**BNU-HKBU United International College**

**DS4013: Data Mining (For DS students)**

**Fall 2021 Assignment2**

**Q1.** In this assignment, you will perform association analysis on the following transactional database.

Table1: Transactional Database

|  |  |
| --- | --- |
| Transaction ID | Items Bought |
| T1 | {a,c,d,e} |
| T2 | {a,b,c,d,e} |
| T3 | {a,c,d,e} |
| T4 | {a,b,c,d} |
| T5 | {a,b,d,e} |

1. Compute the support to itemsets {e}, {b,d} and {b,d,e}.
2. Trace the results of using the Apriori Algorithm on the given example with support threshold s=0.8 and confidence threshold c=0.9. For () iteration of the algorithm, show
   1. the candidate ()-itemsets obtained by self-joining the frequent -itemsets (the items are sorted by alphabetic order), before candidate pruning.
   2. the frequent ()-itemsets (with the corresponding support counts).
3. Based on the results in part b), list all strong association rules (s=0.8, c=0.9) with the corresponding support and confidence (sort the rules by decreasing order of confidence).
4. Use the same support threshold as part (b), build a frequent pattern tree (FP-Tree) (if two

items has the same support count, sort by alphabetic order). Use FP-Growth to discover

the frequent itemsets from the constructed FP-Tree. Show the conditional FP-trees for

each step.

**Q2**. The following table contains training examples that help predict whether a patient is likely to have a heart attack.

* Note that the log base of entropy is 2

Table2: Training Set of Patients

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Patient** | **Chest Pain** | **Male** | **Smoke** | **Exercise** | **Heart Attack** |
| 1 | Yes | Yes | No | Yes | Yes |
| 2 | Yes | Yes | Yes | No | Yes |
| 3 | No | No | Yes | No | Yes |
| 4 | No | Yes | No | Yes | No |
| 5 | Yes | No | Yes | Yes | Yes |
| 6 | No | Yes | Yes | Yes | No |

1. Suppose we set the termination condition to be “all the patients of the node belong to the same class”. Using the Information Gain and construct the decision tree of the given datasets. Show the detail steps and plot the constructed decision tree.
2. Considering the following test set with ground truth, use the constructed decision tree to perform classification. Show the classification results and the precision/recall/F1-score of the classification. (The yes class of Heart Attack is our main interest.)

Table3: Test set of Patients

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Patient** | **Chest Pain** | **Male** | **Smoke** | **Exercise** | **Heart Attack** |
| 7 | Yes | Yes | No | Yes | No |
| 8 | Yes | No | Yes | No | Yes |
| 9 | Yes | No | Yes | No | Yes |
| 10 | No | Yes | No | No | No |
| 11 | No | Yes | Yes | Yes | Yes |
| 12 | No | No | Yes | Yes | No |

**Q3**. Instead of using the Decision Tree model, we now use the Naïve Bayes Classifier model and adopt Laplacian correction strategy while computing the conditional probability

Train a Naïve Bayes Classifier on the training set given in Table2 and make prediction on Patient7 and Patient8. List all the details of your prediction procedure.