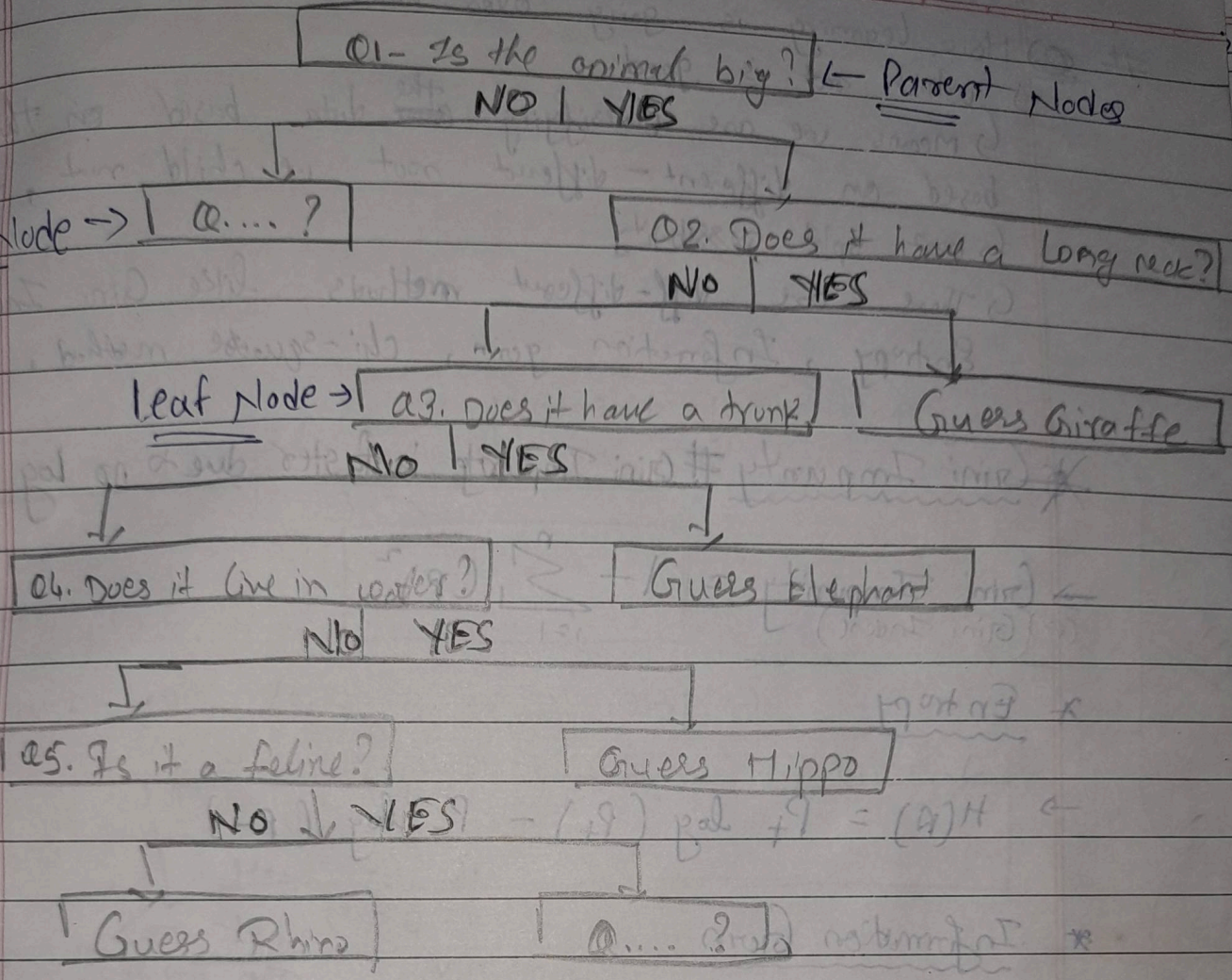
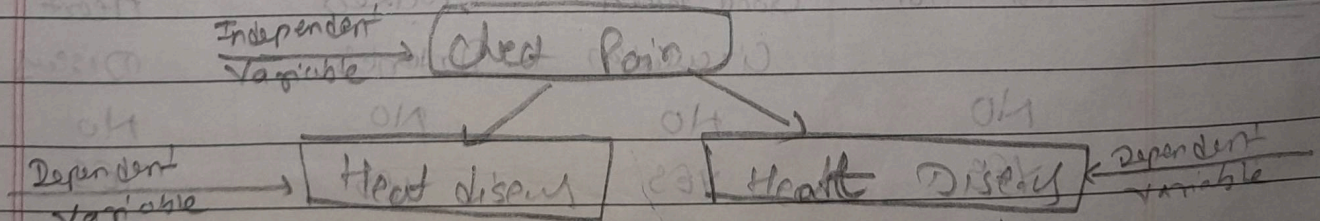
Decision Tree

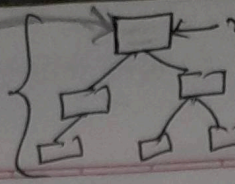


- ~~Information Gain~~ = (base entropy) − (new entropy)
- Gini ~~Index~~ Impurity = 1 − (the Probability of "yes")^2 − (the Pro "no")^2



$$= 1 - \left(\frac{105}{105+39}\right)^2 - \left(\frac{39}{105+39}\right)^2$$

$$= 0.395$$



Q) How learning is going over here?

⇒ which root is good

↳ Means we are classifying the data based on the root, based on different - different root i.e child root, root.

↳ best split possibility

↳ There are diff - different methods like Gini Indexing, Entropy, Information gain, chi-square method.

← gini is half of entropy

* Gini Impurity # Gini Impurity is faster due to no log calculation

Formula

→ Gini Impurity = $1 - \sum_{i=1}^{n} (P)^2$ [Probability]

(or) Gini Index

* Entropy

→ $H(A) = P_+ \log(P_+) - P_- \log(P_-)$

* Information Gain

→ $H(S) - \sum \frac{|S_v|}{|S|} \times H(S_v)$

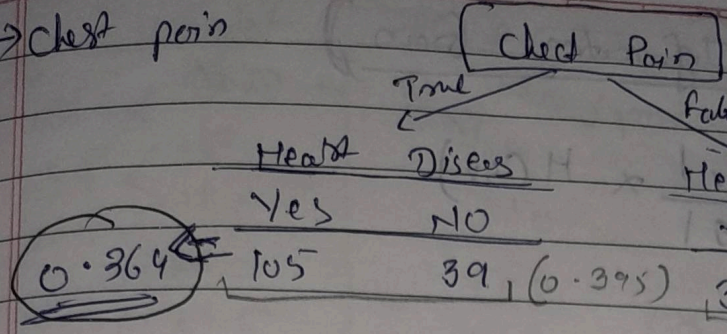
Eg

Dataset ⇒

chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
NO	NO	NO	No
YES	YES	YES	YES
YES	YES	NO	NO
YES	NO	???	YES
etc...	etc..	etc...	etc.

To determine which seperation is best, we need a way to measure and compare "Impurity"

→ chest pain



Calculation

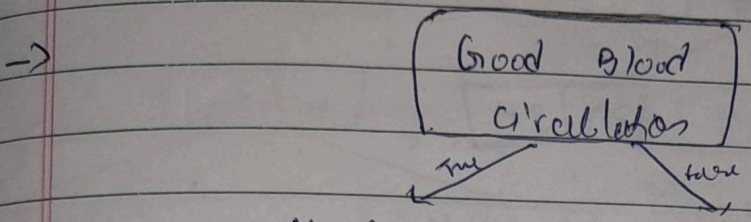
Gini Impurity = $1 - \sum_{i=1}^{n} p^2$

$= 1 - \left(\frac{105}{105+39}\right)^2 - \left(\frac{39}{105+39}\right)^2$

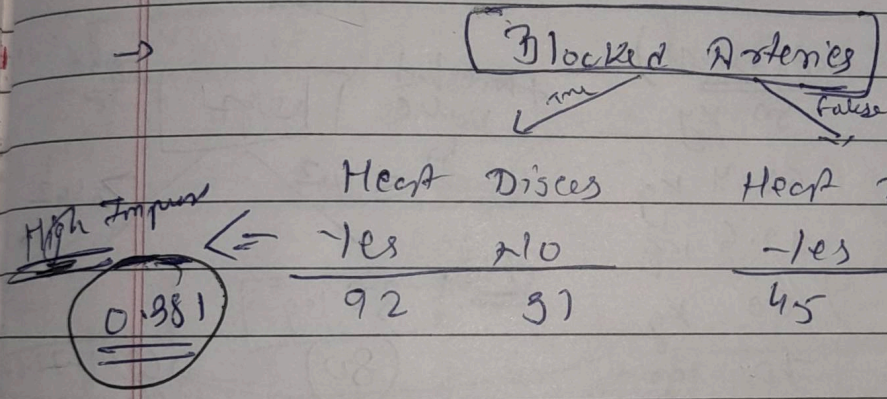
$= 0.395$

then

→



→



Gini Impurity for chest Pain = weighted avg of Gini impurities for the leaf node

$= \left(\frac{144}{144+159}\right) 0.395 + \left(\frac{159}{144+159}\right) 0.336$

$= 0.364$ //

Step 1 ⇒ Calculate Gini Impurity

Step 2 ⇒ weighted avg of Gini Impurity

① Gini Impurity = $1 - \sum_{i=1}^{n} p^2$

② Gini impurity for chest pain = weighted avg of Gini Impurity for the leaf nodes

Pre-Pruning ⇒ Pre-Pruning = we are not reaching to the leaf node in-between, we are going to stop a construction of decision tree.

(It is hyperparameter)

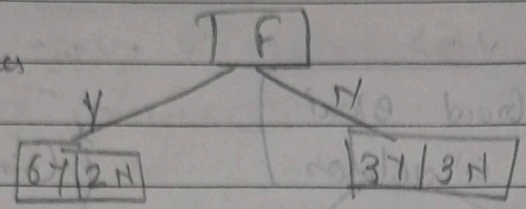
Date :
Page :

* Explanation of (Information Gain)

Root Entry → Sample → 1st level

$$H(S) - \sum_{i=1} \frac{|S_v|}{|S|} \times H(S_v)$$

Calculate using entropy formula



DTR

Step1 → Sort

$$H(S) - \left(\frac{8}{14} \times H(S) + \frac{6}{14} (H S)\right)$$

DTR

Sort (weight)	Height	weight	Sort (weight)
4.2	5.4	80 kg	80 kg
4.3	5.6	70 kg	68.4 kg
4.4	5.8	75 kg	72.6 kg
5.4	6.1	72 kg	60 kg
5.6	4.3	68.4 kg	70 kg
5.8	4.4	72.6 kg	75 kg
6.1	4.2	80 kg	72 kg

mean of adjacent

threshold value → height

≤ 4.2 → 80 kg (80)

> 4.2 → 68.4+72.6+60+70+75+72 / 6 (find mean of these value) (67)

Then find MSE (or) MAE

\# * How we can find MSE?
→ w.r.t Each and Every Data Points we will try to calculate mse.

(80-80) (80-68.4) (80-72.6) (80-60) (80-70) (80-75) (80-72)

Calculation for MSE $\sum (y - \hat{y})^2$

Steps ⇒ ① sort height → lowest to highest
② Calculate the Average (mean) weight for all adjacent

The less mse the best node selection

DTC

Weight	Heart Disease
155	NO
180	YES
190	NO
220	YES
225	YES

threshold value

① Weight
- ≤ 167.5 → 0Y | 1N
- > 167.5 → 3Y | 2N

②
- ≤ 185 → 1Y | 1N
- > 185 → 2Y | 1N

③
- ≤ 205 → 2N | 1Y
- > 205 → 2Y | 0N

④
- ≤ 222.5 → 2Y | 2N
- > 222.5 → 0N | 1Y

Now we can calculate
① Gini impurity
② Entropy

Cloth size (X)	Height (Y)
M	5.4
M	5.6
L	6.1
M	6.2
L	6.3
L	4.3
M	4.1
M	5.3

X
- M → 5.4 + 5.6 + 6.2 + 4.1 + 5.2 → mean → 5.44
- F → 6.1 + 6.3 + 4.3 → mean → 4.9

* Calculate MSE value

$(5.4-5.44)^2 + (5.6-5.44)^2 + (6.2-5.44)^2$

$(4-4.9)^2$

CART ⇒ Gini Index (By Default) Impurity we are going to build DT (Classification Regression Tree, based on Gini)

ID3 ⇒ Information Gain (Based on Information Gain we are going to build DT)

Interview Ques

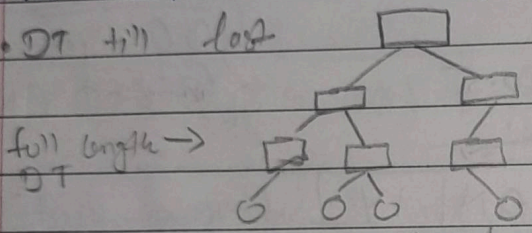
DT → overfitting, train ✓ test ✗, low bias, high variance

Hyperparameters of DT

→ Criterion : {"gini", "entropy"}, default = "gini"
(Gini Impurity / for Information Gain)
minimum Gini Impurity we are going to consider

→ Splitter : {"best", "random"}, default = "best"
used to choose the split at each node.
best → to choose the best split
random → to choose the best random split

→ max_depth : int, default = None
• DT till last



• full length — two concept comes into picture → ① Pre-pruning ② Post-pruning

→ min_sample_split : int or float, default = 2

→ min_sample_leaf : int or float, default = 1

→ max_leaf_node

→ max_features

• Splitting :- It is a process of dividing a node into two or more sub-nodes.

• Pruning :- when we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of Splitting.