KDD Cup , Bilgisayar Bilimi ve Veri Madenciliği toplulukları arasında popüler olan bir yarışma serisidir ve genellikle ağ güvenliği ve siber güvenlik alanlarındaki zorlukları içerir.

1. **Kaynak:**
    - Veri seti, 1999 KDD Cup yarışması için düzenlenen yarışmanın bir parçası olarak oluşturulmuştur.
    - UCI Machine Learning Repository üzerinden erişilebilir: KDD Cup 1999 Data
2. **Amaç:**
    - Veri setinin amacı, ağ trafiğini temsil eden bir dizi bağlantıyı içermek ve bu bağlantıların normal veya anormal olup olmadığını sınıflandırmak için makine öğrenimi modelleri oluşturmak için kullanılmaktadır.
    - Veri seti, siber güvenlikteki ağ güvenliği sorunlarına odaklanan bir zorluk olan ağ saldırılarını tespit etme amacını taşımaktadır.
3. **Siber Güvenliğe Olan İlgisi:**
    - Veri seti, siber güvenlik alanındaki temel bir sorunu ele almaktadır: ağ trafiğindeki anormal durumları tespit etme.
    - Bu tür veri setleri, siber güvenlik uzmanlarının ve araştırmacılarının ağ saldırılarına karşı savunma geliştirmeleri ve güvenlik önlemleri alabilmeleri için önemlidir.
    - Model oluşturma ve eğitme süreçleri, saldırı tespiti ve ağ güvenliği konularında siber güvenlik uygulamalarına yönelik çeşitli tekniklerin geliştirilmesine katkıda bulunabilir.

Model eğitmeden önce bir veri ön işlemi yapıldı.Siber güvenlik veri setleri üzerinde yapılan etkili veri ön işleme, modelin daha güvenilir, hızlı ve dengeli bir şekilde öğrenmesine olanak tanır. Bu adımlar, siber güvenlik tehditlerini daha etkili bir şekilde tespit etmek ve analiz etmek için kullanılan modellerin güvenilirliğini artırabilir. Veri ön işleme adımlarının doğru bir şekilde uygulanması, siber güvenlik uzmanlarının daha etkili güvenlik önlemleri almasına da yardımcı olabilir.Aşağıda  veri ön işleminde bazı krtik noktalar belirtilmiştir.

**Veri Ön İşlemenin Kritik Noktaları:**

**1-)Aykırı Değerlerin İdentifikasyonu ve İşlenmesi:**

Siber güvenlik veri setlerindeki aykırı değerler, saldırıları veya anormal durumları temsil edebilir. Bu değerlerin doğru bir şekilde tanımlanması ve işlenmesi, modelin güvenilir sonuçlar üretmesini sağlar.

**2-)Eksik Verilerin Yönetimi:**

Eksik veriler, modelin yanıltıcı sonuçlar üretmesine neden olabilir. Bu nedenle, eksik verilerin etkili bir şekilde doldurulması veya çıkarılması önemlidir.

**3-)Kategorik Değişkenlerin İşlenmesi:**

Kategorik değişkenlerin sayısallaştırılması veya kategorize edilmesi gerekebilir. Bu, modelin bu değişkenleri anlayabilmesi için önemlidir.

**4-)Normalizasyon ve Standartizasyon:**

Sayısal özelliklerin farklı ölçeklerde olması, modelin performansını olumsuz etkileyebilir. Bu nedenle, normalizasyon veya standartizasyon adımları önemlidir.

**5-)Anlamsız ve Gereksiz Özelliklerin Çıkarılması:**

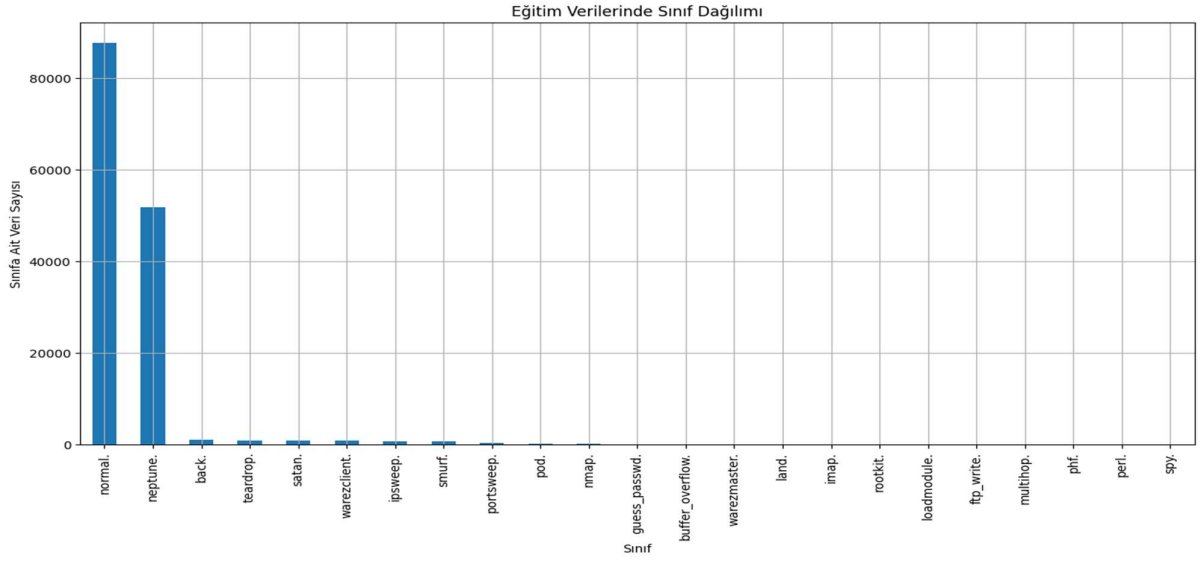
Model için gereksiz özelliklerin çıkarılması, modelin daha hızlı eğitilmesine ve daha iyi genelleme yapmasına olanak tanır.

Veri ön işleme kapsamında daha anlaşılır olması açısından ,ilk önce veri setine ilgili sütun isimleri verildi. Daha sonra  veri temizleme kapsamında null değer kontrolü yapılarak, veri setinde herhangi bir eksik veri olup olmadığı kontrol edildi.Duplicate değerler, veri setinden çıkarılarak veri setinin temizlenmesi sağlandı.NaN değer içeren sütunlar, veri setinden çıkarılarak temizlik sağlandı.

```
# Veri Temizleme
# NULL değerleri kontrol etme
print('Veri setindeki NULL değer sayısı', len(df[df.isnull().any(1)]))
print('=' * 40)
```

```
Veri setindeki NULL değer sayısı 0
========================================
<ipython-input-12-7b8883ef9b3f>:3: FutureWarning: In a future version of pandas all arguments of DataFrame.any and Series.any will be keyword-only.
  print('Veri setindeki NULL değer sayısı', len(df[df.isnull().any(1)]))
```

Daha sonra EDA işlemi ile birlikte veri setinin sınıf dağılımı görselleştirildi. Bu, veri setindeki sınıfların (etiketlerin) sayısal dağılımını gösterir.



Daha sonra model için gerekli olmayan bazı özellikler (protocol_type, service, flag, vb.) veri setinden çıkarılarak modelin karmaşıklığını azaltma amaçlandı.Sınıflandırma etiketleri, "normal" ve "abnormal" olarak iki kategoriye düzenlendi. Saldırı tespiti yapılacaksa "abnormal" durumları içerecek şekilde düzenleme yapıldı.

Son olarak Z-skor normalizasyonu kullanılarak sayısal özellikler standart bir normal dağılıma dönüştürüldü. Bu, modelin sayısal değerlerin farklı ölçeklerinden etkilenmesini engeller.

```
Normalleştirilmiş Özellikler:
   duration  src_bytes  dst_bytes  wrong_fragment    urgent      hot  \
0  -0.10785  -0.004293   0.042596       -0.084394 -0.004737 -0.07021
1  -0.10785  -0.004261  -0.039036       -0.084394 -0.004737 -0.07021
2  -0.10785  -0.004263  -0.025042       -0.084394 -0.004737 -0.07021
3  -0.10785  -0.004272  -0.025042       -0.084394 -0.004737 -0.07021
4  -0.10785  -0.004273  -0.013613       -0.084394 -0.004737 -0.07021

   num_failed_logins  num_compromised  root_shell  su_attempted  ...  \
0          -0.018022        -0.007905    -0.01944     -0.008613  ...
1          -0.018022        -0.007905    -0.01944     -0.008613  ...
2          -0.018022        -0.007905    -0.01944     -0.008613  ...
3          -0.018022        -0.007905    -0.01944     -0.008613  ...
4          -0.018022        -0.007905    -0.01944     -0.008613  ...

   dst_host_count  dst_host_srv_count  dst_host_same_srv_rate  \
0       -1.740383           -1.054224                0.979272
1       -1.639472           -0.967051                0.979272
2       -1.538562           -0.879878                0.979272
3       -1.437651           -0.792705                0.979272
4       -1.336741           -0.705532                0.979272

   dst_host_diff_srv_rate  dst_host_same_src_port_rate  \
0               -0.417555                     0.071230
1               -0.417555                    -0.177606
2               -0.417555                    -0.260552
3               -0.417555                    -0.260552
4               -0.417555                    -0.302025

   dst_host_srv_diff_host_rate  dst_host_serror_rate  \
0                     -0.31531             -0.644428
1                     -0.31531             -0.644428
2                     -0.31531             -0.644428
3                     -0.31531             -0.644428
4                     -0.31531             -0.644428
```