National Institute of Technology Kurukshetra
Department Of Computer Engineering
Machine Learning and Data Analytics (CSIC-221)
End Term Examination

Time- 3 Hrs.

Max.Mark

Q-1: Write short notes on the following: **(1.5 * 4 = 6 Marks)**

a) Curse of Dimensionality
b) Overfitting vs Underfitting
c) Supervised vs Unsupervised
d) Classification and Regression

Q-2: Explain Inferential and Descriptive Statistics in details and further classify each type. **(4 Marks)**

Q-3: Solve the question based on below data using univariate normal distribution, suppose the engineer measures the compressive strengths of five different concrete samples. The recorded values (in Mpa) are given below. **(5 Marks)**

Sample	1	2	3	4	5
Values (in MPa)	38	42	39	40	41

Q-4: Calculate the Pearson and Spearman correlation using the following dataset: **(5 Marks)**

Observation	1	2	3	4	5
Xi (Hour Studied)	2	4	6	8	10
Yi (Test Score)	50	55	60	65	70

Q-5: Suppose you are tasked with predicting the final exam score of students based on their hours of study using linear regression. The dataset consists of the following data for 5 students, Using the below data, perform a linear regression to find the equation of the regression line and predict the final exam score for a student who studies for 6 hours. **(5 Marks)**

No. of Students	1	2	3	4	5
Xi (Hours of Studied)	2	4	6	8	10
Yi (Test Score)	50	55	60	65	70

Q-6: Discuss the role of clustering in image compression, particularly in algorithms such as K-Means. How does clustering help reduce the data size without significant loss of information? **(5 Marks)**

Or

A company claims that the average lifespan of their light bulbs is 1,200 hours. A consumer group randomly selects 36 light bulbs and finds that the sample mean lifespan is 1,180 hours, with a standard deviation of 150 hours. Test the company's claim at a 1% significance level ($\alpha = 0.01$). **(5 Marks)** $z = -2.33$

P.T.O

Q-7: Given the following dataset with attributes Weather, Temperature, and the target variable Play. Calculate the Information Gain (IG) for splitting the dataset using the Weather attribute as the root node and use the calculated IG to decide if Weather is a good choice for the root node of a decision tree. **(5 Marks)**

Day	Weather	Temperature	Play?
1	Sunny	Hot	No
2	Cloudy	Hot	Yes
3	Sunny	Mild	Yes
4	Cloudy	Mild	Yes
5	Rainy	Mild	No
6	Rainy	Cool	No
7	Rainy	Mild	Yes
8	Sunny	Hot	No
9	Cloudy	Hot	Yes
10	Rainy	Mild	No

Q-8: A company collects data on two features, Height (cm) and Weight (kg), from five individuals, as shown in the table below. Find the principal components and explain which eigen vector corresponds to the most variance in the data. **(5 Marks)**

Individual	1	2	3	4	5
Height (cm)	160	165	170	175	180
Weight (kg)	55	60	65	70	75

Q-9: Given the following dataset with two features (*X1* and *X2*) and corresponding class labels. *You are tasked with classifying a new data point* (X1=5, X2=6) using the K-Nearest Neighbors (KNN) algorithm with $K = 3$.

(5 Marks)

X1	1	2	3	6	7
X2	2	3	3	7	8
Class	A	A	B	B	B

Q-10: Describe the various stages in the data science lifecycle. In your answer, explain the key activities involved in each stage and how they contribute to the overall goal of solving a data-driven problem. **(5 Marks)**