Methodology

1. **Model Implementation**: Implemented Logistic Regression, Decision Tree, and Random Forest models.
2. **Model Tuning**: Tuned hyperparameters using GridSearchCV and cross-validation.
3. **Model Evaluation**: Evaluated models using accuracy, F1-score, classification reports, and confusion matrices.

Findings

Model	Accuracy	F1-Score	Confusion Matrix
Logistic Regression	0.8301	0.5575	[[10596, 637], [1852, 1568]]
Decision Tree	0.8628	0.6616	[[10676, 557], [1454, 1966]]
Random Forest	0.8624	0.6825	[[10470, 763], [1253, 2167]]

Confusion Matrix Summary Table

Model	True Negatives	False Positives	False Negatives	True Positives
Logistic Regression	10596	637	1852	1568
Decision Tree	10676	557	1454	1966
Random Forest	10470	763	1253	2167

Interpretation

- **Best Performing Model**: Random Forest achieved the highest F1-score (0.6825) and high accuracy (86.24%), indicating its robustness and better handling of class imbalances.
- **Logistic Regression**: Provided a good baseline with 83.01% accuracy but struggled with classifying the minority class.
- **Decision Tree**: Performed well with 86.28% accuracy but had a slightly lower F1-score compared to Random Forest.

- **Random Forest** has a higher number of true positives (2167) and a lower number of false negatives (1253) compared to Logistic Regression (TP: 1568, FN: 1852) and Decision Tree (TP: 1966, FN: 1454). This indicates that Random Forest is better at correctly identifying positive instances (income > 50K), which is crucial in many applications.

Insights

- **Random Forest**: Its ensemble approach led to better generalization and reduced overfitting.
- **Hyperparameter Tuning**: Significantly improved model performance across all models.