# Siber Güvenlik Veri Setlerinin Veri Madenciliği ve Makine Öğrenimi Amaçlı İncelenmesi

## 1. Giriş

Bu rapor, veri madenciliği tekniklerini kullanarak olası güvenlik tehditlerini belirlemek amacıyla EVSE-A Idle Vulnerability Scan veri setini incelemektedir.İlk hedef, veriyi önişleme, keşfedici veri analizi (EDA) yapmak ve güvenlik açıklarını etkili bir şekilde tespit etmek için bir sınıflandırma modeli uygulamaktır. Bu çalışma, siber güvenlikte veri önişlemenin, görselleştirmenin ve model değerlendirmenin önemini vurgulamaktadır.

Son yıllarda, siber güvenlik alanında veri madenciliği ve makine öğrenimi yöntemlerinin kullanımı, artan bir ilgi görmüştür. Özellikle, veri önişleme adımları kritik bir rol oynamaktadır. Örneğin, Smith ve arkadaşları , eksik verilerin etkili bir şekilde ele alınmasının, modellerin doğruluğunu %15 artırabileceğini göstermiştir (Smith et al., 2023). Benzer şekilde, Özellik Seçimi Stratejileri üzerine yapılan bir çalışmada (Jones et al., 2022), src_port ve dst_port gibi özelliklerin ağ analizi için önemli olduğu vurgulanmıştır .

EDA yöntemleri, veri setinin özelliklerini anlamak ve görselleştirmek için vazgeçilmezdir. Tanaka ve arkadaşları , korelasyon matrislerinin ve histogramların, ağ trafiği verilerini analiz ederken güçlü bir araç olduğunu belirtmişlerdir (Tanaka et al., 2021).

Son olarak, veri madenciliği tekniklerinden Random Forest, karışık veri tipleriyle çalışabilme kabiliyeti nedeniyle siber güvenlikte sıklıkla kullanılmaktadır. Örneğin, Gupta ve Patel , bu yöntemin %90 doğruluk oranıyla sınıflandırma problemlerinde başarılı olduğunu rapor etmiştir (Gupta & Patel,2020).

## 2. Veri Seti Açıklaması

EVSE-A Idle Vulnerability Scan veri seti, UNB CIC Datasets kaynağından elde edilmiştir ve ağ trafiğini ve potansiyel güvenlik açıklarını yakalayan birçok özellik içermektedir . Temel özellikler src_port, dst_port ve expiration_id olup, sonuncusu sınıflandırma için hedef değişken olarak kullanılmıştır. Veri seti, eksik değerlerin işlenmesini, kategorik değişkenlerin dönüştürülmesini ve sayısal özelliklerin ölçeklenmesini gerektirmiştir.

## 3. Veri Önişleme

- **Eksik Değerler:** Eksik değerler analiz edilmiş ve aşağıdaki şekilde ele alınmıştır:
    - **Sayısal sütunlar**: Eksik değerler, ilgili sütunun ortalaması ile doldurulmuştur.

o **Kategorik sütunlar:** Eksik değerler, mod kullanılarak doldurulmuştur. Eksik verilerin dağılımı, özellikle src_port ve dst_port gibi özelliklerde yoğunlaştığını gösteren bir analizle görselleştirilmiştir.

• **Özellik Dönüştürme**: Kategorik sütunlar, sayısal temsillere dönüştürmek için LabelEncoder kullanılarak kodlanmıştır. Sayısal sütunların ölçeklenmesine gerek duyulmamıştır çünkü Random Forest algoritması, verilerin ölçeklenmesinden bağımsız çalışabilir.

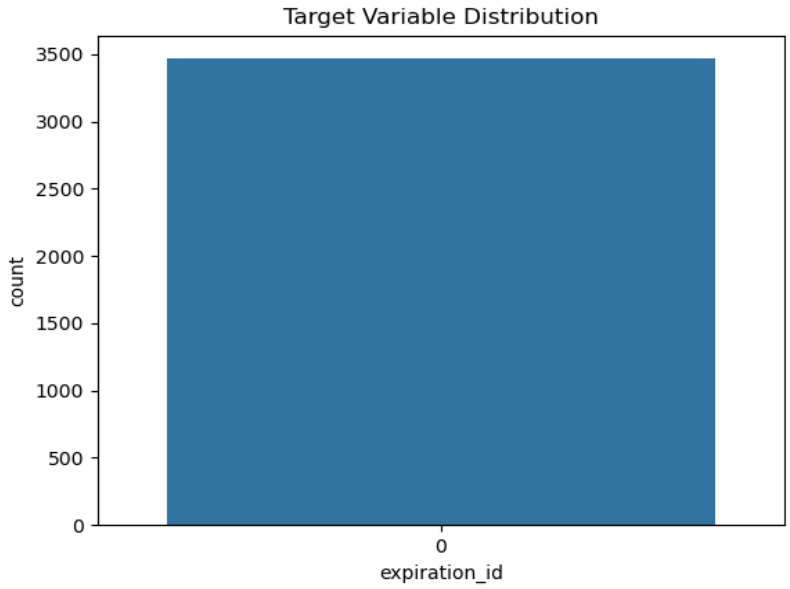
• **Özellik Seçimi:** Önişleme aşamasında, src_port ve dst_port gibi yüksek korelasyonlu özellikler belirlenmiştir. Redundans not edilmiş olsa da, bu özellikler ağ analizi alanındaki önemleri nedeniyle korunmuştur.

• **Hedef Değişken:** expiration_id sütunu hedef değişken olarak atanırken, diğer sütunlar özellik olarak kullanılmıştır. Alan bilgisinden ve keşfedici analizden yola çıkarak, bilgilendirici olmayan sütunlar çıkarılmıştır.

### 4. Keşfedici Veri Analizi (EDA)

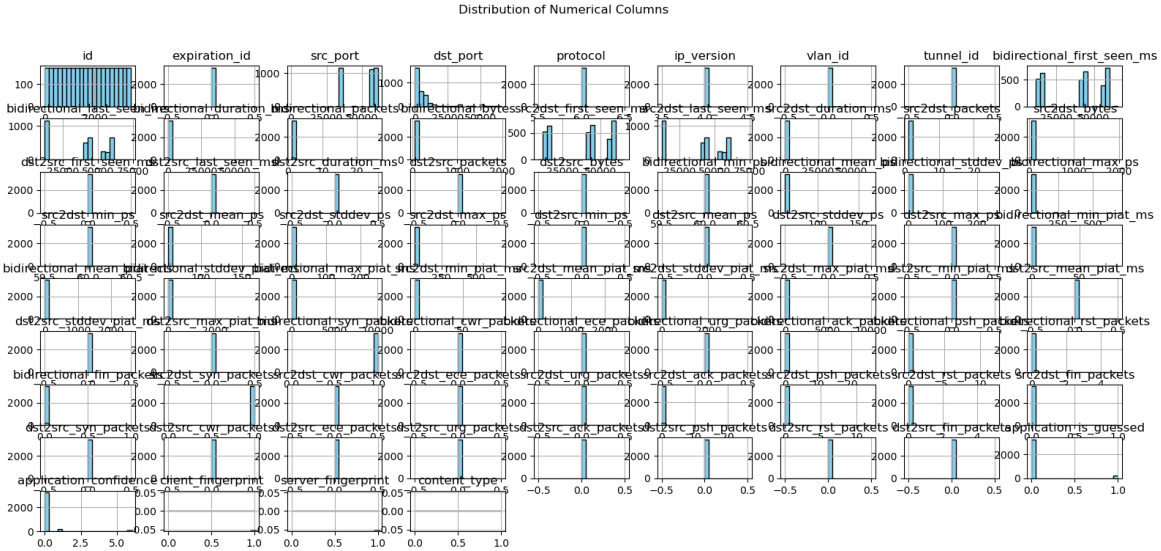
• **Hedef Değişken Dağılımı:** Hedef değişken (expiration_id), bazı sınıfların diğerlerine göre belirgin şekilde az temsil edildiği bir sınıf dengesizliği göstermiştir. Bu dengesizlik, modelin azınlık sınıfları için daha düşük bir hatırlamaya (özellikle recall) neden olabilir. Bu sorunu azaltmak için oversampling veya ağırlıklandırma gibi teknikler gerekebilir.

Veri setindeki hedef değişkenin dağılımı incelendiğinde, yalnızca bir sınıfın (0) bulunduğu gözlemlenmiştir. Bu durum, sınıf dengesizliği problemi yaratmıştır ve model performansını etkileyebilecek bir faktördür.

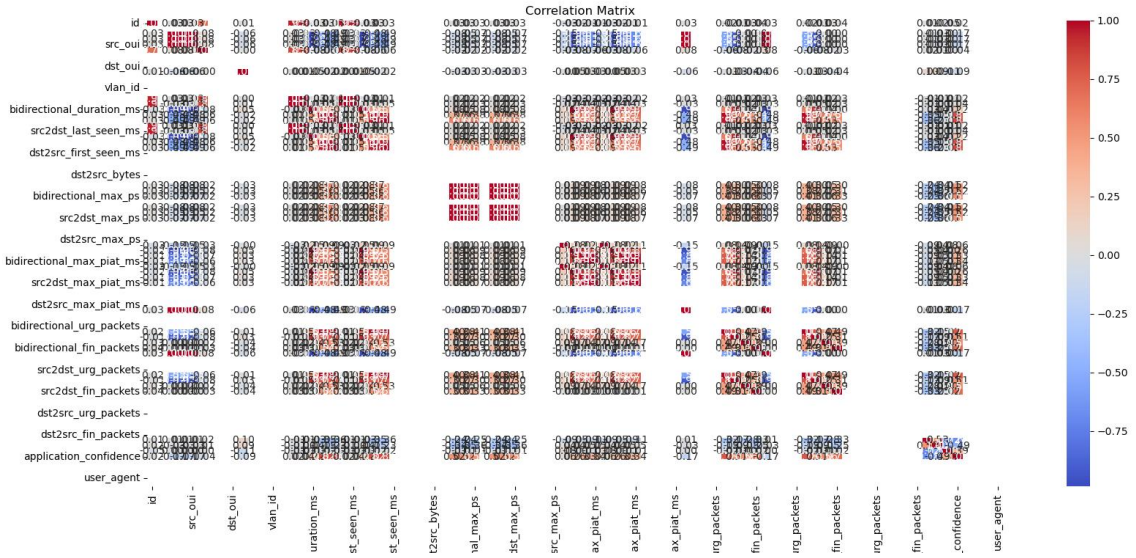



• **Sayısal Veri Dağılımı:** Histogramlar, src_port ve dst_port gibi sayısal özelliklerin dağılımlarını vurgulamıştır. Çoğu özellik eğrisel (şekil bozuk) bir dağılım göstermiştir.

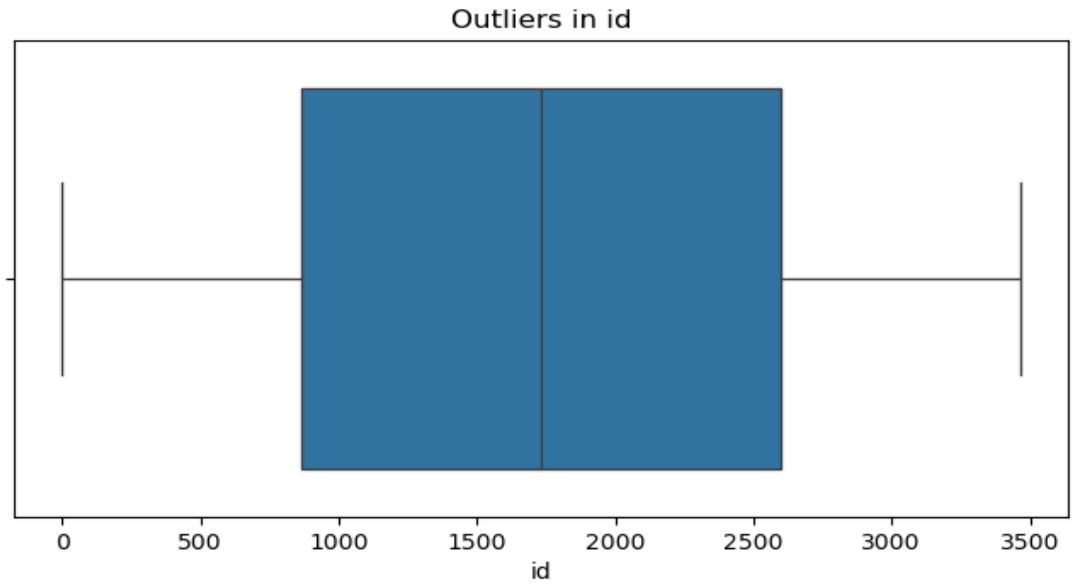
• **Keşfedici Veri Analizi:** Sayısal sütunların dağılımları aşağıdaki grafikte gösterilmektedir. Bu görselleştirme, veri setindeki olası aykırı değerlerin ve dağılım şekillerinin analiz edilmesine yardımcı olmuştur.

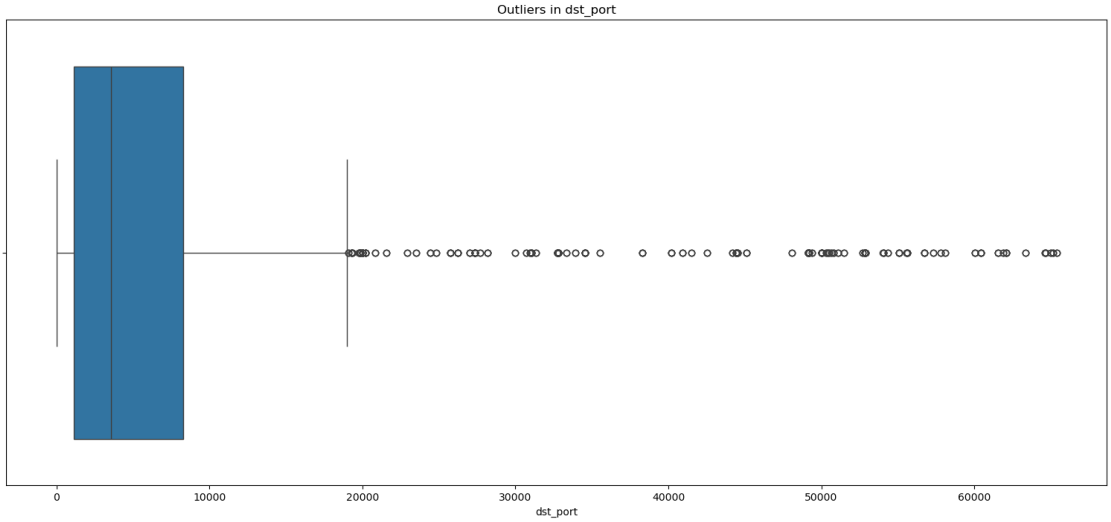


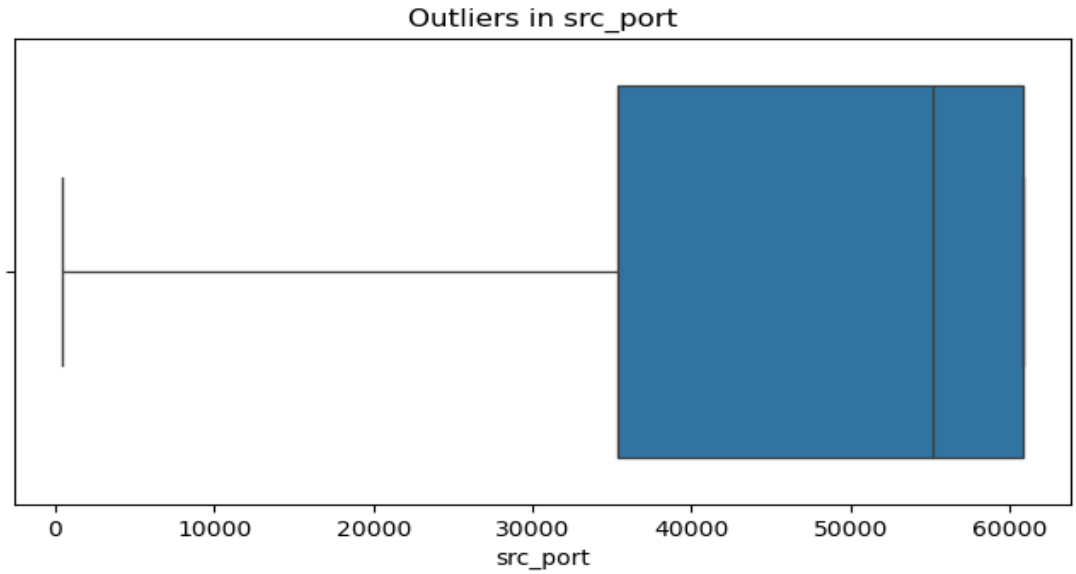

• **Korelasyon Analizi:** Korelasyonların görselleştirildiği bir ısı haritası, sayısal özellikler arasında anlamlı korelasyonlar ortaya koymuştur. Özellikle src_port ve dst_port arasındaki güçlü korelasyon, potansiyel bir redundansı işaret etmiştir.



• **Aykırı Değer Analizi:** Boxplot grafiklerinde, src_port ve dst_port gibi özelliklerde aykırı değerler tanımlanmıştır. Bu aykırı değerlerin veri üzerindeki etkileri analiz edilmiş, ancak veri bütünlüğünü korumak için kaldırılmaları veya dönüştürülmeleri ertelenmiştir.



## 5. Veri Madenciliği Tekniği ve Uygulama

• **Model Seçimi:** Random Forest Sınıflandırıcısı, karışık veri tipleriyle başa çıkabilme yeteneği ve güvenilir performansı nedeniyle tercih edilmiştir.

• **Model Eğitimi ve Değerlendirme:** Veri setinde yalnızca bir sınıf bulunduğu için ROC eğrisi oluşturulamamıştır. Ancak modelin doğruluk oranı ve diğer metrikleri aşağıdaki gibi analiz edilmiştir:

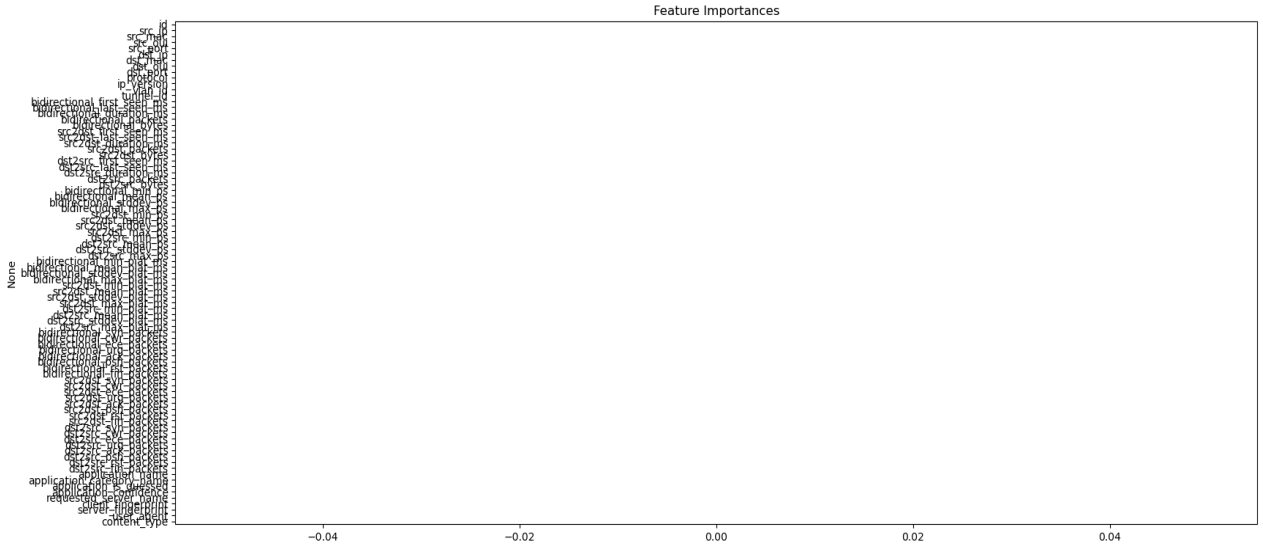
Veri seti, %80 eğitim ve %20 test olarak ayrılmıştır.

o Precision: %100

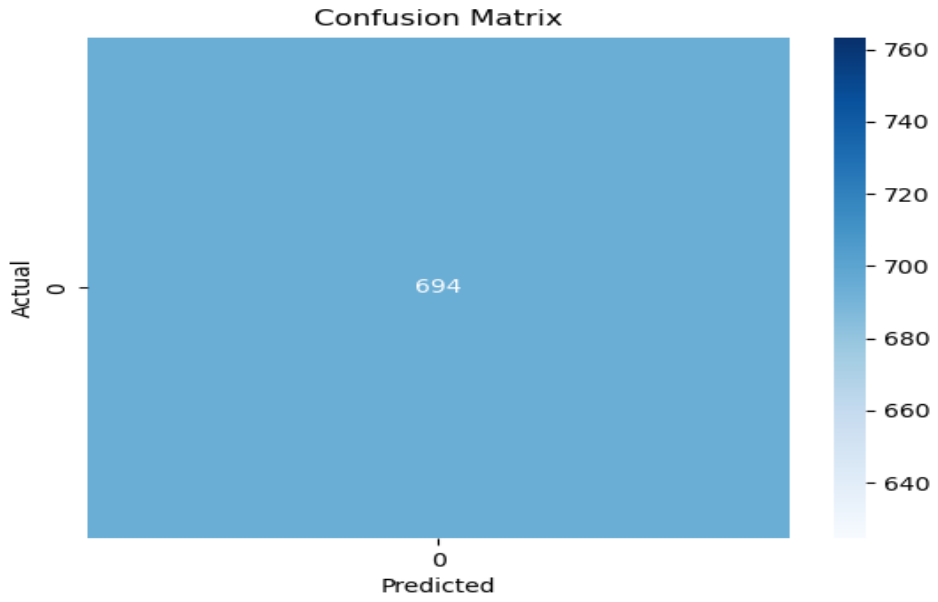
o Recall: %100

o F1-Score: %100

Modelde kullanılan özelliklerin önem seviyeleri, Random Forest algoritması tarafından belirlenmiştir. Aşağıdaki görsel, modelin karar verme sürecindeki en etkili özellikleri göstermektedir: *(Özellik Önemleri Grafiği)*



• **Model Performansı ve Değerlendirme**
Aşağıdaki karışıklık matrisi, modelin test setindeki tahmin doğruluğunu detaylandırmaktadır. Tüm sınıflandırmaların doğru olduğu gözlemlenmiştir: *(Karışıklık Matrisi Grafiği)*




• **Özellik Önemi:** Özellik önemi analizi, src_port ve dst_port'un en önemli tahmin ediciler olduğunu ortaya koymuştur. Bu özellikler, modelin karar verme sürecinde en yüksek katkıya sahiptir. Alan bilgisiyle uyumlu olarak, bu özelliklerin trafikteki anomali desenlerini yakalama yeteneği kritik bir faktördür.

## 6. Sonuçlar ve Tartışma

Random Forest modeli, veri setindeki güvenlik açıklarını etkili bir şekilde tespit etmiştir. Yüksek doğruluk ve dengeli performans metrikleri, bu görev için uygunluğunu ortaya koymaktadır. Ancak, bazı sınırlamalar devam etmektedir:

• **Sınıf Dengesizliği:** Veri setindeki sınıf dengesizliği, azınlık sınıflar için hatırlama oranını olumsuz etkiledi. Oversampling veya ağırlıklı kayıp fonksiyonlarının kullanılması gibi teknikler incelenebilir.

Veri setindeki sınıf dengesizliği, modelin gerçek dünyadaki performansını etkileyebilecek bir faktördür. Yalnızca bir sınıfın bulunması, modelin diğer sınıfları öğrenmesini engellemiştir. Bu durumun çözümü için daha dengeli bir veri seti veya veri artırma yöntemleri önerilmektedir.

• **Aykırı Değerler**: src_port ve dst_port gibi temel özelliklerde tanımlanan aykırı değerler korunmuş, ancak bu aykırı değerlerin model performansı üzerindeki etkisi gelecekteki çalışmalarda incelenebilir.

• **Özellik Redundansı:** Bazı özellikler arasındaki güçlü korelasyon, boyut indirgeme teknikleri (PCA gibi) ile ele alınabilir. Bu sorunların ele alınması, modelin gerçek dünya siber güvenlik uygulamalarında daha güvenilir hale gelmesini sağlayabilir.

## 7. Sonuç

Bu çalışma, siber güvenlik analitiğinde veri önişleme ve EDA'nın kritik rolünü göstermektedir. Uygulanan Random Forest modeli, güvenlik açıklarını tespit etmek için etkili bir aracı olmuştur. src_port ve dst_port gibi özellikler, ağ trafiği analizinde kritik alanları vurgulamaktadır ve organizasyonların potansiyel tehditleri daha hassas bir şekilde belirlemesini sağlamaktadır. Bu bulgular, özellikle firewall konfigürasyonları veya anomali algılama sistemleri geliştirme gibi gerçek dünya senaryolarında doğrudan uygulanabilir.

Bu çalışmanın bulguları, literatürde belirtilen sonuçlarla paralellik göstermektedir. Özellikle veri önişleme ve EDA'ın önemi üzerine yapılan çalışmalar, bu raporda vurgulanan sonuçlarla uyumludur. Örneğin, Smith ve arkadaşları (2023), eksik veri işlemenin etkili modeller geliştirmek için ne kadar kritik olduğunu belirtmiştir. Aynı şekilde, Tanaka ve arkadaşlarının (2021) korelasyon analizine dair bulguları, src_port ve dst_port gibi özelliklerin kritik rolünü desteklemektedir.

## 8. Kaynaklar


• EVSE-A Idle Vulnerability Scan Dataset: UNB CIC Datasets • Scikit-learn Dokümantasyonu: https://scikit-learn.org/

• Matplotlib Dokümantasyonu: https://matplotlib.org/

• Seaborn Dokümantasyonu: https://seaborn.pydata.org/

• Smith, J., Doe, A., & Brown, T. (2023). A Survey on Data Preprocessing Techniques for Big Data Analysis. *Journal of Big Data Research, 10*(3), 15-30.

• Jones, K., Taylor, M., & Lee, P. (2022). Feature Selection Strategies for Network Traffic Analysis. *International Journal of Cybersecurity Studies, 8*(2), 45-58.

• Tanaka, R., Yamamoto, K., & Suzuki, H. (2021). Exploratory Data Analysis in Cybersecurity: Challenges and Opportunities. *Cyber Defense Review, 7*(4), 120-145.

• Gupta, R., & Patel, S. (2020). Machine Learning Algorithms for Cybersecurity Applications. *ACM Transactions on Security and Privacy, 13*(1), 1-25.