[This question paper contains 8 printed pages.]

Your Roll No.....

Sr. No. of Question Paper : 7802 **K**

Unique Paper Code : 6202453501

Name of the Paper : Machine Learning

Name of the Course : **B.Voc. Software Development**

Semester : V

Duration : 3 Hours Maximum Marks : 90

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. The paper has **two** sections. **Section A** is compulsory. Each question is of **5** marks.
3. Attempt any **four** questions from **Section B**. Each question is of **15** marks.

P.T.O.

Section A

1. (a) Given the following real-world applications, classify each into Supervised, Unsupervised, or Reinforcement Learning. Justify each in one line. (5)

(i) A bank detects fraudulent transactions using past labeled data.

(ii) A chatbot improves its responses by receiving a reward when the user gives positive feedback.

(iii) A movie streaming app groups viewers based on their watch preferences.

(iv) A handwriting recognition model trained with labeled examples of digits.

(v) A delivery robot learns the best path to deliver packages by trial and error.

(b) A spam email classifier was tested on a dataset of 6000 emails. Out of these, 2000 were spam and 4000 were legitimate. The classifier correctly identified 1700 spam emails, but incorrectly marked 300 legitimate emails as spam.

Construct the confusion matrix and compute :

(i) Precision

(ii) Sensitivity

(iii) Specificity

(iv) F1-Score (5)

(c) With the help of an error vs. model complexity diagram. Explain overfitting and underfitting. How do regularization methods improve model generalization and prevent overfitting? (5)

(d) What is the importance of feature engineering in ML? Explain any two feature subset selection techniques to deal with high dimensionality problem. (5)

(e) Apply the Naive Bayes Classifier to predict whether a car will be stolen or not stolen for the following input features : (5)

Color: RED, Type: SUV, Origin: Domestic

Use the dataset given below :

Color	Type	Origin	Stolen
RED	SPORTS	DOMESTIC	YES
RED	SPORTS	DOMESTIC	NO
RED	SPORTS	DOMESTIC	YES
YELLOW	SPORTS	DOMESTIC	NO
YELLOW	SPORTS	IMPORTED	YES
YELLOW	SUV	IMPORTED	NO
YELLOW	SUV	IMPORTED	YES
YELLOW	SUV	DOMESTIC	NO
RED	SUV	IMPORTED	NO
RED	SPORTS	IMPORTED	YES

(f) What are the two popular types of clustering approaches? Describe Hierarchical clustering? (5)

Section B

2. (a) Explain why standardization is needed before PCA. How does the PCA algorithm help reduce dimensionality in machine learning? Write the steps of the PCA algorithm. (7)

(b) Given the following dataset of weather conditions and whether a person plays tennis, calculate the Information Gain for the attribute "Outlook".

Outlook	Temperature	PlayTennis
Sunny	Hot	No
Sunny	Mild	No
Overcast	Hot	Yes
Rainy	Cool	Yes
Rainy	Mild	Yes
Rainy	Cool	No

(8)

3. (a) What type of problems can logistic regression solve? How is it different from linear regression? What is the role of the sigmoid function in logistic regression? (7)

(b)

Marketing Cost (₹1000) (x)	Customers Gained (y)
2	20
4	40
6	50
8	70
10	85

Use the least squares method to compute the best-fitting regression line. Estimate the number of customers if Rs. 7000 thousand is spent. (8)

4. (a) (i) What is the role of activation functions in a neural network?

(ii) Explain why non-linear activation functions are essential in multi-layer networks.

(iii) Give examples of commonly used activation functions. (7)

(b) How does hold out validation differ from k-fold cross validation? For k-10 and datapoints- D1, D2, D3, D4, D5, D6, D7, D8, D9, and Dl0 in the dataset, mention one possible dataset distribution between training and test partition for k-fold cross validation. (8)

5. (a) (i) What's the difference between hard margin and soft margin SVM?

(ii) Explain how SVM can handle non-linearly separable data using the kernel trick.

(iii) Mention two common kernel functions and their purposes. (7)

(b) List two commonly used distance measures in k-NN algorithm. Why is feature scaling important in the k-NN algorithm? Also, list two main limitations of the k-NN classifier. (8)

6. (a) A data analyst fits a linear regression model to predict a student's course grade using class participation and assignment scores as predictors. After evaluation, the following results are obtained:

Mean Squared Error (MSE): 49.0

Coefficient of Determination (R^2): 0.76

Define both MSE and R^2 with their mathematical formulas.

Explain why R^2 is often preferred over MSE when comparing two regression models trained on the same dataset. (7)

(b) The following 1D dataset contains data points:

$$X = \{2,3,4,15,16,17\}$$

Perform one iteration of K-Means clustering ($k = 2$) using initial centroids $C_1 = 3$, $C_2 = 16$. (8)

7. (a) (i) What are the two types of supervised learning? Describe in detail their two applications each.

(ii) Name any two ML techniques can be used for a supermarket chain problem where customers want to find products that best match their needs quickly? (7)

(b) (i) Define True Positive Rate (TPR) and False Positive Rate (FPR). How are they used to construct the ROC curve?

(ii) Discuss the significance of the ROC curve and how the Area Under the Curve (AUC) helps in comparing classifier performance. (8)