LOYOLA COLLEGE (AUTONOMOUS) CHENNAI – 600 034

M.Sc. DEGREE EXAMINATION – **MATHEMATICS**

THIRD SEMESTER – **NOVEMBER 2024**

PMT3ME01 – AN INTRODUCTION TO DATA SCIENCE

Date: 19-11-2024 Dept. No. Max. : 100 Marks

Time: 01:00 pm-04:00 pm

	SECTION A – K1 (CO1)
	Answer ALL the questions **(5 x 1 = 5)**
1	**Answer the following.**
a)	Write an example of how Big Data analytics can be used in the healthcare industry.
b)	Define centroid in clustering.
c)	Point out the purpose of logistic regression.
d)	Write the primary advantage of using the Naïve Bayes algorithm in classification.
e)	Recall the main goal of text normalization.
	SECTION A – K2 (CO1)
	Answer ALL the questions **(5 x 1 = 5)**
2	**MCQ**
a)	What type of data is best described as structured and stored in rows and columns in a database? i. Unstructured data ii. Semi-structured data iii. Structured data iv. Bigdata
b)	What is the formula for Confidence in association rules? i. $Confidence(X \rightarrow Y) = \frac{Support(X \wedge Y)}{Support(X)}$ ii. $Confidence(X \rightarrow Y) = \frac{Support(X \vee Y)}{Support(X)}$ iii. $Confidence(X \rightarrow Y) = \frac{Support(X \wedge Y)}{Support(Y)}$ iv. $Confidence(X \rightarrow Y) = \frac{Support(X \vee Y)}{Support(Y)}$
c)	In linear regression, the relationship between the dependent and independent variables is assumed to be: i. Non-linear ii. Linear iii. Exponential iv. Logarithmic
d)	Which of the following is a key assumption of the Naïve Bayes classifier? i. All features are dependent on each other Linear ii. All features are independent of each other iii. The model is non-probabilistic iv. The data is always linearly separable
e)	Which of the following is a step in text normalization? i. Tokenization

	ii. TF-IDF calculation iii. Lowercasing text iv. Sentiment analysis

SECTION B – K3 (CO2)

	Answer any THREE of the following	**(3 x 10 = 30)**
3	Illustrate the various drivers of big data.	
4	By using K-means clustering, separate the following dataset into three optimal clusters and find the corresponding centroids: (1, 3), (1, 5), (1, 9), (2, 3), (2, 7), (3, 10), (4, 5), (4, 9), (5, 10) and (6, 10).	
5	Derive the coefficients of a linear regression model by applying the Ordinary Least Squares (OLS) technique.	
6	Evaluate the performance of classifier with precision and recall measures.	
7	Illustrate data sources and formats for text analysis.	

SECTION C – K4 (CO3)

	Answer any TWO of the following	**(2 x 12.5 = 25)**
8	Analyse business intelligence and data science.	
9	Criticize Naïve Bayes theorem.	
10	A company is evaluating a new chemical compound to determine its effectiveness in removing stains. They tested the compound at various concentrations and recorded whether the stains were completely removed or not. The concentrations varied from 1% to 10%. The results are summarized in the following table. Draw ROC curve for the given data.	

Concentration (%)	Stains removed	Stains not removed
1	50	30
2	65	25
3	80	20
4	90	15
5	95	10
6	100	5
7	98	7
8	85	15
9	70	25
10	60	30

11	Analyse corpora in natural language processing.	

SECTION D – K5 (CO4)

	Answer any ONE of the following	**(1 x 15 = 15)**
12	Explain Association rules with an example. Also write the Apriori algorithm to find the frequent itemsets.	
13	Interpret decision tree with suitable example.	

SECTION E – K6 (CO5)

	Answer any ONE of the following	**(1 x 20 = 20)**
14	Compile emerging Big Data ecosystem and propose how new analytical approaches are reshaping data-driven decision-making.	
1	Compile the concepts tokenization and case folding in text analysis with suitable examples.	

5	

$$$$$$$$$$$$$$$