Question Paper

Exam Date & Time: 04-May-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

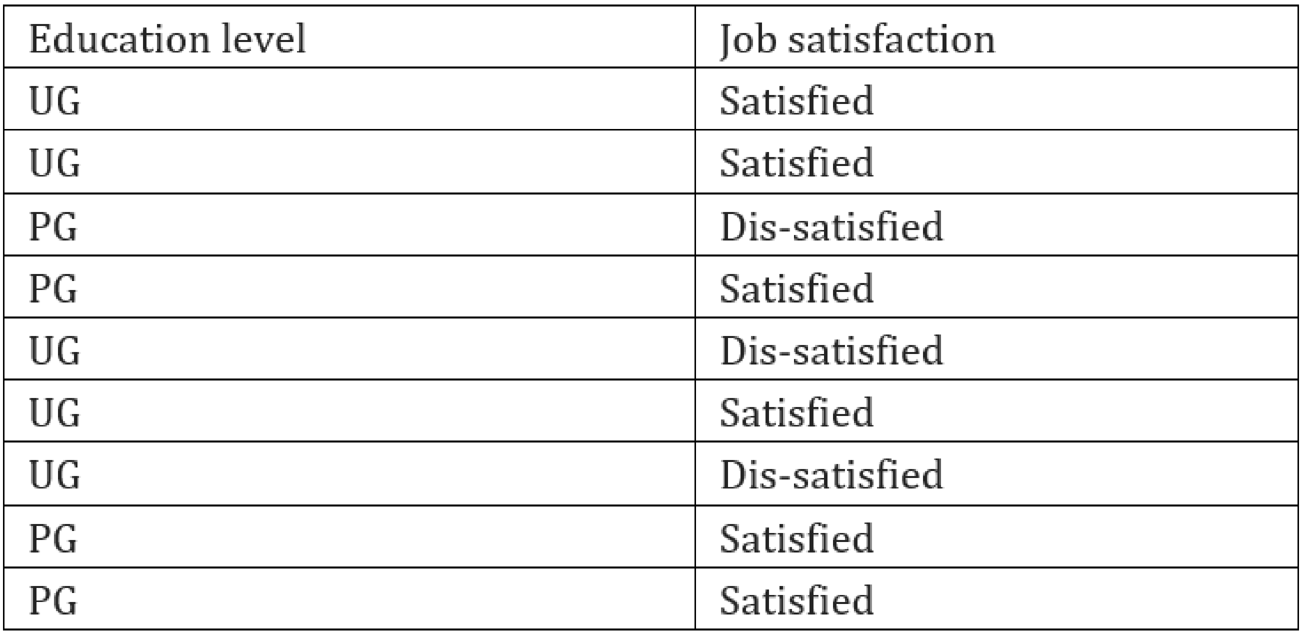
SIXTH SEMESTER B.TECH. (INFORMATION TECHNOLOGY) DEGREE EXAMINATIONS - APRIL / MAY 2024
SUBJECT: ICT 3253/ICT_3253 - DATA WAREHOUSING AND DATA MINING

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

Answer all the questions.

1A) Cluster { (1,2), (2,2), (3,2), (3,3), (8,7), (7,8), (9,9) } by using k-medoid algorithm by assuming initial medoids as (2,2) and (9,9). Determine the clusters after swapping (9,9) with (7,8). Compare the costs of initial cluster and the cluster formed after swapping the medoid. (5)

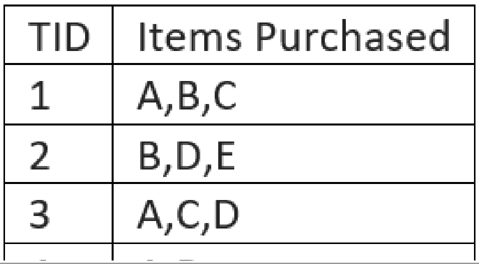
1B) Consider a survey report showing education level and job satisfaction of participants as given in the following table. Construct a contingency table and determine whether education level and job satisfaction have statistically significant relationship by using chi-squared measure and cosine measure. (3)



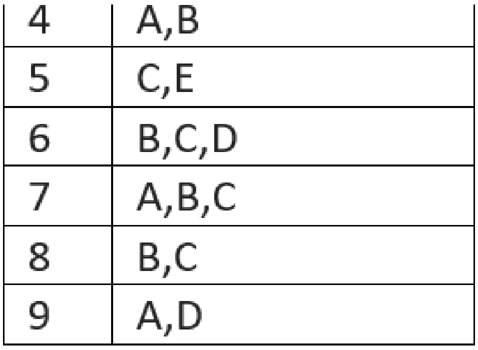
Education level	Job satisfaction
UG	Satisfied
UG	Satisfied
PG	Dis-satisfied
PG	Satisfied
UG	Dis-satisfied
UG	Satisfied
UG	Dis-satisfied
PG	Satisfied
PG	Satisfied

1C) An online retailer collects data on customer browsing patterns, purchase history, product preferences, and feedback. The company aims to use this data to enhance user experience, increase sales, and reduce logistic costs. Identify and describe three specific data mining tasks that could be implemented to meet these objectives. Analyze why these tasks are suitable for the stated goals. (2)

2A) Determine frequent itemsets for the transaction data given below by applying apriori algorithm with the minimum support threshold as 2. Find all strong association rules by assuming minimum confidence threshold as 50% (5)



TID	Items Purchased
1	A,B,C
2	B,D,E
3	A,C,D



4	A,B
5	C,E
6	B,C,D
7	A,B,C
8	B,C
9	A,D

2B) Consider a dataset containing daily foot traffic counts in a mall over three months: [1200, 1180, 1220, 1250, 1300, 1280, 1260, 1250, 1230, 1210, 1150, 1180, 1200, 1190, 1170, 1150, 1130, 1100, 1120, 1150, 1170, 1200, 1220, 1250]: (3)

i) Analyze the effects of binning by means using a bin depth of 7 days on this dataset. Outline the steps for implementing this method and discuss how it may impact the understanding of weekly traffic patterns.

ii) Discuss the use of binning by boundaries to manage noise within the dataset. Evaluate how this method compares to binning by means in terms of preserving data integrity during outlier occurrences.

2C) A multinational retailer's data warehouse contains sales data categorized by Product (Type, Brand), Time (Year, Quarter, Month), and Sales Territory (Region, Country). Consider the following OLAP operations performed on this data cube: (2)

- Drill-down on Time from Year to Quarter.
- Slice for sales in the "Electronics" type.
- Roll-up on Sales Territory from Region to Country.

i) Analyze how the sequence of these OLAP operations affects the analysis of quarterly sales trends for electronics across different countries.
ii) Propose an alternative sequence of OLAP operations that could potentially provide clearer insights into electronics sales performance by country for each quarter. Provide a rationale for the order you propose.

3A) Consider a multinational corporation's human resources database that includes the attributes: Employee, City, State, and Country. The database contains extensive data across many global locations. (5)

i) Assuming the schema already establishes a hierarchy of City < State < Country, define explicit groupings for a portion of this hierarchy to include intermediate regional levels such as "Northeastern United States" and "Southern Europe". Provide examples of how these groupings might be structured.

ii) Discuss how the addition of these intermediate regional levels could enhance data analysis and reporting capabilities within the corporation. Provide specific examples of analytical queries that would benefit from these new hierarchical groupings.

3B) A health clinic has recorded the number of daily patient visits over 30 days: [15, 18, 22, 25, 28, 30, 30, 32, 35, 35, 38, 40, 40, 42, 45, 45, 48, 50, 50, 53, 55, 58, 60, 62, 65, 68, 70, 72, 75, 78]. (3)

i) Compute the range, first quartile Q1), third quartile Q3), and the interquartile range (IQR) for the number of daily visits.
ii) Analyze the quartile spread and IQR within the context of the overall range. What does this analysis reveal about patient visit patterns at the clinic?
iii) Based on your analysis, propose specific operational adjustments the clinic could make to better accommodate the observed patient visit patterns. Justify your proposals with the statistical insights derived.

3C) Consider the following scenarios: (2)

- Clinical Data Analysis: Where even small deviations in patient vital signs can indicate critical conditions that may require immediate medical attention.
- Marketing Data Analysis: Involving customer purchase histories where variations are larger due to consumer behavior diversity and promotional impacts.

Given these scenarios, analyze how the approach to outlier detection would differ between clinical data analysis and marketing data analysis. Specifically:
i) Identify the potential consequences of mislabeling outliers and noise in these two applications.

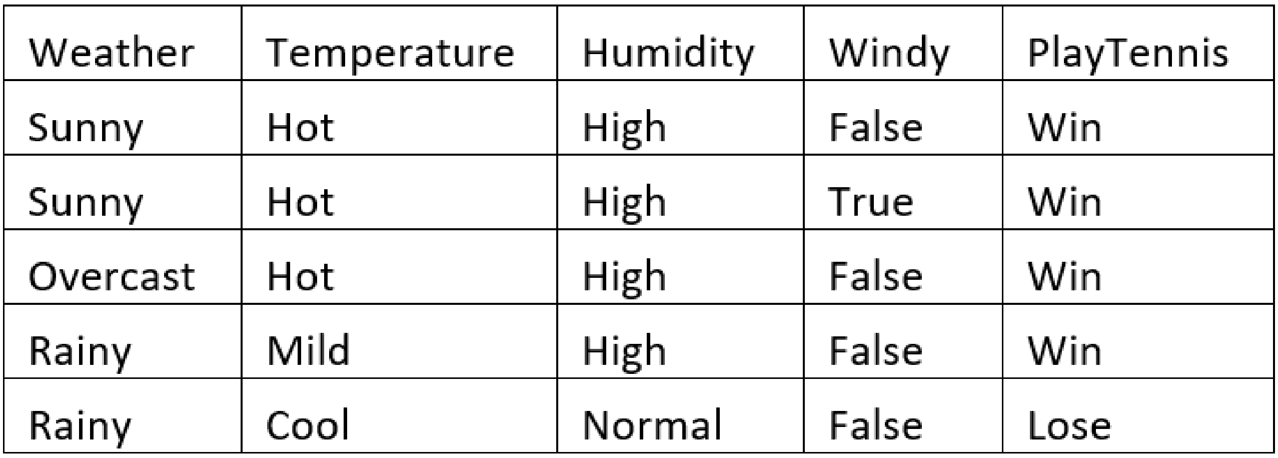
ii) Provide recommendations on the types of outlier detection methods and distance/similarity measures that would be most effective for each scenario, considering their specific requirements and challenges.

4A) Illustrates how the Pincer Search Algorithm can be applied to optimize a dataset by maximizing the commonality among transactions. Consider minimum support as 2. (5)

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

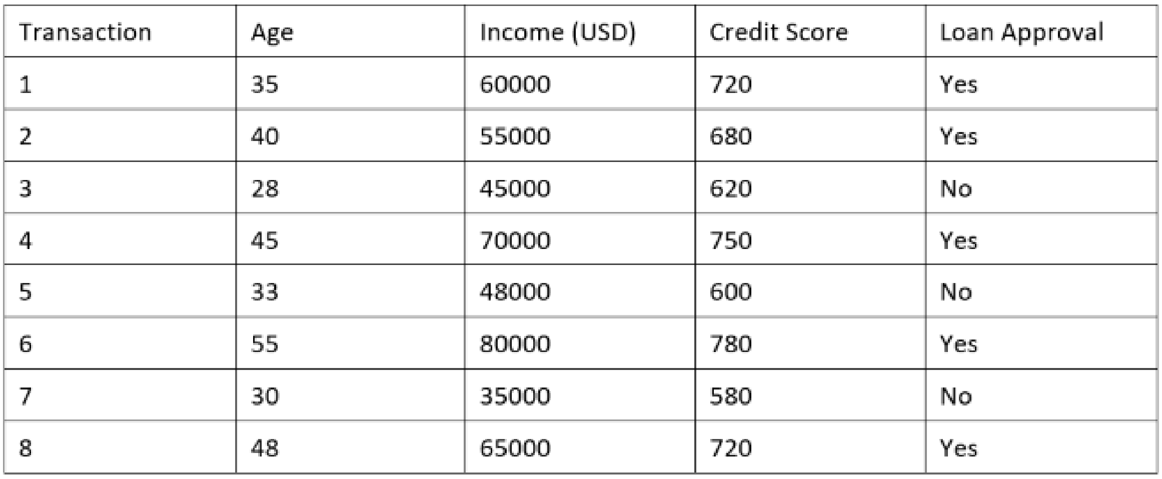
4B) Construct decision tree, by using concepts of gain ratio for the given data. (3)



Weather	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	Win
Sunny	Hot	High	True	Win
Overcast	Hot	High	False	Win
Rainy	Mild	High	False	Win
Rainy	Cool	Normal	False	Lose

4C) Identify the type of attribute (nominal, ordinal, interval, or ratio) for each of the attributes listed above. (2)
i) Age
ii) Gender

5A) Build a predictive model using a decision tree algorithm to classify customers as high-risk or low-risk for loan approval based on their credit history and financial attributes. Evaluate the model's performance using appropriate metrics such as accuracy, precision, and recall. (5) (5)



Transaction	Age	Income (USD)	Credit Score	Loan Approval
1	35	60000	720	Yes
2	40	55000	680	Yes
3	28	45000	620	No
4	45	70000	750	Yes
5	33	48000	600	No
6	55	80000	780	Yes
7	30	35000	580	No
8	48	65000	720	Yes

5B) You are tasked with designing a star schema for a data warehouse to support an online clothing retailer's needs for analyzing sales performance, managing inventory, and refining marketing strategies. The retailer requires detailed analyses on sales trends, customer demographics, product performance, and seasonal impacts. Develop a star schema for the above scenario (3)

5C) Discuss the challenges and limitations associated with outlier detection, particularly in large and high-dimensional datasets. (2)

-----End-----