P P SAVANI UNIVERSITY

Second Semester of M.Sc. CS Examination

June 2020

SSCS7060 Data Mining with Big Data

18.06.2020, Thursday **Time: 10:00 a.m. To 12:30 p.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 Short Question (Any Five) **[05]**

(i) What is Data Mining?

(ii) What is predictive modeling?

(iii) What is data preprocessing?

(iv) What do you mean by training and testing dataset?

(v) What is data discretization?

(vi) What is data integration?

(vii) Enlist types of data in data mining?

Q - 2 (a) Describe the functionalities of Data mining systems. Give real life examples of each DM Functionalities. **[05]**

Q - 2 (b) Explain DM architecture. **[05]**

OR

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%. **[05]**

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

Q - 2 (b) Explain various methods of handling missing values and noisy values in given dataset. **[05]**

Q - 3 (a) What is Decision Tree? Explain how classification is done using decision tree induction.. **[05]**

Q - 3 (b) Using Naive Bayesian classification method, predict class label of X = (age = youth, income = medium, student = yes, credit_rating = fair) using following training dataset. **[05]**

age	income	Student	credit_rating	Class: buys_computer
youth	high	no	Fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	Yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

OR

Q - 3 (a) Consider your own dataset of at least 10 rows and explain How rule based classifier is used for classification. **[05]**

Q - 3 (b) Explain Linear and non-linear support vector machines with proper diagram. **[05]**

Q - 4 Attempt any one. **[05]**

(i) Support Vector Machines.

(ii) Issues in Data Mining.

SECTION – II

Q - 1 Short Question. (Any Five) **[05]**

(i) What is prediction?

(ii) Explain clustering.

(iii) Define big Data.

(iv) What are outliers in clusters?

(v) Enlist 2 differences between clustering and classification.

(vi) What are the different types of clustering in data mining?(names only)

(vii) What do you mean by distributed files in hadoop?

Q - 2 (a) What is r2 in regression? Explain what is SSE, SSR, and SST in regression and how to find out the values of each term? **[05]**

Q - 2 (b) Explain linear regression. **[05]**

OR

Q - 2 (a) Explain non-linear regression in detail. **[05]**

Q - 2 (b) What are ensemble methods? **[05]**

Q - 3 (a) Define "clustering"? Mention any two applications of clustering. **[05]**

Q - 3 (b) Explain map reduce. **[05]**

OR

Q - 3 (a) Explain the V's of big data. **[05]**

Q - 3 (b) Explain how k-means clustering works? **[05]**

Q - 4 Attempt any one. **[05]**

(i) DBSCAN

(ii) Hadoop

****