***PROJECT PROPOSAL***

**Credit Card Fraud Detection using**

**Machine Learning**

1.Decision Tree

2.KNN

3.LOGISTIC REGRESSION

4.SVM

**Submitted By:**

**AVULA MALLISHWARI**

**211EE111**

**ABSTRACT**

Credit card fraud is a major problem, with billions of dollars lost each year. Machine learning can be used to detect credit card fraud by identifying patterns that are indicative of fraudulent transactions. Credit card fraud refers to the physical loss of credit card or loss of sensitive credit card information. Many machine- learning algorithms can be used for detection. This project proposes to develop a machine learning model to detect credit card fraud. The model will be trained on a dataset of historical credit card transactions, and it will be evaluated on a holdout dataset of unseen transactions.

**INTRODUCTION**

'Fraud’ in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Fraud has been increasing drastically with the progression of state-of-art technology and worldwide communication. Credit cards are one of the most popular objectives of fraud but not the only one. Credit card fraud, wide-ranging term for theft and fraud committed or any similar payment mechanism as a fraudulent resource of funds in a transaction. Credit card fraud has been expanding issue in the credit card industry. Detecting credit card fraud is a difficult task when using normal process, so the development of the credit card fraud detection models has become of importance whether in the academic or business organizations currently. Fraud can be avoided in two main ways: prevention and detection. Prevention avoids any attacks from fraudsters by acting as a layer of protection. Detection happens once the prevention has already failed. Therefore, detection helps in identifying and alerting as soon as a fraudulent transaction is being triggered.

Machine learning is this generation's solution which replaces such methodologies and can work on large datasets which is not easily possible for human beings. Machine learning techniques fall into two main categories: supervised learning and unsupervised learning. Fraud detection can be done in either way and only can be decided when to use according to the dataset. Supervised learning requires prior classification to anomalies. During the last few years, several supervised algorithms have been used in detecting credit card fraud. The data which is being used in this study is analyzed in two main ways: as categorical data and as numerical data. The dataset originally comes with categorical data. The raw data can be prepared by data cleaning and other basic preprocessing techniques. First, categorical data can be transformed into numerical data and then appropriate techniques are applied to do the evaluation. Secondly, categorical data is used in the machine learning techniques to find the optimal algorithm.

This project consists of selecting optimal algorithms for fraud patterns through an extensive comparison of machine learning such as Logistic Regression, KNN Neighbors, Decision Tree. Techniques via an effective performance measure for the detection of fraudulent credit card transactions. The rest of this paper is presented as follows. Section 2 presents the literature review. Section 3 provides the experimental methodology including results. Finally, conclusions and discussions of the paper are presented in Section 4.

**LITERATURE REVIEW**

In earlier studies, many approaches have been proposed to bring solutions to detect fraud from supervised approaches, unsupervised approaches to hybrid ones, which makes it a must to learn the technologies associated in credit card frauds detection and to have a clear understanding of the types of credit card fraud. With the analysis of various detection models, past researchers have found many problems regarding fraud detection. Classical algorithms such as Support Vector Machines (SVM), Decision Tree (DT), LR and RF proven useful.

In paper [1], European dataset was also used, and comparison was made between the models based on LR, DT and RF. Among the three models, RF proved to be the best, with accuracy of 95.5%, followed by DT with 94.3% and LR with accuracy of 90%.

According to [2] and [3], k-Nearest neighbors (KNN) and outlier detection techniques can also be efficient in fraud detection. They are proven useful in minimizing false alarm rates and increasing fraud detection rate.

KNN algorithm also performed well in experiment for paper [4], where the authors tested and compared it with other classical algorithms. The paper [5] discussed commonly used supervised techniques and they have provided a thorough evaluation of supervised learning techniques. Also, they have shown that all algorithms change according to the problem area.

Fraud detection system presented in paper [6] is built to handle class imbalance, the formation of labelled and unlabeled, and processing of large datasets. The proposed system was able to overcome all the challenges.

In paper [7] they have highlighted fraud detection cost and lack of adaptability as challenges in the fraud detection process. When considering a system, the cost of fraudulent behavior and the prevention cost should be taken into consideration. Lack of adaptability occurs when the algorithm is exposed to new types of fraud patterns and normal transactions.

**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.

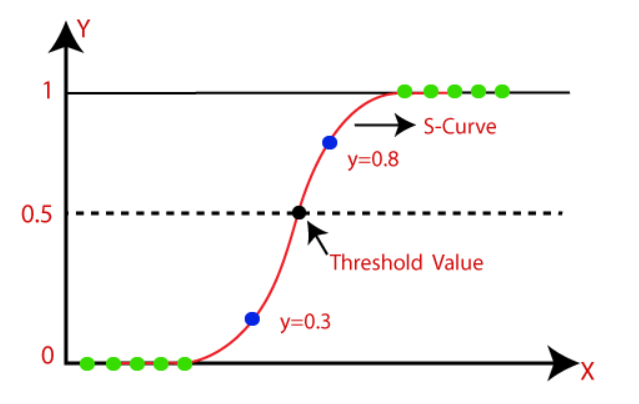


Fig: The logistic regression model

**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.

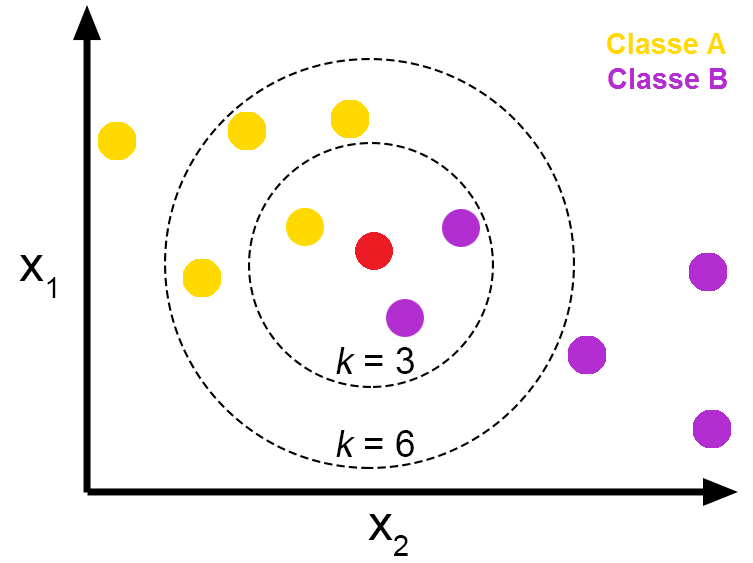


Fig: Pros and Cons of K-Nearest Neighbors - From The GENESIS

**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**

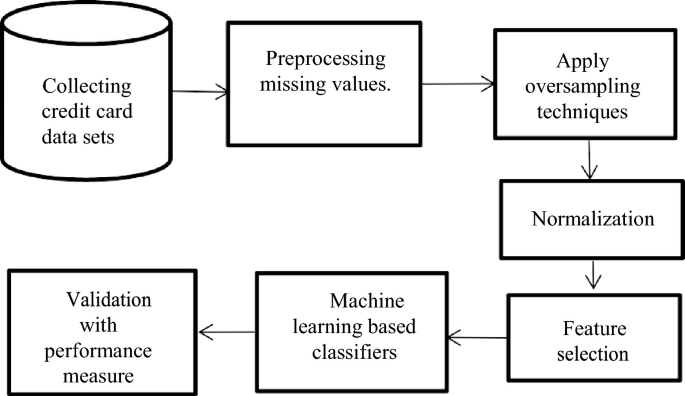


Fig: Project plan.

**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.

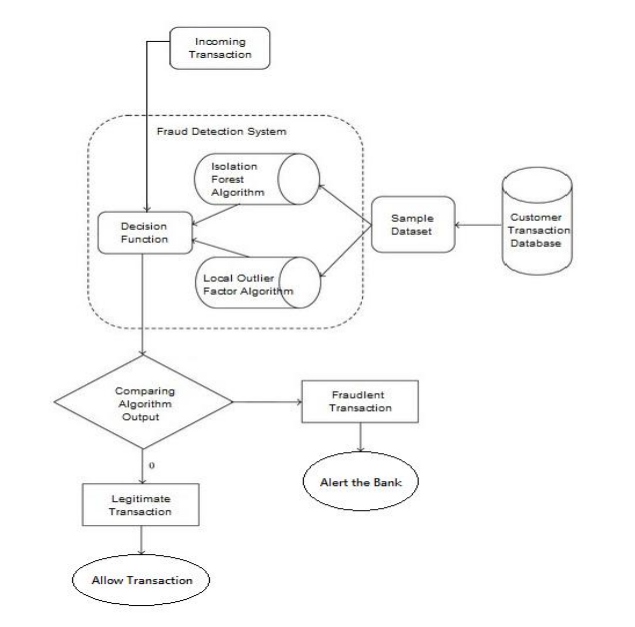
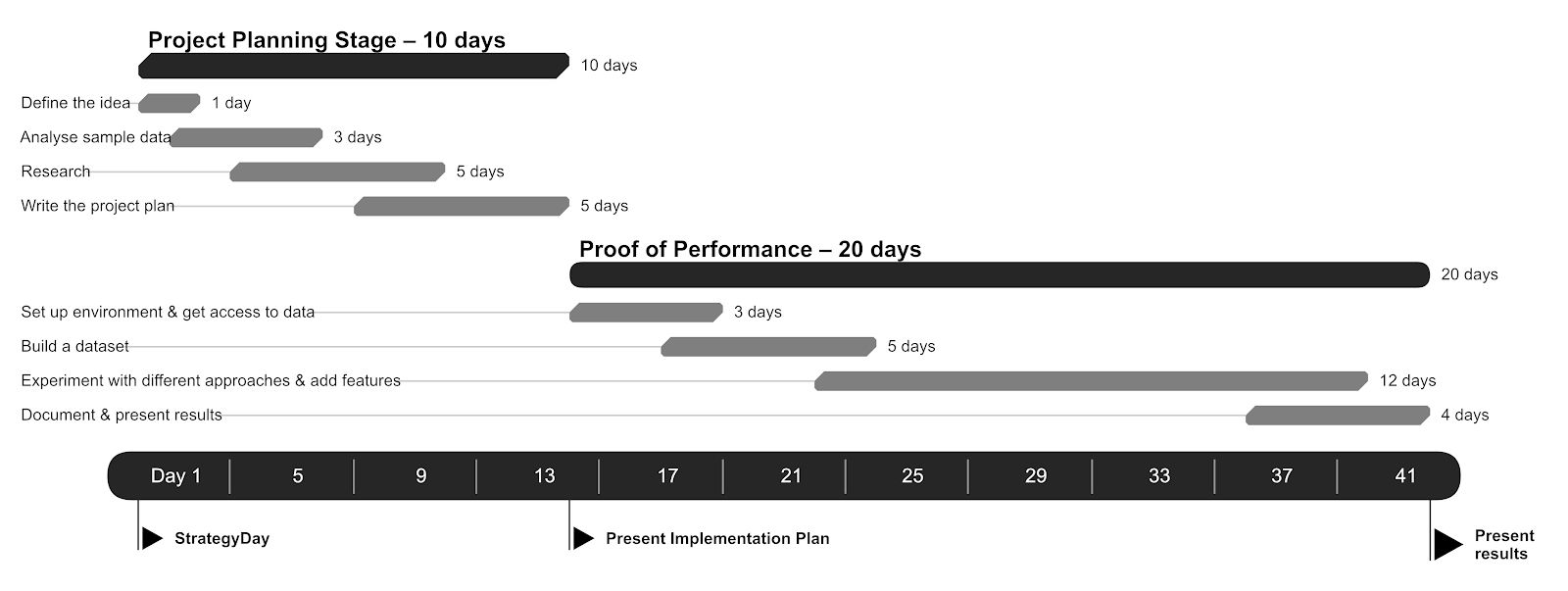
****

Fig: Working Flow of Credit Card Fraud Detection

**Timeline for Our Project:**



**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

**REFERENCES**

[1]. S. V. S. S. Lakshmi, S. D. Kavilla “Machine Learning for Credit Card Fraud Detection System”, unpublished

[2] N. Malini, Dr. M. Pushpa, “Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection “, Advances in Electrical, Electronics, Information, Communication and Bio- Informatics (AEEICB), 2017 Third International Conference on pp. 255-258. IEEE.

[3] Mrs. C. Navamani, M. Phil, S. Krishnan, “Credit Card Nearest Neighbor Based Outlier Detection Techniques”

[4] J. O. Awoyemi, A. O. Adentumbi, S. A. Oluwadare, “Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis”, Computing Networking and Informatics (ICCNI), 2017 International Conference on pp. 1-9. IEEE.

[5] R. Choudhary and H. K. Gianey 2017 Int. Conf. Mach. Learn. Data Sci., pp. 3743, 2017.

[6]. G. E. Melo-Acosta, F. Duitama-Muñoz, and J. D. Arias-Londoño, -supervised

Common. Compute. (COLCOM), 2017 IEEE Colomb. Conf., pp. 16, 2017.

[7]. Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented 26, 2016

[8]. Credit Card Fraud Detection dataset: downloaded from Kaggle, September 2013 by European cardholders.