**Problem Definition:**

The given dataset is about Census Income, and the objective is to predict whether an individual’s income exceeds $50,000 per annum based on various attributes like education, employment, occupation and work class. This is a binary classification problem where the target variable is the “income” column, which has two classes: <=50k and >50k.

This dataset consists of 32560 rows and 15 columns such as Age, Work class, Fnlwgt, Education, Education\_num, Marital\_status, Occupation, Relationship, Race, sex, capital\_loss, Hours\_per\_week, Native\_country, Income.

**Data Analysis:**

Data analysis is the process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information, drawing conclusions and supporting decision making. It involves a series of steps that help to understand the data, identify patterns, and extracts insights.

In this dataset has 32560 entries and 15 columns which has no null values however it has some’?’ those were considered as the missing values. Missing values in the Work class, occupation and native country columns were imputed using mode. Additionally, Duplicated records were removed to avoid any biases that could affect model performance. Outliers are also found in capital gain and capital loss columns are identified and treated to ensure they did not skew the results.

Exploratory data analysis helps to understand the distribution and relationships of the data. Descriptive statistics revealed that most individuals worked 40 hours per week and had a high school level of education. Visualizations such as scatterplot, pair plot, histograms and boxplots were used to inspect the distribution of numerical features and relationships between different variables. Also, bar plots and count plot were used to understand about the categorical features. For instance, it was observed that higher education levels were associated with higher income levels and also most of the individuals in the dataset were employed in private sector and high number of individuals were single in the dataset.

**Exploratory Data Analysis Concluding Remarks:**

**The** Exploratory Data Analysis provides significant insights**:**

* There is a strong corelation between higher Education level and higher income brackets.
* Majority of the individuals in the dataset working fill-time with standard workweek of around 40 hours per week.
* Most of the individuals working in private sector.
* Capital gains and losses has outliers.
* The distribution of Capital gains and capital losses were highly skewed.
* Majority of the people in the dataset below 50 years.
* It is important to note that the majority of the individuals represented are American and white.
* Most of the individuals in the dataset had High school level of education.

These insights indicated that educational level, capital gains and capital losses are vital features for predicting the income. These features will be essential in building and refining the predictive model. Furthermore, the presence of outliers in capital gains and losses suggests that these variables might need special handling during preprocessing.

This dataset is imbalanced which means one of the classes is significantly under represented than the other class. To overcome this there is one popular technique to address all these issues is SMOTE which, stands for Synthetic Minority Over Sampling Technique.

**Pre-processing Pipeline:**

Feature engineering involves creating new features or modifying existing ones to improve the model’s performance. In this analysis, dropping some of the columns like education does not affect the performance of the model because education and education\_num column represents same information.

Data transformation is necessary to prepare the dataset for modelling. Numerical features were normalized to ensure they had a consistent scale, which helps in improving the performance of many machine learning algorithms. Categorical variables were one-hot encoded to convert them into a numerical format that could be fed into the models.

To evaluate the model’s performance effectively, the dataset was split into training and testing sets. An 70-30 splits was used, where 70% of the data was used for training the model and the remaining 30% was used for testing. This split ensures that the model performance can be evaluated on unseen data, providing an estimate of how it would perform in a real-world scenario.

**Building Machine Learning Models:**

The selection of models is a critical step in the analysis, for this problem Logistic Regression, decision tree, adaboost, random forest models were chosen due to their interpretability and proven performance in classification tasks. Logistic Regression is a simple and effective model for binary classification, while decision tree and random forest are powerful for capturing complex interaction between features.

Each model was trained on the training dataset, and hyperparameter tuning was performed using grid search and Cross-validation techniques to avoid the overfit the training data. By using the performance metrics such as accuracy score, confusion matrix, precision, recall, f1-score and classification report to evaluate the performance of each model.

The random forest model outperformed the other models achieving the accuracy of 83%. This indicated that random forest classification model was accurately classify the individual income level with high degree of precision and recall.

**Concluding Remarks:**

**Summary of Results:** The random forest model was the best-performing model in this analysis, accurately classifying income levels based on the provided attributes.