

SPRING SEMESTER 2024

DSA2040: DATA MINING AND WAREHOUSING

INSTRUCTOR: Dr. Mursi

DATE: 20th FEBRUARY 2024

Duration: 1530hrs to 1730hrs

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

MIDSEMESTER

Total marks (30)

Instructions:

1. Attempt all the questions
2. Paste your code and output below each question
3. Show all the working USING R no marks will be awarded for solutions/ outputs only
4. Upload all your work via Blackboard. DON’T EMAIL your work
5. Make sure you upload the right document

## *Question 1 (9mks)*

1. *Differentiate the three Anomaly detection techniques* ***(5mks)***
2. *A financial institution analyzes customer transactions to identify potential fraud. How could anomaly detection be used in this situation? How would you balance the need for accurate fraud detection with the risk of wrongly flagging legitimate transactions?* ***(4mks)***

## *Question 2 (12mks)*

1. *Using the Groceries dataset, find the top 5 most frequent itemsets with a minimum support threshold of 0.09 using the Apriori algorithm. Provide the itemsets and their corresponding support values.* ***(3mks)***
2. *Using the same Groceries dataset, identify the top 5 most frequent itemsets with a minimum support threshold of 0.09 using the ECLAT algorithm. Compare the results with those from the Apriori algorithm.* ***(2mks)***
3. *Using the Groceries dataset, find all association rules where* ***soda*** *is part of the LHS. Use support threshold of 0.05. & confidence of 0.05. What is the highest support value among these rules?* ***(3mks)***
4. *Generate rules from the Groceries dataset where coffee is only in the RHS. List the top 5 rules by their support values.* ***(4mks)***

## *Question 2 (9mks)*

1. *Using the airquality dataset, use the Z-score method to detect outliers in the Ozone column. Consider observations with a Z-score greater than 2 or less than -2 as outliers. How many outliers are there?* ***(3mks)***
2. *Discuss the role of a data warehouse in supporting the data mining process for a global retail chain. How can data warehousing facilitate the extraction of actionable business intelligence? Consider Data Integration, Historical Data Analysis, Data Quality Scalability and Performance in your answer* ***(6mks)***