ML Bölümü Bitirme Ödevi

Hazırlayan: Satı Doğukan Esen,

Ödev Dosyası: <https://www.kaggle.com/dogukanesen/aiseclab-ml-odev>

Bu rapor Elif Parlak’ın verdiği ML eğitimi bitirme ödevinin sunumunu içermektedir.

# Veri Seti

Bu ödev için çalışmamız istenilen veri seti 2 sütundan oluşmaktadır. Bu veri setinde ilk sütunda(name) urller bulunmaktadır. Diğer sütununda(type) ise bu urllerin hangi kötü amaçla kullanıldığı bulunmaktadır. Bizden istenen bu veri setini kullanarak kötü amaçlı urlleri tespit edebilen bir makine öğrenmesi modeli geliştirmemizdir.

# EDA

Veri setinde 651191 adet veri vardır. Url sütununda 651191 adet benzersiz veri varken type sütununda 4 adet benzersiz veri vardır. Veri setinde null veri yoktur. Type sütunundaki 4 tür benign, defacement, malware, phishing’ dir. Benign’ den 428103 tane, defacement’ ten 96457 tane, malware’ dan 32520 tane ve phishing’ den 94111 tane veri bulunmaktadır. Verileri işleyebilmek için label encoding ile verilerimizi numeric hale getirilmiştir. Ardından korelasyon matrisi çizmekteyiz. Bu matriste veriler arasında düşük yüzde olduğunu görmekteyiz. Daha sonra çizdiğimiz jointplot’ ta verilerin dağılımını görmekteyiz.

# Makine Öğrenmesi

Url sütunu type sütununu etkilediği için bağımsız değişken olarak type sütununu seçmekteyiz. Daha sonra veriler eğitim ve test verisi olarak ayrılır. Burada eğitim verileri, verilerin %70i iken test verileri, verilerin %30’udur. Oluşturacağım modelin doğruluk oranını minumum %80 olarak düşündüm.

## Lojistik Regresyon

IBM’ in lojistik regresyon yazısından(<https://www.ibm.com/docs/tr/spss-statistics/saas?topic=regression-logistic> ) elde ettiğim bilgilerle modelimi bu yöntemle oluşturmaya karar verdim. Modeli eğitip test ettikten sonra modelin doğruluk skoru 0.658959448806806 di ve bu benim istediğim değerden daha düşük olduğu için bu model başarızdı.

## Karar Ağacı

IBM’ in Karar ağacı yazısından(<https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes>. ) elde ettiğim bilgilere göre modelim için bu yöntemin daha iyi olduğuna karar verdim. Modeli eğitip test ettikten sonra model doğruluk skoru 0.9177305254967802 oldu. Bu doğruluk oranı benim için yeterliydi ve model başarılıydı.

## Random Forest

Bu yöntemi daha yüksek bir doğruluk oranı için deneme kararı aldım. Modeli eğitip test ettiğimde doğruluk skoru 0.9178073076096193 çıktı. Yani karar ağacından çok az daha yüksek bir sonuç oldu. Bu modelde başarılıydı.

## KMeans

Kmeans denememin sebebi modelin denetimsiz öğrenmede nasıl çalıştığını görmekti. Kmeans modeli eğittim fakat silhouette\_score hesaplayan kodum takıldı o yüzden başarı durumunu öğrenemedim. Ayrıca modeli standarilize ederekte denedim.

# Sonuç

İlgili veri setinde ön işleme yaptıktan sonra random forest ve karar ağacı yöntemlerinin başarılı olduğu sonucuna vardım.