

Bansilal Ramnath Agarwal Charitable Trust's
VISHWAKARMA INSTITUTE OF TECHNOLOGY, PUNE - 411037.
 (An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Examination: ESE

Year: S.Y. Common

Branch: SEDA

Subject: Data science

Subject Code: MD 2201

Max. Marks: 60

Total Pages of Question Paper: 2

Day & Date: Wednesday, 6.12.23

Time: 10.00 a.m. to 12.00 p.m.

Instructions to Candidate

1. All questions are compulsory.
2. Neat diagrams must be drawn wherever necessary.
3. Figures to the right indicate full marks.

Q.N o.	CO No	BT No		Max marks																								
Q.1.	1	1	What are qualitative and quantitative variables? What are their types ? Illustrate with examples.	12																								
Q. 2. (A)	2	2	End semester scores of Mathematics subject are distributed normally, with mean 78 and standard deviation 12 . Student 'A' scored 80 marks in end semester mathematics subject . What is his z-score?	4																								
(B)	2	2	The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a sample mean of 134 calories with a standard deviation of 17 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips? We have verified the independence, sample size, and skew conditions are satisfied. Take alpha as 5% .	6																								
Q. 3. (A)	3	1	Calculate the distance between points A(1,-3,2) and B(3,4,5) using i. Manhattan Distance metric and ii. Euclidean Distance metric	4																								
(B)	3	2	For a given univariate function $f(x) = 3x^2 + 5y^2 + 5x + 4y + 5$ find out the optimal local minimum and global minimum	4																								
Q. 4. (A)	4	4	For a certain model of the car, information is gathered on the resale price in dollars and the number of miles driven (in thousands of miles). The linear regression equation for this model is found to be $\hat{y} = 24784 - 343.58x$ with $r^2 = +0.9452$. Here \hat{y} denotes price of the car and x denotes miles driven. Estimate the resale price of the car of this model which has been driven 35000 miles. Also interpret the model coefficients and r^2	8																								
(B)	4	1	What is logistic/sigmoid function? Draw its graph and explain how it is used in logistic regression.	4																								
Q. 5. (A)	5	4	Apply Nearest neighbor algorithm to the following data to classify the given point. <table><tr><th>Brightness</th><th>Saturation</th><th>class</th></tr><tr><td>60</td><td>10</td><td>Red</td></tr><tr><td>10</td><td>25</td><td>Red</td></tr><tr><td>60</td><td>90</td><td>Blue</td></tr><tr><td>70</td><td>70</td><td>Blue</td></tr><tr><td>60</td><td>80</td><td>Blue</td></tr><tr><td>40</td><td>20</td><td>Red</td></tr><tr><td>50</td><td>50</td><td>?</td></tr></table>	Brightness	Saturation	class	60	10	Red	10	25	Red	60	90	Blue	70	70	Blue	60	80	Blue	40	20	Red	50	50	?	6
Brightness	Saturation	class																										
60	10	Red																										
10	25	Red																										
60	90	Blue																										
70	70	Blue																										
60	80	Blue																										
40	20	Red																										
50	50	?																										
Q. 5. (B)	5	4	Define Gini Impurity, Entropy and misclassification Error used in decision trees.	4																								
OR alternate option for Q.5 (A) AND Q. 5 (B) together as Q. 5 (C)																												

Q. 5 (C)	5	4	What is Bayes theorem? Apply Bayes theorem to the following problem- A doctor knows that cold causes fever 60% of the time. Prior probability of any patient getting cold is 1/80000 and prior probability of any patient getting fever is 1/50. Find the probability that a patient has cold given that he has fever.	10
Q. 6. (A)	6	2	Explain training set and test set and why it is required to split the data.	4
(B)	6	3	What is the relationship between F- measure , precision and recall? Find these three measures for a classifier with TP=60, TN=40, FP=30, FN=20.	4

Course Outcomes:

CO1: Apply data processing and data visualization techniques

CO2: Perform descriptive and inferential statistical analysis

CO3: Utilize appropriate distance metrics and optimization techniques

CO4: Implement supervised algorithms for classification and prediction

CO5: Implement unsupervised classification algorithms

CO6: Evaluate the performance metrics of supervised and unsupervised algorithms

*Blooms Taxonomy (BT) Level No:

1. Remembering; 2. Understanding; 3. Applying; 4. Analyzing; 5. Evaluating; 6. Creating