Report on Cleaning the Adult Income Dataset

1. Dataset Chosen and Reason for Selection

The dataset used in this project is the **Adult Income Dataset** from the UCI Machine Learning Repository. This dataset is commonly used for classification tasks, especially for predicting whether a person earns more or less than $50K per year. It includes various attributes such as age, education, occupation, work class, and marital status.

The dataset was selected because it contains both numerical and categorical data, which makes it useful for learning data cleaning techniques. Additionally, it is a real-world dataset that helps in understanding income distribution and social factors related to earnings.

2. Challenges Faced During the Cleaning Process

Several challenges were faced while cleaning the dataset:

- **Missing Values**: Some columns had missing values represented by a? symbol instead of NaN.
- **Inconsistent Data**: Different formats and categorical values made it difficult to process the data directly.
- **Duplicates**: Some rows contained duplicate entries.
- **Encoding Issues**: The dataset contained text-based categorical variables that needed to be converted into numerical values.
- **Imbalanced Classes**: The target column (income) had more people earning <=50K than >50K, affecting model fairness.

3. Steps Followed for Cleaning and Their Impact

The following steps were taken to clean the dataset:

Handling Missing Data

- The missing values in the workclass and occupation columns were replaced with the most common value (mode).
- Rows with missing values in the native-country column were removed.
    - **Impact:** This ensured no missing values remained, improving data consistency.

Removing Duplicates

- Duplicate rows were identified and removed.

    - **Impact:** This reduced redundancy and improved model performance.

Encoding Categorical Data

- One-hot encoding was applied to categorical columns like marital-status, race, and relationships.
- Label encoding was used for the income column, converting <=50K to 0 and >50K to 1.
    - **Impact:** This allowed machine learning models to understand categorical data better.

Feature Scaling

- Numerical columns such as age, and hours-per-week were normalized to have a similar range.
    - **Impact:** This helped prevent models from being biased toward larger numbers.

4. Insights from Cleaning and Dataset Readiness

After cleaning, the dataset became structured, complete, and suitable for further analysis or machine learning modeling. Some insights gained include:

- Most missing values were in categorical columns like work class and occupation.
- Encoding and scaling improved data usability.
- The dataset still has an imbalance in the income classes, which needs to be handled in modeling.

Now, the dataset is ready for feature selection, exploratory data analysis (EDA), and predictive modeling.