Project Report: Predicting Income Levels from Census Data

Introduction:

This project aims to analyse the Census Income dataset obtained from the UCI Machine Learning Repository. The dataset contains income information for over 48,000 individuals based on the 1994 US census. The primary objective of this project is to develop predictive models to determine whether an individual makes over 50,000 a year or less based on demographic and socioeconomic attributes.

Dataset Overview:

- **Dataset Source:** UCI Machine Learning Repository
- **Number of Instances:** 48,000+
- **Attributes:** Various demographic and socioeconomic attributes, including age, education, occupation, marital status, race, gender, etc.
- **Target Variable:** Income level (binary: >50K, <=50K)

Methodology:

Data Preprocessing:

- Handled missing values by dropping rows with missing values.
- Encoded categorical variables using label encoding and one-hot encoding.

Exploratory Data Analysis (EDA):

- Explored the dataset to gain insights into the distribution of income levels.
- Visualized key features such as work class, gender, race education level, income, occupation distribution, etc.

Model Building:

- Trained three different classifiers on the preprocessed dataset:
    - Logistic Regression
    - Decision Tree
    - Random Forest

Model Evaluation:

- Evaluated the performance of each classifier using the following evaluation metrics:
    - Accuracy: The proportion of correctly classified instances.
    - Precision: The proportion of true positive predictions among all positive predictions.

- Recall: The proportion of true positive predictions among all actual positive instances.
- F1-score: The harmonic mean of precision and recall, providing a balanced measure of the classifier's performance.

Results

The evaluation metrics for DT_classifier as follows:

Classifier	Accuracy	Precision	Recall	F1-score
Decision Tree	0.81	0.81	0.82	0.81

Conclusion

- Decision Tree classifiers achieved relatively higher accuracy and F1-score compared to others.
- The Random Forest classifier performed slightly better than the Logistic Regression classifier in terms of accuracy and F1-score.
- The results indicate that demographic and socioeconomic attributes can be effective predictors for determining income levels.
- Further feature engineering and model tuning may improve the predictive performance of the classifiers.