Daffodil International University

Faculty of Science & Information Technology
Department of Computer Science & Engineering
Mid Examination, Fall 2024
Course Code: CSE315, Course Title: Introduction to Data Science
Level: 3 Term: 1 Batch: 63& 62

Time: 01:30 Hrs. **Marks: 25**

Answer ALL Questions

[The figures in the right margin indicate the full marks and corresponding course outcomes. All portions of each question must be answered sequentially.]

1.	a)	A telecom company has gathered information on customer age, contract type, monthly charges, and customer complaints in an effort to reduce retention rates. In summary, **describe** how the company can use these variables to predict which customers are most likely to leave.	2	CO1
	b)	A university wants to conduct a survey to understand student satisfaction across all departments. Since surveying every student is not feasible, they decide to use sampling techniques. **Explain** which sampling method (e.g., random sampling, stratified sampling, or systematic sampling) would be most appropriate for this scenario and why.	3	
2.	a)	Given the following dictionary of employee data: **employee_data = {** **'Name': ['Alice', 'Bob', 'Charlie', 'David'],** **'Age': [25, 30, 35, 40],** **'Department': ['HR', 'Finance', 'IT', 'Marketing'],** **'Salary': [50000, 60000, 70000, 80000]** **}** i. Write Python code to convert this dictionary into a Pandas DataFrame. ii. Display the first two rows of the DataFrame. iii. Write code to find the average salary of the employees.	3	CO2
	b)	Given the following list of ages for a group of people: 8, 9, 91, 100, 96, 5, 39, 2, 34, 25, 28, 22, 54, 68, 80, 11, 74, 28, 13, 6 **Calculate** the first quartile (Q1) and third quartile (Q3) of the data. Use the 1.5 IQR (Interquartile Range) method to **identify** any outliers.	3	
	C)	A company collects the following data on the number of hours employees spend on training and their corresponding performance scores: Hours of Training: 2, 3, 5, 7, 8, 9, 11, 12 Performance Score: 50, 55, 65, 70, 80, 90, 95, 105 i. **Explain** how you would apply linear regression to model the relationship between the hours of training and performance score. ii. Based on this model, **describe** how you can predict the performance score for an employee who spends 6 hours on training.	4	
3.	a)	What is the difference between the generalized Bayes rule and the naïve Bayes rule? **Describe** the meaning of the term "naïve" for the Naive Bayes classifier.	2	CO2

b)	**Explain** why Naive Bayes is a good option and how its assumptions and features match the situation. Talk about real-world applications where Naive Bayes is often utilized.	3	
c)	You are given a randomly selected dataset of ten email messages to build a Naive Bayes classifier to identify whether an email is spam or not based on the occurrence of certain words, including Money, Free, and Win.	5	

Email	"Money" (Yes=1, No=0)	"Free" (Yes=1, No=0)	"Win" (Yes=1, No=0)	Label (Spam=1, Not Spam=0)
1	1	1	0	1
2	0	1	0	0
3	1	0	0	1
4	0	0	0	0
5	1	1	1	1
6	0	0	1	0
7	0	1	1	1
8	0	1	1	0
9	1	1	1	1
10	0	0	0	0
11	0	0	0	0
12	1	1	1	1
13	1	0	0	0
14	1	0	0	1
15	1	1	0	1

A vector of words represents each email. Each email is labeled as either spam (1) or not spam (0).

i) **Determine** the prior probability of each label (spam or not spam).

ii) **Determine** the Likelihood of each attribute (Money, Free, and Win).

iii) **Calculate** the probability that a new email containing the words Money, Win, but not free is spam or not spam.