Code No: R1632052 **R16** SET - 1

III B. Tech II Semester Supplementary Examinations, February-2022

DATA WAREHOUSING AND DATA MINING

(Computer Science and Engineering)

Time: 3 hours Max. Marks: 70

Note: 1. Question Paper consists of two parts (**Part-A** and **Part-B)**
2. Answer **ALL** the question in **Part-A**
3. Answer any **FOUR** Questions from **Part-B**

~~~~~~

PART –A **(14 Marks)**

1. a) Mention the significant applications of data mining. [2M]
   b) What is data discretization and when is it needed? [2M]
   c) How to verify that the classification model is Over fitting or Under fitting? [2M]
   d) Define Prior and Posterior probabilities. [3M]
   e) What are Frequent and Closed item sets in association mining? [3M]
   f) What are the four types of linkages used in Hierarchical clustering? [2M]

PART –B **(56 Marks)**

2. a) What do you mean by data mining? What kind of data and patterns can be mined? What are the major challenges of data mining? [7M]
   b) How do you measure the similarity between data objects? Discuss various measures used for this. [7M]

3. a) Why is data preprocessing became an inevitable phase in the knowledge discovery process? Explain about the major tasks in data preprocessing. [7M]
   b) Explain about Min-Max and Z-score normalization techniques with an example. [7M]

4. a) Explain the step by step approach of ID3 algorithm to build a decision tree classification model. [7M]
   b) How is a splitting point chosen for continuous variables in decision trees? [3M]
   c) Define the following: i) Entropy ii)Information Gain [4M]

5. a) State Bayes' theorem. Explain various classification models based on this theorem. [7M]
   b) Discuss various metrics used to evaluate the performance of a data classification model. [7M]

6. a) Explain the step-by-step approach to generate frequent patterns using FP-growth algorithm. [10M]
   b) Generate association rules from the frequent item set $\{I_1, I_2, I_5\}$ by assuming 0.8 confidence. [4M]

7. a) Demonstrate k-means clustering technique and also discuss its strengths and weaknesses. [7M]
   b) Explain the DBSCAN clustering technique and also derive its time and space complexities. [7M]

*****
~~~~~~