Name: Ahmed Mostafa AbdEl-Rahman Hassan

ID: 20221372883

Census Data

Download Data

First we will download the data from the file

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

Preprocess the Data

Has Missing Values?
Yes

We will Handle missing values by dropping rows with NaN values.

We will encode categorical variables into numerical using one-hot encoding and then map the income column to binary values: 0 for <=50K and 1 for >50K.

Data Splitting

We will split the dataset into features (X) and target variable (y) , Then divide the data into training and testing sets with a test size of 20%

Model Building

We will use Gaussian Naive Bayes classifier and train the classifier using the training data.

Prediction and Evaluation

We will make predictions on the test set then calculate the confusion matrix to evaluate the model's performance.

```
Sensitivity: 0.32017823042647997
Specificity: 0.9514366653176851
```

Sensitivity (True Positive Rate): This value (0.3202) suggests that the classifier correctly identifies about 32.02% of individuals who actually make over 50K a year from the total number of individuals who actually make over 50K a year.

Specificity (True Negative Rate): This value (0.9514) indicates that the classifier correctly identifies about 95.14% of individuals who actually do not make over 50K

a year from the total number of individuals who do not make over 50K a year.

```
Posterior Probability of making over 50K a year: [4.31088775e-03 1.37859620e-02 1.71229441e-02 ... 1.00000000e+00
 6.39925511e-03 6.53278628e-04]
Average posterior probability of making over 50K a year: 0.12424877416671441
```

These values represent the probability assigned by the classifier to each instance belonging to the positive class (income >50K). The probabilities are scaled between 0 and 1, and they represent the confidence of the classifier in its predictions.