**PROPOSED METHODS**

In this project we will use three different Algorithms to find out the prediction of a card to real or fraud. Description of these Algorithms are given blew:

**Logistic Regression:**

This statistical classiﬁcation model based on probabilities detects the fraud using logistic curve. Since the value of this logistic curve varies from 0 to 1, it can be used to interpret class membership probabilities. The dataset fed as input to the model is being classiﬁed for training and testing the model. Post model training, it is tested for some minimum threshold cut-off value for prediction. Since the logistic regression, based on some threshold probabilities can divide the plane using a single line and divides dataset points into exactly two regions.

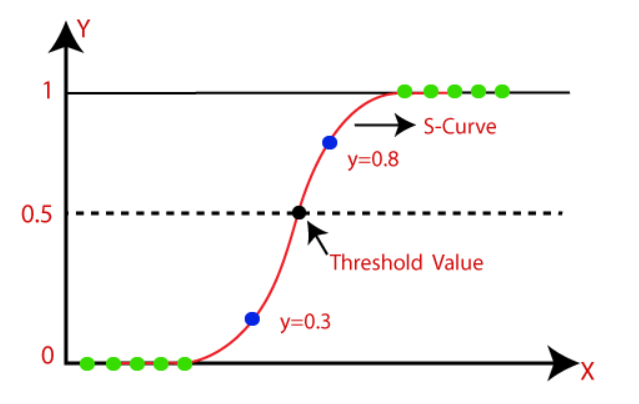


Fig: The logistic regression model

**K-Nearest Neighbor (KNN):**

This is a supervised learning technique that achieves consistently high performance in comparison to other fraud detection techniques of supervised statistical pattern recognition [24]. Three factors majorly affect its performance distance to identify the least distant neighbors, some rule to deduce a categorization from k-nearest neighbor & the count of neighbors to label the new sample. This algorithm classiﬁes any transactions that occurred by computing the least distant point to this particular transaction and if this least distant neighbor is classiﬁed as fraudulent then the new transaction is also labeled as a fraudulent one. Euclidean distance is a good choice to calculate the distances in this scenario. This technique is fast and results in fault alerts. Its performance can be improved by distance metric optimization.

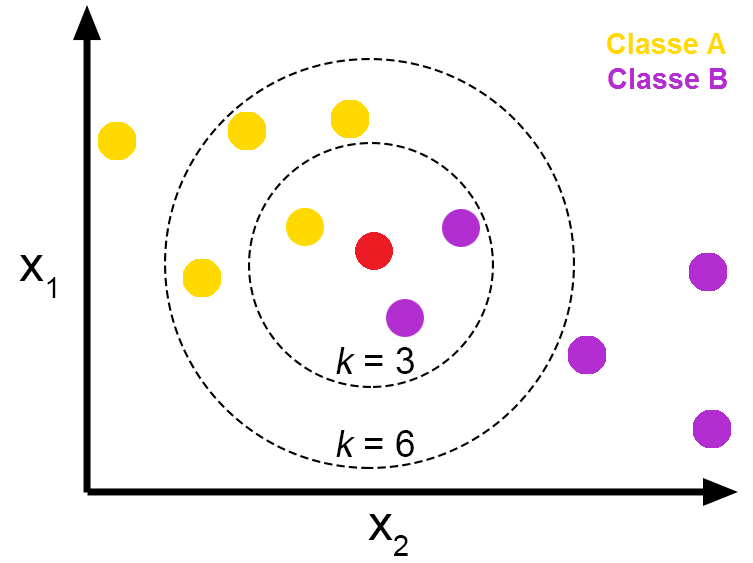


Fig: Pros and Cons of K-Nearest Neighbors - From The GENESIS

**Discission Tree:**

A supervised learning algorithm, A decision tree which is in the form of tree structure, consisting of root node and other nodes split in a binary or multi-split manner further into child nodes with each tree using its own algorithm to perform the splitting process. With the tree growing, there may be possibilities of overﬁtting of the training data with possible anomalies in branches, some errors or noise. Hence pruning is used for improving classiﬁcation performance of the tree by removing certain nodes. Ease in the use, and the ﬂexibility that the decision trees provide to handle different data types of attributes make them quite popular.



Fig: Decision Tree Algorithm in Machine Learning

**Support Vector Machine:**

Support vector machines or SVMs are linear classiﬁers as stated in that work in high dimensionality because in high-dimensions, a non-linear task in input becomes linear and hence this makes SVMs highly useful for detecting frauds. Due to its two most important features that is a kernel function to represent classiﬁcation function in the dot product of input data point projection, and the fact that it tries ﬁnding a hyperplane to maximize separation between classes while minimizing overﬁtting of training data, it provides a very high generalization capability.



Fig: Support Vector Machine algorithm.

**Dataset:**

In this research the Credit Card Fraud Detection dataset was used, which can be downloaded from Kaggle [8]. This dataset contains transactions, occurred in two days, made in September 2013 by European cardholders.

[Credit Card Fraud Detection](https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)

**PROJECT PLAN**

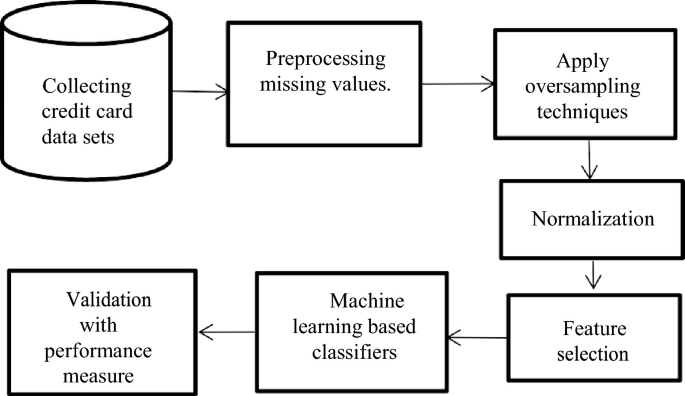


Fig: Project plan.

**The project will be completed in different phases:**

**Data collection**:

The first phase will involve collecting a dataset of historical credit card transactions. The data will be collected from a variety of sources, including banks, credit card companies, and merchants.

**Data Cleaning:**

* Impute the missing values with the mean, median, or mode of the column.
* Drop the rows with missing values.
* Use a machine learning model to predict the missing values like isnull(), heatmap().

**Normalize the data:**

Normalization is the process of scaling the data so that all of the features have a similar range of values. This can help to improve the performance of machine learning models by making the features more comparable.

**Model training:**

The second phase will involve training the machine learning model on the collected data. The model will be trained using a supervised learning algorithm, such as SVM.

**Model evaluation:**

The third phase will involve evaluating the performance of the machine learning model on a holdout dataset of unseen transactions. The performance of the model will be evaluated using metrics such as accuracy, precision, and recall.

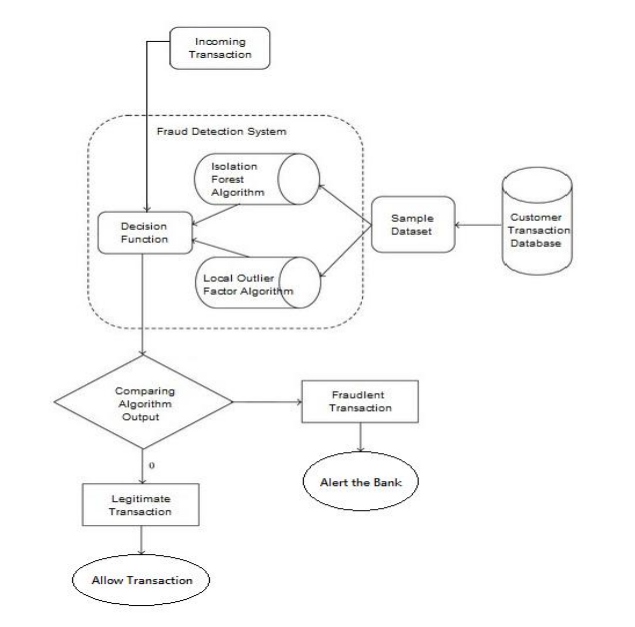
****

Fig: Working Flow of Credit Card Fraud Detection

**Results and Evaluations**

**Expected Result:**

* A machine learning model that can detect credit card fraud with high accuracy.
* A better understanding of the patterns that are indicative of fraudulent transactions.
* A framework for using machine learning to detect credit card fraud in real-time.

**Performance Metrics and Evaluation Methodology:**

**Confusion Metrics:**

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.



**Classification Report:**

A screenshot of a computer screen

Description automatically generated