MACHINE LEARNING DA7

Topics covered :-

=> Decision Tree

=> Gini Impurities

=> Entropy

=> Information Gain.

★ Decision Tree :-

Decision tree is used for both.

① Regression Task

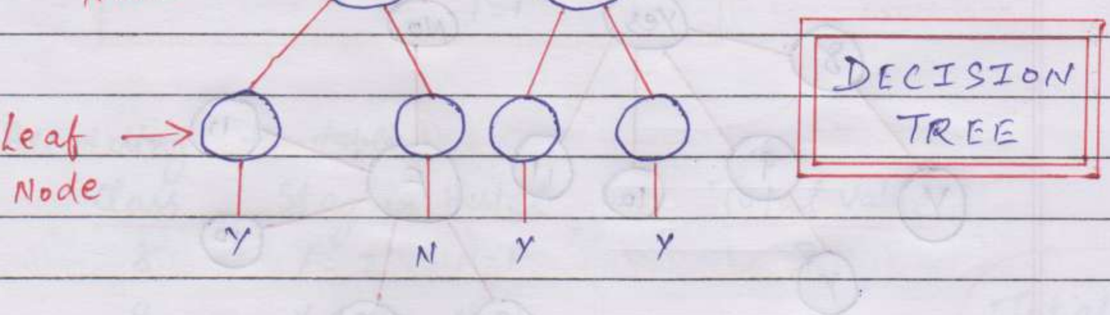
② Classification Task.

In Classification we have

-> Binary Data classification - 0 or 1

-> Multiclass classification - 1,2, 3, 4, 5, 6 or a, b, c, d

In regression, we get real values or continuous data.



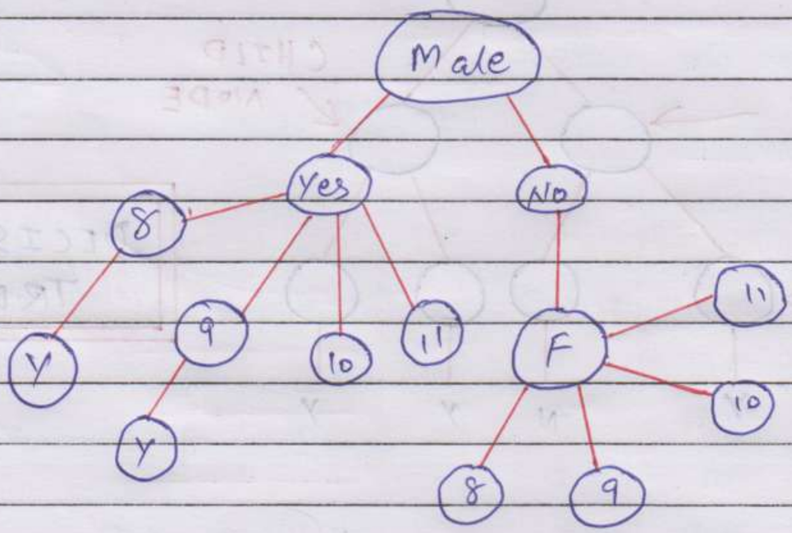
Example: DATASET

Class	Gender	Stay in Hostel
9	M	Y
10	F	N
8	F	Y
8	F	N
9	M	Y
10	M	N
11	F	Y
11	M	Y
8	F	Y
9	M	N
11	M	N
11	M	Y
10	F	N
10	M	Y
8	F	□

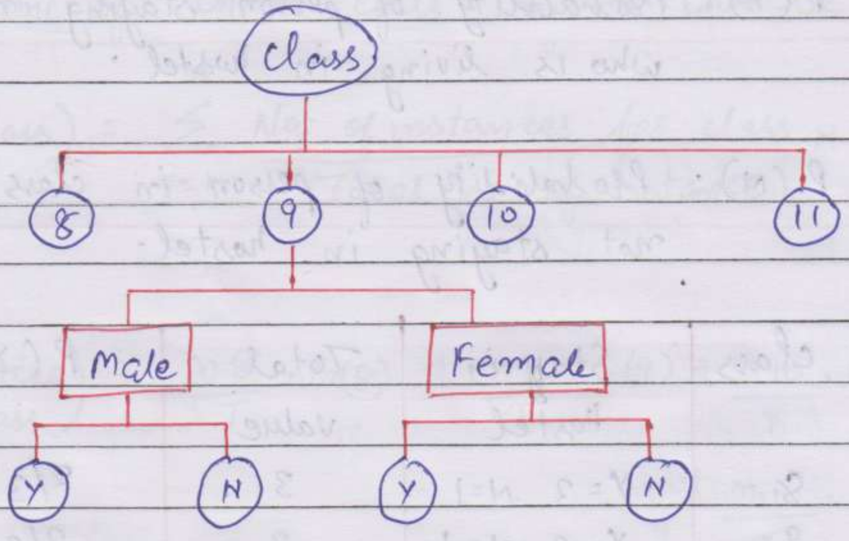
predict ↑

TABLE 1

If we make tree based on Male Column.



Tree Based on class column.



So we can take any attribute as the root node. But which attribute (column) we have to take as root node?

Here are only two attributes class and Gender. Suppose there will be 100 of attributes then how to Know which attributes to select as root node?

For that we have some formulas.

⇒ Gini Impurities:-

$$\boxed{Gini = 1 - \sum_{i=1}^{n} (p_i)^2}$$

According to table 1

Class	Stay in Hostel		Total value
8	Y = 2	N = 1	3
9	Y = 2	N = 1	3
10	Y = 1	N = 3	4
11	Y = 3	N = 1	4
			14

Total 14 dataset

- P(Y): Probability of person staying in class 8 and who is living in hostel.

- P(N): Probability of person in class 8 and who is not staying in hostel.

class	Stay in Hostel	Total value	P(Y)	P(N)
8	Y = 2 N = 1	3	2/3	1/3
9	Y = 2 N = 1	3	2/3	1/3
10	Y = 1 N = 3	4	1/4	3/4
11	Y = 3 N = 1	4	3/4	1/4
		14		

Now we will calculate Gini impurity for each and every individual classes.

$$Gini = 1 - \sum_{i=1}^{n} (P_i)^2$$

$$Gini(8) = 1 - P(Y)^2 - P(N)^2$$
$$= 1 - (2/3)^2 - (1/3)^2 = \boxed{4/9}$$

$$Gini(9) = 1 - (2/3)^2 - (1/3)^2 = 1 - 4/9 - 1/9 = \boxed{4/9}$$

$$Gini(10) = 1 - (1/4)^2 - (3/4)^2 = 1 - 1/16 - 9/16 = \boxed{3/8}$$

$$Gini(11) = 1 - (3/4)^2 - (1/4)^2 = 1 - 9/16 - 1/16 = \boxed{3/8}$$

Now Gini of entire class column will be

$$\boxed{\text{Gini}(\text{class}) = \sum_{i=1}^{n} \frac{\text{No. of instances for class}}{\text{Total no. of instance}} \times \text{Gini}(c')}$$

$$\text{Gini}\left(\begin{matrix}\text{Entire}\\ \text{Class}\end{matrix}\right) = \frac{n8}{T} \cdot G(8) + \frac{n9}{T} \cdot G(9) + \frac{n10}{T} \cdot G(10) + \frac{n_{11}}{T} \cdot G(11)$$

$$= \frac{3}{14} \cdot \frac{2}{3} + \frac{3}{14} \cdot \frac{4}{9} + \frac{4}{14} \cdot \frac{3}{8} + \frac{4}{14} \cdot \frac{3}{8}$$

$$= 0.66 + 0.44 + 0.375 + 0.375$$

Gini (Entire class) = 0.404

This whole calculation is for class column only. Now we have to calculate gini for Gender column.

Gender	Stay in Hostel	Total Value	P(Y)	P(N)
Male	Y = 5 N = 3	8	5/8	3/8
Female	Y = 3 N = 3	6	3/6	3/6
		14		

$$\text{Gini (male)} = 1 - \left(5/8\right)^2 - \left(3/8\right)^2 = 1 - \frac{25}{64} - \frac{9}{64}$$

$$= 0.468$$

$$\text{Gini (Female)} = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$$

$$\text{Gini}\begin{pmatrix}\text{Gender}\\\text{Column}\end{pmatrix} = \frac{8}{14} \cdot 0.468 + \frac{6}{14} \cdot 0.5 = \underline{\underline{0.4817}}$$

$$\text{Gini}\begin{pmatrix}\text{Class}\\\text{column}\end{pmatrix} = 0.404$$

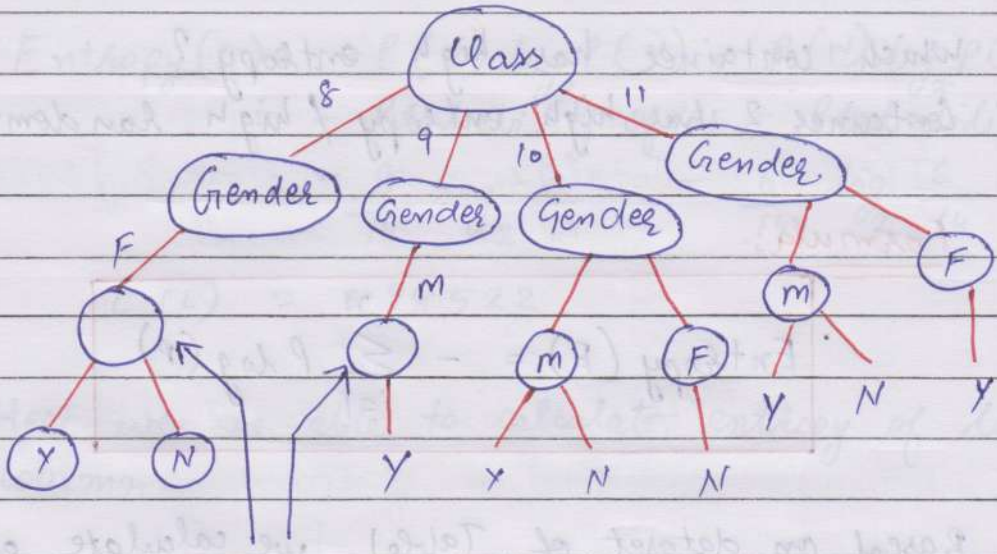
⇒ Out of Gender column and class column Gini of Gender column is more. Gini is actualy Gini impurity.

Note:- Here Gini Impurity (Gender column) is more compared to Gini Impurity (Class column). So we have to take class column as the root node or Parent node.

We have to select the attribute which is giving low gini impurities. So we can define probability much accurately in class column as compared to gender column.

Now in practical senario, suppose we have 100 columns, we will calculate gini of every column and use that column as our root node which has less Gini impurities.

Note:- Gini is an algorithm that works well with categorical data. Not continuous data.



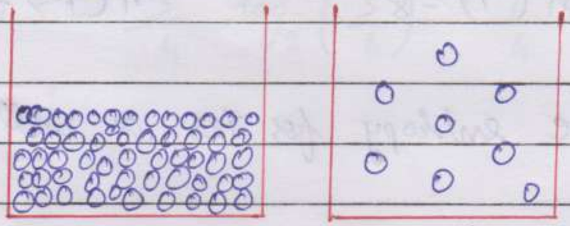
Leaf Nodes

- If we have to predict that in class 11 their is a female is she goin to stay in hostel or not. So by seeing the above tree we can predict that yes.

- If a new student in class 11 is male, then wheather he is going to stay in hostel or not? Answer is Yes because most of time male students are staying in hostel of class 11.

★ Entropy And Information Gain :-

Entropy means Randomness
Entropy → Degree of Freedom.



Container 1 Container 2

Que Which containee has high entropy?

Ans. Containee 2 has high entropy / high randomness.

Formula:-

$$\boxed{\text{Entropy}(E) = - \sum_{i=1}^{n} P \log(P)}$$

Based on dataset of Table1 we calculate entropy and based on entropy we try to calculate Information Gain.

Information Gain:- A measure of how much information a feature provides about a class.

$$\text{Information Gain}(IG) = E_{\text{Before}} - E_{\text{after}}$$

The difference between the entropy before creating a tree and after creating a tree is called as Information gain

Explanation:-

We have 14 records and 2 class in our dataset. out of these 14 records, how many instances gives Yes as the output.

14 Records $n(Y) = 8$ $n(N) = 6$

Try to calculate entropy for Yes as well as for No.

$$\text{Entropy}(L) = -P(Y)\log_2 P(Y) - P(N)\log_2 P(N)$$

$$= -\text{Probability of yes} - \text{Probability of No.}$$

$$= -\frac{8}{14}\log_2\frac{8}{14} - \frac{6}{14}\log_2\frac{6}{14}$$

$$E(L) = 0.98522$$

Here we are able to calculate entropy of label column.

Now we can try to calculate entropy for class column and Gender column.

Entropy for class column:

$$E(8) = -P(Y)\log_2 P(Y) - P(N)\log_2 P(N)$$

$$\boxed{= -2/3\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)}$$

$$E(8) = 0.918296$$

$$E(9) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.918296$$

$$E(10) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = 0.811$$

$$E(11) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$$

Information Gain from Class column.

$$I(\text{Class}) = \frac{\text{Total records of Class 8}}{\text{Total no. of records}} \cdot \left[\text{Entropy of class 8}\right]$$

$$+$$

$$\frac{\text{Total records of class 9}}{\text{Total no. of records}} \cdot \left[\text{Entropy of class 9}\right]$$

$$+$$

$$\vdots$$

$$I(\text{Class}) = \left(\frac{3}{14} \cdot 0.918\right) + \left(\frac{3}{14} \cdot 0.918\right) + \left(\frac{4}{14} \cdot 0.811\right) + \left(\frac{4}{14} \cdot 0.811\right)$$

$$I(\text{Class}) = 0.8574 \leftarrow \text{Total Gain}$$

Now

$$\text{Information Gain (IG)} = E_{\text{Before}} - E_{\text{after}}$$

Here $E_{\text{before}} \Rightarrow E(\text{Label Column})$

$E_{\text{after}} \Rightarrow E(\text{Class column})$

$$\therefore IG = 0.98522 - 0.8574 = \boxed{0.12782}$$

0.12782 will be the total Information gain, means this is the total difference between the entropy of label column and enthopy of class column.

Information Gain for Gender column

$$\text{Entropy }(m) = -P(Y)\log_2 P(Y) - P(N)\log_2 P(N)$$

$$= -\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right)$$

$$E(m) = 0.9544.$$

$$\text{Entropy }(F) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) = 1$$

$$\text{Entropy }(\text{Gender}) = \left(\frac{8}{14}\cdot 0.9544\right) + \left(\frac{6}{14}\cdot 1\right)$$

$$= 0.973943$$

← Total Gain

Information Gain IG = E before − E after

For Gender column

$$= E(\text{Label}) - E(\text{Gender})$$

$$= 0.98522 - 0.973943$$

Information Gain (Gender) = 0.011277.

Information Gain (Class) = 0.12782.