Extra (IT)

K. J. Somaiya Institute of Technology, Sion, Mumbai
(An Autonomous Institute Permanently Affiliated to the University of Mumbai)
End Semester Exam
May-June 2024

Program: B.Tech. (Information Technology) **Scheme:** II

Regular Examination: TY - Semester VI
Course Code: ITC601 **Course Name:** Data Mining and Business Intelligence

Date: May 15, 2024 **Duration:** 2.5 Hours **Max. Marks:** 60

Instructions:
(1) All questions are compulsory.
(2) Draw neat diagrams wherever applicable.
(3) Assume suitable data, if necessary.

Ques. No.	Question	Max. Marks	CO	BT Level
Q1.	**Solve any six questions out of eight:**	**12**		
i)	Explain the concept of Data Warehouse.	2	CO1	U
ii)	Explain various types of attributes.	2	CO2	U
iii)	Explain the assumption made by Naïve Bayes classifier.	2	CO3	U
iv)	Sketch and explain Decision Tree as a classifier.	2	CO3	U
v)	Explain various points in Density-based Clustering algorithm.	2	CO4	U
vi)	State the Apriori principle.	2	CO4	U
vii)	Explain Lift measure to evaluate association rule interestingness.	2	CO5	U
viii)	Explain the key components of a Business Intelligence system.	2	CO6	U
Q2.	**Solve any four questions out of six:**	**16**		
i)	Sketch and explain the process of Knowledge Discovery from Data.	4	CO1	U
ii)	Differentiate Symmetric and Skewed data with plots and examples.	4	CO2	A
iii)	Consider a binary classification problem where we are trying to predict whether emails are spam or not spam. Based on the confusion matrix below, calculate Accuracy and Precision: 			Spam	Not Spam	 	Spam	85	10	 	Not Spam	5	95		4	CO3	A
iv)	Explain types of outliers with example.	4	CO4	U
v)	Explain multilevel association rules with example.	4	CO5	U
vi)	Explain phases of decision-making process.	4	CO6	U
Q3.	**Solve any two questions out of three:**	**16**		
i)	Compare OLAP and OLTP.	8	CO1	U
ii)	Consider the runs scored by a cricket team in a league of 12 matches – 100, 120, 110, 150, 110, 140, 130, 170, 120, 220, 140, 110. Sketch a box plot. Further, calculate the mean, median, and standard deviation.	8	CO2	A
iii)	Suppose that the data mining task is to cluster the points A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9) representing coordinates of location (x, y) into 3 clusters. Suppose initially A1, B1, and C1 are represented as the centers of each clusters respectively. Use Euclidean distance and apply the k-means algorithm to show the three cluster centers after first round of execution.	8	CO4	A

Q4.	Solve any two questions out of three:	16		
i)	Apply Linear Regression to calculate *Income* corresponding to *Age* of 45 considering the data below: 	Age	Income (in Lakhs)
---	---
24	4
25	5
28	7
30	8
34	11
38	17		8	CO3	A
ii)	Consider a transactional dataset in the form <TID: Items> as: <T1:I1, I2, I4>, <T2:I2, I3, I4>, <T3:I1, I2>, <T4: I2, I4>, <T5: I1, I2, I3>. If minimum support = 20% and minimum confidence = 70%, apply Apriori algorithm to find all possible association rules and use lift measure to evaluate the rules.	8	CO5	A
iii)	Consider the case of Weather prediction and apply KDD process to derive Business Intelligence. Clearly explain each phase / operation in the KDD process with respect to the stated application.	8	CO6	A

****