17/5/22

K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai
(An Autonomous Institute Affiliated to the University of Mumbai)

End Semester Exam
April – May 2022

B.Tech. (Information Technology)

Examination: TY - Semester VI

Course Code: 1UITC601 **Course Name:** Data Mining and Business Intelligence

Duration: 03 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 Drill Down operation with example.	2	CO1	U
ii)	Compute the Dissimilarity Matrix for 04 instances of an attribute 'Ratings' represented as: (Good, Fair, Average, Good).	2	CO2	U
iii)	List different approaches for handling missing data.	2	CO3	U
iv)	Normalize data 15, 250, 250000 in the range [0, 1] using Decimal Scaling.	2	CO3	U
v)	Explain two-step process of Classification.	2	CO4	U
vi)	Explain Market Basket Analysis.	2	CO4	U
vii)	Explain Support And Confidence with example.	2	CO5	U
viii)	Explain Decision Support Systems.	2	CO6	U
Q2.	**Solve any four questions out of six:**	**16**		
i)	Explain Data Warehousing.	4	CO1	U
ii)	Sketch a boxplot and identify the outliers in the data of *age* of participants in an event: 12, 51, 56, 60, 65, 70, 77.	4	CO2	A
iii)	Sketch a Concept Hierarchy for *Computer Accessories*.	4	CO3	A
iv)	Consider data of 5 objects characterized by a single continuous feature *score*: Player 1: 10, Player 2: 20, Player 3: 40, Player 4: 50, Player 5: 60 Assume that there are two clusters: C1: {a, b}, and C2: {c, d, e}. Calculate the distance matrix. Also calculate single link, complete link, and average distance between C1 and C2 to evaluate clusters' quality.	4	CO4	A
v)	State the Apriori principle. Consider the below 2-itemsets in vertical data format and identify the acceptable 3-itemsets: {A, B}: {T1, T4, T8, T9} {A, C}: {T5, T7, T8, T9} {A, D}: {T4} {A, E}: {T1, T8} {B, C}: {T3, T6, T8, T9} {B, D}: {T2, T4} {B, E}: {T1, T8} {C, E}: {T8}	4	CO5	A
vi)	Explain the cycle of Business Intelligence analysis.	4	CO6	U

Q3.	Solve any two questions out of three:	16		
i)	Sketch and explain the process of Knowledge Discovery from Data.	8	CO1	U
ii)	Compute the highest similarity between two financial articles with term frequency vectors as follows using Cosine Similarity:	8	CO2	A

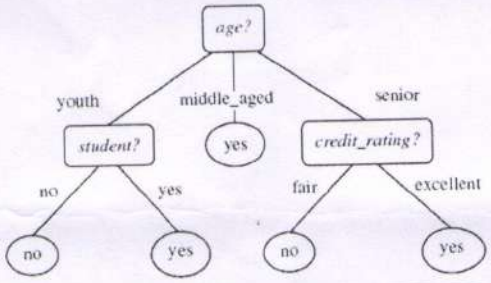
Article No.	Digital Rupee	Budget	Blockchain	Digital Assets	Income	GDP	Disinvestments
1	2	4	3	2	3	2	1
2	5	0	3	4	0	0	2
3	3	0	6	3	0	0	3
4	1	3	4	2	7	4	2

iii)	A survey was conducted to analyse if *discount sale* has a correlation with the ratings given by customers. Apply Chi-square test to analyze whether the attributes *Discount_Sale* and *Customer_ Ratings* are correlated or not:	8	CO3	A

Discount_Sale		*Customer_Ratings*		Total
		Good	Average	
	Yes	250	300	550
	No	50	1100	1150
Total		300	1400	1700

Consider that for 1 degrees of freedom, the chi-square value to reject the hypothesis at 0.001 significance level is 10.828.

Q4.	Solve any two questions out of three:	16		
i)	Suppose that the data mining task is to cluster 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: a. The three cluster centers after the first round of execution. b. The final three clusters.	8	CO4	A
ii)	Consider an *Electronics* store that surveyed customers likely to buy computer from them, based on the data of past customers. Apply the given Decision Tree classifier on the below mentioned test dataset and calculate Accuracy, Precision, and Recall, Sensitivity, Specificity.	8	CO5	A



Age	Income	Student	Credit Rating	Buys Computer
Youth	High	No	Fair	No
Senior	High	No	Excellent	No
Middle	High	No	Fair	No
Senior	Medium	No	Fair	Yes
Senior	Low	Yes	Excellent	Yes
Middle	Low	Yes	Excellent	Yes

iii)	Consider the case of classifying credit card transactions as legitimate or fraudulent. 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

****