1. Data tuples
    a. Confusion matrix (TP, TN, FP, FN)
        i. Threshold: Prob >= 0.55 (positive) else negative
        ii. From the table
            1.

Tuple #	Class (True)	Prob (classifier)	Prediction
1	p	0.95	Positive
2	n	0.85	Positive
3	p	0.78	Positive
4	p	0.66	Positive
5	n	0.60	Positive
6	p	0.55	Positive
7	n	0.53	Negative
8	n	0.52	Negative
9	n	0.51	Negative
10	p	0.40	Negative

            2. True Positives (TP): Correctly classified positive cases
                a. Tuples: 1, 3, 4, 6
                b. Count: 4
            3. False Positives (FP): Incorrectly classified negative cases as positive (n classified as p)
                a. Tuples: 2, 5
                b. Count: 2
            4. True Negatives (TN): Correctly classified negative cases (n)
                a. Tuples: 7, 8, 9
                b. Count: 3
            5. False Negatives (FN): Incorrectly classified positive cases as negative (p classified as n)
                a. Tuple: 10
                b. Count: 1
            6.

Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive (p)	TP = 4	FN = 1
Actual Negative (n)	FP = 2	TN = 3

        iii. Metrics calculation

    1. 1. Accuracy

       $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{4+3}{4+3+2+1} = \frac{7}{10} = 0.7$$

       2. Sensitivity (True Positive Rate)

       $$\text{Sensitivity (TPR)} = \frac{TP}{TP + FN} = \frac{4}{4+1} = \frac{4}{5} = 0.8$$

       3. Specificity (True Negative Rate)

       $$\text{Specificity (TNR)} = \frac{TN}{TN + FP} = \frac{3}{3+2} = \frac{3}{5} = 0.6$$

    2. 4. Precision

       $$\text{Precision} = \frac{TP}{TP + FP} = \frac{4}{4+2} = \frac{4}{6} = 0.67$$

       5. Recall

       Recall is the same as sensitivity:

       $$\text{Recall} = \frac{TP}{TP + FN} = 0.8$$

       6. F1 Score

       $$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.67 \times 0.8}{0.67 + 0.8} = 2 \times \frac{0.536}{1.47} \approx 0.73$$

    3. Summary of results:
        a. Accuracy: 0.7
        b. Sensitivity (TPR): 0.8
        c. Specificity (TNR): 0.6
        d. Precision: 0.67
        e. Recall: 0.8
        f. F1 Score: 0.73
2. Threshold to make positive / negative calls
    a. Key
        i. p (positive) and n (negative) are the true classes
        ii. A threshold decides if a probability should be classified as positive or negative

iii. For each threshold, classify probabilities accordingly and calculate the confusion matrix values (TP, FP, TN, FN)

iv. Calculations

1. Threshold = 0.9
    a. Prediction rule: Only probabilities ≥ 0.9 are classified as Positive.
    b. Tuples 1 (p) is classified as Positive.
    c.

Tuple	Class	Prob	Prediction	Outcome
1	p	0.95	Positive	TP
2	n	0.85	Negative	TN
3	p	0.78	Negative	FN
4	p	0.66	Negative	FN
5	n	0.60	Negative	TN
6	p	0.55	Negative	FN
7	n	0.53	Negative	TN
8	n	0.52	Negative	TN
9	n	0.51	Negative	TN
10	p	0.40	Negative	FN

    d. TP = 1, FP = 0, TN = 5, FN = 4
    e. TPR (Sensitivity) = TP / (TP + FN) = 1 / (1 + 4) = 0.2
    f. TNR (Specificity) = TN / (TN + FP) = 5 / (5 + 0) = 1
    g. FPR = 1 - TNR = 0
2. Threshold = 0.8
    a. Prediction rule: Probabilities ≥ 0.8 are classified as Positive.
    b. Tuples 1 (p), 2 (n) are classified as Positive.

c.

Tuple	Class	Prob	Prediction	Outcome
1	p	0.95	Positive	TP
2	n	0.85	Positive	FP
3	p	0.78	Negative	FN
4	p	0.66	Negative	FN
5	n	0.60	Negative	TN
6	p	0.55	Negative	FN
7	n	0.53	Negative	TN
8	n	0.52	Negative	TN
9	n	0.51	Negative	TN
10	p	0.40	Negative	FN

d. TP = 1, FP = 1, TN = 4, FN = 4

e. TPR = 0.2, TNR = 4 / (4 + 1) = 0.8, FPR = 0.2

3. Threshold = 0.7
    a. Prediction rule: Probabilities ≥ 0.7 are classified as Positive.
    b. Tuples 1 (p), 2 (n), 3 (p) are classified as Positive.
    c.

Tuple	Class	Prob	Prediction	Outcome
1	p	0.95	Positive	TP
2	n	0.85	Positive	FP
3	p	0.78	Positive	TP
4	p	0.66	Negative	FN
5	n	0.60	Negative	TN
6	p	0.55	Negative	FN
7	n	0.53	Negative	TN
8	n	0.52	Negative	TN
9	n	0.51	Negative	TN
10	p	0.40	Negative	FN

    d. TP = 2, FP = 1, TN = 4, FN = 3
    e. TPR = 2 / (2 + 3) = 0.4, TNR = 0.8, FPR = 0.2
4. Threshold = 0.65

a. Prediction rule: Probabilities ≥ 0.65 are classified as Positive.
b. Tuples 1 (p), 2 (n), 3 (p), 4 (p) are classified as Positive.
c.

Tuple	Class	Prob	Prediction	Outcome
1	p	0.95	Positive	TP
2	n	0.85	Positive	FP
3	p	0.78	Positive	TP
4	p	0.66	Positive	TP
5	n	0.60	Negative	TN
6	p	0.55	Negative	FN
7	n	0.53	Negative	TN
8	n	0.52	Negative	TN
9	n	0.51	Negative	TN
10	p	0.40	Negative	FN

d. TP = 3, FP = 1, TN = 4, FN = 2
e. TPR = 3 / (3 + 2) = 0.6, TNR = 0.8, FPR = 0.2

5. Threshold = 0.6
   a. Prediction rule: Probabilities ≥ 0.6 are classified as Positive.
   b. Tuples 1 (p), 2 (n), 3 (p), 4 (p), 5 (n) are classified as Positive.
   c.

Tuple	Class	Prob	Prediction	Outcome
1	p	0.95	Positive	TP
2	n	0.85	Positive	FP
3	p	0.78	Positive	TP
4	p	0.66	Positive	TP
5	n	0.60	Positive	FP
6	p	0.55	Negative	FN
7	n	0.53	Negative	TN
8	n	0.52	Negative	TN
9	n	0.51	Negative	TN
10	p	0.40	Negative	FN

        d. TP = 3, FP = 2, TN = 3, FN = 2
        e. TPR = 0.6, TNR = 3 / (3 + 2) = 0.6, FPR = 0.4
    6. Threshold = 0.55
        a. TP = 4, FP = 2, TN = 3, FN = 1
        b. TPR = 0.8, TNR = 0.6, FPR = 0.4
    7. Threshold = 0.5
        a. TP = 4, FP = 3, TN = 2, FN = 1
        b. TPR = 0.8, TNR = 0.4, FPR = 0.6
    8. Threshold = 0.4
        a. TP = 4, FP = 5, TN = 0, FN = 1
        b. TPR = 0.8, TNR = 0, FPR = 1

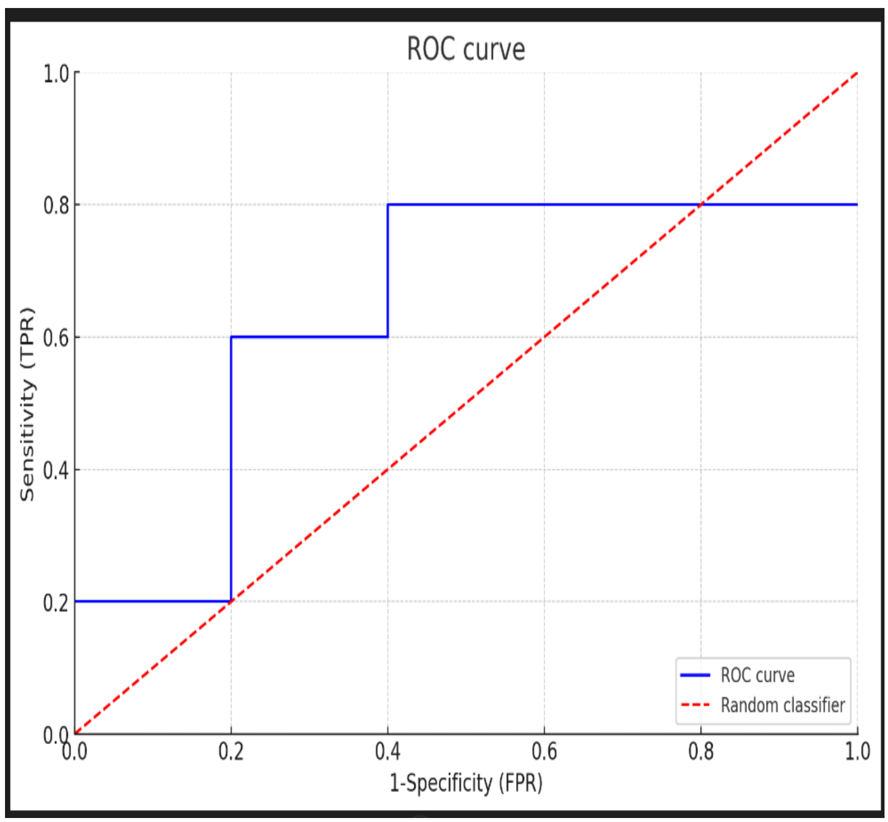
v.

Threshold	TP	FP	TN	FN	TPR (Sensitivity)	TNR (Specificity)	FPR
0.9	1	0	3	1	0.20	1.00	0
0.8	1	1	2	1	0.20	0.67	0.33
0.7	2	1	2	0	0.40	0.67	0.33
0.65	3	1	2	0	0.60	0.67	0.33
0.6	3	2	1	0	0.60	0.33	0.67
0.55	4	2	1	0	0.80	0.33	0.67
0.5	4	3	0	0	0.80	0.00	1.00
0.4	4	3	0	0	0.80	0.00	1.00

b. ROC curve
    i. We can plot the ROC curve based on the calculated TPR (Sensitivity) and FPR values for each threshold. The red line in the plot represents a random classifier, where the classifier's performance would be equal to random guessing.

ii. **ROC Curve Data Points (FPR vs TPR for each threshold):**

Threshold	TPR	FPR
0.9	0.2	0
0.8	0.2	0.2
0.7	0.4	0.2
0.65	0.6	0.2
0.6	0.6	0.4
0.55	0.8	0.4
0.5	0.8	0.6
0.4	0.8	1

c. AUC
   a. Maximum AUC value: The maximum AUC value is 1, which represents a perfect classifier
   b. For a classifier with a poor performance like this, the AUC value would likely be close to 0.5 (the diagonal line), suggesting the classifier isn't significantly better than random guessing.
      a. 

3. Steps:
   a. Calculate prior probabilities

i.
$$P(+) = \frac{\text{Number of } + \text{ records}}{\text{Total records}} = \frac{5}{10} = 0.5$$
$$P(-) = \frac{\text{Number of } - \text{ records}}{\text{Total records}} = \frac{5}{10} = 0.5$$

b. Calculate conditional probabilities

i. For $A = 0$, given class $+$ and class $-$:
$$P(A=0|+) = \frac{\text{Number of } A=0 \text{ in } +}{\text{Total number of } +} = \frac{3}{5} = 0.6$$
$$P(A=0|-) = \frac{\text{Number of } A=0 \text{ in } -}{\text{Total number of } -} = \frac{3}{5} = 0.6$$

ii. For $B = 1$, given class $+$ and class $-$:
$$P(B=1|+) = \frac{3}{5} = 0.6$$
$$P(B=1|-) = \frac{3}{5} = 0.6$$

iii. For $C = 0$, given class $+$ and class $-$:
$$P(C=0|+) = \frac{2}{5} = 0.4$$
$$P(C=0|-) = \frac{2}{5} = 0.4$$

c. Calculate posterior probabilities

i.
For class $+$:

$$P(+|A=0,B=1,C=0) = P(A=0|+)P(B=1|+)P(C=0|+)P(+)$$
$$= 0.6 \times 0.6 \times 0.4 \times 0.5 = 0.072$$

For class $-$:

$$P(-|A=0,B=1,C=0) = P(A=0|-)P(B=1|-)P(C=0|-)P(-)$$
$$= 0.6 \times 0.6 \times 0.4 \times 0.5 = 0.072$$

d. Conclusion
   i. Since the posterior probabilities for both classes + and − are the same, the classifier may not be able to decisively classify the record based on this data alone.
   ii. However, by convention or tie-breaking, you could choose either class, though additional data or information could be needed to make a better decision.

4. Gini index:

a.
Steps to Calculate Gini Index:

1. **Calculate the Gini Index for the Parent Node:** Gini Index is calculated as:

$$Gini = 1 - \sum(p_i^2)$$

where p_i is the proportion of samples in class i.

For the parent node (the root node, before splitting), we have:

- $C0$ has 10 instances
- $C1$ has 10 instances

The Gini for the root node is:

$$Gini(\text{Parent}) = 1 - \left(\left(\frac{10}{20}\right)^2 + \left(\frac{10}{20}\right)^2\right) = 1 - (0.25 + 0.25) = 0.5$$

2. **Choose Splitting Attributes:** We have three attributes to consider for splitting:
   - Gender
   - Car Type
   - Shirt Size
3. **Calculate the Gini Index for each possible split:** We will calculate the Gini Index for splitting by each attribute, and then choose the attrib with the highest Gini gain.

b. Gini gain calculations:

i.

1. Split by Gender:

- For $\text{Gender} = M$: 7 samples are $C0$, 3 samples are $C1$

$$Gini(\text{M}) = 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2 = 1 - 0.49 - 0.09 = 0.42$$

- For $\text{Gender} = F$: 3 samples are $C0$, 7 samples are $C1$

$$Gini(\text{F}) = 1 - \left(\frac{3}{10}\right)^2 - \left(\frac{7}{10}\right)^2 = 1 - 0.09 - 0.49 = 0.42$$

The weighted Gini for Gender is:

$$Gini(\text{Gender}) = \frac{10}{20} \times 0.42 + \frac{10}{20} \times 0.42 = 0.42$$

ii.

2. Split by Car Type:

- For $\text{Car Type} = \text{Family}$: 2 samples are $C0$, 3 samples are $C1$

$$Gini(\text{Family}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 1 - 0.16 - 0.36 = 0.48$$

- For $\text{Car Type} = \text{Sports}$: 5 samples are $C0$, 0 samples are $C1$

$$Gini(\text{Sports}) = 1 - \left(\frac{5}{5}\right)^2 - 0 = 0$$

- For $\text{Car Type} = \text{Luxury}$: 3 samples are $C0$, 7 samples are $C1$

$$Gini(\text{Luxury}) = 1 - \left(\frac{3}{10}\right)^2 - \left(\frac{7}{10}\right)^2 = 0.42$$

The weighted Gini for Car Type is:

$$Gini(\text{Car Type}) = \frac{5}{20} \times 0.48 + \frac{5}{20} \times 0 + \frac{10}{20} \times 0.42 = 0.045 + 0 + 0.21 = 0.255$$

iii.

3. Split by Shirt Size:

- For $\text{Shirt Size} = \text{Small}$: 3 samples are $C0$, 3 samples are $C1$

$$Gini(\text{Small}) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 1 - 0.25 - 0.25 = 0.5$$

- For $\text{Shirt Size} = \text{Medium}$: 3 samples are $C0$, 5 samples are $C1$

$$Gini(\text{Medium}) = 1 - \left(\frac{3}{8}\right)^2 - \left(\frac{5}{8}\right)^2 = 1 - 0.14 - 0.39 = 0.47$$

- For $\text{Shirt Size} = \text{Large}$: 1 sample is $C0$, 2 samples are $C1$

$$Gini(\text{Large}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.44$$

- For $\text{Shirt Size} = \text{Extra Large}$: 3 samples are $C0$, 0 samples are $C1$

$$Gini(\text{Extra Large}) = 1 - \left(\frac{3}{3}\right)^2 - 0 = 0$$

The weighted Gini for Shirt Size is:

$$Gini(\text{Shirt Size}) = \frac{6}{20} \times 0.5 + \frac{8}{20} \times 0.47 + \frac{3}{20} \times 0.44 + \frac{3}{20} \times 0 = 0.15 + 0.188 + 0.066 + 0 = 0.404$$

iv. Best split:

a. Based on the Gini index calculations:
   - **Car Type** has the lowest Gini value (0.255), so it is the best attribute to split on first.

b. This means the first node of the decision tree should split on **Car Type**, and the tree can further split based on the remaining attributes.