ML Assignment 2 – Income Classification

a) 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 attributes. This is a binary classification problem where the target variable represents income category (<=50K or >50K).

The project demonstrates an end-to-end machine learning workflow including: - Data preprocessing - Model training - Model evaluation - Performance comparison - Deployment using Streamlit

---

b) Dataset Description

The Adult Income dataset is obtained from the UCI Machine Learning Repository. It contains demographic and employment-related attributes such as age, workclass, education, occupation, marital status, hours-per-week, and others.

- **Number of Instances:** ~48,000+
- **Number of Features:** 14
- **Target Variable:** Income (<=50K or >50K)
- **Problem Type:** Binary Classification

The dataset satisfies the assignment requirements of having more than 12 features and more than 500 instances.

---

c) Models Used and Evaluation Metrics

The following six classification models were implemented on the same dataset:

1. Logistic Regression
2. Decision Tree Classifier
3. K-Nearest Neighbors (kNN)
4. Naive Bayes (Gaussian)

5. Random Forest (Ensemble)
6. XGBoost (Ensemble)

Each model was evaluated using the following metrics:

- Accuracy
- AUC Score
- Precision
- Recall
- F1 Score
- Matthews Correlation Coefficient (MCC)

Comparison Table

ML Model Name	Accuracy	AUC	Precision	Recall	F1	MCC
Logistic Regression	0.814283	0.841913	0.698166	0.441570	0.540984	0.449689
Decision Tree	0.807208	0.742946	0.610080	0.615522	0.612789	0.484452
kNN	0.818925	0.846963	0.654082	0.571811	0.610186	0.494839
Naive Bayes	0.783109	0.823390	0.628676	0.305085	0.410811	0.326139
Random Forest (Ensemble)	0.847225	0.898880	0.733189	0.603033	0.661772	0.568801
XGBoost (Ensemble)	0.865355	0.923469	0.770042	0.651204	0.705655	0.622813

Model Performance Observations

ML Model Name	Observation about model performance
Logistic Regression	Logistic Regression achieved good overall accuracy but relatively low recall. This indicates that while it makes precise predictions for high-income individuals, it misses a significant portion of actual positive cases. Being a linear model, it may not capture complex relationships in the dataset.
Decision Tree	The Decision Tree model shows moderate performance with balanced precision and recall. However, the lower AUC suggests possible overfitting compared to ensemble methods.
kNN	The kNN model performs better than Logistic Regression in terms of F1 and MCC score, indicating that neighborhood-based classification captures nonlinear patterns in the data.
Naive Bayes	Naive Bayes performs the weakest among all models, particularly in recall and F1 score. This suggests that the feature independence assumption does not hold well for this dataset.
Random Forest (Ensemble)	Random Forest significantly improves performance over a single Decision Tree, demonstrating the strength of ensemble learning. It achieves higher AUC, F1, and MCC scores due to reduced variance and improved generalization.
XGBoost (Ensemble)	XGBoost achieves the best overall performance across all evaluation metrics. The boosting mechanism effectively reduces bias and variance, allowing it to model complex feature interactions efficiently. It is the best-performing model for this dataset.

Streamlit Deployment

The trained models were deployed using Streamlit Community Cloud.

The web application includes: - CSV test dataset upload - Model selection dropdown - Display of evaluation metrics - Confusion matrix visualization

Final Notes

- All six models were implemented on the same dataset.

- All required evaluation metrics were calculated.

- The Streamlit app meets all assignment requirements.