Şevval Çakıcı

CSSM502

Income Prediction Using Machine Learning Models

Dataset and Objective

This study predicts people's income levels using the Adult Income dataset from the UCI Machine Learning Repository. The dataset includes details like age, education, occupation, working hours, income and gender. Income is a binary variable, showing if someone earns more than $50K per year.

Data Preparation

At the beginning the dataset consisted of 48,842 data points. In the dta cleaning process, rows with missing values (indicated by question marks) in categorical fields were removed, leaving 45,222 data points. Categorical variables were transformed using one-hot encoding. Numerical variables were standardized. The data were split into 80% for training and 20% for testing using stratified sampling.

Methodology

I compared three models: 1. Logistic Regression, 2. Decision Tree, and 3. Random Forest. Logistic Regression was the base model. I added a Decision Tree to capture nonlinear patterns (it can sometimes show overfitting). I trained and tuned the Random Forest model using GridSearchCV and cross-validation to improve its generalization.

Results

Model performance was measured using accuracy and F1-score. (With extra focus on the F1-score because the classes are imbalanced). The results:

Model	Accuracy	F-1 Score
Logistic Regression	0.846	0.658
Decision Tree	0.806	0.617
Random Forest (Tuned)	0.855	0.682

Discussion and Conclusion

Logistic Regression was a good starting point but had trouble identifying high-income individuals. The Decision Tree did worse on the test data. The improved Random Forest model gave the best results, highest accuracy and F1-score. By combining many decision trees, the Random Forest avoided overfitting and found more complex patterns in the data. Overall, using several models together worked best for predicting income levels in this dataset.