Report: Predicting Adult Income (UCI Dataset)

1. Introduction

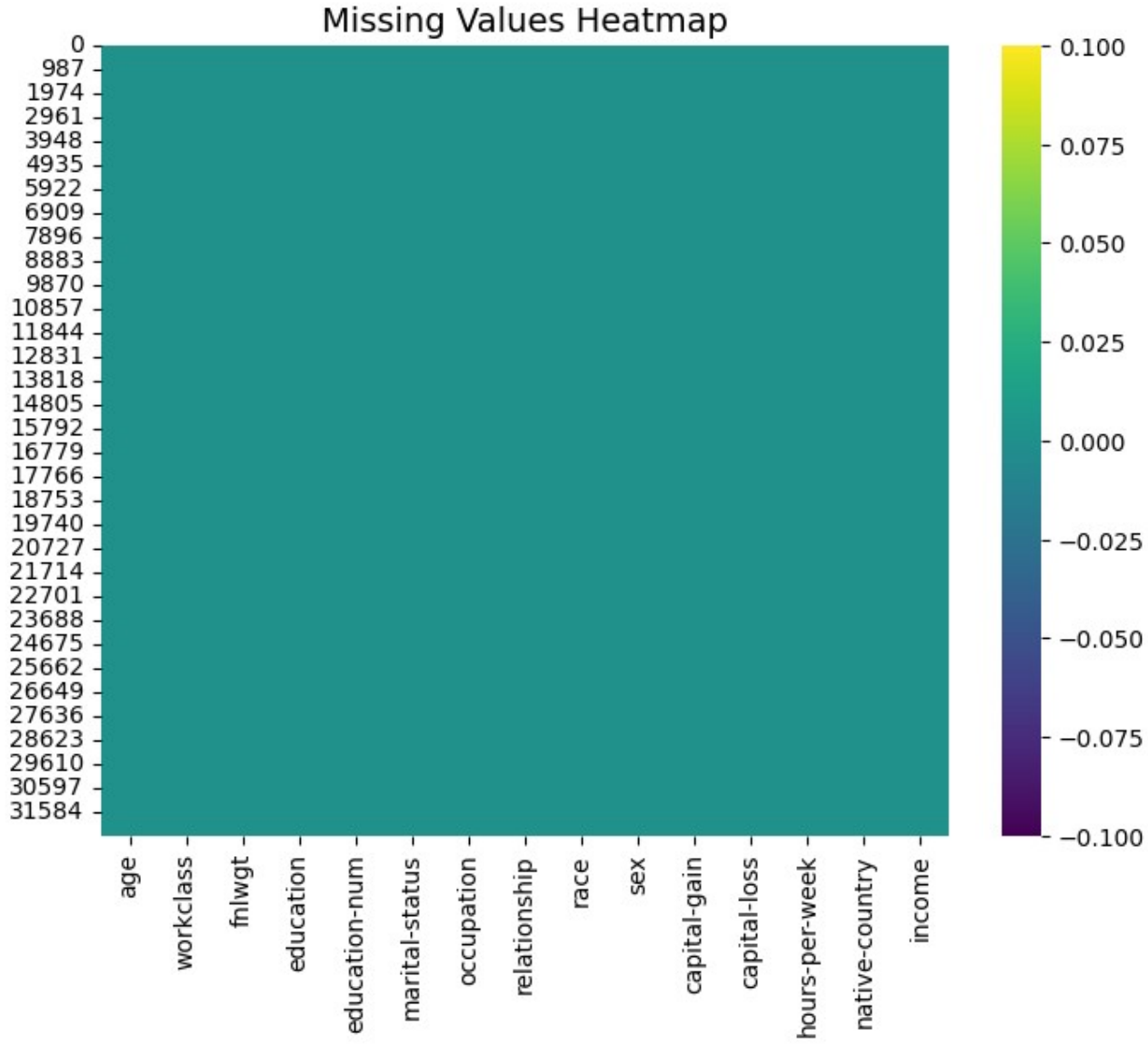
This report summarizes our machine learning pipeline for predicting whether income >50K using the UCI Adult dataset.

2. Dataset

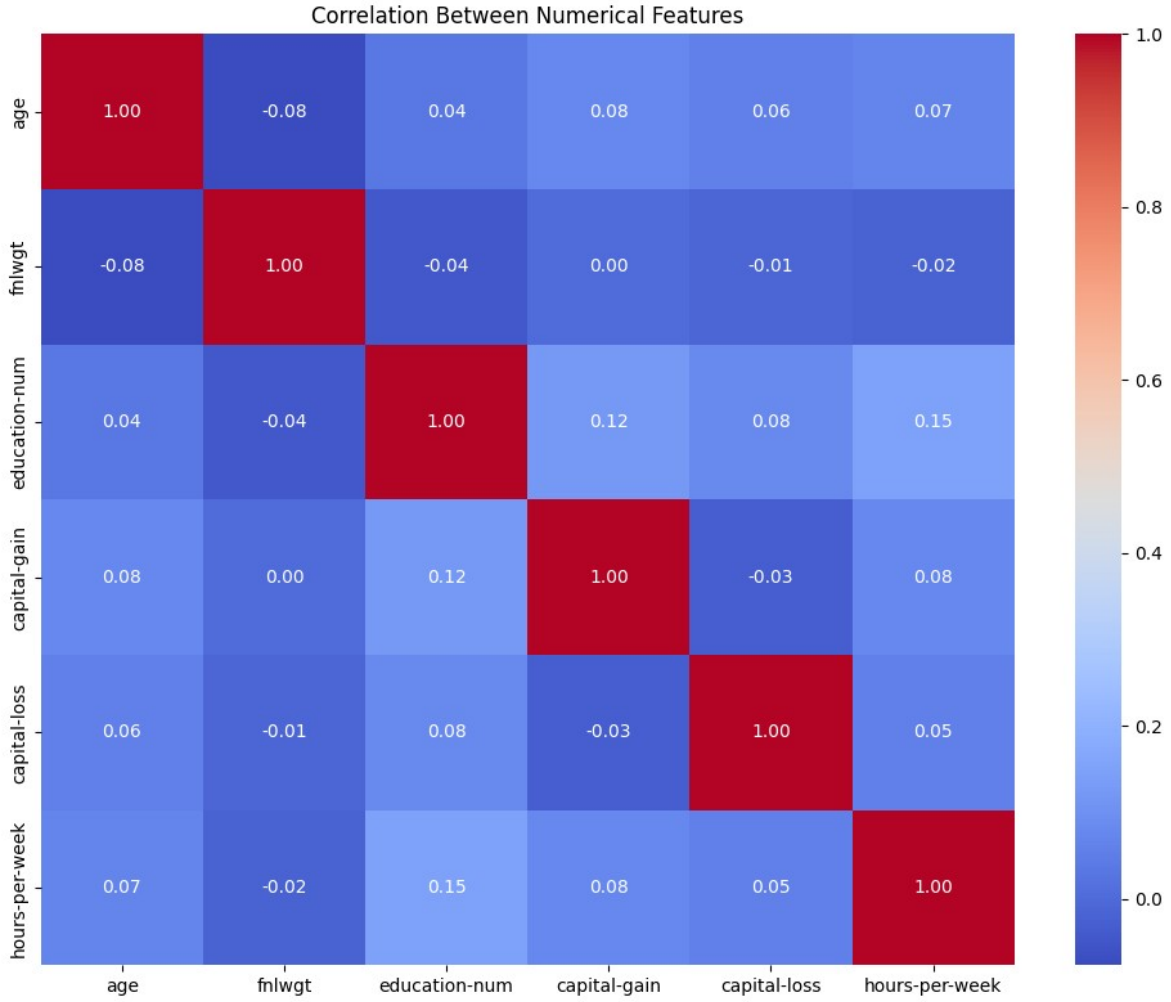
Samples: 32561, Features: 15

3. Exploratory Data Analysis

Missing Values



Correlation Heatmap



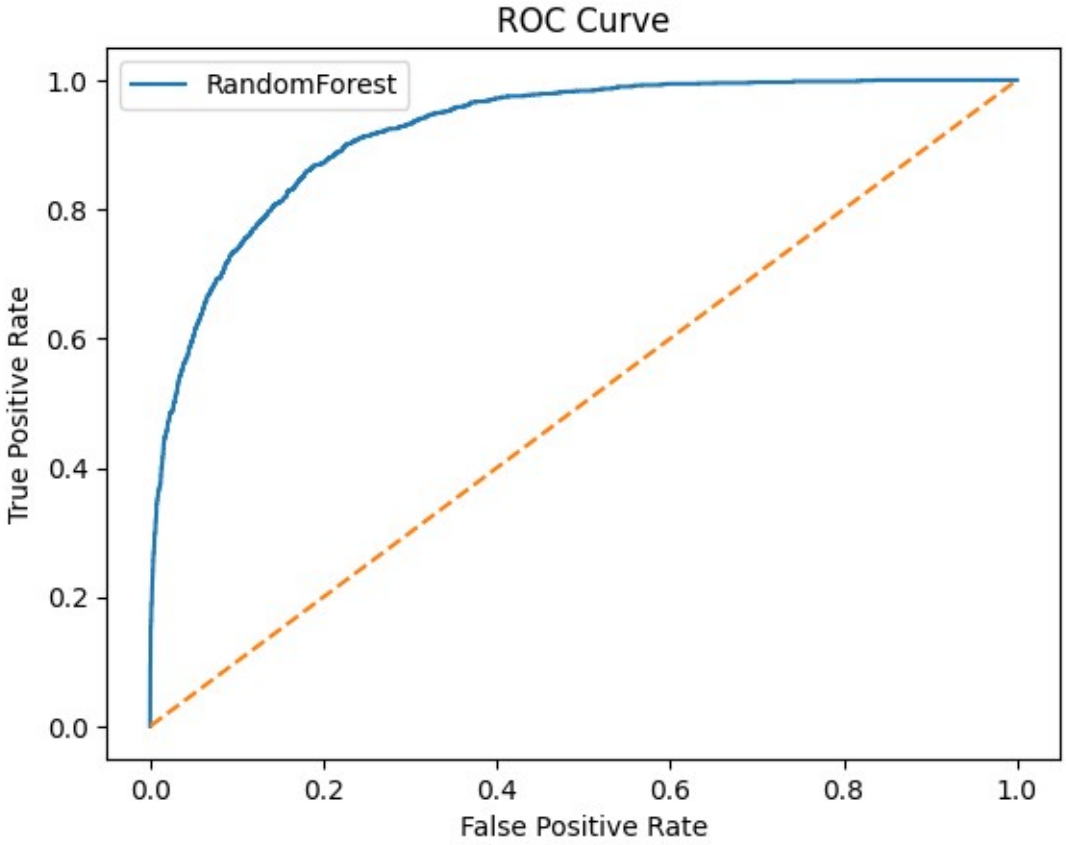
4. Modeling & Results

We trained Logistic Regression, Decision Tree, Random Forest, and SVM.

Model Performance

Model	Accuracy	F1-score	ROC-AUC
LogisticRegression	0.85598	0.674532	0.909838
RandomForest	0.857209	0.68216	0.906689

ROC Curve



5. Key Insights

- Random Forest achieved the best ROC-AUC.
- Accuracy alone is misleading due to class imbalance, F1 and ROC-AUC are better metrics.
- Education, hours-per-week, and age are strong predictors.
- Gender bias is visible in dataset.

6. Conclusion

The best performing model was Random Forest. Future work includes hyperparameter tuning, fairness analysis, and deployment.