1.What are the different types of machine learning techniques? List any five major applications of machine learning.(5 + 5)

1 ANSWER—--

Types of Machine Learning Techniques:(5 M)

1. **Supervised Learning** – The model is trained on labeled data (input-output pairs).
    ○ Example: Spam email detection
2. **Unsupervised Learning** – The model identifies patterns in unlabeled data.
    ○ Example: Customer segmentation
3. **Semi-Supervised Learning** – A mix of labeled and unlabeled data to improve learning.
    ○ Example: Fraud detection in banking
4. **Reinforcement Learning** – The model learns by interacting with an environment and receiving rewards or penalties.
    ○ Example: Game-playing AI (e.g., AlphaGo)
5. **Self-Supervised Learning** – The model generates its own labels from raw data.
    ○ Example: Natural language processing (NLP)

Five Major Applications of Machine Learning:(5 M)

1. **Healthcare** – Disease diagnosis, personalized treatment, medical imaging analysis
2. **Finance** – Fraud detection, stock market prediction, credit scoring
3. **E-commerce** – Recommendation systems, price optimization, customer sentiment analysis
4. **Autonomous Vehicles** – Object detection, path planning, self-driving cars

5. **Cybersecurity** – Malware detection, intrusion detection, phishing prevention

2.List and explain the Data Preprocessing. techniques . Explain the steps to improve the quality of data.

Applying various binning techniques and show the results by considering the following set

S={12,14,19,22,24,26,28,31,34}

2 ANSWER-

.Data preprocessing is a crucial step in machine learning that involves transforming raw data into a clean and usable format. The goal is to improve data quality, ensure consistency, and enhance the model's performance. The key techniques include:------------------4 M

1. **Data Cleaning**
   - Handling missing values (e.g., imputation, removal)
   - Removing duplicate records
   - Correcting inconsistent data entries
2. **Data Transformation**
   - Normalization (scaling values between 0 and 1)
   - Standardization (scaling data to have zero mean and unit variance)
   - Encoding categorical variables (e.g., one-hot encoding, label encoding)
3. **Data Reduction**
   - Feature selection (choosing relevant variables)
   - Feature extraction (reducing dimensionality, e.g., PCA)
   - Sampling (reducing the dataset size while maintaining patterns)

4. **Data Integration**
   - Merging multiple datasets (e.g., joining customer data from different sources)
   - Resolving schema conflicts (ensuring consistency in data formats)
5. **Data Discretization & Binning**
   - Converting continuous data into categorical bins (e.g., age groups)
   - Reducing data complexity while maintaining interpretability

---

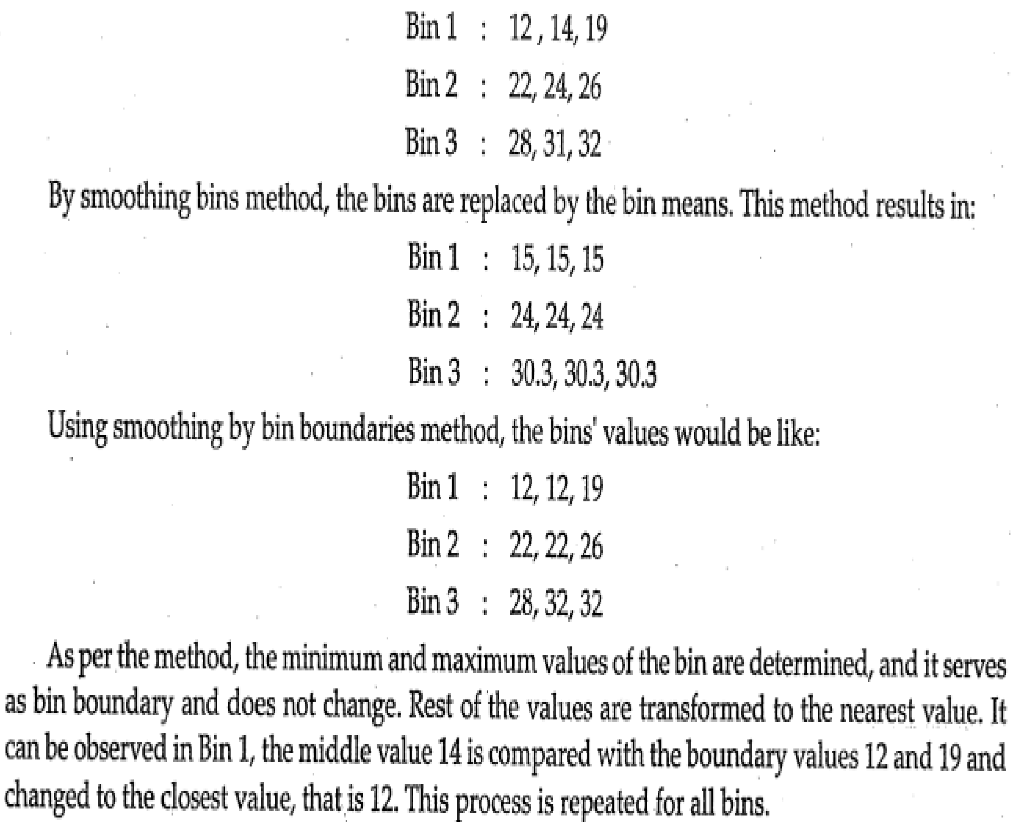
Steps to Improve Data Quality—--------------------------4 M

1. **Understand the Data**
   - Conduct exploratory data analysis (EDA)
   - Identify missing values, outliers, and inconsistencies
2. **Handle Missing Data**
   - Remove rows/columns with excessive missing values
   - Use imputation methods (mean, median, mode, or predictive models)
3. **Remove Duplicates and Inconsistent Data**
   - Identify duplicate records and remove them
   - Correct inconsistent data entries (e.g., different date formats)
4. **Standardize and Normalize Data**
   - Apply normalization or standardization to ensure uniform data scales
   - Handle categorical variables properly with encoding techniques
5. **Detect and Handle Outliers**

- ○ Use visualization techniques (box plots, histograms) to detect outliers
- ○ Apply statistical methods (Z-score, IQR) to remove or transform outliers

6. **Ensure Data Consistency and Accuracy**
   - ○ Validate data using domain knowledge
   - ○ Cross-check with multiple sources if necessary
7. **Balance the Dataset (if applicable)**
   - ○ Handle imbalanced datasets using oversampling (SMOTE) or undersampling
   - ○ Ensure fair representation of all classes in classification problems
8. **Feature Engineering**
   - ○ Create new relevant features based on existing data
   - ○ Remove irrelevant or redundant features

—------------------------2 M—---



Bin 1 : 12, 14, 19
Bin 2 : 22, 24, 26
Bin 3 : 28, 31, 32

By smoothing bins method, the bins are replaced by the bin means. This method results in:

Bin 1 : 15, 15, 15
Bin 2 : 24, 24, 24
Bin 3 : 30.3, 30.3, 30.3

Using smoothing by bin boundaries method, the bins' values would be like:

Bin 1 : 12, 12, 19
Bin 2 : 22, 22, 26
Bin 3 : 28, 32, 32

As per the method, the minimum and maximum values of the bin are determined, and it serves as bin boundary and does not change. Rest of the values are transformed to the nearest value. It can be observed in Bin 1, the middle value 14 is compared with the boundary values 12 and 19 and changed to the closest value, that is 12. This process is repeated for all bins.

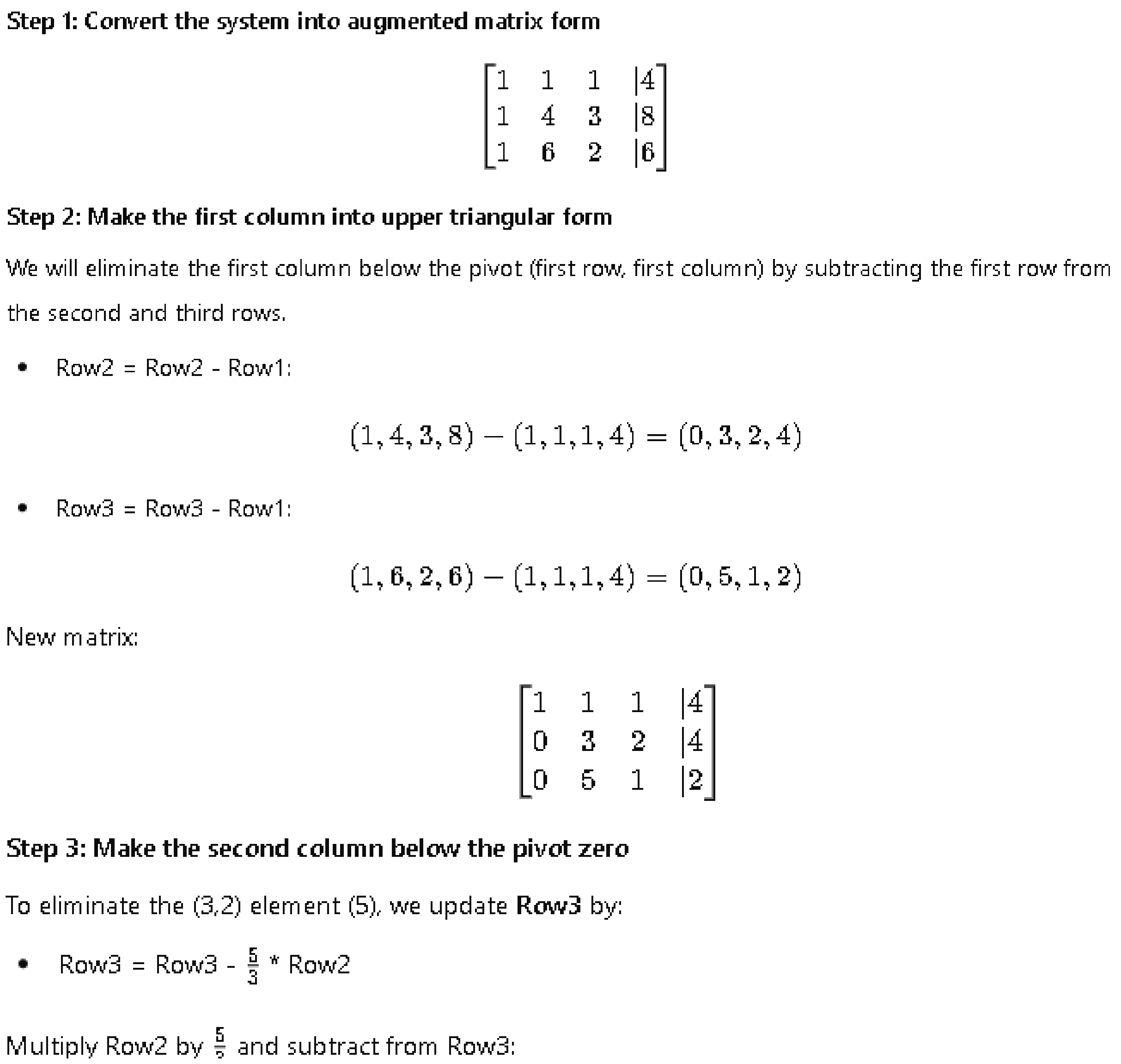
3. **Given system of equations:**

$x + y + z = 4$

$x + 4y + 3z = 8$

$x + 6y + 2z = 6$

—-------------------10 M—---



Step 1: Convert the system into augmented matrix form

$$\begin{bmatrix} 1 & 1 & 1 & |4 \\ 1 & 4 & 3 & |8 \\ 1 & 6 & 2 & |6 \end{bmatrix}$$

Step 2: Make the first column into upper triangular form

We will eliminate the first column below the pivot (first row, first column) by subtracting the first row from the second and third rows.

- Row2 = Row2 - Row1:

$$(1,4,3,8) - (1,1,1,4) = (0,3,2,4)$$

- Row3 = Row3 - Row1:

$$(1,6,2,6) - (1,1,1,4) = (0,5,1,2)$$

New matrix:

$$\begin{bmatrix} 1 & 1 & 1 & |4 \\ 0 & 3 & 2 & |4 \\ 0 & 5 & 1 & |2 \end{bmatrix}$$

Step 3: Make the second column below the pivot zero

To eliminate the (3,2) element (5), we update **Row3** by:

- Row3 = Row3 - $\frac{5}{3}$ * Row2

Multiply Row2 by $\frac{5}{3}$ and subtract from Row3:

$$(0,5,1,2) - \left(\frac{5}{3} \times (0,3,2,4)\right)$$

$$(0,5,1,2) - (0,5,\frac{10}{3},\frac{20}{3})$$

$$(0,0,-\frac{7}{3},-\frac{14}{3})$$

New matrix:

$$\begin{bmatrix} 1 & 1 & 1 & |4 \\ 0 & 3 & 2 & |4 \\ 0 & 0 & -\frac{7}{3} & |-\frac{14}{3} \end{bmatrix}$$

Step 4: Back Substitution

From the last row:

$$-\frac{7}{3}z = -\frac{14}{3}$$

$$z = 2$$

From the second row:

$$3y + 2(2) = 4$$

$$3y + 4 = 4$$

$$y = 0$$

From the first row:

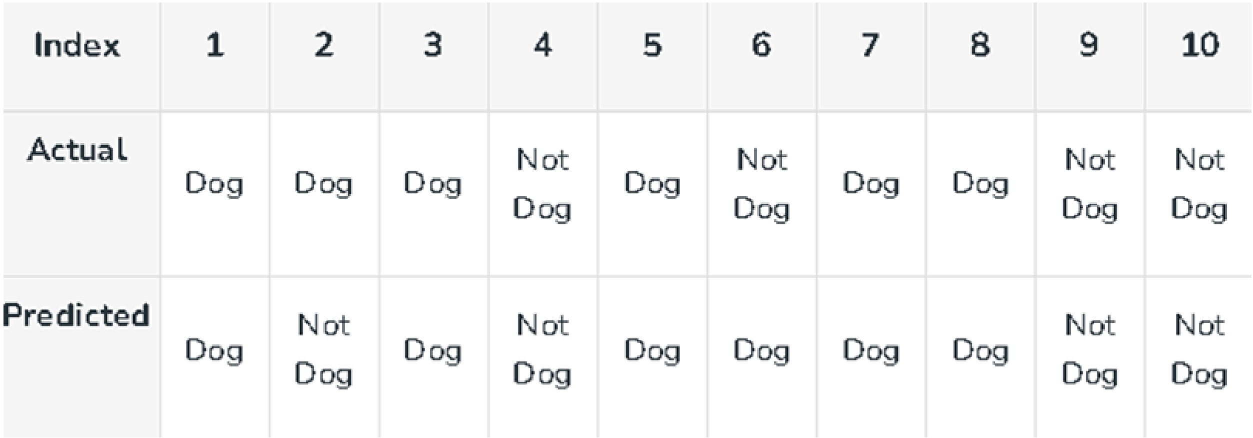
$$x + 0 + 2 = 4$$

$$x = 2$$

Final Solution:

$$x = 2, \quad y = 0, \quad z = 2$$

4.What is a confusion matrix.



Index	1	2	3	4	5	6	7	8	9	10
Actual	Dog	Dog	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Not Dog	Not Dog
Predicted	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Dog	Dog	Not Dog	Not Dog

Calculate the accuracy, precision, recall, sensitivity and F1 score from the matrix.

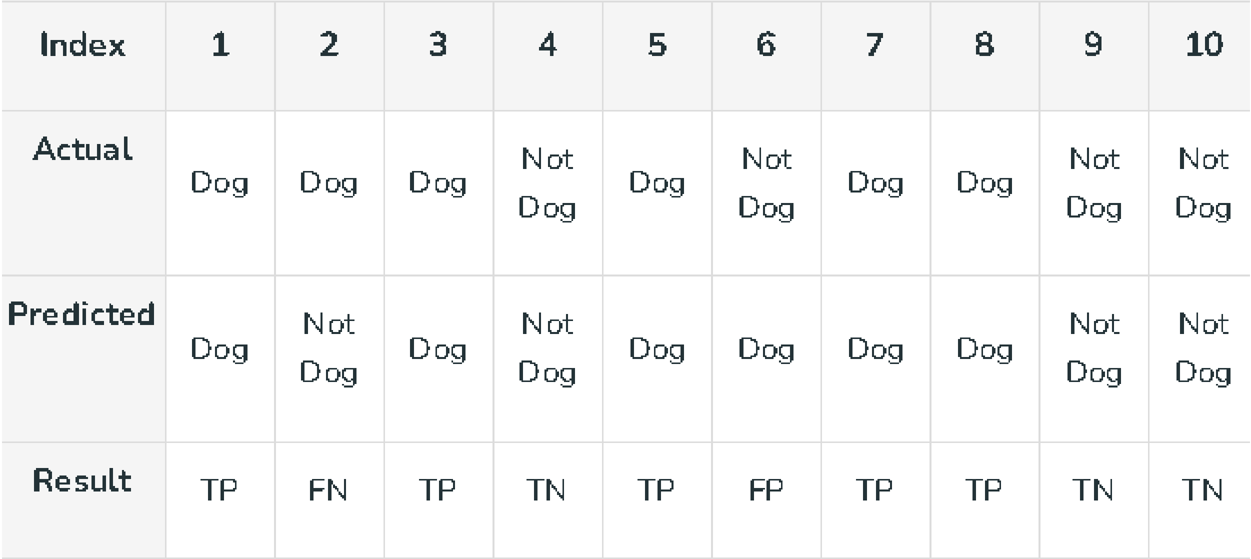
4 ANSWER-

A confusion matrix is a simple table that shows how well a classification model is performing by comparing its predictions to the actual results. It breaks down the predictions into four categories: correct predictions for both classes (true positives and true negatives) and incorrect predictions (false positives and false negatives). This helps you understand where the model is making mistakes, so you can improve it.

The matrix displays the number of instances produced by the model on the test data.

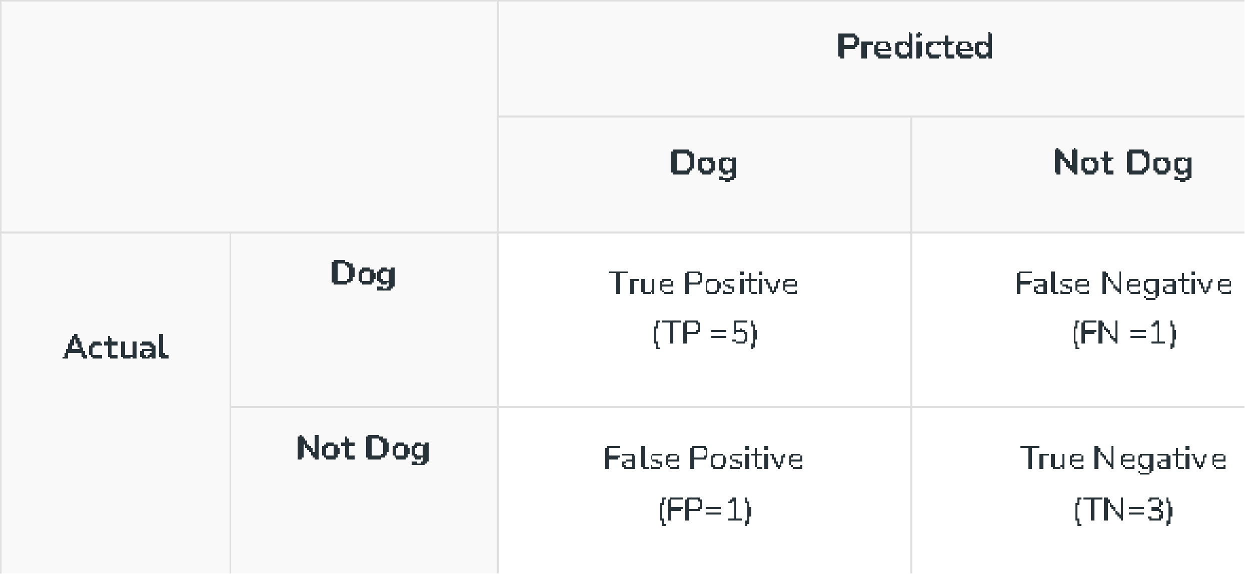
- **True Positive (TP): The model correctly predicted a positive outcome (the actual outcome was positive).**
- **True Negative (TN): The model correctly predicted a negative outcome (the actual outcome was negative).**

- False Positive (FP): The model incorrectly predicted a positive outcome (the actual outcome was negative). Also known as a Type I error.
- False Negative (FN): The model incorrectly predicted a negative outcome (the actual outcome was positive). Also known as a Type II error.



Index	1	2	3	4	5	6	7	8	9	10
Actual	Dog	Dog	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Not Dog	Not Dog
Predicted	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Dog	Dog	Not Dog	Not Dog
Result	TP	FN	TP	TN	TP	FP	TP	TP	TN	TN

- Actual Dog Counts = 6
- Actual Not Dog Counts = 4
- True Positive Counts = 5
- False Positive Counts = 1
- True Negative Counts = 3
- False Negative Counts = 1



		Predicted	
		Dog	Not Dog
Actual	Dog	True Positive (TP =5)	False Negative (FN =1)
	Not Dog	False Positive (FP=1)	True Negative (TN=3)

Accuracy = 80%

Precision = 83.33%

Recall (Sensitivity) = 83.33%

F1 Score = 83.33%

5 Write an algorithm for PCA. Apply PCA for the following matrix and prove that it works.

$$X = \begin{bmatrix} 2 & 3 \\ 3 & 4 \\ 4 & 5 \end{bmatrix}$$

ANSWER-

PCA is a dimensionality reduction technique that transforms data into a new coordinate system where the axes (principal components) capture the maximum variance.--------------------2 Marks

PCA Algorithm Steps:

1. **Standardize the Dataset**
    - Subtract the mean from each feature.
    - Scale the features to have unit variance (optional).
2. **Compute the Covariance Matrix**
    - Calculate the covariance between each pair of features.
3. **Compute the Eigenvalues and Eigenvectors**
    - Solve for eigenvalues and eigenvectors of the covariance matrix.
    - The eigenvectors represent principal components.
    - The eigenvalues indicate the amount of variance captured by each principal component.
4. **Sort and Select the Top k Principal Components**
    - Rank eigenvalues in descending order.
    - Select the top k eigenvectors corresponding to the largest eigenvalues.
5. **Transform the Data**
    - Project the original data onto the new lower-dimensional space using the selected principal components.

—------------------8 M—---------------------------------

$$X = \begin{bmatrix} 2 & 3 \\ 3 & 4 \\ 4 & 5 \end{bmatrix}$$

Step 1: Compute the Mean of Each Feature

Compute the column-wise mean:

$$\mu_1 = \frac{2+3+4}{3} = 3, \quad \mu_2 = \frac{3+4+5}{3} = 4$$

Mean vector:

$$\mu = [3, 4]$$

Step 2: Center the Data (Subtract Mean)

$$X_{\text{centered}} = \begin{bmatrix} 2-3 & 3-4 \\ 3-3 & 4-4 \\ 4-3 & 5-4 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$$

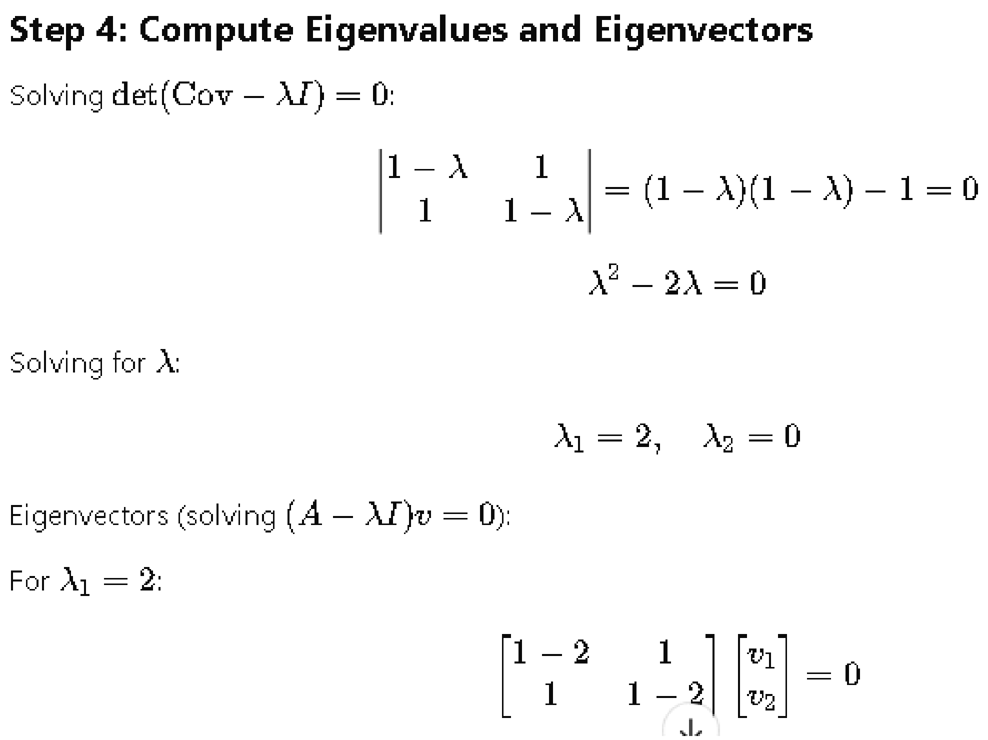
Step 3: Compute the Covariance Matrix

Covariance formula:

$$\text{Cov}(X) = \frac{1}{n-1} X^T X$$

$$\text{Cov}(X) = \frac{1}{2} \begin{bmatrix} (-1)(-1) + (0)(0) + (1)(1) & (-1)(-1) + (0)(0) + (1)(1) \\ (-1)(-1) + (0)(0) + (1)(1) & (-1)(-1) + (0)(0) + (1)(1) \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$



Step 4: Compute Eigenvalues and Eigenvectors

Solving $\det(\mathrm{Cov} - \lambda I) = 0$:

$$\begin{vmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix} = (1-\lambda)(1-\lambda) - 1 = 0$$

$$\lambda^2 - 2\lambda = 0$$

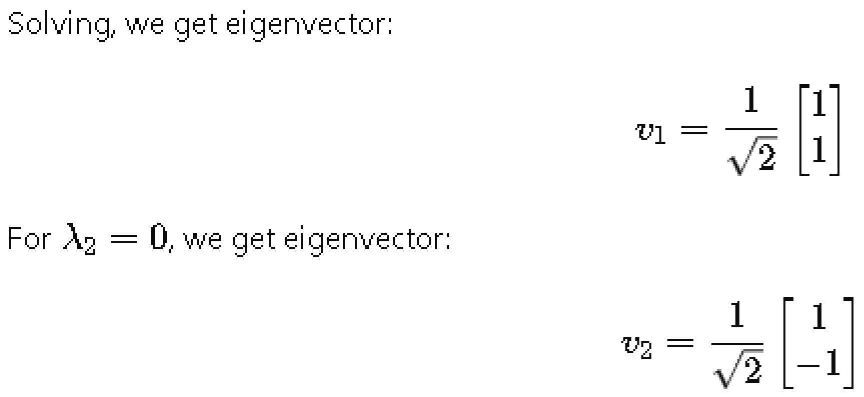
Solving for λ:

$$\lambda_1 = 2, \quad \lambda_2 = 0$$

Eigenvectors (solving $(A - \lambda I)v = 0$):

For $\lambda_1 = 2$:

$$\begin{bmatrix} 1-2 & 1 \\ 1 & 1-2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

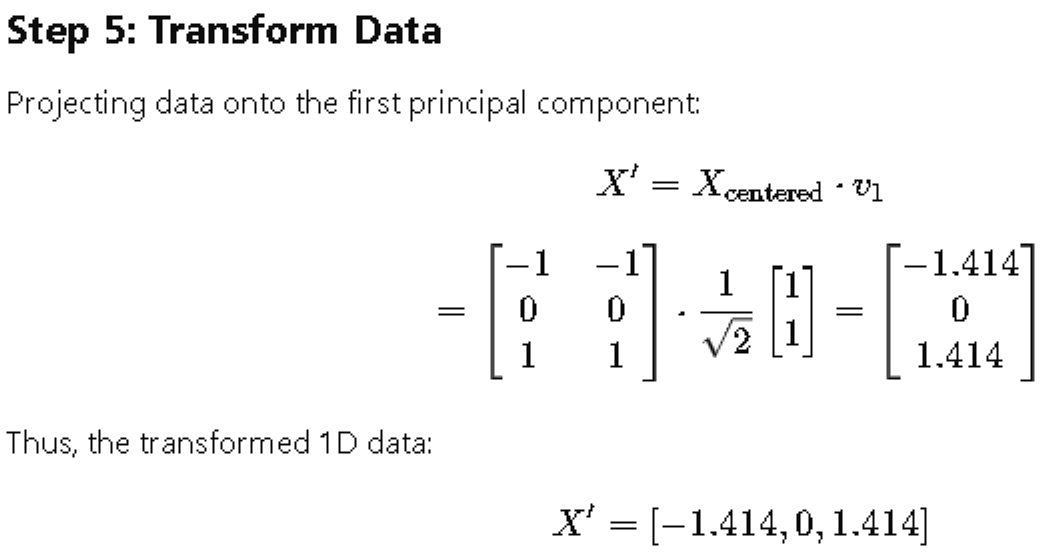


Solving, we get eigenvector:

$$v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

For $\lambda_2 = 0$, we get eigenvector:

$$v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$



Step 5: Transform Data

Projecting data onto the first principal component:

$$X' = X_{\text{centered}} \cdot v_1$$

$$= \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1.414 \\ 0 \\ 1.414 \end{bmatrix}$$

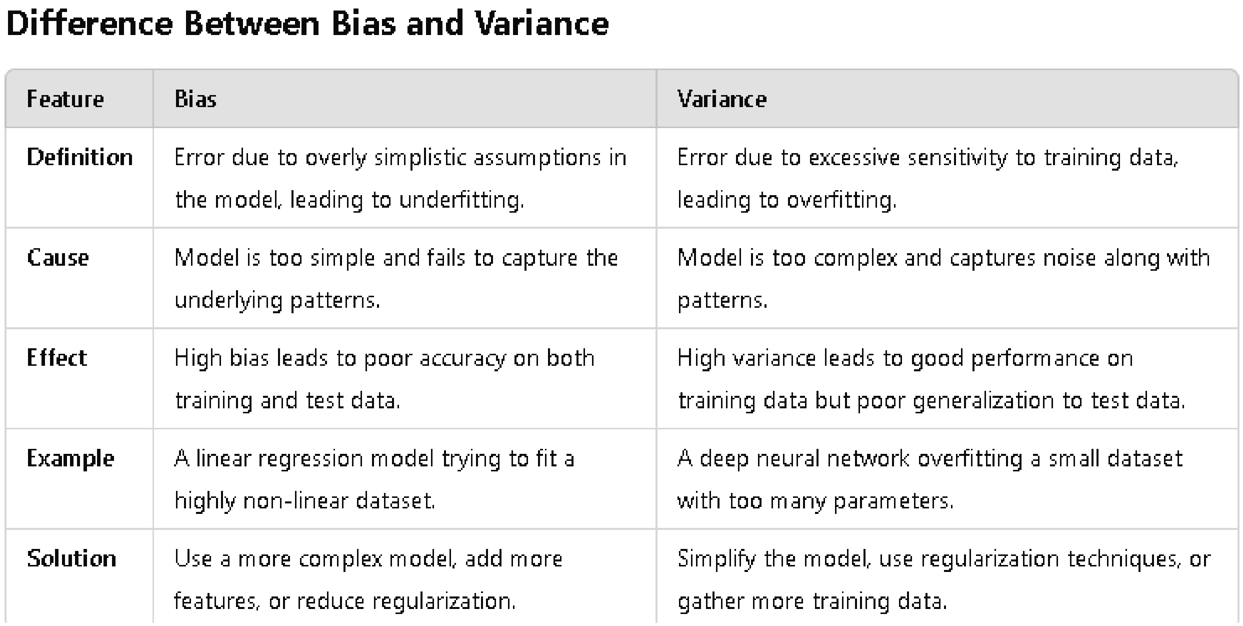
Thus, the transformed 1D data:

$$X' = [-1.414, 0, 1.414]$$

6 a)Differentiate between

1.Training and Testing Datasets

2.Bias and variance

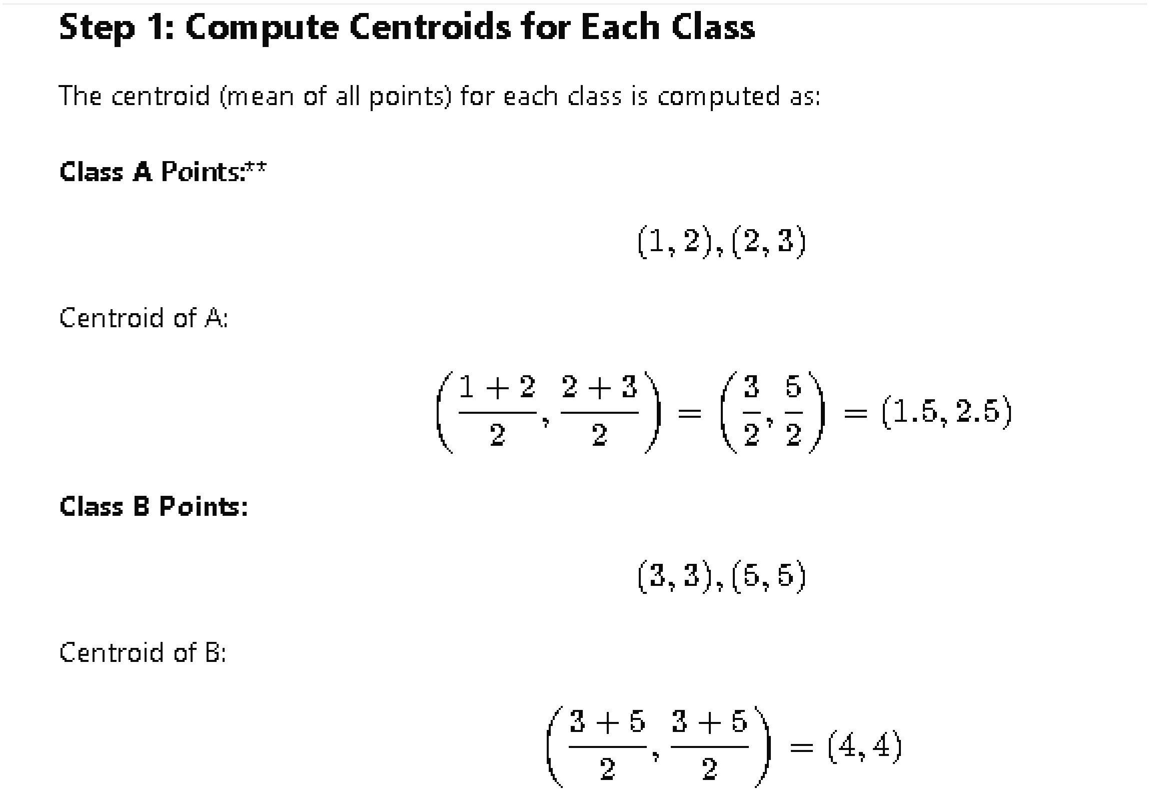


Difference Between Bias and Variance

Feature	Bias	Variance
Definition	Error due to overly simplistic assumptions in the model, leading to underfitting.	Error due to excessive sensitivity to training data, leading to overfitting.
Cause	Model is too simple and fails to capture the underlying patterns.	Model is too complex and captures noise along with patterns.
Effect	High bias leads to poor accuracy on both training and test data.	High variance leads to good performance on training data but poor generalization to test data.
Example	A linear regression model trying to fit a highly non-linear dataset.	A deep neural network overfitting a small dataset with too many parameters.
Solution	Use a more complex model, add more features, or reduce regularization.	Simplify the model, use regularization techniques, or gather more training data.

6.b)Consider the data sample having two features x and y. The target variable has two classes A or B. Predict the class using Nearest Centroid Classifier. Classify a new point (3,4) using K=3 (3-Nearest Neighbour) with the nearest centroid classifier.

(X,Y)	class
(1,2)	A
(2,3)	A
(3,3)	B
(5,5)	B



Step 1: Compute Centroids for Each Class

The centroid (mean of all points) for each class is computed as:

Class A Points:**

$$(1, 2), (2, 3)$$

Centroid of A:

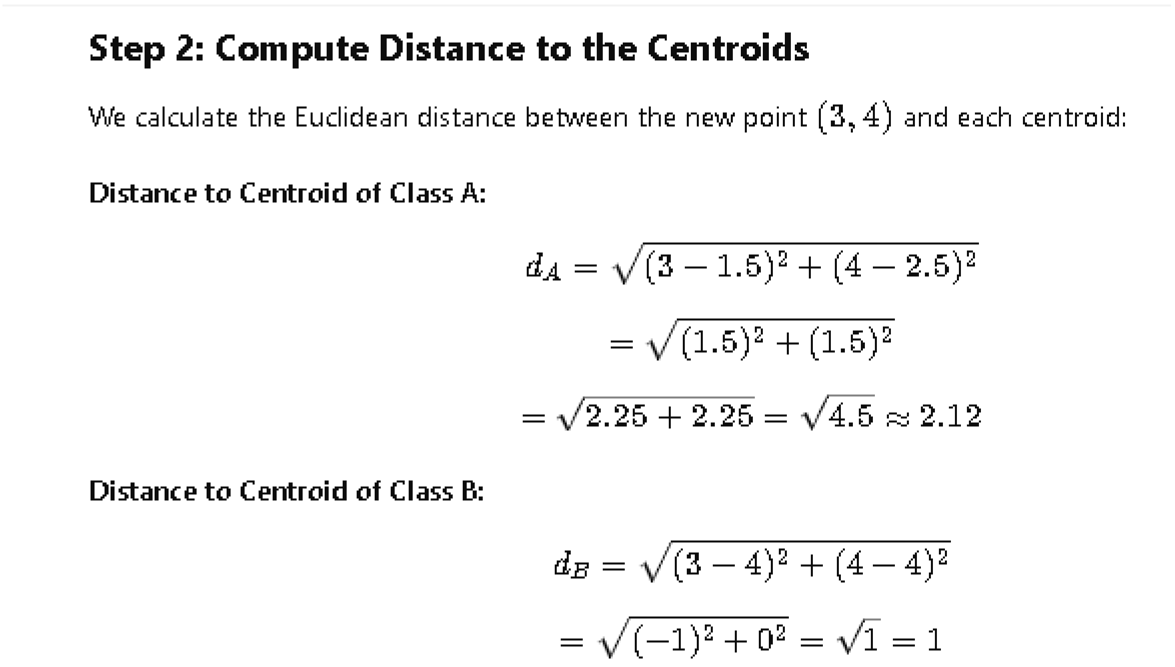
$$\left(\frac{1+2}{2}, \frac{2+3}{2}\right) = \left(\frac{3}{2}, \frac{5}{2}\right) = (1.5, 2.5)$$

Class B Points:

$$(3, 3), (5, 5)$$

Centroid of B:

$$\left(\frac{3+5}{2}, \frac{3+5}{2}\right) = (4, 4)$$



Step 2: Compute Distance to the Centroids

We calculate the Euclidean distance between the new point $(3, 4)$ and each centroid:

Distance to Centroid of Class A:

$$d_A = \sqrt{(3-1.5)^2 + (4-2.5)^2}$$

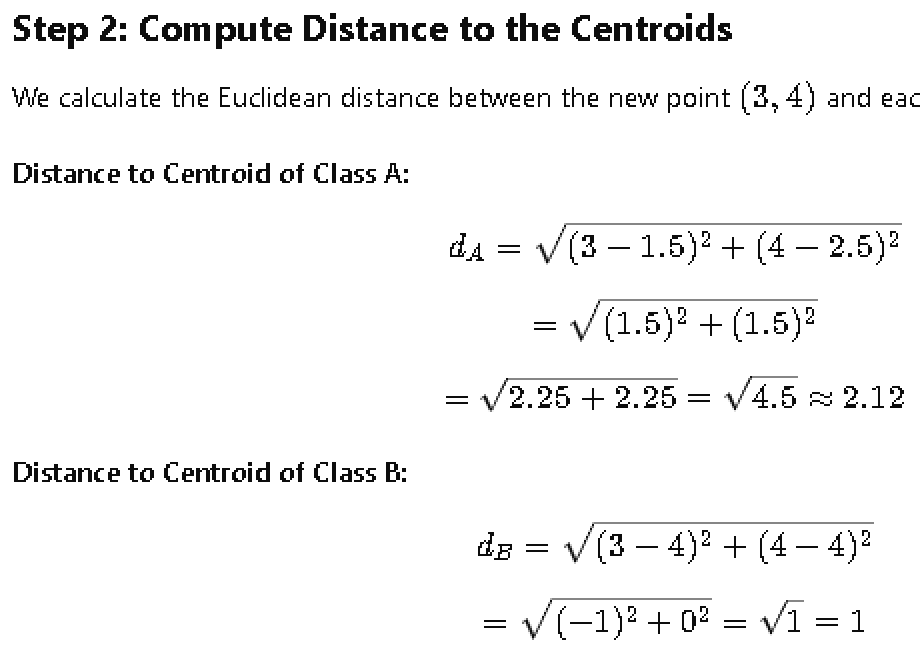
$$= \sqrt{(1.5)^2 + (1.5)^2}$$

$$= \sqrt{2.25 + 2.25} = \sqrt{4.5} \approx 2.12$$

Distance to Centroid of Class B:

$$d_B = \sqrt{(3-4)^2 + (4-4)^2}$$

$$= \sqrt{(-1)^2 + 0^2} = \sqrt{1} = 1$$



Step 2: Compute Distance to the Centroids

We calculate the Euclidean distance between the new point $(3, 4)$ and each centroid:

Distance to Centroid of Class A:

$$d_A = \sqrt{(3-1.5)^2 + (4-2.5)^2}$$

$$= \sqrt{(1.5)^2 + (1.5)^2}$$

$$= \sqrt{2.25 + 2.25} = \sqrt{4.5} \approx 2.12$$

Distance to Centroid of Class B:

$$d_B = \sqrt{(3-4)^2 + (4-4)^2}$$

$$= \sqrt{(-1)^2 + 0^2} = \sqrt{1} = 1$$

Step 3: Classification Decision

Since the new point $(3, 4)$ is **closer to the centroid of Class B** ($d_B = 1$ vs. $d_A = 2.12$), it is classified as:

$$\textbf{Class B}$$

Step 4: K-Nearest Neighbors (K=3) Approach

For **K=3 Nearest Neighbors**, we determine the three closest points to $(3, 4)$ from the dataset:

Point	Distance to (3,4)
(3,3) B	$\sqrt{(3-3)^2 + (4-3)^2} = \sqrt{1} = \mathbf{1}$
(2,3) A	$\sqrt{(3-2)^2 + (4-3)^2} = \sqrt{2} \approx \mathbf{1.41}$
(5,5) B	$\sqrt{(3-5)^2 + (4-5)^2} = \sqrt{5} \approx \mathbf{2.24}$

The **3 nearest neighbors** are:

1. (3,3) **B**
2. (2,3) **A**
3. (5,5) **B**

Since **two out of three neighbors belong to Class B**, the point $(3, 4)$ is classified as:

$$\textbf{Class B}$$

Final Answer:

- **Using Nearest Centroid Classifier: B**
- **Using K=3 Nearest Neighbors: B**