Purpose

The goal of this project is to predict whether an individual's income exceeds 50K based on various demographic features. The dataset used is the UCI Adult dataset, and the project involves data preprocessing, model training, and evaluation of multiple machine learning algorithms.

Steps

Data Loading & Cleaning:

The dataset is loaded from a CSV file. Missing values represented as '?' are replaced and rows with missing data are dropped.

Duplicates are also removed to ensure clean data.

Feature Encoding:

Categorical features are converted into numerical values using one-hot encoding (pd.get_dummies), allowing models to process these features.

Feature Scaling:

Numerical features like age, capital-gain, capital-loss, and hours-per-week are scaled to a range of 0 to 1 using MinMaxScaler to normalize their values.

Data Visualization:

A correlation heatmap is plotted to show relationships between features.

A boxplot for the age column is used to check for outliers.

Model Training:

The data is split into training and validation sets.

Three models are trained: Logistic Regression, Decision Tree, and Random Forest. Each model is evaluated using key metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Model Evaluation & Comparison:

The models are compared based on their performance metrics, which are printed and visualized in a bar chart.