
# MAKİNE ÖĞRENMESİ İLE BELGE TANIMA

**2024**
**YÜKSEK LİSANS TEZİ**
**BİLGİSAYAR MÜHENDİSLİĞİ**

**İsa YURDABAKAN**

**Tez Danışmanı**
**Dr. Öğr. Üyesi Muhammet ÇAKMAK**

# MAKİNE ÖĞRENMESİ İLE BELGE TANIMA / DOCUMENT RECOGNITION WITH MACHINE LEARNING

**İsa YURDABAKAN**

**Tez Danışmanı**
**Dr. Öğr. Üyesi Muhammet ÇAKMAK**

**T.C.**
**Karabük Üniversitesi**
**Lisansüstü Eğitim Enstitüsü**
**Bilgisayar Mühendisliği**
**Yüksek Lisans Tezi**
**Olarak Hazırlanmıştır**

**KARABÜK**
**Temmuz 2024**

İsa YURDABAKAN tarafından hazırlanan "MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANARAK BELGE ALGILAMA" başlıklı bu tezin Yüksek Lisans Tezi olarak uygun olduğunu onaylarım.

Dr. Öğr. Üyesi Muhammet ÇAKMAK .....
Tez Danışmanı, Bilgisayar Mühendisliği Anabilim Dalı

Bu çalışma, jürimiz tarafından Seçiniz ile Anabilim Dalınız Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir. 16/05/2024

Ünvanı, Adı SOYADI (Kurumu) İmzası

Üye : Doç. Dr. Zafer ALBAYRAK (SUBÜ) .....

Üye : Dr. Öğr. Üyesi Muhammet ÇAKMAK (SNÜ) .....

Üye : Dr. Öğr. Üyesi İsa AVCI (KBÜ) .....

KBÜ Lisansüstü Eğitim Enstitüsü Yönetim Kurulu, bu tez ile, Yüksek Lisans derecesini onamıştır.

Doç. Dr. Zeynep ÖZCAN .....
Lisansüstü Eğitim Enstitüsü Müdürü

*"Bu tezdeki tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu; ayrıca bu kuralların ve ilkelerin gerektirdiği şekilde, bu çalışmadan kaynaklanmayan bütün atıfları yaptığımı beyan ederim."*

İsa YURDABAKAN

# ÖZET

**Yüksek Lisans Tezi**

**MAKİNE ÖĞRENMESİ İLE BELGE TANIMA**

**İsa YURDABAKAN**

**Karabük Üniversitesi**
**Lisansüstü Eğitim Enstitüsü**
**Bilgisayar Mühendisliği**

**Tez Danışmanı:**
**Dr. Öğr. Üyesi Muhammet ÇAKMAK**
**Temmuz 2024, 42 sayfa**

Gümrük, bir ülkenin sınırlarından eşya ve malların giriş ve çıkışlarında denetimlerin yapıldığı ve vergilerin ödendiği bir kamu kurumu olarak tanımlanabilir. Gümrük beyannameleri, bu süreçte önemli bir rol oynar çünkü gümrük vergisinin toplanmasını sağlar. Gümrük beyannameleri üzerinden alınan gümrük vergisi, uluslararası ticarette eşyanın ithali ya da ihracına bağlı olarak öngörülen vergilerden sadece birini oluşturur. Bu, gümrük işlemlerinin düzgün bir şekilde yürütülmesi ve uluslararası ticaretin düzenli bir şekilde gerçekleştirilmesi için kritik öneme sahiptir. Gümrük beyannamelerinin hatasız bir şekilde yazılması, bu sürecin etkinliği ve doğruluğu açısından büyük önem taşır. Beyannamenin yazılmasında kullanılan metod çeşitliliği ve karmaşıklığı, insan kaynaklı hataların oluşmasına neden olabilir. Bu nedenle, bu sürecin otomatikleştirilmesi ve makine öğrenmesi teknolojilerinin kullanılması, bu tür hataları önlemeye yardımcı olabilir.

Makine öğrenmesi teknolojileri, özellikle gümrük beyannamelerinin doğruluğunu artırarak, uluslararası ticaretin daha düzenli ve güvenli bir şekilde yürütülmesine katkıda bulunmaktadır. Bu teknolojiler, aynı zamanda veri işleme sürecini hızlandırmakta ve insan kaynaklı hataları minimize etmektedir. Dijital dönüşüm ve sayısallaştırma süreçleri, kurumların daha verimli çalışmasını sağlarken, maliyetleri düşürmekte ve bilgiye erişimi kolaylaştırmaktadır.

Bu tezde makine öğrenmesi yöntemleri ile taranan doküman üzerindeki metinlerin en iyi doğruluk oranı ile sınıflandırılması yapılmıştır. Çalışmada, destek vektör makinesi (SVM), k-en yakın komşu (Knn), karar ağaçları (DT), rastgele orman (RF), eXtreme Gradient Boosting (XGBoost) ve ANN yöntemleri incelenmiştir. En iyi yöntem olarak XGBoost uygulanmış olup, 97.49% başarı oranı elde edilmiştir.

**Anahtar Sözcükler :** Makine Öğrenmesi Algoritmaları, Veri İşleme, Belge Tanıma
**Bilim Kodu** **:** 92431

## ABSTRACT

**Master Thesis**

**DOCUMENT DETECTION USING MACHINE LEARNING ALGORITHMS**

**İsa YURDABAKAN**

**Karabük University**
**Institute of Graduate Programs**
**Computer Engineering**

**Thesis Advisor:**
**Assist. Prof. Dr. Muhammet ÇAKMAK**
**July 2024, 42 pages**

Customs can be defined as a public institution where inspections and tax payments occur for goods entering or leaving a country's borders. Customs declarations play a crucial role in this process as they facilitate the collection of customs duties. Customs duties collected through customs declarations constitute only one of the taxes envisaged depending on whether goods are imported or exported in international trade. Therefore, it is of critical importance for customs procedures to be conducted properly and for international trade to be conducted regularly. The accurate completion of customs declarations is crucial for the efficiency and accuracy of this process. The variety and complexity of methods used in drafting declarations can lead to human errors. Hence, automating this process and utilizing machine learning technologies can help prevent such errors.

Machine learning technologies contribute significantly to enhancing the accuracy of customs declarations, thereby facilitating more orderly and secure international trade.

These technologies also expedite data processing and minimize human errors. Digital transformation and digitization processes enable institutions to operate more efficiently, reduce costs, and facilitate access to information.

In this study, text classification of scanned documents using machine learning methods was conducted to achieve the highest accuracy rate. The study examined Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees (DT), Random Forests (RF), eXtreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN). XGBoost was implemented as the best method, achieving a success rate of 97.49%.

**Key Word** **:** Machine Learning Algorithms, Data Processing, Document Recognition

**Science Code** **:** 92431

## TEŞEKKÜR

Bu tez çalışmasının planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen ve her zaman destek olan değerli hocam Dr. Öğr. Üyesi Muhammet ÇAKMAK'a, arkadaşlarıma ve sevgili aileme de minnettarım. Onların desteği olmasaydı bu tez çalışması mümkün olmazdı.

Son olarak, bu çalışmanın gerçekleşmesine yardımcı olan herkese teşekkürlerimi sunarım. Bu süreçte bana sabır, anlayış ve cesaret veren herkese derin minnettarlığımı ifade ederim.

Teşekkürler,

## İÇİNDEKİLER

**Sayfa**

# ŞEKİLLER DİZİNİ

**Sayfa**

## ÇİZELGELER DİZİNİ

**Sayfa**

# KISALTMALAR DİZİNİ

## KISALTMALAR

**YZ** : Yapay Zeka

**ML** : Machine Learning

**AI** : Artificial Intelligence

**DT** : Decision Tree

**RF** : Random Forest

**KNN** : K-Nearest Neighbour

**ANN** : Yapay Sinir Ağları

**MAE** : Ortalama Mutlak Hata

# BÖLÜM 1

# GİRİŞ

Gümrük, ihracat ve ithalat işlemlerine tabi olan eşyaların denetlendiği ve vergilendirildiği bir kurumdur. Ticarete konu olan eşyanın ithalatı ve ihracatına bağlı olarak alınan vergiler, beyanname açısından farklı kanunlar ya da yasal metinler ile düzenlenmiştir [1]. Gümrük vergilendirilmesi için kullanılan beyannamelerin hatasız sunulması önemlidir [2]. Gümrük beyannamesi, çeşitli unsurların bir araya gelmesiyle oluşan karmaşık bir yapıyı temsil eder. Unsurlar, gümrük tarifeleri, ithalat ve ihracat düzenlemeleri, ürün sınıflandırmaları ve daha pek çok faktörü içerir. Bileşenlerin her biri, beyannamenin genel işlevselliği ve doğruluğu üzerinde önemli bir etkiye sahiptir. Gümrük beyannamesi, gümrük işlemlerinin merkezinde yer alır ve dikkatli bir şekilde oluşturulması gereken bir belgedir. Bu süreç, önemli ölçüde yetenekli iş gücü ve zaman gerektirir. Uygulama yazılımının geliştirilmesi süreci, teknik yetenek ve bilgiye ek olarak, iş süreçlerinin ve mevzuatın derinlemesine anlaşılmasını gerektirir. Mevzuatlar, her geçen gün yenilenen ve sürekli takip edilmesi gereken dinamik bir özellik taşır. İşin bu tür karmaşıklıkları yanında, yetkin ve yeterli insan kaynağının eksikliği, beyanname yazımında ciddi hataların ortaya çıkmasına yol açmaktadır. Bu tür hatalar, yüklü cezalarla karşılaşılmasına neden olurken, aynı zamanda operasyon maliyetlerini ve süreçlerini de olumsuz yönde etkilemektedir. Ayrıca, yasal ve hukuki bazı olumsuz sonuçları da beraberinde getirebilmektedir.

Günümüzde, Nesnelerin İnterneti (IoT) ve Makineler Arası İletişim (M2M) gibi ileri teknolojiler, modern dünyada geniş bir uygulama yelpazesine sahip olup, sürekli gelişmekte ve evrilmekte olup, bilişim teknolojilerindeki hızlı ilerlemeler yaşamın hemen hemen her yönünü etkilemiştir [3]. Özellikle lojistik sektörü, bu yeni teknolojilerden biri olan blok zincir teknolojisinin dönüştürücü etkilerini deneyimlemektedir [4]. Dijitalleşme ve dijital dönüşüm süreçleri, insan emeği, iş gücü

dinamikleri ve çalışma koşulları üzerinde belirgin değişiklikler yaratmaktadır. Bu, çalışma türlerinin çeşitliliğini ve yapısını da etkilemektedir.

Makine öğrenmesi yöntemleri ile beyanname üzerinde bulunan unsurların sınıflandırlması, gümrük teknolojisindeki en büyük ilerlemelerden birini temsil etmektedir. Gelişmiş algoritmalar ve makine öğrenmesi modellerinden yararlanılarak, gümrük beyannamelerinin çeşitli özelliklerini - örneğin, ürün kodları, beyanname alıcısı, beyanname göndericisi, MRN numarası ve toplam ürün ağırlığı - analiz edebilir ve böylece, doğru beyanname oluşturma ve uygun gümrük vergilendirmesini sağlanabilir. Öte yandan, genellikle görüntülerden elde edilen bu morfolojik özelliklere ek olarak, makine öğrenme algoritmaları ile özellikler hesaplayarak gümrük beyannamesi tespit sistemlerini modellemek, verimli sonuçlar elde etmeye yardımcı olabilir. Tespit sistemleri, sadece operasyonel verimliliği artırmakla kalmaz, aynı zamanda gümrük beyanname yazım uygulamalarını da destekler. Gereksiz kaynak kullanımını en aza indirerek, süreçlerin daha maliyet etkin hale gelmesine yardımcı olur. Bütünüyle bakıldığında, makine öğrenmesi tekniklerinin gümrük beyanname tespit süreçlerine dahil edilmesi, beyanname oluşturma sürecinde verimlilik, kalite ve sürdürülebilirliği artırmakta ve özellikle bu alanda faaliyet gösteren sektörlerde yeni potansiyel uygulama alanlarını keşfetmekte önemli bir rol oynamaktadır.

Bu çalışmada, karma vektörleştirici (HashingVectorizer), terim frekansı ve ters belge frekansı (TF-IDF), sayım vektörleştirici (CountVectorizer), temel bileşenler analizi (PCA) ve Otokodlayıcı (Autoencoder) olmak üzere 5 farklı özellik çıkarma yöntemi, gümrük beyannamesinde bulunan unsurların özellik çıkarımı ve sınıflandırılması amacıyla kullanılmıştır. Bu bölüm, gümrük beyannamesi sınıflandırmasına odaklanan araştırmanın sonuçlarını sunmaktadır. Karma vektörleştirici (HashingVectorizer), terim frekansı ve ters belge frekansı (TF-IDF), sayım vektörleştirici (CountVectorizer), temel bileşenler analizi (PCA) ve Otokodlayıcı (Autoencoder) mimarileri kullanılarak çıkarılan özellikler, daha sonra destek vektör makinesi (SVM), k-en yakın komşu (KNN), karar ağaçları (DT), rastgele orman (RF), eXtreme Gradient Boosting (XGBoost) makine öğrenme algoritmaları ve yapay sinir ağı (ANN) kullanılarak sınıflandırma sürecinde kullanılmıştır. Makine öğrenme yaklaşımları ile

gerçekleştirilen sınıflandırma süreçlerinde, bu özellikler 97.49% doğruluk oranı ile ayırt edilmiştir.

Çalışma kapsamında, aşağıdaki önemli katkılar sağlanmıştır:

1. Çok sınıflı gümrük beyannamelerinin özelliklerinin çıkarılması ve sınıflandırılması için hibirt bir (X model) makine öğrenme yöntemi geliştirilmiştir.
2. Önerilen model, beyannamelerin çeşitli unsurlarını - örneğin, ürün kodları, beyanname alıcısı, beyanname göndericisi, MRN numarası ve toplam ürün ağırlığı - analiz edebilir ve böylece, doğru beyanname oluşturma ve uygun gümrük vergilendirmesini belirlemeyi sağlar.
3. Önerilen hibrit ve geleneksel makine öğrenmesi modellerin başarıları accuracy, precision, F1-score, recall ve MEA değerlerine göre karşılaştırılmıştır.
4. Makine öğrenmesi algoritmalarının gümrük beyannamesi tespit için kullanılması beyanname oluşturma sürecinde hızı, verimlilik ve kaliteyi artırarak sürdürülebilirliğe katkı sağlamaktadır.

Önerilen çalışmada, manuel özellik çıkarımı ve derin özelliklerin birlikte kullanıldığı hibrit (TF-IDF, PCA ve XGBoost) yaklaşımı kullanılmıştır.

İkinci bölümde, literatür çalışması verilmiştir. Üçüncü bölümde veri seti ve metodoloji tanıtılmaktadır. Dördüncü bölüm, deneysel çalışma ve sonuçları sunar. Beşinci bölümde, tartışma ve sonuçlar ele alınmaktadır. Son olarak, sonuçlar ve gelecek çalışmalar bölümü verilmişti

# BÖLÜM 2

# LİTERATÜR

Teknolojideki hızlı gelişmelerle beraber beyanname bileşenlerinin tespit edilmesi için kullanılan bilgisayar destekli sistemlerle yapılan çalışmalar yoğunlaşmıştır. Literatürde metin sınıflandırması üzerine çok sayıda çalışma bulunmaktadır. Beyanname hata tespitleri üzerinde çeşitli sınıflandırma çalışmaları mevcuttur. Bu çalışmalarda, sınıflandırma işlemleri farklı yöntemlerle gerçekleştirilmiştir. Araştırmacılar, sınıflandırma işlemleri için çeşitli makine öğrenmesi algoritmaları ve derin öğrenme modelleri kullanmışlardır.

Literatürde metin özellik çıkarımı ve sınıflandırması üzerine yapılan bazı çalışmalar aşağıda verilmiştir. Bu çalışmalar, çeşitli amaçlar için gerçekleştirilmiştir. Bu amaçlar arasında metin üzerinde özellik çıkarımı [5], büyük boyuttaki verileri daha düşük boyuta indirgeme [6], beyanname bileşenlerinin makine öğrenme yöntemleri üzerinde değerlendirilmesi [7], gümrük dolandırıcılık tespiti [8] ve ticarette gümrük beyannamesinin önemi ve insan hatalarının kaçınılmaz cezalara nasıl yol açtığı [9] bulunmaktadır. Ayrıca, özellik çıkarımı, sınıflandırması ve beyanname tespiti üzerine çeşitli çalışmalar da bulunmaktadır. Bu çalışmalar aşağıda daha detaylı bir şekilde tartışılmaktadır.

Agustina ve diğerleri [10], Twitter'daki 13297 metin üzerinde Destek Vektör Makineleri (SVM) yöntemini uygulamıştır. Özellik çıkarma yöntemi olarak terim frekansı-ters belge frekansı (TF-IDF) ve Word2vec algoritmasını kullanmışlardır. SVM+TF-IDF kombinasyonu %85 precision, %86 recall ve %84 f1 score ile en iyi sonucu üretmiştir. Semary ve diğerleri [11], duygu analizi görevlerinin performansını artırmak için en uygun bir özellik çıkarma yöntemi üzerinde çalışmışlardır. Çalışmalarında Amazon yorum veri seti ve Twitter ABD havayolları veri setlerini kullanmışlardır. Rastgele Orman (RO) algoritması üzerinde terim frekansı-ters belge

frekansı (TD-IDF) yöntemi kullanılarak, Amazon yorum veri setinde %99 ve Twitter ABD havayolları setinde %96 doğruluk oranları elde etmişlerdir.

Doo ve Kim [12], aynı anda kümeleme ve özellik seçimi gerçekleştiren ve özellikler arasındaki göreceli önemi yorumlayan sıralama K-concrete otokodlayıcıyı önermişlerdir. Önerilen yöntem, verileri birleşik bir DNN ve K-means kümeleme çerçevesi ile kümeler ve giriş katmanından sonra bir beton seçici katman ekleyerek önemli özellikleri seçmiştir. Natha ve Rajeswari [13], hem otomatik hem de manuel özellik çıkarma metodolojilerinin gücünden yararlanarak cilt kanserini tespit etmek için tasarlanmış XGBoost, Lojistik Regresyon, Uzun Kısa Süreli Bellek (LSTM), CatBoost, Çok Katmanlı Algılayıcı (MLP), Bayesian Model Averaging (BMA) ve Bayesian Model Kombinasyonu (BMC) algoritmaları tanıtmışlardır. Önerilen model hibirit bir PCA ve Autoencoder modeli kullanmaktadır. Catboost 0.61 ile en yüksek doğruluk oranın elde etmiştir.

Avcı [14], makine öğrenmesi ve özellik seçimi tekniklerine dayalı etkili bir Saldırı Tespit Sistemi (IDS) oluşturmayı amaçlamıştır. Saldırı tespiti için uygun tekniği bulmak amacıyla dört farklı makine öğrenme tekniği olan Rastgele Orman (RF), K-En Yakın Komşular (KNN) ve Destek Vektör Makinesi'nin (SVM) performansı karşılaştırmıştır. New Brunswick Üniversitesi KDD'99 veri setinin temiz ve rafine edilmiş bir versiyonu olan NSL-KDD veri setini kullanarak Rastgele Orman algoritmasında 99.72% doğruluk oranı elde etmiştir.

Seck [15], Senegal'deki gümrük beyannamesi kontrol sistemini iyileştirmek için, Yapay Sinir Ağları (MLP), Destek Vektör Makinesi (SVM), Rastgele Orman (RF) ve eXtreme Gradient Boosting (XGBoost) gibi makine öğrenmesi yöntemleri ile inşa edilmiş dolandırıcılık riski tahminlerini sınıflandırmıştır. Rastgele Orman (RF) %96 doğruluk değeri en iyi başarı değerini elde etmiştir.

Uluer ve diğerleri [16], Sınır Ağ Geçidi Protokolü (BGP) anomalilerinin tespitinde makine öğrenmesi ve derin öğrenme algoritmaları kullanılarak bir sınıflandırma modeli önermişlerdir. Önerilen model karar ağacı, rastgele orman ve çok katmanlı algılayıcı algoritmalarına dayalı olarak geliştirilmiştir. Modelin değerlendirilmesinde

dolaylı BGP anomalileri ve bağlantı hatası anomalileri, doğruluk ve F1-puan ölçütleri kullanılmıştır. Slammer veri seti kullanılarak gerçekleştirdikleri çalışmada en iyi sonucun %99,47 doğruluk ve %98,85 F1-Puanı değeriyle Hibrit Model ile elde edildiği görülmüştür.

Aisyah [17], finans sektöründe kritik bir rol oynayan kredi durumunun tahmin edilmesi konusunda Karar Ağacı Sınıflandırıcısının etkinliğini incelemiştir. Bu çalışma, %82 doğruluk oranı elde ederek, Karar Ağacı'nın bu tür veriler için uygun bir model olduğunu göstermiştir. HIDAJAT [18] ise, Endonezya'daki tasarruf ve kredi kooperatiflerinde standart bir kredi puanı değerlendirmesi olmadığından, kooperatifler için bir kredi puanı modeli önermiştir. Bu model, Logistic Regression Classifier, Support Vector Machine Classifier, K-Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier ve Light Gradient Boosting Machine Classifier olmak üzere 7 farklı makine öğrenimi algoritmasını kullanmıştır. Logistic Regression Classifier ve Destek Vektör Makinesi (SVM) Sınıflandırıcı, %98 doğruluk oranı elde etmiştir.

Isnayni [19] tarafından yürütülen bir başka çalışmada, kahve dükkanı incelemeleri için Rastgele Orman Sınıflandırıcısı yöntemi kullanılmıştır. Ön işleme aşaması, büyük-küçük harf dönüşümü, belirteçleme, durak kelime kaldırma ve kök çıkarma işlemlerini içermiştir. Bu çalışmanın sonuçları, Rastgele Orman Sınıflandırıcısı yöntemiyle kahve dükkanı incelemelerinin %79 doğruluk oranı bulmuştur.

Djeldjli ve diğerleri [20] tarafından yürütülen bir çalışma, Güney Cezayir'deki altı şehir için yapılmış olup, destek vektör makineleri (SVM), yapay sinir ağları (ANN) ve yeni bir hibrid ateşböceği algoritması tabanlı model (FFA-ANN) olmak üzere üç modelin doğruluğunu, on bir yıllık bir dönem boyunca küresel güneş ışıması tahmin edilirken incelemiştir. ANN ve SVM modelleri umut verici sonuçlar üretmiş olsa da, önerilen FFA-ANN hibrid modeli, en iyi değerlerle (R = 0.9321, rRMSE = 9.35% ve MAPE = 6.29%) üç istatistiksel faktör olan korelasyon katsayısı, göreceli kök ortalama karesel hata ve ortalama mutlak yüzde hatası kullanarak tek başına ANN tabanlı modeli geride bırakmıştır. Bulgular, FFA-ANN'in tüm bölgelerde günlük küresel

güneş ışıması tahmin edilirken optimize edilmiş SVM ve ANN modellerine tercih edildiğini göstermiştir.

Son olarak, Salih ve Zeebaree [21] tarafından yürütülen bir çalışma, Ekstra Ağaçlar Sınıflandırıcısı'nın performansını değerlendirmeyi ve özellikle CPU paralel işlemenin sınıflandırma doğruluğu ve hesaplama verimliliği üzerindeki etkisini incelemiştir. Moda MNIST, giysi öğelerini temsil eden gri tonlamalı görüntülerin bir koleksiyonu, bu çalışmanın temel veri seti olarak kullanılmıştır. Ekstra Ağaçlar Sınıflandırıcısı Paralel İşlem Olmadan, %88,23'lük bir doğruluk elde edilmiştir. Paralel İşlemle Ekstra Ağaçlar Sınıflandırıcısı, CPU paralel işlemi kullanarak, %88,43'lük bir doğruluk elde etmiştir.

Literatürde makine öğrenmesi ve derin öğrenme üzerine önemli sayıda çalışma olmasına rağmen, beyanname sınıflandırması üzerine yeterli çalışma bulunmamaktadır. Bu çalışmada Morfolojik özelliklere ek olarak, daha yüksek bir veri temsili seviyesine sahip derin özellikler de kullandık. Özellikle gümrük beyannamesinin hatasız olması üzerine yapılan çalışmalar, gümrük sektöründe beyanname bileşenlerin sınıflandırılmasıda makine öğrenmesi algoritmalarının kullanımı potansiyelini ortaya koymaktadır. Bu alan geliştikçe, gümrük vergilendirmesindeki hata oranını ve beyanname üzerindeki hataların iyileştirmesinde önemli ilerlemeler olacağı öngörülmektedir.

# BÖLÜM 3

# MATERYAL VE YÖNTEMLER

## 3.1. Veri Seti

Çalışma için özel bir veri seti kullanılmıştır. Kullanılan veri seti 2020 yılına ait olan ve TOBB şirketine ait beyannamelerin OCR ile çıkarıldığı 'Gönderici Adı', 'Gönderici Adresi', 'Gönderici Posta Kodu', 'Gönderici Şehir', 'Gönderici Ülke', 'eori Numarası', 'Alıcı Adı', 'Alıcı Adresi', 'Alıcı Posta Kodu', 'Alıcı Şehir', 'Alıcı Ülke', 'bölge', 'mrn Numarası', 'sayfa Numarası', 'toplam Sayfa Sayısı', 'toplam Ürün Sayısı', 'toplam Paket Sayısı', 'toplam Brüt Ağırlık', 'çıkış Ülke Kodu', 'varış Ülke Kodu' özellikleri içeren bir veri setidir.

## 3.2. Öznitelik Çıkartma ve Öznitelik Seçimi

Bu çalışmada, özellikle belirli veri setleri için özellik çıkarma araçları olarak kullanılan HashingVectorizer, Terim Frekansı ve Ters Belge Frekansı (TF-IDF) ve Sayım Vektörleştirici (CountVectorizer) gibi yöntemler ele alınmıştır. Bu yöntemler, ham metinleri sayısal vektörlere dönüştürmek için temel rol oynamaktadır ve bu vektörler daha sonra makine öğrenme modelleri tarafından işlenebilmektedir [22]. CountVectorizer, metin belgelerini bir token sayısı matrisine dönüştürerek, her kelimenin belge içindeki frekansını etkili bir şekilde yakalamaktadır. TfidfVectorizer ise, CountVectorizer'ın faydalarını Term Frekansı-Ters Belge Frekansı (TF-IDF) ağırlıkları ile birleştirerek, sık tekrarlanan ve dolayısıyla daha az bilgi sağlayan belirteçlerin etkisini azaltır. HashingVectorizer ise, her kelimenin metindeki tekrar sayısını dikkate alarak verilen metni bir vektöre dönüştürmektedir [23].

1. CountVectorizer: metin verilerini etkili bir şekilde destekleyen token'ların frekans matrisine dönüştürerek, metni öğrenme modelleri için uygun hale getirir.
2. TfidfVectorizer: metindeki her bir token için TF-IDF puanlarını hesaplayarak, token'ların belgelerin genelindeki nadirliğine karşı sıklığını dengeleyerek anlamlı kelimeleri tanıma yeteneğini artırır.
3. HashingVectorizer: terim frekansı sayılarını bir karma işlevine uygular ve bellekte bir kelime dağarcığını depolama ihtiyacını ortadan kaldırarak büyük ölçekli metin işleme için verimli hale getirir. Bu sayede, metin verileri daha etkin bir şekilde işlenebilir.

Özellik seçimi, belirli bir algoritmanın uygulanmasıyla gerçekleştirilir. Bu süreç, belirli bir probleme en uygun ve önemli özellikleri belirleyerek veri setinin boyutunu azaltma stratejisi olarak hizmet eder. Başlangıçta, orijinal veri setinden bir özellik alt kümesi çıkarılır. Daha sonra, değerlendirme süreci çeşitli kriterler veya matematiksel formülasyonlar kullanılarak düşünülen özelliklerin uygunluğunu değerlendirir. Bu değerlendirmeye dayanarak, belirli bir özelliğin bir özellik alt kümesine dahil edilip edilmeyeceğine ilişkin bir karar verilir. Seçilen özellikler ilgili alt kümeye dahil edilir ve bu süreç, algoritmanın önceden tanımlanmış durma kriterleri karşılanana kadar devam eder. Özellik seçimi, aynı zamanda nitelik veya değişken seçimi olarak da bilinir ve orijinal veri kümesini etkili bir şekilde karakterize eden değişkenlerin optimal alt kümesini belirlemede kritik bir rol oynar. Temel bileşen analizi (PCA), veri kümesinin boyutunu bilgi kaybı olmaksızın azaltmak için özellikleri birbirine dik yeni düzlemlere yansıtır. Bu süreç, veri kümesinin yorumlanabilirliğini ve boyut indirgeme avantajlarını artırır [24]. Autoencoder ise temel olarak bir kodlayıcı ve bir dekoderden oluşan özel bir sinir ağı modelidir [25]. Kodlayıcı, giriş verisini düşük boyutlu, yoğun bir gizli temsile dönüştürmekten sorumludur; dekoder ise bu sıkıştırılmış temsilden orijinal veriyi geri yüklemeye çalışır.

### 3.3. Makine Öğrenimi Algoritmaları

Makine öğrenimi modelleri ve algoritmaları oldukça geniş bir alanı kapsar, bu nedenle bir makalede tüm detaylarıyla ele almak imkansızdır. Bu bölüm, temel denetimli

modellere ve algoritmalarına odaklanarak bu iş akışı alt modülünü temsil etmeyi amaçlar ve diğerleriyle entegrasyonu sağlar. Denetimli öğrenme, hedef değişken değerleri regresyon (sürekli aralık hedef değerleri) veya sınıflandırma problemleri (önceden belirlenmiş belirli sınıfları gösteren hedef değerleri) varsa mümkündür. Makine öğrenimi modelleri parametrik olarak nitelendirilirse, belirli bir işlev formuna ve parametrelere sahip olurlar ve bu parametrelerin değerleri bir veri kümesiyle belirlenebilir [26].

Makine öğrenimi algoritmaları, veri analizi ve model oluşturma süreçlerinde kullanılan matematiksel yöntemlerdir. Bu yöntemler, verilerdeki desenleri tanımlamak, öğrenmek ve tahmin yapmak için kullanılır. Makine öğrenimi algoritmaları genellikle aşağıdaki kategorilere ayrılır:

1. Denetimli Öğrenme Algoritmaları: girdi verileri ile birlikte doğru çıkışları içeren etiketli veri kümelerini kullanır. Denetimli öğrenme altında yaygın olarak kullanılan algoritmalar şunlardır:
   Destek Vektör Makineleri (SVM), K-En Yakın Komşu (KNN), Karar Ağaçları (DT), Rastgele orman (RF), eXtreme Gradient Boosting (XGBoost), Yapay Sinir Ağları (ANN)
2. Denetimsiz Öğrenme Algoritmaları: etiketlenmemiş veri kümelerinden desenler çıkarır ve verilerdeki yapıları keşfeder. Denetimsiz öğrenme altında yaygın olarak kullanılan algoritmalar şunlardır:

Principal Component Analysis (PCA), Autoencoder

### 3.4. Destek Vektör Makineleri (SVM)

Corinna Cortes ve Vapnik, 1995 yılında SVM (Destek Vektör Makineleri) modelini oluşturdular. Bu model, nonlineer ve yüksek boyutlu desen tanıma için bir dizi özel avantaj sunar. Küçük örneklerin çözümü için de kullanılabilir. SVM, diğer makine öğrenimi problemleri üzerinde genişletilebilir. Girdi vektörlerinin noktalarını bölmek için hiperdüzlemi kullanarak gerekli katsayıları bulur. En büyük marjı olan, yani hiperdüzlem ile en yakın girdi nesneleri arasındaki mesafe, en iyi hiperdüzlemdir. Destek vektörleri, hiperdüzlem tarafından tanımlanan girdi noktalarıdır [27].

### 3.5. K-En Yakın Komşu (KNN)

Bu teknik, sınıflandırma ve regresyon işlemlerinde kullanılır. K-en yakın komşular (KNN) algoritması, mevcut tüm verileri toplamak ve depolamak ilkesine dayanır, ardından yeni bir veri noktasını sınıflandırmak için k-en yakın komşularının çoğunluk oylamasını kullanır. Bu, en basit algoritmalardan biridir. Bir veri noktası için tahmin edilen sınıf, k-en yakın komşuların arasında en yaygın olan sınıftır, mesafe fonksiyonuyla ölçülen. Sürekli fonksiyonlar için Öklid mesafesi ve kategorik değişkenler için Hamming mesafesi gibi mesafe fonksiyonları kullanılabilir. K = 1 durumunda, en yakın komşusuna atanmış bir sınıf vardır. KNN modellemesi sırasında bazen k değerinin belirlenmesi zorlayıcı bir problemdir [28].

### 3.6. Karar Ağaçları (DT)

Karar ağacı algoritmaları, makine öğreniminde yaygın olarak kullanılan ve hiyerarşik yapıları ve karar verme süreçlerini açık bir şekilde tasvir eden algoritmalardır. Bu algoritmalar, değerli içgörülerin çıkarılmasını kolaylaştırır ve karar süreci boyunca derinlemesine analiz sağlar. Karar ağaçlarının yorumlanabilirliği, çeşitli veri kümelerinin esnek bir şekilde işlenmesiyle birleştiğinde, farklı alanlarda geniş çapta benimsenmelerinin nedenidir [29].

### 3.7. Rastgele Orman (RF)

Rastgele Orman Sınıflandırıcısı, bir ensemble öğrenme metodolojisi olarak, sağlamlığı ve çok yönlülüğü nedeniyle makine öğrenimi alanında önemli bir ilgi toplamıştır. Bu sınıflandırıcı, karar ağacı birleştirme prensibine dayanır. Birden çok karar ağacının toplu çıktısı, tahmin doğruluğunu ve istikrarını artırmak için kullanılır. Tekil yöntemlerden ensemble metodolojilere olan bu paradigm değişikliği, makine öğrenimi tekniklerinde önemli bir gelişmeyi işaret eder [30].

### 3.8. Extreme Gradient Boosting (XGBoost)

XGBoost, sınıflandırma ve regresyon tahmin problemlerini modellemek için kullanılan bir makine öğrenimi algoritması sınıfıdır. Ayrıca, yüksek performansı ve olağanüstü tahmin yetenekleri ile bilinen bir ağaç tabanlı ensemble modelidir. XGBoost'un gradyan artırma yöntemi, önceki model hatalarından yeni modeller oluşturur ve kalan modelleri final tahminler yapmak için ekler. Bu yöntem, bir sonraki ağacın önceki ağaca bağlı olduğu bir koleksiyon karar ağacından oluşan bir artırma yöntemidir. Artırma yöntemi kullanılarak, model ardışık olarak karar ağaçlarını eğitir ve karmaşık veri desenlerini yakalamak için zayıf ağaçları birleştirerek tek bir kararlı ve sağlam ağaç oluşturur. Eğitim prosedürünün her aşamasında, bir sonraki basit ağaçtan kalan tahmin artıklarını telafi ederek kayıp fonksiyonunu azaltmak için yeni bir ağaç oluşturulur [31].

### 3.9. Yapay Sinir Ağları (ANN)

Bir Yapay Sinir Ağı (YSA), insan beyninin davranışını taklit eden soyut bir hesaplama yaklaşımıdır [32]. Her bir gizli katman için, her bir nöron, giriş sinyali $y_i$'nin ağırlığını hesaplar, ardından çıktı sinyali $u_j$'yi üretmek için bir doğrusal olmayan aktivasyon fonksiyonuna uygulanır. Bu fonksiyonun yapısı:

$$u_j = \sum_{i=0}^{n} W_{ij} y_i \tag{3.1}$$

Arka yayılım algoritması (BP) kullanılarak çok katmanlı ileri beslemeli sinir ağı (MLF), beyanname üzerinde bulunan unsurları tahmin etmek için kullanılan yöntemlerden biridir. Bu yaklaşım, doğrusal olarak ayrılabilir olmayan problemleri temsil etmek için kullanılır. Giriş katmanı, bir dizi gizli katman ve bir nihai çıkış katmanı MLF'yi oluşturur. Her bir katmanı bağlayan ağırlıklar $W_{jk}$ ile $W_{ij}$, her bir nöronun çıktı oluşturmadan önce toplama bir eşik terimi veya bir sapma eklediği yerdir. Düğümün aktivasyon fonksiyonu, bu doğrusal olmayan geçişe verilen bir terimdir. MLF'lerde, gizli ve çıkış katmanlarının genellikle sırasıyla logistik sigmoid (Denklem 2) ve lineer fonksiyonlar (Denklem 3) kullandığı bilinmektedir.[33].

$$f(W) = \frac{1}{1+e^{-W}} \tag{3.2}$$

$$f(x) = x \tag{3.3}$$

Giriş katmanından çıkış katmanına giden giriş x ile temsil edilir, girişin ağırlıklı toplamı ise w ile gösterilir. Çıkış katmanında bir hata hesaplandığında ve bu hata giriş katmanına geriye doğru yayıldığında, buna geriye yayılım (BP) denir [34].

### 3.10. Hiperparametre Optimizasyonu

Makine öğrenimi alanında, "optimizasyon" terimi, hiperparametreleri ince ayarlamak için kullanılır. Bu, düzenleme, çekirdekler ve öğrenme yoğunluğu gibi parametrelerin ayarlanmasını içerir [35]. Hiperparametreler, ML sınıflandırıcılarının doğruluğunu artırmada kritik bir rol oynar ve öğrenme, yapılandırma ve değerlendirme aşamaları boyunca etkiler. "Model keşfi" ve "hiperparametre ayarlama" terimleri, ML sınıflandırıcıları için optimal hiperparametrelerin belirlenme sürecini özetler [36]. Her algoritmanın bir dizi parametre ile geldiği ve bunların çoğunun optimize edilebilir olduğu göz önüne alındığında, bu parametrelerin ayarlanması kritik hale gelir. Bu optimizasyon, daha yüksek bir model puanı elde etmeyi ve sınıflandırıcının genel performansını artırmayı amaçlar.
Makine öğrenimi modellerinin ayarlanması, fonksiyonun kaybını en aza indiren veya doğruluğu en üst düzeye çıkaran optimal değer kombinasyonunu belirlemek için çeşitli hiperparametrelerin ince ayarlanmasını içeren bir optimizasyon zorluğunu örnekler. Hiperparametre optimizasyonu için kullanılan yöntemlerden biri Manuel Arama'dır. Bu yaklaşımda, uygulayıcılar hiperparametreleri seçmek için uzmanlıklarına güvenir, ardından model eğitimi, doğruluk değerlendirmesi ve yinelemeli iyileştirme yapılır. Bu döngüsel süreç, modelin performansı ve doğruluğunun belirtilen gereksinimlerle uyumlu hale gelene kadar devam eder [37].

Rastgele arama yöntemi, önceden belirlenmiş birkaç kombinasyonu rastgele deneyerek, ardından hiperparametreler değerlendirilir ve en iyi sonuçlar alınır [39]. Rastgele arama verimlidir ve büyük boyutlu verileri iyi işleyebilir [40]. Rastgele aramanın işleyişi şu şekildedir:

- Parametre kombinasyonunun iterasyon sayısını başlatma
- Parametrelerin tüm değerlerini başlatma
- İterasyon sayısına dayalı olarak parametre değerlerinin rastgele kombinasyonlarını döndürme
- Eğitim verilerinde makine öğrenme algoritmalarını kullanarak eğitim yapma
- Elde edilen sınıflandırmaları test verileriyle değerlendirme
- Sınıflandırma sonucundan en iyi değeri ve en iyi parametre değer kombinasyonunu saklama

## 3.11. Model Yapılandırmaları ve Performans Ölçümleri

### 3.11.1. Doğruluk

Sınıflandırma algoritmasının veri kümesindeki sınıfları doğru bir şekilde öngörebilme yeteneği, doğruluk tarafından ima edilir. Bu, gerçek veya teorik değer ile beklenen değer arasındaki örtüşmenin ölçümüdür [41]. Doğruluk genellikle doğru tahminlerin toplam oluşum sayısına oranı olarak ifade edilir. Denklem 4 doğruluk denklemini gösterir.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{3.4}$$

### 3.11.2. Kesinlik

Her beklenen değerden doğru bir şekilde tahmin edilen değerler, hassasiyeti ölçmek için kullanılır [42]. Sınıflandırıcıların hassasiyeti, negatif bir örneği pozitif olarak sınıflandıramama yetenekleri ile ölçülür. Çünkü makro ortalaması, her sınıfa aynı ağırlığı atar, bu nedenle çoklu sınıfl sınıflandırması için kullanılır. Denklem 5, makro ortalama hassasiyet denklemini gösterir.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.5}$$

### 3.11.3. Geri Çağırma

Pozitif değerlerin doğru bir şekilde sınıflandırılma oranına "recall" denir. Gerçek pozitiflerin ne kadarının doğru bir şekilde sınıflandırıldığı, recall tarafından yanıtlanır. Denklem 6, hatırlama denklemini gösterir. Modellerin recall değeri, makro ortalama kullanılarak belirlendiğinden, makro ortalama recall, Denklem 6'da gösterilen formül kullanılarak hesaplanır.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3.6}$$

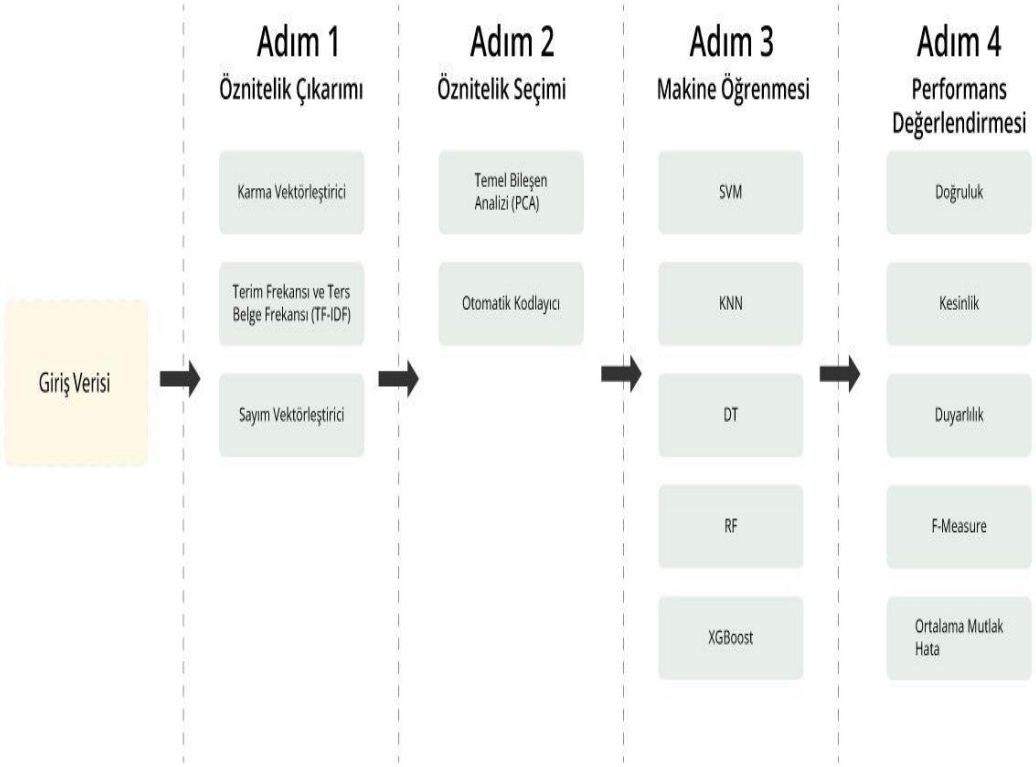
### 3.11.4. F1 Skoru

F-ölçütü, hatırlama ve hassasiyetin harmonik ortalaması olan F1-skoru olarak sıkça adlandırılır [43]. Denklem 7, F1-skoru denklemini gösterir.

$$\text{F1 Score} = \frac{\text{Precision}+\text{Recall}}{2*(\text{Precision}*\text{Recall})} \tag{3.7}$$

### 3.11.5. Ortalama Mutlak Hata

MAE, tahmin edilen ve gerçek değerler arasındaki mutlak farkların test örneği üzerinde ortalama hesaplamasını yapar [44], her fark eşit öneme sahip olarak kabul edilir. Parametreler yi, yp ve n sırasıyla gerçek değer, tahmin edilen değer ve gözlem sayısıdır.




Şekil 1 Veri Analizi ve Model Geliştirme Süreci Akış Şeması

# BÖLÜM 4

# DENEYLER

Deneysel değerlendirmede, elma türlerinin sınıflandırılması üç farklı yolla özellik çıkarımı yapılarak gerçekleştirildi. Tablo 1, Hashing Vectorizer, TF-IDF ve Count Vectorizer yöntemlerine göre çıkarılan özelliklerin destek vektör makinesi (SVM), k-en yakın komşu (Knn), karar ağaçları (DT), rastgele orman (RF), eXtreme Gradient Boosting (XGBoost) makine öğrenimi yöntemlerine göre sonuçlarını göstermektedir. Ayrıca, Tablo 2, PCA ve Autoencoder gibi boyut indirgeme yöntemlerine göre çıkarılan özelliklerin vektör makinesi (SVM), k-en yakın komşu (Knn), karar ağaçları (DT), rastgele orman (RF), eXtreme Gradient Boosting (XGBoost) gibi makine öğrenimi yöntemlerine göre sonuçlarını göstermektedir. Tüm deneyler train_test_split yöntemi ile %20 test, %80 eğitim veri setine göre gerçekleştirildi.

Tablo 1. Farklı Morfolojik Vektörleştirme Yöntemlerine Göre Makine Öğrenmesi Algoritmalarının Performans Karşılaştırması

| Morphological Method | Feature Number | ML Algorithm | Accuracy | MAE | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|---|
| Hashing Vectorizer | 1008 | **KNN** | 90.21% | 0.2689 | 90.09% | 90.21% | 90.00% |
| | | **DT** | 94.97% | 0.1572 | 94.92% | 94.97% | 94.92% |
| | | **SVM** | 80.63% | 0.6091 | 76.62% | 80.63% | 76.34% |
| | | **RF** | 96.15% | 0.1193 | 96.12% | 96.15% | 96.10% |
| | | **XGBoost** | 95.95% | 0.125 | 95.92% | 95.95% | 95.88% |
| Count Vectorizer | 2600 | **KNN** | 91.03% | 0.2418 | 91.26% | 91.03% | 91.00% |
| | | **DT** | 94.92% | 0.1480 | 95.08% | 94.92% | 94.97% |
| | | **SVM** | 80.94% | 0.5568 | 76.83% | 80.94% | 76.73% |
| | | **RF** | 95.74% | 0.1142 | 95.79% | 95.74% | 95.74% |
| | | **XGBoost** | 95.79% | 0.1219 | 95.83% | 95.79% | 95.80% |
| TF-IDF Vectorizer | 2339 | **KNN** | 90.52% | 0.2674 | 90.88% | 90.52% | 90.42% |
| | | **DT** | 95.85% | 0.1265 | 95.83% | 95.85% | 95.83% |
| | | **SVM** | 80.37% | 0.6260 | 76.38% | 80.37% | 76.16% |
| | | **RF** | 96.56% | 0.1096 | 96.55% | 96.56% | 96.54% |
| | | **XGBoost** | 96.26% | 0.1398 | 96.25% | 96.26% | 96.24% |

Tablo 1'de, HashingVectorizer, TF-IDF ve CountVectorizer olarak adlandırılan farklı morfolojik özellik çıkarma yöntemleri, çeşitli makine öğrenimi algoritmaları kullanılarak doğruluk, kesinlik, hatırlama ve F-skoru açısından değerlendirilmiştir. Makine öğrenimi algoritmaları SVM, KNN, DT, XGBoost ve RF performansları açısından değerlendirilmiştir. Tablo 1 incelendiğinde, TF-IDF doğruluk, kesinlik, hatırlama ve F-skoru açısından diğer morfolojik yöntemlerden daha iyi performans gösterdiği gözlemlenebilir. Tablo 1'deki morfolojik özellik çıkarma yöntemleri arasında, rastgele orman (RF) sürekli olarak doğruluk, kesinlik, hatırlama ve F-skoru için en yüksek değerleri elde etmektedir. XGBoost ve DT de iyi performans sergilemektedir, ancak genellikle rastgele ormana kıyasla biraz daha düşük performans göstermektedirler.

TF-IDF özellik çıkarımı için:

- KNN, %90.52 doğruluk, %90.88 kesinlik, %90.52 hatırlama, %90.42 F-skoru ve 0.2674 MAE elde etti.

- DT, %95.85 doğruluk, %95.83 kesinlik, %95.85 hatırlama, %95.83 F-skoru ve 0.1265 MAE elde etti.

- SVM, %80.37 doğruluk, %76.38 kesinlik, %80.37 hatırlama, %76.16 F-skoru ve 0.6260 MAE elde etti.

- RF, %96.56 doğruluk, %95.55 kesinlik, %96.56 hatırlama, %96.54 F-skoru ve 0.1096 MAE elde etti.

- XGBoost, %96.26 doğruluk, %96.25 kesinlik, %96.26 hatırlama, %96.24 F-skoru ve 0.1398 MAE elde etti.

Sonuç olarak, TF-IDF yöntemi kullanıldığında, SVM, KNN, RSS, RF ve XGBoost algoritmaları yüksek performans sergilemektedir. Özellik çıkarma yöntemleri ve algoritmalarının seçimi, istenilen performans metriklerine göre yapılmalıdır.

Tablo 2. Farklı Morfolojik Vektörleştirme Yöntemleri ve Özellik Seçimi Tekniklerine Göre Makine Öğrenmesi Algoritmalarının Performans Karşılaştırması

| Morphological Method | Feature Selection | ML Algorithm | Accuracy | MAE | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|---|
| Hashing | Auto Encoder | **KNN** | 90.52% | 0.2674 | 90.88% | 90.52% | 90.42% |
| Vectorizer | | **DT** | 95.85% | 0.1265 | 95.83% | 95.85% | 95.83% |
| | | **SVM** | 80.37% | 0.6260 | 76.38% | 80.37% | 76.16% |
| | | **RF** | 96.56% | 0.1096 | 96.55% | 96.56% | 96.54% |
| | | **XGBoost** | 96.26% | 0.1398 | 96.25% | 96.26% | 96.24% |
| Hashing | PCA | **KNN** | 83.79% | 0.4822 | 85.12% | 83.79% | 82.17% |
| Vectorizer | | **DT** | 58.10% | 2.3873 | 55.64% | 58.10% | 54.80% |
| | | **SVM** | 68.37% | 0.9762 | 61.89% | 68.37% | 62.17% |
| | | **RF** | 42.29% | 3.3913 | 32.46% | 42.29% | 33.60% |
| | | **XGBoost** | 59.68% | 2.7944 | 51.57% | 59.68% | 54.66% |
| Count | Auto Encoder | **KNN** | 74.07% | 0.9129 | 73.71% | 74.07% | 73.69% |
| Vectorizer | | **DT** | 84.73% | 0.4415 | 84.78% | 84.73% | 84.70% |
| | | **SVM** | 46.77% | 2.5906 | 37.23% | 46.77% | 36.88% |
| | | **RF** | 86.47% | 0.3919 | 86.46% | 86.47% | 86.35% |
| | | **XGBoost** | 86.93% | 0.3831 | 86.98% | 86.93% | 86.87% |
| Count | PCA | **KNN** | 88.78% | 0.3237 | 88.85% | 88.78% | 88.65% |
| Vectorizer | | **DT** | 66.59% | 1.0906 | 66.20% | 66.59% | 64.88% |
| | | **SVM** | 85.14% | 0.4810 | 81.51% | 85.14% | 82.91% |
| | | **RF** | 56.40% | 2.8171 | 54.42% | 56.40% | 48.91% |
| | | **XGBoost** | 74.64% | 1.2873 | 77.51% | 74.64% | 71.11% |
| TF-IDF | Auto Encoder | **KNN** | 73.25% | 0.9492 | 73.21% | 73.25% | 73.01% |
| Vectorizer | | **DT** | 82.83% | 0.4805 | 83.09% | 82.83% | 82.90% |
| | | **SVM** | 45.38% | 2.6444 | 36.26% | 45.38% | 35.71% |
| | | **RF** | 83.81% | 0.4728 | 83.98% | 83.81% | 83.84% |
| | | **XGBoost** | 83.96% | 0.4579 | 84.08% | 83.96% | 83.95% |
| TF-IDF | PCA | **KNN** | 90.11% | 0.3596 | 91.12% | 90.11% | 89.04% |
| Vectorizer | | **DT** | 55.73% | 1.7272 | 56.03% | 55.73% | 52.29% |
| | | **SVM** | 72.33% | 0.9960 | 66.05% | 72.33% | 66.92% |
| | | **RF** | 33.20% | 3.8181 | 26.81% | 33.20% | 26.38% |
| | | **XGBoost** | 65.61% | 1.8577 | 61.69% | 65.61% | 62.18% |

Tablo 2, farklı morfolojik yöntem kombinasyonları, özellik seçme teknikleri ve makine öğrenimi algoritmaları hakkında bilgi sağlamaktadır. Tablo 2'de, Hashing Vectorizer, TF-IDF ve Count Vectorizer istatistiksel özellikleri çıkarılmıştır. Elde edilen sayısal özelliklere iki farklı özellik seçme yöntemi uygulanmıştır. Son aşamada, en ideal özellikler üzerinde makine öğrenimi yöntemleriyle doğruluk, kesinlik, duyarlılık ve F-skoru açısından değerlendirilmiştir. Tablo 2 incelendiğinde, TF-IDF yönteminin, PCA özellik seçme yöntemi ile diğer yöntemlerden daha iyi performans gösterdiği görülmektedir. Şimdi, bazı spesifik değerleri yorumlayalım:

- "TF-IDF" satırında, PCA özellik seçme yöntemi kullanılarak, KNN algoritması %90.11 doğruluk, %91.12 kesinlik, %90.11 hatırlama ve %89.04 F-skoru elde etmiştir.

- "HashingVectorizer" satırında, PCA özellik seçme yöntemi kullanılarak, KNN algoritması %83.79 doğruluk, %85.12 kesinlik, %83.79 hatırlama ve %82.17 F-skoru elde etmiştir.

- "CountVectorizer" satırında, PCA özellik seçme yöntemi kullanılarak, KNN algoritması %88.78 doğruluk, %88.85 kesinlik, %88.78 hatırlama ve %88.65 F-skoru elde etmiştir.

Tablo 3. Farklı Morfolojik Vektörleştirme Yöntemlerine Göre Makine Öğrenmesi Algoritmalarının Hiperparametre Ayarları

| **Morphological Method** | **KNN** | **DT** | **SVM** | **RF** | **XGBoost** |
|---|---|---|---|---|---|
| **TF-IDF Vectorizer** | {'n_neighbors': 9, 'metric': 'manhattan'} | {'min_samples_leaf': 1, 'max_depth': 20, 'criterion': 'entropy'} | {'gamma': 0.0007, 'C': 46} | {'n_estimators': 145, 'max_leaf_nodes': 50, 'max_depth': 48} | {'subsample': 0.8, 'max_depth': 6} |
| **Count Vectorizer** | {'n_neighbors': 5, 'metric': 'manhattan'} | {'min_samples_leaf': 1, 'max_depth': 18, 'criterion': 'gini'} | {'gamma': 0.0006, 'C': 50} | {'n_estimators': 145, 'max_leaf_nodes': 49, 'max_depth': 48} | {'subsample': 0.6, 'max_depth': 8} |
| **Hashing Vectorizer** | {'n_neighbors': 1, 'metric': 'euclidean'} | {'min_samples_leaf': 1, 'max_depth': 17, 'criterion': 'entropy'} | {'gamma': 0.0009, 'C': 43} | {'n_estimators': 145, 'max_leaf_nodes': 49, 'max_depth': 47} | {'subsample': 0.8, 'max_depth': 7} |

Tablo 3'te, Hashing Vectorizer, TF-IDF ve Count Vectorizer gibi farklı morfolojik özellik çıkarma yöntemleri kullanılarak elde edilen verilerle, çeşitli makine öğrenimi algoritmalarının hiperparametreleri hesaplanmıştır. Bu yöntemler, algoritmaların performansını optimize etmek için kritik öneme sahip olan parametrelerin en uygun değerlerini belirlemek amacıyla kullanılmaktadır. Hiperparametrelerin belirlenmesinde RandomizedSearchCV kullanılmıştır. Bu yöntem, geniş parametre uzayında etkili bir arama yaparak optimal hiperparametre kombinasyonlarını bulmayı sağlar.

Tablo 4. Farklı Morfolojik Vektörleştirme Yöntemleri ve Özellik Seçimi Tekniklerine Göre Makine Öğrenmesi Algoritmalarının Hiperparametre Ayarları

| Morphological Method | Feature Selection | KNN | DT | SVM | RF | XGBoost |
|---|---|---|---|---|---|---|
| **TF-IDF Vectorizer** | Auto Encoder | {'n_neighbors': 1, 'metric': 'euclidean'} | {'min_samples_leaf': 1, 'max_depth': 20, 'criterion': 'entropy'} | {'gamma': 0.0009, 'C': 47} | {'n_estimators': 140, 'max_leaf_nodes': 48, 'max_depth': 40} | {'subsample': 0.6, 'max_depth': 9} |
| **TF-IDF Vectorizer** | PCA | {'n_neighbors': 3, 'metric': 'manhattan'} | {'min_samples_leaf': 4, 'max_depth': 19, 'criterion': 'gini'} | {'gamma': 0.0006, 'C': 44} | {'n_estimators': 130, 'max_leaf_nodes': 50, 'max_depth': 43} | {'subsample': 0.8, 'max_depth': 2} |
| **Count Vectorizer** | Auto Encoder | {'n_neighbors': 2, 'metric': 'euclidean'} | {'min_samples_leaf': 3, 'max_depth': 16, 'criterion': 'entropy'} | {'gamma': 0.0009, 'C': 46} | {'n_estimators': 135, 'max_leaf_nodes': 50, 'max_depth': 47} | {'subsample': 0.8, 'max_depth': 8} |
| **Count Vectorizer** | PCA | {'n_neighbors': 5, 'metric': 'manhattan'} | {'min_samples_leaf': 3, 'max_depth': 18, 'criterion': 'entropy'} | {'gamma': 0.0006, 'C': 47} | {'n_estimators': 140, 'max_leaf_nodes': 46, 'max_depth': 47} | {'subsample': 1, 'max_depth': 8} |
| **Hashing Vectorizer** | Auto Encoder | {'n_neighbors': 1, 'metric': 'minkowski'} | {'min_samples_leaf': 1, 'max_depth': 10, 'criterion': 'gini'} | {'gamma': 0.0006, 'C': 47} | {'n_estimators': 145, 'max_leaf_nodes': 50, 'max_depth': 49} | {'subsample': 1, 'max_depth': 3} |
| **Hashing Vectorizer** | PCA | {'n_neighbors': 3, 'metric': 'euclidean'} | {'min_samples_leaf': 3, 'max_depth': 18, 'criterion': 'entropy'} | {'gamma': 0.0006, 'C': 41} | {'n_estimators': 120, 'max_leaf_nodes': 43, 'max_depth': 41} | {'subsample': 1, 'max_depth': 6} |

Tablo 4, Hashing Vectorizer, TF-IDF ve Count Vectorizer kullanılarak istatistiksel özelliklerin çıkarılmasını detaylandırmaktadır. Elde edilen sayısal özelliklere, özellik seçimi için Auto Encoder ve PCA yöntemleri uygulanmıştır. Bu yöntemlerle, çeşitli makine öğrenimi algoritmalarının performansını optimize etmek amacıyla hiperparametreler hesaplanmıştır. Hiperparametrelerin belirlenmesinde RandomizedSearchCV yöntemi kullanılmıştır. Bu geniş parametre aralıklarında etkin bir arama yaparak en uygun hiperparametre kombinasyonlarını tespit etmeyi sağlar. Bu süreç, algoritmaların doğruluğunu ve genel performansını artırmak için kritik öneme sahiptir.

Tablo 5. Hiperparametre Ayarları Uygulanarak Farklı Morfolojik Vektörleştirme Yöntemlerine Göre Makine Öğrenmesi Algoritmalarının Performans Karşılaştırması

| Morphological Method | Feature Number | ML Algorithm | Accuracy | MAE | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|---|
| Hashing | 1008 | **KNN** | 90.88% | 0.2674 | 94.18% | 93.75% | 93.94% |
| Vectorizer | | **DT** | 96.05% | 0.1188 | 96.75% | 96.36% | 96.51% |
| | | **SVM** | 91.39% | 0.3012 | 94.80% | 93.61% | 94.17% |
| | | **RF** | 90.67% | 0.2996 | 96.55% | 89.53% | 90.71% |
| | | **XGBoost** | 96.61% | 0.1162 | 97.04% | 96.89% | 96.95% |
| Count | 2600 | **KNN** | 92.93% | 0.2182 | 94.99% | 95.34% | 95.10% |
| Vectorizer | | **DT** | 94.77% | 0.1659 | 96.50% | 95.38% | 95.83% |
| | | **SVM** | 91.80% | 0.3032 | 95.20% | 94.64% | 94.87% |
| | | **RF** | 83.45% | 0.7464 | 90.46% | 75.30% | 76.20% |
| | | **XGBoost** | 96.36% | 0.1111 | 96.78% | 97.21% | 96.96% |
| TF-IDF | 2339 | **KNN** | 91.80% | 0.2443 | 94.48% | 94.20% | 94.32% |
| Vectorizer | | **DT** | 96.00% | 0.1434 | 96.74% | 95.37% | 95.98% |
| | | **SVM** | 70.95% | 1.8355 | 95.83% | 70.06% | 79.32% |
| | | **RF** | 80.02% | 1.0696 | 89.93% | 69.90% | 69.74% |
| | | **XGBoost** | 96.77% | 0.1024 | 97.49% | 96.72% | 97.08% |

Tablo 5'de, yukarıda çıkarılan hiper parametreler uygulanarak Hashing Vectorizer, TF-IDF ve Count Vectorizer olarak adlandırılan farklı morfolojik özellik çıkarma yöntemleri, çeşitli makine öğrenimi algoritmaları kullanılarak doğruluk, kesinlik, hatırlama ve F-skoru açısından değerlendirilmiştir. Makine öğrenimi algoritmaları KNN, DT, SVM, RF ve XGBoost performansları açısından değerlendirilmiştir. Tablo 5 incelendiğinde, TF-IDF doğruluk, kesinlik, hatırlama ve F-Skoru açısından diğer morfolojik yöntemlerden daha iyi performans gösterdiği gözlemlenebilir. Tablo 5'deki morfolojik özellik çıkarma yöntemleri arasında, XGBoost sürekli olarak doğruluk, kesinlik, hatırlama ve F-skoru için en yüksek değerleri elde etmektedir. KNN ve DT de iyi performans sergilemektedir, ancak genellikle XGBoost kıyasla biraz daha düşük performans göstermektedirler.

TF-IDF özellik çıkarımı için:

- KNN, %91.80 doğruluk, %94.48 kesinlik, %94.20 hatırlama, %94.32 F-skoru ve 0.2443 MAE elde etmiştir.

- DT, %96 doğruluk, %96.74 kesinlik, %95.37 hatırlama, %95.98 F-skoru ve 0.1434 MAE elde etmiştir.

- SVM, %70.95 doğruluk, %95.83 kesinlik, %70.06 hatırlama, %79.32 F-skoru ve 1.8355 MAE elde etmiştir.

- RF, %80.02 doğruluk, %89.93 kesinlik, %69.90 hatırlama, %69.74 F-skoru ve 1.0696 MAE elde etmiştir.

- XGBoost, %96.77 doğruluk, %97.49 kesinlik, %96.72 hatırlama, %97.08 F-skoru ve 0.1024 MAE elde etmiştir.

Sonuç olarak, TF-IDF yöntemi kullanıldığında, DT, KNN ve XGBoost algoritmaları yüksek performans sergilemektedir. Özellik çıkarma yöntemleri ve algoritmalarının seçimi, istenilen performans metriklerine göre yapılmalıdır.

Tablo 6. Hiperparametre Ayarları Uygulanarak Farklı Morfolojik Vektörleştirme Yöntemleri ve Özellik Seçimi Tekniklerine Göre Makine Öğrenmesi Algoritmalarının Performans Karşılaştırması

| Morphological Method | Feature Selection | ML Algorithm | Accuracy | MAE | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|---|
| Hashing | Auto | **KNN** | 69.96% | 1.1106 | 80.93% | 81.46% | 80.20% |
| Vectorizer | Encoder | **DT** | 75.88% | 0.9328 | 81.94% | 84.31% | 82.45% |
| | | **SVM** | 50.59% | 2.2845 | 40.70% | 45.43% | 39.30% |
| | | **RF** | 81.02% | 0.6956 | 86.98% | 86.30% | 86.24% |
| | | **XGBoost** | 82.21% | 0.6442 | 88.87% | 87.50% | 87.73% |
| Hashing | PCA | **KNN** | 86.16% | 0.4743 | 93.04% | 87.54% | 87.86% |
| Vectorizer | | **DT** | 66.00% | 2.1897 | 55.49% | 59.79% | 55.01% |
| | | **SVM** | 57.31% | 3.6442 | 44.04% | 39.55% | 40.10% |
| | | **RF** | 23.71% | 4.1936 | 15.53% | 16.13% | 14.91% |
| | | **XGBoost** | 74.70% | 1.0711 | 64.33% | 64.17% | 61.09% |
| Count | Auto | **KNN** | 82.37% | 0.5809 | 87.25% | 87.93% | 87.48% |
| Vectorizer | Encoder | **DT** | 84.98% | 0.4446 | 90.93% | 91.21% | 91.03% |
| | | **SVM** | 34.27% | 3.7090 | 03.34% | 09.83% | 04.93% |
| | | **RF** | 78.48% | 0.6644 | 85.86% | 83.51% | 84.38% |
| | | **XGBoost** | 87.44% | 0.3627 | 93.85% | 92.25% | 92.96% |
| Count | PCA | **KNN** | 91.95% | 0.2371 | 95.46% | 94.38% | 94.89% |
| Vectorizer | | **DT** | 73.97% | 1.0850 | 50.57% | 61.85% | 54.27% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | **SVM** | 90.31% | 0.2674 | 94.57% | 91.49% | 92.70% |
| | | **RF** | 56.19% | 2.9938 | 49.50% | 37.68% | 37.15% |
| | | **XGBoost** | 71.77% | 1.7197 | 70.43% | 60.32% | 58.64% |
| TF-IDF | Auto | **KNN** | 81.45% | 0.5881 | 86.98% | 86.83% | 86.85% |
| Vectorizer | Encoder | **DT** | 85.29% | 0.3919 | 91.77% | 91.13% | 91.42% |
| | | **SVM** | 58.65% | 1.7832 | 49.67% | 55.59% | 51.92% |
| | | **RF** | 77.30% | 0.6956 | 84.23% | 83.02% | 83.39% |
| | | **XGBoost** | 87.03% | 0.3550 | 93.46% | 92.17% | 92.70% |
| TF-IDF | PCA | **KNN** | 87.35% | 0.5335 | 86.50% | 89.92% | 88.01% |
| Vectorizer | | **DT** | 64.03% | 1.0513 | 52.72% | 59.57% | 52.33% |
| | | **SVM** | 79.44% | 0.7272 | 71.80% | 75.83% | 73.46% |
| | | **RF** | 49.80% | 2.6284 | 39.77% | 32.45% | 30.90% |
| | | **XGBoost** | 76.67% | 1.2134 | 82.50% | 79.03% | 79.54% |

Tablo 6’da, yukarıda çıkarılan hiper parametreler uygulanarak farklı morfolojik yöntem kombinasyonları, özellik seçme teknikleri ve makine öğrenimi algoritmaları hakkında bilgi sağlamaktadır. Tablo 6'da, Hashing Vectorizer, TF-IDF ve Count Vectorizer istatistiksel özellikleri çıkarılmıştır. Elde edilen sayısal özelliklere iki farklı özellik seçme yöntemi uygulanmıştır. Son aşamada, en ideal özellikler üzerinde makine öğrenimi yöntemleriyle doğruluk, kesinlik, duyarlılık ve F-skoru açısından değerlendirilmiştir. Tablo 6 incelendiğinde, Count Vectorizer yönteminin, PCA özellik seçme yöntemi ile diğer yöntemlerden daha iyi performans gösterdiği görülmektedir. Şimdi, bazı spesifik değerleri yorumlayalım:

- "CountVectorizer" satırında, PCA özellik seçme yöntemi kullanılarak, KNN algoritması %91.95 doğruluk, %95.46 kesinlik, %94.38 hatırlama ve %94.89 F-skoru elde etmiştir.

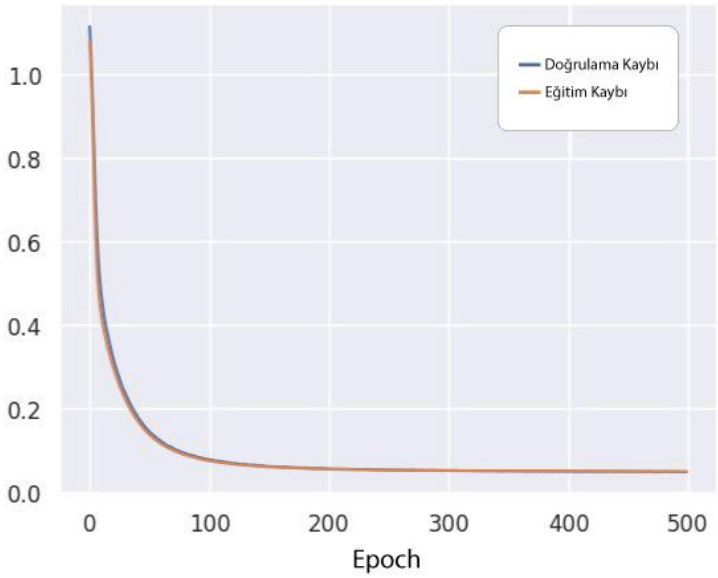
- "HashingVectorizer" satırında, PCA özellik seçme yöntemi kullanılarak, KNN algoritması %86.16 doğruluk, %93.04 kesinlik, %87.54 hatırlama ve %87.96 F-skoru elde etmiştir.

- "TF-IDF" satırında, PCA özellik seçme yöntemi kullanılarak, KNN algoritması %87.35 doğruluk, %86.50 kesinlik, %89.92 hatırlama ve %88.01 F-skoru elde etmiştir.

Şekil 2’de, Yapay Sinir Ağı (YSA) için uygulanan eğitim ve doğrulama kayıplarının grafiği sunulmuştur. Bu grafik, YSA’nın performansını değerlendirmek için önemli bir araçtır. Toplamda 500 eğitim turu gerçekleştirilmiştir ve bu süreç boyunca modelin doğruluk oranı %84.80 olarak belirlenmiştir.

Veri setinin boyutu dikkate alındığında, klasik makine öğrenme yöntemlerinin düşük veri hacimlerinde daha yüksek performans gösterdiği görülmektedir. Ancak, veri setinin büyüklüğü arttıkça, YSA’nın doğruluk oranının da arttığı gözlemlenmiştir. Bu durum, YSA’nın büyük veri setlerini işleme kapasitesinin bir göstergesi olabilir.

Sonuç olarak, YSA’nın performansı ve doğruluk oranı, eğitim turu sayısı ve veri setinin boyutu gibi faktörlere bağlıdır. Bu nedenle, bir YSA modelinin etkinliğini değerlendirmek için bu faktörlerin dikkate alınması önemlidir. Aşağıdaki grafik, bu faktörlerin YSA’nın performansı üzerindeki etkisini göstermektedir. Bu bilgiler, YSA modelinin gelecekteki uygulamaları için değerli içgörüler sağlar.




Şekil 2. Epoch Sayısına Göre Eğitim ve Doğrulama Kaybının Azalışı

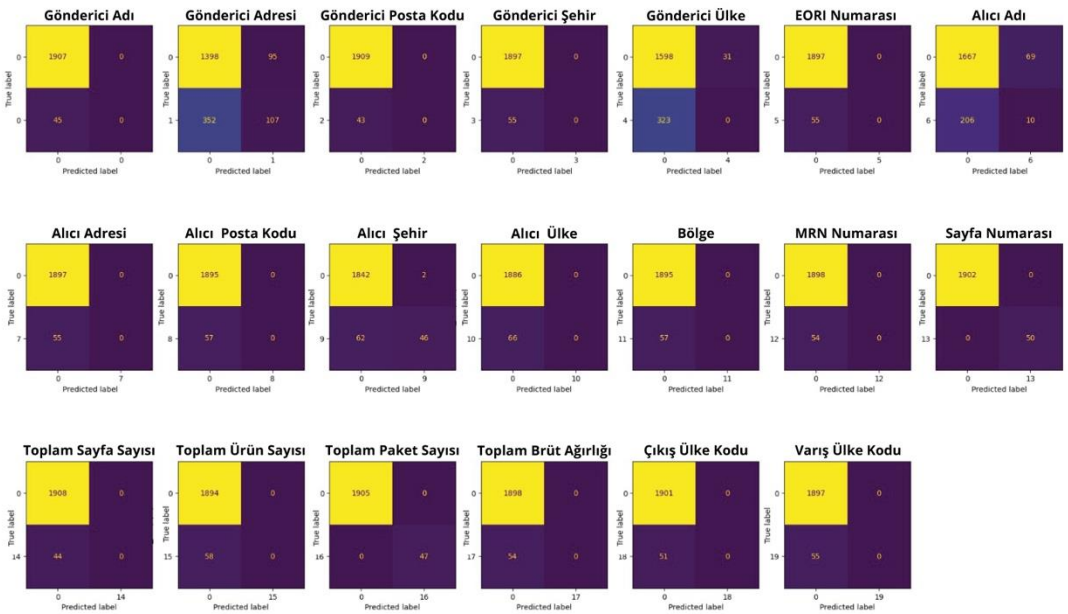
Şekil 3’te, Yapay Sinir Ağı (YSA) için oluşturulan karışıklık matrisi sunulmuştur. Karışıklık matrisi, bir sınıflandırma modelinin performansını değerlendirmek için

yaygın olarak kullanılan bir tekniktir. Bu matris, modelin tahminlerini ve bu tahminlerin gerçek değerlerle ne kadar iyi eşleştiğini gösterir.

YSA modeli için, beyannamedeki unsurların hedef niteliğe ait tahminlerini ve bu tahminlerin gerçek değerlerle karşılaştırmasını içerir. Bu, modelin doğruluğunu, hassasiyetini, hatırlama oranını ve F1 skorunu hesaplama yeteneği sağlar. Bu metrikler, modelin genel performansını değerlendirmek için önemlidir.

Karışıklık matrisi, modelin sınıflandırma hatalarını belirlemek için de kullanılabilir. Örneğin, bir sınıfın örneklerini yanlışlıkla başka bir sınıfa atayan modelin hatalarını belirlemek mümkündür. Bu bilgiler, modelin performansını iyileştirmek için kullanılabilir.

Sonuç olarak, karışıklık matrisi, YSA modelinin performansını değerlendirmek için önemli bir araçtır. Bu matris, modelin güçlü ve zayıf yönlerini belirlemek ve modelin gelecekteki uygulamalarını iyileştirmek için değerli içgörüler sağlar.




Şekil 3. Karmaşıklık Matrisi Performans Göstergesi

# BÖLÜM 5

# SONUÇ

Gümrük beyannamesinde bulunan unsurları tanımlama süreci, gümrük alanında büyük önem taşımaktadır. Farklı gümrük beyannamesinde bulunan unsurları ayırt etmenin karmaşık görevi, gümrük vergilendirilmesi için kullanılan beyannamelerin hatasız sunulmasına yardımcı olmanın yanı sıra, operasyon maliyetlerini ve süreçlerine de olumlu yönde etkileri vardır. Gümrük işlemleri için hatasız beyanname yazımı, ihracat ve ithalat işlemlerinin ve gümrük vergilendirilmesini sağlamak için çok önemlidir. Ayrıca gümrük beyannamesindeki unsurlarının doğru tanımlanması, yetkin ve yeterli insan kaynağının faydalı kullanıma ve her geçen gün yenilenen mevzuatın sürekli takip edilmesine katkıda bulunur. Beyanname üzerindeki unsurlarının geleneksel tanımlama yöntemlerindeki insan hatası potansiyeli göz önüne alındığında, makine öğrenimi gibi ileri teknolojilerin entegrasyonu, gümrük beyannamesinde bulunan unsurları tanımlama sürecinin doğruluğunu ve verimliliğini artırmak için umut verici bir yaklaşım olarak ortaya çıkmıştır. Bu, ithalat ve ihracat işlemlerinde çeşitli avantajlar sunmaktadır. Son yıllarda, makine öğrenimi algoritmaları, uzman yargısına geleneksel bağımlılığı, ki bu yöntemler emek yoğun, öznel ve insan hatasına açık olabilir, değiştirerek gümrük alanında beyanname üzerindeki unsurları sınıflandırmak için etkili araçlar olarak ortaya çıkmıştır. Sunulan sonuçlar, özellik çıkarma görevine uygulandığında çeşitli sınıflandırma algoritmalarının ve özellik kombinasyonlarının performansı hakkında içgörüler sunmaktadır. Bu, özellik çıkaramı, TF-IDF ve XGBoost yöntemlerini birleştiren bir yaklaşımdır. XGBoost algoritması etkileyici bir doğruluk oranı olan %96.77 ile öne çıkmaktadır. Ayrıca, sırasıyla %97.49 ve %96.72 hassasiyet ve geri çağırma puanları ile dikkat çekici bir F-skoru olan %97.08 göstermektedir.

## KAYNAKLAR


1. Kaya, M., & Doğan, A. (2020). Dış Ticarete Konu Eşyanın Vergilendirilmesinde Gümrük Kıymetinin Rolü, Beyanı ve Kontrolü. Gümrük Ve Ticaret Dergisi, 7(19), 10-24.

2. Gültekin, R. (2023). Gümrük Mevzuatında İspat Rejimi. IBANESS: International Journal of Business, Economics and Social Studies. Erişim tarihi: 9 Temmuz 2024, https://www.ibaness.org/bnejss/2023_09_04/15_Gultekin.pdf

3. Sahinaslan, O., & Sahinaslan, E. (2019, April 2). Cross-object information security: A study on new generation encryption. AIP Conference Proceedings, 2086(1), 030034. https://doi.org/10.1063/1.5099206

4. Koh, L. D. (2020). Blockchain in transport and logistics – paradigms and transitions. International Journal of Production Research, 58(7), 2054-2062. https://doi.org/10.1080/00207543.2020.1736428

5. Semary, N. A., Ahmed, W., Amin, K., Pławiak, P., & Hammad, M. (2024). Enhancing machine learning-based sentiment analysis through feature extraction techniques. PLOS ONE. https://doi.org/10.1371/journal.pone.0294968

6. Tseng, C. H., Tsaur, W. J., & Shen, Y. M. (2024). Classification Tendency Difference Index Model for Feature Selection and Extraction in Wireless Intrusion Detection. Future Internet, 16(1), 25. https://doi.org/10.3390/fi16010025

7. Günerkan, M., Şahinaslan, E., & Şahinaslan, Ö. (2022). Gümrük beyannamesi sürecinde öğrenmeye dayalı algoritmaların etkinliğinin incelenmesi. Acta Infologica

8. Seck, D. A. N. (2024). Building Machine Learning Models for Fraud Detection in Customs Declarations in Senegal. WSEAS Transactions on Information Science and Applications, 21, 208-215.

9. Thao, P. N. P., Uyen, H. H. N., Vinh, H. N. K., & Hien, D. N. (2024). Control data mismatching between customs declare sheet and engineering warehouse system: A case study at semiconductor component factory. International Research Journal of Modernization in Engineering Technology and Science, 6(1). https://doi.org/10.56726/IRJMETS48191

10. Agustina, C. N., Novita, R., Mustakim, & Rozanda, N. E. (2024). The implementation of TF-IDF and Word2Vec on booster vaccine sentiment analysis using Support Vector Machine algorithm. Procedia Computer Science, 234, 156-163. https://doi.org/10.1016/j.procs.2024.02.162

11. Semary, N. A., Ahmed, W., Amin, K., Pławiak, P., & Hammad, M. (2024). Enhancing machine learning-based sentiment analysis through feature extraction techniques. PLOS ONE, 19(2), e0294968.

12. Doo, W., & Kim, H. (2024). Simultaneous deep clustering and feature selection via K-Concrete autoencoder. IEEE Transactions on Knowledge and Data Engineering, 36(6), 2629-2642. https://doi.org/10.1109/TKDE.2023.3323580

13. Natha, P., & Rajeswari, P. R. (2023). Advancing skin cancer prediction: A deep dive into hybrid PCA-autoencoder. International Journal of Intelligent Systems and Applications in Engineering, 12(8s), 442-449. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/4144

14. Avcı, İ., & Koca, M. (2023). Cybersecurity Attack Detection Model, Using Machine Learning Techniques. Acta Polytechnica Hungarica, 20(7), 29-44.

15. Seck, D. A. N. (2024). Building Machine Learning Models for Fraud Detection in Customs Declarations in Senegal. WSEAS Transactions on Information Science and Applications, 21, 208-215.

16. Uluer, A. F., Albayrak, Z., Özalp, A. N., Çakmak, M., & Altunay, H. C. (2022, May). BGP Anomali Tespitinde Hibrit Model Yaklaşımı. In 2022 30th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

17. Aisyah, S. (2024). Loan status prediction using decision tree classifier. Power Elektronik: Jurnal Orang Elektro, 13(1), 68-70.

18. Hidajat, T., & Ismail, A. (2024). Machine learning algorithm in credit scoring to prevent bad debt in cooperatives.

19. Isnayni, B. N., Saputra, N., & Hastono, T. (2024). Sentiment analysis of coffee shop reviews using random forest classifier method. JTH: Journal of Technology and Health, 2(2), 16-27.

20. Djeldjli, H., Benatiallah, D., Tanougast, C., & Benatiallah, A. (2024). Solar radiation forecasting based on ANN, SVM and a novel hybrid FFA-ANN model: A case study of six cities south of Algeria. AIMS Energy, 12(1).

21. Hussein, N., & Zeebaree, S. R. (2024). Performance evaluation of Extra Trees Classifier by using CPU parallel and non-parallel processing. Indonesian Journal of Computer Science, 13(2).

22. Liu, D., Wang, M., & Catlin, A. G. (2024). Detecting anti-Semitic hate speech using Transformer-based large language models. arXiv preprint arXiv:2405.03794.

23. Qarshiev, Z. A., & Sapaeva, D. I. (2024). Ma'lumotlarni qidirish usullaridan foydalan-gan holda veb-sahifalarni tasniflash modelini ishlab chiqish. O'zbekistonda fanlararo innovatsiyalar va ilmiy tadqiqotlar jurnali, 3(29), 345-352.

24. Nwokoma, F., Foreman, J., & Akujuobi, C. M. (2024). Effective data reduction using discriminative feature selection based on principal component analysis. Machine Learning and Knowledge Extraction, 6(2), 789-799.

25. Zhang, L., Wang, J., Chang, R., & Wang, W. (2024). Investigation of the effectiveness of a classification method based on improved DAE feature extraction for hepatitis C prediction. Scientific Reports, 14(1), 9143.

26. Kampezidou, S. I., Tikayat Ray, A., Bhat, A. P., Pinon Fischer, O. J., & Mavris, D. N. (2024). Fundamental Components and Principles of Supervised Machine Learning Workflows with Numerical and Categorical Data. Eng, 5(1), 384-416.

27. Çakmak, M. (2024, April). Machine learning approach for predicting obesity levels. In 2nd International Conference on Scientific and Innovative Studies (pp. 845-852).

28. Sharma, D., Chauhan, U., & Khan, H. (2024). Human Activity Recognition using Extremely Fast Decision Tree Classifier. International Journal of Intelligent Systems and Applications in Engineering, 12(16s), 664-671.

29. Terulun, T. (2024). Transient stability assessment of grid-connected inverters using decision tree classifier.

30. Çakmak, M. (2022). AFCC-r: Adaptive Feedback Congestion Control Algorithm to Avoid Queue Overflow in LTE Networks. Mobile Networks and Applications. (27/5, 2138-2152)

31. Çakmak, M. (2024, March). Classification of apple quality using XGBoost machine learning model. In 4th International Conference on Innovative Academic Studies (ICIAS), 12-13 March 2024, Konya, Turkey.

32. Avcı, İ., & Yıldırım, M. (2021). Görme Engelli Bireyler İçin Derin Öğrenme Tabanlı Nesne Tanıma Modeli. Avrupa Bilim ve Teknoloji Dergisi, (28), 220-227.

33. Djeldjli, H., Benatiallah, D., Tanougast, C., & Benatiallah, A. (2024). Solar radiation forecasting based on ANN, SVM and a novel hybrid FFA-ANN model: A case study of six cities south of Algeria. AIMS Energy, 12(1).

34. Salih, M. M. M., & Çakmak, M. (2022, July). Neural Network Approach For Classification And Detection Of Chest Infection. In 2022 2nd International Conference on Computing and Machine Intelligence (ICMI) (pp. 1-5). IEEE.

35. Claesen, M., Simm, J., Popovic, D., & Moor, B. (2014, September). Hyperparameter tuning in python using optunity. In Proceedings of the international workshop on technical computing for machine learning and mathematical engineering (Vol. 1, p. 3).

36. Khan, F., Kanwal, S., Alamri, S., & Mumtaz, B. (2020). Hyper-parameter optimization of classifiers, using an artificial immune network and its application to software bug prediction. Ieee Access, 8, 20954-20964.

37. Al-Fraihat, D., Sharrab, Y., Al-Ghuwairi, A. R., Alshishani, H., & Algarni, A. (2024). Hyperparameter Optimization for Software Bug Prediction Using Ensemble Learning. IEEE Access.

38. Anggoro, D. A., & Mukti, S. S. (2021). Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure. International Journal of Intelligent Engineering & Systems, 14(6).

39. Putatunda, S., & Rama, K. (2018, November). A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. In Proceedings of the 2018 international conference on signal processing and machine learning (pp. 6-10).

40. Shanmugarajeshwari, V., & Ilayaraja, M. (2024). Intelligent Decision Support for Identifying Chronic Kidney Disease Stages: Machine Learning Algorithms. International Journal of Intelligent Information Technologies (IJIIT), 20(1), 1-22.

41. Avcı, İ., & Alzabaq, A. (2023). A New Respiratory Diseases Detection Model in Chest X-Ray Images Using CNN. Traitement du Signal, 40(1).

42. Altunay, H. C., & Albayrak, Z. (2021). Network intrusion detection approach based on convolutional neural network. Avrupa Bilim ve Teknoloji Dergisi, (26), 22-29.

43. Ozalp, A. N., & Albayrak, Z. (2022). Detecting cyber attacks with high-frequency features using machine learning algorithms. Acta Polytechnica Hungarica.

44. Yurdabakan, İ., & Çakmak, M. (2023). Document detection with machine learning algorithms. In Proceedings of the 9th International Mardin Artuklu Scientific Researches.

# ÖZGEÇMİŞ

Zonguldak Ereğli'de ilköğretim ve ortaöğretimini tamamlayan İsa YURDABAKAN, 2010 yılında Abant İzzet Baysal Üniversitesi Fizik Bölümü'nden mezun oldu. Mezuniyetinin ardından profesyonel kariyerine Dolunay Ar-GE Ltd. Şti.'nde yazılım uzmanı olarak başladı.

Kısa süre sonra, Microsoft A.R. şirketlerinden birinde yazılım geliştirme takım lideri olarak görev yapma fırsatı buldu. Ardından Kardemir şirketinde Bilgi Teknolojileri Departmanı'nda yazılım geliştirme takım lideri olarak görev aldı. İsa YURDABAKAN, kariyerine TOBB bünyesinde yazılım departmanı yöneticisi olarak devam etti. Şu anda Kumport şirketinde yazılım geliştirme bölümü müdürü olarak görev yapmaktadır.

Akademik kariyerini daha da genişletmek isteyen İsa YURDABAKAN, yüksek lisans derecesi almayı kararlaştırdı. 2019 yılında Karabük Üniversitesi Bilgisayar Mühendisliği Bölümü'nde tezsiz yüksek lisans çalışmalarına başladı ve 2021 yılında 3,60 ortalama ile mezun oldu.

2021 yılında, İsa YURDABAKAN, Karabük Üniversitesi Bilgisayar Mühendisliği Bölümü'nde tezli yüksek lisans programına geçiş hakkı kazandı.