Decision_trees_2

05 March 2025 20:36

Gini Impurity

→ Formula: $G(y) = 1 - \sum [P(y_i)]^2$

<u>Background</u>: If our dataset has 'd' features & 'n' datapoints then:

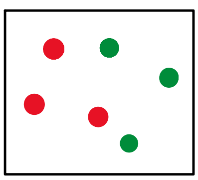
① Calculate Entropy for each feature (eg, Gender | Education)

② Find out Information Gain for each feature

③ Chose the feature with maximum IG to split our dataset

→ How to compute Gini Impurity in our example?

$$G(y) = 1 - \sum [P(y_i)]^2 \longrightarrow G(y) = 1 - \left[P(y-g)^2 + P(y-r)^2 \right]$$

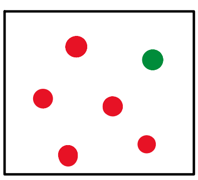


$P(-ve) = 1/2$

$P(+ve) = 1/2$

$$G(y) = 1 - \left[(0.5)^2 + (0.5)^2\right] = 1 - 0.5$$

$$G(y) = 0.5$$

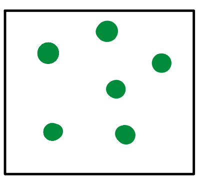


$P(-ve) = 5/6$

$P(+ve) = 1/6$

$$G(y) = 1 - \left[\left(\frac{5}{6}\right)^2 + \left(\frac{1}{6}\right)^2 \right] = 1 - \left[\frac{26}{36}\right] = \frac{36-26}{36}$$

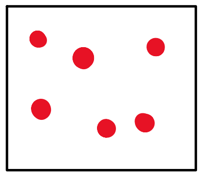
$$G(y) = 10/36 = 0.28$$



$P(-ve) = 0$

$P(+ve) = 1$

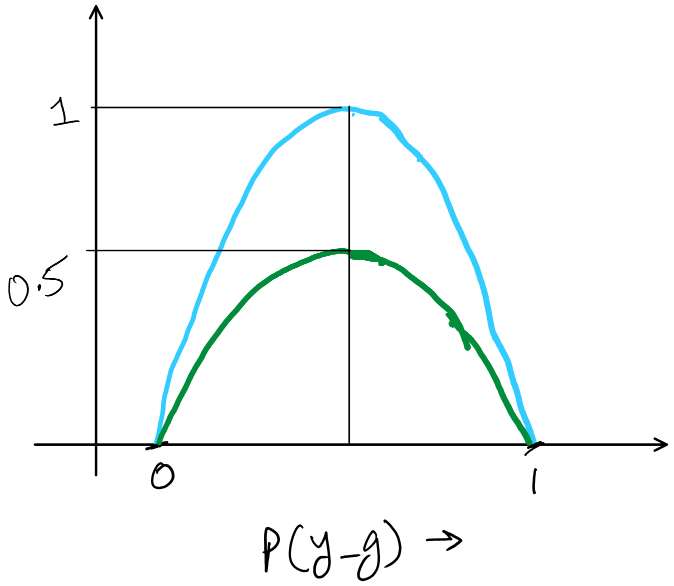
$$G(y) = 1 - [0^2 + 1^2] = 0$$



$P(-ve) = 1$

$P(+ve) = 0$

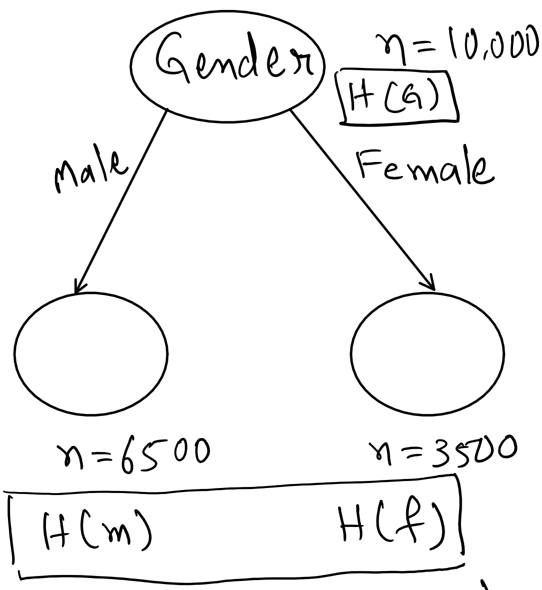
$$G(y) = 1 - [1^2 + 0^2] = 0$$



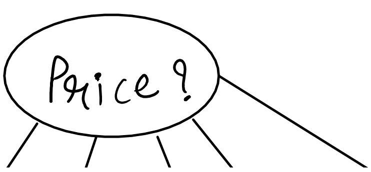
☆ **Can we use this same approach to numerical columns? Ans - No. Why?**

In case of categorical columns:

Gender	Education
male	Non Grad
female	Non Grad
female	Grad
female	Non Grad
male	Grad

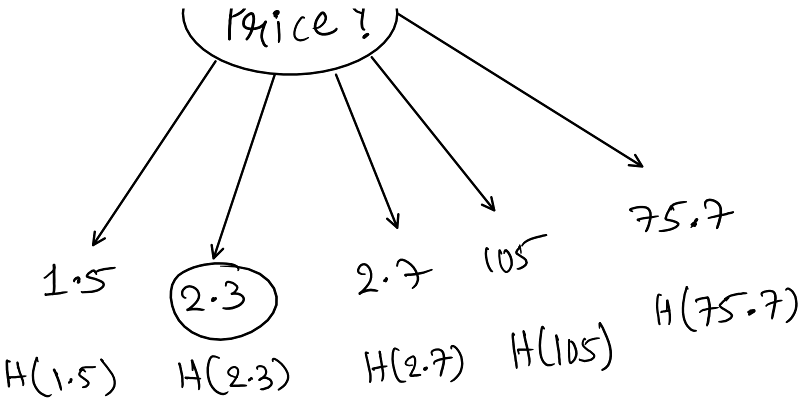


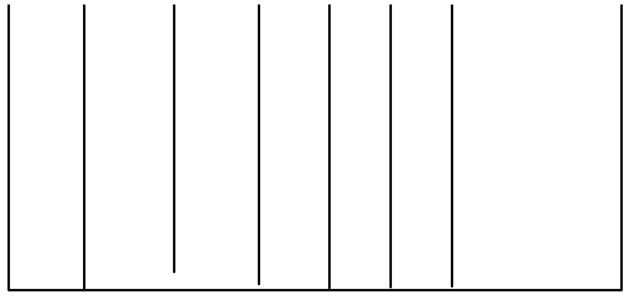
For a numeric column:



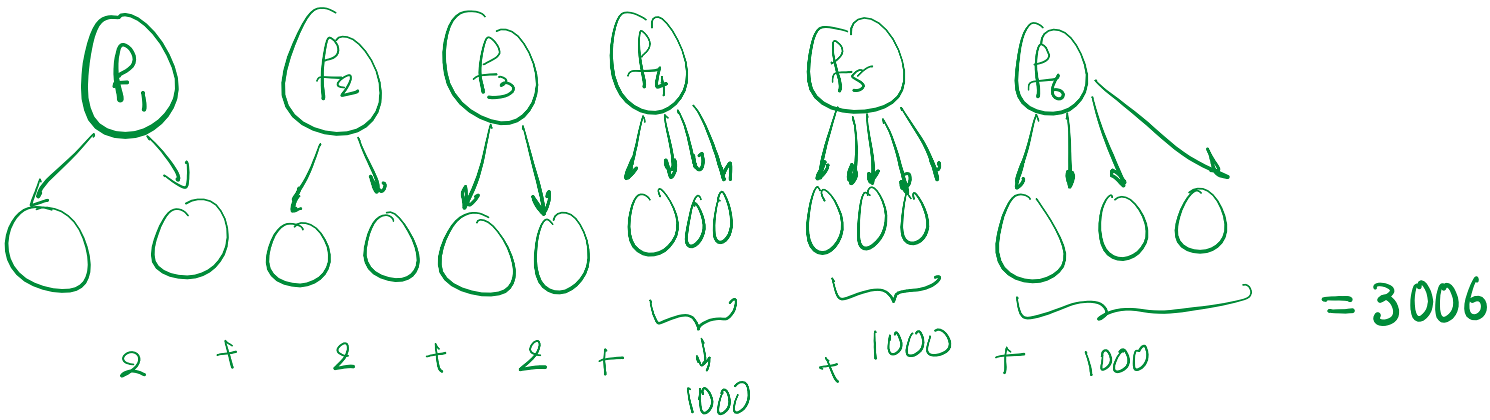
categorical — Numerical with 1000 unique vals each

f_1	f_2	f_3	f_4	f_5	f_6	





How many Entropy calculations are needed ?



Steps – ① Sort the data in ascending order of numerical column

② For each unique value, calculate Entropy

③ Compute IG for all the thresholds

④ Find the question with maximum IG.

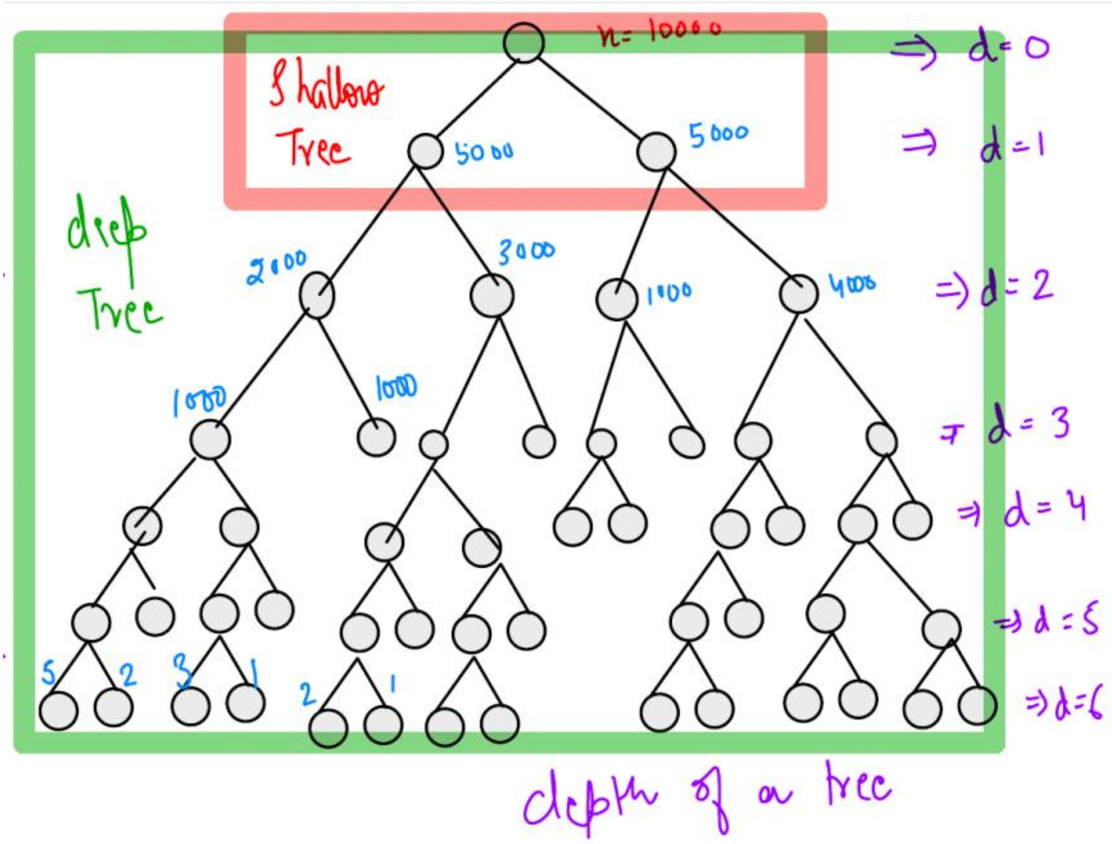
Disadvantage –

A lot expensive computationally.

Then how can we find Entropy of numerical columns?

Ans: Creating bins on the numerical column and calculating Entropy of each bin rather than calculating for each unique value.

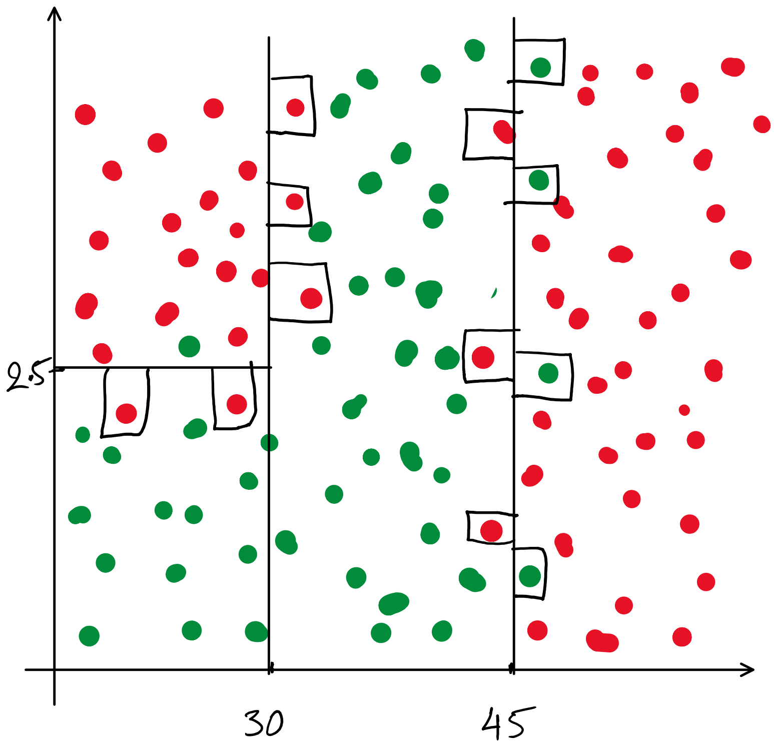
☆ An entire view:



In context of Underfitting & Overfitting:

Underfitting:
→ The model doesn't learn enough.

Overfitting:
→ Our model tries to learn each datapoint



← Each question creates an axis-parallel boundary.

Very Shallow Tree leads to Underfitting

Extremely Deep trees lead to overfitting.

→ ∴ Won't it be a good idea to control 'd'?

Hence, d is one of the Hyperparameter while implementing Decision Trees

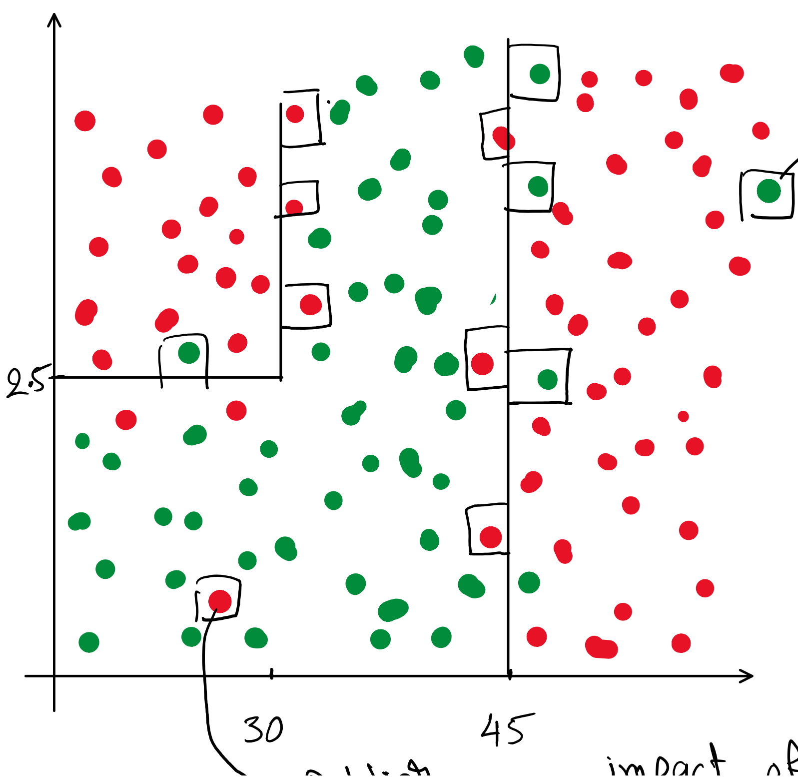
Hence, d is one of the ...

implementing Decision Trees.

$d\uparrow\uparrow \rightarrow$ overfitting $d\downarrow\downarrow \rightarrow$ Underfitting.

Depth ('d')	Training Accuracy	Validation Accuracy	Comments
1	↓↓	↓↓	underfit
2	↓↓	↓↓	
3	↓↓	↓↓	
5	↑	↑	Best accuracy at validation (Good choices of d)
7	↑	↑	
10	↑	↑	
50	↑↑	↓	Overfit
100	↑↑	↓	

✪ Do outliers impact Decision Tree?

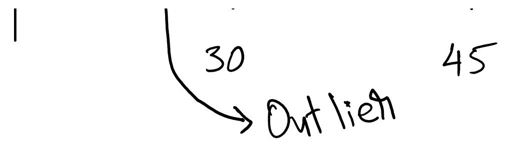


$d = 1$

→ for shallow trees, impact of outliers is negligible

→ For extreemly deep trees, impact of outliers will be significantly high.

→ For 'best' values of d impact of outliers will be there but



→ For 'best' values of d impact of outliers will be there but not very significant.

☆ Do we need feature scaling in decision trees? (Normalisation | Standardisation)

Ans – No. Why?

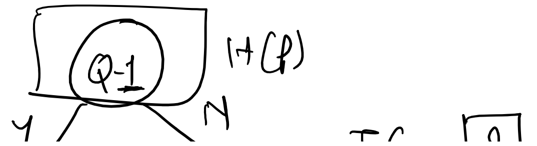
1. We divide the data by asking a question that splits the data into two parts instead of finding distances of each point and hence, we don't need to shrink the numbers.
2. While calculating Entropy | Gini Impurity, we compute probabilities

$$H(Y) = -\sum_{i=1}^{n} P(Y)\cdot\log(P(Y)) \qquad G(Y) = 1 - \sum P(Y)^2$$

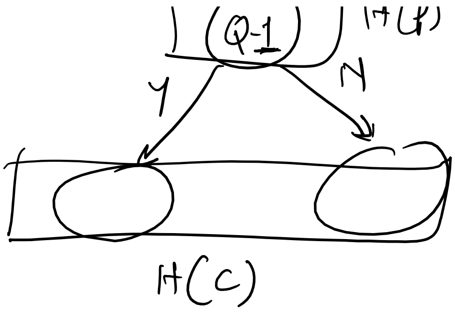
And probability considers 'frequency' of the point not the value of that point.

But, should we Normalise | Standardise ? → Yes!

☆ Should we use Decision Trees for high dimension data? eg, d = 1,00,000 ⇒ Ans: NO!



→ Because it will be very slow



IG = $\boxed{?}$

slow

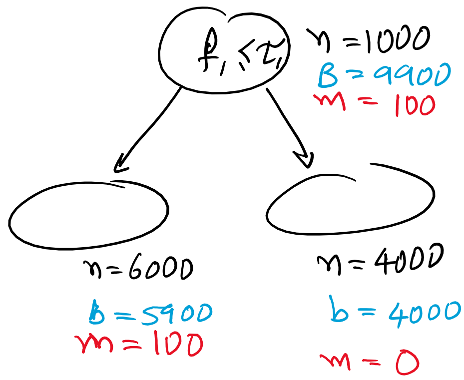
→ Lot's of computation are needed.

☆ Will imballanced data affect decision trees?

(Is it needed to do data-rebalancing while using decision trees?) ⇒ Yes!

$\boxed{\text{Tumor}}$ → Benign - Non-cancerous → 99% ⎫ 9,900
↑
10,000
↓
→ Malignant - Cancerous — 1% ⎭ 100

Question → Entropy | Gini Impurity
↳ p(y) ↵



$P(m) = \frac{100}{6000}$

$P(b) = 5900/6000$

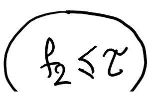
Data Rebalancing - Under sampling, Over Sampling, Class Weights,
SMOTE

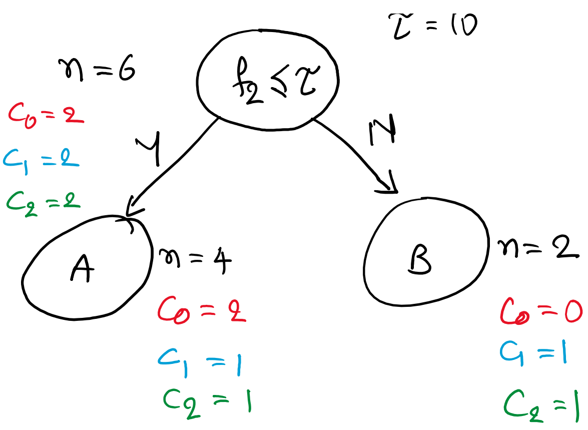
☆ Can we use Decision Trees in multiclass-classification?

f_1	f_2	f_3	f_4	Y
-	2	-	-	0
-	5	-	-	1

$\tau = 10$

n = 6





–	2	–	–	
–	5	–	–	1
–	11	–	–	2
–	15	–	–	1
–	10	–	–	2
–	8	–	–	0

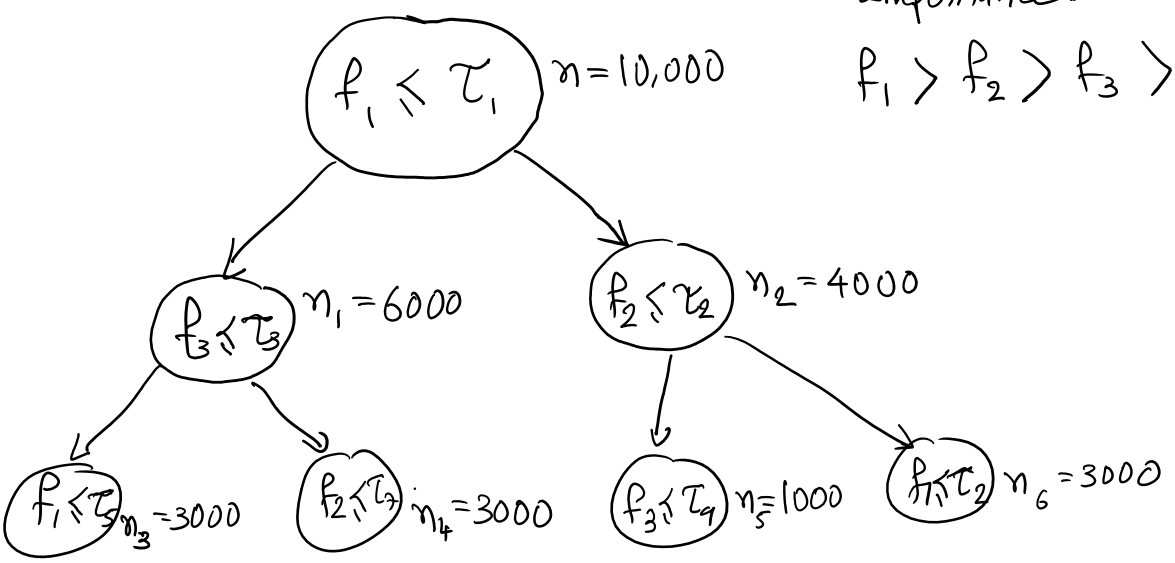
$$G(A) = 1 - [P(C_0)^2 + P(C_1)^2 + P(C_2)^2]$$

Yes! We can use Decision Trees for multiclass classification.

☆ **How will we calculate feature importance through Decision Trees?**

Importance:

$f_1 > f_2 > f_3 > f_4$



We compute **Normalised Information Gain** of each feature and then the feature with highest NIG is the most important feature.

$$\text{NIG of } f_2 = \frac{n_2}{n} \cdot \text{IG of } f_2 \text{ (at d = 2)} + \frac{n_4}{n} \cdot \text{IG of } f_2 \text{ (at d = 3)}$$

$$\text{NIG of } f_2 = \frac{n_2}{n} \cdot \underset{(\text{at } d=2)}{\text{IG of } f_2} + \frac{n_{14}}{n} \cdot \underset{(\text{at } d=3)}{\text{IG of } f_2}$$