Reg. No.															

Question Paper Code	13502

B.E. / B.Tech. - DEGREE EXAMINATIONS, APRIL / MAY 2025

Sixth Semester

Computer Science and Engineering

20CSEL605 - PREDICTIVE MODELING

Regulations - 2020

Duration: 3 Hours | Max. Marks: 100

PART - A (MCQ) (10 × 1 = 10 Marks)

Answer ALL Questions

		Marks	K – Level	CO
1.	The goal of supervised learning is to train the model so that it can predict the output when it is given new _____. (a) patterns (b) data (c) features (d) labels	1	K1	CO1
2.	In data analysis, what does smoothing refer to? (a) Combining databases (b) Remove noise from data (c) data cube construction (d) Concept hierarchy	1	K2	CO1
3.	A data set contains four variables, X1, X2, X3, and X4. If a researcher wants to reduce the dimensionality of the data using principal component analysis, what is the maximum number of principal components that can be extracted? (a) 1 (b) 2 (c) 3 (d) 4	1	K3	CO2
4.	Which of the following is a symptom of overfitting in a predictive model? (a) The model has a high training error and a low testing error. (b) The model has a low training error and a low testing error. (c) The model has a low training error and a high testing error. (d) The model has a high training error and a high testing error.	1	K2	CO2
5.	A multiple regression model has _____ (a) only one independent variable (b) more than one dependent variable (c) more than one independent variable (d) exactly one dependent variable	1	K1	CO3
6.	Logistic regression uses _____ function or logistic function which is a complex cost function. (a) quadratic (b) sigmoid (c) lasso (d) linear	1	K2	CO3
7.	The effectiveness of an SVM depends upon: (a) Selection of Kernel (b) Kernel Parameters (c) Soft Margin Parameter C (d) All of the mentioned	1	K2	CO4
8.	A perceptron works by taking in some numerical inputs along with what is known as _____ and _____. (a) weights, bias (b) threshold, bias (c) weighted sum, sigmoid (d) sigmoid, bias	1	K2	CO4
9.	Which PMML tag is used to specify the input data fields? (a) <TransformationDictionary> (b) <DataDictionary> (c) <DataField> (d) <Output>	1	K2	CO5
10.	What is the output for the following code? >myString<- "Hello, World!" > print (myString) (a) myString (b) HelloWorld (c) Hello,World! (d) error	1	K3	CO6

PART - B (12 × 2 = 24 Marks)

Answer ALL Questions

11.	Define Data discretization.	2	K1	CO1
12.	Write about the SEMMA approach.	2	K1	CO1

13.	Write the difference between training, validation, and test sets.	2	K2	CO2
14.	Define overfitting in machine learning.	2	K1	CO2
15.	What inference can you formulate with Bayes theorem?	2	K3	CO3
16.	Define correlation and market basket analysis.	2	K2	CO3
17.	Write about the kernel trick in Support Vector Machines (SVM)	2	K1	CO4
18.	Define a Time Series Model. Give one example.	2	K1	CO4
19.	List the key components of a PMML file.	2	K1	CO5
20.	Describe the role of the <DataDictionary>element in PMML.	2	K2	CO5
21.	List any two features of RapidMiner.	2	K1	CO6
22.	Write the difference between IBM SPSS Statistics and SPSS Modeler?	2	K2	CO6

PART - C (6 × 11 = 66 Marks)

Answer ALL Questions

23.	a) i)	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) : 13, 15, 16, 16, 19, 20, 23, 29, 35, 41, 44, 53, 62, 69, 72 Use min-max normalization to transform the value of 45 for age onto the range [0,1]	6	K3	CO1
	ii)	Use Z-Score normalization to transform the value 45 for age where the standard deviation of age is 20.64 years**.**	5	K3	CO1
		OR			
	b)	Elaborate in detail about building a Statistical model in predictive modeling.	11	K3	CO1
24.	a)	Explain in detail about Multiple Regression by highlighting their purpose, advantages and limitations.	11	K2	CO2
		OR			
	b)	What is oversampling? Explain the Synthetic Minority Over-sampling Technique (SMOTE) with an example. How does it help in addressing data imbalance?	11	K2	CO2
25.	a)	Develop an algorithm for classification using decision trees. Illustrate the algorithm with a relevant example.	11	K3	CO3
		OR			
	b)	Explain and Apply the Apriori algorithm for discovering frequent item sets of the table.	11	K3	CO3

Trans ID	Items Purchased
101	Milk,bread,eggs
102	Milk,juice
103	Juice,butter
104	Milk,bread,eggs
105	Coffee,eggs
106	Coffee
107	Coffee,Juice
108	Milk,bread,cookies,eggs
109	Cookies,butter
110	Milk,bread

26.	a)	Describe about the process of multi layer feed forward neural network classification using back propagation learning.	11	K2	CO4
		OR			
	b)	Explain in detail about Time Series model.	11	K2	CO4

27.	a)	Explain the structure of a typical PMML document with suitable diagram.	*11*	*K2*	*CO5*
		OR			
	b)	Explain how PMML supports ensemble modeling using the <MiningModel> element with suitable examples.	*11*	*K2*	*CO5*
28.	a)	Explain about the features, interface, and use cases of the Rapid Miner tool in predictive modeling.	*11*	*K2*	*CO6*
		OR			
	b)	Explain the capabilities of IBM SPSS Statistics in predictive modeling. How does it differ from IBM SPSS Modeler?	*11*	*K2*	*CO6*