final clustering scheme using a dendrogram. The dendrogram should clearly show the order in which the data points are merged.

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

(b) What is an out-of-bag error? How is the error rate of a base classifier in AdaBoost computed? (5)

[This question paper contains 12 printed pages.]

Your Roll No..............

Sr. No. of Question Paper : 1195 **I**

Unique Paper Code : 2343010009

Name of the Paper : Data Mining – II

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

Semester : IV

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. **Section A** (Question No. **1**) is compulsory.
3. Attempt any **four** questions from **Section B** (Questions **2** to **7**).
4. The use of a simple calculator is allowed.
5. Parts of the question must be answered together.

Section A

1. (a) State two applications where 'anomalous' data points are of considerable interest. (2)

(b) Using appropriate libraries, write Python statements to generate daily timestamps from '2024-04-01' to '2024-06-01'. (2)

(c) Differentiate between Parametric and Non-Parametric methods for anomaly detection. Give one suitable example of each. (2)

(d) State the time complexity of the DBSCAN algorithm. Is the space complexity of the DBSCAN algorithm linear, even for data with high dimensions? State the reason for your answer. (3)

(b) Consider the following Python Code : (4)

```
using pandas as pd

using numpy as np

rng = pd.date_range('2024-03-20', periods=12,
freq='D')

ts = pd.Series(np.arange(12), index=rng)
```

(i) Show the content of time series ts.

(ii) Find the output of ts.resample('M') .min().

(c) What are the limitations of STREAM clustering that can be overcome by the CluStream algorithm? Discuss how these limitations are addressed by the CluStream algorithm. (8)

7. (a) Consider thefollowing distance matrix for four data points A, B, C, and D to compute clusters using complete link agglomerative clustering. Show the

P.T.O.

D1: Machine Learning can be very interesting

D2: Machine Learning has several applications

(b) Consider the following text in Document D3.

D3: "Natural Language Processing (NLP) is amazing! language research in NLP has helped in enabling the era of generative AI."

Perform the given text mining preprocessing tasks on the document D3 after applying a mandatory step. Also, show the final pre-processed results.

(i) Stop word removal

(ii) Removal of punctuation marks (5)

6. (a) What is the significance of the symbolic aggregate approximation (SAX) approach? What are the two steps involved in it? (3)

(e) State any one weakness of K-means clustering algorithm. Also, write a proposed method to handle it. (3)

(f) State any two characteristics that are unique to "text data" due to which modifications are required before the application of traditional multidimensional data mining techniques to such data. (3)

(g) Document D has 100 words where the term fc appears 2 0 times. The corpus contains 10,000 documents, out of which 100 documents contain the term t. Calculate tf-idf for t. (3)

(h) The random forests ensemble method is robust to overfitting problems. Comment. (3)

(i) Why is boosting considered a "sequential" ensemble model? How do the weights of "samples" change in each boosting iteration? (4)

(j) Given that a stream is running starting at a clock time of 1 with α = 2 and 1 = 2, with current clock time of 64. Compute

(i) Maximum order of any snapshot stored at the 64 th time unit.

(ii) Maximum number of snapshots maintained at the 64th time unit since the beginning of the stream. (4)

Section B

2. (a) Given the following data points in one dimensional plane :

12, 3, 20, 30, 11, 25. Show the clusters obtained using the K- means algorithm after two iterations. Assume K = 3, and the initial centres as (2), (4), (6). Show the updated centroids for the second iteration. (6)

(b) Which distance measure is appropriate to determine anomalies in multivariate data? Write the formula for computing this metric. Given a dataset following Gaussian distribution with centre (0,0) and inverse of a co-variance matrix as

$S^{-1} = \begin{bmatrix} 2 & 4 \\ 4 & 1 \end{bmatrix}$. Find the distance of a point (2,2) from the given distribution. (5)

(c) Compare and contrast the strengths and weaknesses of statistical approaches and proximity-based approaches for anomaly detection. (4)

5. (a) Consider the following two documents D1 and D2, create their vector representation using Bag-of-words model. Ignore any pre-processing. Compute the Cosine similarity and Jaccard coefficient between these two documents.

P.T.O.

top three anomalous points using knn method where k is the number of nearest neighbours and can take any value from (1, 5, 8, 10). Comment on the effect of different values of k on detecting anomalous points from the given data.

Point No.	X	Y	1st NN	5th NN	8th NN	10th NN
1	1.02	5.04	0.03	0.09	0.13	0.15
2	4.00	2.50	1.41	1.92	1.97	2.01
3	4.02	5.39	0.05	0.27	0.40	0.42
4	3.31	3.73	0.49	0.98	1.06	1.13
5	1.11	5.09	0.06	0.11	0.15	0.16
6	0.94	4.48	0.36	0.50	0.55	0.56
7	4.07	5.38	0.05	0.27	0.35	0.38
8	4.17	5.85	0.18	0.48	0.60	0.72
9	4.63	3.92	0.52	0.75	0.80	0.90

(b) Consider five data points P1, P2, P3, P4, and P5. They belong to two clusters C1 = {P1, P2, P4}and C2 = {P3, P5} and two classes L1 = {P1, P2} and L2 = {P3, P4, P5}. Compute the ideal cluster matrix and class similarity matrix. Also, find the Rand statistic using these matrices. (5)

(c) Given two sets of 100 points that fall within the unit square. One set of points is arranged so that the points are randomly spaced. In the other set, points are generated from a uniform distribution. What will be the behaviour of DBSCAN on the:

(i) uniform data set

(ii) random data set (4)

3. (a) What is the significance of the Bagging algorithm in ensemble learning? How does Bagging combine multiple weak learners to create a strong classifier?

P.T.O.

Consider the following dataset with six data points (P1 – P6), create two bootstrap samples consisting of points sampled from the given data. (6)

P1	P2	P3	P4	P5	P6

(b) For the given the confusion matrix :

	Predicted		
Actual		+	–
	+	510	110
	–	210	210

Compute the values of the following measures by stating proper formulas :

(i) Precision

(ii) F1 measure

(iii) G measure

(c) Rohan recently came across the concept of ensemble methods in a data mining class and decided to apply it to stock market prediction. To predict whether the stock market would rise or fall on a given day, he decided to flip a coin 1000 times and predicted the stock market would go up if heads turned up in the majority and vice-versa. He thinks that this approach could get him a better prediction of the stock market because an ensemble of independent classifiers could potentially obtain a better prediction. Do you agree with him? Give a brief justification. (3)

4. (a) Consider the following table with nine points in a two-dimensional space (X, Y). The table also shows the distances of the points to their 1st, 5th, 8th and 10th nearest neighbours (NN). Find the