Title : Question Paper **FF No. 868**

Reg.No.	1	2	2	1	1	5	8	0

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:**

Subject: Data science **Subject Code: MD 2201**

Max. Marks:60 **Total Pages of Question Paper: 1**

Day & Date: Wed : 22/11/23 **Time:** 10.30 am -12.30 pm

Instructions to Candidate

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

<table>
<tr><th>Q.No.</th><th>CO No</th><th>BT No</th><th></th><th>Max marks</th></tr>
<tr><td>Q.1.</td><td>1</td><td>1</td><td>Part of an annual transparency report published by a leading multinational technology company is as shown below –

<table>
<tr><th>Country</th><th>CR_ req</th><th>CR_ compl in %</th><th>UD_ req</th><th>UD_ compl in %</th><th>Hemi</th><th>HDI</th></tr>
<tr><td>Austria</td><td>21</td><td>100</td><td>134</td><td>32</td><td>Southern</td><td>High</td></tr>
<tr><td>Belgium</td><td>10</td><td>33</td><td>361</td><td>73</td><td>Northern</td><td>High</td></tr>
<tr><td>Brazil</td><td>224</td><td>67</td><td>703</td><td>82</td><td>Southern</td><td>Medium</td></tr>
<tr><td>Somalia</td><td>104</td><td>31</td><td>227</td><td>61</td><td>Southern</td><td>Poor</td></tr>
<tr><td>USA</td><td>92</td><td>63</td><td>5950</td><td>93</td><td>Northern</td><td>High</td></tr>
</table>
Where, the variables are as follows -
CR_req: Content removal requests
CR_comp: Content removal compliance in %
UD_req: User Data requests
UD_compl: User Data compliance in %
Hemi: Hemisphere
HDI: Human Development Index

Identify each variables as Discrete Numerical, Continuous Numerical, Ordinal Categorical or Regular Categorical with justification</td><td>12</td></tr>
<tr><td>Q. 2. (A)</td><td>2</td><td>2</td><td>What are type-I and type-II errors in hypothesis testing? Which error should be minimized while giving a judgement in the court of law? Justify</td><td>4</td></tr>
<tr><td>(B)</td><td>2</td><td>2</td><td>A sample of 50 news paper readers were asked about the total hours they spend per week in reading the newspaper. The group in the sample had an average of 3.2 hours with a standard deviation of 1.74. Calculate the 95% confidence interval range, based on this data.</td><td>6</td></tr>
<tr><td>Q. 3. (A)</td><td>3</td><td>1</td><td>Calculate the distance between points A (2,7,4) and B (3.2,4.8,5.8) using i. Manhattan Distance metric and ii. Euclidean Distance metric</td><td>4</td></tr>
<tr><td>(B)</td><td>3</td><td>2</td><td>For a given univariate function $f(x) = 2x^4 - 8x^3 - 112x^2 + 1717$ find out the optimal local minimum and global minimum</td><td>4</td></tr>
<tr><td>Q. 4. (A)</td><td>4</td><td>4</td><td>As an outcome of a linear regression process, following performance values are obtained. SSR also known as sum of squares due to regression = 92.48; SSE also known as sum of squares due to error = 12.87. Calculate the value of R^2 and comment if this regression fit is a good fit or not.</td><td>4</td></tr>
</table>

(B)	4	1	A group of 10 customers, having yearly savings between 0 to 6 lacs are considered in the following logistic regression example. Loan status as 0 indicates a loan defaulter and 1 indicates non-defaulter. Amount in Savings (in lacs) / Loan Status 1.00 / 0 1.25 / 0 1.5 / 0 1.8 / 1 2.25 / 0 2.4 / 0 3.4 / 0 4.6 / 0 5.2 / 1 5.8 / 1 The logistic regression function after Maximum Likelihood estimation is given as $\Pi = (1 / 1 + e^{-(-1.07778 + 1.5046 * \text{Savings})})$. Answer the following – 1. If a loan applicant with annual savings of Rs. 2.5 Lacs approaches the bank, should the application be processed for loan disbursement or rejected? 2. Based on the predictions made using logistic regression for all 10 data, how many times the predictions would match with the actual loan status? What will be the classification accuracy of the logistic regression as a classifier?	8
Q. 5. (A)	5	4	The training data for a supervised classification is as follows – X_1= (1.8,1.6, 1), X_2= (2,1.8, 1), X_3= (3.2,2.4, 1), X_4= (2.4,2.6, 1), X_5= (6.5,4.2, 2), X_6= (7.3,4.3, 2), X_7= (6.5, 4.2, 2), X_8= (7.0,4.8, 2). The test datapoint is at (4.6, 3.2). Use 1. Nearest Neighbor assign appropriate class to the test point 2. K - Nearest Neighbor with k = 3 to assign appropriate class to the test point 3. Weighted / Modified K - Nearest Neighbor with k = 3 to assign appropriate class to the test point	6
Q. 5. (B)	5	4	Consider 50 patterns that are split in 3 classes with 12, 28 and 10 samples respectively. Calculate the Entropy impurity, Gini impurity and Misclassification impurity at the node.	4
			OR alternate option for Q.5 (A) AND Q. 5 (B) together as Q. 5 (C)	
Q. 5 (C)	5	4	The table below gives the following training data – Sr. No. / Worker / Mood / Job / Time / Good Work Quality? 1 / Sam / Bad / Painting / Morning / Yes 2 / Sam / Good / Plumbing / Evening / Yes 3 / Ashwin / Bad / Painting / Morning / No 4 / Ashwin / Bad / Plumbing / Evening / No 5 / Ashwin / Good / Washing / Morning / Yes 6 / Sam / Good / Washing / Evening / Yes 7 / Sham / Good / Painting / Morning / Yes 8 / Sham / Bad / Plumbing / Evening / No 9 / Ashwin / Bad / Washing / Morning / No 10 / Ashwin / Good / Washing / Evening / Yes 11 / Sam / Good / Painting / Evening / Yes 12 / Sham / Bad / Washing / Morning / No 13 / Sham / Good / Washing / Evening / Yes 14 / Sam / Bad / Plumbing / Morning / Yes 15 / Sham / Good / Painting / Morning / No Using Naïve Bays Classification, estimate the class for 'Good Work Quality' for the given feature vector X = { Worker = Sham, Mood = Bad, Job = Painting, Time = Morning}	10
Q. 6. (A)	6	2	Explain the 'Hold-out' approach used in planning the training and testing data	4
(B)	6	3	A Confusion matrix as shown below, is observed for a machine learning algorithm, used to detect authentic messages and spam messages. Consider the positive class as – Authentic and Spam as negative class / Predicted authentic / Predicted Spam Actual authentic / 1202 / 5 Actual Spam / 29 / 54 Calculate Accuracy, precision, recall and f-score	4