**Birla Institute of Technology & Science, Pilani**
**Work-Integrated Learning Programmes Division**
**First Semester 2019-2020**
**M.Tech (Data Science and Engineering)**
**Mid-Semester Test (EC-2 Make-up)**

| | |
|---|---|
| Course No. | : DSECLZC415 |
| Course Title | : DATA MINING |
| Nature of Exam | : Closed Book |
| Weightage | : 30% |
| Duration | : 90 Minutes |
| Date of Exam | : 06/07/2019 (AN) |

| | |
|---|---|
| No. of Pages | = 3 |
| No. of Questions | = 4 |

Note:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

**Answer All the Questions (only in the pages mentioned against questions. if you need more pages, continue remaining answers from page 21 onwards)**

**Question 1: [2 + 2 + 2 = 6 marks]** **[to be answered only in pages 2-6]**

a) Given the following two objects with four binary attributes. **[1+1= 2 marks]**

| | **Attribute1** | **Attribute2** | **Attribute3** | **Attribute4** |
|---|---|---|---|---|
| Object1 | 1 | 1 | 0 | 0 |
| Object2 | 1 | 0 | 1 | 0 |

1) What is the distance between the objects if all variables are symmetric?
2) What is the distance between the objects if all variables are asymmetric?

b) Data Mining will automatically clean up our messy databases. Justify or Invalidate.

c) Suppose you have collected a set of 1,000,000 labeled data points and you build a decision tree classifier from them. You then choose 100 of these points at random, and find that your classifier returns the correct answer on all of them. Can you conclude that your algorithm works with a 0% error rate on any input? Why or why not?

**Question 2: [3+3 = 6 Marks]** **[to be answered only in pages 7-11]**

a) Given the term frequency vectors of the documents, identify the most similar ones using appropriate similarity measure:

| Document | Term frequency vector |
|---|---|
| D1 | 10011101010101 |
| D2 | 11001001000101 |
| D3 | 10111101010111 |

b) Ramesh is an investor. His portfolio primarily tracks the performance of the Nifty and Ramesh wants to add the stock of ABC Corp. Before adding the stock to his portfolio, he wants to assess the directional relationship between the stock and the Nifty. Ramesh does not want to increase the unsystematic risk of his portfolio. Thus, he is interested in owning securities in the portfolio that tend to move in the same direction.
Considering the data set given below, what would you suggest whether Ramesh should invest in ABC Corp. stock? Justify your answer

| **Year** | **2013** | **2014** | **2015** | **2016** | **2017** |
|---|---|---|---|---|---|
| **Nifty** | 1692 | 1978 | 1884 | 2151 | 2519 |
| **ABC Corp** | 68 | 102 | 110 | 112 | 154 |

**Question 3: [6 + 4 = 10 Marks]** **[to be answered only in pages 12-16]**

a) Suppose we train a model to predict whether an email is Spam or Not Spam. After training the model, we apply it to a test set of 200 new email messages (also labelled) and the model produces the contingency table below. **[1.5*4= 6 marks]**

| | | Predicted Class | |
|---|---|---|---|
| | | Spam | Not Spam |
| True Class | Spam | 60 | 0 |
| | Not Spam | 120 | 20 |

1) Compute the precision of this model with respect to the "Spam" class and with respect to the "Not Spam" class.
2) Compute the recall of this model with respect to the "Spam" class and with respect to the "Not Spam" class.

Suppose we have two users (Radha and Seetha) with the following preferences.

- Radha hates seeing spam messages in her inbox! However, she doesn't mind periodically checking the "junk" folder for messages incorrectly marked as spam.
- Seetha doesn't even know where the "junk" folder is. She would prefer to see spam messages in her inbox than to miss genuine messages without knowing!

3) Would Radha like this classifier? Justify your answer.
4) Would Seetha like this classifier? Justify your answer.

b) Consider the following training data set (with two attributes and two possible classes Y, N) for a binary class problem. The attributes are nominal with two possible values. We intend to create decision tree model. **[2+1+1=4 marks]**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Income | H | H | H | H | H | L | L | L | H | H |
| Education | L | H | H | L | H | L | L | L | H | L |
| Will Buy | N | N | N | Y | N | Y | Y | Y | Y | Y |

1) Calculate the gain in the Gini index when splitting on *Income* and *Education*.
2) Which attribute would the decision tree induction algorithm choose?
3) Can Gini Index help you get the most optimal decision tree?

**Question 4: [4+2+2 = 8 Marks]** **[to be answered only in pages 17-20]**

a) A user wants to mine association rules of form X → Y where X and Y are one-item sets that maximize the measure called lift. Identify 3 rules that have highest lift for the given set of transactions.

| Transaction ID | Items |
|---|---|
| 1 | a,b,c,d |
| 2 | b,c,d |
| 3 | a,c,d,e |
| 4 | b,c,d,e, |
| 5 | c,d |
| 6 | a,b,c |

b) Consider the following set of frequent 3-itemsets:
{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}.
Assuming that there are only five items {1,2,3,4,5} in the data set, identify all the possible frequent 4-itemsets that satisfy the Apriori principle.

c) Suppose you are dealing with a large transactional dataset of size 10 gb. You need to perform frequent Itemset mining. Which algorithm (FP-Growth or Apriori algorithm) would you prefer? Justify your answer.