

[This question paper contains 12 printed pages.]

Your Roll No.....22075570085-

Sr. No. of Question Paper : 4192

H

Unique Paper Code : 2343012005

Name of the Paper : Data Mining – I

Name of the Course : **B.Sc. (Hons.) Computer Science**

Semester : IV

Duration : 3 Hours Maximum Marks : 90

### **Instructions for Candidates**

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. **Section A** (Question No. 1) is compulsory.
3. Attempt any **four** questions from **Section B** (Questions 2 to 7).
4. The use of a simple calculator is allowed.
5. Parts of the question must be answered together.

**Section A**

1. (a) Differentiate between the unsupervised and supervised evaluation measures used for cluster validity. (3)
- (b) What is the anti-monotone property of the support measure in association rule mining? Does the confidence measure follow anti-monotone property? (3)
- (c) Consider a dataset with two class labels, News and Entertainment, and six labeled documents D1-D6. A new document, D7, is to be classified. The similarity values of D7 with D1, D2, D3, D4, D5 and D6 are 0.75, 0.85, 0.66, 0.87, 0.70 and 0.84 respectively. Using the k-Nearest Neighbor classifier, predict the class label that should be assigned to D7 when k=3. Will the predicted class label change with k=5? (4)

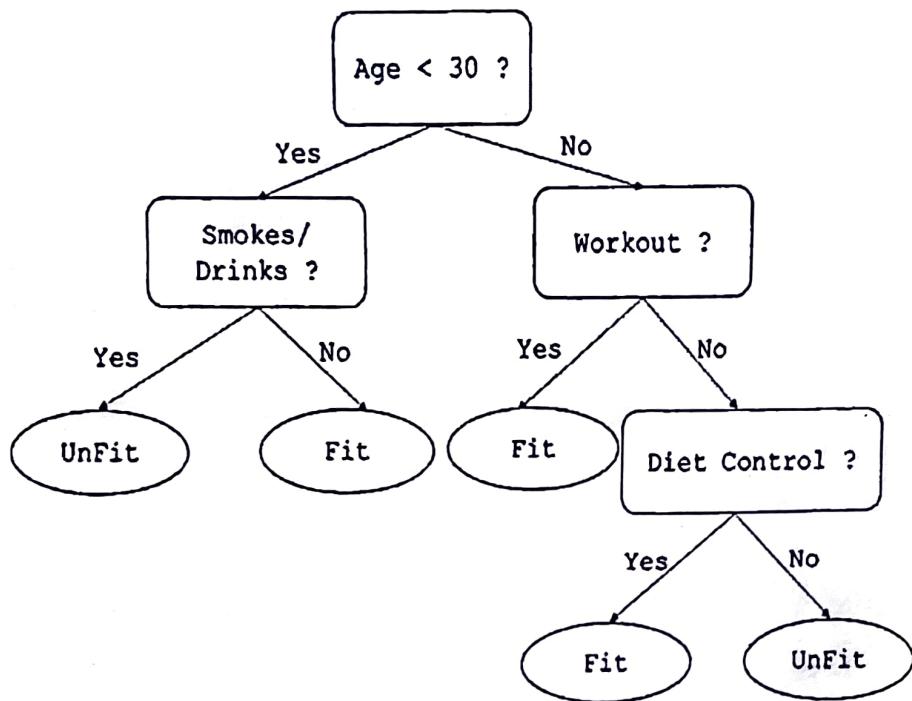
Document	Class Label
D1	News
D2	Entertainment ✓
D3	Entertainment
D4	News ✓
D5	News
D6	Entertainment ✓

(d) Consider the given dataset, which contains six objects, each with two attributes: Age and Salary. K-means clustering is used to cluster the given objects. Do you see any issue with applying K-means to the given dataset? If yes, then state the issue. Also apply the appropriate preprocessing technique to overcome it. If no, state explicitly that no preprocessing technique is required. (4)

	Age (in years)	Salary (in rupees)
Object 1	40	62000
Object 2	24	48000
Object 3	30	54000
Object 4	35	67000
Object 5	46	80000
Object 6	34	66000

(e) Define the curse of dimensionality. The Iris flower dataset comprises of 150 data points and four features, namely sepal length, sepal width, petal width, and petal length. Is it a high-dimensional data or low-dimensional data? Justify your answer. (4)

(f) Consider a decision tree to classify the health of an individual as Fit or Unfit given below :



(i) Extract all classification rules from the decision tree.

(ii) Classify the following object:

Age = 50, Workout = No, Smokes/Drinks = No, Diet Control = No, Health = ? (4)

(g) Classify the following tasks as “predictive” or “descriptive”. Justify your answer. (4)

(i) Foretelling whether an online user will shop on Flipkart for a specific item.

- (ii) Grouping the customers of a company according to their buying interests.
- (iii) Finding a group of genes such that genes in each group have related functionality.
- (iv) Using historical data from previous financial statements to project sales, revenue, and expenses for a company.
- (h) Given two objects  $X = (22, 1, 42, 10)$  and  $Y = (20, 0, 36, 8)$ , compute the distance between these two objects using the following distance measures :
- (i) Euclidean Distance
- (ii) Manhattan Distance (4)

## Section B

2. (a) Given the following training dataset, compute all class conditional and prior probabilities. Use the Naive Bayes approach to predict the class label (Salary) for the test instance : (12)

Education Level = PG, Career = Management,  
 Years of Experience = 3 to 10

Education Level	Career	Years of Experience	Salary
UG	Management	Less than 3	Low
UG	Management	3 to 10	Low
PG	Management	Less than 3	High
PG	Service	More than 10	Low
UG	Service	3 to 10	Low
PG	Service	3 to 10	High
PG	Management	More than 10	High
PG	Service	Less than 3	Low
UG	Management	More than 10	High
UG	Service	More than 10	Low

$$P(X=1|Y) = P(Y=1) \times P(X=1|Y)$$

(b) A data mining application uses a particular type of data. Give one application for each of the following type : (3)

(i) Sparse dataset

(ii) Spatio-Temporal data

(iii) Graph-based data

3. (a) Consider the following dataset having details about different departments of a company :

ID	Dept. Name	Location	Established On	Size	Annual Budget
DP12	Finance	Nehru Place	5-01-2020	Large	460
DP19	Marketing	Nehru Place	8-08-2020	Medium	300
DP21	Human Resource	Hauz Khas	2-01-2020	Medium	240
DP27	Production		2-02-2020	Medium	290
DP33	Research & Development	Nehru Place	4-07-2021	Small ✓	90
DP39	Information Technology	Hauz Khas	6-08-2020	Medium	210
DP41	Sales	Nehru Place	9-09-2020	Large	510
DP52	Customer Service	Hauz Khas	2-10-2020	Medium	
DP55	Public Relations	Nehru Place	3-03-2021	Large	900

\* Annual Budget is In Lakhs

- (i) Identify the type of attributes ID, Dept. Name, Location, Established On, Size, and Annual Budget as nominal, ordinal, interval, or ratio. Give justification for each. (6)
- (ii) Suggest a technique for dealing with missing values in the attribute Location. Will the same technique apply to the attribute Annual Budget? Justify. (3)

(iii) What is an outlier? Spot an outlier in the provided dataset. (3)

(b) What is the need for sampling in data mining ? What problems arise if the sample size is too small or too large? (3)

4. Consider the following transactional data of a grocery store :

Transaction ID	Items
T1	Boots, Hoodie, Gloves
T2	Boots, Hoodie
T3	Hoodie, Coat, Cardigan
T4	Cardigan, Coat
T5	Cardigan, Gloves
T6	Hoodie, Coat, Cardigan

(a) What is the maximum number of rules that can be extracted from this data (including rules that have zero support).  $3^d - 2^{d+1} + 1$  (3)

(b) Use the Apriori algorithm on the given transactional dataset and compute the candidate and frequent itemsets for each dataset scan. Assume a support threshold of 33.34%. (6)

(c) Enumerate all association rules generated from the largest frequent itemset found in each dataset scan. Compute the confidence of each generated rule. Assuming that the minimum confidence threshold is 70%, find all the strong association rules.

(6)

~~5.~~ (a) A medical team develops classification models for predicting the occurrence of a "genetic disorder" using Classifier A and Classifier B. Patients having genetic disorders are considered positive instances. In contrast, negative instances are ones with the absence of genetic disorders. The classifiers were tested on data from 500 patients and then obtained the result as :

	Actual Label	
	Presence of Genetic Disorder	Absence of Genetic Disorder
Classifier A, predicted "presence of genetic disorder"	131	155
Classifier A, predicted "absence of genetic disorder"	19	195
Classifier B, predicted "presence of genetic disorder"	82	72
Classifier B, predicted "absence of genetic disorder"	68	278

(i) List the confusion matrix for “Classifier A” and “Classifier B”. Find the accuracy, precision, sensitivity, recall and specificity for each classifier. (8)

(ii) What problem may occur if the provided training dataset of 500 patients had only 15 positive instances and the remaining negative instances? Which performance measure would you choose to evaluate the classifiers in such a scenario? Which is the better classifier between Classifier A and Classifier B in such a scenario?

(4)

(b) Consider a categorical attribute Grade with three values {A, B, and C}. Convert this attribute to asymmetric binary attributes. (3)

6. Consider the given COVID-19 dataset of ten patients.

ID	Age	Fever	BD	Outcome
P1	Young	Yes	High	In ICU ✓
P2	Young	No	High	Hospitalized ✗
P3	Elderly ✓	Yes	High	In ICU ✓
P4	Middle aged	Yes	Moderate	In ICU ✓
P5	Middle aged	No	High	Home Care
P6	Middle aged	Yes	Moderate	In ICU ✓
P7	Elderly ✓	No	Moderate	In ICU
P8	Elderly ✓	No	High	Deceased
P9	Elderly ✓	Yes	High	In ICU ✓
P10	Young	No	High	Hospitalized ✗

BD: Breathing Difficulty

- (a) Compute the Gini Index of Age, Fever, and BD attributes. Given that you construct a decision tree using the Gini Index as the splitting criteria, which of the three attributes would you choose at the root? Justify your choice. (9)
- (b) Compute the Gini Index of ID. Why should it not be used as a splitting attribute for constructing a decision tree? (3)
- (c) Given ten objects in the dataset (P1 – P10), mention all train and test distributions for performing k-fold cross-validation. Assume the value of k = 5. (3)

**4192**

**12**

- Given a dataset with six records about startup companies, each record has two fields: Number of Clients and Annual Turnover. Assuming that  $k=2$  and initial cluster centres as the first two records, compute the cluster centres of the resulting clusters until the stopping criterion is met. Use Euclidean distance as the distance metric. Also, compute the SSE (Sum of Squared Error) of each generated cluster.

Number of Clients	Annual Turnover (in Lakhs)
185	72
170	56
168	60
179	68
182	72
188	77

(15)

(1000)