Use the Naive Bayes classification rule to classify a Red Domestic SUV. Show all the steps involved in the process which include calculating prior probabilities, conditional probabilities, and posterior probabilities and comparing the posterior probabilities.

(b) Differentiate between bagging and boosting ensemble techniques with suitable examples. (5)

[This question paper contains 12 printed pages.]

Your Roll No.....

Sr. No. of Question Paper : 1199 **I**

Unique Paper Code : 2343010013

Name of the Paper : Data Mining for Knowledge Discovery (DSE-5)

Name of the Course : **B.Sc. (Prog.) Computer Science**

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.
3. All questions in **'Section A'** are compulsory.
4. Attempt any **four** questions from **'Section B'**.
5. Parts of a question must be answered together.
6. Non-programmable, basic calculators are allowed.

SECTION A

1. (i) Indicate whether the following activities can be identified as a Data Mining Task or a Database Query : (2)

   (a) Calculating the revenue for each product category from sales data.

   (b) Predicting the future stock price of a company using historical data.

   (c) Generating a list of all students who have 12^{th} standard percentage higher than 80.

   (d) Clustering customers into different groups based on their buying patterns to target marketing campaigns more effectively.

   (ii) Determine the attribute type of the following : (2)

Use the k-Nearest Neighbor (k-NN) classifier with K = 5 and proximity measure as Euclidean distance to classify a test record (A = 5, B = 5). Which class would the k-NN classifier assign to the test record?

7. (a) Consider the following training data set for car theft : (10)

Sample No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Compute the Gini index for the overall collection of training data. Compute the Gini index for attribute A and B. Which attribute would the decision tree algorithm choose for the split at the root node? Give reason.

(b) Consider the following dataset, which consists of 10 records, with two features, A and B, and a binary target class Z. (7)

Record No.	A	B	Z
R1	5	10	+
R2	4	6	+
R3	4	2	+
R4	3	8	+
R5	3	9	-
R6	2	4	-
R7	2	7	-
R8	1	2	-
R9	1	7	+
R10	1	5	+

(a) Calender Year such as 2000, 2010, 2020, 2030.

(b) Payment methods such as "cash", "credit card", "debit Card".

(c) Speed such as 5 km/h, 10 km/h, 50 km/h, 100 km/h

(d) Severity of Disease such as "mild", "moderate", "severe".

(iii) What is the difference between predictive tasks and descriptive tasks? (2)

(iv) What is the difference between noise and outlier? List any two ways to handle the missing values. (3)

(v) What is the difference between model overfitting and model underfitting? (3)

(vi) How is an eager learner classifier different from a lazy learner classifier? Support your answer with an example from both category of classifiers. (3)

(vii) Consider the following data for the feature "Income": (3)

[20, 40, 10, 30, 60, 100, 200, 300, 400], Discretize the data into three bins using the equal width and equal depth approach.

(viii) Suppose an attribute marital status has three distinct values {single, married, divorced}. Illustrate the three different ways to perform attribute test condition for the binary split. Also perform for the multiway split. (3)

(ix) Consider a set of 10 data points in a 2D space given as follows P1 = (1, 2), P2 = (2, 2), P3 = (2, 3), P4 = (5, 6), P5 = (5, 7), P6 = (6, 6), P7 = (7, 7), P8 = (10, 10), P9 = (10, 11), P10 = (20, 20). Determine the core points, with $\varepsilon = 2$ and MinPts = 3 using DBSCAN clustering technique. (4)

(b) Draw the logical view of the ensemble learning method. Discuss how ensemble of classification can be constructed by manipulating the input features. (5)

6. (a) Consider the following dataset for binary classification problem (8)

Instance	A	B	C	Target Class
1	T	T	5	+
2	T	T	7	+
3	F	T	8	-
4	F	F	3	+
5	T	F	7	-
6	T	F	4	-
7	F	F	5	-
8	F	T	6	+
9	T	F	1	-

(b) Describe the situation under which DBSCAN is the preferred method for clustering. Discuss its strengths and weaknesses. (5)

5. (a) Use the distance matrix given below to perform the hierarchical clustering using single link (MIN) and complete link (MAX). Draw the dendrograms which should clearly show the order in which points are merged. (10)

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

(x) The confusion matrix of a 2-class problem is given below : (5)

		Predicted Values	
		Class = 1	Class = 0
Actual Values	Class = 1	100	25
	Class = 0	50	75

Calculate the accuracy, sensitivity, specificity, true negative rate, and false positive rate.

SECTION B

2. (a) Compute the Minkowski distance from the given table, for r = 1 (Manhattan distance), and r = ∞ (Supremum distance). (8)

point	X coordinate	Y coordinate
p1	2	4
p2	4	2
p3	3	3
p4	7	3
p5	1	4

(b) Draw the Knowledge Discovery in Databases (KDD) diagram and describe its phases. Discuss briefly four core data mining tasks. (7)

3. (a) For evaluating the performance of a classifier, how does holdout method differ from cross validation?

Consider the data points- D1, D2, D3, D4, D5, D6, D7, D8, D9, and D10 in the dataset. (9)

(i) For a train test split (70, 30) mention a holdout dataset distribution.

(ii) For k = 5 mention one possible dataset distribution between training and test partition for k-fold cross-validation.

(b) Discuss the following data preprocessing strategies and techniques with suitable examples. (6)

(i) Aggregation

(ii) Sampling

(iii) Dimensionality Reduction

4. (a) Consider the eight data points P1(3, 11), P2(3, 6), P3(9, 5), P4(6, 9), P5(8, 6), P6(7, 5), P7(2, 3), P8(5, 10). Assume initial centroids (seed points) are P1, P4, and P7. Use k-means algorithm and Euclidean Distance to make these clusters. What will be the three clusters and new centroids after the second iteration? (10)