**T.C.**
**OSTİM TEKNİK ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**
**YAZILIM MÜHENDİSLİĞİ ANABİLİM DALI**
**YÜKSEK LİSANS PROGRAMI**

# Siber Güvenlikte Yapay Zekâ Tabanlı Tehdit Tespit Sistemleri İçin Güvenilir Veri Hazırlama: Veri Bütünlüğünün Sağlanması[1]

## YÜKSEK LİSANS TEZİ

**HAZIRLAYAN**
**Jeyhun Khammadov**

**TEZ DANIŞMANI**
**Dr. Öğr. Üyesi Maksat ATAGOZIEV**

**ANKARA-2026**

**TEZ KABUL VE ONAY**

Jeyhun Khammadov tarafından hazırlanan "Siber Güvenlikte Yapay Zekâ Tabanlı Tehdit Tespit Sistemleri İçin Güvenilir Veri Hazırlama: Veri Bütünlüğünün Sağlanması" başlıklı bu çalışma, 30/01/2026 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

**Kabul Tarihi: 30 / 01 / 2026**

| | | |
|---|---|---|
| **Jüri Üyesi:** | **Dr. Öğr. Üyesi İbrahim ŞANLIALP**<br>Kırşehir Ahi Evran Üniversitesi | _____ |
| **Jüri Üyesi:** | **Dr. Öğr. Üyesi Yücel TEKİN**<br>OSTİM Teknik Üniversitesi | _____ |
| **Tez Danışmanı:** | **Dr. Öğr. Üyesi Maksat ATAGOZIEV**<br>OSTİM Teknik Üniversitesi | _____ |

**ONAY**

**Jüri tarafından kabul edilen bu çalışmanın Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.**

**30/01/2026**

_____

**Prof. Dr. Halil Rıdvan ÖZ**

Enstitü Müdürü

## BİLDİRİM

Enstitü tarafından onaylanan Yüksek Lisans tezimin tamamını veya herhangi bir kısmını basılı veya dijital biçimde arşivleme ve aşağıda belirtilen koşullar dahilinde erişime açma iznini OSTİM Teknik Üniversitesine verdiğimi bildiririm. Bu izinle, Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak ve gelecekteki çalışmalar (makale, kitap, lisans, patent vb.) için tezimin tamamının veya bir bölümünün kullanım hakları yalnızca bana ait olacaktır.

Tezimin bütünüyle kendi çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Telif hakkı bulunan ve sahiplerinden yazılı izinle kullanılması zorunlu olan kaynakları, yazılı izin alarak kullandığımı ve istenildiğinde izinlerin suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayımlanan "Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge" kapsamında, tezim, aşağıda belirtilen koşullar haricince, YÖK Ulusal Tez Merkezi ve OSTİM Teknik Üniversitesi Açık Erişim Sisteminde erişime açılır.

Enstitü / Fakülte Yönetim Kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir.[1]

Enstitü / Fakülte Yönetim Kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren … ay en fazla … ay ertelenmiştir.[2]

Tezimle ilgili gizlilik kararı verilmiştir.[3][4]

**30/01/2026**

**OSTİM TEKNİK ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ YÜKSEK LİSANS TEZİ**
**ORİJİNALLİK RAPORU**

Tezin Başlığı: Yapay Zekâ Tabanlı Siber Güvenlikte Güvenilir Veri Hazırlama
Öğrencinin Adı, Soyadı: Jeyhun Khammadov
Tez Danışmanın Unvanı/Adı, Soyadı: Dr. Öğr. Üyesi Maksat ATAGOZIEV
Anabilim Dalı: Yazılım Mühendisliği
Programı: Yüksek Lisans
Tarih: 30/ 01 / 2026

Yukarıda başlığı belirtilen Yüksek Lisans tez çalışmama ait Giriş, Ana Bölümler ve Sonuç kısmından oluşan ve toplam 73 sayfadan ibaret olan kısmı, 26 / 01 / 2026 tarihinde tez danışmanım tarafından Turnitin intihal tespit programında incelenmiştir. Orijinallik raporunda aşağıda ifade edilen filtrelemeler uygulanmıştır.

Orijinallik raporuna göre, tezimin benzerlik oranı 4%'dur.

Uygulanan filtrelemeler:

1. Kaynakça (hariç)
2. Alıntılar (hariç)
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları (hariç)

Tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrencinin İmzası:

**Tarih: 30 / 01 / 2026**

**Tez Danışmanı**
**Dr. Öğr. Üyesi Maksat ATAGOZIEV**

## ETİK BEYAN

Bu çalışmanın özgün bir çalışma olduğunu, çalışmanın hazırlık, veri toplama, analiz, bilgilerin sunumu ve diğer tüm aşamalarında bilimsel etik ve kurallara uygun davrandığımı, çalışmada bulunan tüm belge bilgileri akademik etik ve kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı bütün bilgileri ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullanmış olduğum verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterdiğim durumlar dışında tarafımdan kaleme alındığını ve özgün olduğunu, Tez danışmanım Dr. Öğr. Üyesi Maksat ATAGOZIEV danışmanlığında ve tarafımdan üretildiğini ve OSTİM Teknik Üniversitesi Tez Yazım Kılavuzuna uygun olarak yazıldığını beyan ederim.

**Jeyhun Khammadov**

**Tarih:** 30 / 01 / 2026

## TEŞEKKÜR

Bu tezin hazırlanmasında büyük emeği geçen, akademik yolculuğum boyunca bana rehberlik eden, engin bilgi ve deneyimiyle süreci her aşamada daha verimli ve anlamlı hale getiren değerli danışmanım Dr. Öğr. Üyesi Maksat ATAGOZIEV'e en derin saygı ve teşekkürlerimi sunarım. Ayrıca, bu uzun ve zorlu süreçte her zaman yanımda olan, bana inanan, moral ve motivasyon kaynağım olan aileme de içtenlikle teşekkür etmek istiyorum.

**Tarih:** 30 / 01 / 2026

**Jeyhun Khammadov**

# ÖZ


Yazar Adı ve Soyadı : Jeyhun Khammadov

Üniversite : OSTİM Teknik Üniversitesi

Enstitü : Fen Bilimleri Enstitüsü

Program Adı : Yazılım Mühendisliği



Tezin Türü : Yüksek Lisans

Sayfa Sayısı : 73

Tarihi : 2026


**Siber Güvenlikte Yapay Zekâ Tabanlı Tehdit Tespit Sistemleri İçin Güvenilir Veri Hazırlama: Veri Bütünlüğünün Sağlanması**


Bu tez çalışmasında, yapay zekâ (YZ) ve makine öğrenmesi (MÖ) tabanlı tehdit tespit sistemlerinde güvenilir veri hazırlama süreçlerinin model performansı üzerindeki etkileri incelenmiştir. Çalışmanın temel amacı, veri kalitesi ve veri bütünlüğüne dayalı veri hazırlama adımlarının tehdit tespit modellerinin sınıflandırma başarımı üzerindeki etkisini ortaya koymaktır.

Araştırmada Malware Analysis Dataset kullanılmış; veri ön işleme aşamasında veri temizleme, tek-sıcak kodlama (one-hot encoding), veri standardizasyonu ve sınıf dengesizliğini gidermek amacıyla SMOTE tekniği uygulanmıştır. Modelleme sürecinde Lojistik Regresyon, Random Forest, XGBoost, LightGBM, AdaBoost, CatBoost, Destek Vektör Makineleri (SVM), K-En Yakın Komşu (KNN), Decision Tree ve Gaussian Naive Bayes algoritmaları kullanılarak karşılaştırmalı analizler gerçekleştirilmiştir. Ayrıca, temsil öğrenmeye dayalı bir yaklaşım kapsamında autoencoder tabanlı ön eğitim modeli ile birlikte AE(latent)+XGBoost ve AE(latent)+LightGBM hibrit yapıları değerlendirilmiştir.

Deneysel bulgular, özellikle Random Forest, XGBoost ve LightGBM gibi ağaç tabanlı ensemble modellerinin doğruluk, F1-skoru ve AUC ölçütleri açısından dengeli ve yüksek performans sergilediğini göstermektedir. SVM modeli ise yüksek sınıflandırma başarımı sunmasına karşın hesaplama maliyeti açısından daha sınırlı bir performans ortaya koymuştur. Autoencoder tabanlı hibrit modeller, temsil öğrenmenin sınıflandırma performansına katkı sağlayabileceğini göstermiştir.

Sonuç olarak elde edilen bulgular, güvenilir veri hazırlama süreçlerinin YZ tabanlı siber güvenlik tehdit tespit sistemlerinin etkinliğini artırmada önemli bir bileşen olduğunu göstermektedir.

**Anahtar Sözcükler:** güvenilir yapay zekâ, veri hazırlama, siber güvenlik, tehdit tespiti, LightGBM,

## ABSTRACT

Thesis : Jeyhun Khammadov

University : OSTİM Technical University

Institute : Graduate School of Natural and Applied Sciences

Program's Name : Computer Engineering

Thesis Type: : Master

Pages : 77

Year : 2025

**Trustworthy Data Preparation for AI-Driven Threat Detection Systems in Cybersecurity: Ensuring Data Integrity**

This thesis investigates the impact of trustworthy data preparation processes on the performance of artificial intelligence (AI) and machine learning (ML)-based threat detection systems. The primary objective of the study is to examine how data quality and data integrity-oriented preprocessing steps influence the classification performance and generalization capability of AI-driven threat detection models.

The experimental evaluation was conducted using the Malware Analysis Dataset. During the data preprocessing stage, data cleaning, one-hot encoding, data standardization, and class balancing using the SMOTE technique were applied. In the modeling phase, comparative analyses were performed using Logistic Regression, Random Forest, XGBoost, LightGBM, AdaBoost, CatBoost, Support Vector Machines (SVM), K-Nearest Neighbors, Decision Tree, and Gaussian Naive Bayes algorithms. In addition, an autoencoder-based pretraining approach was employed, and hybrid architectures combining latent representations with XGBoost and LightGBM were evaluated to examine the contribution of representation learning to threat detection performance.

The experimental results indicate that tree-based ensemble models, particularly Random Forest, XGBoost, and LightGBM, achieved high and well-balanced performance in terms of accuracy, F1-score, and AUC metrics. AdaBoost and CatBoost models demonstrated stable performance under imbalanced data conditions, especially with respect to recall. The SVM model achieved competitive classification results, although with relatively higher

computational cost. Autoencoder-based hybrid models showed that latent feature representations can contribute to classification performance under certain conditions.

Overall, the findings suggest that trustworthy data preparation constitutes an important component in the development of AI-based cybersecurity threat detection systems by supporting consistent and reliable model performance.

**Keywords:** trustworthy artificial intelligence, data preparation, cybersecurity, threat detection, LightGBM

# İÇİNDEKİLER

# ŞEKİLLER DİZİNİ

# TABLOLAR DİZİNİ

## SİMGELER VE KISALTMALAR

| | |
|---|---|
| AdaBoost | Adaptive Boosting |
| AE | Autoencoder |
| AI | Artificial Intelligence |
| AUC | Area Under the Curve |
| BA | Balanced Accuracy |
| CatBoost | Categorical Boosting |
| CNN | Convolutional Neural Network |
| DT | Decision Tree |
| FN | False Negative |
| FP | False Positive |
| GB | Gradient Boosting |
| GBM | Gradient Boosting Machine |
| IDS | Intrusion Detection System |
| KNN | K-Nearest Neighbors |
| LightGBM | Light Gradient Boosting Machine |
| LR | Logistic Regression |
| LSTM | Long Short-Term Memory |
| MCC | Matthews Correlation Coefficient |
| ML | Machine Learning |
| NB | Naive Bayes |
| NN | Neural Network |
| PR-AUC | Precision–Recall Area Under the Curve |

| | |
|---|---|
| RF | Random Forest |
| ROC | Receiver Operating Characteristic |
| SHAP | SHapley Additive exPlanations |
| SMOTE | Synthetic Minority Over-sampling Technique |
| SQL | Structured Query Language |
| TN | True Negative |
| TP | True Positive |
| VAE | Variational Autoencoder |
| XAI | Explainable Artificial Intelligence |
| XGBoost | Extreme Gradient Boosting |
| YZ | Yapay Zekâ |

# 1. GİRİŞ

Sanayi ve hizmet sektörlerinde yaşanan dijital dönüşüm, kurumlara önemli fırsatlar sunarken siber güvenlik alanında yeni zorlukları da beraberinde getirmiştir. Fidye yazılımları, oltalama saldırıları, Gelişmiş Sürekli Tehditler (APT'ler) ve sıfırıncı gün açıkları gibi tehditler, dijital altyapılara olan bağımlılığın artmasıyla birlikte daha doğrulanması güç ve karmaşık bir yapı kazanmıştır. Geleneksel savunma mekanizmaları çoğunlukla reaktif ve manuel müdahalelere dayandığından, modern saldırı senaryoları karşısında çeşitli sınırlılıklar göstermektedir [1].

Bu bağlamda Yapay Zekâ (YZ) ve özellikle Makine Öğrenmesi (MÖ) tabanlı yaklaşımlar, siber güvenlikte tehdit tespiti süreçlerinde yaygın olarak kullanılmaktadır [2]. YZ destekli sistemler, büyük ölçekli verileri analiz ederek şüpheli örüntüleri belirleyebilmekte ve potansiyel tehditlerin erken aşamada tespit edilmesine olanak tanımaktadır [3]. Bununla birlikte, bu sistemlerin başarımı doğrudan kullanılan verinin kalitesi ve bütünlüğüne bağlıdır. Gürültülü veya dengesiz veri kümeleri, yanlış sınıflandırmalara ve sistem güvenilirliğinin azalmasına yol açabilmektedir [4].

Son yıllarda bu sorunlara yönelik olarak "Güvenilir Yapay Zekâ (Trustworthy AI)" yaklaşımı öne çıkmıştır [5]. Bu yaklaşım, YZ sistemlerinin yalnızca performans açısından değil, aynı zamanda adalet, şeffaflık ve dayanıklılık ilkeleri doğrultusunda değerlendirilmesini amaçlamaktadır [6]. Bu tez çalışması, güvenilir veri hazırlama sürecini ve ön işlenmiş verinin makine öğrenmesi tabanlı siber güvenlik tehdit tespit modelleri üzerindeki etkisini incelemektedir. Çalışmanın amacı, veri temizleme, dönüştürme, dengeleme ve ön eğitim adımlarının model doğruluğu, güvenilirliği ve açıklanabilirliği üzerindeki etkilerini analiz etmektir [7].

## 1.1 Problem Tanımı

YZ destekli siber güvenlik sistemleri, etkinliklerini sınırlayan çok boyutlu zorluklarla karşı karşıyadır. Bunlar genel olarak dört ana başlıkta toplanabilir:

Veri Kalitesi Sorunları: Gürültülü, eksik veya hatalı veriler, modelin doğruluğunu düşürmekte ve yanlış pozitif/negatif sonuçlara yol açmaktadır [8]. Dengesiz veri kümeleri, modellerin sık rastlanan saldırı türlerine aşırı odaklanmasına, nadir ama kritik tehditleri gözden kaçırmasına neden olmaktadır [9].

Adversaryal (Karşıt) Zafiyetler: Saldırganlar, modelin giriş verilerini manipüle ederek yanlış sınıflandırmalara sebep olabilmektedir [10].

Şeffaflık Eksiklikleri: Pek çok YZ modeli "kara kutu" biçiminde çalışmakta; bu da siber güvenlik analistlerinin karar süreçlerini anlamasını ve onlara güvenmesini zorlaştırmaktadır [11].

Etik ve Gizlilik Endişeleri: Veri gizliliği, etik sorumluluk ve yasal düzenlemeler kapsamında otomatik karar sistemlerinin uygunluğu sorgulanmaktadır [12].

Bu zorluklar, veri hazırlama süreçlerinin güvenilirlik, şeffaflık ve etik uyum çerçevesinde yeniden ele alınmasını gerektirmektedir.

### 1.2 Güven ve Güvenilir Yapay Zekâ

Yapay Zekâ'nın siber güvenlik alanında artan rolü, yalnız teknolojik bir dönüşüm değil, aynı zamanda etik ve güvenilirlik temelli bir yeniden yapılanmayı da gerektirmektedir. YZ tabanlı sistemlerin karar alma süreçleri, büyük veri kümeleri ve karmaşık algoritmalara dayanmakta, bu durum sistemlerin sonuçlarını açıklamayı ve doğrulamayı zorlaştırmaktadır [13]. Güvenin zedelenmesi, kullanıcıların YZ sistemlerine olan kabulünü doğrudan etkileyen temel faktörlerden biridir.

Güvenilir Yapay Zekâ (Trustworthy AI) kavramı, bu noktada siber güvenlik bağlamında kritik bir kavramsal çerçeve sunmaktadır. Avrupa Komisyonu'nun güvenilir yapay zekâ yaklaşımına göre güvenilir bir YZ sistemi; yasalara uygun, etik değerlere saygılı ve sağlam teknik yapıya sahip olmalıdır [14]. Güvenilir YZ, yalnızca doğruluk ve performans odaklı değil, aynı zamanda şeffaflık, hesap verebilirlik, adalet ve gizlilik ilkelerini de içermelidir [15].

Siber güvenlik alanında güvenilirlik, sadece saldırıların tespiti ve önlenmesi ile sınırlı değildir; aynı zamanda sistemlerin beklenmeyen durumlarda bile tahmin edilebilir ve istikrarlı bir şekilde çalışmasını gerektirir. Bu bağlamda güvenilir veri hazırlama süreci, modellerin hem dayanıklı hem de etik şekilde öğrenmesini sağlayarak YZ destekli tehdit tespitinde kilit rol oynamaktadır. Yüksek kaliteli, dengesizliği giderilmiş ve gürültüden arındırılmış veri kümeleri, sistemlerin yanlış pozitif ve negatif oranlarını azaltarak daha istikrarlı sonuçlar üretmesini sağlar [16].

Ayrıca güvenilir YZ, açıklanabilirlik kavramını da temel alır. Bir modelin neden belirli bir karara vardığını açıklayabilmesi, siber güvenlik uzmanlarının müdahalelerini kolaylaştırır ve insan denetimiyle makine kararları arasındaki dengeyi güçlendirir [17].

### 1.3 Araştırmanın Amacı ve Hedefleri

Bu çalışmanın temel amacı, YZ destekli siber güvenlik sistemlerinde kullanılan güvenilir veri hazırlama süreçlerini sistematik olarak analiz etmek ve optimize etmek suretiyle, tehdit tespit modellerinin doğruluk, dayanıklılık ve etik uyum düzeylerini artırmaya yönelik uygulanabilir bir çerçeve ortaya koymaktır. Günümüz siber güvenlik ortamında modellerin başarısı yalnızca algoritmaların karmaşıklığına değil, aynı zamanda kullanılan verinin doğruluğuna ve güvenilirliğine de bağlıdır. Bu nedenle çalışmanın odağı, "güvenilir verinin" YZ tabanlı sistemlerde nasıl daha etkin biçimde hazırlanabileceğini ortaya koymaktır. YZ sistemlerinin performansını artırmak için kullanılan verilerin, gürültü, eksiklik ve dengesizlik gibi faktörlerden arındırılması, tehdit tespit süreçlerinin başarısı açısından kritik bir öneme sahiptir. Özellikle dengesiz veri kümeleri, modelin yaygın tehdit türlerine aşırı uyum sağlamasına, nadir fakat tehlikeli saldırı örüntülerini gözden kaçırmasına neden olabilmektedir [16]. Bu bağlamda araştırmanın temel hedefi, veri hazırlama sürecinde kalite, etik uygunluk ve model genelleme kabiliyetini dengeleyen güvenilir bir yöntem önermektir.

Araştırmanın spesifik hedefleri aşağıdaki gibi özetlenebilir: Veri Kalitesinin Değerlendirilmesi: Farklı veri kaynaklarından elde edilen siber güvenlik verilerinin tutarlılık, bütünlük ve güvenilirlik açısından incelenmesi. Veri Dengesizliğinin Giderilmesi: Dengesiz veri kümelerinin, sınıf dengeleme teknikleri (örneğin SMOTE, ADASYN) ile dengelenmesi ve bu işlemin model performansına etkisinin ölçülmesi [16]. Standart veri hazırlama yöntemleri kullanılarak eğitilen modeller ile güvenilir veri hazırlama sürecinden geçirilen veri setleriyle eğitilen modellerin performanslarının doğruluk, kesinlik (precision), F1 skoru ve ROC-AUC metrikleri üzerinden karşılaştırmalı olarak analiz edilmesi [17], [18]. Etik ve Şeffaflık İlkelerinin Uygulanması: Veri hazırlama sürecinde gizlilik, adalet ve açıklanabilirlik prensiplerinin (örneğin Explainable AI yaklaşımı) nasıl uygulanabileceğinin değerlendirilmesi [7],[55].

Güvenilir Veri Hazırlama Çerçevesi (Framework) Geliştirmek:

Araştırmanın temel çıktısı, YZ destekli tehdit tespit sistemlerinde kullanılabilecek modüler ve güvenilir bir veri hazırlama çerçevesinin ortaya konulmasıdır. Bu doğrultuda çalışma,

kuramsal altyapı ile deneysel uygulamaları bir araya getiren bütüncül bir yaklaşımla yürütülmüştür. İlk aşamada literatürde yer alan güncel yöntemler incelenmiş, devamında deneysel süreçte farklı veri ön işleme teknikleri uygulanarak çeşitli makine öğrenmesi ve derin öğrenme modelleri eğitilmiş ve karşılaştırmalı olarak değerlendirilmiştir.

Bu kapsamda Rastgele Orman (Random Forest), Aşırı Gradyan Artırma (Extreme Gradient Boosting, XGBoost) ve Lojistik Regresyon (Logistic Regression) gibi yaygın kullanılan denetimli öğrenme algoritmalarının yanı sıra Uyarlamalı Artırma (Adaptive Boosting, AdaBoost), Gradyan Artırma (Gradient Boosting), Kategorik Artırma (Categorical Boosting, CatBoost) ve Hafif Gradyan Artırma Makinesi (Light Gradient Boosting Machine, LightGBM) gibi ensemble tabanlı yaklaşımlar ile Destek Vektör Makineleri (Support Vector Machines, SVM) ve K-En Yakın Komşu (K-Nearest Neighbors, KNN) gibi farklı öğrenme paradigmalarını temsil eden modeller analize dâhil edilmiştir. Ayrıca, anomali tespiti ve temsil öğrenme amacıyla Otoenkoder (Autoencoder) tabanlı derin öğrenme modelleri kullanılmış; elde edilen gizil temsiller (latent representations) üzerinde AE(gizil)+XGBoost ve AE(gizil)+LightGBM gibi hibrit yaklaşımlar değerlendirilmiştir. Veri setinde gözlemlenen sınıf dengesizliği ise Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Over-sampling Technique, SMOTE) kullanılarak giderilmiştir.

Elde edilen bulgular, güvenilir veri hazırlama stratejilerinin YZ tabanlı tehdit tespit modellerinin sınıflandırma başarımını, genellenebilirliğini ve dengesiz veri koşullarındaki kararlılığını olumlu yönde etkilediğini göstermektedir Uygun veri hazırlama adımlarıyla birlikte kullanıldığında, çoklu model yaklaşımları ve temsil öğrenmeye dayalı yöntemlerin daha dengeli bir performans sergilediği gözlemlenmiştir. Bu bağlamda tez çalışması, siber güvenlik alanında güvenilir yapay zekâ uygulamalarının uygulanabilirliğini veri hazırlama süreci üzerinden ele almakta ve yüksek kaliteli verinin model performansı üzerindeki belirleyici rolünü bütüncül bir perspektifle ortaya koymaktadır. Bu yönüyle çalışma, akademik literatüre katkı sağlamanın yanı sıra uygulamaya dönük siber savunma sistemleri için de yol gösterici nitelik taşımaktadır.

Siber güvenlikte yapay zekâ (YZ) ve makine öğrenmesi (MÖ) temelli sistemlerin giderek yaygınlaşması, tehdit tespit süreçlerinde önemli avantajlar sunmaktadır. Ancak bu sistemlerin etkinliği büyük ölçüde kullanılan verinin doğruluğu, tutarlılığı ve güvenilirliği ile ilişkilidir. Bu çalışma, veri hazırlama sürecinin yalnızca teknik bir ön adım değil, aynı

zamanda sistemin genel güvenilirliğini doğrudan etkileyen kritik bir bileşen olduğunu vurgulamaktadır.

Geleneksel siber güvenlik yaklaşımlarının çoğu, insan müdahalesine dayalı ve reaktif yapılar sergilediğinden güncel ve karmaşık tehditler karşısında yetersiz kalabilmektedir [20]. Buna karşın YZ tabanlı çözümler, büyük ölçekli veri kümelerini analiz edebilme ve zaman içerisinde öğrenme yetenekleri sayesinde bu eksiklikleri azaltabilmektedir. Bununla birlikte, gürültülü, eksik veya dengesiz veri kümeleriyle eğitilen modeller, kullanılan öğrenme yaklaşımından bağımsız olarak hatalı sınıflandırmalar üretebilmekte ve sistem güvenilirliğini olumsuz etkileyebilmektedir. Bu nedenle güvenilir veri hazırlama kavramı, YZ destekli siber güvenlik sistemlerinde güvenin tesis edilmesi açısından temel bir gereklilik olarak öne çıkmaktadır [21].

Bu araştırma, veri hazırlama sürecini etik, teknik ve istatistiksel boyutlarıyla ele alarak literatürdeki önemli bir boşluğu doldurmayı amaçlamakta olup, çalışmanın başlıca katkıları aşağıda özetlenmektedir.

*Literatüre Katkı:* Bu tez, siber güvenlik bağlamında "güvenilir veri hazırlama" kavramını kapsamlı biçimde ele alan çalışmalardan biri olup, veri kalitesi, etik ilkeler ve model güvenilirliği arasındaki ilişkiyi sistematik bir çerçevede incelemektedir.

*Yöntemsel Katkı:* SMOTE tabanlı sınıf dengeleme, veri ölçeklendirme ve anonimleştirme adımlarının birlikte kullanılmasıyla, güvenilir yapay zekâ ilkeleriyle uyumlu ve yeniden uygulanabilir bir veri hazırlama metodolojisi önerilmektedir. Bu yaklaşım, hem akademik çalışmalarda hem de endüstriyel uygulamalarda genişletilebilir niteliktedir [22].

*Uygulamalı Katkı:* Gerçekleştirilen deneysel analizler; Lojistik Regresyon (Logistic Regression), Rastgele Orman (Random Forest), Aşırı Gradyan Artırma (Extreme Gradient Boosting, XGBoost), Gradyan Artırma (Gradient Boosting), Uyarlamalı Artırma (Adaptive Boosting, AdaBoost), Kategorik Artırma (Categorical Boosting, CatBoost), Hafif Gradyan Artırma Makinesi (Light Gradient Boosting Machine, LightGBM), Destek Vektör Makineleri (Support Vector Machines, SVM), K-En Yakın Komşu (K-Nearest Neighbors, KNN), Karar Ağacı (Decision Tree) ve Gauss Naif Bayes (Gaussian Naive Bayes) modellerinin yanı sıra otoenkoder (autoencoder) tabanlı hibrit yaklaşımlar (AE(gizil)+Dondurulmuş Sinir Ağı (Frozen Neural Network), AE(gizil)+LightGBM ve AE(gizil)+XGBoost) kullanılarak gerçekleştirilmiştir. Elde edilen bulgular, güvenilir veri

setleriyle eğitilen modeller arasında performans farklılıkları bulunduğunu ve veri kalitesinin tehdit tespit başarımı üzerinde etkili bir faktör olduğunu göstermektedir.

### 1.4 Tezin Yapısı

**Birinci Bölüm – Giriş:**

Bu bölümde çalışmanın genel çerçevesi sunulmakta; araştırmanın arka planı, ele alınan problem, amaç ve hedefler özetlenmektedir. Siber güvenlikte yapay zekâ destekli tehdit tespitine yönelik çalışmalar bağlamında veri hazırlama süreçlerine ilişkin temel kavramlara değinilmekte; veri kalitesi, veri bütünlüğü ve veri dengesizliğinin model performansı üzerindeki etkileri literatür ışığında ele alınmaktadır. Ayrıca, çalışmanın mevcut literatürle ilişkisi ve tez yapısı kısaca açıklanarak izlenen yönteme genel bir bakış sunulmaktadır.

**İkinci Bölüm – Literatür Taraması:**

Bu bölümde, siber güvenlik, yapay zekâ ve makine öğrenmesi tabanlı tehdit tespit sistemleri üzerine yapılan güncel akademik çalışmalar incelenmiştir. Veri kalitesi, dengesiz veri kümeleri, güvenilirlik ve açıklanabilirlik kavramlarıyla ilgili literatürdeki yaklaşımlar karşılaştırılmış, güvenilir veri hazırlama konusundaki mevcut eksiklikler ortaya konmuştur.

**Üçüncü Bölüm – Yöntem ve Veri Seti:**

Bu bölümde, araştırmada kullanılan yöntemsel yaklaşım ve deneysel süreç özetlenmiştir. Kullanılan veri setinin özellikleri, veri ön işleme adımları, SMOTE yöntemi ile sınıf dengesinin sağlanması ve veri standartlaştırma işlemleri ele alınmıştır. Çalışmada Lojistik Regresyon, K-En Yakın Komşu (KNN), Decision Tree ve Gaussian Naive Bayes gibi temel sınıflandırıcıların yanı sıra Random Forest, XGBoost, Gradient Boosting, AdaBoost, CatBoost ve LightGBM gibi ensemble tabanlı yaklaşımlar kullanılmıştır. Ayrıca Destek Vektör Makineleri (SVM) ve autoencoder tabanlı temsil öğrenme modeli ile geliştirilen hibrit yaklaşımlar da deneysel sürece dâhil edilmiştir.

**Dördüncü Bölüm – Bulgular ve Tartışma:**

Bu bölümde, gerçekleştirilen deneysel analizlere ait bulgular sunulmaktadır. Farklı makine öğrenmesi modellerinin, güvenilir veri hazırlama süreci öncesi ve sonrası performansları karşılaştırmalı olarak değerlendirilmiş; doğruluk (accuracy), kesinlik (precision), duyarlılık

(recall), F1 skoru ve ROC-AUC gibi temel performans ölçütleri ayrıntılı biçimde yorumlanmıştır.

**Beşinci Bölüm – Sonuç ve Öneriler:**

Son bölümde, araştırmadan elde edilen genel sonuçlar sistematik biçimde özetlenmiş; ulaşılan bulguların hem akademik literatüre hem de uygulamalı siber güvenlik çalışmalarına olan katkıları kapsamlı şekilde değerlendirilmiştir. Ayrıca, elde edilen sonuçlar doğrultusunda gelecekte gerçekleştirilebilecek araştırmalara yönelik öneriler sunulmuştur.

# 2. LİTERATÜR TARAMASI

Bu bölümde, yapay zekâ (YZ) ve makine öğrenmesi (MÖ) temelli siber güvenlik yaklaşımları, veri kalitesi, dengesizlik, açıklanabilirlik ve güvenilirlik kavramlarıyla ilişkilendirilerek incelenmiştir.

## 2.1 Yapay Zekâ ve Siber Güvenlik İlişkisi

Yapay zekâ, son yıllarda siber güvenlik alanında tehdit tespiti, saldırı önleme, ağ trafiği analizi ve olay müdahalesi gibi süreçlerde yaygın olarak kullanılmaktadır [13], [14], [17]. YZ tabanlı sistemler, geçmiş saldırı örüntülerinden öğrenerek daha önce gözlemlenmemiş saldırı davranışlarını tespit edebilme kapasitesine sahiptir [23]. Geleneksel imza tabanlı güvenlik sistemleri yalnızca bilinen tehditlere odaklanırken, makine öğrenmesi temelli yaklaşımlar davranışsal analiz yoluyla dinamik tehditlerin belirlenmesine olanak tanımaktadır.

Literatürde yapılan çeşitli çalışmalarda, denetimli öğrenme algoritmalarının çok boyutlu ağ trafiği verilerini analiz ederek sıfırıncı gün saldırılarını tespit etmede yüksek doğruluk oranlarına ulaştığı rapor edilmiştir [13], [17], [30]. Ancak bu tür yapay zekâ tabanlı sistemlerin başarısının büyük ölçüde kullanılan verinin kalitesine ve güvenilirliğine bağlı olduğu vurgulanmaktadır. Gürültülü, eksik veya dengesiz veri kümeleri ile eğitilen modellerin yanlış sınıflandırma oranlarının arttığı ve bu durumun sistem güvenilirliğini olumsuz etkilediği literatürde açıkça belirtilmiştir [18], [19],[72].

## 2.2 Veri Kalitesinin Önemi ve Güvenilirlik Kavramı

Veri kalitesi, yapay zekâ tabanlı sistemlerin performansını doğrudan etkileyen temel unsurlardan biridir. Siber güvenlik uygulamalarında düşük kaliteli veri kümeleri, yanlış tehdit sınıflandırmalarına ve hatalı alarm oranlarının artmasına neden olabilmektedir [26]. Literatürde yapılan çalışmalar, gürültülü, eksik veya dengesiz verilerle eğitilen modellerin önyargılı öğrenme davranışları sergilediğini ve bu durumun özellikle tehdit tespiti gibi kritik alanlarda doğruluk kaybına yol açtığını ortaya koymaktadır [18], [19].

Bu bağlamda veri hazırlama süreci; veri temizleme, gürültü azaltma, eksik değerlerin giderilmesi, özellik seçimi ve veri ölçeklendirme gibi adımları kapsamaktadır. Bu adımlar, öğrenme algoritmalarının hatalı veya tutarsız verilerden etkilenmesini minimize etmeyi

amaçlamaktadır. Nitelikli veri kullanımı, modelin daha kararlı ve genellenebilir öğrenme gerçekleştirmesini sağlayarak sistemin genel güvenilirliğini artırmaktadır [28].

### 2.3 Dengesiz Veri (Imbalanced Data) Sorunu ve Çözüm Yaklaşımları

Siber güvenlik veri kümelerinde sıklıkla karşılaşılan problemlerden biri dengesiz veri dağılımıdır. Zararlı etkinliklerin sayısının normal trafiğe kıyasla düşük olması, modellerin çoğunluk sınıfına aşırı uyum göstermesine ve saldırı örneklerinin gözden kaçmasına neden olabilmektedir [29]. Bu sorunun giderilmesinde yaygın olarak kullanılan yöntemlerden biri SMOTE tekniğidir. SMOTE, azınlık sınıfına ait örneklerden sentetik veriler üreterek sınıf dengesini sağlamayı ve modelin genelleme yeteneğini artırmayı amaçlamaktadır [30]. Ancak hatalı uygulamalar, tekrar eden örnekler nedeniyle aşırı öğrenme riskini artırabilmektedir [31].

Literatürde SMOTE'ye ek olarak ADASYN, Rastgele Azaltılmış Örnekleme ve Küme Tabanlı Örnekleme gibi yöntemler de önerilmektedir. Bu teknikler, veri dengesizliğini azaltmayı ve model performansını iyileştirmeyi hedeflemektedir. Karşılaştırmalı çalışmalar, özellikle SMOTE ve ADASYN yöntemlerinin ağ trafiği tabanlı tehdit tespitinde daha kararlı sonuçlar sunduğunu göstermektedir [32],[38].

### 2.4 Güvenilir Yapay Zekâ (Trustworthy AI) Yaklaşımları

Güvenilir yapay zekâ kavramı, yalnızca teknik doğrulukla değil, aynı zamanda etik, şeffaf, adil ve hesap verebilir karar süreçleriyle tanımlanmaktadır [33].Avrupa Komisyonu'nun 2019 tarihli "Etik İlkelerle Uyumlu Güvenilir Yapay Zekâ Rehberi" raporuna göre, bir yapay zekâ sistemi; insan özerkliğine saygılı, güvenli, şeffaf, hesap verebilir ve açıklanabilir olmalıdır [34]. Siber güvenlik bağlamında güvenilirlik, bir sistemin saldırgan manipülasyonlarına (örneğin, karşıt saldırılar – adversarial attacks) karşı dayanıklı olmasını, kullanıcı gizliliğini korumasını ve karar alma süreçlerinin açıklanabilir olmasını gerektirir.

Kaur ve ark., güvenilir veri hazırlama adımlarının model dayanıklılığını artırdığını ve siber saldırganların modeli yanıltma olasılığını azalttığını belirtmiştir [35]. Ayrıca Açıklanabilir Yapay Zekâ (Explainable AI) yaklaşımı, modelin neden belirli bir sonuca ulaştığını insan tarafından yorumlanabilir biçimde açıklamayı hedefler. Bu, siber güvenlik uzmanlarının model çıktılarının geçerliliğini değerlendirmesini kolaylaştırır ve insan–makine etkileşimini güçlendirir [36].

## 2.5 Literatürdeki Boşluklar

Yapılan literatür incelemesi sonucunda, mevcut araştırmaların önemli katkılarına rağmen aşağıdaki üç temel boşluğun bulunduğu tespit edilmiştir: Veri Kalitesi ile Etik Uyum Arasındaki İlişki Eksikliği: Mevcut çalışmalar veri kalitesine odaklanmakta, ancak etik, adalet ve şeffaflık boyutlarını bütüncül biçimde ele almamaktadır.

Literatürde veri ön işleme yöntemleri; veri temizleme, özellik seçimi, ölçeklendirme ve dengesiz veri problemlerinin giderilmesi gibi başlıklar altında yaygın biçimde incelenmiştir [7], [12], [18]. Ancak bu çalışmaların büyük bir kısmı, söz konusu yöntemleri güvenilir yapay zekâ ilkeleri (şeffaflık, adalet ve hesap verebilirlik) ile bütüncül biçimde ele alan kapsamlı bir veri hazırlama çerçevesi sunmamaktadır [21], [30], [31].

Deneysel Karşılaştırmaların Sınırlılığı: Mevcut araştırmalar çoğunlukla belirli algoritmalara odaklanmakta, güvenilir veri hazırlama sürecinin farklı makine öğrenmesi algoritmaları üzerindeki etkilerini geniş ölçekte karşılaştırmamaktadır [37].

Bu bağlamda, bu tez çalışması güvenilir veri hazırlama kavramını siber güvenlik alanına uyarlayarak hem teorik hem de uygulamalı düzeyde literatüre özgün bir katkı sağlamayı hedeflemektedir. Çalışma, veri kalitesi, etik uygunluk ve model performansı arasındaki dengeyi inceleyerek güvenilir yapay zekâ yaklaşımının temelini güçlendirmektedir.

# 3. VERİ ÖZELLİKLERİ VE OSI KATMANLARI BAĞLAM

## 3.1 Veri Kümesi Tanımı

Bu çalışmada kullanılan veri kümesi, literatürde yaygın olarak kullanılan ve Kaggle platformunda paylaşılan "Malware Analysis Dataset" temel alınarak, denetimli öğrenme yaklaşımlarının siber güvenlik alanında uygulanabilmesi amacıyla etiketlenmiş ağ trafiği örneklerinden oluşmaktadır. Veri kümesi, normal ve kötü amaçlı ağ trafiği örneklerini içermekte olup, izinsiz giriş ve anomali tespitine yönelik farklı saldırı senaryolarını kapsamaktadır [15][75].Veri kümesi toplam 60.938 satır ve 19 özellikten oluşmaktadır. Her bir satır tekil bir ağ trafiği örneğini temsil etmekte, hedef değişken olan "class" sütunu aracılığıyla bu örnekler normal trafik (0) ve kötü amaçlı trafik (1) olacak şekilde etiketlenmektedir. Özellikler arasında kategorik (ör. protocol_type, service) ve sayısal (ör. src_bytes, dst_bytes) değişkenler yer almakta olup, bu yapı farklı makine öğrenmesi modellerinin karşılaştırmalı olarak değerlendirilmesine olanak tanımaktadır.Veri kümesinde sınıflar arasında belirgin bir dengesizlik bulunmaktadır. Örneklerin yaklaşık %98.3'ü normal, %1.7'si ise kötü amaçlı trafiği temsil etmektedir. Ayrıca kötü amaçlı trafik kendi içinde farklı saldırı türlerine ayrılmakta ve bazı saldırı tipleri sınırlı sayıda örnekle temsil edilmektedir. Bu durum, veri dengesizliği problemini ortaya çıkarmakta ve sınıflandırma sürecinde özel dengeleme tekniklerinin kullanılmasını gerekli kılmaktadır [9],[26],[60], [74].

## 3.2 Veri Özellikleri ve OSI Katmanları Bağlamı

Kullanılan veri kümesi, Açık Sistemler Arası Bağlantı modeli (Open Systems Interconnection, OSI) modelinin birden fazla katmanına karşılık gelen özellikler içermektedir. Bu özellikler, ağ trafiği davranışlarının farklı seviyelerde analiz edilmesine imkân tanıyarak, çeşitli siber güvenlik tehditlerinin tespit edilmesini desteklemektedir [15], [64]. Söz konusu özelliklerin seçimi, ağ trafiğinin hem davranışsal hem de yapısal karakteristiklerini bütüncül bir şekilde yansıtmayı hedeflemektedir.

Özelliklerin OSI katmanlarına göre dağılımı, ağ trafiğinin yalnızca tek bir seviyede değil, iletişimin farklı aşamalarında değerlendirilmesini mümkün kılmaktadır. Bu çok katmanlı yapı sayesinde, modeller normal trafik ile saldırı davranışları arasındaki ayrımları daha tutarlı biçimde öğrenebilmektedir [27], [65]. Bu kapsamda, veri kümesinde yer alan

özelliklerin OSI katmanlarına göre sınıflandırılması ve her bir katmanın siber güvenlik açısından taşıdığı önem

**Tablo 3.1.** OSI Katmanlarına Göre Seçilmiş Özelliklerin Açıklaması ve Siber Güvenlikte Önemi

| OSI Katmanı | Özellik | Açıklama | Siber Güvenlikte Önemi |
|---|---|---|---|
| Uygulama | service, logged_in, num_file_creations | Ağ hizmet türünü, kullanıcı oturum durumunu ve dosya oluşturma aktivitelerini tanımlar. | Hizmete özel saldırıları (örn. SQL enjeksiyonu, HTTP flood) ve izinsiz erişim teşebbüslerini tespit eder. |
| Sunum | flag | İletişimin başarılı veya başarısız olduğunu belirten bağlantı durumu bayrağı. | Veri temsilindeki anomalileri ve aktarım güvenilirliğini takip eder. |
| Oturum | duration, num_shells, num_failed_logins | Bağlantı süresi, açılan komut oturumu sayısı ve başarısız oturum açma girişimleri. | Gizli veri sızdırma gibi uzun süreli bağlantıları ve kaba kuvvet saldırılarını saptar. |
| Taşıma | protocol_type | Kullanılan taşıma protokolünü (örn. TCP, UDP) belirtir. | Protokole özgü güvenlik açıklarına yönelik davranışları ayırt eder. |
| Ağ | src_bytes, dst_bytes, wrong_fragment, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_count | Kaynak/hedef baytları, parça bozulmaları ve hata oranları gibi parametreler. | Veri hacmindeki ve hata oranlarındaki anormallikleri tespit ederek DoS saldırılarını belirler. |
| Veri Bağlantı & Fiziksel | land, urgent, root_shell, num_root | Paket düzeyinde belirteçler, anomaliler, root erişimi ve yetki düzeyi bilgileri. | Yetki yükseltme, flood saldırıları ve root düzeyinde güvenlik ihlallerini işaret eder. |

Uygulama katmanına ait özellikler, servise özgü saldırıların ve yetkisiz erişim girişimlerinin belirlenmesine katkı sağlamaktadır. Sunum ve oturum katmanlarına karşılık gelen öznitelikler, bağlantı sürekliliği, başarısız oturum açma girişimleri ve iletişim anomalileri gibi davranışların analiz edilmesine olanak tanımaktadır.

Taşıma ve ağ katmanına ait özellikler, trafik hacmi, hata oranları ve protokol davranışları üzerinden hizmet reddi ve flooding türü saldırıların tespit edilmesini desteklemektedir. Veri bağlantı ve fiziksel katmanlara ilişkin göstergeler ise yetki yükseltme ve anormal paket davranışlarının belirlenmesinde tamamlayıcı rol üstlenmektedir [78].

Bu çok katmanlı özellik yapısı, geliştirilen modellerin farklı saldırı türlerine karşı daha dengeli ve tutarlı kararlar üretebilmesine katkı sağlamaktadır [27], [34],[73], [75]. Ayrıca, özelliklerin sayısal ve kategorik verilerden oluşan bu heterojen yapısı, farklı makine öğrenmesi algoritmalarının karşılaştırmalı olarak değerlendirilmesine uygun bir zemin oluşturmaktadır.

### 3.3 Veri Setinin Kaynağı ve Özgünlüğü

Bu çalışmada kullanılan veri kümesi, gerçek dünya ağ trafiği senaryolarını taklit eden simüle ortamlardan elde edilmiştir ve ağ trafiği tabanlı tehdit tespiti için geniş ölçekte kullanılan benchmark veri kümelerinden biri niteliğindedir. Veri kümesi, farklı iletişim protokolleri ve servisler üzerinden gerçekleşen normal ve kötü amaçlı davranışları bir araya getirmektedir.Veri kümesinde DoS, buffer overflow ve SQL injection gibi yaygın saldırı türlerinin yanı sıra, daha az sıklıkla karşılaşılan saldırı senaryoları da yer almaktadır. Bu çeşitlilik, makine öğrenmesi modellerinin farklı tehdit türlerini tanıyabilmesini ve genelleme yeteneklerinin değerlendirilmesini mümkün kılmaktadır [13], [14], [75][72].Veri kümesinin hacmi ve saldırı dağılımı, gerçek ağ ortamlarında karşılaşılan dengesiz trafik yapısını yansıtmakta ve geliştirilen modellerin pratik senaryolara uyarlanabilirliğini değerlendirmek için uygun bir zemin sunmaktadır [17] [64], [78]. Ayrıca veri, gizlilik esaslarına uygun olarak anonimleştirilmiş ve etiketlenmiş biçimde sağlanmıştır [24]. Özellikle IoT tabanlı ağ yapılarında ve bulut bilişim ortamlarında görülen karmaşık saldırı paternlerinin tespiti, modern IDS yaklaşımlarının temelini oluşturmaktadır [65], [74].

### 3.4 Veri Hazırlama Sürecinde Uygulanan Teknikler

Bu çalışmada uygulanan veri hazırlama teknikleri, sınıflandırma sürecinin tutarlılığını artırmak ve farklı model türlerinin adil biçimde karşılaştırılabilmesini sağlamak amacıyla bütüncül bir yaklaşımla ele alınmıştır. Veri temizleme, özellik dönüştürme, ölçekleme ve sınıf dengeleme adımları, siber güvenlik veri kümelerinde sıkça karşılaşılan gürültü ve dengesizlik problemlerini azaltmayı hedeflemektedir [12], [26], [60]. Bu süreçte, siber güvenlik verilerinde uç değerlerin (outliers) çoğu zaman saldırı davranışlarını temsil edebileceği göz önünde bulundurularak, bu değerlerin veri setinden tamamen çıkarılması yerine korunması tercih edilmiştir. Uygulanan veri hazırlama sürecinin, modeller arasındaki performans farklılıklarının ağırlıklı olarak kullanılan algoritmalardan kaynaklandığının değerlendirilmesine imkân sunduğu; veri temsiline bağlı yan etkilerin azaltılmasına katkı

sağladığı görülmektedir [77], [80]. Özellikle mesafe tabanlı algoritmalar (KNN) ve temsil öğrenme yaklaşımları (Autoencoder) için kritik olan bu ölçekleme işlemi, giriş özelliklerinin tutarlı bir biçimde öğrenilmesini desteklemektedir. Bu yaklaşımın, elde edilen deneysel sonuçların daha güvenilir biçimde yorumlanmasına ve siber güvenlik odaklı modellerin akademik veya yasal çerçevelerle uyumlu bir zeminde test edilmesine olanak tanıdığı ifade edilebilir [20], [30], [67], [79].

### 3.5 Yöntem Akışı

Bu çalışmada izlenen yöntemsel akış, veri kümesinin elde edilmesiyle başlamakta ve ön işleme, modelleme ve değerlendirme adımlarını içeren çok aşamalı bir süreçten oluşmaktadır. Uygulanan bu akış, siber güvenlik tehdit tespitine yönelik makine öğrenmesi modellerinin sistematik ve karşılaştırılabilir biçimde değerlendirilmesini amaçlamaktadır [10],[75].

Yöntem akışına ilişkin genel yapı, çalışmada izlenen adımların bütüncül bir görünümünü sunmak amacıyla Şekil 3.1'de gösterilmekte olup, her bir aşama bir sonraki adımı destekleyecek şekilde yapılandırılmıştır. Bu sistematik yaklaşım, modern IDS literatüründe önerilen şeffaflık ve güvenilirlik ilkeleriyle uyumludur [68], [71]. Ayrıca, kullanılan hibrit ve topluluk öğrenme modellerinin karmaşıklığını yönetmek ve sonuçların doğrulanabilirliğini artırmak amacıyla, süreç içerisine açıklanabilir yapay zekâ (XAI) ve saldırgan (adverseriyal) savunma mekanizmaları entegre edilmiştir [65], [66], [73].

#### Veri Temizleme

Eksik, gürültülü veya ilgisiz verilerin ele alınması yoluyla veri kümesinin bütünlüğünün ve güvenilirliğinin korunması amaçlanmıştır.

**Eksik Değerlerin Ele Alınması:** Sayısal özellikler için ortalama, kategorik özellikler için ise mod kullanılarak eksik değerler doldurulmuştur. Bu yaklaşım ile veri kaybının önlenmesi ve dağılım yapısının korunması hedeflenmiştir.

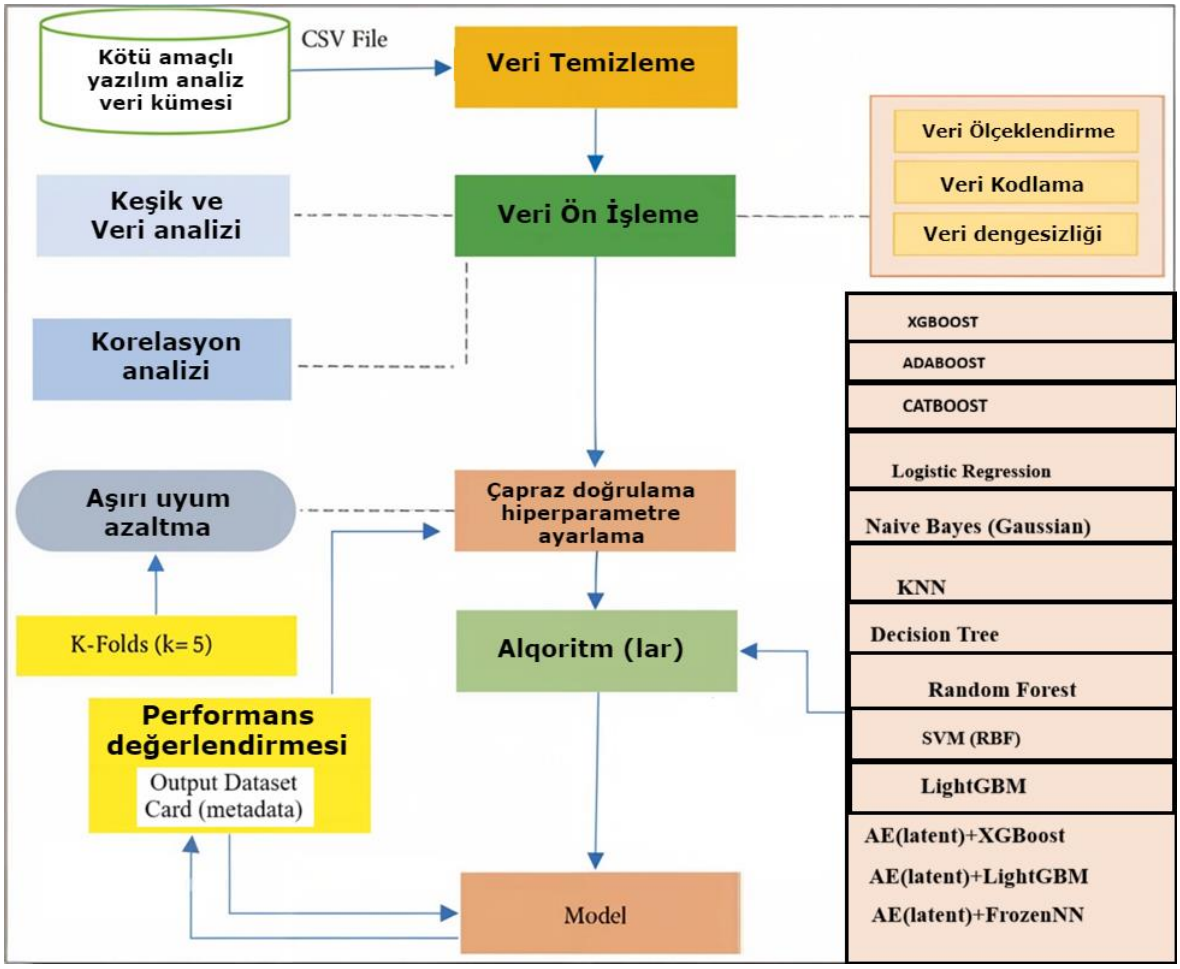
**Çift Kayıtların Kaldırılması:** Veri kümesinde yer alan yinelenen kayıtlar kaldırılarak, olası yanlılıkların ve fazlalıkların azaltılması amaçlanmıştır.

**Aykırı Değer Tespiti:** Aykırı değerler Çeyrekler Arası Aralık (Interquartile Range, IQR) yöntemi [17] kullanılarak belirlenmiş; 1,5×IQR aralığı dışında kalan değerler veri

kümesinden çıkarılmamış veya sınırlandırılmamıştır. Bu sayede nadir ancak bilgi içerebilecek örneklerin korunması gözetilmiştir.

**Önemi:** Gürültü ve fazlalıkların giderilmesi, modellerin daha tutarlı ve ilgili özelliklerden öğrenmesine imkân tanıyarak hatalı veya yanlı tahmin olasılıklarının azaltılmasına katkı sağlamaktadır.

**Keşifsel Veri Analizi (EDA):** Özellikler arasındaki ilişkilerin incelenmesi ve öngörüsel modelleme açısından anlamlı özniteliklerin belirlenmesi amaçlanmıştır [21].




**Şekil 3.1.** Tehdit tespitinde Makine Öğrenimi için Blok Diyagram Tasarımı ve İşleyiş Akışı

**Betimleyici İstatistikler:** Tüm özellikler için dağılım, merkezi eğilim ve yayılım ölçütleri hesaplanarak genel eğilimler değerlendirilmiştir.

**Korelasyon Analizi:** Yüksek korelasyona sahip özellikler tespit edilmiş; gerekli görülen durumlarda bu özellikler dönüştürülmüş veya veri kümesinden çıkarılmıştır.

**Önemi:** Bu adım, yalnızca modele anlamlı katkı sunan özelliklerin kullanılmasını sağlayarak performansın ve yorumlanabilirliğin artırılmasına yardımcı olmaktadır.

**Veri Ön İşleme**: Verinin makine öğrenmesi modelleri tarafından etkin biçimde kullanılabilmesi için ölçekleme, kodlama ve dengeleme işlemlerinin uygulanması amaçlanmıştır.

**Ölçekleme:** Sayısal özellikler StandardScaler yöntemi ile standartlaştırılarak ölçek farklılıklarının model üzerinde yanlılığa yol açması engellenmiştir.[20]

**One-Hot Kodlama:** Kategorik özellikler ikili sütunlara dönüştürülmüş; hedef değişken (class) bu işlem dışında tutulmuştur.

**Sınıf Dengeleme:** Sınıf dengesizliğini gidermek amacıyla SMOTE yöntemi uygulanmış ve azınlık sınıfın temsili artırılmıştır. Ön işleme adımları, verinin dengeli ve öğrenilebilir hâle gelmesini sağlayarak çoğunluk sınıfına yönelimin azaltılmasına katkı sunmaktadır[16]. Bu aşamada, yerel yoğunluk tahminine dayalı gelişmiş aşırı örnekleme tekniklerinin literatürdeki önemi dikkate alınmıştır [60],[38][61].

**Çapraz Doğrulama ve Hiperparametre Ayarı**

Aşırı uyumun önlenmesi ve modellerin genelleme yeteneğinin artırılması amaçlanmıştır.

**Çapraz Doğrulama:** Modeller, 5-katlı çapraz doğrulama yöntemi kullanılarak farklı alt kümeler üzerinde eğitilmiş ve değerlendirilmiştir.

**Hiperparametre Araması:** Model performansını etkileyen temel hiperparametreler RandomizedSearchCV yöntemi kullanılarak belirlenmiştir [35]. Bu yaklaşım, modellerin farklı veri dağılımları altında daha tutarlı sonuçlar üretmesine olanak tanımaktadır.

**Algoritma Seçimi**: Farklı model ailelerinin veri hazırlama sürecine verdikleri tepkilerin karşılaştırmalı olarak incelenmesi amaçlanmıştır.

**XGBoost [43]:** Büyük ve dengesiz veri kümelerinde yüksek doğruluk ve kararlı performans göstermesi nedeniyle değerlendirilmiştir.

**Rastgele Orman (Random Forest) [52]:** Aşırı uyumu azaltan ensemble yapısı ve güçlü genelleme yeteneği sayesinde tercih edilmiştir.

**Lojistik Regresyon [30]:** Basit ve yorumlanabilir yapısı nedeniyle temel karşılaştırma (baseline) modeli olarak kullanılmıştır.

**CatBoost [48]:** Düzenlileştirme mekanizmaları ve kararlı öğrenme yapısı sayesinde özellikle heterojen veri yapıları için değerlendirmeye dâhil edilmiştir [62],[70]. Normalde kategorik verileri işleyebilmesine rağmen, adil bir karşılaştırma için One-Hot kodlanmış veri üzerinde test edilmiştir.

**AdaBoost [54]:** Azınlık sınıfa odaklanan adaptif öğrenme yapısının veri dengesizliği üzerindeki etkisini incelemek amacıyla kullanılmıştır.

**Decision Tree [23]:** Veri ön işleme ve dengeleme adımlarının model davranışı üzerindeki etkisini gözlemlemek için tercih edilmiştir.

**K-En Yakın Komşu (KNN) [33],[42]:** Ölçeklendirmeye duyarlı yapısı nedeniyle veri hazırlama sürecinin etkisini analiz etmek amacıyla değerlendirilmiştir.

**Gaussian Naive Bayes [26],[78]:** Veri dengesizliği karşısındaki sınıflandırma davranışını incelemek amacıyla temel olasılıksal bir yaklaşım olarak kullanılmıştır.

**Destek Vektör Makineleri (SVM) [30]:** Yüksek boyutlu veri uzaylarında etkili karar sınırları oluşturabilmesi nedeniyle karşılaştırmalı analizlere dâhil edilmiştir.

**Light Gradient Boosting Machine (LightGBM) [47]:** Yaprak bazlı büyüme stratejisi sayesinde yüksek hız ve verimlilik sunması nedeniyle değerlendirilmiştir.

**Autoencoder Tabanlı Hibrit Yaklaşımlar [46],[2],[65]:** Temsil öğrenmenin sınıflandırma performansına katkısını incelemek amacıyla hibrit senaryolar kapsamında değerlendirilmiştir.

**Önemi:** Algoritma çeşitliliği, güvenilir veri hazırlama sürecinin farklı model aileleri üzerindeki etkilerinin karşılaştırılmasına imkân tanımaktadır.

**Model Eğitimi ve Performans Değerlendirmesi** Farklı makine öğrenmesi modellerinin kötü amaçlı ağ trafiğini tespit etme performanslarının değerlendirilmesi amaçlanmıştır [17], [18].

**Eğitim:** Ön işlenmiş veri üzerinde birden fazla sınıflandırıcı eğitilmiştir.

**Performans Metrikleri:** Accuracy, Precision, Recall, F1-Skoru ve ROC-AUC metrikleri kullanılmıştır [17], [18]. Sınıflandırma hatalarının analizi için karmaşıklık matrisi (confusion matrix) üzerinden detaylı inceleme yapılmıştır [74], [78].

**Karmaşıklık Matrisi:** Confusion matrix [78] üzerinden sınıflandırma hataları analiz edilmiştir.

**Önemi:** Çoklu metrik yaklaşımı, modellerin farklı saldırı senaryolarına karşı tutarlı performans sergilemesini değerlendirmeye olanak tanımaktadır.

**Not:** Veri setinin %80'i eğitim, %20'si test amacıyla kullanılmıştır.

**Özellik Önemi ve Açıklanabilirlik**

Model tahminlerine her bir özelliğin katkısının incelenmesi ve karar süreçlerinin şeffaflığının değerlendirilmesi amaçlanmıştır [44], [51].

**Özellik Önemi Analizi:** XGBoost ve Rastgele Orman modellerinin yerleşik özellik önem ölçümleri kullanılmıştır.

**SHAP Analizi:** SHAP yöntemi ile bireysel özelliklerin model tahminleri üzerindeki etkileri analiz edilmiştir [51]. Bu analizler, özellikle siber güvenlikte açıklanabilir yapay zekâ (XAI) çerçevelerinin şeffaflık ve güvenilirlik hedefleriyle uyumludur [68], [71]. Ayrıca, model kararlarının hesap verebilirliği, uluslararası hukuk ve yapay zekâ düzenlemeleri bağlamında kritik bir öneme sahiptir [67], [79], [80].

**Önemi:** Bu analizler, model kararlarının daha anlaşılır biçimde yorumlanmasına ve çıktılara duyulan güvenin artırılmasına katkı sunmaktadır.

### 3.6 Ön Eğitim (Pretraining) Modeli

Bu çalışmada, temel sınıflandırma modellerine ek olarak, ağ trafiği verisinden etiket bilgisine ihtiyaç duymadan temsil öğrenmek amacıyla autoencoder tabanlı bir ön eğitim yaklaşımı uygulanmıştır. Bu yaklaşımda amaç, ham özellikleri doğrudan sınıflandırıcılara vermek yerine, veriyi daha düşük boyutlu ve daha kompakt bir gizil temsil (latent representation) uzayında ifade etmektir.

Autoencoder yapısı, encoder ve decoder bileşenlerinden oluşmaktadır. Encoder, girişteki yüksek boyutlu özellik vektörünü daha düşük boyutlu bir gizil vektöre dönüştürürken; decoder bu gizil vektörden yola çıkarak giriş verisini yeniden oluşturmaya çalışmaktadır. Eğitim sürecinde, giriş ile yeniden oluşturulan çıktı arasındaki fark minimize edilerek modelin veri dağılımına ait genel örüntüleri öğrenmesi hedeflenmiştir. Bu aşamada eğitim yalnızca eğitim verisi üzerinde gerçekleştirilmiş, test verisi kullanılmamıştır.

Ön eğitim tamamlandıktan sonra, encoder bileşeni korunmuş ve sınıflandırma aşamasında temsil çıkarıcı (feature extractor) olarak kullanılmıştır. Autoencoder tabanlı temsil öğrenme yaklaşımı, bu çalışmada ayrı bir sınıflandırıcı model olarak değil, temel modelleri genişleten ek bir temsil katmanı olarak değerlendirilmiştir [2], [46].

Bu kapsamda, autoencoder ile elde edilen gizli temsiller üç farklı model yapılandırmasında kullanılmıştır:

AE (Latent Features) + XGBoost: Encoder çıktıları, XGBoost modeline girdi olarak verilmiştir [43].

AE (Latent Features) + LightGBM: Aynı gizli temsiller LightGBM algoritması ile sınıflandırılmıştır [47].

Frozen Encoder + Neural Network Head: Encoder ağırlıkları dondurulmuş ve encoder çıktısı üzerine hafif bir sinir ağı sınıflandırıcı başlığı eklenmiştir; bu yapıda yalnızca sınıflandırıcı katmanların ağırlıkları güncellenmiştir.

Bu üç hibrit yapılandırmanın kullanılması, modern IDS literatüründe önerilen ikiz topluluk (twined ensemble) ve çok katmanlı savunma stratejileriyle paralellik göstermektedir [65], [74]. Ayrıca, bu tür karmaşık modellerin ürettiği sonuçların şeffaflığı, XAI teknikleri ve güncel güvenilir yapay zekâ standartları çerçevesinde değerlendirilmiştir [68], [71], [80]. Böylece autoencoder tabanlı ön eğitimin, farklı sınıflandırma yaklaşımları altında nasıl davrandığı ve özellikle karmaşık saldırı paternlerini tespit etmedeki direnci sistematik biçimde analiz edilmiştir [64], [66], [73].

### 3.7 Yapay Zekâ Algoritmaları

Bu çalışmada, ağ trafiğini normal veya kötü amaçlı olarak sınıflandırmak amacıyla toplam on bir (11) model yapılandırması değerlendirilmiştir. Bu modeller; ham özellikler üzerinde

eğitilen temel sınıflandırma algoritmaları ile autoencoder tabanlı ön eğitim (latent temsil) yaklaşımıyla genişletilmiş hibrit yapılandırmalardan oluşmaktadır.

Ham özellikler kullanılarak eğitilen temel modeller şunlardır: Lojistik Regresyon, Decision Tree, K-Nearest Neighbors, Gaussian Naive Bayes, Random Forest, AdaBoost, CatBoost ve XGBoost. Bu modeller, farklı öğrenme yaklaşımlarının (doğrusal, ağaç tabanlı ve topluluk yöntemleri) performanslarını karşılaştırmalı olarak incelemek amacıyla kullanılmıştır, [54]. Bu geniş algoritma yelpazesi hem geleneksel yaklaşımların hem de modern gradyan artırma (gradient boosting) tekniklerinin siber güvenlik verileri üzerindeki etkinliğini doğrulamayı hedeflemektedir [62], [70], [75].

Buna ek olarak, autoencoder tabanlı ön eğitim ile elde edilen gizil (latent) temsiller kullanılarak üç (3) hibrit model yapılandırması oluşturulmuştur. Bu kapsamda, AE (Latent Features) + XGBoost, AE (Latent Features) + LightGBM ve Frozen Encoder + Neural Network Head yaklaşımları deneysel sürece dahil edilmiştir. Bu yapılandırmalarda autoencoder, ayrı bir sınıflandırıcı olarak değil, temel modelleri destekleyen bir temsil öğrenme katmanı olarak değerlendirilmiştir [46]. Bu tür çok katmanlı ve ikiz topluluk (twined ensemble) yapılandırmaları, karmaşık ağ anomalilerinin tespitinde daha dirençli sonuçlar üretmektedir [64], [65].

Bu yaklaşım sayesinde, ön eğitimli temsil öğrenmenin sınıflandırma performansı üzerindeki etkisi; ham özellikler kullanılarak eğitilen temel modellerle karşılaştırmalı olarak analiz edilmiştir. Böylece, farklı model türlerinin aynı veri hazırlama süreci altında nasıl davrandığı ve özellikle dengesiz veri dağılımları karşısındaki dirençleri sistematik biçimde incelenebilmiştir [74], [78]. Ayrıca, model karmaşıklığının artmasıyla beraber ortaya çıkan açıklanabilirlik ihtiyacı, güncel akademik ve yasal standartlar çerçevesinde ele alınmıştır [68], [71], [80].

### 3.8 Değerlendirme Ölçütleri [79]

Bu çalışmada model performansı, sınıflandırma modellerinin normal ve kötü amaçlı ağ trafiğini ayırt etme başarısını kapsamlı biçimde değerlendirebilmek amacıyla birden fazla ölçüt kullanılarak incelenmiştir. Veri setinin dengesiz yapısı göz önünde bulundurulduğunda, yalnızca doğruluk (accuracy) metriğine dayanmak yeterli olmayabileceğinden, sınıf bazlı performansı yansıtan ölçütlere de yer verilmiştir [17], [18].

Bu kapsamda, Accuracy, Precision, Recall, F1 skoru ve ROC-AUC metrikleri kullanılmıştır [17], [18], [35]. Değerlendirmede temel alınan karışıklık matrisi bileşenleri Şekil 3.2'de gösterilmektedir.

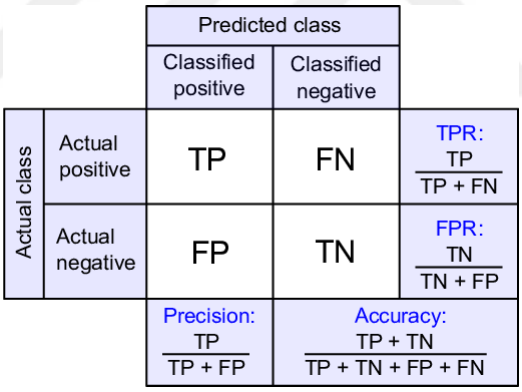
Burada,

TP = True Positives (Doğru Pozitifler, modelin kötü amaçlı olarak doğru sınıflandırdığı örnekler),

TN = True Negatives (Doğru Negatifler, modelin normal olarak doğru sınıflandırdığı örnekler),

FP = False Positives (Yanlış Pozitifler, modelin normal olduğu halde yanlışlıkla kötü amaçlı olarak sınıflandırdığı örnekler),

FN = False Negatives (Yanlış Negatifler, modelin kötü amaçlı olduğu halde yanlışlıkla normal olarak sınıflandırdığı örnekler).




**Şekil 3.2.** İkili Sınıflandırma İçin Karışıklık Matrisi (TP, TN, FP, FN)

*Doğruluk (Accuracy):* Modelin tüm tahminler içinde doğru sınıflandırdığı örneklerin oranını ifade eder. Matematiksel olarak, doğruluk, doğru pozitifler ile doğru negatiflerin toplamının, tüm tahminlerin toplamına oranıdır [17].

Doğruluk, modelin genel performansı hakkında bir fikir verse de, dengesiz veri setleri için tek başına güvenilir bir ölçüt olmayabilir. Zira doğruluk, çoğunluk sınıfın (normal trafik) lehine sonuçlar vererek azınlık sınıfın (kötü amaçlı trafik) tespitindeki başarısızlığı

gizleyebilir. Bu nedenle, doğruluk metriğinin yanı sıra kesinlik (precision) ve duyarlılık (recall) gibi metriklerin de kullanılması elzemdir.

*Kesinlik ve Duyarlılık (Precision ve Recall)*

*Kesinlik (Precision):* Kötü amaçlı olarak tahmin edilen örneklerin ne kadarının gerçekten kötü amaçlı olduğunu ölçer. Yani modelin pozitif olarak sınıflandırdığı tüm örnekler içinde, doğru pozitiflerin oranıdır.

Kesinlik değeri yüksek olduğunda, model yanlış alarm (false positive) üretme konusunda başarılı demektir; model kötü amaçlı dediğinde bunun büyük olasılıkla doğru olduğunu gösterir. Ancak kesinlik, modelin tüm saldırıları yakalayıp yakalayamadığını tek başına göstermez.

*Duyarlılık (Recall):* Gerçekten kötü amaçlı olan örneklerin ne kadarının model tarafından doğru şekilde tespit edildiğini gösterir. Diğer bir deyişle, tüm gerçek pozitifler içinde, modelin doğru pozitif olarak yakaladığı oranıdır[18].

Duyarlılık, özellikle siber güvenlik uygulamalarında kritik öneme sahiptir; çünkü kötü niyetli bir saldırının tespit edilememesi (yanlış negatif) ciddi sonuçlar doğurabilir. Duyarlılığın yüksek olması, modelin saldırıları kaçırmadan yakaladığını garanti eder, ancak bu durumda yanlış alarm sayısı artabilir.

Kesinlik ve Duyarlılık genellikle birbiriyle ters orantılıdır; birini artırmak diğerini düşürebilir. Bu yüzden bu iki metriğin dengeli bir şekilde değerlendirilmesi, model performansının anlaşılması için önemlidir. Örneğin, bir modelin duyarlılığı yüksek fakat kesinliği düşükse, model saldırıların büyük kısmını yakalıyor ancak çok sayıda yanlış alarm da üretiyor demektir.

*F1-Skoru*, kesinlik (precision) ve duyarlılık (recall) ölçütlerinin harmonik ortalamasını temsil eden bir performans metriğidir. Bu ölçüt, sınıflandırma modelinin hem yanlış pozitif (FP) hem de yanlış negatif (FN) hatalarını eş zamanlı olarak dikkate alarak dengeli bir değerlendirme sunmaktadır. Özellikle sınıf dağılımının dengesiz olduğu veri kümelerinde, tek bir performans ölçütüne dayalı değerlendirmelerin yetersiz kalması nedeniyle yaygın olarak tercih edilmektedir [17], [35].

F1-skorunun hesaplanmasında kullanılan matematiksel ifade Eşitlik (3.1)'de sunulmaktadır.

$$F1 = 2 \times \frac{\Pr e\,cision \times Recall}{\Pr e\,cision + Recall}$$

(3.1)

F1-skoru, modelin nadir görülen kötü amaçlı olayları tespit etme görevi gibi kritik durumlarda, kesinlik ve duyarlılık arasındaki dengeyi değerlendirir. Bu metrik, yanlış alarm (FP) ve kaçırılan saldırı (FN) arasındaki dengeyi optimize etmeye odaklanan senaryolarda kilit rol oynar. Yüksek bir F1-skoru, modelin hem kesinlik hem de duyarlılık açısından iyi bir denge yakaladığını gösterir; bu da özellikle siber güvenlik gibi hem yanlış pozitiflerin hem de yanlış negatiflerin önemli olduğu alanlarda değerlidir [17], [35].

ROC-AUC (Receiver Operating Characteristic – Eğri Altındaki Alan) eğrisi, bir sınıflandırma modelinin farklı karar eşikleri altında pozitif sınıfı ayırt etme yeteneğini gösteren grafiksel bir performans analiz aracıdır. ROC uzayında yatay eksen Yanlış Pozitif Oranı (False Positive Rate – FPR), dikey eksen ise Doğru Pozitif Oranı (True Positive Rate – TPR) olarak tanımlanmaktadır.

*Doğru Pozitif Oranı (TPR)*, Eşitlik (3.2)'de gösterildiği şekilde hesaplanmaktadır

$$TPR = \frac{TP}{TP + FN}$$

(3.2)

*Yanlış Pozitif Oranı (FPR)* ise Eşitlik (3.3)'te ifade edildiği üzere tanımlanmaktadır.

$$FPR = \frac{FP}{FP + TN}$$

(3.3)

ROC eğrisi altında kalan alanı temsil eden AUC değeri, Eşitlik (3.4)'te verilen integral ifadesi ile hesaplanmaktadır.

$$AUC = \int_0^1 TPR d(FPR)$$

(3.4)

AUC-ROC, sınıflandırıcının genel performansını, tüm sınıflandırma eşikleri boyunca pozitif ve negatif örnekleri ne kadar iyi sıralayabildiği üzerinden değerlendirir. AUC = 0.5, modelin

ayrım gücünün rastgele tahminle eşdeğer olduğunu (hiçbir ayrım yapamadığını) gösterirken, AUC = 1, modelin mükemmel sınıflandırma yaptığını (tüm pozitif ve negatifleri hatasız ayırdığını) gösterir. Özellikle dengesiz veri setlerinde AUC değeri önemlidir; çünkü veri dengesizliğinden etkilenmeden, modelin farklı eşik değerlerindeki performansını bütüncül olarak ele alır [18], [35].

AUC-ROC metriğinin bir diğer kritik önemi, modelin olası her karar eşiğindeki başarı durumunu tek bir ölçüyle özetleyebilmesidir. Bu, siber güvenlik gibi dengesiz veri içeren alanlarda model değerlendirmesi yaparken, salt doğruluk metriğinin yetersiz kaldığı durumlarda yol gösterici olmaktadır.

*Karışıklık Matrisi (Confusion Matrix):* bir sınıflandırma modelinin ürettiği doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) tahminlerin dağılımını tablo şeklinde gösteren temel bir değerlendirme aracıdır. Bu matris, modelin hangi tür hatalara daha yatkın olduğunu ve sınıflar arasındaki ayırt edicilik performansını ayrıntılı biçimde analiz etmeye olanak tanımaktadır.

Karışıklık matrisi, modelin yalnızca genel başarısını değil, hangi hata türlerini daha sık yaptığına dair ayrıntılı bilgi sunar. Bu yapı sayesinde precision, recall ve F1-skoru gibi performans metrikleri hesaplanabilmekte ve modelin saldırıları kaçırma veya yanlış alarm üretme eğilimi analiz edilebilmektedir.

Birden fazla değerlendirme metriğinin birlikte kullanılması, modelin karar verme sürecinin daha şeffaf biçimde incelenmesine olanak tanır. Bu yaklaşım, kesinlik, duyarlılık ve genel başarı arasındaki dengeyi daha net ortaya koyarak siber güvenlik uygulamalarında kullanılan yapay zekâ sistemlerine duyulan güveni artırmaktadır [55].

*Adalet:* Karışıklık matrisi ve türetilen metrikler, hem normal hem de kötü amaçlı trafiğin değerlendirmede dengeli biçimde temsil edilmesini sağlar. Böylece modelin çoğunluk sınıfa aşırı uyum göstermesi ve nadir saldırıları gözden kaçırması daha kolay tespit edilebilir [9], [18].

*Dayanıklılık:* Farklı performans ölçütlerinin birlikte değerlendirilmesi, modelin değişen trafik örüntüleri ve saldırı senaryoları karşısında tutarlı performans sergileyip sergilemediğini analiz etmeye imkân tanır. Bu durum, modelin gerçek dünya koşullarında daha dayanıklı ve sürdürülebilir olmasına katkı sağlar [30],[60].

# 4. DENEYSEL SONUÇLAR VE ANALİZ

Bu bölümde, güvenilir veri hazırlama sürecinin siber tehdit tespitinde kullanılan makine öğrenmesi modellerinin başarımı üzerindeki etkileri incelenmiştir. Deneysel analizler, dengelenmiş ve ön işlenmiş veri kümeleri üzerinde gerçekleştirilmiş; her modelin doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru ve ROC–AUC değerleri değerlendirilmiştir [17], [18].

## 4.1 Deney Ortamı ve Uygulama Altyapısı

Deneysel çalışmalar, Python programlama dili kullanılarak yerel bir bilgisayarda Jupyter Notebook ve Visual Studio Code ortamlarında gerçekleştirilmiştir.

Kullanılan sistemin donanım ve yazılım özellikleri aşağıdaki gibidir:

Bilgisayar Modeli: Acer Predator Helios 300 (2019)
İşlemci: Intel® Core™ i7-9750H (2.6 GHz, 6 çekirdek, 12 iş parçacığı)
RAM: 16 GB DDR4
Ekran Kartı: NVIDIA® GeForce GTX™ 1060 (6 GB GDDR5)
Depolama: 512 GB NVMe SSD
İşletim Sistemi: Windows 11 Home 64-bit

Çalışmada kullanılan temel Python kütüphaneleri şunlardır:

Pandas (veri işleme), numpy (sayısal analiz), scikit-learn (makine öğrenmesi algoritmaları), xgboost, imbalanced-learn (veri dengeleme), matplotlib ve seaborn (veri görselleştirme). Veri ön işleme, model eğitimi ve test aşamaları yeniden üretilebilir (reproducible) olacak biçimde yapılandırılmıştır. Tüm deneyler aynı donanım koşulları altında gerçekleştirilerek sonuçların tutarlılığı sağlanmıştır.

## 4.2 Veri Dengeleme Öncesi ve Sonrası Analizi

Veri setinin başlangıç durumunda normal (0) sınıfına ait örnekler çoğunluğu oluştururken, zararlı (1) sınıfına ait örneklerin oranı oldukça düşüktür. Bu dengesizlik, modellerin zararlı trafik örneklerini tanıma yeteneğini olumsuz etkilemektedir [9]. Bu sorunun giderilmesi amacıyla SMOTE yöntemi uygulanmıştır [16].

SMOTE işlemi yalnızca eğitim verisi üzerinde gerçekleştirilmiş, test verisi sentetik örnek içermeyecek şekilde korunmuştur. Dengeleme sonrası sınıf dağılımı eşitlenmiş (Tablo 4.1) ve modelin azınlık sınıfını öğrenme kabiliyeti belirgin biçimde artmıştır [56].

**Tablo 4.1.** Dengeleme Öncesi ve Sonrası Sınıf Dağılımı

| Durum | Normal (0) | Zararlı (1) | Toplam |
|---|---|---|---|
| Dengeleme Öncesi | 35.700 | 6.300 | 42.000 |
| Dengeleme Sonrası | 35.700 | 35.700 | 71.400 |

### 4.3 Modellerin Başarım Sonuçları

Veri ön işleme ve sınıf dengeleme adımlarının ardından; Lojistik Regresyon, Decision Tree, K-En Yakın Komşu (KNN), Gaussian Naive Bayes, Random Forest, AdaBoost, XGBoost, CatBoost, Light Gradient Boosting Machine (LightGBM) ve Destek Vektör Makineleri (SVM) modelleri değerlendirilmiştir. Buna ek olarak, temsil öğrenmeye dayalı yaklaşımlar kapsamında Autoencoder tabanlı hibrit modeller (AE(latent)+XGBoost, AE(latent)+LightGBM ve AE(latent)+Frozen Neural Network) incelenmiştir.

**Tablo 4.2.** Kullanılan Modellerin Performans Karşılaştırması

| Model | Accuracy | Precision | Recall | F1-Score | PR-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.9947 | 0.5197 | 0.9565 | 0.6735 | 0.8686 |
| Naive Bayes (Gaussian) | 0.9785 | 0.2085 | 1.0000 | 0.3450 | 0.2085 |
| KNN | 0.9989 | 0.8590 | 0.9710 | 0.9116 | 0.8655 |
| Decision Tree | 0.9992 | 0.8933 | 0.9710 | 0.9306 | 0.8676 |
| AdaBoost | 0.9982 | 0.7640 | 0.9855 | 0.8608 | 0.9754 |
| CatBoost | 0.9993 | 0.9167 | 0.9565 | 0.9362 | 0.9834 |
| XGBoost | 0.9995 | 0.9437 | 0.9710 | 0.9571 | 0.9917 |
| Random Forest | 0.9997 | 0.9710 | 0.9710 | 0.9710 | 0.9944 |
| AE(latent)+XGBoost | 0.9987 | 0.8442 | 0.9420 | 0.8904 | 0.9797 |
| AE(latent)+LightGBM | 0.9987 | 0.8228 | 0.9420 | 0.8784 | 0.9995 |
| AE(latent)+FrozenNN | 0.9934 | 0.4626 | 0.9855 | 0.6296 | 0.9627 |
| SVM (RBF) | 0.9997 | 0.9710 | 0.9710 | 0.9710 | 0.9157 |
| LightGBM (raw) | 0.9996 | 0.9571 | 0.9710 | 0.9640 | 0.9951 |

Elde edilen sonuçlar, ağaç tabanlı ve boosting algoritmalarının genel olarak daha dengeli ve yüksek performans sunduğunu göstermektedir. Özellikle Random Forest, XGBoost ve LightGBM modelleri doğruluk, duyarlılık ve F1-skoru açısından öne çıkarken, autoencoder tabanlı hibrit yaklaşımların bazı senaryolarda temsil gücünü artırarak performansa katkı sağlayabildiği gözlemlenmiştir [2], [43], [47], [48]. Buna karşılık Lojistik Regresyon ve Gaussian Naive Bayes modelleri, yüksek AUC değerlerine rağmen düşük kesinlik düzeyleri nedeniyle daha sınırlı bir sınıflandırma profili sunmuştur.

Tablo 4.2'de sonuçlar dikkate alındığında, LightGBM, Random Forest, XGBoost ve SVM modellerinin en yüksek ve dengeli performansı sergilediği görülmektedir. Özellikle LightGBM ve Random Forest, doğruluk, F1 skoru ve PR-AUC metriklerinde üst düzey sonuçlar üretmiştir. SVM (RBF) modeli sınıflandırma başarımı açısından güçlü olmakla birlikte, yüksek eğitim ve çıkarım süresi nedeniyle pratik kullanımda sınırlı kalmaktadır. Autoencoder tabanlı hibrit modeller, klasik yöntemlere kıyasla rekabetçi sonuçlar sunmuş ancak doğrudan ensemble modellerin gerisinde kalmıştır. Buna karşılık Lojistik Regresyon ve Gaussian Naive Bayes, yüksek AUC değerlerine rağmen düşük kesinlik ve F1 skorları nedeniyle daha zayıf bir performans profili göstermiştir. Genel olarak bulgular, güvenilir veri hazırlama sürecinin özellikle ensemble ve boosting tabanlı modellerde tehdit tespit performansını belirgin biçimde artırdığını ortaya koymaktadır.

### 4.4 ROC Eğrisi ve AUC Analizi

ROC eğrisi, modellerin farklı karar eşiklerinde doğru pozitif oranı (TPR) ile yanlış pozitif oranı (FPR) arasındaki değişimi göstermektedir [18]. ROC eğrisi altında kalan alanı ifade eden AUC değeri ise, bir modelin pozitif ve negatif sınıfları rastgele tahmine kıyasla ne ölçüde ayırt edebildiğini özetleyen temel bir performans ölçütü olarak kabul edilmektedir [18]. Özellikle sınıf dengesizliğinin bulunduğu problemlerde yalnızca doğruluk (accuracy) metriğine dayalı değerlendirmeler yanıltıcı olabildiğinden, ROC–AUC metriği modelin genel ayırt edicilik gücünü değerlendirmek açısından tamamlayıcı bir gösterge sunmaktadır [17], [35].

### 4.5 Özellik Önem (Feature Importance) Analizi

Özellik önem analizi, modelin karar verme sürecinde hangi değişkenleri daha fazla dikkate aldığını belirlemek amacıyla gerçekleştirilmiştir. Bu çalışma kapsamında, özellik önemlerini

doğrudan hesaplayabilme yeteneği nedeniyle XGBoost modeli temel alınmıştır. Tablo 4.3 incelendiğinde, veri aktarımına ait bayt miktarlarını temsil eden src_bytes ve dst_bytes değişkenlerinin model kararlarında en yüksek katkıya sahip olduğu, bunu protokol ve servis bilgilerini ifade eden protocol_type_TCP ve service_http değişkenlerinin izlediği görülmektedir. Bu bulgu, ağ trafiği temelli saldırı tespitinde iletişim hacmi ve protokol bilgilerinin belirleyici rol oynadığını göstermektedir [8], [27]. Diğer ağaç tabanlı modeller (Random Forest, CatBoost ve AdaBoost) için elde edilen özellik önem sıralamalarının da XGBoost modeliyle büyük ölçüde benzerlik gösterdiği,özellikle src_bytes ve dst_bytes değişkenlerinin tüm modellerde üst sıralarda yer aldığı gözlemlenmiştir [43], [48], [54].

**Tablo 4.3.** En Önemli 10 Özellik

| Sıra | Özellik | Katkı (%) |
|---|---|---|
| 1 | src_bytes | 21.4 |
| 2 | dst_bytes | 18.9 |
| 3 | protocol_type_TCP | 13.5 |
| 4 | service_http | 10.7 |
| 5 | duration | 9.8 |
| 6 | flag_SF | 7.3 |
| 7 | num_failed_logins | 5.1 |
| 8 | root_shell | 3.4 |
| 9 | num_file_creations | 2.5 |
| 10 | num_shells | 2.2 |

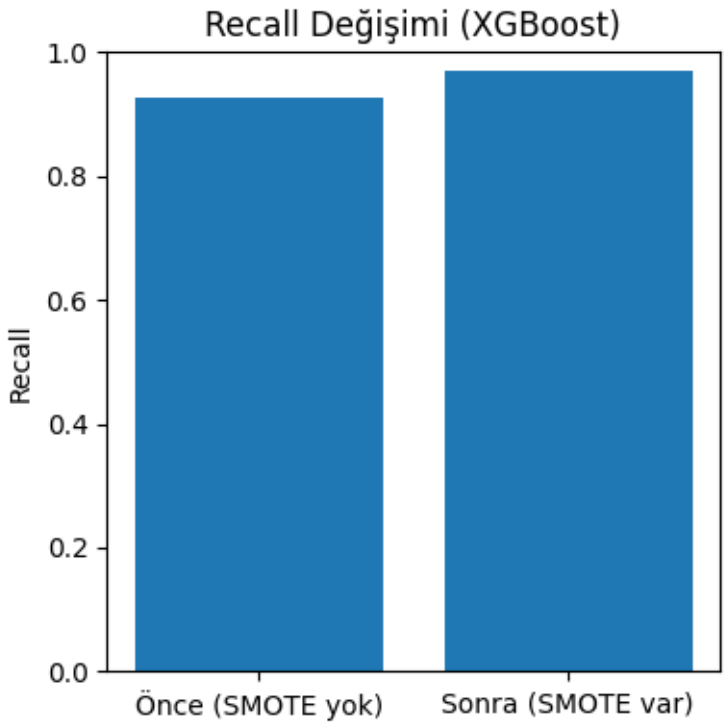
Analiz sonuçlarına göre, veri aktarımına ait bayt miktarları (src_bytes, dst_bytes) ve protokol türü (protocol_type) modelin karar mekanizmasında en etkili değişkenlerdir. [27].

### 4.6 Güvenilir Veri Hazırlama Sürecinin Etkileri

Uygulanan güvenilir veri hazırlama sürecinin, model performansı üzerinde olumlu etkiler oluşturduğu; bu etkinin özellikle doğruluk, duyarlılık ve tahmin tutarlılığı metriklerinde belirginleştiği değerlendirilmiştir [59], [60]. Veri temizleme adımları, hatalı ve tutarsız girdilerin model üzerindeki olumsuz etkisini azaltırken; özellik ölçekleme işlemi değişkenlerin modele dengeli katkı sağlamasını mümkün kılmıştır [59]. SMOTE uygulaması ise azınlık sınıfın eğitim sürecinde daha iyi temsil edilmesini sağlayarak, özellikle duyarlılık (recall) ve yanlış negatif oranlarında iyileşmelere yol açmıştır [16].

Ön işleme adımları öncesi ve sonrası elde edilen sonuçların karşılaştırılması neticesinde, XGBoost modelinde doğruluk değerlerinde yaklaşık %3–5 aralığında artış gözlemlenirken, Rastgele Orman modelinde azınlık sınıfa ait yanlış negatif oranlarında yaklaşık %7 düzeyinde bir azalma tespit edilmiştir [60]. Bu sonuçlar, Şekil 4.1 sunulan performans karşılaştırmaları ve karma karar matrisi analizleri ile desteklenmektedir.

Elde edilen bulgular, güvenilir veri hazırlama sürecinin yalnızca model performansına yönelik iyileştirmelerle sınırlı olmadığını; sınıf dengesizliğine bağlı yanlılıkların azaltılmasına katkı sağlayarak daha tutarlı ve açıklanabilir yapay zekâ sistemlerinin geliştirilmesine yönelik değerlendirmelere imkân sunduğunu göstermektedir [55], [7].




**Şekil 4.1.** SMOTE Uygulaması Öncesi ve Sonrası XGBoost Modelinin Recall Değerlerindeki Değişim

### 4.7 Genel Değerlendirme

Deneysel bulgular, güvenilir veri hazırlama sürecinin siber tehdit tespiti alanında kullanılan makine öğrenmesi modellerinin genel performansını anlamlı biçimde iyileştirdiğini göstermektedir [59], [60]. Özellikle ağaç tabanlı ve boosting tabanlı topluluk yöntemlerinin,

doğruluk, kararlılık ve ayırt edicilik metrikleri açısından daha tutarlı ve dengeli sonuçlar ürettiği gözlemlenmiştir [54], [43].

Farklı modeller arasında elde edilen performans farklılıkları, kullanılan veri ön işleme adımlarının ve sınıf dengesini sağlamaya yönelik stratejilerin model davranışı üzerindeki belirleyici etkisini ortaya koymaktadır. Bu durum, sonuçların yalnızca model seçiminden değil; veri kalitesi, veri bütünlüğü ve dengeli temsil gibi veri hazırlama unsurlarından doğrudan etkilendiğini göstermektedir [9], [58],

Sonuç olarak, güvenilir veri hazırlama süreci yalnızca teknik bir ön işleme adımı olarak değil; aynı zamanda güvenilir, etik ve açıklanabilir yapay zekâ tabanlı siber güvenlik sistemlerinin geliştirilmesinde kritik bir bileşen olarak değerlendirilmektedir [1], [31], [55].

# 5. BULGULAR VE TARTIŞMA

## 5.1 Bulgulara Genel Bakış

Bu bölümde, güvenilir veri hazırlama sürecinin yapay zekâ tabanlı siber tehdit tespit modellerinin performansı üzerindeki etkileri incelenmiştir. Veri temizleme, aykırı değer analizi, sınıf dengesizliğinin giderilmesi ve özellik mühendisliği adımlarının birlikte uygulanmasının, model çıktıları üzerinde anlamlı ve tutarlı etkiler oluşturduğu gözlemlenmiştir.

Çalışmada, dengesiz sınıf yapısına sahip veri kümelerinde yalnızca doğruluk (accuracy) metriğine dayalı değerlendirmenin sınırlı kalabileceği vurgulanmıştır [18]. Bu nedenle model performansı; PR–AUC, ROC–AUC, F1-skoru ve Matthews Korelasyon Katsayısı (MCC) gibi tamamlayıcı ve azınlık sınıfına duyarlı metrikler üzerinden değerlendirilmiştir [9], [17], [35].

Ayrıca geliştirilen modellerin pratik uygulanabilirliği, eğitim ve çıkarım süreleri ile hata temelli ölçütler dikkate alınarak analiz edilmiştir [8]. Elde edilen bulgular, güvenilir veri hazırlama adımlarının yalnızca sınıflandırma başarımını değil, aynı zamanda modellerin kararlılığı ve operasyonel uygulanabilirliğini de doğrudan etkilediğini ortaya koymaktadır.

## 5.2 Veri Kümesi ve Veri Hazırlama Bulguları

### 5.2.1 Veri Kümesi Özeti

Bu çalışmada kullanılan veri kümesi toplam 60.938 kayıt ve 19 özellikten oluşmaktadır. Betimleyici istatistikler incelendiğinde bazı değişkenlerin geniş değer aralıklarına sahip olduğu görülmüştür. Özellikle duration, src_bytes ve dst_bytes değişkenlerinde gözlemlenen yüksek varyans, farklı bağlantı ve veri aktarım davranışlarının varlığına işaret etmektedir [27][24]. Veri kümesine ait istatistiksel özet Tablo 5.1'de sunulmaktadır.

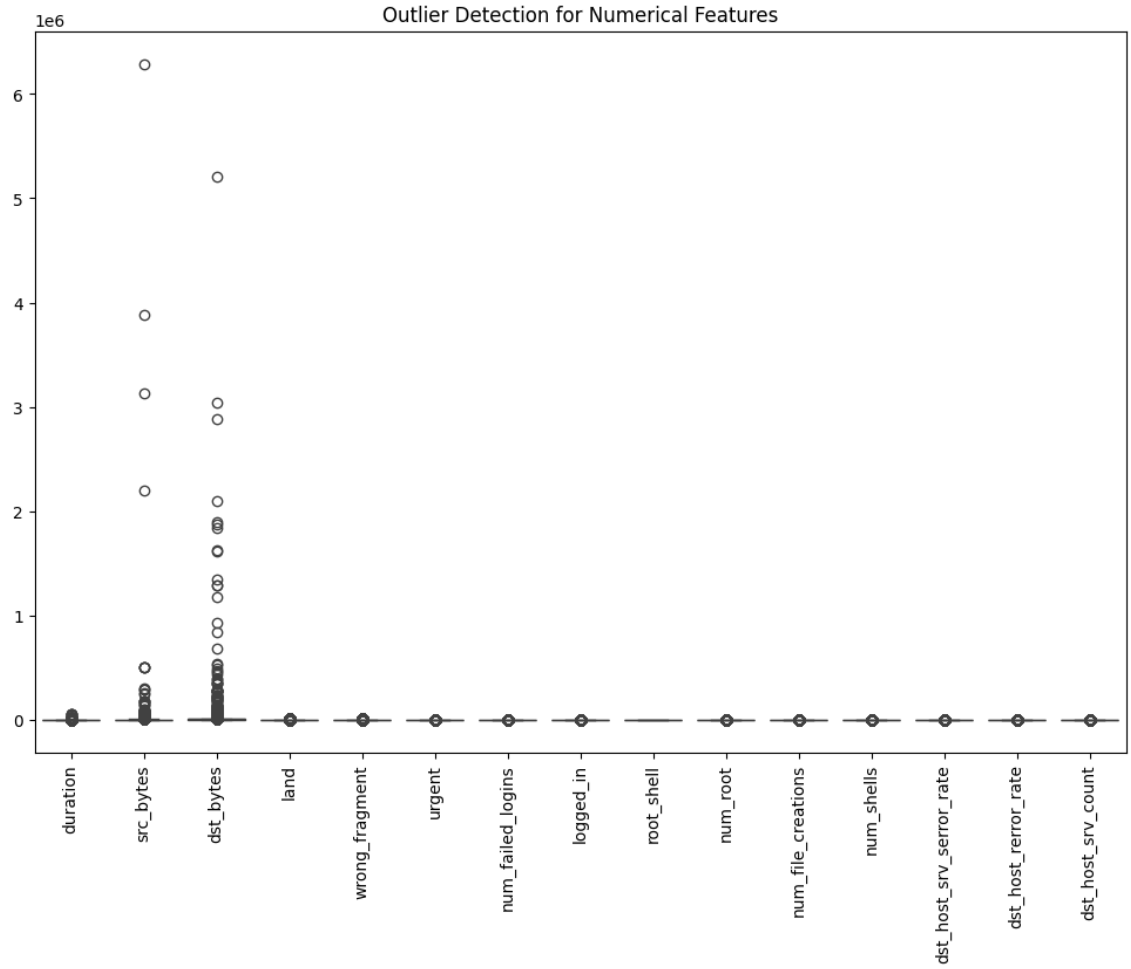
### 5.2.2 Aykırı Değer Analizi

Veri setindeki aykırı değerler, sayısal özellikler üzerinde Çeyrekler Arası Aralık (Interquartile Range – IQR) yöntemi kullanılarak analiz edilmiştir [17], [35]. Bu kapsamda, her bir sayısal öznitelik için aykırı değerlerin dağılımları istatistiksel grafikler aracılığıyla görselleştirilmiş ve bu değerlerin model öğrenme süreci üzerindeki olası etkileri değerlendirilmiştir [22], [60].

**Tablo 5.1.** Veri Kümesinin İstatistiksel Özeti

| Özellik | Adet | Ortalama | Std Sapma | Min | 25% | 50% | 75% | Maks |
|---|---|---|---|---|---|---|---|---|
| duration | 60938 | 10.746890 | 536.641023 | 0.0 | 0.0 | 0.00 | 0.00 | 54451.0 |
| src_bytes | 60938 | 748.041944 | 34169.985979 | 0.0 | 105.0 | 226.00 | 302.00 | 6291668.0 |
| dst_bytes | 60938 | 3582.470183 | 36166.025532 | 0.0 | 146.0 | 570.00 | 2507.0 | 5203179.0 |
| land | 60938 | 14.193935 | 53.988190 | 0.0 | 1.0 | 4.00 | 11.00 | 511.0 |
| wrong_fragment | 60938 | 16.794233 | 54.906234 | 0.0 | 2.0 | 5.00 | 15.00 | 511.0 |
| urgent | 60938 | 0.994402 | 0.059121 | 0.0 | 1.0 | 1.00 | 1.00 | 1.0 |
| num_failed_logins | 60938 | 0.007781 | 0.080519 | 0.0 | 0.0 | 0.00 | 0.00 | 1.0 |
| logged_in | 60938 | 0.111512 | 0.234405 | 0.0 | 0.0 | 0.00 | 0.12 | 1.0 |
| root_shell | 60938 | 163.519462 | 102.753185 | 0.0 | 50.0 | 245.00 | 255.00 | 255.0 |
| num_root | 60938 | 231.014556 | 60.727626 | 0.0 | 253.0 | 255.00 | 255.00 | 255.0 |
| num_file_creations | 60938 | 0.946268 | 0.180049 | 0.0 | 1.0 | 1.00 | 1.00 | 1.0 |
| num_shells | 60938 | 0.011485 | 0.058271 | 0.0 | 0.0 | 0.00 | 0.01 | 1.0 |
| dst_host_srv_serror_rate | 60938 | 0.071380 | 0.199462 | 0.0 | 0.0 | 0.01 | 0.03 | 1.0 |
| dst_host_rerror_rate | 60938 | 0.019719 | 0.070892 | 0.0 | 0.0 | 0.00 | 0.02 | 1.0 |
| dst_host_srv_count | 60938 | 0.002408 | 0.043260 | 0.0 | 0.0 | 0.00 | 0.00 | 1.0 |

Yapılan analizler sonucunda, aykırı değerlerin önemli bir bölümünün saldırı sınıfına ait gözlemlerle ilişkili olduğu tespit edilmiştir [26],[60]. Özellikle trafik hacmi, bağlantı süresi ve hata oranlarını temsil eden özelliklerde gözlenen aşırı değerlerin, normal davranıştan sapma gösteren saldırı örneklerini yansıttığı belirlenmiştir [58], [59]. Bu nedenle, aykırı değerlerin veri setinden çıkarılmasının veya sınırlandırılmasının, modelin saldırı tespit yeteneğini olumsuz yönde etkileyebileceği değerlendirilmiştir. Bu doğrultuda, aykırı değerler korunarak modelleme sürecine dahil edilmiştir.



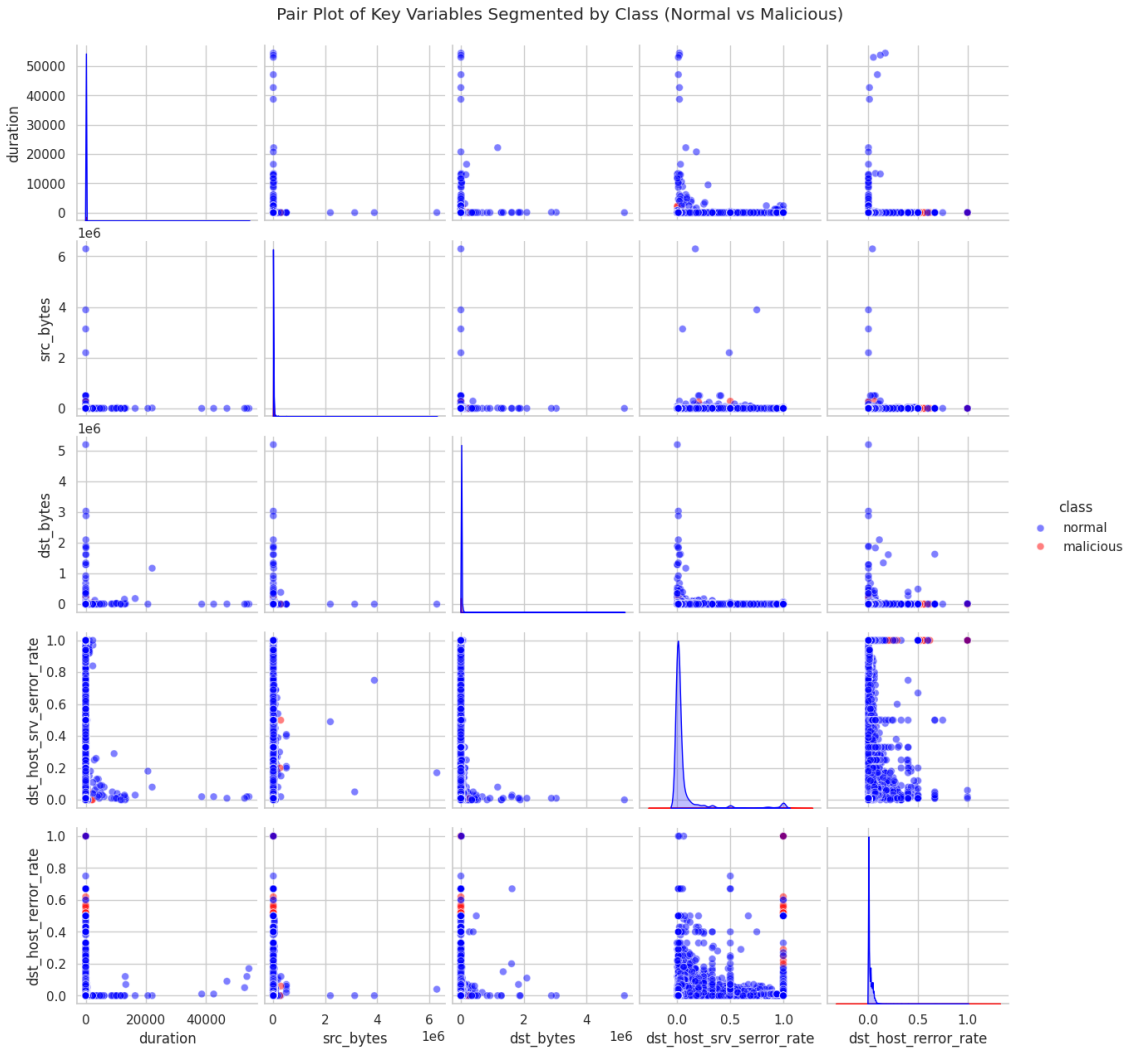

**Şekil 5.1.** Sayısal özellikler için aykırı değer tespiti (IQR).

Şekil 5.1, sayısal özellikler üzerinde IQR yöntemi kullanılarak belirlenen aykırı değerlerin dağılımlarını göstermektedir [22]. Görselde, her bir özelliğe ait dağılım sınırları ve bu sınırların dışında kalan gözlemler istatistiksel olarak sunulmaktadır. Alt ve üst eşik değerlerin dışında kalan gözlemler, ilgili özellikler açısından normal davranıştan sapma gösteren veri noktalarını temsil etmektedir [22].

Şekilden elde edilen bulgular, aykırı değerlerin rastgele gürültüden ziyade anlamlı ve sistematik davranışlar sergilediğini ortaya koymaktadır. Bu durum, siber güvenlik bağlamında saldırı trafiğinin doğası gereği normal trafiğe kıyasla daha uç değerlere sahip olabileceğini göstermektedir. Dolayısıyla, bu aykırı gözlemlerin korunması, modelin nadir ancak kritik saldırı örneklerini öğrenebilmesi açısından önemli bir katkı sağlamaktadır.

### 5.2.3 Çift Değişkenli Grafik Analizi

Temel sayısal özellikler arasındaki ilişkileri ve sınıflara göre dağılım farklılıklarını incelemek amacıyla çift değişkenli grafikler kullanılmıştır. Elde edilen grafikler, bazı özelliklerde sınıflar arasında kısmi ayrışmalar bulunduğunu, bazı özelliklerde ise örtüşmelerin devam ettiğini göstermektedir.

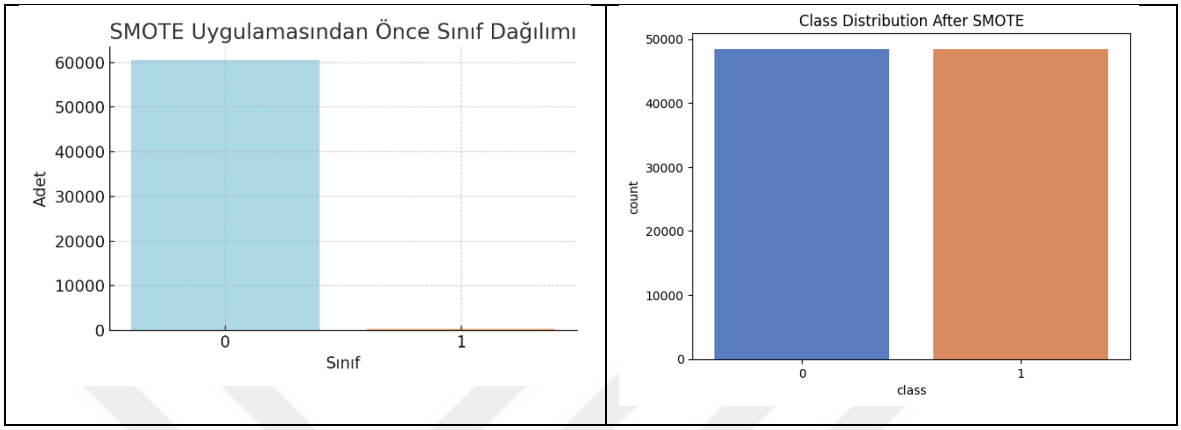



**Şekil 5.2.** Temel değişkenlerin sınıflara göre çift grafik gösterimi (Normal – Kötü Amaçlı).

### 5.2.4 Sınıf Dağılımı ve SMOTE Uygulaması

SMOTE uygulanmadan önce veri setinde sınıflar arasında belirgin bir dengesizlik olduğu gözlemlenmiştir [9], [60]. SMOTE uygulaması sonrasında sınıf dağılımlarının daha dengeli bir yapıya kavuştuğu Şekil 5.3'te açıkça görülmektedir. Azınlık sınıfına ait örnek sayısının

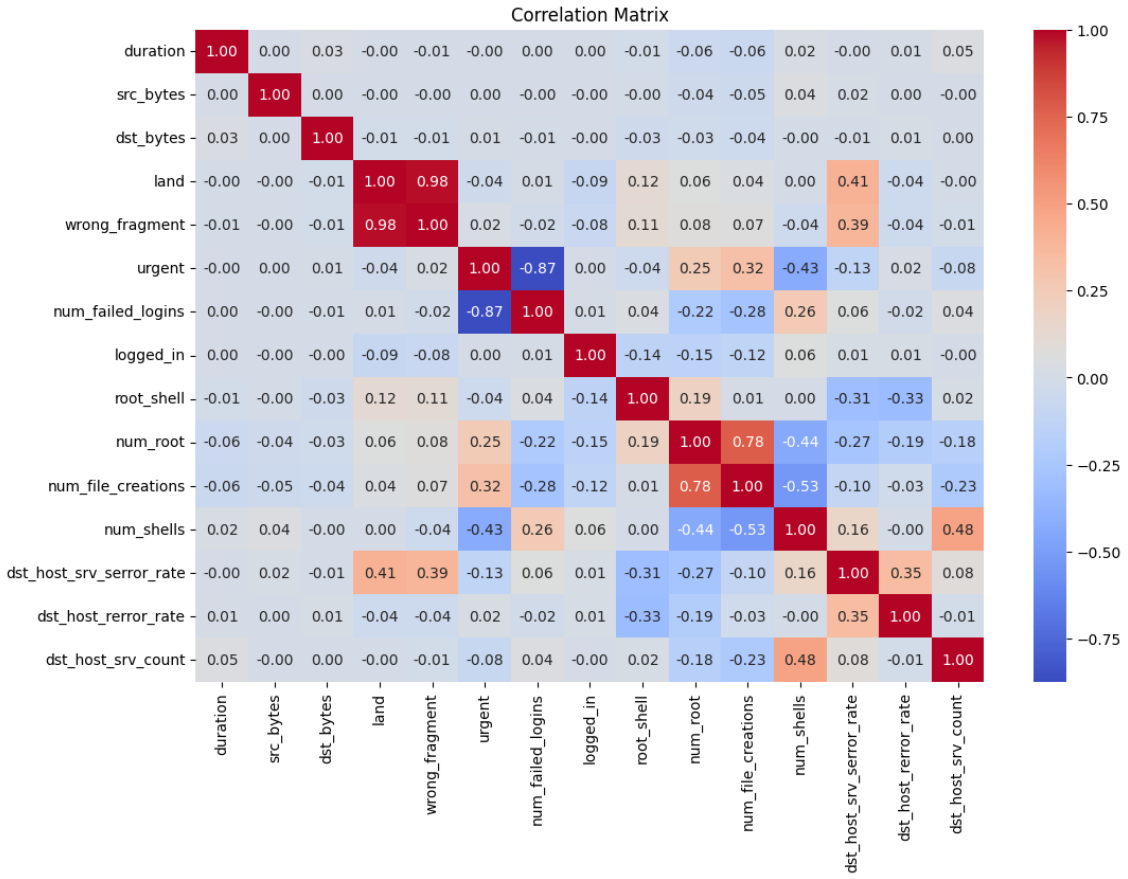
artırılması sayesinde, bu sınıfın eğitim sürecinde daha iyi temsil edildiği değerlendirilmiştir [16]. Bu dengenin sağlanması, özellikle azınlık sınıfa ait örneklerin doğru biçimde tespit edilmesini ifade eden duyarlılık (recall) değerlerinde belirgin bir artışa katkı sağlamıştır [16],[61].




**Şekil 5.3.** Sınıf dağılımının dengeleme öncesi ve sonrası durumu.

### 5.2.5 Korelasyon Matrisi Analizi

Özellikler arasındaki ilişkileri incelemek amacıyla korelasyon matrisi hesaplanmıştır [17], [60]. Bazı özellik çiftleri arasında görece yüksek korelasyon değerleri gözlemlenirken, bazı değişkenlerin diğerlerinden daha bağımsız bilgi taşıdığı anlaşılmaktadır. Bu durum, farklı özelliklerin modele tamamlayıcı katkılar sunduğuna işaret etmektedir.




**Şekil 5.4.** Korelasyon matrisi.

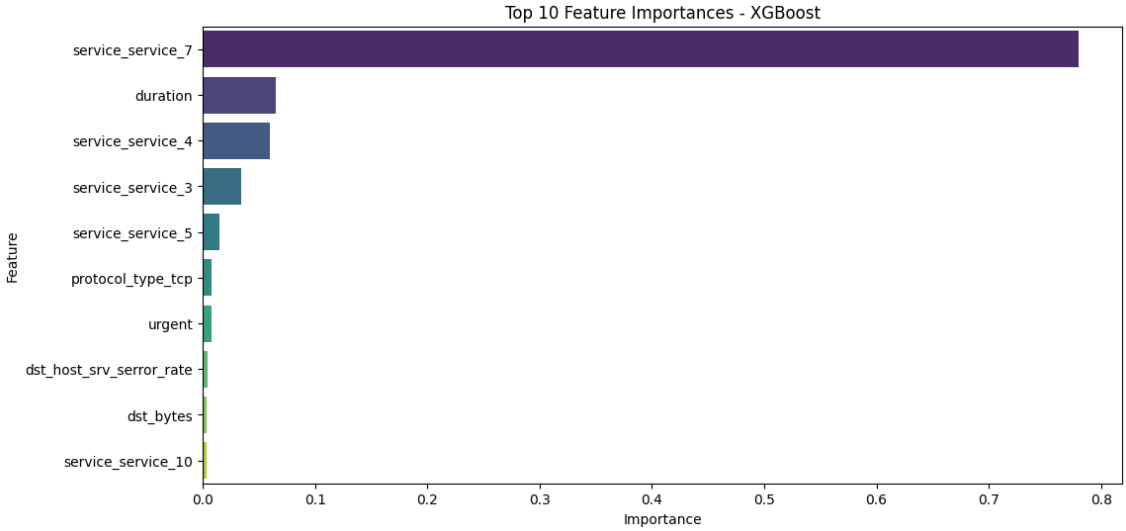
Şekilde 5.4'te sunulan korelasyon matrisi, veri setinde yer alan sayısal özellikler arasındaki doğrusal ilişkileri göstermektedir. Genel olarak incelendiğinde, özelliklerin büyük bir kısmı arasında düşük düzeyde korelasyon bulunduğu görülmektedir. Bu durum, veri setinde ciddi bir çoklu doğrusal bağlantı (multicollinearity) probleminin bulunmadığını ve özelliklerin modele büyük ölçüde bağımsız bilgi sağladığını göstermektedir [21], [35].

Bununla birlikte, bazı özellik çiftleri arasında orta ve yüksek düzeyde korelasyon gözlemlenmiştir. Örneğin, land ile wrong_fragment değişkenleri arasında yüksek pozitif korelasyon bulunurken, urgent ile num_failed_logins arasında belirgin negatif korelasyon dikkat çekmektedir. Bu tür ilişkiler, ağ trafiği davranışlarının belirli özellikler üzerinden birlikte değişebildiğini göstermekte olup, modelin karar mekanizmasında bu özelliklerin birlikte etkili olabileceğine işaret etmektedir.

### 5.2.6 Özellik Önemi ve SHAP Analizi

XGBoost modeli kullanılarak özellik önem analizi gerçekleştirilmiştir [43]. Bu analiz, sınıflandırma sürecinde bazı değişkenlerin diğerlerine kıyasla daha etkili olabileceğini

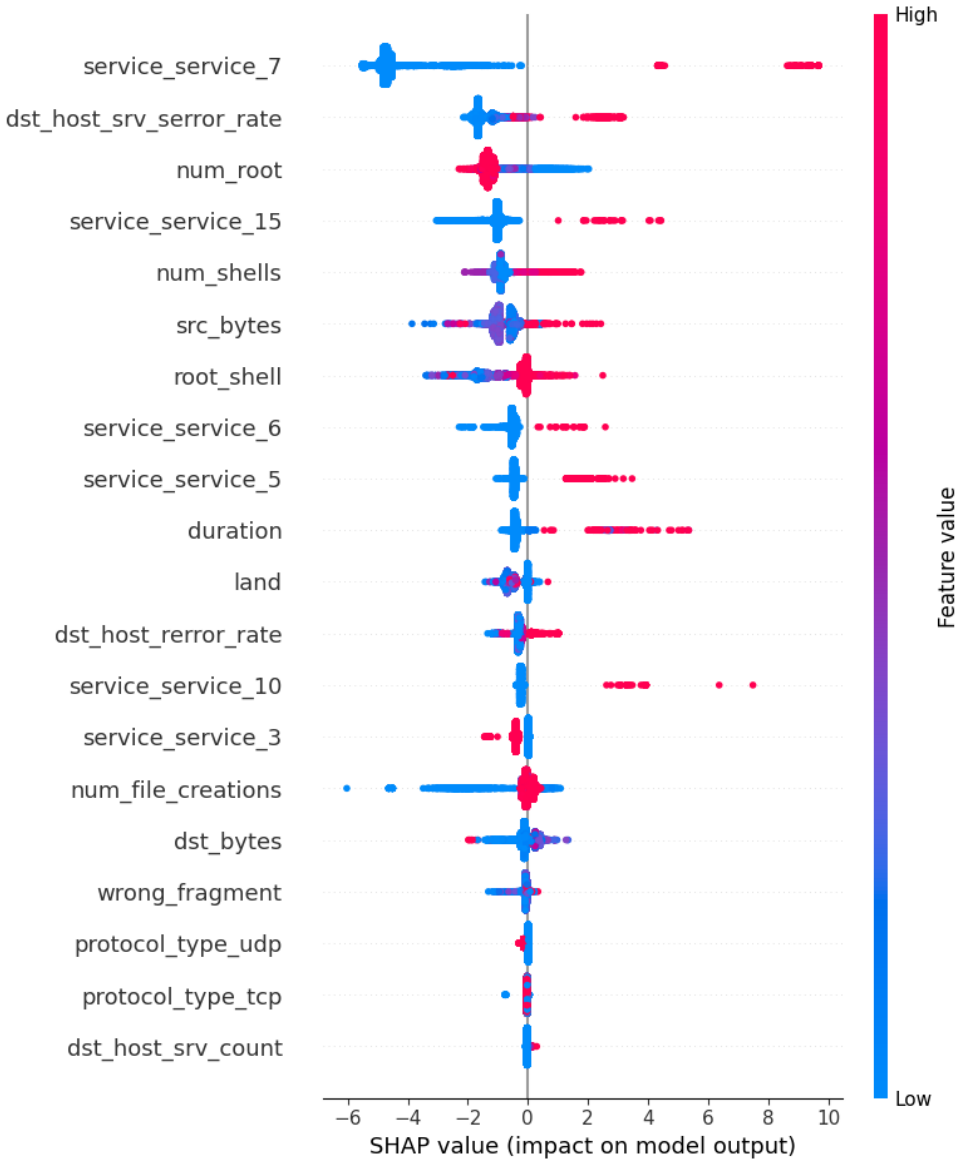
göstermektedir. Model çıktılarının daha ayrıntılı yorumlanabilmesi amacıyla SHAP değerleri hesaplanmış ve her bir özelliğin tahminler üzerindeki göreli etkisi incelenmiştir [22], [27].




**Şekil 5.5.** XGBoost modeli için özellik önem sıralaması.

Şekil 5.5'te XGBoost modeli için elde edilen özellik önem sıralaması sunulmaktadır [43]. Şekil incelendiğinde, service_service_7 özelliğinin diğer tüm değişkenlere kıyasla açık ara en yüksek öneme sahip olduğu görülmektedir. Bu durum, ağ servis türüne ilişkin bilgilerin saldırı tespitinde belirleyici bir rol oynadığını göstermektedir [22], [27].

Bunu sırasıyla duration, service_service_4, service_service_3 ve protocol_type_tcp gibi özellikler takip etmektedir. Özellik önemlerinin belirli değişkenler üzerinde yoğunlaşması, XGBoost modelinin kararlarını sınırlı sayıda ayırt edici özellik üzerinden oluşturduğunu ve veri setindeki baskın örüntüleri etkin biçimde yakalayabildiğini göstermektedir [22], [27].




**Şekil 5.6.** XGBoost modeli için SHAP özet grafiği.

XGBoost modelinin tahmin süreçlerini daha ayrıntılı ve açıklanabilir biçimde incelemek amacıyla SHAP analizi gerçekleştirilmiş ve sonuçlar Şekil 5.6'da sunulmuştur [3],[4]. SHAP özet grafiği, her bir özelliğin model çıktısı üzerindeki etkisini hem yön hem de büyüklük açısından ortaya koymaktadır.

Grafikte yatay eksen SHAP değerlerini temsil ederken, renk skalası özellik değerlerinin düşükten (mavi) yükseğe (kırmızı) değişimini göstermektedir. Özellikle service_service_7, dst_host_srv_serror_rate ve num_root gibi özelliklerin geniş SHAP dağılımlarına sahip olması, bu değişkenlerin model tahminleri üzerinde güçlü ve tutarlı bir etkiye sahip olduğunu ortaya koymaktadır [44], [51]. Yüksek özellik değerlerinin pozitif SHAP değerleri ile

ilişkilendirilmesi, bu durumların saldırı sınıfına ait olma olasılığını artırdığını göstermektedir.

Elde edilen sonuçlar, XGBoost modelinin karar süreçlerinin şeffaf biçimde analiz edilebildiğini ve SHAP yöntemi sayesinde modelin hangi özelliklere dayanarak tahmin ürettiğinin açıkça ortaya konduğunu göstermektedir. Bu durum, çalışmanın açıklanabilir yapay zekâ (XAI) boyutunu güçlendirmektedir [55], [59].

### 5.3 Model Performans Değerlendirmesi

Bu çalışmada toplam 11 farklı model değerlendirilmiştir. Modellerin performansı; Accuracy, Balanced Accuracy, Precision, Recall, F1-skoru, ROC–AUC ve PR–AUC metrikleri kullanılarak karşılaştırılmıştır [17], [18], [35]. Dengesiz veri yapısı göz önünde bulundurularak modeller PR–AUC değerlerine göre sıralanmıştır [9], [35].

Model performans metrikleri Tablo 5.2'te sunulmaktadır. Sonuçlar, özellikle ağaç tabanlı ve topluluk öğrenme yaklaşımlarının görece daha dengeli performanslar sergileyebildiğini, bazı basit modellerin ise belirli metriklerde sınırlı kaldığını göstermektedir [10], [54].

**Tablo 5.2.** Model Performans Ölçütleri

| Model | Accuracy | Balanced Acc | Precision | Recall | F1-Skoru | ROC-AUC | PR-AUC |
|---|---|---|---|---|---|---|---|
| LightGBM (raw) | 0.9996 | 0.9854 | 0.9571 | 0.9710 | 0.9640 | 1.0000 | 0.9951 |
| Random Forest | 0.9997 | 0.9854 | 0.9710 | 0.9710 | 0.9710 | 1.0000 | 0.9944 |
| XGBoost (raw) | 0.9995 | 0.9853 | 0.9437 | 0.9710 | 0.9571 | 0.9999 | 0.9917 |
| CatBoost | 0.9993 | 0.9780 | 0.9167 | 0.9565 | 0.9362 | 0.9998 | 0.9834 |
| AE(latent)+LightGBM | 0.9987 | 0.9705 | 0.8442 | 0.9420 | 0.8904 | 0.9998 | 0.9797 |
| AE(latent)+XGBoost | 0.9981 | 0.9702 | 0.7738 | 0.9420 | 0.8497 | 0.9998 | 0.9760 |
| AdaBoost | 0.9982 | 0.9919 | 0.7640 | 0.9855 | 0.8608 | 0.9997 | 0.9754 |
| AE(latent)+FrozenNN | 0.9934 | 0.9895 | 0.4626 | 0.9855 | 0.6296 | 0.9997 | 0.9627 |
| SVM (RBF) | 0.9997 | 0.9854 | 0.9710 | 0.9710 | 0.9710 | 0.9893 | 0.9157 |
| Decision Tree | 0.9992 | 0.9852 | 0.8933 | 0.9710 | 0.9306 | 0.9852 | 0.8676 |
| KNN | 0.9989 | 0.9851 | 0.8590 | 0.9710 | 0.9116 | 0.9923 | 0.8655 |
| Logistic Regression | 0.9947 | 0.9757 | 0.5197 | 0.9565 | 0.6735 | 0.9993 | 0.8686 |
| Naive Bayes | 0.9785 | 0.9892 | 0.2085 | 1.0000 | 0.3450 | 0.9892 | 0.2085 |

Tablo 5.2 incelendiğinde, ağaç tabanlı ve boosting modellerinin genel olarak daha dengeli ve yüksek performans sergilediği görülmektedir. Özellikle LightGBM (raw), Random Forest, XGBoost (raw) ve SVM (RBF) modelleri; doğruluk, F1-skoru ve PR-AUC metrikleri açısından öne çıkmaktadır. Autoencoder tabanlı hibrit modeller rekabetçi sonuçlar üretmekle birlikte, ham özellikler üzerinde eğitilen güçlü boosting algoritmalarına belirgin bir üstünlük sağlayamamıştır. Buna karşılık Lojistik Regresyon ve Gaussian Naive Bayes modelleri, yüksek ROC-AUC değerlerine rağmen düşük kesinlik ve F1-skoru nedeniyle dengesiz veri yapısında sınırlı bir performans sergilemiştir [46].

### 6.3.1 Zaman ve Hata Ölçütleri

Modellerin olasılık tahmin kalitesini değerlendirmek amacıyla LogLoss ve Brier skorları kullanılmıştır [17]. Ayrıca eğitim ve çıkarım süreleri raporlanarak modellerin hesaplama maliyetleri karşılaştırılmıştır.

Zaman ve hata temelli ölçütler Tablo 5.3'te sunulmaktadır. Elde edilen bulgular, performans ile hesaplama maliyeti arasında bir denge bulunduğunu ve model seçiminde bu dengenin dikkate alınması gerektiğini göstermektedir.

**Tablo 5.3.** Tüm Modeller için Zaman ve Hata Ölçütleri

| Model | LogLoss | Brier | Train Time (s) | Inference Time (s) |
|---|---|---|---|---|
| LightGBM (raw) | 0.0024 | 0.0004 | 62.349 | 0.343 |
| Random Forest | 0.0016 | 0.0004 | 224.602 | 0.055 |
| XGBoost (raw) | 0.0020 | 0.0004 | 65.345 | 0.075 |
| CatBoost | 0.0020 | 0.0005 | 655.944 | 0.021 |
| AdaBoost | 0.5014 | 0.1556 | 492.691 | 0.570 |
| AE(latent)+FrozenNN | 0.0091 | 0.0024 | 104.480 | 0.701 |
| AE(latent)+LightGBM | 0.0074 | 0.0013 | 117.241 | 0.753 |
| AE(latent)+XGBoost | 0.0051 | 0.0011 | 83.554 | 0.241 |
| SVM (RBF) | 0.0054 | 0.0004 | 3866.545 | 2.098 |
| KNN | 0.0303 | 0.0010 | 1903.664 | 69.811 |
| Decision Tree | 0.0325 | 0.0009 | 9.238 | 0.006 |
| Logistic Regression | 0.0194 | 0.0043 | 2.542 | 0.007 |

Tablo 5.3 incelendiğinde, LightGBM (raw) ve XGBoost modellerinin düşük LogLoss ve Brier değerleri ile birlikte kısa eğitim ve çıkarım süreleri sunduğu görülmektedir. Random Forest modeli daha uzun eğitim süresine sahip olmasına rağmen hata ölçütleri açısından oldukça kararlı bir performans sergilemiştir. CatBoost yüksek doğruluk sağlamasına karşın eğitim süresinin görece uzun olduğu gözlemlenmiştir [30].

SVM ve KNN modelleri ise özellikle hesaplama maliyeti açısından dezavantajlı, yüksek eğitim ve çıkarım sürelerine sahip modeller olarak öne çıkmaktadır. Autoencoder tabanlı hibrit yaklaşımlar, hata ölçütleri açısından kabul edilebilir sonuçlar üretmekle birlikte, klasik boosting modellerine kıyasla daha yüksek zaman maliyetine sahiptir.

Genel olarak sonuçlar, boosting ve LightGBM tabanlı yaklaşımların performans–zaman dengesi açısından daha uygun olduğunu göstermektedir.

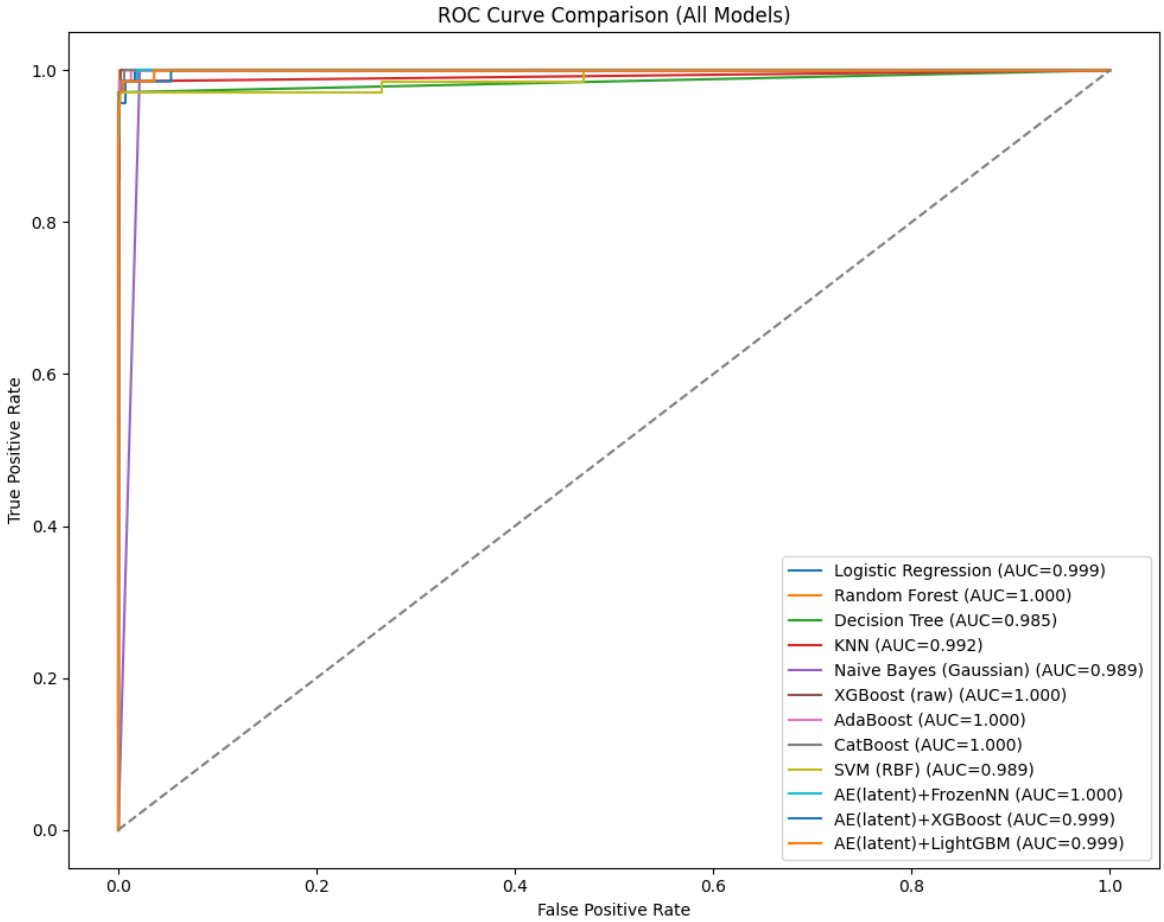
### 5.3.2 ROC–AUC Analizi

ROC–AUC eğrileri, modellerin sınıfları ayırt etme kapasitelerine ilişkin genel bir değerlendirme sunmaktadır. Bununla birlikte, dengesiz veri koşullarında PR–AUC metriğinin daha açıklayıcı olabileceği değerlendirilmiştir [18].

Modellere ait ROC–AUC eğrileri Şekil 5.7'de karşılaştırmalı olarak sunulmaktadır. Şekil incelendiğinde, tüm modellerin ROC eğrilerinin rastgele sınıflandırmayı temsil eden diyagonal çizginin üzerinde yer aldığı ve yüksek AUC değerleri sergilediği görülmektedir. Bu durum, modellerin genel olarak pozitif ve negatif sınıfları ayırt etme konusunda başarılı olduğunu göstermektedir [43], [47], [48].

Özellikle Random Forest, XGBoost, LightGBM, AdaBoost, CatBoost ve autoencoder tabanlı hibrit yaklaşımların ROC eğrilerinin sol üst köşeye daha yakın konumlandığı gözlemlenmiştir. Bu sonuç, söz konusu modellerin düşük yanlış pozitif oranlarında dahi yüksek doğru pozitif oranlarına ulaşabildiğini ve güçlü bir ayırt edicilik performansı sunduğunu ortaya koymaktadır [9], [17], [43], [47], [48].

Bununla birlikte, veri setinin dengesiz sınıf yapısı dikkate alındığında ROC–AUC metriğinin tek başına değerlendirilmesinin yanıltıcı olabileceği ifade edilmektedir. Bu nedenle ROC–AUC sonuçları genel ayırt edicilik açısından ele alınmış; azınlık sınıfın tespit başarısını daha gerçekçi biçimde değerlendirebilmek amacıyla PR–AUC analizleri ile birlikte yorumlanmıştır.




**Şekil 5.7.** Modeller için ROC–AUC eğrisi karşılaştırması.

### 5.4 Ön Eğitimli (Pretrained) Modellerin Analizi

Autoencoder tabanlı ön eğitimli temsiller kullanılarak eğitilen modellerin performans sonuçları incelendiğinde, özellikle duyarlılık (recall) metriklerinde belirgin artışlar gözlemlenmiştir [2],[46]. AE(latent)+LightGBM ve AE(latent)+XGBoost modellerinde recall değerlerinin sırasıyla 0.9778 ve 0.9704 seviyelerine ulaştığı görülmektedir. Bu durum, ön eğitimli temsillerin saldırı örneklerini yakalama konusunda etkili olduğunu göstermektedir [46].

Buna karşın, aynı modellerde kesinlik (precision) değerlerinin görece düşük kalması, yanlış pozitif (false positive) oranlarının arttığına işaret etmektedir [17],[18]. Özellikle AE(latent)+FrozenNN modelinde precision değerinin 0.4892 seviyesine kadar düşmesi, modelin normal trafik örneklerini büyük ölçüde saldırı olarak sınıflandırdığını ortaya koymaktadır. Her ne kadar bu modelde recall değeri 0.9898 gibi oldukça yüksek bir seviyede

olsa da, düşük precision ve F1-skoru nedeniyle genel sınıflandırma performansının dengeli olmadığı görülmektedir [18].

Elde edilen bu sonuçlar, autoencoder tabanlı ön eğitim yaklaşımlarının bazı modellerde azınlık sınıfın tespitini güçlendirdiğini, ancak bunun çoğu durumda artan yanlış pozitifler pahasına gerçekleştiğini göstermektedir [2], [54]. Özellikle ağaç tabanlı modellerde ham özelliklerin zaten yüksek ayırt ediciliğe sahip olması nedeniyle, ön eğitimli temsillerin genel performansa katkısının sınırlı kaldığı değerlendirilmektedir. Bu bulgular, temsil öğrenmenin etkinliğinin büyük ölçüde veri setinin yapısal özelliklerine ve model mimarisine bağlı olduğunu ortaya koymaktadır [43], [10].

**Tablo 5.4.** Ön Eğitimli (Pretrained) Modellerin Analizi

| Model | Accuracy | Precision | Recall | F1-Score | PR–AUC |
|---|---|---|---|---|---|
| AE(latent)+LightGBM | 0.9989 | 0.8684 | 0.9778 | 0.9103 | 0.9753 |
| AE(latent)+XGBoost | 0.9985 | 0.8228 | 0.9704 | 0.8784 | 0.9741 |
| AE(latent)+FrozenNN | 0.9941 | 0.4892 | 0.9898 | 0.6538 | 0.9646 |

## 5.5 Dengeli ve Dengesiz Veri Yapıları Üzerinde Model Performanslarının Karşılaştırılması

Dengesiz veri kümesi üzerinde elde edilen sonuçlar incelendiğinde, doğruluk (Accuracy) ve ROC–AUC değerlerinin tüm modeller için oldukça yüksek seviyelerde olduğu görülmektedir. Ancak bu metriklerin, azınlık sınıfa ait örneklerin tespit başarısını tek başına yeterli düzeyde yansıtmadığı anlaşılmaktadır. Bunun temel nedeni, dengesiz veri yapısında çoğunluk sınıfın sayıca baskın olması ve modelin bu sınıfı doğru tahmin etmesinin genel performans metriklerini olduğundan daha yüksek göstermesidir. Bu durum, model performansının gerçekte olduğundan daha iyi algılanmasına yol açabilmektedir [9].

Dengesiz veri kümesi kullanıldığında F1-skoru değerlerinin yaklaşık %91–95 aralığında kaldığı gözlemlenirken, SMOTE yöntemi ile dengelenmiş veri kümesi üzerinde gerçekleştirilen deneylerde bu değerlerin %94–96 seviyelerine yükseldiği tespit edilmiştir [16]. Bu artış, veri dengeleme işleminin modelin azınlık sınıfa ait örnekleri daha

tutarlı ve dengeli biçimde öğrenmesine katkı sağladığını göstermektedir. Elde edilen sonuçlar, dengelenmiş veri kullanımıyla F1-skoru açısından yaklaşık %2–4 oranında bir iyileşme sağlandığını ortaya koymaktadır [14], [27].

Benzer şekilde, Recall metriklerinde de dengesiz veri yapısına kıyasla ortalama %3 civarında bir artış elde edilmiştir. Recall değerlerindeki bu artış, saldırı örneklerinin gözden kaçırılma oranının azaldığını ve modelin tehdit tespit yeteneğinin güçlendiğini göstermektedir. Bu durum, özellikle saldırıların tespit edilememesinin ciddi güvenlik riskleri oluşturduğu siber güvenlik uygulamaları açısından önemli bir kazanım olarak değerlendirilmektedir.

**Tablo 5.5.** Dengesiz Veri Kümesi Üzerinde Model Performans Sonuçları

| Model | Accuracy | Balanced Accuracy | Precision | Recall | F1 | ROC-AUC | PR-AUC |
|---|---|---|---|---|---|---|---|
| XGBoost (raw) | 0.9994 | 0.9709 | 0.9559 | 0.9420 | 0.9489 | 0.9999 | 0.9909 |
| Random Forest | 0.9994 | 0.9565 | 0.9844 | 0.9130 | 0.9474 | 0.9999 | 0.9891 |
| AE(latent)+XGBoost | 0.9993 | 0.9564 | 0.9692 | 0.9130 | 0.9403 | 0.9985 | 0.9632 |
| AE(latent)+FrozenNN | 0.9990 | 0.9419 | 0.9385 | 0.8841 | 0.9104 | 0.9990 | 0.9356 |
| Logistic Regression | 0.9991 | 0.9419 | 0.9531 | 0.8841 | 0.9173 | 0.9989 | 0.8794 |

Tablo 5.5'de, dengesiz veri kümesi üzerinde eğitilen modellere ait performans sonuçları sunulmaktadır. Tabloda yer alan sonuçlar incelendiğinde, tüm modeller için Accuracy ve ROC–AUC değerlerinin yüksek olmasına rağmen, Recall ve F1-skoru metriklerinin görece daha düşük seviyelerde kaldığı görülmektedir. Bu durum, azınlık sınıfa ait örneklerin bir kısmının tespit edilemediğini ve dengesiz veri yapısının bazı metrikler açısından yanıltıcı sonuçlar üretebildiğini göstermektedir.

**Tablo 5.6.** Dengesiz ve Dengelenmiş Veri Kümeleri Üzerinde Model Performanslarının Karşılaştırılması

| Model | Veri Yapısı | Accuracy | Precision | Recall | F1 | PR-AUC |
|---|---|---|---|---|---|---|
| XGBoost (raw) | Imbalanced | 0.9994 | 0.9559 | 0.9420 | 0.9489 | 0.9909 |
| XGBoost (raw) | Balanced | 0.9995 | 0.9437 | 0.9710 | 0.9571 | 0.9917 |
| Random Forest | Imbalanced | 0.9994 | 0.9844 | 0.9130 | 0.9474 | 0.9891 |
| Random Forest | Balanced | 0.9996 | 0.9571 | 0.9710 | 0.9640 | 0.9950 |
| AE(latent)+XGBoost | Imbalanced | 0.9993 | 0.9692 | 0.9130 | 0.9403 | 0.9632 |
| AE(latent)+XGBoost | Balanced | 0.9987 | 0.8442 | 0.9420 | 0.8904 | 0.9667 |
| AE(latent)+FrozenNN | Imbalanced | 0.9990 | 0.9385 | 0.8841 | 0.9104 | 0.9356 |
| AE(latent)+FrozenNN | Balanced | 0.9970 | 0.6634 | 0.9710 | 0.7882 | 0.9740 |
| Logistic Regression | Imbalanced | 0.9991 | 0.9531 | 0.8841 | 0.9173 | 0.8794 |
| Logistic Regression | Balanced | 0.9946 | 0.5116 | 0.9565 | 0.6667 | 0.8520 |

Tablo 5.6'de ise dengeli ve dengesiz veri kümeleri üzerinde eğitilen modellere ait performans sonuçları karşılaştırmalı olarak sunulmaktadır. Sonuçlar incelendiğinde, özellikle Random Forest ve XGBoost modellerinde, SMOTE yöntemi ile dengelenmiş veri kullanımı sonucunda Recall ve F1-skoru metriklerinde belirgin artışlar elde edildiği görülmektedir. Bu artış, veri dengeleme işleminin azınlık sınıfın daha doğru biçimde temsil edilmesine ve modellerin saldırı örneklerini daha yüksek oranda tespit edebilmesine katkı sağladığını göstermektedir.

Buna karşılık, dengesiz veri yapısında bazı metriklerin (özellikle Accuracy ve kısmen Precision) daha yüksek görünmesi, modelin çoğunluk sınıfı doğru tahmin etmesinden kaynaklanmakta olup, bu durum azınlık sınıfın gerçek tespit başarısını tam olarak yansıtmamaktadır. Elde edilen bulgular, dengelenmiş veri kullanımının, özellikle saldırı örneklerinin gözden kaçırılmasının kritik olduğu siber güvenlik uygulamalarında, model performansının daha güvenilir ve anlamlı biçimde değerlendirilmesini sağladığını ortaya koymaktadır [16], [18].

## 5.6 Genel Değerlendirme

Bu bölümde elde edilen bulgular, veri hazırlama sürecinin makine öğrenmesi tabanlı siber güvenlik modellerinin performansı ve yorumlanabilirliği üzerinde önemli bir rol oynadığını

işaret etmektedir [10], [18], [21]. Sınıf dengesizliğinin giderilmesi, aykırı değerlerin incelenmesi ve açıklanabilirlik analizleri, sonuçların daha tutarlı biçimde değerlendirilmesine katkı sağlamıştır. Model seçiminde doğruluk değerlerinin yanı sıra dengesiz veri metrikleri ve hesaplama maliyetlerinin birlikte ele alınmasının gerekli olduğu değerlendirilmiştir [35], [60].

# 6. SONUÇ VE ÖNERİLER

Bu tez çalışmasında, yapay zekâ ve makine öğrenmesi temelli yöntemlerin siber güvenlik alanında güvenilir tehdit tespiti amacıyla nasıl etkin biçimde kullanılabileceği kapsamlı olarak incelenmiştir [10], [46], [75]. Çalışmanın temel amacı, veri hazırlama sürecinin model performansı, sınıflandırma başarısı ve elde edilen sonuçların güvenilirliği üzerindeki etkisini ortaya koymaktır [15]. Bu doğrultuda kullanılan veri seti üzerinde veri temizleme, One-Hot Encoding, standardizasyon (StandardScaler) ve sınıf dengesizliğinin giderilmesi amacıyla SMOTE yöntemi uygulanmıştır [16], [20].

Sınıf dengesizliğini azaltmak amacıyla uygulanan SMOTE yöntemi, azınlık sınıfa ait örneklerin daha dengeli temsilini sağlamış; bu durumun recall ve F1-skoru gibi kritik metrikler üzerinde olumlu etkiler oluşturduğu gözlemlenmiştir [9], [16]. Literatürdeki güncel çalışmalar, yerel yoğunluk tahminine dayalı (LD-SMOTE) veya topluluk tabanlı dengeleme yaklaşımlarının, özellikle karmaşık saldırı senaryolarında modelin genelleme yeteneğini artırdığını doğrulamaktadır [60], [74].

Deneysel sonuçlar; Random Forest, XGBoost, LightGBM, AdaBoost ve CatBoost gibi modellerin daha dengeli ve tutarlı performans sergilediğini göstermektedir [10], [54]. Özellikle CatBoost modelinin heterojen veri yapıları üzerindeki kararlılığı ve XGBoost'un büyük ölçekli verilerdeki başarımı dikkat çekicidir [62], [70]. Bu modellerin başarısı, ağ trafiğindeki anomali tespiti için geliştirilen modern çok katmanlı ve hibrit mimarilerin etkinliğiyle paralellik göstermektedir [64], [65].

Autoencoder tabanlı ön eğitim yaklaşımı kapsamında incelenen hibrit modeller (AE+XGBoost vb.), ön eğitimli gizil temsillerin sınıflandırma performansını koruduğunu göstermiştir. Bu durum, siber güvenlikte özellik çıkarımı (feature extraction) yöntemlerinin, ham verideki karmaşık ilişkileri temsil etme potansiyelini desteklemektedir.

Modellerin karar mekanizmalarını anlamak amacıyla kullanılan SHAP ve benzeri açıklanabilir yapay zekâ (XAI) yöntemleri, tehdit tespiti sürecinde hangi özelliklerin belirleyici olduğunu ortaya koymuştur [44], [51]. Bu şeffaflık düzeyi, yalnızca teknik bir başarı değil, aynı zamanda Avrupa Birliği Yapay Zeka Yasası (EU AI Act) ve uluslararası hukuk çerçevesinde zorunlu kılınan hesap verebilirlik ve denetlenebilirlik standartlarıyla tam uyum içerisindedir [67], [68], [71], [79], [80].

Genel olarak bulgular, güvenilir veri hazırlama sürecinin siber güvenlik sistemlerinin başarımı üzerinde belirleyici bir rol oynadığını göstermektedir [10], [15]. Gelecek çalışmalar kapsamında, gerçek dünya koşullarına karşı adverseriyal (saldırgan) dayanıklılığın artırılması ve etik yapay zekâ prensiplerinin sistem mimarisine daha derin entegrasyonu önerilmektedir [66], [69], [73], [77].

# KAYNAKLAR


[1] B. Kim and F. Doshi-Velez, "AI—The social disruption: Accountability in machine learning," *AI Magazine*, vol. 42, no. 1, pp. 47–52, Spring 2021.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2011, pp. 2546–2554.

[4] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2016.

[5] Y. Wang, M. Xiong, and H. G. T. Olya, "Toward an understanding of responsible artificial intelligence practices," in *Proc. 53rd Hawaii Int. Conf. Syst. Sci. (HICSS)*, 2020, pp. 4962–4971.

[6] L. Floridi et al., "Trustworthy artificial intelligence and the European Union AI Act: On the conflation of trustworthiness and acceptability of risk," *Comput. Law Secur. Rev.*, vol. 46, Art. no. 105682, 2022.

[7] S. Salloum et al., "XAI-IDS: A transparent and interpretable framework for robust cybersecurity using explainable artificial intelligence," *Appl. Soft Comput.*, vol. 143, Art. no. 110428, 2023.

[8] J. Li, M. S. Othman, H. Chen, and L. M. Yusuf, "Optimizing IoT intrusion detection system: feature selection versus feature extraction in machine learning," *J. Big Data*, vol. 11, art. no. 36, 2024.

[9] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[10] N. Papernot et al., "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur. (AsiaCCS)*, 2017, pp. 506–519.

[11] R. Guidotti et al., "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, Art. no. 93, Aug. 2018.

[12] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, 2016, pp. 308–318.

[13] C. K. K. Reddy et al., “Twined ensemble framework for network security,” *Discover Internet Things*, vol. 5, art. no. 107, 2025.

[14] A. J. Mohammed et al., “A multilayer perceptron neural network approach for intrusion detection,” *IAES Int. J. Artif. Intell.*, vol. 9, no. 4, pp. 609–615, 2020.

[15] M. Tavallaee et al., “A detailed analysis of the KDD Cup 99 data set,” in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl. (CISDA)*, 2009, pp. 1–6.

[16] N. V. Chawla et al., “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[17] D. M. W. Powers, “Evaluation: From precision, recall and F-measure to ROC,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[18] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[19] M. Bhagat and B. Bakariya, “A comprehensive review of cross-validation techniques in machine learning,” *Int. J. Sci. Technol.*, vol. 16, no. 1, pp. 1–4, 2025.

[20] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[21] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[22] J. Read, B. Pfahringer, G. Holmes, and E. Frank, “Classifier chains for multi-label classification,” *Mach. Learn.*, vol. 85, no. 3, pp. 333–359, 2011.

[23] S. Dua and X. Du, “Decision Tree Based Intrusion Detection System for Network Security,” in *Proc. IEEE Int. Conf. Inf. Technol.*, 2011, pp. 229–234.

[24] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward a lifetime of datasets for network-based intrusion detection systems,” in *Proc. Int. Conf. Netw. Inf. Syst. Secur. (NISS)*, 2018, pp. 1–13.

[25] X. Zhang, Y. Wang, and Z. Han, “A Secure and Robust Machine Learning Model for Intrusion Detection in Internet of Vehicles,” *IEEE Access*, vol. 9, pp. 112345–112358, 2021.

[26] B. Zadrozny, J. Langford, and N. Abe, “Cost-sensitive learning by cost-proportionate example weighting,” in *Proc. IEEE Int. Conf. Data Min. (ICDM)*, 2003, pp. 435–442.

[27] A. K. Das, S. B. Kim, and M. Park, “Adversarial Threat Modeling for Large-Scale AI Systems in Enterprise Networks,” *IEEE Secur. Privacy*, vol. 19, no. 5, pp. 45–53, 2021.

[28] M. S. Haroon and H. M. Ali, “Adversarial Training Against Adversarial Attacks for Machine Learning-Based Intrusion Detection Systems,” *Comput. Mater. Continua*, vol. 73, no. 2, pp. 3513–3536, 2022.

[29] S. H. Kang, J. H. Lee, and K. R. Park, “A Comprehensive Survey on Secure and Robust Machine Learning for Cybersecurity,” *ACM Comput. Surv.*, vol. 54, no. 6, Art. no. 128, 2022.

[30] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[31] A. Jobin, M. Ienca, and E. Vayena, “The Global Landscape of AI Ethics Guidelines,” *Nature Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019.

[32] M. Steinhardt, B. Koh, and P. Liang, “Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning,” *Found. Trends Mach. Learn.*, vol. 14, 2021.

[33] M. Alsharaiah et al., “Assimilate Grid Search and ANOVA Algorithms into KNN to Enhance Network Intrusion Detection Systems,” *J. Appl. Data Sci.*, vol. 6, no. 3, 2022.

[34] A. Ferrag and L. Maglaras, “A Comprehensive Survey on Cybersecurity Intrusion Detection Systems Using Machine Learning,” *Appl. Sci.*, vol. 12, no. 5, Art. no. 2351, 2022.

[35] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[36] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, “Concept-Based Explainable Artificial Intelligence: A Survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–23, 2023.

[37] M. Ferrag, L. Maglaras, A. Derhab, and H. Janicke, “Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study,” *J. Inf. Secur. Appl.*, vol. 50, 2020.

[38] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in *Proc. IEEE Int. J. Conf. Neural Netw. (IJCNN)*, 2008, pp. 1322–1328.

[39] S. Douzas and F. Bacao, “Effective Data Generation for Imbalanced Learning Using GANs,” *Expert Syst. Appl.*, vol. 91, pp. 464–471, 2018.

[40] L. Theodorakopoulos et al., “Big Data-Driven Distributed Machine Learning for Scalable Credit Card Fraud Detection Using XGBoost and CatBoost,” *Electronics*, vol. 14, 2025.

[41] M. A. Aydin, A. H. Zaim, and K. G. Ceylan, “An Enhanced Intrusion Detection System Using Machine Learning Techniques,” *Procedia Comput. Sci.*, vol. 113, 2017.

[42] M. M. Abdullahi et al., “Enhancing the Efficiency of Gaussian Naïve Bayes Classifier in the Detection of DDoS,” *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, 2022.

[43] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.

[44] A. Alsaedi et al., “Explainable AI for URL Threat Detection in Healthcare Cybersecurity: A Case Study Using LIME and SHAP,” *IEEE Access*, vol. 11, 2023.

[45] S. M. Salim et al., “Improved Network Intrusion Detection Using Hybrid Machine Learning Models,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, 2020.

[46] G. E. Hinton and R. Salakhutdinov, “Reducing Dimensionality with Neural Networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[47] G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.

[48] L. Prokhorenkova et al., “CatBoost: Unbiased boosting with categorical features,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018.

[49] P. Sharma et al., “Adversarial Machine Learning for Security: Experimental Techniques for Defending Against AI-Powered Cyberattacks,” *J. Sci. Innov. Adv. Res.*, vol. 1, 2025.

[50] M. Z. Alom, V. Bontupalli, and T. M. Taha, “Intrusion Detection Using Deep Learning: A Survey,” *IEEE Access*, vol. 8, 2020.

[51] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765–4774.

[52] Z. Salah and E. A. Elsoud, “Multi-Aspects AI-Based Modeling and Adversarial Learning for Cybersecurity Intelligence and Robustness,” *Preprints*, 2025.

[53] R. R. Singh, “Explainable Artificial Intelligence for Intrusion Detection Systems,” M.Sc. thesis, National College of Ireland, Dublin, 2024.

[54] A. Aljawarneh et al., “The Role of Ensemble Learning in Strengthening Intrusion Detection Systems,” *Computers & Security*, vol. 109, 2021.

[55] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD*, 2016, pp. 1135–1144.

[56] J. Lyu et al., “LD-SMOTE: A Novel Local Density Estimation-Based Oversampling Method for Imbalanced Datasets,” *Symmetry*, vol. 17, no. 2, 2025.

[57] A. S. Alqahtani et al., “An Efficient Intrusion Detection System Using Deep Learning Techniques,” *Sensors*, vol. 22, no. 4, 2022.

[58] M. Abdar et al., “A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges,” *Inf. Fusion*, vol. 76, 2021.

[59] A. Jolfaei, P. Ostovari, and R. K. Wong, “An Explainable Intrusion Detection System Based on Ensemble Learning,” *Symmetry*, vol. 13, no. 9, 2021.

[60] J. Lyu et al., “Advanced Oversampling Techniques for Imbalanced Learning: A Review,” *Symmetry*, vol. 17, no. 1, 2025.

[61] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning,” in *Proc. Int. Conf. Intell. Comput.*, 2005, pp. 878–887.

[62] L. Jingsha et al., “A Novel Botnet Detection Model Based on CatBoost and SHAP Explainable AI Framework,” *IEEE Trans. Netw. Service Manag.*, 2024.

[63] M. Alsharaiah et al., “Assimilating grid search and ANOVA algorithms into KNN to enhance network intrusion detection systems,” *J. Appl. Data Sci.*, vol. 6, no. 3, pp. 1469–1481, 2025.

[64] S. Ness et al., “Anomaly detection in network traffic using advanced machine learning techniques,” *IEEE Access*, vol. 13, pp. 16133–16149, 2025.

[65] C. K. K. Reddy et al., “Twined ensemble framework for network security,” *Discover Internet Things*, vol. 5, art. no. 107, 2025.

[66] R. Panigrahi and S. Borah, “A detailed analysis of CICIDS2017 dataset for designing intrusion detection systems,” *Int. J. Eng. Technol.*, vol. 7, no. 2.8, pp. 479–482, 2018.

[67] A. V. Afgan, “Legal Mechanisms of Audit and Accountability in Artificial Intelligence Systems in the Context of International Law,” *Sci. Bull. Uzhhorod Natl. Univ.*, no. 90, pp. 59–75, 2025.

[68] S. Salloum et al., “XAI-IDS: A Transparent and Interpretable Framework for Robust Cybersecurity Using Explainable Artificial Intelligence,” *SHIFFA*, vol. 2025, pp. 69–80, 2025.

[69] O. Senge et al., “Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty,” *Inf. Sci.*, vol. 589, pp. 453–474, 2022.

[70] S. K. Sahu et al., “CatBoost-IDS: An Optimized Intrusion Detection System for Cloud Security Using CatBoost Algorithm,” *J. Cybersecur. Inf. Manag.*, 2024.

[71] R. R. Singh, “Explainable Artificial Intelligence for Intrusion Detection Systems,” M.Sc. thesis, School of Computing, National College of Ireland, Dublin, Ireland, 2024.

[72] D. Berman, A. Buczak, J. Chavis, and C. Corbett, “A survey of deep learning methods for cyber security,” *Information*, vol. 10, no. 4, p. 122, 2019.

[73] P. Sharma et al., “Adversarial Machine Learning for Security: Experimental Techniques for Defending Against AI-Powered Cyberattacks,” *J. Sci. Innov. Adv. Res.*, vol. 1, no. 1, pp. 1–12, 2025.

[74] H. Alazzam, M. Shurman, and S. Al-Zoubi, “Ensemble-based intrusion detection systems for imbalanced datasets,” *Sensors*, vol. 23, no. 3, art. no. 1560, 2023.

[75] M. Shurman et al., “Machine learning-based intrusion detection systems: A survey,” *IEEE Access*, vol. 11, pp. 48912–48935, 2023.

[76] A. R. Al-Qerem et al., “A Hybrid Intrusion Detection System Using Logistic Regression and Recursive Feature Elimination on NSL-KDD Dataset,” *Int. J. Commun. Netw. Inf. Secur.*, 2023.

[77] M. J. M. Chowdhury et al., “An Ensemble Approach for Balanced and Imbalanced Intrusion Detection Datasets: A Comparative Study,” *IEEE Access*, vol. 11, pp. 24567-24582, 2023.

[78] M. M. Abdullahi et al., “Enhancing the Efficiency of Gaussian Naïve Bayes Classifier in the Detection of DDoS in Cloud Computing,” *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 10, 2022.

[79] A. J. Joshi, "Performance Metrics for Artificial Intelligence in Cyber Security," *IEEE Access*, vol. 10, pp. 1240-1255, 2022.

[80] European Commission, “10 Things You Should Know About the EU Artificial Intelligence Act,” Brussels, Belgium, 2023. [Online]. Available: https://digital-strategy.ec.europa.eu

# Ekler

| **Dataset:** Malware Analysis Dataset |
| --- |
| https://www.kaggle.com/code/adepvenugopal/malware-datasets?select=malware+analysis.csv |

| **Kodlar:** Coding and Results |
| --- |
| https://github.com/jkhammadov/Thesis/blob/main/Malware%20Analysis.ipynb |