Machine Learning (BCS602)
Problems

MODULE 1

Q1. What is Binning? Consider the following set S={12,14,19,22,24,26,28,31,34}. Apply various binning techniques and show the result.

Simple Explanation with Real-Life Analogy:

Think of a supermarket arranging fruits by price.
Let's say you have 9 fruits costing ₹12, ₹14, ₹19, ₹22, ₹24, ₹26, ₹28, ₹31, ₹34.
If someone says:
"I don't want exact prices. Just tell me which range the fruits fall into — like low, medium, or high price."
That's exactly what **binning** does.

What is Binning in Machine Learning?

Binning is a **data preprocessing technique** used to **smooth noisy data** or **group numeric values** into **categories or bins**.
It helps in:

- Reducing the effect of minor observation errors or variations
- Simplifying models by converting continuous features to categorical ranges

---

Technically, What Are We Doing in Binning?

We are **dividing the data range into intervals (bins)** and then replacing actual values with:

- Mean of the bin
- Median of the bin
- Or keeping values but grouped within a bin

Now let's apply the three main **binning techniques** on your set:
Given:
S = {12, 14, 19, 22, 24, 26, 28, 31, 34}
We divide it into **3 bins (equal-frequency bins)** → Each bin has 3 elements.
So we have:

- Bin 1: {12, 14, 19}
- Bin 2: {22, 24, 26}
- Bin 3: {28, 31, 34}

---

1. Binning by Mean

In this method, each element in a bin is replaced with the **mean** (average) of that bin.
Calculations:

- Bin 1 mean = (12 + 14 + 19) / 3 = 15
- Bin 2 mean = (22 + 24 + 26) / 3 = 24
- Bin 3 mean = (28 + 31 + 34) / 3 = 31

Result:
S = {15, 15, 15, 24, 24, 24, 31, 31, 31}

---

2. Binning by Median

In this method, each element in a bin is replaced with the **median** (middle value) of that bin.

Calculations:

- Bin 1 median = 14
- Bin 2 median = 24
- Bin 3 median = 31

Result:

S = {14, 14, 14, 24, 24, 24, 31, 31, 31}

---

3. Binning by Boundary Values

Here, each value is replaced by the **nearest boundary value** (first or last value in the bin).

Bin 1: {12, 14, 19} → boundaries are 12 and 19

- 12 → 12
- 14 → closer to 12 → 12
- 19 → 19

Bin 2: {22, 24, 26} → boundaries 22 and 26

- 22 → 22
- 24 → closer to 22 → 22
- 26 → 26

Bin 3: {28, 31, 34} → boundaries 28 and 34

- 28 → 28
- 31 → closer to 28 → 28
- 34 → 34

Result:

S = {12, 12, 19, 22, 22, 26, 28, 28, 34}

Summary Table

Technique	Result
Binning by Mean	{15, 15, 15, 24, 24, 24, 31, 31, 31}
Binning by Median	{14, 14, 14, 24, 24, 24, 31, 31, 31}
Binning by Boundary	{12, 12, 19, 22, 22, 26, 28, 28, 34}

Question 2:
For S = {5, 10, 15, 20, 25, 30}, Find Mean, Median, Mode, Range, Standard Deviation, and Variance

Simple Explanation:

Let’s first understand what each of these terms means in a simple way:

- **Mean:** The average value (like total marks divided by number of subjects)
- **Median:** The middle value when numbers are sorted
- **Mode:** The most frequent value (like most common size in a shop)
- **Range:** Difference between highest and lowest value
- **Standard Deviation (SD):** How much the data values vary from the mean (spread of data)
- **Variance:** Square of the standard deviation (also tells spread)

Given Data:

S = {5, 10, 15, 20, 25, 30}

Total numbers, n = 6

1. Mean

Mean = (Sum of all values) / n
= (5 + 10 + 15 + 20 + 25 + 30) / 6
= 105 / 6 = **17.5**

---

2. Median

Since n = 6 (even), median = average of middle two values
Middle values = 15 and 20
Median = (15 + 20) / 2 = **17.5**

---

3. Mode

All values occur only once → No repeating value
Mode = No mode

---

4. Range

Range = Maximum value – Minimum value
= 30 – 5 = **25**

---

5. Variance (σ^2)

Step 1: Find the mean = 17.5
Step 2: Calculate (xi – mean)^2 for each xi

xi	xi – mean	(xi – mean)^2
5	-12.5	156.25
10	-7.5	56.25
15	-2.5	6.25
20	2.5	6.25
25	7.5	56.25
30	12.5	156.25

Sum of squares = 437.5
Variance = 437.5 / 6 = **72.92**

---

6. Standard Deviation (σ)

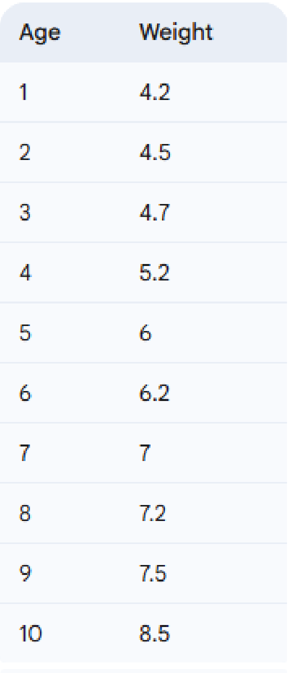
SD = $\sqrt{\text{Variance}} = \sqrt{72.92} \approx$ **8.54**

Final Answers:

Measure	Value
Mean	17.5
Median	17.5
Mode	No mode
Range	25
Variance	72.92
Standard Deviation	8.54

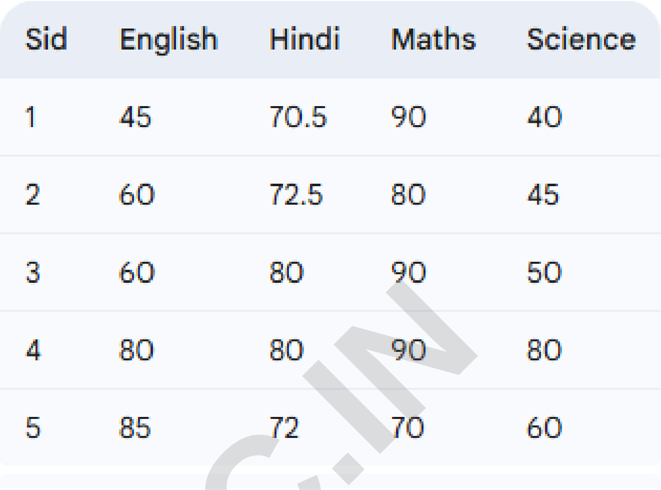
Q3. Considering the following table

Table 2.6: Sample Data



Age	Weight
1	4.2
2	4.5
3	4.7
4	5.2
5	6
6	6.2
7	7
8	7.2
9	7.5
10	8.5

Table 2.7: Students Marks Table



Sid	English	Hindi	Maths	Science
1	45	70.5	90	40
2	60	72.5	80	45
3	60	80	90	50
4	80	80	90	80
5	85	72	70	60

For univariate attribute weight, English, and maths, find the following:
i. Mean, median, mode
ii. Variance and standard deviation
iii. Five-point summary
iv. Skewness and kurtosis
v. Covariance between English and hindi marks

Ans:

Mean, Median, Mode

A. Weight : **Data**: 4.2, 4.5, 4.7, 5.2, 6, 6.2, 7, 7.2, 7.5, 8.5

- **Mean** = (Sum of all values) / Count
  = (4.2 + 4.5 + 4.7 + 5.2 + 6 + 6.2 + 7 + 7.2 + 7.5 + 8.5) / 10 = **6.1**
- **Median** = Average of 5th and 6th values (in sorted data)
  = (6 + 6.2) / 2 = **6.1**
- **Mode** = No value repeats → **No mode**

B. English Marks

Data: 45, 60, 60, 80, 85

- **Mean** = (45 + 60 + 60 + 80 + 85) / 5 = **66**
- **Median** = Middle value = **60**
- **Mode** = Value repeated most = **60**

C. Maths Marks

Data: 90, 80, 90, 90, 70

- **Mean** = (90 + 80 + 90 + 90 + 70) / 5 = **84**
- **Median** = Sorted: 70, 80, 90, 90, 90 → Middle value = **90**
- **Mode** = Most frequent = **90**

ii. Variance and Standard Deviation

Formulas:

- Variance = $\sum(x - \text{mean})^2 / (n - 1)$
- Standard Deviation = $\sqrt{\text{Variance}}$

A.Weight

- Mean = 6.1
- Squared deviations:
  $(4.2-6.1)^2 = 3.61$
  $(4.5-6.1)^2 = 2.56$
  $(4.7-6.1)^2 = 1.96$
  $(5.2-6.1)^2 = 0.81$
  $(6-6.1)^2 = 0.01$
  $(6.2-6.1)^2 = 0.01$
  $(7-6.1)^2 = 0.81$
  $(7.2-6.1)^2 = 1.21$
  $(7.5-6.1)^2 = 1.96$
  $(8.5-6.1)^2 = 5.76$
- Total = 19.7
- Variance = 19.7 / 9 = **2.1889**
- Standard Deviation = $\sqrt{2.1889}$ = **1.48**

B. English

- Mean = 66
- Squared deviations:
  $(45-66)^2 = 441$
  $(60-66)^2 = 36$
  $(60-66)^2 = 36$
  $(80-66)^2 = 196$
  $(85-66)^2 = 361$
- Total = 1070
- Variance = 1070 / 4 = **267.5**
- Standard Deviation = $\sqrt{267.5}$ = **16.36**

C. Maths

- Mean = 84
- Squared deviations:
  $(90-84)^2 = 36$
  $(80-84)^2 = 16$
  $(90-84)^2 = 36$
  $(90-84)^2 = 36$
  $(70-84)^2 = 196$
- Total = 320
- Variance = 320 / 4 = **80**
- Standard Deviation = $\sqrt{80}$ = **8.94**

---

iii. Five-Point Summary (Consists of: Minimum, Q1, Median, Q3, Maximum)

A.Weight

- Sorted Data: 4.2, 4.5, 4.7, 5.2, 6, 6.2, 7, 7.2, 7.5, 8.5
- Q1 = Median of lower half = 4.7
- Q3 = Median of upper half = 7.2
- **Five-Point Summary**: 4.2, 4.7, 6.1, 7.2, 8.5

B. English Marks

- Sorted: 45, 60, 60, 80, 85
- Q1 = (45 + 60) / 2 = 52.5
- Q3 = 80
- **Five-Point Summary**: 45, 52.5, 60, 80, 85

C. Maths Marks

- Sorted: 70, 80, 90, 90, 90
- Q1 = (70 + 80) / 2 = 75
- Q3 = (90 + 90) / 2 = 90
- **Five-Point Summary**: 70, 75, 90, 90, 90

iv. Interpretation of Skewness and Kurtosis

A.Weight

- Skewness ≈ 0.317 → Slightly **right-skewed**
- Kurtosis ≈ -0.857 → **Platykurtic** (flatter than normal distribution)

B. English

- Skewness ≈ 0.427 → Slightly **right-skewed**
- Kurtosis ≈ -1.483 → **Platykurtic**

C. Maths

- Skewness ≈ -1.342 → **Left-skewed**
- Kurtosis ≈ +1.000 → **Leptokurtic** (more peaked than normal)

---

v. Covariance between English and Hindi Marks

Formula:

$Cov(X, Y) = \sum(xi - \bar{x})(yi - \bar{y}) / (n - 1)$

Mean(English) = 66

Mean(Hindi) = (70.5 + 72.5 + 80 + 80 + 72) / 5 = **75**

Sid	xi (Eng)	yi (Hindi)	xi – 66	yi – 75	Product
1	45	70.5	-21	-4.5	94.5
2	60	72.5	-6	-2.5	15
3	60	80	-6	5	-30
4	80	80	14	5	70
5	85	72	19	-3	-57

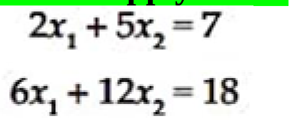
- Total Product = 92.5
- Covariance = 92.5 / (5 – 1) = **23.125**

Interpretation:

Covariance is positive, so there is a **direct relationship** between English and Hindi marks. When English marks increase, Hindi marks also tend to increase.

MODULE 2

Q1. Explain Gaussian Elimination Method and apply it on :



$$2x_1 + 5x_2 = 7$$

$$6x_1 + 12x_2 = 18$$

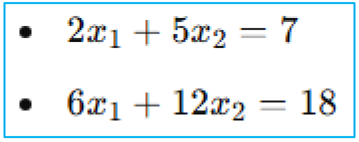
Gaussian Elimination Method:

Gaussian Elimination is a systematic method for solving systems of linear equations. It transforms the system's **augmented matrix** into **Row Echelon Form (REF)** using **elementary row operations**, then solves for variables using **back-substitution**.

Key Steps:

1. Form the augmented matrix from the equations.
2. Apply row operations to create zeros below the pivot (leading 1s).
3. Perform back-substitution to find variable values.

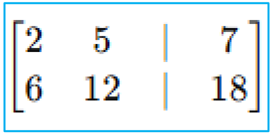
Given Equations:



- $2x_1 + 5x_2 = 7$
- $6x_1 + 12x_2 = 18$

Step-by-Step Solution:

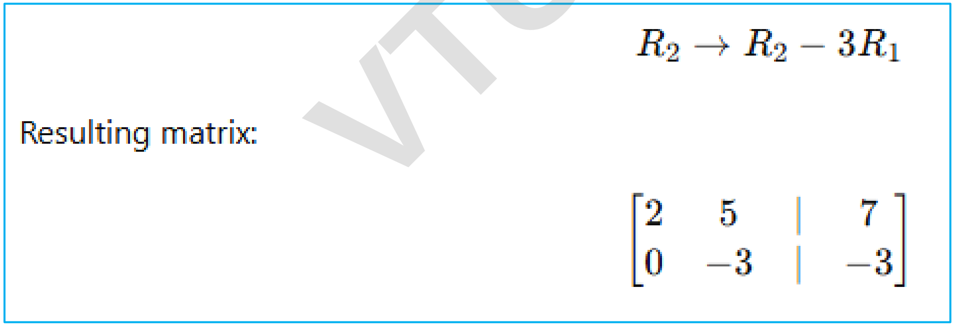
Step 1: Write the Augmented Matrix



$$\begin{bmatrix} 2 & 5 & | & 7 \\ 6 & 12 & | & 18 \end{bmatrix}$$

Step 2: Eliminate x1 from Equation 2

We want to create a **zero** below the first pivot (element in Row 1, Column 1).
Since 6÷2=3, perform the row operation:

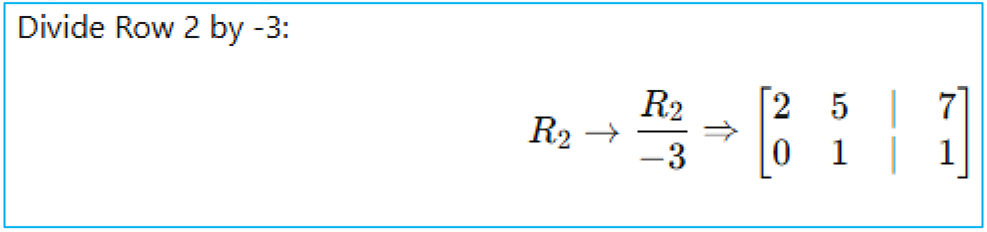


$$R_2 \to R_2 - 3R_1$$

Resulting matrix:

$$\begin{bmatrix} 2 & 5 & | & 7 \\ 0 & -3 & | & -3 \end{bmatrix}$$

Step 3: Make the pivot in Row 2 a 1



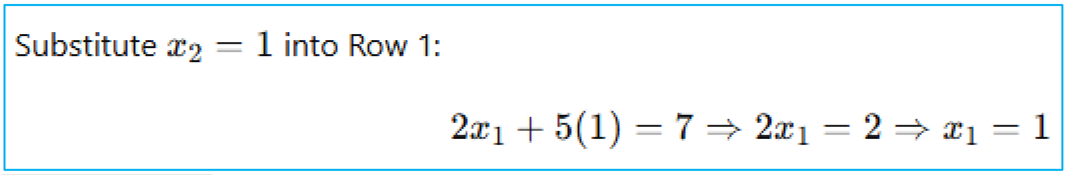
Divide Row 2 by -3:

$$R_2 \to \frac{R_2}{-3} \Rightarrow \begin{bmatrix} 2 & 5 & | & 7 \\ 0 & 1 & | & 1 \end{bmatrix}$$

Step 4: Back-Substitution

From Row 2:

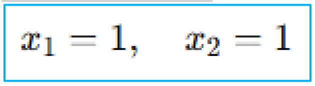
$$x_2 = 1$$



Substitute $x_2 = 1$ into Row 1:

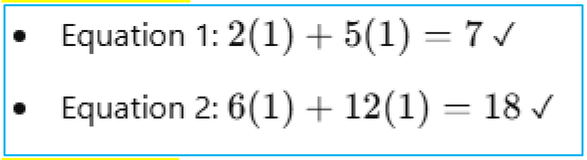
$$2x_1 + 5(1) = 7 \Rightarrow 2x_1 = 2 \Rightarrow x_1 = 1$$

Final Answer:



$$x_1 = 1, \quad x_2 = 1$$

Verification:

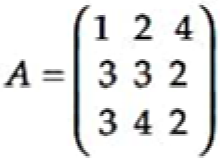


- Equation 1: $2(1) + 5(1) = 7 \checkmark$
- Equation 2: $6(1) + 12(1) = 18 \checkmark$

Conclusion:

- **Row operations** preserve the original solutions while simplifying the matrix.
- **Pivots (leading 1s)** help isolate variables during back-substitution.
- Always **verify the solution** by substituting back into original equations.

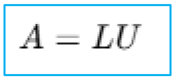
Q2. Apply LU decomposition for the given matrix:



$$A = \begin{pmatrix} 1 & 2 & 4 \\ 3 & 3 & 2 \\ 3 & 4 & 2 \end{pmatrix}$$

Goal:

We want to decompose matrix **A** into a product of two matrices:

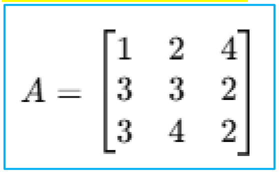


$$A = LU$$

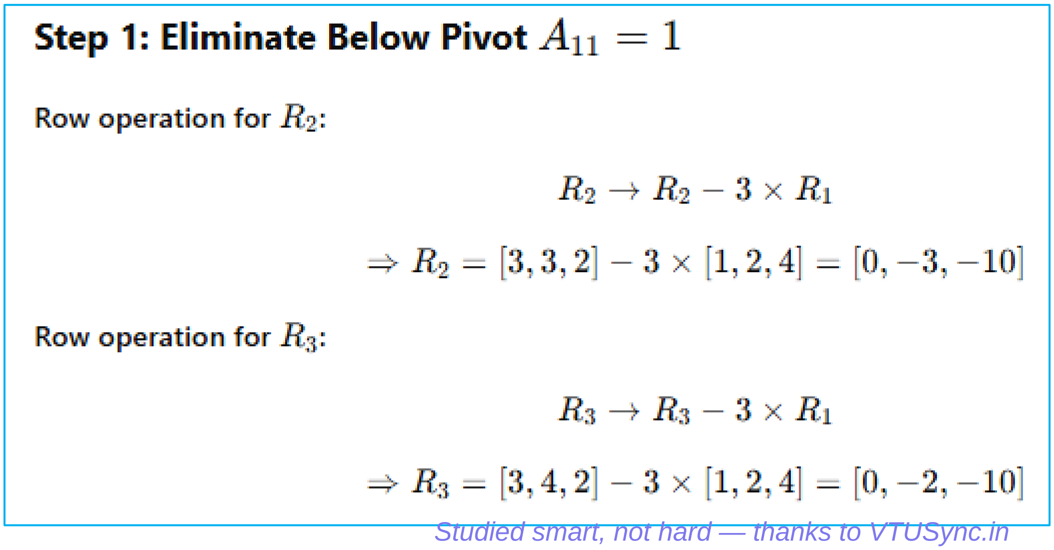
Where:

- **L** = Lower triangular matrix (1s on diagonal)
- **U** = Upper triangular matrix

Given Matrix A:



$$A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 3 & 2 \\ 3 & 4 & 2 \end{bmatrix}$$



Step 1: Eliminate Below Pivot $A_{11} = 1$

Row operation for R_2:

$$R_2 \rightarrow R_2 - 3 \times R_1$$

$$\Rightarrow R_2 = [3, 3, 2] - 3 \times [1, 2, 4] = [0, -3, -10]$$

Row operation for R_3:

$$R_3 \rightarrow R_3 - 3 \times R_1$$

$$\Rightarrow R_3 = [3, 4, 2] - 3 \times [1, 2, 4] = [0, -2, -10]$$

Matrix after step 1:

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 0 & -2 & -10 \end{bmatrix}$$

Step 2: Eliminate Below Pivot $A_{22} = -3$

Row operation for R_3:

$$\text{Multiplier: } l_{32} = \frac{-2}{-3} = \frac{2}{3}$$

$$R_3 \rightarrow R_3 - \frac{2}{3} \times R_2$$

Calculations:

- Second element: $-2 - \frac{2}{3} \times (-3) = -2 + 2 = 0$
- Third element: $-10 - \frac{2}{3} \times (-10) = -10 + \frac{20}{3} = \frac{-30+20}{3} = \frac{-10}{3}$

Final Upper Triangular Matrix U:

$$U = \begin{bmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 0 & 0 & -\frac{10}{3} \end{bmatrix}$$

Step 3: Constructing Lower Triangular Matrix L

Multipliers used:

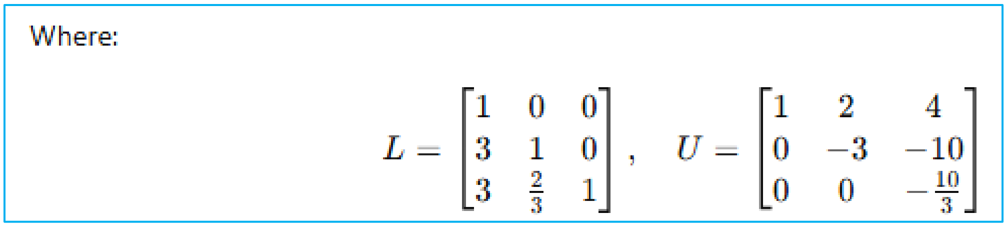
- $l_{21} = 3$
- $l_{31} = 3$
- $l_{32} = \frac{2}{3}$

Matrix L:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & \frac{2}{3} & 1 \end{bmatrix}$$

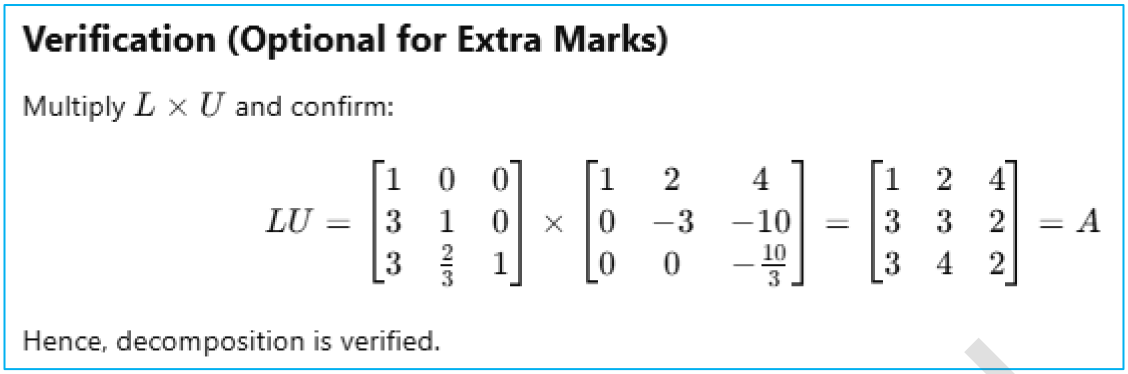
Final LU Decomposition Result

$$A = LU$$



Where:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & \frac{2}{3} & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 0 & 0 & -\frac{10}{3} \end{bmatrix}$$



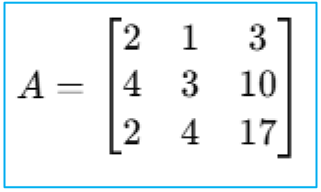
Verification (Optional for Extra Marks)

Multiply $L \times U$ and confirm:

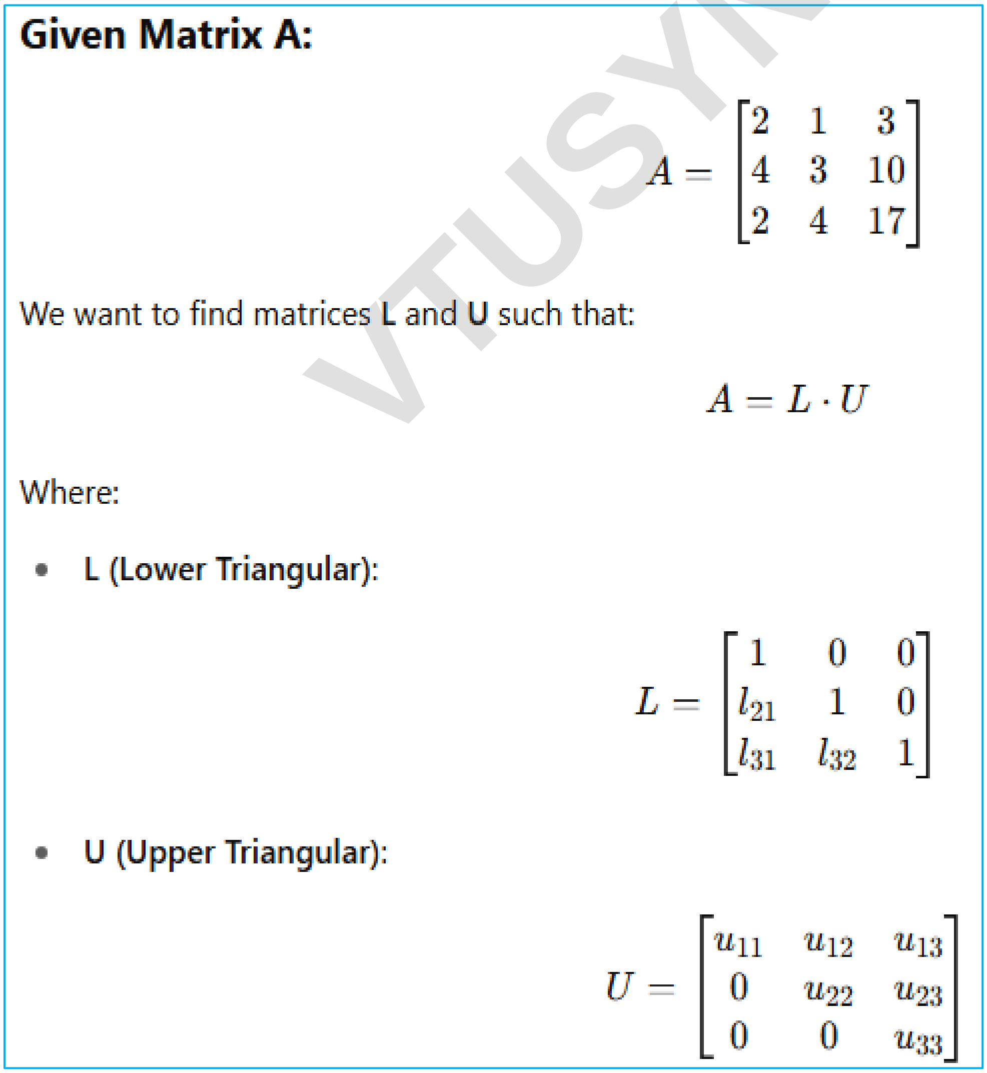
$$LU = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & \frac{2}{3} & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 0 & 0 & -\frac{10}{3} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 3 & 2 \\ 3 & 4 & 2 \end{bmatrix} = A$$

Hence, decomposition is verified.

Q3. Apply LU decomposition for the given matrix:



$$A = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & 10 \\ 2 & 4 & 17 \end{bmatrix}$$



Given Matrix A:

$$A = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & 10 \\ 2 & 4 & 17 \end{bmatrix}$$

We want to find matrices **L** and **U** such that:

$$A = L \cdot U$$

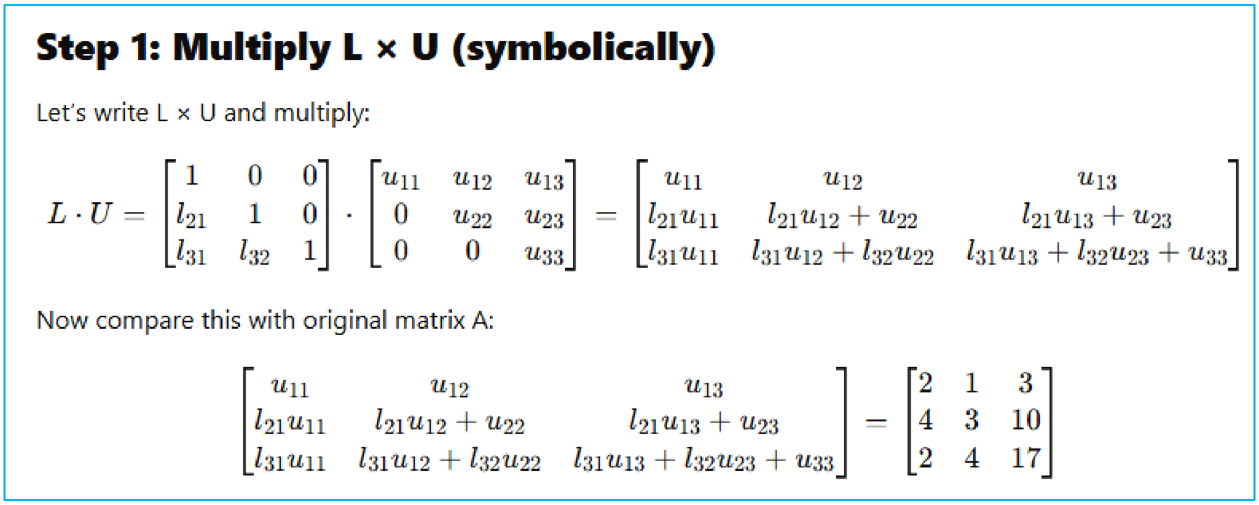
Where:

- **L (Lower Triangular):**

$$L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix}$$

- **U (Upper Triangular):**

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$



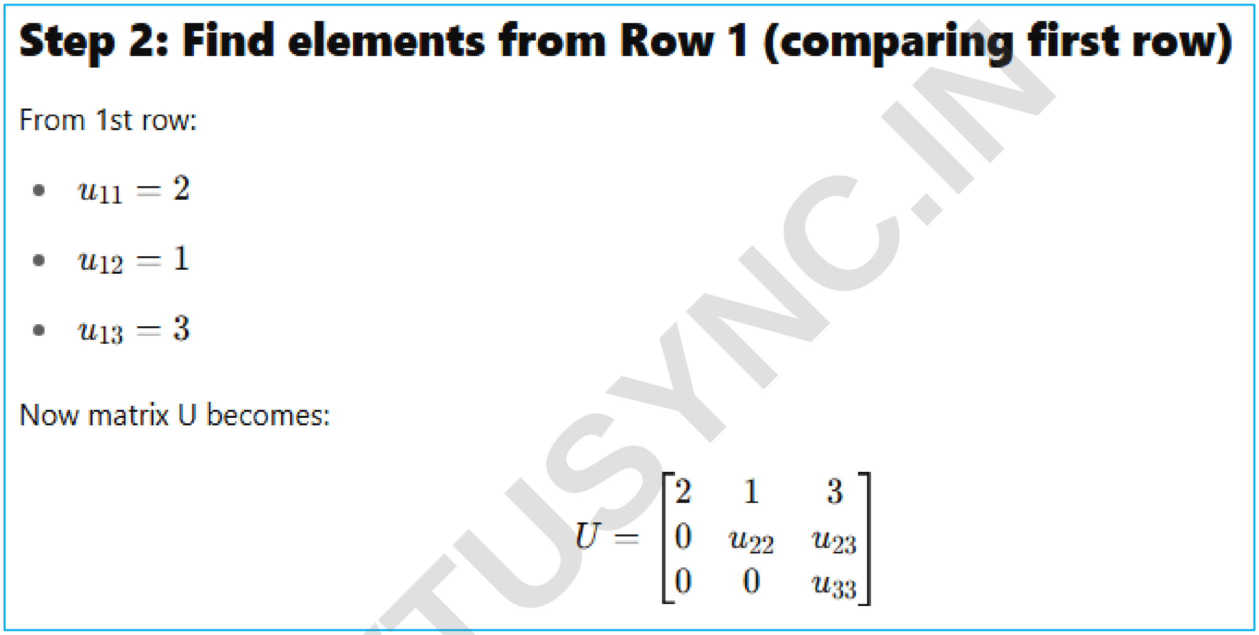
Step 1: Multiply L × U (symbolically)

Let's write L × U and multiply:

$$L \cdot U = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \cdot \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ l_{21}u_{11} & l_{21}u_{12} + u_{22} & l_{21}u_{13} + u_{23} \\ l_{31}u_{11} & l_{31}u_{12} + l_{32}u_{22} & l_{31}u_{13} + l_{32}u_{23} + u_{33} \end{bmatrix}$$

Now compare this with original matrix A:

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ l_{21}u_{11} & l_{21}u_{12} + u_{22} & l_{21}u_{13} + u_{23} \\ l_{31}u_{11} & l_{31}u_{12} + l_{32}u_{22} & l_{31}u_{13} + l_{32}u_{23} + u_{33} \end{bmatrix} = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & 10 \\ 2 & 4 & 17 \end{bmatrix}$$



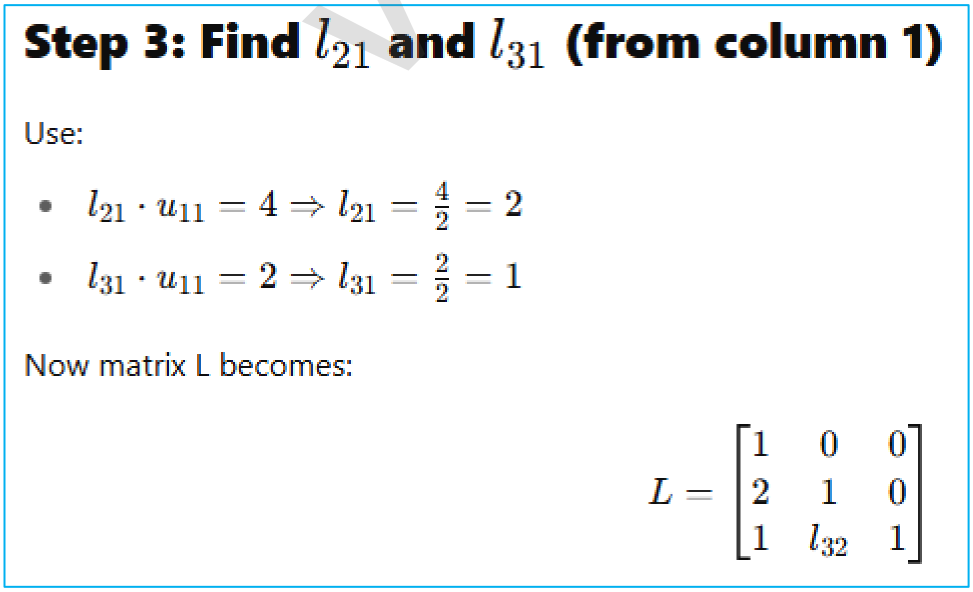
Step 2: Find elements from Row 1 (comparing first row)

From 1st row:

- $u_{11} = 2$
- $u_{12} = 1$
- $u_{13} = 3$

Now matrix U becomes:

$$U = \begin{bmatrix} 2 & 1 & 3 \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$



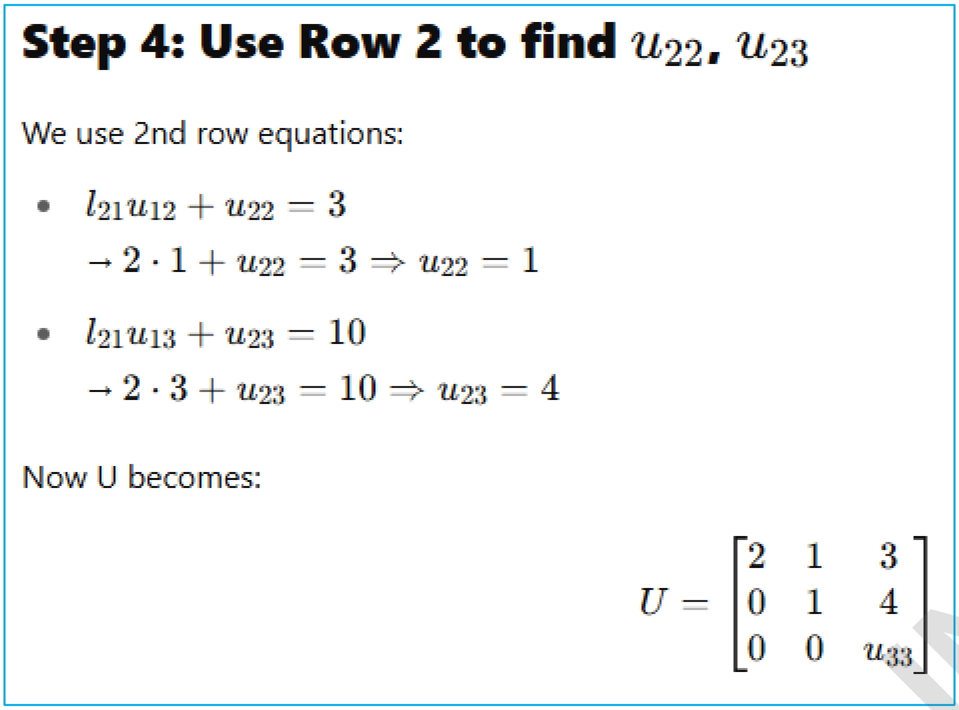
Step 3: Find l_{21} and l_{31} (from column 1)

Use:

- $l_{21} \cdot u_{11} = 4 \Rightarrow l_{21} = \frac{4}{2} = 2$
- $l_{31} \cdot u_{11} = 2 \Rightarrow l_{31} = \frac{2}{2} = 1$

Now matrix L becomes:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & l_{32} & 1 \end{bmatrix}$$



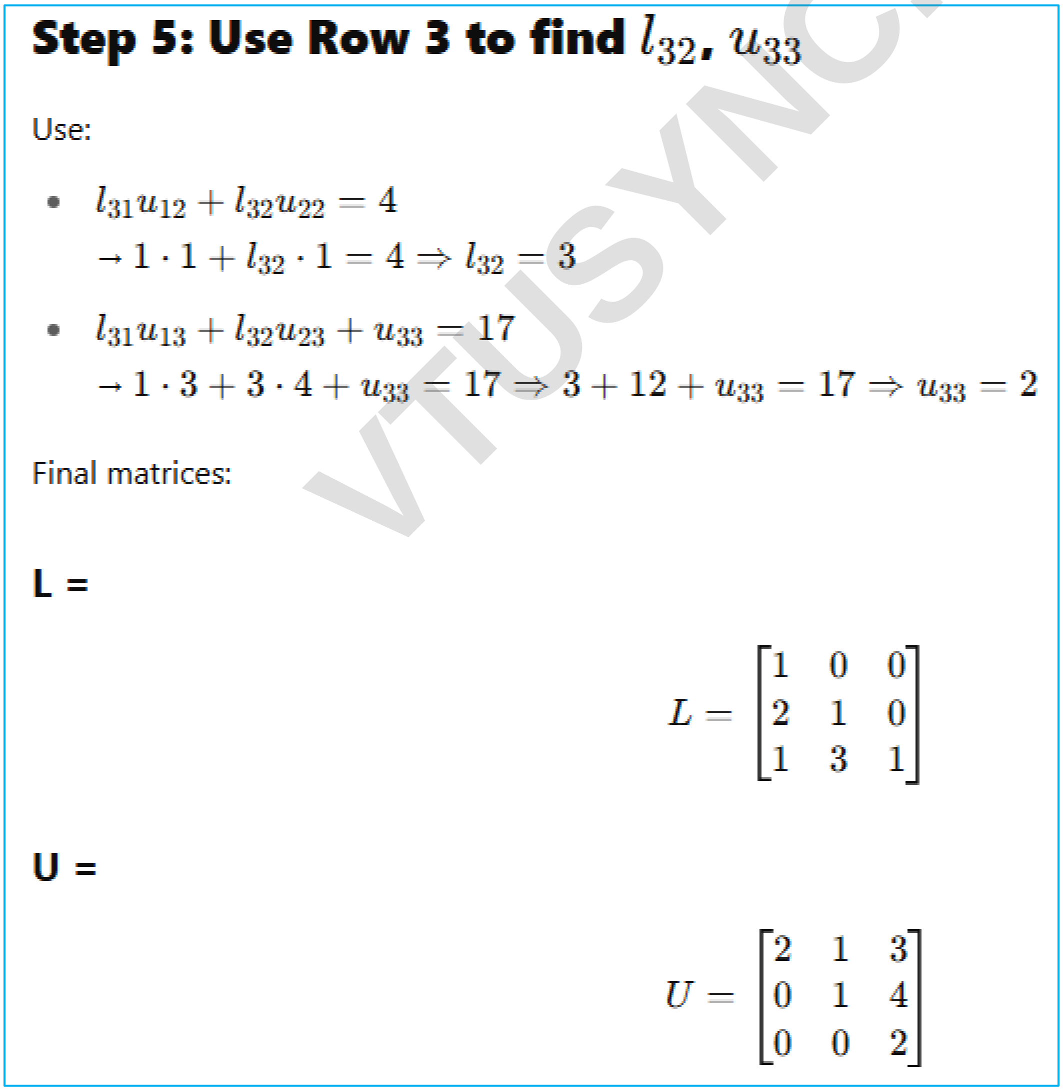
Step 4: Use Row 2 to find u_{22}, u_{23}

We use 2nd row equations:

- $l_{21}u_{12} + u_{22} = 3$
  $\rightarrow 2 \cdot 1 + u_{22} = 3 \Rightarrow u_{22} = 1$
- $l_{21}u_{13} + u_{23} = 10$
  $\rightarrow 2 \cdot 3 + u_{23} = 10 \Rightarrow u_{23} = 4$

Now U becomes:

$$U = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & u_{33} \end{bmatrix}$$



Step 5: Use Row 3 to find l_{32}, u_{33}

Use:

- $l_{31}u_{12} + l_{32}u_{22} = 4$
  $\rightarrow 1 \cdot 1 + l_{32} \cdot 1 = 4 \Rightarrow l_{32} = 3$
- $l_{31}u_{13} + l_{32}u_{23} + u_{33} = 17$
  $\rightarrow 1 \cdot 3 + 3 \cdot 4 + u_{33} = 17 \Rightarrow 3 + 12 + u_{33} = 17 \Rightarrow u_{33} = 2$

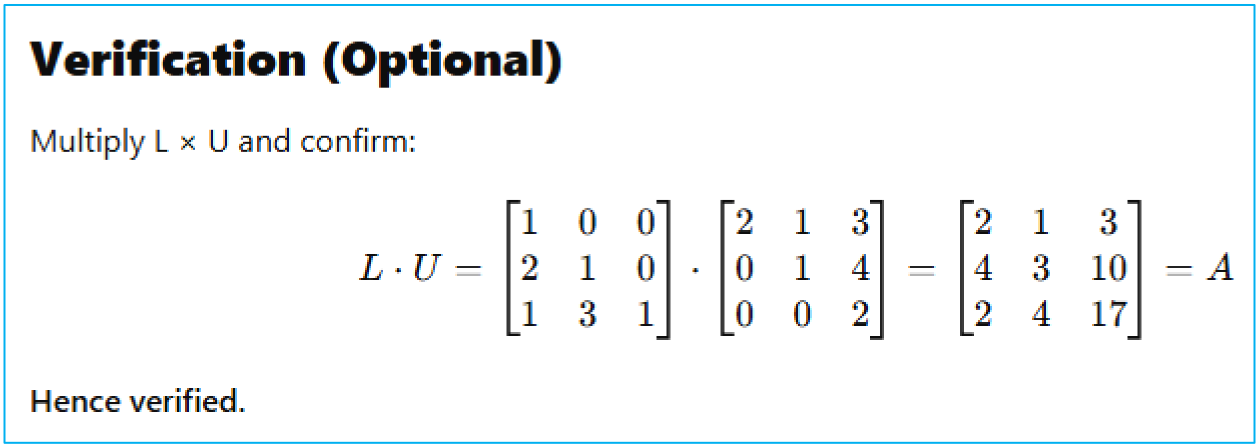
Final matrices:

L =

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 3 & 1 \end{bmatrix}$$

U =

$$U = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 2 \end{bmatrix}$$



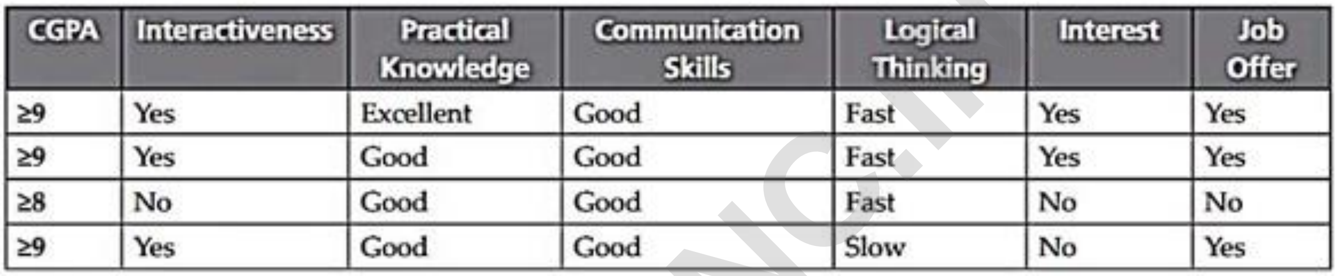
Verification (Optional)

Multiply L × U and confirm:

$$L \cdot U = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 3 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 1 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & 10 \\ 2 & 4 & 17 \end{bmatrix} = A$$

Hence verified.

Q4. Explain and apply candidate elimination algorithm for the given dataset



CGPA	Interactiveness	Practical Knowledge	Communication Skills	Logical Thinking	Interest	Job Offer
≥9	Yes	Excellent	Good	Fast	Yes	Yes
≥9	Yes	Good	Good	Fast	Yes	Yes
≥8	No	Good	Good	Fast	No	No
≥9	Yes	Good	Good	Slow	No	Yes

Step 1: Understanding the Dataset

CGPA	Interactiveness	Practical Knowledge	Communication	Logical Thinking	Interest	Job Offer
≥9	Yes	Excellent	Good	Fast	Yes	**Yes**
≥9	Yes	Good	Good	Fast	Yes	**Yes**
≥8	No	Good	Good	Fast	No	**No**
≥9	Yes	Good	Good	Slow	No	**Yes**

Step 2: Initial Hypotheses

- **S** = Most **Specific Hypothesis**
  Start with the first positive example:
  ```
  ini

  S = [≥9, Yes, Excellent, Good, Fast, Yes]
  ```
- **G** = Most **General Hypothesis**
  ```
    G = [?, ?, ?, ?, ?, ?]
  ```

Step 3: Go Through Each Example

Example 1 → Positive

Already used to initialize `S`.

Example 2 → Positive

```
 [≥9, Yes, Good, Good, Fast, Yes]
```

Compare it with current `S = [≥9, Yes, Excellent, Good, Fast, Yes]`

→ Difference at Practical Knowledge (`Excellent` ≠ `Good`)

So we generalize S:

```
S = [≥9, Yes, ?, Good, Fast, Yes]
```

G remains `[?, ?, ?, ?, ?, ?]` (still matches positive)

Example 3 → Negative

`[≥8, No, Good, Good, Fast, No] → No`

Now check which hypotheses in G **match this negative**:

- G has `[?, ?, ?, ?, ?, ?]` → matches everything → □ must be specialized!

Specialize G using current S:

Current S = `[≥9, Yes, ?, Good, Fast, Yes]`

Negative = `[≥8, No, Good, Good, Fast, No]`

Let's compare and specialize each differing attribute:

Attribute	S	Negative	Specialization
CGPA	≥9	≥8	[≥9, ?, ?, ?, ?, ?]
Interactiveness	Yes	No	[?, Yes, ?, ?, ?, ?]
Interest	Yes	No	[?, ?, ?, ?, ?, Yes]

Add these to G:

```
G = {
  [≥9, ?, ?, ?, ?, ?],
  [?, Yes, ?, ?, ?, ?],
  [?, ?, ?, ?, ?, Yes]
}
```

Now **filter G**: Remove those that are **too specific** or **don't match earlier positives**.

□ All 3 are consistent with the positive examples → keep them.

□ Example 4 → Positive

`[≥9, Yes, Good, Good, Slow, No]`

Compare with current `S = [≥9, Yes, ?, Good, Fast, Yes]`

Only **Logical Thinking** and **Interest** differ:

- `Fast` ≠ `Slow` → generalize `Logical Thinking` to `?`
- `Yes` ≠ `No` → generalize `Interest` to `?`

Updated S:

`S = [≥9, Yes, ?, Good, ?, ?]`

Now check each hypothesis in G:

- [≥9, ?, ?, ?, ?, ?] → matches this → keep
- [?, Yes, ?, ?, ?, ?] → matches → keep
- [?, ?, ?, ?, ?, Yes] → doesn't match because this example has `Interest = No` → □ Remove this one

Updated G:

```
G = {
  [≥9, ?, ?, ?, ?, ?],
  [?, Yes, ?, ?, ?, ?]
}
```

Final Hypothesis Version

Hypothesis	Final Value
S	[≥9, Yes, ?, Good, ?, ?]
G	{[≥9, ?, ?, ?, ?, ?], [?, Yes, ?, ?, ?, ?]}

Final Summary

Step	S (Specific)	G (General)
Init	First positive	[? ? ? ? ? ?]
E2 (Yes)	[≥9, Yes, ?, Good, Fast, Yes]	same
E3 (No)	same	[≥9, ?, ?, ?, ?, ?], [?, Yes, ?, ?, ?, ?], [?, ?, ?, ?, ?, Yes]
E4 (Yes)	[≥9, Yes, ?, Good, ?, ?]	[≥9, ?, ?, ?, ?, ?], [?, Yes, ?, ?, ?, ?]

Q5. Apply Candidate Elimination on this dataset

Sky	Temp	Humidity	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Step 0: Initialize

- **S (Specific Hypothesis)** = Most specific
  `S = [∅, ∅, ∅, ∅, ∅, ∅]`
- **G (General Hypothesis)** = Most general
  `G = [?, ?, ?, ?, ?, ?]`

Step 1: First Example → Yes

Example: `Sunny, Warm, Normal, Strong, Warm, Same`

- Update S to match this example:
  `S = [Sunny, Warm, Normal, Strong, Warm, Same]`
- G remains unchanged:
  `G = [?, ?, ?, ?, ?, ?]`

Step 2: Second Example → Yes

Example: `Sunny, Warm, High, Strong, Warm, Same`

Compare with S = `[Sunny, Warm, Normal, Strong, Warm, Same]`

- Only difference: **Humidity**
- So replace Humidity with `?` in S

New S:

`S = [Sunny, Warm, ?, Strong, Warm, Same]`

G remains unchanged.

Step 3: Third Example → No

Example: `Rainy, Cold, High, Strong, Warm, Change`

(Negative example)

- S remains unchanged.
- Specialize G to eliminate this negative example.

Current G:

`G = [?, ?, ?, ?, ?, ?]` (matches everything → too general)

Use S to specialize G:

S = [Sunny, Warm, ?, Strong, Warm, Same]

Specializations:

1. [Sunny, ?, ?, ?, ?, ?]
2. [?, Warm, ?, ?, ?, ?]
3. [?, ?, ?, Strong, ?, ?]
4. [?, ?, ?, ?, Warm, ?]
5. [?, ?, ?, ?, ?, Same]

Now remove any that still match the negative example:

Negative = [Rainy, Cold, High, Strong, Warm, Change]

Check each:

- [Sunny, ?, ?, ?, ?, ?] → □ Valid (doesn't match)
- [?, Warm, ?, ?, ?, ?] → □ Valid (doesn't match)
- [?, ?, ?, Strong, ?, ?] → □ Matches →**REMOVE**
- [?, ?, ?, ?, Warm, ?] → □ Matches →**REMOVE**
- [?, ?, ?, ?, ?, Same] → □ Valid (doesn't match)

New G set:

1. [Sunny, ?, ?, ?, ?, ?]
2. [?, Warm, ?, ?, ?, ?]
3. [?, ?, ?, ?, ?, Same]

Step 4: Fourth Example → Yes

Example: Sunny, Warm, High, Strong, Cool, Change

Current S = [Sunny, Warm, ?, Strong, Warm, Same]

Compare:

- Sky = Sunny → ok
- Temp = Warm → ok
- Humidity = High → ok
- Wind = Strong → ok
- Water = Cool ≠ Warm → change to ?
- Forecast = Change ≠ Same → change to ?

New S:

S = [Sunny, Warm, ?, Strong, ?, ?]

Now update G:

Remove any general hypotheses that **don't match this example**:

Current G:

1. [Sunny, ?, ?, ?, ?, ?] → □ Matches
2. [?, Warm, ?, ?, ?, ?] → □ Matches
3. [?, ?, ?, ?, ?, Same] → □ Doesn't match →**REMOVE**

Final G:

1. [Sunny, ?, ?, ?, ?, ?]
2. [?, Warm, ?, ?, ?, ?]

Final Answer:

- **Specific Boundary (S):**
  S = [Sunny, Warm, ?, Strong, ?, ?]

- **General Boundary (G):**
  `G = [Sunny, ?, ?, ?, ?, ?]`
  `G = [?, Warm, ?, ?, ?, ?]`

What it means:

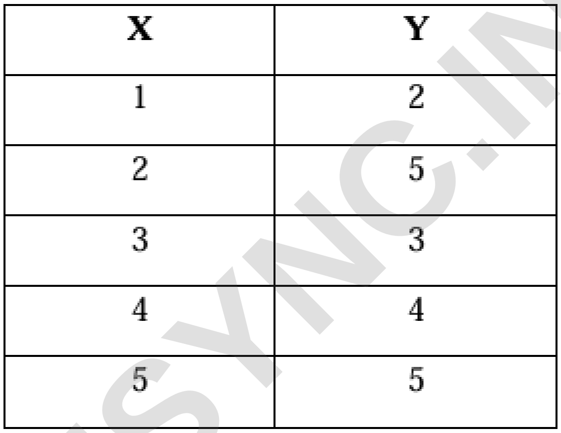
Your learning system has now narrowed down consistent hypotheses.
It predicts `EnjoySport = Yes` when:

- Either **Sky is Sunny**, or
- **Temp is Warm**

And the most specific case includes:
Sunny, **Warm**, and **Strong** conditions as common features.

MODULE 3

1Q. The values of independent variable x and dependent value y are given as:



X	Y
1	2
2	5
3	3
4	4
5	5

Apply Linear Regression and evaluate the value of y when x=6 and x=9

Soln: Given:

x	y
1	2
2	5
3	3
4	4
5	5

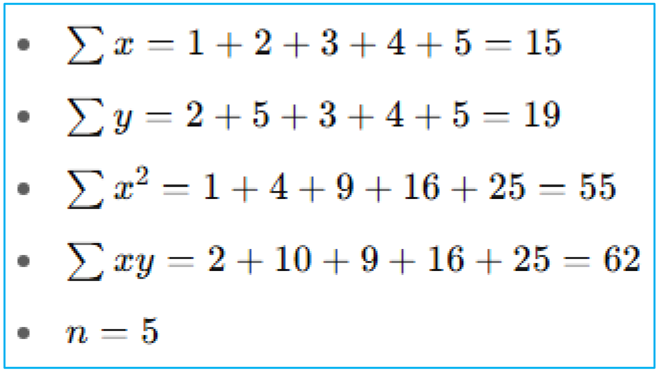
We are to:

1. Apply **linear regression**
2. Find the **regression line equation**: y=a + bx
3. Use it to **predict y** when x=6 and x=9

Step 1: Create Table for Calculations

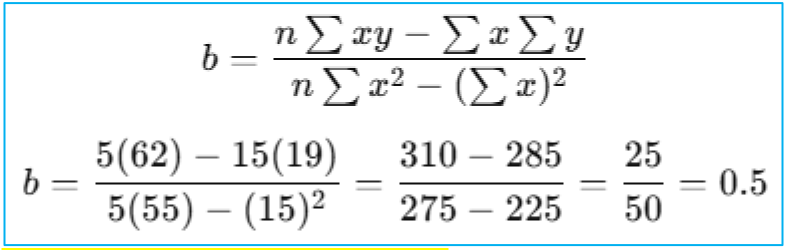
x	y	x^2	xy
1	2	1	2
2	5	4	10
3	3	9	9
4	4	16	16
5	5	25	25

Now compute the totals:



- $\sum x = 1 + 2 + 3 + 4 + 5 = 15$
- $\sum y = 2 + 5 + 3 + 4 + 5 = 19$
- $\sum x^2 = 1 + 4 + 9 + 16 + 25 = 55$
- $\sum xy = 2 + 10 + 9 + 16 + 25 = 62$
- $n = 5$

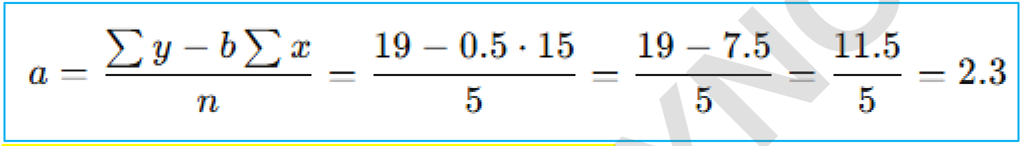
Step 2: Calculate Slope (b)



$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

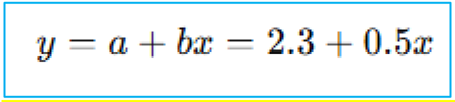
$$b = \frac{5(62) - 15(19)}{5(55) - (15)^2} = \frac{310 - 285}{275 - 225} = \frac{25}{50} = 0.5$$

Step 3: Calculate Intercept (a)



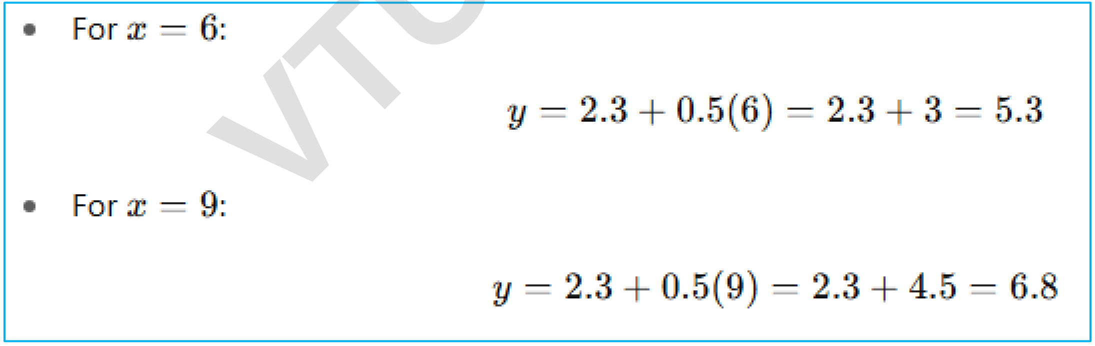
$$a = \frac{\sum y - b\sum x}{n} = \frac{19 - 0.5 \cdot 15}{5} = \frac{19 - 7.5}{5} = \frac{11.5}{5} = 2.3$$

Step 4: Final Linear Regression Equation



$$y = a + bx = 2.3 + 0.5x$$

Step 5: Predict y when x = 6 and x = 9



- For $x = 6$:

$$y = 2.3 + 0.5(6) = 2.3 + 3 = 5.3$$

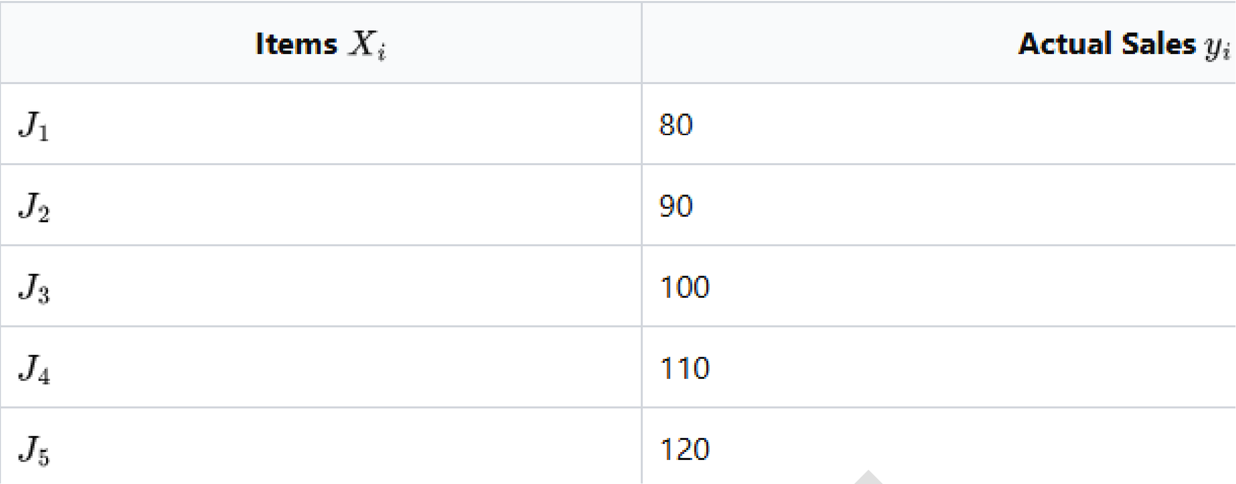
- For $x = 9$:

$$y = 2.3 + 0.5(9) = 2.3 + 4.5 = 6.8$$

Final Answers:

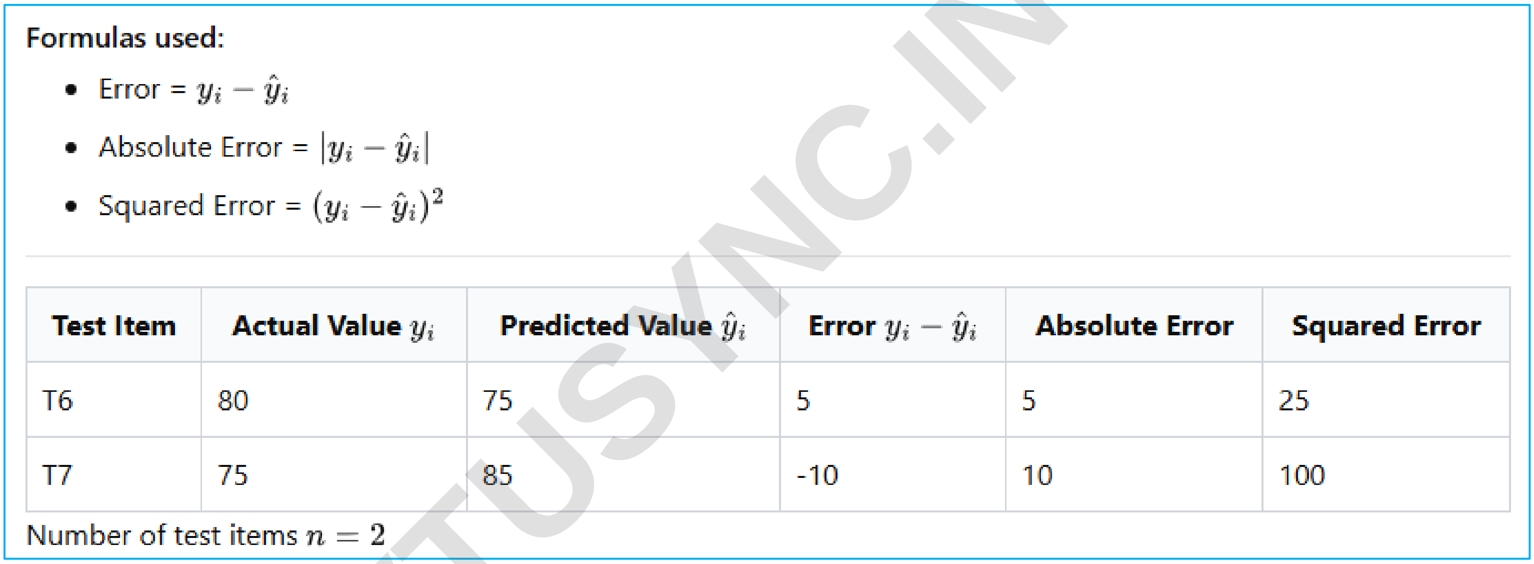
- **Regression Equation**: y=2.3+0.5x
- **When x = 6, y = 5.3**
- **When x = 9, y = 6.8**

2Q. Consider the following dataset for predicting the sales of items:



Items X_i	Actual Sales y_i
J_1	80
J_2	90
J_3	100
J_4	110
J_5	120

Ans:

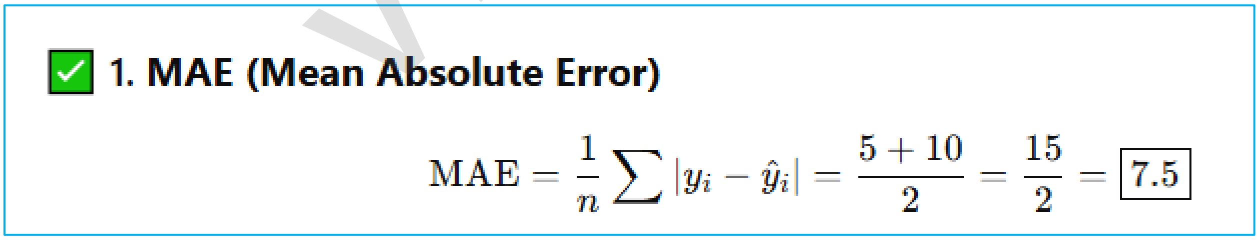


Formulas used:

- Error = $y_i - \hat{y}_i$
- Absolute Error = $|y_i - \hat{y}_i|$
- Squared Error = $(y_i - \hat{y}_i)^2$

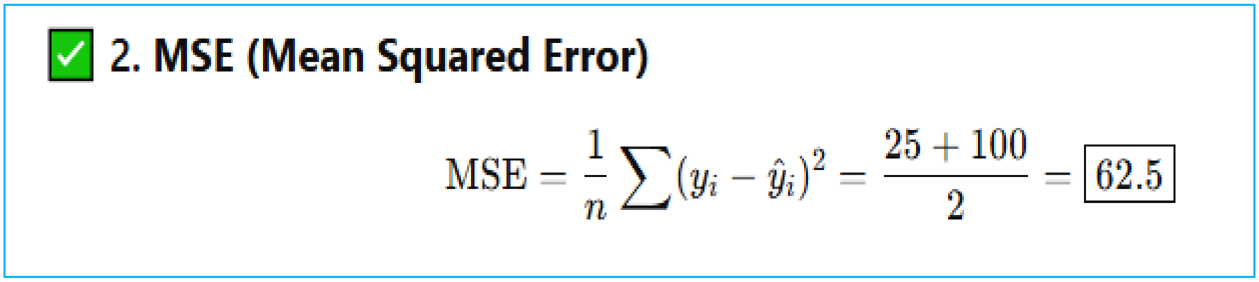
Test Item	Actual Value y_i	Predicted Value $\hat{y}_i$	Error $y_i - \hat{y}_i$	Absolute Error	Squared Error
T6	80	75	5	5	25
T7	75	85	-10	10	100

Number of test items $n = 2$



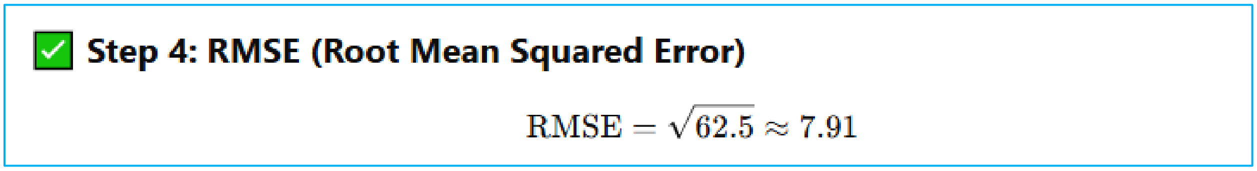
✅ 1. MAE (Mean Absolute Error)

$$\text{MAE} = \frac{1}{n}\sum|y_i - \hat{y}_i| = \frac{5+10}{2} = \frac{15}{2} = \boxed{7.5}$$



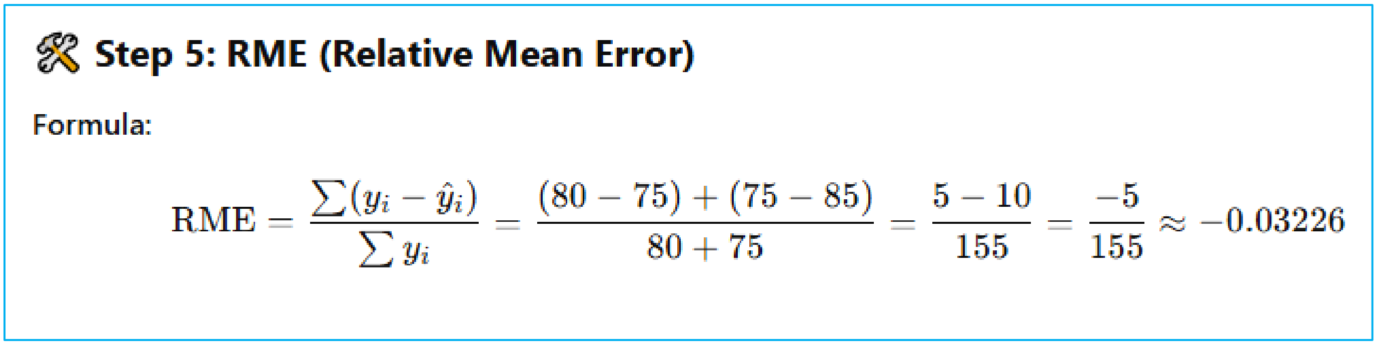
✅ 2. MSE (Mean Squared Error)

$$\text{MSE} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 = \frac{25+100}{2} = \boxed{62.5}$$



✅ Step 4: RMSE (Root Mean Squared Error)

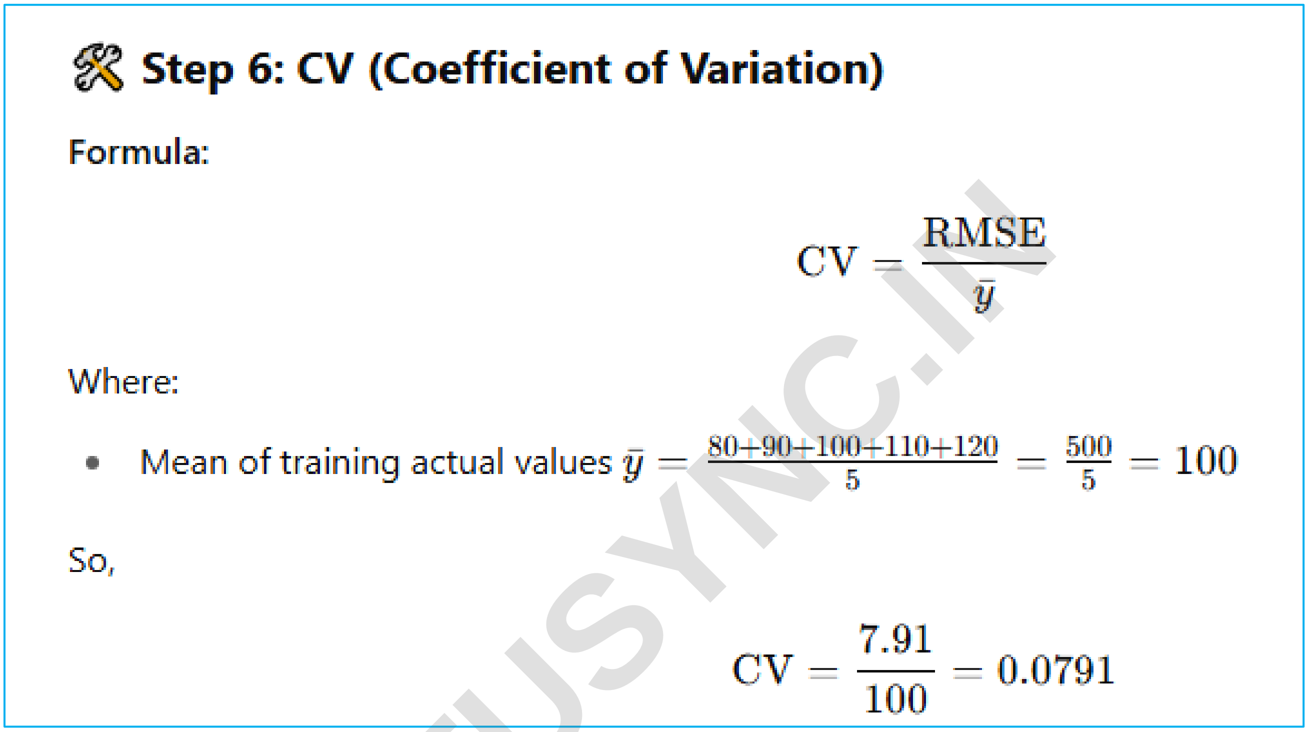
$$\text{RMSE} = \sqrt{62.5} \approx 7.91$$



🛠️ Step 5: RME (Relative Mean Error)

Formula:

$$\text{RME} = \frac{\sum(y_i - \hat{y}_i)}{\sum y_i} = \frac{(80-75)+(75-85)}{80+75} = \frac{5-10}{155} = \frac{-5}{155} \approx -0.03226$$



🛠️ Step 6: CV (Coefficient of Variation)

Formula:

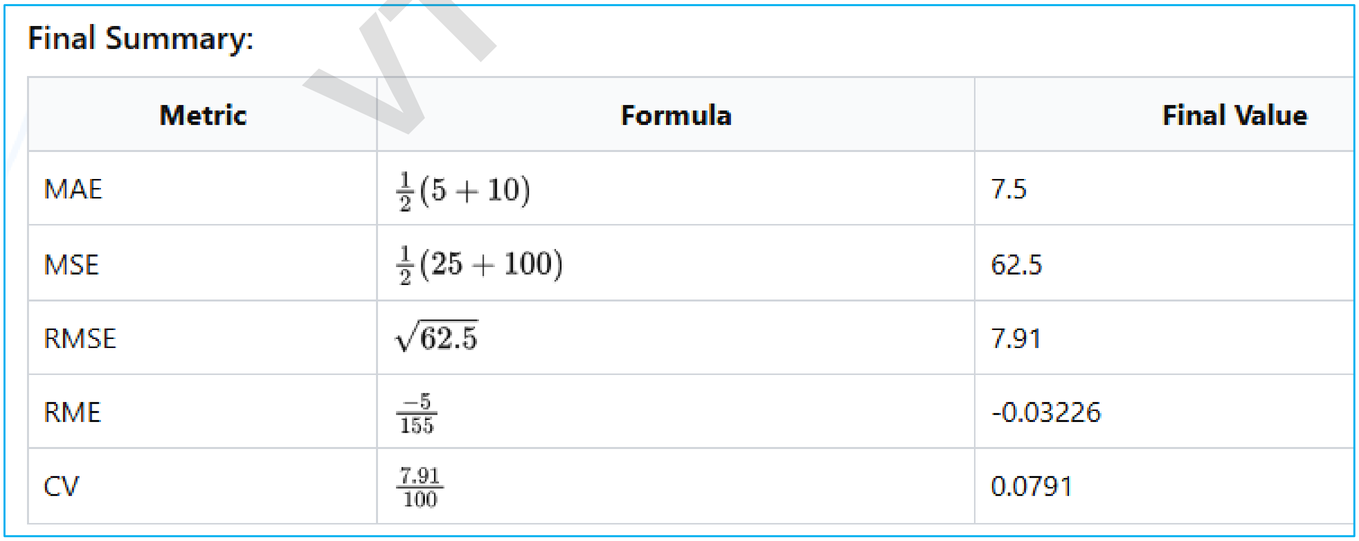
$$\text{CV} = \frac{\text{RMSE}}{\bar{y}}$$

Where:

- Mean of training actual values $\bar{y} = \frac{80+90+100+110+120}{5} = \frac{500}{5} = 100$

So,

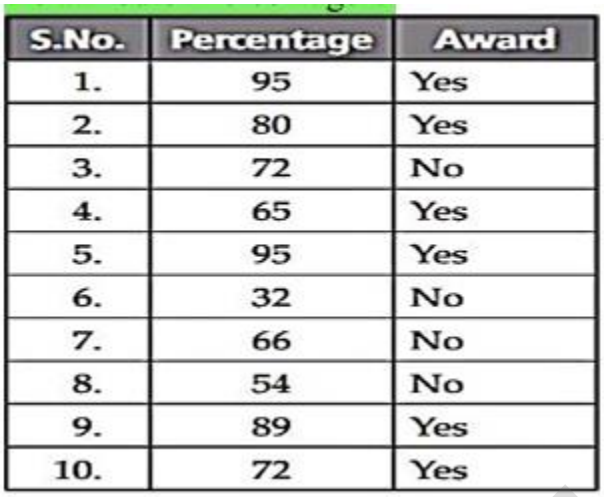
$$\text{CV} = \frac{7.91}{100} = 0.0791$$



Final Summary:

Metric	Formula	Final Value
MAE	$\frac{1}{2}(5+10)$	7.5
MSE	$\frac{1}{2}(25+100)$	62.5
RMSE	$\sqrt{62.5}$	7.91
RME	$\frac{-5}{155}$	-0.03226
CV	$\frac{7.91}{100}$	0.0791

3Q. How are continuous attributes discretized? Consider the training dataset and discretize the attribute “Percentage”:



S.No.	Percentage	Award
1.	95	Yes
2.	80	Yes
3.	72	No
4.	65	Yes
5.	95	Yes
6.	32	No
7.	66	No
8.	54	No
9.	89	Yes
10.	72	Yes

Ans:

Simple Explanation: What is Discretization?

In real life, many values like **temperature**, **percentage**, or **salary** are **continuous** — meaning they can take any value like 65.5, 72.8, etc.

But some **machine learning algorithms** (like decision trees) work better if the values are **discrete** — like categories: **High**, **Medium**, **Low**.

So, **Discretization** means:

Converting continuous values into categories or ranges.

Techniques to Discretize Continuous Attributes

There are 3 main methods:

1. **Equal Width Binning** → divide the full range into equal-size intervals
2. **Equal Frequency Binning** → each bin gets almost same number of data points
3. **Supervised Discretization** → based on class labels (like Yes/No)

In this question, we’ll use **Supervised Discretization** (like in Decision Tree ID3) — since we have a target label: **Award (Yes/No)**.

Given Dataset:

S.No	Percentage	Award
1	95	Yes
2	80	Yes
3	72	No
4	65	Yes
5	95	Yes
6	32	No
7	66	No
8	54	No
9	89	Yes
10	72	Yes

Goal:
Discretize the “Percentage” attribute into intervals **based on class label**.
We'll follow **Entropy-based discretization** (used in decision trees like ID3).

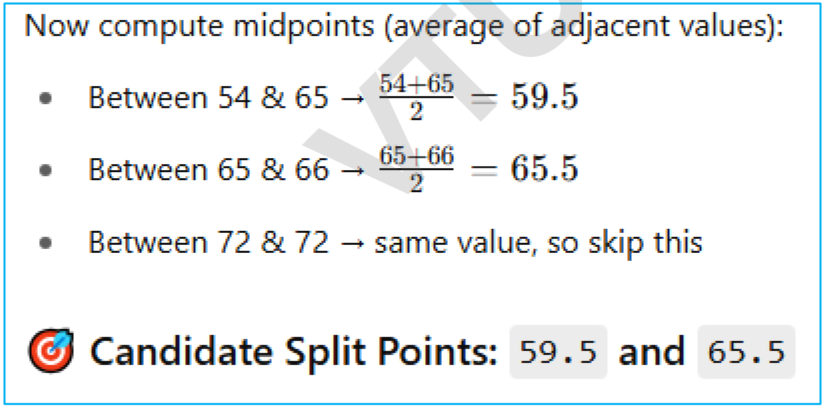
Step 1: Sort the data by Percentage

%	Award
32	No
54	No
65	Yes
66	No
72	No
72	Yes
80	Yes
89	Yes
95	Yes
95	Yes

Step 2: Find candidate split points

We only consider a **split between rows where class changes**. So we check where `Award` changes from No → Yes or Yes → No.
Let’s check pairs:

- Between 54 (No) & 65 (Yes) → Yes changes → split
- Between 65 (Yes) & 66 (No) → No changes → split
- Between 72 (No) & 72 (Yes) → Yes changes → split



Now compute midpoints (average of adjacent values):

- Between 54 & 65 → $\frac{54+65}{2} = 59.5$
- Between 65 & 66 → $\frac{65+66}{2} = 65.5$
- Between 72 & 72 → same value, so skip this

🎯 **Candidate Split Points:** `59.5` **and** `65.5`

Step 3: Choose the best split using Information Gain (Entropy reduction)

We calculate **entropy before and after split**, and pick the one with highest **Information Gain**.
□ **Entropy Formula:**

$$Entropy(S) = -p_{yes}\log_2(p_{yes}) - p_{no}\log_2(p_{no})$$

Entropy of Full Dataset (10 records):

- 6 Yes, 4 No

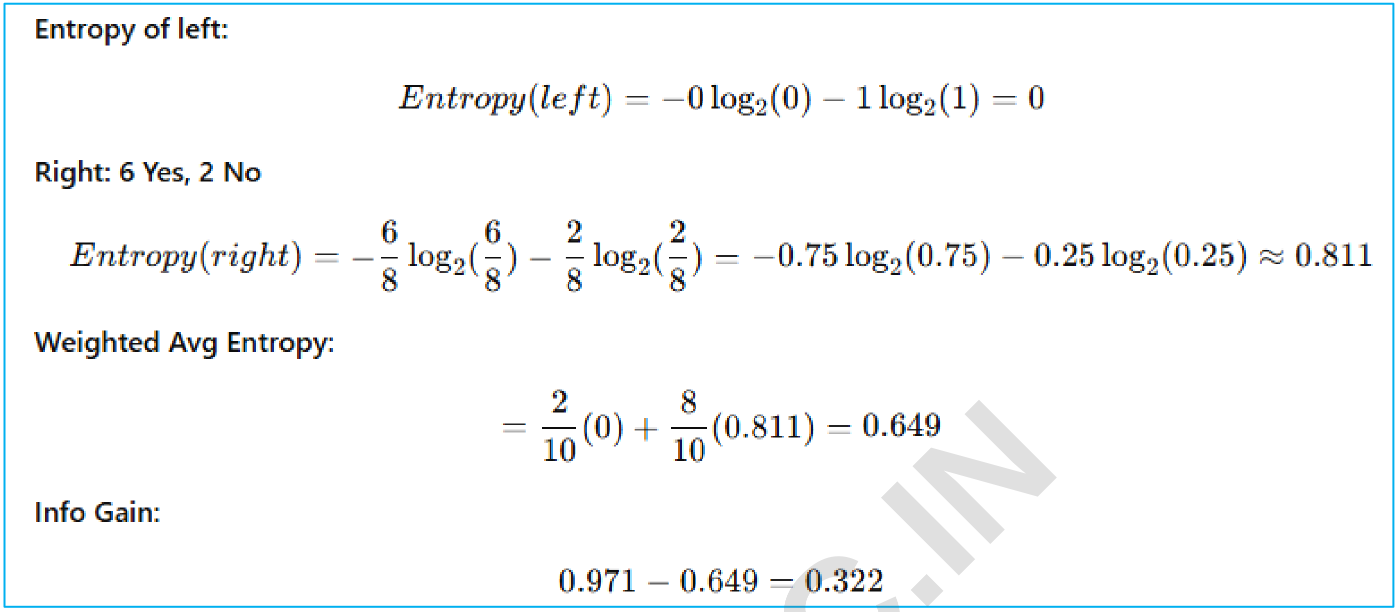
$$Entropy(S) = -\frac{6}{10}\log_2(\frac{6}{10}) - \frac{4}{10}\log_2(\frac{4}{10}) = -0.6\log_2(0.6) - 0.4\log_2(0.4) \approx 0.971$$

Now we check Entropy after split at **59.5** and **65.5**

Try Split at 59.5

Split into:

- **Left (<=59.5):** 32 (No), 54 (No) → 2 records → 0 Yes, 2 No
- **Right (>59.5):** 65,66,72,72,80,89,95,95 → 8 records



Entropy of left:

$$Entropy(left) = -0\log_2(0) - 1\log_2(1) = 0$$

Right: 6 Yes, 2 No

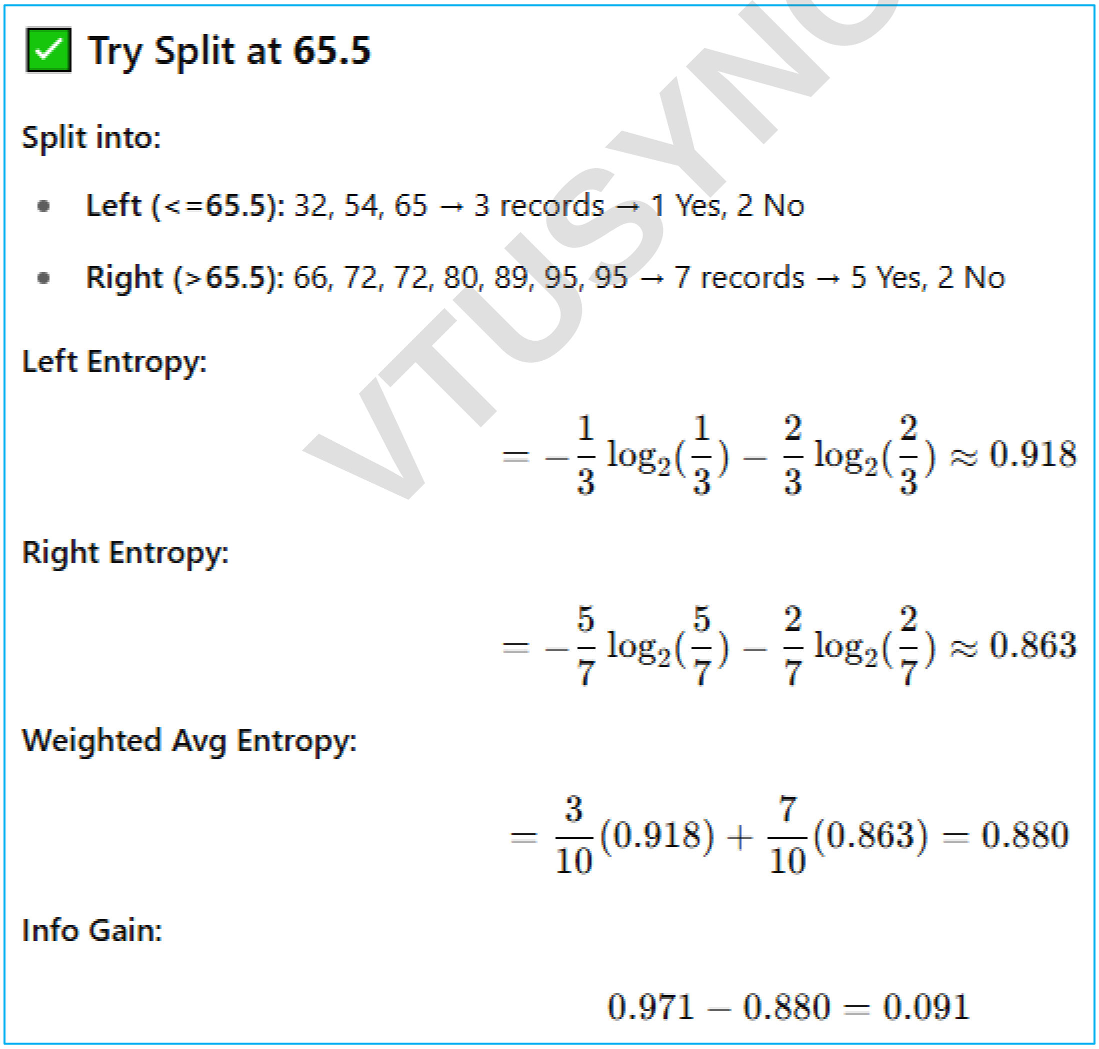
$$Entropy(right) = -\frac{6}{8}\log_2(\frac{6}{8}) - \frac{2}{8}\log_2(\frac{2}{8}) = -0.75\log_2(0.75) - 0.25\log_2(0.25) \approx 0.811$$

Weighted Avg Entropy:

$$= \frac{2}{10}(0) + \frac{8}{10}(0.811) = 0.649$$

Info Gain:

$$0.971 - 0.649 = 0.322$$



✅ Try Split at 65.5

Split into:

- **Left (<=65.5):** 32, 54, 65 → 3 records → 1 Yes, 2 No
- **Right (>65.5):** 66, 72, 72, 80, 89, 95, 95 → 7 records → 5 Yes, 2 No

Left Entropy:

$$= -\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3}) \approx 0.918$$

Right Entropy:

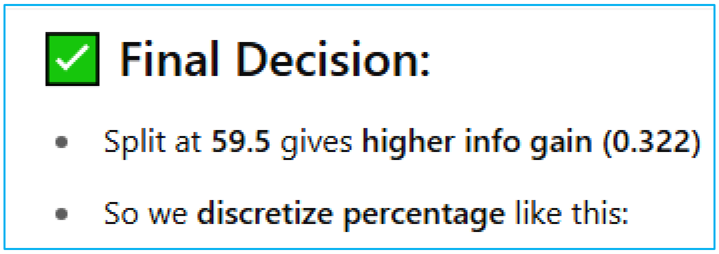
$$= -\frac{5}{7}\log_2(\frac{5}{7}) - \frac{2}{7}\log_2(\frac{2}{7}) \approx 0.863$$

Weighted Avg Entropy:

$$= \frac{3}{10}(0.918) + \frac{7}{10}(0.863) = 0.880$$

Info Gain:

$$0.971 - 0.880 = 0.091$$



✅ Final Decision:

- Split at **59.5** gives **higher info gain (0.322)**
- So we **discretize percentage** like this:

Final Discretized Attribute:

Range	Category
<= 59.5	Low
> 59.5	High

Final Table with Discretized "Percentage"

S.No	Percentage	Category	Award
1	95	High	Yes
2	80	High	Yes
3	72	High	No
4	65	High	Yes
5	95	High	Yes
6	32	Low	No
7	66	High	No
8	54	Low	No
9	89	High	Yes
10	72	High	Yes

MODULE 4

Q1. Consider a perceptron to represent the boolean function AND with the initial weights w1=0.3 and w2=-0.2, learning rate ∝ = 0.2 and bias ϴ = 0.4. The activation function used is step function f(x), which gives the output as binary. If the value of f(x) is greater than or equal to 0 then the output is 1 else it is 0. Design a perceptron that performs the boolean function AND and update the weights until the boolean function gives the desired output.

Question:

Design a perceptron that performs the Boolean AND function using the following given data:

- Initial weights: w1 = 0.3, w2 = -0.2
- Bias (θ) = 0.4
- Learning Rate (α) = 0.2
- Activation function = **Step function**:
  If net ≥ 0, output = 1
  Else output = 0

Truth Table for AND Function:

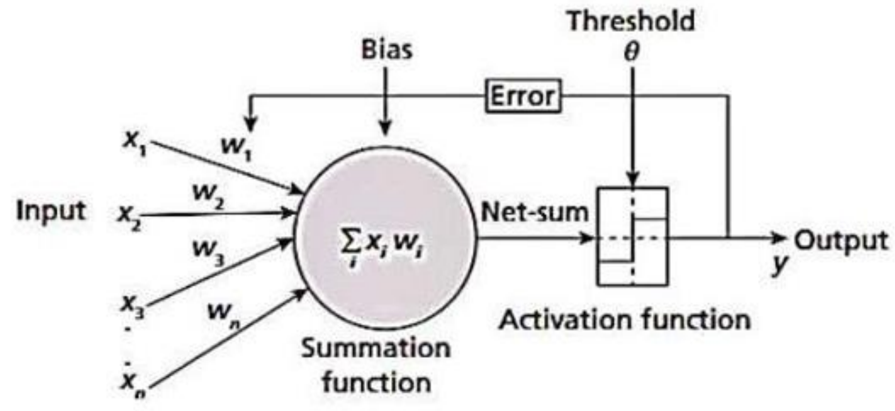
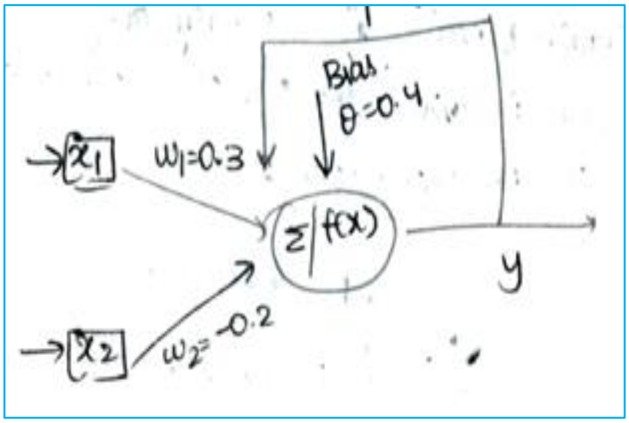


Figure 10.5: Perceptron Model



x1	x2	Target Output (t)
0	0	0
0	1	0
1	0	0
1	1	1

EPOCH 1

Weights: w1 = 0.3, w2 = -0.2

x1	x2	Net Input (w1·x1 + w2·x2 - θ)	Output
0	0	-0.4	0
0	1	-0.6	0
1	0	-0.1	0
1	1	-0.3	0 □ (wrong)

Correction:

- Error = 1 - 0 = **+1**
- Update:
  w1 = 0.3 + (0.2 × 1 × 1) = 0.5
  w2 = -0.2 + (0.2 × 1 × 1) = 0.0

New weights: **w1 = 0.5, w2 = 0.0**

EPOCH 2

Weights: w1 = 0.5, w2 = 0.0

x1	x2	Net Input	Output
0	0	-0.4	0
0	1	-0.4	0
1	0	0.1	1 □ (wrong)
1	1	0.1	1 □

Corrections:

- Row 3: Error = 0 - 1 = **-1**
  w1 = 0.5 - (0.2 × 1) = 0.3
  w2 remains = 0.0
- Row 4: Output correct, but recalculate net input with updated weights:
  0.3×1 + 0.0×1 - 0.4 = **-0.1** → Output = 0 □
- Error = 1 - 0 = +1
  w1 = 0.3 + (0.2 × 1) = 0.5
  w2 = 0.0 + (0.2 × 1) = 0.2

□ New weights:**w1 = 0.5, w2 = 0.2**

EPOCH 3

Weights: w1 = 0.5, w2 = 0.2

x1	x2	Net Input	Output
0	0	-0.4	0
0	1	-0.2	0
1	0	0.1	1 □
1	1	0.3	1 □

Corrections:

- Row 3: Error = 0 - 1 = **-1**
  w1 = 0.5 - 0.2 = 0.3
  w2 remains = 0.2

□ New weights:**w1 = 0.3, w2 = 0.2**

□ EPOCH 4

Weights: w1 = 0.3, w2 = 0.2

x1	x2	Net Input	Output
0	0	-0.4	0 □
0	1	-0.2	0 □
1	0	-0.1	0 □
1	1	0.1	1 □

□ All outputs are correct. Training complete.

Final Weights:

- w1 = **0.3**
- w2 = **0.2**
- Bias θ = **0.4**

Final Output:

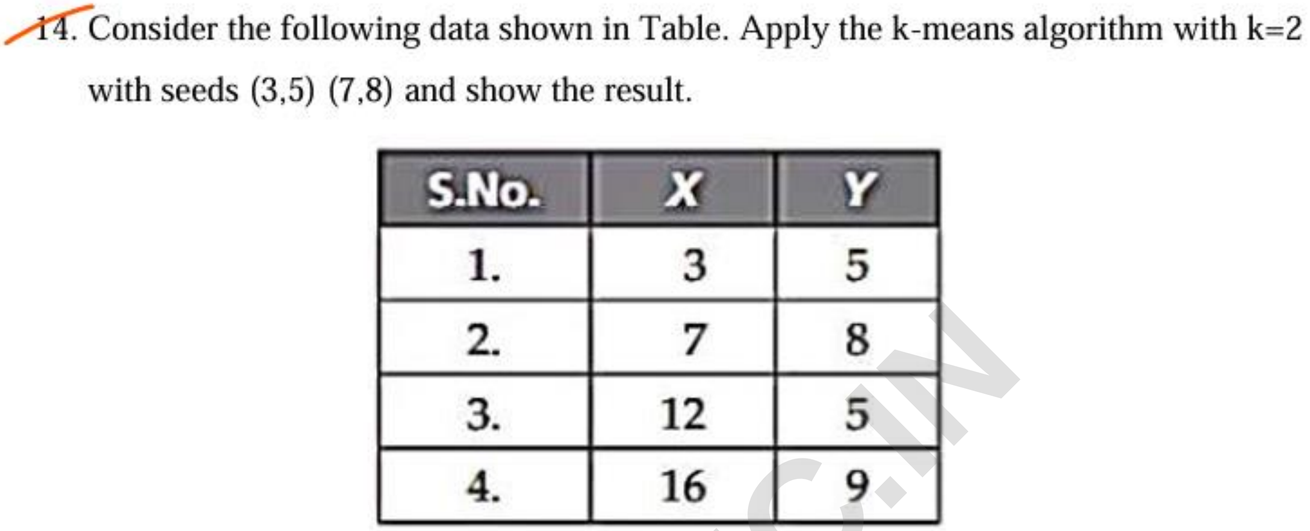
Perceptron correctly learns the AND function in **4 epochs**.

****

Machine Learning (BCS602)

IA 2 - Problems

Module - 5



14. Consider the following data shown in Table. Apply the k-means algorithm with k=2 with seeds (3,5) (7,8) and show the result.

S.No.	X	Y
1.	3	5
2.	7	8
3.	12	5
4.	16	9

□ Step 1: Understand the Question

We are given **4 points**:

S.No	X	Y
1	3	5
2	7	8
3	12	5
4	16	9

And we need to cluster them using **K-Means with K = 2** clusters.
Initial **cluster centers (centroids)**:

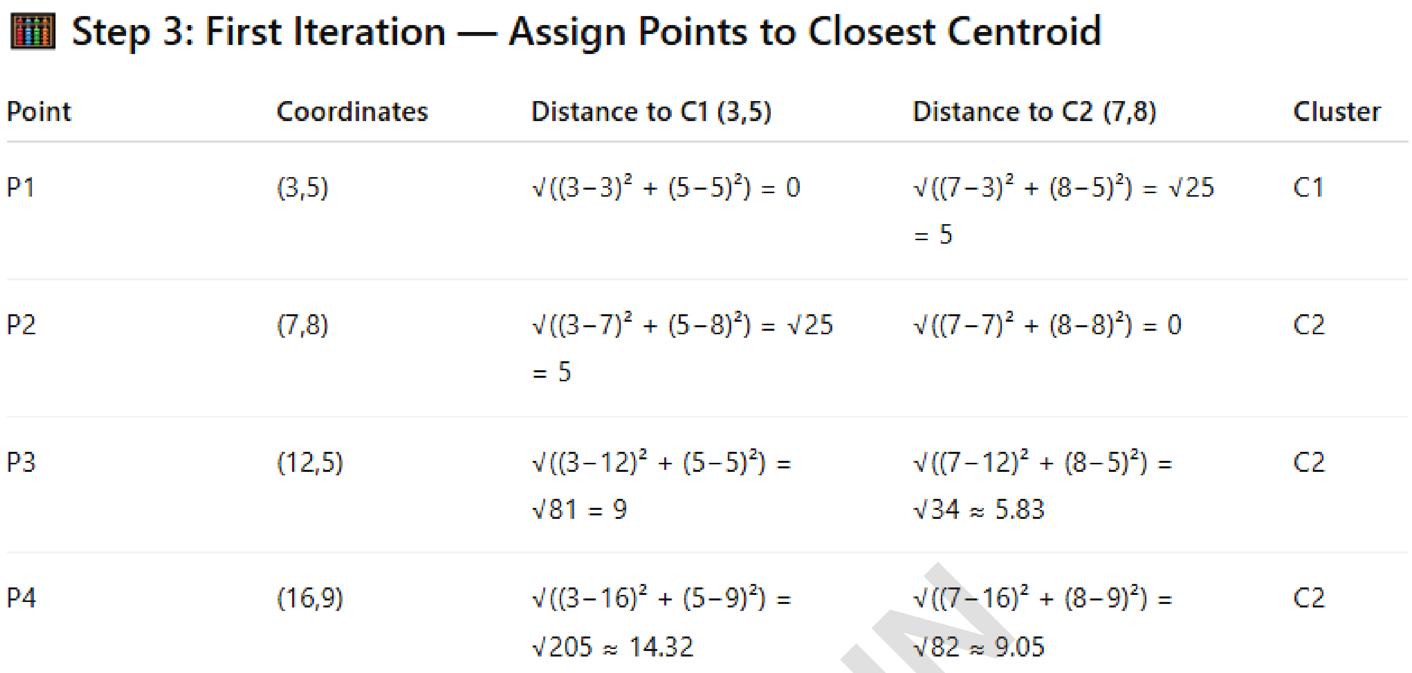
- **C1 = (3,5)**
- **C2 = (7,8)**

□ Step 2: Formula to Remember

We use **Euclidean distance** formula to assign each point to the nearest cluster:

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

We'll use this to find distance of each point from C1 and C2.



🧮 Step 3: First Iteration — Assign Points to Closest Centroid

Point	Coordinates	Distance to C1 (3,5)	Distance to C2 (7,8)	Cluster
P1	(3,5)	$\sqrt{(3-3)^2 + (5-5)^2} = 0$	$\sqrt{(7-3)^2 + (8-5)^2} = \sqrt{25} = 5$	C1
P2	(7,8)	$\sqrt{(3-7)^2 + (5-8)^2} = \sqrt{25} = 5$	$\sqrt{(7-7)^2 + (8-8)^2} = 0$	C2
P3	(12,5)	$\sqrt{(3-12)^2 + (5-5)^2} = \sqrt{81} = 9$	$\sqrt{(7-12)^2 + (8-5)^2} = \sqrt{34} \approx 5.83$	C2
P4	(16,9)	$\sqrt{(3-16)^2 + (5-9)^2} = \sqrt{205} \approx 14.32$	$\sqrt{(7-16)^2 + (8-9)^2} = \sqrt{82} \approx 9.05$	C2

Clusters after Iteration 1:

- C1: P1 (3,5)
- C2: P2 (7,8), P3 (12,5), P4 (16,9)

□ Step 4: Recalculate Cluster Centroids

New C1 = Mean of P1
= (3,5)

New C2 = Mean of P2, P3, P4

$$X = \frac{7+12+16}{3} = \frac{35}{3} \approx 11.67, \quad Y = \frac{8+5+9}{3} = \frac{22}{3} \approx 7.33$$

→ New C2 ≈ (11.67, 7.33)

🧮 Step 5: Second Iteration — Assign Points Again

Point	Coordinates	Distance to New C1 (3,5)	Distance to New C2 (11.67, 7.33)	Cluster
P1	(3,5)	0	$\sqrt{(11.67-3)^2 + (7.33-5)^2} \approx \sqrt{79.29} \approx 8.9$	C1
P2	(7,8)	$\sqrt{25} = 5$	$\sqrt{(11.67-7)^2 + (7.33-8)^2} \approx \sqrt{22.58} \approx 4.75$	C2
P3	(12,5)	9	$\sqrt{(11.67-12)^2 + (7.33-5)^2} \approx \sqrt{5.21} \approx 2.28$	C2
P4	(16,9)	14.32	$\sqrt{(11.67-16)^2 + (7.33-9)^2} \approx \sqrt{20.22} \approx 4.49$	C2

Clusters remain same as before:

- C1: P1
- C2: P2, P3, P4

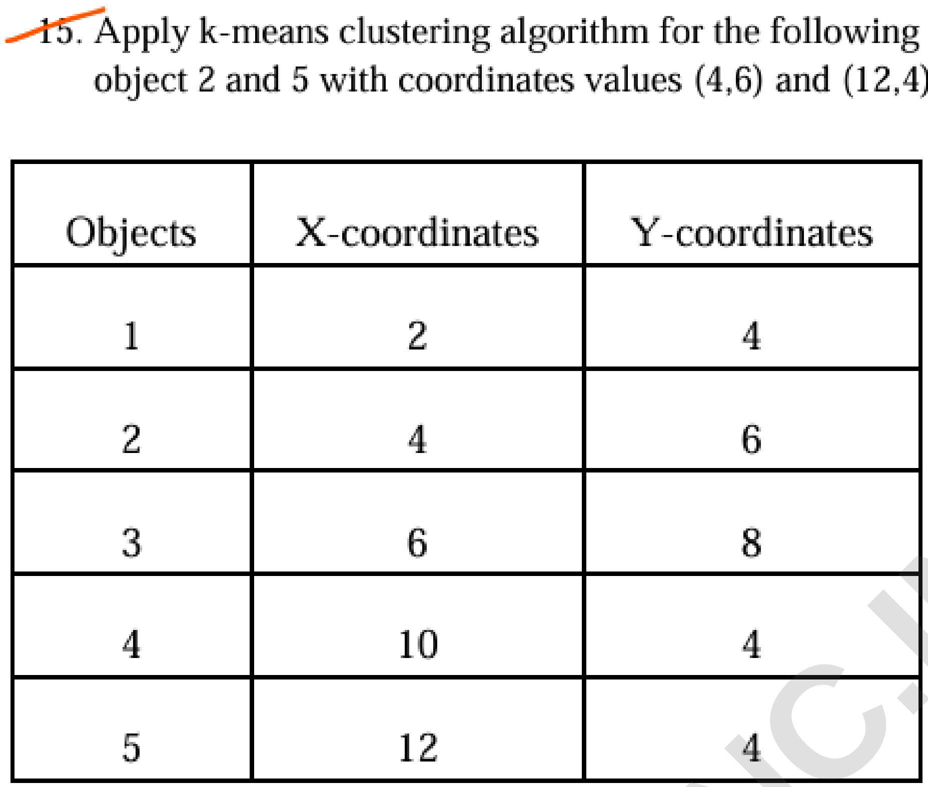
✅ **Converged! Final Clusters found**

✅ Final Answer

Cluster	Points	Centroid
C1	P1 (3,5)	(3,5)
C2	P2 (7,8), P3 (12,5), P4 (16,9)	(11.67, 7.33)

✍️ Quick Revision Points

Step	What to Do	Formula Used
Assign Points	Use Euclidean distance	$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
Update Centroid	Take average of X and Y in cluster	Mean = $(x_1+x_2+...+x_n)/n$
Repeat	Until cluster assignment doesn't change	—



15. Apply k-means clustering algorithm for the following dataset with initial value of object 2 and 5 with coordinates values (4,6) and (12,4) as initial seed

Objects	X-coordinates	Y-coordinates
1	2	4
2	4	6
3	6	8
4	10	4
5	12	4

□ Step 1: Understand the Dataset and the Question

We are given **5 data points**:

Object	X	Y
1	2	4
2	4	6
3	6	8
4	10	4
5	12	4

We are asked to perform **K-Means Clustering** with:

- **k = 2**
- Initial seeds:
  - **C1 = (4,6)** (object 2)
  - **C2 = (12,4)** (object 5)

□ Step 2: Euclidean Distance Formula

To assign each point to the nearest centroid, we use:

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

🧮 Step 3: First Iteration — Assign Points to Closest Cluster

We calculate distance of all 5 points from both centroids C1 and C2.

Object	Point (x,y)	Dist to C1 (4,6)	Dist to C2 (12,4)	Assigned Cluster
1	(2,4)	$\sqrt{[(4-2)^2 + (6-4)^2]} = \sqrt{8} \approx 2.83$	$\sqrt{[(12-2)^2 + (4-4)^2]} = \sqrt{100} = 10$	C1
2	(4,6)	$\sqrt{0} = 0$	$\sqrt{[(12-4)^2 + (4-6)^2]} = \sqrt{68} \approx 8.25$	C1
3	(6,8)	$\sqrt{[(6-4)^2 + (8-6)^2]} = \sqrt{8} \approx 2.83$	$\sqrt{[(12-6)^2 + (4-8)^2]} = \sqrt{52} \approx 7.21$	C1
4	(10,4)	$\sqrt{[(10-4)^2 + (4-6)^2]} = \sqrt{40} \approx 6.32$	$\sqrt{[(12-10)^2 + (4-4)^2]} = \sqrt{4} = 2$	C2
5	(12,4)	$\sqrt{[(12-4)^2 + (4-6)^2]} = \sqrt{68} \approx 8.25$	0	C2

□ Cluster Assignment After Iteration 1

- **Cluster C1**: Points 1, 2, 3
- **Cluster C2**: Points 4, 5

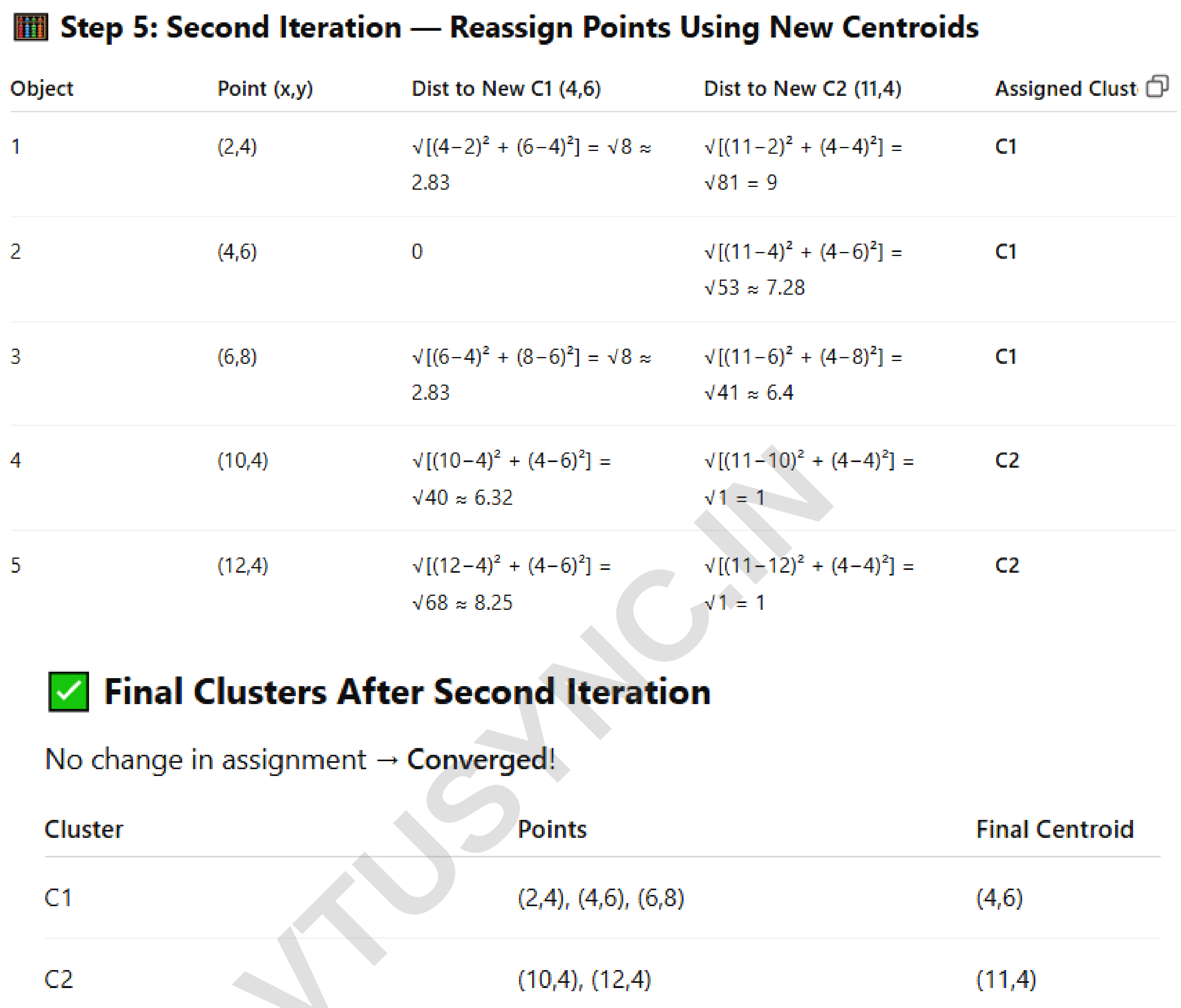
🔄 Step 4: Recalculate New Centroids

New C1: Mean of points (2,4), (4,6), (6,8)

$$X = \frac{2+4+6}{3} = \frac{12}{3} = 4, \quad Y = \frac{4+6+8}{3} = \frac{18}{3} = 6 \Rightarrow \text{New C1} = (4,6)$$

New C2: Mean of points (10,4), (12,4)

$$X = \frac{10+12}{2} = 11, \quad Y = \frac{4+4}{2} = 4 \Rightarrow \text{New C2} = (11,4)$$



🧮 Step 5: Second Iteration — Reassign Points Using New Centroids

Object	Point (x,y)	Dist to New C1 (4,6)	Dist to New C2 (11,4)	Assigned Clust
1	(2,4)	$\sqrt{[(4-2)^2 + (6-4)^2]} = \sqrt{8} \approx 2.83$	$\sqrt{[(11-2)^2 + (4-4)^2]} = \sqrt{81} = 9$	**C1**
2	(4,6)	0	$\sqrt{[(11-4)^2 + (4-6)^2]} = \sqrt{53} \approx 7.28$	**C1**
3	(6,8)	$\sqrt{[(6-4)^2 + (8-6)^2]} = \sqrt{8} \approx 2.83$	$\sqrt{[(11-6)^2 + (4-8)^2]} = \sqrt{41} \approx 6.4$	**C1**
4	(10,4)	$\sqrt{[(10-4)^2 + (4-6)^2]} = \sqrt{40} \approx 6.32$	$\sqrt{[(11-10)^2 + (4-4)^2]} = \sqrt{1} = 1$	**C2**
5	(12,4)	$\sqrt{[(12-4)^2 + (4-6)^2]} = \sqrt{68} \approx 8.25$	$\sqrt{[(11-12)^2 + (4-4)^2]} = \sqrt{1} = 1$	**C2**

✅ Final Clusters After Second Iteration

No change in assignment → **Converged!**

Cluster	Points	Final Centroid
C1	(2,4), (4,6), (6,8)	(4,6)
C2	(10,4), (12,4)	(11,4)

▯ Final Answer

Cluster C1

- Objects: 1, 2, 3
- Centroid: (4,6)

Cluster C2

- Objects: 4, 5
- Centroid: (11,4)

✍️ Summary Table for Revision

Step	Action	Formula / Value
Distance	Euclidean	$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
New Centroid	Mean	$\frac{x_1+x_2+\dots}{n}, \frac{y_1+y_2+\dots}{n}$
Stop Condition	Same clusters after iteration	—