**Assignment 3**

**Task I. Basic concepts (10 points each, total 20 points)**

1. Briefly outline the idea and major steps of decision tree classification.

Decision tree classification is basically a binary tree that represents a set of decision for every outcome that gives out a certain result. Each decision should lead to a yes or no question. The major steps for a decision tree are:

1. Understanding the dataset.
2. Divide and conquer
   1. Split into subsets.
   2. Is the dataset pure.
   3. If yes: stop
   4. If no: repeat
3. With the test data set see where the new data falls into with the tree.
4. What is the basic idea of k-nearest neighbor? How to compute distance between two points?

The basic idea of k-nearest neighbor that’s its classification technique that represents the attributes into groups. The aim for the k-nearest neighbor algorithm is to find the most similar data points in the training set data and make an educated guess based of the test dataset.

**Task II. Naïve Bayes classification (30 points)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Manufacturer** | **RAM** | **CAPACITY** | **WARRANTY** | **BATTERY** | **COST** | **BUY** |
| DELL | 4GB | 500GB | 0 | 4hrs | LOW | **NO** |
| TOSHIBA | 8GB | 1TB | 1 | 12hrs | HIGH | **YES** |
| SAMSUNG | 16GB | 1TB | 1 | 4hrs | LOW | **YES** |
| LENEVO | 8GB | 2TB | 0 | 12hrs | LOW | **NO** |
| TOSHIBA | 16GB | 500GB | 2 | 12hrs | LOW | **YES** |
| SONY | 16GB | 1TB | 2 | 4hrs | HIGH | **YES** |
| SAMSUNG | 4GB | 1TB | 0 | 12hrs | LOW | **NO** |
| LENEVO | 8GB | 500GB | 1 | 8hrs | LOW | **NO** |
| APPLE | 4GB | 2TB | 2 | 12hrs | LOW | **YES** |
| DELL | 16GB | 2TB | 0 | 8hrs | LOW | **YES** |
| SONY | 8GB | 500GB | 1 | 18hrs | HIGH | **NO** |
| TOSHIBA | 4GB | 4TB | 2 | 18hrs | VERY HIGH | **NO** |
| HP | 16GB | 2TB | 2 | 4hrs | LOW | **NO** |
| APPLE | 4GB | 1TB | 0 | 18hrs | HIGH | **NO** |
| ASUS | 16GB | 500GB | 3 | 8hrs | HIGH | **YES** |
| SAMSUNG | 8GB | 2TB | 0 | 18hrs | HIGH | **NO** |
| DELL | 4GB | 4TB | 1 | 8hrs | LOW | **YES** |
| HP | 16GB | 500GB | 2 | 8hrs | LOW | **YES** |
| SAMSUNG | 4GB | 2TB | 0 | 18hrs | HIGH | **NO** |
| DELL | 8GB | 4TB | 1 | 18hrs | HIGH | **YES** |
| HP | 16GB | 1TB | 3 | 4hrs | HIGH | **NO** |
| ASUS | 4GB | 500GB | 0 | 12hrs | HIGH | **NO** |
| SONY | 16GB | 4TB | 1 | 8hrs | LOW | **YES** |
| HP | 8GB | 2TB | 3 | 12hrs | LOW | **YES** |
| SAMSUNG | 4GB | 4TB | 2 | 8hrs | LOW | **YES** |
| TOSHIBA | 16GB | 500GB | 3 | 18hrs | HIGH | **YES** |
| APPLE | 4GB | 1TB | 0 | 4hrs | HIGH | **NO** |
| ACER | 4GB | 4TB | 1 | 12hrs | LOW | **YES** |
| ASUS | 4GB | 8TB | 3 | 12hrs | LOW | **YES** |
| DELL | 8GB | 500GB | 3 | 4hrs | LOW | **NO** |
| HP | 16GB | 4TB | 3 | 18hrs | VERY HIGH | **YES** |
| ACER | 8GB | 4TB | 0 | 8hrs | HIGH | **NO** |
| HP | 16GB | 8TB | 3 | 8hrs | VERY HIGH | **YES** |
| HP | 8GB | 1TB | 1 | 18hrs | LOW | **YES** |
| APPLE | 8GB | 500GB | 1 | 8hrs | LOW | **NO** |
| TOSHIBA | 8GB | 8TB | 2 | 18hrs | VERY HIGH | **YES** |
| HP | 16GB | 8TB | 3 | 12hrs | VERY HIGH | **YES** |
| ACER | 8GB | 1TB | 0 | 4hrs | LOW | **NO** |
| APPLE | 4GB | 8TB | 3 | 12hrs | HIGH | **YES** |
| SONY | 16GB | 8TB | 2 | 4hrs | HIGH | **YES** |
| LENEVO | 16GB | 1TB | 3 | 18hrs | LOW | **YES** |
| ACER | 8GB | 8TB | 3 | 12hrs | HIGH | **YES** |
| HP | 4GB | 2TB | 1 | 8hrs | LOW | **YES** |
| ASUS | 8GB | 8TB | 3 | 18hrs | VERY HIGH | **NO** |
| HP | 8GB | 4TB | 2 | 4hrs | LOW | **NO** |
| ACER | 16GB | 8TB | 2 | 12hrs | HIGH | **YES** |
| SONY | 4GB | 2TB | 0 | 4hrs | LOW | **NO** |
| LENEVO | 16GB | 4TB | 3 | 8hrs | VERY HIGH | **YES** |
| ASUS | 16GB | 2TB | 1 | 12hrs | HIGH | **YES** |
| HP | 4GB | 8TB | 2 | 4hrs | LOW | **NO** |

**Figure 1: Training Data Set**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Manufacturer** | **RAM** | **CAPACITY** | **WARRANTY** | **BATTERY** | **COST** | **BUY (Actual Label)** |
| DELL | 8GB | 500GB | 0 | 4hrs | LOW | NO |
| LENEVO | 16GB | 1TB | 3 | 8hrs | HIGH | YES |
| HP | 4GB | 2TB | 1 | 12hrs | LOW | YES |
| APPLE | 8GB | 4TB | 2 | 4hrs | HIGH | NO |
| ASUS | 16GB | 500GB | 0 | 18hrs | VERY HIGH | YES |
| DELL | 8GB | 8TB | 2 | 8hrs | LOW | YES |
| TOSHIBA | 4GB | 1TB | 1 | 12hrs | VERY HIGH | NO |
| ACER | 16GB | 4TB | 2 | 8hrs | HIGH | YES |
| SONY | 8GB | 2TB | 3 | 18hrs | VERY HIGH | YES |
| SAMSUNG | 4GB | 8TB | 2 | 8hrs | HIGH | NO |

**Figure 2: Testing Data Set**

Consider the above training and testing data sets. The training data set contains 50 data points and the testing data set contains 10 data points.

All the attributes in the data set are nominal. The attributes along with their possible nominal values are shown below.

**'Manufacturer' {'DELL', 'LENEVO', 'HP', 'APPLE', 'ASUS', 'TOSHIBA', 'ACER', 'SONY', 'SAMSUNG'}**

**'RAM' {'4GB', '8GB', '16GB'}**

**'CAPACITY' {'500GB', '1TB', '2TB', '4TB', '8TB'}**

**'WARRANTY' {'0', '1', '2', '3'}**

**'BATTERY' {'4hrs', '8hrs', '12hrs', '18hrs'}**

**'COST' {'LOW', 'HIGH', 'VERY HIGH'}**

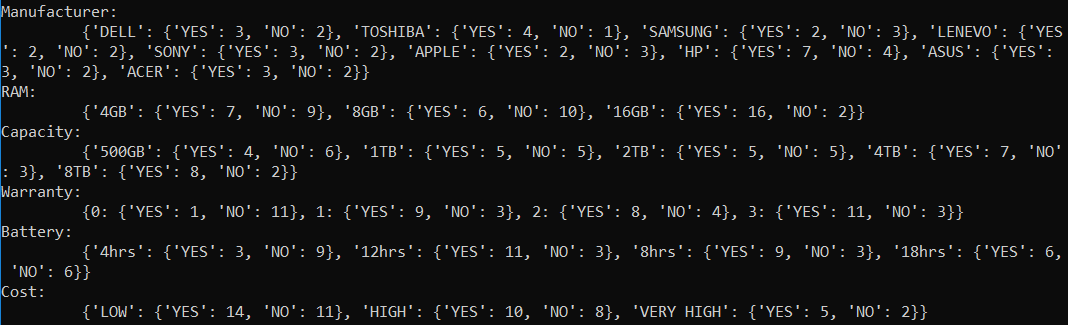
**'BUY' {'YES', 'NO'}**

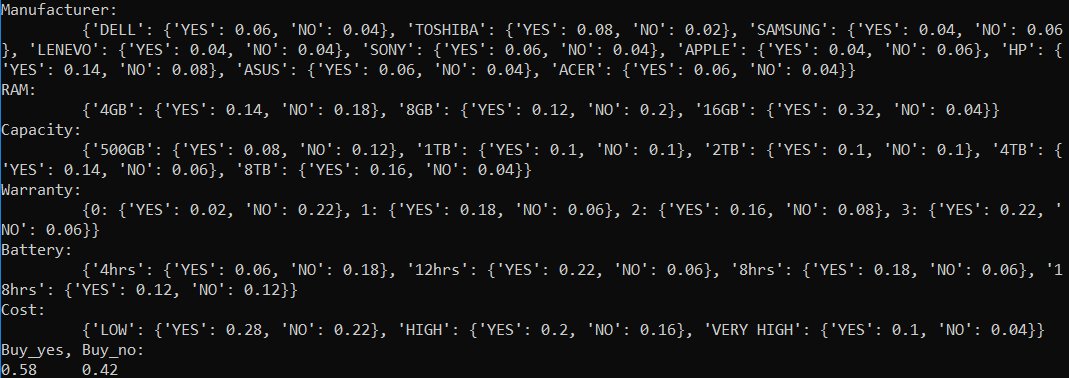
The ‘BUY’ attribute is the label that tells us if a customer would buy a laptop or not considering the values of the other attributes.

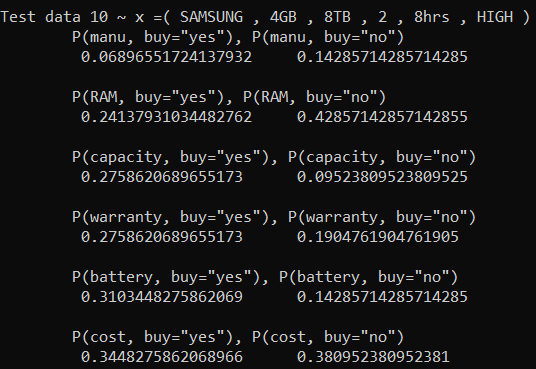
1. **(20 total)** Use the Naïve Bayes Algorithm to build a model using the training data set and use this model to predict the label, which is the ‘BUY’ attribute for the testing data set.

**I built a python script to do the Naïve Bayes Algorithm for the following training & test data set.**

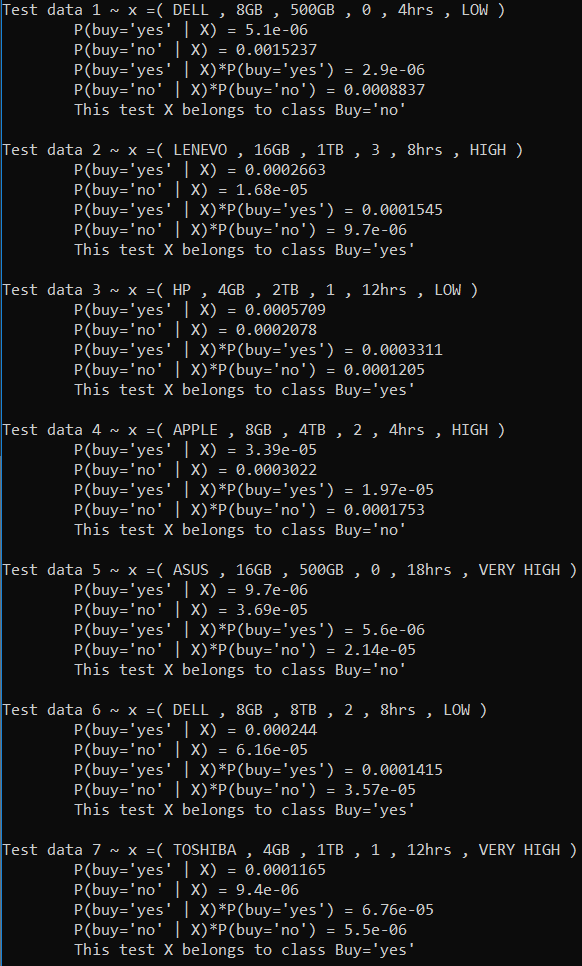
* 1. From the training set data get the all the count of yes or no for each nominal value.

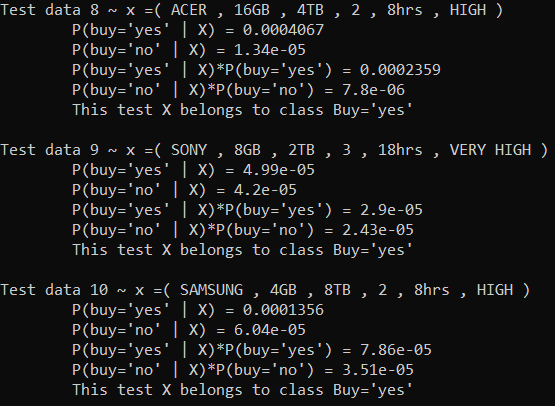


* 1. Compute P(X and Ci) & P(Ci)
     1. 
  2. Compute P(X|Ci) for each class. X = one attribute Ci= either Buy ‘yes’ or ‘no’
     1. Result for the last test data set of P(X | Ci) for each class.



* 1. Find the probability for each test should be for Buy.





1. **(10 total)** After predicting the label for the testing data set, you will have both the predicted and the actual label for the data set. Use this information to calculate the Accuracy, Precision and Recall evaluating the performance of the model that you just created.

Please indicate each step on how you predicted the label of each data point in the test data set as discussed in the slides.