Siber Güvenlik İçin Veri Madenciliği
-Ara Sınav Ödevi-

Bölüm 1: Veri Seti Seçimi

CICIDS2017 adlı veri seti, ağ güvenliği ve siber saldırı tespiti alanında kullanılan bir veri setidir. Bu veri seti, ağ trafiğini izlemek için kullanılmıştır ve farklı günlerdeki farklı ağ aktivitelerini içermektedir.

Bu veri setinin temel amaçlarından bazıları şunlar olabilir:

1. **Siber Saldırıları İzleme ve Tanımlama:** Pazartesi hariç olmak üzere salı, çarşamba, perşembe ve cuma günlerinde farklı türde siber saldırılar gerçekleştirilmiştir. Bu saldırılar, örneğin, Brute Force, DoS/DDoS, Heartbleed, Web Saldırıları gibi çeşitli siber saldırıları içermektedir.
2. **Benign Ağ Trafiği ile Kıyaslama:** Pazartesi günü sadece normal ağ aktivitelerini içeren veri, diğer günlerde gerçekleştirilen saldırıların etkisini daha net bir şekilde anlamak için karşılaştırma yapmak üzere kullanılabilir.
3. **Ağ Trafiğinin Profilenmesi ve Analizi:** Veri seti, ağ trafiğini analiz etmek için CICFlowMeter adlı bir araç kullanılarak etiketlenmiş akışlar şeklinde sunulmuştur.
4. **Makine Öğrenimi ve Derin Öğrenme Amaçlı CSV Dosyaları:** Veri seti, makine öğrenimi ve derin öğrenme gibi yapay zeka teknikleri için CSV dosyaları içermektedir.

Veri setinin kaynağı, Iman Sharafaldin, Arash Habibi Lashkari ve Ali A. Ghorbani'nin 4th International Conference on Information Systems Security and Privacy (ICISSP), Portekiz, Ocak 2018'de sundukları bir araştırma çalışmasına dayanmaktadır. Veri setinin amacı, güvenlik uzmanları, araştırmacılar ve geliştiricilerin siber güvenlik sistemlerini eğitmek, test etmek ve geliştirmek için güvenilir bir veri kaynağı sunmaktır.

Bu veri seti, farklı saldırı türlerini içermesi ve gerçek dünya ağ trafiği benzeri olmasıyla, siber güvenlik alanında modelleme, eğitim ve test amaçlı kullanılabilir. Özellikle, bu tür veri setleri siber güvenlik araştırmaları ve siber güvenlik uygulamaları için algoritmaların geliştirilmesinde önemli bir kaynak olabilir.

Bölüm 2: Veri Ön İşleme

Siber güvenlik veri setlerinde veri ön işleme kritik önem taşır çünkü:

1. **Veri Temizliği:** Hatalı, eksik verilerin düzeltilmesi analiz doğruluğunu artırır.
2. **Özellik Seçimi:** Gereksiz özelliklerin kaldırılması, model performansını artırır.
3. **Gizlilik Koruma:** Kişisel verilerin gizliliği sağlanır.
4. **Veri Dengeleme:** Dengesizlik, yanlış sınıflandırmayı azaltmak için düzeltilir.
5. **Veri Formatlama:** Doğru yapılandırma, saldırı tespiti ve önleme için gereklidir.

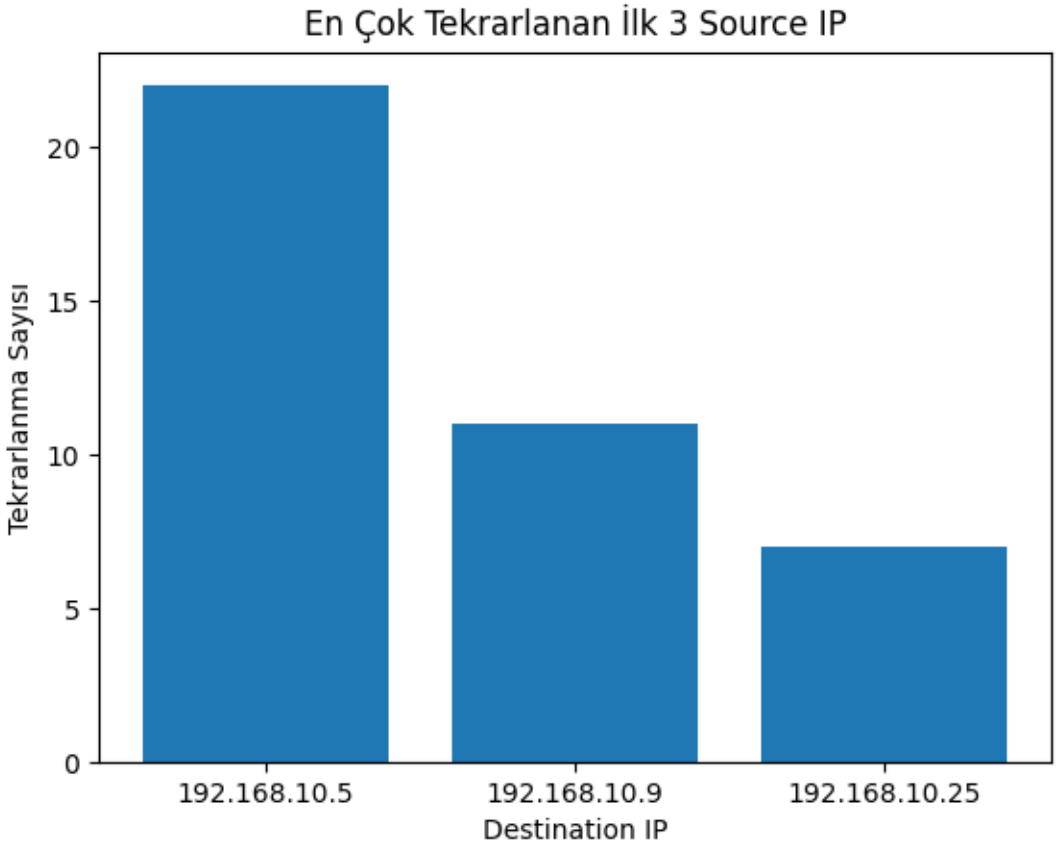
Bu adımlar, güvenlik analistlerinin daha doğru sonuçlar elde etmelerini sağlar.

Bölüm 3: Keşif Amaçlı Veri Analizi (EDA)

Bir bütün olarak ele aldığımız veri setini aşamalar halinde değerlendirdik. Kaynaktan çıkan atakların çıktığı kaynağın, portun tespiti, hangi zamanda gerçekleştiği, hangi hedefe ulaştığı ve hangi port üzerinden işlem gerçekleştirildiği gibi bilgileri bölümlere ayırarak farklı tablolar üzerinde anlattık.

İncelediğimiz veri setinin analizinde ilk aşama olarak hedef(destination) IP'lere yönelik yapılan saldırıların nitelik olarak sayısı ve en yoğun saldırıların olduğu IP adreslerini saptamamız bulunuyor. Yazdığımız analiz hangi hedef IP'sine kaç atak bulunduğunu tespit edip grafik üzerinde

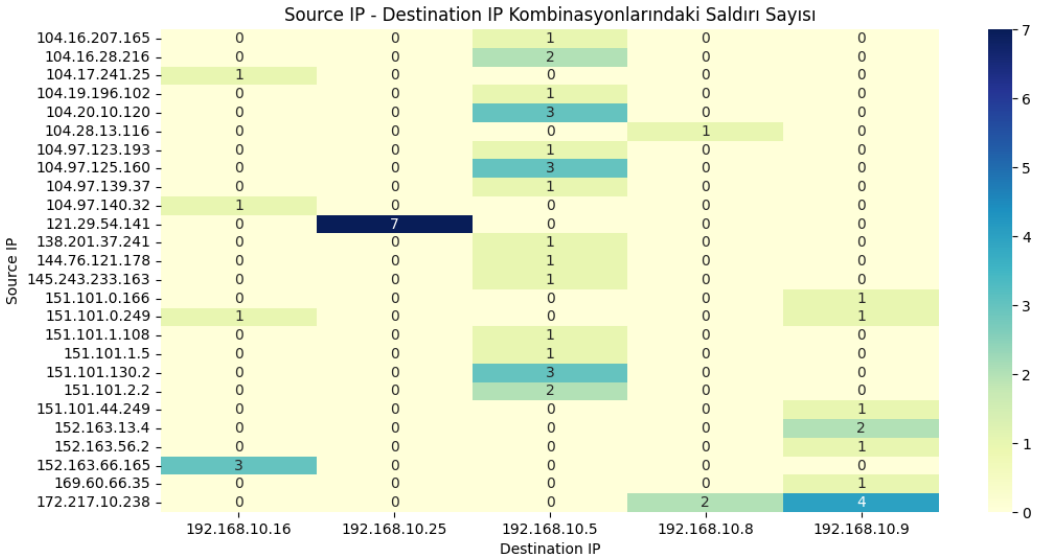
belirtilmesini sağlıyor.



(Figüre 1.1)

Figür 1.1'de görünğü gibi en çok atak gerçekleştirilen IP adresleri belirlenmiş ve riskli görülen IP'ler saptanmıştır.

Kodun ikinci aşamasında incelediğimiz kısımda kaynak IP'den hedef IP'ye gerçekleştirilen atakların sayısının saptanması sayesinde atak yoğunluğu belirlenmiştir. Bu sayede en yoğun saldırıda bulunan IP adreslerinin en yoğun hedef aldığı IP adresleri saptanabilmektedir. Örnek olarak 121.29.54.141 numaralı IP'ye sahip saldırgan 192.168.10.25 IP adresine sahip hedefe toplam yedi saldırı düzenlemesi elimizdeki verilere dayanarak en zayıf halkayı tespit etmemizi sağladı.



(Figür 1.2)

Figüre 1.2'de göründüğü üzere kaynak IP ve hedef IP adresleri arasında gerçekleşen saldırı kombinasyonları belirgindir ve hangi hedeflerin hangi kaynaklara kaç defa atak gerçekleştirdiği gibi sorulara cevap vermektedir.

Ve bu verilere dayanarak diğer olasılıkların hesaplanmasıyla tüm atak ihtimalleri belirlenip en yoğun olarak atak alan konularda öngörü oluşturularak sistem güvenliği sağlanabilir.