**Lab 6**

**Machine Learning Models**

The machine learning models that were used in Lab 6 to predict whether a person’s income will exceed $50K/yr are:

1. Decision Tree - The Decision Tree algorithm partitions data into subsets based on attributes, forming a hierarchical tree structure. It predicts outcomes by traversing branches according to feature values, making it intuitive for classification and regression tasks.
2. Naïve Bayesian - The Naïve Bayesian classifier operates under the assumption that features are independent, which simplifies probability calculations by treating each feature separately. Despite its straightforward nature, this classifier excels in classification, utilizing Bayes' theorem to make accurate probabilistic predictions.

**Size of Testing and Training data**

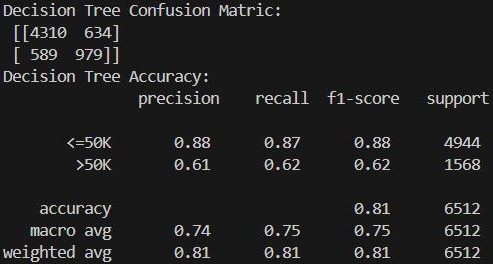
We used the following Python code to split the entire data into testing and training sets:

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, stratify=y)

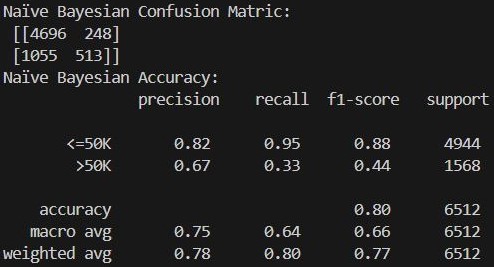
Here, **test\_size=0.2** indicates that 20% of the data is being used for testing and the remaining 80% is used for training. Since we have 32560 rows, the testing and training data quantity are 6512 and 26048 respectively.

**Best Model**

The performance metrics for the Decision Tree algorithm are as follows:



And, the performance metrics for the Naïve Bayesian classifier are as below:



If we were more concerned with the exactness of our model, we would consider the precision matric based on the result we highly value i.e. if we really want to predict whether a person’s income will exceed $50K/yr and we highly value the target class “>50K”, then Naïve Bayesian will be the best as its precision is higher than that of Decision Tree.

Similarly, if we were more concerned with the completeness of our model, we would consider the recall metric based on the result we highly value i.e. if we really want to predict whether a person’s income will exceed $50K/yr and we highly value the target class “>50K”, then Decision Tree will be the best as its recall is higher than that of Naïve Bayesian.

But, since we don’t know whether we consider exactness or completeness as our primary criteria, we consider F1-score matric which is the harmonic mean between precision and recall. **Therefore, the Decision Tree is the best compared to Naïve Bayesian as the F1-score for the target class “>50K” is greater in the Decision Tree.**

**Based on the confusion matrix**

For Decision Tree:

TP = 4310

TN = 979

FP = 589

FN = 634

For Naïve Bayesian:

TP = 4696

TN = 513

FP = 1055

FN =248

Therefore, the Decision Tree correctly predicted 4310+979 = 5289 data whereas Naïve Bayesian correctly predicted 4696+513 = 5209. **Hence, the Decision Tree is better.**