CS 4622 - Machine Learning
Lab 01 - Report

Name : Poorna S. Cooray
Index number : 190110V

Python notebook :
https://colab.research.google.com/drive/1sCkqhGGDP-2PUOQz2kK5cTljaNRKxG8Q?usp=sharing

Introduction

This lab exercise focuses on refining feature engineering and selection techniques, including feature scoring and PCA. The training dataset contains 28,520 rows, 256 features, and 4 target labels. Notably, label 2 has missing values, and label 4's distribution is uneven. According to the observations, label 2 was identified as a prediction problem and others as classification problems. Mitigating these challenges, Xgboost addresses label 2, while RandomForestClassifier suits the other labels.

Method

Data preprocessing:

During the data preprocessing phase, all instances of missing values were eliminated by dropping them, and the features underwent rescaling using scikit-learn's StandardScaler().

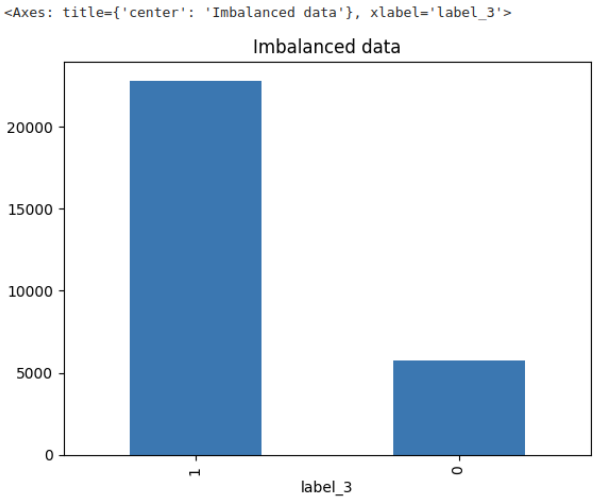
Feature Engineering:

For all the labels, PCA and feature elimination using feature importance values were used. The following table shows the PCA variance value and feature importance value threshold used in every label along with the final number of features and the final accuracy.

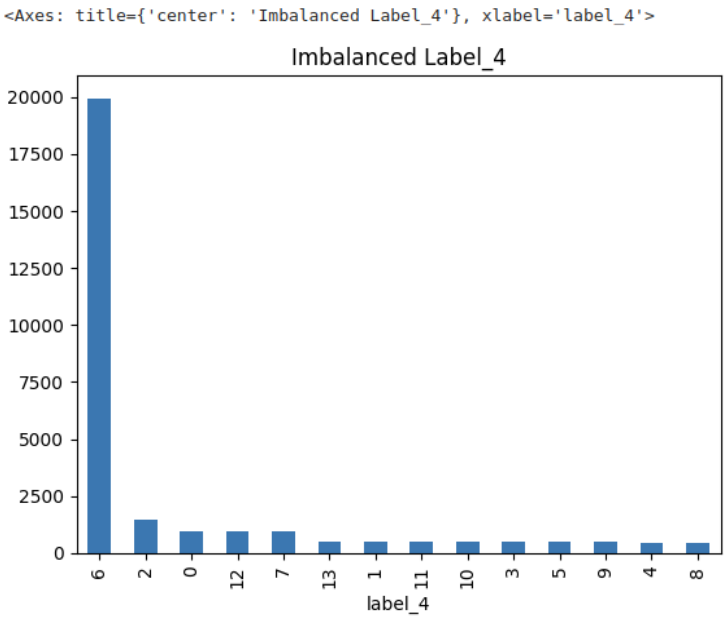
Feature name	PCA- variance value	Feature importance threshold	Final number of features	Final accuracy
label_1	0.97	0.009	21	0.97
label_2	0.8	-	31	3.62 (MSE)
label_3	0.98	0.008	21	1.0

label_4	0.97	0.015	21	1.0

Furthermore, label 3 and label 4 were imbalanced. Hence resampled using the imbalanced-learn package.



Label 3



Label 4