Project Report On

**Credit Card Fraud Detection**

****

Submitted in the partial fulfillment for the award of Post Graduate Diploma in Big Data Analytics (PG-DBDA)

from Know-IT ATC, CDAC ACTS, Pune

**Guided by:**

Mr. Milind Kapse

Submitted By:

Gopal Chavan (250243025010)

Megharani Pol (250243025023)

Saurav Mahto (250243025036)

Arbaj Shaikh (250243025037)

# CERTIFICATE

**TO WHOMSOEVER IT MAY CONCERN**

### This is to certify that

Gopal Chavan (250243025010)

Megharani Pol (250243025023)

Saurav Mahto (250243025036)

Arbaj Shaikh (250243025037)

### have successfully completed their project on

Credit Card Fraud Detection

**Under the guidance of Mr. Milind Kapse**

# ACKNOWLEDGEMENT

The Credit Card Fraud Detection is designed to detect fraudulent activities in credit card transactions. This project has provided us with valuable expertise in applying advanced data preprocessing, feature engineering, and machine learning algorithms to identify suspicious patterns in financial transactions with high accuracy.

We are all grateful to Milind Kapase Sir and for their invaluable help while working on this project. Their advice and support enabled us to overcome a variety of challenges and complexities during the course of the project.

We are grateful to Mr. Shrinivas Jadhav (Vice-President, Know-it, Pune) for his assistance and support throughout the Post Graduate Diploma in Big Data Analytics (PGDBDA) course at CDAC ACTS, Pune.

Our heartfelt gratitude goes to Mrs. Dhanashree (Course Coordinator, PG-DBDA), who provided all of the necessary support and kind coordination to provide all of the necessities, such as required hardware, internet access, and extra lab hours, to complete the project and throughout the course up to the last day here at CDAC KnowIT, Pune.

#### From:

Gopal Chavan (250243025010)

Megharani Pol (250243025023)

Saurav Mahto (250243025036)

Arbaj Shaikh (250243025037)

## TABLE OF CONTENTS

**ABSTRACT**

1. **INTRODUCTION**
2. **SYSTEM REQUIREMENTS**
   1. Software Requirements
   2. Hardware Requirements
3. **FUNCTIONAL REQUIREMENTS**
4. **SYSTEM ARCHITECTURE**
5. **METHODOLOGY**
6. **MACHINE LEARNING ALGORITHMS**
7. **DATA VISUALIZATION AND REPRESENTATION**
8. **CONCLUSION AND FUTURE SCOPE**

## ABSTRACT

In this project, Credit card fraud poses a significant challenge to the financial industry, leading to substantial monetary losses and undermining customer trust. With the increasing volume of digital transactions, traditional rule-based detection methods often fall short in identifying sophisticated fraudulent patterns. This project aims to develop a robust machine learning–based fraud detection system capable of accurately classifying transactions as fraudulent or legitimate. The dataset, sourced from Kaggle, contains over 1.2 million transaction records, including features such as transaction amount, merchant category, transaction time, distance, customer demographics, and location-based metrics. Data preprocessing and feature engineering were performed using PySpark, including the creation of derived features such as amt\_per\_capita, is\_far, and city\_freq. Class imbalance was addressed using SMOTE-Tomek to improve model generalization. Several machine learning algorithms were evaluated, with XGBoost achieving the highest performance, recording an accuracy of 99% and an AUC of 0.99. The system was deployed using Streamlit for real-time predictions and integrated with Tableau dashboards for monitoring and analysis. The results demonstrate the effectiveness of combining scalable data preprocessing, engineered features, and advanced algorithms to detect fraudulent credit card transactions, offering a practical framework that can be adapted for broader fraud detection applications.

## INTRODUCTION

The banking and financial sector is a fast-paced and highly sensitive domain, where identifying and preventing fraudulent credit card transactions is critical to ensuring customer trust and minimizing financial losses. With the rapid advancement of machine learning techniques, it is now possible to process large-scale transaction data and build highly accurate models for fraud detection. By leveraging historical data on transactions, customer demographics, merchant information, location, and spending behavior, machine learning algorithms can uncover hidden patterns that indicate potential fraud. This enables banks and payment processors to take proactive measures, enhance security, and optimize fraud prevention strategies, thereby improving both efficiency and reliability in financial systems.

* In this project, we present a study on credit card fraud detection using machine learning techniques

applied to a large transaction dataset. The dataset contains details such as transaction amount,

merchant, category, location, time of transaction, customer profile, and a fraud label.

* We will use this dataset to develop and evaluate machine learning models capable of

classifying transactions as fraudulent or legitimate based on the available features.

* The main objective of this study is to compare different machine learning algorithms for credit card

fraud detection and identify the one that achieves the best balance between fraud detection rate

(recall) and minimizing false positives (precision). We will also explore feature engineering

techniques to enhance model accuracy and investigate the most influential features in detecting fraud.

* Overall, this project aims to provide valuable insights into the application of machine learning

for detecting fraudulent credit card transactions and understanding the behavioral patterns that

contribute to such activities in real-world financial systems.

### 

### DATA COLLECTION AND FEATURES

**Data Source:**

The primary data source for this project is the **Kaggle Credit Card Transactions Dataset**, a

publicly available dataset containing anonymized transaction records from a major financial institution.

The dataset includes transaction details such as amount, merchant category, transaction time,

location, customer demographic information, and a fraud indicator. This rich set of features enables

effective preprocessing, feature engineering, and the development of robust machine learning models

for detecting fraudulent transactions.

**Dataset Size:**

The dataset comprises a total of **1,200,000 transaction records**, which includes both fraudulent

and legitimate entries. This comprehensive dataset combines information on transaction amounts,

merchant categories, timestamps, customer demographics, and geolocation details, resulting in

a rich and diverse collection of data. This extensive dataset provides a robust foundation for in-depth

analysis and the development of accurate machine learning models for detecting fraudulent activities in

credit card transactions.

**Features / Attributes**

Here is an overview of the key features (attributes) within our dataset:

1. Transaction Amount (amt)

* Represents the monetary value of the transaction in USD.
* Helps identify unusually high or low spending patterns that could indicate fraud.
* Continuous numerical feature.

1. Transaction Category (category)

* Denotes the type of merchant or purchase, such as food\_dining, gas\_transport, grocery\_pos, home, kids\_pets, shopping\_net, shopping\_pos.
* Useful for detecting unusual purchases that do not align with the cardholder’s typical spending behavior.
* Encoded as multiple binary category columns for modeling.

1. Transaction Time (hour\_bucket\_encoded)

* Represents the hour of the day when the transaction took place, grouped into time buckets.
* Fraudulent activities often occur at unusual hours compared to normal spending patterns.
* Encoded as numerical buckets for model input.

1. Gender (gender\_M)
   * Indicates the gender of the customer (Male / Female).
   * Encoded into binary format for model processing.
2. Amount per Capita (amt\_per\_capita)

* Engineered feature representing transaction amount relative to household size or average spend.
* Helps identify unusually large purchases compared to normal spending behavior.

1. City Transaction Frequency (city\_freq)

* Number of transactions made in the same city by the customer.
* Unusual spikes in city-based transactions may indicate fraudulent activity.

1. Distance (distance)

* The geographical distance between the customer’s home location and the transaction location.
* Larger distances can indicate suspicious activity, especially when inconsistent with prior transactions.
* Additional binary feature is\_far created for distances above the 95th percentile.

1. Customer Age (age)

* Age of the customer at the time of transaction.
* Certain age groups may have distinct spending patterns, aiding in anomaly detection.

1. Fraud Label (is\_fraud)

* Target variable indicating whether the transaction is fraudulent or legitimate.
* 1 → Fraudulent transaction
* 0 → Legitimate transaction
* Used as the dependent variable for training and evaluating machine learning models.

## 

## SYSTEM REQUIREMENTS

#### Hardware Requirements:

* Platform – Windows 7 or above
* RAM – Recommended 8 GB of RAM
* Peripheral Devices – Keyboard, Monitor, Mouse
* WiFi connection with minimum 2 Mbps speed

#### Software Requirements:

* Language: Python 3
* Machine Learning
* Tableau
* OS – Windows
* Pyspark

## 

## FUNCTIONAL REQUIREMENTS

#### Python 3:

* + Python is a high-level programming language that is easy to learn and use.
  + Python is an interpreted language, which means that code can be executed on the fly, without the need for compilation.
  + Python is open source and free to use, with a large and active community of developers contributing to its development and maintenance.
  + Python has a vast collection of third-party libraries and packages, such as NumPy, Pandas, Matplotlib, and Scikit-learn, among others, that make it easy to perform data analysis.

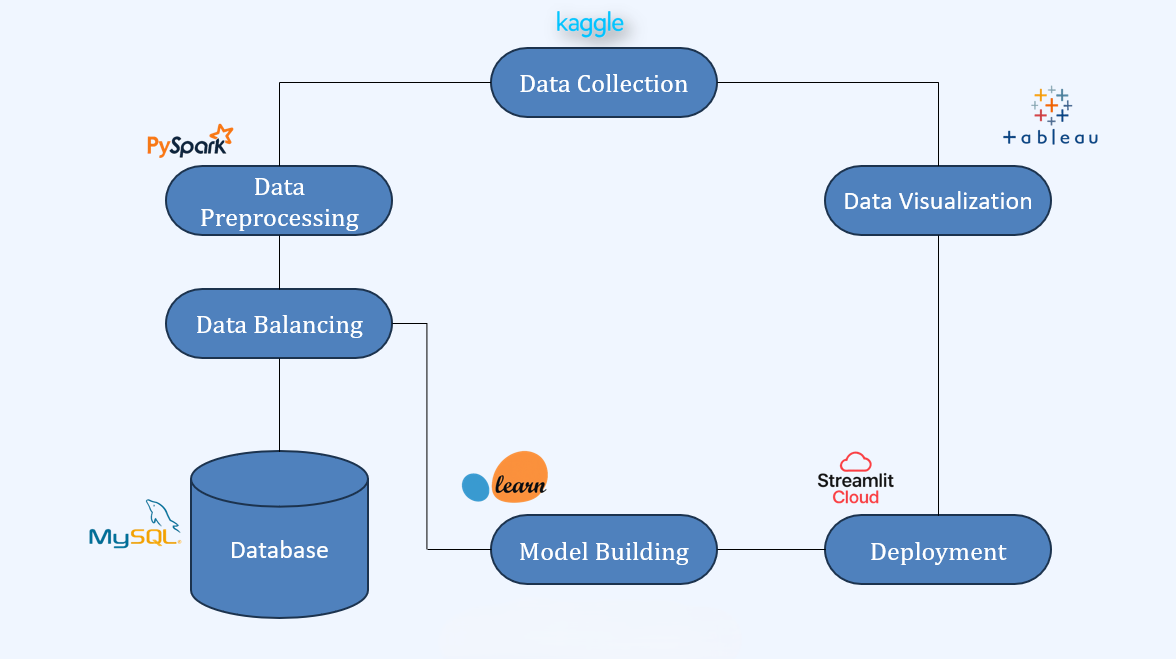
#### Tableau:

* Tableau is a data visualization and business intelligence software that allows users to connect, analyse, and share data in a visual and interactive way.
* It offers a user-friendly drag-and-drop interface that enables users to create interactive dashboards, reports, and charts without the need for complex coding or programming.
* Tableau supports various data sources, including spreadsheets, databases, cloud services, and bigdata platforms, such as Hadoop and Spark.

1. **Data Cleaning:**

* Data cleaning is a crucial step in the data mining process and significantly impacts model development. Despite its importance, it is often overlooked. Ensuring data quality is essential for effective information management, and data cleaning addresses various quality issues.
* Without proper data cleaning, analysis and modeling can produce erroneous or biased results, leading to serious consequences for businesses and organizations. Thus, it is a vital part of data preparation that influences the accuracy and reliability of insights and decisions.
* By improving data quality, organizations can gain a clearer understanding of their operations, customers, and market trends, enabling more informed and effective decision-making.

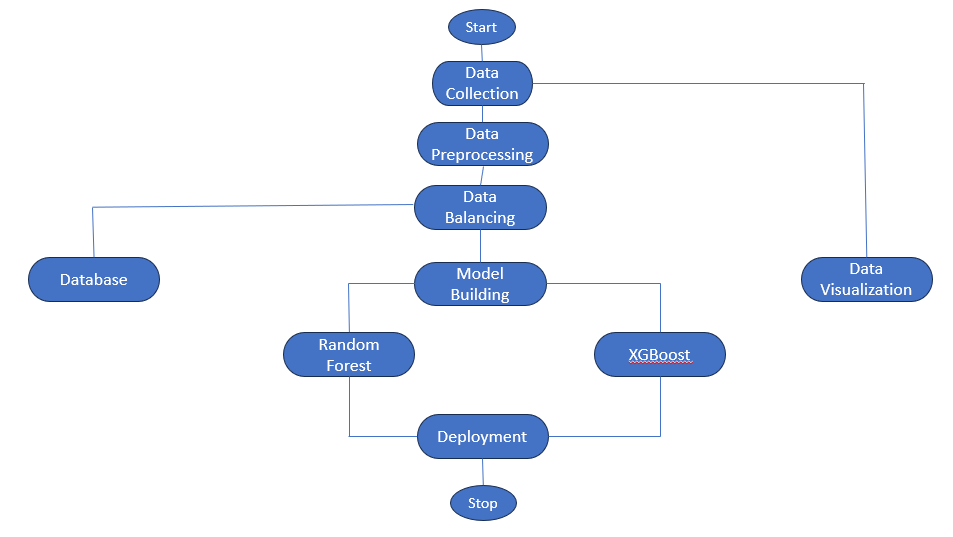
**SYSTEM ARCHITECTURE**



**Fig: System Architecture of Credit Card Fraud Detection**

# 

# METHODOLOGY

****

**Fig: Methodology of Credit Card Fraud Detection**

# MACHINE LEARNING ALGORITHMS

* Machine learning is a subfield of artificial intelligence that involves developing algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed. The goal of machine learning is to enable computers to improve their performance over time by learning from experience and feedback.
* In our project, we applied various Classification Algorithms such as Random Forest, CatBoost, XGBoost and Logistic Regression. After the implementation, we analyzed the accuracy of all the algorithms on our data.

.

1. **Random Forest –**

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions for classification tasks or the mean prediction for regression tasks. It enhances predictive accuracy and controls overfitting.

**Pros:**

* Random Forest is robust to overfitting, especially with large datasets, due to its ensemble

nature.

* It can handle a mix of numerical and categorical features and is effective for high-dimensional data.
* It provides feature importance scores, helping to identify the most influential variables.

**Cons:**

* The model can be complex and less interpretable compared to simpler models like decision

trees.

* It may require more computational resources and time for training, especially with large

datasets.

* Random Forest can struggle with imbalanced datasets, leading to biased predictions.

1. **XGBoost**

XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting framework designed for

speed and performance. It builds models in a sequential manner, where each new model corrects the errors of the previous ones.

**Pros:**

* XGBoost is highly efficient and can handle large datasets with high dimensionality.
* It often provides superior predictive performance compared to other algorithms due to its
* regularization techniques.
* It includes built-in cross-validation, making it easier to tune hyperparameters.

**Cons:**

* The model can be complex and may require careful tuning of hyperparameters to achieve optimal performance.
* It can be sensitive to noisy data and outliers, which may affect the model's accuracy.
* XGBoost may require more computational resources compared to simpler models.

## DATA VISUALIZATION AND REPRESENTATION

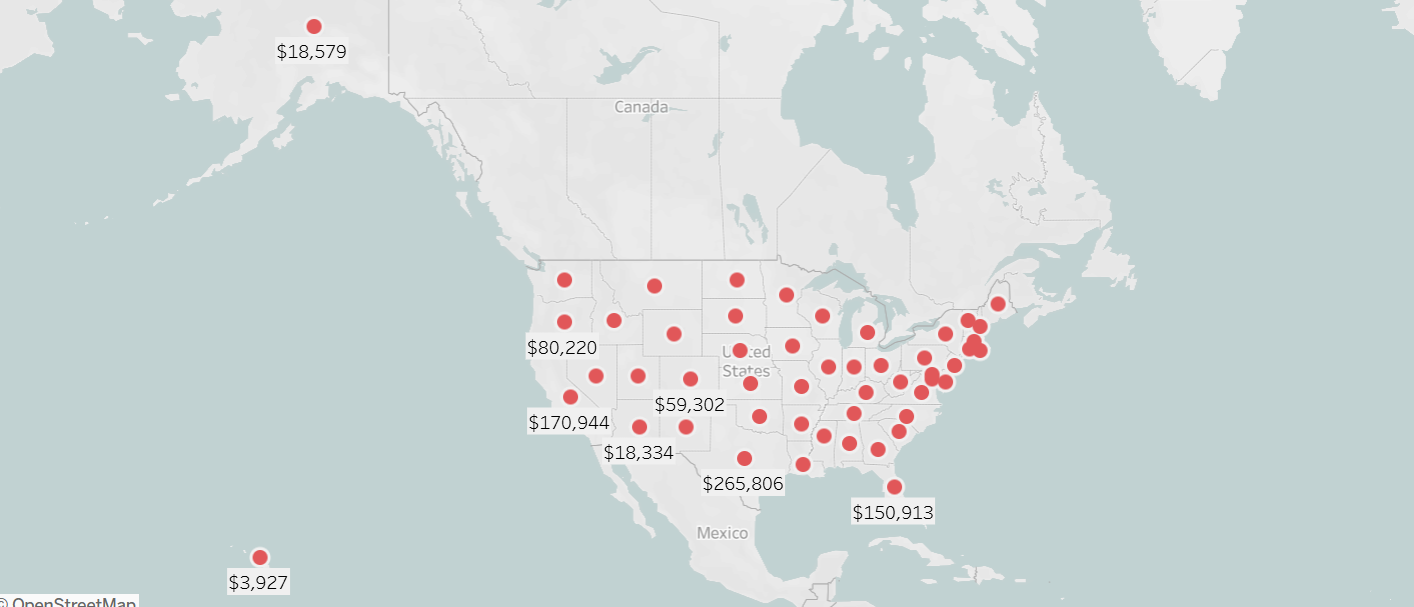


Fig: Statewise Fraud count

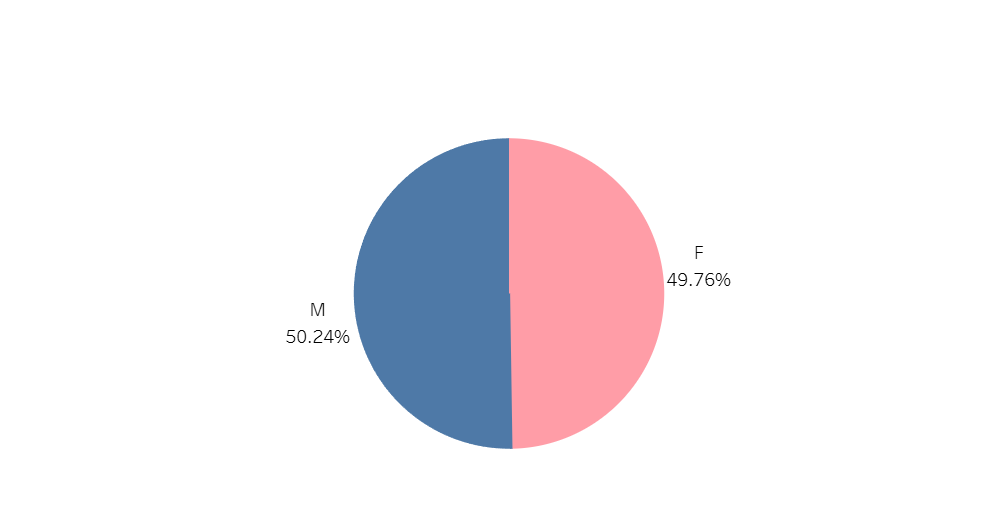


Fig: Fraud % by Provider Gender

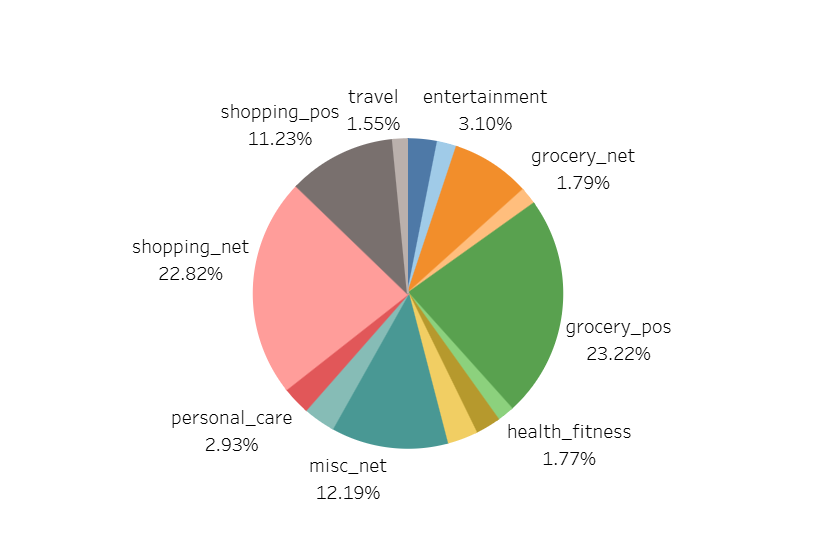


Fig: Fraud % by Provider Category

# 

# CONCLUSION AND FUTURE SCOPE

* The Card Card Fraud Detection System has shown promising results, with the XGBoost model achieving an accuracy of 99% and an AUC of 0.99.
* This strong performance underscores the effectiveness of the model in identifying fraudulent credit card transactions. The success of the system can be attributed to a combination of robust data preprocessing in PySpark, effective feature engineering, and meticulous hyperparameter tuning, all of which have enhanced the model’s predictive capabilities.
* The insights derived from the analysis not only empower banks and financial institutions to make informed decisions but also provide a strategic framework for mitigating fraudulent activities in real time.
* By identifying suspicious spending patterns and anomalies in transaction behavior, the system can help organizations implement proactive measures to reduce the risk of fraud, ultimately leading to reduced financial losses and improved customer trust.
* Moreover, the successful implementation of the Card Card Fraud system can serve as a model for other sectors facing similar challenges with fraud detection
* The methodologies and best practices developed through this project can be adapted and applied to various industries, thereby broadening the impact of the work done.

Future work on this study could focus on several areas to improve the accuracy of the predictions even further. Some potential areas for future research include:

# Integration of Advanced Algorithms: Exploring other machine learning algorithms, such as XGBoost and CatBoost, could further improve detection rates and model performance. Ensemble methods that combine multiple algorithms may also yield better results.

# Real-Time Fraud Detection: Developing capabilities for real-time monitoring and detection of fraudulent activities can significantly enhance the system's responsiveness. Implementing streaming data analysis could allow for immediate alerts and interventions.

# Incorporation of Deep Learning: Investigating deep learning techniques, such as neural networks, could provide additional layers of complexity and improve the model's ability to capture intricate patterns in the data.

# Continuous Learning and Adaptation: Implementing a feedback loop that allows the model to learn from new data and adapt to evolving fraud tactics will ensure its long-term effectiveness. Regular updates and retraining of the model will be essential.

# User -Friendly Dashboard: Developing an intuitive dashboard for stakeholders to visualize insights and trends in real-time can enhance decision-making processes. This tool could provide actionable recommendations based on the model's predictions.

# Collaboration with Regulatory Bodies: Partnering with regulatory agencies and other stakeholders in the healthcare sector can facilitate data sharing and improve the overall effectiveness of fraud detection efforts.

# Focus on Ethical AI: Ensuring that the model is transparent and fair is crucial. Implementing measures to mitigate bias and enhance interpretability will build trust among users and stakeholders.