P P SAVANI UNIVERSITY

2nd Semester of M.S.C. CS Examination

August 2022

SSDS7051 Data Mining with Big Data

12.08.2022, Friday **Time: 09:00 a.m. To 11:30 a.m.** **Maximum Marks: 60**

Instructions:

1. The question paper comprises of two sections.
2. Section I and II must be attempted in separate answer sheets.
3. Make suitable assumptions and draw neat figures wherever required.
4. Use of scientific calculator is allowed.

SECTION – I

Q - 1 Answer the Following: (Any five) **[05]**

(i) Enlist the functionalities of Data mining systems.

(ii) What is nominal data?

(iii) List out major issues in Data Mining.

(iv) Explain data integration.

(v) Define Data Preprocessing.

(vi) What is association rule mining?

(vii) What is noisy data?

Q - 2 (a) Consider the following dataset and find frequent item sets and generate association rules for them using Apriori Algorithm. Minimum support count is 2 minimum confidence is 60%. **[10]**

TID	Items
T1	1,3,4
T2	2,3,5
T3	1,2,3,5
T4	2,5
T5	1,3,5

OR

Q - 2 (a) Explain data mining architecture in detail with suitable diagram. **[05]**

Q - 2 (b) Explain: Entropy, Information Gain, Gini Index **[05]**

Q - 3 (a) Explain the Back Propagation with Proper Diagram. **[05]**

Q - 3 (b) Explain how to handle missing data. **[05]**

OR

Q - 3 (a) Explain Linear Regression along with error accuracy measure. **[05]**

Q - 3 (b) Explain Data Discretization and Concept Hierarchy [05]

Q - 4 Attempt anyone.

(i) Explain ID3 Algorithm with example.

(ii) Correlation analysis.

SECTION – II

Q - 1 Answer the Following: (Any five)

(i) Define: Lazy Learners [05]

(ii) Define: Hadoop

(iii) Define: Outlier analysis

(iv) Define: challenges of Big Data

(v) Which of the following statements is incorrect about the hierarchal clustering?

A. The hierarchal type of clustering is also known as the HCA
B. The choice of an appropriate metric can influence the shape of the cluster
C. In general, the splits and merges both are determined in a greedy manner
D. All of the above

(vi) Cluster is

A. Group of similar objects that differ significantly from other objects.
B. Operations on a database to transform or simplify data in order to prepare it for a machine -learning algorithm
C. Symbolic representation of facts or ideas from which information can potentially be extracted.
D. None of these

(vii) Assume that you want to perform supervised learning and to predict number of newborns according to size of stork's population, it is an example of

A. Classification
B. Regression
C. Clustering
D. Structural equation modeling

Q - 2 (a) Differentiate Clustering and Classification. What is the need of clustering? List out the applications of clustering. [05]

Q - 2 (b) Explain Hadoop Cluster Components. [05]

OR

Q - 2 (a) Give the Difference between K-Mean Clustering and Density Based Clustering. [05]

Q - 2 (b) Explain Hadoop Architecture. [05]

Q - 3 (a) Why Is MapReduce Necessary? How Does MapReduce Work? [05]

Q - 3 (b) Explain Hadoop Ecosystem. [05]

OR

Q - 3 (a) Solve below example through Naïve Bayes Classifier. Attributes are Color, Type, Origin, and the subject, stolen can be either yes or no. **We want to classify a Red Domestic SUV. (Calculate the Answer with Yes or No)** [05]

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Q - 3 (b) Can Support Vector Machine feasible to apply for Non-Linear Data? Justify Your Answer. **[05]**

Q - 4 Attempt anyone. **[05]**

(i) Differentiate between Big Data and traditional Relational data.

(ii) What is the difference in Prediction and Classification?

****