Reg. No.: 20BCE1848

Name : Mohd. Alhafeez Pathan

VIT

Vellore Institute of Technology

Continuous Assessment Test I – January 2023

Programme	:	**B.Tech (CSE,AI & ML, AIR, CPS), M.Tech(Int. SE)**	Semester	:	**Winter 22-23**
Course Title	:	**Essentials of Data Analytics**	Code	:	**CSE 3506**
			Class Nbr(s)	:	**CH2022235000980, CH2022235000979, CH2022235000983, CH2022235000982, CH2022235000981, CH2022235000984**
Faculty (s)	:	**Asnath Phamila Y, Vergin Raja Sarobin M, N. M. Elango, Rajalakshmi R, A Sheik Abdullah, Trilok Nath Pandey**	Slot	:	**G2**
Time	:	**1½ Hours**	Max. Marks	:	**50**

Answer all the Questions

1. The Election Commission of India (ECI) is a constitutional body. The Constitution of India established it to conduct and regulate elections in the country. Thus, the Election Commission is an all-India body because it is common to both the central and state governments.

The commission uses the following steps to carry out the election process in a smooth and good way of conduct. It has to be noted that, each stage in the election process maintains different forms of data based on the election rules and guidelines.

The ECI assigns a data analyst at each stage to monitor and manage the data at each segment because each attribute considered plays an essential role in maintaining the voters' details. As a data analyst, from data collection to modelling stage in generating the voter ID, illustrate the following: 10

- Type of data (Categorical, Continuous, Text and Image)
- Attributes involved in each stage
- Need for information gathering
- Need for data pre-processing
- Data modelling
- Interpretation and evaluation

2. The salesperson wants to know about the number of smart TVs that he can sell based on the lockdown days during that year. You are given the data as shown in the below Table. 15

Number of lock down days	Sale of smart TV
82	15
92	25
83	17
97	28
131	41

i) Calculate the appropriate correlation coefficient between the number of lock down days and sale of smart TVs. Comment on the relationship. (5 marks)

ii) Obtain the regression equation to predict the sale of smart TVs (y) as a function of the

number of lockdown days (x). [8 marks]

iii) Using the obtained regression equation, predict the sale of smart TVs, if the number of lockdown days is 150. [2 marks]

3. The following set 'S' shows the stock price for 8 months, in terms of $. 10

S={140, 145, 150, 156, 162, 167, 175, 182}

a. Determine the correlation between the above given data to understand the time series data pattern. [6 Marks]

b. Identify and elaborate the efficient time series forecast method which combines past variable values and lagged forecast errors. [4 Marks]

4. An experiment was carried out by a pharmaceutical company to analyze the effectiveness of different drugs for treating hypertension. The following table shows the drugs prescribed for treating hypertension among 15 patients and their corresponding recovery time, in days. The objective is to assess whether consuming a specific type of drug will result in fast recovery of hypertension. 15

Patient ID	Drug	Recovery Time (Days)
Id101	A	75
Id102	A	95
Id103	C	85
Id104	C	45
Id105	A	60
Id106	B	70
Id107	B	45
Id108	B	75
Id109	C	65
Id110	C	75
Id111	B	55
Id112	B	65
Id113	C	75
Id114	A	65
Id115	A	75

- State the Null and Alternate hypothesis for this study. What probability level can be chosen for this study and why? [2 marks]
- List the assumptions made for your analysis. [3 marks]
- Is there a significant difference in recovery time between the drugs? [Note $F_{Critical}$= 3.6823 , Significance level=0.05] [8 marks]
- Interpret your answer. [2 marks]

⇔⇔⇔

Reg. No.:

Name :

VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Continuous Assessment Test II – March 2023

Programme	:	**B.Tech (CSE,AI & ML, AIR, CPS), M.Tech(Int. SE)**	Semester	:	**Winter 22-23**
Course Title	:	**Essentials of Data Analytics**	Code	:	**CSE3506**
			Class Nbr(s)	:	**CH2022235000980, CH2022235000979, CH2022235000983, CH2022235000982, CH2022235000981, CH2022235000984**
Faculty (s)	:	**Asnath Phamila Y, Vergin Raja Sarobin M, N. M. Elango, Rajalakshmi R, A.Sheik Abdullah, Trilok Nath Pandey**	Slot	:	**G2**
Time	:	**1½ Hours**	Max. Marks	:	**50**

Answer all the Questions

Q. No	Question Text	Marks
1.	a) Compare linear regression and logistic regression for a business use case, with suitable sketches. **[4 Marks]**	
	b) Find the probability of renting a cab with the following attributes namely, Sedan model (with value 1), AC (with value 1) and rent (Rs. 3000). For calculating the above probability, assume the bias (B0), and the coefficients of the attributes (car model, AC and rent) as 0.25, 0.93, 0.78, 0.85 respectively. **[6 Marks]**	10
2.	Assume that an online-shopping store wanted to know about the potential customers who may later become VIP customers. They have given their data (with n number of features) which reveals that only 30% are VIP customers and the remaining 70% are non-VIP customers. The task is to design a suitable predictive model to identify the future VIP customers. Team A has designed a model which performed badly with more mis-classification. Team B tried with another classifier, but that too resulted in a poor performance. Now, you are assigned to complete this task to improve the performance by combining the advantages of individual classifiers. Elaborate the identified technique that you will apply, and illustrate with a neat sketch.	10
3.	Consider the following dataset with respect to sanctioning the loan for the customers of a XYZ bank.	
	a. Apply suitable splitting criteria to construct a tree-based classifier for the given dataset. **[8 Marks]**	15

b. Derive all the decision rules from the constructed model **[4 Marks]**

c. Mention the conditions for a node to be pure. Also, list the pure nodes at each level. **[3 Marks]**

User Id	Area	House Type	Income	Previous_Customer	Loan Sanction Status
U 1	Suburban	Detached	High	No	No
U 2	Suburban	Detached	High	Yes	No
U 3	Rural	Detached	High	No	Yes
U 4	Urban	Semi-detached	High	No	Yes
U 5	Urban	Semi-detached	Low	No	Yes
U 6	Urban	Semi-detached	Low	Yes	No
U 7	Rural	Semi-detached	Low	Yes	Yes
U 8	Suburban	Terrace	High	No	No
U 9	Suburban	Semi-detached	Low	No	Yes
U 10	Urban	Terrace	Low	No	Yes
U 11	Suburban	Terrace	Low	Yes	Yes
U 12	Rural	Terrace	High	Yes	Yes

4. The annual income and spending score on a scale of 1 to 10 is listed below for a sample of six different customers.

Customer Id	Annual Income (Lakhs)	Spending Score(1-10)
D1	8	7
D2	14	2
D3	15	6
D4	19	1
D5	22	8
D6	25	4

15

a. Illustrate how this customer dataset can be subdivided into two customer groups with similar characteristics, using appropriate method that is robust to outliers. Analyse its time complexity. [Hint: In iteration-1 assume that, the Customer 1 (D1) and Customer 4 (D4) don't have any common characteristics. Try at least two iterations]. **[5 Marks]**

b. Apply bottom-up model based Hierarchical clustering on the above dataset and analyse its time complexity. Draw the corresponding dendrogram and mention the customer IDs in each cluster. **[10 Marks]**

⇔⇔⇔