Ques 1:

Features: Travel Distance (km)
Number of transfers (integer)
Day of week (weekday / weekend)
bus stop ID (numeric code)
average traffic delay (minutes)
weather condition (clear / rain / snow)

a). Classification problem since the output is a categorical value (Yes/No).

b). Feature 'bus stop ID' should be removed as it doesnot carry predictive information.

c). Feature Scaling is important because features like travel distance and average traffic delay have different ranges, which can distort algorithms sensitive to magnitude.
Eg: KNN requires scaling since it uses distance calculations while Decision Trees donot because they split on thresholds and are unaffected by feature scales.

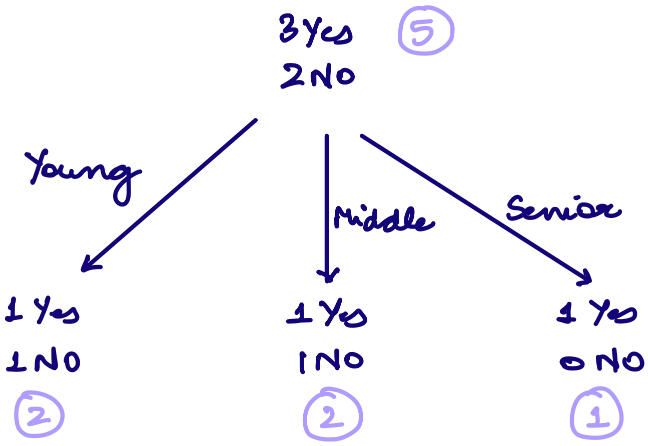
d). Linear Regression is not appropriate because it assumes a continuous target variable, which is not the case in this question. More suitable choice will be logistic Regression or Random forest (for non linear)

Ques 2:

a). Original Dataset:

Age Group (x)	Income Level (x)	Purchase (y)
Young	Low	No
Young	High	Yes
Middle	Low	No
Middle	High	Yes
Senior	High	Yes

x_j : Age Group



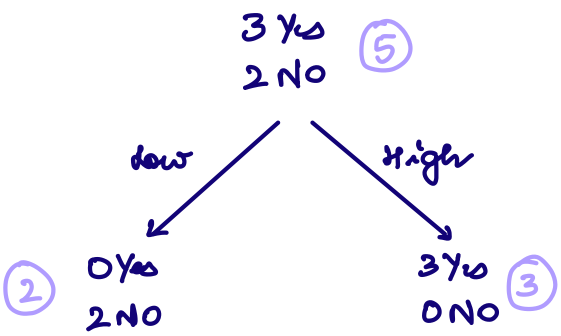
$$Gini(y|x_j = Young) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$$
$$= 0.5$$

$$Gini(y|x_j = Middle) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$$
$$= 0.5$$

$$Gini(y|x_j = Middle) = 1 - \left[\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right]$$
$$= 0$$

$$Gini(y|x_j) = \frac{2}{5} \times 0.5 + \frac{2}{5} \times 0.5 + \frac{1}{5} \times 0 = 0.4$$

x_j : Income Level



$$Gini(y|x_j = Low) : 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$Gini(y|x_j : High) : 1 - \left[\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right] = 0$$

$$Gini(y|x_j) = \frac{2}{5} \times 0 + \frac{3}{5} \times 0 = 0$$

Root chosen: Income Level

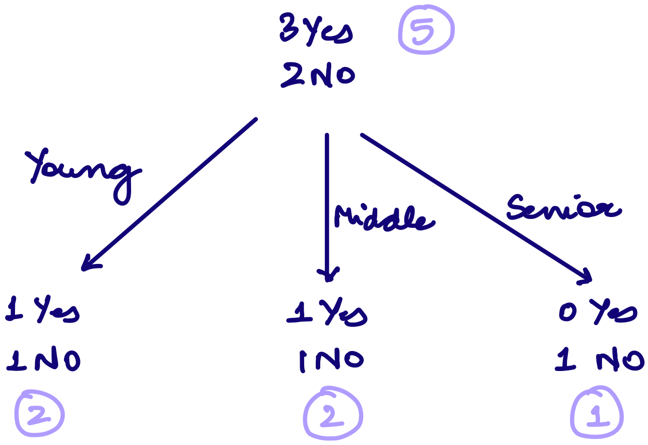
b). Corrected Dataset

$\longleftarrow x \longrightarrow$ $\longleftarrow y \longrightarrow$

Age Group	Income Level	Purchase
Young	Low	NO
Young	High	Yes
Middle	Low	NO

Middle	High	Yes
Senior	High	NO

x_j : Age Group



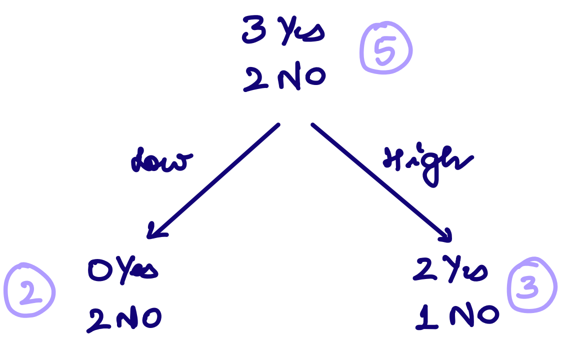
$$Gini(y|x_j = Young) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$$
$$= 0.5$$

$$Gini(y|x_j = Middle) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$$
$$= 0.5$$

$$Gini(y|x_j = Middle) = 1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right]$$
$$= 0$$

$$Gini(y|x_j) = \frac{2}{5} \times 0.5 + \frac{2}{5} \times 0.5 + \frac{1}{5} \times 0 = 0.4$$

x_j : Income Level



$$Gini(y|x_j = Low) : 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$Gini(y|x_j : High) : 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.44$$

$$Gini(y|x_j) = \frac{2}{5} \times 0 + \frac{3}{5} \times 0.44 = 0.264$$

Root chosen: Income Level

c). In original dataset, splitting on Income perfectly separated Yes and NO, giving a highly accurate tree.
In corrected dataset, split on Income didnot separate the outcomes perfectly

d). This example highlights the instability (high variance) of decision trees, a small change in data can cause large structural differences. In real world tasks like churn prediction or loan approval, such instability can lead to inconsistent decisions, reducing trust & reliability of model.

Ques 3:

Table II: Confusion Matrix with $C = 0.1$

	Predicted Diseased (1)	Predicted Healthy (0)
Actual Diseased (1)	40 TP	20 FN
Actual Healthy (0)	5 FP	135 TN

Table III: Confusion Matrix with $C = 100$

	Predicted Diseased (1)	Predicted Healthy (0)
Actual Diseased (1)	55 TP	5 FN
Actual Healthy (0)	25 FP	115 TN

a). $C = 0.1$

$$\text{Precision} : \frac{TP}{TP+FP} = \frac{40}{40+5} = \frac{40}{45} = 0.89$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{40}{40+20} = \frac{40}{60} = 0.67$$

$C = 100$

$$\text{Precision} : \frac{TP}{TP+FP} = \frac{55}{55+25} = \frac{55}{80} = 0.69$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{55}{55+5} = \frac{55}{60} = 0.92$$

b). Missing a diseased plant – Plant is actually diseased but we predicted healthy – False Negatives are very costly. We want a system where we have less no. of False Negatives. In $C=0.1$, there are 20 FN whereas in $C=100$

there are 5 FN. So, we will pick C= 100.

c). Costly operation - actual healthy but predicted diseased - false positives should be minimized. With C=0.1, there are 5 FP and with C=100 there are 25 FP. So, C=0.1 is better.

Ques 4:

Customer	True Label (y)	Predicted Probability (p)	$t=0.5$	$t=0.7$
C1	1	0.90	1	1
C2	0	0.80	1	1
C3	1	0.70	1	1
C4	0	0.60	1	0
C5	1	0.40	0	0
C6	0	0.30	0	0

a).

$t=0.5$

		Predicted	
		1	0
Actual	1	(C1, C3) 2 TP	(C5) 1 FN
	0	(C2, C4) 2 FP	(C6) 1 TN

Recall / Benefit

$$TPR = \frac{TP}{TP+FN} = \frac{2}{2+1} = 0.67$$

$$FPR = \frac{FP}{FP+TN} = \frac{2}{2+1} = 0.67$$

Specificity / Cost

$t=0.7$

		Predicted	
		1	0
Actual	1	(C1, C3) 2 TP	(C5) 1 FN
	0	(C2) 1 FP	(C4, C6) 2 TN

$$TPR = \frac{TP}{TP+FN} = \frac{2}{2+1} = 0.67$$

$$FPR = \frac{FP}{FP+TN} = \frac{1}{1+2} = 0.33$$

b). Raising the threshold from 0.5 to 0.7 reduces FPR, while recall stayed the same.

In Banking context, threshold of 0.7 is better as it reduces FPR.