Internship Project on AI/ML & Data Science Technologies

**Project Name:** *Fraud Detection on Bank Payments*

### Students Names

1. *M.Srija - 22WH1A05B7*
2. *CH.D.P.Bhoomika - 22WH1A05B3*
3. *V.Aayushri - 22WH1A1251*
4. *M.Likhita Sree - 22WH1A1252*
5. *N.Keerthi - 22WH1A1260*

**Guide Name:** *Ms.Vidyullatha*

Table of Contents

* 1. [Introduction to Machine Learning 4](#_TOC_250040)
     1. [What is Machine Learning? 4](#_TOC_250039)
     2. [Importance of Machine Learning in different Industry Sector 4](#_TOC_250038)
     3. [Basic Concepts and Terminology 4](#_TOC_250037)
  2. [Problem Definition and Scope 4](#_TOC_250036)
     1. [Identification of Business Problem 4](#_TOC_250035)
     2. [Scope and Boundaries of the Project 4](#_TOC_250034)
     3. [Data Sources and Availability 5](#_TOC_250033)
     4. [Success Criteria 5](#_TOC_250032)
  3. [Data Collection and Preparation 6](#_TOC_250031)
     1. [Gathering Data 6](#_TOC_250030)
     2. [Data Types and Formats 6](#_TOC_250029)
     3. [Exploring and Understanding the Data 6](#_TOC_250028)
  4. [Data Preprocessing 6](#_TOC_250027)
     1. [Data Cleaning 6](#_TOC_250026)
     2. [Handling Missing Values 6](#_TOC_250025)
     3. [Feature Selection and Engineering 6](#_TOC_250024)
     4. [Data Transformation and Scaling 6](#_TOC_250023)
  5. [Model Selection and Development 7](#_TOC_250022)
     1. [Introduction to Machine Learning Algorithms 7](#_TOC_250021)
     2. [Choosing the Right Algorithm 7](#_TOC_250020)
     3. [Splitting Data into Training and Testing Sets 7](#_TOC_250019)
     4. [Training the Model 7](#_TOC_250018)
  6. [Model Evaluation and Validation 8](#_TOC_250017)
     1. [Performance Metrics 8](#_TOC_250016)
     2. [Cross-Validation Techniques 8](#_TOC_250015)
     3. [Interpreting Model Results 8](#_TOC_250014)
  7. [Model Deployment 8](#_TOC_250013)
     1. [Exporting the Trained Model 8](#_TOC_250012)
     2. [Integrating the Model into GITHUB 8](#_TOC_250011)
  8. [Documentation and Reporting 9](#_TOC_250010)
     1. [Model Documentation 9](#_TOC_250009)
     2. [Technical Reports 9](#_TOC_250008)
  9. [Project Conclusion 9](#_TOC_250007)
     1. [Summary of Key Findings 9](#_TOC_250006)
     2. [Lessons Learned 9](#_TOC_250005)
     3. [Recommendations for Future Work 10](#_TOC_250004)

[Appendix A - Glossary of Terms 10](#_TOC_250003)

[Appendix B - References 10](#_TOC_250002)

[Appendix C - Data Dictionary 10](#_TOC_250001)

[Appendix D - Code Repositories and Documentation 11](#_TOC_250000)

# Introduction to Machine Learning

Machine learning involves creating algorithms and statistical models that enable computers to perform tasks without explicit instructions. It plays a crucial role in various industries by extracting insights from data and making predictions or decisions.

## What is Machine Learning?

Machine learning refers to the process of training algorithms to learn patterns and relationships from data, allowing computers to make predictions or decisions without being explicitly programmed.

## Importance of Machine Learning in different Industry Sector

Machine learning is essential in various industry sectors, including finance, healthcare, retail, and cybersecurity, where it helps in fraud detection, personalized medicine, recommendation systems, and threat detection.

## Basic Concepts and Terminology

Basic concepts in machine learning include supervised learning, unsupervised learning, and reinforcement learning, along with terms like features, labels, training, testing, and evaluation metrics.

# Problem Definition and Scope

## Identification of Business Problem

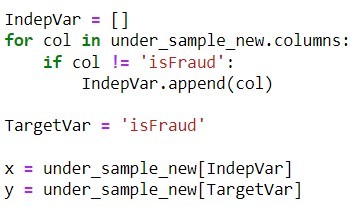
The project focuses on fraud detection in bank payments, aiming to develop models that can accurately identify fraudulent transactions to prevent financial losses.

## Scope and Boundaries of the Project

The scope includes collecting and preprocessing transaction data, selecting appropriate machine learning algorithms, training and evaluating models, and deploying the best-performing model for fraud detection.

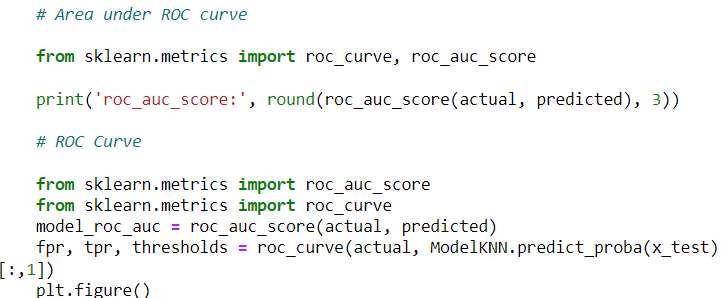
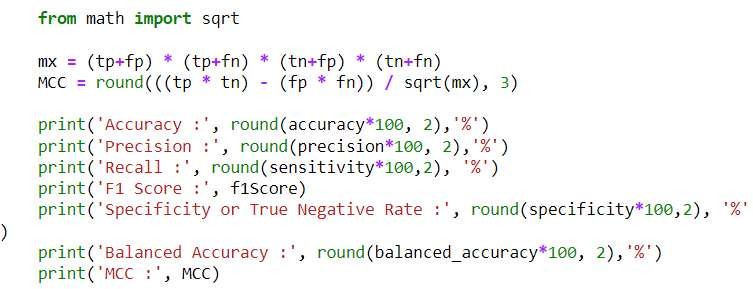
## Data Sources and Availability

Transaction data from a bank's payment system (kaggle) is used for this project. The dataset contains features such as transaction amount, type, step, and target variable indicating fraudulent or non-fraudulent transactions.



## Success Criteria

The success of the project will be evaluated based on the model's performance metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve.



# Data Collection and Preparation

## Gathering Data

Transaction data is collected from the bank's payment system and stored in a structured format for further analysis.

## Data Types and Formats

The data consists of numerical and categorical variables, including transaction amount, type of transaction.

## Exploring and Understanding the Data

Exploratory data analysis is conducted to gain insights into the distribution of features, identify patterns, and detect any anomalies or missing values.

# Data Preprocessing

## Data Cleaning

Data cleaning involves handling missing values, removing duplicates, and addressing any inconsistencies in the dataset.

## Handling Missing Values

Missing values are imputed or removed using appropriate techniques to ensure data quality. There are no missing values in the data set.

## Feature Selection and Engineering

Relevant features are selected or engineered to improve model performance and reduce dimensionality.

## Data Transformation and Scaling

Data transformation techniques such as normalization or standardization are applied to scale features and improve model convergence.

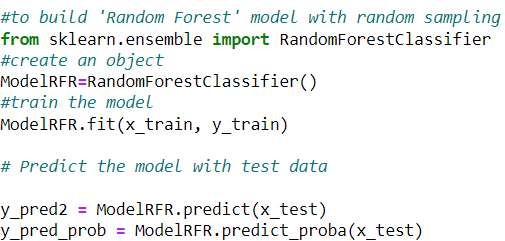
# Model Selection and Development

## Introduction to Machine Learning Algorithms

Various machine learning algorithms such as K-Nearest Neighbors (KNN), Decision Trees, Random Forest, Extra Trees, and Support Vector Machines (SVM) are considered for fraud detection.

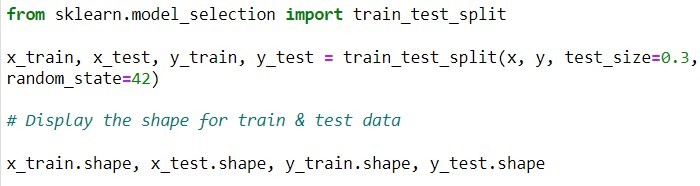
## Choosing the Right Algorithm

The most suitable algorithm is selected based on the nature of the problem, data characteristics, accuracy, and performance evaluation. we have chosen a Random Forest algorithm.



## Splitting Data into Training and Testing Sets

The dataset is divided into training and testing sets to train the model on a subset of data and evaluate its performance on unseen data. Splitting data into training and testing sets in the ratio 7:3.



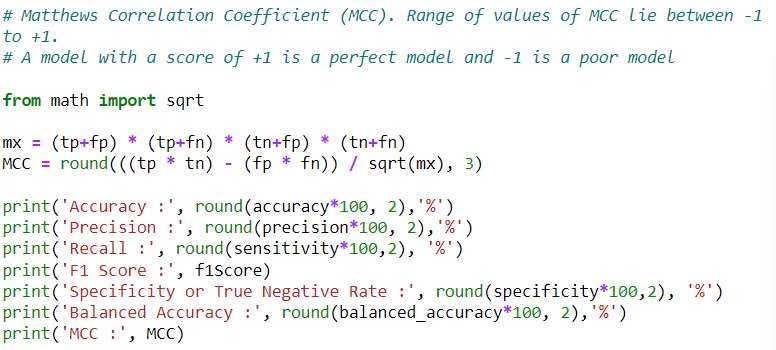
## Training the Model

Random forest algorithm is trained on the training dataset using appropriate parameters.

# Model Evaluation and Validation

## Performance Metrics

Performance metrics such as accuracy, precision, recall, F1-score, specificity, balanced accuracy, and Matthews Correlation Coefficient (MCC) are calculated to evaluate the model's performance.



## Cross-Validation Techniques

Cross-validation techniques such as k-fold cross-validation are used to validate the model and assess its robustness.

## Interpreting Model Results

The model results are interpreted to understand its strengths, weaknesses, and areas for improvement.

# Model Deployment

## Exporting the Trained Model

The trained model is exported and saved for deployment in production environments.

## Integrating the Model into GITHUB

First, create a new repository on GitHub where you'll host your fraud detection project. Link your local repository to the GitHub repository you created earlier.

Use <https://github.com/LikhitaSree52/Bank-Transaction-Fraud-Detection-System> to add the remote

repository, then push your changes using the master branch).

**git push -u origin master**

# Documentation and Reporting

(Assuming you're working

## Model Documentation

Fraud detection is a critical task in various industries, including finance, healthcare, and e-commerce. This document presents the documentation for our fraud detection model, designed to identify fraudulent activities and transactions.

## Technical Reports

This report provides a comprehensive overview of fraud detection techniques, including rule-based systems, anomaly detection, and machine learning approaches. It explores the strengths, weaknesses, and applications of each method in different fraud detection scenarios.

# Project Conclusion

Our fraud detection project represents a significant advancement in detecting and mitigating fraudulent activities within transaction data. By leveraging innovative technologies and methodologies, we have developed a robust and scalable solution capable of effectively combating fraud while maintaining transparency and accountability. As we continue to evolve and adapt to emerging threats, our commitment to safeguarding financial integrity remains unwavering.

## Summary of Key Findings

Importance of Feature Engineering, Model Performance Metrics, Class Imbalance Challenges, Interpretability and Explainability, Continuous Learning and Adaptability, Integration of Domain

Knowledge, Ethical and Regulatory Considerations, Collaboration and Information Sharing.

## Lessons Learned

We have learned Importance of Data Quality, Continuous Model Evaluation, Interpretability and Transparency, Adaptability to Emerging Threats, Education and Awareness,

## Recommendations for Future Work

Exploration of Advanced Techniques, Ensemble Learning and Model Stacking, Explainable AI and Model Interpretability, Continuous Learning and Adaptive Systems, Integration of Domain Knowledge and Expert Systems.

# Appendix A - Glossary of Terms

Fraud Detection, Machine Learning, Feature Engineering, Anomaly Detection, Class Imbalance, Precision, Explainable AI, Continuous Learning, Ethical Considerations.

# Appendix B - References

* + 1. Bank Payment Fraud Dataset\*, Kaggle. Retrieved from https://[www.kaggle.com/code/shankeshrajums/fraudulent-transaction-model](http://www.kaggle.com/code/shankeshrajums/fraudulent-transaction-model)
    2. Pedregosa, F., et al. (2011). \*Scikit-learn: Machine Learning in Python\*. Journal of Machine Learning Research, 12, 2825–2830.
    3. Breiman, L. (2001). \*Random Forests\*. Machine Learning, 45(1), 5-32.
    4. A Comprehensive Guide to Data Preprocessing\* retrieved from https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning a9fa83a5dc9d

# Appendix C - Data Dictionary

1. **TransactionID:** A unique identifier assigned to each transaction in the dataset. This is a definite feature used to differentiate between individual transactions.
2. **Amount:** This represents the amount of money involved in the transaction. It is a numerical feature and is a critical variable in identifying potentially fraudulent activities.
3. **Type:** Indicates the type of transaction, such as transfer, withdrawal, or deposit. This is a categorical feature that helps in distinguishing the nature of each transaction.
4. **Time:** Refers to the specific time at which the transaction occurred. This numerical feature is

often used in identifying patterns over time, such as detecting fraud that occurs at unusual hours.

1. **AccountBalance:** The balance remaining in the account after the transaction is completed. This numerical feature can provide insights into whether a transaction is suspicious based on the account's behavior.
2. **IsFraud:** A binary categorical feature that indicates whether the transaction was identified as fraudulent (1) or non-fraudulent (0). This is the target variable that the machine learning model aims to predict.

# Appendix D - Code Repositories and Documentation

**GitHub Repository:** Fraud-Detection-on-bank-payments

**Link:** https://github.com/LikhitaSree52/Bank-Transaction-Fraud-Detection-System

**Description:** This repository contains the full codebase for the Fraud Detection project, including data preprocessing scripts, model training, evaluation notebooks, and deployment scripts.

### Documentation:

**README:** The repository includes a detailed README file outlining the setup instructions, dependencies, and usage of the project code.

**Notebooks:** Jupyter notebooks that document the exploratory data analysis, feature engineering, and model development processes.

**Scripts:** Python scripts used for preprocessing data, training models, and deploying the final model.