Register No.: ……………………………… Name: ………………………………………………………….

SAINTGITS COLLEGE OF ENGINEERING (AUTONOMOUS)

(AFFILIATED TO APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY, THIRUVANANTHAPURAM)

THIRD SEMESTER M.C.A DEGREE EXAMINATION (Regular), FEBRUARY 2022

(2020 SCHEME)

Course Code: **20MCAT201**

Course Name: **Data Science & Machine Learning**

Max. Marks: **60** **Duration: 3 Hours**

PART A

(Answer all questions. Each question carries 3 marks)

1. With a neat diagram explain CRISP - data mining framework.
2. Define data science.
3. List any three applications of machine learning.
4. Compare supervised learning and unsupervised learning.
5. Explain the OLS method in regression.
6. How does the Laplace estimator work when the posterior probabilities are zero?
7. What are activation functions?
8. Define precision, recall and F-measure.
9. What are support vectors in SVM?
10. Explain bootstrap sampling.

PART B

(Answer one full question from each module, each question carries 6 marks)

MODULE I

11. Data Science problems can be classified into different tasks. Explain with a neat diagram. (6)

OR

12. Explain in detail univariate visualization and multivariate visualization. (6)

MODULE II

13. From the given attributes determine whether the new tissue paper is good or not. Attributes of the new tissue are x1=3 and x2=7. Assume k=3. (6)

Acid durability(x1)	Strength(x2)	classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

OR

14. Estimate conditional probabilities of each attributes {Color, Legs, Height, Smelly} for the species classes: {M, H} using the data given in the table. Using these probabilities estimate the probability values for the new instance-(Color=Green, Legs=2, Height=Short and Smelly=No (Use Naïve Bayes Classifier). (6)

No	Color	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

MODULE III

15. Obtain a linear regression for the data given in the table below assuming that y is the independent variable. (6)

Study Time(X)	Grade(Y)
1	2
2	4
3	5
4	4
5	5

OR

16. a) What is divide-and-conquer strategy? How well it is utilized in the construction of a decision tree? (4)

b) Discuss the strengths and weaknesses of C5.0 decision tree algorithm. (2)

MODULE IV

17. Define an artificial neuron. What are the characteristics of an artificial neural network? (6)

OR

18. a) What is Maximum Margin Hyperplane (MMH)? Explain the characteristics of MMH. (3)

b) Define kernel function. Explain the kernel trick to construct a classifier for a dataset that is not linearly separable. (3)

MODULE V

19. A database contains 80 records on a particular topic of which 55 are relevant to a certain investigation. A search was conducted on that topic and 50 records were retrieved. Of the 50 records retrieved, 40 were relevant. Construct the confusion (6)

matrix for the search and calculate the precision and recall scores for the search.

OR

20. a) Describe the k-fold cross validation scheme for estimating the model performance. (3)

b) Explain k-means clustering algorithm. (3)

****