# BRAINWARE UNIVERSITY

## Term End Examination 2024-2025
## Programme – Dip.CSE-2022
## Course Name – Data Warehousing and Data Mining
## Course Code - DCSE-PE602B
## ( 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) Choose the correct option-- the term dimensional modeling refers to data warehousing.
   a) Modeling 3D graphics for visualization
   b) Designing databases for online transactions
   c) Organizing data for analytical purposes
   d) Processing real-time data streams

(ii) Select from the following is NOT a component of a data warehouse architecture?
   a) Data visualization tools
   b) ETL processes
   c) Data warehouse server
   d) Metadata repository

(iii) Identify the purpose of a data mart in data warehousing?
   a) Storing raw data
   b) Providing a subset of data for a specific business unit
   c) Processing online transactions
   d) Generating real-time reports

(iv) Select from below the terms refers to the process of transforming data into a common format for storage in a data warehouse?
   a) Data extraction
   b) Data transformation
   c) Data loading
   d) Data modeling

(v) Identify the characteristics of data warehouses compared to operational databases?
   a) Optimized for transaction processing
   b) Subject-oriented
   c) Supports day-to-day operations
   d) Contains current data only

(vi) Choose the main disadvantage of the Naive-Bayes Classifier.
   a) it assumes independence between features
   b) It is sensitive to outliers
   c) It requires a large amount of training data
   d) It cannot handle categorical variables

(vii) Choose the main advantage of decision trees.
   a) They are robust to noisy data
   b) They can handle large datasets efficiently
   c) They require little data preprocessing
   d) They are not affected by the curse of dimensionality

(viii) Choose the method is used to handle missing data in classification.

a) Removing instances with missing data
b) Replacing missing values with the mean or median
c) Ignoring missing values during training
d) Using regression to predict missing values

(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) Choose the kernel function ,commonly used in Support Vector Machines for non-linear classification tasks.
a) Linear kernel
b) Polynomial kernel
c) Sigmoid kernel
d) Radial Basis Function (RBF) kernel

(xi) Select the data mining technique suitable for finding hidden patterns in large datasets.
a) Regression analysis
b) Association rule mining
c) Decision trees
d) K-nearest neighbors

(xii) Select, In the KDP process, stage involves extracting patterns that have meaningful interpretation.
a) Data Integration
b) Data Mining
c) Data Reduction
d) Concept Hierarchy Generation

(xiii) Select from 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) Select Data integration involves with which one
a) Transforming data into a suitable format for analysis
b) Combining data from different sources to provide a unified view
c) Reducing the amount of data without losing significant information
d) Creating a hierarchical structure for categorical data

(xv) Select from the following the primary purpose of data preprocessing in KDP.
a) To improve the quality of data
b) To deploy machine learning models
c) To visualize data patterns
d) To generate new data concepts

## Group-B

(Short Answer Type Questions) 3 x 5=15

2. Describe the role of distance metrics in clustering algorithms. (3)
3. Describe the role of regression analysis in data mining. (3)
4. Define Data Warehouses and their types in detail. (3)
5. Name some popular data warehousing tools and technologies. (3)
6. Describe the concept of outlier detection. (3)

**OR**

Analyze the role of PCA in clustering. (3)

## Group-C

(Long Answer Type Questions) 5 x 6=30

7. Describe the ethical considerations in web mining and how they can be addressed. (5)
8. Write down the key steps involved in the text mining process. (5)
9. Explain the PCA algorithm and its objective. (5)
10. Differentiate fact and Dimensional table in Dataware house with example (5)
11. Justify the role of kernel functions in Support Vector Machines (SVMs) and provide examples of commonly used kernels. (5)
12. Analyze some common methods for identifying outliers in a dataset. (5)

**OR**

Explain some common applications of dimension reduction techniques in data mining. (5)

**********