Decision Trees

⇒ Distance-Based Algorithm.

- Linear Reg → Distance
- Logistic Reg → Distance
- K-Means → "
- KNN → "
- SVM → "

When we use Distance Based Algorithm, we will scale our data, such that there will be no biasness.

Decision Trees

→ This is not a distance based Algorithm.

→ It is a Supervised Machine Learning algorithm, which can be used for Classification and Regression.

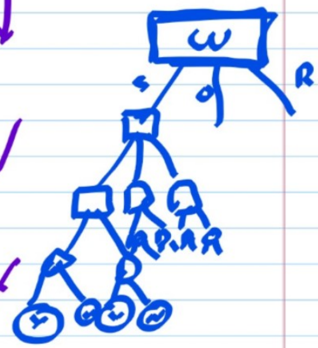
→ It uses set of if-else condition.

Independent Feature → Dependent Feature (Label)

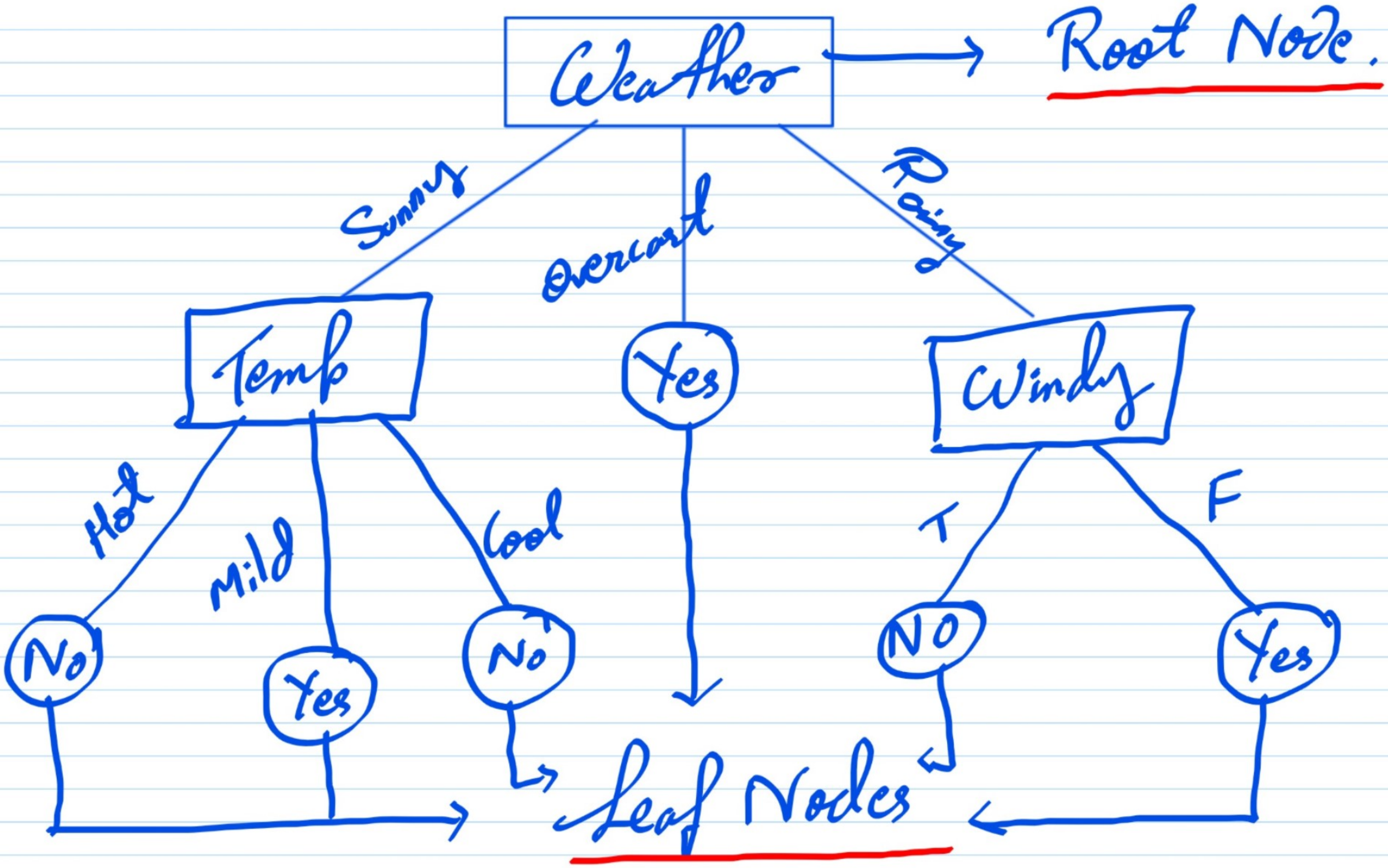
e.g →

Th= 0.18

Weather	Temperature	Humidity	Windy	Play
Sunny	Hot	High	F	N
Sunny	Hot	High	T	N
Overcast	Hot	High	F	Y
Rainy	Mild	High	F	Y
Rainy	Cool	Normal	F	Y
Rainy	Cool	Normal	T	N
Overcast	Cool	Normal	T	Y



Creates a set of if-else statement.



Q) How to select the root node?

Entropy → Measure of Randomness.

→ The more the Randomness, the greater the entropy of our System.

$$\rightarrow \left[H(S) = -\sum_{i=1}^{K} P(y_i) \times \log P(y_i) \right]$$

e.g. →

Feature 1	Feature 2	Feature 3
Y	Y	Y
Y	Y	N
Y	Y	Y
Y	N	Y
Y	N	Y
	N	N

Pure split (Feature 1)

Impure Split (Feature 2)

Feature 2: Y → 3, N → 3

Feature 3: Y → 4, N → 2

No - Randomness | High Randomness | Randomness.

$$\rightarrow \left[H(S) = -P_{(+)} \log_2 (P_+) - P_{(-)} \log_2 (P_-) \right]$$

$P_{(+)}$ → Probability of positive class

$P_{(-)}$ → Probability of Negative class.

$$P(Y) = \frac{\text{Total no. of favourable outcomes}}{\text{Total no. of Outcomes}}$$

$$\Rightarrow \frac{4}{7} = 0.57$$

$$P(N) \Rightarrow \frac{3}{7} = 0.42$$

[Entropy] $H(s) = -0.57 \times \log(0.57) - 0.42 \times \log(0.42)$

$= 0.29$

$H(s) = 0$, it is a best split
$H(s) = 1$, it is a worst split

feature -1

$5 \rightarrow Y$

$0 \rightarrow N$

$$H(s) = -\frac{5}{5} \times \log(1)$$

$$= -\log(1) = 0$$

It has no-randomness

feature - 2

Y → 3

N → 3

$$H(s) = -\frac{3}{6} \times \log\left(\frac{3}{6}\right) - \frac{3}{6} \times \log\left(\frac{3}{6}\right)$$

$$= -\frac{1}{2} \times \log\left(\frac{1}{2}\right) - \frac{1}{2} \times \log\left(\frac{1}{2}\right)$$

$$= -0.5 \times (-0.30) - 0.5 \times (-0.30)$$

$$= +0.5 \times 0.30 + 0.5 \times 0.30$$

$$= 0.15 + 0.15$$

$$= 0.30$$

feature - 3

H(s) → It has some randomness.

Information Gain

Category Sunny

Weather	Temperature	Humidity	Wind	Rainy
Sunny	Hot	High	F	N
Sunny	Hot	High	T	N

$$H(s) = -\cancel{0 \times \log(0)} - \frac{2}{2} \times \log(1)$$

$$= -\log(1) = 0$$

weighted avg $= \frac{2}{7} = 0.28$

~~Category~~ Rainy

Rainy	Mild	High	F	Y
Rainy	Cool	Normal	F	N
Rainy	Cool	Normal	T	N

$$H(s) = -P\left(\frac{1}{3}\right) \times \log\left(\frac{1}{3}\right) - \frac{2}{3} \times \log\left(\frac{2}{3}\right)$$

$$= 0.27$$

w. avg $= \frac{3}{7} = 0.42$

~~Category~~ Overcast

Overcast	Hot	High	F	Y
Overcast	Cool	Normal	T	Y

$$H(s) = 0$$

w. avg $= \frac{2}{7} = 0.28$

Information Gain = $H(S) - [w\,avg \times H(f_1) + w.avg \times H(f_2) + w.avg \times H(f_3)]$

$$IG = 0.29 - [0 + 0.27 \times 0.42 + 0]$$

$$= 0.29 - 0.11 = 0.18$$

→ Information Gain for weather is 0.18

The root node should have the $\boxed{\text{highest}}$ Information Gain.

Gini Impurity (Computationally Efficient)

$$GI = 1 - \sum_{i=1}^{n} (P_i)^2$$

$$= 1 - [(P_+)^2 + (P_-)^2]$$

~~Category~~ Sunny

$$\begin{array}{cc} Y & N \\ + & - \end{array}$$

$$GI = 1 - \left[0 + \left(\frac{2}{2}\right)^2\right]$$

$$= 1 - 1 = 0$$

$$w.avg = \frac{2}{7} = 0.28$$

~~Category~~ Rainy

$$GI = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right]$$

$$= 1 - \left[(0.3)^2 + (0.6)^2\right]$$

$$= 0.44$$

$$w.avg = \frac{3}{7} = 0.42$$

~~Category~~ Overcast

$$GI = 1 - \left[\left(\frac{2}{2}\right)^2 + 0\right]$$

$$= 1 - 1 = 0$$

$$w.avg = \frac{2}{7} = 0.28$$

$$G.I = \frac{2}{7} \times 0 + \frac{3}{7} \times 0.44 + \frac{2}{7} \times 0 \Rightarrow 0.188$$

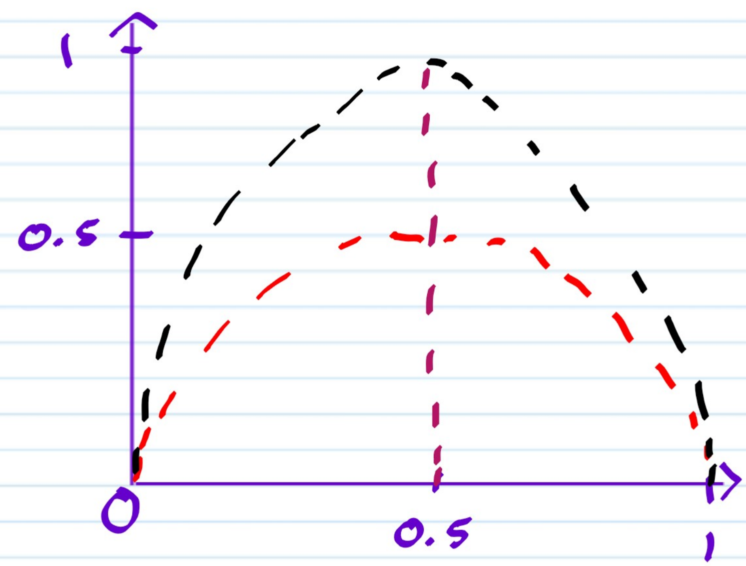
Total Gini Impurity = w.avg × G.I(f_1) + w.avg × G.I(f_2) + w.avg × G.I(f_3)

$$w.avg = \frac{\text{Total no. of data points in a particular category}}{\text{Total no. of data points in the entire system.}}$$

Total Gini Impurity = 0.28 × 0 + 0.42 × 0.44 + 0.28 × 0

⇒ 0.1848.

Gini Impurity Vs Entropy Graph



– → Entropy (0 to 1)

– → Gini Impurity (0 to 0.5)

Impure Split

$Y \rightarrow 3$

$N \rightarrow 3$

$$G.I = 1 - \sum_{i=1}^{n} (P_i)^2$$

$$= 1 - [(P_Y)^2 + (P_N)^2]$$

$$= 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$$

$$= 1 - [(0.5)^2 + (0.5)^2] = 0.5$$

Pure Split

$Y \rightarrow 6$

$N \rightarrow 0$

$$G.I = 1 - [(P_y)^2 + (P_N)^2]$$

$$= 1 - \left[\left(\frac{6}{6}\right)^2 + \left(\frac{0}{6}\right)^2\right]$$

$$= 1 - 1 = 0$$

The root node should have the <u>lowest</u> Gini Impurity

Gini Impurity is Computationally Efficient as it doesn't have to do log computation.

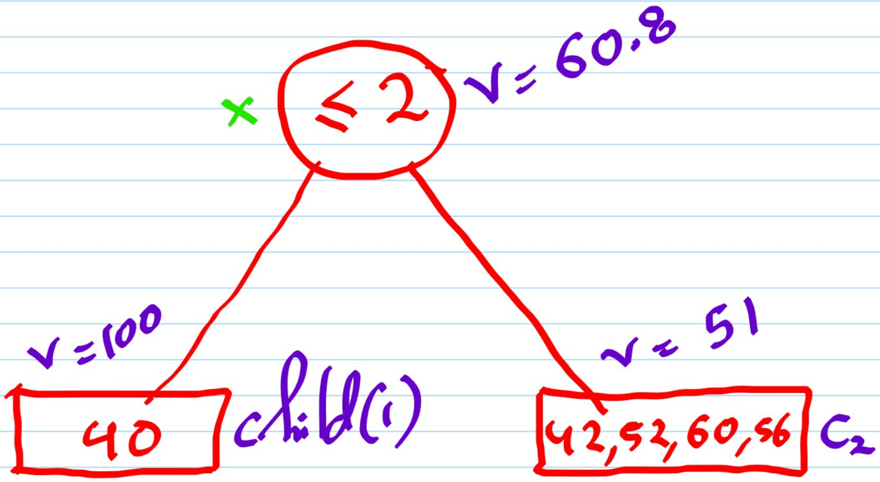
(CART), (ID3), → DT Algorithms

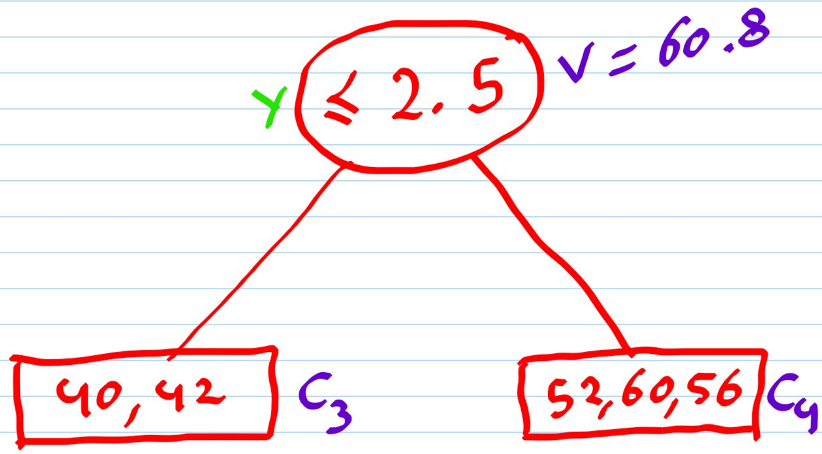
Decision Tree Regressor

e.g:-

	Exp	Gap	Salary(K) (o/p) (label)	
x →	2	Y	40	→ VR → 0
y →	2.5	Y	42	→ VR →
z →	3	N	52	→ VR →
A →	4	N	60	→ VR →
B →	4.5	Y	56	→ VR →

$$Avg = \frac{40 + 42 + 52 + 60 + 56}{5} = 50$$





Variance Reduction (We choose the node, whose Variance Reduction is High.)

Variance (Root) $\Rightarrow \frac{1}{n} \sum_{i=1}^{n} (y - \hat{y})^2$

$\rightarrow$ Error.

$\hat{y}$ = Average Output = 50

$n \rightarrow$ No. of data points

$y \rightarrow$ Each individual data points.

$$\text{Variance} = \frac{1}{5}\left[(40-50)^2 + (42-50)^2 + (52-50)^2 + (60-50)^2 + (56-50)^2\right]$$

$$= \frac{1}{5}\left[100 + 64 + 4 + 100 + 36\right] = 60.8$$

$$v(C_1) = \frac{1}{1}\left[(40-50)^2\right] = 100$$

$$v(C_2) = \frac{1}{4}\left[(42-50)^2 + (52-50)^2 + (60-50)^2 + (56-50)^2\right]$$

$$= \frac{1}{4}\left[64 + 4 + 100 + 36\right] = 51$$

$$\underset{(x)}{\text{Variance Reduction}} = Var(Root) - \sum w_i \, Var(\text{child node})$$

$$= 60.8 - \left[\frac{1}{5} \times 100 + \frac{4}{5} \times 51\right]$$

$$= 60.8 - 20 - 40.8$$

$$= 0$$

$$\boxed{\begin{aligned} w_i &= \text{weights} \\ &= \text{Probability} \end{aligned}}$$