GINI IMPURITY IN DECISION TREES

- BY CLOUDYML

* GINI IMPURITY:

→ The Gini Impurity measure is one of the measures/methods used in decision tree algorithms to decide the optimal split from a root node, and subsequent splits.

→ Gini Impurity tells us what is the probability of misclassifying an observation.

Gini Index is calculated as follows:

$$G_I = \sum_{i=1}^{K} p_i (1 - p_i)$$

p_i = probability of finding a point with label i.

K = numbes of classes

EXAMPLE:

Consider a table having columns:

SEX	CHOLESTROL	HEART DISEASE

Sex(1) = Male Heart disease(0) = NO
Sex(0) = Female Heart disease(1) = YES

Suppose you are given the data for 100 patients and the target variable has two classes: class 0 having 60 people and class 1 having 40 people.

→ Expressing this in terms of probabilities:

$$P_0 = \frac{60}{60+40} = 0.6 \qquad P_1 = \frac{40}{60+40} = 0.4$$

→ Calculate the Gini Index before making any splits:

$$= P_0(1-P_0) + P_1(1-P_1) = 0.6(1-0.6) + 0.4(1-0.4) = \underline{0.48}$$

→ Information about data:

Features/ Classes	Sex		Total
	M	F	
No disease	50	10	60
Disease	20	20	40
Total	70	30	100

Features/ Classes	Cholestrol		Total
	<250	>250	
No disease	50	10	60
Disease	10	30	40
Total	60	40	100

→ ## Split Based on SEX

Probabilities of two classes within MALE subset:

$$P_0 = \frac{50}{50+20} = 0.714 \qquad P_1 = \frac{20}{70} = 0.286$$

Gini Impurity for Males:

$$0.714(1-0.714) + 0.286(1-0.286)$$
$$= 0.41$$

Probabilities of two classes within FEMALE subset:

$$P_0 = \frac{10}{10+20} = 0.333 \qquad P_1 = \frac{20}{10+20} = 0.667$$

Gini Impurity for Females:

$$0.333(1-0.333) + 0.667(1-0.667)$$
$$= 0.44$$

$$P_{male} = \frac{70}{100} = \underline{0.7} \qquad P_{female} = \frac{30}{100} = \underline{0.3}$$

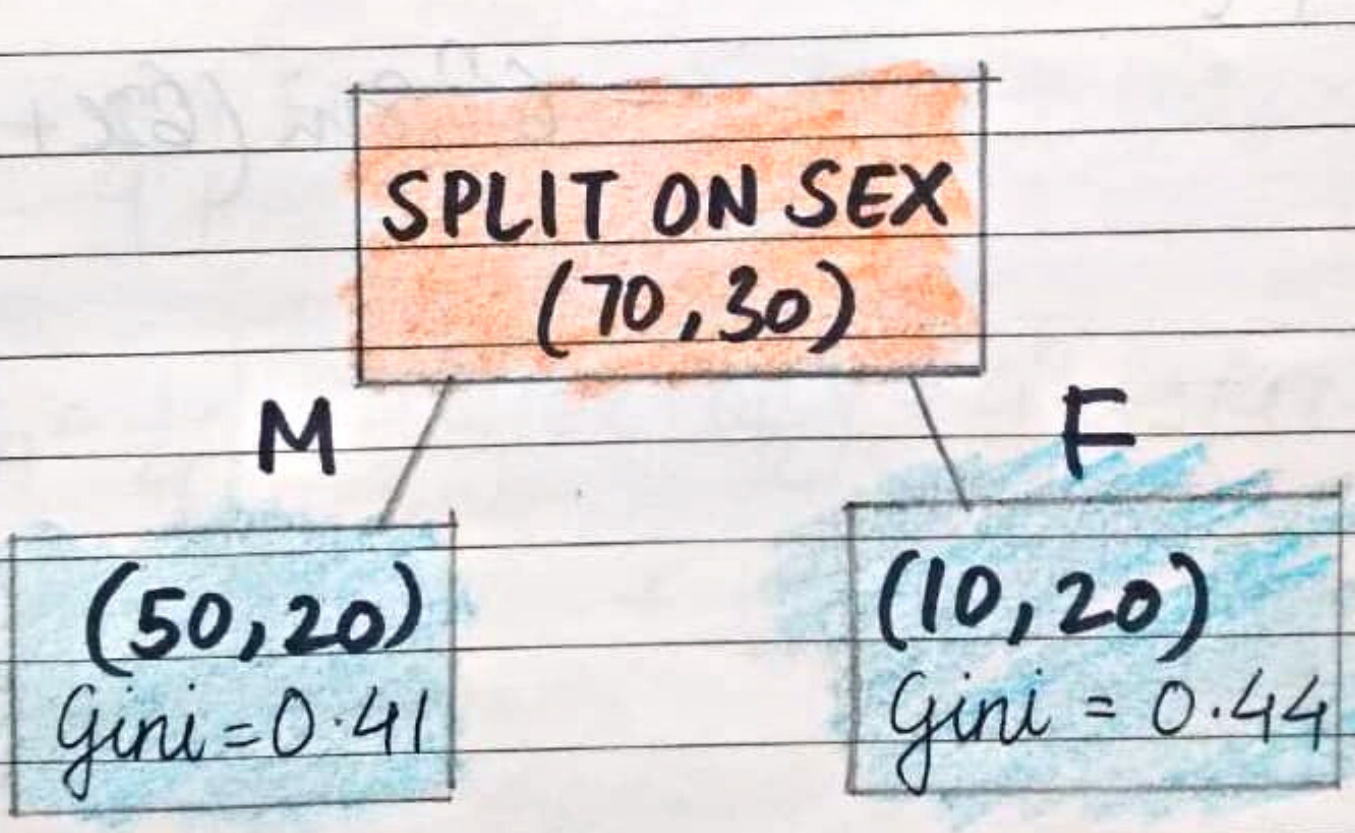
(5)

Gini Impurity after the split based on Sex :

$$0.7 \times 0.41 + 0.3 \times 0.44 = 0.42$$

Gini Impurity before split : 0.48
Gini Impurity after split : 0.42

Reduction in Gini Impurity: 0.06



$$\Delta Gini_{sex} = \underline{0.06}$$

→ Split Based on Cholesterol

Probabilities of two classes within the low cholesterol (<250) subset:

$$P_0 = \frac{50}{50+10} = 0.833 \quad P_1 = \frac{10}{50+10} = 0.167$$

Gini impurity for low cholesterol:

$$0.833(1-0.833) + 0.167(1-0.167) \approx 0.27$$

Probabilities of two classes within the high cholesterol (>250) subset:

$$P_0 = \frac{10}{30+10} = 0.25 \quad P_1 = \frac{30}{30+10} = 0.75$$

Gini impurity for high cholesterol:

$$0.25(1-0.25) + 0.75(1-0.75) \approx 0.37$$

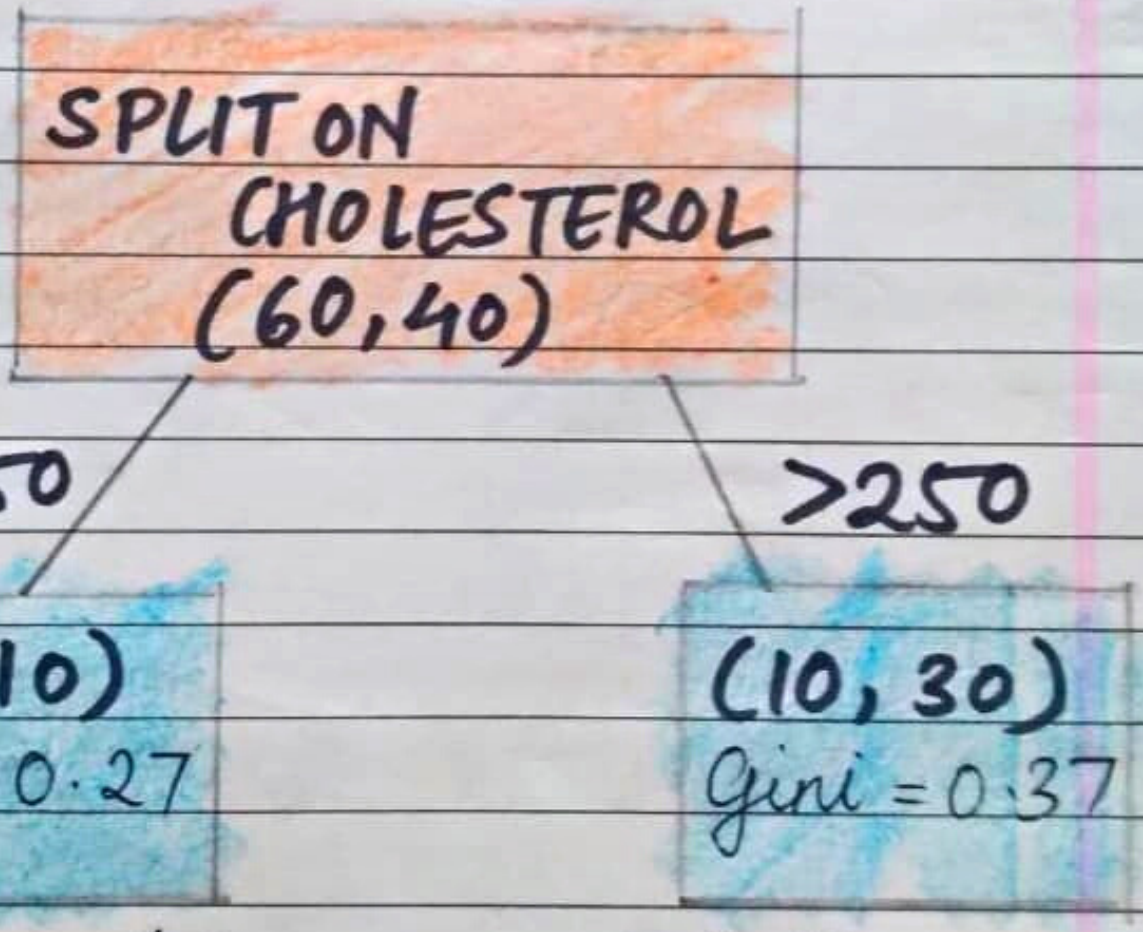
$$P_{low} = \frac{60}{100} = 0.6 \quad P_{high} = \frac{40}{100} = 0.4$$

Gini impurity after the split based on Cholesterol:

$0.6 \times 0.27 + 0.4 \times 0.37 \approx 0.3$

Gini impurity before split = 0.48
Gini impurity after split = 0.3

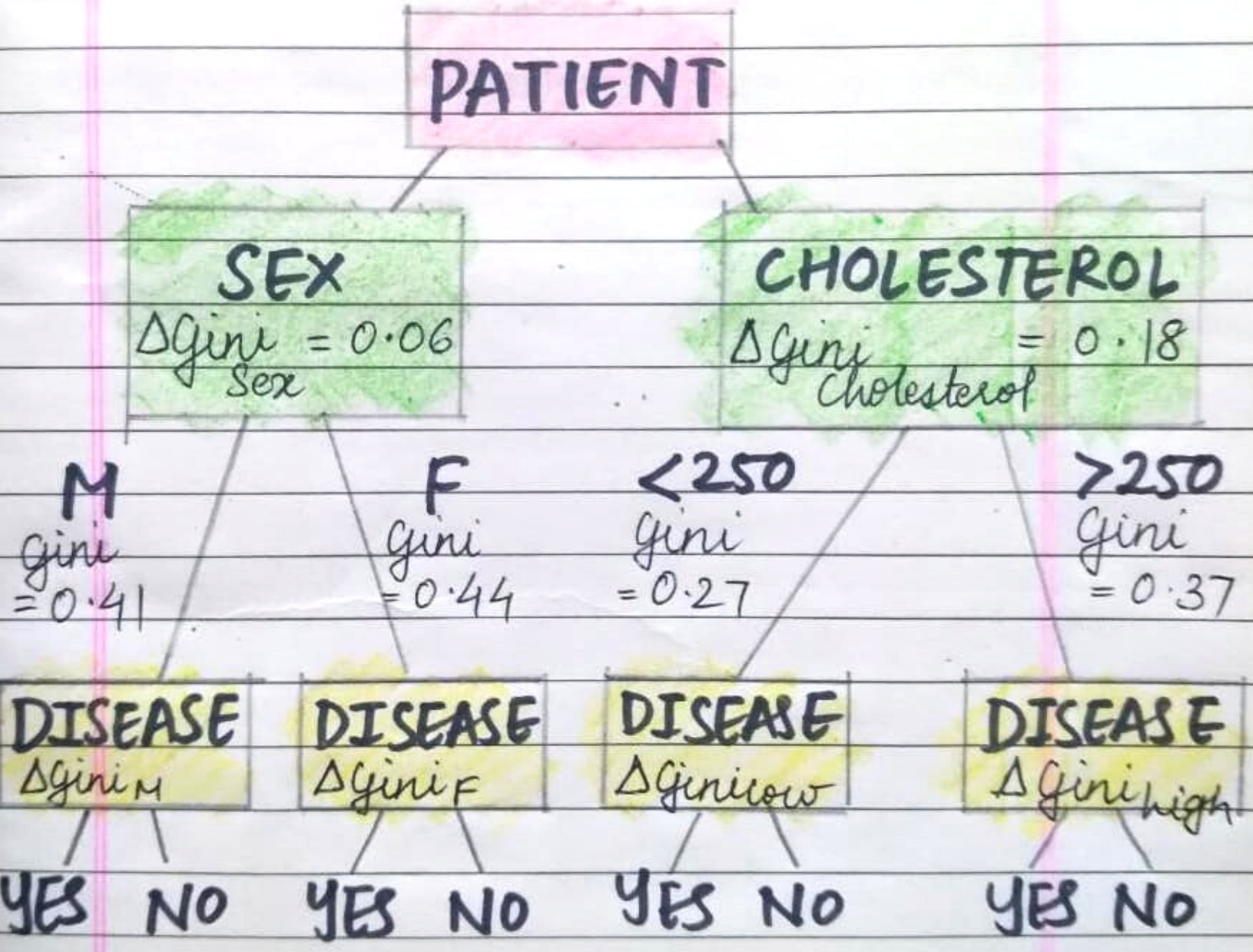
Reduction in Gini impurity = 0.18



$\Delta Gini_{cholesterol} = \underline{0.18}$

* $\Delta Gini_{cholesterol} > \Delta Gini_{sex}$
∴ The tree is splitted based on Cholesterol.

→ The final decision or predicting the value of target variable can be done by following the same procedure ahead considering whether the patient has the disease or not.



* The branches giving the highest reduction in Gini Impurity will be considered for splitting the tree further.

CLOUDYML

TRUSTED BY 7500+ HAPPY LEARNERS

MAKE YOUR CAREER IN DATA SCIENCE & ANALYTICS

- Hands-on Practical Learning
- 1-1 Doubt Clearance Support
- Capstone End-to-End Projects

Python
SQL
Statistics
Machine Learning
Deep Learning
ML Projects
DL Projects
AWS Sagemaker
Tableau
Excel
PowerBi
Amazon QuickSight
Google Data Studio
Interview QnA PDF

Complete Package Is Available @ Just ~~₹24885~~ **₹6999**

**GST ALREADY INCLUDED

Link Given in Comments Section

Visit CloudyML Website Now