Machine Learning Assignment 2

Name: Sukriti Sukhija
Student ID: 2025ab05062

GitHub Repository Link:

https://github.com/SukritiSukhija/ml_assignment_2.git

Live Streamlit App Link:

https://mlassignment2-gre6jxyroevqdam4qubx6a.streamlit.app/

1. Problem Statement

The objective of this project is to build and compare multiple machine learning classification models to predict whether an individual earns more than $50K per year based on demographic and employment-related features.

2. Model Comparison Table (Test Data)

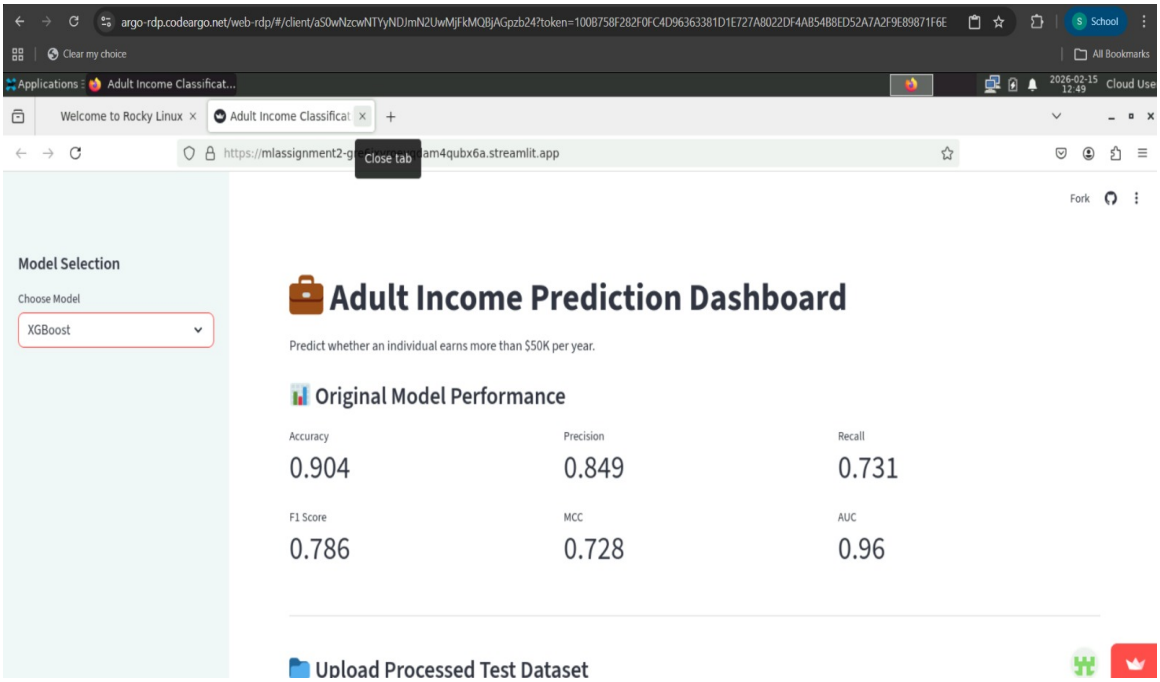
Model	Accuracy	AUC	Precision	Recall	F1	MCC
Logistic Regression	0.81	0.91	0.58	0.85	0.69	0.58
Decision Tree	0.87	0.92	0.79	0.62	0.69	0.62
KNN	0.88	0.94	0.77	0.68	0.72	0.65
Naive Bayes	0.47	0.70	0.31	0.97	0.47	0.28
Random Forest	0.88	0.95	0.85	0.62	0.72	0.66
XGBoost	0.90	0.96	0.85	0.73	0.79	0.73

3. Model-wise Observations

- Logistic Regression: Strong baseline model but limited in capturing non-linear relationships.
- Decision Tree: Captures non-linear patterns but prone to overfitting compared to ensembles.
- KNN: Strong performance but sensitive to feature scaling and dimensionality.
- Naive Bayes: High recall but low precision due to independence assumption limitations.
- Random Forest: Stable and balanced performance due to ensemble averaging.
- XGBoost: Best overall performer capturing complex feature interactions effectively.

4. Application Screenshots

Training Performance (XGBoost):



Test Performance with Confusion Matrix (Logistic Regression):