16.05.2025 (E)

SOMAIYA VIDYAVIHAR UNIVERSITY

Maximum Marks: 100	Semester: January 2025 – April 2025 Examination: ESE Examination	Duration: 3 Hrs.
Programme code: 54 Programme: Honour - Data Science and Analytics	Class: TY	Semester: (SVU 2020) VI
Name of the Constituent College/School: K. J. Somaiya School of Engineering	Name of the department: Computer	
Course Code: 116h54C601	Name of the Course: Advanced Data Mining	
Instructions: 1) Draw neat diagrams 2) All questions are compulsory 3) Assume suitable data wherever necessary		

Que. No.	Question	Max. Marks
Q1	Solve any **Four**	20
i)	What steps are involved in the Knowledge Discovering in Databases (KDD) process?	5
ii)	What is Frequent Pattern Analysis in data mining?	5
iii)	Explain the difference between multilevel pattern mining and multidimensional pattern mining.	5
iv)	What is Hadoop? Give basic Hadoop components.	5
v)	How can we improve the efficiency of the Apriori algorithm?	5
vi)	You are analyzing sales data for an online electronics store and want to analyze the association between two products: Laptops (L) and Headphones (H). The following contingency table shows the data: Calculate the Lift measure for the association between Laptops and Headphones.	5

	Laptops Present	Laptops Absent
Headphones Present	120	40
Headphones Absent	30	210

Que. No.	Question	Max. Marks
Q2 A	Solve the following	10
i)	A Bloom filter with m = 500 cells is used to store n=200 items, with k=3 hash functions. Calculate the false positive probability of this Bloom filter instance. If the number of hash functions is increased to k = 5, explain how it will affect the false positive probability.	5
ii)	Explain how data stream mining is useful in the following application? • Sensor networks • Network traffic analysis	5
	OR	
Q2 A	How Flajolet Martin (FM) Algorithm approximates the number of unique elements in a data stream? Explain with a suitable example.	10
Q 2 B	Solve any **One**	10
i)	Explain Compact Pattern Stream Tree Algorithm with suitable example.	10

ii)	Given the following transaction dataset, apply the Apriori algorithm to find the frequent itemsets and strong association rules with a support threshold of 30% and 70% confidence. **Transaction No.** — **Itemsets** 1 — {b, a, c, e} 2 — {a, d, c} 3 — {b, a, c, e} 4 — {b, d, c, e} 5 — {a, d, c, e} 6 — {a, c, d, e} 7 — {d, c} 8 — {b, a, d} 9 — {b, c, e} 10 — {a, d}	10

Que. No.	Question	Max. Marks
Q3	Solve any **Two**	**20**
i)	Explain the concept of aligning two time series using the dynamic time warping (DTW) method with an example.	10
ii)	Consider the following Web graph with five pages: A, B, C, D, and E, and the directed links as follows: $A \rightarrow B$ $B \rightarrow A, C$ $C \rightarrow B, D$ $D \rightarrow A$ Assume that the PageRank value for any page m at iteration 0 is $PR(m) = 1$ and the teleportation factor for iterations is $\beta = 0.85$. Perform the PageRank algorithm and determine the rank for every page at iteration 2.	10
iii)	Identify five social networks in popular use. What type of social network do they represent? Attempt to capture their essence using a social graph.	10

Que. No.	Question	Max. Marks
Q4	Solve any **Two**	**20**
i)	Given the following query and documents, calculate the cosine similarity using the TF-IDF vectors. Query: "machine learning data" Document 1: "machine learning" Document 2: "data science"	10
ii)	Explain generalized sequential pattern steps in detail.	10
iii)	You have a set of reviews (Documents) and their classification:	10

	Doc_ID	Text	Class
Training set	1	I enjoyed the film	+
	2	I disliked the film	-
	3	Wonderful film, great actors	+
	4	Bad direction	-
	5	Amazing story, good direction.	+
Test set	6	I disliked the bad direction	?

Estimate the parameters of the Naive Bayes classifier & classify test document.

Que. No.	Question	Max. Marks
Q5	(Write notes / Short question type) on any **four**	20
i)	Distributed data mining	5
ii)	Recommendation systems	5
iii)	K-means clustering	5
iv)	"Edge Betweenness" Measure for Graph Clustering	5
v)	Fast Update (FUP) algorithm in association rule mining	5
vi)	Application of sequential pattern mining	5