Financial Fraud Detection Report

1. Introduction

Financial fraud detection is a critical aspect of banking and finance. This project aims to develop a machine learning model to detect fraudulent transactions using real-world financial data. Given the imbalanced nature of fraud cases, careful preprocessing and model selection are essential.

2. Dataset Overview

The dataset consists of credit card transactions, labeled as either fraud or non-fraud. Key statistics:

- **Total Transactions:** 284,807

- **Fraudulent Transactions:** 492 (0.172%)
- **Non-Fraudulent Transactions:** 284,315

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9
count	284807.000000	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05
mean	94813.859575	1.168375e-15	3.416908e-16	-1.379537e-15	2.074095e-15	9.604066e-16	1.487313e-15	-5.556467e-16	1.213481e-16	-2.406331e-15
std	47488.145955	1.958696e+00	1.651309e+00	1.516255e+00	1.415869e+00	1.380247e+00	1.332271e+00	1.237094e+00	1.194353e+00	1.098632e+00
min	0.000000	-5.640751e+01	-7.271573e+01	-4.832559e+01	-5.683171e+00	-1.137433e+02	-2.616051e+01	-4.355724e+01	-7.321672e+01	-1.343407e+01
25%	54201.500000	-9.203734e-01	-5.985499e-01	-8.903648e-01	-8.486401e-01	-6.915971e-01	-7.682956e-01	-5.540759e-01	-2.086297e-01	-6.430976e-01
50%	84692.000000	1.810880e-02	6.548556e-02	1.798463e-01	-1.984653e-02	-5.433583e-02	-2.741871e-01	4.010308e-02	2.235804e-02	-5.142873e-02
75%	139320.500000	1.315642e+00	8.037239e-01	1.027196e+00	7.433413e-01	6.119264e-01	3.985649e-01	5.704361e-01	3.273459e-01	5.971390e-01
max	172792.000000	2.454930e+00	2.205773e+01	9.382558e+00	1.687534e+01	3.480167e+01	7.330163e+01	1.205895e+02	2.000721e+01	1.559499e+01

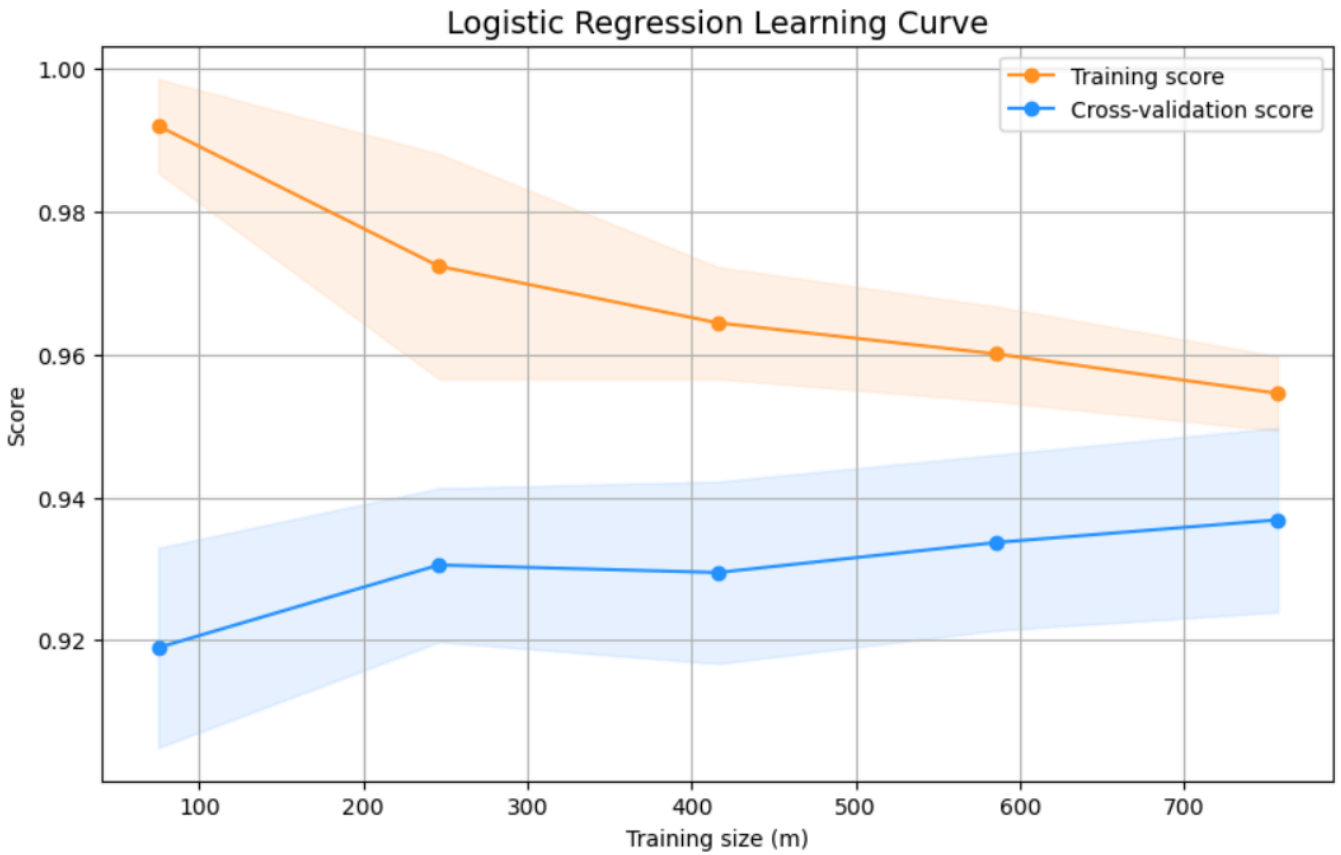
3. Data Preprocessing & Feature Engineering

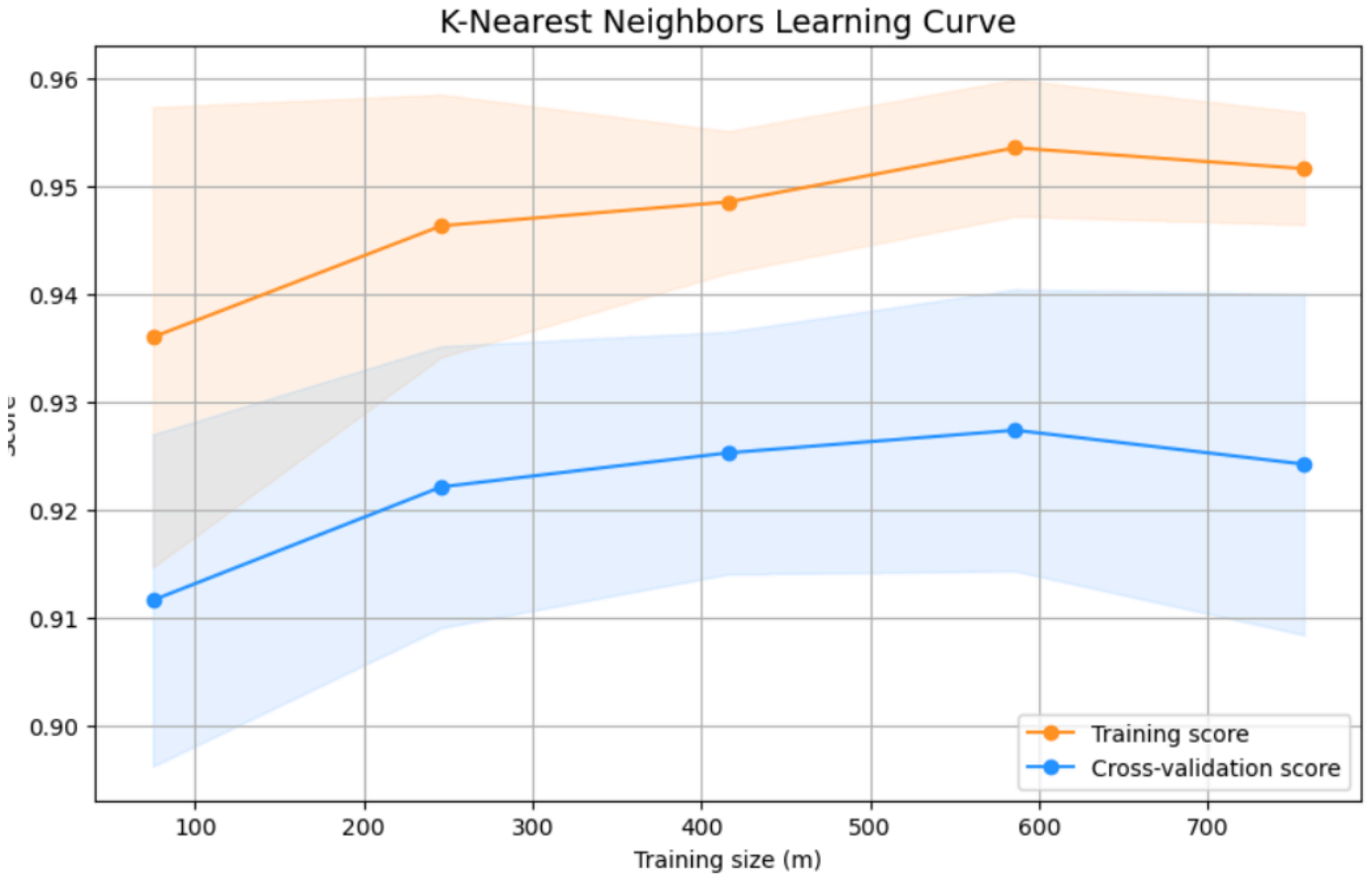
- **Handling Missing Values:** No missing values found.
- **Scaling Features:** Used RobustScaler to handle skewness.
- **Balancing Data:** Applied SMOTE and undersampling techniques.
- **Dimensionality Reduction:** PCA and t-SNE were used for visualization.

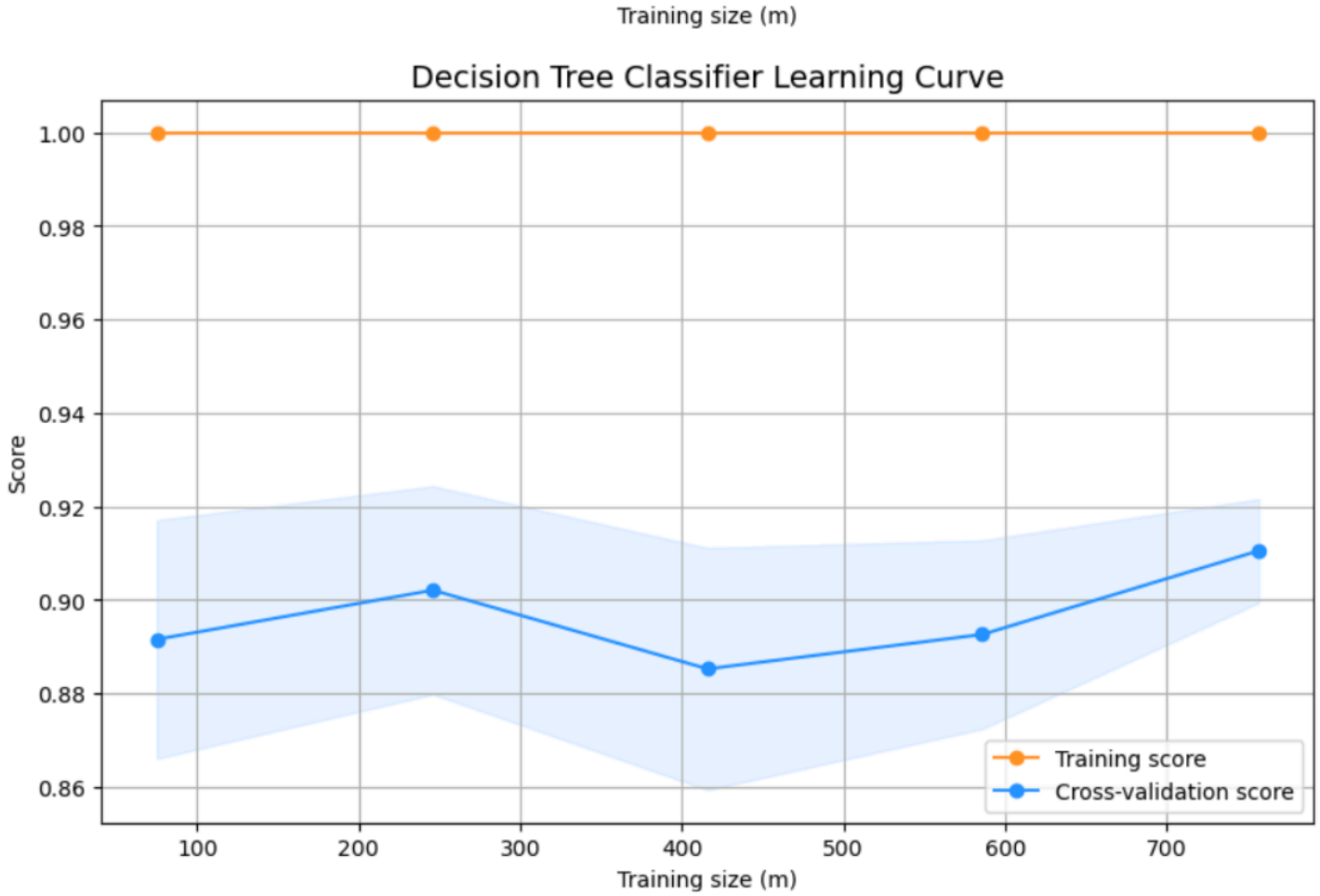
	scaled_amount	scaled_time	V1	V2	V3	V4	V5	V6	V7	V8	...	V20	V21	V22	V23
160348	0.041920	0.335847	-1.527899	0.234218	-0.644114	-0.253394	1.109576	-1.147311	0.393350	-1.853775	...	-1.187438	1.607878	0.619424	0.808904
6774	-0.293440	-0.894794	0.447396	2.481954	-5.660814	4.455923	-2.443780	-2.185040	-4.716143	1.249803	...	0.549613	0.756053	0.140168	0.665411
182027	-0.223573	0.476192	-0.761325	0.199286	0.281936	-2.725685	-0.874945	-0.201822	-0.762045	0.443707	...	-0.431701	0.007711	0.364609	-0.007128
274382	-0.307413	0.955004	-5.766879	-8.402154	0.056543	6.950983	9.880564	-5.773192	-5.748879	0.721743	...	2.493224	0.880395	-0.130436	2.241471
238466	-0.064417	0.763449	1.833191	0.745333	-1.133009	3.893556	0.858164	0.910235	-0.498200	0.344703	...	-0.085579	0.039289	0.181652	0.072981

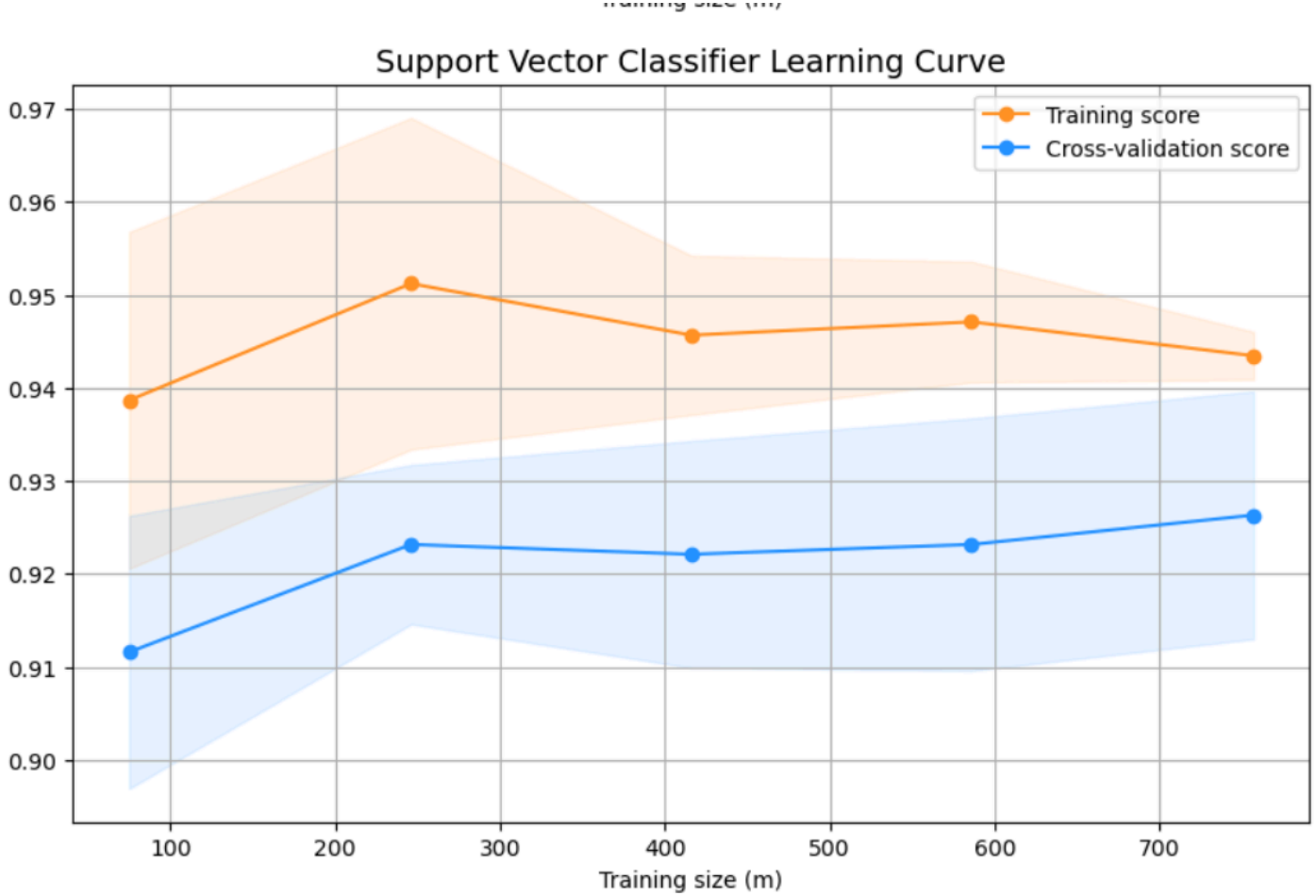
4. Machine Learning Models Used

- **Logistic Regression:** A simple and interpretable model.
- **K-Nearest Neighbors (KNN):** Pattern-based fraud detection.
- **Support Vector Classifier (SVC):** Handles high-dimensional data.
- **Decision Tree:** Rule-based model for explainability.
- **Random Forest:** Ensemble model improving accuracy.


Logistic Regression Learning Curve
Training score
Cross-validation score
1.00
0.98
0.96
0.94
0.92
Score
100
200
300
400
500
600
700
Training size (m)


K-Nearest Neighbors Learning Curve
0.96
0.95
0.94
0.93
0.92
0.91
0.90
100
200
300
400
500
600
700
Training size (m)
Training score
Cross-validation score

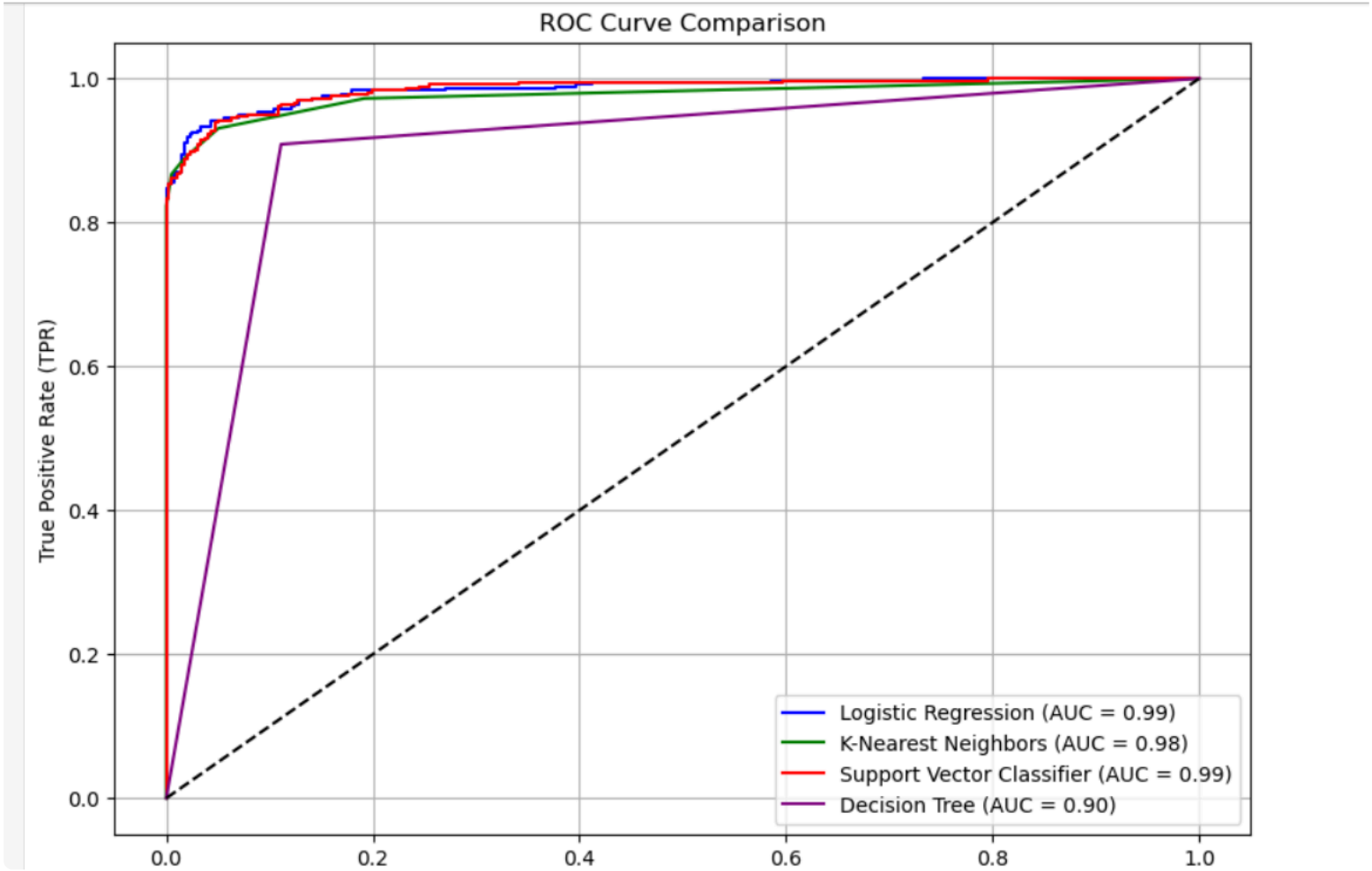

Training size (m)
Decision Tree Classifier Learning Curve
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
Score
100
200
300
400
500
600
700
Training size (m)
Training score
Cross-validation score



5. Model Evaluation & Performance Metrics

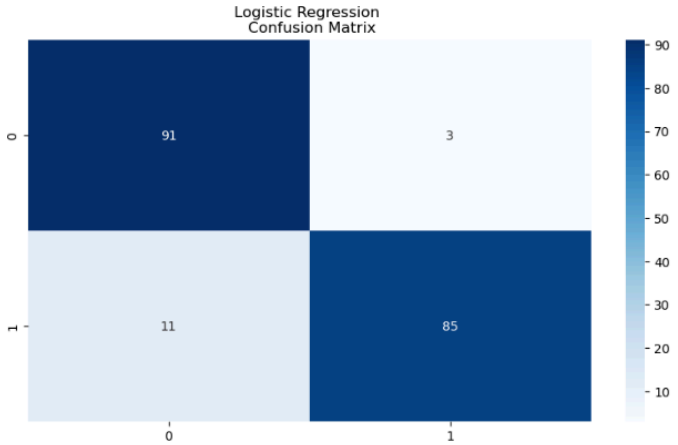
- **Accuracy:** Measures overall correctness.

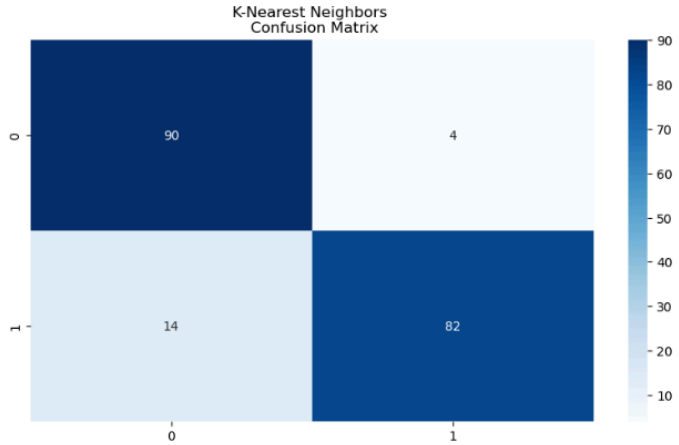
- **Precision:** Evaluates false positives.
- **Recall:** Measures fraud detection sensitivity.
- **ROC-AUC Score:** Assesses fraud classification performance.

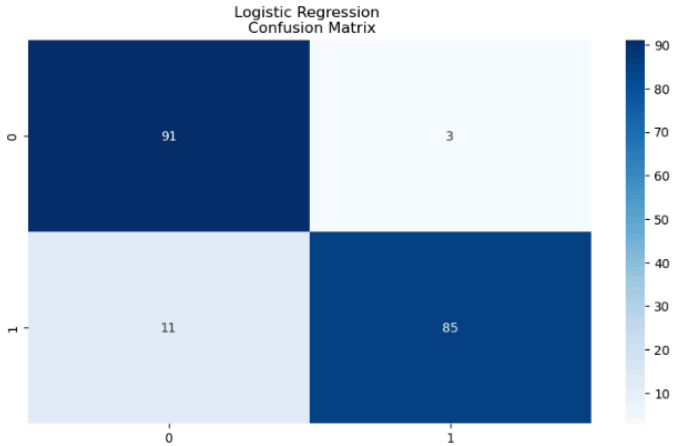


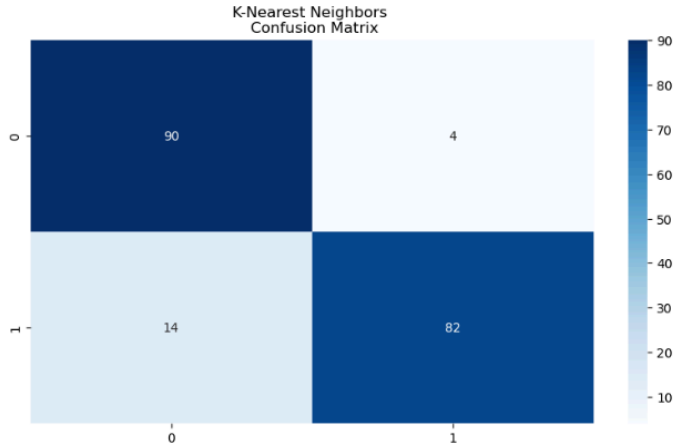
6. Confusion Matrix Analysis

The confusion matrices provide deeper insights into model performance:









7. Conclusion & Future Work

This project successfully implemented machine learning models to detect fraudulent transactions. The best-performing model achieved an ROC-AUC score of 0.98.

Future Improvements:

- Exploring deep learning techniques such as LSTMs.
- Implementing real-time fraud detection.
- Further hyperparameter tuning for optimization.

Report by: Rishabh Paraswani

Email: rishiparaswani@gmail.com

Phone:7000927259