Sanjivani Rural Education Society's

Sanjivani College of Engineering, Kopargaon-423603

(An Autonomous Institute Affiliated to Savitribai Phule Pune University, Pune)

(NAAC 'A' Grade Accredited, ISO 9001:2015 Certified)

Department of Computer Engineering

(NBA Accredited)

Report on

# **Finantial Fruad Detection via Data Mining**

Submitted by:

Roll No. 07 : Arote Pranjal Prakash

Roll No. 53: Gholap Rutuja Pravin

Roll No. 55: Gore Chaitanya Appasaheb

Roll No. 59 : Gupta Dewanshi Lokesh

Division A

T.Y.B.Tech

(Computer Dept.)

Guided By:

Prof. A.S. Bodhe

Sanjivani College of Engineering, Kopargaon-423603

(An Autonomous Institute Affiliated to Savitribai Phule Pune University, Pune)

CERTIFICATE

This is to certify that the report writing based ***Financial Fraud Detection System*** on being submitted by Arote Pranjal, Gholap Rutuja, Gore Chaitanya, Gupta Dewanshi, is a record of bonafide work carried out by her under the supervision and guidance of Prof. A.S. Bodhe in partial fulfilment of the requirement for T.Y.B.Tech. (Computer Engineering) of Savitribai Phule Pune University, Pune in the academic year 2023-24.

Date: /04/2024

Place: Kopargaon

**Under the Guidance of**

| | | |
|---|---|---|
| Prof. A.S. Bodhe | Dr. D. B. kshirsagar | Dr. A.G. Thakur |
| Guide | (HOD)<br>Computer Engineering | Director |

# ACKNOWLEDGEMENT

First and foremost, I express my deep sense of gratitude, sincere thanks to Prof. A.S. Bodhe, Department of Computer Engineering, Sanjivani College Of Engineering, Kopargaon. Your availability at any time throughout the semester, encouragement and support tremendously boosted this project work. Lots of thanks to Head of Computer Engineering Department, Dr. D.B. Kshirsagar for providing us the best support. I would like to express my sincere gratitude to Dr. A.G. Thakur, Director, Sanjivani College Of Engineering, Kopargaon for providing great platform to complete the project within the scheduled time.

07 | Arote Pranjal Prakash

53| Gholap Rutuja Pravin

55| Gore Chaitanya Appasaheb

59| Gupta Dewanshi Lokesh

## Index

# 1. Abstract

Fraudulent financial statements occur when individuals or organizations manipulate financial information by inflating income, assets, sales, and profits while downplaying expenses, debts, or losses. It is crucial to fight against financial crime, fraud, and cyberattacks. With the rise of digital technology, financial crime has shifted to cyberspace, where cybercriminals employ sophisticated tactics, including hacking and social engineering, to exploit weaknesses in financial and corporate institution security. In a world where wireless communications are critical for transferring massive quantities of data while protecting against interference, the growing possibility of financial fraud has become a significant concern. To uncover fraudulent financial statements, traditional methods like manual audits are expensive, imprecise, and slow. Intelligent methods, can be a game-changer for auditors, especially when dealing with a large volume of financial data. In report, we examine and consolidate the current research on intelligent fraud detection in corporate financial statements. Specifically, we have to explore machine learning and data mining techniques, as well as the diverse datasets used for identifying financial fraud. Also paper uses various machine learning algorithms such as Navie Bayes Algorithm, K-Nearest Neighbor Algorithm, Decision Tree Algorithm, Logistic Regression for fraud detection, tried to find accuracy by comparing these algorithms.

## 2. Introduction

In the current digital era, financial fraud is a chronic and changing problem that presents serious difficulties for financial institutions and their clients. The integrity and reliability of financial systems depend heavily on the detection of fraudulent activity. The goal is to improve financial fraud detection skills by utilizing data mining techniques. In an age where financial fraud poses an ever-growing threat to individuals, businesses, and economies, the imperative for robust detection systems has never been more critical. Introducing our cutting-edge Financial Fraud Detection System, a pioneering solution designed to safeguard against illicit activities in the dynamic landscape of finance.

Leveraging advanced algorithms and machine learning techniques, our system stands at the forefront of proactive Defence, tirelessly analyzing vast streams of data to identify anomalies, patterns, and aberrations indicative of fraudulent behaviour. With a commitment to precision, reliability, and adaptability, our platform offers unparalleled protection, empowering institutions to mitigate risks, preserve trust, and uphold the integrity of financial transactions. Join us in the fight against financial fraud, where vigilance meets innovation to forge a safer, more secure future. The exploration and analysis of this synthetic dataset have provided valuable insights into the intricacies of financial transactions, allowing for the development of sophisticated models aimed at detecting fraudulent activities.

## 3. Objective

i. To develop algorithms and models capable of accurately identifying fraudulent transactions while minimizing false positives is the main objective of this financial fraud detection via data mining techniques.

ii. Conduction of the data thorough exploration and Preprocess the data to ensure its suitability for data mining techniques.

iii. Identification of relevant features and engineer new ones that enhances the detection of fraudulent patterns within financial transactions.

iv. Implementation and train data mining models, such as machine learning algorithms, to analyze the synthetic data and detect potential instances of financial fraud.

v. The system should be able to analyze historical data and detect patterns or anomalies that suggest fraudulent behaviour. This could include unusual transaction amounts, frequency, or deviations from typical user behaviour.

vi. Assess the performance of the developed models using appropriate metrics. Validate the models to ensure their effectiveness in distinguishing between legitimate and fraudulent transactions.

vii. Document the entire process, including methodologies, findings, and recommendations. Provide a comprehensive report that outlines the capabilities and limitations of the proposed fraud detection system.

## 4. Scope

i. This project's scope includes using data mining techniques to examine artificial financial data.

ii. By analyzing historical transaction data and user behaviour patterns, the system can identify deviations from normal behaviour, flagging potentially fraudulent activities.

iii. Finding patterns, anomalies, and trends that might point to fraudulent activity within the fictitious financial transactions will be the main goal.

iv. Utilizing machine learning algorithms, the system learns from past instances of fraud to recognize new patterns and adapt its detection capabilities accordingly.

v. With the help of this research, financial institutions will be able to improve their fraud prevention strategies by gaining useful insights and ideas that can be implemented in practical situations.

vi. By addressing these aspects, the project aims to establish a robust foundation for financial fraud detection using synthetic data, with the potential for practical application in real-world financial systems.

# 5. Requirements

1. **Normal Requirements Specifications:**

   i. Data Collection: Collect data from various sources, including payment processors and banking systems.

   ii. Real-time Processing: To analyze transactions as they occur.

   iii. User Authentication and Access Control: To prevent unauthorized access to the fraud detection system.

   iv. Alert Generation: Generate alerts for potentially fraudulent activities, providing details on the nature of the suspicious behaviour and the affected accounts.

   v. Reporting and Analytics: Provide comprehensive reporting and analytics features for stakeholders to monitor the effectiveness of the fraud detection system.

   vi. Security: Encryption of sensitive data and secure communication channels ensures security.

   vii. User Training: Provide training for system users and administrators to use and manage the fraud detection system effectively.

2. **Expected Requirements Specifications:**

   i. Machine Learning Models: Utilize machine learning models to enhance detection capabilities, allowing the system to learn and adapt to new fraud patterns over time.

   ii. Historical Data Analysis: Analyze historical transaction data to identify patterns and trends that may indicate fraudulent behaviour.

   iii. Case Management: Facilitate the management and tracking of fraud cases from detection through investigation and resolution.

   iv. Scalability: Design the system to scale horizontally and vertically to handle increasing transaction volumes and data processing requirements.

   v. Documentation: Develop comprehensive documentation, including user manuals, system architecture documents, and guidelines for maintenance and updates.

   vi. Continuous Improvement: Implement mechanisms for continuous monitoring, feedback, and improvement of the fraud detection algorithms and processes.

3. **Excited Requirements Specifications:**

   i. Behavioural Analysis: Implement advanced behavioural analysis techniques to identify patterns and deviations from normal user behaviour.

   ii. Predictive Modeling: Explore predictive modeling to anticipate potential fraud based on historical data and emerging trends.

iii. Blockchain Integration: Investigate integration with blockchain technology for enhanced security and transparency in financial transactions.

iv. Advanced Threat Intelligence: Incorporate advanced threat intelligence feeds to stay ahead of evolving fraud tactics and techniques.

v. Biometric Authentication: Explore the use of biometric authentication to enhance user verification and reduce the risk of identity theft.

vi. Explainable AI: Implement explainable AI models to enhance transparency and understanding of the decision-making process behind fraud detection alerts.

# 6. System Architecture

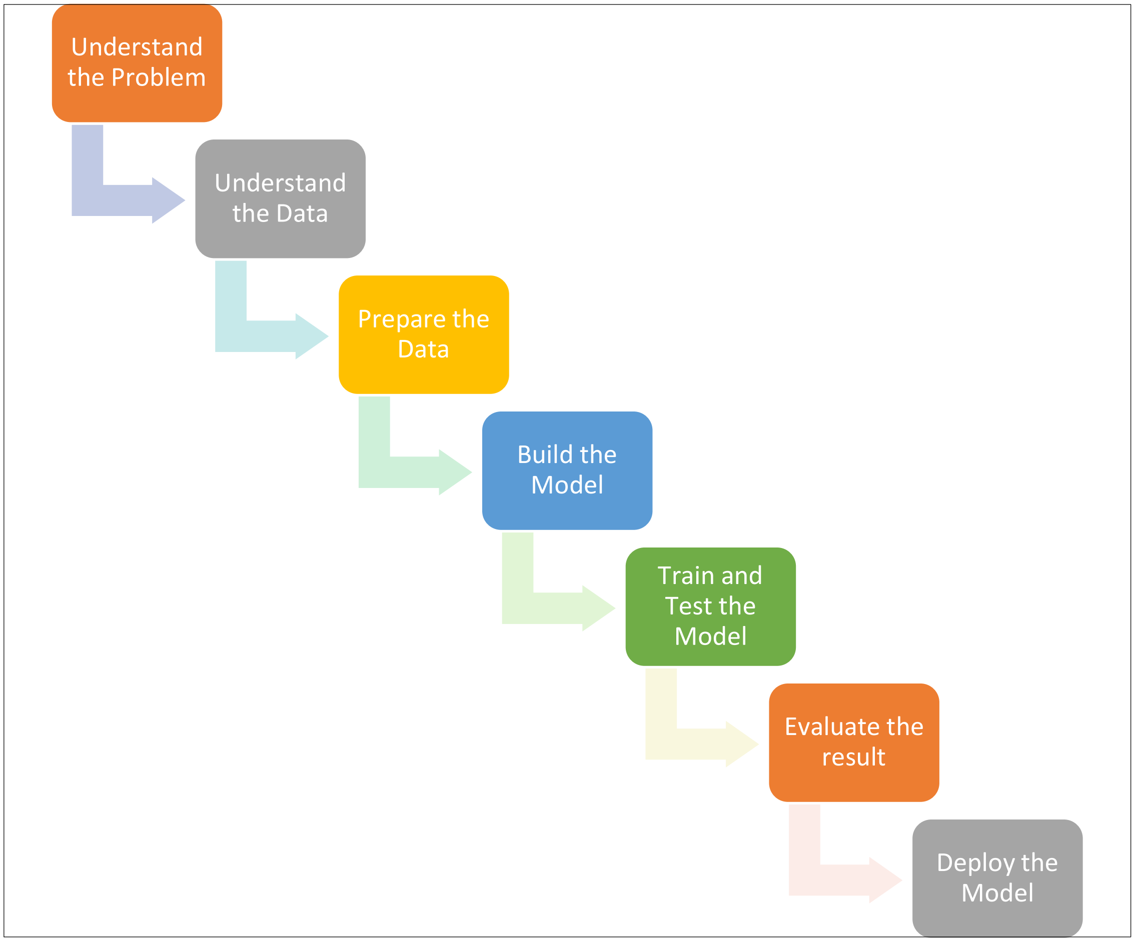


Fig 6.1: System Architecture for Financial Fraud Detection via Data Minning

**Explanation :**

Figure 6.1 shows flow of the financial fraud detection. The system architecture of a financial fraud detection system typically involves multiple components working together to analyze, monitor, and detect fraudulent activities.
Here's an overview of the key components and their interactions:

1. Understand the Problem Statement (Analyz the topic): Before touching, extracting, cleaning or analysing any data we need to understand the problem statement properly. Understanding financial fraud detection starts with recognizing the different ways people cheat in finance, like stealing identities or laundering money. Fraud hurts individuals and

the economy by causing financial losses and destroying trust. Catching fraud is tough because financial data is complex, rules are strict, and crooks keep changing tactics.

2. Understand the Data: To understand the data present in financial fraud detection dataset:
    - i. Data Sources: Identify where the financial data comes from, like transaction records and customer info.
    - ii. Data Quality: Check if the data is accurate and complete. Bad data can lead to wrong fraud alerts.
    - iii. Data Preprocessing: Clean up the data by removing duplicates and fixing errors.
    - iv. Feature Engineering: Choose or create the best data bits to help spot fraud, like transaction patterns.
    - v. Data Imbalance: Remember that fraud is rare compared to normal transactions. Adjust models to handle this.
    - vi. Data Privacy: Keep sensitive data safe and follow privacy laws.
    - vii. Scalability: Make sure systems can handle lots of data without slowing down.
    - viii. Real-Time Processing: Detect fraud quickly by analysing data as it comes in.
3. Prepare the Data: In data preparation for financial fraud detection:
    - i. Data Collection: Gather information from various sources like transactions and customer data.
    - ii. Data Cleaning: Remove any errors or inconsistencies in the data, like missing values or duplicates.
    - iii. Data Formatting: Standardize the format of the data to make it consistent and easy to analyse.
    - iv. Feature Selection: Choose the most relevant information that could help detect fraud, such as transaction amounts or frequency.
    - v. Data Transformation: Convert the data into a suitable format for analysis, like numerical values or categories.
    - vi. Data Sampling: If needed, adjust the size or balance of the data to ensure fair representation and accurate analysis.
    - vii. Data Splitting: Divide the data into training and testing sets to train the fraud detection model and evaluate its performance.
4. Build the Model: In building a model for financial fraud detection:
    - i. Choose a Model: Decide on a method or algorithm to use, like logistic regression or decision trees.

ii. Train the Model: Teach the model what fraud looks like by showing it examples of both fraudulent and legitimate transactions.

iii. Test the Model: Check how well the model can spot fraud by giving it new data it hasn't seen before.

iv. Adjust the Model: Fine-tune the model to make it better at catching fraud without flagging too many innocent transactions.

v. Evaluate Performance: Measure how well the model performs using metrics like accuracy, precision, and recall.

5. Evaluate the Result: In evaluating financial fraud detection results:

i. Check Accuracy: Verify correct identification of fraud cases and legitimate transactions.

ii. Assess False Alarms: Determine instances of incorrect flagging of transactions as fraudulent.

iii. Calculate Precision: Measure the accuracy of flagged fraud cases.

iv. Determine Recall: Evaluate the system's ability to identify actual fraud cases.

6. Deploy the Model:

i. Activate the model to analyse transactions in real-time.

ii. Integrate it into existing systems for transaction monitoring.

iii. Monitor performance and adjust as needed.

iv. Ensure scalability for handling large transaction volumes.

# 7. Literature Review

| S.n. | Title | Author | Publication | Advantages | Disadvantages | Limitations |
|---|---|---|---|---|---|---|
| 1. | Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review | Matin N. Ashtiani And Bijan Raahemi | IEEE | ML algorithms are efficient in detecting fraudulent financial activities, offering accurate predictions and classifications.<br>They save time and resources compared to manual methods, allowing continuous monitoring and analysis.<br>These systems can handle large data volumes and adapt to new fraud patterns, continuously improving their effectiveness over time. | Requires expertise and resources for setup and maintenance.<br>Fraud detection accuracy depends on data quality and relevance.<br>Costly development and maintenance, especially for smaller organizations<br>Risk of false positives leading to potential disruptions. | Search process may overlook relevant studies and keywords.<br>Unstructured data analysis may be challenging.<br>Inappropriate feature selection can weaken ML/DM algorithms.<br>Limited use of sampling techniques can impact model performance. |
| 2. | Credit Card Fraud Detection Using AdaBoost and Majority Voting | Kuldeep Randhawa, Chu Kiong Loo,Manjeevan Seera, Chee Peng Lim, And Asoke K. Nandi | IEEE Access | Insight into machine learning application.<br>Real-world data sets used for evaluation.<br>Hybrid models potential for improved fraud detection accuracy. | Lack of detailed explanation on algorithmic implementations.<br>In-depth analysis of proposed models' computational efficiency is not done. | Lack of detailed information on specific algorithms.<br>Lack of discussion on implementation challenges<br>Absence of scalability for handling large credit card transaction volumes. |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| 3. | Financial fraud detection based on the part-of-speech features of textual risk disclosures in financial reports | Science Direct | Hao Suna,Jianping Lic , Xiaoqian Zhuc | Faster and easier fraud detection.<br>Quicker text scanning than word-by-word reading.<br>Early detection can prevent negative outcomes. | Difficulty in understanding computer's fraud perception due to keyword confusion.<br>Potential for misinterpretation of fraud when it's not.<br>Need for adapting methods to language changes. | Importance of program accuracy in text comprehension.<br>Potential for missing important details due to individual word focus.<br>Method may not suit every company or industrial data. |
| 4. | Internet Financial Fraud Detection Based on a Distributed Big Data Approach With Node2vec | IEEE Access | Hangjun Zhou , Guang Sun, Sha Fu , Linli Wang , Juan Hu , And Ying Gao | Using a distributed approach makes it easier to manage huge amounts of financial data quickly.<br>Node2vec helps pull out important features from complex financial networks, making it easier to spot fraudulent activities.<br>By using advanced techniques like Node2vec, the system can catch more frauds accurately, even those with subtle patterns. | Requires extensive expertise and resources.<br>Privacy and security concerns arise with sensitive financial data.<br>Large-scale system analyses consume significant computing power and storage space. | Reliance on complete, accurate data for effectiveness.<br>Potential for false alarms or missing frauds due to incorrect or missing data.<br>Not suitable for all financial data types due to potential fraud patterns.<br>Potential issues with scaling up for large data amounts or sudden transaction increases. |
| 5. | Online Payment Fraud Detection Model Using Machine Learning Techniques | IEEE Access | Abdul Wahab Ali Almaz Roi , Nasir Ayub | Shows exceptional capabilities in handling modern financial fraud.<br>Outperforms existing solutions, improving detection accuracy. | Requires robust infrastructure and efficient algorithms for large volumes.<br>Requires sophisticated algorithms for meaningful insights. | Data imbalance can lead to overfitting to minority class, affecting reliability of conclusions. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | Addresses inherent inefficiencies of traditional approaches.<br>Conducted performance evaluation comparing model with conventional and deep learning techniques. | Challenges include low latency and high computational resources.<br>Addresses security and privacy concerns.<br>Scaling to handle increasing data volumes and user loads.<br>Complex integration with existing IT infrastructure and data sources. | Linear nature of PCA may result in unintended elimination of important characteristics, impacting model performance.<br>Limited generalizability due to study's reliance on IEEE CIS dataset. Strict procedures to prevent overfitting pose risk, especially in real-world applications. |
| 6. | ▪ CoDetect: Financial Fraud Detection With Anomaly Feature Detection | IEEE Access | Dongxu Huang , Dejun Mu, Libin Yang , And Xiaoyan Ca | Utilizes graph-based similarity and feature matrix.<br>Identifies nature of financial activities from fraud patterns to suspicious properties.<br>Provides interpretable fraud identification on sparse matrix<br>Experimental results show effective fraud detection.<br>Allows executives to trace original fraud with suspicious features. | Potential for false positives leading to unnecessary investigations.<br>Lack of interpretability in flagging fraudulent transactions.<br>Scalability issues as transaction volumes increase.<br>Imbalanced data handling reduces fraud sensitivity.<br>Advanced attacks by fraudsters altering behavior to appear normal. | Effectiveness depends on quality and representativeness of training data.<br>CoDetect may struggle with evolving fraud techniques.<br>Regulatory Compliance: Ensuring compliance with data privacy and financial regulations is challenging.<br>Implementation and maintenance may be prohibitive for smaller organizations.<br>Despite automation, |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  |  |  |  |  |  | human oversight is necessary. |
| 7. | ▪ Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances | Science Direct | Waleed Hilal, S. Andrew Gadsden, John Yawney | Highlights importance of developing fraud detection strategies.<br>Studies machine learning, deep learning, and statistical models.<br>Focuses on graph-based techniques for in-depth analysis. | Potential overlooking of other effective methods in financial fraud detection.<br>Lack of comparative analysis with non-graph-based approaches.<br>Complexity and interpretability of implementing and interpreting graph-based anomaly detection techniques. | Imbalanced Datasets: Rare fraudulent cases can lead to biased models and reduced detection accuracy.<br>Lack of Publicly Available Datasets: Scarcity of publicly available datasets hinders the development and benchmarking of effective detection systems.<br>Generalizability and Scalability |
| 8. | Financial Cybercrime: A Comprehensive Survey of Deep Learning Approaches to Tackle the Evolving Financial Crime Landscape | IEEE | Jack Nichols, Aditya Kuppa , And Nhien-An Le-khac | Hybrid Approach for Fraud Detection<br>Ensemble learning for social attacks.<br>Bipartite graph model for point of compromise detection. | Limited detection and prevention of fraudulent activities.<br>While some techniques compete with advanced algorithms, hybrid approaches may not achieve optimal performance.<br>Lack of Holistic Understanding. | Evolving battle against sophisticated financial cybercriminals.<br>Challenges in latency-sensitive applications and real-time data processing.<br>Need for research on social engineering attacks. |

# 8. Flowchart

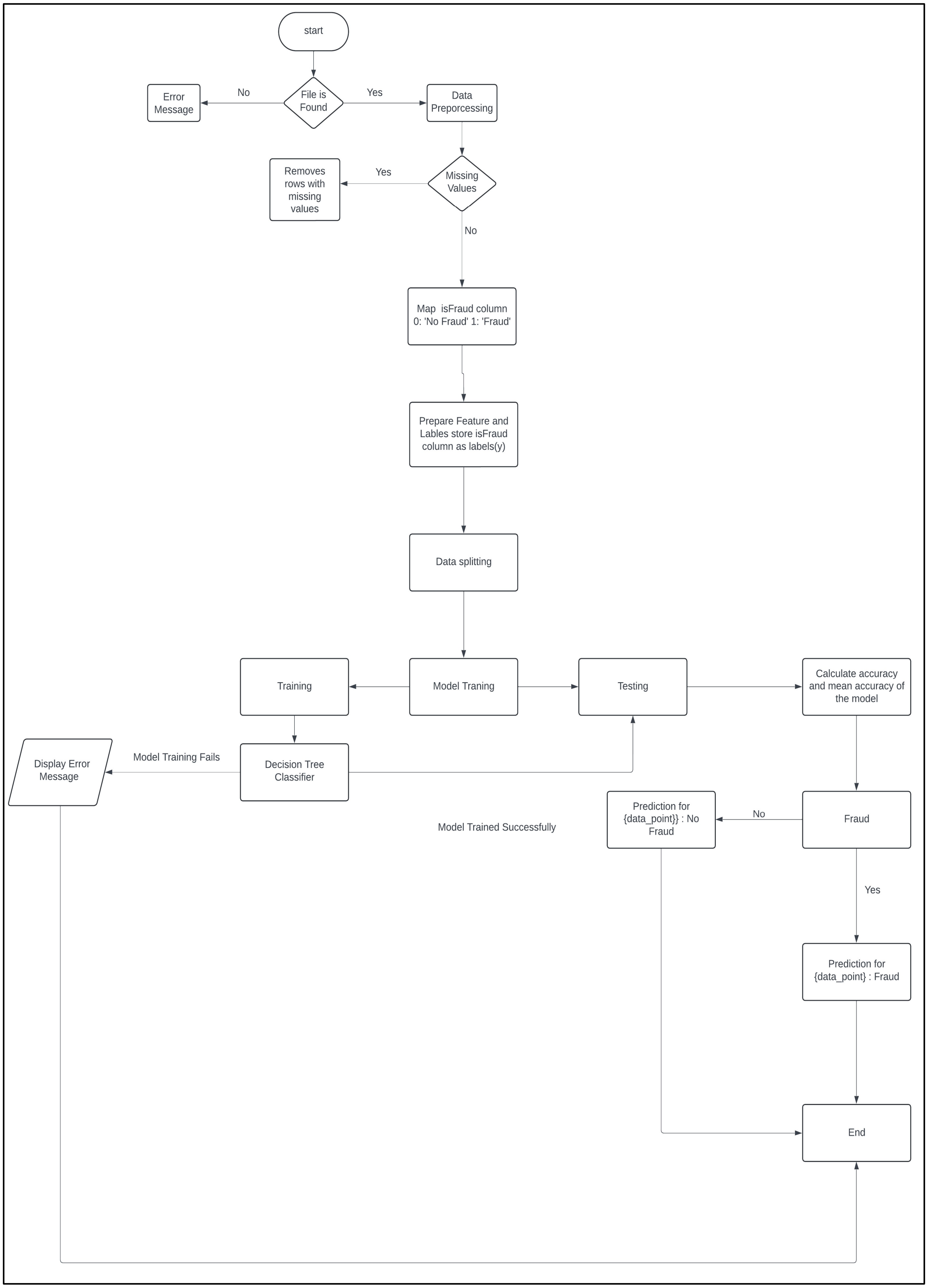
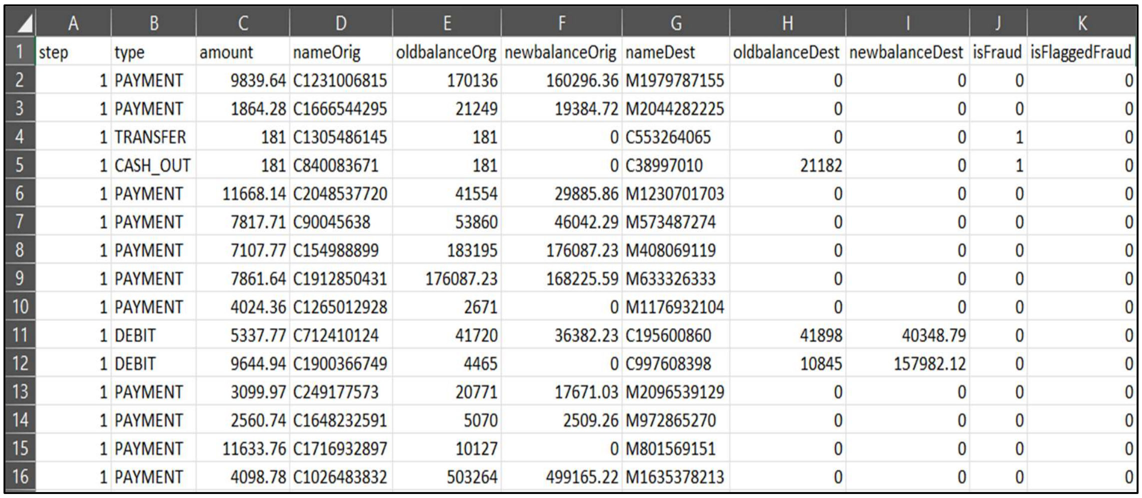

start
File is Found
No
Error Message
Yes
Data Preporcessing
Missing Values
Yes
Removes rows with missing values
No
Map isFraud column 0: 'No Fraud' 1: 'Fraud'
Prepare Feature and Lables store isFraud column as labels(y)
Data splitting
Training
Model Traning
Testing
Calculate accuracy and mean accuracy of the model
Display Error Message
Model Training Fails
Decision Tree Classifier
Model Trained Successfully
Prediction for {data_point}} : No Fraud
No
Fraud
Yes
Prediction for {data_point} : Fraud
End

*Fig. 8.1: Flowchart for financial fraud detection*

# 9. Dataset

- Dataset Name: Financial Fraud Detection



| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | step | type | amount | nameOrig | oldbalanceOrg | newbalanceOrig | nameDest | oldbalanceDest | newbalanceDest | isFraud | isFlaggedFraud |
| 2 | 1 | PAYMENT | 9839.64 | C1231006815 | 170136 | 160296.36 | M1979787155 | 0 | 0 | 0 | 0 |
| 3 | 1 | PAYMENT | 1864.28 | C1666544295 | 21249 | 19384.72 | M2044282225 | 0 | 0 | 0 | 0 |
| 4 | 1 | TRANSFER | 181 | C1305486145 | 181 | 0 | C553264065 | 0 | 0 | 1 | 0 |
| 5 | 1 | CASH_OUT | 181 | C840083671 | 181 | 0 | C38997010 | 21182 | 0 | 1 | 0 |
| 6 | 1 | PAYMENT | 11668.14 | C2048537720 | 41554 | 29885.86 | M1230701703 | 0 | 0 | 0 | 0 |
| 7 | 1 | PAYMENT | 7817.71 | C90045638 | 53860 | 46042.29 | M573487274 | 0 | 0 | 0 | 0 |
| 8 | 1 | PAYMENT | 7107.77 | C154988899 | 183195 | 176087.23 | M408069119 | 0 | 0 | 0 | 0 |
| 9 | 1 | PAYMENT | 7861.64 | C1912850431 | 176087.23 | 168225.59 | M633326333 | 0 | 0 | 0 | 0 |
| 10 | 1 | PAYMENT | 4024.36 | C1265012928 | 2671 | 0 | M1176932104 | 0 | 0 | 0 | 0 |
| 11 | 1 | DEBIT | 5337.77 | C712410124 | 41720 | 36382.23 | C195600860 | 41898 | 40348.79 | 0 | 0 |
| 12 | 1 | DEBIT | 9644.94 | C1900366749 | 4465 | 0 | C997608398 | 10845 | 157982.12 | 0 | 0 |
| 13 | 1 | PAYMENT | 3099.97 | C249177573 | 20771 | 17671.03 | M2096539129 | 0 | 0 | 0 | 0 |
| 14 | 1 | PAYMENT | 2560.74 | C1648232591 | 5070 | 2509.26 | M972865270 | 0 | 0 | 0 | 0 |
| 15 | 1 | PAYMENT | 11633.76 | C1716932897 | 10127 | 0 | M801569151 | 0 | 0 | 0 | 0 |
| 16 | 1 | PAYMENT | 4098.78 | C1026483832 | 503264 | 499165.22 | M1635378213 | 0 | 0 | 0 | 0 |

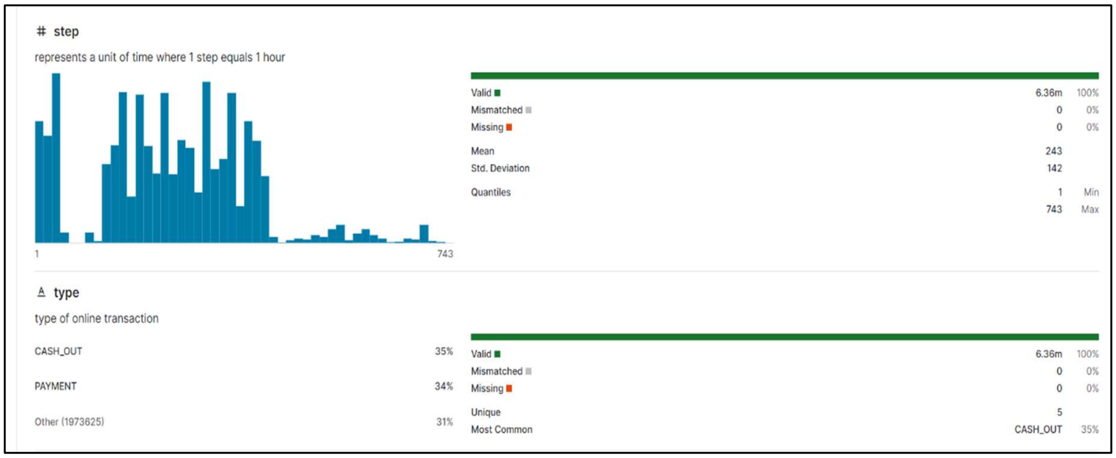
*Fig. 9.1: Description of the Dataset with attributes and Tuples*

- Description of the attributes:

| Sr.No. | Attributes | Description |
|---|---|---|
| 1. | step | Represents a unit of time where 1 step equals 1 hour. |
| 2. | type | Type of online transaction |
| 3. | amount | The amount of the transaction |
| 4. | nameOrig | Customer starting the transaction |
| 5. | oldbalanceOrg | Balance before the transaction |
| 6. | newbalanceOrig | Balance after the transaction |
| 7. | nameDest | recipient of the transaction |
| 8. | oldbalanceDest | Initial balance of recipient before the transaction |
| 9. | newbalanceDest | The new balance of recipient after the transaction |
| 10. | isFraud | Fraud transaction |

- Descriptive analysis:

  Below is a detailed description of the  above-mentioned attributes .



# step

represents a unit of time where 1 step equals 1 hour

Valid 🟩 6.36m 100%
Mismatched ▨ 0 0%
Missing 🟥 0 0%

Mean 243
Std. Deviation 142

Quantiles 1 Min
743 Max

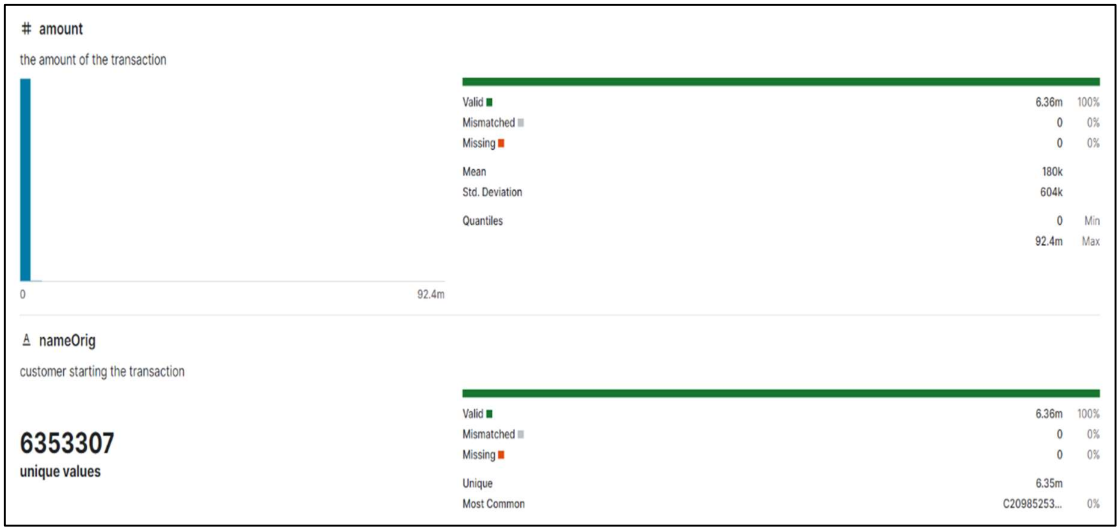
A type

type of online transaction

CASH_OUT 35%
PAYMENT 34%
Other (1973625) 31%

Valid 🟩 6.36m 100%
Mismatched ▨ 0 0%
Missing 🟥 0 0%

Unique 5
Most Common CASH_OUT 35%



# amount

the amount of the transaction

Valid 🟩 6.36m 100%
Mismatched ▨ 0 0%
Missing 🟥 0 0%

Mean 180k
Std. Deviation 604k

Quantiles 0 Min
92.4m Max

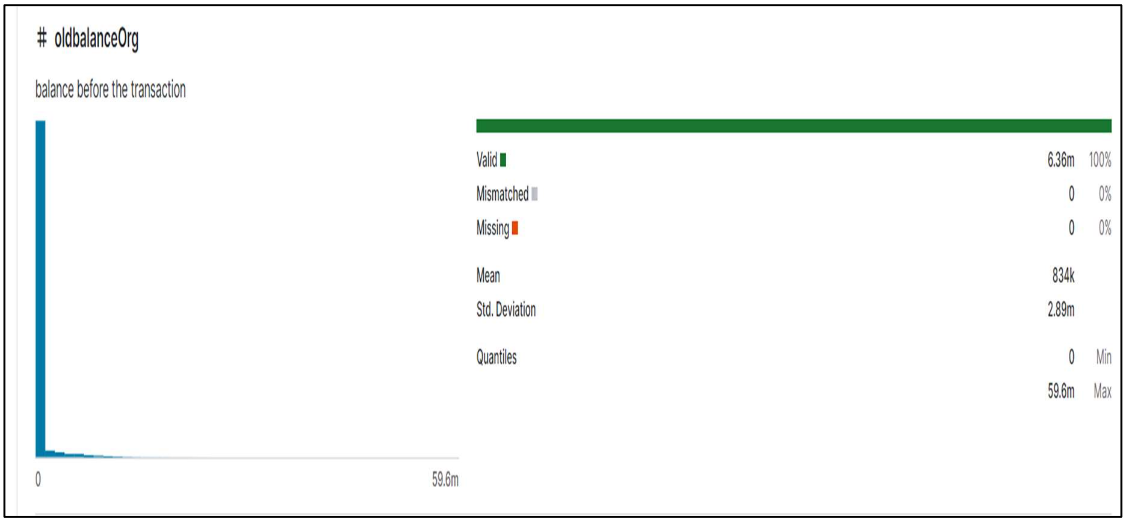
A nameOrig

customer starting the transaction

**6353307**
**unique values**

Valid 🟩 6.36m 100%
Mismatched ▨ 0 0%
Missing 🟥 0 0%

Unique 6.35m
Most Common C20985253... 0%



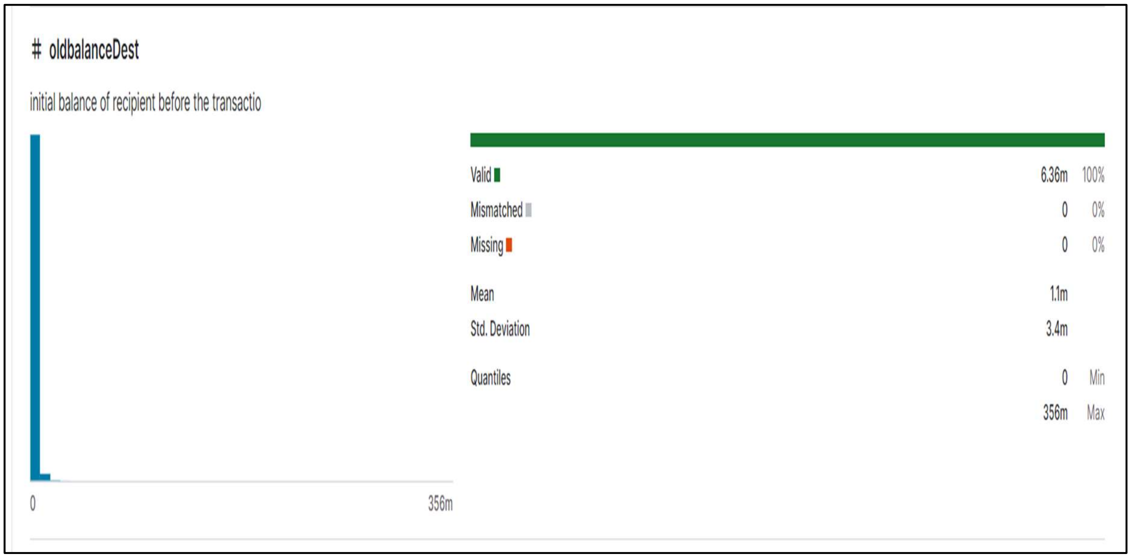
# oldbalanceOrg

balance before the transaction

Valid 🟩 6.36m 100%
Mismatched ▨ 0 0%
Missing 🟥 0 0%

Mean 834k
Std. Deviation 2.89m

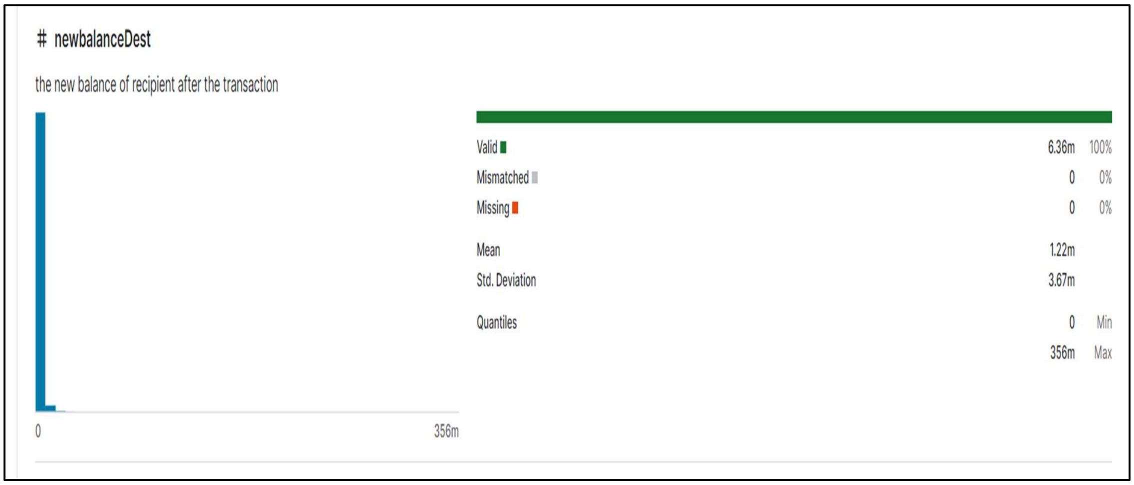
Quantiles 0 Min
59.6m Max



# oldbalanceDest

initial balance of recipient before the transactio

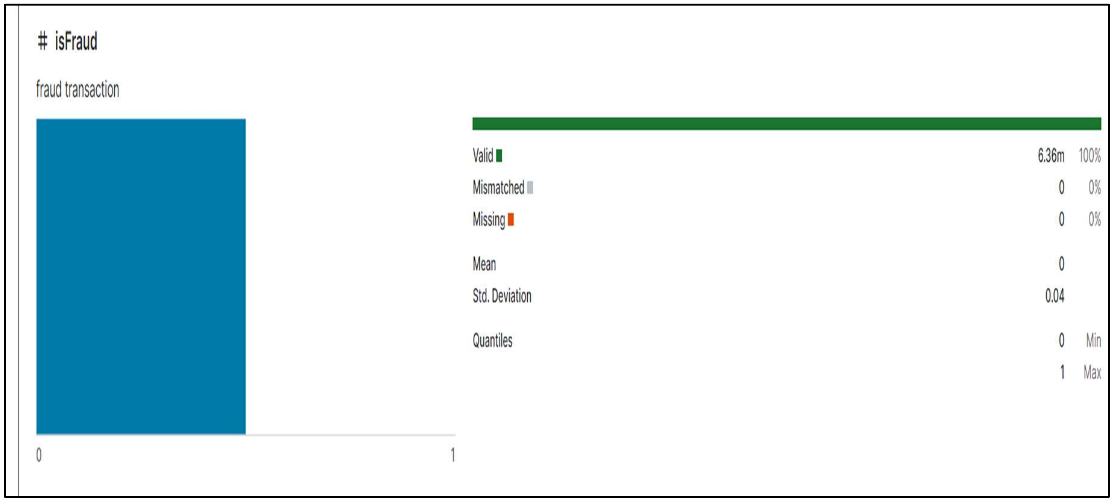
| | | |
|---|---|---|
| Valid 🟩 | 6.36m | 100% |
| Mismatched ▨ | 0 | 0% |
| Missing 🟥 | 0 | 0% |
| Mean | 1.1m | |
| Std. Deviation | 3.4m | |
| Quantiles | 0 | Min |
| | 356m | Max |



# newbalanceDest

the new balance of recipient after the transaction

| | | |
|---|---|---|
| Valid 🟩 | 6.36m | 100% |
| Mismatched ▨ | 0 | 0% |
| Missing 🟥 | 0 | 0% |
| Mean | 1.22m | |
| Std. Deviation | 3.67m | |
| Quantiles | 0 | Min |
| | 356m | Max |



# isFraud

fraud transaction

| | | |
|---|---|---|
| Valid 🟩 | 6.36m | 100% |
| Mismatched ▨ | 0 | 0% |
| Missing 🟥 | 0 | 0% |
| Mean | 0 | |
| Std. Deviation | 0.04 | |
| Quantiles | 0 | Min |
| | 1 | Max |

# 10. Implementation Details and Algorithms used

## A. Decision Tree Algorithm:

1. Decision tree algorithm falls under the category of super-vised algorithm. It can also be used for solving regression and classification problems. It creates a tree like structure, where each node represents a decision, the branch represents the result of the decision and the leaf node specifies the class label.
2. Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.
3. The goal of using a Decision Tree is to create a training model that can use to predict the class and value of the target variable by learning simple decision rules inferred from prior data (training data).
4. In Decision Trees, for predicting a class label for a record we start from the tree.
5. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Important Terminology related to Decision Trees:

- Root Node: It represents the entire population or sample and this further gets divided into
- two or more homogeneous sets.
- Splitting: It is a process of dividing a node into two or more sub
- Decision Node: When a sub-node.
- Leaf / Terminal Node: Nodes do not split is called Leaf or Terminal node.
- Pruning: When we remove sub
- can say the opposite process of splitting.
- Branch / Sub-Tree: A subsection of the entire tree is called branch or sub
- Parent and Child Node: A node, which is divided into sub
- sub-nodes whereas sub-nodes are the child of a parent node.

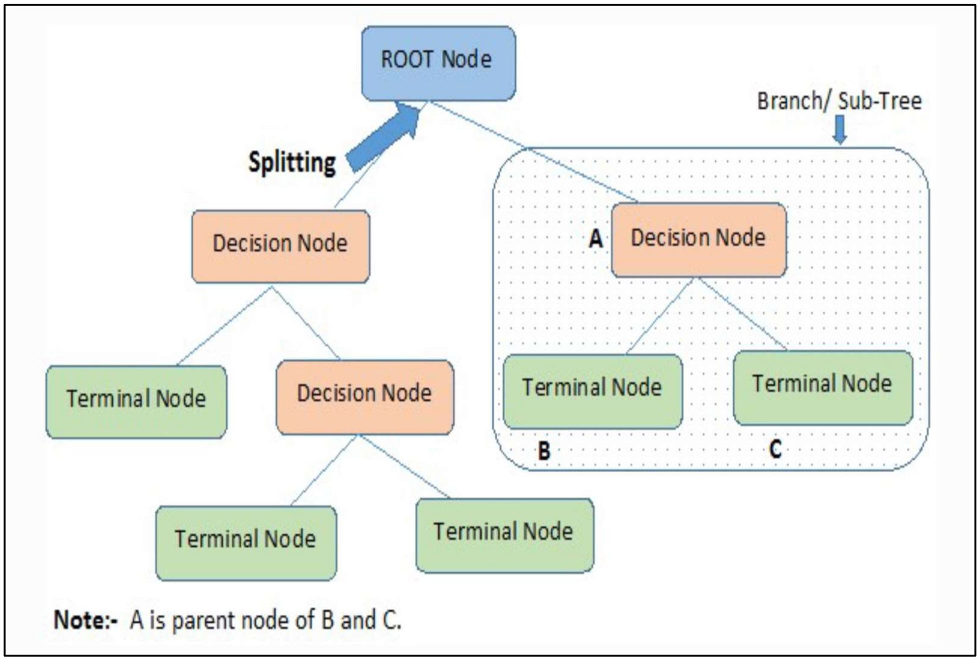


*Fig 10.1::Diagram of Decision Tree*

In this algorithm, we split the whole dataset starting from the root node till the leaf node. The splitting is done based on the comparison between the root attribute to the record's attribute. It is easy to visualize and interpret. The splitting is carried out recursively until maximum depth of the tree is reached. It can handle high dimensional data and numerical or categorical data. We calculate Entropy which is the measure of randomness using 1 .

Once we calculate the entropy then we calculate the information gain using 2

$$E(S) = \sum_{i=1}^{n} p_i \, log_2(p_i)$$

**B. KNN Algorithm:**

1. K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
2. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
3. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
4. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

5. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
6. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
7. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

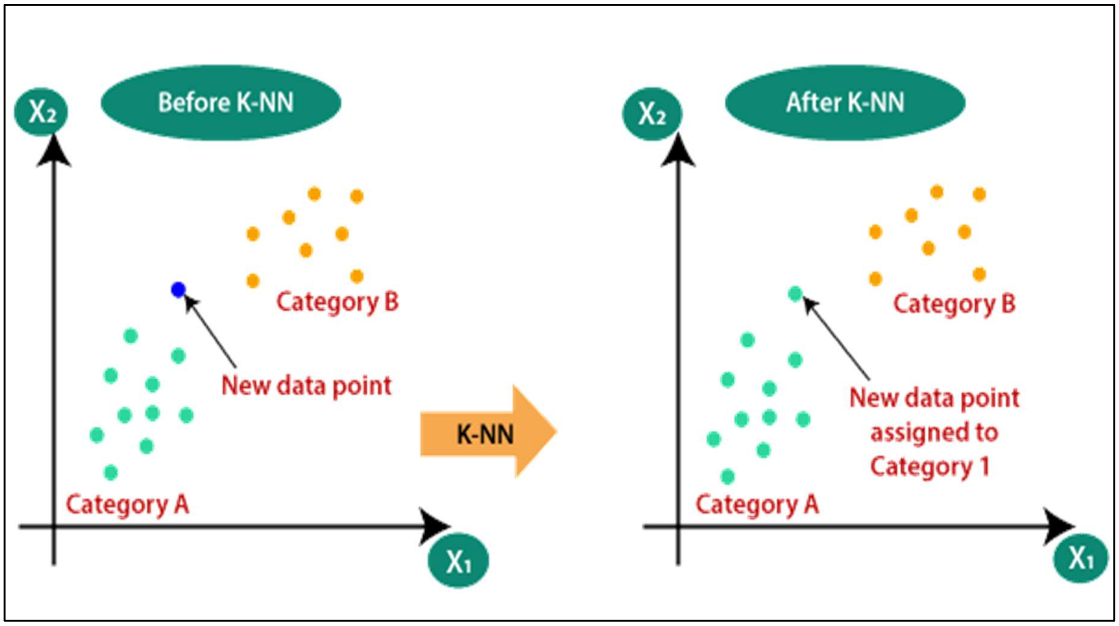


*Fig. 10.2 : Data Points Before and After KNN*

KNN(K- Nearest Neighbor) ,a type of supervised algorithm, is one of the most important type of classification algorithm. KNN doesn’t accept any parameters like other algorithms rather we give two sets of data points one is for training and the other is for testing. KNN is simple , versatile , easy to implement and widely used. It handles both numerical and categorical data. The most important task in KNN is to choose the value of k based on the given inputs. After that we calculate the Euclidean distance using 3 or the Manhattan distance using 4. After that we find the nearest neighbors by arranging the distance in ascending order and choosing the k smallest suitable data points. At last we apply the label to the most frequent among the nearest K neighbors.

$$\text{Euclidean Distance} = \sqrt{}(x2 - x1)2 + (y2 - y1)2 \; (3)$$

$$\text{Manhattan distance, } d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

### C. Logistic Regression

Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors.

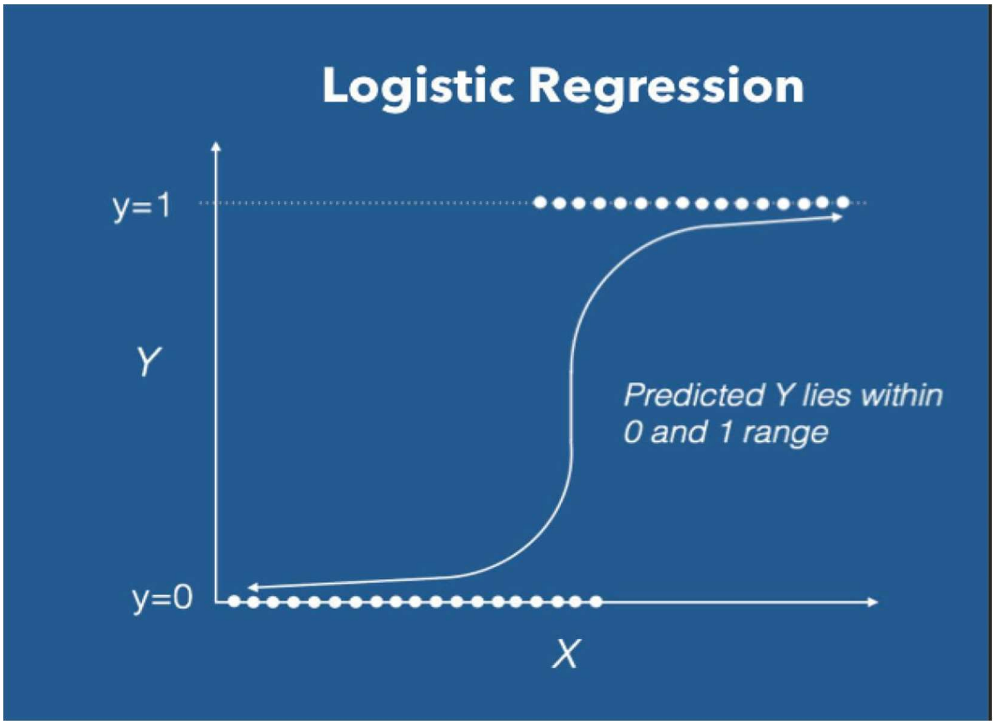


*Fig. 10.3: Graph of Logistic Regression 1*

Logistic regression falls under the category of supervised algorithm. It is used for binary classification. We give in-dependent variables as a input and it gives output in form of probability value as 0 or 1.It handles categorical and discrete value. 0 depicts ‘No’ and 1 depicts ‘Yes’, rather than giving the values as 0 or 1 it gives a probabilistic value between 0 and 1. If the outcome is 0.5 or above then it is considered as 1 else it is considered as 0. Logistic regression is easily implementable and it gives a valuable insights. Logistic regression is of three types namely binomial, Multinomial and ordinal We use sigmoid function in logistic regression, which is a mathematical function that maps predicted values to probability. Sigmoid function is defined as given in 5 .Thus the equation of logistic regression is given by

$$f(x) = \frac{1}{1+e^{-x}}$$

$$y = \frac{e^{(b_0+b_1X)}}{1+e^{(b_0+b_1X)}}$$

**D. Naive Bayes:**

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Naïve Bayes algorithm is used for classification problems. It is highly used in text classification. In text classification tasks, data contains high dimension (as each word represent one feature in the data). It is used in spam filtering, sentiment detection, rating classification etc. The advantage of using naive Bayes is its speed. It is fast and making prediction is easy with high dimension of data.

Naive Bayes is a classification algorithm that uses Bayes' theorem to determine class labels from input features. The model's assumption is that features are independent of each other when considering the class label, making calculations straightforward. It computes the probability of each class and selects the one with the highest probability as the prediction. Variants of Naive Bayes include Gaussian (for continuous data), Multinomial (for discrete data such as word counts), and Bernoulli (for binary data). The algorithm is efficient and well-suited for tasks like text classification and fraud detection.

Gaussian Naive Bayes classifier

In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution When plotted(see Fig. 10.4) , it gives a bell-shaped curve which is symmetric about the mean of the feature values as shown below:

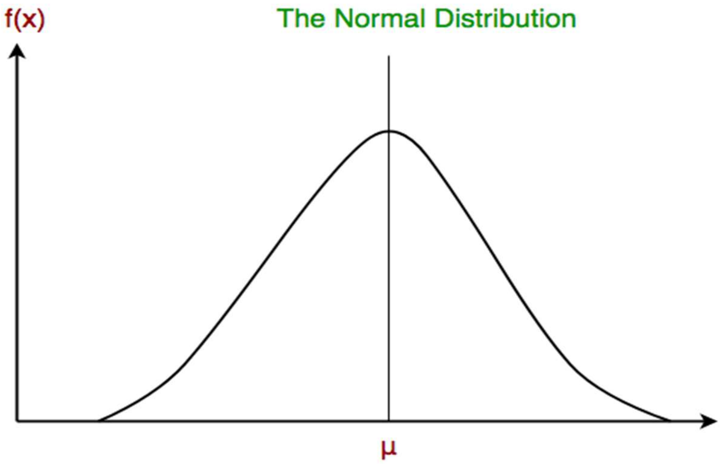


*Fig. 10.4: Normal Distribution Graph*

Updated table of prior probabilities for outlook feature is as following:

The likelihood of the features is assumed to be Gaussian, hence, conditional probability is given by:

$$P(xi|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

# 11. Result

1. **Comparison with respect to the accuracy of different classification models:**

   Summization of all above algorithms are shown using bar grapgh in Figure 5, which indicates that decision tree algorithms have maximum accuracy as compared to other classification algorithms

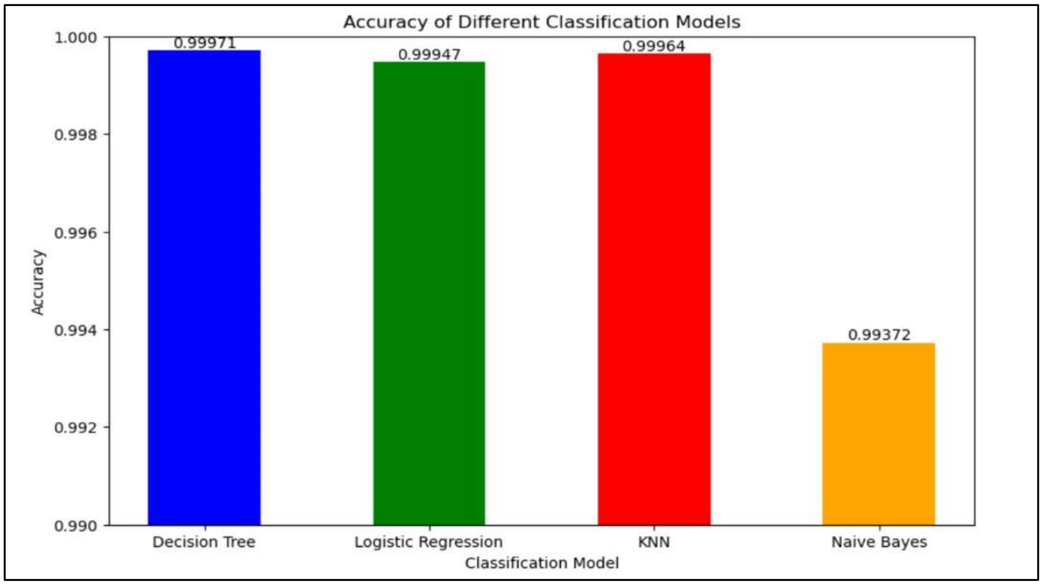


*Fig. 10.5: Graph showing accuracy of classification models*

2. Below graph (Fig.) is showing feature importance of attributes. From graph, oldbalanceOrg have maximum importance than all other attributes.

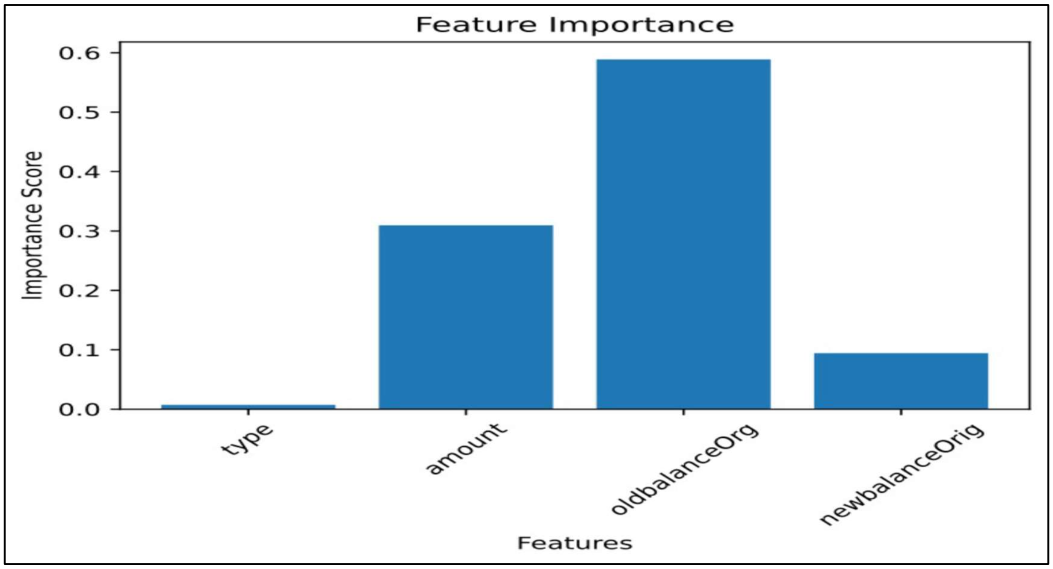


*Fig. 10.6: Feature Importance*

3. **Confusion Matrix:**

The entries in the confusion matrix are defined as the following:

- ✓ False Positive: The total number of incorrect predictions classified as positive.
- ✓ False Negavtive: The total number of incorrect predictions classified as negative.
- ✓ True Positive: The total number of true predictions classified as positive.
- ✓ True Negative: The total number of true predictions classified as negative.

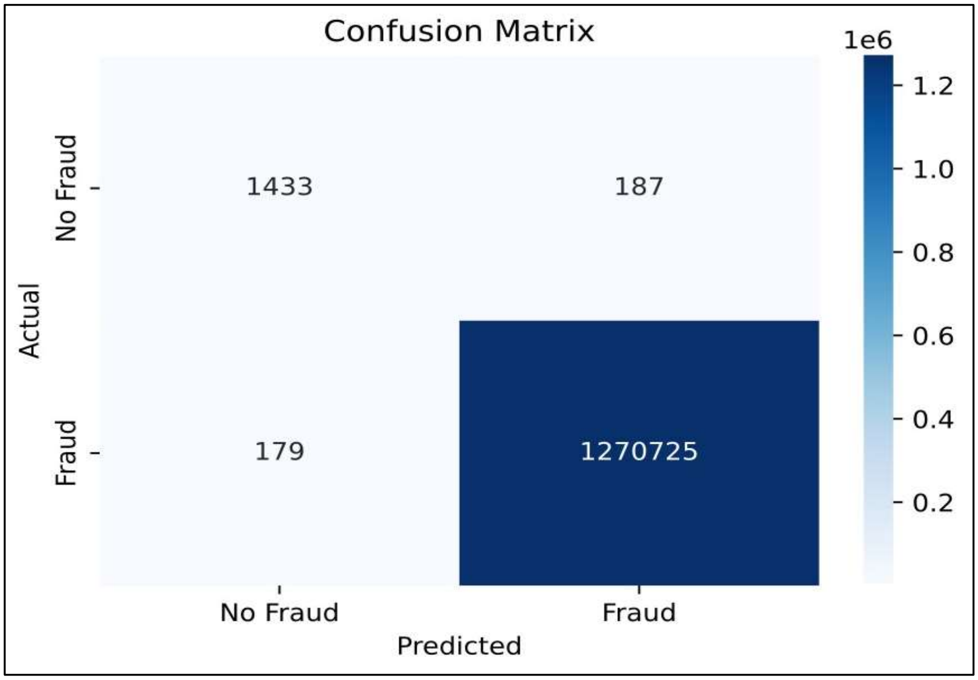


Figure: Confusion Matrix

- Accuracy: A measurement metric, measures the ratio of the total number of correct predictions of fraud to the total number of predictions (both fraud and not fraud). It is calculated as

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

- Precision: Measures the ratio of correctly classified fraud transactions (TP) to the total transactions predicted to be fraud transactions (TP + FP). It is calculated as

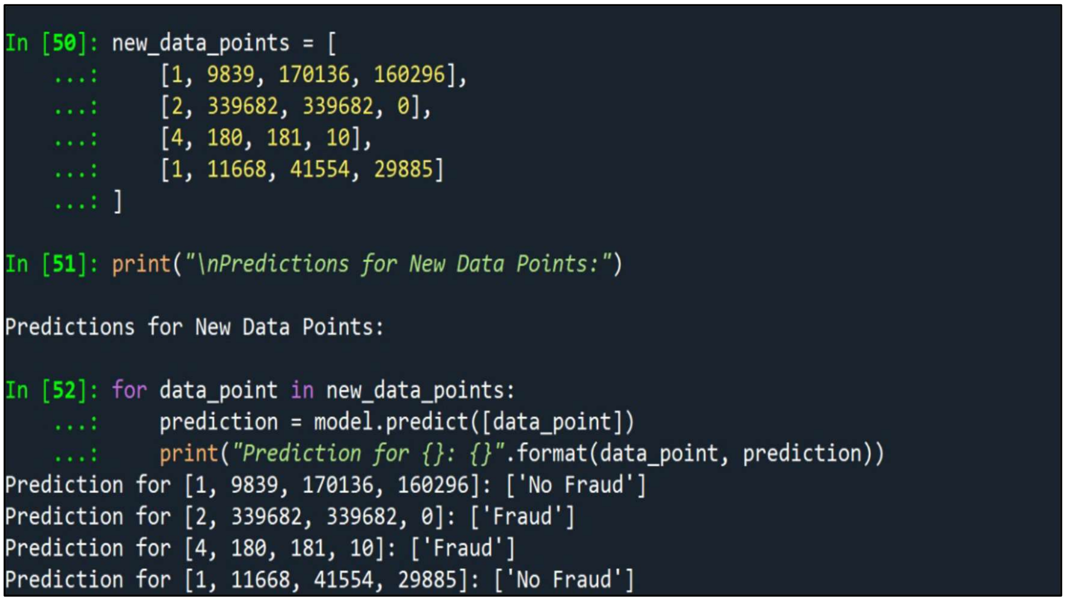
$$Precision = \frac{TP}{FP + TP}$$

- Recall: Measures the ratio of correctly classified fraud transactions (TP) to the total number of fraud transactions [45].

  It is calculated as

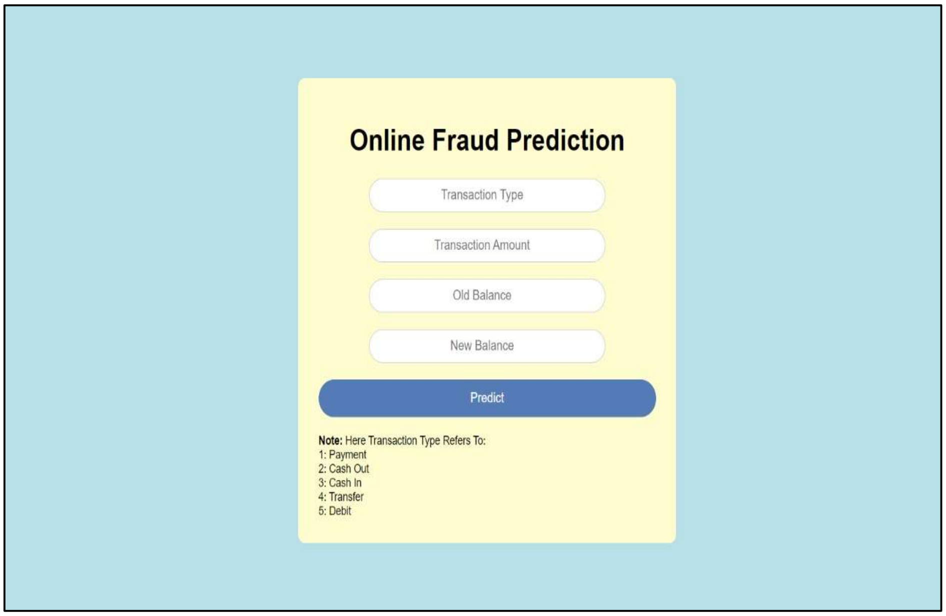
$$Recall = \frac{TP}{TP + FN}$$

### 4. Prediction of Fraud for given data points:



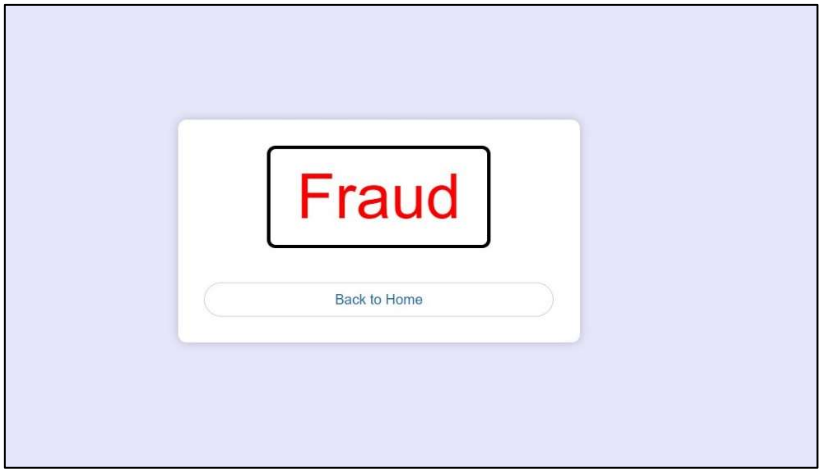
```
In [50]: new_data_points = [
    ...:     [1, 9839, 170136, 160296],
    ...:     [2, 339682, 339682, 0],
    ...:     [4, 180, 181, 10],
    ...:     [1, 11668, 41554, 29885]
    ...: ]

In [51]: print("\nPredictions for New Data Points:")

Predictions for New Data Points:

In [52]: for data_point in new_data_points:
    ...:     prediction = model.predict([data_point])
    ...:     print("Prediction for {}: {}".format(data_point, prediction))
Prediction for [1, 9839, 170136, 160296]: ['No Fraud']
Prediction for [2, 339682, 339682, 0]: ['Fraud']
Prediction for [4, 180, 181, 10]: ['Fraud']
Prediction for [1, 11668, 41554, 29885]: ['No Fraud']
```

### 5. GUI of the Financial Fraud Detection System



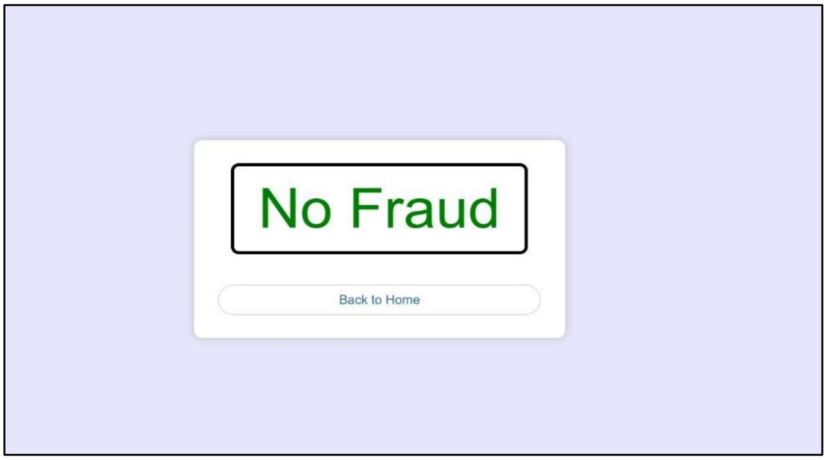
**Note:** Here Transaction Type Refers To:
1: Payment
2: Cash Out
3: Cash In
4: Transfer
5: Debit

- Fraud is Detected:



Fraud

- No Fraud:



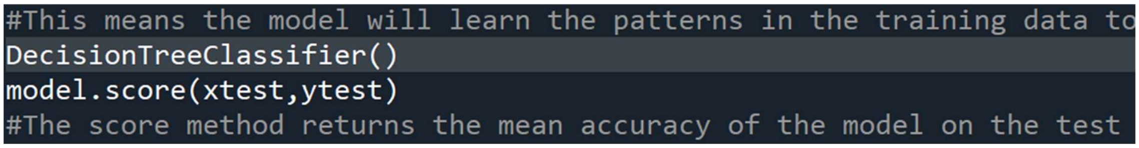
No Fraud

# 12. Test Cases

## 1. Model Accuracy on Test Data

**Theory:** Accuracy measures the proportion of correct predictions over the total number of predictions. It is a good general measure of model performance.

**Code:**

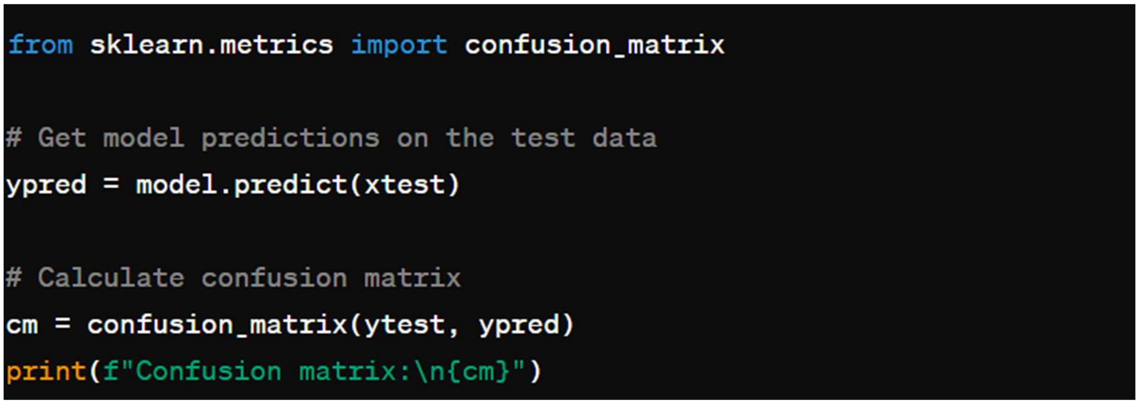


```
#This means the model will learn the patterns in the training data to
DecisionTreeClassifier()
model.score(xtest,ytest)
#The score method returns the mean accuracy of the model on the test
```

## 2. Confusion Matrix

**Theory:** A confusion matrix provides a detailed view of the model's performance. It shows the number of true positives, false positives, true negatives, and false negatives.

**Code:**

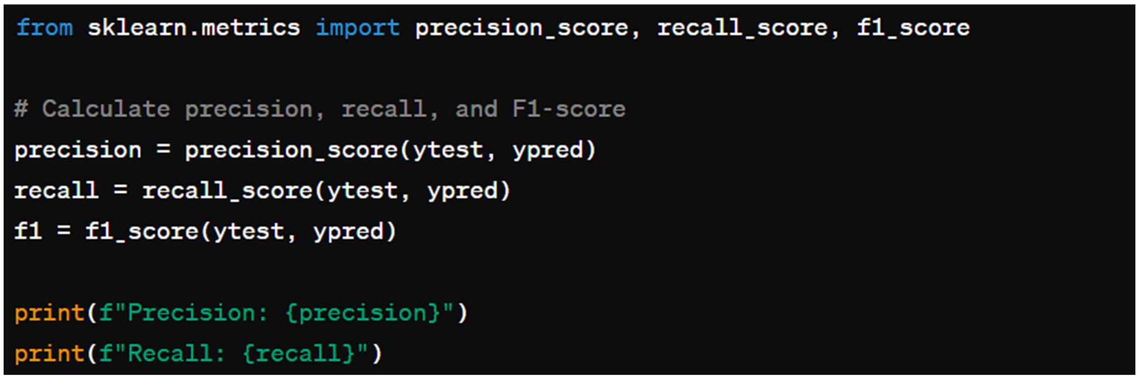


```
from sklearn.metrics import confusion_matrix

# Get model predictions on the test data
ypred = model.predict(xtest)

# Calculate confusion matrix
cm = confusion_matrix(ytest, ypred)
print(f"Confusion matrix:\n{cm}")
```

## 3. Precision, Recall

**Theory:** Precision measures the proportion of true positives out of all positive predictions. Recall measures the proportion of true positives out of all actual positives.

**Code:**

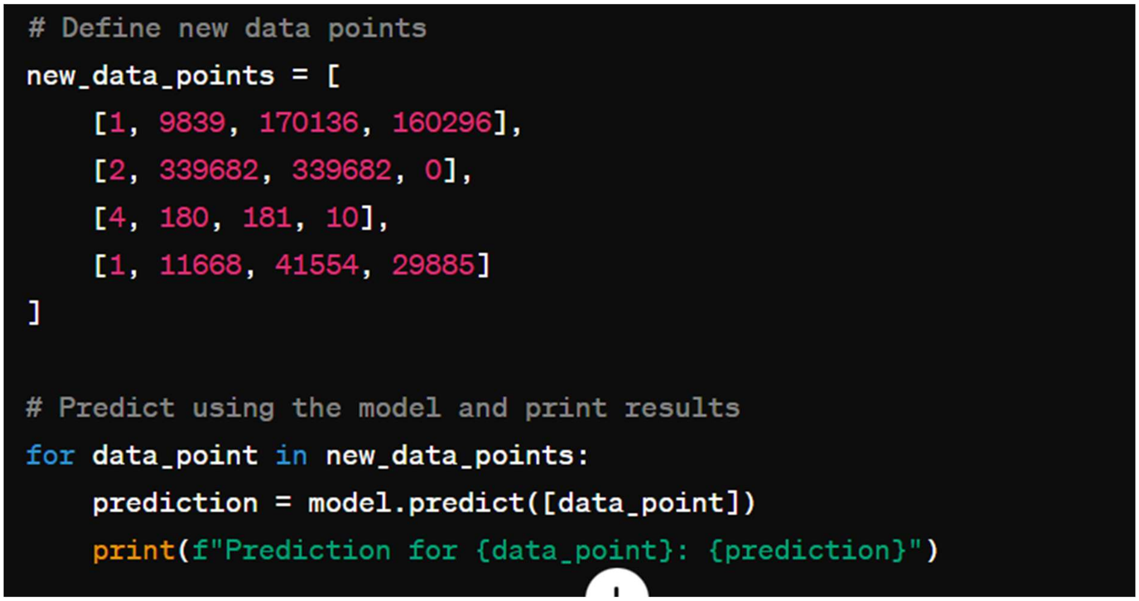


```
from sklearn.metrics import precision_score, recall_score, f1_score

# Calculate precision, recall, and F1-score
precision = precision_score(ytest, ypred)
recall = recall_score(ytest, ypred)
f1 = f1_score(ytest, ypred)

print(f"Precision: {precision}")
print(f"Recall: {recall}")
```

### 4. Predictions on New Data Points

**Theory:** This test case checks the model's performance on unseen data points.

**Code:**

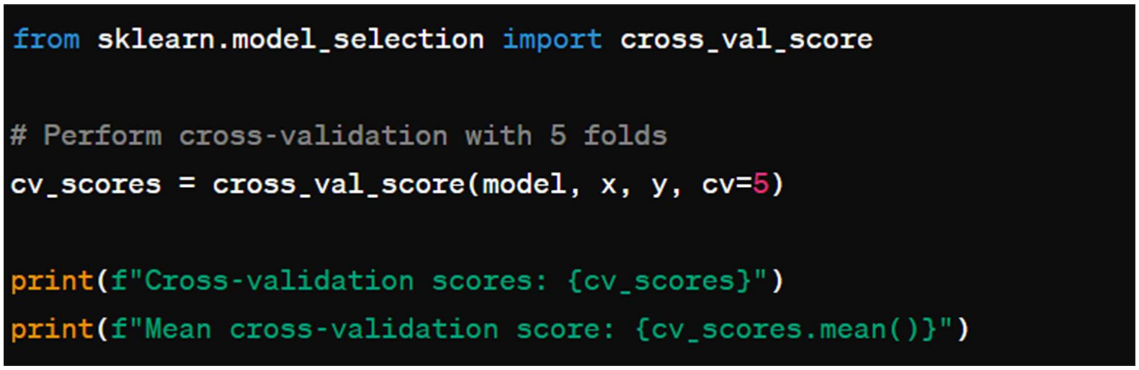


```
# Define new data points
new_data_points = [
    [1, 9839, 170136, 160296],
    [2, 339682, 339682, 0],
    [4, 180, 181, 10],
    [1, 11668, 41554, 29885]
]

# Predict using the model and print results
for data_point in new_data_points:
    prediction = model.predict([data_point])
    print(f"Prediction for {data_point}: {prediction}")
```

### 5. Cross-Validation

**Theory:** Cross-validation assesses how well the model performs across multiple different training and testing data splits. This helps check model robustness.

**Code:**



```
from sklearn.model_selection import cross_val_score

# Perform cross-validation with 5 folds
cv_scores = cross_val_score(model, x, y, cv=5)

print(f"Cross-validation scores: {cv_scores}")
print(f"Mean cross-validation score: {cv_scores.mean()}")
```

In addition to these tests, consider monitoring the model's performance over time if you deploy it in a real-world application.

These tests and metrics provide a comprehensive view of the model's performance, which allows us to assess the strengths and weaknesses across various scenarios.

# 13. Future Scope

The future scope of financial fraud detection is evolving rapidly, driven by advancements in technology, changes in financial ecosystems, and emerging fraud tactics.

Here are some key areas where the future of financial fraud detection is likely headed:

- Advancements in AI and ML algorithms will enable more sophisticated fraud detection models capable of analyzing vast amounts of data in real-time. Deep learning techniques, such as neural networks, will continue to be applied to detect complex patterns and anomalies indicative of fraud.
- The integration of behavioural biometrics, such as keystroke dynamics, mouse movement patterns, and voice recognition, will enhance fraud detection by adding an additional layer of authentication and verification beyond traditional methods.
- Financial institutions will leverage advanced analytics techniques, including data mining, predictive analytics, to uncover hidden insights and identify emerging fraud trends from large and diverse datasets.
- The shift towards real-time payment systems and transactions will necessitate the development of real-time fraud detection capabilities to quickly identify fraudulent activities as they occur, minimizing potential losses.
- Collaborative Intelligence: Financial institutions will increasingly collaborate and share information to combat fraud collectively. Collaborative intelligence platforms and information-sharing networks will facilitate the exchange of insights and threat intelligence to proactively identify and respond to fraudulent activities.
- Integration with cybersecurity solutions will become increasingly important to address the growing overlap between financial fraud and cyber threats. Unified fraud and cybersecurity platforms will provide holistic protection against both external and internal threats.

Overall, the future scope of financial fraud detection will be characterized by technological innovation, collaboration, regulatory compliance, and a continued focus on improving detection accuracy and efficiency to stay ahead of evolving fraud threats.

# 14. Conclusion

Through the integration of advanced technologies such as artificial intelligence, machine learning, and big data analytics, financial institutions are continuously enhancing their capabilities to detect and prevent fraud in real-time. The comparison of Navie Bayes, Random Forest, Decision Tree, Logistic Regression, K- Nearest Neighbor is done for finding accuracy of classification models. The finding indicate that Decision Tree algorithm (refer Fig. 10.5) has achieved highest accuracy in comparison with all other classification models. These machine learning techniques can analyse complex datasets for finding fraudulent activities. Through the integration of advanced technologies such as artificial intelligence, machine learning, and big data analytics, financial institutions are continuously enhancing their capabilities to detect and prevent fraud in real-time.

# 15. References

[1] M. N. Ashtiani and B. Raahemi, "Intelligent fraud detection in finan- cial statements using machine learning and data mining: A systematic literature review," IEEE Access, vol. 10, pp. 72504-72525, 2022.

[2] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, "Credit card fraud detection using adaboost and majority voting," IEEE Access, vol. 6, pp. 14277-14284, 2018.

[3] H. Sun, J. Li, and X. Zhu, "Financial fraud detection based on the part-of-speech features of textual risk disclosures in financial reports," Procedia Computer Science, vol. 221, pp. 57-64, 2023, tenth International Conference on Information Technology and Quantitative Management (ITQM 2023). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923007056

[4] H. Zhou, G. Sun, S. Fu, L. Wang, J. Hu, and Y. Gao, "Internet financial fraud detection based on a distributed big data approach with node2vec," IEEE Access, vol. 9, pp. 43378-43 386, 2021.

[5] A. A. Almazroi and N. Ayub, "Online payment fraud detection model using machine learning techniques," IEEE Access, vol. 11, pp. 137 188- 137 203, 2023.

[6] D. Huang, D. Mu, L. Yang, and X. Cai, "Codetect: Financial fraud detection with anomaly feature detection," IEEE Access, vol. 6, pp. 19 161-19 174, 2018.