Name: **Enrolment No:**	

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, May 2023

Course: Predictive Modelling	**Semester: IV**
Program: MBA (BA)	**Time : 03 hrs.**
Course Code: DSBA 8003	**Max. Marks: 100**

Instructions: Attempt all sections

SECTION A
10Qx2M=20Marks

S. No.		Marks	CO
Q 1	Attempt all multiple choice questions		**CO1**
a.	The purpose of applying data reduction is a) to generate a larger set of variables b) to remove negative values c) to use a smaller set of variables that capture maximum information d) None of the above	**2**	**CO1**
b.	Which of the following is required by K-means clustering? a) defined distance metric b) number of clusters c) initial guess as to cluster centroids d) all of the mentioned	**2**	**CO1**
c.	Movie recommendation systems are an example of: 1. Classification 2. Clustering 3. Reinforcement Learning 4. Regression a) 2 only is correct b) 1 and 2 are correct c) 1 and 3 are correct d) 1, 2 and, 3 are correct e) All are correct	**2**	**CO1**
d.	The main benefit of standardizing a dataset is a) it makes multiple variables of a dataset come to a common scale. b) eliminates negative data values c) makes data interpretation easier.	**2**	**CO1**
e.	What is an outlier? a) data point most proximal to mean b) data point that falls outside the overall pattern.	**2**	**CO1**

	c) data point above or below 3 standard deviations of the mean.		
f.	Which of the following IS NOT a component for a time series plot? a) Seasonality b) Trend c) Cyclical d) Noise e) None of the above	**2**	**CO1**
g.	Financial fraud detection is an example of: a) Prediction problem b) Clustering problem c) Outlier detection problem d) None of these	**2**	**CO1**
h.	What kind of target variables are we dealing with in simple linear regression? a) continous b) binary c) categorical	**2**	**CO1**
i.	On what stage of data exploration are the missing values handled? a) Data transformation b) Data reduction c) Data cleaning d) All of the above	**2**	**CO1**
j.	Which of the following is not a necessary condition for weakly stationary time series? a) Mean is constant and does not depend on time b) Autocovariance function depends on s and t only through their difference |s-t| (where t and s are moments in time) c) The time series under considerations is a finite variance process d) Time series is Gaussian	**2**	**CO1**
SECTION B **4Qx5M= 20 Marks**			
Q2.	What do you understand by data clustering? Explain the difference between different types clustering methods.	**5**	**CO2**
Q3.	What is dimensionality reduction? Explain the difference between feature extraction and feature extraction.	**5**	**CO2**
Q4.	What is curse of dimensionality?	**5**	**CO2**
Q5.	Explain the difference between feature extraction and feature selection.	**5**	**CO2**

<table>
<tr><td colspan="4">SECTION-C
3Qx10M=30 Marks</td></tr>
<tr><td>Q6.</td><td>Explain in detail the steps in Principle component Analysis.</td><td>10</td><td>CO3</td></tr>
<tr><td>Q7.</td><td>What do you understand by a time series? What is stationarity? How do you know if a given time series is stationary or not?</td><td>10</td><td>CO3</td></tr>
<tr><td>Q8.</td><td>A. What do you understand by CART and CHAID? What is the difference between the two?

OR

B. What is data mining? What are the different techniques used in data mining?</td><td>10</td><td>CO3</td></tr>
<tr><td colspan="4">SECTION-D
2Qx15M= 30 Marks</td></tr>
<tr><td>Q9.</td><td>Considering the following confusion matrix, define and compute the following:
a) Accuracy
b) Precision
c) Recall
d) F1 score
e) Sensitivity

<table>
<tr><td colspan="2" rowspan="2">N=165</td><td colspan="2">Predicted</td><td></td></tr>
<tr><td>No</td><td>Yes</td><td></td></tr>
<tr><td rowspan="2">Actual</td><td>No</td><td>50</td><td>10</td><td>60</td></tr>
<tr><td>Yes</td><td>5</td><td>100</td><td>105</td></tr>
<tr><td></td><td></td><td>55</td><td>110</td><td></td></tr>
</table>
</td><td>15</td><td>CO4</td></tr>
<tr><td>Q10.</td><td>A data scientist is trying to build a Breathalyzer. A breathalyzer registers someone's blood alcohol content to tell if they are "over the limit" or "under the influence" of alcohol. It is typically used at roadside police stops to determine if someone is legally able to drive. Answer the following questions:

a. What is the positive class?
b. What would a recall of 70% mean?
c. What would a precision of 90% mean?
d. If the recall is 80% and the precision is 75%, what is the FPR?</td><td>15</td><td>CO4</td></tr>
</table>