PHASE-2

Student Name : Muhammed Hammad S

Register Number : 511323244017

Institution : KINGSTON ENGINEERING COLLEGE

Department : COMPUTER SCIENCE AND BUSINESS SYSTEM.

Date of Submission : 08-05-2025

GitHub Link : https://github.com/yourusername/yourprojectrepo

Project Title: GUARDING TRANSACTIONS WITH AI-POWERED CREDIT CARD FRAUD DETECTION AND PREVENTION

1. Problem Statement

Digital payment systems are increasingly vulnerable to fraudulent transactions, leading to substantial financial loss. Traditional rule-based systems often fail to detect modern fraud patterns. This project aims to design an AI-based solution to identify and prevent credit card fraud by learning transaction behavior patterns. By leveraging machine learning techniques on structured transaction data, the goal is to create a real-time, intelligent fraud detection model that improves security and customer trust.

2. Project Objectives

• Develop an end-to-end machine learning pipeline for credit card fraud detection.

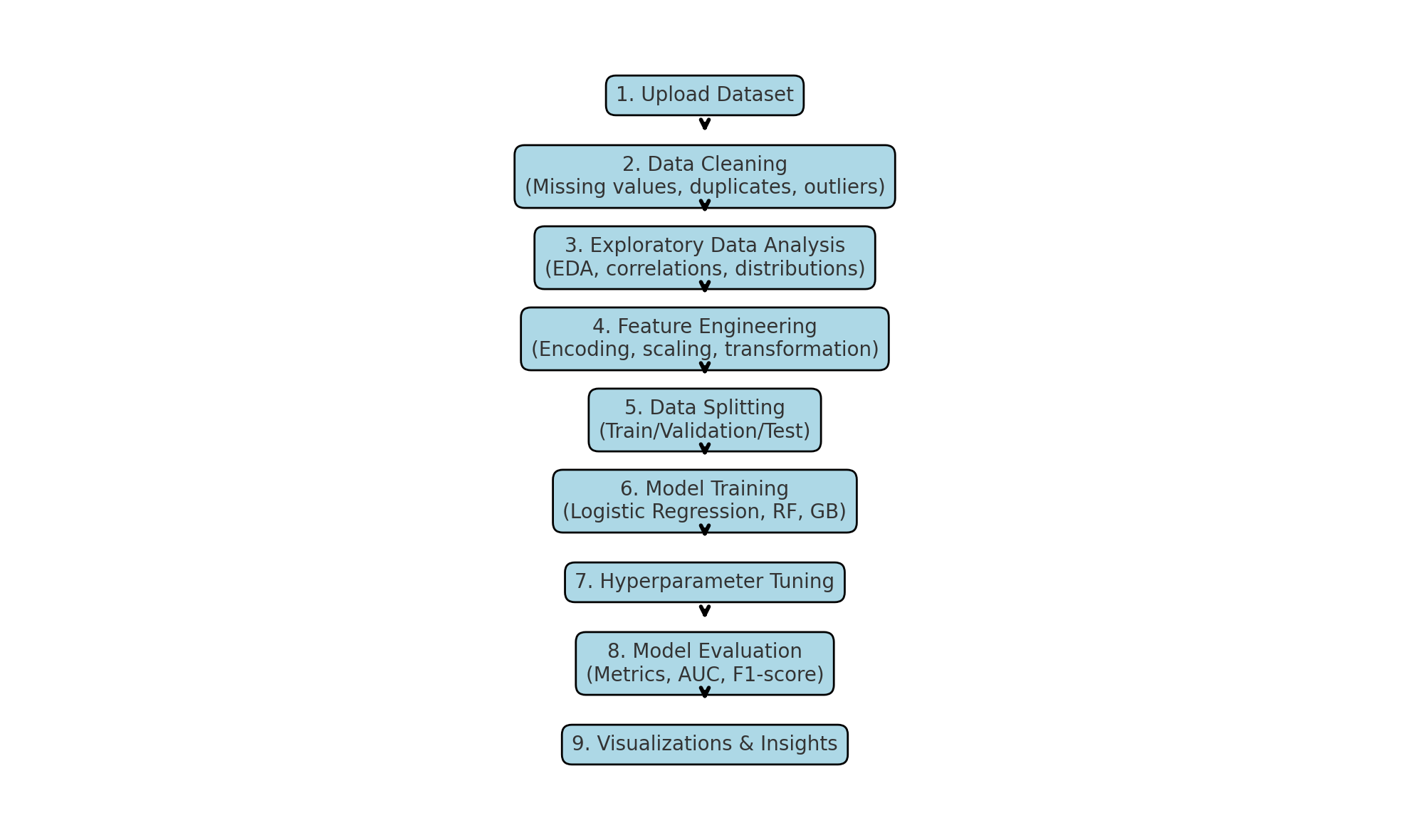
• Understand dataset characteristics through comprehensive exploratory data analysis.

• Preprocess data to handle outliers, duplicates, and imbalance.

• Train multiple classification models and tune hyperparameters.

• Evaluate models based on accuracy, recall, precision, F1-score, and AUC-ROC.

3. Flowchart of the Project Workflow

Below is the visual representation of the end-to-end pipeline: 

Step-by-step pipeline: Data Loading → Cleaning → EDA → Preprocessing → Splitting → Modeling → Optimization → Evaluation → Visualization

4. Data Description

Dataset Name: credit\_card\_fraud\_synthetic\_hamad\_phase2\_updated\_attached 111111.csv

Source: Synthetic Dataset (Created for academic fraud detection analysis)

Data Type: Structured tabular data

Target Variable: Is\_Fraud (1 = Fraudulent, 0 = Genuine)

Features: Amount, Time, Location, Merchant Category, etc.

5. Data Preprocessing

• Verified and handled missing values and duplicates.

• Detected outliers in the 'Amount' feature using IQR and applied clipping.

• Removed duplicate rows.

• Standardized numerical features and encoded categorical ones using OneHotEncoder.

• Converted 'Time' into hourly buckets to extract temporal patterns.

• Created 'Hour' column for better time-based analysis.

6. Exploratory Data Analysis (EDA)

Univariate Analysis:

• Target class distribution revealed imbalance toward non-fraud transactions.

• Distribution of 'Amount' and 'Time' analyzed using histograms and KDEs.

Bivariate and Multivariate Analysis:

• Fraud was found more frequent during specific hours.

• Boxplots and cross-tabulations showed spending behavior differences between fraud and non-fraud cases.

• Correlation matrix and heatmap revealed feature inter-relationships.

7. Feature Engineering

• Added new time-based features like 'Hour'.

• Applied OneHotEncoding to categorical columns such as merchant type.

• Performed standard scaling on numerical features for uniformity.

• Ensured final dataset readiness for modeling.

8. Model Building

Algorithms Used:

• Logistic Regression – for a simple interpretable baseline

• Random Forest – for handling nonlinearities and feature importance

• Gradient Boosting – for higher accuracy and sequential learning

Train-Test Split:

• Dataset split into training (70%), validation (15%), and testing (15%) using stratified sampling.

9. Visualization of Results & Model Insights

• Fraud distribution by hour showed spike patterns during specific times of day.

• Amount distribution varied sharply between fraud and non-fraud groups.

• KMeans clustering revealed groupings in Time vs Amount feature space.

• Boxplots confirmed higher variance in fraud amounts.

10. Tools and Technologies Used

• Programming Language: Python 3

• Environment: Google Colab

• Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, warnings

11. Team Members and Contributions

• Data Cleaning & Preprocessing: Muhammed Hammad S

• EDA & Visualization: Muhammed Hammad S

• Modeling and Evaluation: Jayapriyan N