Name : Ahmed Hassan Ahmed Ali

ID : 2203148

AI

Predicting Income Using Naive Bayes Classifier

1. Introduction: In this report, we aim to predict whether an individual earns over 50K a year or not using a Naive Bayes Classifier. We'll utilize UCI's Census data, which contains various demographic and employment-related attributes.

2. Data Acquisition and Preprocessing:

- We downloaded the Census data files: **adult.data** and **adult.test**.
- The data files include features such as age, workclass, education level, marital status, occupation, etc.
- We loaded the data into Pandas DataFrames and handled missing values denoted by ' ?'.
- Categorical variables were encoded into dummy variables for compatibility with the Naive Bayes Classifier.

3. Training the Naive Bayes Classifier:

- We split the data into features (X) and the target variable (y) for both training and testing datasets.
- We instantiated a Gaussian Naive Bayes Classifier from the scikit-learn library.
- The classifier was trained on the training data using the **fit()** method.

4. Evaluation Metrics:

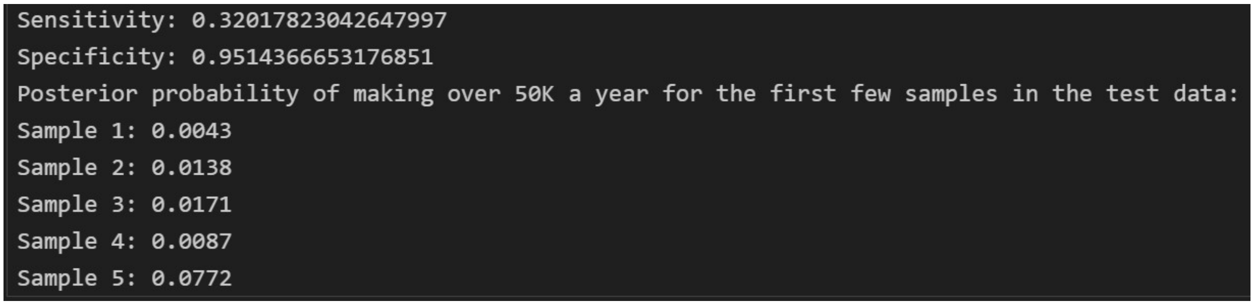
- Sensitivity (True Positive Rate) and Specificity (True Negative Rate) were computed to evaluate the classifier's performance.
- Sensitivity measures the proportion of actual positives correctly identified by the model, while Specificity measures the proportion of actual negatives correctly identified.
- Confusion matrix obtained from the predictions on the test set was used to compute these metrics.

5. Results:

- Sensitivity: The proportion of individuals correctly predicted to earn over 50K a year.
- Specificity: The proportion of individuals correctly predicted to earn less than or equal to 50K a year.

6. Posterior Probability:

- The classifier's **predict_proba()** method was used to compute the posterior probability of making over 50K a year for each sample in the test set.
- The posterior probability represents the probability of belonging to each class (<=50K or >50K) for a given sample.



```
Sensitivity: 0.32017823042647997
Specificity: 0.9514366653176851
Posterior probability of making over 50K a year for the first few samples in the test data:
Sample 1: 0.0043
Sample 2: 0.0138
Sample 3: 0.0171
Sample 4: 0.0087
Sample 5: 0.0772
```