GALGOTIAS UNIVERSITY

ADMISSION NUMBER

School of Business

Bachelor of Business Administration
Semester End Examination - Jun 2024

Duration : 180 Minutes
Max Marks : 100

Sem IV - D1UF401T - Data Mining

General Instructions
Answer to the specific question asked
Draw neat, labelled diagrams wherever necessary
Approved data hand books are allowed subject to verification by the Invigilator

1) Illustrate the need of data mining process in online retail business. Give suitable example. K3(3)

2) How do knowledge elicitation techniques contribute to analyzing domain knowledge for knowledge engineering? K4(4)

3) Illustrate how you would teat missing values in a dataset available for data mining models. Give Example. K3(6)

4) The following is result of neural network model to predict attrition at company of any employee with the help of age, distance from home, Education, hourly rate, monthly income, percent salary hike, total working years, years in current role, years since last promotion, years with current manager, Gender and marital status etc. Develop an approach using neural network algorithm to achieve this result. K6(6)

Education	input	→ Hidden 1	1	0.389
EmployeeCount	input	→ Hidden 1	1	0.736
HourlyRate	input	→ Hidden 1	1	1.407
MonthlyIncome	input	→ Hidden 1	1	-1.160
MonthlyRate	input	→ Hidden 1	1	0.625
PercentSalaryHike	input	→ Hidden 1	1	2.015
StandardHours	input	→ Hidden 1	1	0.894
TotalWorkingYears	input	→ Hidden 1	1	0.901
YearsAtCompany	input	→ Hidden 1	1	-0.229
YearsInCurrentRole	input	→ Hidden 1	1	0.484
YearsSinceLastPromotion	input	→ Hidden 1	1	0.072
YearsWithCurrManager	input	→ Hidden 1	1	-1.272
Intercept		→ No	output	0.242
Hidden 1	1	→ No	output	1.402
Intercept		→ Yes	output	0.133
Hidden 1	1	→ Yes	output	-1.780

Note. The weights are input for the logistic sigmoid activation function.

5) Design an approach using CRISP-DM framework to building predictive models in the Modeling phase to identify customers at K6(6)

risk of churn in Telecom organization. Which machine learning algorithms would you consider using, and how would you evaluate their performance? Discuss its role in systematically guiding the stages of data mining, fostering a structured approach to extracting valuable insights and informing strategic decisions.

6) You are tasked to support an ecommerce company in their analysis of sales data. Analyse the key concepts of collaborative filtering techniques that can be applied in this context. K4(8)

7) Refer to the given dataset and answer: how does a local dairy store apply the concept of support to enhance sales of butter? What is the value of support for product butter? K3(9)

Transaction ID	Customer ID	Product ID	Product Name	Quantity
1	101	1001	Butter	1
1	101	1002	Bread	2
2	102	1001	Butter	1
2	102	1003	Milk	1
3	103	1002	Bread	1
3	103	1004	Cheese	1
4	104	1001	Butter	1
4	104	1005	Eggs	1
...	...	...	...	...

8) As the leader of a project team striving to elevate customer satisfaction in a prestigious hotel chain, how would you organize the utilization of rule-based systems and their applications in the hospitality sector to enhance guest experiences through personalized services? K3(9)

9) Determine the best choice of relationships from the following data and results of simple linear regression and multiple linear regression covering ticket prices, attendance rates, and revenue generation for entertainment events to optimize pricing strategies. K5(10)

Event	Ticket Price ($)	Attendance	Revenue ($)
Event 1	20	1000	20000
Event 2	25	950	23750
Event 3	30	900	27000
Event 4	35	850	29750
Event 5	40	800	32000
Event 6	45	750	33750
Event 7	50	700	35000
Event 8	55	650	35750
Event 9	60	600	36000
Event 10	65	550	35750

Simple Linear Regression Results

Ticket Price vs Revenue
Coefficient: 525.0
Intercept: 10000.0
R-squared: 0.9504

Attendance vs Revenue
Coefficient: -25.0
Intercept: 37500.0
R-squared: 0.903

Ticket Price and Attendance vs Revenue

Ticket Price Coefficient: 425.0
Attendance Coefficient: -250.0
Intercept: 17500.0
R-squared: 0.9796477156437992

10) A dataset with 1000 data points is split into a training set of 80% (800 data points) and a test set of 20% (200 data points). A decision tree algorithm was trained on the training set and K4(12)

achieved an accuracy of 90%. Calculate the expected accuracy on the test set if the algorithm is unbiased and performs equally well on both sets. Explain the results. What will change when training set of 75% and test set of 25% are split?

11) A hospital wants to use data mining to identify patients at risk of developing a certain disease. The hospital has a dataset of patient medical records and demographics. K6(12)

a)Discuss the data mining techniques that can be used to develop a predictive model for disease risk (4 marks).

b)Explain the importance of data preprocessing and feature selection for building an accurate predictive model (4marks).

c) Describe the ethical considerations that need to be taken into account when using data mining for healthcare applications (4 marks).

Patient ID	Age	Gender	Blood Pressure	Cholesterol Level	Family History	Disease Risk
1	45	Male	130/80	200	Yes	High
2	50	Female	140/90	220	No	High
3	35	Male	120/75	180	Yes	Low
4	60	Female	150/95	240	Yes	High
5	40	Male	125/80	190	No	Low
6	55	Female	135/85	210	No	Medium
7	48	Male	128/82	195	Yes	High
8	42	Female	132/88	205	No	Medium
9	58	Male	145/92	230	Yes	High
10	38	Female	127/79	185	No	Low

Metric	Value
Accuracy	0.8
Precision	0.8667
Recall	0.8
F1 Score	0.819
ROC AUC	0.8333

12) Critically appraise the interpretability of cluster analysis results from the following dataset in identifying meaningful customer segments for targeted marketing strategies. K5(15)

Customer ID	Age	Gender	Income	Total Transactions	Total Amount Spent	Cluster
1	35	0	50000	10	500	1
2	45	1	70000	20	1000	2
3	30	1	40000	5	300	0
4	60	0	80000	15	800	2
5	40	1	60000	12	600	1
6	55	0	75000	25	1200	2
7	28	0	35000	8	400	0
8	42	1	65000	18	900	2
9	48	0	72000	22	1100	2
10	33	1	45000	7	350	1