Name:

Enrolment No:

UPES

End Semester Examination, May 2023

Course: Applied Data Science **Semester: 8**
Program: B.TECH Big data, Devops
Time: 03 hrs.
Course Code: CSDV4004 **Max. Marks: 100**
Instructions: Attempt every questions.

SECTION A
(5Qx4M=20Marks)

S. No.		Marks	CO
Q 1	Explain the difference between a population and a sample in statistics? How is a census related to the concept of a population, and what are some common sampling methods used to obtain a representative sample from a population?	**4**	**CO1**
Q 2	Explain the concept of overfitting in regression analysis. How can overfitting be avoided in practice?	**4**	**CO3**
Q 3	Describe the difference between null and alternative hypotheses in hypothesis testing? How are these hypotheses related to each other?	**4**	**CO2**
Q 4	Explain difference between discrete and continuous random variables? Provide an example of each type of random variable.	**4**	**CO4**
Q 5	How can variables be classified in statistics? Explain the difference between independent and dependent variables, and describe how the relationship between variables can be examined using statistical methods.	**4**	**CO1**

SECTION B
(4Qx10M= 40 Marks)

Q 6	Explain the concept of measuring accuracy and error metrics in classification. Why is accuracy alone not always a good measure of performance? Describe at least two additional metrics commonly used to evaluate classification performance, and explain their significance.	**10**	**CO4**

	Finally, provide an example of a scenario where accuracy is not a suitable metric for evaluating the performance of a classifier.		
Q 7	Explain the concept of correlation and regression in statistics. Discuss the differences between Karl Pearson and Spearman’s rank order correlation methods, and their suitability for different types of data. Describe the common types of regression models used in practice, including simple linear regression, multiple regression, and polynomial regression. Discuss the challenges of overfitting and variable scaling, and explain how they can be addressed in regression analysis.	**10**	**CO2**
Q 8	Explain the steps involved in hypothesis testing, including the definition of the null and alternative hypotheses, the selection of an appropriate test statistic, the determination of the level of significance, and the calculation of the p-value. Provide an example of how this process can be used to test a population mean, and describe the differences between testing a large versus a small sample.	**10**	**CO4**
Q 9	Explain the concept of making inferences about populations from samples. Discuss the difference between estimators and estimates, and provide an example of how these concepts can be used to estimate a population mean. Finally, describe the process of constructing a confidence interval for a population mean using a large sample size, and explain the role of the level of confidence and the standard error of the mean in this process. OR Explain how the singular value decomposition (SVD) can be used to solve the least squares problem in optimization, and discuss the advantages and limitations of this approach compared to other optimization methods.	**10**	**C04**
SECTION-C **(2Qx20M=40 Marks)**			
Q 10	Compare and contrast the different classification algorithms. Choose at least three algorithms from Naive Bayes, Decision Trees, Random Forest, Maximum Entropy, and Neural Networks. Explain the strengths and weaknesses of each algorithm, and how they differ in their approach to classification. Use appropriate metrics to evaluate the performance of each algorithm, and explain your choice of metrics. Or	**20**	**CO5**

	Explain the differences between Cluster Analysis, Factor Analysis, and Multidimensional Scaling, including their underlying assumptions, purposes, and methods of data analysis. Provide examples of situations where each of these techniques could be useful in social science research.		
Q 11	Discuss the importance of summarizing data using graphical methods, measures of central tendency, and measures of dispersion, and explain the different levels of measurement used in statistics. Describe the role of random variables and probability distributions in statistical analysis, and differentiate between discrete and continuous random variables. Finally, provide examples to illustrate key concepts throughout your answer.	**20**	**CO2**