CREDIT CARD FRAUD DETECTION

DAA PROJECT BY---

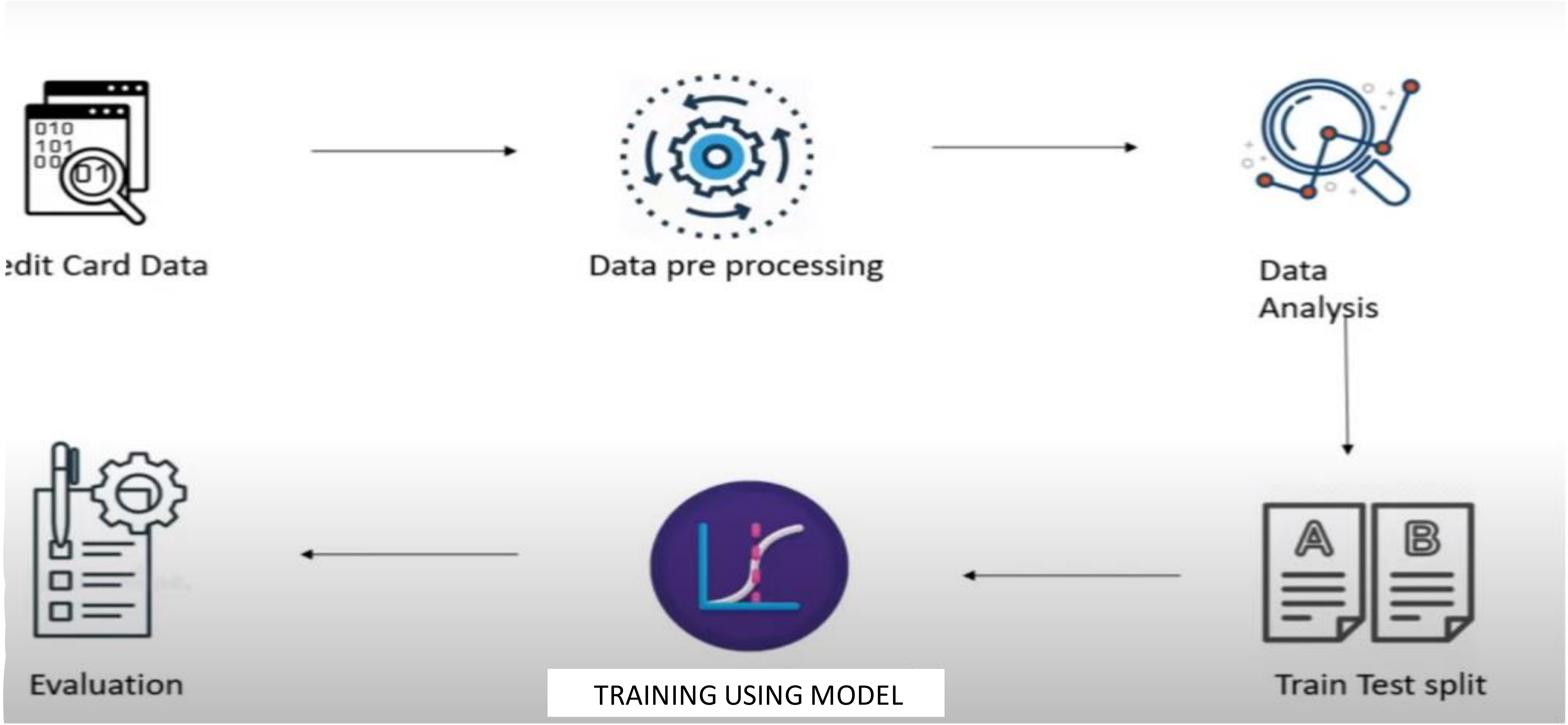
SAYAK HATUI SAPTASWA MISTRI SUBHAM PRADHAN

ARGHYAJYOTI MONDAL

INTRODUCTION AND IMPORTANCE TO DETECTING FRAUDULENT TRANSACTIONS

- Credit Card Fraud Detection is a critical aspect of financial security in the modern world. It involves identifying and preventing unauthorised or fraudulent transactions made using credit cards. With the increasing reliance on digital payments, the risk of fraudulent activities has also escalated. Detecting these activities in real-time is crucial to protect both financial institutions and customers from potential losses. This presentation will delve into various models and techniques used to effectively identify and prevent credit card fraud.

WORK FLOW DIAGRAM



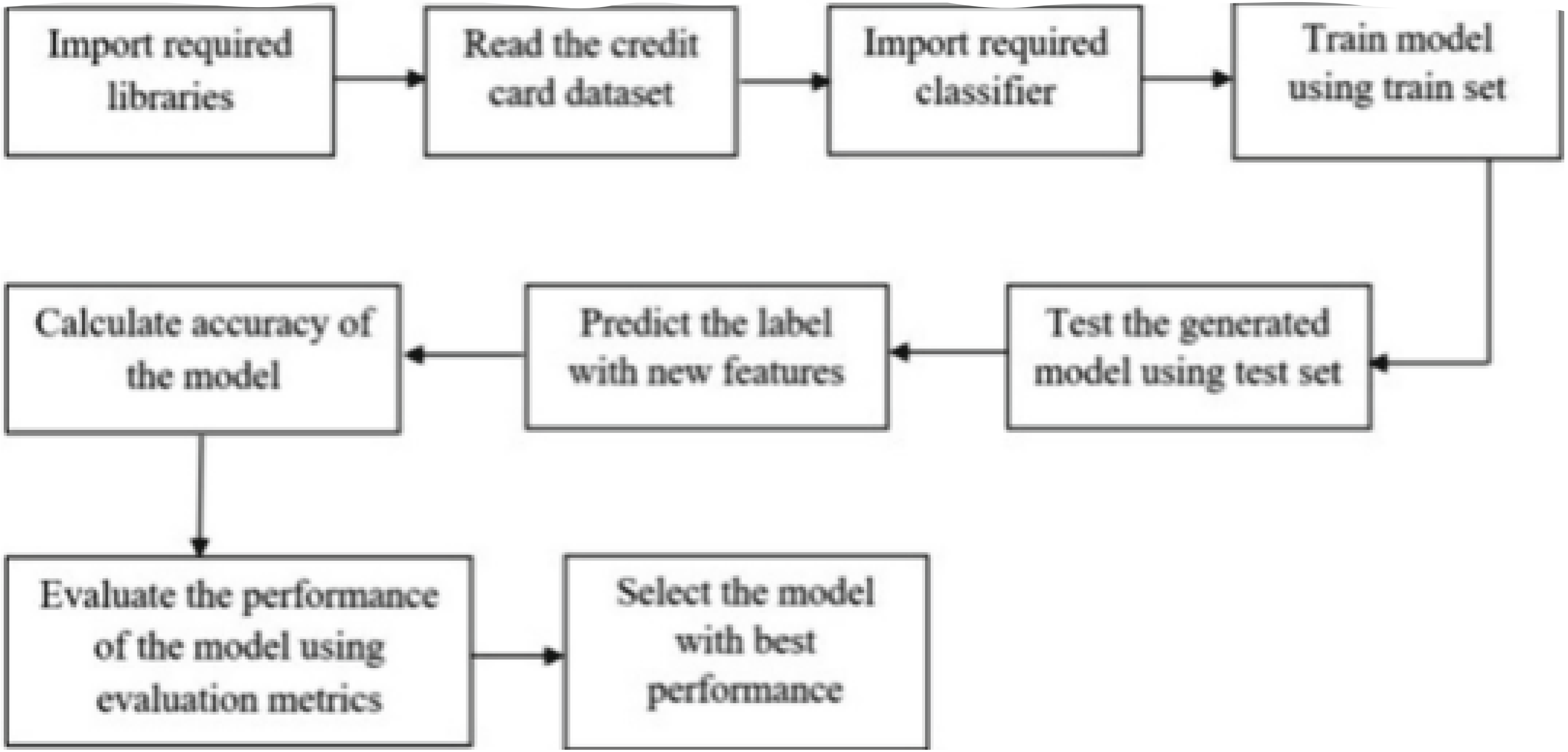
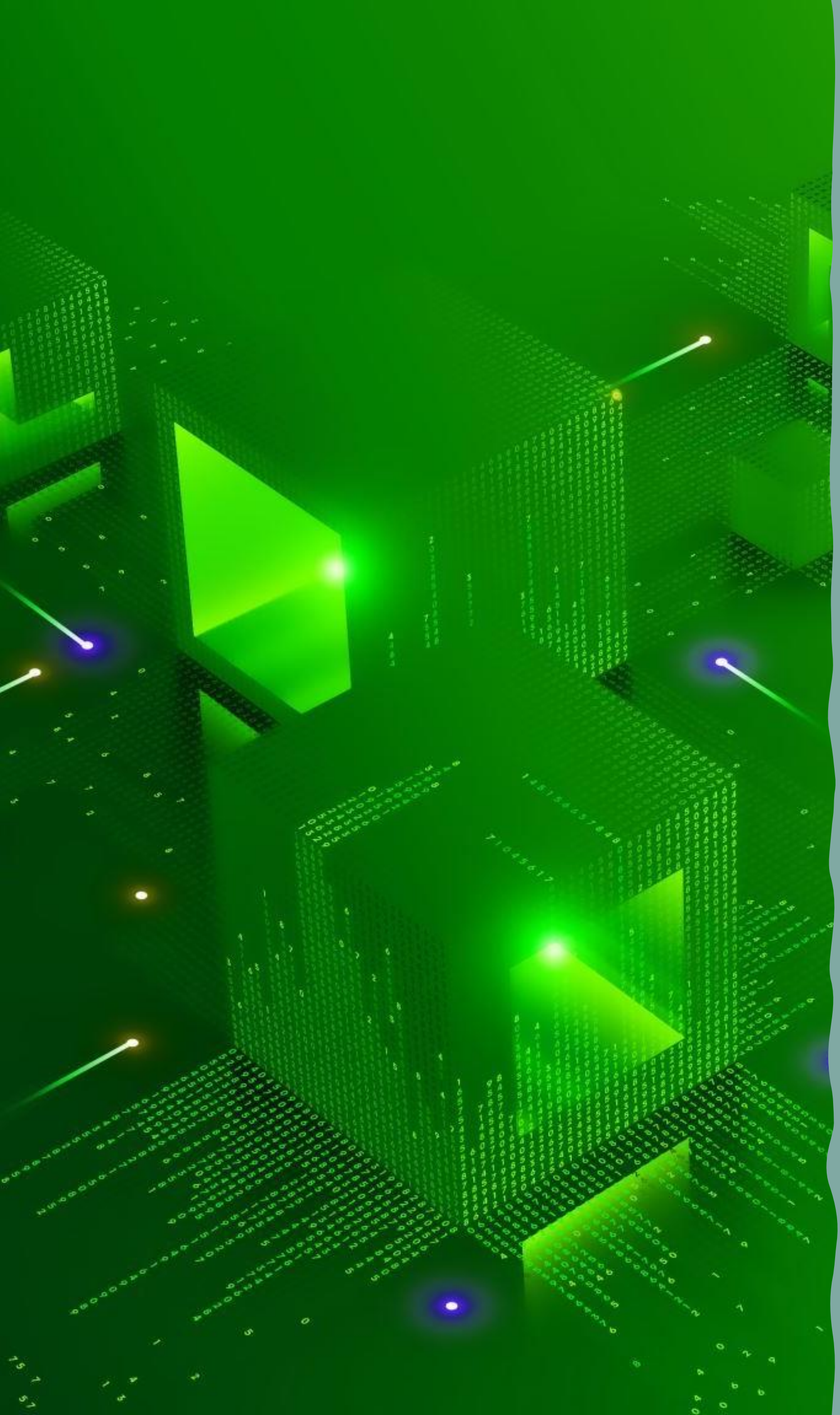
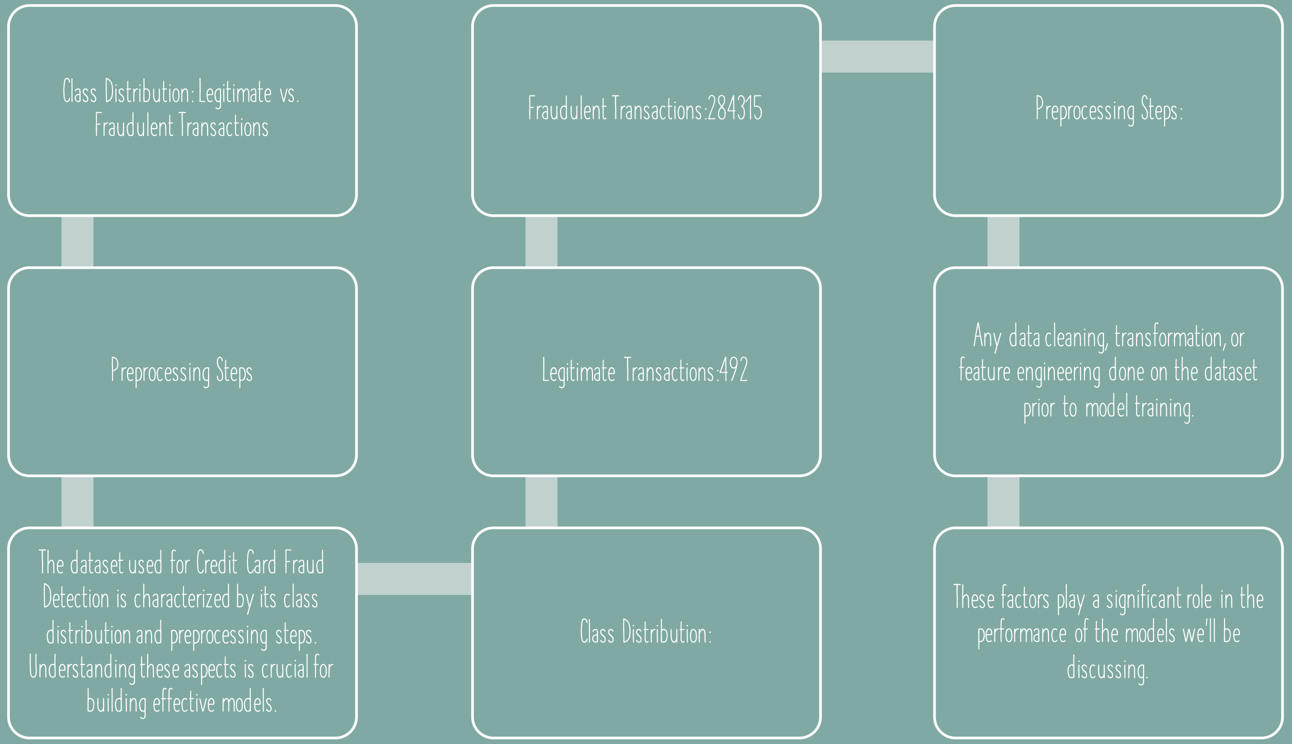

Import required libraries
Read the credit card dataset
Import required classifier
Train model using train set
Test the generated model using test set
Predict the label with new features
Calculate accuracy of the model
Evaluate the performance of the model using evaluation metrics
Select the model with best performance

Fig. 1. Block diagram to build the model



DATASET CHARACTERISTICS



Class Distribution: Legitimate vs. Fraudulent Transactions

Preprocessing Steps

The dataset used for Credit Card Fraud Detection is characterized by its class distribution and preprocessing steps. Understanding these aspects is crucial for building effective models.

Class Distribution:

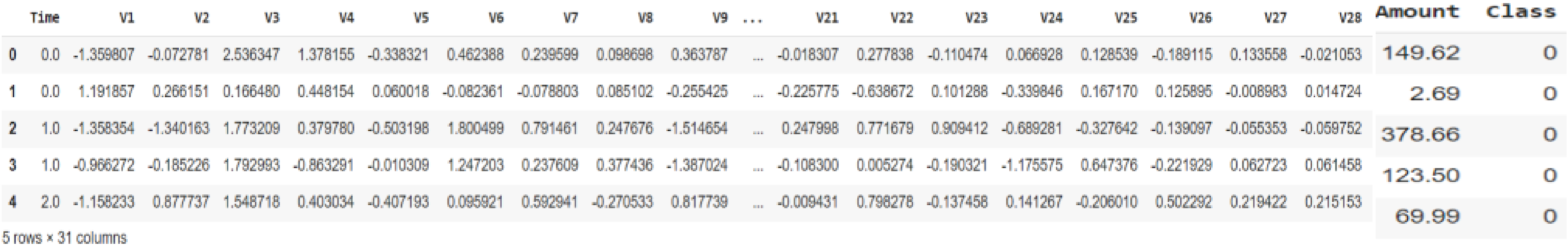
Legitimate Transactions:492

Fraudulent Transactions:284315

Preprocessing Steps:

Any data cleaning, transformation, or feature engineering done on the dataset prior to model training.

These factors play a significant role in the performance of the models we'll be discussing.

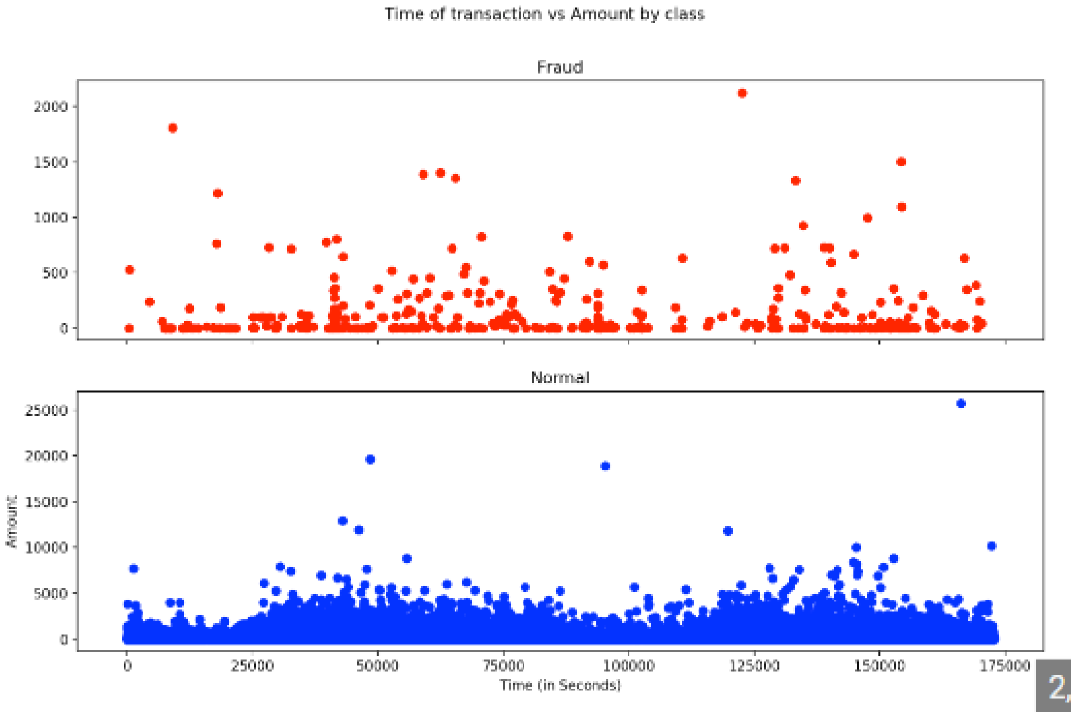


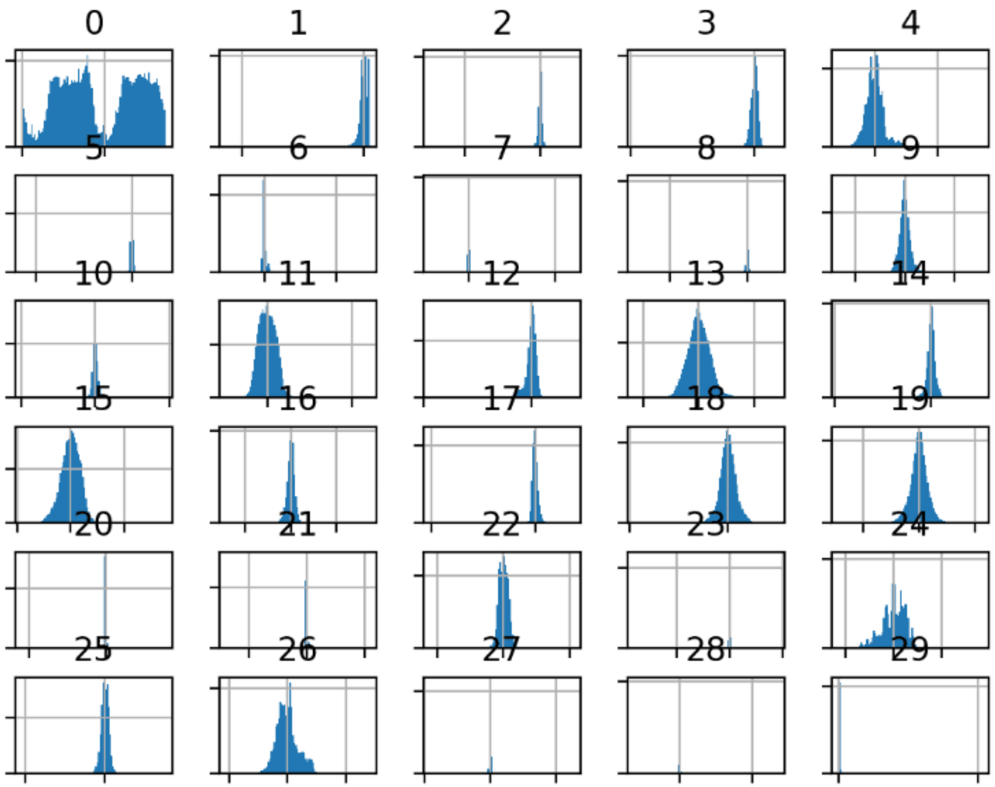
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

5 rows × 31 columns

FEATURES :

- The Data has 32 features from V1-V28 which are unknown for confidentiality, TIme, Amount and Class
- The input features are V1-V28, Time and Amount
- The target variable is Class
- The Data does not have any missing values as evident from the below mentioned code, thus need not be handled
- The Data consists of all numerical features, and only the Target Variable Class is a categorical feature.
  - Class 0: Legitimate Transaction
  - Class 1: Fraud Transaction


Time of transaction vs Amount by class
Fraud
2000
1500
1000
500
0
Normal
25000
20000
15000
10000
5000
0
Amount
0
25000
50000
75000
100000
125000
150000
175000
Time (in Seconds)


0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

PREPROCESSING TECHNIQUES

- Preprocessing techniques are essential for handling imbalanced datasets. These methods help balance the class distribution, which is crucial for training accurate fraud detection models

Undersampling-

- Description: Reduces instances in the majority class to balance the dataset.
- Advantages: Faster training, reduced computational resources.
- Considerations: May lead to loss of information.

Oversampling

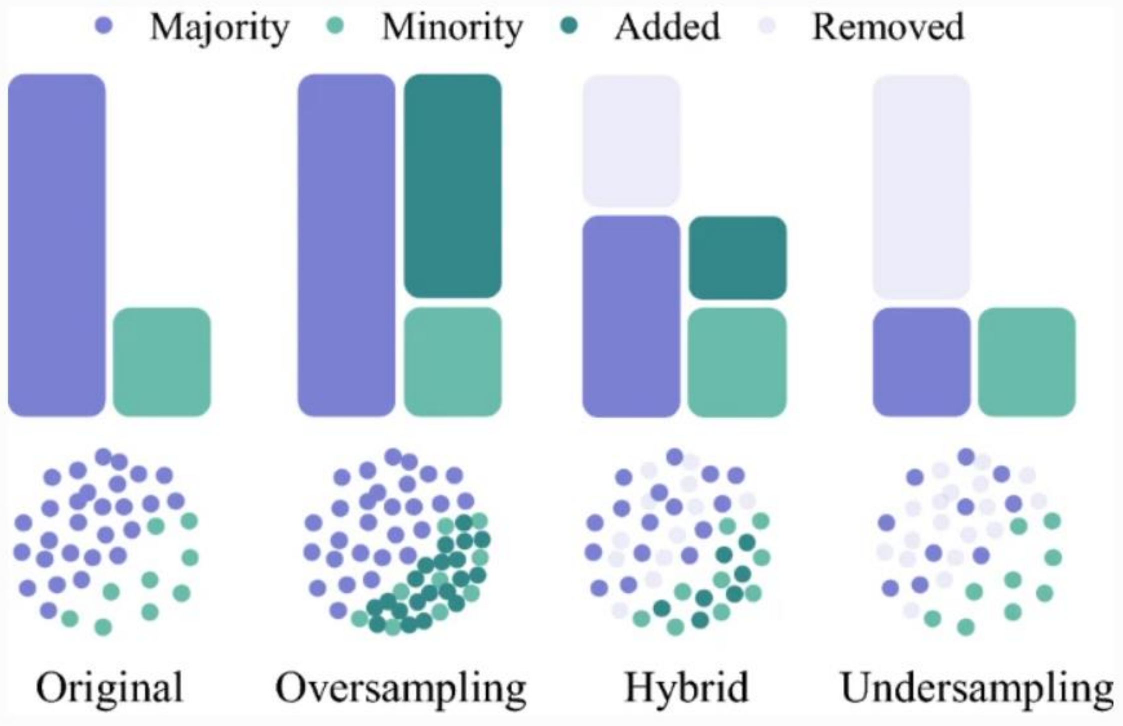
Description: Increases instances in the minority class to balance the dataset.

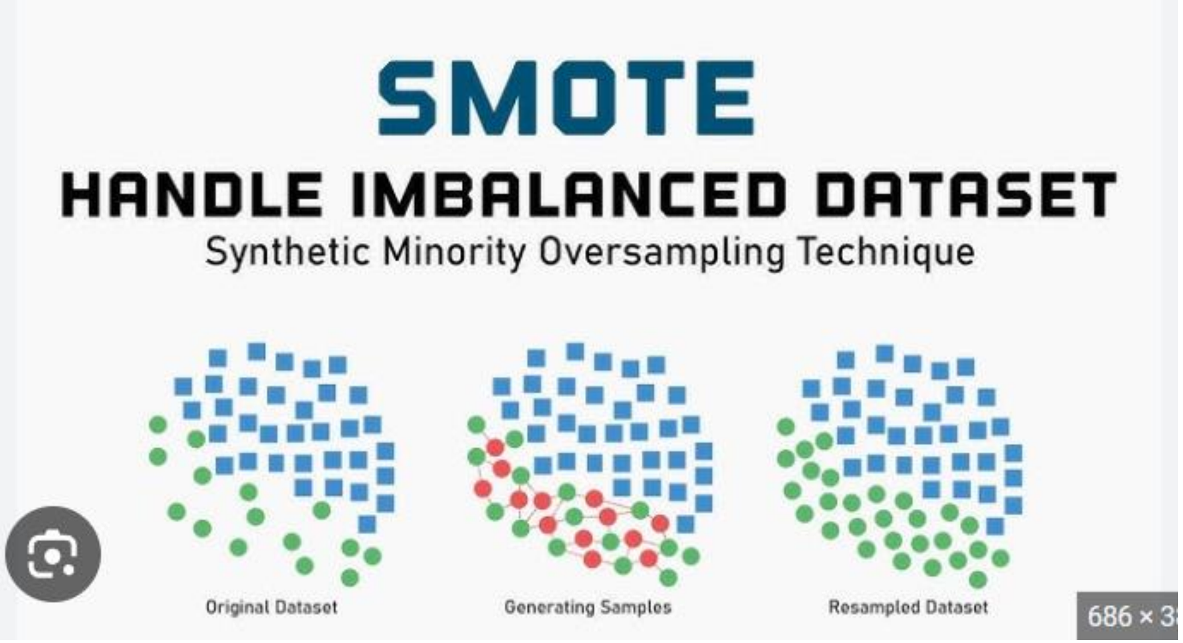
Advantages: Helps prevent loss of information in the minority class.

Considerations: Careful implementation needed to avoid overfitting.

SMOTE (Synthetic Minority Over-sampling Technique):
Description: Generates synthetic samples for the minority class using interpolation.
Advantages: Addresses overfitting, creates realistic synthetic samples.
Considerations: May introduce noise if not applied appropriately.

ig. 1




SMOTE

HANDLE IMBALANCED DATASET

Synthetic Minority Oversampling Technique

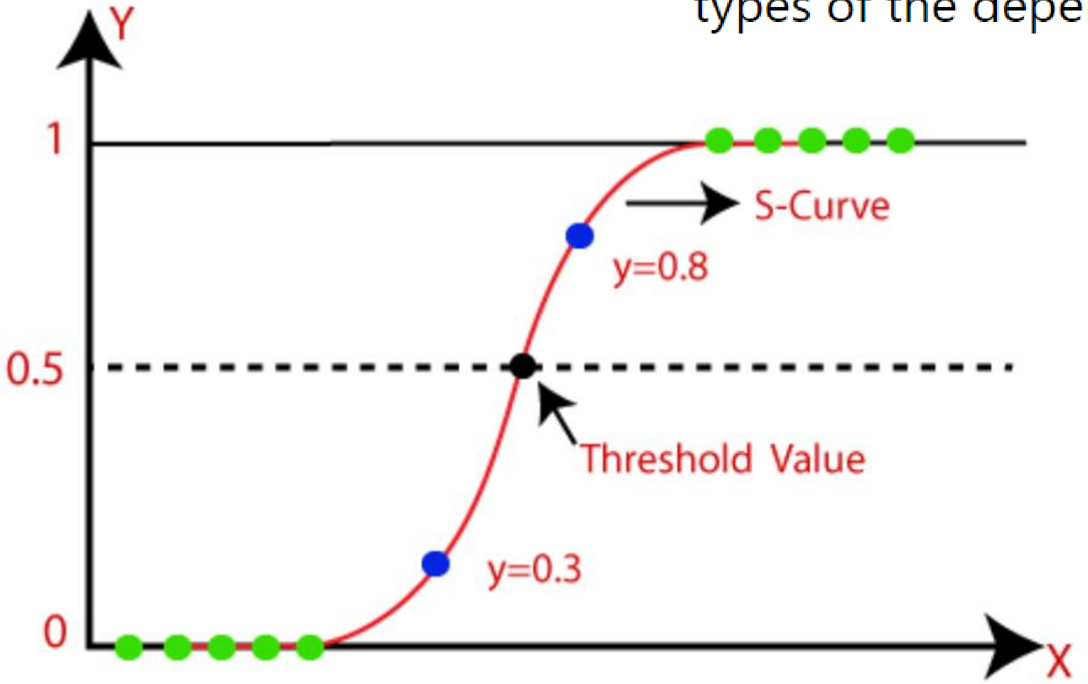
HANDLING IMBALANCED DATASETS

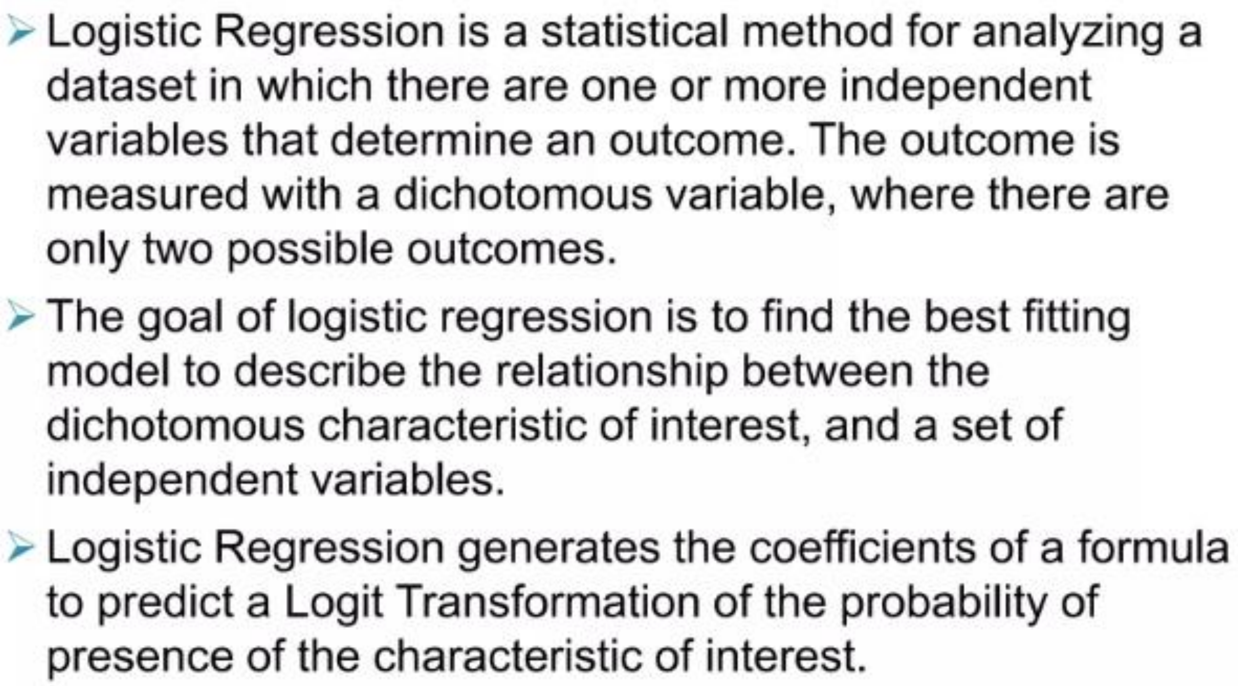
DATA MODELLING (ML TECHNIQUES AND MODELS/CLASSIFIER)

- Logistic Regression: Logistic Regression is a simple yet effective classifier used for binary and multiclass classification tasks. It models the probability of a data point belonging to a particular class and is widely used in applications like medical diagnosis and spam email detection.
- K-Nearest Neighbors (KNN): KNN is a non-parametric classifier that assigns a class label to a data point based on the classes of its k-nearest neighbors in the feature space. It's intuitive and easy to understand, making it suitable for applications such as recommendation systems and anomaly detection.
- Random Forest: Random Forest is an ensemble classifier that combines multiple decision trees. It reduces overfitting by considering random subsets of data and features, offering high accuracy and robustness. It's applicable in finance, healthcare, and image recognition.

LOGISTIC REGRESSION

- **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.





- Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable, where there are only two possible outcomes.
- The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest, and a set of independent variables.
- Logistic Regression generates the coefficients of a formula to predict a Logit Transformation of the probability of presence of the characteristic of interest.

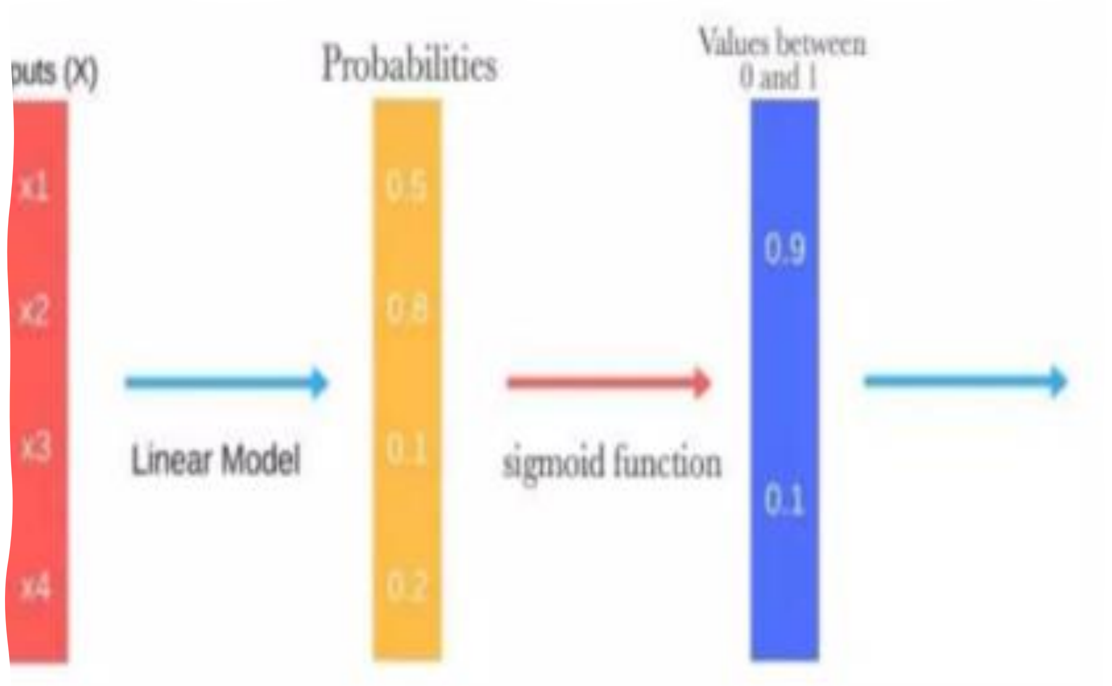
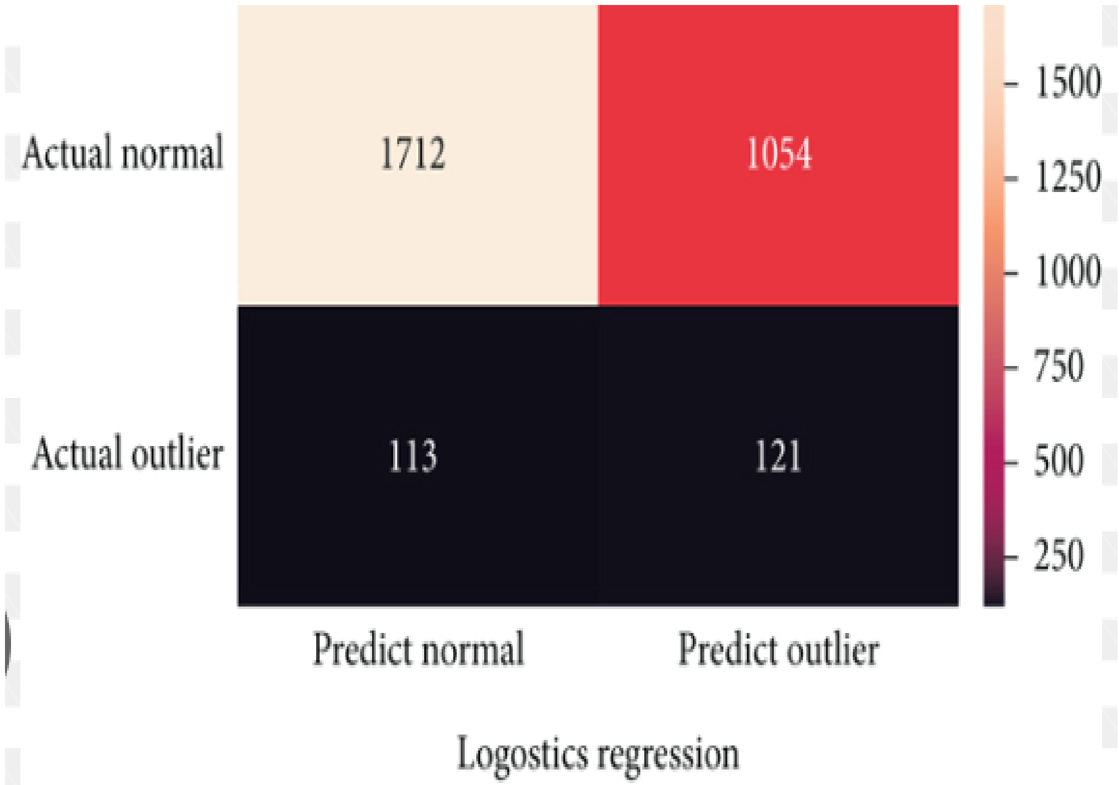
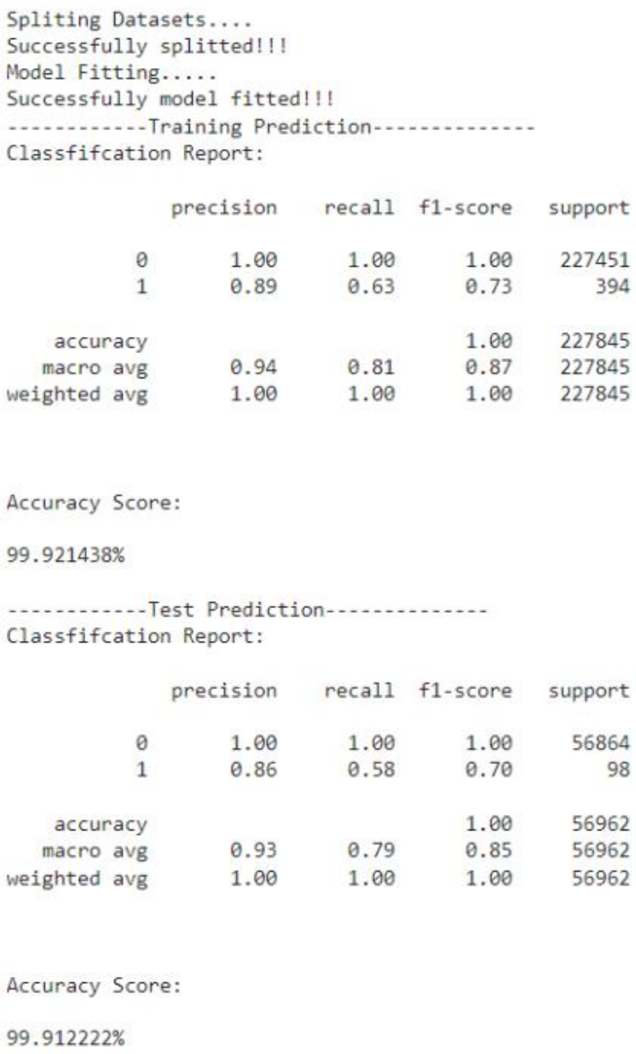


Fig.3 Working of Logistic Regression Model

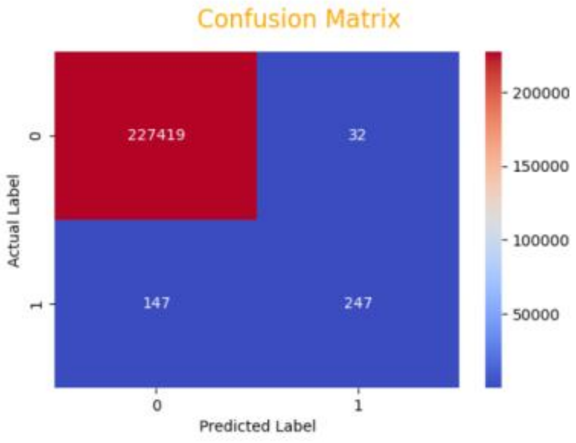


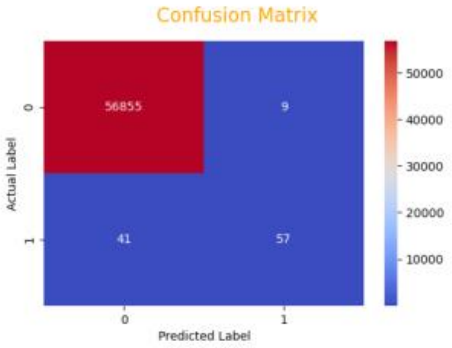
As the performances are shown in Figure 10, although the SMOTE could detect outliers, its accuracy is relatively low. We believe it is because the dataset is super high dimensional, so even though with the technique like SMOTE to oversample the dataset, it still seems very complex for classifiers to process efficiently. However, IForest, according to its theory, could not only detect outliers with high efficiency but also maintain a very high level of accuracy and performance. Moreover, the IForest process the dataset within one second, which is much lower than SMOTE which required ten seconds.

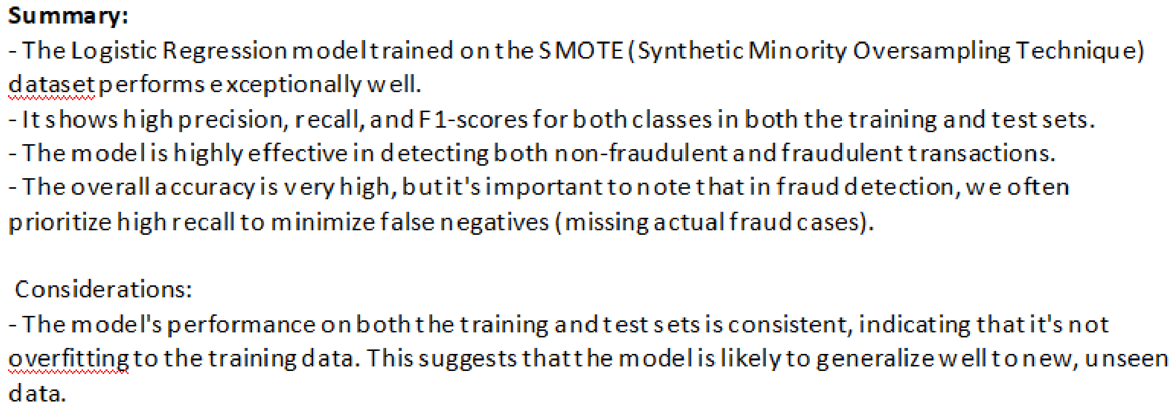
PERFORMANCE ANALYSIS



```
Spliting Datasets....
Successfully splitted!!!
Model Fitting.....
Successfully model fitted!!!
------------Training Prediction-------------
Classficfation Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    227451
           1       0.89      0.63      0.73       394

    accuracy                           1.00    227845
   macro avg       0.94      0.81      0.87    227845
weighted avg       1.00      1.00      1.00    227845


Accuracy Score:

99.921438%

------------Test Prediction-------------
Classficfation Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56864
           1       0.86      0.58      0.70        98

    accuracy                           1.00     56962
   macro avg       0.93      0.79      0.85     56962
weighted avg       1.00      1.00      1.00     56962


Accuracy Score:

99.912222%
```







Summary:

- The Logistic Regression model trained on the SMOTE (Synthetic Minority Oversampling Technique) dataset performs exceptionally well.
- It shows high precision, recall, and F1-scores for both classes in both the training and test sets.
- The model is highly effective in detecting both non-fraudulent and fraudulent transactions.
- The overall accuracy is very high, but it's important to note that in fraud detection, we often prioritize high recall to minimize false negatives (missing actual fraud cases).

Considerations:

- The model's performance on both the training and test sets is consistent, indicating that it's not overfitting to the training data. This suggests that the model is likely to generalize well to new, unseen data.

CONCLUSION:

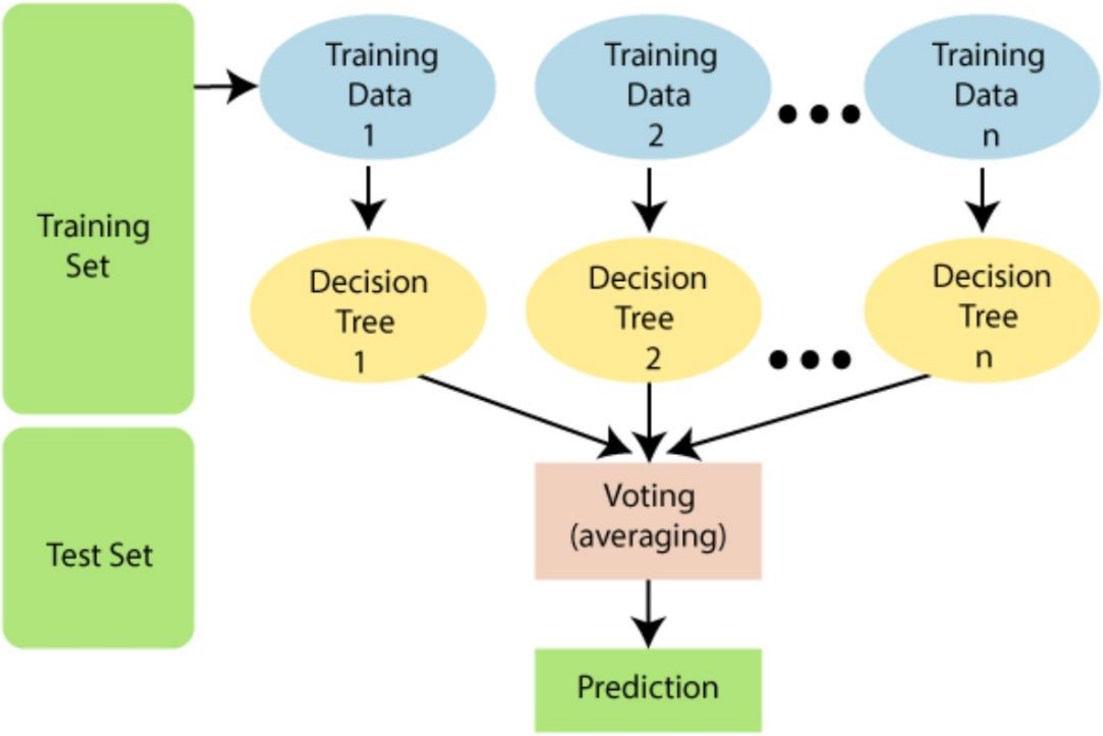
a) Logistic Regression doesn't work efficiently for this imbalanced datasets.

b) It takes around 1-2 minutes for training.

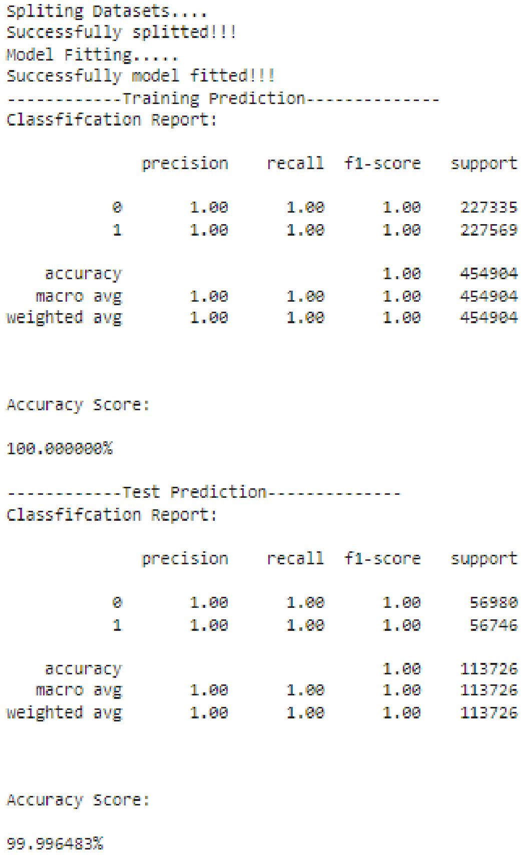
c) Maximum Accuracy of 99.912222% and Macro Average of F1-Score of 0.85 acheived with StandardScaled datasets.

RANDOM FOREST CLASSIFIER

- Random Forest is an ensemble learning technique used for classification and regression tasks. It combines multiple decision trees, each trained on random subsets of data and features, reducing overfitting. By aggregating predictions, it provides accurate results. Its benefits include reduced overfitting, high accuracy, implicit feature selection, and robustness to outliers. It finds wide applications in fields like finance, healthcare, and image recognition, but can be computationally expensive for a large number of trees and less effective on very high-dimensional data.

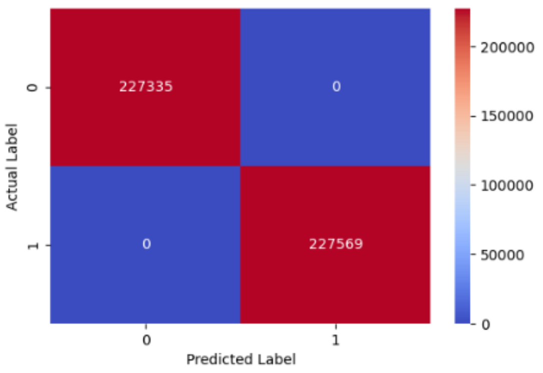


PERFORMANCE ANALYSIS

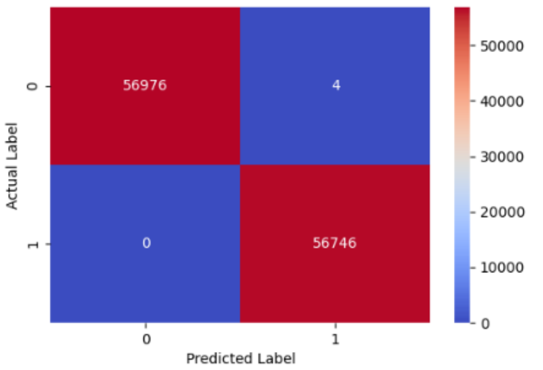


```
Spliting Datasets....
Successfully splitted!!!
Model Fitting.....
Successfully model fitted!!!
------------Training Prediction-------------
Classfifcation Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    227335
           1       1.00      1.00      1.00    227569

    accuracy                           1.00    454904
   macro avg       1.00      1.00      1.00    454904
weighted avg       1.00      1.00      1.00    454904


Accuracy Score:

100.000000%

------------Test Prediction-------------
Classfifcation Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56980
           1       1.00      1.00      1.00     56746

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726


Accuracy Score:

99.996483%
```

Confusion Matrix of Training Datasets



Confusion Matrix of Testing Datasets



CONCLUSION:

a) Undersampling doesn't work efficiently for Large majority class datasets as it ignore many valuable tuples. But, can be efficient for small majority class datasets

b) RandomForest works even efficiently for this imbalanced datasets.

c) RandomForest takes around 10-15 minutes for training.

d) Maximum `Accuracy` of `99.996483%` and `macro-average of F1-Score` of `1.00` acheived with `Oversampling` technique.

K- NEAREST NEIGHBOUR CLASSIFIER MODEL

Formula to calculate Euclidean distance

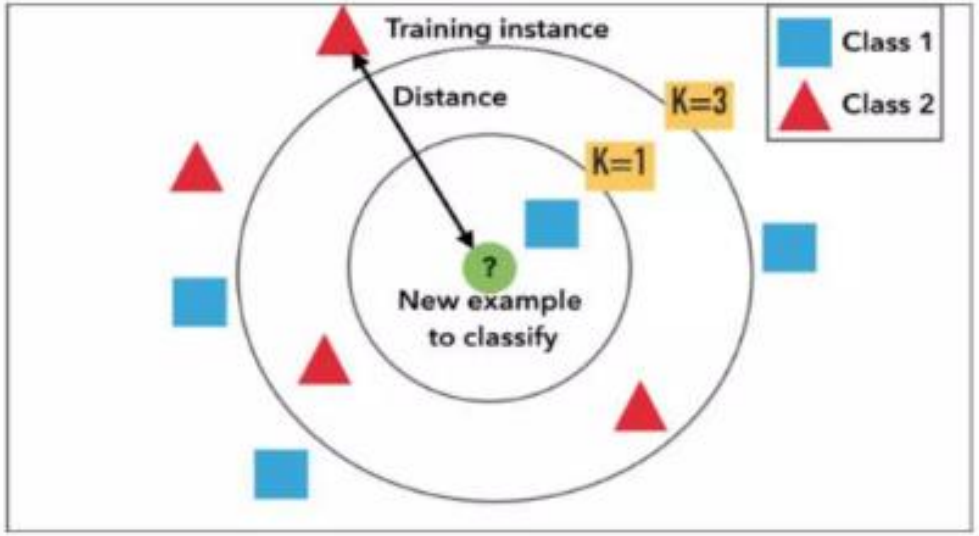
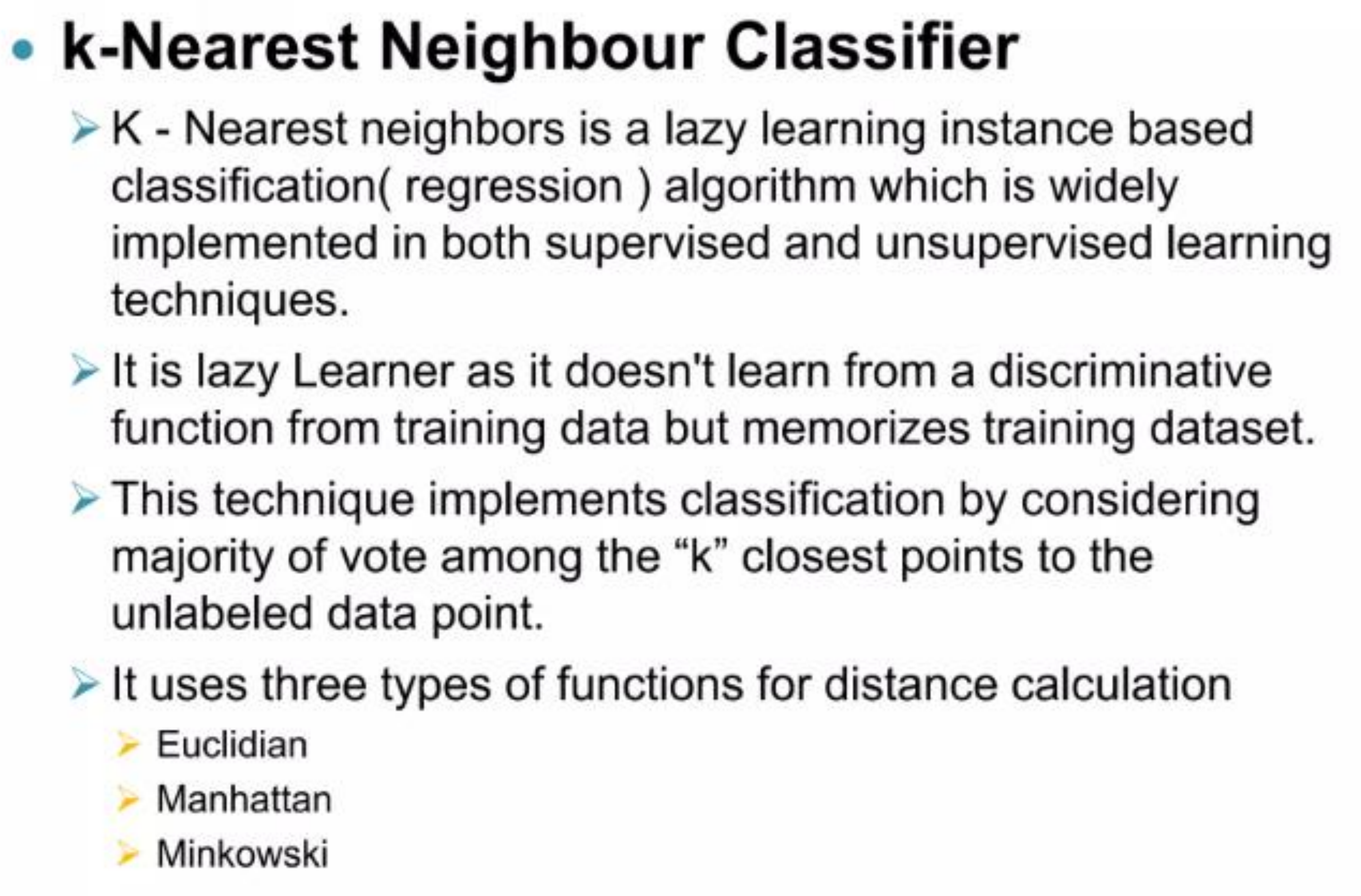
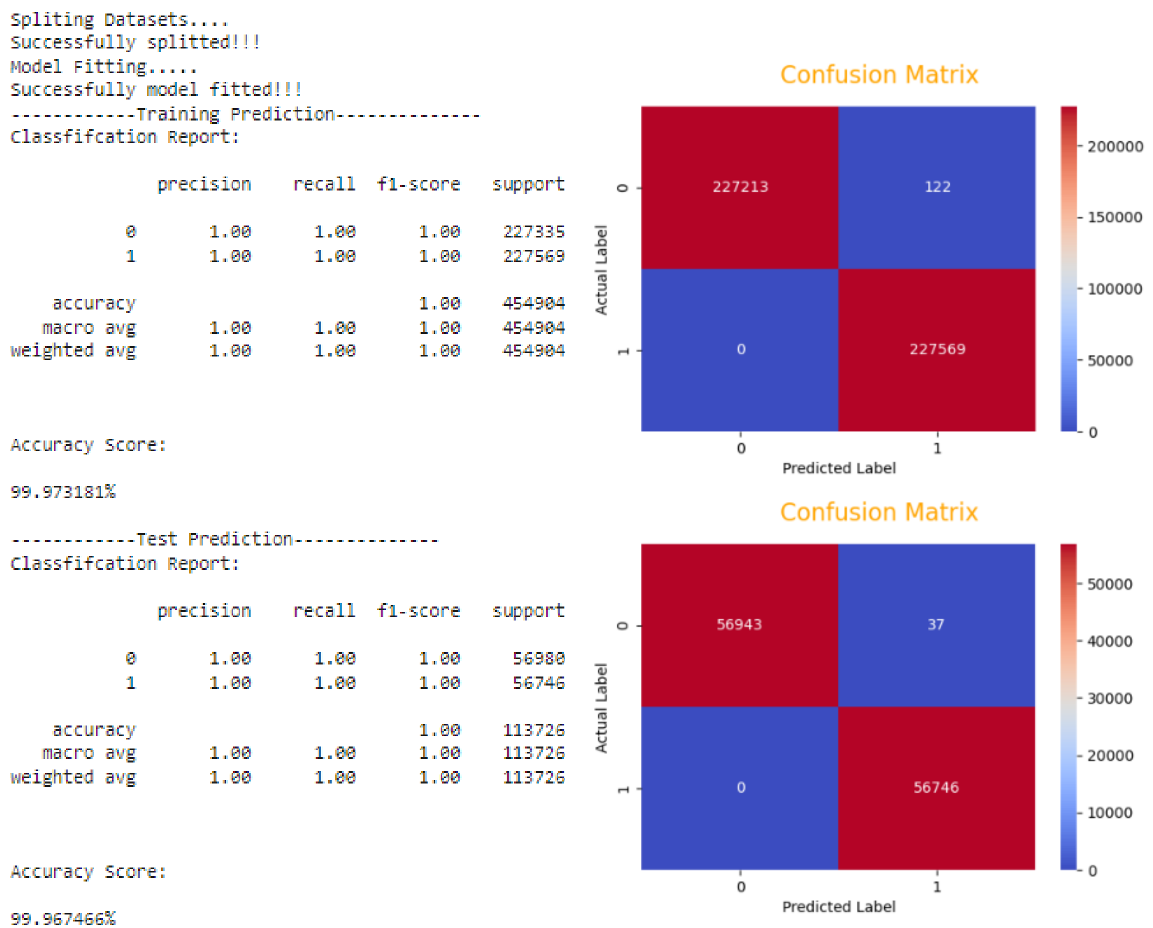


Fig 2. Example of k-NN classification



- **k-Nearest Neighbour Classifier**
  - K - Nearest neighbors is a lazy learning instance based classification(regression) algorithm which is widely implemented in both supervised and unsupervised learning techniques.
  - It is lazy Learner as it doesn't learn from a discriminative function from training data but memorizes training dataset.
  - This technique implements classification by considering majority of vote among the “k” closest points to the unlabeled data point.
  - It uses three types of functions for distance calculation
    - Euclidian
    - Manhattan
    - Minkowski



```
Spliting Datasets....
Successfully splitted!!!
Model Fitting.....
Successfully model fitted!!!
------------Training Prediction-------------
Classficcation Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    227335
           1       1.00      1.00      1.00    227569

    accuracy                           1.00    454904
   macro avg       1.00      1.00      1.00    454904
weighted avg       1.00      1.00      1.00    454904


Accuracy Score:

99.973181%

------------Test Prediction-------------
Classficcation Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56980
           1       1.00      1.00      1.00     56746

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726


Accuracy Score:

99.967466%
```

CONCLUSION:

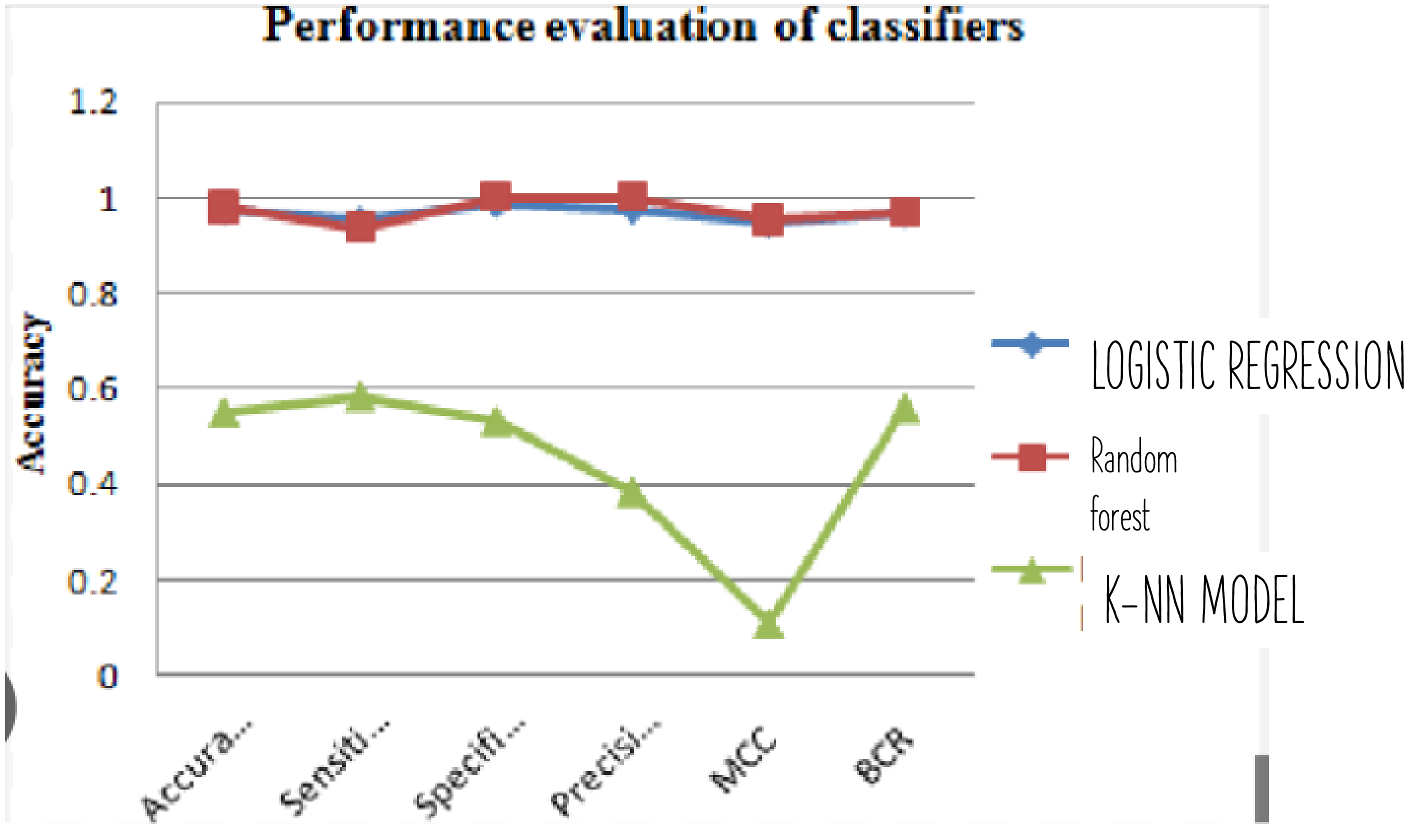
a) k-Neighbors works even efficiently for this imbalanced datasets.

b) It takes around 3-5 minutes for training.

c) Maximum `Accuracy` of `99.967466 %` and `Macro Average of F1-Score` of `1.00` acheived with `Oversampling` Technique

PERFORMANCE ANALYSIS

OVERALL ANALYSIS



SUMMARY

- The models achieved high accuracy across the board, but F1-Score for the minority class is more critical in this imbalanced setting. Random Forest and K-NN outperform Logistic Regression in this aspect.
- Computation Time:
- Logistic Regression is the fastest to train, followed by K-NN. Random Forest, while highly accurate, is more computationally intensive.
- Recommended Model:
- For this imbalanced dataset, the Random Forest Classifier with oversampling or SMOTE is recommended. It achieves near-perfect accuracy and F1-Scores, indicating robust performance in identifying the minority class.

CONCLUSION

LOGISTIC REGRESSION OFFERS INTERPRETABILITY AND EFFICIENCY BUT ASSUMES LINEAR RELATIONSHIPS AND MAY NOT HANDLE COMPLEX PATTERNS WELL.

K-NEAREST NEIGHBORS (KNN) IS NON-PARAMETRIC AND VERSATILE, BUT CAN BE COMPUTATIONALLY INTENSIVE AND SENSITIVE TO OUTLIERS.

RANDOM FOREST CLASSIFIER IS AN ENSEMBLE METHOD PROVIDING HIGH ACCURACY, CAPTURING NON-LINEAR RELATIONSHIPS, AND OFFERING FEATURE IMPORTANCE. HOWEVER, IT CAN BE COMPUTATIONALLY EXPENSIVE AND LESS INTERPRETABLE.

FOR THE SPECIFIC TASK OF CREDIT CARD FRAUD DETECTION, THE RANDOM FOREST CLASSIFIER APPEARS TO BE THE MOST PROMISING CHOICE DUE TO ITS ABILITY TO HANDLE COMPLEX PATTERNS AND PROVIDE VALUABLE INSIGHTS THROUGH FEATURE IMPORTANCE. HOWEVER, CAREFUL DATA PREPROCESSING AND HYPERPARAMETER TUNING ARE CRUCIAL FOR OPTIMAL PERFORMANCE.

THANK YOU