# Fraud Detection Project Documentation

## Table of Contents

1. Overview  
2. Datasets  
3. Data Preprocessing  
4. Exploratory Data Analysis (EDA)  
5. Models  
 - Supervised Learning  
 - Unsupervised Learning  
6. Evaluation Metrics  
7. Results and Insights  
8. Screenshots  
9. Conclusion

## Overview

This project aims to detect fraudulent financial transactions using machine learning techniques.  
It leverages supervised learning for classification tasks and unsupervised learning for anomaly detection.  
The workflow includes:  
- Data analysis and visualization to uncover patterns and insights.  
- Feature engineering to enhance model performance.  
- Implementation of machine learning models to predict and identify fraudulent transactions.

## Datasets

**Files**

* **FraudTrain.csv:** Contains labeled training data.
* **FraudTest.csv:** Contains labeled testing data.

**Features**

* **Transaction Details:** Information like `amt` (transaction amount), `unix\_time`, `lat`/`long` coordinates, and merchant details (`merch\_lat`, `merch\_long`).
* **Fraud Label:** The `fraud` column indicates whether a transaction is fraudulent (`1`) or not (`0`).

**Data Overview**

* **Total number of records:** ~400,000 transactions.
* **Label distribution:**  - Fraudulent: ~1.5%.  
   - Non-fraudulent: ~98.5%.

## Data Preprocessing

**Steps:**

1. \*\*Datetime Conversion\*\*:  
 - Converted `unix\_time` into `year`, `month`, `day`, `hour`, and `day\_of\_week`.  
  
2. \*\*Geographical Calculations\*\*:  
 - Computed the distance between customer (`lat`, `long`) and merchant (`merch\_lat`, `merch\_long`) using the Haversine formula.  
  
3. \*\*Logarithmic Transformation\*\*:  
 - Applied to `amt` (transaction amount) to normalize its distribution.  
  
4. \*\*Encoding and Scaling\*\*:  
 - Encoded categorical features using `LabelEncoder`.  
 - Scaled numerical features with `StandardScaler`.  
  
5. \*\*Feature Creation\*\*:  
 - \*\*High-Value Transactions\*\*: Flagged transactions exceeding a certain threshold.

## Exploratory Data Analysis (EDA)

EDA provides insights into the data and identifies important patterns that influence fraud detection.  
  
 **Key Findings**  
- Fraudulent transactions are significantly fewer than non-fraudulent ones.  
- High-value transactions show a higher likelihood of fraud.  
- Fraudulent transactions are often concentrated at specific times of the day.  
  
**Visualizations**  
**1. \*\*Fraud Distribution (Pie Chart)\*\*:**  
 - Shows the percentage split between fraudulent and non-fraudulent transactions.

A green circle with a bar and a bar graph

Description automatically generated  
  
**2. \*\*Correlation Heatmap\*\*:**  
 - Highlights relationships between features.A screenshot of a graph

Description automatically generated

**3. \*\*Transaction Trends\*\*:**  
 - Scatter plots to analyze the relationship between transaction amounts and geographical distances.

A map of the united states

Description automatically generated

## Models

**Supervised Learning**  
1. \*\*Algorithm\*\*:  
 - Random Forest Classifier.  
   
2. \*\*Process\*\*:  
 - Used the full dataset, with the `fraud` column as the target variable.  
 - Split data into training (80%) and testing (20%) subsets.  
  
3. \*\*Metrics\*\*:  
 - Accuracy, F1 Score, Precision, Recall, Confusion Matrix.  
  
**Unsupervised Learning**  
1. \*\*Approach\*\*:  
 - Removed the `fraud` column and used clustering techniques for anomaly detection.  
   
2. \*\*Goal\*\*:  
 - Detect potential fraud without labeled data.

## Evaluation Metrics

- \*\*Accuracy\*\*: Measures the overall correctness of predictions.  
- \*\*Precision\*\*: Measures how many predicted frauds are actual frauds.  
- \*\*Recall\*\*: Measures how many actual frauds were correctly identified.  
- \*\*F1 Score\*\*: Balances precision and recall.  
- \*\*Confusion Matrix\*\*: Shows true positives, false positives, true negatives, and false negatives.

## Results and Insights

**Supervised Learning**  
- \*\*Accuracy\*\*: ~99%.  
- \*\*Precision\*\*: High due to imbalanced data handling.  
- \*\*Recall\*\*: Captures most fraudulent transactions.  
  
 **Unsupervised Learning**  
- Identified clusters of suspicious transactions effectively.  
- Useful for flagging anomalies in unlabeled datasets.

## Conclusion

This project demonstrates the application of machine learning in fraud detection.  
While supervised learning excels with labeled data, unsupervised techniques prove valuable for anomaly detection in unlabeled datasets.  
Future work can explore deep learning models for improved performance.