FF No. 868

Reg.No.							

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.

8

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 $\Pi = (1 / 1 + e^{-(-4.07778 + 1.5046 * \text{Savings})})$.

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?

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 –

10

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}

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

4

	Predicted authentic	Predicted Spam
Actual authentic	1202	5
Actual Spam	29	54

Calculate Accuracy, precision, recall and f-score

Title : Question Paper

FF No. 868

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

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 + 1**

Day & Date: Thursday, 11/05/23 | **Time:** 11.00 am -1.00 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.

Q.No.	CO No	BT No		Max marks
Q.1.	1	1	Match the pair and explain each term on written on left side (shown in numbers) with an example – 1.Validity — A. Degree to which data is consistent 2. Accuracy — B. Degree to which data conforms to constraints 3. Uniformity — C. Degree to which data is close to true values 4. Consistency — D. Degree to which data is specified using same Units	8
Q. 2. (A)	2	2	What is a 'Choose function' in Binomial distribution. What is the probability of getting a desired outcome of '3', 4 times when an unbiased dice is rolled 10 times?	6
(B)	2	2	A group of researchers want to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of 3-yr old children born to mothers who were on this medication during pregnancy. Previous studies suggest that the SD of IQ scores is 18 points. How many such children should the researchers sample in order to obtain a 95% confidence interval with a margin of error ≤ 4 points? *(Z* for 95% Confidence Interval = 1.96)*	6
Q. 3. (A)	3	1	Calculate the distance between points A (4,5,6) and B (7.2,2.8,6.8) using i. Manhattan Distance metric and ii. Euclidean Distance metric	6
(B)	3	2	For a given univariate function $f(x) = 3x^4 - 4x^3 - 12x^2 + 3$ find out the optimal local minimum and global minimum value.	4
Q. 4. (A)	4	4	Two wheeler sales of a leading brand are modelled by a Multiple Linear Regression equation – Sales = - 22036.74 – 1007.61 Unit Cost + 358.88 Ad Exp + 674. 92 Prom Exp. Unit Cost of the Vehicle is in 'Thousand Rs, Ad Exp. is Rs. in Lac and Prom Exp. is also Rs. in Lac. Interpret the four coefficients.	6
(B)	4	1	How is the figure of merit R^2 calculated in regression? Explain with a neat sketch. Also justify that value of R^2 tending to 1 indicates a good fit of the regression model and a value tending to 0 indicates a bad fit	4
Q. 5. (A)	5	4	The training data for a supervised classification is as follows – X_1= (1.8,1.4, 1), X_2= (2.2,1.8, 1), X_3= (3.0,2.4, 1), X_4= (2.2,2.4, 1), X_5= (6.2,4.4, 2), X_6 = (7.2,4.4, 2), X_7= (6.6, 4.4, 2), X_8= (7.0,4.6, 2). The test datapoint is at (4.4, 3.3). Use K-nn approach with k = 3 to assign appropriate class to the test point	6

<table>
<tr><td>Q. 5.
(B)</td><td>5</td><td>4</td><td>Consider 40 patterns that are split in 3 classes with 8, 20 and 12 samples respectively. Calculate the Entropy impurity, Gini impurity and Misclassification impurity at the node.</td><td>6</td></tr>
<tr><td></td><td></td><td></td><td>OR
*alternate option for Q.5 (A) AND Q. 5 (B) together as Q. 5 (C)</td><td></td></tr>
<tr><td>Q. 5
(C)</td><td>5</td><td>4</td><td>The table below gives the following training data –
<table>
<tr><th>Sr. No.</th><th>Cook</th><th>Mood</th><th>Cuisine</th><th>Time</th><th>Tasty</th></tr>
<tr><td>1</td><td>Sita</td><td>Bad</td><td>Indian</td><td>Lunch</td><td>Yes</td></tr>
<tr><td>2</td><td>Sita</td><td>Good</td><td>Mexican</td><td>Dinner</td><td>Yes</td></tr>
<tr><td>3</td><td>Asha</td><td>Bad</td><td>Indian</td><td>Lunch</td><td>No</td></tr>
<tr><td>4</td><td>Asha</td><td>Bad</td><td>Mexican</td><td>Dinner</td><td>No</td></tr>
<tr><td>5</td><td>Asha</td><td>Good</td><td>Thai</td><td>Lunch</td><td>Yes</td></tr>
<tr><td>6</td><td>Sita</td><td>Good</td><td>Thai</td><td>Dinner</td><td>Yes</td></tr>
<tr><td>7</td><td>Sham</td><td>Good</td><td>Indian</td><td>Lunch</td><td>Yes</td></tr>
<tr><td>8</td><td>Sham</td><td>Bad</td><td>Mexican</td><td>Dinner</td><td>No</td></tr>
<tr><td>9</td><td>Asha</td><td>Bad</td><td>Thai</td><td>Lunch</td><td>No</td></tr>
<tr><td>10</td><td>Asha</td><td>Good</td><td>Thai</td><td>Dinner</td><td>Yes</td></tr>
<tr><td>11</td><td>Sita</td><td>Good</td><td>Indian</td><td>Dinner</td><td>Yes</td></tr>
<tr><td>12</td><td>Sham</td><td>Bad</td><td>Thai</td><td>Lunch</td><td>No</td></tr>
<tr><td>13</td><td>Sham</td><td>Good</td><td>Thai</td><td>Dinner</td><td>Yes</td></tr>
<tr><td>14</td><td>Sita</td><td>Bad</td><td>Mexican</td><td>Lunch</td><td>Yes</td></tr>
<tr><td>15</td><td>Sham</td><td>Good</td><td>Indian</td><td>Lunch</td><td>No</td></tr>
</table>
Using Naïve Bays Classification, estimate the class for ‘Tasty’ for the given feature vector X = { Cook = Sham, Mood = Bad, Cuisine = Indian, Time = Lunch }</td><td>12</td></tr>
<tr><td>Q. 6.
(A)</td><td>6</td><td>2</td><td>Explain the ‘k-fold cross validation‘ approach used in planning the training and testing data</td><td>4</td></tr>
<tr><td>(B)</td><td>6</td><td>3</td><td>A Confusion matrix for a classification exercise returns the following values – TP = 0.942, TN = 0.912, FP = 0.14, FN = 0.06.
Calculate Accuracy, precision, recall and f-score</td><td>4</td></tr>
</table>

Title : Question Paper

FF No. 868

Reg.No.						

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: Monday, 19/12/22	**Time:** 8.30 am -10.30 am

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.No.	CO No	BT No		Max marks
Q.1. (A)	1	1	What is Code book or meta data? Explain with an example.	4
(B)	1	1	Explain with examples the terms – raw data and processed data	4
Q. 2. (A)	2	2	Distinguish between point estimate and confidence interval	3
(B)	2	2	What is the importance of significance level? How it regulates the possibility of occurrence of type 1 and type 2 errors?	6
(C)	2	2	How are the Margin of Error and Standard error related with each other?	3
Q. 3. (A)	3	1	State the formula for Lp norm. Show with the help of an example, L1 metric distance is always larger than L2 metric distance	6
(B)	3	2	Draw a typical 'n x n' hessian matrix. How is it used in optimization?	4
Q. 4. (A)	4	4	In a certain regression activity, following scores are obtained. SSR = 18.12, SSE = 2.21. What is the value of R2? Is this regression a good fit?	6
(B)	4	1	What are dichotomous variables in the context of Logistic regression? Give some examples	4
Q. 5. (A)	5	4	Cluster the following eight points (with (x, y) representing locations) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9). Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). **OR**	6
Q. 5. (A)	5	4	Apply K-nn and predict the class for the test point (3,7) for k=3. Training points with class are (x,y,class) (7,7,2), (7,4,2), (3,4,1), (1,4,1),(2,5,2),(3,8,1)	
Q.5. (B)	5	3	How do you define Genie impurity and entropy impurity? What will their values be, for the purest node?	4
Q. 6. (A)	6	2	How would you execute the k-fold cross-validation strategy? Why is Leave-one -out method its specialization?	4
(B)	6	3	A Confusion matrix for a classification exercise returns the following values – TP = 0.962, TN = 0.93, FP = 0.12, FN = 0.07. Calculate Accuracy, precision, recall, sensitivity, specificity and f-score	6