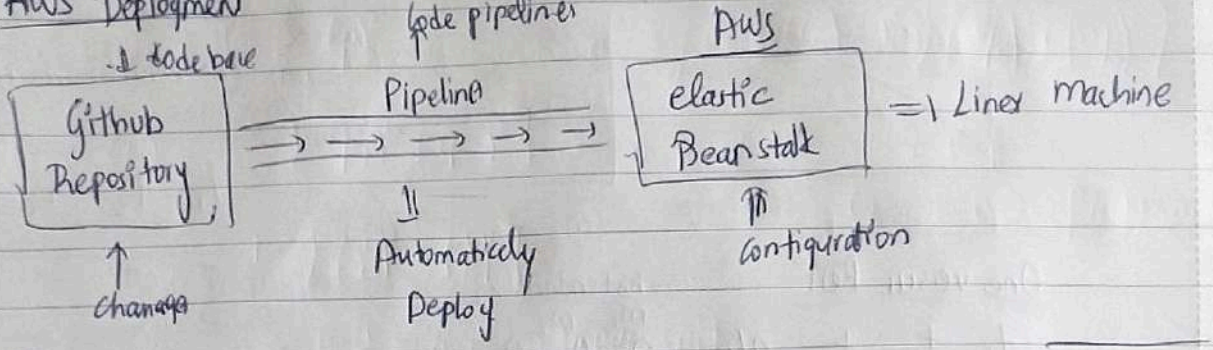
AWS Deployment

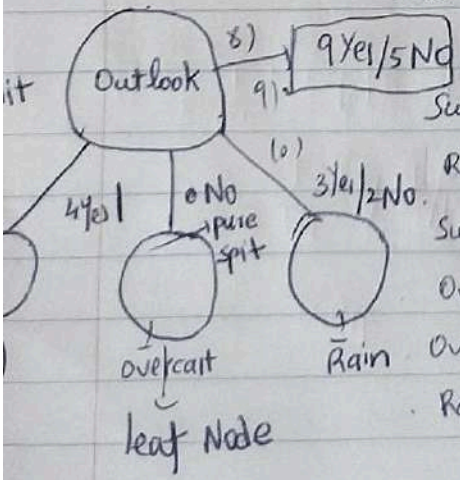


Decision Tree

Dataset

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1)	Sunny	Hot	High	Weak	No
2)	Sunny	Hot	High	Strong	No
3)	Overcast	hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5)	Rain				Yes
6)	Rain				No
7)	Over.				Yes
8)	9Yes/5No				No
9)	Sunny				Yes
	Rain				Yes
	Sunney				Yes
	Overcast				Yes
	Ove.cast				Yes
	Rain				NO

14)



① Purity Spit check :- Pure split (or) Impure split

- Entropy
- Gini Impurity

} measure of purity

② What feature you need to select to start the split → Information Gain

① Purity Check

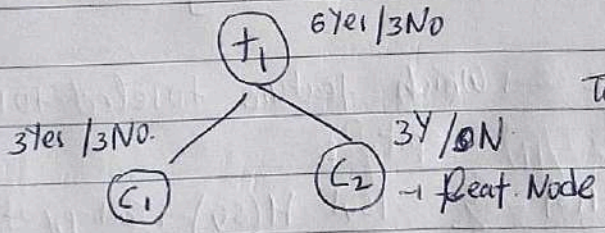
Binary Classification

① Entropy

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

P_+ = probability of positive category

P_- = probability of negative category



$P_+ = \frac{3}{6} = \frac{1}{2}$ ← Total positive count; Total count ← 6

$P_- = \frac{3}{6} = \frac{1}{2}$ ← Total negative count; Total count

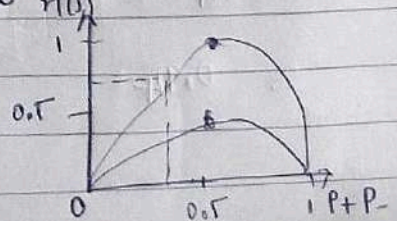
$$H(C_1) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$= \frac{-3}{6} \log_2 - 3/6 \log 3/6$$

$\boxed{H(C_1) = 1}$ → This is impure

$$H(C_2) = \frac{-3}{3} \log_2 3/3 - \frac{0}{3} \log 0/3$$

$\boxed{H(C_2) = 0}$ pure split



② Gini Impurity

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

② Gini Impurity → max → 0.5

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

Gini $(G_1) = 1 - [(p_+)^2 + (p_-)^2]$

$= 1 - [(\frac{1}{2})^2 + (\frac{1}{2})^2]$

$= 1 - \frac{1}{2} = 0.5$ → Impure spit

3 Yes 0 No

$G.I(C_2) = 1 - [(3/3)^2 + (0/3)^2]$

$= 1 - 1$

$C_2 = 0$ → Purity spilt

Multiclass classification Problem : 3 categories In O/P

$$H(S) = -P_{C_1} \log_2 P_{C_1} - P_{C_2} \log_2 P_{C_2} - P_{C_3} \log_2 P_{C_3}$$

$$G.I = 1 - [(P_{C_1})^2 + (P_{C_2})^2 + (P_{C_3})^2]$$

② Information Gain : → Which feature to select to start the spit ?

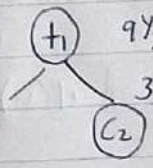
$$\text{Gain}(S, t_1) = H(S) - \sum_{v \in val} \frac{|S_v|}{|S|} H(S_v)$$ → Entropy of child node

$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$

$= -\frac{9}{14} \log 9/14 - \frac{5}{14} \log (5/14)$ 6Y 2N

$\simeq 0.94$

t_1	t_2	t_3	O/P
			9Y 5N ← H(S) = 0.94
			3Y 3No



$$H(C_1) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) \simeq 0.81$$

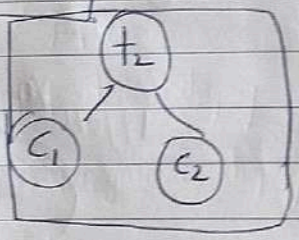
$$H(C_2) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right) = 1$$

$$\text{Gain}(S, f_1) = H(S) - \sum_{v \in \text{val}} \frac{|S_v|}{|S|} H(S_v) \quad \text{Entropy of categories}$$

$$= 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

(Total no. others / Total no. this; $\frac{8}{14}$: Total count; $H(C_1)$; $\frac{6}{14}$: Total count; $H(C_2)$)

$$\text{Gain}(S, f_1) = 0.049$$



$\Rightarrow$ Information Gain = 0.01

$$\text{Gain}(S, t_2) = 0.01 \quad > \quad \text{Gain}(S, t_1) = 0.049$$

We need to start splitting using f_2 feature

~~Entropy~~

Entropy vs Gini Impurity

When dataset is small → Entropy [log formula]

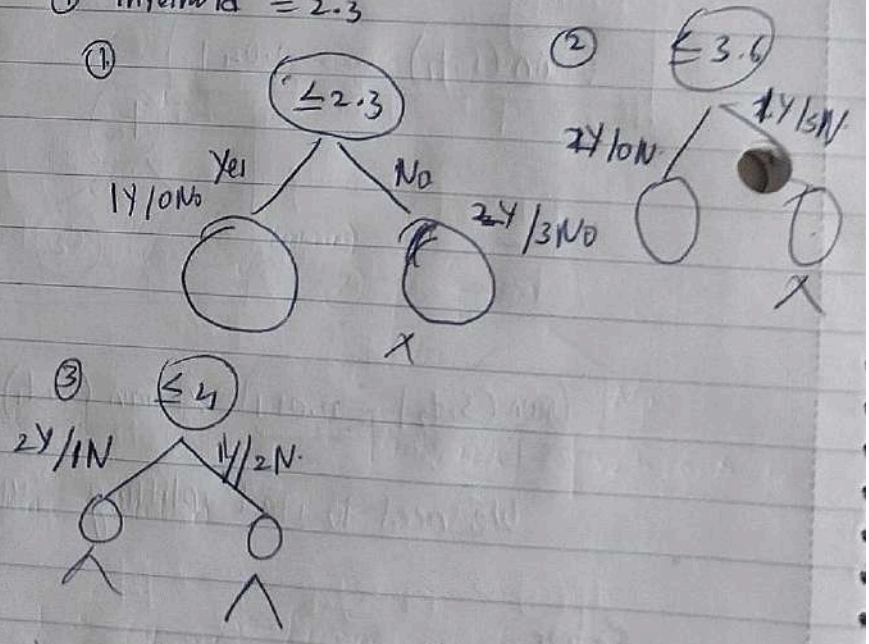
When dataset is huge → Gini Impurity [simple math]

What if my feature is continuous

f_1	O/P
2.3	Yes
3.6	Yes
4	No
5.2	No
6.7	Yes
7.8	No

① Sort the feature f_1

① Threshold = 2.3



Time Complexity is huge ~~then~~ when dataset is huge

```
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier()
```

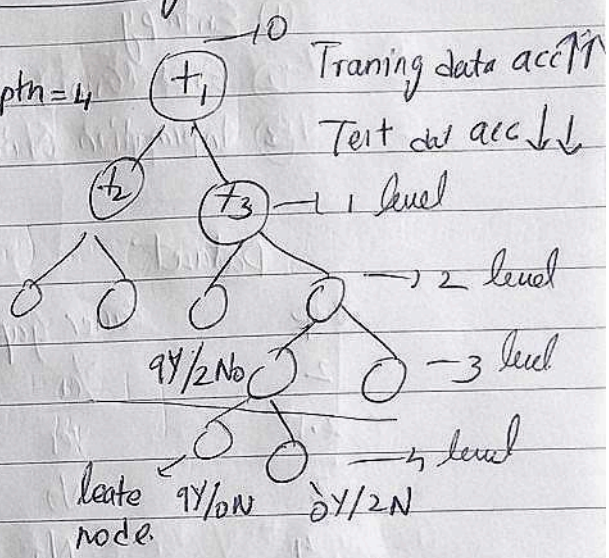
Decision Tree post pruning and pre pruning [Reduce overfitting]

Training Data

t_1 t_2 t_3 O/P

Generalized model

Reduce overfitting



① Post Pruning

① Construct the entire decision Tree to complete leaf node.

② Pruning the decision

③ ~~for~~ Suitable for smaller Dataset

② Pre pruning

① No Hyperparameter Tuning to select Best parameters
(GridSCV · Random RCV)

Decision Tree Classifier

1. Entropy
2. G.I (Gini)
3. Information Gain

Decision Tree Regressor

1. Variance Reduction
2. Variance

Dataset

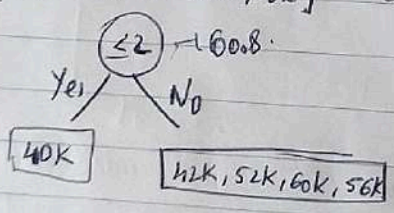
o/p → continuous.

exp	career Gap	Salary
2	Yes	40k
2.5	Yes	42k
3	No	52k
4	No	60k
4.5	Yes	56k

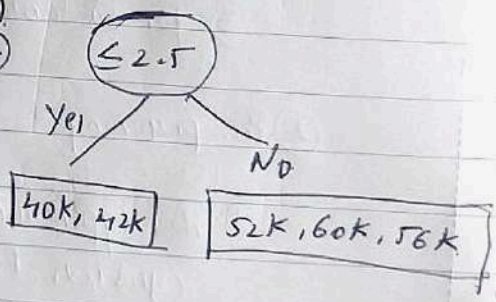
$\bar{y} = 50k$ ← avg. salary (mean)

Variance Reduction

① [40k, 42k, 52k, 60k, 56k]



②



Final aim → Variance reduction

$$\text{Variance or error} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y_i})^2 \quad \text{[mean squared error]}$$

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

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

$$= 60.8$$

Variance of left Ch = $\frac{1}{1}[(40-50)^2$

$= 100$ //

Variance of right = $\frac{1}{4}\left((42-50)^2+(52-50)^2+(60-60)^2+(50-56)^2\right.$

$= 51$ //

Variance of Reduction = Var (~~Reduction~~ Root) $- \sum w_i$ Var (child)

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

$= 0$

Variance Reduction = 0

Slect the gretest Varience Reduction