Question Paper

Exam Date & Time: 20-Jun-2024 (02:30 PM - 05:30 PM)

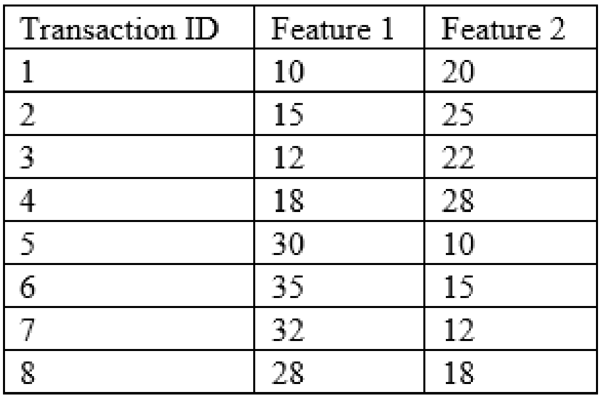
MANIPAL ACADEMY OF HIGHER EDUCATION

SIXTH SEMESTER B.TECH. (INFORMATION TECHNOLOGY) DEGREE EXAMINATIONS - JUNE 2024
SUBJECT: ICT 3253/ICT_3253 - DATA WAREHOUSING AND DATA MINING

Marks: 50 **Duration: 180 mins.**

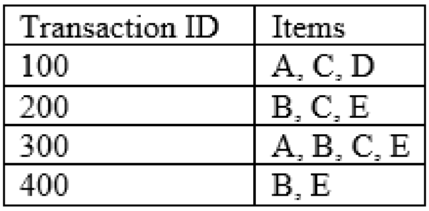
Answer all the questions.

1A) Apply K-mediod clustering, for the given dataset. Consider TID 2 and 7 as initial centroids, whether swapping 7 with 8 improves efficiency? Justify. (5)



Transaction ID	Feature 1	Feature 2
1	10	20
2	15	25
3	12	22
4	18	28
5	30	10
6	35	15
7	32	12
8	28	18

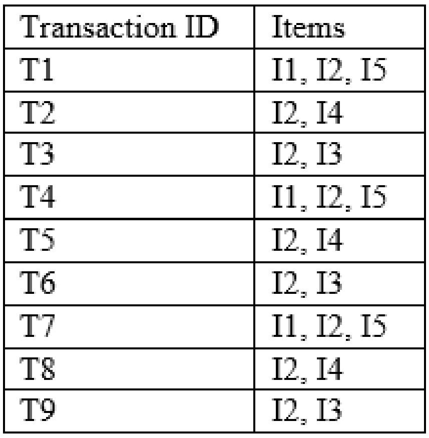
1B) Consider the Data set D. Given the minimum support 2, apply apriori algorithm on this dataset (3)



Transaction ID	Items
100	A, C, D
200	B, C, E
300	A, B, C, E
400	B, E

1C) Justify how underfitting can occur and discuss strategies to prevent it. (2)

2A) In an online retail store, you have transactional data containing customer purchases. Partitioning Algorithm can be applied to identify frequent itemsets. Consider minimum support as 2 and minimum confidence as 70% Also find strong association rules. (5)



Transaction ID	Items
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I5
T5	I2, I4
T6	I2, I3
T7	I1, I2, I5
T8	I2, I4
T9	I2, I3

2B) A data set for analysis includes only one attribute X: (3)
X={ 7,12,5,8,5,9,13,12,19,7,12,12,13,3,4,5,13,8,7,6}
i) What is the mean and median of the data set X?
ii) Find the standard deviation for X
iii) Give the five-number summary of the data.

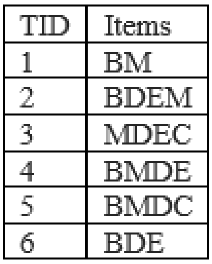
2C) In a sales data cube with dimensions for Time (Year, Quarter, Month), Product (Category, Subcategory), and Region (Country, State), provide a scenario where a dice operation would be useful for analysis. (2)

3A) Suppose a dataset containing the attributes: City, Country, Street, and Postal Code, along with their respective counts of distinct values: City (40), Country (4), Street (200), and Postal Code (500) is provided. (5)
i) Construct a concept hierarchy based on the heuristic that an attribute with fewer distinct values indicates a higher-level concept. Illustrate this hierarchy with a diagram and explain how this structure could optimize query processing and data analysis in a geographic information system (GIS). Provide specific examples to support your explanation.
ii) Discuss the limitations of using this heuristic for generating concept hierarchies and evaluate potential strategies to overcome these challenges. Include how domain expertise and manual adjustments might enhance the practical utility of the hierarchy.

3B) You have been provided with the test scores of 20 students from a mathematics course, listed as follows: [55, 58, 62, 65, 68, 70, 72, 75, 78, 80, 82, 85, 88, 90, 92, 94, 96, 98, 100, 102]. (3)
i) Calculate the range, first quartile (Q1), third quartile (Q3), and the interquartile range (IQR) of the test scores. Illustrate your process and results.
ii) Analyze the calculated quartiles and IQR in relation to the overall range to determine the distribution and concentration of test scores. What does this analysis tell you about the performance variation among the students?
iii) Based on the statistical measures and their analysis, recommend targeted educational interventions that could address the specific needs of students in different performance tiers. Justify.

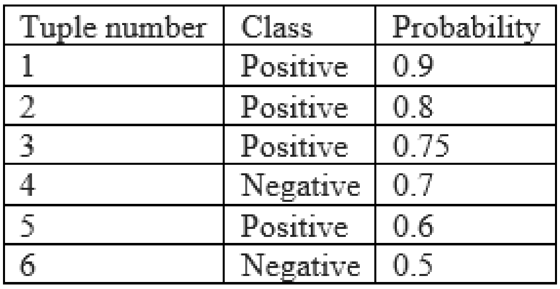
3C) Assume you are provided with the average daily rainfall data (in millimeters) for the month of August over 10 years in a specific city. The data, in value-ascending order, are: 0.0, 0.5, 0.5, 0.8, 1.0, 1.0, 1.1, 1.1, 1.2, and 5.0. Analyze this dataset under the assumption that it follows a normal distribution: (2)
i) Calculate the mean (μ) and variance (σ^2) of this dataset using the maximum likelihood estimation method.
ii) Identify any outliers in the dataset using the calculated mean and standard deviation. Consider a point an outlier if it lies outside the μ ± 3σ range.

4A) Determine frequent itemsets for the transaction data given below by applying Pincer search algorithm with the minimum support threshold as 2. (5)



TID	Items
1	BM
2	BDEM
3	MDEC
4	BMDE
5	BMDC
6	BDE

4B) Consider the results returned by a probabilistic classifier as given in the following table. Compute true positive, false positive, true negative, false negative, true positive rate, and false positive rate for each tuple. Draw a neat ROC curve and comment on the accuracy of the model based on the ROC curve. (3)



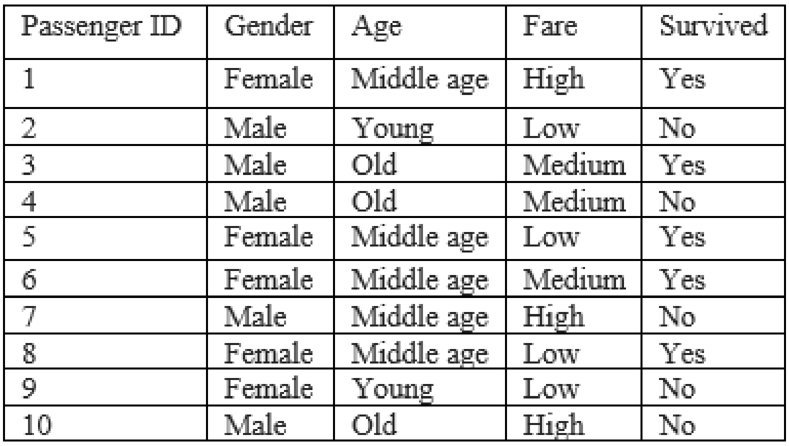
Tuple number	Class	Probability
1	Positive	0.9
2	Positive	0.8
3	Positive	0.75
4	Negative	0.7
5	Positive	0.6
6	Negative	0.5

4C) Consider a dataset from a public health study that includes an attribute representing "Blood Pressure" of participants measured during various visits. The "Blood Pressure" readings are categorized into 'Normal', 'Elevated', 'Stage 1 Hypertension', and 'Stage 2 Hypertension'. (2)

i) Discuss potential classifications(attribute type) for the "Blood Pressure" attribute and provide your rationale for each classification possibility.

ii) Analyze how the interpretation of this attribute's type might influence the choice of statistical methods for studying the relationship between blood pressure categories and the effectiveness of different hypertension treatments.

5A) Consider a decision tree or predicting whether a passenger on the titanic survived or not based on the features given in the following table. Find the best feature to be considered as the root for this decision tree using information gain measure. (5)



Passenger ID	Gender	Age	Fare	Survived
1	Female	Middle age	High	Yes
2	Male	Young	Low	No
3	Male	Old	Medium	Yes
4	Male	Old	Medium	No
5	Female	Middle age	Low	Yes
6	Female	Middle age	Medium	Yes
7	Male	Middle age	High	No
8	Female	Middle age	Low	Yes
9	Female	Young	Low	No
10	Male	Old	High	No

5B) A data warehouse for Revenue consists of the four dimensions, Dealer, Date, Branch_location, and Product, and the two measures, units_sold and income, where income is the revenue that is obtained by selling the product on a given date. Dealers may be UG_students, PG_students, or Employees, with each category having its own income rate. (3)

i) Draw the star schema for the Revenue data warehouse.

ii) What are the OLAP operations (from base cuboid) needed to find the total income from UG_students at Manipal in 2024.

5C) Given the classifications of outlier detection methods into supervised, semi-supervised, and unsupervised approaches: (2)

Analyze the advantages and disadvantages of each outlier detection method in the context of a large, diverse dataset typically found in social media analytics for sentiment analysis. Specifically:

i) Discuss how the inherent imbalance between normal objects and outliers impacts the effectiveness of supervised methods in this context.

ii) Evaluate the potential for high false positive rates in unsupervised methods when applied to social media data, which often lacks clear clustering of normal objects.

-----End-----