BE6-R4- DATA WAREHOUSING AND DATA MINING

NOTE:

1. **Answer question 1 and any FOUR from questions 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

Time: 3 Hours **Total Marks: 100**

1.

a) Define the term *data mining*? Write down two major challenges to data mining regarding user interaction.

b) What is hypothesis testing? Explain its usage with one example?

c) Define nominal attribute. Give the formula used for computing dissimilarity between two objects having nominal attributes only.

d) What do you mean by data sampling? Write down one advantages of using sampling over other data reduction techniques.

e) Describe the data warehouse model that is used to store or manage the data needed by a specific set of users.

f) Why is clustering termed as unsupervised learning technique?

g) What are Feed Forward Neural Networks?

(7x4)

2. Differentiate between (give examples for each case): -

a) Operational database system and data warehouse

b) Bootstrap method and K-fold cross validation

c) Dendrogram and Decision tree

(6+6+6)

3.

a) Why is cosine similarity referred as nonmetric measure?. Using the term-wise frequency vectors given for three documents, find out two most similar documents.

Document	Marks	Repeat	Math	Score	Student	Grade
Document 1	5	0	3	0	2	0
Document 2	3	0	2	0	1	1
Document 3	2	5	0	2	1	1

b) What is the need of data reduction in data pre-processing stage? Explain any two strategies used for data reduction.

c) Write any two ways to improve efficiency of Apriori algorithm.

(9+6+3)

4.

a) What is a data cube? How many cuboids are there in an n-dimensional cube?

b) For the given Weather dataset with Play as a class label, do the following:

Day	Outlook	Temperature	Humidity	Windspeed	**Play**
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes

D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes

i) Compute impurity of the dataset using Gini index. Also compute the best binary split for attribute Temperature.

ii) Consider the given rule R1 generated by rule-based classifier. Compute coverage and accuracy of rule R1.

R1: (outlook = rain) and (Temperature=Mild) ⟹ (Play = yes)

(6+[6+6])

5.

a) How Crossover and Mutation is performed in Genetic Algorithm? Explain with example.

b) How the clustering is different in case of large databases? Explain in detail?

(9+9)

6.

a) What is Naive bayes classifier? Write down the assumption used while building this classifier. How does this assumption hamper its effectiveness?

b) How is clustering different from classification? Write down four major categories of clustering methods. Name the category to which following algorithms belong to:

i) k-Means
ii) BIRCH
iii) STING
iv) DENCLUE

(9+9)

7.

a) Give an overview of the IBM data mining tools explaining the basic layout and model which is Supported by the tool.

b) Explain Spatial Rules? Also explain how spatial classification, clustering and association is done in spatial mining?

c) Define the term OLAP. Differentiate between roll-up and drill-down operations.

(6+6+6)