahmin edilen sınıfı verecektir. Son olarak oob örneği çapraz doğrulama için kullanılır ve bu tahmin sonlandırılır [68]. Şekil 3.2'de Random Forest için bir örnek yer almaktadır.



Şekil 3.2 Random Forest için bir örnek [77]

3.6. LOJİSTİK REGRESYON

Lojistik Regresyon (LR), sınıflandırma problemlerini öğrenmek için geliştirilmiş denetimli bir makine öğrenmesi algoritmasıdır. Hedef değişken kategorik bir değişken olduğunda sınıflandırma öğrenme sorunları ortaya çıkar [24]. Lojistik regresyon, bir kategorik değişkenin (bağımlı değişken) belirli bir durumunu tahmin etmek için bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modellemek için kullanılır [69]. Lojistik Regresyon'un amacı, yeni bir örneğin hedef kategorilerden birine ait olma olasılığını tahmin etmek için veri kümesi özelliklerinin bir fonksiyonunu bir hedefe eşlemektir [24],[25].

Lojistik Regresyon, tahminleri ve bunların olasılıklarını haritalamak için sigmoid işlevi adı verilen bir lojistik işlevi kullanır. Sigmoid işlevi, herhangi bir gerçek değeri 0 ile 1 arasındaki bir aralığa dönüştüren S şeklinde bir eğriyi ifade eder [70]. Sigmoid fonksiyonuna lojistik regresyon için aktivasyon fonksiyonu denir ve denklem (3.6)'daki gibi tanımlanır [70].

$$f(x) = 1/1 + e^{-x} \tag{3.6}$$

- e, Euler sabitini (yaklaşık olarak 2.71828) ifade eder.

- x, giriş değerini ifade eder.
- f(x), çıktıyı ifade eder.

Denklem (3.7)'de Lojistik Regresyon modelinin denklemi yer almaktadır [70].

$$y = e^{(b_0+b_1X)}/1 + e^{(b_0+b_1X)} \quad (3.7)$$

- x, giriş değeri
- y, tahmin edilen çıktı
- b0, sapma veya kesme terimi
- b1, giriş katsayısı (x)

Kategorik yanıta dayalı olarak tanımlanan üç tür lojistik regresyon modeli vardır [69]-[71].

3.6.1. Binary Lojistik Regresyon

İkili lojistik regresyon, bağımsız ve ikili bağımlı değişkenler arasındaki ilişkiyi tahmin eder. Bu regresyon türünün çıktısının bazı örnekleri başarı/başarısızlık, 0/1 veya doğru/yanlış olabilir [70], [71].

3.6.2. Multinomial Lojistik Regresyon

Kategorik bir bağımlı değişken, multinomial regresyon tipinde iki veya daha fazla ayrık sonuca sahiptir. Bu, bu regresyon tipinin ikiden fazla olası sonuca sahip olduğu anlamına gelir [70], [71].

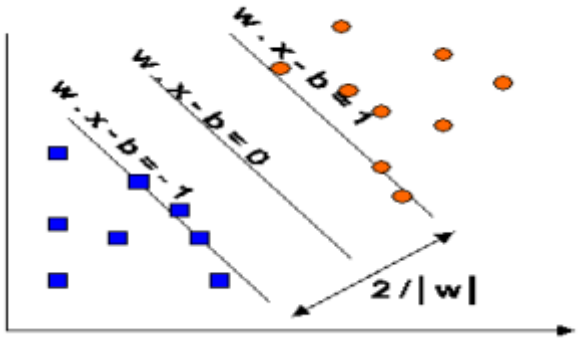
3.6.3. Ordinal Lojistik Regresyon

Bağımlı değişken sıralı durumda (yani sıralı) olduğunda sıralı lojistik regresyon uygulanır. Bağımlı değişken (y), iki veya daha fazla kategori veya seviyeye sahip bir sırayı belirtir. [70], [71].

3.7. DESTEK VEKTÖR MAKİNESİ

DVM özellikle regresyon ve sınıflandırma problemlerinde kullanılmakta olan güçlü bir makine öğrenme algoritmasıdır. Temel amacı, veri kümesindeki öğeleri farklı sınıflara ayrıştırmak için en uygun hiper düzlemi bulmaktır. Bu hiper düzlemi belirlerken DVM,

bir veri noktasının kendi sınıfına olan mesafesini maksimuma çıkarmaya çalışır, böylece sınıflar arasındaki ayrım maksimuma çıkar. Bu, veri noktalarını daha iyi sınıflandırmak ve özetlemek için etkili bir yol sağlar [26], [27]. Örnek olarak Şekil 3.3'te verilen iki grup incelenecektir.



Şekil 3.3 İki sınıftan örneklerle eğitilen bir DVM için maksimum marj hiper düzlemleri [75]

Yukarıdaki her iki sınıf arasında oluşan aralık tolerans olarak adlandırılır. Denklem (3.8)'de bu düzlemde yer alan her noktanın tanımı yer almaktadır [73], [75].

$$D = \{(x_i, c_i) | x_i \in R^p, c_i \in \{-1,1\}\}_{i=1}^{n} \quad (3.8)$$

Her (x_i, c_i) çifti için X vektör uzayındaki bir noktayı ve c'nin X vektör uzayındaki noktanın -1 veya $+1$ olduğunu gösteren bir değeri temsil etmektedir. Bu noktalar kümesi $i = 1$ 'den n'e kadar uzanmaktadır. Denklem (3.8) ifadesi, bir aşırı düzlem (hyperplane) üzerinde olduğu varsayıldığında, her noktanın denklemi denklem (3.9) ile temsil edilir [73], [75].

$$wx + b = 0 \quad (3.9)$$

Denklem (3.9)'daki w, aşırı düzleme dik olan normal vektörü, x ise noktanın değişen parametresini ve b ise kayma oranını temsil eder. Denklem (3.9)'a göre $b/||w||$ değeri iki grup arasındaki mesafe farkını vermektedir. Bu mesafe farkı denklemine göre mesafeyi en yüksek değere çıkarmak için şekil 3.3'te gösterilen 0, -1 ve +1 değerlerine sahip 3 doğruyu veren denklemde $2/||w||$ formülü kullanılmaktadır. Denklem (3.9)'a göre elde edilen iki doğru denklemi denklem (3.10) ve denklem (3.11)'de yer almaktadır [73], [75].

$$wx + b = -1 \tag{3.10}$$

$$wx + b = 1 \tag{3.11}$$

DVM algoritması, linear DVM ve non-linear DVM olmak üzere iki çeşittir [72], [73], [75].

3.7.1. Linear DVM

Linear DVM', farklı sınıfları ayıran düz çizgili bir karar sınırı oluşturmak için doğrusal bir çekirdek kullanır. Veriler doğrusal olarak ayrılabilir olduğunda veya doğrusal bir yaklaşım yeterli olduğunda etkilidirler. Linear DVM'ler hesaplama açısından verimlidir ve karar sınırı giriş özelliği uzayındaki bir hiper düzlem olduğundan iyi yorumlanabilirliğe sahiptir [72], [73], [75].

3.7.2. Non-Linear DVM

Non-Linear DVM, verilerin giriş özelliği alanında düz bir çizgiyle ayrılamadığı senaryoları ele alır. Bunu, verileri doğrusal bir karar sınırının bulunabileceği daha yüksek boyutlu bir özellik alanına örtülü olarak eşleyen çekirdek işlevlerini kullanarak elde ederler. Bu tür DVM'de kullanılan popüler çekirdek işlevleri arasında polinom çekirdeği, Gaussian (RBF) çekirdeği ve sigmoid çekirdeği bulunur. Doğrusal olmayan DVM'ler karmaşık modelleri yakalayabilir ve doğrusal DVM'lerle karşılaştırıldığında daha yüksek sınıflandırma doğruluğu elde edebilir [72], [73], [75].

3.8. DEĞERLENDİRME METRİKLERİ

Bu bölümde makine öğrenmesi modellerinin değerlendirilmesinde kullanılan accuracy (doğruluk), precision (kesinlik), recall (duyarlılık), f1-score (f1-skoru) metrikleri açıklanacaktır.

3.8.1. Doğruluk

Doğruluk, iyi dengelenmiş ve çarpık olmayan veya sınıf dengesizliği olmayan sınıflandırma problemleri için geçerli bir değerlendirme ölçütüdür. Yüksek accuracy, modelin doğru sınıflandırma yeteneğini gösterir. Ancak, sınıflar arasında dengesiz veri dağılımı gibi vb. durumlar varsa doğruluk değerinin yüksek olması aldatıcı olabilir [76]. Doğruluk doğru tahmin edilen veri noktalarının (true positive ve true negative) toplam

veri noktalarına (true positive, true negative, false positive ve false negative) oranıdır [76]. Denklem (3.12)'de doğruluk denklemi yer almaktadır.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Data\ Points} \tag{3.12}$$

3.8.2. Kesinlik

Kesinlik, sınıflandırma modelinin pozitif olarak tahminlediği değerlerin gerçekten kaç adedinin pozitif olduğunu gösterir. [76]. Kesinlik, özellikle yanlış pozitiflerin maliyetinin yüksek olduğu durumlarda önemlidir. Örneğin kanser teşhisi gibi bir uygulamada yanlış pozitifler ciddi sonuçlara yol açabilir. Bu nedenle kesinlik, model performansını değerlendirirken dikkate alınmalıdır [78]. Denklem (3.13)'te kesinlik denklemi yer almaktadır.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3.13}$$

3.8.3. Duyarlılık

Duyarlılık, gerçek pozitif veri noktalarının pozitif olarak tahmin edilen veri noktalarına oranıdır [76]. Duyarlılık, özellikle yanlış negatiflerin maliyetinin yüksek olduğu durumlarda önemlidir. Örneğin kanser teşhisi gibi bir uygulamada yanlış negatifler ciddi sonuçlara yol açabilir. Bu nedenle duyarlılık, model performansını değerlendirirken dikkate alınmalıdır [78]. Denklem (3.14)'te kesinlik denklemi yer almaktadır.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3.14}$$

3.8.4. F1-Skoru

Genel olarak doğruluk, özellikle dengesiz veri kümelerinde tehlikeli bir şekilde aşırı iyimser şişirilmiş sonuçlar gösterebilir [76]. F1-skoru recall ve precision'un harmonik ortalamasıdır ve bu nedenle F1-skoru sınıflandırıcı için recall ve precision arasında bir denge sağlar [76]. F1-skoru, yanlış pozitifleri ve yanlış negatifleri birlikte hesaba katan bir doğruluk ölçüsüdür. F1-skoru, özellikle dengesiz sınıflandırma için genellikle doğruluktan daha kullanışlıdır. [76]. Denklem (3.15)'te f1-skoru denklemi yer almaktadır.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3.15)$$

3.9. TF-IDF

TF-IDF, metin verilerini sayısal bir temsile dönüştürmek için kullanılan bir yöntemdir. Bu vektörleştirici, metinleri anlamlı sayılar olarak göstermenin bir yoludur ve aynı zamanda vektör gösterimi olarak da bilinir. Bir belgedeki t_k terimi için, TF-IDF ağırlığı $w(t_k)$, bir belgede genellikle metin sınıflandırmasında denklem (3.16)'daki gibi temsil edilir [79].

$$w(t_k) = t \int k * log\ (\frac{N}{d \int k}) \quad (3.16)$$

Denklem (3.16)'da $t \int k$, t_k'nın belgede geçtiği terim sıklığıdır. $t \int k$ ise t_k'nın belge sıklığıdır yani $t_k$içeren belgelerin sayısıdır. N ise derlemdeki toplam belge sayısıdır.

4. MATERYAL VE YÖNTEM

Çalışma kapsamında Türkçe maillerden oluşan iki veri seti kullanılmıştır. Bu veri setlerinden ilki kaggle'da yer alan Türkçe spam veri setidir. Bu veri seti 330 spam e-mail ve 496 normal e-mailden olmak üzere toplamda 823 Türkçe e-mailden oluşmaktadır. Çalışma kapsamında kullanılan ikinci veri seti kaggle'da bulunan Türkçe mail veri setidir. Bu veri seti 502 normal e-mail ve 515 spam e-mail olmak üzere toplamda 1017 Türkçe e-mailden oluşmaktadır. Elde edilen veri setlerindeki spam ve normal mailler birleştirilerek spam mail veri seti ve normal mail veri seti olmak üzere iki veri seti oluşturulmuştur.

Elde edilen veriler makine öğrenmesi yöntemleri ile çalışılabilmesi için veri önişleme adımlarından geçirilmiştir. Böylelikle makine öğrenmesi tekniklerinin ve duygu analizinin uygulanmasına veriler hazırlanmıştır. Çalışma kapsamında sırasıyla veri önişleme, görselleştirme, kümeleme, duygu analizi ve sınıflandırma işlemleri yapılmıştır.

4.1. VERİ ÖNİŞLEME

Makine öğrenmesi tekniklerinin başarısı genellikle üzerinde çalıştıkları verinin kalitesine bağlıdır. Eğer veri kümesi yetersiz, gereksiz veya ilgisiz verilerden oluşuyorsa makine öğrenmesi tekniklerinin başarı oranı düşmektedir [28],[29]. Veri küme içerisinden olabildiğince ilgisiz, tekrarlanan gürültüye sebep veren verilerin temizlenmesi gereklidir. Bu işlemlerde veri önişleme başlığı altında gerçekleştirilmektedir. Veri önişleme aşamasında aşağıdaki işlemler yapılmıştır.

a) Verilerin küçük harflere dönüştürülmesi.

b) Web sitesi adresleri "website" ve e-mail adresleri "email" olarak yeniden adlandırıldı.

c) Noktalama işaretleri ve sayısal ifadelerin kaldırılması.

Şekil 4.1'de spam maillerden oluşan veri setine veri önişleme adımları uygulanmadan ve uygulandıktan sonraki veri setinin durumu yer almaktadır.

	Data
0	Sayın Yetkili ;\n Gelişen ve değişen günümüz ...
1	Sayın Yetkili,\n \n28 Kasım 2010 tarihli KPSS ...
2	Sayın Yetkili ;\n Gelişen ve değişen günümüz ...
3	T.C.\nURLA KAYMAKAMLIĞI\nURLA HAKAN ÇEKEN ANAD...
4	Sayın Yetkili ;Hızla büyüyen ve gelişmekte ola...
..	...
510	Yılbaşı Özel İndirimleri Başladı! 16,95 TL'den...
511	Kaçırılmayacak Şahane İndirimler Başladı! Pier...
512	Kadınlar Günü sürprizin hazır mı? Çiçek& Bonny...
513	Aradığın Sürprize En Uygun Hediyeler ÇiçekSepe...
514	Canlı yayında alışveriş etkinliği için son saa...

[844 rows x 1 columns]

	Data
0	sayın yetkili \n gelişen ve değişen günümüz \...
1	sayın yetkili\n \n kasım tarihli kpss sınavı ...
2	sayın yetkili \n gelişen ve değişen günümüz \...
3	tc\nurla kaymakamliği\nurla hakan çeken anadol...
4	sayın yetkili hızla büyüyen ve gelişmekte olan...
..	...
839	yılbaşı özel indirimleri başladı tlden başlay...
840	kaçırılmayacak şahane indirimler başladı pierr...
841	kadınlar günü sürprizin hazır mı çiçek bonnyfo...
842	aradığın sürprize en uygun hediyeler çiçeksepe...
843	canlı yayında alışveriş etkinliği için son saa...

[844 rows x 1 columns]

Şekil 4.1 Spam mail veri setine veri önişleme uygulanmadan önce ve sonrası

4.2. VERİ GÖRSELLEŞTİRME

Veri önişleme adımları uygulanan veri setleri kelime bulutu yöntemiyle görselleştirilmiştir. Verilerin görselleştirilmesi ile sonuçlar daha kolay okunabilir ve daha anlaşılabilir hale gelmiştir.

Kelime bulutu ya da etiket bulutu, metin verilerini görselleştirmenin anlaşılır bir yoludur. Bir kelime bulutuna bakan gözlemci, metnin yoğunluğu, sıklığı ve dağılımı konusunda kolaylıkla çıkarım yapabilir [8],[9].

Şekil 4.2'de veri ön işleme adımlarına ek olarak, bağlaçların ve Türkçe kökenli olmayan kelimelerin veri setinden temizlenmesi ile oluşturulan kelime bulutu yer almaktadır.

Şekil 4.2 Verilerden kelime bulutu elde edilmesi

4.3. VERİLERİN KÜMELENMESİ

Kelime bulutu ile görselleştirilen spam ve normal veri setleri hakkında kelime bulutu ve bar grafiği sonuçlarıyla bir çıkarım yapılabilir ancak bu çıkarımlar tek başına yeterli değildir. Spam maillerin ve normal maillerin kümelenmesi ile yönelimleri arasında daha tutarlı çıkarımlar yapmak mümkündür. Verilerin kümelenmesinde K-means ve Isolation

Forest kümeleme yöntemleri kullanılmıştır.

4.4. DUYGU ANALİZİ

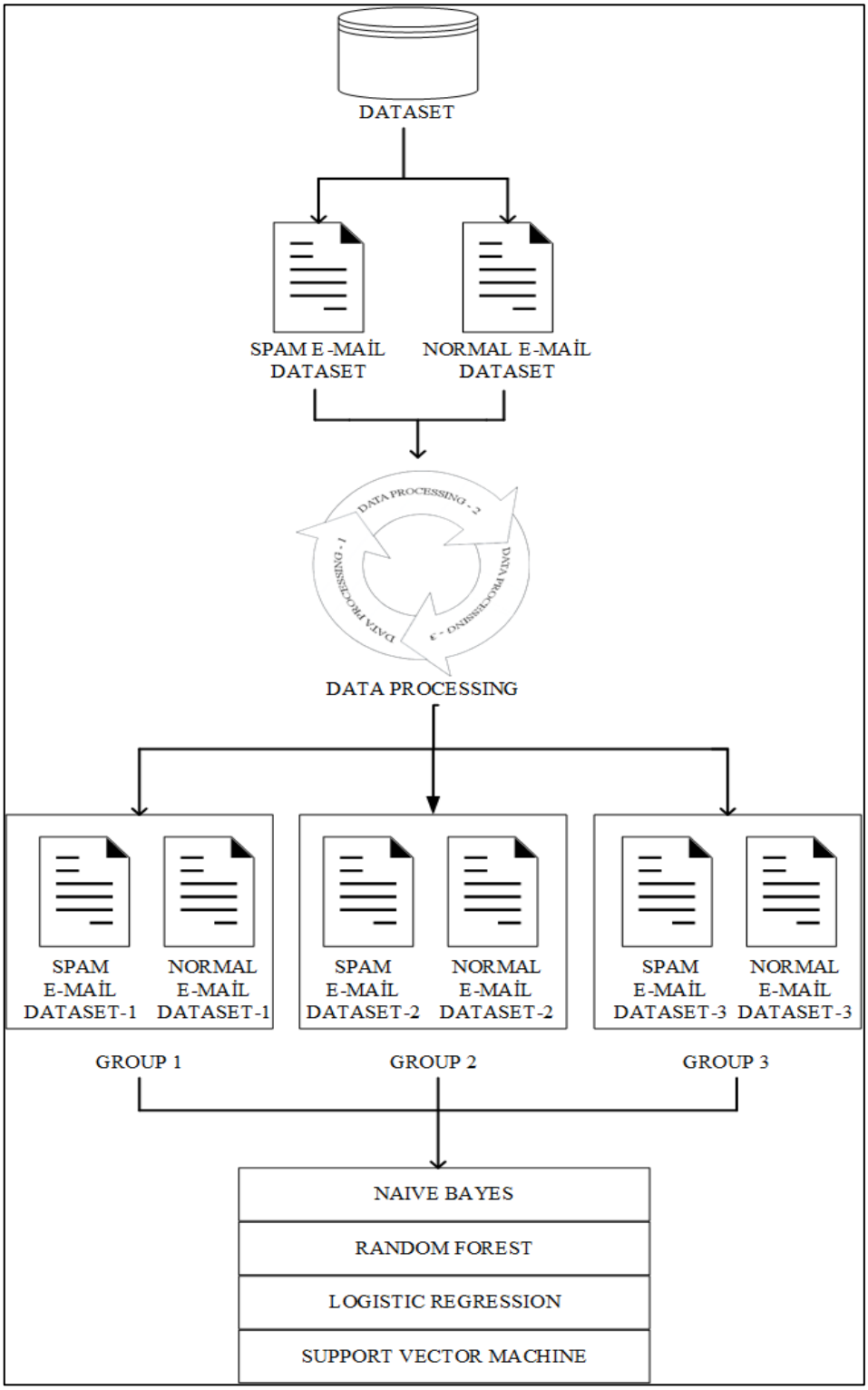
Spam ve normal maillerin tespitinde duygu tonunun belirlenmesi oldukça önemlidir. Spam mailler dolandırıcılık, kişisel bilgilerin çalınması vb. amaçlı kullanılabilmektedir. Dolandırıcılık gibi ciddi suç amacı barındıran spam mailler içerikleri yönünden muhatabını korkutmak, caydırmak amaçlı olduklarından genel itibariyle olumsuz duygu tonuna sahiptirler [31],[32]. Ama aynı zamanda olumlu duygu durumuna sahip mailler aracılığıyla bu amaca hizmet eden spam mailler mevcuttur [30].

Çalışma kapsamında elde edilen veri setleri genel kanıya göre dolandırıcılık gibi suç amacı güden spam mailler, normal maillere göre duygu tonu doğrusal mı ya da doğrusal olmayan bir eğilim mi sergilemektedir incelenmiştir.

4.5. VERİLERİN SINIFLANDIRILMASI

Çalışma kapsamında elde edilen veri setlerinin sınıflandırılmasında Naive Bayes, DVM, Random Forest, Lojistik Regresyon yöntemleri kullanılmıştır. Bu kullanılan algoritmalarının performansları karşılaştırılmış ve aralarından en iyi yöntem saptanmıştır. Çalışma kapsamın veriler normal ve spam veri setleri olarak ikiye ayrılmıştır. Bu ayrılan veri setleri üç işlem adımına tabi tutulmuştur. Bu tabi tutulma işlemi sonrası, üç normal ve üç spam mail veri seti olmak üzere toplamda altı veri seti elde edilmiştir. İlk işlem adımında veriler üzerinde temel veri önişleme adımları uygulanmıştır. İkinci işlem adımında ilk işlem adımına ek olarak verilerden Türkçe kökenli olmayan kelimeler, dört harften kısa kelimeler ve dört kelimeden kısa cümleler çıkarılmıştır. Üçüncü işlem adımında ikinci işlem adımı sonucu veri setinden çıkarılan veri satırları, ilk işlem adımı sonucu oluşturulan veri setinden de çıkarılmıştır. İlk işlem adımına ait veri setleri grup 1'de, ikinci işlem adımına ait veriler grup 2'de, üçüncü işlem adımına ait veriler grup 3'te listelenmiştir. Her veri grubu sırasıyla Naive Bayes, Random Forest, Lojistik Regresyon ve DVM yöntemleri ile sınıflandırılmıştır.

Şekil 4.3'te yapılan bu işlem adımlarına ait blok şema yer almaktadır.



Şekil 4.3 Veri sınıflandırma işlemi aşamasında yapılan işlemlerin blok şeması

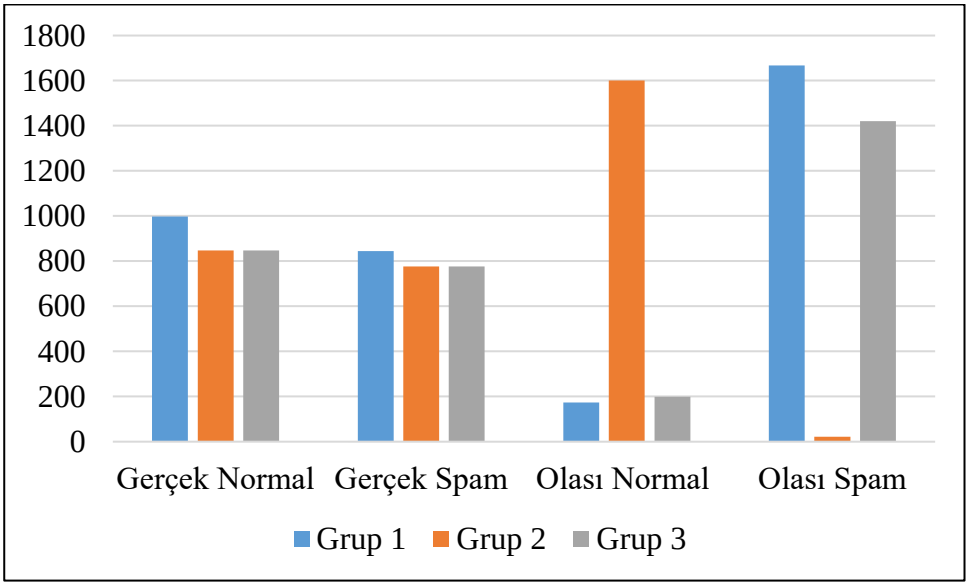
5. BULGULAR VE TARTIŞMA

Bu bölümde tez çalışması kapsamın kullanılan makine öğrenmesi modellerinin sonuçlarına dair bulgular yer almaktadır.

5.1. K-MEANS ile VERİLERİN KÜMELENMESİ

K-means ile veriler kümelendirilirken normal ve spam mailler bir veri seti altında birleştirilmiştir. Çalışma kapsamında denetimsiz algoritmaların performansının da karşılaştırılması için K-means ile etiketlenmemiş verilerin 2 küme altında dağılımı gözlemlenmiştir.

Şekil 5.1'de K-means ile grup 1, grup 2 ve grup 3'e ait verilerin kümelenmesi sonucu gerçek spam, gerçek normal, olası spam ve olası normal dağılımları gösterilmiştir.



Şekil 5.1 K-means ile kümeleme sonuçları

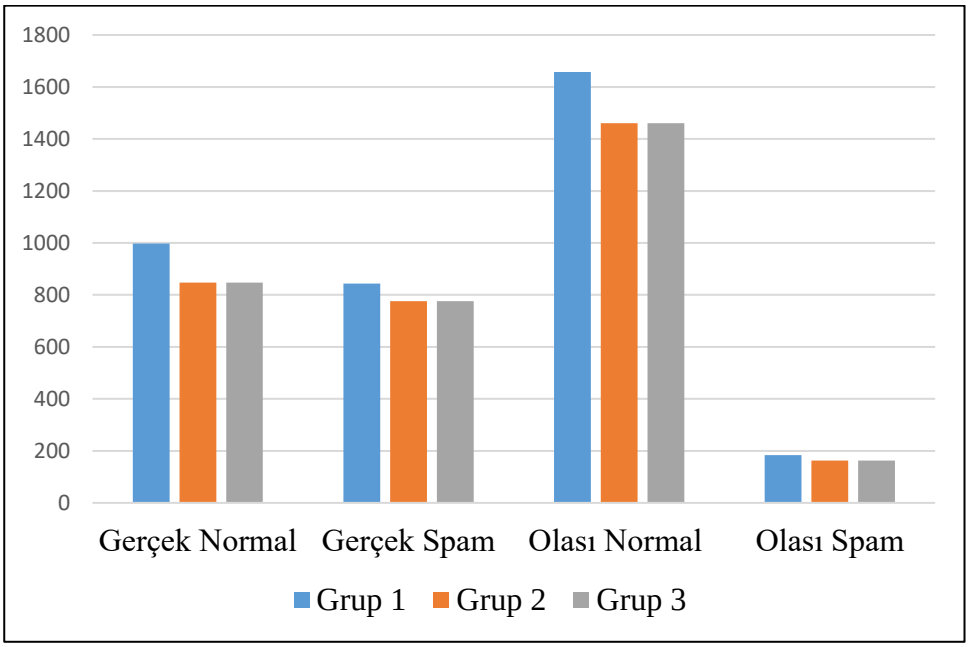
Şekil 5.1'deki veriler incelendiğinde K-means sonucu verilerin kümelenmesi ekseriyetle grup 1'de olası spam, grup 2'de olası normal ve grup 3'te olası spam üzerinde olmuştur. K-means ile kümeleme sonucunda veriler gerçek normal ve gerçek spam dağılımlarına yakın bir kümelenme gerçekleşmemiştir. Veri setinden bağlaçların, Türkçe kökenli olmayan kelimelerin ve dört kelimeden kısa cümlelerin çıkarılması kümelenme eğilimini

değiştirmiştir. K-means verilerin olası spam veya olası normal kümelenmesinde, gerçek dağılıma nazaran başarısız olmuştur.

5.2. ISOLATION FOREST ile VERİLERİN KÜMELENMESİ

Isolation Forest anomali tespiti için kullanılan bir algoritmadır. Normal veriler, tipik veya beklenen davranışları sergileyen verileri ifade eder. Anomali ise beklenmeyen, aykırı davranışta bulunan verileri ifade eder. Çalışma kapsamında bu anomali tespitinden yararlanarak grup 1, grup 2 ve grup 3'teki verilerin normal maillerin ve spam maillerin dağılımlarının anomali (olası spam) ve normal olarak dağılımları karşılaştırılmıştır. Isolatin Forest algoritmasının contamination değeri 0.1, eşik değeri ise sıfır olarak ayarlanmıştır.

Şekil 5.2'de grup 1, grup 2 ve grup 3'e ait normal ve spam e-maillerin Isolation Forest ile kümelenmesi sonucu dağılımları yer almaktadır.



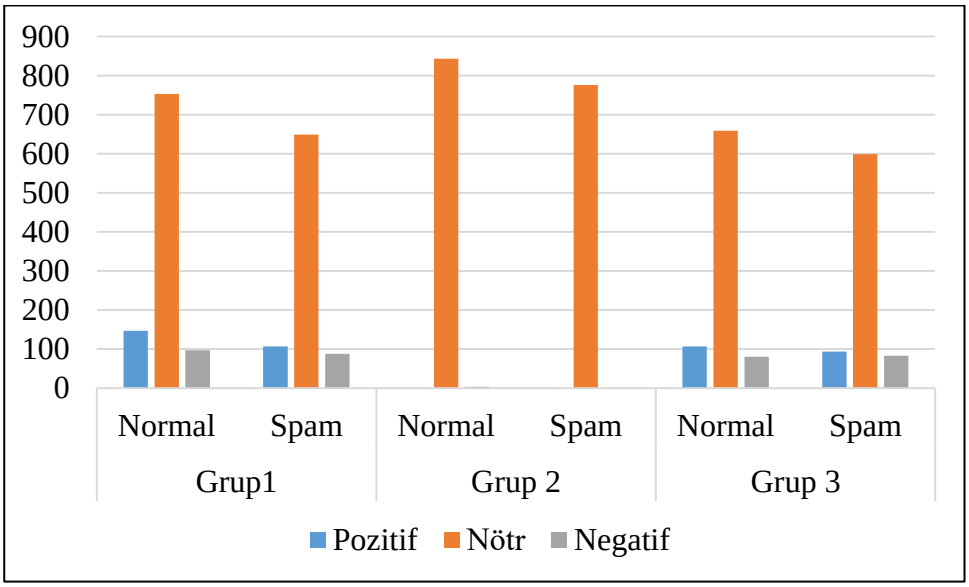
Şekil 5.2 K-means ile kümeleme sonuçları

Şekil 5.2'deki veriler incelendiğinde anomali ve normal verilerin dağılımları spam ve normal etiketlerinden gözle görülür bir şekilde farklı dağılmıştır. Normal ve spam etiketlerinde dağılımlar daha homojen yapıya sahipken Isolation Forest kümelemesi her üç grup içinde olası normal ekseninde yığılmıştır. Sonuç olarak veri setinden, Türkçe kökenli olmayan kelimelerin, dört kelimeden kısa cümlelerin ve bağlaçların veri setinden çıkarılması olası spam oranını %0,05 artırmıştır.

5.3. DUYGU ANALİZİ

Normal ve spam mail tespitinde duygu analizi maillerin duygusal tonunun belirlenmesinde kullanılır. Kişisel bilgilerin çalınmasında veya dolandırıcılık amacıyla kullanılan spam mailler çoğunlukla olumsuz duygu tonuna sahiptirler. Aksi bir durumda söz konuda mevcuttur. Bu gibi sebeplerden spam ve normal maillerde duygu tonunun karşılaştırmalı olarak incelenmiştir.

Çalışma kapsamında elde edilen grup 1, grup 2 ve grup 3'e ait veri setlerinde sadece spam ve normal etiketleri yer almaktadır. Bu verilere farklı çercevelereden yaklaşılması amacıyla spam ve normal verilere NLTK kütüphanesine ait Sentiment Intensity Analyzer sınıfı ile duygu analizi yapılmıştır. Bu sınıf, metinlerin duygu analizinin yapılmasına olanak sağlar. Metin verilerini sayısal özelliklere dönüştürmek TfidfVectorizer kullanılmıştır. TfidfVectorizer, max_features değerini 5000, max_iter değeri 100 ve random_state değeri 42 olarak ayarlanmıştır. Bu analiz sonucu veriler pozitif, negatif ve nötr olmak üzere sınıflandırılmıştır. Şekil 5.3'te verilerin duygu analizi sonucu dağılımı gösterilmektedir. İlk işlem adımına ait veri setleri grup 1'de, ikinci işlem adımına ait veriler grup 2'de, üçüncü işlem adımına ait veriler grup 3'te listelenmiştir.



Şekil 5.3 Duygu analizi sonuçları

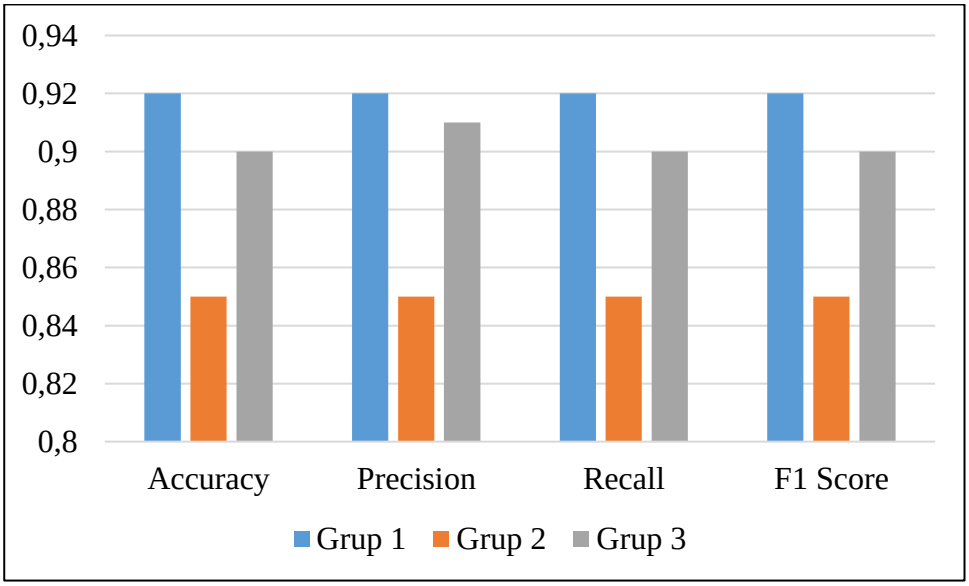
Şekil 5.3'teki verileri incelediğimizde göze çarpan ilk sonuç veri setlerinden Türkçe olmayan kelimelerin, bağlaçların ve dört kelimeden kısa maillerin çıkarılması, duygu dağılımının nötr ekseninde yoğunlaşmasına sebep olmuştur. Şekil 5.3'ten çıkartılan bir başka sonuçta normal ve spam verilerin duygu durumlarının paralel bir dağılım içerisinde

olmuş olmasıdır. Normal ve spam mailler duygu analizi sonucunda aynı duygu grupları ekseninde toplanmıştır.

5.4. NAIVE BAYES ile VERİLERİN SINIFLANDIRILMASI

Verilerden elde edilen altı veri seti çalışma kapsamında ilk olarak Naive Bayes ile sınıflandırılmıştır. Naive Bayes ile sınıflandırma sırasında aynı işlem adımları uygulanan veri setleri birleştirilmiş ve veriler spam ya da normal olarak etiketlenmiştir. Burada ki amaç veriler üzerinde yapılan işlem adımlarının sonuçlar üzerine etkisini gözlemlenmesidir. Naive Bayes sonucu modelin performansını değerlendirmek için accuracy, precision, recall ve f1 score parametreleri kullanılmıştır.

Şekil 5.4'te veri setlerine Naive Bayes uygulanması sonucu accuracy, precision, recall ve f1 score parametrelerinin sonuçları gösterilmiştir.



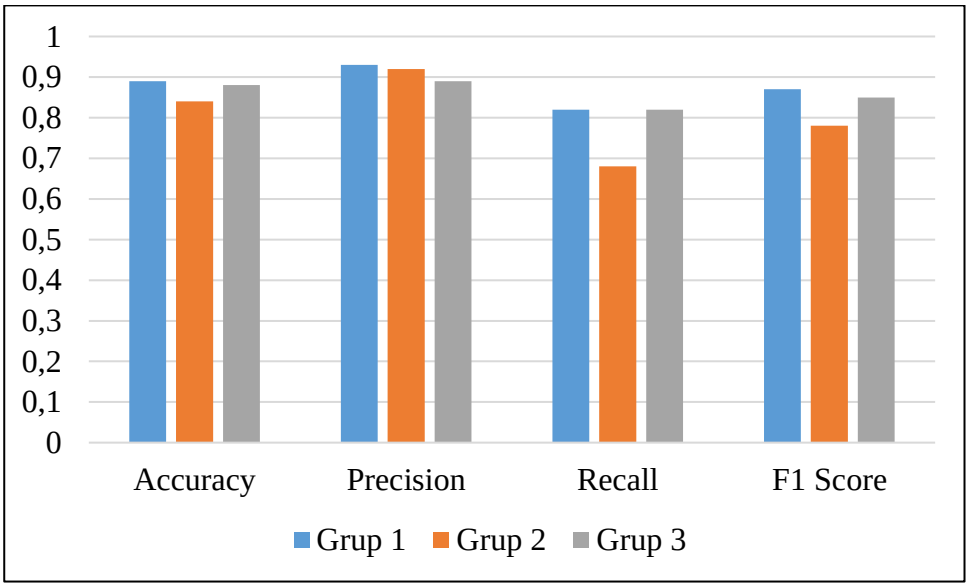
Şekil 5.4 Naive Bayes uygulanması sonucu accuracy, precision, recall ve f1 score parametrelerinin sonuçları

Şekil 5.4'teki veriler incelendiğinde en yüksek performansa sahip model grup 1 olmuştur. En düşük performansa sahip model ise grup 2 olmuştur. Bu değerler göz önüne alındığında veriler içerisinden Türkçe kökenli olmayan kelimelerin, bağlaçların ve 4 kelimeden kısa cümlelerin çıkarılması modelin performansını olumsuz etkilemektedir.

5.5. RANDOM FOREST ile VERİLERİN SINIFLANDIRILMASI

Random Forest modeli oluşturulurken sırasıyla yapılan işlemler, bağımsız değişkenler ve bağımlı değişkenler belirlenmiştir. Bağımsız değişkenler mail verilerimizdir, bağımlı değişkenlerimiz ise mail etiketleridir. Etiket değerleri sayısal değerlere dönüştürülmüştür. Verilerin %80'i eğitim %20'si test verisi olacak şekilde ayarlanmıştır. Metin verilerini sayısal özelliklere dönüştürmek için TfidfVectorizer kullanılmıştır. TfidfVectorizer, max_features değerini 5000, n_estimators değeri 100 ve random_state değeri 42 olarak ayarlanmıştır. Modelin sonuçlarını değerlendirmek için accuracy, precision, recall, f1-score parametrelerinden faydalanılmıştır.

Şekil 5.5'te Random Forest ile eğitilen modelin accuracy, precision, recall ve f1-score parametrelerine göre değerlendirme sonuçları yer almaktadır.



Şekil 5.5 Random Forest ile eğitilen modelin accuracy, precision, recall ve f1 score parametrelerinin sonuçları

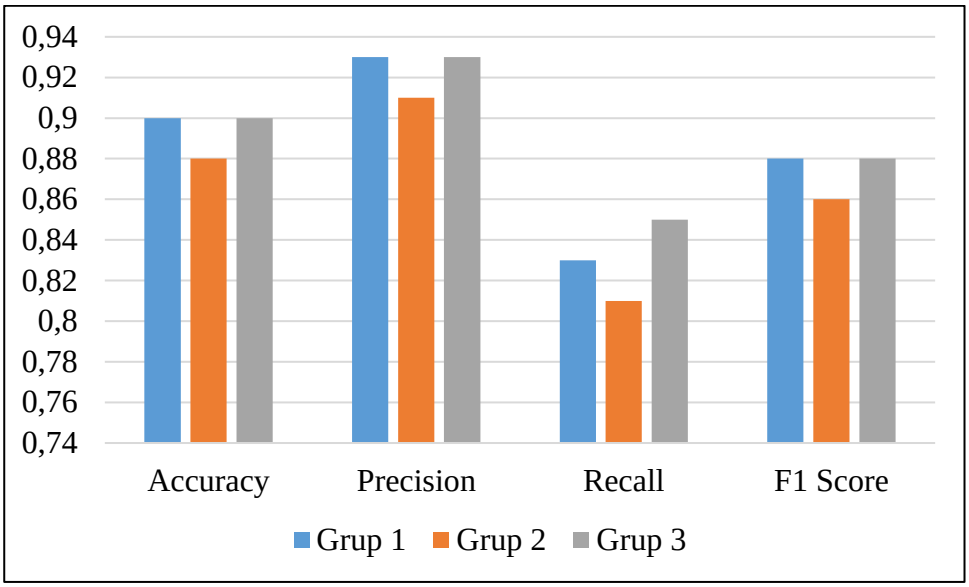
Şekil 5.5'teki ki veriler incelendiğinde grup 1 altında ki veriler diğer gruplara nispeten daha yüksek sonuçlar vermiştir. Türkçe kökenli olmayan kelimelerin, bağlaçların ve 4 kelimeden kısa cümlelerin veri setinden çıkarılması modelin performansını olumsuz etkilemektedir.

5.6. LOJİSTİK REGRESYON ile VERİLERİN SINIFLANDIRILMASI

Lojistik Regresyon modeli oluşturulurken sırasıyla yapılan işlemler, metin verileri (X) ve

etiketler (y) olarak belirlenmiştir. Etiket değerleri label encoder ile sayısal değerlere dönüştürülmüştür. Verilerin %80'i eğitim %20'si test verisi olacak şekilde ayarlanmıştır. Metin verilerini sayısal özelliklere dönüştürmek TfidfVectorizer kullanılmıştır. TfidfVectorizer, max_features değerini 5000, max_iter değeri 100 ve random_state değeri 42 olarak ayarlanmıştır. Modelin sonuçlarını değerlendirmek için accuracy, precision, recall, f1-score parametrelerinden faydalanılmıştır.

Şekil 5.6'da Lojistik Regresyon ile eğitilen modelin accuracy, precision, recall ve f1-score parametrelerine göre değerlendirme sonuçları yer almaktadır.



Şekil 5.6 Lojistik Regresyon ile eğitilen modelin accuracy, precision, recall ve f1 score parametrelerinin sonuçları

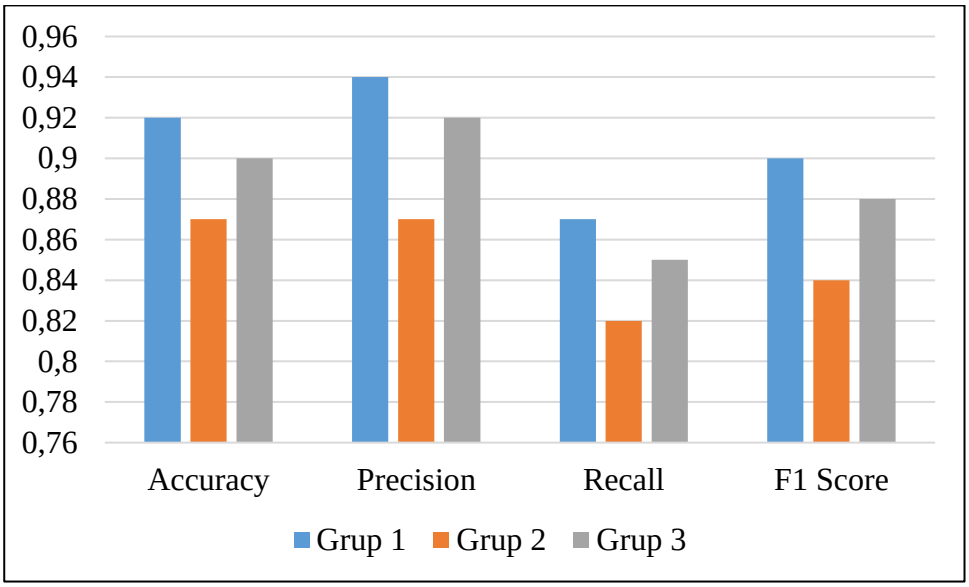
Şekil 5.6'da ki veriler incelendiğinde grup 3 altındaki veriler diğer gruplara nispeten daha yüksek sonuçlar vermiştir. 3 grup altındaki veriler incelendiğinde modellerin doğruluk değerleri birbirine oldukça yakındır. Türkçe kökenli olmayan ve 4 kelimeden kısa cümlelerin veri setinden çıkarılması modelin performansını olumsuz etkilemektedir. Ancak grup 2 aşamasında veri setinden çıkarılan veri satırları grup 1'den çıkarılması ile oluşturulan grup 3'te ki modelin performansını artırmaktadır. Dolaylı yoldan Türkçe kökenli olmayan ve 4 kelimeden kısa cümlelerin çıkarılması modelin performansını artırmıştır.

5.7. DESTEK VEKTÖR MAKİNESİ ile VERİLERİN SINIFLANDIRILMASI

DVM modeli oluşturulurken sırasıyla yapılan işlemler, metin verileri (X) ve etiketler (y)

olarak belirlenmiştir. Etiket değerleri label encoder ile sayısal değerlere dönüştürülmüştür. Verilerin %80'i eğitim %20'si test verisi olacak şekilde ayarlanmıştır. Metin verilerini sayısal özelliklere dönüştürmek TfidfVectorizer kullanılmıştır. TfidfVectorizer, max_features değerini 5000, max_iter değeri 100 ve random_state değeri 42 olarak ayarlanmıştır. Model çekirdeği olarak linear seçilmiştir. Linear çekirdek veriler doğrusal olarak ayrılabilir olan veri setlerinde kullanılabilir bir çekirdektir. Çalışma kapsamında veriler normal ve spam olarak ayrıştırılabilir durumdadır. Modelin sonuçlarını değerlendirmek için accuracy, precision, recall, f1-score parametrelerinden faydalanılmıştır.

Şekil 5.7'de DVM ile eğitilen modelin accuracy, precision, recall ve f1-score parametrelerine göre değerlendirme sonuçları yer almaktadır.



Şekil 5.7 DVM ile eğitilen modelin accuracy, precision, recall ve f1 score parametrelerinin sonuçları

Şekil 5.7'de ki veriler incelendiğinde grup 1 altındaki veriler diğer gruplara nispeten daha yüksek sonuçlar vermiştir. Türkçe kökenli olmayan kelimelerin, bağlaçların ve 4 kelimeden kısa cümlelerin veri setinden çıkarılması modelin performansını olumsuz etkilemektedir.

6. SONUÇ

Bu tez çalışmasında Türkçe e-maillerden oluşan veri setleri, K-means, Isolation Forest, duygu analizi, Naive Bayes, Random Forest, Lojistik Regresyon ve destek vektör makinesi yöntemleri ile test edilmiştir. Test edilen veri setine uygulanan üç işlem adımından sonra üç veri seti elde edilmiştir. Elde edilen ilk veri seti, verilerin küçük harflere dönüştürülmesi, web sitesi adreslerinin "website" ve e-mail adresleri "email" olarak yeniden olarak yeniden adlandırılması, noktalama işaretlerinin ve sayısal ifadelerin kaldırılması ile elde edilmiştir. İkinci veri seti ilk veri setinden bağlaçların, Türkçe kökenli olmayan kelimelerin ve dört kelimeden kısa cümlelerin çıkarılması ile elde edilmiştir. Üçüncü veri seti ilk veri seti ile ikinci veri setinin kesişim kümesidir.

Bu üç veri seti, K-means algoritması ile test edilerek olası spam ve olası normal e-mailler olarak sınıflandırılmıştır. Elde edilen bu veri setlerinde sırasıyla grup 1'de 998 normal 844 spam, grup 2'de 847 normal 776 spam ve grup 3'te 847 normal 776 spam verilerden oluşmaktadır. K-means sonucunda grup 1'de yer alan 1842 verinin 174'ü olası spam 1668'i olası normal, grup 2'de yer alan 1623 verinin 1601'i olası spam 22'si olası normal, grup 3'te yer alan 1623'te yer alan 199'u olası spam 1421'ı olası normal olarak etiketlenmiştir.. Isolation forest ile veri setleri üzerinde olası spam tespiti yaptığımızda grup 1'de yer alan 1842 verinin 184'ü olası spam 1658'i olası normal, grup 2'de yer alan 1623 verinin 163'ü olası spam 1460'ı olası normal, grup 3'te yer alan 1623'te yer alan 163'ü olası spam 1460'ı olası normal olarak etiketlenmiştir. Türkçe kökenli olmayan kelimelerin veri setlerinden çıkarılması spam k-means davranışını önemli ölçüde etkilerken isolation forest sonuçlarında etkisi olmamıştır. Duygu analizi sonuçlarında ise veri setlerinden Türkçe olmayan kelimelerin, bağlaçların ve dört kelimeden kısa maillerin çıkarılması, duygu dağılımının nötr ekseninde yoğunlaşmasına sebep olmuştur. Naive Bayes, Random Forest, Lojistik Regresyon ve destek vektör makinesi yöntemleri veri setlerinin test edilmesi sonucunda ilk veri setini üzerinde Naive Bayes ve DVM 0,92 doğruluk değeri ile en başarılı sonucu verirken Lojistik Regresyon ile 0,90 ve Random Forest ile 0,89 doğruluk değerleri elde edilmiştir. İkinci veri seti üzerinde ise Lojistik Regresyon 0,88 doğruluk değeri ile en başarılı yöntem olurken, DVM 0,87, Naive Bayes 0,85 ve Random Forest 0,84 doğruluk değerleri elde edilmiştir. Üçüncü ve son veri

setinde ise Naive Bayes, DVM ve Lojistik Regresyon 0,90 doğruluk değerleri ile en başarılı sonucu vermiş Random Forest ile 0,88 doğruluk değeri elde edilmiştir.

Veri setinden bağlaçların, Türkçe kökenli olmayan kelimelerin ve dört kelimeden kısa cümlelerin çıkarılması Naive Bayes, Random Forest, Lojistik Regresyon ve destek vektör makinesi modellerinin doğruluk değerini düşürmektedir. Sadece mail içeriklerinin değerlendirilmesinden ziyade konu başlıklarının değerlendirilmesi spam ve normal maillerin eğilimlerinin daha geniş bir çerçeveden değerlendirilmesine olanak sağlayabilir.

7. KAYNAKLAR

[1] Prodanoff, J. T. . *21 Must-read stats about how many emails are sent per day*. [Online]. Erişim: https://webtribunal.net/blog/how-many-emails-are-sent-per-day/.

[2] Z. A. Güven, "Türkçe e-postalarda spam tespiti için makine öğrenme yöntemlerinin ve dil modellerinin analizi," *Avrupa Bilim ve Teknoloji Dergisi*, c. 47, 1-6, 2023.

[3] E:E. Eryılmaz, & E. Kılıç, "İstenmeyen maillerin tespiti için kullanılan yöntemlerin incelenmesi," *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi,* c. 11(3), 977-987, 2020.

[4] C. Özdemir, M. Ataş & A. B. Özer, "Classification of Turkish spam e-mails with artificial immune system," *In 2013 21st Signal Processing and Communications Applications Conference (SIU)*, 1-4, 2013

[5] B. A. Kumari & C. Nagaraju, "Robust machine learning technique for detection and classification of spam mails," *EasyChair*, 1-8, 2023.

[6] S. Thanarattananakin, S. Bulao, B. Visitsilp & M. Maliyaem, "Spam detection using word embedding-based LSTM," *In 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, 227-231, 2022.

[7] C. T. Udogwu, "Ensemble Classification Method for Email Spam Prediction", Doktora tezi,Dublin İrlanda Ulusal Koleji , 2021.

[8] C. A. DePaolo & K. Wilkinson K, "Get your head into the clouds: using word clouds for analyzing qualitative assessment data," *TechTrends, vol. 58, no. 3*, 38–44, 2014.

[9] R. Atenstaedt R. "Word cloud analysis of the BJGP," *The British Journal of General Practice : the Journal of the Royal College of General Practitioners, vol. 62, no. 596*, 148–148, 2012.

[10] W. Medhat, A. Hassan & H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Snormals engineering journal, 5(4)*, 1093-1113, 2014.

[11] B. Liu, “Sentiment analysis and opinion mining,” *Springer Nature*, 2022.

[12] A. Agarwal, B. Xie, I. Vovsha, O. Rambow & R. J. Passonneau, “Sentiment analysis of twitter data,” *In Proceedings of the workshop on language in social media (LSM 2011)*, 30-38, 2011.

[13] H. Saif, Y. He & H. Alani, H, “Semantic sentiment analysis of twitter,” In *The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11*, 508-524, 2012.

[14] B. Bostancı & A. Albayrak, “Duygu analizi ile kişiye özel içerik önermek”, *Veri Bilimi, 4(1)*, 53-60, 2021.

[15] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 14*, 281-297, 1967.

[16] J. Wang, X. Zhang & H. Zhou, “A genetic k-means algorithm for spatial clustering,” *Computer Engineering, 3*, 188-190, 2006.

[17] Z. Zhang, J. Zhang & H. Xue, “Improved K-means clustering algorithm,” *In 2008 Congress on Image and Signal Processing Vol. 5*, 169-172, 2008.

[18] F. T. Liu,K. M. Ting & Z. H. Zhou, “Isolation forest,” *In 2008 eighth ieee international conference on data mining*, 413-422, 2008.

[19] L. Zhang, L. Jiang, C. Li & G. Kong, “Two feature weighting approaches for naive Bayes text classifiers,” *Knowledge-Based Systems, 100*, 137-144, 2016.

[20] P. Langley & S. Sage, “Induction of selective Bayesian classifiers,” *In Uncertainty Proceedings 1994*, 399-406, 1994.

[21] P. Bermejo, J. A. Gámez & J.M. Puerta, “Speeding up incremental wrapper feature subset selection with Naive Bayes classifier,” *Knowledge-Based Systems, 55*, 140-147, 2014.

[22] S. Chen, G. I. Webb, L. Liu & X. Ma, “A novel selective naïve Bayes algorithm,” *Knowledge-Based Systems, 192, 105361,* 1-12, 2020

[23] G. Biau & E. Scornet, "A random forest guided tour," *Test, 25*, 197-227, 2016.

[24] S. S. Keerthi, K. B. Duan, S. K. Shevade & A. N. Poo, "A fast dual algorithm for kernel logistic regression," *Machine learning, 61*, 151-165, 2005.

[25] X. Zou, Y. Hu, Z. Tian & K. Shen, "Logistic regression model optimization and case analysis," *In 2019 IEEE 7th international conference on computer science and network technology (ICCSNT),* 135-139, 2019.

[26] L. Wang, "Support vector machines: theory and applications," *Springer Science & Business Media*, 2005

[27] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt & B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications, 13(4)*, 18-28, 1998.

[28] J. Huang, Y. F. Li & M. Xie, "An empirical analysis of data preprocessing for machine learning-based software cost estimation," *Information and Software Technology, 67*, 108-127, 2015.

[29] S. B. Kotsiantis, D. Kanellopoulos & P. E. Pintelas, "Data preprocessing for supervised leaning," *International Journal of Computer Science, 1(2),* 111-117, 2006.

[30] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti & M. Alazab, "A comprehensive survey for intelligent spam email detection," *IEEE Access, 7*, 168261-168295, 2019.

[31] Q. Peng & M. Zhong, "Detecting spam review through sentiment analysis," *J. Softw., 9(8),* 2065-2072, 2014.

[32] E. Ezpeleta, I. Velez de Mendizabal, J. M. G. Hidalgo & U. Zurutuza, "Novel email spam detection method using sentiment analysis and personality recognition," *Logic Journal of the IGPL, 28(1)*, 83-94, 2020.

[33] M. Takaoğlu & F. Takaoğlu, "K-means ve hiyerarşik kümeleme algoritmanın weka ve matlab platformlarında karşılaştırılması," *İstanbul Aydın Üniversitesi Dergisi,11(3)*, 303-317, 2019.

[34] S. Hariri, M. C. Kind & R. J. Brunner, "Extended isolation forest," *IEEE transactions on knowledge and data engineering, 33(4),* 1479-1489, 2019.

[35] W. A. Awad & S. M. Elseuofi, "Machine learning methods for spam e-mail classification," *International Journal of Computer Science & Information Technology (IJCSIT)*, 3(1), 173-184, 2011.

[36] N. Kumar & S. Sonowal, "Email spam detection using machine learning algorithms," *In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)* pp. 108-113, 2020.

[37] Q. Yaseen, "Spam email detection using deep learning techniques," *Procedia Computer Science*, 184, 853-858, 2021.

[38] P. Sharma & U. Bhardwaj, "Machine Learning based Spam E-Mail Detection," *International Journal of Intelligent Engineering & Systems*, 11(3), 2018.

[39] L. GuangJun, S. Nazir, H. U. Khan & A. U. Haq, "Spam detection approach for secure mobile message communication using machine learning algorithms," *Security and Communication Networks*, 1-6, 2020.

[40] Z. B. Siddique, M. A. Khan, I. U. Din, A. Almogren, I. Mohiuddin & S. Nazir, "Machine learning-based detection of spam emails," *Scientific Programming*, 1-11, 2021.

[41] M. Gupta, A. Bakliwal, S. Agarwal & P. Mehndiratta, "A comparative study of spam SMS detection using machine learning classifiers," *In 2018 eleventh international conference on contemporary computing*,1-7, 2018

[42] S. Nandhini & J. M. KS, J. M., "Performance evaluation of machine learning algorithms for email spam detection," *In 2020 International Conference on Emerging Trends in Information Technology and Engineering*,1-4, 2020.

[43] Y. Kontsewaya, E. Antonov & A. Artamonov, "Evaluating the effectiveness of machine learning methods for spam detection," *Procedia Computer Science*, 190, 479-486, 2021.

[44] S. Gibson, B. Issac, L. Zhang & S.M. Jacob, "Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms," *Ieee Access*, 8, 187914-187932, 2020.

[45] S. K. Trivedi, "A study of machine learning classifiers for spam detection," *In 2016 4th international symposium on computational and business intelligence*, 176-180, 2016.

[46] D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi & P. L. Surya, "Spam classification based on supervised learning using machine learning techniques," *In 2011 International Conference on Process Automation, Control and Computing* , 1-7, 2011.

[47] A.S. Aski & N.K. Sourati, "Proposed efficient algorithm to filter spam using machine learning techniques," *Pacific Science Review A: Natural Science and Engineering*, 18(2), 145-149, 2016.

[48] M. Bassiouni, M. Ali & E. A. El-Dahshan, E. "Ham and spam e-mails classification using machine learning techniques," *Journal of Applied Security Research*, 13(3), 315-331, 2018.

[49] P. Charanarur, H. Jain, G. S. Rao, D. Samanta, S. S. Sengar & C.T. Hewage, "Machine-Learning-Based Spam Mail Detector," *SN Computer Science*, 4(6), 858, 2023.

[50] P.K. Panigrahi, "A comparative study of supervised machine learning techniques for spam e-mail filtering" *In 2012 Fourth International Conference on Computational Intelligence and Communication Networks*,506-512, 2012.

[51] B. Yu & Z. B. Xu, "A comparative study for content-based dynamic spam classification using four machine learning algorithms," *Knowledge-Based Systems*, 21(4), 355-362, 2008.

[52] A. Rayan, "Analysis of e-Mail Spam Detection Using a Novel Machine Learning-Based Hybrid Bagging Technique," *Computational Intelligence and Neuroscience*, 2022.

[53] S. Srinivasan, V. Ravi, M. Alazab, S. Ketha, A. M. Al-Zoubi & S. Kotti Padannayil, "Spam emails detection based on distributed word embedding with deep learning," *Machine intelligence and big data analytics for cybersecurity applications*, 161-189, 2021.

[54] F. Hossain, M. N. Uddin & R. K. Halder, "Analysis of optimized machine learning and deep learning techniques for spam detection," *In 2021 IEEE International IOT, Electronics and Mechatronics Conference*, pp. 1-7, 2021.

[55] C. C. Lai, "An empirical study of three machine learning methods for spam filtering," *Knowledge-Based Systems*, 20(3), 249-254, 2007.

[56] T. Vyas, P. Prajapati & S. Gadhwal, "A survey and evaluation of supervised machine learning techniques for spam e-mail filtering," *In 2015 IEEE international conference on electrical, computer and communication technologies*, 1-7, 2015.

[57] A. S. Mashaleh, N. F. B. Ibrahim, M. A. Al-Betar, H.M. Mustafa & Yaseen, Q. M., "Detecting spam email with machine learning optimized with Harris Hawks optimizer (HHO) algorithm," *Procedia Computer Science*, 201, 659-664, 2022.

[58] M. Wankhade, A. C. S. Rao & C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, 55(7), 5731-5780, 2022.

[59] P. Bhatia, Y. Ji & J. Eisenstein, "Better document-level sentiment analysis from rst discourse parsing," *arXiv preprint arXiv:1509.01599*, 2015.

[60] B. Yang & C. Cardie, "Context-aware learning for sentence-level sentiment analysis with posterior regularization," *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 325-335, 2014.

[61] T. T. Thet, J. C. Na, J. C., & C. S. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards," *Journal of information science*, 36(6), 823-848, 2010.

[62] K. Schouten & F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE transactions on knowledge and data engineering*, 28(3), 813-830, 2015.

[63] B. Lu, M. Ott, C. Cardie & B. .K. Tsou, B. K., "Multi-aspect sentiment analysis with topic models," *In 2011 IEEE 11th international conference on data mining workshops*, 81-88, 2011.

[64] I. Rish, "An empirical study of the naive Bayes classifier," *In IJCAI 2001 workshop on empirical methods in artificial intelligence*, c. 3, 41-46, 2001.

[65] G. I. Webb, E. Keogh & R. Miikkulainen, "Naïve Bayes," *Encyclopedia of machine learning*, c. 15(1), 713-714, 2010.

[66] Ray, S. (2024). *Naive Bayes classifier explained: applications and practice problems of Naive Bayes classifier*. [Online], Erişim: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.

[67] Tafralı, S. (2022). *Yapay öğrenme: Rastgele Orman*. [Online], Erişim: https://medium.com/machine-learning-türkiye/yapay-öğrenme-rastgele-orman-e8debdc886e7.

[68] IBM. *What is random forest?*. [Online], Erişim: https://www.ibm.com/topics/logistic-regression.

[69] IBM. *What is logistic regression?*. [Online], Erişim: https://www.ibm.com/topics/random-forest.

[70] Kanece, V. (2022). *What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices.* [Online], Erişim: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

[71] A. Şerbetçi, "Sıralı lojistik regresyon analizi ile istatistik ve ekonometri, derslerinde başarıyı etkileyen faktörlerin belirlenmesi: Atatürk Üniversitesi İktisadi ve İdari Bilimler Fakültesi öğrencileri üzerine bir uygulama," *Kahramanmaraş Sütçü İmam Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, c. 3(1), 89-110, 2013.

[72] Saini, A. (2024). *Guide on support vector machine (SVM) algorithm.* [Online], Erişim: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinesSVM-a-complete-guide-for-beginners/

[73] Şeker, Ş. E. (2008). *DVM (suppor vector machine, Destekçi Vektör Makinesi)*. [Online], Erişim: https://bilgisayarkavramlari.com/2008/12/01/DVM-support-vector-machine-destekci-vektor-makinesi/

[74] Tabsharami, F. *support vector machine (SVM).* [Online], Erişim: https://www.techtarget.com/whatis/definition/support-vector-machine-SVM

[75] H. Bhavsar & M. H. Panchal. "A review on support vector machine for data classification," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, c. 1(10), 185-189, 2012.

[76] S. Seo, Y. Kim, H. J. Han, W. C. Son, Z. Y. Hong, I. Sohn, ... & C. Hwang, "Predicting successes and failures of clinical trials with outer product–based convolutional neural network," *Frontiers in Pharmacology*, 12, 670670., 2021

[77] Kirubasagar V. (2023).*Random Forest.* [Online], Erişim: https://www.linkedin.com/pulse/random-forest-kirubasagar-v/

[78] Öğündür G. (2019). *Doğruluk (Accuracy) , Kesinlik(Precision) , Duyarlılık(Recall) ya da F1 Score ?*. [Online], Erişim: https://medium.com/@gulcanogundur/doğruluk-accuracy-kesinlik-precision-duyarlılık-recall-ya-da-f1-score-300c925feb38

[79] K. Chen, Z. Zhang, J. Long & H. Zhang. "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Systems with Applications*, c. 66, 245-260., 2016.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Yunus Emre PALAVAR

Yabancı Dili : İngilizce

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	Siber Güvenlik A.B.D.	Düzce Üniversitesi	2024
Lisans	Bilgisayar Müh.	Düzce Üniversitesi	2021
Lise	Tıbbi Laboratuvar Teknisyenliği	Anadolu Sağlık Meslek Lisesi	2016

YAYINLAR

Tez Kapsamında Çıkan Yayınlar

1. Y. E. Palavar & A. Albayrak, "Türkçe e-maillerin duygu analizi ve makine öğrenmesi yöntemleri ile morfolojik analizi," *Akıllı sistemlerin endüstriyel uygulaması I,* 1. baskı, Ankara, Türkiye: BİDGE Yayınları, 2023, böl. 8, ss. 187-209.

Diğer Yayınlar

1. Y. E. Palavar, Z. Çelik, A. Albayrak, E. Özçelik, & M. Erdil, "Analysis of the COVID-19 process in terms of health managers," *In 2022 2nd International Conference on Computing and Machine Intelligence* ,1-5, 2022.