Satyandra Pal
22234768023
Vikas Sir

Master of Computer Applications

MCAC 204: Data Analysis and Visualization
Unique Paper Code: 223421204
Semester II
August-2023

Time: Three Hours **Maximum marks: 70**

Note: Answer all the questions. Attempt all parts of a question together.

1. You are a data scientist working for an organization that sells life insurance policies to its customers. In order to make any new policy popular, the company sends product details to its existing customers, where sending a message has an associated cost. The organization has recently realized that a large amount can be saved by sending messages to only selected customer groups having similar purchasing behaviors. Your task is to apply the k-means clustering algorithm to segment the customers provided in Table 1, which will later be used for targeted marketing. The table contains ten customer details over the attributes: Age, Annual Income (in lakh), and Number of Policies Purchased in the Last Year.

Table 1: Customer Details

Customer	Age	Annual Income	Policies Purchased
C1	25	45	2
C2	35	75	5
C3	45	60	3
C4	50	55	4
C5	28	50	1
C6	55	85	6
C7	40	70	3
C8	30	65	2
C9	32	70	4
C10	48	80	5

You have to apply k-means clustering algorithm to segment the customers into three groups based on their purchasing behaviors. What are the values of c_1, c_2, and c_3 after the second iteration of k-means? Initialize the k-means clustering algorithm with 3 cluster centers $c_1 = C2$, $c_2 = C6$, and $c_3 = C8$ and show the steps of the algorithm, including the assignment of customers to clusters and the centroid updates.

2. In the game of cricket, players' fitness is one of the major concerns. Most often, players get injured because of bad playing conditions. To overcome this, the club has hired you to forecast the playing conditions based on the weather data. The club also provided you with the weather data for the last seven days (Table 2, Decision attribute: *Play*). You then decided to build a decision tree classifier that is at least two levels deep (root at level 1) with information gain as an attribute test condition.

   (a) Write the step-by-step decision tree construction process. [10]

   (b) Suppose that the data entry operator has missed entering the first row's value for Humidity. Describe two ways in which you can handle the missing value. [4]

Table 2: Weather Dataset

Humidity	Outlook	Windy	Play
Low	Overcast	No	Yes
Low	Overcast	Yes	Yes
High	Sunny	No	Yes
High	Sunny	Yes	No
Low	Sunny	No	No
Low	Sunny	Yes	Yes
High	Overcas	No	No

3. Consider a dataset that contains information about sales transactions over the attributes: *date*, *product_name*, *product_category*, *quantity_sold* and *sales_amount*. Your task is to load the data into a pandas DataFrame and perform the following basic data analysis.

   (a) Display the first 10 rows of the DataFrame. [3]

   (b) Find the average quantity sold per transaction. [3]

   (c) Add a new column "total_revenue" by multiplying "sales_amount" with "quantity_sold".[3]

   (d) Generate a bar chart displaying the total sales amount for each "product_category".[3]

   (e) How will you fill the missing values with appropriate data? [3]

4. You have to analyze a dataset containing information about a retail company's sales over the past year. The dataset includes details about products, sales dates, quantities sold, and prices. The company is interested in understanding sales trends, identifying top-selling products, and predicting future sales.

   (a) Explain the importance of exploratory data analysis in the context of this dataset. [6]

   (b) Outline the steps you would take in the data science process to address the company's objectives of understanding sales trends, identifying top-selling products, and predicting future sales. [9]

5. You have a dataset containing information about students' test scores. The dataset includes columns namely *Subject*, *Topic*, *StudentID*, and *Score*.

   (a) Create a hierarchical index on the DataFrame using 'Subject' and 'Topic' as levels. Explain the benefits of using hierarchical indexing in this scenario. [5]

   (b) Calculate the average score for each 'Subject' and 'Topic' using the hierarchical index you created. [3]

   (c) Identify the top-3 highest-scoring subjects based on the average score. [3]

   (d) Create a pivot table to display the average score for each 'Subject' and 'Topic', using 'Subject' as rows and 'Topic' as columns. [4]