DATA MINING PROJECT

CANBERK ARICI -171044062

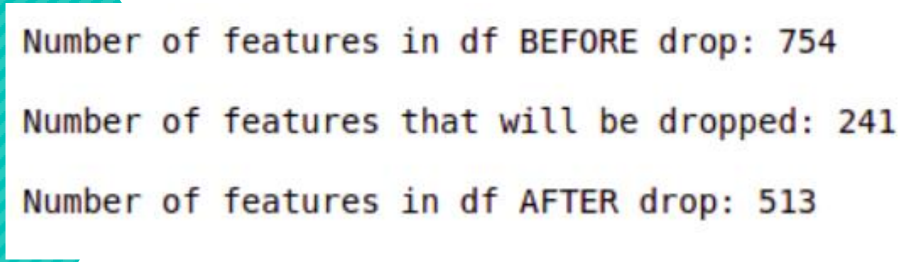
PREPROCESSING

Data Cleaning

Dropping "id" column

Feature Extraction

- 754 Features in Dataset
- Correlation checked between Features
- If higher than 0.95, drop them

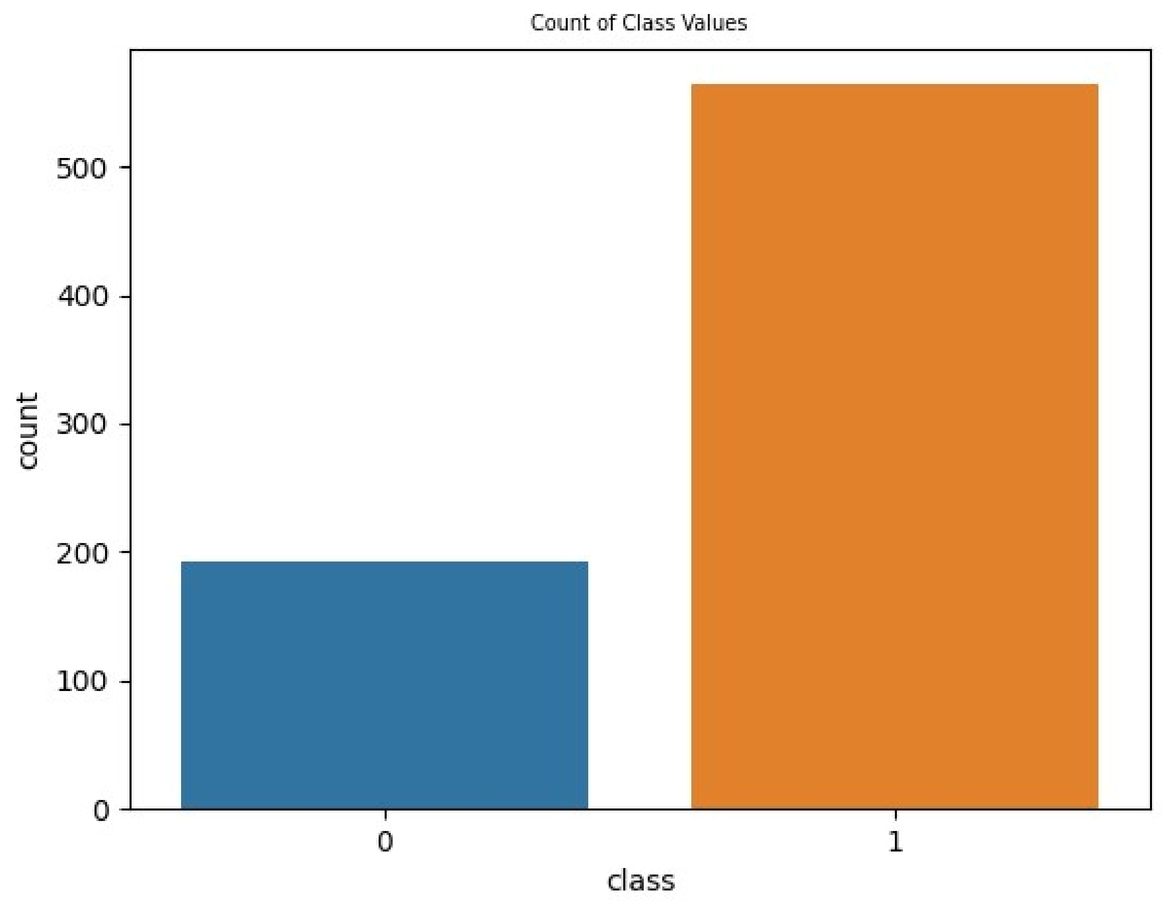


```
Number of features in df BEFORE drop: 754

Number of features that will be dropped: 241

Number of features in df AFTER drop: 513
```

Features of Dataset

- There is no categorical value in dataset therefore no need to turn them into numerical values.
- There is no NaN value.

Balance of Target Feature



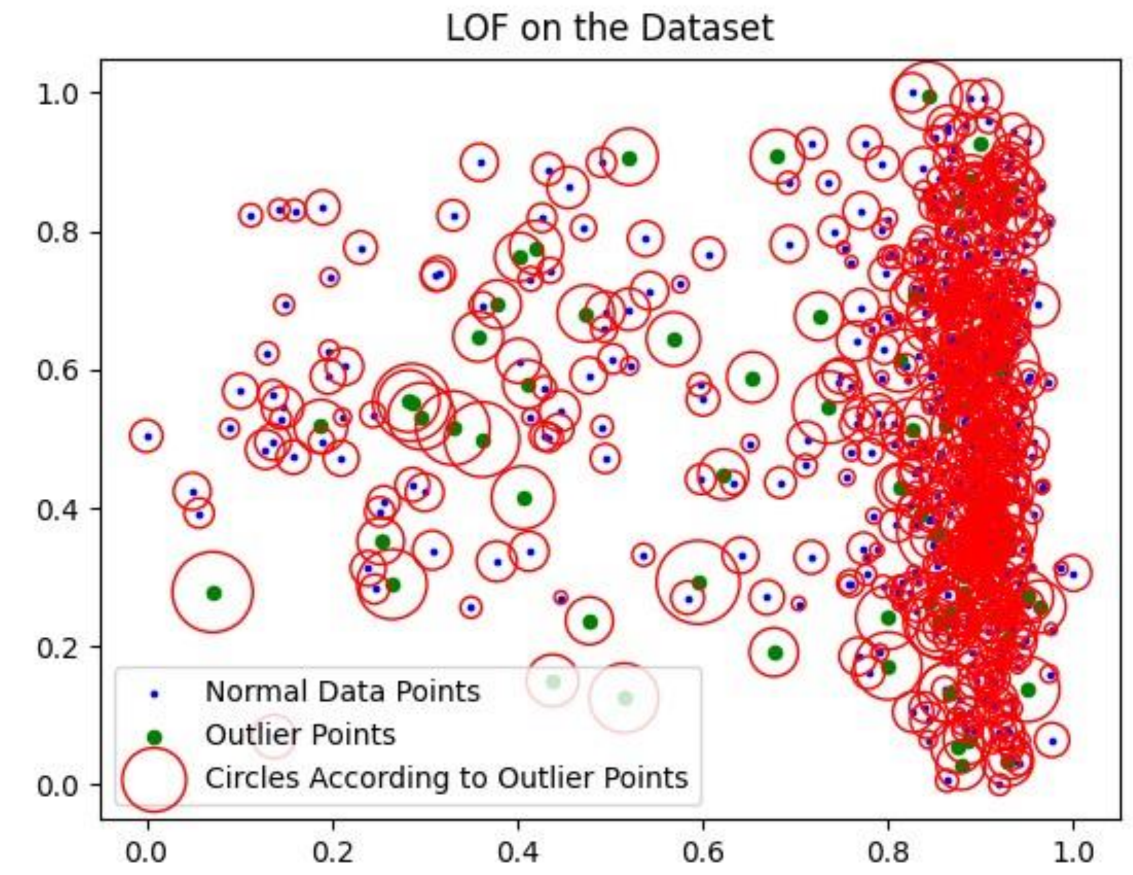
Normalization

There are values such as 0.000000135 and 4451980.807 in the dataset

Their affect on classification result will be very different

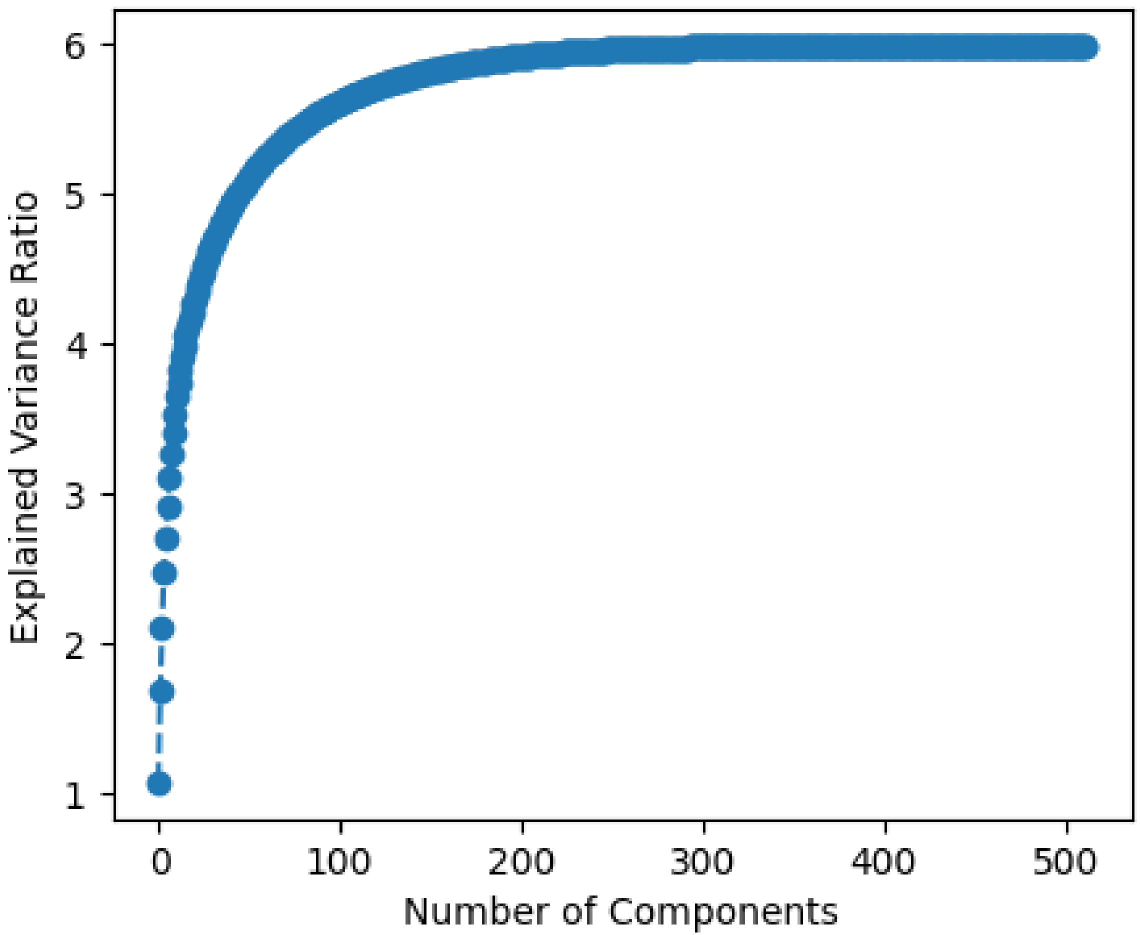
I used MinMaxScaler() for normalization in order to fit values between 0 and 1

Outlier Detection with Local Outlier Factor (LOF)



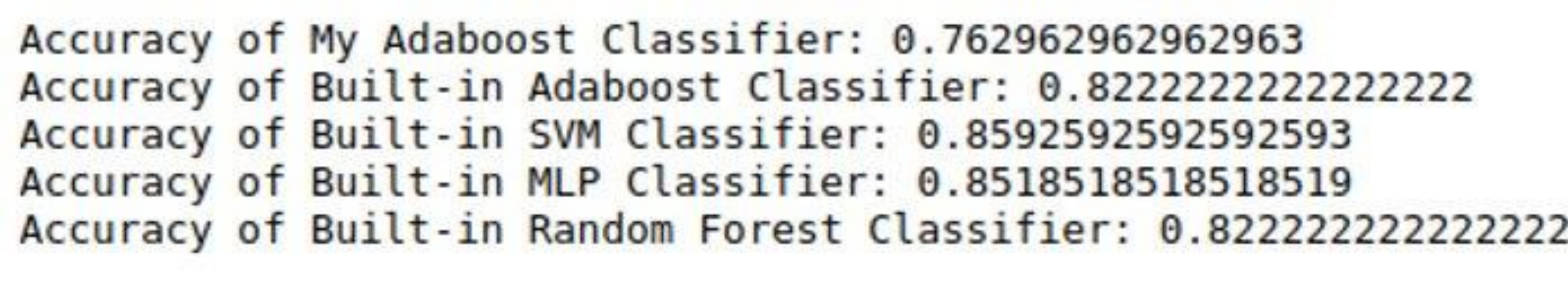
- **I selected threshold as 1.25 because it was very different from other data points' values**

PCA



- We will need roughly 60 components to keep 90% of the information.

RESULTS



```
Accuracy of My Adaboost Classifier: 0.762962962962963
Accuracy of Built-in Adaboost Classifier: 0.8222222222222222
Accuracy of Built-in SVM Classifier: 0.859259259259593
Accuracy of Built-in MLP Classifier: 0.851851851851519
Accuracy of Built-in Random Forest Classifier: 0.822222222222222
```