UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, May 2022

Course: ML **Semester: 2nd**
Program: Master of Technology (CSE) **Time: 03 hrs.**
Course Code: CSA 7007P **Max. Marks: 100**

Instructions:

SECTION A
(5Qx4M=20Marks)

S. No.		Marks	CO
Q 1	Write all steps of learning system in Machine learning	**4**	**CO1**
Q 2	What is the difference between linear model and non-linear model? Can we transform non-linear model to linear model? Justify your answer	**4**	**CO1**
Q 3	Define information gain and entropy. A dice is rolled several times and it is observed that there is 75% chance of getting 5 and chance of getting rest of numbers are equal distributed. Find the entropy	**4**	**CO3**
Q 4	Let us consider the following dataset. Identify all the outliers exist in the dataset. Dataset: X = [21, 10, 51, 52, 5, 11, 91, 50, 54, 55, 57, 100].	**4**	**CO2**
Q 5	Distinguish between the supervised and un-supervised learning system. Is it possible to apply supervised algorithm in un-labelled data? Justify your answer	**4**	**CO3**

SECTION B
(4Qx10M= 40 Marks)

Q 1	Develop a regression model and calculate the performance of the model.		**CO3**

Year	Chemical composition	Profit in thousand
2015	5.5	70
2016	5.0	65
2017	4.5	60
2018	3.5	55
2019	3.0	50
2020	6.0	67
2021	6.5	62

Q 2	Consider a classification model ¿ $\frac{e^y}{1+e^y}$) for the following dataset In addition, calculate the performance of the model. a) Find the accuracy of the model if the threshold value is considered as 0.58, 0.65		**CO3**

b) What fraction of actual positive and negative class is correctly predicted?
c) What fraction of predicted positive and negative class is correctly predicted?
d) Find the F1- score

Chemical composition (X)	Usable (yes) or not usable(no)	Prediction of model
5.5	yes	0.55
5.0	no	0.60
4.5	no	0.65
3.5	yes	0.70
3.0	yes	0.85
6.0	no	0.75
6.5	no	0.70

Q 3 — Given the below Confusion Matrix, predict the accuracy, precision, recall, and F1-score — **C03**

		Predicted	
	Class	1	0
Actual	1	30	10
	0	20	30

Q 4 — Briefly explain the Applications of ML. Can an unknown function be approximated? If so, justify — **CO2**

OR

Briefly explain the procedure of data pre-processing. Can a known function be approximated? If so, justify — **CO2**

SECTION-C
(2Qx20M=40 Marks)

Q 1 — Write the algorithm of k-means clusters through flow chart. Implement k-means algorithm for the given dataset for creating two clusters. Let us consider the centroid (1.0,1.0) for cluster 1 and (5.0,7.0) for cluster 2. — **CO3**

Individual	Var1	Var2

	1	1.0	1.0		
	2	1.5	2.0		
	3	3.0	4.0		
	4	5.0	7.0		
	5	3.5	5.0		
	6	4.5	5.0		
	7	3.5	4.5		

Q 2 | **CO3**

Let us consider the following dataset. Apply Naïve Bayes classifier to predict the output class for some X = [Outlook=Sunny, Temperature=Cool, Humidity=High, Windy=True, Play =?]

Instance number	Outlook	Temperature	Humidity	Windy	Play
1	Overcast	Mild	High	True	Yes
2	Overcast	Mild	Normal	True	Yes
3	Sunny	Mild	Normal	False	Yes
4	Rainy	Mild	High	True	Yes
5	Rainy	Hot	Normal	False	No
6	Overcast	Cool	High	True	Yes
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Hot	High	False	No
11	Sunny	Hot	High	False	Yes
12	Sunny	Hot	High	True	Yes
13	Overcast	Hot	Normal	False	No
14	Sunny	Cool	Normal	False	No

Q 3 | **CO3**

Suppose you developed a Logistic regression model, which is defined as $z=\frac{e^y}{1+e^y}$ for the following training data set, where y is the linear regression model. First develop the regression model

X (Independent variable)	Y (Dependent variable)
2	yes
2.5	yes
1.5	no
0.5	no
1.75	no
3.5	yes

		5	yes		
		5.5	yes		
		Calculate the performance of the model in train and test phase. Find the model's prediction for x=5.25? Are you confident about the model's prediction? Justify your answer.			