Veri Madenciliği Tekniği Seçimi:

Veri kümesini analiz etme sürecinde, analiz amacına uygun bir veri madenciliği tekniği seçildi. Bu süreçte sınıflandırma tekniği tercih edildi. Sınıflandırma, veri kümesindeki örnekleri belirli kategorilere sınıflandırmak amacıyla kullanılan etkili bir tekniktir.

Teknik Seçiminin Gerekçesi ve Siber Güvenlikle İlgisi:

Sınıflandırma tekniğinin tercih edilme gerekçeleri şu şekildedir:

- **Çeşitli Saldırı Türlerinin Sınıflandırılması:** Veri setindeki ağ trafiği örneklerini normal ve anormal aktiviteler olarak sınıflandırarak farklı saldırı türlerini belirlemek.
- **Tespit ve Önleme Yeteneği:** Sınıflandırma, zararlı aktiviteleri tespit ederek siber güvenlik önlemlerini geliştirmek ve uygulamak için kullanılabilir.
- **Hızlı Karar Alma:** Sınıflandırma, gerçek zamanlı veri analizi ile anında müdahale imkanı sağlayarak güvenlik önlemlerini optimize etme avantajı sunabilir.

Bu sınıflandırma tekniği ile, veri setindeki çeşitli etkinlikleri tanımlayarak siber güvenlik tehditlerini daha etkili bir şekilde ele almayı hedefliyoruz. Beklenen sonuçlar, belirli saldırı türlerini doğru bir şekilde sınıflandırarak bu bilgiyi güvenlik önlemlerini güçlendirmek için kullanmaktır. Ayrıca, modelin doğruluğunu ve performansını değerlendirerek sınıflandırma işleminin etkinliğini ölçeceğiz.

Güvenlik Tehdidi Tespiti Uygulama:

Seçilen sınıflandırma tekniği, veri seti içindeki potansiyel güvenlik tehditlerini ve anormallikleri tespit etmek amacıyla uygulandı. Bu aşamada, model, ağ trafiği verilerini analiz ederek normal ve anormal aktiviteleri sınıflandırdı.

Sonuçlar:
KNN (K-Nearest Neighbors) Sınıflandırma Tekniği Kullanılarak Elde Edilen Sonuçlar Şu Şekildedir:

```
KNN Doğruluk Oranı: 0.9944775668324313
KNN Sınıflandırma Raporu:
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_class
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_class
  _warn_prf(average, modifier, msg_start, len(result))
                   precision    recall  f1-score   support

            back.       1.00      0.99      0.99       254
buffer_overflow.       1.00      0.57      0.73         7
       ftp_write.       0.00      0.00      0.00         3
    guess_passwd.       1.00      0.94      0.97        18
            imap.       1.00      1.00      1.00         2
         ipsweep.       0.83      0.95      0.89       170
            land.       1.00      0.67      0.80         3
         neptune.       1.00      1.00      1.00     12923
            nmap.       0.55      0.23      0.32        48
          normal.       1.00      1.00      1.00     21980
            perl.       0.00      0.00      0.00         1
             phf.       1.00      1.00      1.00         1
             pod.       1.00      1.00      1.00        47
       portsweep.       0.90      0.58      0.71       114
         rootkit.       0.00      0.00      0.00         1
           satan.       0.96      0.80      0.87       217
           smurf.       0.96      0.97      0.96       157
             spy.       0.00      0.00      0.00         1
        teardrop.       0.99      1.00      0.99       245
     warezclient.       0.96      0.98      0.97       201
     warezmaster.       1.00      1.00      1.00         4

         accuracy                           0.99     36397
        macro avg       0.77      0.70      0.72     36397
     weighted avg       0.99      0.99      0.99     36397
```

Doğruluk Oranı: KNN'nin doğruluk oranı 0.9944 olarak hesaplanmıştır, yani %99.44 başarı sağlamıştır.

Sınıflandırma Raporu:

- **Precision (Kesinlik):** Sınıflandırma raporunda belirtilen precision değerleri, tahmin edilen pozitif örneklerin gerçek pozitif örneklerle oranını gösterir. Örneğin, "normal" sınıfı için precision %100 olarak raporlanmıştır, yani "normal" olarak tahmin edilen örneklerin tamamı gerçekten "normal"dir.
- **Recall (Duyarlılık):** Recall değerleri, gerçek pozitif örneklerin tespit edilme oranını ifade eder. Örneğin, "ipsweep" sınıfı için recall değeri 0.95'tir, yani gerçek ipsweep örneklerinin %95'i doğru bir şekilde tespit edilmiştir.
- **F1-Score:** F1-score, precision ve recall'in harmonik ortalamasıdır ve bu iki metrik arasındaki dengeyi gösterir. Örneğin, "teardrop" sınıfının f1-score'u 0.99'dur, bu da precision ve recall arasında güçlü bir denge olduğunu gösterir.

Önemli Notlar:

- Bazı sınıflar için precision ve recall değerleri, özellikle az örnek içeren veya dengesiz sınıflandırma problemleri olan durumlarda sıfır olabilir. Bu durum, ilgili sınıfa ait örneklerin model tarafından doğru bir şekilde sınıflandırılamadığını gösterir.
- Özellikle "rootkit" ve "perl" gibi nadir sınıflar, az sayıda örnek içerdiğinden dolayı sınıflandırma metrikleri zayıf olabilir.

Random Forest Sınıflandırma Tekniği Kullanılarak Elde Edilen Sonuçlar Şu Şekildedir:

```
Random Forest Doğruluk Oranı: 0.9989834327004973
Random Forest Sınıflandırma Raporu:
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_class
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_class
  _warn_prf(average, modifier, msg_start, len(result))
                   precision    recall  f1-score   support

             back.       1.00      1.00      1.00       254
buffer_overflow.       1.00      0.57      0.73         7
       ftp_write.       0.00      0.00      0.00         3
    guess_passwd.       1.00      0.94      0.97        18
             imap.       0.67      1.00      0.80         2
          ipsweep.       0.98      1.00      0.99       170
             land.       1.00      0.67      0.80         3
          neptune.       1.00      1.00      1.00     12923
             nmap.       1.00      0.85      0.92        48
           normal.       1.00      1.00      1.00     21980
             perl.       0.00      0.00      0.00         1
              phf.       1.00      1.00      1.00         1
              pod.       1.00      1.00      1.00        47
        portsweep.       1.00      0.97      0.99       114
          rootkit.       0.00      0.00      0.00         1
            satan.       1.00      1.00      1.00       217
            smurf.       1.00      0.99      0.99       157
              spy.       0.00      0.00      0.00         1
         teardrop.       1.00      1.00      1.00       245
      warezclient.       0.97      0.97      0.97       201
      warezmaster.       0.80      1.00      0.89         4

          accuracy                           1.00     36397
         macro avg       0.78      0.76      0.76     36397
      weighted avg       1.00      1.00      1.00     36397
```

Doğruluk Oranı: Random Forest'ın doğruluk oranı 0.9990 olarak hesaplanmıştır, yani %99.90 başarı sağlamıştır.

Sınıflandırma Raporu:

- **Precision :** Örneğin, "normal" sınıfı için precision %100 olarak raporlanmıştır, yani "normal" olarak tahmin edilen örneklerin tamamı gerçekten "normal"dir.
- **Recall :** Örneğin, "ipsweep" sınıfı için recall değeri 1.00'dir, yani gerçek ipsweep örneklerinin tamamı doğru bir şekilde tespit edilmiştir.
- **F1-Score**:Örneğin, "teardrop" sınıfının f1-score'u 1.00'dur, bu da precision ve recall arasında güçlü bir denge olduğunu gösterir.

Önemli Notlar:

- "rootkit" ve "perl" gibi nadir sınıfların sınıflandırma metrikleri, az sayıda örnek içerdiğinden dolayı sıfır olabilir. Ancak, bu durum Random Forest'ın genel başarısını etkilememiştir.
- Random Forest, genel olarak yüksek doğruluk oranları ve sınıflandırma performansı ile öne çıkan güçlü bir sınıflandırma tekniğidir.
- Bu sonuçlara dayanarak, Random Forest sınıflandırma tekniğinin genel olarak yüksek başarı sağladığı ve çeşitli sınıfları başarılı bir şekilde tanımladığı görülmektedir. Bu teknik, özellikle dengesiz sınıflandırma problemleri için etkili bir çözüm sunabilir.

Lojistik Regresyon Sınıflandırma Tekniği Kullanılarak Elde Edilen Sonuçlar Şu Şekildedir:

```
Lojistik Regresyon Doğruluk Oranı: 0.9313404950957497
Lojistik Regresyon Sınıflandırma Raporu:
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_cl
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_cl
  _warn_prf(average, modifier, msg_start, len(result))
                   precision    recall  f1-score   support

            back.       0.00      0.00      0.00       254
buffer_overflow.       0.00      0.00      0.00         7
       ftp_write.       0.00      0.00      0.00         3
    guess_passwd.       0.00      0.00      0.00        18
            imap.       0.00      0.00      0.00         2
         ipsweep.       0.00      0.00      0.00       170
            land.       0.00      0.00      0.00         3
         neptune.       0.89      1.00      0.94     12923
            nmap.       0.00      0.00      0.00        48
          normal.       0.96      0.95      0.96     21980
            perl.       0.00      0.00      0.00         1
             phf.       0.00      0.00      0.00         1
             pod.       0.00      0.00      0.00        47
       portsweep.       0.00      0.00      0.00       114
         rootkit.       0.00      0.00      0.00         1
           satan.       0.00      0.00      0.00       217
           smurf.       0.00      0.00      0.00       157
             spy.       0.00      0.00      0.00         1
        teardrop.       0.00      0.00      0.00       245
     warezclient.       1.00      0.06      0.11       201
     warezmaster.       1.00      1.00      1.00         4

         accuracy                           0.93     36397
        macro avg       0.18      0.14      0.14     36397
     weighted avg       0.90      0.93      0.91     36397
```

Doğruluk Oranı: Lojistik Regresyon'un doğruluk oranı 0.9313 olarak hesaplanmıştır, yani %93.13 başarı sağlamıştır.

Sınıflandırma Raporu:

- **Precision:**Birçok sınıf için precision değeri 0.00 olarak raporlanmıştır, yani tahminlerin çoğu doğru değildir.
- **Recall**:Yine birçok sınıf için recall değeri 0.00'dir, yani gerçek pozitif örneklerin çoğu tespit edilememiştir.
- **F1-Score:**. Aynı şekilde,birçok sınıf için f1-score değeri 0.00 olarak raporlanmıştır.

Önemli Notlar:

- Lojistik Regresyon, bu veri kümesindeki dengesiz sınıflar ve karmaşıklık nedeniyle düşük başarı elde etmiştir. Ayrıca, "warezclient" sınıfındaki precision değeri oldukça yüksek görünmektedir, ancak bu sınıfın küçük bir örnek sayısı olabilir.

Uyarılar:

- "UndefinedMetricWarning" uyarıları, bazı sınıfların tahmin edilmemiş olması nedeniyle precision ve f1-score gibi değerlerin hesaplanamadığını belirtir. Bu durum, özellikle nadir sınıfların veya kötü temsil edilen sınıfların olduğu durumlarda karşımıza çıkabilir.

Lojistik Regresyon'un bu özel veri kümesindeki performansı, Random Forest gibi diğer sınıflandırma tekniklerine kıyasla düşüktür. Bu durum, veri setinin özellikleri, sınıf dengesizliği ve karmaşıklığı göz önüne alındığında değerlendirilmelidir.

Genel Değerlendirme Yapacak Olursak;

1. **K-En Yakın Komşular (KNN):** Yüksek doğruluk oranına sahip olmasına rağmen, bazı sınıflar için düşük precision ve recall değerleri dikkat çekmektedir. Bu, özellikle nadir sınıfların tespitinde zorluk yaşandığını gösterebilir.
2. **Rastgele Orman (Random Forest):** Yüksek doğruluk oranı ve genel olarak iyi performans sergilemiştir. Ancak, bazı sınıflar için düşük precision ve recall değerleri, modelin belirli durumlarda zorlandığını gösterir.
3. **Lojistik Regresyon:** Düşük doğruluk oranına sahiptir ve birçok sınıf için precision ve recall değerleri sıfırdır. Bu, modelin genel etkinliğinin düşük olduğunu gösterir.

Daha Önce Yaptığımız Literatürdeki Çalışmaların Etkisi:

Daha önce literatürde yapılan çalışmalarda, özellikle dengesiz veri kümeleri üzerindeki sınıflandırma zorluklarına odaklanmıştır. Benzer şekilde, bu çalışma da siber güvenlik veri kümesindeki dengesizlikle başa çıkmak ve nadir olayları doğru bir şekilde tanımlamak amacıyla çeşitli yöntemleri ele almıştır.Örneğin, literatürde öne çıkan bir içgörü, sınıflandırma modellerinin dengesiz veri kümelerindeki nadir sınıfları doğru bir şekilde öğrenemediği ve bu durumun genel performansı olumsuz etkilediğidir. Bu bağlamda, bu çalışmada da benzer bir zorlukla karşılaşılmış ve özellikle 'teardrop' gibi nadir saldırı türlerinin doğru bir şekilde sınıflandırılması için özel stratejiler geliştirilmiştir.Ayrıca, literatürdeki anormallik tespiti çalışmalarından biri, nadir olayları tanımlamak üzere özel algoritmaların kullanılmasını vurgulamıştır. Bu noktada, bu çalışmada da veri madenciliği teknikleri arasında Random Forest'in nadir olayları tespit etme konusundaki etkinliği üzerinde durulmuş ve algoritmanın bu bağlamdaki başarısı öne çıkarılmıştır.