Satyug Darshan Institute of Technology,Faridabad		
	Programme: B.Tech	**Year:4th Semester: 7th**
	Subject Name : Data Analytics	**Subject Code:PEC-CS-D-702 (1)**
	Department:CSE	**AY:2025-2026**
Prepared By: Ms.Shivani Bansal		
Unit -1		
Short Questions		
S. No.	**Questions**	**Bloom Taxonomy**
1	Define Descriptive Statistics.	BTL–1
2	Define Inferential Statistics?	BTL–1
3	State the null and alternative hypothesis.	BTL–2
4	Explain what is a probability distribution.	BTL–1
5	Discuss the purpose of permutation testing?	BTL–2
6	Explain Descriptive Statistics and Inferential Statistics with suitable examples.	BTL–2
7	Describe different types of probability distributions used in statistical analysis.	BTL–2
8	Explain the steps involved in hypothesis testing with an appropriate example.	BTL–3

9	Analyze the concept of permutation and randomization tests and compare them with parametric tests.	BTL–4
10	Discuss the difference between Type I error and Type II error?	BTL–2
11	Calculate the median for the data set: 12, 7, 3, 10, 8.	BTL–3
12	Compute the mode of the following data: 4, 6, 8, 6, 10, 6.	BTL–3
13	Find the range of the data: 18, 25, 10, 32, 20.	BTL–3
14	Calculate the variance for the data: 2, 4, 6, 8.	BTL–3
15	Find the standard deviation of the data: 5, 5, 5, 5.	BTL–3
Long Questions		
1	Explain Descriptive Statistics and Inferential Statistics with suitable examples.	BTL–2
2	Describe different types of probability distributions used in statistical analysis.	BTL–2
3	Explain the steps involved in hypothesis testing with an appropriate example.	BTL–3
4	Analyze the concept of permutation and randomization tests and compare them with parametric tests.	BTL–4
5	Evaluate the importance of Inferential Statistics in real-world decision-making. The following data shows the marks obtained by students: 12, 15, 18, 20, 22, 25, 30. (a) Calculate the mean, median, and mode. (b) Comment on the nature of the distribution.	BTL–5
6	A random sample of size 100 has a mean of 40 and a standard deviation of 5. Test at 5% level of significance whether the population mean is 42. | BTL–5 |	BTL–5
7	A manufacturer claims that the average lifetime of a bulb is 1200 hours. A sample of 50 bulbs gives a mean life of 1150 hours with a standard deviation of 100 hours. Test the manufacturer's claim at 5% level of significance.	BTL–5
8	Using permutation testing, test whether the difference between the means of the following two samples is significant: Sample A: 10, 12, 14, 16 Sample B: 8, 9, 11, 13 | BTL–6 |	BTL–5
9	Two coins are tossed 100 times and the following results are obtained: OutcomeHHHTTHTT Frequency30202525 Using a Chi-square test, test whether the coins are fair.	BTL–5
10	Describe probability distributions (Normal, Binomial, Poisson) and their applications.	BTL-2
11	On average, a call center receives 4 calls per minute. Using the Poisson distribution, find the probability that in a given minute the center receives: 1.Exactly 2 calls 2.At most 3 calls 3.More than 4 calls	BTL-3
Unit -2		
Short Questions		

S. No.	Questions	Bloom Taxonomy
1	Mention one difference between Ridge and Lasso regression.	BTL–2
2	Describe overfitting in regression models?	BTL–1
3	Define the role of regularization in machine learning.	BTL–1
4	Define K-Nearest Neighbour (KNN) algorithm.	BTL–1
5	Identify the value of K in KNN and why is it important?	BTL–2
6	Discuss the difference between regression and classification?	BTL–2
7	Define model-based learning framework.	BTL–1
8	Define the algorithmic learning approach.	BTL–1
9	State one application of regression in real-world problems.	BTL–3
10	Explain what do you meant by bias in machine learning models?	BTL–1
11	Calculate the Euclidean distance between points A(1,2) and B(4,6) used in KNN.	BTL–3
12	Define the k-Nearest Neighbors (KNN) algorithm and list the main steps involved in using it for classification. Given the following training data with two features and a new sample point, use K = 3 to classify the new sample. Training data: (1.0, 2.0) → Class A (2.0, 3.0) → Class A (3.0, 3.0) → Class B (6.0, 5.0) → Class B (7.0, 7.0) → Class B New sample: (2.5, 2.5) Show your distance calculations, identify the 3 nearest neighbours and then state the predicted class.	BTL–3
13	Apply your understanding of regularization techniques to compare Ridge Regression and Lasso Regression.	BTL-4
	Long Questions	
1	A teacher wants to check if three different teaching methods have significantly different effects on student performance. The marks of students taught by the three methods are given below: •Method A: 65, 70, 75 •Method B: 60, 65, 70 •Method C: 55, 60, 65 At a 5% significance level, test whether there is any significant difference among the teaching methods. (Fcritical≈5.14)	BTL-2
2	Apply the concept of regression vs. classification in machine learning. Provide examples of each.	BTL-4
3	Analysis the steps in conducting of Hypothesis Test	BTL-2
5	A pharmaceutical company claims that its new drug reduces the mean recovery time of patients from a standard treatment mean of 14 days to less than 12 days. As a biostatistician, you are asked to critically evaluate the validity of this claim using a clinical trial. Design how you would conduct the hypothesis test. Evaluate the claim based on hypothetical data: from a sample of 64 patients using the new drug, mean recovery time=11.2 days, with known population standard deviation=4 days.	BTL-5

6	Determine the standard phases of the data analytics lifecycle.	BTL-4
7	A dataset contains 10 features. In a model-based learning framework, explain how feature selection and regularization improve prediction accuracy, supporting your answer with a numerical illustration.	BTL–5
8	The following data represents study hours (X) and marks (Y): X: 3, 5, 7, 9, 11 Y: 40, 50, 65, 70, 80 (a) Compute the correlation coefficient. (b) Fit a linear regression model and predict marks for 10 hours of study.	BTL–5
9	For a regression model with two predictors, the ANOVA values are given as: SSR = 150, SSE = 30, and total observations n = 12. (a) Calculate SST. (b) Find the Mean Square Regression (MSR) and Mean Square Error (MSE). (c) Compute the F-statistic.	BTL–5
10	Consider a KNN regression problem with K = 3. Training data: (1,2), (2,4), (3,6), (6,12), (7,14) Estimate the output for input x = 4 using Euclidean distance and averaging.	BTL–5
11	Explain the Apriori Algorithm in detail. Describe its steps, assumptions, and applications in association rule mining.	BTL–4

Unit -3

Short Questions

S. No.	Questions	Bloom Taxonomy
1	Define supervised learning.	BTL–1
2	Explain the bias–variance dichotomy?	BTL–2
3	Define model validation.	BTL–1
4	Describe the Logistic Regression?	BTL–1
5	State one application of Logistic Regression.	BTL–2
6	Explain the Linear Discriminant Analysis (LDA)?	BTL–1
7	How does Quadratic Discriminant Analysis (QDA) differ from LDA?	BTL–2
8	State the decision tree classifier?	BTL–1
9	Define entropy in the context of decision trees.	BTL–1
10	Explain the Random Forest?	BTL–1
11	State one advantage of ensemble methods.	BTL–2
12	Describe the Support Vector Machine (SVM)?	BTL–1

13	Define hyperplane in SVM.	BTL–1
14	State what do you meant by kernel trick in SVM?	BTL–2
15	Define the term neural network.	BTL–1
16	Define deep learning.	BTL–1
	Long Questions	
1	A Random Forest consists of 15 trees. If 9 trees predict Class A and 6 predict Class B, determine the final predicted class.	BTL–3
2	Apply the concept of active learning to design a model for text classification where labeled data is limited. Explain how the system selects the most informative samples for labeling	BTL–4
3	Construct and explain the process of calculating support, confidence, and lift for association rules with a suitable example dataset	BTL–3
4	Evaluate the effectiveness of Support Vector Machines (SVMs) in solving classification problems. Discuss the role of Kernel functions in SVM and justify why different kernel types are used for various data patterns.	BTL-5
5	Analyze the performance of the following two models trained on the same dataset. Model A: Low training error and low test error. Model B: Very low training error but high test error. Identify which model likely has high bias or high variance and explain how you would adjust model complexity and other hyperparameters to improve generalisation.	BTL-4
6	A Random Forest classifier uses 30 trees. If the majority threshold is more than 50%, calculate the minimum number of trees needed to predict a class.	BTL–3
7	Explain the bias–variance dichotomy in detail. Discuss its impact on model performance with suitable examples.	BTL–4
8	Analyze the relationship between bias, variance, underfitting, and overfitting in supervised learning models.	BTL–5
9	Using suitable diagrams, explain how the bias–variance tradeoff affects generalization error.	BTL–5
10	Evaluate different techniques such as regularization, ensemble learning, and cross-validation in controlling bias and variance.	BTL–5
	Unit -4	
	Short Questions	
S. No.	**Questions**	**Bloom Taxonomy**
1	Define unsupervised learning.	BTL–1
2	State the term clustering in machine learning?	BTL–1
3	Define K-means clustering.	BTL–1
4	Define hierarchical clustering.	BTL–1
5	Define dimensionality reduction.	BTL–1

6	Illustrate Principal Component Analysis (PCA)?	BTL–1
7	State one application of unsupervised learning.	BTL–2
8	Define Big Data.	BTL–1
9	Mention any two challenges of Big Data Analytics.	BTL–2
10	Explain the term Active Learning?	BTL–1
11	Discuss Reinforcement Learning?	BTL–1
12	Define reward in Reinforcement Learning.	BTL–1
13	Describe what is exploration–exploitation tradeoff?	BTL–2
14	Describe the sequence of steps a single Decision Tree follows from root node to leaf in an ensemble setting like Random Forest.	BTL–1
15	Diagrammatically express the Support vectors in machine learning.	BTL–2
16	Explain the three V's of Big Data**.**	BTL–1
	Long Questions	
1	Explain the concept of Unsupervised Learning. Discuss its importance and applications in Big Data Analytics.	BTL–2
2	Describe K-means clustering algorithm in detail. Explain its advantages, limitations, and applications.	BTL–3
3	Explain Hierarchical Clustering techniques with suitable examples and dendrograms.	BTL–3
4	Compare K-means and Hierarchical clustering methods in terms of working principle, complexity, and use cases.	BTL–4
5	Discuss Dimensionality Reduction techniques with emphasis on Principal Component Analysis (PCA).	BTL–4
6	Analyze the challenges of Big Data Analytics with respect to volume, velocity, and variety.	BTL–5
7	Explain the role of Unsupervised Learning in handling high-dimensional and unstructured data.	BTL–5
8	Describe Active Learning and explain how it helps in creating efficient datasets for analytics.	BTL–4
9	Explain Reinforcement Learning concepts such as agent, environment, reward, and policy with suitable examples.	BTL–4
10	Evaluate the impact of Active Learning and Reinforcement Learning in solving real-world Big Data problems.	BTL–5
11	For the dataset: 10, 12, 15, 18, 20, 22 (a) Compute the mean (b) Perform dimensionality reduction using PCA by finding the first principal component (conceptual numerical explanation).	BTL–5
12	Apply the K-means clustering algorithm for the following data points: (2,2), (4,4), (6,6), (8,8). Assume K = 2 and initial centroids as (2,2) and (8,8). Show all iterations and final clusters.	BTL–4
13	Explain how Apriori Algorithm is used in market basket analysis with a real-world example.	BTL–4