QP Code: RM23MTECH127 | Reg. No | | | | | | | | | | | AY 23

GIET UNIVERSITY, GUNUPUR – 765022

M. Tech (Second Semester) Examinations, May – 2024

MPECS2031 - Data Preparation and Analysis

(CSE)

Time: 3 Hrs | Maximum: 70 Marks

(The figures in the right hand margin indicate marks.)

PART – A **(2 x 10 = 20 Marks)**

Q.1. Answer all questions	CO#	Blooms Level
a. Explain the importance of data gathering and preparation in the context of machine learning.	CO1	K2
b. Perform a hypothesis test to determine if the mean age of two groups is significantly different, given the following data: Group A (mean=35, standard deviation=5, n=100) and Group B (mean=38, standard deviation=6, n=120).	CO1	K3
c. Name one technology or framework that helps achieve scalability in Big Data processing.	CO1	K2
d. What are MAR and MNAR, and how do they relate to missing data in datasets?	CO2	K2
e. Describe the difference between feature scaling and feature engineering in the context of data transformation.	CO2	K2
f. Write down the steps of gradient descent optimization in terms of multiple linear regression.	CO4	K2
g. Explain Typ1- and Type-2 errors.	CO1	K2
h. What is the perceptron algorithm, and how does it work?	CO2	K2
i. Why Data cleaning is an essential pre-processing step in data analysis?	CO2	K2
j. Define the bias variance tradeoff .		

PART – B **(10 x 5=50 Marks)**

Answer ***ANY FIVE*** questions0

		Marks	CO#	Blooms Level
2. a.	Consider the dataset : Explain the significance of each error metric (MSE, RMSE, RMSLE, MAE) in evaluating the performance of a regression model. Provide examples of scenarios where each metric would be preferred over others.	5	CO1	K3

Actual value	Predicted value
100	130
150	170
200	220
250	260
300	325

b.	Two random samples were drawn from two normal populations and their values are:- A: 16,17,25,26,32,34,38,40,42 B : 14,16,24,28,32,35,37,42,43,45,47 Test whether two populations have the same variance at 5% level of significance ($F_{0.05}$ =3.35) using the Hypothetical F-Test.	5	CO1	K3
3.a.	Compare and contrast the mechanisms of MAR (Missing at Random) and MNAR (Missing Not at Random) in the context of missing data. Discuss the implications of each mechanism for data analysis and potential strategies for handling missing data.	3	CO2	K2

b. What is data transformation? Find the min-max normalization of the following tabular data: - 7 CO2 K3

Slno	Midtem Mark	CGPA
1	84	4
2	63	3.2
3	77	2.6
4	78	2.1
5	90	3.2
6	75	3.7
7	49	2.1

4. a. Explain what is KNN classifier and given a dataset use the KNN algorithm to classify a new flower sample. Choose k=3 and determine the class label for the new sample based on the majority vote of its three nearest neighbors. 5 CO3 K3

Slno	Age	Income	Class
1	30	40	A
2	35	45	A
3	25	30	B
4	40	55	B
5	28	35	A

b. Explain Support Vector machine briefly and consider a dataset containing the following data points how SVM works on the dataset to estimate the followings: - 5 CO3 K3

Data Point	Feature 1	Feature 2	Class
1	2	3	A
2	3	4	A
3	5	6	B
4	7	8	B
5	9	10	A
6	11	12	B

(i) Determine the optimal hyperplane that separates the classes.
(ii) Calculate the margin between the classes.
(iii) Discuss the significance of the margin in SVM and its impact on model generalization.

5.a. Explore the correlations between the number of hours spent studying and exam scores in a class of students. Create a plot to visualize the relationship between these variables and calculate the correlation coefficient. Interpret the correlation and discuss any outliers. 3 CO4 K3

Student	Hours studied	Exam Score
1	2	65
2	4	75
3	6	85
4	8	95

b. Consider a dataset containing the following values: 7 CO4 K3
[10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60].
(i) Create a boxplot to visualize the distribution of these values.
(ii) Interpret the components of the boxplot, including the median, quartiles, and any outliers.

6. a. Suppose we have four different brands of fertilizer (X, Y, Z, and W) and we want to determine if there is a significant difference in the crop yield achieved using these fertilizers. The crop yield data for each brand are as follows: 7 CO5 K3

Fertilizer X: [20, 25, 30, 35, 40]
Fertilizer Y: [18, 22, 26, 30, 34]
Fertilizer Z: [15, 20, 25, 30, 35]
Perform a one-way ANOVA to test the null hypothesis that there is no significant difference in the mean crop yield among the four fertilizer brands. Use a significance level of α=0.05

b. What is clustering, and why is it useful in exploratory analysis? 3 CO5 K3

7. a. Explain how a multi layered perceptron works? In the given dataset, a MLP with one hidden layer for binary classification, input layer, output layer. Apply multilayer perceptron of the below-mentioned dataset and compute the following. 7 CO5 K3

1. Initialization
2. Forward propagation
3. Loss calculation
4. Backpropagation
5. update weights and biases

X1	X2	Y
0.1	0.5	0
0.2	0.4	1

b. What is principal component analysis (PCA), and how does it help in reducing the dimensionality of a dataset? 3 CO2 K2

8. a. A binary classification model that predicts whether a patient has a certain disease based on some medical test results. The model has been applied to a dataset containing 100 samples, out of which 30 samples are positive (patients with the disease) and 70 samples are negative (patients without the disease). After applying the model, we obtain the following results:

Actual Class Predicted Probability
Positive 0.7
Negative 0.3
Positive 0.6
Negative 0.4

Given the predicted probabilities for each sample, calculate the True Positive Rate (TPR) and False Positive Rate (FPR) for different threshold values. Then, plot the ROC curve for the classification model.

b. Write a short note on the following: 5 CO3 K2

(i) Association and its rules
(ii) Naïve Bayes classification

--- End of Paper ---