QP Code: RN22BTECH251 | Reg. No | | | | | | | | | | | AR 22

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

B. Tech (Fifth Semester - Regular) Examinations, November – 2024

22BCMPC35002 – Data Mining and Predictive Modeling

(CSE-AIML)

Time: 3 hrs | Maximum: 70 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	CO #	Blooms Level
a. Describe data preprocessing and why it is important.	CO1	K2
b. Differentiate between filter, wrapper, and embedded approaches.	CO2	K4
c. What are ensemble methods, and why are they useful?	CO3	K1
d. What is cross-validation, and why is it used?	CO4	K1
e. Explain the use of ROC curves and their interpretation.	CO4	K2

PART – B **(15 x 4 = 60 Marks)**

Answer ***All*** the questions	Marks	CO #	Blooms Level
2. a. Discuss the steps in KDD with a neat diagram.	8	CO1	K2
b. A dataset contains the following missing values and outliers: Age: [25, 27, 29, NaN, 35, 1000, 30, NaN, 33, 28] (i) Identify the outliers and replace them with suitable values. (ii) Handle the missing values using mean imputation and median imputation. Discuss the impact of each method on the dataset.	7	CO1	K3
(OR)			
c. Discuss the steps in Predictive Modeling with a neat diagram.	8	CO1	K2
d. Normalize the following dataset using Min-Max normalization and Z-score normalization: Values: [25, 30, 35, 40, 45, 50, 55, 60] (i) Perform Min-Max normalization to scale values between 0 and 1. (ii) Perform Z-score normalization and interpret the transformed data values.	7	CO1	K2
3.a. Given a dataset with two features for five samples: Feature X: [2, 4, 5, 6, 8] Feature Y: [1, 3, 3, 5, 7] (i) Calculate the covariance matrix. (ii) Derive the eigenvalues and eigenvectors. (iii) Perform PCA and project the data onto the first principal component.	15	CO2	K3
(OR)			
b. Explain the steps in Apriori Growth Algorithm along with its limitations. Apply the FP Growth Algorithm with a minimum support count of 3 for the given data set.	15	CO2	K3

TID	T1	T2	T3	T4	T5
Items Bought	f,a,c,d,g,i, m,p	a,b,c,f,l,m, o	b,f,h,j,o	b,c,k,s,p	a,f,c,e,l,p, m,n

4.a. A logistic regression model is used to predict whether a patient has a particular disease (1 for disease, 0 for no disease) based on their age and cholesterol level. A small dataset of five patients is provided below: 8 CO3 K3

Patient	Age(Years)	Cholesterol Level (mg/dL)	Disease (1 = Yes, 0 = No)
1	45	210	0
2	50	220	0
3	55	250	1
4	60	260	1
5	65	270	1

Given this data, answer the following questions:

(i) Set Up the Logistic Regression Model:

(ii) Estimate the Parameters $\beta_0, \beta_1, \beta_2$

(iii) Calculate the Probability of Disease for Each Patient

(iv) Interpret the Coefficients

(v) Predict the Probability of a new patient

b. Describe how a neural network works in predictive modeling. 7 CO3 K2

(OR)

c. Consider a dataset with two features and a binary outcome: 8 CO3 K3

Feature X: [2, 3, 10, 11], Feature Y: [5, 4, 6, 7], Outcome: [0, 0, 1, 1]

(i) Construct two simple decision trees using subsets of this data.

(ii) Explain how Random Forest would combine these trees and the benefit of using ensemble methods in this context.

d. Discuss the steps involved in building a collaborative filtering recommender system. 7 CO3 K2

5.a. A binary classifier yields the following confusion matrix on a test set: 8 CO4 K2

True Positive (TP): 50, True Negative (TN): 30, False Positive (FP): 10, False Negative (FN): 5

(i) Calculate accuracy, precision, recall, F1 score, and specificity.

(ii) Interpret each metric and discuss which metric would be most important if the model were used for detecting fraud.

b. Describe the ARIMA model and its application in time series forecasting. 7 CO4 K2

(OR)

c. Discuss the impact of overfitting and how it can be controlled in predictive modeling. 8 CO4 K2

d. Write a short note on Smoothing Techniques 7 CO4 K1

(i) Moving Average

(ii) Exponential Average

--- End of Paper ---