DSBDA UNIT – 5 PYQ 'S

➢ *MAY / JUN 2022*

Q5)

a) Write short note on :

i) Time Series Analysis

Time series analysis is a statistical technique that deals with time-ordered data points. It is used to analyze trends, patterns, and seasonal variations over a period of time to make future predictions or detect anomalies.

Key Components:

1. **Trend:** Long-term movement in the data (increasing, decreasing, or constant).
2. **Seasonality:** Repeating short-term cycles at regular intervals (e.g., monthly sales spikes).
3. **Cyclic Patterns:** Long-term oscillations not of fixed period (like business cycles).
4. **Irregularity (Noise):** Random variations due to unforeseen events.

Applications:

- Stock market predictions
- Sales and revenue forecasting
- Climate and temperature monitoring

Example Table:

Date & Time	Temperature (°C)
2025-05-01 09:00	30
2025-05-01 10:00	31

ii) TF-IDF (Term Frequency – Inverse Document Frequency)

TF-IDF is a **text mining** technique used to measure the importance of a word in a document compared to a collection (corpus). It's widely used in **information retrieval, NLP, and search engines**.

1. Term Frequency (TF):

Indicates how frequently a term appears in a document.

Formula:

$$TF = \frac{\text{Number of times term t appears in document d}}{\text{Total terms in document d}}$$

2. Inverse Document Frequency (IDF):

Penalizes common terms that appear in many documents.
Formula:

$$IDF = \log\left(\frac{\text{Total number of documents}}{1 + \text{Number of documents containing term t}}\right)$$

3. TF-IDF:

$$TFIDF = TF \times IDF$$

Example:

If the term "*data*" appears 3 times in a document of 100 words, and is present in 10 out of 1000 documents:

$$TF = \frac{3}{100} = 0.03, \quad IDF = \log\left(\frac{1000}{1+10}\right) \approx 1.99, \quad TFIDF \approx 0.03 \times 1.99 = 0.0597$$

Use Cases:

- Ranking pages in search engines (Google)
- Text classification and sentiment analysis

Q5)

b) What is clustering? With suitable example explain the steps involved in K-Means algorithm.[9]

1. Clustering

Clustering is an **unsupervised learning technique** used to group a set of data points into **clusters** such that data points in the same group (cluster) are more **similar to each other** than to those in other groups.
It is used when **no predefined labels** exist for the data.

K-Means Algorithm (with Example) :

sr_no	age	amount
C1	20	500
C2	40	1000
C3	30	800
C4	18	300
C5	28	1200
C6	35	1400
C7	45	1800

Input:

- Dataset with attributes (e.g., age, amount)
- Number of clusters: K = 2

Step 1 : Initialize Clusters

- Choose initial centroids randomly or logically.
  From image:
  - **Cluster 1 centroid = C1 → (20, 500)**
  - **Cluster 2 centroid = C2 → (40, 1000)**

Step 2: Assign each point to the nearest centroid

For each point (e.g., C3 = (30, 800)), calculate **Euclidean distance** to both centroids:

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

✅ Example: Classify C3 (30, 800)

Distance from Cluster 1 (C1 = 20, 500):

$$\sqrt{(30-20)^2+(800-500)^2} = \sqrt{100+90000} = \sqrt{90100} \approx 300$$

Distance from Cluster 2 (C2 = 40, 1000):

$$\sqrt{(30-40)^2+(800-1000)^2} = \sqrt{100+40000} = \sqrt{40100} \approx 200$$

🔸 C3 is closer to Cluster 2, so assign it to Cluster 2

Step 3: Repeat assignment for all points

- Use the same method (distance calculation) for C4, C5...C7

Step 4: Recalculate centroids

- After assigning all points, compute new centroid for each cluster:

$$\text{New Centroid} = \left(\frac{\text{Sum of Age}}{\text{Points}}, \frac{\text{Sum of Amount}}{\text{Points}}\right)$$

Example from image:

Cluster 2 → Avg of C2 & C3 →

$$\text{Age} = \frac{40+30}{2} = 35, \quad \text{Amount} = \frac{1000+800}{2} = 900$$

Step 5: Repeat Steps 2–4

- Keep reassigning and updating centroids until:
  - **No change in cluster assignment**, or
  - **Centroids stabilize**

Q6)

a) Write short note on [9]

i)Confusion matrix

ii) AVC - ROC curve

i) Confusion Matrix

A **Confusion Matrix** is a tabular representation used to evaluate the performance of a classification model by comparing the predicted class labels against the actual labels. It contains four key elements

- **True Positive (TP):** Cases where the model correctly predicts the positive class.
- **True Negative (TN):** Cases where the model correctly predicts the negative class.
- **False Positive (FP):** Cases where the model incorrectly predicts positive when it is actually negative (Type I error).
- **False Negative (FN):** Cases where the model incorrectly predicts negative when it is actually positive (Type II error).

The confusion matrix is useful for calculating performance metrics like:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

ii) ROC Curve

The **ROC (Receiver Operating Characteristic) Curve** is a graphical plot that illustrates the diagnostic ability of a binary classifier as the classification threshold changes. It plots:

- **True Positive Rate (TPR)** or **Sensitivity** on the Y-axis, which measures how many actual positives are correctly identified.
- **False Positive Rate (FPR)** (which is 1 - Specificity) on the X-axis, representing the proportion of negatives incorrectly classified as positives.

The curve shows the trade-off between sensitivity and specificity. The **Area Under the Curve (AUC)** quantifies the model's overall ability to discriminate between positive and negative classes. A higher AUC (closer to 1) indicates better classification performance, while an AUC of 0.5 suggests no discriminative power (random guessing).

ROC curves are widely used in fields like healthcare and machine learning model evaluation to select the best threshold for classification.

ADD EXAMPLE TO IT ANY !!

b) Discuss Holdout method and Random Sub Sampling methods.[9]

1. Holdout Method:

The **Holdout method** is a simple technique used to evaluate the performance of machine learning models. It involves dividing the dataset into two or three mutually exclusive sets:

- **Training Set:** Used to train the model.
- **Testing Set:** Used to evaluate model performance.
- (Optionally, a **Validation Set** may be used for parameter tuning.)

Typical Split Ratio:

- 70% training, 30% testing
- or 60% training, 20% validation, 20% testing

Example:

If a dataset contains 1000 samples and a 70:30 split is used, then:

- 700 samples are used for training.
- 300 samples are used for testing.

Advantages:

- Simple and fast.
- Suitable for large datasets.

2. Random Subsampling Method:

Random Subsampling is an extension of the holdout method. It involves performing the holdout process **multiple times** with different random splits of the data.

Process:

- Repeatedly split the dataset into training and testing sets randomly.
- Train and evaluate the model for each split.
- Finally, compute the **average performance** over all iterations.

Example:

- Run the holdout method 5 times with a 70:30 split.
- Suppose the accuracies are: 85%, 83%, 84%, 86%, 82%.
- The **average accuracy = 84%**.

Advantages:

- Reduces variance and bias caused by a single train-test split.
- Flexible Number of Trials

Disadvantages:

- Increased computational cost due to multiple training and testing cycles.
- Some data points may never be selected in training/testing.

➢ NOV / DEC 2022

Q5)
a) Suppose that the given data the taste is to cluster points (With (x.y) representing location) into three cluster, where the points are. A1(2,10), A2(2,5), A3(8,4), B1 (5,8) B2(7,5) B3(6,4), C1(1,2), C2(4,9) The distance function is Euclidean distance suppose initially we assign A1, B1 and C1 as the center of each cluster, respectively. use the k means algorithm to show only the three cluster centers after the first round of execution with steps. [9]

Given:

- Points:
  A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9)
- Initial centroids:
  - Cluster 1 → A1 (2,10)
  - Cluster 2 → B1 (5,8)
  - Cluster 3 → C1 (1,2)
- Distance Metric: **Euclidean Distance**

Step 1: Assign Each Point to Nearest Centroid

We'll compute distance from each point to all 3 centroids.

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

A2 (2,5):

- To A1: √[(2-2)² + (5-10)²] = √(0 + 25) = **5**
- To B1: √[(2-5)² + (5-8)²] = √(9 + 9) = √18 ≈ **4.24**
- To C1: √[(2-1)² + (5-2)²] = √(1 + 9) = √10 ≈ **3.16** → **Cluster 3**

A3 (8,4):

- To A1: √[(8-2)² + (4-10)²] = √(36 + 36) = √72 ≈ **8.49**
- To B1: √[(8-5)² + (4-8)²] = √(9 + 16) = √25 = **5** → **Cluster 2**
- To C1: √[(8-1)² + (4-2)²] = √(49 + 4) = √53 ≈ **7.28**

B2 (7,5):

- To A1: √[(7-2)² + (5-10)²] = √(25 + 25) = √50 ≈ **7.07**
- To B1: √[(7-5)² + (5-8)²] = √(4 + 9) = √13 ≈ **3.61** → **Cluster 2**
- To C1: √[(7-1)² + (5-2)²] = √(36 + 9) = √45 ≈ **6.71**

B3 (6,4):

- To A1: √(16 + 36) = √52 ≈ **7.21**
- To B1: √(1 + 16) = √17 ≈ **4.12** → **Cluster 2**
- To C1: √(25 + 4) = √29 ≈ **5.38**

C2 (4,9):

- To A1: √(4 + 1) = √5 ≈ **2.24**
- To B1: √(1 + 1) = √2 ≈ **1.41** → **Cluster 1**
- To C1: √(9 + 49) = √58 ≈ **7.62**

B1 (5,8) → Itself → Cluster 2

C1 (1,2) → Itself → Cluster 3

✅ Step 2: Cluster Assignment (after 1st round)

Cluster	Points Assigned
C1	A1(2,10)
C2	A3(8,4), B1(5,8), B2(7,5), B3(6,4), C2(4,9)
C3	A2(2,5), C1(1,2)

✅ Step 3: Calculate New Centroids

🔹 Cluster 1 (C1):

Only A1 → Centroid = **(2,10)**

🔹 Cluster 2 (C2):

Points: (8,4), (5,8), (7,5), (6,4), (4,9)

- $\bar{X}$ = (8+5+7+6+4)/5 = **6.0**
- $\bar{Y}$ = (4+8+5+4+9)/5 = **6.0**
  → New Centroid = **(6,6)**

🔹 Cluster 3 (C3):

Points: (2,5), (1,2)

- $\bar{X}$ = (2+1)/2 = **1.5**
- $\bar{Y}$ = (5+2)/2 = **3.5**
  → New Centroid = **(1.5, 3.5)**

✅ Final Answer: New Centroids after 1st Round

Cluster	New Centroid
C1	(2, 10)
C2	(6, 6)
C3	(1.5, 3.5)

Q5 b) Explain the following text analysis steps with suitable example. [8 Marks]

i) Part of Speech (POS) Tagging
ii) Lemmatization
iii) Stemming

i) Part of Speech (POS) Tagging (3 Marks)

- POS tagging assigns grammatical labels (noun, verb, adjective, etc.) to each word in a sentence.
- It helps in understanding the structure and meaning of a sentence.
- **Used in:** parsing, named entity recognition, lemmatization.

Example:
Sentence: *"The cat sat on the mat."*
POS Tags:

- The/DT (determiner),
- cat/NN (noun),
- sat/VBD (verb),
- on/IN (preposition),
- the/DT,
- mat/NN

ii) Lemmatization (2.5 Marks)

- Lemmatization reduces words to their **root/base form (lemma)** by considering the **context and part of speech**.

- It uses a vocabulary and morphological analysis.
- More accurate than stemming.

Example:

- *"am", "are", "is" → "be"*
- *"running", "ran" → "run"*

iii) Stemming (2.5 Marks)

- Stemming is a **rule-based method** that chops off word endings to get the root form.
- It may not always give meaningful words (can be rough).

Example:

- *"playing", "played", "plays" → "play"*
- *"universities" → "univers"*

Feature	Lemmatization	Stemming
Accuracy	High	Lower
Output Word	Dictionary word	Root may not be valid word
Example	"better" → "good"	"playing" → "play"

Q6

a) Calculate Accuracy, Precision, Recall, and Error Rate using the given confusion matrix. [8]

Confusion Matrix :

	Predicted: Risk-Yes	**Predicted: Risk-No**
Actual: Risk-Yes	**80** (TP)	**220** (FN)
Actual: Risk-No	**150** (FP)	**9500** (TN)

Formulas :

- **Accuracy** = (TP + TN) / (TP + FP + TN + FN)
- **Precision** = TP / (TP + FP)
- **Recall (Sensitivity)** = TP / (TP + FN)
- **Error Rate** = (FP + FN) / Total

Step-by-Step Calculations :

- TP = 80
- TN = 9500
- FP = 150
- FN = 220
- Total = 80 + 220 + 150 + 9500 = 9950

✅ **1. Accuracy**

$$\frac{TP + TN}{Total} = \frac{80 + 9500}{9950} = \frac{9580}{9950} \approx \boxed{0.9627 \text{ or } 96.27\%}$$

✅ **2. Precision**

$$\frac{TP}{TP + FP} = \frac{80}{80 + 150} = \frac{80}{230} \approx \boxed{0.3478 \text{ or } 34.78\%}$$

✅ **3. Recall**

$$\frac{TP}{TP + FN} = \frac{80}{80 + 220} = \frac{80}{300} \approx \boxed{0.2667 \text{ or } 26.67\%}$$

✅ **4. Error Rate**

$$\frac{FP + FN}{Total} = \frac{150 + 220}{9950} = \frac{370}{9950} \approx \boxed{0.0372 \text{ or } 3.72\%}$$

Interpretation (Description on Heart Attack Risk)

- The **accuracy** of the model is high (96.27%), indicating most predictions are correct.
- However, **precision (34.78%)** and **recall (26.67%)** for detecting *Heart Attack Risk-Yes* are low.
- This means the model is **poor at identifying people at risk** of heart attacks.
- In real-world applications, **low recall** is dangerous because many actual positive cases (220) are being missed

b) Explain the TF-IDF (Term Frequency – Inverse Document Frequency) in Text Analysis with suitable example

➔ **Already done ! means Repeated !**
just add at starting :
TF-IDF is a popular technique in **text analysis** used to evaluate how important a word is to a

document relative to a corpus. It helps filter out common terms and highlight unique, meaningful words.

➢ **MAY / JUN 2023**

Q5)

a) What is text processing? Explain TF-IDF with example. [8]
→ REPEATED

b) With suitable example ,explain the steps involved in k-means algorithm. [9]

→ REPEATED

6 a) Define the following terms with respect to confusion matrix: [8 Marks]

i) Accuracy
ii) Precision
iii) Recall
iv) AUC–ROC

i) **Accuracy**

Accuracy is the ratio of correctly predicted instances (both positive and negative) to the total instances in the dataset.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

It indicates the overall effectiveness of the model but can be misleading in imbalanced datasets where one class dominates. Therefore, accuracy alone is not sufficient for evaluating models on skewed data.

---

ii) **Precision**

Precision is the ratio of correctly predicted positive instances to the total instances predicted as positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision means that the model has a low false positive rate, which is important in applications like spam detection. It helps to minimize the cost of false alarms.

iii) Recall (Sensitivity or True Positive Rate)

Recall is the ratio of correctly predicted positive instances to all actual positive instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

High recall ensures that most of the actual positive cases are identified, which is critical in disease diagnosis. It helps to reduce the chance of missing important positive cases.

---

iv) AUC–ROC (Area Under the Receiver Operating Characteristic Curve)

The ROC curve plots the **True Positive Rate (Recall)** against the **False Positive Rate** at different threshold levels.

The AUC quantifies the model's ability to distinguish between classes across all thresholds, with higher values indicating better discrimination. It is especially useful for comparing models regardless of classification thresholds.

b) Explain k-fold Cross Validation & Random Subsampling [9]

1. k-fold Cross Validation

Definition:
k-fold Cross Validation is a robust method to evaluate the performance of a machine learning model by dividing the dataset into **k equal-sized folds (subsets)**.

Process:

- The dataset is split into *k* folds (e.g., k = 5 or 10).
- The model is trained on **k-1 folds** and tested on the **remaining 1 fold**.
- This process is repeated *k* times, each time using a different fold as the test set.
- The overall performance is the average of the metrics from all *k* trials.

Example:
If k = 5, the dataset is divided into 5 parts. Each part acts as a test set once, while the remaining 4 parts form the training set.
The model is trained and evaluated 5 times, and the average accuracy or error is calculated.

Advantages:

- Efficient use of data, as all samples are used for training and testing.
- Reduces bias and variance in performance estimates compared to a single train-test split.

2. Random Subsampling (Repeated Holdout Method)

Definition:
Random Subsampling involves randomly splitting the dataset into training and testing sets multiple times to evaluate model performance.

Process:

- The dataset is randomly divided into training and testing sets (e.g., 70% training, 30% testing).
- The model is trained and tested once on this split.
- This process is repeated several times (e.g., 10 times) with different random splits.
- The final performance is the average of all iterations.

Example:
For 10 random splits, the model's accuracy on each split is calculated and averaged to estimate overall performance.

Advantages:

- Provides a more reliable estimate than a single holdout by averaging over multiple splits.
- Easy to implement and understand.

k-fold Cross Validation is a preferred method for balanced and reliable model evaluation, especially for small to medium datasets.

Random Subsampling provides flexibility and a quick performance check but may suffer from variability and uneven data usage.

- **NOV / DEC 2023**

Q5)

a) Suppose that the given data the task is to cluster points (with (x, y) representing location) into three clusters, where the points are A1 (2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A1, B1 and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only show only the first round of execution with cluster center. [8]

b) Explain the following Text Analysis steps with suitable example i) ii) Part-of-speech(POS)tagging Lemmatization [9]

➔ **BOTH a) , b) are repeated !!**

Q6

a) Given the confusion matrix, calculate Accuracy, Precision, Recall, and Error Rate with description on Diabetic Risk. [8]

	Predicted: Yes	**Predicted: No**
Actual: Diabetic Risk – Yes	90 (TP)	210 (FN)
Actual: Diabetic Risk – No	140 (FP)	9560 (TN)

Formulas:

- **Accuracy** = (TP + TN) / (TP + TN + FP + FN)
- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- **Error Rate** = (FP + FN) / Total

Values:

- TP = 90
- FN = 210
- FP = 140
- TN = 9560
- Total = 90 + 210 + 140 + 9560 = 10000

✅ **1. Accuracy**

$$\frac{TP + TN}{Total} = \frac{90 + 9560}{10000} = \frac{9650}{10000} = \boxed{96.5\%}$$

✅ **2. Precision**

$$\frac{TP}{TP + FP} = \frac{90}{90 + 140} = \frac{90}{230} \approx \boxed{39.13\%}$$

✅ **3. Recall**

$$\frac{TP}{TP + FN} = \frac{90}{90 + 210} = \frac{90}{300} = \boxed{30.0\%}$$

✅ **4. Error Rate**

$$\frac{FP + FN}{Total} = \frac{140 + 210}{10000} = \frac{350}{10000} = \boxed{3.5\%}$$

Interpretation (Diabetic Risk Description):

- The model has high accuracy (96.5%), meaning overall performance is good.
- But low recall (30%) shows it misses many true diabetic cases.
- Also, precision (39.13%) indicates that many predicted diabetic cases are actually not diabetic.
- In real-world diabetic screening, higher recall is crucial to avoid missing risky patients.

b) Explain the Text Preprocessing Steps with Suitable Example [9]

Text preprocessing refers to a series of steps used to clean and prepare raw text data before feeding it into a machine learning or NLP model. It removes noise and standardizes the text format for better performance.

1. Tokenization

- Splits text into smaller units (words or sentences).
- Helps in processing each word individually.

Example:
Sentence: "Data Science is powerful."
Word tokens: ['Data', 'Science', 'is', 'powerful']

2. Lowercasing

- Converts all text to lowercase to maintain uniformity.
  Example:
  "Data" → "data", "SCIENCE" → "science"

3. Removing Stop Words

- Removes commonly used words (is, the, an, etc.) that add little meaning.
- These are considered **noise**.
  Example:
  "This is a pen" → ["pen"]

4. Stemming

- Reduces words to their **root form** by removing suffixes.
- May not return valid words.
  Example:
  "playing", "played", "plays" → "play"

5. Lemmatization

- Converts words to their **base dictionary form** using grammar rules.
- More accurate than stemming.
  Example:
  "am", "are", "is" → "be"
  "better" → "good"

6. Removing Punctuation and Special Characters

- Eliminates symbols like .,!?@# which are not needed in most models.
  Example:
  "Hello!!! How are you?" → "Hello How are you"

7. Removing Numbers (Optional)

- Removes digits when numerical values are not relevant.
  Example:
  "I have 2 cats" → "I have cats"

8. Final Output

After applying all steps, the text becomes clean, consistent, and ready for feature extraction (e.g., TF-IDF).

➢ **MAY / JUN 2024**

Q5

a) Perform K-Means clustering (K = 2) using initial centroids (2,3) and (8,6). [9 Marks]

Step 1: Given Dataset

Point	X	Y
A	2	3
B	4	7
C	3	5
D	6	9
E	8	6
F	7	8

Initial Centroids:

- Cluster 1 (C1) → (2, 3)
- Cluster 2 (C2) → (8, 6)

Step 2: Compute Distance of Each Point from Both Centroids (Euclidean)

🔸 Point A (2,3)

- To C1 (2,3):

$$\sqrt{(2-2)^2+(3-3)^2} = \sqrt{0+0} = 0$$

- To C2 (8,6):

$$\sqrt{(2-8)^2+(3-6)^2} = \sqrt{36+9} = \sqrt{45} \approx 6.71$$

✅ Cluster 1

🔸 Point B (4,7)

- To C1:

$$\sqrt{(4-2)^2+(7-3)^2} = \sqrt{4+16} = \sqrt{20} \approx 4.47$$

- To C2:

$$\sqrt{(4-8)^2+(7-6)^2} = \sqrt{16+1} = \sqrt{17} \approx 4.12$$

✅ Cluster 2

🔸 Point C (3,5)

- To C1:

$$\sqrt{(3-2)^2+(5-3)^2} = \sqrt{1+4} = \sqrt{5} \approx 2.24$$

- To C2:

$$\sqrt{(3-8)^2+(5-6)^2} = \sqrt{25+1} = \sqrt{26} \approx 5.10$$

✅ Cluster 1

🔸 Point D (6,9)

- To C1:

$$\sqrt{(6-2)^2+(9-3)^2} = \sqrt{16+36} = \sqrt{52} \approx 7.21$$

- To C2:

$$\sqrt{(6-8)^2+(9-6)^2} = \sqrt{4+9} = \sqrt{13} \approx 3.61$$

✅ Cluster 2

🔸 **Point E (8,6)**

- To C1:

$$\sqrt{(8-2)^2+(6-3)^2} = \sqrt{36+9} = \sqrt{45} \approx 6.71$$

- To C2:

$$\sqrt{(8-8)^2+(6-6)^2} = \sqrt{0+0} = 0$$

✅ Cluster 2

🔸 **Point F (7,8)**

- To C1:

$$\sqrt{(7-2)^2+(8-3)^2} = \sqrt{25+25} = \sqrt{50} \approx 7.07$$

- To C2:

$$\sqrt{(7-8)^2+(8-6)^2} = \sqrt{1+4} = \sqrt{5} \approx 2.24$$

✅ Cluster 2

Step 3: New Cluster Assignments

- Cluster 1 (C1): A, B, C
- Cluster 2 (C2): D, E, F

Step 4: Recalculate Centroids

C1 :

- $\bar{X}$ = (2 + 4 + 3)/3 = 3.0
- $\bar{Y}$ = (3 + 7 + 5)/3 = 5.0
  → **New C1 = (3.0, 5.0)**

C2 :

- $\bar{X}$ = (6 + 8 + 7)/3 = 7.0
- $\bar{Y}$ = (9 + 6 + 8)/3 = 7.67
  → **New C2 = (7.0, 7.67)**

Step 5 – Reassign Points Using New Centroids

From previous steps (Step 4), we had these new centroids:

- **C1:** (3.0, 5.0)

- **C2:** (7.0, 7.67)

Now, we'll recalculate the distance from each point to these new centroids and reassign them.

🔸 **Point A (2,3)**

- To C1:

$$\sqrt{(2-3)^2+(3-5)^2} = \sqrt{1+4} = \sqrt{5} \approx 2.24$$

- To C2:

$$\sqrt{(2-7)^2+(3-7.67)^2} = \sqrt{25+21.8} = \sqrt{46.8} \approx 6.84$$

✅ Cluster 1

🔸 **Point B (4,7)**

- To C1:

$$\sqrt{(4-3)^2+(7-5)^2} = \sqrt{1+4} = \sqrt{5} \approx 2.24$$

- To C2:

$$\sqrt{(4-7)^2+(7-7.67)^2} = \sqrt{9+0.45} = \sqrt{9.45} \approx 3.07$$

✅ Cluster 1

🔸 **Point C (3,5)**

- To C1:

$$\sqrt{(3-3)^2+(5-5)^2} = \sqrt{0+0} = 0$$

- To C2:

$$\sqrt{(3-7)^2+(5-7.67)^2} = \sqrt{16+7.1} = \sqrt{23.1} \approx 4.81$$

✅ Cluster 1

🔸 **Point D (6,9)**

- To C1:

$$\sqrt{(6-3)^2+(9-5)^2} = \sqrt{9+16} = \sqrt{25} = 5$$

- To C2:

$$\sqrt{(6-7)^2+(9-7.67)^2} = \sqrt{1+1.77} = \sqrt{2.77} \approx 1.66$$

✅ Cluster 2

🔸 **Point E (8,6)**

- To C1:

$$\sqrt{(8-3)^2+(6-5)^2} = \sqrt{25+1} = \sqrt{26} \approx 5.10$$

- To C2:

$$\sqrt{(8-7)^2+(6-7.67)^2} = \sqrt{1+2.8} = \sqrt{3.8} \approx 1.95$$

✅ Cluster 2

🔸 **Point F (7,8)**

- To C1:

$$\sqrt{(7-3)^2+(8-5)^2} = \sqrt{16+9} = \sqrt{25} = 5$$

- To C2:

$$\sqrt{(7-7)^2+(8-7.67)^2} = \sqrt{0+0.11} = \sqrt{0.11} \approx 0.33$$

✅ Cluster 2

Final Cluster Assignments after Step 5

Cluster	Points
C1	A, B, C
C2	D, E, F

b) How do you handle noise and irrelevant information in text data during preprocessing? Explain the terms bag of words and TF IDF in text analytics.[9]

1. Handling Noise and Irrelevant Information in Text Preprocessing

To clean noisy text data, we apply the following steps during **text preprocessing**:

Remove Punctuation and Special Characters

- E.g., “Hello!!!” → “Hello”

Remove Stop Words

- Common words like *is, the, and, was* are removed as they add little meaning.

Lowercasing

- Converts all text to lowercase for consistency: “Text” → “text”

Remove Numbers *(optional)*

- If numbers don't carry value, they are removed.

Spelling Correction / Normalization

- Fixing misspelled words to reduce redundancy.

Lemmatization / Stemming

- Reduce words to their root form for uniformity.

2. Bag of Words (BoW)

- Bag of Words is a **simple feature extraction technique** used in text analysis.
- It represents text as a collection (bag) of words **without considering order**.
- It uses the **frequency (count)** of each word to represent a document.

  Example:
  Document: “I love data science”

  Vocabulary: [“I”, “love”, “data”, “science”]

  BoW vector: [1, 1, 1, 1]

- If “data” appears twice: [1, 1, 2, 1]

3.TF-IDF (Term Frequency – Inverse Document Frequency)

Already given !!

Q6

a) Explain how hierarchical clustering can be used for visualizing hierarchical relationships in data with a suitable example. What are some real-world applications of hierarchical clustering? [9 Marks]

1. Hierarchical Clustering Overview

- Hierarchical clustering builds a **tree-like structure** (called a **dendrogram**) to show nested grouping of data points.
- It does not require pre-specifying the number of clusters (unlike K-means).
- Two types:
    - **Agglomerative (Bottom-Up)**: Start with each point as its own cluster and merge step-by-step.
    - **Divisive (Top-Down)**: Start with one cluster and split recursively.

2. Visualizing Hierarchical Relationships

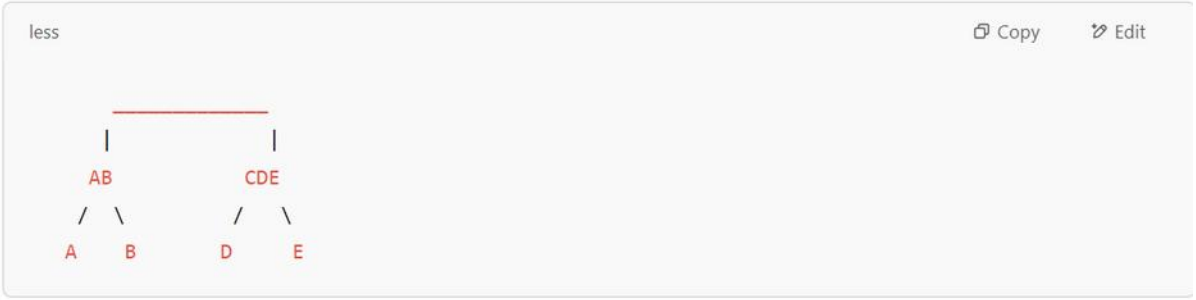
- The result is shown using a **dendrogram**, a tree diagram where:
    - Leaves = individual data points
    - Branches = cluster merges
- You can **cut the dendrogram at different heights** to choose the number of clusters.

Example:
Consider these 5 data points with similarities:

- A, B (similar), C (less similar), D & E (different)

Dendrogram View:

less Copy Edit



From the above, AB are closely linked, while D and E are far apart.

3. Real-World Applications of Hierarchical Clustering *(4 Marks)*

Customer Segmentation:

- Grouping customers by purchase behavior, age, income, etc.

Document Classification:

- Cluster articles based on topic similarity.

Gene Expression Analysis (Bioinformatics):

- Cluster genes that behave similarly across conditions.

Image Segmentation:

- Grouping pixels with similar colors or textures.

Social Network Analysis:

- Detecting communities or hierarchical relations between users.

b) What is the holdout method, and how does it work? Explain the difference between training set, validation set, and test set in the holdout method.[9]

Holdout Method:

The **Holdout Method** is a simple and widely used technique for evaluating the performance of a machine learning model. In this method, the available dataset is divided into two or three separate subsets:

- The **training set** (used to train the model),
- The **validation set** (used to tune parameters and prevent overfitting), and
- The **test set** (used to evaluate the final performance of the model).

How Holdout Method Works:

1. The dataset is **randomly split** into subsets, commonly:
    - 60-70% for training,
    - 10-20% for validation,
    - 20-30% for testing.
2. The **training set** is used by the algorithm to learn the model parameters.
3. The **validation set** is used to tune hyperparameters and select the best model configuration by assessing performance on unseen data during training.
4. Finally, the **test set** is used once at the end to evaluate the generalization ability of the trained model on completely new data.

Difference between Training, Validation, and Test Sets:

Set	Purpose	When Used
Training Set	Used to train the machine learning model by adjusting weights and parameters based on data.	During the model learning phase.
Validation Set	Used to tune hyperparameters (e.g., learning rate, number of layers) and perform model selection by checking performance on unseen data during training.	During training but separate from training data.
Test Set	Used to assess the final performance and generalization of the model after training and validation are complete. It acts as unseen, unbiased data.	Only after the model is fully trained and tuned.

- The **holdout method** helps to estimate how well a model will perform on unseen data.
- It prevents overfitting by separating data for training and testing.
- Using a **validation set** helps improve model tuning without biasing the final evaluation.
- The **test set** provides an unbiased estimate of model performance for real-world use.

➢ **NOV / DEC 2024**

Q5)

a) What is Hierarchical clustering? Explain hierarchical clustering algorithms. [9]

Hierarchical Clustering:

Hierarchical clustering is a **unsupervised machine learning** technique used to group similar data points into clusters based on their distance or similarity. It builds a hierarchy or tree of clusters, called a **dendrogram**, showing how clusters are merged or split at different levels.

Unlike flat clustering methods (like k-means), hierarchical clustering does **not require specifying the number of clusters in advance**.

Types of Hierarchical Clustering Algorithms:

There are two main types of hierarchical clustering:

1. **Agglomerative Hierarchical Clustering (Bottom-Up Approach):**
    - Starts with each data point as a separate cluster.
    - Iteratively merges the two closest clusters until only one cluster remains (or until a stopping criterion).
    - Produces a dendrogram showing merges at different similarity levels.
2. **Divisive Hierarchical Clustering (Top-Down Approach):**

- Starts with all data points in one single cluster.
- Recursively splits clusters into smaller clusters until each data point is alone or a stopping condition is met.
- Less commonly used due to computational complexity.

Agglomerative Clustering Algorithm Steps:

1. Assign each data point to its own cluster.
2. Compute the distance (or similarity) between all clusters.
3. Merge the two closest clusters into one.
4. Update the distance matrix to reflect the merge.
5. Repeat steps 2-4 until only one cluster remains or desired number of clusters is reached.

Distance/Linkage Criteria Used:

- **Single Linkage:** Distance between the closest points of two clusters.
- **Complete Linkage:** Distance between the farthest points of two clusters.
- **Average Linkage:** Average distance between all points of two clusters.
- **Ward's Method:** Minimizes the variance within clusters after merging.

Aspect	Description
Type	Unsupervised, hierarchical grouping
Approaches	Agglomerative (bottom-up), Divisive (top-down)
Output	Dendrogram showing nested clusters
No. of clusters needed	Not required beforehand
Use cases	Gene expression, image segmentation, market segmentation

b) Write a note on: [8 Marks]

i) Holdout Method
ii) k-Fold Cross-Validation

i) Holdout Method

The holdout method is a simple technique for evaluating machine learning models by splitting the dataset into distinct subsets: typically, a **training set** and a **test set**.

- The **training set** is used to train the model.
- The **test set** is used to evaluate its performance on unseen data.

Sometimes, a **validation set** is also used to tune model parameters. The holdout method is fast and straightforward but can be sensitive to how the data is split, potentially causing variance in performance results if the split is not representative.

ii) k-Fold Cross-Validation

k-Fold Cross-Validation is a more robust evaluation method that splits the dataset into **k equal-sized folds** (subsets). The process is:

- The model is trained on **k-1 folds** and tested on the **remaining fold**.
- This is repeated **k times**, each time using a different fold as the test set.
- The results are averaged to provide a more reliable estimate of model performance.

This method reduces bias and variance compared to the holdout method, especially for small datasets, by using all data points for both training and testing.

Method	Description	Pros	Cons
Holdout	Single split into train/test	Simple, fast	High variance, biased split possible
k-Fold Cross-Validation	Multiple train/test splits (k times)	More reliable, less biased	More computationally expensive

Q6) a) What is Text Analysis? Explain the different steps involved in text analysis. [9]

Text Analysis

Text analysis, also known as **text mining** or **natural language processing (NLP)**, is the process of extracting meaningful information and insights from unstructured textual data. It involves transforming raw text into structured data that can be analyzed to identify patterns, trends, sentiments, or topics.

Text analysis is widely used in applications like sentiment analysis, spam detection, topic modeling, and chatbots.

Steps Involved in Text Analysis:

1. **Text Collection:**
   Gathering raw textual data from various sources such as documents, social media, emails, or web pages.
2. **Text Preprocessing:**
   Cleaning and preparing text data by:
   - Removing noise (punctuation, numbers, special characters)
   - Converting text to lowercase
   - Removing stop words (common words like “the”, “is”)
   - Tokenization (splitting text into words or phrases)
   - Stemming or Lemmatization (reducing words to their root forms)
3. **Text Transformation:**
   Converting cleaned text into a structured format like:
   - Bag of Words (BoW) model
   - Term Frequency-Inverse Document Frequency (TF-IDF)
   - Word embeddings (Word2Vec, GloVe)
4. **Feature Extraction:**
   Selecting relevant features (words, phrases, or patterns) for analysis, reducing dimensionality if needed.
5. **Text Analysis and Modeling:**
   Applying machine learning, statistical methods, or linguistic rules to analyze text:
   - Classification (e.g., spam or not spam)
   - Clustering (grouping similar documents)
   - Sentiment analysis
   - Topic modeling
6. **Interpretation and Visualization:**
   Presenting the results in a comprehensible form such as charts, word clouds, or reports for decision-making.

Text analysis helps convert unstructured text into meaningful insights using structured techniques. It is widely used in industries for better decision-making, automation, and understanding customer feedback.

b) Write a note on Social Network Analysis. What are the applications of Social Network Analysis? [8]

Social Network Analysis (SNA):

Social Network Analysis (SNA) is a technique used to study relationships and structures within a network of social entities such as people, organizations, or systems.

It focuses on understanding how individuals (nodes) are connected through relationships (edges) and how these connections influence behaviour , information flow, and group dynamics.

SNA uses concepts from graph theory to model social structures, where entities are represented as **nodes**, and their connections as **edges or links**. It helps analyze the **strength, direction, and pattern** of relationships.

Key Concepts in SNA:

- **Node (Vertex):** Represents an individual or entity in the network.
- **Edge (Link):** Represents a relationship or interaction between two nodes.
- **Degree Centrality:** Number of connections a node has.
- **Betweenness Centrality:** How often a node appears on the shortest path between other nodes.
- **Closeness Centrality:** How close a node is to all other nodes in the network.

Applications of Social Network Analysis:

1. **Marketing and Influence Detection:**
   Identify key influencers and target audiences in social media platforms.
2. **Fraud Detection:**
   Analyze transaction patterns and detect suspicious connections in financial networks.
3. **Healthcare:**
   Understand disease transmission patterns and patient referral networks.
4. **Cybersecurity:**
   Detect threats by analyzing communication patterns in a network.
5. **Organizational Management:**
   Improve internal communication by analyzing employee interaction networks.
6. **Recommendation Systems:**
   Improve product or friend suggestions based on user connections and behaviors.

Social Network Analysis provides deep insights into the structure and behavior of complex networks. It is a powerful tool used across various domains like business, healthcare, and cybersecurity for strategic decision-making and problem-solving.