# Random Forest Model for Predicting Income Levels

**1. Introduction**

In this project, we built a Random Forest model to predict whether a person's income will be greater than $50K or less than or equal to $50K. We utilized the Census Income dataset, which contains various demographic and socioeconomic attributes of individuals.

**2. Data Preprocessing**

Before building the model, we performed necessary data preprocessing steps:

Handled missing values: We dropped rows with missing values to ensure data quality.

Encoded categorical variables: We used Label Encoding to convert categorical variables into numerical format, which is required for training machine learning models.

**3. Model Building and Tuning**

We built a Random Forest model using scikit-learn's RandomForestClassifier. The model was trained on the preprocessed data. To optimize its performance, we experimented with hyperparameter tuning using Grid Search. We searched over different combinations of hyperparameters such as the number of trees (n\_estimators), maximum depth of trees (max\_depth), and minimum number of samples required to split a node (min\_samples\_split). The best combination of hyperparameters was selected based on accuracy.

**4. Model Evaluation**

After training the model, we evaluated its performance on a separate testing dataset. We calculated various performance metrics including accuracy, precision, recall, and F1-score using scikit-learn's classification\_report. The model achieved an accuracy of approximately 86.17%, indicating its ability to correctly predict income levels for the majority of individuals.

**5. Discussion of Findings**

Important features for prediction: Based on the Random Forest model, important features for predicting income levels include age, education level, occupation, marital status, and capital gain. These features provide valuable insights into the socioeconomic status of individuals and their potential income levels.

**6. Conclusion**

In conclusion, we successfully built and trained a Random Forest model to predict income levels using the Census Income dataset. The model demonstrated good performance in distinguishing between individuals with income greater than $50K and those with income less than or equal to $50K. Further improvements could be made by exploring additional features, experimenting with different machine learning algorithms, and fine-tuning hyperparameters.

**7. Future Work**

In future iterations of this project, we could explore advanced techniques such as feature engineering, handling class imbalance, and ensemble learning methods to further enhance the model's performance. Additionally, collecting more data and conducting more in-depth analysis could provide deeper insights into the factors influencing income levels.