# CREDIT CARD FRAUD DETECTION

*Dissertation submitted to Shri Ramdeobaba College of Engineering and Management, Nagpur in partial fulfillment of requirement for the award of degree of*

**Bachelor of Technology (B.Tech)**

In

**COMPUTER SCIENCE AND ENGINEERING**

*By*

**Reema Khandelwal (16)**

**Rashmi Tiwari (15)**

**Vinit Tiwari (72)**

**Yash Agrawal (73)**

*Guide*

**Ms. Wani Bisen**

RCOEM
Shri Ramdeobaba College of Engineering and Management, Nagpur

**Computer Science and Engineering**
**Shri Ramdeobaba College of Engineering & Management, Nagpur 440 013**

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University Nagpur)

**Nov 2023**

**SHRI RAMDEOBABA COLLEGE OF ENGINEERING & MANAGEMENT, NAGPUR**

**(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University Nagpur)**

**Department of Computer Science and Engineering**

# CERTIFICATE

This is to certify that the Thesis on "**CREDIT CARD FRAUD DETECTION**" is a bonafide work of "Reema Khandelwal, Rashmi Tiwari, Yash Agrawal, Vinit Tiwari" submitted to the Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur in partial fulfillment of the award of a Bachelor of Engineering, in Computer Science has been carried out at the Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur during the academic year 2022-2023.

Date:

Place:

Ms. Wani Bisen
Project Guide
Department of CSE

Dr. R. Hablani
H.O.D.
Department of CSE

Dr. R. S. Pande
Principal

# DECLARATION

We, hereby declare that the thesis titled "**CREDIT CARD FRAUD DETECTION**" submitted herein, has been carried out in the Department of Computer Science and Engineering of Shri Ramdeobaba College of Engineering & Management, Nagpur. The work is original and has not been submitted earlier as a whole or part for the award of any degree/diploma at this or any other institution / University.

Date:

Place:

Reema Khandelwal
Roll no. 16

Rashmi Tiwari
Roll no. 15

Vinit Tiwari
Roll no. 72

Yash Agrawal
Roll no. 73

# Approval Sheet

This report entitled **CREDIT CARD FRAUD DETECTION** by Reema Khandelwal, Rashmi Tiwari, Yash Agrawal, Vinit Tiwari is approved for the degree of B.Tech in Computer Science and Engineering.

Name & Signature of Supervisor(s)

…….

……..

Name & Signature of External Examiner(s)

……..

……..

Name & signature of HOD

…..……..

..…….

Date:

Place:

# ACKNOWLEDGEMENT

We would want to take this opportunity to thank everyone who was involved in this initiative, whether they were directly associated or not. Firstly, we would take this opportunity to express our deepest gratitude to **Dr. Rajesh S Pande**, Principal of Shri Ramdeobaba College of Engineering and Management. We would also like to thank **Dr Ramchand Hablani,** Head of Department of Computer Science for giving us this wonderful opportunity to work on this project and our project guide **Prof. Wani Bisen** for mentoring and supporting us throughout the project. We are very appreciative of them.

Secondly, Our sincere gratitude and appreciation go out to all of our colleagues and team members for their collaboration, support, and constructive feedback during the entire project. Every team member has given the project their all in terms of innovation, creativity, and enthusiasm.

Date:

Place:

Reema Khandelwal (16)

Rashmi Tiwari (15)

Vinit Tiwari (72)

Yash Agrawal (73)

# ABSTRACT

As online transactions become more prevalent, the risk of fraud associated with online payment methods also increases. Credit card fraud is a major threat to individuals, institutions and the broader economy and the development and deployment of reliable fraud detection algorithms is critical. This report provides a comparative analysis of five distinct algorithms, namely Random Forest,SVC, Autoencoder, Ensemble learning and Logistic Regression, to evaluate their performance in identifying fraudulent transactions in credit card data.

The study commences by examining the current landscape of credit card fraud and the importance of robust fraud detection methods to mitigate its impact. Subsequently, it delves into the intricacies of data preprocessing and feature engineering procedures, pivotal for the meticulous preparation of credit card transaction datasets, thereby accentuating the critical role played by data quality in optimizing algorithmic performance.

The research paper conducts a detailed examination of the five chosen algorithms. Random Forest, an ensemble technique renowned for its adeptness in managing intricate, high-dimensional data, is assessed. Additionally, Autoencoder, a neural network variant, is investigated due to its proficiency in capturing intricate transaction data patterns and anomalies. Logistic Regression, a firmly established linear classification method, is also incorporated for its straightforwardness and interpretative qualities. SVC which classifies data linearly and Ensemble learning model , which comprises 3 different algorithms namely, Logistic Regression, Random Forest and SVC.

# Table of Contents

# List of Figures

# Chapter 1: Introduction

## 1.1 Background:

With the increasing prevalence of e-commerce, online banking, and electronic payment systems, credit card transactions have become a fundamental part of everyday life.While these digital transactions are convenient, they also allow criminals to exploit loopholes. Credit card transactions involve various frauds such as CNP or card-not-present fraud, account takeover, identity theft, etc. This leads to adverse financial effects causing victims to suffer from financial losses, financial institutions may incur substantial costs related to fraud investigations, customer reimbursements, and implementing security measures. Our solution proposes detecting the fraud when it occurs and alerting the card holder to save them from financial losses and unnecessary troubles.

## 1.2 Problem Statement:

Credit card fraud is a growing threat in today's digital economy, resulting in significant financial losses and security risks. In a period of rising online transactions and credit card usage, effective fraud detection is critical. Our solution aims to protect consumers, financial institutions, and merchants from fraudulent activity while also ensuring the integrity and security of the payment ecosystem.

## 1.3 Aim:

The aim of this project, "Credit Card Fraud Detection" is to develop and implement an advanced fraud detection system that leverages supervised techniques such as Random Forest classification, Logistic Regression, SVC and unsupervised autoencoder neural networks to enhance the accuracy and efficiency of identifying fraudulent credit card transactions. This study attempts to improve overall financial transaction security by developing a model that can learn the underlying patterns of lawful transactions and detect abnormalities associated with fraudulent behavior. The study also aims to compare the effectiveness of the autoencoder-based system to standard fraud detection approaches, as well as to evaluate its adaptability to emerging fraud patterns in real-world scenarios. Ultimately, this research aims to contribute to the

ongoing efforts to mitigate credit card fraud and reduce financial losses for individuals and financial institutions.

## 1.4 Motivation:

The motivation behind the project stems from the escalating threat of credit card fraud, causing significant financial losses and security concerns. Traditional fraud detection methods often struggle to adapt to evolving fraud techniques, leading to false alarms and financial losses. Utilizing the enormous power of autoencoder neural networks, this research aims to develop a more robust and adaptive solution. The motive is to enhance financial security by creating a model capable of learning and identifying subtle patterns in legitimate credit card transactions while minimizing false positives. This initiative uses machine learning improvements to reduce fraud risks, helping both individuals and financial institutions by providing a more accurate and effective means of safeguarding electronic payment networks.

## 1.5 Objectives:

The objectives of the project include designing and implementing an ensemble of supervised and unsupervised techniques leveraging the concept of autoencoder model for learning patterns in legitimate credit card transactions. The project involves collecting and preprocessing a dataset comprising both genuine and fraudulent transactions. The model's performance is evaluated against traditional fraud detection methods, with a focus on precision, recall, and F1-score. Additionally, the project tests the model's adaptability to changing fraud patterns using real-world data. Optimization of model parameters, user-friendly integration, thorough documentation, and knowledge dissemination are also key goals. Ultimately, the objective is to enhance financial security by creating a more accurate and adaptable fraud detection system.

# Chapter 2:Literature Review

[1]The "Credit Card Fraud Detection using Machine Learning and Data Science" was published in the International Research Journal of Engineering and Technology (IRJET) .

The literature review provides an in-depth exploration of credit card fraud detection methods, emphasizing the role of machine learning and data science in tackling this issue. It delves into the challenges of detecting fraudulent transactions, such as class imbalance and evolving patterns, and outlines the workflow of real-world fraud detection systems. The review cites various algorithms and techniques, including artificial neural networks, genetic algorithms, decision trees, and more. The methodology section discusses the approach used in this specific study, employing Local Outlier Factor and Isolation Forest algorithms for anomaly detection. The results highlight the performance of these algorithms, with an emphasis on accuracy and precision, demonstrating that while achieving 100% accuracy remains challenging, increased dataset size can enhance precision and reduce false positives. Future enhancements, such as the integration of additional algorithms and larger datasets, are proposed to further improve credit card fraud detection systems.

[2]The "A machine learning based credit card fraud detection using the GA algorithm for feature selection" described in the Journal of Big Data highlights related work in the field of credit card fraud detection using machine learning methods. Previous research efforts have employed a variety of machine learning algorithms, including Decision Trees, Random Forest, Naive Bayes, Logistic Regression, and Artificial Neural Networks. These studies have utilized datasets, such as the one containing credit card transactions by European cardholders, to evaluate the effectiveness of their proposed methods. Addressing the challenge of class imbalance in these datasets, some authors have implemented techniques like Synthetic Minority Oversampling Technique (SMOTE) to improve model performance. Feature selection methods, including Genetic Algorithms, have been applied to optimize the choice of attributes for credit card fraud detection. The literature review also points out the limitations and areas for improvement in existing approaches, providing context for the proposed research in the paper.

[3]The paper is titled "Supervised Machine Learning Algorithms for Credit Card Fraud Detection " published in IEEE . The literature review here provides an overview of the existing research related to credit card fraud detection, focusing on the application of machine learning

models. Previous work has highlighted the increasing prevalence of credit card fraud, particularly in the context of online and offline transactions, emphasizing the need for effective fraud detection methods. Research has explored various approaches, including supervised and unsupervised machine learning techniques, as well as hybrid models that combine both. Additionally, there has been a notable effort to address the challenge of imbalanced datasets in fraud detection, with methods like under-sampling and oversampling being employed to create more balanced data. The review also touches on the significance of feature engineering, data preprocessing, and the choice of metrics in evaluating the performance of machine learning models. Overall, the literature underscores the importance of developing accurate and efficient credit card fraud detection systems to protect users from financial losses.

[4]The paper titled "Credit Card Fraud Detection Using LSTM Algorithm" published by the Journal of Computer and Mathematic Science state that Credit card fraud is a persistent issue in the financial industry, posing substantial financial losses to both cardholders and financial institutions. In response to this challenge, researchers have turned to machine learning techniques, with a notable focus on Long Short-Term Memory (LSTM) algorithms, which are renowned for their ability to process sequential data effectively.

Traditional and modern credit card fraud detection methods have evolved to address this problem. Rule-based systems have been the norm but often fail to keep pace with the dynamic nature of fraud. This inadequacy has led to a shift toward data-driven approaches . LSTM, as a specialized type of recurrent neural network (RNN), offers the promise of effectively handling sequential data. Its architecture, featuring memory cells and gates, makes it well-suited to capturing long-term dependencies in time series data, which is essential in fraud detection where fraudulent patterns may evolve over time. It concludes on the note that the application of LSTM algorithms in credit card fraud detection signifies a significant step forward in mitigating this persistent problem. The literature reveals a growing body of work that explores the various aspects of using LSTMs for fraud detection, and the potential for further advancements in the field is evident. LSTM models offer an effective means to address the evolving nature of credit card fraud and to enhance the accuracy and timeliness of detection, thereby benefiting both cardholders and financial institutions.

| Sr No | Title | Publication | Algorithm/Method used | Advantages | Future scope |
|---|---|---|---|---|---|
| 1 | Credit Card Fraud Detection using Machine Learning and Data Science | IJERT 09, September-2019 | Isolation forest algorithm(99.67 %), Local outlier factor(99.77%) | Accuracy and Efficiency,Adaptability | To reach 100% accuracy. This model can further be improved with the addition of more algorithms into it. |
| 2 | A machine learning based credit card fraud detection using the GA algorithm for feature selection | Journal of Big Data | Random Forest, Genetic Algorithm | Robust Handling of Large Datasets,reduced Noise and Overfitting | Explore the combination of GA with other optimization techniques, such as particle swarm optimization or simulated annealing, to compare their effectiveness in feature selection for fraud detection. |
| 3 | Supervised Machine Learning Algorithms for Credit Card Fraud Detection | IEEE | CDecision Tree , KNN, Random forest ,Logistic Regression,Naïve bayes | Minimum time taken to detect fraud with help of Decision tree Model | Performance of Decision Tree Model must also be evaluated with the help of unsupervised machine learning models in the future to produce a more conclusive result. |
| 4 | Credit Card Fraud Detection Using LSTM Algorithm | Journal of Computer and Mathematic Science | Recurrent and LSTM Neural Networks | LSTM model is the one that gives the highest accuracy in predicting late fees and mis-payments, and that is why it is best for banks' interests. | Incorporating calculation of H measure as H measure is a measure of the misclassification loss, and this depends on the relative proportion of objects belonging to each class. |

**Table 1. Literature Review**

# Chapter 3: Methodology

Credit card fraud detection systems employ advanced machine learning and data science techniques. These systems utilize various algorithms, including supervised, unsupervised, and deep learning models, coupled with hybrid approaches, to analyze transaction data for irregularities. Data preprocessing involves tasks like handling missing values, scaling features, and encoding categorical attributes. The model selection, hyperparameter tuning, and training with labeled data are pivotal to the process. Performance is assessed using metrics like accuracy, precision, recall, and F1-score. Comparative analysis of multiple algorithms reveals their strengths. Additionally, adaptability and scalability are examined to ensure the system can handle evolving fraud patterns and increasing transaction volumes.

## 3.1. Type of Credit Card Fraud Detection Algorithm:

Based on their features and working, algorithm here can be classified into several different categories:

### 3.1.1. Supervised Algorithm:

Supervised learning is a category of machine learning where an algorithm is trained on a labeled dataset, which means that for each input data point, there is an associated target or output value. The primary goal of supervised learning is to learn a mapping or relationship from input data to the corresponding output, so that the algorithm can make predictions or classifications on new, unseen data.The benefit of using a rule based chatbot is :

- Predictive Accuracy
- Continuous Improvement
- Effective Feature Engineering

### 3.1.1.1 Logistic Regression :

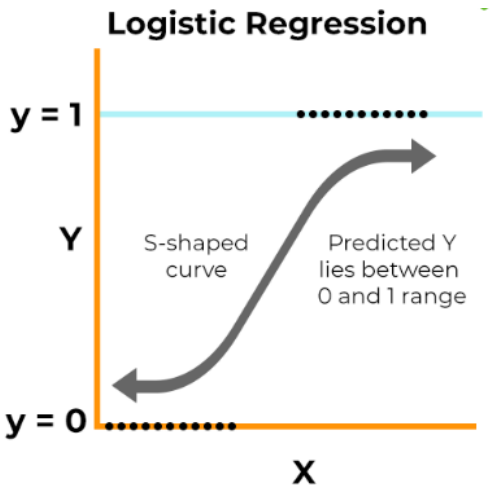



**Fig 1. Logistic Regression**

Logistic regression is a statistical method and a type of supervised learning algorithm used for binary and multi-class classification problems. It is a versatile and widely used technique in machine learning, statistics, and various fields where the goal is to predict the probability of a data point belonging to a particular category or class.

Key Features of Logistic Regression:

- Binary Classification
- Evaluation Metrics
- Sigmoid Function

### 3.1.1.2 Random Forest :

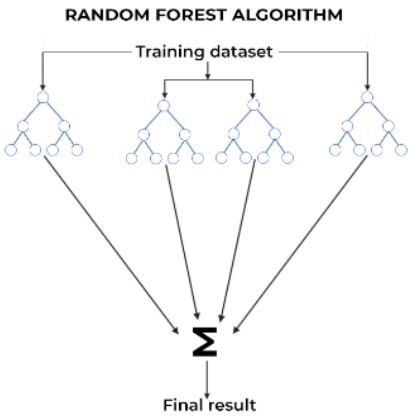



**Fig 2. Random Forest**

Random Forest is a versatile and powerful ensemble learning algorithm used for both classification and regression tasks in machine learning. It is based on the concept of decision trees and combines multiple decision trees to create a more robust and accurate predictive model. Random Forest is known for its ability to handle complex and high-dimensional data while mitigating issues such as overfitting.

Key features of Random Forest:

- Ensemble Learning
- Random Feature Selection
- Wide Applicability

#### 3.1.1.3 SVC:

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for classification and regression tasks. In SVM, the objective is to find a hyperplane that maximizes the margin between different classes in the feature space. Support Vector Classifier (SVC) is the classification variant of SVM, ideal for tasks where data needs to be separated into distinct categories. SVC works by identifying support vectors, data points closest to the decision boundary, to create an efficient classification model.

### 3.1.2 Unsupervised Learning Models

Unsupervised learning models are a category of machine learning techniques where the algorithm learns patterns, structures, or anomalies in data without the need for labeled or categorized examples. Within this category, two notable models for credit card fraud detection are highlighted:

### 3.1.2.1 Isolation Forest :

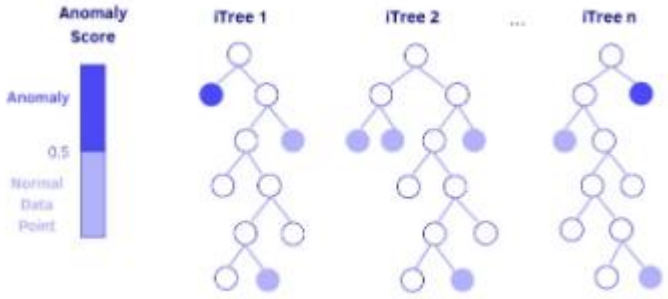



**Fig 3. Isolation Forest**

Isolation Forest is an anomaly detection algorithm that excels at identifying outliers or anomalies within a dataset. It works by creating a random forest of decision trees and isolating instances that require fewer splits to separate from the rest. This algorithm is particularly efficient at detecting rare and novel fraud cases, making it valuable for credit card fraud detection.

### 3.12.2 Local Outlier Factor(LOF) :

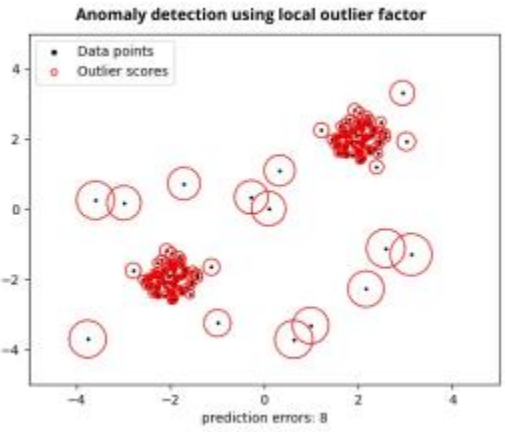



**Fig 4. Local Outlier Factor**

The Local Outlier Factor is another unsupervised anomaly detection algorithm. LOF assesses the local density of data points relative to their neighbors. It identifies instances that have a significantly lower density compared to their neighbors as potential outliers. In the context of credit card fraud detection, LOF can pinpoint transactions that deviate significantly from the typical patterns, making it a useful tool for identifying fraud.

### 3.1.3 Deep Learning Models:

Deep learning models harness intricate neural networks with multiple layers, emulating the human brain's architecture. They excel in processing intricate data, making them vital for tasks like image recognition and natural language processing. In the realm of credit card fraud detection, they bolster security by effectively identifying intricate fraud patterns and adapting to emerging threats, thereby enhancing electronic payment system security.

#### 3.1.3.1 Autoencoders:

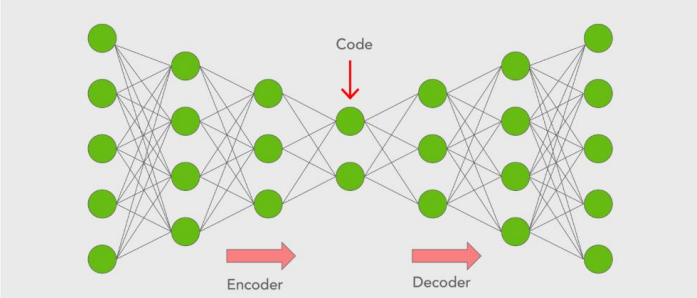



**Fig 5. Autoencoder network diagram**

Autoencoders, a subset of deep learning models, are primarily designed for unsupervised learning and dimensionality reduction. In the context of fraud detection, they become adept at encoding and decoding normal transactions. Their capacity to flag deviations as potential fraud is instrumental. Autoencoders further display adaptability to ever-evolving fraud techniques, making significant contributions to bolstering the security of electronic payment systems.

### 3.1.4 Hybrid Models

Hybrid models in the realm of credit card fraud detection combine different machine learning techniques, often mixing supervised and unsupervised approaches. By integrating both types of methods, these models aim to leverage the strengths of each. This hybridization allows for enhanced accuracy and adaptability, effectively identifying both known fraud patterns and anomalies.

### 3.1.4.1 Combining Supervised and Unsupervised Techniques

In credit card fraud detection, combining supervised and unsupervised techniques entails merging the power of labeled data with the flexibility of unsupervised learning. This approach uses labeled fraud cases to train models, enhancing their ability to recognize known fraud patterns. Simultaneously, it applies unsupervised methods to identify unknown anomalies, which might represent emerging fraud techniques.

### 3.1.4.2 Ensemble Methods

Ensemble methods in fraud detection combine the outputs of multiple models to improve overall accuracy and reliability. They often integrate various algorithms, such as decision trees, neural networks, or support vector machines. By aggregating the results, ensemble methods reduce the impact of individual model weaknesses and enhance fraud detection by effectively reducing false positives and increasing precision.

# Chapter 4. Implementation details

Outline of the proposed methodology in this paper is demonstrated in Figure 1. It focuses on various steps involved starting from processing the dataset till integrating with our GUI and showing results.

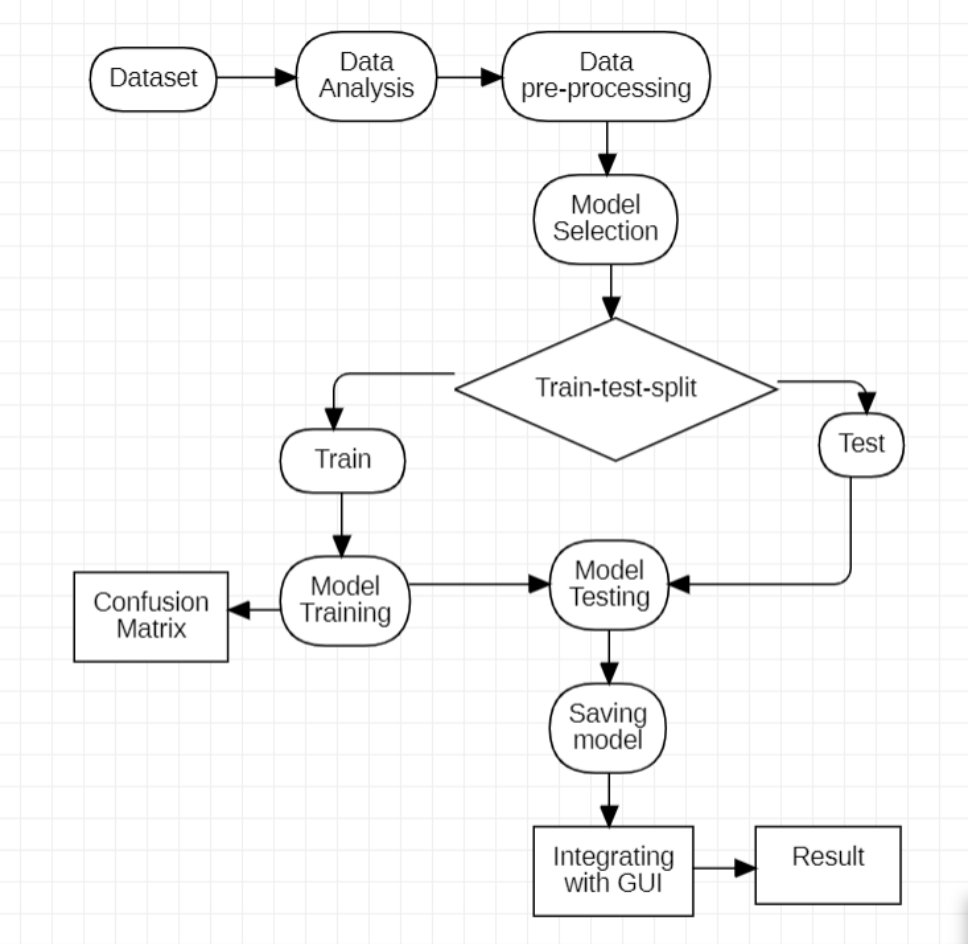



**Fig. 6 Outline of proposed methodology**

## 4.1 Dataset Used

This is a publicly available dataset taken from open-source site ‘Kaggle’.This Dataset consists of simulated credit card transaction dataset containing legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants.

This was generated using Sparkov Data Generation | Github tool created by Brandon Harris. This simulation was run for the duration - 1 Jan 2019 to 31 Dec 2020. The files were combined and converted into a standard format. Total features in this dataset are around 22 including the target class ‘is_fraud’. Total entries (rows) in this dataset are 15593 .

| | Unnamed: 0 | trans_date_trans_time | cc_num | merchant | category | amt | first | last | gender | street | city | state | zip | lat | long | city_pop | job | dob | trans_num | unix_time | merch_lat | merch_long | is_fraud |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 2019-01-01 00:00:18 | 2703186189652095 | fraud_Rippin, Kub and Mann | misc_net | 4.97 | Jennifer | Banks | F | 561 Perry Cove | Moravian Falls | NC | 28654 | 36.0788 | -81.1781 | 3495 | Psychologist, counselling | 1988-03-09 | 0b242abb623afc578575680df30655b9 | 1325376018 | 36.011293 | -82.048315 | 0 |
| 1 | 1 | 2019-01-01 00:00:44 | 630423337322 | fraud_Heller, Gutmann and Zieme | grocery_pos | 107.23 | Stephanie | Gill | F | 43039 Riley Greens Suite 393 | Orient | WA | 99160 | 48.8878 | -118.2105 | 149 | Special educational needs teacher | 1978-06-21 | 1f76529f8574734946361c461b024d99 | 1325376044 | 49.159047 | -118.186462 | 0 |
| 2 | 2 | 2019-01-01 00:00:51 | 38859492057661 | fraud_Lind-Buckridge | entertainment | 220.11 | Edward | Sanchez | M | 594 White Dale Suite 530 | Malad City | ID | 83252 | 42.1808 | -112.2620 | 4154 | Nature conservation officer | 1962-01-19 | e1a22d70485983eac12b5b88dad1cf95 | 1325376051 | 43.150704 | -112.154481 | 0 |
| 3 | 3 | 2019-01-01 00:01:16 | 3534093764340240 | fraud_Kutch, Hermiston and Farrell | gas_transport | 45.00 | Jeremy | White | M | 9443 Cynthia Court Apt. 038 | Boulder | MT | 59632 | 46.2306 | -112.1138 | 1939 | Patent attorney | 1967-01-12 | 6b849c168bdad6f867558c3793159a81 | 1325376076 | 47.034331 | -112.561071 | 0 |
| 4 | 4 | 2019-01-01 00:03:06 | 375534208663984 | fraud_Keeling-Crist | misc_pos | 41.96 | Tyler | Garcia | M | 408 Bradley Rest | Doe Hill | VA | 24433 | 38.4207 | -79.4629 | 99 | Dance movement psychotherapist | 1986-03-28 | a41d7549acf90789359a9aa5346dcb46 | 1325376186 | 38.674999 | -78.632459 | 0 |

**Fig 7. Dataset Used**

## 4.2 Required Libraries

The libraries required for project are :

### 4.1.1 Streamlit

Streamlit is an open-source Python library that simplifies the process of creating web applications for data science and machine learning. With Streamlit, you can transform data scripts into shareable web apps with minimal effort. It boasts features such as rapid development, interactive widgets, and support for various data visualization libraries, making it an ideal choice for creating interactive data dashboards, prototyping machine learning models, and sharing data insights.

### 4.1.2 Pandas

Pandas is a highly popular open-source library for data manipulation and analysis in Python. It provides data structures and functions for working with structured data, including dataframes and series. Pandas is known for its efficiency in data manipulation, data cleaning, transformation, and statistical analysis. It seamlessly integrates with other data analysis libraries, making it indispensable for tasks such as data cleaning, preprocessing, and data preparation for machine learning.

### 4.1.3 NumPy

NumPy (Numerical Python) is a fundamental library for numerical and scientific computing in Python. It stands out for its support of large, multi-dimensional arrays and matrices, along with an extensive collection of mathematical functions for operating on these arrays. NumPy is vital for scientific computing, data analysis, and machine learning applications that involve numerical operations and complex mathematical calculations.

### 4.1.4 Scikit-learn

Scikit-learn is a widely-used machine learning library for Python. It offers a comprehensive set of tools for data mining, data analysis, and model development. Built on other scientific libraries like NumPy and SciPy, Scikit-learn provides machine learning algorithms for classification, regression, clustering, dimensionality reduction, and more. It is also equipped with features for model evaluation, selection, and integration into data science workflows, making it a valuable resource for building and assessing predictive models.

### 4.1.5 Keras

Keras is an open-source deep learning framework that operates on top of other deep learning libraries like TensorFlow and Theano. Keras offers a user-friendly and high-level interface for building and training neural networks. It is known for its simplified API, flexibility in designing various neural network architectures, and compatibility with different backend engines. Keras is commonly employed for deep learning tasks such as image recognition, natural language processing, and constructing neural networks for various applications.

## 4.3 Dataset analysis

The figure below shows that the dataset is highly unbalanced. The dataset consists of large amounts of Non-fraud Transactions as compared to Fraudulent transactions. This will directly affect the result. When we train our models on this dataset, the accuracy will come out to be close to 99 percent as the model is trained on an unbalanced dataset. To solve this , we have to incorporate different sampling techniques. Different sampling methods that we tried are :

1. Oversampling
2. UnderSampling

3. SMOTE

Out of these, we found that our model works best with SMOTE sampling technique.

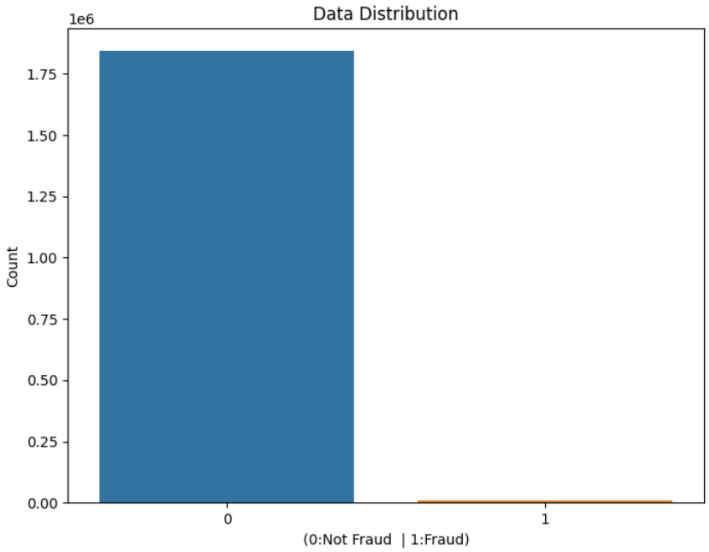



**Fig 8. SMOTE sampling**

### 4.3.1 Analysis of Amount class vs Fraudulent transaction

The figure below shows how much amount is getting transferred when the transaction is fraudulent and Non-fraudulent transactions. The graph depicts that a total $200 or less amount of fraudulent transactions tend to peak around $300 and then hover around the $800–$1000 mark. Thus, we can definitely see some patterns and can include amount class in our feature data.

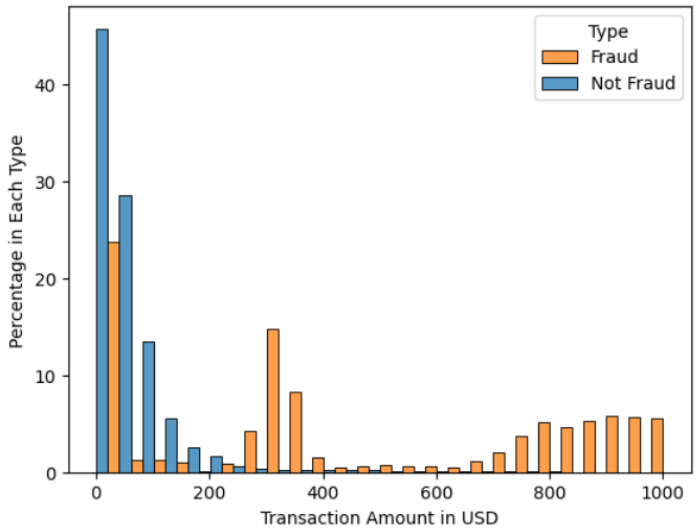



**Fig 9. Analysis of Amount class vs Fraudulent transaction**

### 4.3.2 Analysis of gender class vs Fraudulent transaction

The figure below depicts that fraudulent transactions and non-fraudulent transactions are divided equally in terms of age. No valid pattern can be detected here.

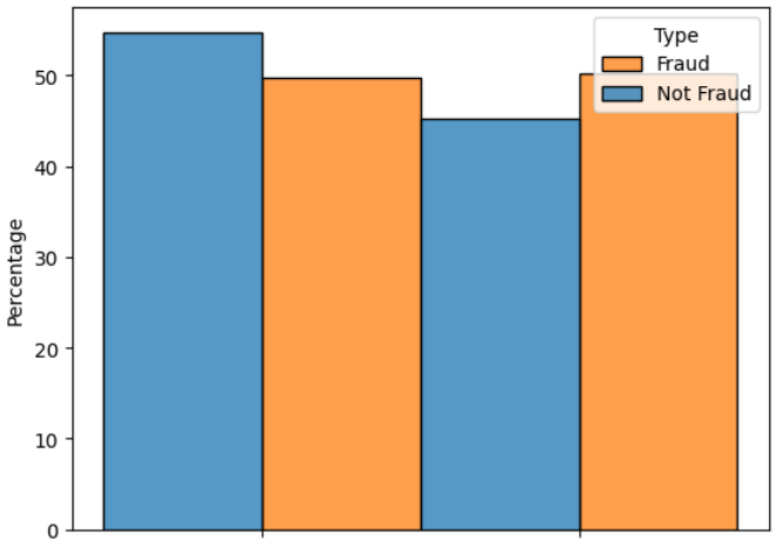



**Fig 10. Analysis of gender class vs Fraudulent transaction**

### 4.3.3 Analysis of spending Category vs Fraudulent transaction

The graph below shows the percentage of fraudulent transactions happening in a particular spending category. Fraud tends to happen more often in 'Shopping_net', 'Grocery_pos', and 'misc_net' while 'home' and 'kids_pets' among others tend to see more normal transactions than fraudulent ones. Thus, we can clearly see a pattern arising here.

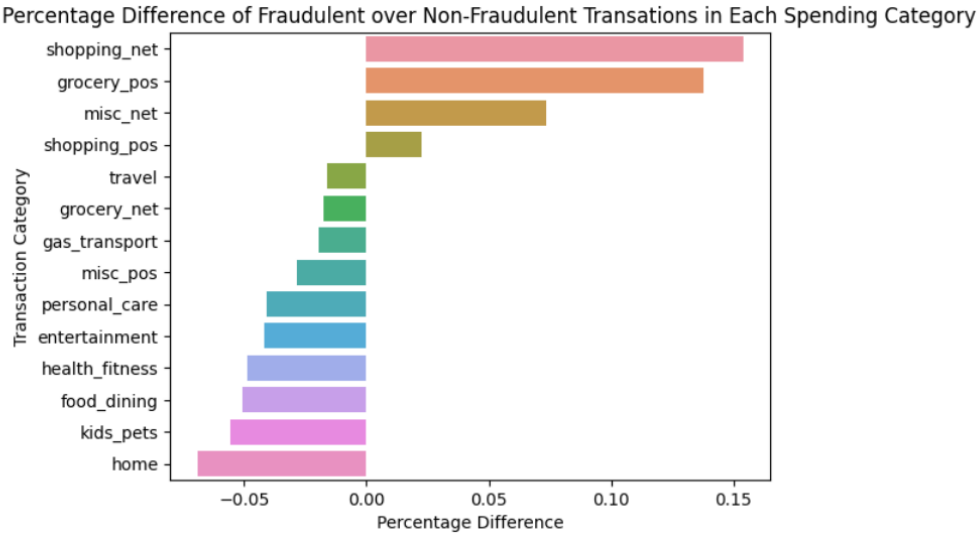



**Fig.11 Analysis of spending Category vs Fraudulent transaction**

### 4.3.4 Analysis of age vs Fraudulent transaction

The age distribution is visibly different between 2 transaction types. In normal transactions, there are 2 peaks at the age of 37-38 and 49-50, while in fraudulent transactions, the age distribution is a little smoother and the second peak does include a wider age group from 50-65. This does suggest that older people are potentially more prone to fraud.

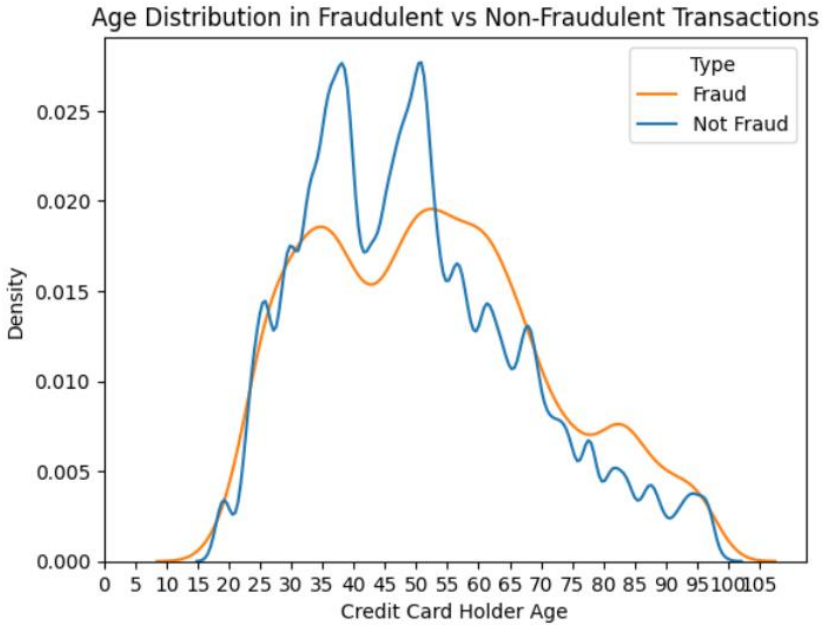



**Fig. 12 : Analysis of age vs Fraudulent transaction**

### 4.3.5 Trends in hours, days, months.

Given below are 3 graphs which show the trend of fraudulent transactions which are made in different hours, days and months.

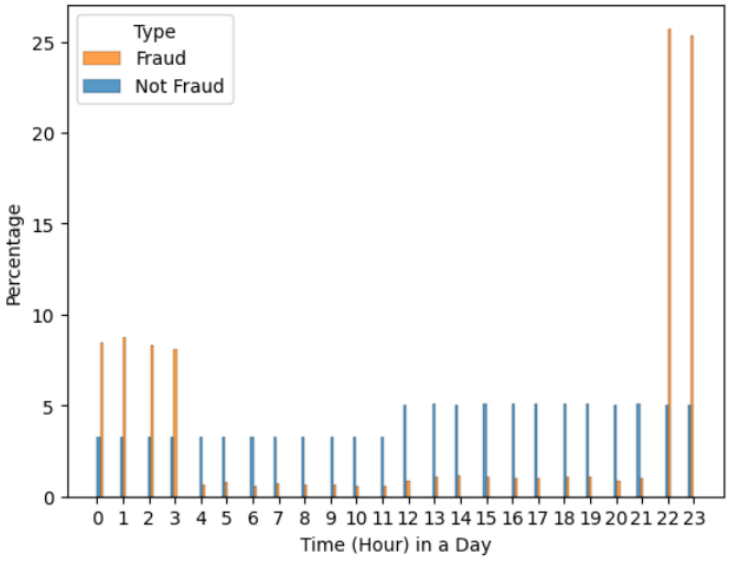



**Fig. 13 Plot of trends in hours**

The trend of hours vs transaction shows that  normal transactions are distributed more or less equally throughout the day. Fraudulent payments happen disproportionately around midnight when most people are asleep!

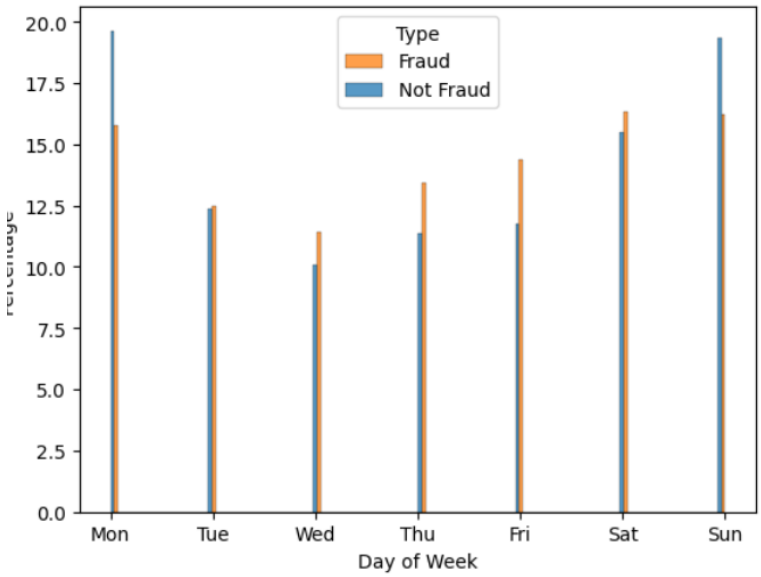



**Fig. 14 Plot of trends in days of week**

The trend of days of week vs transactions shows that Normal transactions tend to happen more often on Monday and Sunday while fraudulent ones tend to spread out more evenly throughout the week.

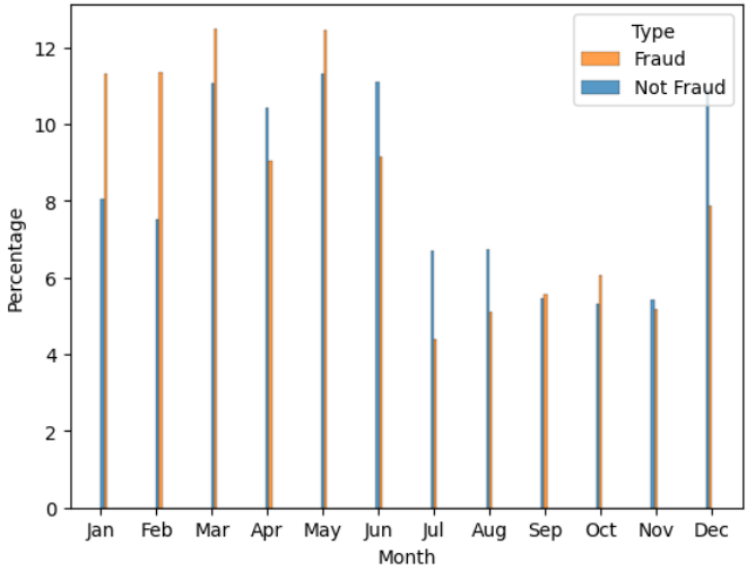



**Fig. 15 Plot of trends vs month**

And the monthly trend depicts that fraudulent transactions are more concentrated in Jan-May.

### 4.3.6 Analysis of Fraudulent transactions on the basis of state features.

The below graph depicts the percentage of fraud happening in different states. As it can be seen, NY and OH among others have a higher percentage of fraudulent transactions than normal ones, while TX and MT are the opposite.

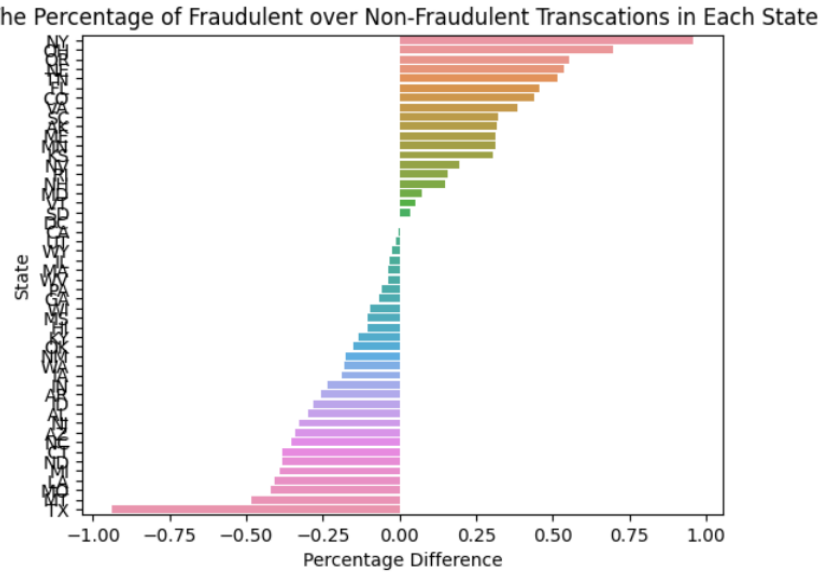



**Fig. 16 Analysis of Fraudulent transactions on the basis of state features.**

## 4.4 Data preprocessing

In this step, we performed some standardized preprocessing steps which are given below:

1. Data Cleaning : Removed duplicates and handled missing values.
2. Feature Engineering : Created some useful columns like hour, days and month to analyze the trend and appended these columns in our dataset to increase accuracy. Finally, we concluded our EDA and took the following features into consideration : category, state , amount ,zip, latitude , longitude , city population , merch_latitude , merch longitude , age , hour, day, month, is_fraud.
3. Applying Data sampling techniques : Used SMOTE (synthetic minority oversampling) to balance our dataset.
4. Data Splitting : Splitting our dataset into training and test sets.

5. Data Normalization : using One-hot Encoder for converting string column like states and categorical variables into Numerical values and using Standard scaler of Scikit-learn library for normalization.

```
preprocessor = ColumnTransformer([
    ('num', StandardScaler(), numeric_columns),
    ('cat', OneHotEncoder(sparse_output=False, drop='first'), categorical_columns)
])
```

**Fig.17 Data preprocessing**

## 4.5 Model Training

For supervised algorithm, we followed the steps given below:

1. In model training, we first trained our models on preprocessed data. We used different models like Logistic regression, Random forest, SVC.
2. Then we defined a pipeline in Python using scikit-learn's Pipeline class. This is a common practice for creating a sequential workflow that includes data preprocessing and a machine learning model.
3. Then we made predictions using the pipeline pipe1 and calculated the accuracy of those predictions. We used our already split data(X_test and y_test) and fitted pipeline (pipe1) that includes both data preprocessing and a logistic regression model.

```
[ ] model1=LogisticRegression(solver='lbfgs', max_iter=400)
    model1.fit(X_train_resampled, y_train_resampled)

    pipe1 = Pipeline(steps=[
        ('step1', preprocessor),
        ('step2', model1)
    ])

    y_pred = pipe1.predict(X_test)
    # Calculate accuracy
    accuracy = accuracy_score(y_test, y_pred)
    print(f"Accuracy: {accuracy}")
```

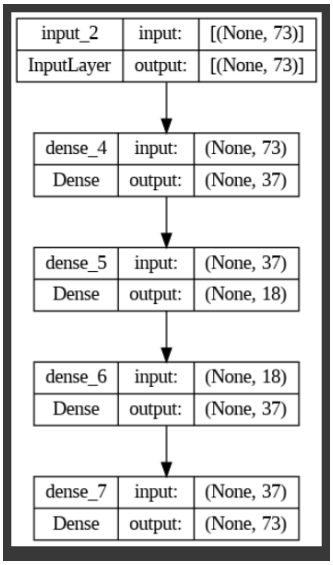
```
▾ LogisticRegression
LogisticRegression(max_iter=400)
```

**Fig. 18 Pipeline creation**

For autoencoder model, we followed the given steps :

1. In autoencoder, we preprocessed and applied SMOTE to balance our dataset.This step resulted in converting our 14 featured dataset to 73 featured dataset.
2. Then we created 1 input layer, followed by 2 encoding and 2 decoding layer.The encoder layers reduce the dimensionality of the input data while capturing important features.

The training process utilizes the RMSPROP optimizer and minimizes the mean squared error loss. Additionally, a ModelCheckpoint callback is implemented to save the model with the highest validation accuracy.




| Layer (type) | Output Shape | Param # |
|---|---|---|
| input_2 (InputLayer) | [(None, 73)] | 0 |
| dense_4 (Dense) | (None, 37) | 2738 |
| dense_5 (Dense) | (None, 18) | 684 |
| dense_6 (Dense) | (None, 37) | 703 |
| dense_7 (Dense) | (None, 73) | 2774 |

Total params: 6899 (26.95 KB)
Trainable params: 6899 (26.95 KB)
Non-trainable params: 0 (0.00 Byte)

**Fig. 19 Layers of autoencoder model**

3. The classification model for fraud detection is defined using the Keras sequential model. The model architecture consists of an initial layer, class_model, which probably works as a feature extractor. This is followed by a dense layer with 256 units and a hyperbolic tangent (tanh) activation function. The last layer is a dense layer with a single unit and a sigmoid activation function, which is common in binary classification tasks.
4. The Model Checkpoint callback is configured to save the model with the highest validation accuracy. The model is compiled using the 'adam' optimizer and the mean squared error (MSE) loss function to enhance accuracy.

For the ensemble model, we used votingClassifier and used 3 models in it. Logistic regression, Random forest and SVC.

```python
# Create a Voting Classifier
ensemble_model = VotingClassifier(estimators=[
    ('linear', logistic_model),
    ('random_forest', random_forest),
    ('svc', svc)
], voting='hard')

# Fit the ensemble model on the training data
ensemble_model.fit(creditCardFeatures_sampled, creditCardLabels_sampled)

# Make predictions on the validation set
ensemble_predictions = ensemble_model.predict(x_test)

# Evaluate the ensemble model
accuracy = accuracy_score(y_test, ensemble_predictions)
print("Ensemble Model Accuracy:", accuracy)
```

**Fig. 20 Ensemble Model**

## 4.6 Integration with User-Interface

The machine learning models was integrated with a user-friendly interface using Streamlit which involves creating a web application that provides an interface for users to interact with our model. Streamlit is a popular Python library for building web applications quickly and easily. Here are the steps to deploy a model using Streamlit:

- First ensure that you have streamlit library installed if it is not present in your system then run **pip install streamlit**
- Now run **streamlit run app.py**

```
PS G:\Credit_card_fraud_detection\Credit_card_fraud_detection> streamlit run app.py

      Welcome to Streamlit!

      If you'd like to receive helpful onboarding emails, news, offers, promotions,
      and the occasional swag, please enter your email address below. Otherwise,
      leave this field blank.

      Email:  tiwarirs@rknec.edu

  You can find our privacy policy at https://streamlit.io/privacy-policy

  Summary:
  - This open source library collects usage statistics.
  - We cannot see and do not store information contained inside Streamlit apps,
    such as text, charts, images, etc.
  - Telemetry data is stored in servers in the United States.
  - If you'd like to opt out, add the following to %userprofile%/.streamlit/config.toml,
    creating that file if necessary:

    [browser]
    gatherUsageStats = false


  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.52.65:8501
```

**Fig. 21 : GUI integration**

# Chapter 5:Result and Discussion

The act of using another person's credit card to make purchases or seek cash advances without the cardholder's knowledge or agreement is known as credit card fraud. The efficiency of several methods for detecting credit card fraud has been revealed through performance evaluation. Random Forest is the highest performance among the models examined, with an astounding accuracy of 99.7%. Random Forest's capacity to manage complicated data linkages and make very accurate predictions has made it an excellent choice for this essential task.

A simpler but still strong model, logistic regression, obtained an accuracy of 96.4%. Its reasonably high accuracy implies that it might be a dependable alternative for fraud detection, with the added benefit of being interpretable and simple to deploy.

The Support Vector Classifier (SVC) achieved a competitive accuracy of 96.6%, demonstrating its efficacy in capturing complicated decision boundaries. While it may necessitate more processing resources than simpler models, its performance justifies its use in situations where precision is critical.

The ensemble model, which incorporates the strengths of numerous models, was 96.1% accurate. While it did not exceed the separate models, it demonstrates the power of combining several algorithms to improve prediction skills.

The Autoencoder model, designed exclusively for anomaly identification, obtained 91.3% accuracy. Although it did not equal the accuracy of the other models, it excels at finding unique patterns and might be a useful addition to other models for detecting uncommon and nuanced fraud situations.

Finally, the unique needs and goals of the application should guide the selection of a credit card fraud detection model. In terms of accuracy, Random Forest emerges as the clear winner, making it an ideal choice when minimizing false positives is critical. Logistic Regression and SVC are both excellent competitors, providing a good combination of accuracy and

interpretability. The ensemble model and Autoencoder are useful tools for enhancing overall detection skills, particularly in circumstances with a wide range of fraud behaviors. Combining these models, or selecting one based on the unique use case, can assist organizations in efficiently combating credit card fraud and protecting their consumers from financial loss.

| Model | Random Forest | Logistic Regression | SVC | AutoEncoders | Ensemble |
|---|---|---|---|---|---|
| Accuracy | 99.77 | 96.44% | 96.601% | 99.93% | 96.15% |

**Table 2 : Comparative analysis of accuracies of various models**

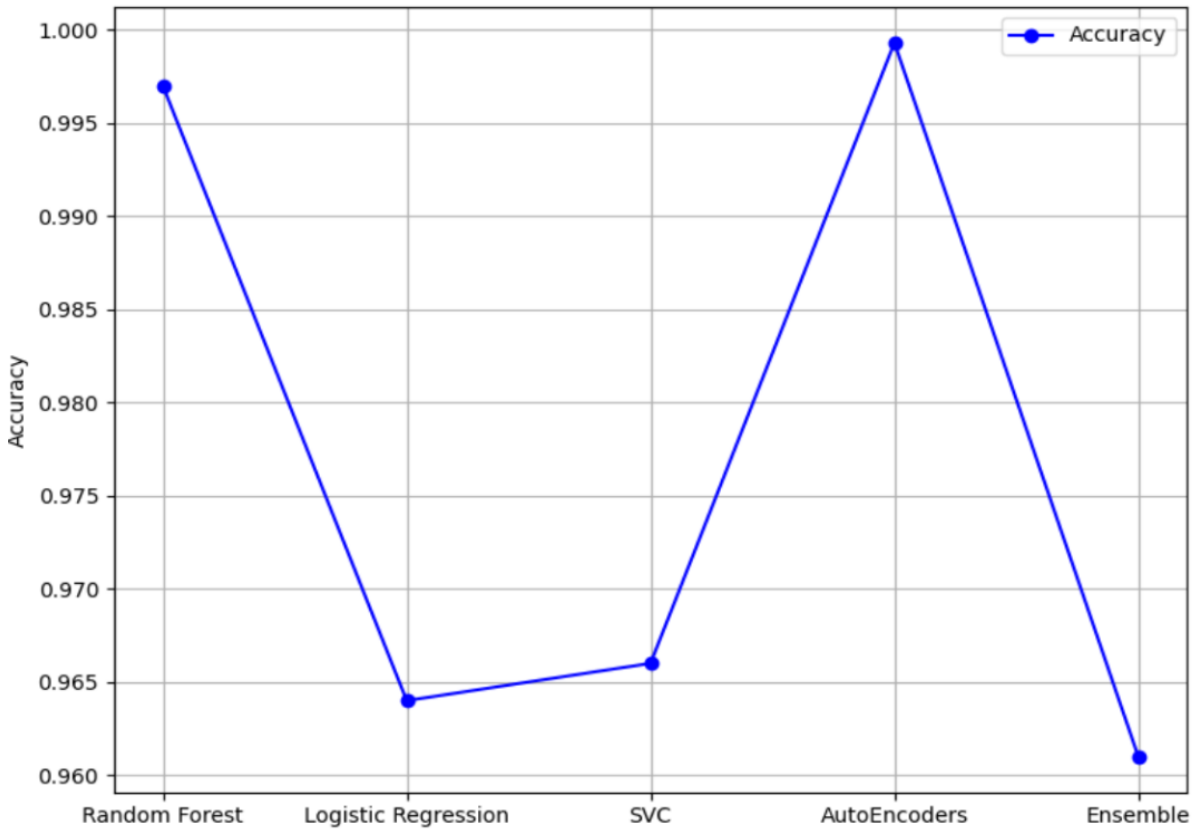



**Fig 22. Algorithm accuracy comparison**

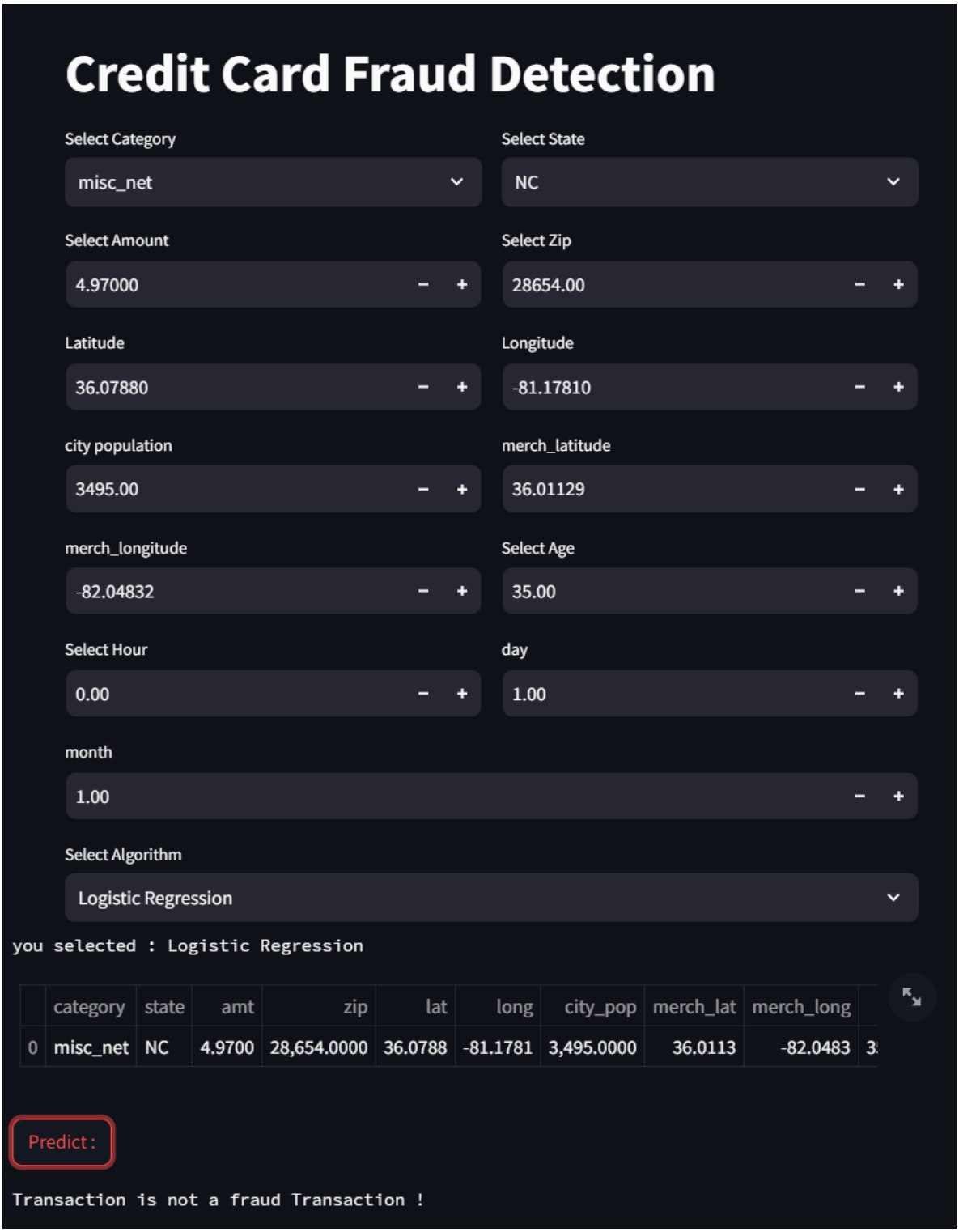



**Fig 23. Demonstration of result**

# Chapter 6 : Conclusion and Future work

Credit card fraud is one of the biggest frauds that are happening right now around the globe. This paper has explained how credit card frauds have been happening and we studied these frauds using a dataset that consists of transactions made in the real world.

We saw how different machine learning algorithms are used to predict the fraud transactions on our dataset and we also addressed the class imbalance issue of our dataset and used SMOTE techniques to address this issue. Later, we trained different models to classify Fraudulent transactions which came up with good accuracy.

Future research in this area should concentrate on staying ahead of these issues and increasing the overall efficacy of fraud detection systems. Consider using sophisticated machine learning and deep learning models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs).

Another area of development in our project is mixing different models using ensemble learning and creating a model whose accuracy can be higher than other models. Also, a notification system can be built which takes user details like name, contact number and sends an SMS if the transaction is fraudulent.

# Chapter 7: Reference

[1] Emmanuel Ileberi, Yanxia Sun ,Zenghui Wang. A machine learning based credit card fraud detection using the GA algorithm for feature selection

[2] Maniraj SP, Saini A, Ahmed S, Sarkar D. Credit card fraud detection using machine learning and data science. Int J Eng Res 2019; 8(09).

[3] Dornadula VN, Geetha S. Credit card fraud detection using machine learning algorithms. Proc Comput Sci. 2019;165:631–41.

[4] Khatri S, Arora A, Agrawal AP. Supervised machine learning algorithms for credit card fraud detection: a comparison. In: 10th international conference on cloud computing, data science & engineering (Confluence); 2020. p. 680-683.

[4] Credit Card Definition.‖ https://www.investopedia.com/terms/c/creditcard.asp(accessed Apr.

[5] V. N. Dornadula and S. Geetha, —Credit Card Fraud Detection using Machine Learning Algorithms,‖ ProcediaComput. Sci., vol. 165, pp. 631–641, 2019, doi: 10.1016/j.procs.2020.01.057.

[6] B. Wickramanayake, D. K. Geeganage, C. Ouyang, and Y.Xu, —A survey of online card payment fraud detection using data mining-based methods,‖ arXiv, 2020.

[7] A. Agarwal, —Survey of Various Techniques used for Credit Card Fraud Detection,‖ Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 7, pp. 1642–1646, 2020, doi:10.22214/ijraset.2020.30614.