Name: **Enrolment No:**	UPES UNIVERSITY OF TOMORROW

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

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.	A graph that uses vertical bars to represent data is called as a) Line graph b) Bar graph c) Scatterplot d) Vertical graph	**2**	**CO1**
c.	Precision is a useful metric in cases where False Positive is a higher concern than False Negatives a) True b) False	**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. c) data point above or below 3 standard deviations of the mean.	**2**	**CO2**
f.	____ are used when you want to visually examine the relationship between two quantitative variables. a) Bar graph b) pie graph c) line graph	**2**	**CO2**

	d) Scatterplot		
g.	Financial fraud detection is an example of: a) Prediction problem b) Clustering problem c) Outlier detection problem d) None of these	**2**	**CO2**
h.	Recall is a useful metric in cases where False Negative trumps False Positive. a) True b) False	**2**	**CO2**
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**	**CO2**
j.	Statement 1: Data transformation works on individual variables. Statement 2: Data reduction works on a set of variables. a) Only statement 1 is true b) Only statement 2 is true c) Both the statements are True d) Both the statements are False	**2**	**CO1**
SECTION B **4Qx5M= 20 Marks**			
Q2.	What do you understand by data cleaning? What is an outlier? Explain the process of outlier detection.	**5**	**CO2**
Q3.	What is dimensionality reduction? Explain the difference between feature extraction and feature extraction.	**5**	**CO1**
Q4.	What is curse of dimensionality?	**5**	**CO2**
Q5.	What is subset selection? Explain forward and backward search.	**5**	**CO1**
SECTION-C **3Qx10M=30 Marks**			
Q6.	Explain in detail the steps in Principle component Analysis.	**10**	**CO2**
Q7.	What do you understand by a time series? What is stationarity? How do you know if a given time series is stationary or not?	**10**	**CO2**
Q8.	A. What do you understand by CART and CHAID? What is the difference between the two?	**10**	**CO2**

	OR B. What is data mining? What are the different techniques used in data mining?		
SECTION-D 2Qx15M= 30 Marks			
Q9.	Considering the following confusion matrix, define and compute the following: a) Accuracy b) Precision c) Recall d) F1 score e) Sensitivity	**15**	**CO3**
Q10.	A. Study the Excel regression output that follows. How many predictors are there? What is the equation of the regression model? Using the key statistics discuss the strength of the model and its predictors. OR B. The following Excel ogive shows toy sales by a company over a 12-month period. As a business analyst what	**15**	**CO3**

N=165		Predicted		
		No	Yes	
Actual	No	50	10	60
	Yes	5	100	105
		55	110	

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.814
R Square	0.663
Adjusted R Square	0.636
Standard Error	51.761
Observations	28

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	131567.02	65783.51	24.55	0.0000013
Residual	25	66979.65	2679.19		
Total	27	198546.68			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	203.3937	67.518	3.01	0.0059
X_1	1.1151	0.528	2.11	0.0448
X_2	–2.2115	0.567	–3.90	0.0006

	conclusions can you reach about toy sales at this company?