Sl. No.:O23/CSE/BTech/139/Set 1

Roll Number												

JIS University

End Semester Examinations - Odd 2023

YCS5003 - Introduction to Data Science

Time: 2 Hrs **Maximum Marks: 50**

Instructions to the candidate:

Figures to the right indicate full marks.
Draw neat sketches and diagram wherever is necessary.
Candidates are required to give their answers in their own words as far as practicable

Part A
Answer any Ten (10x1=10 Marks)

1. The following data is used to apply a linear regression algorithm with the least squares regression line Y=a1X. Then, the approximate value of a1 is given by:(X-Independent variable, Y-Dependent variable) (1) CO1 BL6

a) 27.876
b) 32.650
c) 40.541
d) 28.956

2. Which of the following statements is FALSE about Ridge and Lasso Regression? (1) CO2 BL1

a) These are types of regularization methods to solve the overfitting problem.
b) Lasso Regression is a type of regularization method.
c) Ridge regression shrinks the coefficient to a lower value.
d) Ridge regression lowers some coefficients to a zero value.

3. How do you choose the right node while constructing a decision tree? (1) CO3 BL5

a) An attribute having high entropy
b) An attribute having high entropy and information gain
c) An attribute having the lowest information gain.
d) An attribute having the highest information gain.

4. Which one of the following statements is TRUE for a Decision Tree? (1) CO4 BL1

a) Decision tree is only suitable for the classification problem statement.
b) In a decision tree, the entropy of a node decreases as we go down the decision tree.
c) In a decision tree, entropy determines purity.
d) Decision tree can only be used for only numeric valued and continuous attributes.

5. The robotic arm will be able to paint every corner of the automotive parts while minimizing the quantity of paint wasted in the process. Which learning technique is used in this problem? (1) CO5 BL1

a) Supervised Learning.
b) Unsupervised Learning.
c) Reinforcement Learning.

d) Both (A) and (B).

6. Which of the following statement is TRUE? (1) CO1 BL1

a) Outliers should be identified and removed always from a dataset.
b) Outliers can never be present in the test set.
c) Outliers is a data point that is significantly close to other data points.
d) The nature of our business problem determines how outliers are used.

7. Which of the following statement is False in the case of the KNN Algorithm? (1) CO1 BL1

a) For a very large value of K, points from other classes may be included in the neighborhood
b) For the very small value of K, the algorithm is very sensitive to noise.
c) KNN is used only for classification problem statements.
d) KNN is a lazy learner.

8. What kind of distance metric(s) are suitable for categorical variables to find the closest neighbors? (1) CO1 BL2

a) Euclidean distance.
b) Manhattan distance.
c) Minkowski distance.
d) Hamming distance.

9. In the Naive Bayes algorithm, suppose that the prior for class w1 is greater than class w2, would the decision boundary shift towards the region R1(region for deciding w1) or towards region R2 (region for deciding w2)? (1) CO1 BL3

a) towards region R1
b) towards region R2.
c) No shift in decision boundary.
d) It depends on the exact value of priors.

10. Which of the following is FALSE about Correlation and Covariance? (1) CO1 BL1

a) A zero correlation does not necessarily imply independence between variables.
b) Correlation and covariance values are the same.
c) The covariance and correlation are always the same sign.
d) Correlation is the standardized version of Covariance.

11. In Regression modeling, we develop a mathematical equation that describes how, (Predictor-Independent variable, Response-Dependent variable) (1) CO1 BL1

a) one predictor and one or more response variables are related.
b) several predictors and several response variables response are related.
c) one response and one or more predictors are related.
d) All of these are correct.

12. True or False: In a naive Bayes algorithm, the entire posterior probability will be zero when an attribute value in the testing record has no example in the training set. (1) CO1 BL4

a) True
b) False

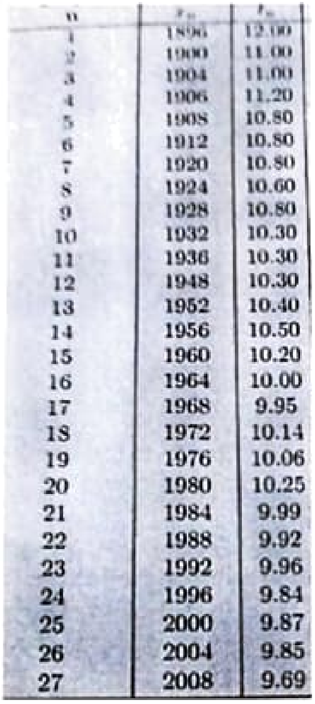
c) Can't be determined
d) None of these

Part B
Answer any Two (2x5=10 Marks)

13. What is linear regression? (3) What is logistic regression? Give an example. (2)	(5)	CO1	BL1
14. What do you understand by the term normal distribution? (3) What is central limit theorem and why is it important? (2)	(5)	CO1	BL1
15. What are the differences between over-fitting and under-fitting? Explain. (5)	(5)	CO1	BL1
16. Explain Bayes theorem. (3) Explan Data cleaning and Data munging. (2)	(5)	CO1	BL1

Part C
Answer any Three (3x10=30 Marks)

17. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. what is the probability that you see at least one shooting star in the period of an hour? (5) A jar has 1000 coins, of which 999 are fair and 1 is tackle headed. Pick a coin at random, and toss it 10 times, given that you see 10 heads, what is the probability that the next toss of that coin is also a head? (5)	(10)	CO1	BL1
18. What is P - value? What is correlation and covariance in statistics? (5 + 5)	(10)	CO1	BL1
19. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls? (5) How can outliers can be treated? (3) What are the drawbacks of the linear model? (2)	(10)	CO1	BL1
20. What is bias-variance trade-off? (5) Explain confusion matrix. (5)	(10)	CO1	BL1
21. The data of the Olympic 100 m dataset is summarised in the table. Applying Least squares fit to the Olympic show prediction for both the 2012 and 2016 year. (You need to draw graph, and derive equation)	(10)	CO5	BL1



n	x_n	t_n
1	1896	12.00
2	1900	11.00
3	1904	11.00
4	1906	11.20
5	1908	10.80
6	1912	10.80
7	1920	10.80
8	1924	10.60
9	1928	10.80
10	1932	10.30
11	1936	10.30
12	1948	10.30
13	1952	10.40
14	1956	10.50
15	1960	10.20
16	1964	10.00
17	1968	9.95
18	1972	10.14
19	1976	10.06
20	1980	10.25
21	1984	9.99
22	1988	9.92
23	1992	9.96
24	1996	9.84
25	2000	9.87
26	2004	9.85
27	2008	9.69