SHRI VISHNU ENGINEERING COLLEGE FOR WOMEN :: BHIMAVARAM

Department of Computer Science and Engineering - Cyber Security

MINI PROJECT - II

SmartSwipe

BATCH-CS15

Guided By:
Dr.P.Kiran Sree
HOD(CSE Department)

Presented By:

B.H.S.L.R.Anjani- 23B01A4607
P.Chetana Srija - 23B01A4644
V.Rukmini - 23B01A4663

Problem Statement:

- With the rise of digital transactions, credit card fraud has become increasingly common. This project aims to develop a machine learning model to detect fraudulent transactions by training on historical data and evaluating it on unseen data.

Objective:

- The main aim of this project is the detection of fraudulent credit card transactions, as it is essential to figure out the fraudulent transactions so that customers do not get charged for the purchase of products that they did not buy.

Functional Requirements

1. **Load and preprocess data**
2. **Train and evaluate model**
3. **Detect fraud**
4. **Visualize results**
5. **Predict new transactions**

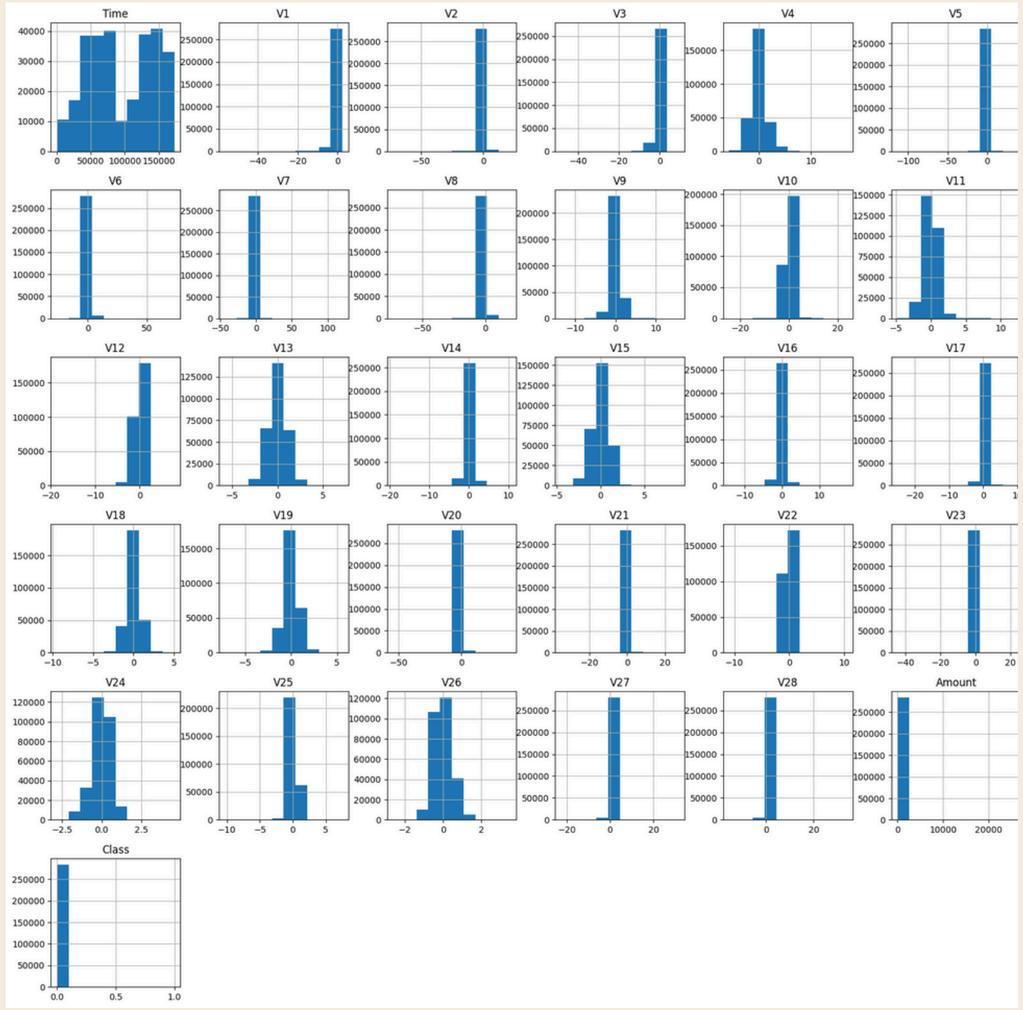
Data Description

The dataset was retrieved from an open-source website, Kaggle.com.
The dataset consists of 31 attributes and 284,808 rows.

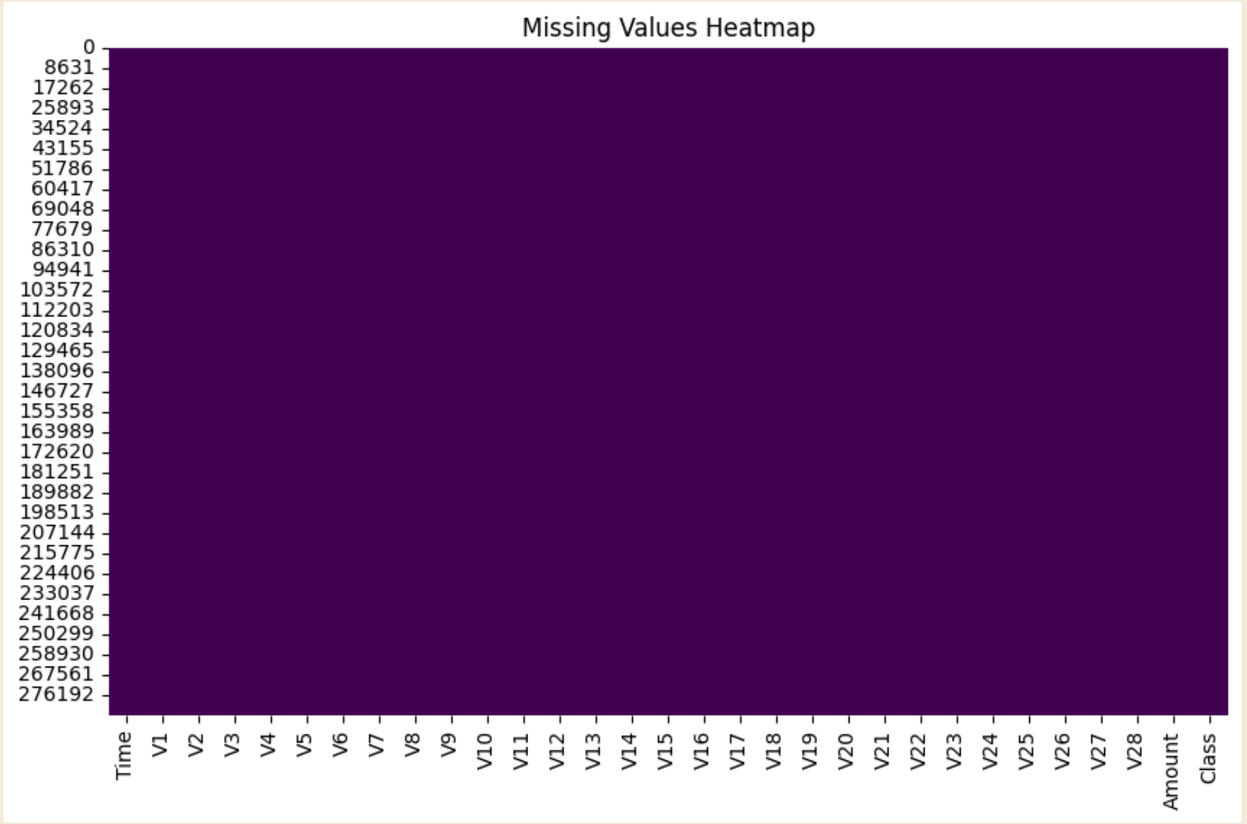
Time: which contains the elapsed seconds between the first and other transactions of each Attribute.
Amount : Which is the amount of each transaction
Class : which contains binary variables where 1 is a case of fraudulent transaction, and 0 is not as case of fraudulent transaction.

Dataset : https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Data Analysis



Histograms



Check for null values

Methodology

1. Preprocessing

- Normalize features (scaling).
- Handle missing values (imputation/removal).

2. Data Preparation

- Apply oversampling (e.g., SMOTE) as data is imbalanced.
- Split data into train-test sets.

3. Model Training

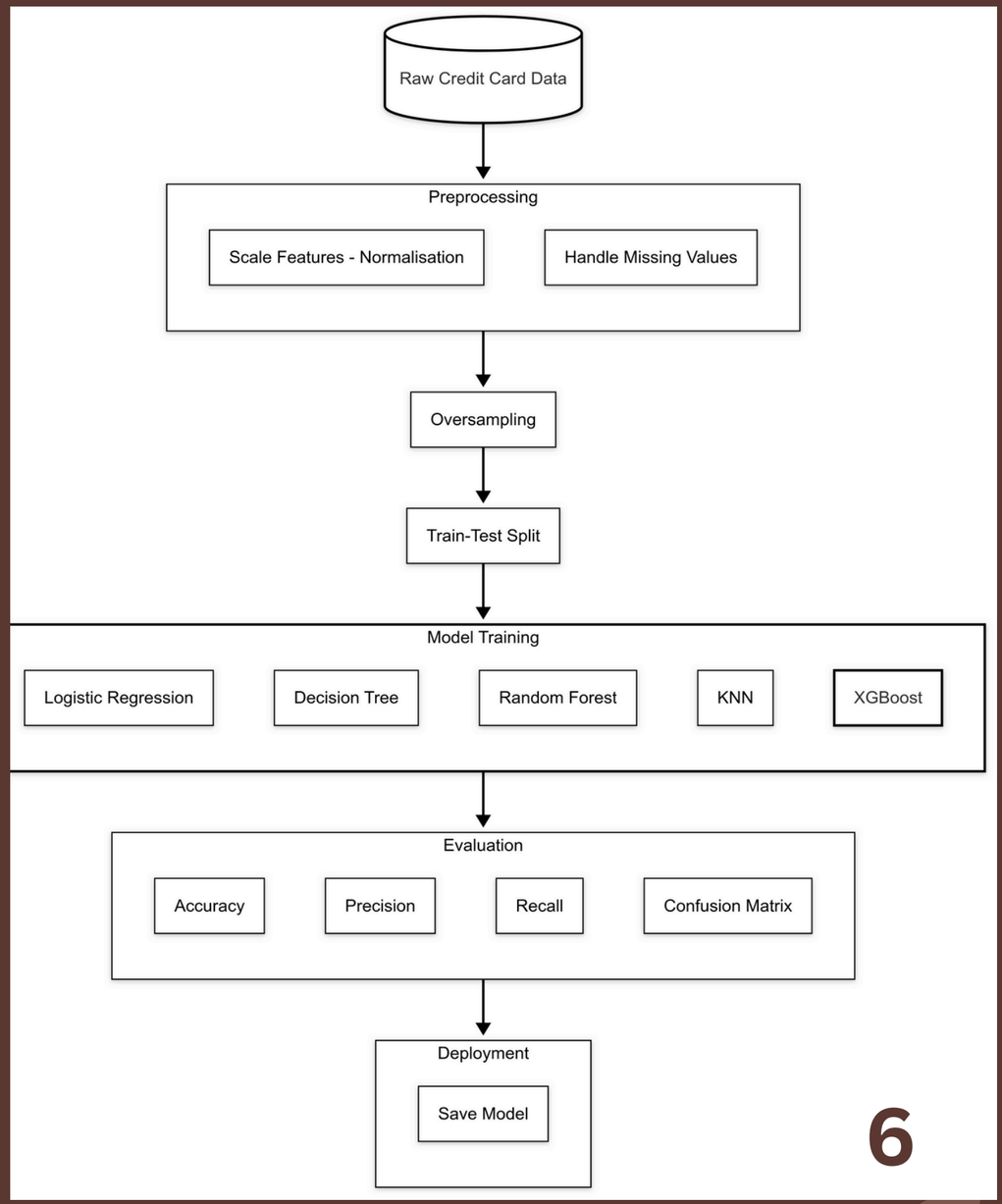
- Logistic Regression
- Decision Tree
- Random Forest
- KNN
- XGBoost

4. Evaluation

- Metrics: Accuracy (1.00), Precision (1.00), Recall (1.00).
- Confusion Matrix analysis.

5. Deployment

- Save the best-performing model for production.

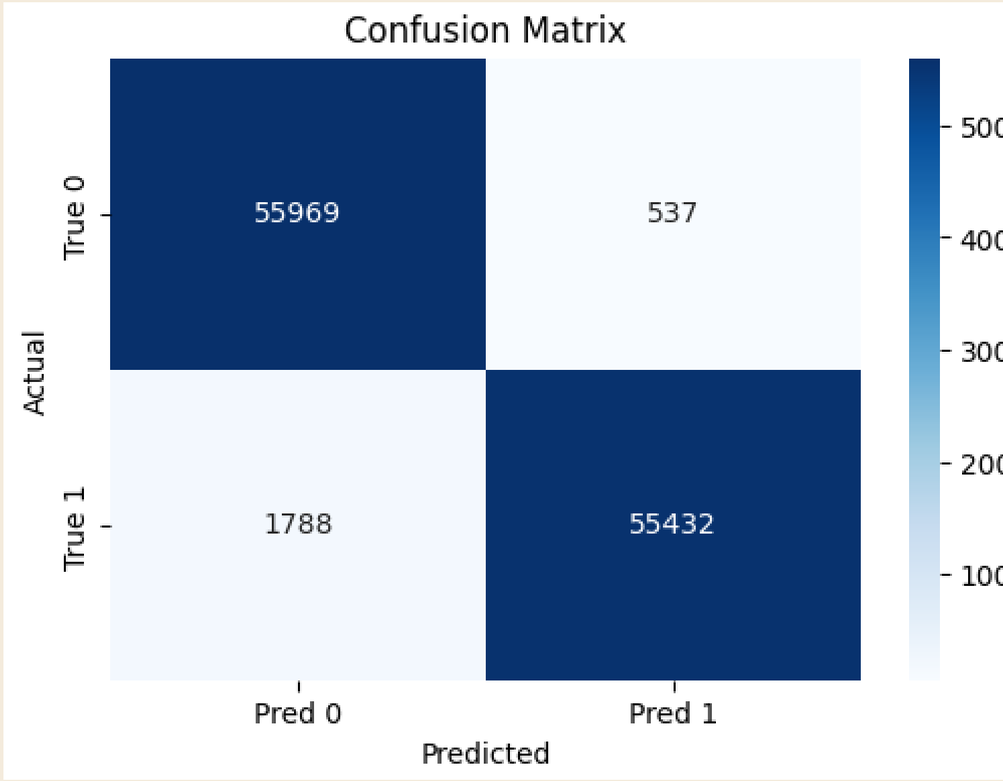


Implementation

LOGISTIC REGRESSION

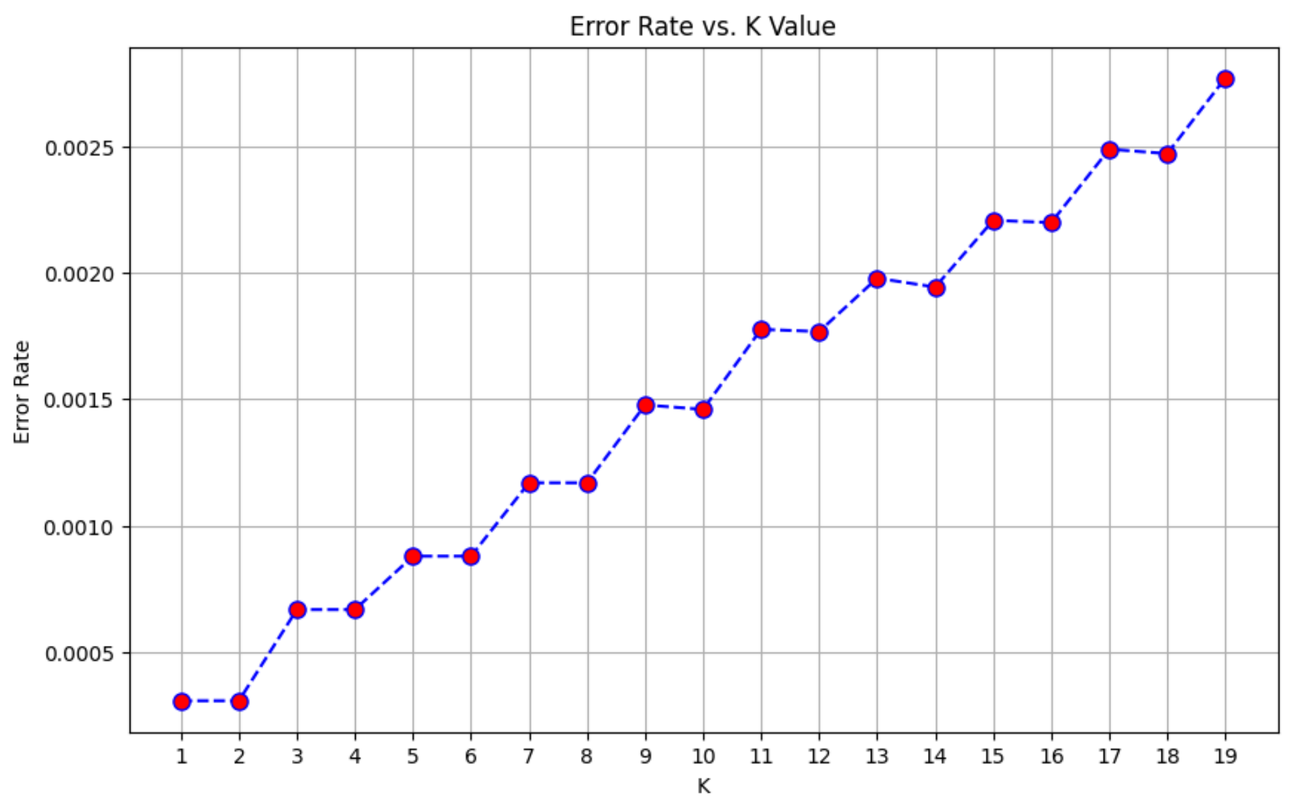
The first model created is Logistic Regression; the model managed to score an Accuracy on Training data of 99.99% , while it scored an Accuracy score on Test Data of 99.95%, as presented in blew Figure.

```
Accuracy on Training data :  0.9799386244413063
Accuracy score on Test Data :  0.9795561261277105
```



KNN

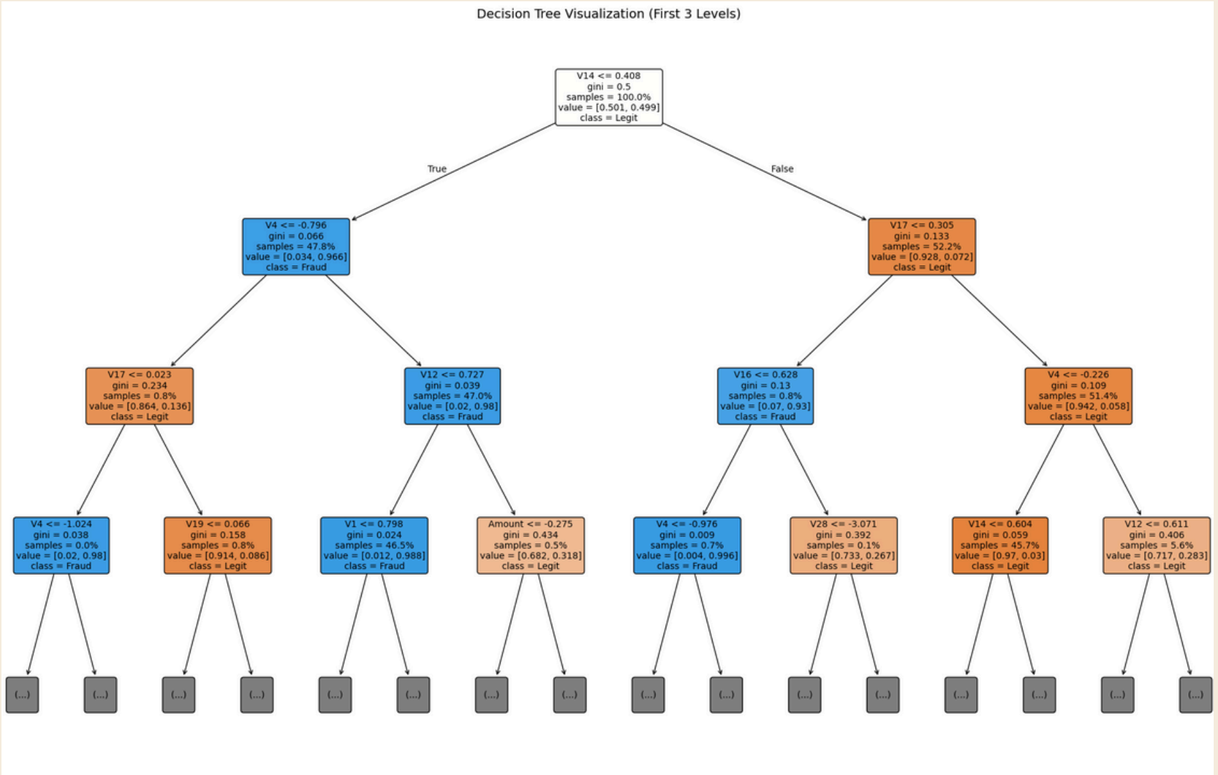
It shows a general upward trend, indicating that the error rate increases as K increases. Lower K values (e.g., K=1 or 2) result in the lowest error rates in this case.




```
====== K-Nearest Neighbors (KNN) ======
Accuracy on traing data: 0.9994658213601112
Accuracy: 0.9991206935968908
Confusion Matrix:
 [[56406   100]
 [    0 57220]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56506
           1       1.00      1.00      1.00     57220

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726
```

Decision Tree

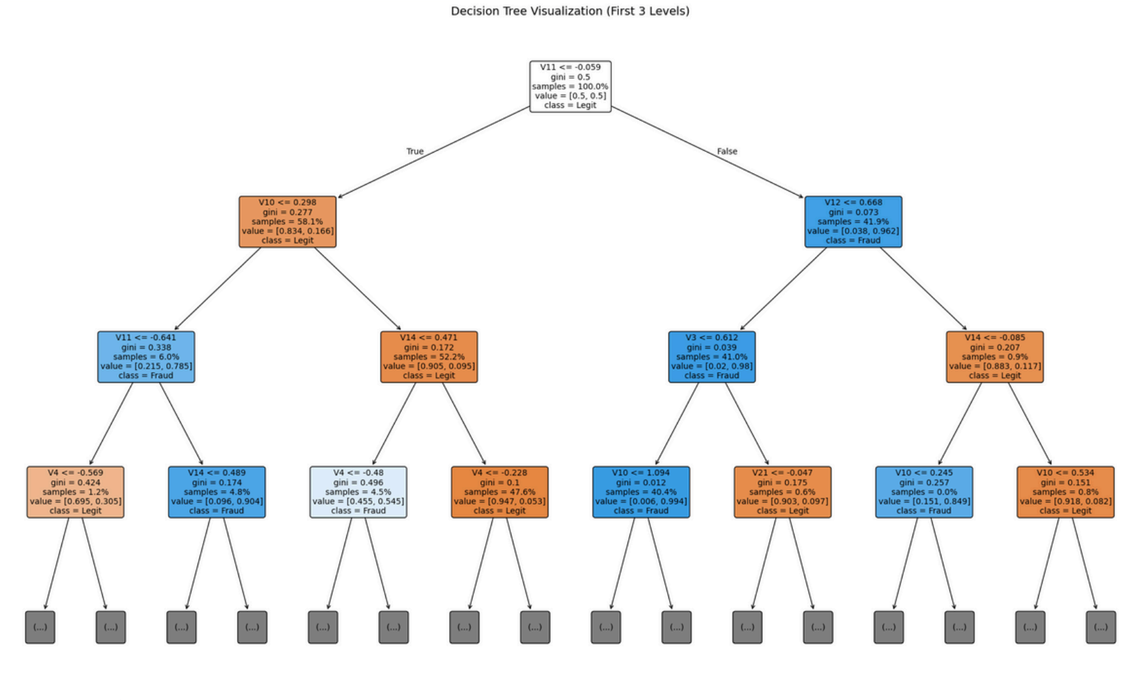
The recall is 0.98 for class 0 and 0.96 for class 1, indicating the model is slightly better at identifying class 0 correctly.

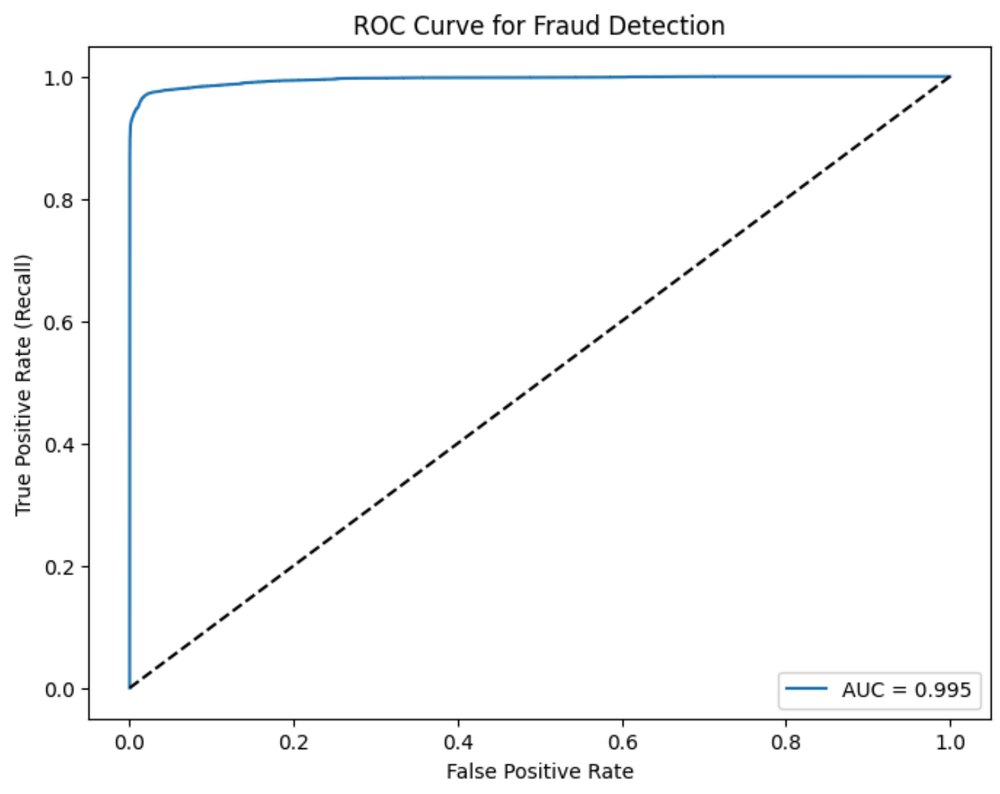



```
Accuracy on Training data :  0.97025306438281 49
Accuracy on Test data 0.969593584580483
Confusion Matrix:
 [[55388  1118]
 [ 2340 54880]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97     56506
           1       0.98      0.96      0.97     57220

    accuracy                           0.97    113726
   macro avg       0.97      0.97      0.97    113726
weighted avg       0.97      0.97      0.97    113726
```

Random Forest

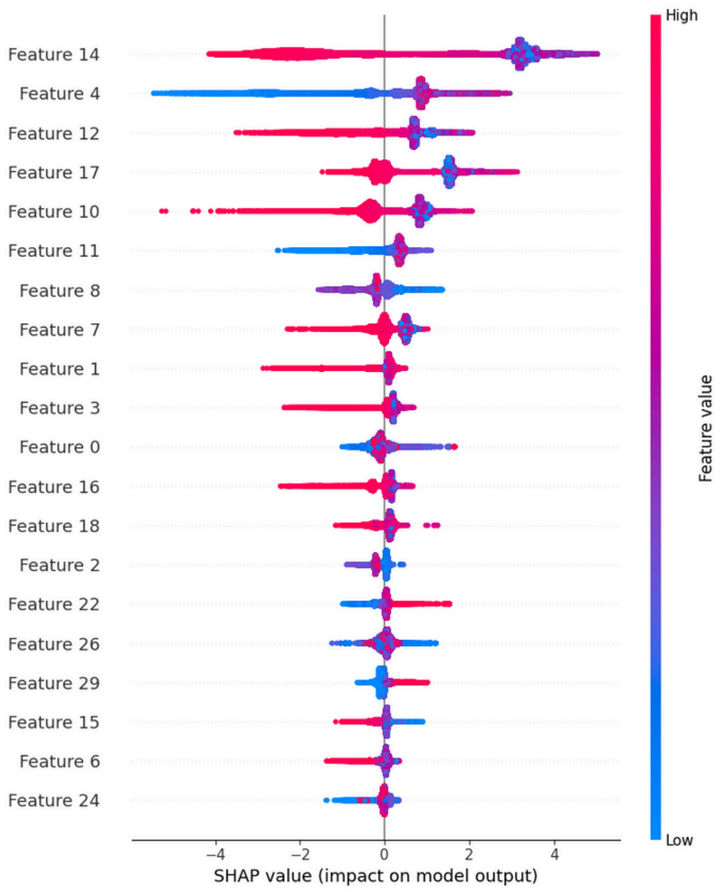
The model shows overall performance with a training accuracy of 96.81% and test accuracy of 96.78%.





XG BOOOST

SHAP stands for Shapley Additive explanations. It is a powerful explainable AI technique that helps you understand how each feature in your dataset impacts the predictions made by a machine learning model



The SHAP summary plot shows that Features 14, 4, and 12 have the highest impact on the model's predictions.

```
Confusion Matrix:
 [[56375   131]
 [  165 57055]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56506
           1       1.00      1.00      1.00     57220

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726
```

Accuracy Table

Model Performance Comparison

Model	Training Accuracy	Test Accuracy	Precision (0/1)	Recall (0/1)	F1-Score (0/1)
XGBoost	0.9978	0.9974	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
Decision Tree	0.9703	0.9696	0.96 / 0.98	0.98 / 0.96	0.97 / 0.97
Random Forest	0.9681	0.9678	0.94 / 0.99	0.99 / 0.94	0.97 / 0.97
KNN	0.9995	0.9991	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
Logistic Regression	**0.9799**	**0.9796**	0.97 / 0.99	0.99 / 0.97	0.98 / 0.98

1. Top Performers
   - KNN & XGBoost: Near-perfect scores (precision=recall=F1=1.00).
     - KNN: Zero missed fraud (FN=0) → Ideal for minimizing fraud risk.
     - XGBoost: Highest true fraud detections (TP=57,055).
2. Logistic Regression
   - Balanced but misses 1,788 fraud cases (recall=97%).
3. Decision Tree
   - Higher false alarms (FP=1,118) than logistic regression.
4. Random Classifier
   - Baseline (accuracy=97%).

Technologies

1. Programming Language

- Python

2. Data Processing & Analysis

- Pandas
- NumPy
- Matplotlib
- Seaborn

3. Machine Learning Libraries

- Scikit-learn
- XGBoost
- Imbalanced-learn

4. Development Environment

- Jupyter Notebook
- Google Colab

Thank You