INDIAN STATISTICAL INSTITUTE

MS in QMS

TEST ON MULTIVARIATE DATA ANALYSIS

Date: 24 April 2024 Time: 3 hours Maximum Marks: 50

Answer as many questions as you can. The maximum you can score is 50

1. What is multicollinearity in linear regression? Why is it problematic? How can we detect and measure multicollinearity? Describe various methods for tackling multicollinearity in linear regression, providing clear explanations of their logic and procedures.

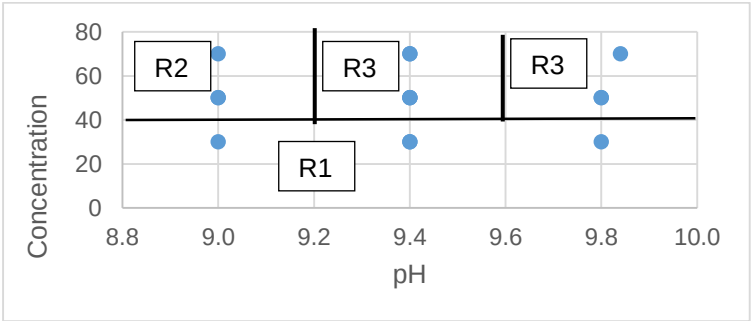
[10]

2. The data provided in the table below is collected to study the impact of filling speed and operating pressure on the outcome of a soft drink filling process. We need to develop a classification tree model to predict the outcome based on these two features. Identify the feature and branching or cut point of the feature for the first split.

SL No	Speed	Pressure	Outcome	SL No	Speed	Pressure	Outcome
1	100	15	Fail	10	200	10	Pass
2	300	20	Pass	11	100	10	Fail
3	100	10	Fail	12	100	20	Fail
4	300	15	Fail	13	300	20	Pass
5	200	15	Fail	14	300	10	Pass
6	100	20	Pass	15	200	20	Pass
7	300	15	Fail	16	100	15	Fail
8	300	10	Pass	17	200	15	Pass
9	200	20	Pass	18	200	10	Pass

[10]

3. A model has been developed to estimate the migration time of the ethambutol hydrochloride peak using buffer pH and buffer concentration as features, employing data partitioning as depicted in the figure below



The data employed to construct the aforementioned model is provided in the table below. Please compute the predicted values of the migration time and the mean square error.

SL No	Concentration	pH	Migration Time	SL No	Concentration	pH	Migration Time
1	70	9.0	2.79	9	30	9.4	3.51
2	50	9.4	2.85	10	70	9.4	4.93
3	30	9.0	2.32	11	30	9.4	1.97
4	50	9.8	2.26	12	50	9.8	4.95
5	70	9.8	3.28	13	30	9.8	2.77
6	50	9.0	1.89	14	50	9.0	3.67
7	50	9.4	2.99	15	70	9.4	2.08
8	50	9.4	2.95				

[12]

4. A k-nearest neighbor model has been developed with k = 3 to estimate yield based on temperature, time, and viscosity. The data provided in the table below is used for this purpose. Please compute the estimated yield for the following conditions: temperature = 160, time = 75, and viscosity = 12

SL No	Temperature	Time	Viscosity	Yield
1	276	63	10	82.1
2	127	73	6	43.2
3	278	66	12	80.6
4	266	88	8	84
5	234	71	8	72.7
6	182	69	9	55.7
7	121	67	13	32.4
8	231	77	13	66.4
9	160	86	15	43.6
10	108	61	15	25.5

[12]

5. What is ensemble learning? What are its pros and cons? Explain different ensemble learning methods, including their logic and procedures.

[10]

6. Write short notes on the following:
   a. Cross validation
   b. Hyper parameter tuning

[5]