Code No: R2031053 **R20** SET - 1

III B. Tech I Semester Regular/Supplementary Examinations, December -2023
DATA WAREHOUSING AND DATA MINING
(Computer Science and Engineering)

Time: 3 hours Max. Marks: 70

Answer any **FIVE** Questions **ONE** Question from **Each unit**
All Questions Carry Equal Marks
*****

UNIT-I

1. a) Draw an architecture of typical data mining system with a neat sketch and explain its components. [7M]
   b) What are the various OLAP operations are used in the multidimensional data model? Explain them in detail with an example. [7M]

(OR)

2. a) Explain the data warehouse implementation. [7M]
   b) Briefly describe data mining functionalities. [7M]

UNIT-II

3. a) Briefly discuss the forms of Data preprocessing with neat diagram. [7M]
   b) What is the use of data discretization? Explain entropy based data discretization? [7M]

(OR)

4. a) What is data integration? Discuss the issues to be considered for data integration. [7M]
   b) Differentiate between data reduction and dimensionality reduction for data discretization [7M]

UNIT-III

5. a) What measures are used to find best split in Decision Tree Induction algorithm? How Can we improve the scalability in Decision Tree Induction algorithm? [7M]
   b) Explain in detail about Attribute Selection methods in Classification [7M]

(OR)

6. a) What is the role of data preprocessing in classification? Give examples of common data preprocessing tasks and explain how they impact the performance of a classification model. [9M]
   b) Discuss about contusion matrix in detail. [5M]

UNIT-IV

7. a) Discuss about basic concepts of frequent item set mining. [7M]
   b) What are the drawbacks of Apriori Algorithm? Explain. [7M]

(OR)

8. a) What are the advantages of FP-Growth algorithm? [7M]
   b) Write about basic concept in Association Rule mining. How many association rules can be generated for a given transactional database? [7M]

UNIT-V

9. a) What is meant by clustering? Explain the partitioning methods with an example. [7M]
   b) What is the drawback of k-means algorithm? How can we modify the algorithm to diminish? That problem? [7M]

(OR)

10. a) Explain K Means clustering method [7M]
    b) What is cluster analysis? Describe the types of data in cluster analysis [7M]

|'''|'|'|''||'''||||

Code No: R2031053 **R20** SET - 2

III B. Tech I Semester Regular/Supplementary Examinations, December -2023
DATA WAREHOUSING AND DATA MINING
(Computer Science and Engineering)

Time: 3 hours Max. Marks: 70

Answer any **FIVE** Questions **ONE** Question from **Each unit**
All Questions Carry Equal Marks
*****

UNIT-I

1. a) Discuss the major issues in data mining. [7M]
   b) Briefly explain about efficient computation of Data Cubes. [7M]

(OR)

2. a) Write the difference between OLTP vs OLAP [7M]
   b) Explain the basic elements of Data warehouse with a neat sketch. [7M]

UNIT-II

3. a) Explain about concept hierarchy generation for categorical data. [7M]
   b) Describe Data Transformation & Data Discretization. [7M]

(OR)

4. a) What is data reduction? Discuss about dimensionality reduction. [7M]
   b) What is Numerosity Reduction? What are the available techniques for numerosity reduction? [7M]

UNIT-III

5. Compare and contrast post-pruning and pre-pruning techniques in the context of data mining. How do these techniques help in addressing over fitting, and what are the potential consequences of not pruning a decision tree? [14M]

(OR)

6. a) Explain the classification by decision tree induction. [7M]
   b) Explain the purpose of "Attribute selection measures" in classification by decision tree induction? How we can use the "Tree pruning" in classification? [7M]

UNIT-IV

7. a) Consider the following transactional data for a commercial shop. [7M]

TID	List of Items with Ids
T1	I2,i4
T2	I1,i2,i5
T3	i2, i3
T4	i1, i3
T5	i1, i2, i4
T6	i2, i3
T7	i1, i3
T8	i1, i2, i3
T9	i1, i2, i3, i5

Generate all the frequent itemsets using Apriori algorithm. Consider the minimum support count is 2. Clearly show your computational steps.

   b) Explain Mining Frequent Patterns using FP-Growth. [7M]

(OR)

8. a) What are the various Constraints in Constraint based Association rule mining? Explain. [7M]

b) Explain how confident-based pruning is used in rule generation. What is the purpose of pruning rules, and how does it impact the quality and interpretability of the rules? [7M]

UNIT-V

9. a) Classify various Clustering methods. [7M]
   b) Write partitioning around mediods algorithm. [7M]

(OR)

10. a) Describe the concept of bi-secting K-means as an extension of the K-means algorithm. How does this technique aim to address some of the limitations of standard K-means? [7M]
    b) What is clustering and what is conceptual clustering? Describe dimensions and measures in a spatial data cube. [7M]

Code No: R2031053 **R20** SET - 3

III B. Tech I Semester Regular/Supplementary Examinations, December -2023
DATA WAREHOUSING AND DATA MINING
(Computer Science and Engineering)

Time: 3 hours Max. Marks: 70

Answer any **FIVE** Questions **ONE** Question from **Each unit**
All Questions Carry Equal Marks
*****

UNIT-I

1. a) Explain ROLAP, MOLAP and HOLAP. [7M]
   b) What are various schemas for multidimensional data models? [7M]

(OR)

2. a) Can you list the characteristic differences between OLAP and OLTP? [7M]
   b) With an example, describe snowflake and fact constellations. [7M]

UNIT-II

3. a) How can we smooth out noise in data cleaning process? Explain [7M]
   b) Normalize the following group of data by using the following techniques. 200, 300, 400, 600, 1000 [7M]
   i) min-max normalization technique ii)z-score normalization iii)Decimal scaling. Write your observations on the above techniques.

(OR)

4. a) What is data reduction? What is dimensionality reduction? What is lossless and lossy dimensionality reduction? [7M]
   b) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. [7M]
   i) Use min-max normalization to transform the value35for age onto the range [0:0; 1:0]. Ii)Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years. Iii)Use normalization by decimal scaling to transform the value 35 for age.

UNIT-III

5. a) What is the Basic Concept of Classification? [4M]
   b) Discuss in detail visual mining for decision tree induction. [10M]

(OR)

6. a) Explain the techniques and strategies used to improve the scalability of decision tree induction in data mining, such as parallelization and distributed computing. [7M]
   b) Describe the importance of visual mining in data mining, with a focus on decision tree induction. How can visual representations of decision trees aid in model interpretation and decision support? [7M]

UNIT-IV

7. a) Apply Apriori Algorithm in tracing all the frequent item datsets [7M]

Transactions	Itemset
T100	1 2 3
T200	2 3 5
T300	1 2 3 5
T400	2 5
T500	1 3 5

#Hint: Consider appropriate minimal support and minimal configure values for generating the rules.

b) Explain the closed item set and maximal item set concepts in the context of compact representation. What distinguishes closed item sets from maximal item sets, and why are they useful? [7M]

(OR)

8. a) Explain in detail the candidate generation procedures. [7M]

b) With an example, explain the Fp-growth algorithm? [7M]

UNIT-V

9. a) What are the main factors to consider when choosing the number of clusters (k) in K-means? Discuss common methods for determining the optimal value of k. [7M]

b) Discuss in detail additional issues of K- Means algorithm? [7M]

(OR)

10. a) Describe density-based clustering and its main idea. How does it identify clusters based on the density of data points rather than distance measures? [7M]

b) Define spherical, ellipsoidal, and arbitrary-shaped clusters. How do the shape and distribution of data points impact the choice of clustering techniques? [7M]

Code No: R2031053 **R20** SET - 4

III B. Tech I Semester Regular/Supplementary Examinations, December -2023
DATA WAREHOUSING AND DATA MINING
(Computer Science and Engineering)

Time: 3 hours Max. Marks: 70

Answer any **FIVE** Questions **ONE** Question from **Each unit**
All Questions Carry Equal Marks
*****

UNIT-I

1. a) What are the OLAP operations? Explain. [7M]
   b) Why data mining functionalities are used? Explain with an example data characterization and data discrimination. [7M]

(OR)

2. a) Illustrate on what kinds of patterns can be mined.. [7M]
   b) Differentiate operational database systems and data warehousing. [7M]

UNIT-II

3. a) Why preprocessing of data is needed? Discuss about data integration in detail. [7M]
   b) In real world data tuples with missing values for some attributes are common occurrence. Describe various methods for handling this problem. [7M]

(OR)

4. a) What is redundancy? Why correlation analysis is useful? Describe how correlation coefficient is computed? [7M]
   b) What are the value ranges of the following normalization methods? [7M]
   (i) min-max normalization (ii) z-score normalization (iii) normalization by decimal scaling

UNIT-III

5. a) How tree pruning in decision tree induction is useful? Explain various methods for pruning decision trees. [7M]
   b) Write a note on attribute selection measures. [7M]

(OR)

6. a) Explain decision tree induction algorithm for classifying data tuples and with suitable example [7M]
   b) Explain the concepts of class labels, features, and training data in a classification problem. How are these components used to train and evaluate classification models? [7M]

UNIT-IV

7. a) Discuss about closed frequent itemsets in detail. [7M]
   b) How can you find frequent itemsets using candidate generation? [7M]

(OR)

8. a) Write and explain the APRIORI algorithm with an example. [7M]
   b) Give examples of real-world applications where association analysis is commonly used. How does identifying associations between items or events benefit these applications? [7M]

UNIT-V

9. a) Discuss the importance of cluster analysis in real-world applications. Provide examples of industries or fields where cluster analysis is a valuable tool for gaining insights and making decisions. [7M]

b) Compare and contrast the advantages and disadvantages of K-means and bi-secting K-means in terms of performance and quality of clustering results. [7M]

(OR)

10. a) Explain the key goals of cluster analysis. How does it help in discovering hidden patterns in data and making data-driven decisions? [7M]

b) Explain the influence of the initial seed point selection on K-means clustering results. What strategies can be used to improve the reliability of K-means? [7M]