


VIT

Vellore Institute of Technology

Final Assessment Test - April 2025

Course: BCSE208L - Data Mining

Class NBR(s): 1781/1784/1791

Time: Three Hours

Slot: F1

Max. Marks: 100

- KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
- DON'T WRITE ANYTHING ON THE QUESTION PAPER

 Answer ALL Questions

(10 X 10 = 100 Marks)

- As a software engineer at All electronics store, you are tasked with designing a data mining system to analyse their sales database. Describe the data mining functionalities and provide an example illustrating their use in this context.
- A researcher wants to determine whether gender (male or female) is independent of preference for a particular product (A, B, or C). The data is summarized as follows:

	Product A	Product B	Product C	Total
Male	30	50	20	100
Female	20	60	20	100
Total	50	110	40	200

Calculate the degrees of freedom and use a significance level (α) of 0.05. The critical value from the chi-square distribution table is approximately 5.991. Apply the chi-square test to analyze the correlation between gender and product preference, and write your inference.

- A database has 5 transactions, Let minimum support = 60% and minimum confidence = 80%. Apply Apriori algorithm and find the frequent itemset, along with the valid association rules.

TID	Items bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

- For the given transaction dataset with min_sup=3, find all frequent itemsets using FP-Growth and Construct FP-tree.

Transactions	Items
Transaction 1	A, B, C, D, E, F
Transaction 2	G, B, C, D, E, F
Transaction 3	A, H, D, E
Transaction 4	A, I, J, D, F
Transaction 5	J, B, B, D, K, E

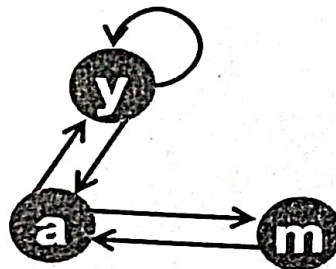
5. Consider the given training data, build the conditional probability table for each attribute and apply Naïve Bayes algorithm to test the data, {Age = youth, cough = medium, Fever = yes, cold = no} and predict the decision class.

Age	Cough	Fever	Cold	COVID-19
youth	high	no	no	no
youth	high	no	yes	no
adult	high	no	no	yes
senior	medium	no	no	yes
senior	low	yes	no	yes
senior	low	yes	yes	no
adult	low	yes	yes	yes
youth	medium	no	no	no
youth	low	yes	no	yes
senior	medium	yes	no	yes
youth	medium	yes	yes	yes
adult	medium	no	yes	yes
adult	high	yes	no	yes
senior	medium	no	yes	no

6. Given the following confusion matrix, evaluate the classifier's performance by calculating the key evaluation metrics: Accuracy, Error Rate, Sensitivity, Specificity, and F1-Score. Clearly present the step-by-step calculations for each metric. Discuss the ways to improve classification accuracy?

Actual Class\Predicted class	cancer = yes	cancer = no	Total
cancer = yes	90	210	300
cancer = no	140	9560	9700
Total	230	9770	10000

7. a) Discuss the key differences between spatial data mining and temporal data mining, highlighting their unique characteristics, data structures and techniques used. [5]
 b) Explain any one statistical data mining method in detail, provide a real-world example of its application. [5]
8. Compute the Page Rank of each page as shown in the web graph given below. Nodes represent individual web pages, while links represent the hyperlinks or relationships between those pages. Use the power iteration to converge within the error limits of double-precision arithmetic (two iterations).



- 9.a) ✓ Sketch the three-tier architecture of a data warehouse to illustrate efficient data integration, storage, and accessibility for business intelligence applications.

OR

- 9.b) A data warehouse comprises of four dimensions: date, spectator, location, and game, and two measures: count and charge, where charge is the fee that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.
- Draw a star schema diagram for the data warehouse. [6]
 - With the base cuboid [date; spectator; location; game], what sequence of OLAP operations should be performed to list the total charge paid by student spectators located at CHEPAK Stadium in 2025? [4]
- 10.a) ✓ Assume that you are given with the dataset consisting of patient ID with two attributes X and Y. For the given patient data apply k-means algorithm to cluster the patients using 1st, 4th and 7th patient as cluster centroid ($k=3$) and perform clustering for two iterations.

Patient ID	x	y
1	2	10
2	2	5
3	8	4
4	5	8
5	7	5
6	6	4
7	1	2
8	4	9

OR

- 10.b)
 - Describe the steps involved in the DBSCAN algorithm, and explain the significance of its two primary parameters and their impact on clustering. [5]
 - Discuss the potential impact of outliers on the dataset and how they might affect model performance. [5]

⇔⇔⇔ F/E/TY ⇔⇔⇔