**Introduction**

The dataset used predicts whether an individual’s income exceeds $50K annually based on features like age, education, and occupation. It contains 32,560 records with both numerical and categorical data, and the target variable is income.

**Data Preparation & Exploration**

**Column Naming**:

To improve readability, the dataset’s column names were manually renamed, mapping them to meaningful labels such as age, workclass, education, etc.

**Data Inspection:**

After renaming, the dataset was checked for duplicates, missing values, and its general structure

Missing values were detected in the workclass and occupation columns. These missing values were encoded as “?” in the dataset and were replaced with NaN for better handling during the cleaning process.

**Handling Missing Values:**

For both the workclass and occupation columns, missing values were imputed using the mode (most frequent value) from the training dataset. This approach ensures minimal information loss while filling in gaps in the data.

**Data Cleaning**

The ‘?’ values in categorical columns such as workclass, occupation, and native\_country were replaced with NaN.

Missing values in these columns were replaced using the mode of the respective column, which is a common and simple imputation technique when dealing with categorical data.

**Feature Encoding**

Given that the dataset contains categorical variables, these were converted to numerical format using One-Hot Encoding. This method creates binary indicator variables for each category, enabling categorical data to be used in machine learning models. Both the training and testing datasets were transformed using this technique.

**Model Training**

The model chosen for this classification problem was Gaussian Naive Bayes (GaussianNB). This model assumes that features are normally distributed and works well for tasks with simple assumptions about data distribution.

**Train-Test Split:**

The dataset was split into training and testing sets using an 80-20 split, with 70% of the data used for training and 30% for testing. This helps in ensuring that the model is trained on a significant portion of the data while being evaluated on unseen data.

**Model Training:**

The Gaussian Naive Bayes model was trained on the encoded training dataset. After training, predictions were made on the test dataset.

**Model Performance Evaluation**

The performance of the model was evaluated using accuracy scores on both the training and test sets:

Model Accuracy: The model achieved an accuracy of 79.21% on the test set.

Training Set Score: The training accuracy was 79.56%, which is consistent with the test accuracy, suggesting that the model generalizes well and is not overfitting.

**Insights & Conclusions**

Data Quality: The dataset had some missing values in key columns like workclass and occupation, which were handled via imputation. However, better results may be achieved by considering more sophisticated imputation techniques or external sources to fill in these missing values.

Model Choice: The Gaussian Naive Bayes classifier provided reasonable accuracy (79%) for this binary classification problem. However, given that the dataset contains a mix of numerical and categorical data, experimenting with more sophisticated models (e.g., decision trees or logistic regression) could potentially improve performance.

Predictive Power: Features like education level, occupation, and hours worked per week are likely influential in predicting income. Further feature analysis and engineering (e.g., feature scaling, interaction terms) could improve the model’s predictive power