Project Design Phase-I
Proposed Solution Template

Date	19 October 2023
Team ID	Team-591653
Project Name	Online Fraud Detection
Maximum Marks	2 Marks

Problem Description:

The proposed methodology aims to predict the legality of transactions and detect fraudulent transactions, making it a classification problem. To achieve the highest possible prediction accuracy, supervised machine learning models will be deployed

Solution Architecture:

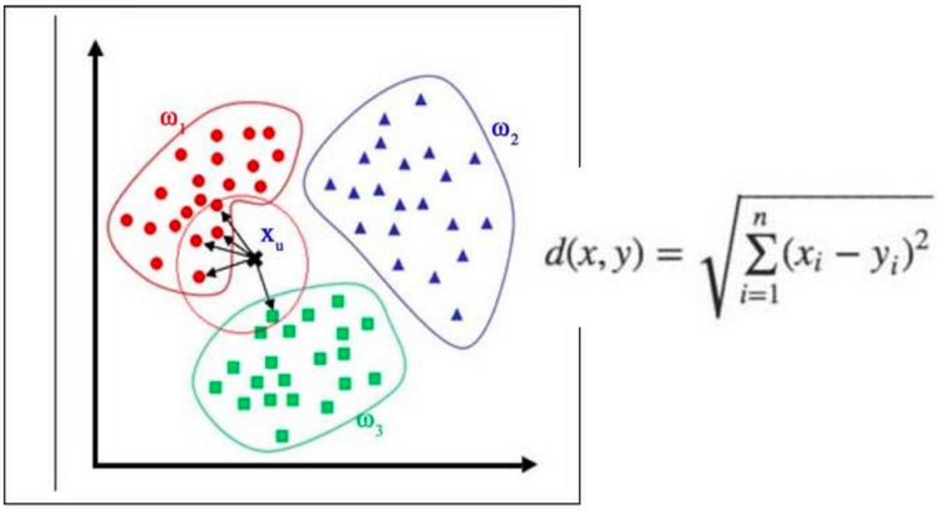
The objective of this study is to predict whether a transaction is legitimate or fraudulent. The proposed solution to address these issues involves undertaking a comprehensive data exploration and cleaning process, followed by selecting an appropriate machine learning algorithm.

- **Data Acquisition**
- **Exploratory Data Analysis**
- **Feature Engineering**
- **Data Processing**
- **Supervised Machine Learning Model Deployment**
- **Results Analysis**

Machine Learning Models:-

K Nearest Neighbor

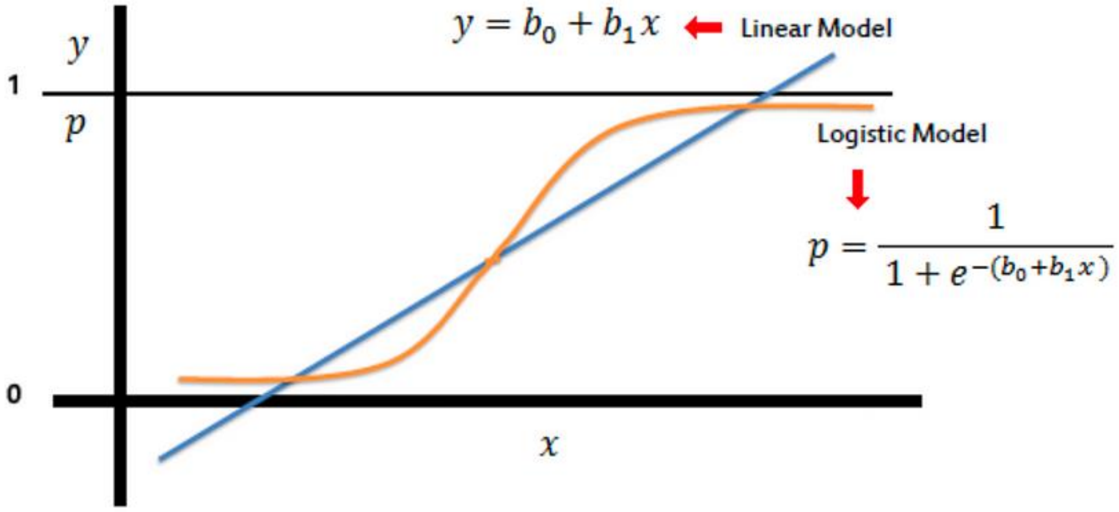
KNN is a non-parametric classification approach for solving classification and regression issues. KNN does not do any generalization, resulting in a relatively quick training procedure. Because of the lack of generalization, the KNN training phase is either small or retains all of the training data. The value *k* (number of nearest neighbors) is user defined.



K Nearest Neighbor algorithm is suitable for classification of fraud transactions, so by selecting the optimal nearest neighbor we can use K nearest neighbor to classify a transaction as legal or fraudulent.

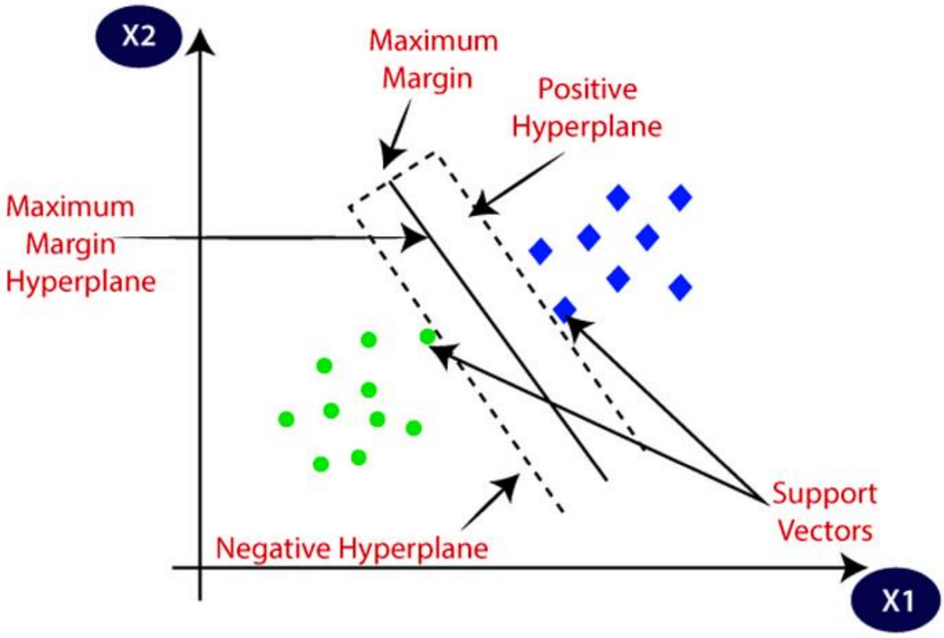
Logistic Regression

Logistic regression is a classification procedure that is used to forecast the likelihood of a target variable. The target or dependent variable has a dichotomous character, which means there are only two potential classes. The representation for logistic regression is an equation. To anticipate an output value, input data are linearly mixed with coefficient values. The output value is represented as a binary value, which distinguishes it from linear regression.



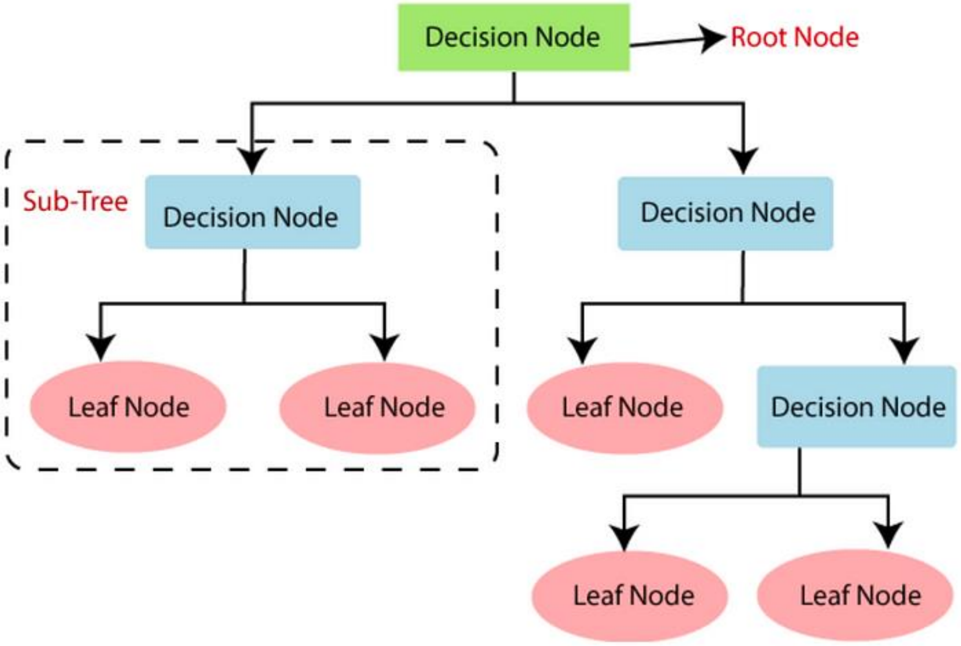
Support Vector Machine

Support vector machine is a set of supervised learning methods used for classification, regression, and outlier detection. Different planes (hyperplanes) could be chosen, to separate the data points into two classes. Given a series of training examples, each labeled as belonging to one of two categories, an SVM training method constructs a model that assigns future instances to one of the two categories, resulting in a non-probabilistic binary linear classifier.



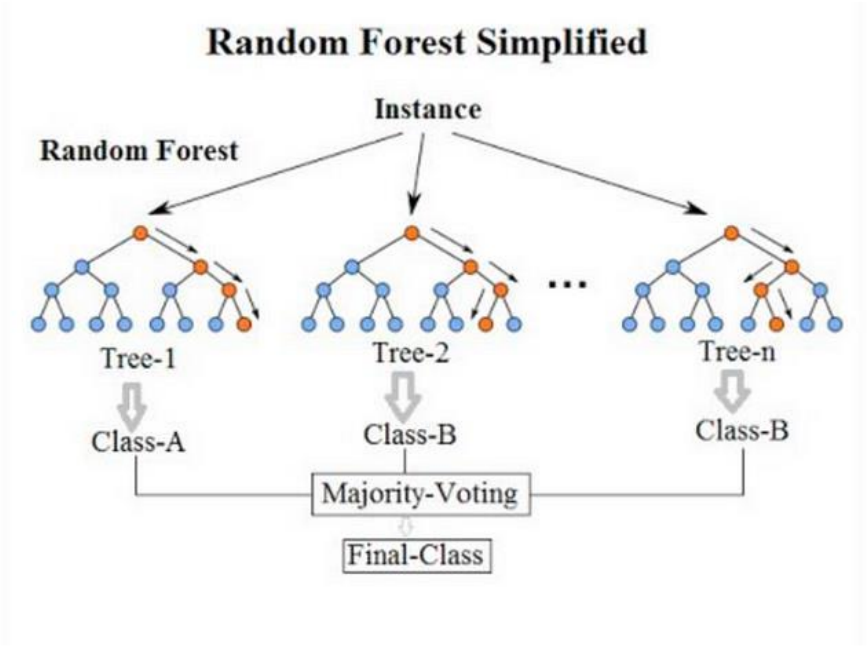
Decision Tree

A decision tree is a decision-making tool that employs a tree-like model of decisions and their potential outcomes, such as chance event outcomes, resource costs, and utility. It is one method of displaying an algorithm that consists solely of conditional control statements. Decision trees are a prominent method in machine learning and are often used in operations research, notably in decision analysis, to assist determine the approach most likely to achieve a goal.



Random Forest

Random forest is an ensemble approach, because of its versatility and simplicity, it is one of the most often used algorithms. This model employs a large number of decision trees. Each of these decision trees separates a class of predictions, and the class with the most votes becomes our model’s final output prediction. While growing trees in a random forest, rather than looking for the most significant characteristics for splitting, it seeks for the best features from a random selection of features for splitting the nodes. This results in a wide range of variety, which will provide us with a more accurate model. Because there is no association between the many models developed, the models provide ensemble forecasts that are more accurate than any of the individual projections. This is due to the fact that although certain trees may be incorrect.



Analysis and Results

Visualizations are performed in each step, in order to highlight new insights about the underlying patterns and relationships contained within the data. The data analysis process for the deployment of classification models is based on the following steps.

i) Data Acquisition

Download data

Upload data in Python environment

ii) Data Exploration

Checking data head, info, summary statistics and null valuee

iii) Feature Engineering

We can see from the above data that only two type of transactions are classified as fraud so we will drop the remaining types to generalize the data and we will only keep Cash_out and Transfer type.

iv) Data Processing

After feature engineering we scaled the data.

We dropped the uninterested and unscaled features from the dataset.

Supervised Machine Learning

Modeling of the classifier system

Validation of the model

Visualization of the model

Visualization and interpretation of the results

Results analysis

1. Deployment of the models
2. Comparing prediction accuracy of ML models
3. Visualization of the results

Conclusions

The goal was to predict whether a transaction is a legal transaction or a fraudulent transaction, this falls under the scope of a classification problem. We intend to deploy Supervised Machine Learning models in order to achieve the highest prediction accuracy.