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

SAINTGITS COLLEGE OF ENGINEERING (AUTONOMOUS)

(AFFILIATED TO APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY, THIRUVANANTHAPURAM)

THIRD SEMESTER MCA DEGREE EXAMINATION (S), FEBRUARY 2024

(2021 SCHEME)

Course Code: **21CA301**

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

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

PART A

(Answer all questions. Each question carries 3 marks)

1. List any three machine learning applications.
2. Explain the importance of data preparation in machine learning.
3. Explain how to apply K-NN classifier in data science problem.
4. State Naive Bayes algorithm to build classification problems.
5. Decision trees have the disadvantage of overfitting at the time of training. How can we solve the issue of this overfitting?
6. Explain the OLS method in regression.
7. Explain the idea behind backpropagation algorithm.
8. Describe any three activation functions.
9. Explain the methods to perform cross validation.
10. Differentiate bagging, boosting and voting.

PART B

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

MODULE I

11. Explain the data science classification and the data science tasks. (6)

OR

12. Explain how data is being visualized using scatterplot. (6)

MODULE II

13. a) Consider the following datasets with attributes height, weight and class labels underweight and normal. Suppose that a new observation is added with height 170 and weight 57. Find class label of the new observation using K-NN algorithm. (4)

Height (in cm)	Weight (in cm)	Class Label
167	51	Underweight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Underweight
169	58	Normal
173	57	Normal
170	55	Normal
170	57	?

b) Describe the key concepts of nearest neighbour classifiers, and explain why they are considered as "lazy learners". (2)

OR

14. The following data set contains factors that determine whether tennis is played or not. Using Naive Bayes classifier, find the play prediction for the day if <Sunny, Cool, High and Strong> as input. (6)

DAY	OUTLOOK	TEMP	HUMIDITY	WIND	PLAY
DAY 1	Sunny	Hot	High	Weak	NO
DAY 2	Sunny	Hot	High	Strong	NO
DAY 3	Overcast	Hot	High	Weak	YES
DAY 4	Rain	Mild	High	Weak	YES
DAY 5	Rain	Cool	Normal	Weak	YES
DAY 6	Rain	Cool	Normal	Strong	NO
DAY 7	Overcast	Cool	Normal	Strong	YES
DAY 8	Sunny	Mild	High	Weak	NO
DAY 9	Sunny	Cool	Normal	Weak	YES
DAY 10	Rain	Mild	Normal	Weak	YES
DAY 11	Sunny	Mild	Normal	Strong	YES
DAY 12	Overcast	Mild	High	Strong	YES
DAY 13	Overcast	Hot	Normal	Weak	YES
DAY 14	Rain	Mild	High	Strong	NO

MODULE III

15. What are the benefits of pruning in decision tree induction? Explain any one approach of tree pruning. (6)

OR

16. Give the steps in simple linear regression for prediction using the straight-line equation $y = \alpha + \beta x$ .The explanation should include the computation of best estimates of $\alpha \text{ and } \beta$. (6)

MODULE IV

17. Explain how artificial neural networks mimic human brain to model arbitrary functions. How can it be applied to real-world problems? (6)

OR

18. How can the support vector machine be used for the classification of linearly separable data? Explain. (6)

MODULE V

19. Suppose 10000 patients get tested for flu; out of them, 9000 are actually healthy and 1000 are sick. For the sick people, a test was positive for 620 and negative for 380. For the healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the precision and recall for the data. (6)

OR

20. Describe the random forest algorithm to improve classifier accuracy with an example. (6)

****