Gabriel Queiroz Teles - 2311064 - Ciência da Computação
Relatório Machine Learning - Bank

Analisando rapidamente os dados:

- Percebi que era um Dataset pequeno de apenas 600 linhas
- Notei que muitas colunas eram do tipo ‘object’

Sabendo disso eu substitui os dados no Dataset de forma lógica para que as colunas que eram ‘object’ fossem do tipo ‘int64’

Após essa análise rápida, fiz uma análise mais profunda para entender como as variáveis do Dataset se comportavam. Fiz:

- Histograma (serviu para ver a proporção dos dados em cada coluna já que a maioria era binária)
- Boxplot (serviu para identificar se alguma variável possuía outliers
- Gráfico de Correlação (serviu para ver como a coluna ‘pep’ se correlaciona com as outras colunas)

Com isso eu percebi principalmente que as colunas de maneira geral não possuíam uma correlação boa diretamente com a coluna ‘pep’. Devido a este “problema”, eu resolvi não remover muitas colunas já que se eu removesse todas com correlação ruim, não sobraria nenhuma coluna no Dataset. Além disso, com várias variáveis os algoritmos poderiam identificar algo que os ajudaria a acertar juntando as colunas.

Tendo terminado a fase de análise, comecei a preparar o Dataset para o treinamento. Eu com isso eu:

- Apliquei o SMOTE (para criar dados sintéticos para equilibrar a distribuição de classes no conjunto de dados)
- Separei o Dataset em train e test
- Apliquei o StandardScaler (para normalizar os dados)

Além disso, criei funções para mostrar as métricas (acurácia, precisão, recall, f1-score, matriz de confusão e curva ROC)

Com tudo isso feito, comecei o treinamento utilizando os seguintes algoritmos:

Regressão Logística;
Árvores de Decisão;
Random Forest;
ExtraTrees;
k-NN;
SVM(SVC);
BernoulliNB;

O que se saiu melhor entre esses foi o Random Forest. Por isso, eu o escolhi para tentar otimizar com os algoritmos BayesSearch, RandomizedSearch e GridSearch. Após a otimização percebeu-se que os resultados pioraram um pouquinho, mas ainda estão altos e que, portanto, o modelo serve para resolver o problema de classificação.