Decision Trees 2

02 April 2025 18:15

Gini - Another measure to quantify impurity of a node

Formula: $\boxed{G(Y) = 1 - \Sigma (P(y_i))^2}$

Background:

Suppose our dataset has 'n' rows & 'd' features then the steps will be:

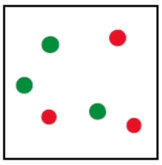
1. Calculating Entropy for each feature (e.g., gender = "female")
2. Compute Information Gain for each feature
3. Choose the feature with maximum IG & split the dataset on the basis of that feature

The formula for entropy is a little complicated and that's why the Gini Impurity concept was introduced.

How to compute Gini Impurity in our example?

$$G(Y) = 1 - \Sigma (P(y_i))^2 \Rightarrow G(Y) = 1 - \left[P(y \in G)^2 + P(y \in R)^2 \right]$$

A



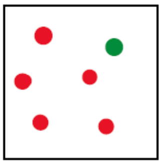
$P(-ve) = \frac{1}{2}$

$P(+ve) = \frac{1}{2}$

$$G(A) = 1 - \left[(\tfrac{1}{2})^2 + (\tfrac{1}{2})^2 \right] = 1 - (\tfrac{1}{4} + \tfrac{1}{4})$$

$$\boxed{G(A) = \tfrac{1}{2} = 0.5} \leftarrow$$ Max possible value of Gini

B



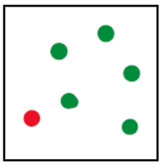
$P(-ve) = 5/6$

$P(+ve) = 1/6$

$$G(B) = 1 - \left[(\tfrac{5}{6})^2 + (\tfrac{1}{6})^2 \right] = 1 - \left[\frac{25-1}{36} \right] = 1 - \frac{\cancel{24}\ 2}{\cancel{36}\ 3}$$

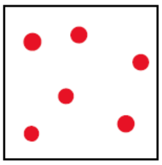
$$\frac{3-2}{3} = \frac{1}{3} \quad \boxed{G(B) = \tfrac{1}{3} = 0.33}$$

C



$$\boxed{G(C) = \tfrac{1}{3} = 0.33}$$

D



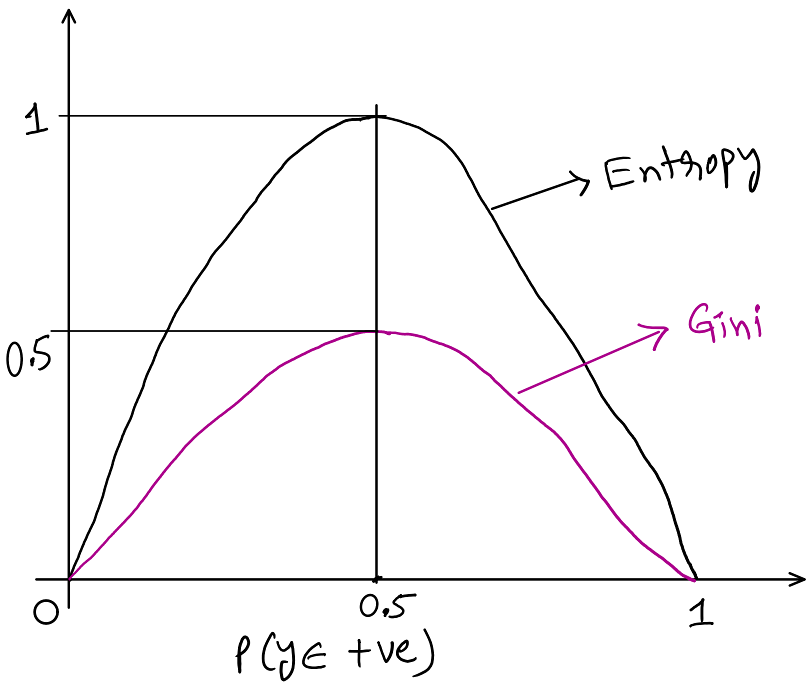
$P(-ve) = 1$

$P(+ve) = 0$

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

$$\boxed{G(D) = 0} \leftarrow$$ Minimum possible value of Gini

Plotting Entropy & Gini both together:

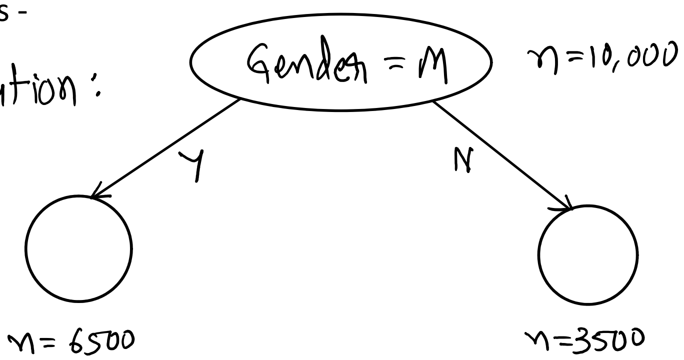


Can we use the same approach for a numerical column?
Ans: No

But why?
Because if we use this method to make question on a numerical column then we will have to ask too many questions.
Example:
Let's consider two categorical columns -

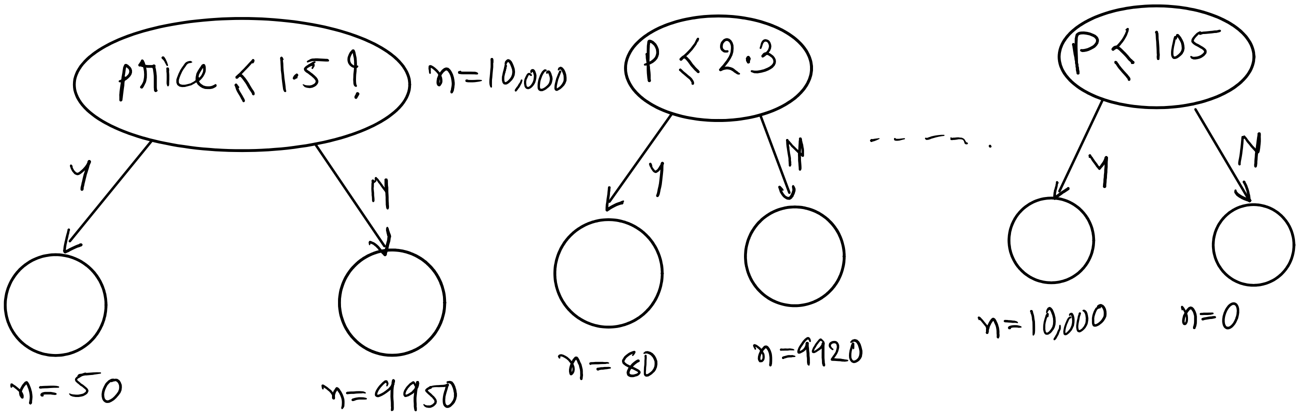
Gender	Education
M	Non-grad
F	Non-grad
F	Grad
F	Non-grad
M	Grad

Solution:



But if we consider a numerical column -

Price
1.5
2.3
2.7
105
75
1.5
2.7
75



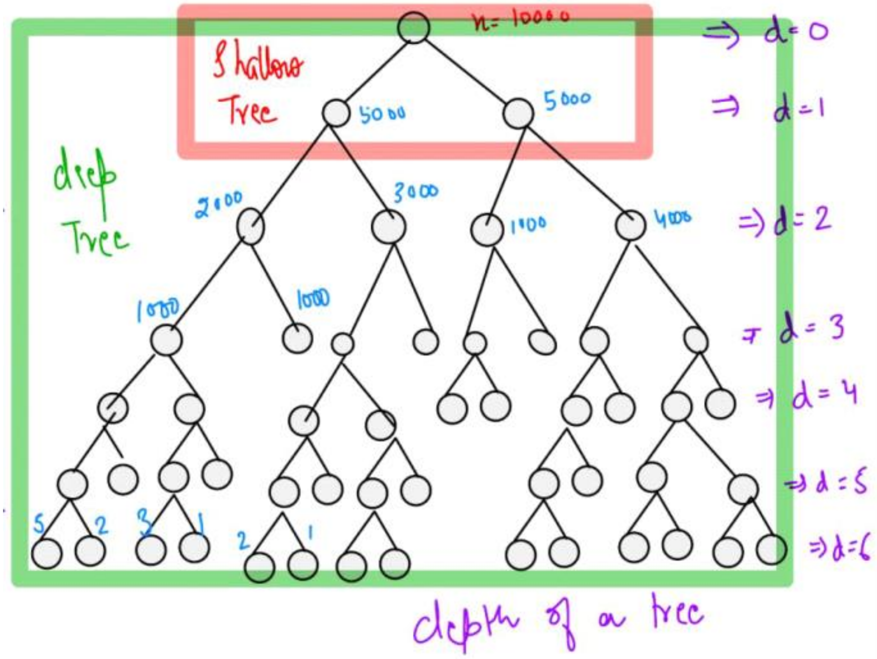
So, if we want to apply the same strategy for "Price" column then we need to perform the following steps:

1. Sort the entire data in the ascending order of that categorical column
2. For each unique value (threshold) in that column, calculate Entropy
3. Find the IG for each unique threshold
4. Identify the question with highest IG

Disadvantage of this approach: Computationally expensive

Then how can we find Entropy of numerical columns?
Sol: We may create bins on the numerical column & then calculate Entropy for each bin rather than on each unique value.

The Entire Picture of a Decision Tree:



What is underfitting?

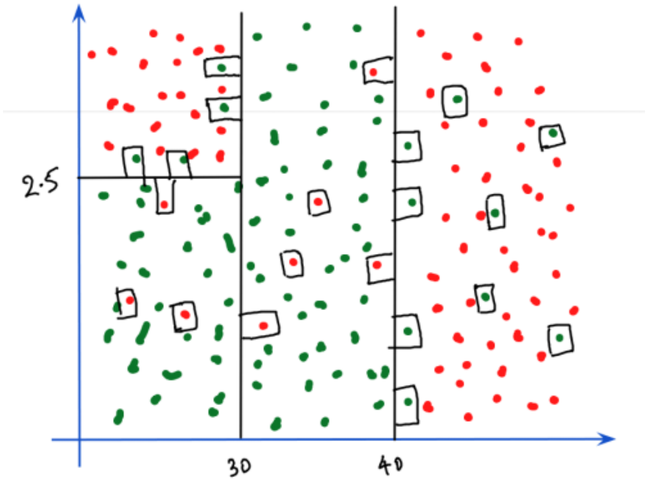
Ans: The model has not learned enough.

If we ask too less questions, our model will not learn enough and hence the trees which are "Shallow" usually underfits.

What is overfitting?

Ans: The model has learned each and every data point.

Each question that we ask to our data, creates an Axis-parallel decision boundary. Therefore, more questions we ask, more decision boundaries will be created and if we ask too many questions then there will so many axis-parallel boundaries that they will try to isolate each and every edge cases as shown beside. Hence trees those are very "deep" (asks too many questions) usually end up getting overfitted.



Therefore, it seems to be a very good thought to control & fine tune the depth "d" of a decision tree. Hence "d" is a very important hyper-parameter of a decision tree.

Depth d	Training Accuracy	Validation Accuracy
1	Very low	Very low
2	Very low	Very low
3	Very low	Very low
.	.	.
5	Decent	Decent
7	Decent	Decent
10	Decent	Decent
.	.	.
50	Very High	Low
100	Very High	Low
500	Very High	Low

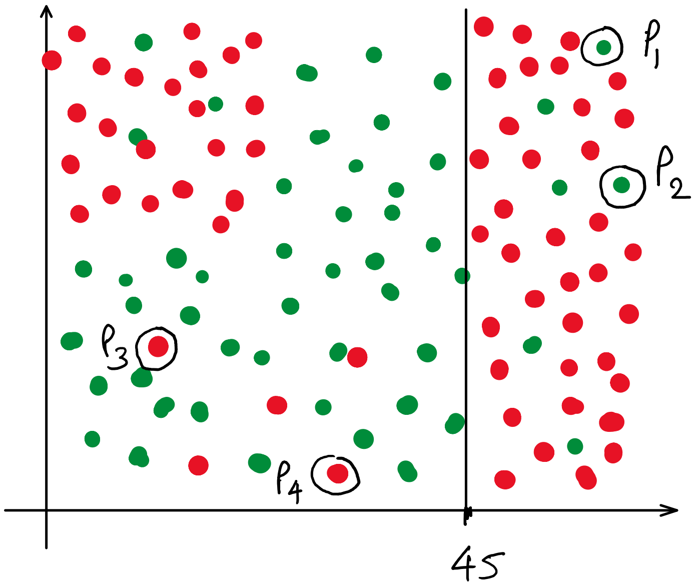
} Underfit

} Perfect fit (Where we get best Validation Accuracy)

} Overfit

Miscellaneous Questions about Decision Trees :

1. Do the outliers impact the output of a decision tree?



Case - 1: Shallow Tree (e.g., d = 1)
Question: age > 45?

```
If yes:
    Leave
Else:
    Stay
```

In this case, p1, p2, p3 & p4 are the outliers. If p1 & p2 were Red points (not outliers), our model would have classified them as Red only. And even if they are Green, our model is not changing its decision. Hence, our model is not impacted by these outliers.

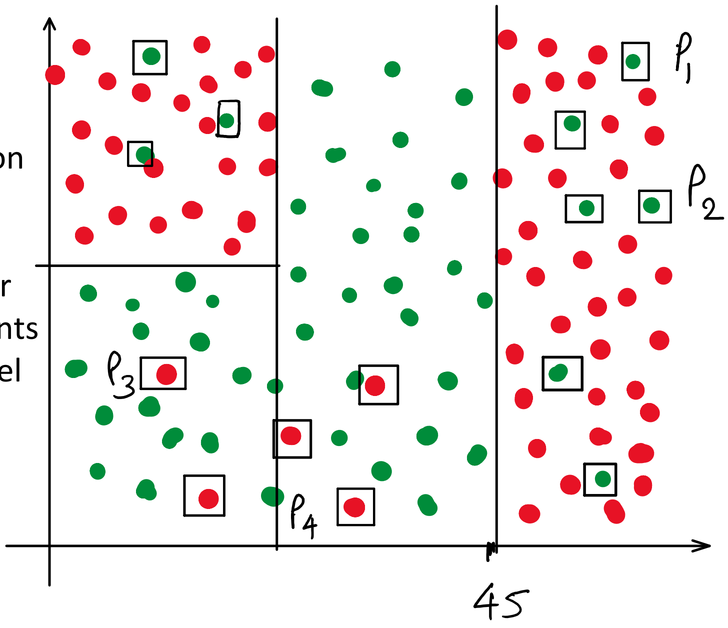
Similarly, if p3 & p4 were not outliers (if they were Green) then also our model would have classified them as Green just like they are classified right now. Hence, our model is not getting impacted by outliers for d=1

Case - 2: Deep Tree (e.g., d = 100)

Hence our model will create boundary as soon as a point of different class is encountered.

Therefore, if P_1 & P_2 were red points then our model would have classified them as red points but as in this case, if they are green our model will create boundaries around them and will classify them as green points.

So, our model is getting impacted by the outliers in this case.



Depth d	Training Accuracy	Validation Accuracy	Type of fit	Impact of outliers
1	Very low	Very low	Underfit	Extremely Low
2	Very low	Very low	Underfit	Extremely Low
3	Very low	Very low	Underfit	Extremely Low
.	.	.		
5	Decent	Decent	Good Fit	Somewhat
7	Decent	Decent	Good Fit	Somewhat
10	Decent	Decent	Good Fit	Somewhat
.	.	.		
50	Very High	Low	Overfit	Extremely High
100	Very High	Low	Overfit	Extremely High
500	Very High	Low	Overfit	Extremely High

2. Is feature scaling essential in decision trees? (normalization/standardization)

Ans: Let's ask ourself why do we do normalization/standardization? To get all the data on the same scale & to minimize the large numbers. There are two reasons why normalization/standardization are not essential for decision trees:

1. But here in decision trees, we don't compute the number with one another hence it is not necessary to get them on the same scale, nor they have to be minimized. Because instead of finding distances of each & every datapoint from a line, we are just asking some questions that split the data into two parts. So even if a number is very large (let's say 1,000,000) then also our comparison will not need additional computation power (e.g., n > 50) and hence, we do not need to shrink the numbers.

2. While computing Entropy/Gini impurity, we compute probabilities and probability does not take magnitude of a number in account but it considers the number of occurrences (frequency) of that number. (For example, if we want to find probability of getting "10" card while drawing a card from the standard deck, the magnitude of number 10 is not important rather, we are interested in finding 'how many "10" numbered cards are there in a standard deck?')

Hence we don't have to do standardization/normalization with decision trees.
But, do we do standardization/normalization?
Ans: Yes, we still do it just to ensure that we may also use other techniques along with decision trees.

3. Is using decision trees a good idea for a data with very high dimensionality?
(e.g., what will happen if our data is having 1000 dimensions?)

Ans: No! But why?
Because it will be very slow.
Let's assume all of 1000 features are categorical just with 3 distinct values in each one of them. Still, we have to compute IG for 3000 questions and find the best question at every node in our decision tree. Furthermore, if some of the columns are numerical then it will need even more computations.

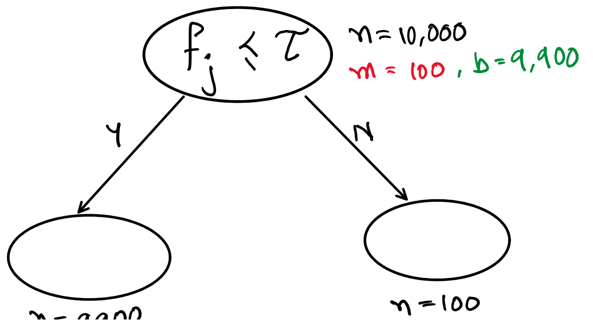
4. Will imbalanced data impact decision trees?
Let's take example of data for predicting a cancerous tumor.

Tumor → Benign = Non-cancerous - 99% - 9900
Tumor → Malignant = Cancerous - 1% - 100

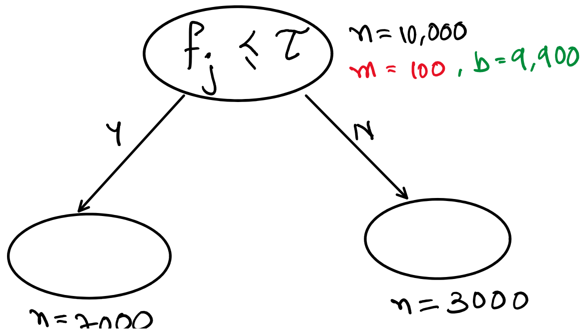
----

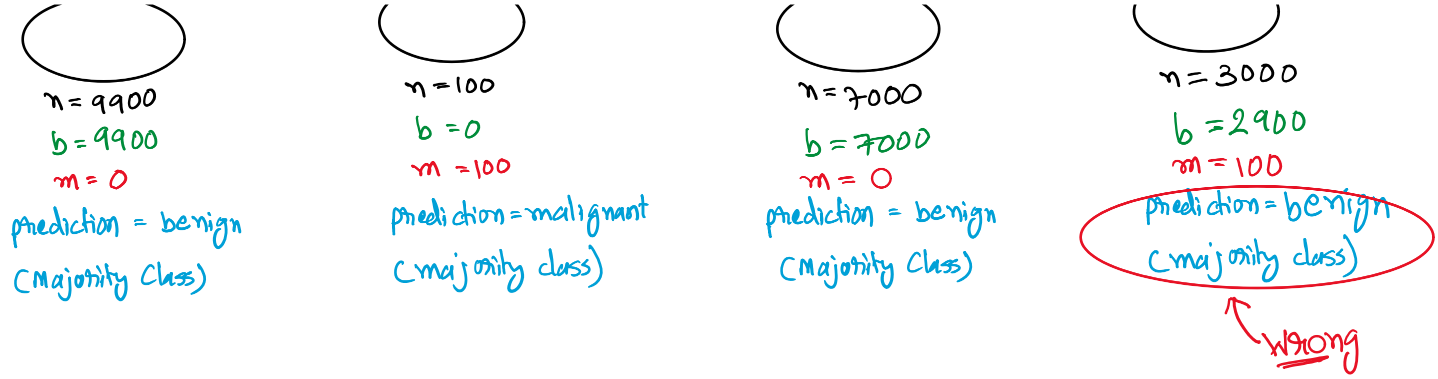
n = 10,000

A Hypothetical Scenario



Practical Scenario



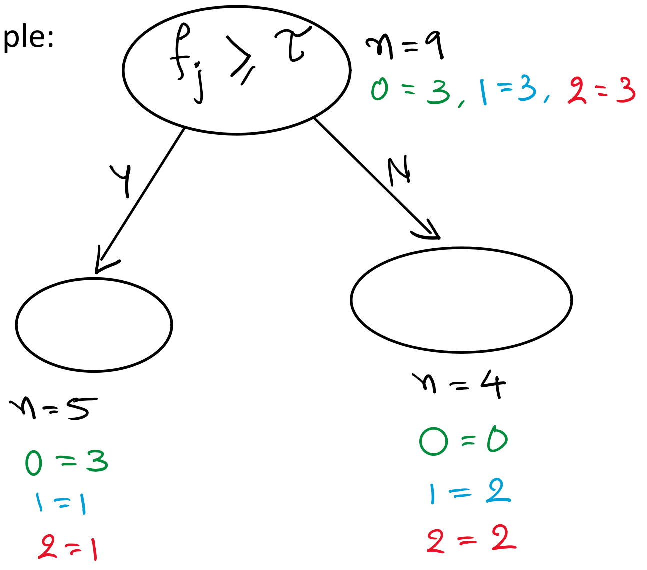


Hence we can clearly see that the decision trees are **impacted by imbalanced data**. So, we should consider rebalancing our data using techniques like SMOTE/Over sampling/Under sampling/Class weights etc.

5. Can we use decision trees in multiclass-classification scenarios? And how?

Again, let's take the following example:

f1	f2	f3	f4	…	Y
					0
					0
					2
					1
					0
					1
					2
					2
					1



$$G = 1 - \sum (P(y))^2$$

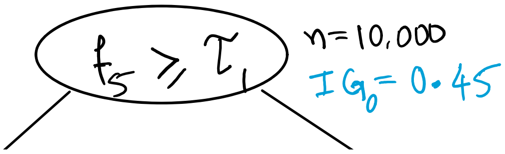
$$G = 1 - \left[0 + \left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$$

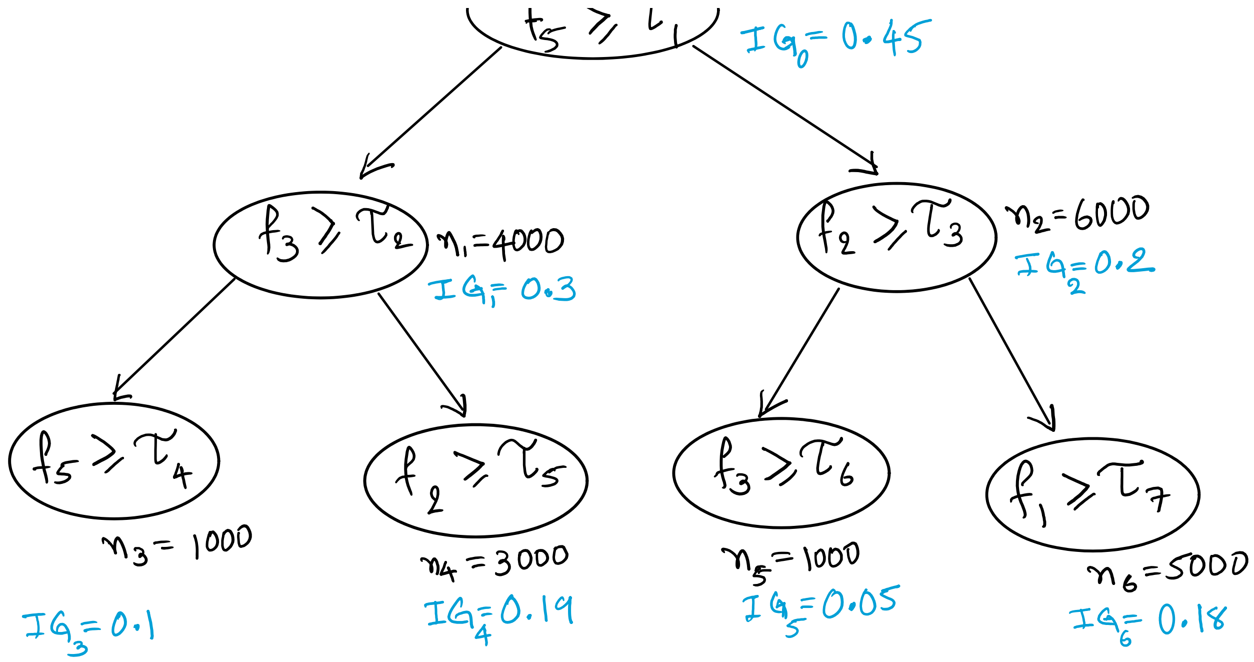
$$G = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right]$$

So, we did not face any problem in calculating Gini/Entropy/IG. Everything is just as it is as in case of binary classification. Therefore, **we can use decision trees for multiclass-classification problems** without any issues.

6. How to calculate the feature importance from the decision trees?

Again, let's try to get answer to this question by taking an example as below:





Now, how will we determine the feature importance? What will be the sequence of the features if they are sorted in descending order of their importance?

It is absolutely clear that f_5 is the most important feature. But how will we decide which is more important feature among f_2 & f_3? Because both of them are asked twice and IG of f_3 are 0.3 & 0.05 while IG of f_2 are 0.2 & 0.19. So, what should we do?

Ans: Taking weighted means of their IG that is known as **Normalized Information Gain**

Formula goes like this:

$$\text{NIG of } f_2 = \frac{n_2}{n} \cdot IG_2 + \frac{n_4}{n} \cdot IG_4 = \frac{6\cancel{000}}{10,\cancel{000}} \cdot 0.2 + \frac{3}{10} \cdot 0.19$$

$$= 0.6 * 0.2 + 0.3 * 0.19 = 0.175$$

$$\text{NIG of } f_3 = \frac{n_1}{n} \cdot IG_1 + \frac{n_5}{n} \cdot IG_5 = 0.4 * 0.3 + 0.1 * 0.05 = 0.125$$

Therefore, the importance of features are as under:

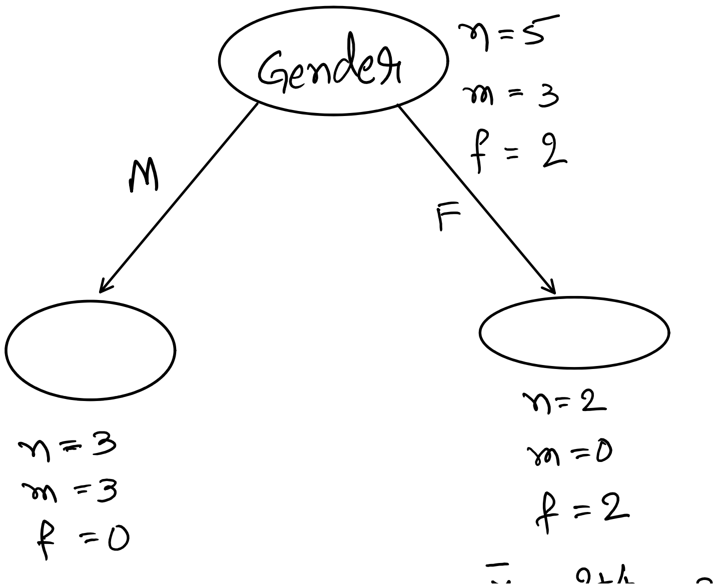
$f_5 > f_2 > f_3 > f_1 > f_4$.

Decision Tree Regressor:

$\rightarrow Y$

Gender	Education	Salary		
F	G	2	3	-1
M	NG	3	4.67	-1.67
F	NG	4	3	+1
M	NG	5	4.67	+0.33
M	G	6	4.67	+1.33

$$\boxed{\bar{Y} = \frac{20}{\ } = 4}$$



$m = 3$

$$\boxed{\bar{Y} = \frac{20}{5} = 4}$$

$f = 0$

$$\bar{Y} = \frac{3+5+6}{3} = 4.67$$

$f = 2$

$$\bar{Y} = \frac{2+4}{2} = 3$$

Instead of using raw errors, if use MSE (Mean Squared Error) as our metric, we will have the following table:

Gender	Education	Salary	Predicted	Error
F	G	2	3	-1
M	NG	3	4.67	-1.67
F	NG	4	3	1
M	NG	5	4.67	0.33
M	G	6	4.67	1.33

$$MSE_M = \frac{1}{3}\left((-1.67)^2 + (0.33)^2 + (1.33)^2\right) = 1.55$$

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

$$MSE_{child\text{-}level} = \frac{3}{5} MSE_M + \frac{2}{5} MSE_F$$

$$= \frac{3}{5} \times 1.55 + \frac{2}{5} \times 1$$

$$\boxed{MSE_c = 1.33}$$

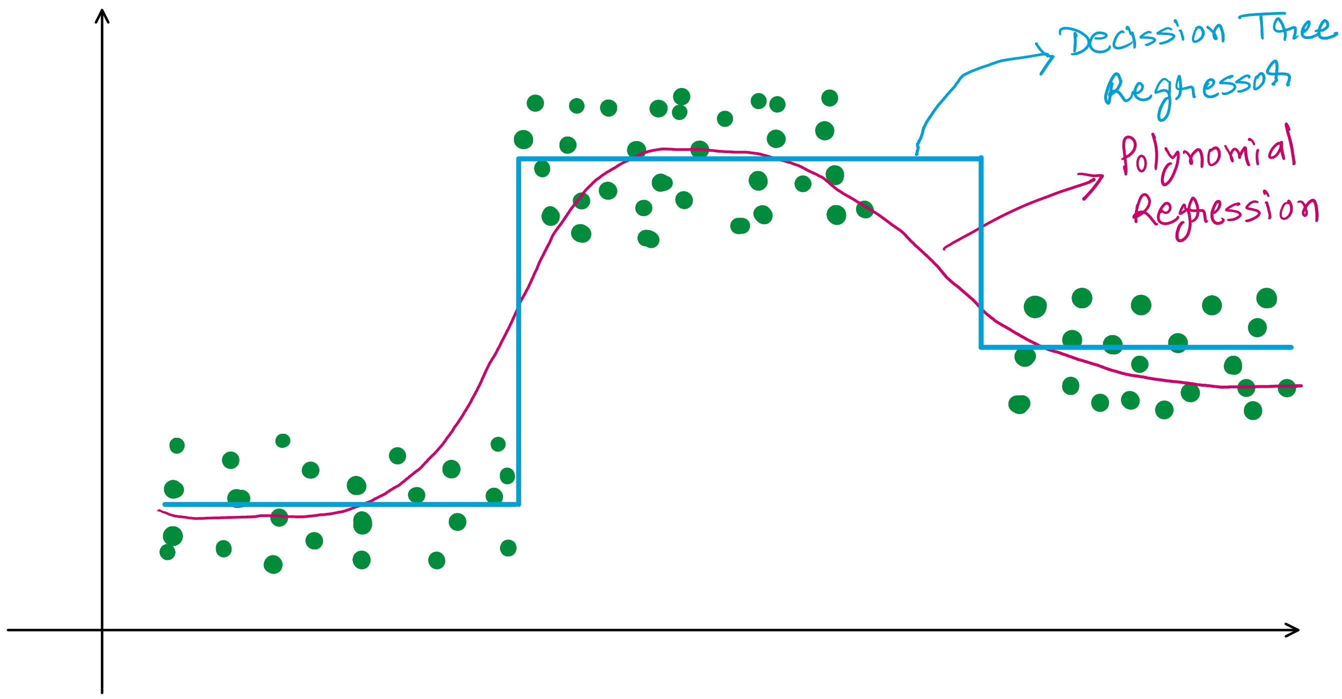
$$MSE_{parent\text{-}level} = \frac{1}{5} \sum (y - \bar{y})^2$$

$$= \frac{1}{5}\left[(2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2\right]$$

$$= \frac{1}{5}\left[4 + 1 + 0 + 1 + 4\right]$$

$$\boxed{MSE_p = 2}$$

Visualization:


Decission Tree Regressor
Polynomial Regression