EPCET LIBRARY BANGALORE-49

CBCS SCHEME

USN | | | | | | | | | | |

BAD702

Seventh Semester B.E./B.Tech. Degree Examination, June/July 2026
Statistical Machine Learning for Data Science

Time: 3 hrs. Max. Marks: 100

Note: 1. Answer any FIVE full questions, choosing ONE full question from each module.
2. M : Marks , L: Bloom's level , C: Course outcomes.

		Module – 1	M	L	C
Q.1	a.	Explain Exploratory Data Analysis (EDA) and discuss how estimates of location and variability are used to interpret the characteristics of a dataset.	10	L2	CO1
	b.	Explain how boxplots , frequency tables , histograms and density plots help in making sense of data patterns and sup-porting decision making.	10	L2	CO1
		OR			
Q.2	a.	Describe the various techniques for exploring binary and categorical data and illustrate them with an example.	10	L2	CO1
	b.	A University collected data on 2,000 students including Department , Gender , Attendance % , Final marks and Teaching Method. Briefly explain i) How a contingency table can be used to study categorical relationships. ii) How a hexagonal binning and contour plots can reveal pattern between attendance and marks. iii) How a violin plot can compare score distributions across teaching methods?	10	L2	CO1
		Module – 2			
Q.3	a.	Explain the concepts of population and sample and describe how different sampling methods can be applied in real – world with examples.	10	L2	CO2
	b.	Describe the major types of statistical distributions and their significance in data analysis.	10	L2	CO2
		OR			
Q.4	a.	Define Bootstrap and describe the steps involved in the bootstrap algorithm.	10	L2	CO2
	b.	Discuss how Chi – Square distribution and F – distribution support hypothesis testing and variance analysis in research and experimental studies.	10	L2	CO2
		Module – 3			
Q.5	a.	Discuss how A/B testing and hypothesis testing are used to make data – driven business or research decisions.	8	L2	CO3
	b.	Explain the role of resampling techniques and P – value in improving the reliability of statistical conclusions.	6	L2	CO3

VTU-20-05-2026 08:49:49am EP - EP - EP -

	c.	Discuss how t – test and degrees of freedom help in testing population parameter when sample sized are small.	6	L2	CO3
		OR			
Q.6	a.	Discuss the key steps involved in the classical statistical inference pipeline and explain their role in drawing valid conclusion from data.	8	L2	CO3
	b.	Discuss how Type – I error and Type – II error arise during the process of statistical significance testing and explain their practical implications.	6	L2	CO3
	c.	Compare the functions of the Null hypothesis and the alternative hypothesis in hypothesis testing and illustrate their roles with a suitable example.	6	L2	CO3
		Module – 4			
Q.7	a.	Explain the working of the multi – armed bandit algorithm and illustrate its application with a suitable example.	10	L2	CO4
	b.	Discuss how factor variables are used in regression analysis and their role in interpreting categorical data.	10	L2	CO4
		OR			
Q.8	a.	Discuss how polynomial regression can be applied to model nonlinear relationships between variables in real – world scenarios.	10	L2	CO4
	b.	Discuss the working and advantages of spline regression in fitting flexible curves to data.	10	L2	CO4
		Module – 5			
Q.9	a.	Explain and discuss how Fisher's linear discriminant is applied for dimensionality reduction and classification in real – world problems.	10	L2	CO5
	b.	Compare the role of the covariance matrix in Discriminant analysis with its role in other multivariate techniques such as Principal component analysis. Then, for the dataset X = [2, 4, 6, 8] , Y = [1, 3, 5, 7]. Construct the 2 × 2 convariance matrix and interpret the relationship between X and Y.	10	L3	CO5
		OR			
Q.10	a.	Discuss the key terms used for evaluating classification models such as accuracy , precision , recall , specificity and F1 – score. Then for a classification problem with the following results : True Positives (TP) = 40, False Positives (FP) = 10 , True Negatives (TN) = 30 , False Negatives (FN) = 20. Construct the confusion matrix and calculate accuracy , precision and recall.	10	L3	CO5
	b.	Discuss different strategies for handling imbalanced data , such as oversampling , undersampling and synthetic data generation and explain their impact on model performance.	10	L2	CO5

* * * * *