ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION **SUMMER SEMESTER, 2021-2022**

DURATION: 3 HOURS **FULL MARKS: 150**

CSE 6293: Data Warehousing and Mining

Programmable calculators are not allowed. Do not write anything on the question paper.

Answer **all 6 (six)** questions. Figures in the right margin indicate full marks of questions.

---

1. a) Consider the following set of frequent 3-itemsets where there are only five items in the data set: 13

{1, 2, 3}, {1, 2, 4), {1, 2, 5}, {1, 3, 4},{ 1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5 } .

i. List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

ii. List all candidate 4-itemsets obtained by the candidate generation procedure using *Apriori* algorithm.

iii. List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.

b) Answer the following considering the market basket transactions provided in Table1. 12

i. What is the maximum number of association rules that can be extracted from this data?

ii. What is the maximum size of frequent itemsets that can be extracted (assuming $minsup > 0$)?

iii. Find an itemset (of size 2 or larger) that has the largest support.

Table 1: Market basket transactions for Question 1b).

Transaction ID	Items Bought
1	{Milk, Beer, Diapers]
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

2. a) What are the different sources of classification error. Mention two disadvantages of decision tree based classification. 6

b) What are the causes of model overfitting? Explain different techniques to handle the class imbalance problem. 4+10

c) What are the outcomes of choosing a smaller or larger K value in the KNN algorithm? 5

b) Table2 shows the training examples for a binary classification problem. Considering the data answer the following questions:

i. Compute the Gini index for the overall collection of training examples.
ii. Compute the Gini index for the *Custoner ID* attribute.
iii. Compute the Gini index for the *Gender* attribute.
iv. Compute the Gini index for the *Car Type* attribute using multiway split.
v. Which attribute is better, *Gender*, *Car Type*, or *Shirt Size*?

Table 2: Data set for Question 3b).

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	CO
2	M	Sports	Medium	CO
3	M	Sports	Medium	CO
4	M	Sports	Large	CO
5	M	Sports	Extra Large	CO
6	M	Sports	Extra Large	CO
7	F	Sports	Small	CO
8	F	Sports	Small	CO
9	F	Sports	Medium	CO
10	F	Luxury	Large	CO
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

4. a) Mention the data preprocessing techniques used in text mining. 1
   b) With an example explain the Bag of Words (BoW) technique for text mining. Mention the limitations of this technique. 10+

5. a) Bitmap indexing is useful in data warehousing. Mention the advantages and problems of using a bitmap index structure.
   b) A data warehouse can be modeled by either a *star schema* or a *snowflake schema*. Briefly describe the similarities and the differences of the two models.
   c) What are the differences among *information processing*, *analytical processing*, and *data mining*? Discuss the motivation behind OLAP mining (OLAM). 7+

6. a) Explain the main building blocks of a multi-Tiered data warehouse architecture? 1
   b) Suppose that $x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$ and $y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$ are the two term-frequency vectors. Compute the cosine similarity between the two vectors.
   c) What are the various types of data sampling method? What are the benefits of applying dimensionality reduction to a dataset? 3+