Written Part-

Date __/__/____

Saathi

Written part

1. Gini Impurity Index

Steps -
1. Find Gini Impurity
2. Find Best Split
3. Make tree

1. $Gini(D) = 1 - \sum_{i=1}^{n} P_i^2$, n = classes, P_i = probability of class

$= 1 - (P(IS)^2 + P(IV)^2)$ IS = Iris setosa, IV = Iris virginica

$= 1 - \left(\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right) = 0.5$

2. Checking Best split for each feature

a. Sepal Length - median = $\frac{5.1 + 5.8}{2} = 5.45$

Left - 3 IS, 0 IV
Right - 0 IS, 3 IV

Gini (Left) $= 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right) = 0$

Gini (Right) $= 1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right) = 0$

Gini split = 0

b. Sepal width - median = $\frac{3.0 + 3.2}{2} = 3.1$

Left : 1 IS, 2 IV
Right : 2 IS, 1 IV

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

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

Gini split $= \frac{3}{6} * 0.44 + \frac{3}{6} * 0.44 = 0.44$

c. Petal Length - median = $(1.4 + 5.1)/2 = 3.25$

Left : 3 IS, 0 IV
Right : 0 IS, 3 IV

Gini (left) : $1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right) = 0$ $\Rightarrow$ Gini split = 0

Gini (Right) : $1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right) = 0$

Page No.

Date __/__/____

Saathi

d) Petal width : ⟶ median $= \left(\frac{0.2+1.9}{2}\right) = 1.05$

left : 3 IS , 0 IV

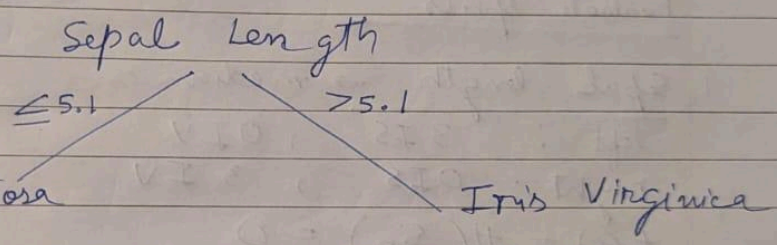
Right : 0 IS , 3 IV

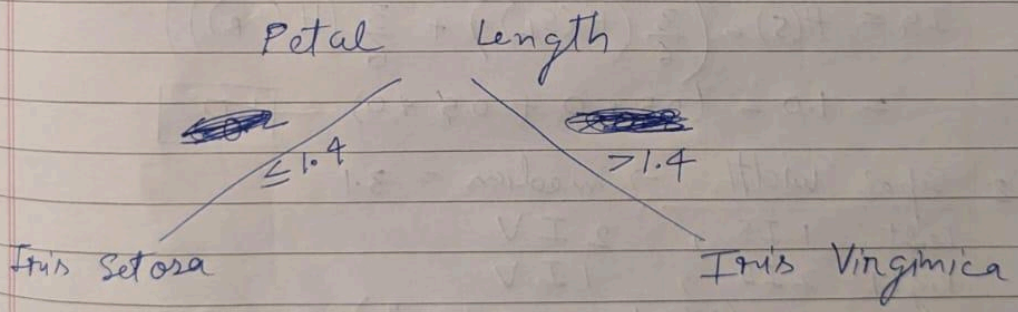
Gini (left) $= 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right) = 0 \Rightarrow$ **Gini split = 0**

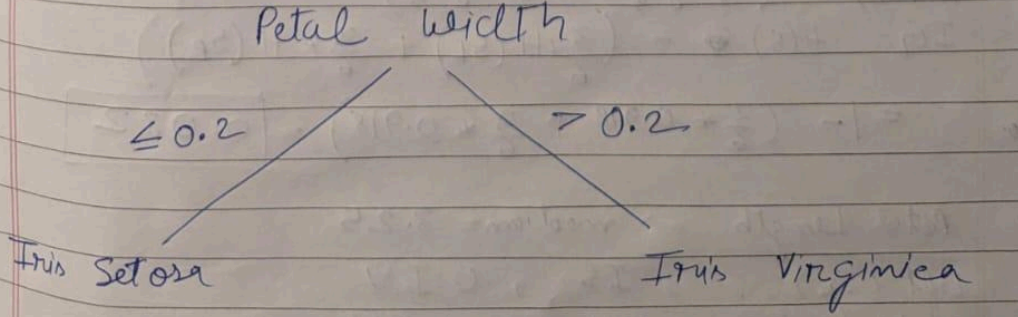
Gini (Right) $= 1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right) = 0$

So, we get gini impurity = 0 for features Sepal Length, Petal length and Petal width, all 3 can be used for optimal split.

Trees -







Date ___/___/___

2. Information Gain

Entropy (H): $H(S) = -\sum_{i=1}^{c} P_i \log_2 (P_i)$

$IG(S, A) = H(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} H(S_i)$

Initial entropy:

$H(S) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right)$

$= -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1.0$

Evaluate splits

1. Sepal length → median =

left : 3 IS , 0 IV

right : 0 IS , 3 IV

$H(S_L) = H(S_R) = 0$

$IG = H(S) - \left(\frac{3}{6}(H(S_L)) + \frac{3}{6}(H(S_R))\right)$

$= 1.0 - (0.5 * 0 + 0.5 * 0) = \boxed{1.0}$

2. Sepal width → median = 3.1

Left: 1 IS , 2 IV

Right: 2 IS , 1 IV

$H(S_L) = -\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right) = 0.918$

$H(S_R) = -\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right) = 0.918$

$IG = H(S) - \left(\frac{3}{6} * H(S_L) + \frac{3}{6} * H(S_R)\right)$

$= 1 - \left(\frac{3}{6} * 0.918 + \frac{3}{6} * 0.918\right) = \boxed{0.082}$

3. Petal Length → median = 3.25

Left (S_L) : 3 IS , 0 IV

Right (S_R) : 0 IV , 3 IV

$H(S_L) = H(S_R) = 0$

$I_G = H(S) - \left(\frac{3}{6} H(S_L) + \frac{3}{6} H(S_R)\right)$

$\boxed{I_G = 1.0}$

Page No.

Date __/__/____ Saathi

1 Petal width → median = 1.05

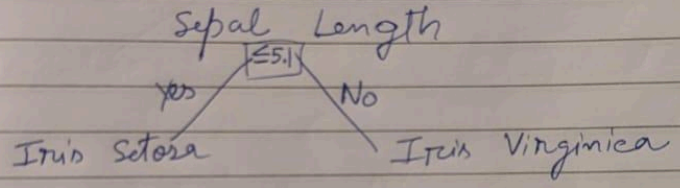
Left (S_L): 3 IS , 0 IV

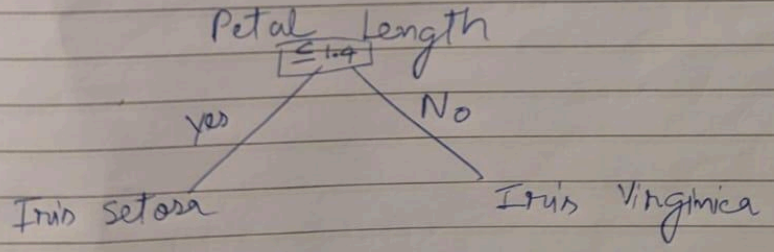
Right (S_R): 0 IS , 3 IV

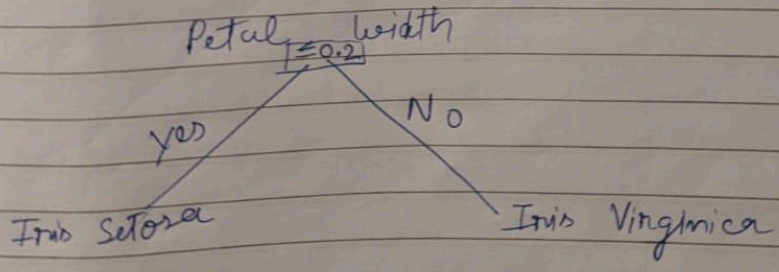
$H(S_L) = H(S_R) = 0$

$IG = H(S) - \left(\frac{3}{6} H(S_L) + \frac{3}{6} H(S_R)\right) = 1.0$

Since, we're getting Information Gain = 1.0 for features Sepal Length, ~~Sepo~~ Petal Length, Petal width, we can use any of these for optimal split and tree creation.







Saathi

Date ___/___/_____

Features	Split Value	Information Gain	Gini
Petal Length	1.4	1.0	0.5
Petal Width	0.2	1.0	0.5
Sepal Length	5.1	1.0	0.5

All 3 features provide perfect class separation; Max IG = 1.0 and Max Gini Gain 0.5 and Pure nodes (Gini = 0 for both children). So any of 3 features is equally valid choice.