Credit Card Fraud Classification

Toshani Rungta   
Computer Science Department  
PES UniversityBangalore, India  
toshanirungta@gmail.com

Vismaya Ganesan Iyer  
Computer Science Department  
PES UniversityBangalore, India   
vismayagiyer@gmail.com

*Abstract* — With the surge in online shopping and services, payment systems have also moved online. The methods of online payment include credit card transactions, and with evolving times, people prefer e-transactions and plastic money over cash. This has led to a huge rise in credit card payments and consequentially led to a swell in the number of frauds happening in this sector. Credit card fraud generally happens when the card was stolen for any of the unauthorized purposes or even when the fraudster uses the credit card information for his use. Therefore, banks and financial institutions offer credit card fraud detection applications much value and demand [1]. This project aims to focus mainly on machine learning algorithms. We have used a dataset containing transactions made by credit cards in September 2013 by European cardholders. The algorithms used are SMOTE-ing and KNN. The results of the two algorithms are based on accuracy, precision, recall, and F1-score. Each parameter signifies the efficiency and usefulness of the model.

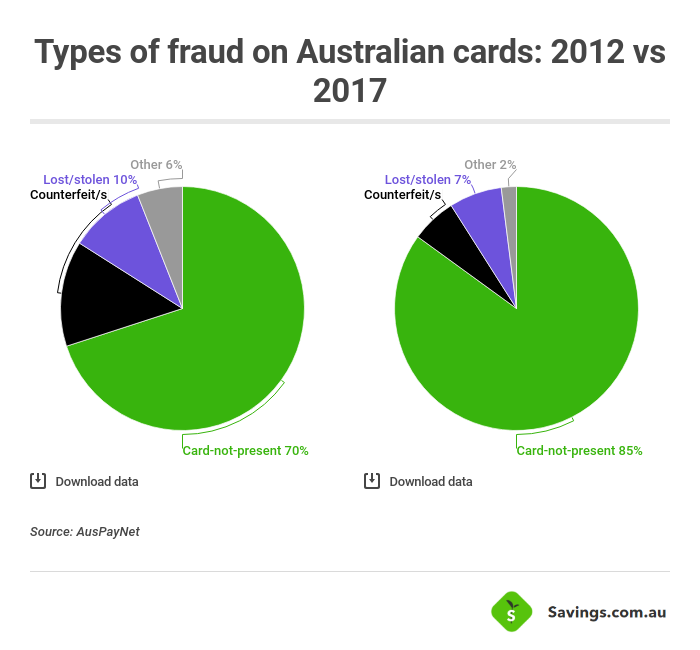


Figure 1: Credit Card Fraud Types

Keywords — KNN, SMOTE-ing, Machine Learning, Credit Card

# Introduction

The data set contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. [2]

# RELATED WORK

## Exploratory Data Analysis (EDA)

In statistics, exploratory data analysis is an approach to analysing data sets to summarize their main characteristics, often with visual methods. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.  
In our project we have used not just the correlation matrix but also the heat map to understand and interpret the data.  
  
Heat Map – A heat map is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

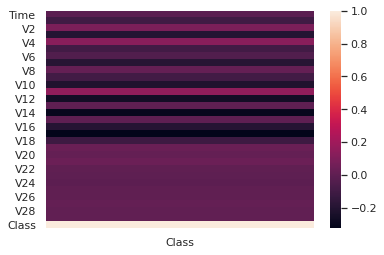


Figure 2: Heat Map for Chosen Data Set

Data Imbalance – While classifying credit card transactions as fraud and non-fraud, the biggest challenge faced is the severe imbalance in data as the fraud class in this dataset constitutes only 0.17% of all transactions.

This results in overfitting, with the formation of a model that is highly biased towards the majority class, leaving many frauds undetected.

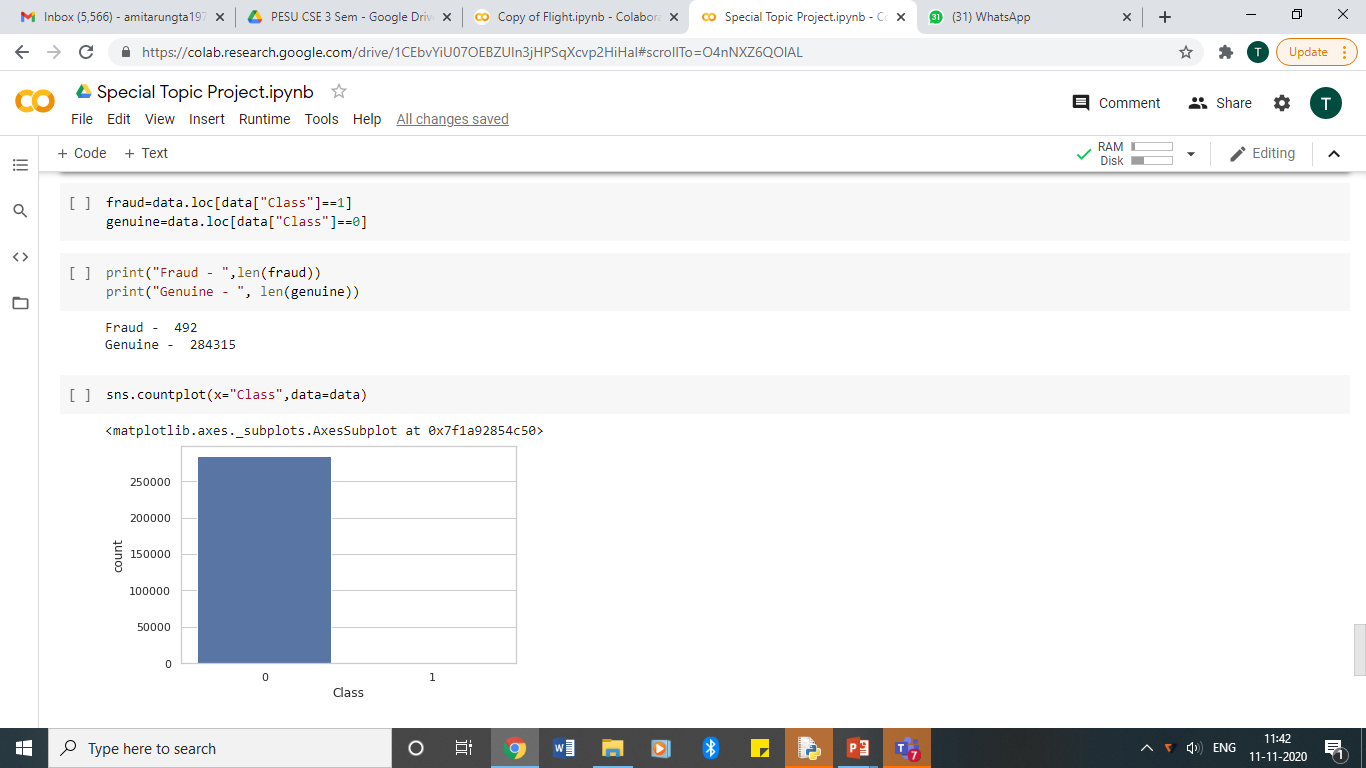


Figure 3: Data Imbalance in Chosen Data Set

## Balancing the data – To tackle the issue of overfitting that the imbalance causes, the data needs to first be balanced.So as to prevent data loss, oversampling is preferred to undersampling. Synthetic Minority Oversampling Technique (SMOTE) is a technique used for oversampling, SMOTE synthesizes new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.where synthetic samples of the minority class are created.

## This algorithm helps overcome the overfitting caused by random oversampling. For the implementation of SMOTE, the class SMOTE from the python package imblearn is used, after which there are 255882 records in each class.

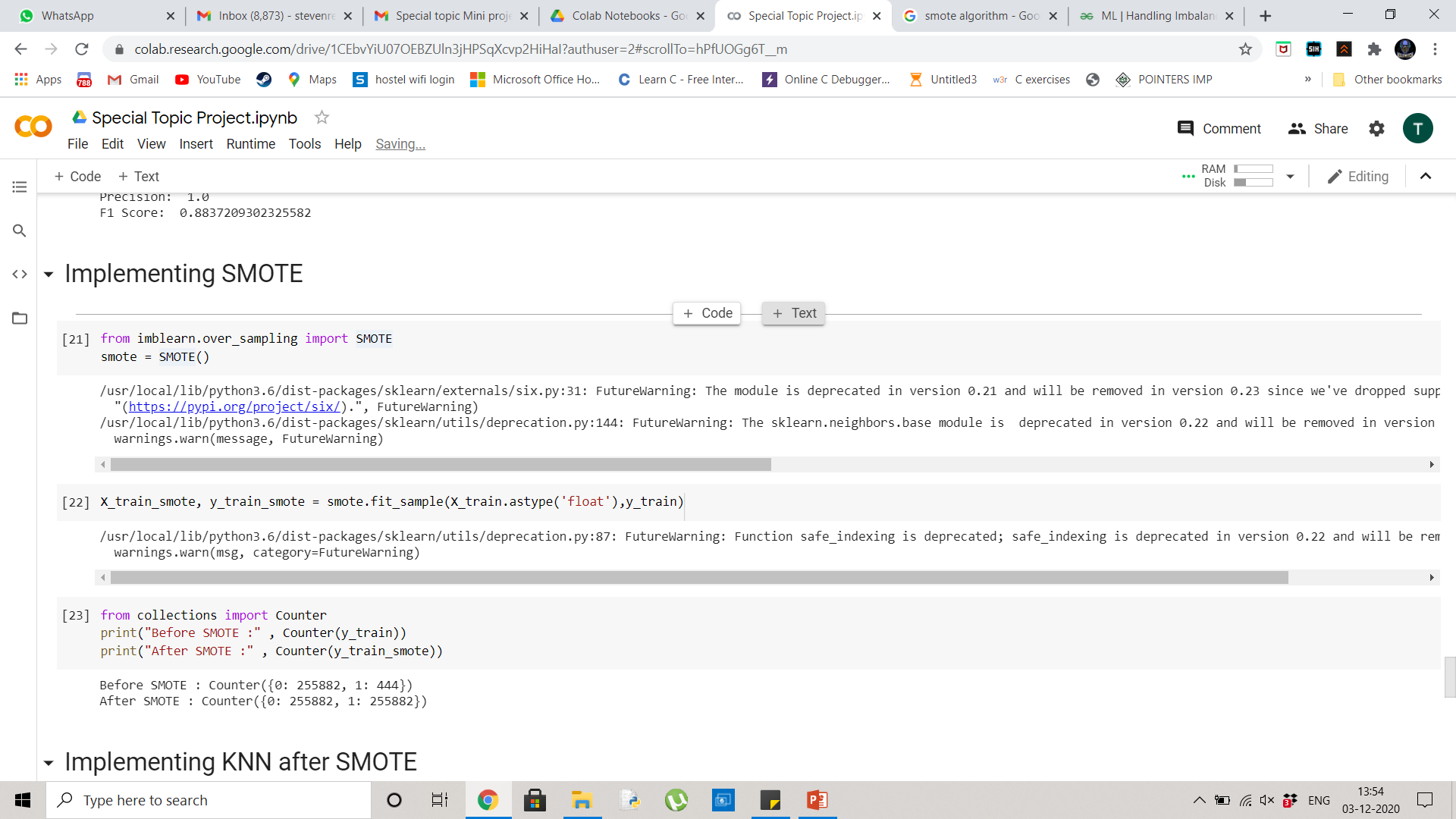


Figure 4: Data Distribution before & after SMOTE-ing

Classification – The technique used for classification is k-nearest neighbors (KNN), where the case is classified by a majority vote of its nearest neighbors. For the fitting of the model using KNN, we use KNeighborsClassifier from the python package sklearn. [3]

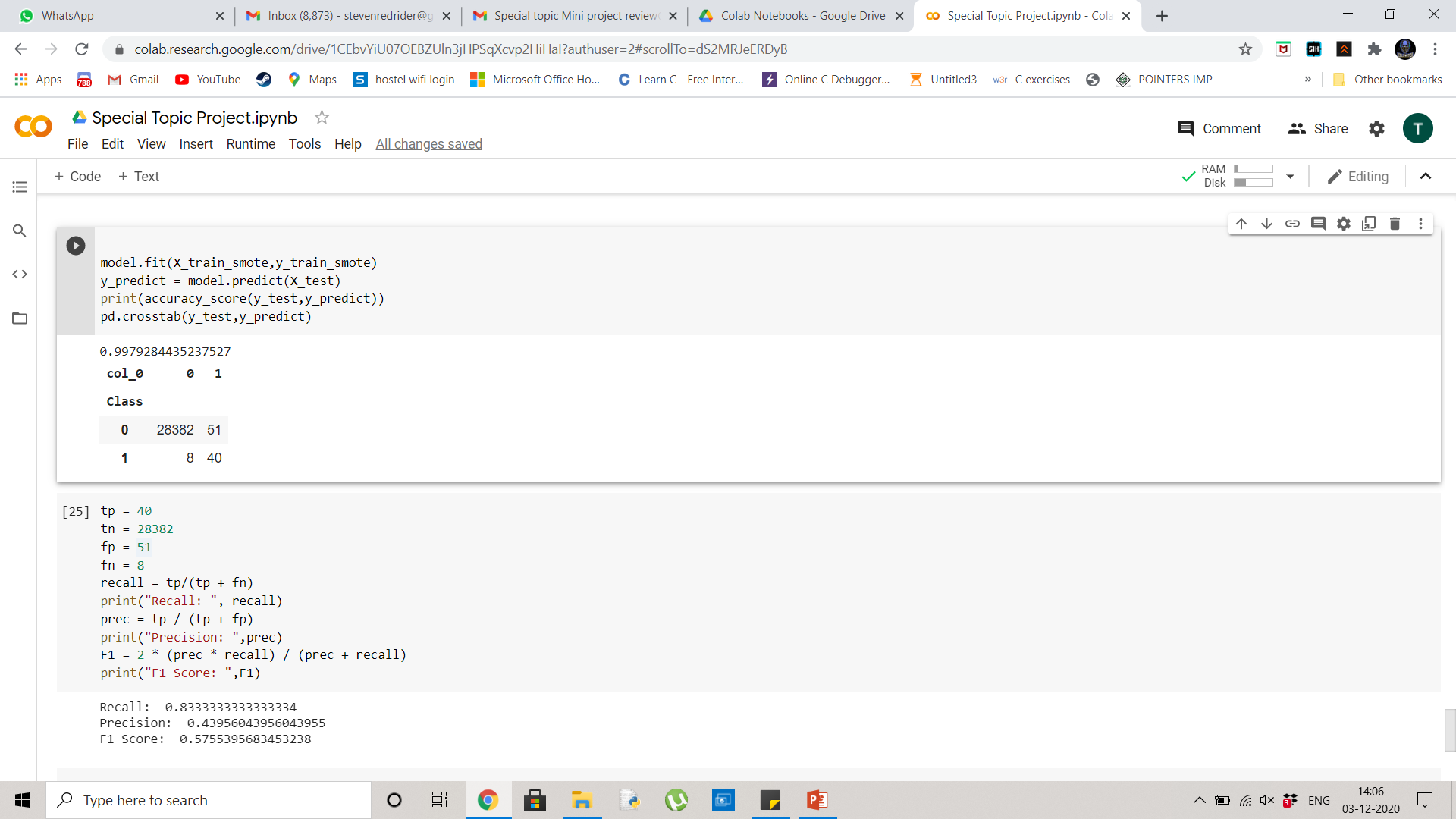


Figure 5: KNN Code block used in the project

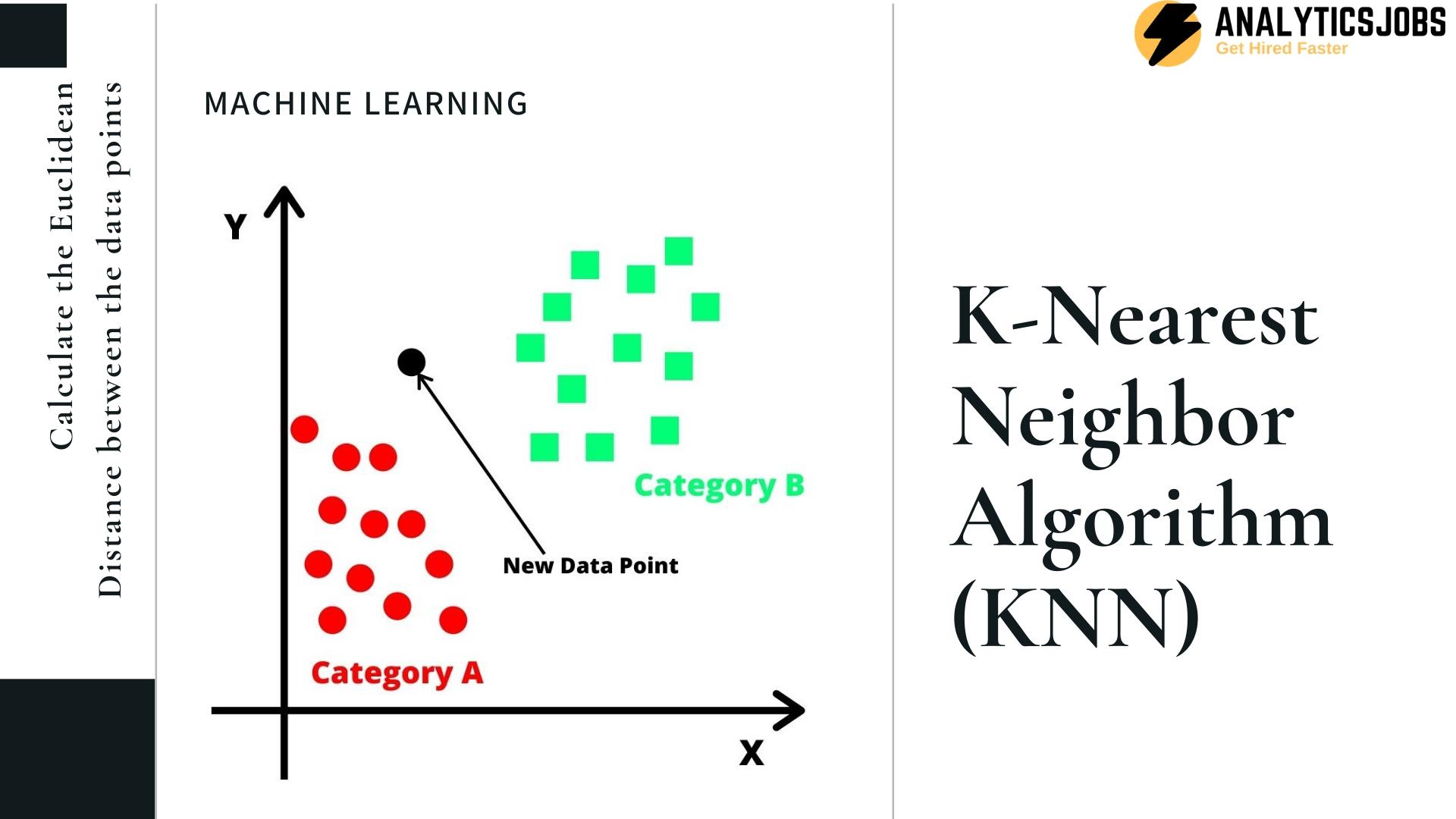


Figure 6: Diagrammatic Representation of KNN Algorithm

# EVALUATION AND RESULT ANALYSIS

The accuracy is found to be 0.9979284435237527, while the recall, precision and f1 score are 0.8333333333333334 0.43956043956043955 and 0.5755395683453238 respectively. Though we see a high accuracy, the other parameters offering more insight into the result show that there is a high number of non-fraudulent transactions being flagged as fraud, which results in the low precision and a considerable number of frauds left undetected, which explains the recall.

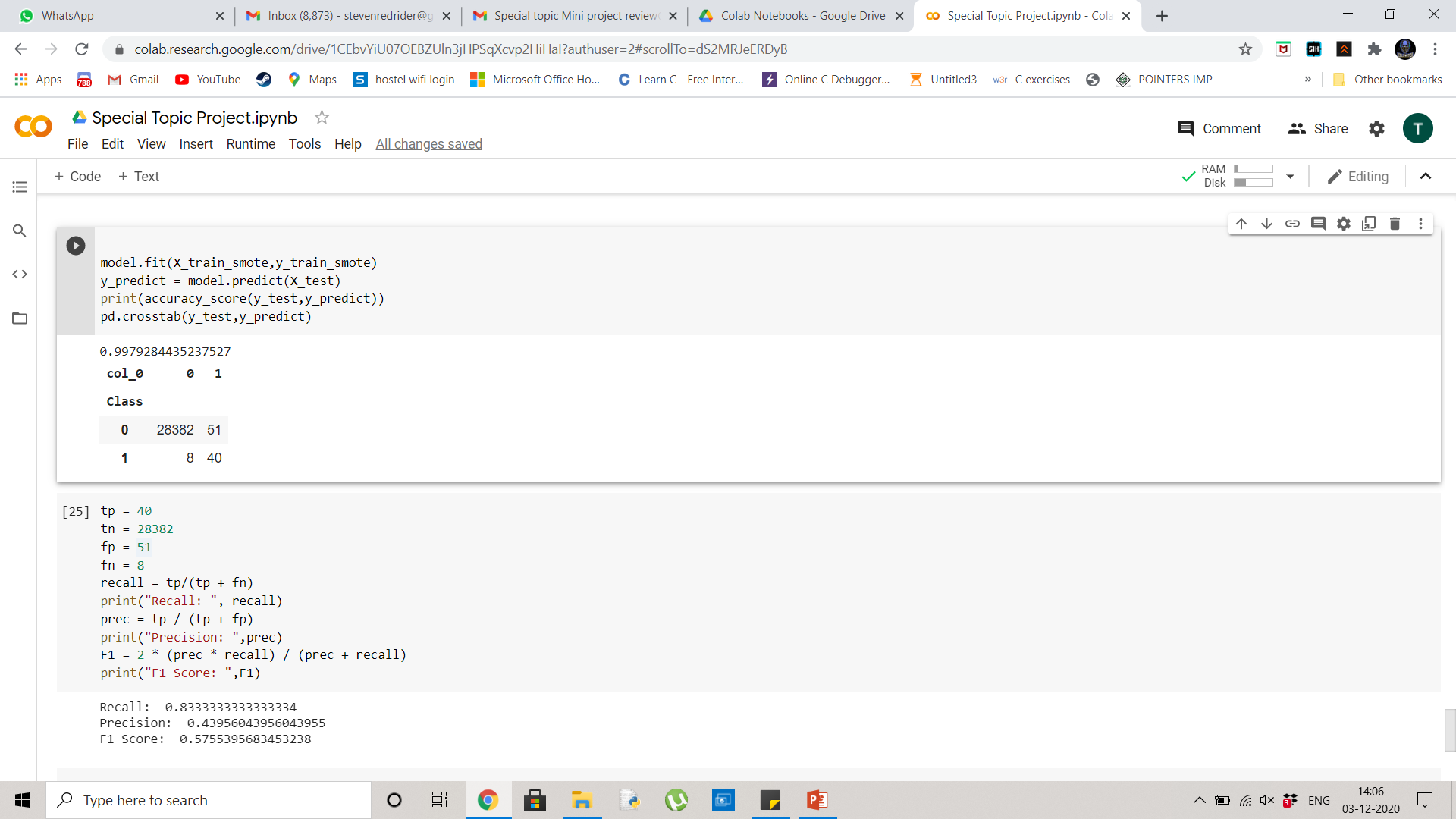


Figure 7: Results of the Model

# CONCLUSION

With this, we see that KNN model is not fit for classification in our data set even though it is a well proven algorithm. We are getting a very low precision and F1 score depicting that the training set fails to help classify the test set. As such, it would be better to try and experiment with models such as Random Forest and verify the same with k-fold cross validation.

# Acknowledgment *(Heading 5)*

We thank the CSE Department, PES University for giving us the wonderful opportunity to do a project on machine learning. We further thank and are grateful for the guidance and support of our mentor Prof. Kakoli Bora. Last but not the least, we thank the panel members who always gave us immaculate and important suggestions at each step. This has been an insightful educational experience and extend our heartfelt gratitude for the same.

# References

1. https://ieeexplore.ieee.org/document/8776942/references#references
2. https://www.kaggle.com/mlg-ulb/creditcardfraud?select=creditcard.csv
3. https://colab.research.google.com/drive/1\_Vsr4zxMETBhvoYs1s9ZebM52o6\_4Xpv?usp=sharin