**CS 634 DATA MINING FINAL PROJECT**

**NAME:- SAILESH ALLA**

**UCID –** [**SA276@NJIT.EDU**](mailto:SA276@NJIT.EDU)

**Project :-**

This is a basic implementation in Python of the Supervised Data Mining (Classification) Algorithms i.e. Linear Logistic Regression, Decision Tree and Random Forest

**Requirements:**

**Software:** Python 3.7

Jupyter Notebook

**Libraries:** Pandas library is used to import the CSV file.

Itertools used to iterate over data structures that can be stepped over using a for-loop

NumPy used for working with arrays.

matplotlib.pyplot used to create figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc

seaborn is Python data visualization library based on matplotlib.

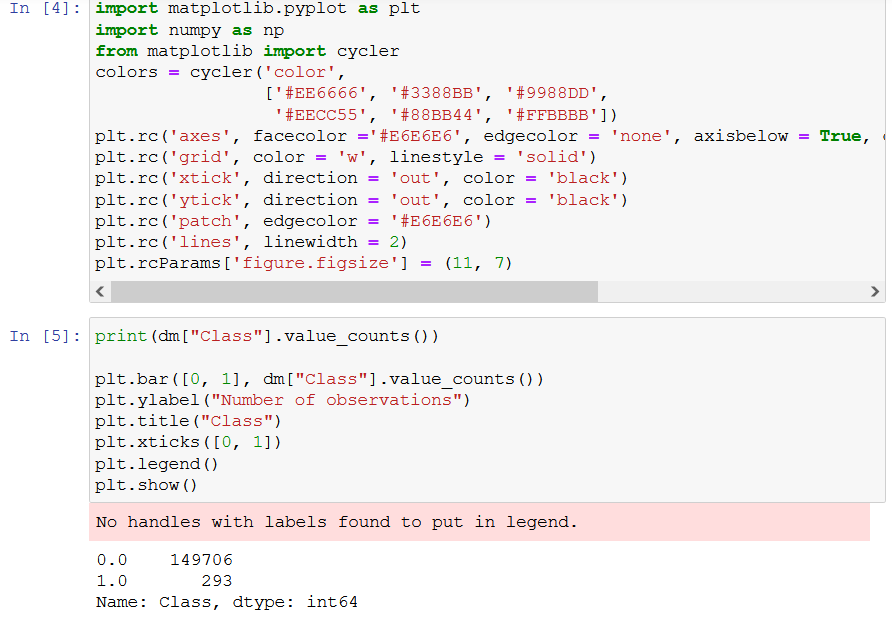
**Installation:**

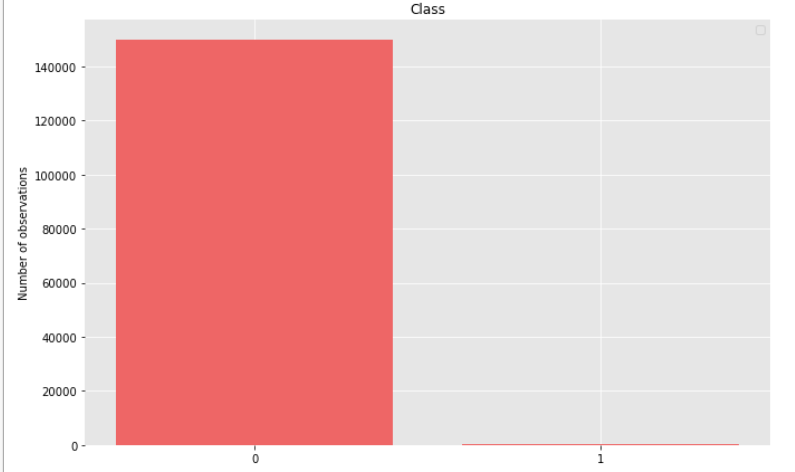
1. Install the necessary packages numpy, pandas, csv, matplotlib.pyplot, seaborn.
2. Once the packages are installed start with the imports.
3. Once the libraries are imported, import and explore the dataset

Graphical user interface, text, application, email

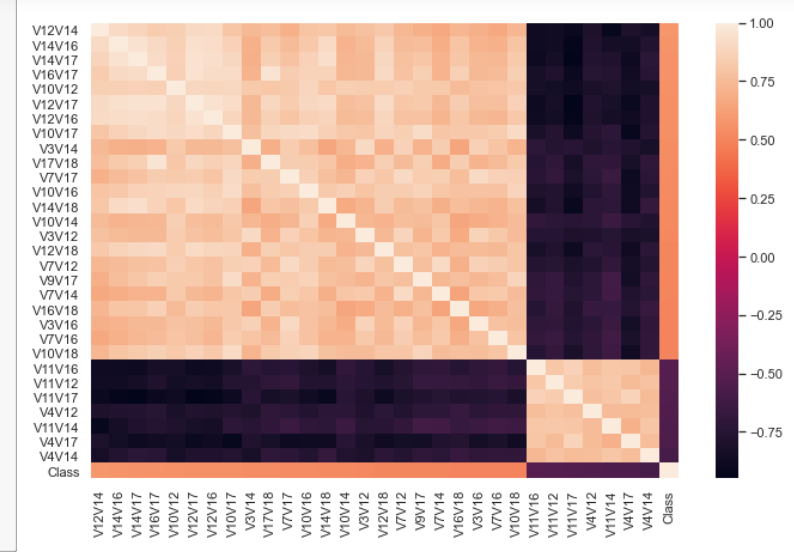
Description automatically generated

1. Generating histogram

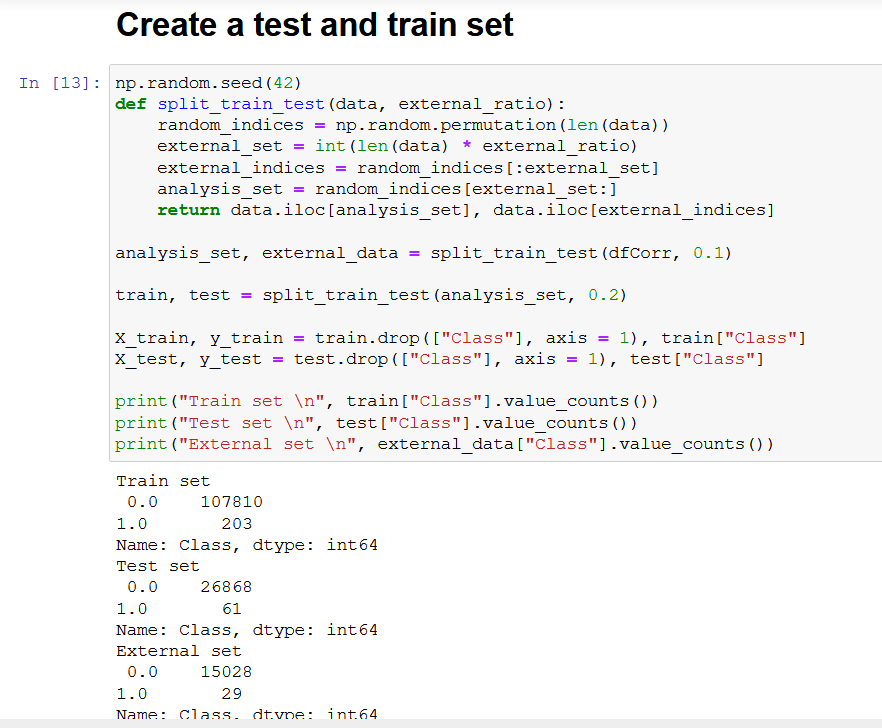




1. Function to find a strong correlation

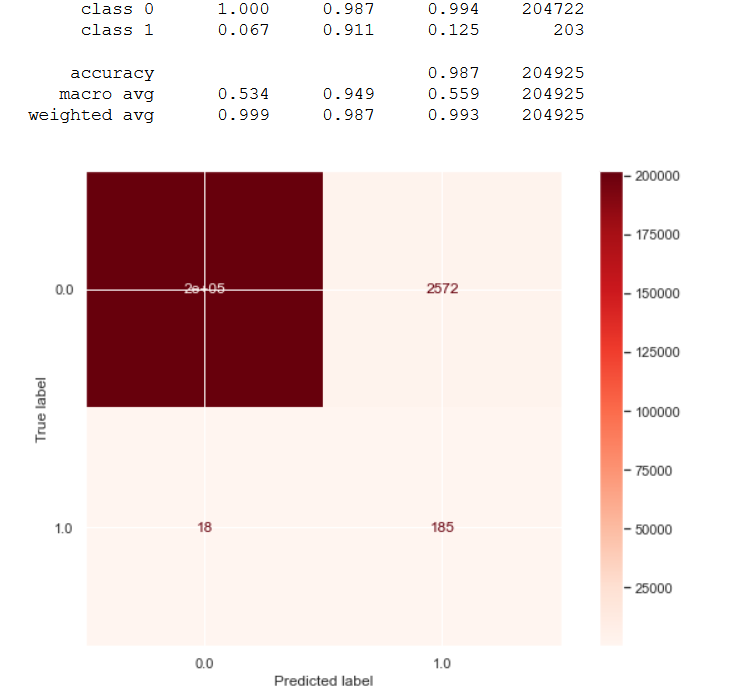


1. Started prepping the data creating train and test set for the analysis

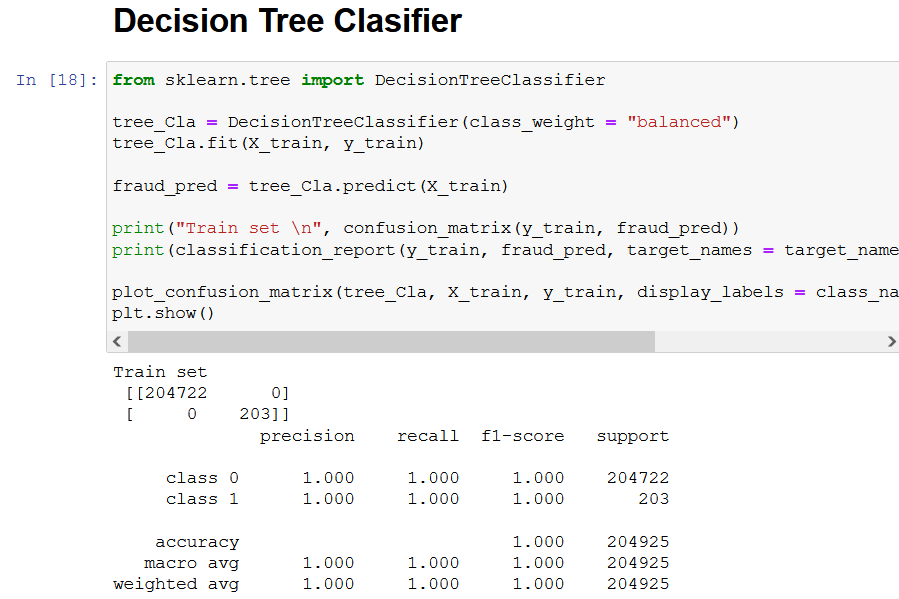


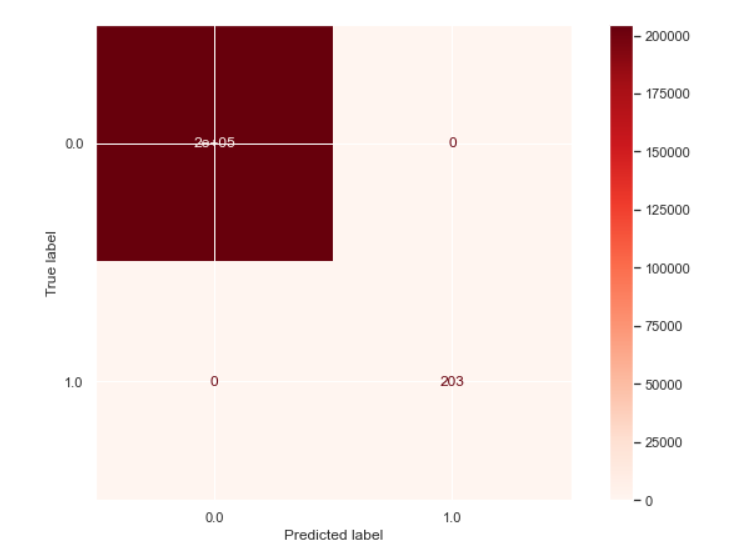
1. Logistic Regression



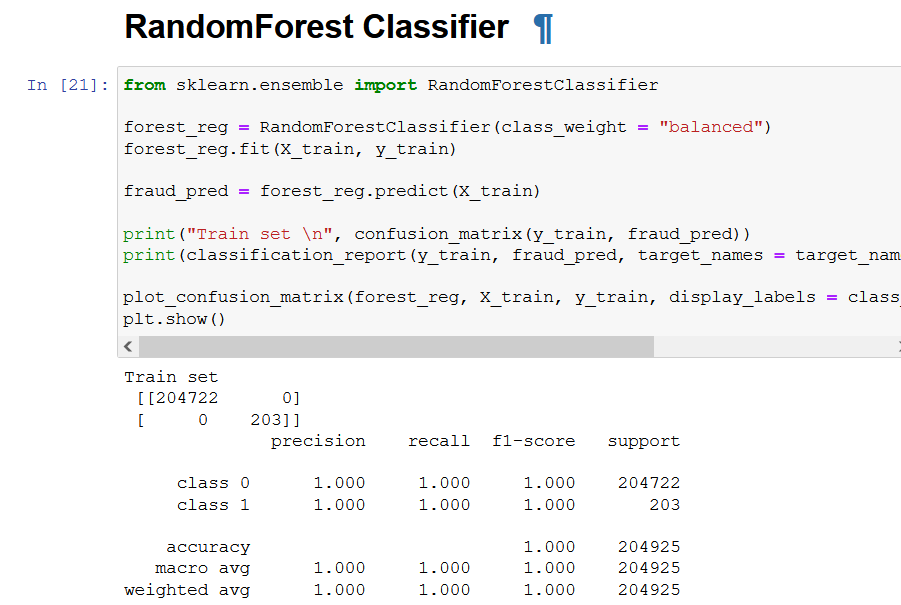


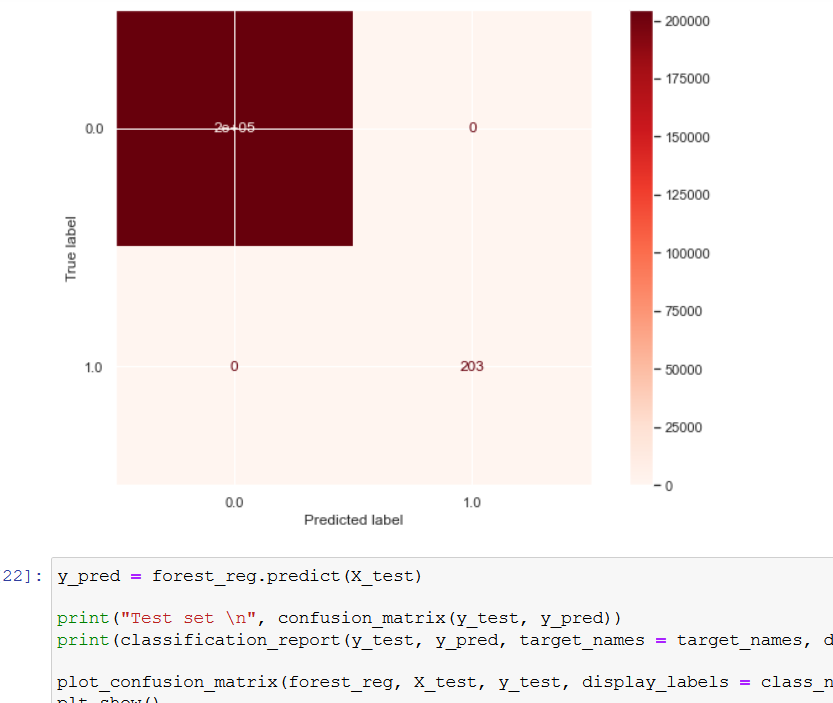
1. Decision Tree Classifier Model





1. RandomForest Classifier Model





**Conclusion:**

* In this notebook a comparison of classification algorithms (Linear Logistic Regression, Decision Tree and Random Forest) was presented.
* A function was created to carry out binary combinations (multiplication of attributes) by returning a dataframe with the combinations that meet the required value of correlation interest. In this case, those combinations that present a Pearson correlation greater than 0.5 were sealed.
* From the principal components, binary combinations were carried out to improve the linear correlation with respect to the Class attribute.
* Of the three classification models presented, the logistic regression model presents better results, with random forest model and decision tree model producing almost similar results.
* In the logistic regression, a recall for fraud detection was higher than 0.90 in the train set 0.89 with the test set and 0.95 recall with the outer set.