T.C.
SAKARYA ÜNİVERSİTESİ
İŞLETME ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ İLE METİN SINIFLANDIRMA: BAKIM YÖNETİM SİSTEMİ ÖRNEĞİ

YÜKSEK LİSANS TEZİ

İbrahim Burak TOSUN

Enstitü Anabilim Dalı : Yönetim Bilişim Sistemleri

Tez Danışmanı: Dr.Öğr.Üyesi Çağla EDİZ

MAYIS – 2021

İbrahim Burak Tosun tarafından hazırlanan “Makine Öğrenmesi ile Metin Sınıflandırma: Bakım Yönetim Sistemi Örneği” başlıklı bu tez, 29/06/2021 tarihinde Sakarya Üniversitesi Lisansüstü Eğilim ve Öğretim Yönetmeliği'nin ilgili maddeleri uyarınca yapılan Tez Savunma Sınavı sonucunda başarılı bulunarak, jürimiz tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Danışman: Dr.Öğr.Üyesi Çağla EDİZ
Sakarya Üniversitesi

Jüri Üyeleri: Prof Dr. Aykut Hamit TURAN
Sakarya Üniversitesi

Doç.Dr. Derya BİRANT
Dokuz Eylül Üniversitesi

SAKARYA ÜNİVERSİTESİ	**T.C.** **SAKARYA ÜNİVERSİTESİ** **İŞLETME ENSTİTÜSÜ** **TEZ SAVUNULABİLİRLİK VE ORJİNALLİK BEYAN FORMU**	Sayfa : 1/1

Öğrencinin

Adı Soyadı	:	**İbrahim Burak TOSUN**
Öğrenci Numarası	:	**Y176054006**
Enstitü Anabilim Dalı	:	**Yönetim Bilişim Sistemleri**
Enstitü Bilim Dalı	:	**Yönetim Bilişim Sistemleri**
Programı	:	☑ YÜKSEK LİSANS ☐ DOKTORA
Tezin Başlığı	:	**Makine Öğrenmesi ile Metin Sınıflandırma: Bakım Yönetim Sistemi Örneği**
Benzerlik Oranı	:	**%5**

ENSTİTÜSÜ MÜDÜRLÜĞÜNE

☑ Sakarya Üniversitesi İşletme Enstitüsü Lisansüstü Tez Çalışması Benzerlik Raporu Uygulama Esaslarını inceledim. Enstitünüz tarafından Uygulalma Esasları çerçevesinde alınan Benzerlik Raporuna göre yukarıda bilgileri verilen tez çalışmasının benzerlik oranının herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi beyan ederim.

31/05/2021
İmza

☑ Sakarya Üniversitesi İşletme Enstitüsü Lisansüstü Tez Çalışması Benzerlik Raporu Uygulama Esaslarını inceledim. Enstitünüz tarafından Uygulalma Esasları çerçevesinde alınan Benzerlik Raporuna göre yukarıda bilgileri verilen öğrenciye ait tez çalışması ile ilgili gerekli düzenleme tarafımca yapılmış olup, yeniden değerlendirilmek üzere gsb@sakarya.edu.tr adresine yüklenmiştir.

Bilgilerinize arz ederim.

31/05/2021
İmza

Uygundur

Danışman
Unvanı / Adı-Soyadı: Dr.Öğr.Üyesi Çağla EDİZ

Tarih:

İmza:

☐ KABUL EDİLMİŞTİR ☐ REDDEDİLMİŞTİR **EYK Tarih ve No:**	**Enstitü Birim Sorumlusu Onayı**

İÇİNDEKİLER

KISALTMALAR

AI : Artificial Intelligence (Yapay Zeka)

ANN : Artificial Neural Network (Yapay Sinir Ağları)

BoW : Bag of Words (Kelime Torbası)

DT : Decision Tree (Karar Ağacı)

FP : False Positive (Yanlış Pozitif)

FN : False Negative (Yanlış Negatif)

KNN : K-Nearest Neighbors (K-En Yakın Komşu)

LR : Logistic Regression (Lojistik Regresyon)

MaxEnt : Maximum-Entropy Classification (Maksimum-Entropi Sınıflandırması)

NB : Naive Bayes

NLP : Natural Language Processing (Doğal Dil İşleme)

TF-IDF : Term Frequency - Inverse Document Frequency (Terim Frekansı – Ters Doküman Frekansı)

RNN : Recurrent Neural Network (Yinelenen Sinir Ağları)

RF : Random Forest (Rastgele Orman)

SMO : Sequential Minimal Optimization (Sıralı Minimum Optimizasyon)

SVC : Support Vector Classification (Destek Vektörü Sınıflandırması)

SVM : Support Vector Machine (Destek Vektörü Makinesi)

SGDC : Stochastic Gradient Descent – Classifier (Stokastik Gradyan Azalma Sınıflandırıcısı)

TP : True Positive (Gerçek Pozitif)

TN : True Negative (Gerçek Negatif)

TABLO LİSTESİ

ŞEKİL LİSTESİ

Tezin Başlığı: Makine Öğrenmesi ile Metin Sınıflandırma: Bakım Yönetim Sistemi Örneği

Tezin Yazarı: İbrahim Burak TOSUN **Danışman:** Dr.Öğr.Üyesi Çağla EDİZ

Kabul Tarihi: 29.06.2021 **Sayfa Sayısı:** vii (ön kısım)+ 37(tez)

Anabilim Dalı: Yönetim Bilişim Sistemleri

Günümüzde firmalar için bilgiyi yönetmek, süreçleri kişilerden bağımsız hale getirmek açısından önem arz etmektedir. Bir çalışan tarafından elde edilen bilgi diğer çalışanlara da ulaştırılabilmeli ve çalışan iş yerinden ayrılsa dahi bilgi firmada kalıcı olmalıdır. Bu nedenle, firmaların farklı senaryolar için farklı bilgi yönetim sistemlerine ihtiyaçları giderek artmaktadır.

Firmalarda kurulu makine, ekipman, ısıtma ve soğutma sistemleri gibi tüm varlıkların yönetilmesi, bakım süreçlerinin sağlıklı bir şekilde organize edilmesi, bakım işlerinin zamanında yapılması, yapılan bakım işlerinin takip edilebilmesi ve bakımların daha ekonomik yapılabilmesi için elektronik ortamda bir yönetim programı kullanmasının gerekliliği ortaya çıkmaktadır. Bu gerekliliği karşılamaya yönelik yazılan varlık ve bakım yönetimi programları yalnızca etkin bir şekilde kullanılabildiğinde istenen faydaları sağlamaktadır. Çalışmada iplik firmasında kullanılmakta olan varlık ve bakım yönetim programının ana işlevlerinden olan varlıklarda oluşan arızaların kaydedilmesi sürecine odaklanılmaktadır. Kayıt sürecinin kolaylaştırılması, hızlandırılması ve insan hatasının azaltılması üzerine önerilen makine öğrenmesi tekniği anlatılmaktadır. Bu teknik ile kayıt esnasında yazılan arıza açıklamasının makine öğrenmesi ile sınıflandırılarak ilgili arıza kategorisine atanması amaçlanmaktadır. Çalışma ile veri girişi yapanların işlerinin kolaylaşması ve hızlanması, bakımı yapacakların doğru bilgiye ulaşması, sistemin daha etkin ve verimli kullanılmasının süreci iyileştirmesi beklenmektedir.

Bu çalışmada, makine öğrenmesi ile bakım yönetim sistemindeki arıza açıklamalarının, arıza tanımlarına göre sınıflandırılması için öznitelikler kelime torbası yöntemleri ile çıkarılmış ve model lojistik regresyon, lineer destek vektörü sınıflandırması (SVC), stokastik gradyan azalma sınıflandırıcısı (SGDC) ve Naive Bayes (NB) algoritmaları ile eğitilmiştir. Lineer SVC algoritması ve terim frekansı – ters doküman frekansı (TF-IDF) öznitelik çıkarma yöntemi ile bakım yönetim sistemindeki tüm arızaların %72,12'sine denk gelen 46 arıza tanımının tahminini %87 başarı ile yapabilen model oluşturulmuştur.

Anahtar Kelimeler: Makine Öğrenmesi, Metin Sınıflandırma, Kelime Torbası

Title of Thesis: Text Classification with Machine Learning: An Example of a Maintenance Management System
Author of Thesis: İbrahim Burak TOSUN **Supervisor:** Assist. Prof. Çağla EDİZ
Accepted Date: 29.06.2021 **Np:** vii (pre text) + 37(main body)
Department: Management İnformation Systems

Today, it is important for companies to manage information and to make processes independent from individuals. Information obtained by an employee should be able to be conveyed to other employees and the information should be permanent in the company even if the employee leaves the workplace. For this reason, the needs of companies for information management systems are increasing.

It is necessary to use an electronic management program in order to manage all assets such as machinery, equipment, heating and cooling systems installed in companies, to organize maintenance processes in a healthy way. Asset and maintenance management programs written to meet this requirement provide the desired benefits only when they can be used effectively. This study focuses on the process of recording malfunctions in assets, which is one of the main functions of the asset and maintenance management program used in the yarn company. The recommended machine learning technique for facilitating and accelerating the registration process and reducing human error is explained. With this technique, it is aimed to classify the fault description written during the recording with machine learning and assign it to the relevant fault category. With the study, it is expected that the work of those who enter data will become easier and faster, those who will perform maintenance have access to the right information, and the more effective and efficient use of the system will improve the process.

In order to classify the fault descriptions in the maintenance management system according to the fault definitions with machine learning, in this study, the attributes were extracted with bag of word methods and the model was trained with logistic regression, linear support vector classification (SVC), stochastic gradient descent - classifier (SGDC), Naive Bayes (NB) algorithms. With the Linear SVC algorithm and the TF-IDF feature extraction method, a model was created that can predict 46 fault definitions corresponding to 72.12% of all failures in the maintenance management system with 87% success.

Keywords: Machine Learning, Text Classification, Bag of Words

GİRİŞ

Sürekli artan rekabet ortamında artık firmaların ürünlerinin kalitelerini arttırmaya ve fiyatlarını düşürmeye çalışmaları tek başına yeterli olmamaktadır. Kendi içlerindeki veya tedarikçileri ve müşterileri gibi dışlarındaki kaynaklardan elde edecekleri verilerden bilgi üretmeleri zorlu rekabet ortamında ayakta kalabilmeleri açısından önem arz etmektedir.

Bakım masrafları firmalarda en büyük giderlerden olup bu masrafların azaltılmasında günümüzde bilgi yönetim sistemlerinin kullanılması zorunluluk haline gelmektedir. Bu amaçla hazırlanan varlık ve bakım yönetim programları firmalardaki varlıklara ait bakım ve arızlarının kayıt altına alınmasını sağlamakta, bu kayıtlarla ise makine ve bakım personeli performansı gibi birçok bilgi elde edilebilmektedir. Programlardan istenilen faydanın sağlanabilmesi için etkin kullanılmaları gerekmektedir. Etkin kullanılmalarının önünde personel isteksizliği, eksik veya yanlış veri girilmesi gibi engeller bulunmaktadır.

Çalışma kapsamında varlık ve bakım yönetim sisteminin etkin kullanılmasındaki engellerden olan tekrarlı işlerin bulunmasından dolayı oluşan personel isteksizliği, verilerin eksik veya yanlış girilmesi problemlerinin azaltılabilmesi için makine öğrenmesi ile metin sınıflandırma yöntemi önerilmektedir. Bu yöntem ile arıza kaydı esnasında yazılan arıza açıklamasının makine öğrenmesi ile sınıflandırılarak ilgili arıza kategorisine atanması amaçlanmaktadır.

Çalışmanın ilk bölümünde bakım yönetim sistemleri hakkında genel bilgiler verilmekte, makine öğrenmesi ve türleri açıklanmakta, metinlerdeki doğal dilin nasıl işleneceğinden, çalışma kapsamında verilere uygun olarak seçilen makine öğrenmesi modellerinden ve performans ölçülerinden bahsedilmektedir. İkinci bölümde alandaki çalışmalar incelenmekte ve özet tablo halinde verilmektedir. Üçüncü bölümde çalışmanın metodolojisi ve uygulaması açıklanmaktadır. Sonuç bölümde ise çalışmanın sonuçlarından ve önerilerden bahsedilmektedir.

Araştırmanın Amacı

Varlık ve bakım yönetim sisteminde kayıtlı bulunan personelin yazdığı arıza açıklamalarından, arızanın ait olduğu arıza tanımı kategorisinin tahmin edilebileceği makine öğrenmesi sınıflandırma modeli oluşturulması amaçlanmaktadır. Bu sayede sistemi kullanan personelin arıza kayıt sürelerinin azalması, verilerin eksik ve yanlış girilmesinin azalması ve personelin tekrarlı kayıt adımlarının kaldırılması hedeflenmektedir. Varlık ve bakım yönetim sisteminde sanal asistan şeklinde yalnızca ses ile arıza kayıtların oluşturulabileceği bir gelecekte ise bu çalışmanın sesten yazıya dönüşmüş metnin ilgili arıza sınıfına atanmasına zemin hazırlaması öngörülmektedir.

Araştırma Problemi ve Soruları

Varlık ve bakım yönetimi sisteminde bir varlık için arıza kaydı oluştururken; öncelikle varlık listesinden ilgili varlığın seçilmesi, üretimin durumu, arızanın aciliyeti, arıza açıklaması yazılması ve listeden ilgili arıza tanımı seçilmesi gerekmektedir. Analizlerin listelerdeki arıza tanımlarına göre yapılıyor olması nedeniyle arıza tanımı seçilmesi zorunlu olmakla beraber, açıklama yazıldıktan sonra tekrar tanımın seçiliyor olması kayıt oluşturma süresini uzatmakta ve personelin motivasyonunu düşürmektedir. Aynı zamanda bazı personellerin açıklama kısımlarını es geçmelerine yani çok kısa açıklama yazarak arızanın anlaşılamamasına neden olmaktadır. Ek olarak arıza tanımlarının yanlış seçildiği durumlar olmakta ve bunlar bakım ekibi tarafından tespit edilebilirse kayıt iptal edilmekte, tespit edilemez ise yanlış arızanın çözülmeye çalışılması ile iş gücü kaybı oluşturmaktadır. Bahsedilen problemlerin önüne geçmek için sistemde bulunan arıza kayıtları ile arıza açıklamaları metinlerinden ilgili arıza tanımının tahmini yapan modelin oluşturulabilirliği sorgulanmaktadır.

Araştırmanın Önemi

Çalışma ile varlık ve bakım yönetim sistemine arıza kaydı girme süresi kısalacaktır. Yılda yüz binin üzerinde kayıt girildiği düşünüldüğünde önemli ölçüde iş gücü süresi kazanılacaktır. Tanımlama hatalarının azalması ile iptal edilen kayıtlar ile yeniden doğru kayıt oluşturmada harcanan vakit azalacak ve daha doğru analizler elde edilebilecektir. İlerleyen süreçte bakım yönetim sistemine dahil edilmesi planlanan ses ile arıza kaydı oluşturma yöntemine (sesten metne dönüştürülen arıza açıklaması ile ilgi tanımın eşleştirilmesinin) zemin hazırlayacağı ön görülmektedir.

Araştırmanın Yöntemi

Varlık ve bakım yönetim sisteminde arıza kaydı oluştururken arıza açıklaması yazıldığında ya da gelecekte planlandığı gibi dikte edildiğinde açıklamayı önceden tanımlı arıza tanımlarına atayan modelin oluşturulması için makine öğrenmesi kullanılacaktır. Arıza tanımları önceden belirli olduğundan problem gözetimli makine öğrenmesi yöntemlerinden sınıflandırma yöntemi ile çözümlenmektedir. Makine öğrenmesi modelinin eğitilmesi için sınıflandırma algoritmaları kullanılarak Python yazılım dilinde kodlama yapılmaktadır. İlk aşama olarak varlık ve bakım yönetim sisteminden arıza kayıtları toplandı. Toplanan veriler ön işleme aşamalarından geçirildi. Standart hale getirilmiş temiz veriler yazılıma aktarıldı. Veri setindeki metinlerden sözlük oluşturularak öznitelikler elde edildi. Öznitelikler ve tahmin edilecek arıza tanımı etiketleri ile makine öğrenmesi sınıflandırma modelleri oluşturuldu, oluşturulan modeller karşılaştırıldı ve sonuçları değerlendirildi.

Araştırmanın Kısıtları

Çalışmanın problemlerinden olan tekrarlı işlerin personelde motivasyon eksikliği yaratması, talep açıklamalarına çok kısa açıklama yazılması ya da hiç açıklama yazmamak şeklinde veri setinde etkisini göstermektedir. Veri ön işleme kısmında 10 karakterin altında yazılan açıklamalar bu nedenle ayıklandı. Benzer olarak arızanın tamının yanlış ya da rastgele seçildiği durumlar veri seti için dezavantaj ortaya çıkarmaktadır. Bazı arızaların diğerlerinden daha sık meydana gelmesinden dolayı veri setindeki tanımlar dengeli dağılmamakta bu da başka bir dezavantaj sayılabilmektedir.

BÖLÜM 1: BAKIM YÖNETİM SİSTEMİ VE MAKİNE ÖĞRENMESİ KAVRAMLARI

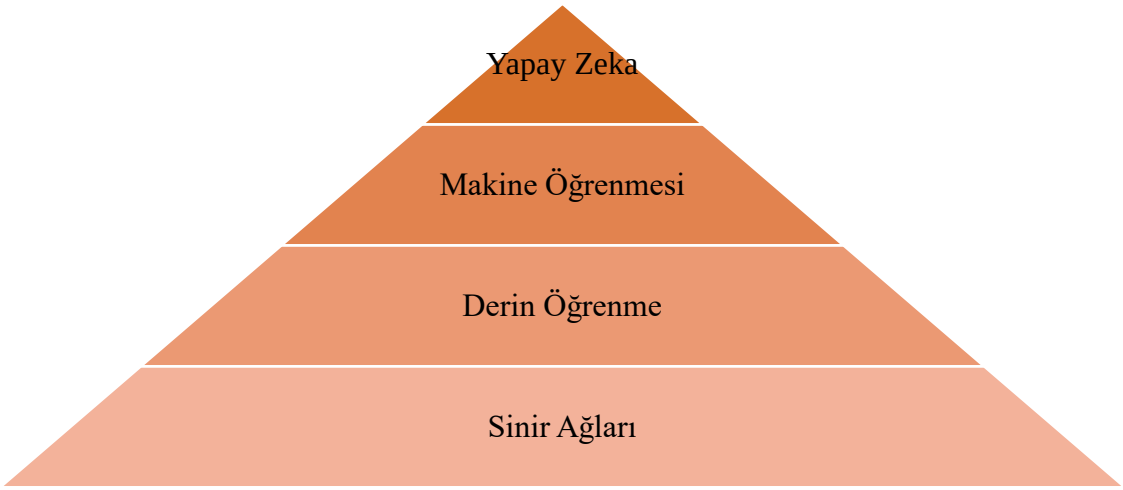
1.1. Bakım Yönetim Sistemi

Kullanılan ekipmanların çalışır durumda kalmasını sağlamak, arızalanmalarını önlemek ve arızalandıklarında ise çalışır hale getirmek için yapılan çalışmalara bakım adı verilmektedir (Özek, 2012). Bir başka ifade ile bakım makine ya da tesisin çalışır durumda kalması için yapılan çalışmalardır şeklinde özetlenebilir (Bal, 2013).

Bakım yönetimi ise işletmelerin tüm teknik ve fiziksel varlıklarının arıza yapmadan sürekli çalışır halde kalmalarını sağlayan fonksiyon faaliyetleridir. Temel fonksiyonları varlıkların korunması, kontrolü, bakımı ve daha verimli bakım hizmeti alınmasının sağlanmasıdır (Taşkın, 2006). Bakım işlemleri varlıklar üzerinden yürütüldüğü için bakım yönetim sistemlerinde aynı zamanda varlıkların yönetilmesi mümkün olmaktadır.

1.2. Makine Öğrenmesi

Makine öğrenmesi, makinelerin görevlerini başarı ile yapabilmeleri için verilerden beslenirken algoritmaları ve istatistiksel öğrenme yöntemlerini kullanan bir yapay zeka (Artificial Intelligence AI) bilimi dalıdır (Mohammed, Khan ve Bashier, 2016). Yapay zeka alt bilim dallarına Şekil 1'de yer verilmektedir.



Şekil 1: Yapay Zeka ve Alt Dalları

Makine öğrenmesindeki temel fikir makinenin öğrenmesini sağlayarak tahmin gibi görevleri gerçekleştirebilmesi için model geliştirmektir (Mohammed, Khan ve Bashier,

2016). Makine öğrenmesi; bireylerin manuel olarak kurallar üreterek ve büyük miktarda veriyi analiz ederek modeller oluşturmasını gerektirmek yerine, tahmine dayalı ve önceden belirlenmiş bir denkleme dayanmadan bilgileri doğrudan verilerden "öğrenmek" için istatiksel hesaplama yöntemlerini kullanır (Raschka, 2015).

Makine öğrenmesi günümüzde sürekli gelişen veri bilimi alanının en önemli bileşenlerinden biri olmakla beraber sınıflandırmalar, kümelemeler ve tahminler yapmak için kullanılmaktadır.

Örnek olarak;

- Finans alanında kredi skorlamada
- Biyoloji alanında tümör tespitinde
- Enerji üretimi alanında tüketim tahminde
- Üretim alanında kestirimci bakım çalışmalarında
- Doğal dil işleme alanında ses tanıma uygulamalarında kullanılmaktadır.

Makine öğrenmesi Gözetimli (Supervised) ve Gözetimsiz (Unsupervised) olmak üzere iki farklı teknikten oluşmaktadır (Şekil 2).

Şekil 2: Makine Öğrenmesi Teknikleri

Gözetimli öğrenme, gelecekteki çıktıları tahmin edebilmek için bilinen girdi ve çıktıları kullanırken gözetimsiz öğrenme ise girdilerdeki gizli örüntüleri ya da iç yapıları bulmaktadır.

1.2.1. Gözetimli Öğrenme

Gözetimli öğrenmede veri seti bir dizi girdi verisi ve bunların bilinen çıktılarından (etiketlerinden) oluşmaktadır. Gözetimli denilmesinin sebebi girdi verilerinin açıklaması (çıktısı) olan etiketlerin gözetmen tarafından verilmiş ya da belirlenmiş olmasıdır (Mohammed, Khan ve Bashier, 2016). Burada hedef veri setinde olmayan yeni veriler geldiğinde makul tahminler oluşturmak için bir model eğitilmesidir.

Gözetimli öğrenmede makine öğrenmesi modeli geliştirmek için sınıflandırma (classfication) ve regresyon (regression) teknikleri uygulanmaktadır.

- Sınıflandırma tekniği belirli bir veri setinde özniteliklere bağlı olarak girdilerin sınıflara ayırılmasında kullanılmaktadır. Kısaca bir ayrım mekanizmasıdır ve amaç bir kategoriyi ya da sınıfı tahmin etmektir. Örneğin; bir e-postanın spam ya da değil olarak sınıflandırılmasında, bir tümörün kanserli ya da iyi huylu olarak ayrılmasında kullanılır.
- Regresyon tekniği ise çıktı olarak sayısal bir değer tahmin edilmeye çalışıldığında kullanılmaktadır (Gollapudi ve Laxmikanth, 2016). Örneğin elektrik şebekesindeki tüketimin tahmininde veya bir ekipmanın arıza yapmasına kalan süresinin tahmininde kullanılır.

1.2.2. Gözetimsiz Öğrenme

Gözetimsiz öğrenmede etiketsiz veri setleri analiz edilir. Buradaki amaç veri setindeki gizli yapıların, örüntülerin veya veri gruplarının ortaya çıkarılmasıdır (Mohammed, Khan ve Bashier, 2016). Benzer özellikteki verilerden kümeler oluşturabildikleri için pazarlama stratejileri oluşturulmasında veya müşteri segmantasyonlarında kullanılırlar.

1.3. Doğal Dil İşleme

Doğal Dil İşleme (Natural Language Processing NLP), sosyal medya paylaşımları veya film diyalogları gibi günlük konuşma dilinin otomatik (veya yarı otomatik) olarak işlenmesi olarak tanımlanmaktadır. Doğal dil işleme bir bilgisayar bilimi alanı olup, makinenin insan doğal dilini işleme yeteneğidir (Thanaki, 2017).

1.3.1. Özellik Çıkarımı

Özellikler makine öğrenmesi algoritmaları için giriş parametreleridir, algoritmalar bu parametrelere göre çıktı ürettikleri için özellik çıkarımı işlemi model geliştirmenin en önemli adımlarındandır. Algoritmalar doğal dil verisi ile işlevlerini yerine getirememektedirler bu nedenle ham veriden özellik çıkarımı işlemine gerek duymaktadırlar (Thanaki, 2017).

- Kelime Torbası (Bag of Words) (BoW)

Kelime torbası yöntemi kelimelerin veya kelime gruplarının, metinde kaç kere tekrarlandığına veya metinde bulunup bulunmadığına bakılan yöntemdir. Metinde yalnızca kelimelerin bulunup bulunmadığına bakılıyor ise ikili özellik seçimi yöntemi olarak tanımlanırken, kelimelerin metindeki tekrarları sayılıyor ise terim frekans ağırlığı yöntemi olarak tanımlanmaktadır. Her kelime tek başına metin içerisinde aranıyor ise unigram, ikili aranıyor ise bigram ya da n adet grup halinde aranıyorsa n-gram modeli olarak tanımlanmaktadır (Türkalp, 2019).

Yöntemin uygulanmasında ilk olarak veri setindeki tüm benzersiz kelimelerden bir sözlük oluşturulur. Burada metinlerin ön işleme aşamalarından geçmiş temiz metin olmaları önemlidir aksi takdirde örneğin büyük harfle yazılmış bir kelimenin küçük harfle yazılmış halinin farklı olduğu varsayılacak ve sözlükte iki ayrı yer tutacaktır. Oluşturulan sözlükteki her bir kelime için ayrı bir sütun açılarak öznitelik (özellik) matrisi oluşturulur. Bu matrisin satırlarına veri setinin her bir kaydındaki metinlerin içlerindeki kelimelerin kaç kez tekrar ettiğini belirten rakamlar ile sütunlar doldurulur.

- Terim Frekansı – Ters Doküman Frekansı (Term Frequency – Inverse Document Frequency) (TF-IDF)

Kelime torbası metodundaki kelime sayma işlemi yalnızca kayıt bazında kelime sayıları ile ilgilenmekte iken TF-IDF metodu veri setini bir bütün olarak hesaplamaya katmaktadır. Özellik çıkarımı işleminin hedefi algoritma için girdileri hazırlamak olduğundan, veri setinde çok sık tekrar eden bir özelliğin algoritmaya katacağı bilgi azalacağından bu özelliklerden kaçınılmalıdır (Bonaccorso, 2018).

TF–IDF değeri, bir kelimenin kayıtta görünme sayısı ile orantılı olarak artar ve kelimeyi içeren veri setindeki kayıt sayısıyla azalır. Değer aşağıdaki formül ile hesaplanmaktadır (Thanaki, 2018).

$$\text{TF*IDF} = \left(\frac{kelimenin\ kayıtta\ görünme\ sayısı}{kayıttaki\ toplam\ kelime\ sayısı}\right) * \left(\log_{10}\frac{toplam\ kayıt\ sayısı}{kelimenin\ göründüğü\ kayıt\ sayısı}\right)$$

Yöntemin uygulanmasında ilk adım kelime torbası yöntemi ile aynıdır öncelikle sözlük oluşturulur. Oluşturan sözlükteki kelimeler matrisin sütunlarına yazılırken satırlarda her bir kayıtta bulunan kelimelerin TF–IDF değerleri ilgili sütunlara yazılarak matris oluşturulur.

1.4. Makine Öğrenmesi Modelleri

Makine öğrenmesi modelleri, makine öğrenmesi uygulamalarının merkezinde yer almaktadır. Modeller bir sistemde gözlenen verileri tanımlarken, yeni davranışların öğrenilmesine ve bunların tahmin edilmesine yardımcı olan yeni veri kümelerine uygulanmaktadır. Algoritmaların amacı ise mümkün olduğunca en doğru çıktıyı üretebilmektir. Algoritmalar öğrenme alt dallarına (gözetimli/gözetimsiz) ve problem kategorilerine (sınıflandırma, regresyon, kümeleme) göre sınıflandırılabilmektedir (Gollapudi, ve Laxmikanth, 2016).

Python yazılım dili için hazırlanmış olan Scikit-Learn modülü, gözetimli gözetimsiz birçok popüler makine öğrenmesi algoritmasını entegre biçimde sunmaktadır (Pedregosa vd., 2011). Farklı algoritmalar verileri farklı şekillerde işledikleri için kendilerine verilen görevin türüne ve verilerin tipine göre başarı oranları da değişiklik göstermektedir. Şekil 3'de farklı problemler için sıkça kullanılan algoritmalar verilmiş olup, gözetimli makine öğrenmesi sınıflandırma algoritmalarından çalışma kapsamında kullanılan dört algoritma açıklanmaktadır.

GÖZETİMLİ	**SINIFLANDIRMA**	GÖZETİMSİZ	**KÜMELEME**
	Lineer Destek Vektör Makinesi		Bölümleme Yöntemleri (K-Ortalamalar)
	Lojistik Regresyon		Hiyerarşik Yöntemler (AGNES)
	Rastgele Orman		Yoğunluk Tabanlı Yöntemler (DBSCAN, OPTICS)
	Karar Ağaçları		Izgara Tabanlı Yöntemler (STING, WaveCluster)
	Naïve Bayes		Model Tabanlı Yöntemler (COBWEB)
	Sinir Ağları		
	REGRESYON		
	Rastgele Orman		
	Karar Ağaçları		
	Sinir Ağları		
	Lineer Regresyon		

Şekil 3: Makine Öğrenmesi Algoritmaları

- Lojistik Regresyon (Logistic Regression)

İsminde regresyon geçmesine rağmen sınıflandırma problemleri için kullanılan doğrusal bir modeldir. Literatürde Logit-regression, maximum-entropy classification (MaxEnt) ve log-linear classifier isimleriyle de anılmaktadır.

- Lineer Destek Vektörü Sınıflandırılması (Linear Support Vector Classification SVC)

Lineer destek vektörü sınıflandırılması olarak tanımlanmaktadır. Veri setinde çok sınıflı sınıflandırma görevini gerçekleştirebilen bir algoritmadır. SVC'nin aksine lineer SVCbüyük veri setleri ile çalışabilme avantajına sahiptir (https://scikit-learn.org/stable/supervised_learning.html#supervised-learning, 2020).

- Stokastik Gradyan Azalma Sınıflandırıcısı (Stochastic Gradient Descent – Classifier SGDC)

Stokastik Gradyan Azalma Sınıflandırıcısı lojistik regresyon algoritmasına benzer şekilde çalışmaktadır ancak farklı olarak tüm veri setini dikkate almak yerine, güncelleme prosedürü veri setinden rastgele seçilen parçalara uygulanarak işletilmektedir. Bu özelliği ile bilgisayar hafızasına sığmayan veri setleri ile dahi çalışma imkanı sağlamaktadır (Bonaccorso, 2018).

- Naive Bayes Sınıflandırıcısı (Naive Bayes Classifier)

Naive Bayes sınıflandırıcısı, Bayes teoremine dayanan bir olasılık modeli olup çok terimli (multinominal) versiyonu kelime sayısı ya da TF-DIF vektörü gibi ayrık özelliklere sahip veri setlerinde kullanılmaktadır (Thanaki, 2018). Scikit-Learn kütüphanesinde Gaussian, Multinomial and Bernoulli olmak üzere 3 farklı çeşidi bulunan bu algoritmanın çalışma kapsamında Multinomial varyantı kullanılmıştır

1.5. Model Performans Ölçüleri

Modelleri birbirleri ile kıyaslayabilmek ve hedefe ulaşmada ne kadar başarılı olduklarını anlayabilmek için çeşitli metrikler kullanılmaktadır. Çalışmadaki sınıflandırma algoritmalarının başarılarını ölçme ve kıyaslamada bu bölümde açıklanan performans ölçüleri kullanılmıştır.

- Karışıklık Matrisi (Confusion Matrix)

Karışıklık matrisi birden fazla sınıfın olduğu çok sınıflı sınıflandırma problemlerinde, modelin sınıf bazında performansını göstermektedir. Matris her sınıfa ait doğru ve yanlış tahmin edilen kayıt sayısından oluşmaktadır (Thanaki, 2018). Gerçek sınıf ve tahmin edilen sınıf olmak üzere iki eksenden oluşmaktadır. Her eksen ise sınıf sayısı kadar sütun ve satırdan meydana gelmektedir (Şekil 4).

		Gerçek Değerler	
		Sıcaklık Arızası	**Basınç Arızası**
Tahmin Edilen Değerler	**Sıcaklık Arızası**	300	20
	Basınç Arızası	50	400

Şekil 4: Karışıklık Matrisi

- Gerçek Pozitif (True Positive - TP): Gerçek pozitif değeri doğru tahmin edilen şekildeki yeşil bölümleri ifade etmektedir. Örneğin Şekil 4'deki sıcaklık arızası 300 kez doğru olarak sıcaklık arızası sınıfında tahmin edilmiştir.
- Gerçek Negatif (True Negative - TN): Gerçek negatif değeri şekilde sıcaklık arızası dışındaki diğer arızaların (basınç arızası) doğru tahmin edilmesidir. Şekil 4'deki değeri 400 olmaktadır.
- Yanlış Pozitif (False Positive - FP): Yanlış pozitif değeri şekildeki örneğe göre sıcaklık arızasının basınç arızası olarak yanlış tahmin edilmesidir. Şekil 4'de sıcaklık arızasının basınç arızası olarak yanlış tahmin edilmesidir ve değeri 50 olmaktadır.
- Yanlış Negatif (False Negative - FN): Yanlış negatif değeri ise Şekil 4'e göre sıcaklık arızası olmadığı halde sıcaklık arızası olarak tahmin edilen değerlerdir. Sıcaklık arızası olmadığı model basınç arızasını sıcaklık arızası olarak tahmin etmiştir ve değeri 20 olmaktadır.

- Hassasiyet (Precision)

Hassasiyet, pozitif tahminlerin yüzdesi olarak tanımlanmaktadır ve literatürde pozitif tahmin değeri olarak da geçebilmektedir.

TP/(TP+FP) olarak hesaplanmaktadır. Şekil 4'deki örneğe göre sıcaklık arızasının tahmin hassasiyetinde 300/(300+50)=0.86 değeri elde edilmektedir.

- Hatırlama (Recall)

Hatırlatma tüm potansiyel pozitifler arasında (yanlış negatifler dahil) modelin gerçek pozitif örnekleri tespit etme yeteneği olarak tanımlanmaktadır.

TP/(TP+FN) olarak hesaplanmaktadır. Şekil 4'deki örneğe göre sıcaklık arızasının hatırlama değeri 300/(300+20)=0.94 olarak elde edilmektedir.

- F1 Puanı (F1-Score)

F1 puanı hassasiyet (precision) ve hatırlama (recall) değerlerinin harmonik ortalaması alınarak aşağıdaki formül ile hesaplanmaktadır (Bonaccorso, 2018).

$$\text{F1 Puanı} = 2 * \frac{hassasiyet * hatırlama}{hassasiyet + hatırlama}$$

Örnekteki sıcaklık arızasının değerleri kullanılarak F1 puanı hesaplandığında 2*(0.86*0.94)/(0.86+0.94)=0.9 değeri bulunmaktadır. Tüm sınıflar için hesaplandığında Tablo 1'deki değerler elde edilmektedir.

  - Makro Ortalama (Macro Average)

Modelin performansını sınıfların ortalama performans değerlerini alarak belirleyen yöntemdir. Tablo 1'deki örnek üzerinden modelin makro ortalama F1 puanı hesaplanmak istendiğinde (0.90+0.92)/2=0.91 değeri elde edilmektedir.

  - Ağırlıklı Ortalama (Weighted Average)

Modelin performansını sınıfların ağırlıklı ortalama performans değerlerini alarak belirleyen yöntemdir. Her sınıfın ağırlığı o sınıftaki toplam veri sayısıdır. Tablo 1'deki örnek üzerinden modelin ağırlıklı ortalama F1 puanı hesaplanmak istendiğinde ((0.90*350)+(0.92*420))/(350+420) = 0.91 değeri elde edilmektedir.

Tablo 1: Sınıfların Performans Değerleri

Sınıflar	Hassasiyet	Hatırlama	F1-Puanı
Sıcaklık Arızası	0.86	0.94	0.90
Basınç Arızası	0.95	0.89	0.92

BÖLÜM 2: MAKİNE ÖĞRENMESİ İLE METİN SINIFLANDIRMA ALANYAZIN TARAMASI

Metin sınıflandırma, metnin önceden tanımlı sınıflardan veya kategorilerden hangisine ya da hangilerine ait olduğunun tahmin edilmesidir (Tantuğ, 2016). Makine öğrenmesi kullanılarak gerçekleştirilen metin sınıflandırma çalışmaları incelendiğinde; sosyal medya paylaşımlarından veya alışveriş ve otel siteleri gibi sitelerdeki ürün/hizmet yorumlarından duygu analizi çalışmalarına sıkça rastlanmaktadır. Çalışmalardaki diğer bir konu haber metinlerin konularına göre sınıflandırılması gibi içerik sınıflandırma üzerinedir. Aynı zamanda siber zorbalık olarak tanımlanan saldırgan içerikli metinlerin tespiti ve istenmeyen e-postalar gibi metinlerin tespiti de metin sınıflandırma çalışmalarında görülmektedir.

Makine öğrenmesi sınıflama algoritmaları kullanarak metin sınıflandırma işlemi, bir web sayfasının e-ticaret sitesi olduğunun veya olmadığının tahmin edilmesi gibi tek etiketli (Kaşıkçı ve Gökçen, 2014) ya da Türkçe haberleri içeriklerine göre finans, kültür vb. kategoriler halinde çok etiketli (Çelik ve Koç, 2021; Uslu ve Akyol, 2021) olabilmektedir. Göçgün ve Onan (2021) çalışmalarında Amazon alışveriş sitesinden İngilizce ürün yorumlarını olumlu/olumsuz olarak sınıflandırırken, Tuzcu (2020) çalışmasında bir kitap satış sitesinden hazırladığı veri seti ile kitap yorumlarını olumlu/olumsuz şeklinde sınıflandırmıştır. Mengutayci ve Temurtaş (2021) ile Ahmetoğlu ve Resul (2020) çalışmalarında otel sitelerinden elde ettikleri Türkçe değerlendirmeler ile duygu analizi çalışırlarken, Aksu ve Karaman (2020) turistik mekan değerlendirmeleri üzerine duygu analizi çalışmışlardır. Yılmaz vd. Türkiye'de Twitter paylaşımlarında saldırgan dil kullanımının tespiti üzerine çalışma gerçekleştirmişlerdir (Yılmaz, Özer, ve Gökçen, 2021). Benzer şekilde Eryılmaz vd. Türkçe e-postaların spam olup olmadığının tespiti üzerine çalışmışlardır (Eryılmaz, Şahin ve Kılıç, 2020). Safalı (2020) farklı sosyal medya platformlarından topladığı Türkçe kişi paylaşımlarını siyasi ittifakı destekleyip desteklememelerine göre sınıflandırmıştır. Çılgın vd. paylaşılan İngilizce kripto para hakkındaki tweetleri önce duygularına göre sınıflandırmışlar sonra bu duygular ile kripto para değeri arasındaki ilişkiyi araştırmışlarıdır (Çılgın vd., 2020). Edwards vd. makine öğrenmesinde kümeleme ve sınıflandırma yönetimlerini bir arada kullanarak, baraj pompa istasyonunda 12 sene boyunca tutulan arıza kayıtlarından

yapılan bakımın planlı bakım mı plansız bakım mı olduğunun tahmini üzerine çalışma gerçekleştirmişlerdir (Edwards, Zatorsky ve Nayak, 2008).

2020 ve 2021 yıllarına ait literatürde Makine Öğrenmesi ile Metin Sınıflandırma alanında yapılan çalışmalardan bazıları Tablo 2'de özet olarak verilmiştir.

Tablo 2: Makine Öğrenmesi ile Metin Sınıflandırma Alanyazın Taraması

Yazar(lar)	Tarih	Araştırmanın Amacı	Öznitelik Çıkarım Yöntemi	Kullanılan Modeller
Çelik, Ö., & Koç, B. C.	2021	Türkçe haber metinlerinin içeriklerine göre sınıflandırılması	• Tfidfvectorizer • Word2Vec • FastText	• SVM • NB • LR • RF • ANN
Uslu, O., & Akyol, S.	2021	Türkçe haber metinlerinin içeriklerine göre sınıflandırılması	• Countvectorizer	• SVM • NB • RF
Göçgün, Ö. F., & Onan, A.	2021	İngilizce ürün değerlendirme metinleri ile duygu analizi	• Tfidfvectorizer	• SVM • BN • LR • RF
Yilmaz, Ş. Ş., Özer, İ., & Gökçen, H.	2021	Türkçe tweet paylaşım metinlerinde saldırgan dil tespiti	• Word2Vec	• RNN
Mengutayci, Ü. & Temurtaş, H.	2021	Türkçe otel değerlendirme metinleri ile duygu analizi	• Tfidfvectorizer	• ANN
Ahmetoğlu, H., & Resul, D. A. Ş.	2020	Türkçe otel değerlendirme metinleri ile duygu analizi	• Word2Vec	• RNN

Aksu, M. Ç., & Karaman, E.	2020	Türkçe turistik mekan değerlendirme metinleri ile duygu analizi	• Tfidfvectorizer • FastText	• SVM • NB • KNN
Tuzcu, S.	2020	Türkçe kitap değerlendirme metinleri ile duygu analizi	• Tfidfvectorizer	• SVM • NB • LR • ANN
Safalı, Y.	2020	Türkçe sosyal medya paylaşım metinlerinde siyasi ittifak desteği tespiti	• Tfidfvectorizer	• RF • NB • KNN • SMO
Çılgın, C., Ünal, C., Alıcı, S., Akkol, E., & Gökşen, Y.	2020	İngilizce kriptopara konulu tweet paylaşım metinlerinde duygu analizi	• Countvectorizer	• SVM • NB • LR • ANN
Eryılmaz, E. E., Şahin, D. Ö., & Kılıç, E.	2020	Türkçe metinler ile istenmeyen e-posta tespiti	• Tfidfvectorizer	• SVM • NB • LR • RF • DT • KNN • ANN

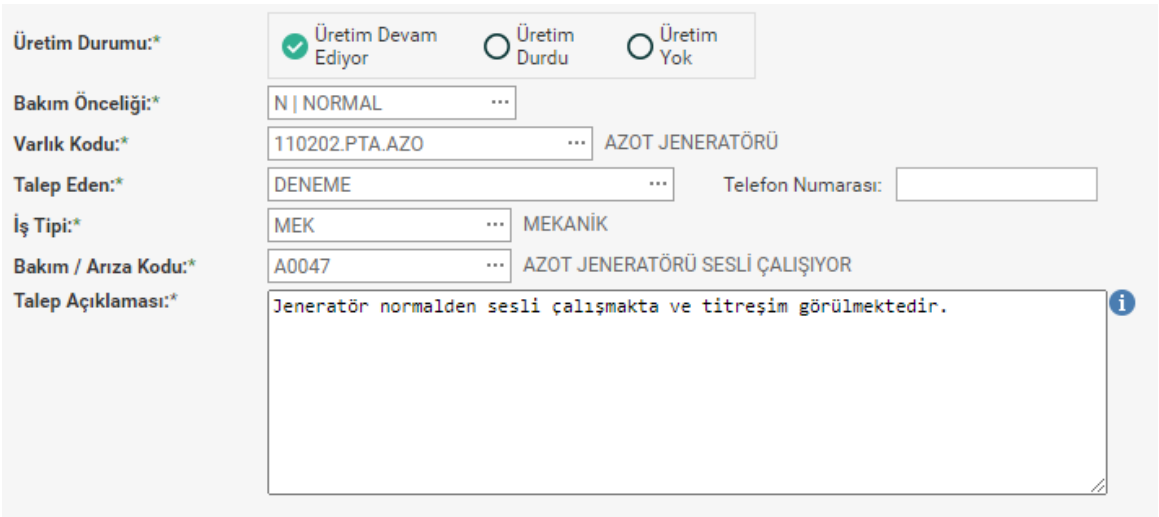
BÖLÜM 3: ARAŞTIRMANIN METODOLOJİSİ

İplik üretimi yapan firmada satın alınmış ticari bir program olan varlık ve bakım yönetim sistemi çalışmanın temel veri kaynağını oluşturmaktadır. Çalışma kapsamında sistemin firmaya sağladığı faydalar ile kullanılmasındaki zorluklar ve problemli süreçler incelendi.

Makine öğrenmesi türlerinden olan sınıflandırma yöntemi arıza açıklamalarının arıza tanımlarına atanmasında kullanıldı. Sınıflandırma yönteminin hedefi her arıza açıklaması için tek bir arıza tanımın tahmin edilmesi olarak belirlendi. Bu amaçla, Python yazılım dilinde 2 farklı öznitelik seçim yöntemi ve 4 farklı algoritma denenerek en başarılı sınıflandırma yöntemi arandı.

3.1. Bakım Yönetim Sisteminin İncelenmesi

Bakım yönetim sisteminden sağlanabilecek fayda sistemin etkin kullanımıyla doğru orantılıdır. Etkin kullanım verilerin sisteme zamanında, doğru şekilde ve eksiksiz işlenmesi ile sağlanabilmektedir. Sistemdeki veriler ile elde yapılan analizler sonucu elde edilen bilgilerin ve verilen kararların doğruluğunu sürecin en başındaki veri girişi aşaması belirlemektedir. Şekil 5'de sistemde arıza kaydının girildiği ekran verilmiştir.



Şekil 5: Bakım Yönetim Sistemi Arıza Giriş Ekranı

Çalışmada incelenen bakım yönetim sisteminde arıza kaydının oluşturulması 7 aşamada gerçekleşmektedir. İlk olarak üretim devam ediyor/durdu/yok şeklinde arıza sonrası üretimin durumu seçilmekte, arızanın önceliği normal/acil olarak seçilmekte sonrasında arızalanan varlık seçilmekte ve arızayı bildiren kişi kendi adını listeden seçmektedir.

Son olarak arıza tanımları listesinden ilgili tanım seçilmekte ve arızanın açıklaması yazılmaktadır.

Sistemde tanımlı 625 farklı arıza tanımı bulunmaktadır. İlgili arıza tanımının listeden bulunması vakit almakla beraber yanlış seçildiği durumlarda bulunmaktadır. Birbirine yakın anlamlı tanımlar karıştırılabilmekte veya dikkatsizlik nedeniyle yanlış tanım seçilebilmektedir. Bazı durumlarda ise tanım seçilip arıza açıklamasının es geçilmesi ya da üstünkörü açıklanması ile karşılaşılmaktadır. Bu problemler bakım yönetim sistemi ile yapılan analizlerin doğruluğunu azaltmakta ve yanlış arıza bildirimi gibi durumlarda bakım personelinin işini zorlaştırmaktadır.

İş Tipi Kodu	Bakım / Arıza Kodu	Bakım / Arıza Tanımı	Talep Açıklaması	Sarf Yeri Tanımı	Kısım Kodu	Varlık Kodu	Varlık Tanımı
MEK	A0047	AZOT JENERATÖRÜ SESLİ ÇALIŞIYOR	Jeneratör normalden sesli çalışmakta ve titreşim görülmektedir.	POLY 1	110202.PTA	110202.PTA.AZO	AZOT JENERATÖRÜ
MEK	A0377	PİSTON ARIZASI	ARIZANIN GIDERILMESI. BILGINIZE.	TEKSTÜRE 3.İŞLETME	110404.T3B	110404.T3B.BTB46.B.PNC17.201	BTB46 POZİSYON-201
MEK	A0101	CAĞLIK BORUSU TIKALI ARIZASI	DIRSEKLER YERINDEN CIKMIS MUDAHALE EDILMESI.	TEKSTÜRE 4.İŞLETME	110412.T4SNG	110412.T4SNG.BTE12.A.PNC01.012	BTE12 POZİSYON-12
ELO	A0482	T2 FIRIN ISISI ARIZASI	7.PEN T2 ISILARI SUREKLI TOLARANS DIŞI ALARMI VERIP ON/OFF OLMAKTADIR.170 C DEN 200 C DERECEYE CIKTIGI GORULDU.	TEKSTÜRE 3.İŞLETME	110404.T3A	110404.T3A.BTA27.A.PNC07	BTA27 PENCERE-7
MEK	A0015	AGREGAT ARIZASI	52. POZISYON AGREGAT ALT MOTORU CALISMIYOR.	TEKSTÜRE 4.İŞLETME	110412.T4MLT	110412.T4MLT.BTEM4.A.PNC05.052	BTEM4 POZİSYON-52

Şekil 6: Bakım Yönetim Sistemi Arıza Kayıtları Listesi

Oluşturulan arıza kayıtları bakım personeli tarafından liste halinde görülebilmektedir (Şekil 6). Arıza tanımlarına ve açıklamalarına göre aciliyet durumu belirlenmekte ve bakım personeli belirlediği sıraya göre, gerekli ekipmanlar ile ilgili varlığa müdahale için harekete geçmektedir.

Problemler göz önüne alındığında yalnızca arızanın yalın bir açıklaması ile arıza tanımın tahmin edilmesi hem bakım personelinin dikkatlice yazılmış açıklamadan arızayı daha net anlayıp bakım başarısını yükselteceği hem de sistem üzerinden yapılan analizlerde daha yüksek doğrulukla sonuçlara ulaşılacağı öngörülmektedir.

3.2. Sistemden Veri Toplanması

Çalışmanın ilk aşamasında veri seti oluşturulması amacıyla bakım yönetim sisteminden arıza kayıtları Excel formatında dışarı aktarıldı. Sistemden kayıtlar çekilirken 3 ayrı filtre uygulandı.

I. Tarih filtresi 2020 yılı olarak filtrelendi.

II. Veri setini iptal edilmiş ve henüz sonuca ulaştırılmamış kayıtlardan arındırmak için arıza kayıt durumları (açık, kapalı veya iptal olabilmektedir) kapalı olan kayıtlar filtrelendi.

III. Arıza nedeni "arızaya rastlanmadı" olarak belirtilen kayıtlar, hatalı ya da eksik açılmış olma ihtimaline karşı veri setinin dışında bırakıldı.

Filtrelemeler sonrası dışarı aktarılan veri seti hakkında özet bilgiler Tablo 3'de verilmektedir.

Tablo 3: 2020 Yılı Arıza Kayıt Sayıları

Toplam Kayıt Sayısı	101.715
Aylık Ortalama Kayıt Sayısı	8.476,25
Günlük Ortalama Kayıt Sayısı	278,67

Veri setinde, çalışma kapsamında kullanacağımız türü metin olan arıza açıklaması sütünü ve türü kategorik olan tahmin etmeye çalışacağımız arıza tanımı sütunu dışındaki tüm sütunlar silinmiştir. İşlem öncesi veri setinin örnek ilk 5 satırı Tablo 4'de işlem sonrası örnek ilk 5 satırı Tablo 5'de verilmektedir.

Tablo 4: Veri Setinin Örnek 5 Sütunu

…	**Arıza Tanımı**	**Arıza Nedeni Tanımı**	**Arıza Çözüm Tanımı**	**Toplam Maliyet (Döviz)**	**Arıza Açıklaması**	**Yapılan İşin Açıklaması**	**Talep Eden**	**Sarf Yeri Tanımı**	…
…	ENERJİ KESİNTİSİ ARIZASI	İÇ KAYNAKLI ENERJİ KESİNTİSİ	KONTROL EDİLDİ	0,137462642	ARIZALI OLAN TERAZİNİN AKÜSÜNÜN ONARILMASI	ARIZA SERVİSLİK	ŞAHİN MERCAN	BÜKÜM	…
…	SPINING TEX ARIZASI	DTEX PROBLEMİ	CİHAZ DEĞİŞTİRİLDİ	7,172120836	20 mak 12 kanalın dtx leri döşük geliyor eriyik pompasının değiştirilmesi	DETEXSSEB İLE ERİYİK POMPASI DEGİŞİM YAPILDI	HACI MAHMUT KARAHAN	POY 2	…
…	FIRIN SOPASI ARIZASI	REGÜLATÖR PROBLEMİ	PERİYODİK BAKIM YAPILDI	0,063522156	18 P.Z FIRIN SOPASI KIRIK MUDALE EDILMESI BILGINIZE	PERİYODİK BAKIM	ÜZEYİR SAĞRI	TEKSTÜRE 4.İŞLETME	…

…	BAKALİT SARIĞI ARIZASI	REGÜLATÖR PROBLEMİ	PERİYODİK BAKIM YAPILDI	0,063522156	230.pozisyonda bakalitte sarik olmasi nedeniyle kapatildi	PERİYODİK BAKIM	SAKİN ERGÜN	TEKSTÜRE 4.İŞLETME	…
…	BAKALİT SARIĞI ARIZASI	REGÜLATÖR PROBLEMİ	PERİYODİK BAKIM YAPILDI	0,063522156	BAKALITTE SARIK VAR POZISYON KAPATILDI	PERİYODİK BAKIM	UFUK YILMAZ	TEKSTÜRE 4.İŞLETME	…

Tablo 5: Veri Setinin Metin ve Etiket Sütunları

Arıza Açıklaması	Arıza Tanımı
ARIZALI OLAN TERAZİNİN AKÜSÜNÜN ONARILMASI	ENERJİ KESİNTİSİ ARIZASI
20 mak 12 kanalın dtx leri döşük geliyor eriyik pompasının değiştirilmesi	SPINING TEX ARIZASI
18 P.Z FIRIN SOPASI KIRIK MUDALE EDILMESI BILGINIZE	FIRIN SOPASI ARIZASI
230.pozisyonda bakalitte sarik olmasi nedeniyle kapatildi	BAKALİT SARIĞI ARIZASI
BAKALITTE SARIK VAR POZISYON KAPATILDI	BAKALİT SARIĞI ARIZASI

Veri setindeki arıza tanımları incelendiğinde; sistemde kayıtlı 625 farklı arıza tanımından yıl içinde 460 tanım için arıza kaydı oluşturulduğu görüldü. Mevcut yapıda metin olarak saklanan arıza tanımları; veri setinde etiket sütunu oluşturulması amacıyla, en sık tekrarlayan arıza tanımına 1 rakamı atanarak en az tekrarlayana doğru 460 rakamına kadar numaralandırıldı. Tablo 6'da son 4 arıza tanımı, sayıları ve atanan arıza numaraları verilmektedir.

Tablo 6: Veri Setinin Metin, Etiket ve Arıza Numarası Sütunları

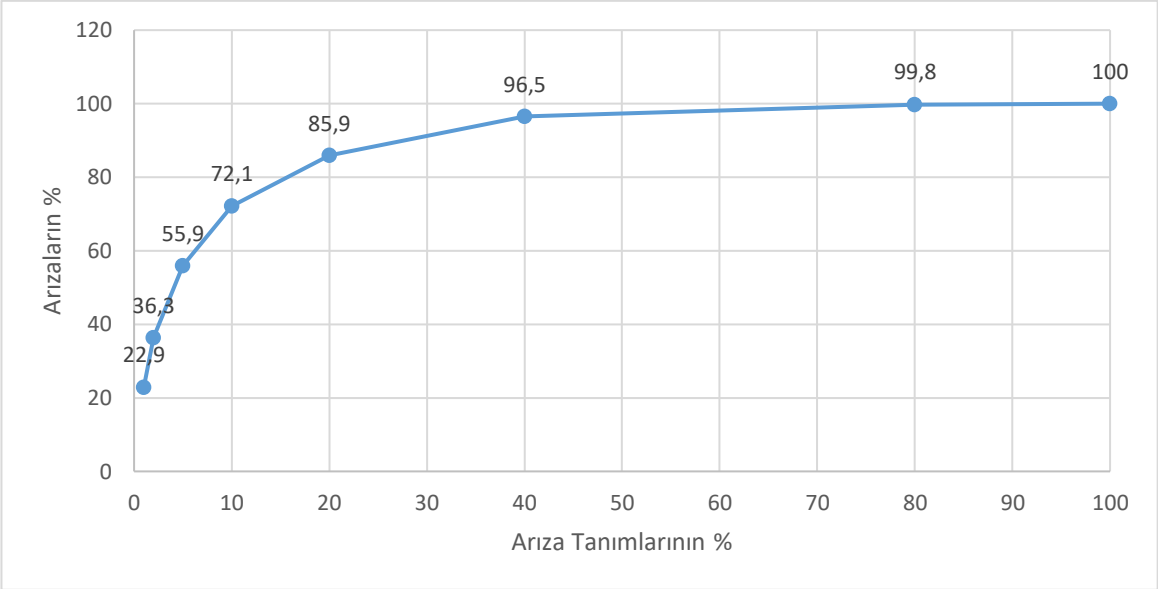
Arıza Tanımı	Arıza Sayısı	Arıza No
UPS ARIZASI	1	457
YAĞ MANDALI KİLİTLİ POZİSYON	1	458
YAĞLAYICI KLAVUZ GERİLİM ARIZASI	1	459
YAZILIM ARIZASI	1	460

Her arıza tanımı için kaç adet kayıt oluşturulduğu büyükten küçüğe sıralanarak Tablo 7'de bu sıralamanın ilk 23 (tüm tanımların %5'i) kalemine yer verilmektedir. Ayrıca arıza tanımlarının toplam içerisindeki payları yüzdelik ve kümülatif yüzdelik olarak verilmektedir. Örneğin 1 numaralı arıza tanımına ait 9001 kayıt bulunmakta olup, toplam arıza kayıtlarının %8,8'ine denk gelmektedir.

Tablo 7: Büyükten Küçüğe Sıralı Arıza Tanım Sayıları

Arıza No	Arıza Sayısı	Toplam İçerisindeki %	Kümülatif %
1	9001	8,8%	8,8%
2	3832	3,8%	12,6%
3	3706	3,6%	16,3%
4	3421	3,4%	19,6%
5	3313	3,3%	22,9%
6	3133	3,1%	26,0%
7	2738	2,7%	28,7%
8	2664	2,6%	31,3%
9	2600	2,6%	33,8%
10	2493	2,5%	36,3%
11	2296	2,3%	38,5%
12	1953	1,9%	40,5%
13	1935	1,9%	42,4%
14	1821	1,8%	44,1%
15	1796	1,8%	45,9%
16	1608	1,6%	47,5%
17	1572	1,5%	49,0%
18	1397	1,4%	50,4%
19	1174	1,2%	51,6%
20	1145	1,1%	52,7%
21	1136	1,1%	53,8%
22	1080	1,1%	54,9%
23	1023	1,0%	55,9%

En çok tekrar eden hataların toplamdaki payları Şekil 7'deki grafik üzerinde incelendiğinde logaritmik artış görülmektedir.



Şekil 7: En Çok Tekrar Eden Arıza Tanımlarının Toplamdaki Payları

Örneğin, en çok tekrar eden %20'lik arıza tanımları grubu (en çok tekrar eden 92 arıza tanımı) tüm arızaların %86'sını oluşturmaktadır.

3.3. Veri Ön İşleme

Veri setinin ön işlemesi ve temizleme süreçleri 5 adımda gerçekleştirilmiştir.

I. Arıza açıklamaları metinlerinin bulunduğu sütun text (metin) olarak, arıza tanımlarının bulunduğu sütun label (etiket) olarak isimlendirildi (Tablo 8).

Tablo 8: Arıza Kayıtları Örneği

Metin	Etiket
ARIZALI OLAN TERAZİNİN AKÜSÜNÜN ONARILMASI	295
20 mak 12 kanalın dtx leri döşük geliyor eriyik pompasının değiştirilmesi	452
18 P.Z FIRIN SOPASI KIRIK MUDALE EDILMESI BILGINIZE	16
230.pozisyonda bakalitte sarik olmasi nedeniyle kapatildi	21

II. Arıza açıklamalarını, arıza tanımlarının tahminini etkilemeyecek değişikliklerden arındırmak ve öznitelik çıkarımı aşamasında sözlükte tekrarlı ve gereksiz kelimelerin oluşmasının önüne geçmek amacı ile öncellikle;

a. Açıklamalardaki rakamlar kaldırıldı.
b. Rakamlardan sonra noktalama işaretleri kaldırıldı.
c. Eğer varsa yazdırılamayacak karakterler kaldırıldı.
d. Birden fazla boşluklar kaldırıldı.
e. Tüm karakterler küçük harfe çevrildi.

Temizlenmiş arıza açıklama metinleri Tablo 9'da metin sütununda verilmektedir.

Tablo 9: Temizlenmiş Arıza Kayıtları Örneği

Metin	Etiket
arızalı olan terazinin aküsünün onarılması	295
mak kanalın dtx leri döşük geliyor eriyik pompasının değiştirilmesi	452
pz fırın sopası kırık mudale edılmesı bılgınıze	16
pozisyonda bakalitte sarik olmasi nedeniyle kapatildi	21

III. Arıza açıklamalarında tek kelimeden oluşan, özensiz şekilde yazılmış bu nedenle arızanın anlaşılamadığı açıklamaları veri seti dışında bırakmak ve II. Adımdaki ön işlemlerden sonra hiç karakter kalmayan ya da anlam çıkarılamayacak kadar az karakter kalan metinleri elemek için 10 karakterden az olan açıklamalar kaldırıldı.

Tablo 10'da 10 karakterden az metinden oluşan kayıtların örnekleri karakter sayıları ile birlikte verilmiş olup, üçüncü satırda hiç karakter bulunmayan bir kaydın örneği bulunmaktadır.

Tablo 10: Az Sayıda Karakter İçeren Arıza Kayıtları Örneği

Metin	Etiket	Karakter Sayısı
ayarsız	9	7
tıkalı	33	6
	50	0
arızalı	194	7
kırıktır	19	8

IV. Aynı arıza açıklaması birden fazla arıza tanımında bulunmaması gerektiğinden çakışan arıza açıklamalarının en çok kullanıldığı arıza tanımına ait olduğu varsayıldı ve daha az sayıda olan tanımlardan kaldırıldı. Eşit sayıda farklı arıza tanımında olması durumunda ise tüm tanımlardan kaldırıldı.

Tablo 11: Farklı Kategorilerde Az Sayıda Tekrar Eden Arıza Kayıtları Örneği

Metin	Etiket	Aynı Etikette Tekrar Sayısı	Tüm Etiketlerde Tekrar Sayısı	Aynı Etikette Maksimum Tekrar Sayısı
fırın sopası arızalı	263	2	227	223
apron arızalı	75	4	96	92
kopma tekrarı yapıyor	116	3	275	238
klavuz yok	128	9	111	76

Örneğin, Tablo 11'de görüldüğü üzere "fırın sopası arızalı" metni arıza kayıtların toplamda 227 kez geçmiş ancak 263 numaralı arıza tanımında sadece 2 kez tekrar etmiştir. Farklı bir arıza tanımında 223 kez tekrar ettiği görülebildiği için asıl ait olması gereken tanımın 263 numaralı tanım olmadığı anlaşılmaktadır. Bu nedenle veri setinden kaldırılmıştır.

V. Aynı arıza açıklamasının aynı arıza tanımında birden fazla tekrarlandığı durumlar kaldırıldı.

Tablo 12: Aynı Kategoride Tekrar Eden Arıza Kayıtları Örneği

Metin	Etiket	Aynı Etikette Tekrar Sayısı	Tüm Etiketlerde Tekrar Sayısı	Aynı Etikette Maksimum Tekrar Sayısı
img düze değişimi w wx apel arabalarının ayarlanması	3	184	185	184
img düze değişimi w wx apel arabalarının ayarlanması	3	184	185	184
img düze değişimi w wx apel arabalarının ayarlanması	3	184	185	184
img düze değişimi w wx apel arabalarının ayarlanması	3	184	185	184

Örneğin, Tablo 12'de görüldüğü üzere "img düze değişimi w wx apel arabalarının ayarlanması" metni veri setinde toplamda 185 kez tekrar ederken 184 kez 3 numaralı arıza tanımda tekrar etmiştir. Geri kalan farklı kategorideki 1 adetlik kayıt bir önceki aşamada kaldırılmış olup bu aşamada ise 184 tekrardan arındırılıp tek bir kayıt haline dönüştürüldü.

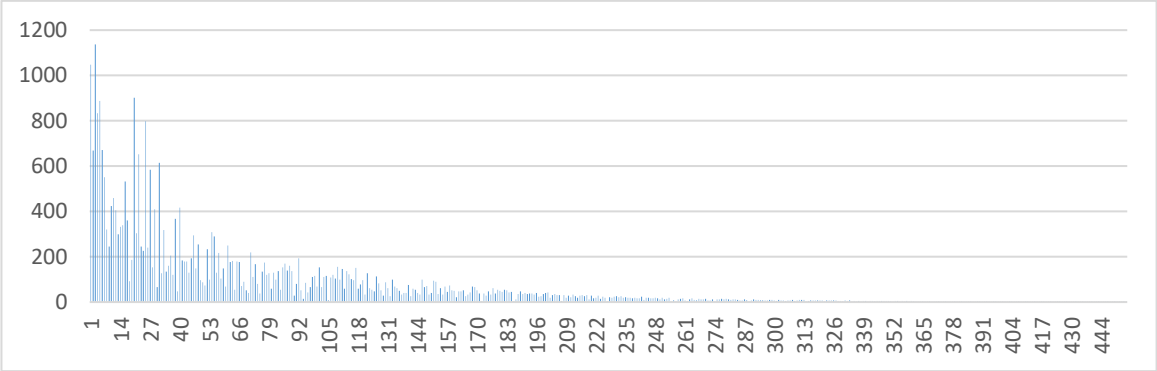
Ön işleme adımlarından sonraki kayıt sayısındaki azalmalar Tablo 13'te verildi. En ciddi azalma tekrarlı kayıtların çıkarılmasından sonra meydana gelmiştir.

Tablo 13: Veri Ön İşleme Adımlarından Sonra Kayıt Sayıları

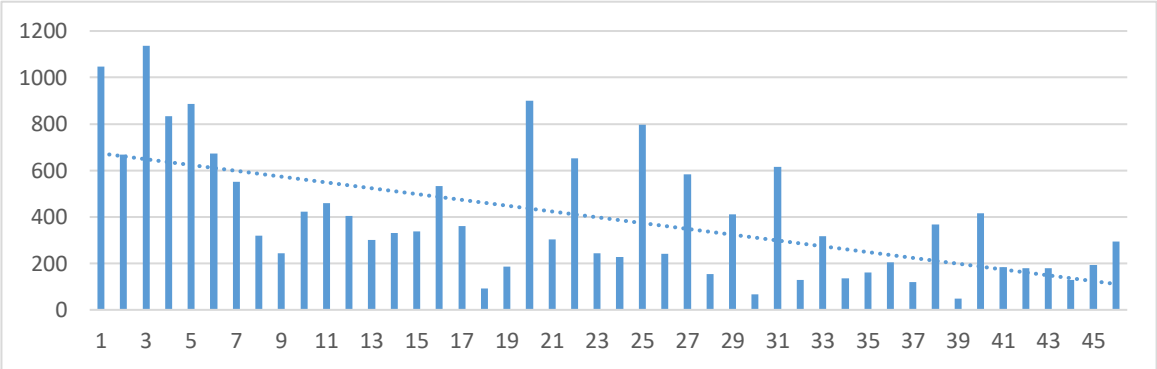
İşlem	Kayıt Sayısı
Veri Ön işleme öncesi toplam kayıt	101.715
Metin temizleme işlemleri ile karakter sınırı sonrası kalan toplam kayıt (II-III)	95.473
Tekrarlı açıklamaların kaldırılması sonrası kalan toplam kayıt (IV-V)	32.889

Gerçekleştirilen ön işleme adımları sonrasında Tablo 7'de büyükten küçüğe sıralanan kayıt sayıları Şekil 8'de görülüğü gibi farklı oranlarda azalmışlardır. Ön işleme sonrası arıza tanımlarına göre arıza kayıt sayıları incelendiğinde (Şekil 8) tüm arıza tanım kategorilerini tahmin etmek için yeterli sayıda kayıt olmadığı görülmektedir.

Kayıt sayısı arttıkça başarı oranı yükseleceğinden çalışma kapsamında en fazla arızanın kapsanması ve en az hata oranı ile modelin çalışması hedefi üzerine arızaların %72,12ine denk gelen (Şekil 7) %10'luk yani 460 hata tanımından en çok kayıt bulunan ilk 46 hata tanımı (Şekil 9) için tahmin modeli oluşturulması planlandı. Bu doğrultuda ön işlenmesi ve temizlemesi gerçekleştirilmiş veri setinde en çok tekrar eden ilk 46 hata tanımı ayrılarak 18037 adet arıza kaydına sahip yeni bir veri seti oluşturuldu.



Şekil 8: Ön İşleme Sonrası Arıza Tanımlarına Ait Kayıt Sayıları



Şekil 9: Ön İşleme Sonrası İlk 46 Arıza Tanımına Ait Kayıt Sayıları

Tahmin edilecek 46 arıza tanımı ile toplamda 18037 adet arıza kaydına sahip olan temiz nihai veri setine ait arıza tanımları ve kayıt sayıları grafiği Şekil 9'da verilmekte ve arıza tanımları ait kayıt sayıları dağılımı görülebilmektedir. Veri setinde ön işleme adımlarından önce en fazla kayda sahip olan arıza tanımı 1 numaralı tanım iken nihai veri seti grafiğinde görüldüğü gibi 1136 adet kayıt ile 3 numaraları arıza tanımı en fazla kayda sahip olan tanımdır. En az kayda sahip olan tanım ise 48 adet kayıt ile 39 numaralı arıza tanımıdır.

3.4. Veri Setini İçeri Aktarma ve Özetleme

Ön işleme adımlarından geçirilen veri seti, Python yazılım dilinde işlenebilmesi için sıkça tercih edilen csv formatına çevrildi. Toplamda 18037 arıza kaydından ve 46 farklı arıza kategorisinde(tanımına) sahip olan veri seti *pandas* kütüphanesi kullanılarak içeri aktarıldı.

Aktarılan veri seti hakkında özet bilgilerin görüntülenebilmesi amacı ile *pandas_profiling* kütüphanesi ile rapor hazırlandı. Şekil 10'daki raporda veri setindeki sütün ve satır sayısı, boş hücre ve tekrarlı satır sayısı ile değişken tipleri görülmektedir.

Veri Seti İstatistikleri	
Değişken Sayısı	2
Kayıt Sayısı	18037
Boş Hücre	0
Boş Hücre (%)	0%
Tekrarlı Satır	0
Tekrarlı Satır (%)	0%

Değişken Tipleri	
Kategorik	1
Nümerik	1

Şekil 10: Veri Seti Özet Bilgileri

Veri setinin 18037 kayıttan ve 2 değişkenden oluştuğu, arıza tanımları rakamsal olarak etiket sütununda ifade edildiği için sistem tarafından nümerik olarak algılandığı ve arıza açıklamaları her biri birinden farklı metin olduğu için kategorik alındığı görülmektedir. Ayrıca veri setinde boş hücre ya da tekrarlı satır olmadığı bilgisine ulaşılabilmektedir.

Farklı	18037
Farklı (%)	100%
Eksik	0
Eksik (%)	0%

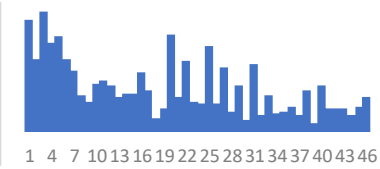
hav tekrarı devam ed...	1
teksture fırın sopası k..	1
tabyatağı kırık	1
Diğer değerler (18034)	18034

Şekil 11: Arıza Açıklamaları Sütunu Özet Bilgileri

Veri setinde text (metin) başlıklı arıza açıklamalarının bulunduğu sütunun 18037 satırdan oluştuğu farklılık oranın %100 olduğu yani tekrar eden açıklamanın bulunmadığı ve eksik açıklamanın yer almadığı Şekil 11'de görülmektedir.

Farklı	46
Farklı (%)	30%
Eksik	0
Eksik (%)	0%
Ortalama	17,89

Minimum	1
Maksimum	46



Şekil 12: Arıza Tanımları Sütunu Özet Bilgileri

Veri setinde etiket başlıklı arıza tanımlarının bulunduğu sütunun 46 farklı kategoriden oluştuğu görülmekte (Şekil 12) ve kategoriler arası dağılım Şekil 10'da ve burada grafik olarak verilmektedir.

	text	label
0	pz fırın sopası kırık mudale edılmesı bılgınıze	16
1	pozisyonda bakalitte sarik olmasi nedeniyle kapatildi	21
2	bakalıtte sarık var pozısyon kapatıldı	21
3	pozıısyonlarda bakalıtte sarık olması nedenıyle kapatıldı	21
4	bakalit sarik	21
5	fskolu kırık	16
6	fikse fırın tıkalı	37
7	nip tiwist arızalı	13
8	pozisyon havlı devam etmektedir	5
9	ve pencere niptwist pzaktarma kayış kopmuş	34

Şekil 13: Kodlamaya Aktarılmış Veri Setinin İlk 10 Satırı

Aktarılmış olan veri setinin ilk 10 satırı metin (text) ve etiket (label) Şekil 13'te görülmektedir.

3.5. Öznitelik Çıkarımı

Değişken tipi metin olan arıza açıklamalarının makine öğrenmesi algoritmalarında kullanılabilmesi için kelime çantası (BoW) ve Terim Frekansı – Ters Doküman Frekansı (TF-IDF) yöntemleri ile önce kelime sözlüğü oluşturuldu daha sonra vektörize edildi. Bu amaçla *sklearn* kütüphanesinin *feature.extraction* modülü kullanıldı.

- Kelime Torbası (BoW) yöntemi ile arıza açıklamalarından (metin sütunundan) kelime sözlüğü oluşturuldu. *Min_df=2* parametresi ile veri setinde en az 2 kez geçen kelimeler sözlüğe eklendi. Oluşan sözlük incelendiğinde (Şekil 14) 4542 kelimenin sözlüğe dahil olduğu görülmektedir.

'Bow-TF :'
(18037, 4542)

	ab	abböl	ablukatör	abol	aböl	abölsaat	acil	acilen	acilll	acilllll	...	şanjuru	şanjurun	şanjür	şanzuman	şatıl	şatıldan	şekilde	şeklinde	şimdi	şiş
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 4542 columns

Şekil 14: Kelime Torbası Yöntemi Sözlüğü

Oluşan sözlük kullanılarak arıza tanımları (etiketleri) ile birlikte vektör tablosu hazırlandı (Şekil 15). Etiket (label) başlıklı ilk sütun arıza tanımlarını ifade ederken devamındaki sütunlar sözlükteki kelimeler yani modelin özniteliklerdir. Her satır bir arıza kaydını ifade etmekte ve ilgili kayıtta bulunan kelimeler 1 ile gösterilirken bulunmayan kelimeler 0 ile gösterilmektedir. Şekil 15'te veri setinin ilk 5 ve son 5 satırı ile etiket sütunu ve öznitelik sütunları görülmektedir.

(18037, 4543)

	label	ab	abböl	ablukatör	abol	aböl	abölsaat	acil	acilen	acilll	...	şanjuru	şanjurun	şanjür	şanzuman	şatıl	şatıldan	şekilde	şeklinde	şimdi	şiş
0	16	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	21	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	21	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	21	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	21	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 4543 columns

	label	ab	abböl	ablukatör	abol	aböl	abölsaat	acil	acilen	acilll	...	şanjuru	şanjurun	şanjür	şanzuman	şatıl	şatıldan	şekilde	şeklinde	şimdi	şiş
18032	30	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
18033	23	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
18034	26	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
18035	21	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
18036	25	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 4543 columns

Şekil 15: Kelime Torbası Yöntemi Tablosu

- Terim Frekansı – Ters Doküman Frekansı (TF-IDF) yönteminde, kelime torbası yöntemi ile aynı şekilde kelime sözlüğü oluşturuldu ve sözlükte aynı sayıda kelime elde edildi. Kelime torbası yönteminden farklı olarak ilgili arıza kaydında geçen/geçmeyen kelimelere 0/1 atamak yerine bu yöntemde her kelimeye tf*idf skorları hesaplanıp yazılmıştır (Şekil 16).

'Bow-TF:IDF :'
(18037, 4542)

	ab	abböl	ablukatör	abol	aböl	abölsaat	acil	acilen	acilll	acillll	...	şanjuru	şanjurun	şanjür	şanzuman	şatıl	şatıldan	şekilde	şeklinde	şimdi	şiş
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 4542 columns

Şekil 16: Terim Frekansı – Ters Doküman Frekansı Yöntemi Sözlüğü

Oluşan sözlük kullanılarak arıza tanımları (etiketleri) ile birlikte vektör tablosu hazırlandı. Şekil 17'de veri setinin ilk 5 ve son 5 satırı ile etiket sütunu ve öznitelik sütunları görülmektedir.

(18037, 4543)

	label	ab	abböl	ablukatör	abol	aböl	abölsaat	acil	acilen	acilll	...	şanjuru	şanjurun	şanjür	şanzuman	şatıl	şatıldan	şekilde	şeklinde	şimdi	şiş
0	16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 4543 columns

	label	ab	abböl	ablukatör	abol	aböl	abölsaat	acil	acilen	acilll	...	şanjuru	şanjurun	şanjür	şanzuman	şatıl	şatıldan	şekilde	şeklinde	şimdi	şiş
18032	30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18033	23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18034	26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18035	21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18036	25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 4543 columns

Şekil 17: Terim Frekansı – Ters Doküman Frekansı Yöntemi Tablosu

3.6. Model Eğitimi ve Performansı

İki farklı öznitelik çıkarma yöntemi ile oluşturulmuş; 18037 kayıta, 4542 özniteliğe ve tahmin edilmesi amaçlanan 46 farklı arıza kategorisine sahip veri setleri makine öğrenmesi modelini eğitebilmek ve sonrasında test edebilmek için eğitim ve test veri setlerine ayrıldı.

İlk olarak %70 eğitim ve %30 test daha sonra %80 eğitim ve %20 test son olarak %90 eğitim ve %10 test seti şeklinde rastgele ayrılan veri setleri 4 farklı algoritma ile eğitildi ve test edildi. Modellerin sonuçları Tablo 14, Tablo 15 ve Tablo 16'da verilmektedir.

Tablo 14: Model Skorlarının Karşılaştırılması %70 Eğitim – %30 Test

Model	**Öznitelik →**	Kelime Torbası			Kelime Torbası: TF-IDF		
	Skorlama	hassasiyet	hatırlama	f1-puanı	hassasiyet	hatırlama	f1-puanı
Logistic Regression	makro ort.	0,85	0,82	0,83	0,86	0,80	0,82
	ağırlıklı ort.	0,85	0,84	0,84	0,84	0,83	0,84
Linear SVC	makro ort.	0,84	0,83	0,83	0,85	0,84	0,84
	ağırlıklı ort.	0,85	0,84	0,84	0,86	0,85	**0,85**
SGDC	makro ort.	0,85	0,83	0,84	0,85	0,83	0,84
	ağırlıklı ort.	0,85	0,85	0,85	0,85	0,85	0,85
Naive Bayes Classifier	makro ort.	0,81	0,77	0,78	0,82	0,78	0,79
	ağırlıklı ort.	0,81	0,80	0,80	0,82	0,81	0,80

Tablo 15: Model Skorlarının Karşılaştırılması %80 Eğitim – %20 Test

Model	**Öznitelik →**	Kelime Torbası			Kelime Torbası: TF-IDF		
	Skorlama	hassasiyet	hatırlama	f1-puanı	hassasiyet	hatırlama	f1-puanı
Logistic Regression	makro ort.	0,84	0,82	0,82	0,85	0,81	0,82
	ağırlıklı ort.	0,84	0,84	0,84	0,85	0,84	0,84
Linear SVC	makro ort.	0,83	0,82	0,82	0,85	0,84	0,84
	ağırlıklı ort.	0,84	0,84	0,84	0,86	0,86	**0,86**
SGDC	makro ort.	0,84	0,83	0,83	0,84	0,81	0,83
	ağırlıklı ort.	0,85	0,85	0,85	0,85	0,84	0,84
Naive Bayes Classifier	makro ort.	0,84	0,73	0,76	0,81	0,61	0,64
	ağırlıklı ort.	0,81	0,80	0,79	0,79	0,74	0,72

Tablo 16: Model Skorlarının Karşılaştırılması %90 Eğitim – %10 Test

Model	**Öznitelik →**	Kelime Torbası			Kelime Torbası: TF-IDF		
	Skorlama	hassasiyet	hatırlama	f1-puanı	hassasiyet	hatırlama	f1-puanı
Logistic Regression	makro ort.	0,85	0,83	0,84	0,87	0,83	0,84
	ağırlıklı ort.	0,86	0,85	0,85	0,86	0,85	0,85
Linear SVC	makro ort.	0,84	0,84	0,84	0,86	0,86	0,85
	ağırlıklı ort.	0,86	0,85	0,85	0,87	0,87	**0,87**
SGDC	makro ort.	0,84	0,84	0,83	0,84	0,84	0,83
	ağırlıklı ort.	0,86	0,85	0,85	0,86	0,85	0,85
Naive Bayes Classifier	makro ort.	0,84	0,76	0,78	0,84	0,63	0,67
	ağırlıklı ort.	0,83	0,81	0,81	0,80	0,75	0,74

Veri setindeki sınıf sayılarında farklılıkların olması nedeniyle makro ortalama sonuçları yerine sınıf büyüklüklerini dikkate alan ağırlıklı ortalama sonuçları üzerinden karşılaştırma yapılması daha faydalı olmaktadır. Modellerin ağırlıklı ortalama F1 puanları incelendiğinde çalışmada en başarılı sonucu veren modelin %87 F1 puanı ile Lineer SVC algoritması çalıştırılan olduğu görülmektedir. Eğitim testi yüzdesi arttırıldığında ise modelin daha geniş veri setine ulaşabilmesi sayesinde başarı oranının arttığı anlaşılmaktadır.

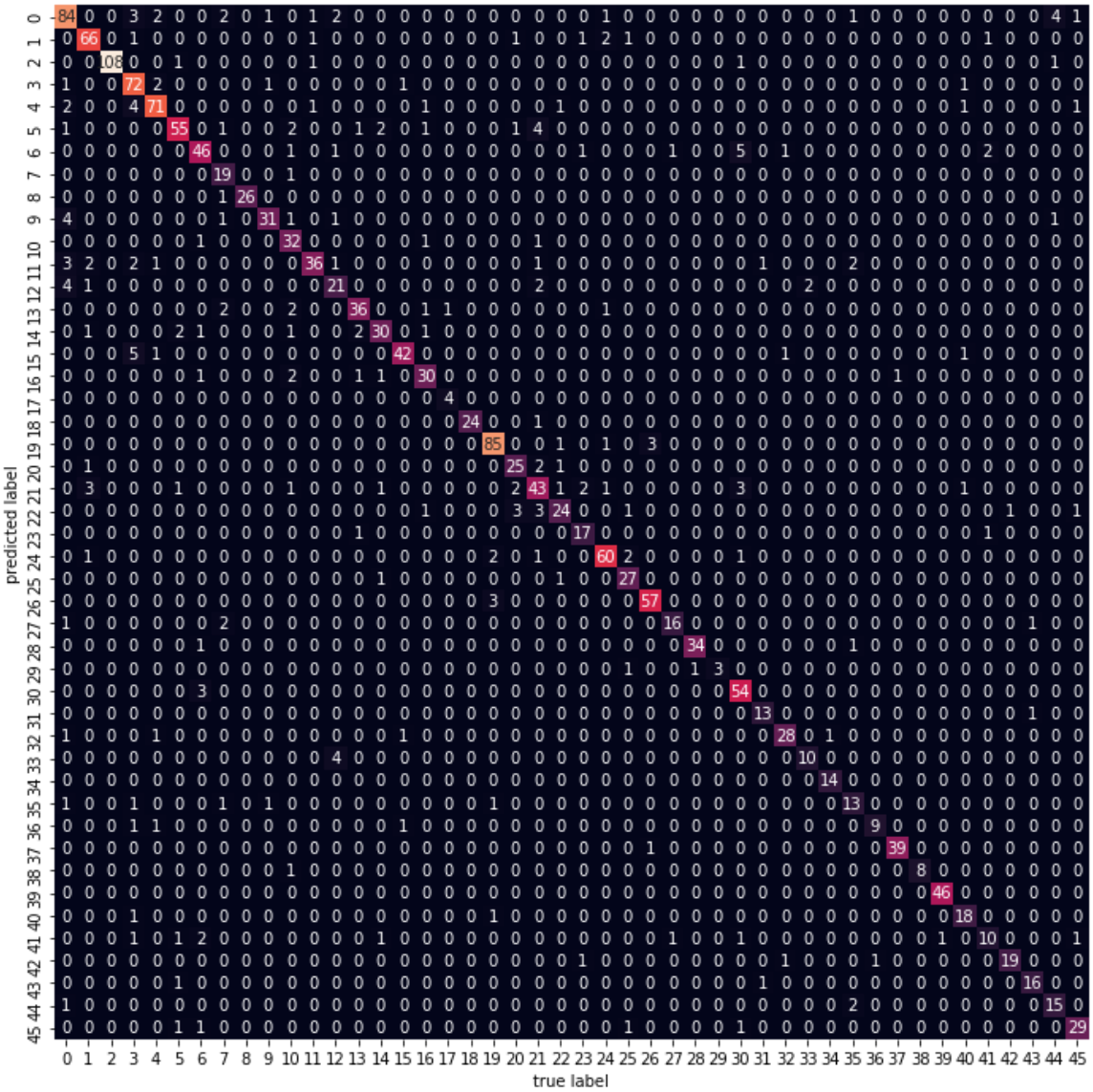
Kelime Torbası ve Terim Frekansı – Ters Doküman Frekansı sözlük oluşturma yöntemlerinin farkına bakıldığında ise TF-IDF yönteminin açıklamalarda geçen kelimeleri sadece sayı olarak belirtmek yerine diğer açıklamalarda geçme oranına göre değer ataması yapması model başarısını arttırmaktadır. Şekil 18'de solda BoW ve sağda TF-IDF metotlarının arıza tanım sınıfları bazında model başarısı görülmektedir.

	precision	recall	f1-score	support
1	0.81	0.81	0.81	102
2	0.88	0.88	0.88	74
3	0.99	0.95	0.97	112
4	0.80	0.87	0.83	78
5	0.86	0.88	0.87	82
6	0.89	0.81	0.85	68
7	0.84	0.79	0.81	58
8	0.70	0.95	0.81	20
9	1.00	0.96	0.98	27
10	0.90	0.72	0.80	39
11	0.78	0.91	0.84	35
12	0.85	0.71	0.78	49
13	0.59	0.73	0.66	30
14	0.90	0.84	0.87	43
15	0.81	0.76	0.78	38
16	0.89	0.84	0.87	50
17	0.76	0.81	0.78	36
18	0.67	1.00	0.80	4
19	0.96	0.92	0.94	25
20	0.88	0.91	0.90	90
21	0.84	0.90	0.87	29
22	0.71	0.69	0.70	59
23	0.79	0.76	0.78	34
24	0.68	0.79	0.73	19
25	0.90	0.90	0.90	67
26	0.79	0.93	0.86	29
27	0.92	0.92	0.92	60
28	0.94	0.80	0.86	20
29	0.97	0.94	0.96	36
30	0.75	0.60	0.67	5
31	0.84	0.95	0.89	57
32	0.83	0.71	0.77	14
33	0.94	0.91	0.92	32
34	0.90	0.64	0.75	14
35	1.00	0.93	0.96	14
36	0.70	0.78	0.74	18
37	0.89	0.67	0.76	12
38	0.95	0.95	0.95	40
39	0.89	0.89	0.89	9
40	1.00	0.98	0.99	46
41	0.86	0.90	0.88	20
42	0.73	0.58	0.65	19
43	0.90	0.82	0.86	22
44	0.76	0.89	0.82	18
45	0.73	0.89	0.80	18
46	0.85	0.88	0.87	33
accuracy			0.85	1804
macro avg	0.84	0.84	0.84	1804
weighted avg	0.86	0.85	0.85	1804

	precision	recall	f1-score	support
1	0.82	0.82	0.82	102
2	0.88	0.89	0.89	74
3	1.00	0.96	0.98	112
4	0.79	0.92	0.85	78
5	0.90	0.87	0.88	82
6	0.89	0.81	0.85	68
7	0.82	0.79	0.81	58
8	0.66	0.95	0.78	20
9	1.00	0.96	0.98	27
10	0.91	0.79	0.85	39
11	0.73	0.91	0.81	35
12	0.90	0.73	0.81	49
13	0.70	0.70	0.70	30
14	0.88	0.84	0.86	43
15	0.83	0.79	0.81	38
16	0.93	0.84	0.88	50
17	0.83	0.83	0.83	36
18	0.80	1.00	0.89	4
19	1.00	0.96	0.98	25
20	0.92	0.94	0.93	90
21	0.78	0.86	0.82	29
22	0.74	0.73	0.74	59
23	0.83	0.71	0.76	34
24	0.77	0.89	0.83	19
25	0.91	0.90	0.90	67
26	0.82	0.93	0.87	29
27	0.93	0.95	0.94	60
28	0.89	0.80	0.84	20
29	0.97	0.94	0.96	36
30	1.00	0.60	0.75	5
31	0.82	0.95	0.88	57
32	0.87	0.93	0.90	14
33	0.90	0.88	0.89	32
34	0.83	0.71	0.77	14
35	0.93	1.00	0.97	14
36	0.68	0.72	0.70	18
37	0.90	0.75	0.82	12
38	0.97	0.97	0.97	40
39	1.00	0.89	0.94	9
40	0.98	1.00	0.99	46
41	0.82	0.90	0.86	20
42	0.71	0.53	0.61	19
43	0.95	0.86	0.90	22
44	0.89	0.89	0.89	18
45	0.71	0.83	0.77	18
46	0.88	0.88	0.88	33
accuracy			0.87	1804
macro avg	0.86	0.86	0.85	1804
weighted avg	0.87	0.87	0.87	1804

Şekil 18: BoW ve TF-IDF Metotlarına Göre Lineer SVC Modelinin Sınıf Bazında Başarı Oranları

TF-IDF öznitelik çıkarma metodu kullanılarak Lineer SVC algoritması ile oluşturulmuş karşılaştırılan en başarılı modelin karışıklık matrisi Şekil 19'da verilmiştir.



Şekil 19: TF-IDF Metodu Kullanılarak Lineer SVC Algoritması ile Oluşturulmuş Modelin Karışıklık Matrisi

Matris incelendiğinde, tahmin edilen ve gerçek değerlerin kesişim noktalarındaki sayıların büyüklüğü ve tam tersi kesişim noktaları dışındaki sayaların küçüklüğü modelin başarısını açıklamaktadır. Örneğin 19 numaralı arıza tanımı 90 kez tahmin edilmiş ve 85 kez gerçek sınıfı doğru olarak tahmin edilebilmiştir.

SONUÇ

Çalışma kapsamında; incelenen bakım yönetim sisteminin etkin kullanılmasındaki engellerden olan arıza açıklamalarının ve arıza tanımlarının hızlı, verimli ve doğru şekilde sisteme dahil edilememesi problemine çözüm olarak sunulan arıza açıklamalarından arıza tanımının tahminini yapabilen makine öğrenmesi modelinin geliştirilmesi hedeflenmiştir.

Çalışmanın hedefi kapsamında; bakım yönetim sistemindeki tüm arızaların %72,12'sine denk gelen 46 arıza tanımının tahminini, arıza açıklamalarına bakarak %87 başarı ile yapabilen model oluşturuldu. Geliştirilen modelin bakım yönetim sistemine entegrasyonu ile en sık karşılaşılan arızaların kayıt edilme işlemi kısalıp kolaylaşırken sürecin hızlanması beklenmektedir. Personelin yeni sistemde yazdığı açıklamanın doğru tahmin edilip edilmediğini kontrol ederken aynı zamanda yanlış arıza tanımı girilmesinin önüne geçileceği ve modelin daha fazla doğru tahmin edebilmesi için özensiz açıklamaların sona ereceği öngörülmektedir.

Gelecek çalışmalarda, bakım yönetim sistemini kullanan personele verilerin sağlıklı tutulması halinde sağlanacak faydalar anlatılarak veri kalitesi arttırılmaya çalışabilir ve daha kaliteli veriler ile daha başarı sonuçlar elde edilebilir. Sistemdeki kayıt sayısının artması ile daha fazla arıza tanımı için aynı çalışma gerçekleştirilebilir. Arıza açıklamaları metinlerinden sözlük oluşturulurken kelime kökü bulma yöntemleri denenerek başarıyı arttırıp arttırmayacağı araştırılabilir.

KAYNAKÇA

Ahmetoğlu, H., & Resul, D. A. Ş. (2020). Türkçe Otel Yorumlarıyla Eğitilen Kelime Vektörü Modellerinin Duygu Analizi ile İncelenmesi. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 24(2),* 455-463.

Aksu, M. Ç., & Karaman, E. (2020). FastText ve Kelime Çantası Kelime Temsil Yöntemlerinin Turistik Mekanlar İçin Yapılan Türkçe İncelemeler Kullanılarak Karşılaştırılması. *Avrupa Bilim ve Teknoloji Dergisi, (20),* 311-320.

Bal, A. (2013). *Üretim Tesisleri İçin Rfid Destekli Bakım Yönetimi* (Yayımlanmamış Doktora Tezi). İstanbul Teknik Üniversitesi / Fen Bilimleri Enstitüsü, İstanbul.

Bonaccorso, G., (2018). *Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning,* Packt Publishing.

Çelik, Ö., & Koç, B. C. (2021). TF-IDF, Word2vec ve Fasttext Vektör Model Yöntemleri ile Türkçe Haber Metinlerinin Sınıflandırılması. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 23(67),* 121-127.

Çılgın, C., Ünal, C., Alıcı, S., Akkol, E., & Gökşen, Y. (2020). Metin Sınıflandırmada Yapay Sinir Ağları ile Bitcoin Fiyatları ve Sosyal Medyadaki Beklentilerin Analizi. *Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi, 4(1),* 106 - 126.

Edwards, B., Zatorsky, M., & Nayak, R. (2008). Clustering and Classification of Maintenance Logs Using Text Data Mining, *Conferences in Research and Practice in Information Technology Series, 87,* 193–199.

Eryılmaz, E. E., Şahin, D. Ö., & Kılıç, E. (2020). Türkçe İstenmeyen E-postaların Farklı Öznitelik Seçim Yöntemleri Kullanılarak Makine Öğrenmesi Algoritmaları ile Tespit Edilmesi. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 13(2),* 57-77.

Gollapudi, S., & Laxmikanth, V. (2016*). Practical Machine Learning.* Packt Publishing.

Göçgün, Ö. F., & Onan, A. (2021). Amazon Ürün Değerlendirmeleri Üzerinde Derin Öğrenme/Makine Öğrenmesi Tabanlı Duygu Analizi Yapılması. *Avrupa Bilim ve Teknoloji Dergisi, (24),* 445-448.

Kaşıkçı, T., Gökçen, H., (2014). Metin Madenciliği ile E-Ticaret Sitelerinin Belirlenmesi, *Bilişim Teknolojileri Dergisi, 7 (1).*

Mengutayci, Ü. & Temurtaş, H. (2021). Yapay Sinir Ağları ile Türkçe Otel Yorumlarının Sınıflandırılması. *Internatıonal Black Sea Coastlıne Countrıes Scıentıfic Research Symposıum, (VI),* 683-687

Mohammed, M., Khan, M., Bashier, E. (2016). *Machine Learning Algorithms and Applications.* CRC Press.

Uslu, O., & AKYOL, S. (2021). Türkçe Haber Metinlerinin Makine Öğrenmesi Yöntemleri Kullanılarak Sınıflandırılması. *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 2(1),* 15-20.

Özek, S. (2012). *Termik santrallarda bakım yönetim sisteminin enerji verimliliğine etkileri.* (Yayımlanmamış Yüksek Lisans Tezi). Gazi Üniversitesi / Fen Bilimleri Enstitüsü, Ankara

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12,* 2825-2830.

Raschka S. (2015). *Python Machine Learning,* Packt Publishing

Safalı, Y. (2020). Sosyal Medya Kullanıcılarının Cumhur İttifakı Hakkındaki Görüşlerinin Veri Madenciliği Teknikleri ile Sınıflandırılması. *Bilgisayar Bilimleri ve Teknolojileri Dergisi, 1(2),* 51-57.

sklearn, (2020). Erişim adresi: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning (Erişim Tarihi: 30.04.2021)

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

Taşkın, M. F. (2006). *Önleyici bakım politikası altında optimum stok miktarının bulanık mantık yöntemiyle belirlenmesi* (Yayımlanmamış Doktora Tezi). Sakarya Üniversitesi/Fen Bilimler Enstitüsü, Sakarya.

Thanaki, J., (2017). *Python Natural Language Processing.* Packt Publishing.

Thanaki, J., (2018). *Machine Learning Solutions: Expert Techniques to Tackle Complex Machine Learning Problems Using Python,* Packt Publishing.

Tuzcu, S. (2020). Çevrimiçi Kullanıcı Yorumlarının Duygu Analizi ile Sınıflandırılması. *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 1(2),* 1 - 5.

Türkalp, M. V. (2019). *Tviter verileri üzerinde sınıflandırma algoritmaları kullanarak hisse senedi değerleri için yön tahmini,* (Yüksek Lisans Tezi). Konya Teknik Üniversitesi / Lisansüstü Eğitim Enstitüsü, Konya

Yilmaz, Ş. Ş., Özer, İ., & Gökçen, H. (2021). Türkçe Metinlerde Derin Öğrenme Yöntemleri Kullanılarak Duygu Analizi. *International Symposium of Scientific Research and Innovative Studies (Vol. 22),* 25.

ÖZGEÇMİŞ

İbrahim Burak Tosun, Lisans eğitimini Sakarya Üniversitesi Mühendislik Fakültesi Endüstri Mühendisliği Bölümü'nde 2017 yılında tamamlamıştır. İkinci lisans eğitimini ise Anadolu Üniversitesi İşletme Fakültesi İşletme Bölümü'nde 2020 yılında tamamlamıştır. Sakarya Üniversitesi Yönetim Bilişim Sistemleri Anabilim Dalı'nda yüksek lisansına 2017 yılında başlamıştır. Web Analitiğinde Sosyal Medya Kullanımı konusunda bilimsel yayını bulunmaktadır.