LNMIIT/B. Tech./CSE/PC/2021-22/ODD/ **CSE-327/ET**

LNMIIT
The LNM Institute of Information Technology

The LNM Institute of Information Technology

Computer Science & Engineering

Introduction to Data Science (CSE327)

Exam Type: End Term
Date: 09 December 2021

Time: 180 Minutes **Max. Marks:**35

Name: VAIBHAV **Enrollment No:** 19UCS181

Instruction:

- **Attempt all the questions. Write the answers sequentially in order.**
- **Marks for each question mentioned against them.**
- **Calculators and Statistical Tables are allowed during the exams.**
- **No doubt clearing at the time of examination. If you have any doubt regarding the question, you can make your assumption. You must write your assumption clearly, before you start attempting that question. If your doubt/ confusion and assumptions are genuine then only Instructor will entertain that assumption and evaluate the question based on the assumptions otherwise your doubt/confusion/assumption will be simply ignored.**

1. Test the hypothesis with 5% significance that the population variances of the systolic blood pressure of healthy subjects (status=0) and subjects with hypertension (status=1) are equal. The dataset contains $n1 = 25$ subjects with status 0 and $n2 = 30$ with status 1 (Dataset in the Table below).
[Given $F_{0.05,\ 24,\ 29} = 1.900$, $F_{0.95,\ 24,\ 29} = 0.514$, $F_{0.025,\ 24,\ 29} = 2.154$, $F_{0.975,\ 24,\ 29} = 0.451$] (4 Marks)

S.No.	Status	mmHg	S.No.	Status	mmHg	S.No.	Status	mmHg	S.No.	Status	mmHg
1	0	120	15	0	114	29	1	127	43	1	152
2	0	115	16	0	105	30	1	141	44	1	135
3	0	94	17	0	115	31	1	149	45	1	134
4	0	118	18	0	134	32	1	144	46	1	161
5	0	111	19	0	109	33	1	142	47	1	130
6	0	102	20	0	109	34	1	149	48	1	125
7	0	102	21	0	93	35	1	161	49	1	141
8	0	131	22	0	118	36	1	143	50	1	148
9	0	104	23	0	109	37	1	140	51	1	153
10	0	107	24	0	106	38	1	148	52	1	145
11	0	115	25	0	125	39	1	149	53	1	137
12	0	139	26	1	150	40	1	141	54	1	147
13	0	115	27	1	142	41	1	146	55	1	169
14	0	113	28	1	119	42	1	159			

2. Mathematically define the clustering technique. (Marks- 4)

3. Specify three drawbacks/ properties of k-means clustering technique which are addressed/ overcome by DBSCAN clustering technique and how it was overcome. (Marks- 3)

4. Specify the mathematical and logical differences of Single linkage and Complete linkage algorithms. (Marks-2)

5. What is support vectors in SVM. Why it is called support vectors? (Marks-2)

Observe the Table below and answer questions 06 to ~~09~~ 08.

Table below shows the small corpus that contains two documents D1 and D2.

Document ID	Text in Document
D1	I Like Book Reading. They Do Not Like Book Reading.
D2	I Do Not Like Outdoor Games. They Like Outdoor Games.

Let $P(w_i)$ denote the probability of a word w_i, and *Count* is the number of times word w_i occurs in the corpus. We have $P(w_i) = Count(w_i)/\text{Total } Number\ of\ words\ in\ corpus.$

6. Suppose Bigram Language Model for the corpus. What is the probability of the sentences:

   (a) "I Like Outdoor Games." (b) "I Do Not Like Book Reading." [2 Marks]

7. Compute One-Hot Encoding, Bag-of-Words, Bag of 2-Gram, and TF-IDF representations of D1 and D2. [6 Marks]

8. Compute Cosine Similarity and Euclidean Distance between Bag-of-Words, Bag of 2-Gram, and TF-IDF representations of D1 and D2. [2 Marks]

9. Describe Parts of Speech tagging Problem (POS Tagging) with an example. How Viterbi Algorithm performs POS Tagging, please describe in details. [3 Marks]

10. Describe the Word-sense disambiguation problem (WSD) with an example. Mathematically describe what is Naive Bayes assumption and how it helps to perform WSD. [3 Marks]

11. Suppose that you are hired as a predictive data analyst for customer churn in the Airtel Company as many customers are leaving Airtel for other Telecom companies (e.g., Jio and BSNL). Describe the major steps of Data Science pipeline for customer churn problems at Airtel. [4 Marks]