QP Code: SN23BTECH201 Reg. No | | | | | | | | | | AY 23

Gandhi Institute of Engineering and Technology University, Odisha, Gunupur (GIET University)

B. Tech (Fifth Semester - Regular) Examinations, November 2025

23BCMPC35001– Data Mining & Predictive Modeling

(CSE-AIML)

Time: 3 hrs Maximum: 60 Marks

Answer ALL questions

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

PART – A **(2 x 5 = 10 Marks)**

Q.1. Answer ***ALL*** questions

	Question	CO #	Blooms Level
a.	What is the role of data visualization in EDA? Give examples.	CO2	K2
b.	Differentiate between the Filter and Wrapper methods.	CO3	K2
c.	Explain the differences between Classification and Regression with examples.	CO4	K2
d.	What does the ROC curve represent in the evaluation of a binary classifier?	CO5	K1
e.	Mention any four ethical issues in Predictive Modeling.	CO6	K2

PART – B **(10 x 5 = 50 Marks)**

<u>Answer ***ALL*** the questions</u>

Marks | CO # | Blooms Level

2. a. Explain briefly the KDD process with a neat diagram. 5 CO1 K1

b. Find the covariance between the two variables X and Y given below: 5 CO2 K3

X	2	4	6	8	10
Y	5	9	12	15	18

(OR)

c. What are the four primary levels of measurement in statistics? Discuss with relevant examples. 5 CO2 K2

d. The following are the weekly sales figures of a store:120, 130, 135, 140, 145, 155, 160, 165, 170, 190. Construct a box plot and identify if there are any outliers. 5 CO2 K3

3.a. Use the Apriori Algorithm to find all frequent itemsets and association rules with minimum support = 50% and minimum confidence = 60% from the given transactions: 5 CO3 K3

Transaction ID	T1	T2	T3	T4	T5
Items Bought	{Milk, Bread, Butter}	{Milk, Bread}	{Bread, Butter}	{Milk, Bread, Butter}	{Milk, Butter}

b. Differentiate between Agglomerative and Divisive Clustering. 5 CO3 K2

(OR)

c. Explain in brief how the ID3 algorithm decides which attribute to split first. A company wants to predict whether a person will buy a computer based on their Age and Income. The following dataset is given: 10 CO3 K3

ID	1	2	3	4	5	6
Age	Young	Young	Middle	Senior	Senior	Middle
Income	High	Medium	High	Low	Medium	Low
Buys_Computer	No	No	Yes	Yes	Yes	No

4.a.	A company recorded monthly advertising (X in \$1000) and sales (Y in \$1000) for 6 months: X (\$1000): 1, 2, 3, 4, 5 Y (Sales): 100, 110, 115, 125, 140 (i) Determine the regression line. (ii) Forecast sales for X = 6. (iii) Find which week has the largest positive deviation.	5	CO4	K3
b.	Write a short note on the following: (i) Bagging (ii) Boosting (iii) Blending (iv) Stacking	5	CO4	K2
	(OR)			
c.	A company wants to predict whether a customer will buy a product (Yes = 1, No = 0) based on age (X in years): X: 22, 25, 30, 35, 40, 45 Y: 0, 0, 1, 1, 1, 1 (i) Fit a logistic regression model. (ii) Predict the probability of purchase for age = 3	5	CO4	K3
d.	Explain the concept of train-test split in machine learning. Discuss its purpose, how the split is performed, and why it is important to keep the test data separate from the training data. Give an example of a typical split ratio and its effect on model evaluation.	5	CO4	K2
5.a.	A disease detection model correctly identified 90% of diseased and 80% of healthy patients. If there were 100 diseased and 200 healthy patients, construct the confusion matrix and find all key metrics.	5	CO5	K3
b.	A dataset records monthly sales over 24 months. (i) How would you smooth the data to reduce short-term fluctuations? (ii) Describe any two techniques for smoothing time series data.	5	CO5	K2
	(OR)			
c.	A company recorded the sales (in units) for the last 6 months: Use Single Exponential Smoothing with a smoothing constant $\alpha = 0.3$ to forecast the sales for Month 7.	5	CO5	K3

Month	1	2	3	4	5	6
Sales	120	130	125	145	150	160

d.	Differentiate between Underfitting and Overfitting in Machine Learning.	5	CO5	K2
6.a.	Explain Big Data and Data Mining in distributed systems using Hadoop and Spark.	5	CO6	K2
b.	Discuss the features and uses of the Weka and Orange tools.	5	CO6	K1
	(OR)			
c.	Describe a case study of a data mining application in healthcare or banking.	5	CO6	K2
d.	List out the Data Mining Tools and explain the role of Python libraries in data mining.	5	CO6	K2

--- End of Paper ---