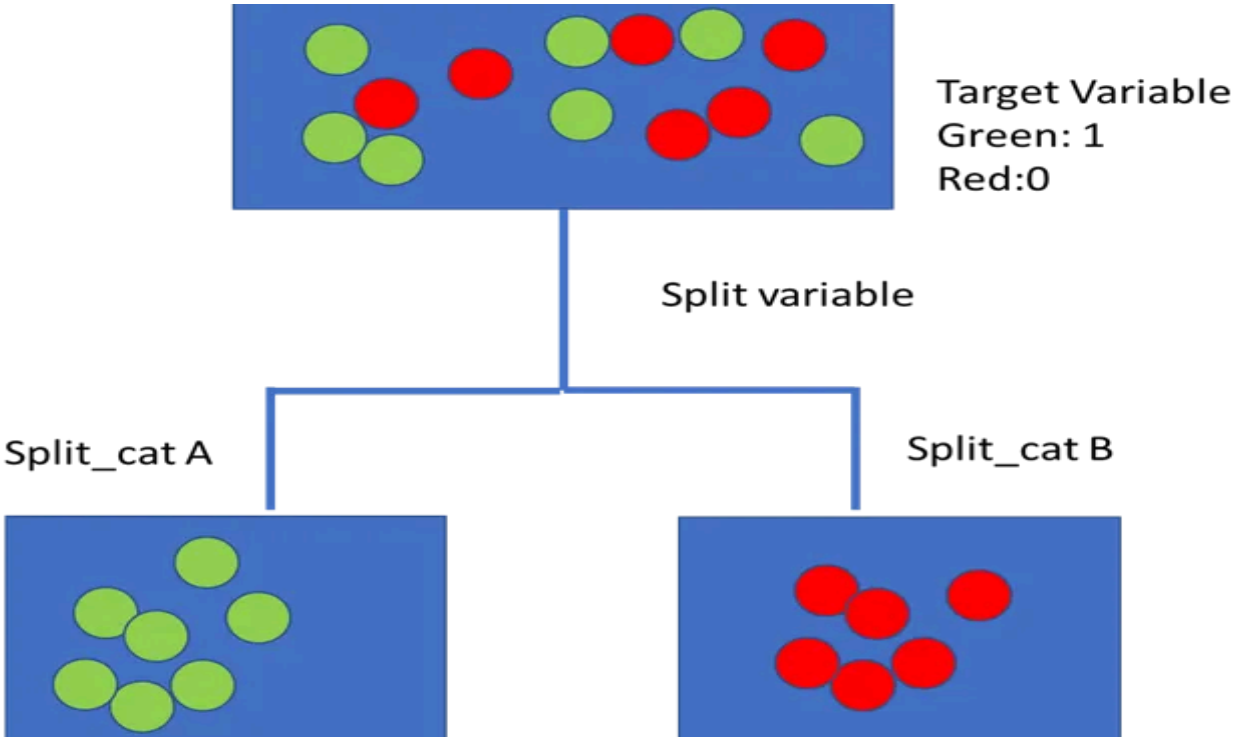
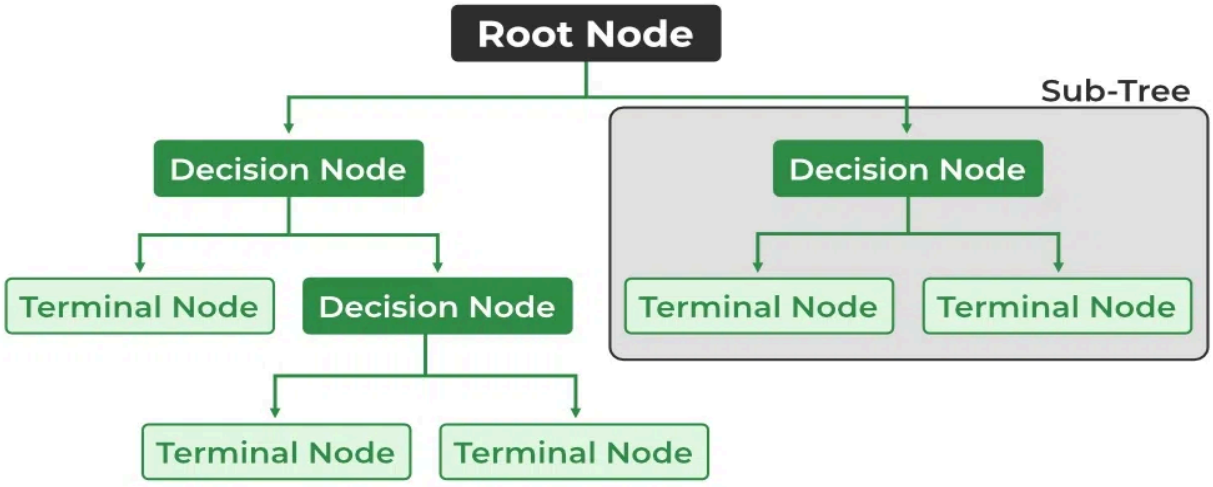
Decision Tree

- A decision tree is a flowchart-like tree structure where each internal node denotes the feature, branches denote the rules and the leaf nodes denote the result of the algorithm.
- Decision tree is a hierarchical data structure that represents data through a divide and conquer strategy.
- A decision tree is a simple model for supervised classification. It is used for classifying a single discrete target feature.

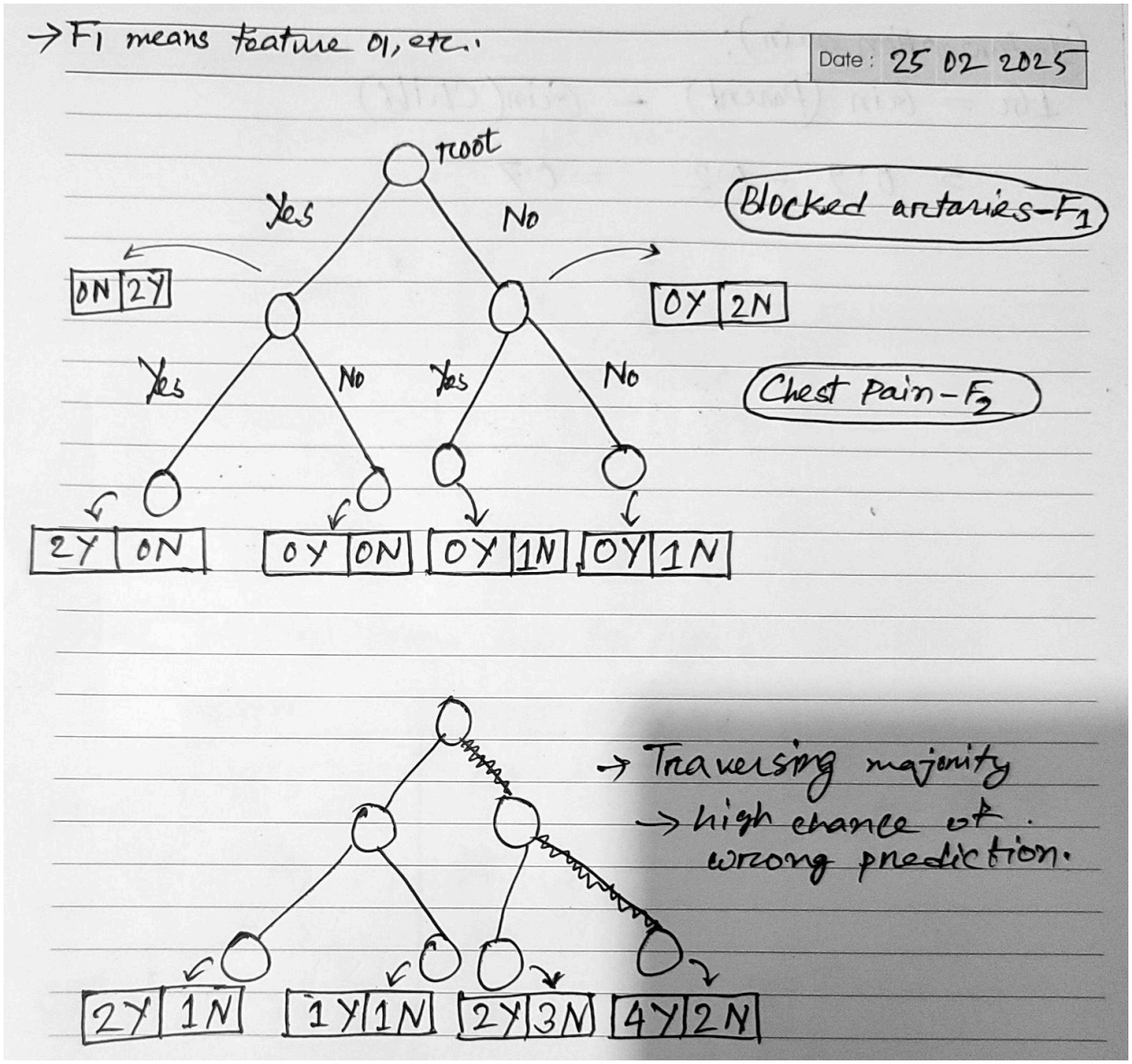
What does a Decision Tree do?

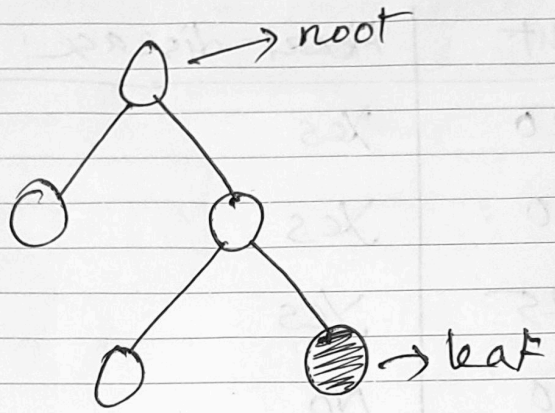


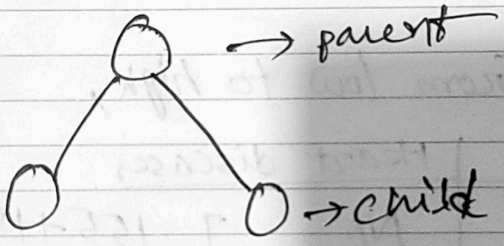
Decision Tree Look like:



X


→ F1 means feature 01, etc.
Date : 25 02 2025
root
Yes
No
Blocked arteries - F1
0N 2Y
0Y 2N
Yes
No
Yes
No
Chest Pain - F2
2Y 0N
0Y 0N
0Y 1N
0Y 1N
→ Traversing majority
→ high chance of wrong prediction.
2Y 1N
1Y 1N
2Y 3N
4Y 2N





→ Decision tree's prediction comes from leaf node..

Impurity:

for each feature (calculation)

↓

lowest impurity feature (good sign)

↓

make decision tree

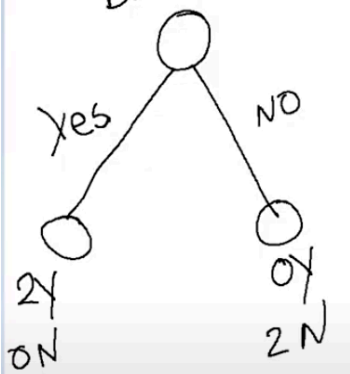
Gini = $1 - \Sigma p(i)^2$ (usally used)

Entropy = $- \Sigma\ pi\ \log_2(Pi)$

chest pain	good blood circulation	Blocked arteries	Heart disease
NO	NO	NO	NO
yes	yes	yes	yes
yes	yes	NO	NO
yes	NO	yes	yes

weight	Heart disease
220	Yes
180	Yes
225	Yes
190	NO
155	NO

Blocked arteries



left

$P_{yes} = \frac{2}{2} = 1$

$P_{NO} = \frac{0}{2} = 0$

$Gini = 1 - 1^2 - 0^2$

$= 0$

Right

$P_{yes} = \frac{0}{2} = 0$

$P_{NO} = \frac{2}{2} = 1$

$Gini = 1 - 0^2 - 1^2$

$= 0$

$$Gini = \left(\frac{3}{4}\right) 0.455 + \frac{1}{4} 0$$

$$= \boxed{0.33}$$ perfect

$$Gini = \frac{2}{4}(0) + \frac{2}{4}(0)$$

$$= 0$$

Decision based on Information Gain:

Decision Trees:

- Think of a decision tree like a flowchart, where each decision leads to more decisions, and finally, to an outcome. In each step, we make a decision based on certain features to get closer to our goal.

Entropy - Measure of Disorder:

- Entropic Playground: Entropy is like a measure of messiness or disorder in our data.
- Low Entropy: Low entropy means our data is more organized or homogeneous.
- High Entropy: High entropy means our data is a bit messy or diverse.

Information Gain - Seeking Order:

- Information Gain is our guide. It helps us decide which features bring more order to our data. We pick the feature that reduces the chaos in our dataset the most.

Classification or Regression:

Once the decision tree is built, it can be used for classification (for categorical target variables) or regression (for numerical target variables). Each path from the root to a leaf node represents a decision rule.

chest pain	good blood circulation	Blocked arteries	Heart disease
NO	NO	NO	NO
yes	yes	yes	yes
Yes	Yes	NO	NO
Yes	NO	Yes	Yes

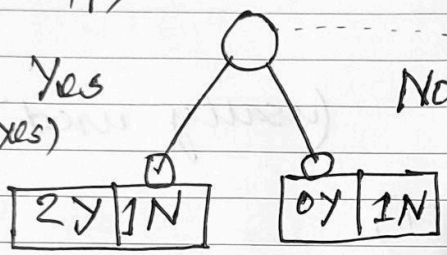
weight	Heart disease
220	Yes
180	Yes
225	Yes
190	NO
155	NO

Feature: Chest pain, good blood circulation, Blocked arteries

Target: Heart disease

Dataset Split:

F1 ⇒ Chest pain



~~Gini = 1 − Σ(Pi)²~~

∴ Gini = $1 - \Sigma P(i)^2$

∴ Gini = $1 - \Sigma P(i)^2$

Left node:

$P_{yes} = 2/3 = 0.66$

$P_{NO} = 1/3 = 0.33$

∴ $Gini_L = 1 - (0.66)^2 - (0.33)^2$

$= 0.455$

Right node:

$P_{yes} = \frac{0}{1} = 0$

$P_{NO} = \frac{1}{1} = 1$

∴ $Gini_R = 1 - (0)^2 - (1)^2$

$= 0$

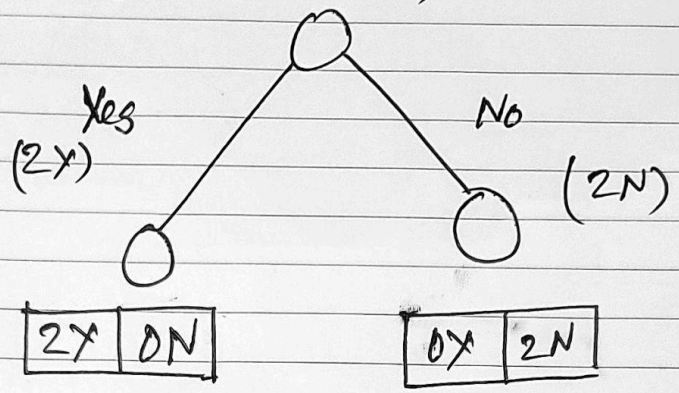
∴ Gini = Gini (Left) × sample + Gini (Right) × sample

$= 3/4 \times 0.455 + 0 \times 1/4$

$= 0.34125$

Date :

(Blocked arteries) t_2



Left:

Gini =

$P_{yes} = 2/2 = 1$

$P_{NO} = 0/2 = 0$

$\therefore Gini_L = 1 - (1)^2 - (0)^2$

$= 0$

Right:

Gini =

$P_{yes} = \frac{0}{2} = 0$

$P_{NO} = \frac{2}{2} = 1$

$\therefore Gini_R = 1 - (0)^2 - (1)^2$

$= 0$

$\therefore Gini_P = 2/4 \times 0 + 2/4 \times 0$

$= \boxed{0}$

For Numerical value,

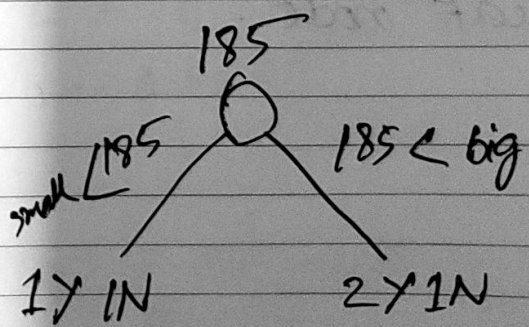
Date :

Weight	Heart disease
220	Yes
180	Yes
225	Yes
190	No
155	No

After sorting from low to high,

	Weight	Heart disease
167.5	155	No
	180	Yes
185	190	No
205	220	Yes
222.5	225	Yes

$\frac{155+180}{2} = 167.5$



যেই Average value এর জন্য Gini এর মান কম আসবে, তাকে root এ the scheme draw করতে হবে।

Decision Tree Colab:

https://colab.research.google.com/drive/1Kvm_faeSyZOfKop6WuSrQ_BTpChS2xrw?usp=sharing
3. SVM classifier, Decision Tree Regressor-Classification.ipynb