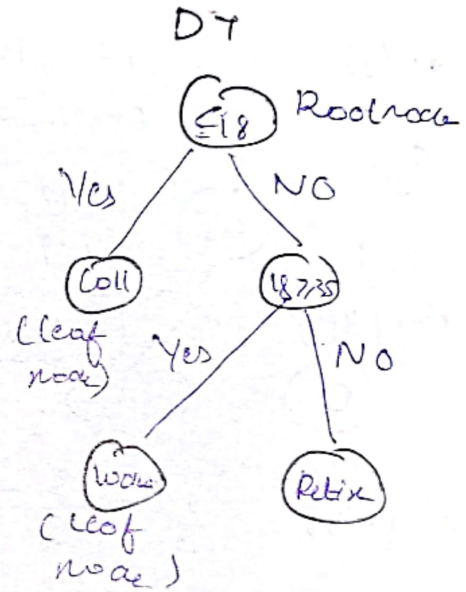
Day-9 Machine Learning Algorithms

Agenda =

1. Decision Tree Classification
2. Decision Tree Regression
3. Practical Implementation
4. Ensemble Technique

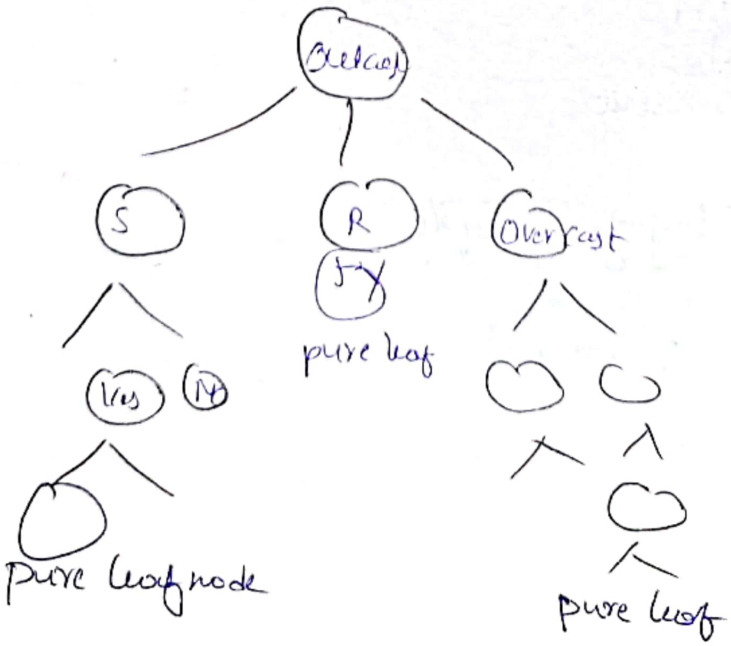
Decision Tree solving many usecases
- Classification
- Regression (It has worst time complexity)

```
if (age <= 18):
    print("college")
elif (age > 18 and age <= 35):
    print("work")
else:
    print("retire")
```



Nested If-else => Decision Tree

Eg:-



① purity → purity Split ?
(it is used for less number of features)

↳ Entropy
↳ Gini Coefficient
(it is used for huge features)

} ways to find purity

② How the features are selected

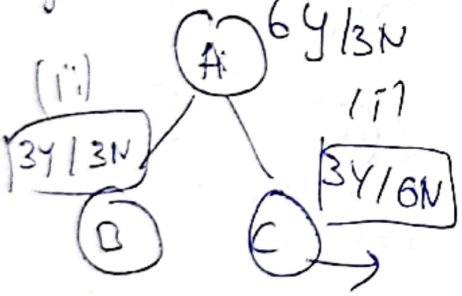
↳ Information Gain ? G

① Entropy

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

(+ = Y, - = No)

Eg:-



i) $H(S) = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$

$= -1 \log_2 1$

$= 0$

↓

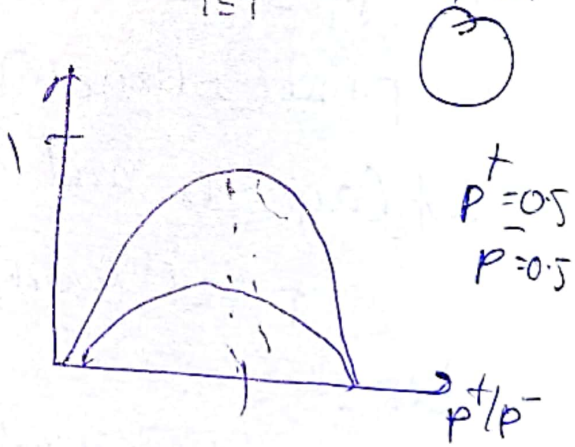
∀ Pure Split

ii) $H(S) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log_2 3/6$

$= 1$ Impure Split

① GINI Impurity

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



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

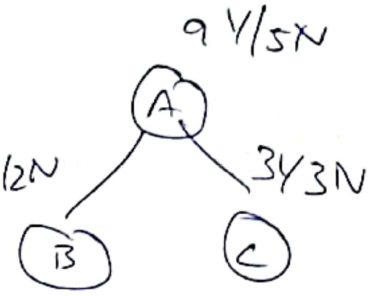
$= 1 - [(\frac{1}{2})^2 + (\frac{1}{2})^2]$

$= 1 - \frac{1}{2}$

$= 0.5$

Information Gain

$$\text{Gain}(S, F_1) = H(S) - \sum_{v \in \text{Val}} \frac{|S_v|}{|S|} H(S_v)$$



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

$$= -\frac{9}{14} \log_2 (9/14) - 5/14 \log_2 (5/14)$$

$$= \boxed{0.94}$$

$$H(S_{v_B}) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8}$$

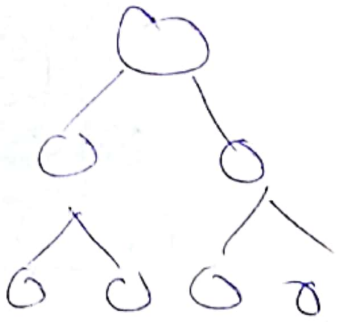
$$\boxed{H(S_{v_B}) = 0.81}$$

$$H(S_{v_C}) = -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6}$$

$$\boxed{H(S_{v_C}) = 1}$$

$$\text{Gain}(S, t_1) = 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

$$\boxed{\text{Gain}(S, t_1) = 0.049}$$



In second feature set

Let Suppose

$$\boxed{\text{Gain}(S_2, t_2) = 0.051}$$

$$\boxed{\text{Gain}(S_2, t_2) > > \text{Gain}(S, t_1)}$$

So select 2nd feature because it have high Information value.