BRAINWARE UNIVERSITY

Term End Examination 2024-2025
Programme – B.Tech.(CSE)-AIML-2021/B.Tech.(CSE)-DS-2021/B.Tech.(CSE)-AIML-2022/B.Tech.(CSE)-DS-2022
Course Name – Data Mining and Data Warehousing
Course Code - PCC-CSM602/PCC-CSD602
(Semester VI)

Full Marks : 60 **Time : 2:30 Hours**

[The figure in the margin indicates full marks. Candidates are required to give their answers in their own words as far as practicable.]

Group-A

(Multiple Choice Type Question) 1 x 15=15

1. *Choose the correct alternative from the following :*

(i) Define the term "granularity" refer to in data warehousing?

a) The size of a data warehouse server b) The level of detail in the data
c) The speed of data retrieval d) The frequency of data updates

(ii) Which of the following OLAP operations involves viewing data from different angles or perspectives?

a) Slice b) Dice
c) Drill Down d) Roll up

(iii) Which is define Web Mining?

a) Extracting gold from the internet b) Extracting useful information or knowledge from the World Wide Web
c) Building websites for data collection d) Creating web-based video games

(iv) Select which algorithm is used for association rule mining in Web Usage Mining?

a) K-means b) Apriori
c) Decision Trees d) Support Vector Machines

(v) Define stemming in text mining.

a) Identifying the root form of words b) Analyzing the sentiment of text
c) Identifying named entities d) Classifying text documents

(vi) Tell which algorithm is commonly used for document clustering in text mining?

a) K-means b) Decision Trees
c) Support Vector Machines d) Naive Bayes

(vii) Give an example of a divisive method in hierarchical clustering.

a) DIANA (Divisive Analysis) b) Single Linkage
c) Ward's Method d) Complete Linkage

(viii) Classify a method used for evaluating the performance of clustering algorithms.

a) Silhouette Score
b) Accuracy
c) F1-Score
d) Mean Squared Error

(ix) Choose the primary limitation of the k-nearest neighbor algorithm.

a) It requires a large amount of memory
b) It is sensitive to the choice of distance metric
c) It cannot handle categorical variables
d) It is computationally expensive during training

(x) Select which of the following steps is NOT part of the data preprocessing stage in KDP?

a) Data Integration
b) Data Reduction
c) Data Mining
d) Discretization

(xi) Select, what is the main goal of data discretization?

a) To convert continuous data into categorical data
b) To remove outliers from the dataset
c) To merge similar data points
d) To standardize the data distribution

(xii) Select, In the KDP process, which stage involves extracting patterns that have meaningful interpretation?

a) Data Integration
b) Data Mining
c) Data Reduction
d) Concept Hierarchy Generation

(xiii) Select which of the following is NOT a step in the Knowledge Discovery Process (KDP)?

a) Data Retrieval
b) Pattern Evaluation
c) Pattern Discovery
d) Data Presentation

(xiv) Choose the step in the KDP process, which one is involves removing irrelevant or redundant attributes from the dataset

a) .Data Transformation
b) Data Reduction
c) Data Mining
d) Data Integration

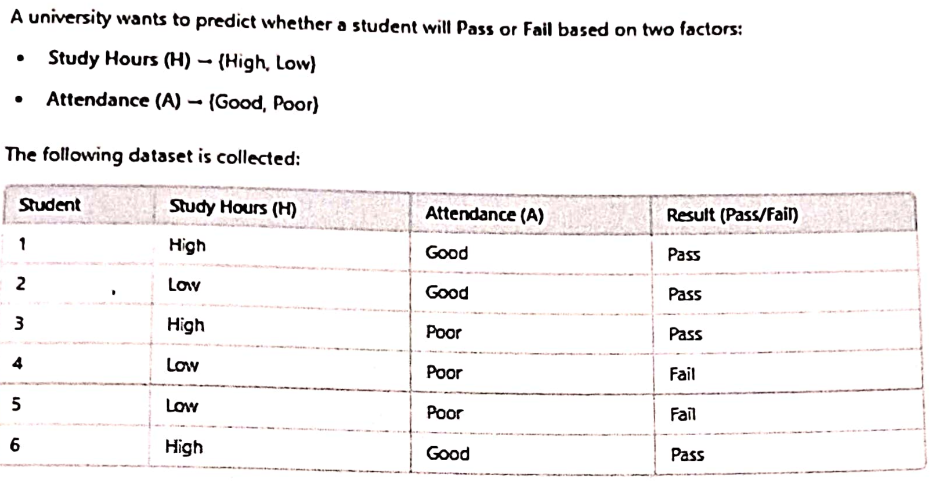
(xv) Select from the following,How would you select the support value for a specific itemset from a large transaction dataset?

a) How would you select the support value for a specific itemset from a large transaction dataset?
b) Calculate the ratio of transactions containing the itemset to the total transactions
c) Use a pre-defined threshold for support d) Apply statistical tests to estimate support
d) none

Group-B

(Short Answer Type Questions) 3 x 5=15

2. Explain the Knowledge Discovery Process (KDP) in data mining. (3)
3. Identify the key steps involved in concept hierarchy generation. (3)
4. Define which one is faster, Multidimensional OLAP or Relational OLAP. (3)
5. Illustrate the concept of a Decision Tree using an experiment on student pass/fail classification. Construct a decision tree based on given conditions and calculate the entropy and information gain for splitting the dataset. (3)



A university wants to predict whether a student will Pass or Fail based on two factors:

- Study Hours (H) → {High, Low}
- Attendance (A) → {Good, Poor}

The following dataset is collected:

Student	Study Hours (H)	Attendance (A)	Result (Pass/Fail)
1	High	Good	Pass
2	Low	Good	Pass
3	High	Poor	Pass
4	Low	Poor	Fail
5	Low	Poor	Fail
6	High	Good	Pass

6. Compare agglomerative and divisive methods in hierarchical clustering. (3)

OR

Differentiate between PCA and t-SNE for dimensionality reduction. (3)

Group-C

(Long Answer Type Questions) 5 x 6=30

7. Estimate the Principal Component Analysis (PCA) algorithm used for in-dimension reduction. (5)
8. Analyze some common methods for identifying outliers in a dataset. (5)
9. Explain the construction process of a decision tree in classification, including the criteria used for node splitting. (5)
10. Explain the Bayesian Belief Networks (BBNs) and discuss their role in classification tasks. (5)
11. Describe how can web mining contribute to knowledge discovery and decision-making in various domains. (5)
12. Justify the role of kernel functions in Support Vector Machines (SVMs) and provide examples of commonly used kernels. (5)

OR

Express the curse of dimensionality in the context of classification and its implications. (5)

****