Date :- 5th November, 2022

MACHINE LEARNING (Day-7)

Agenda

1. Decision Trees.
2. Decision Tree Classifier
3. Decision Tree Regressor
4. Gini Impurities
5. Entropy
6. Information Gain

1) What is Decision Tree ?

⇒ A Decision Tree is a Type of Supervised Machine Learning used to categorize or make predictions based on How a previous set of questions were answered. The model is a form of Supervised Learning, meaning that the model is trained and tested on a set of data that contains the desired Categorizations.

*) How Does the Decision Tree Work ?

⇒ Before we divide into How a Decision Tree works, let's define some key terms of a decision tree.

(a) Root node :- The base of the Decision Tree.

(b) Splitting :- The process of dividing a note into multi sub-nodes.

(c) Decision node :- When a Sub-node is further split into additional sub-nodes.

P.T.O

(d) Leaf Node :- When a sub-node does not further split into additional sub-nodes; represents possible out comes.

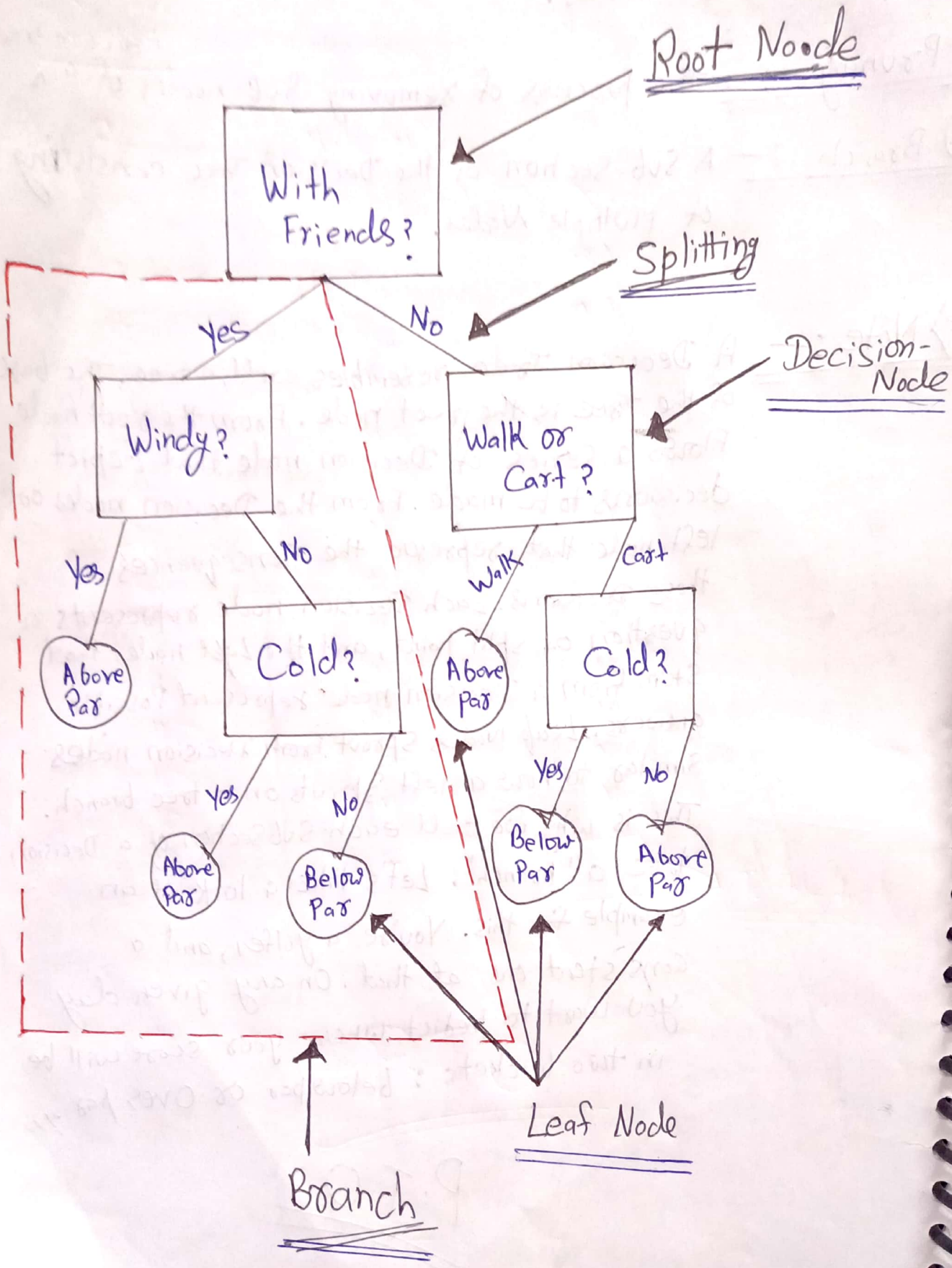
(e) Pruning :- The process of removing sub-nodes of a decision Tree.

(f) Branch :- A sub-section of the Decision Tree consisting of multiple Nodes.

*) Note :- A Decision Tree resembles, well, a tree. The base of the Tree is the root node. From the root node flows a series of Decision node that depict decisions to be made. From the Decision nodes are left node that represent the consequences of those Decisions. Each Decision node represents a question or split point, and the Left nodes that stem from a Decision node represent Possible answers. Leaf Nodes sprout from Decision nodes similar to how a left sprouts on a tree branch. This is why we call each subSection of a Decision tree a "branch". Let's take a look at an example for this. You're a golfer, and a consistent one at that. On any given day you want to predict where your score will be in two buckets : below par or over par ;)

P.T.O

Example of Decision Tree :-



✱) Advantages of a Decision Tree :—

- Works for Numerical or Categorical data and Variables.
- Models Problems with Multiple outputs.
- Tests the reliability of the tree.
- Requires less data cleaning than other data modelling Techniques.
- Easy to explain to those without an Analytical Background.

✱) Dis-Advantages of a Decision Tree :—

- Affected by Noise in the data.
- Not ideals for large datasets.
- Can Disproportionately Value, or Weigh, attributes.
- The Decisions at nodes are limited to binary outcomes, reducing the complexity that the tree can handle.
- Trees can become very Complex when dealing with Uncertainty and numerous Linked outcomes.

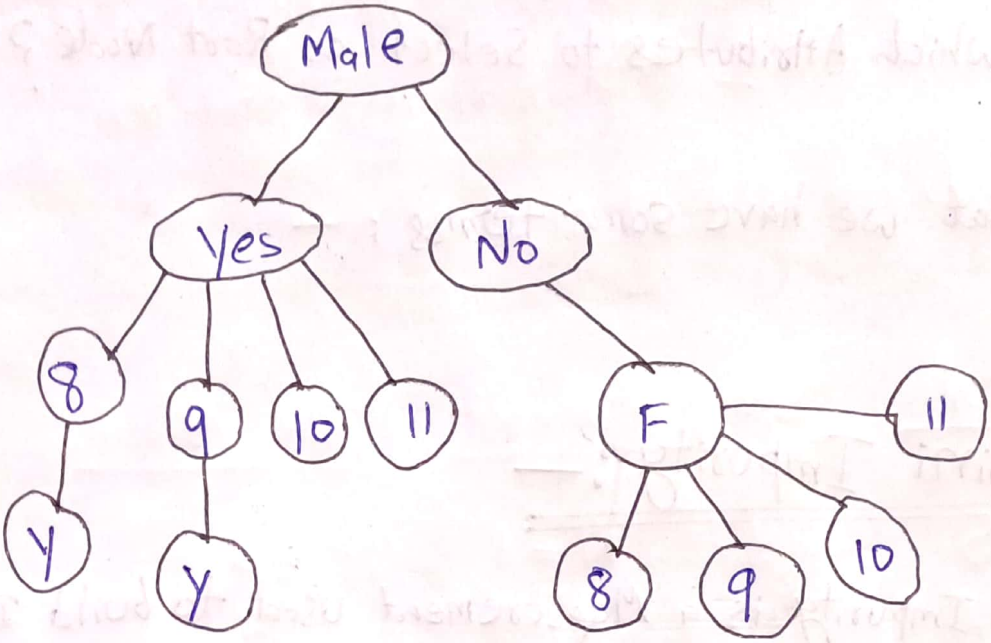
P.T.O

*) Example :- Dataset - 1

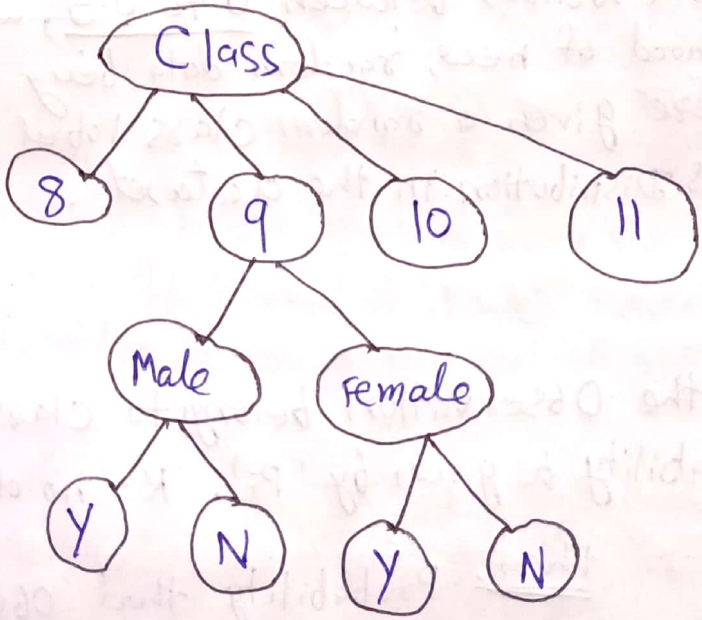
Class	Gender	Stay in Hostel
9	Male (M)	Y
10	Female (F)	N
8	F	Y
8	F	N
9	M	Y
10	M	N
11	F	Y
11	M	Y
8	F	Y
9	M	N
11	M	N
11	M	Y
10	F	N
10	M	Y
8	F	

P.T.O

*) 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. Support there will be 100 of Attributes then How to know which Attributes to Select as Root Node?

*) For that we have some terms :—

*) Gini Impurity :—

⇒ Gini Impurity is a Measurement used to build Decision Trees to determine how the features of a Dataset Should Split nodes to form the Tree.

Most Precisely, The Gini Impurity of a Dataset is a Number between 0 to 0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label According to the Class Distribution in the dataset.

Now,

let the Observation belongs to class 'i' and its Probability be given by 'P_i'. K = no. of Classes.

then Probability that Observation belongs to any other class,

other than 'i',

$\therefore \sum_{K \neq i} P_K = 1 - P_i$

P.T.O//

then,

$$\text{Gini Impurity} = \sum_{i=1}^{J} p_i \times \sum_{K \neq i} p_K$$

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

$$= \sum_{i=1}^{J} (p_i - p_i^2)$$

$$= \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2$$

$$= 1 - \sum_{i=1}^{J} p_i^2$$

$$\boxed{\text{Gini Impurity} = 1 - \sum_{i=1}^{J} p_i^2}$$

Gini Impurity :— Gini Impurity is a measure of how often a randomly choosen element from the set would be incorrectly labelled if it was randomly labelled according to the Distribution of labels in the Subset.

It is calculated by Multiplying the probability that a given observation is classified into the correct class and sum of all the probabilities when that particular Observation is classified into the wrong class.

P.T.O

Gini Impurity value lies between 0 and 1

0 → No Impurity
1 → Random Distribution.

According to Dataset-01,

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 Staying in Class 8 and Who is not Staying in Hostel.

P.T.O

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

Now,

We will calculate Gini Impurity for each and individual Classes.

$$\text{Gini } (D) = 1 - \sum_{i=1}^{I} (P_i)^2$$

$$\text{Gini}(8) = 1 - P(y)^2 - P(N)^2 \Rightarrow 1 - (2/3)^2 - (1/3)^2$$
$$= 4/9 //$$

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

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

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

P.T.O

Now

Gini of Entire Class Column will be :-

$$\text{Gini (class)} = \sum_{i=1}^{n} \frac{\text{No. of instances for Class}}{\text{Total No. of Instances}} \times \text{Gini}_i(c)$$

Now,

$$\text{Gini (Entire Class)} :- \frac{n8}{T} \cdot G(8) + \frac{n9}{T} \cdot G(9) + \frac{n10}{T} G(10) + \frac{n11}{T} G(11).$$

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

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

$$= 0.404$$

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		

Gini

P.T.O

Gini (Male) = $1 - P(y)^2 - P(N)^2$

$= 1 - (5/8)^2 - (3/8)^2$

$= 1 - \frac{25}{64} - \frac{9}{64}$

$= 0.468$

Gini (Female) = $1 - (3/6)^2 - (3/6)^2$

$= 0.5$

Now :-

Gini (Gender Column) = $\frac{8}{14} \times 0.468 + \frac{6}{14} \times 0.5$

$= 0.4817$

And

Gini (Class Column) = 0.404

Out of Gender Column and Class Column Gini of Gender column is More.

(Gini → Gini Impurity)

Here,

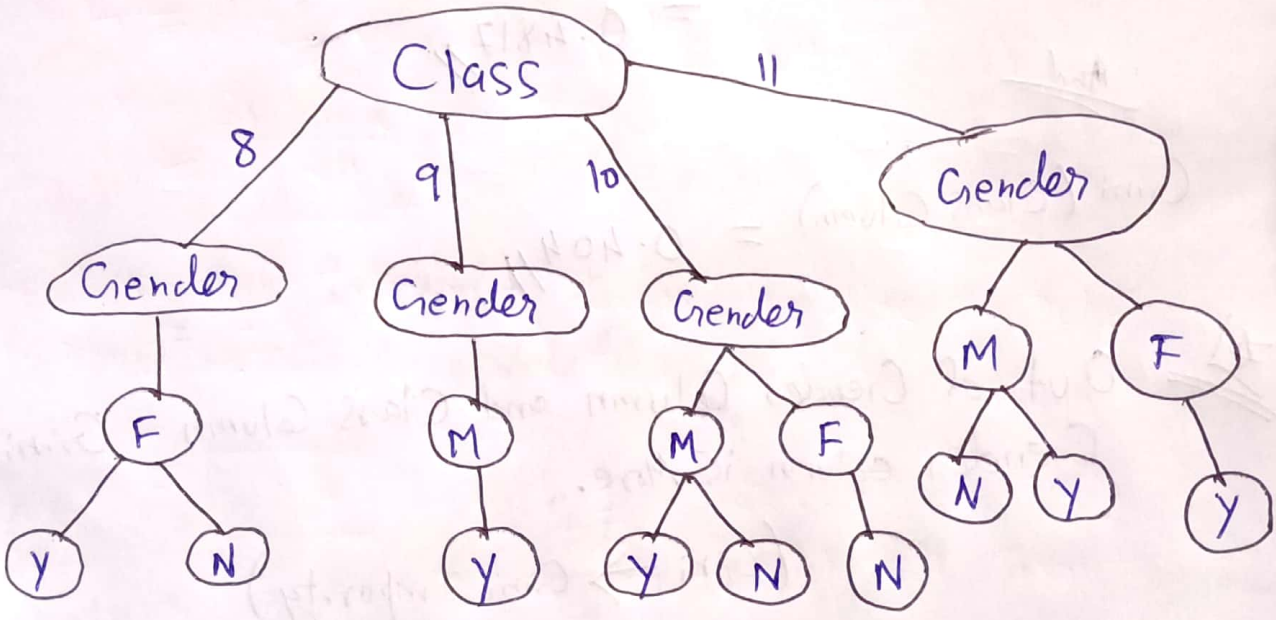
Gini Impurity (Gender Column) is more compared to Gini Impurity (Class Column). So, we have to take Class Column as the Root Node or Parent Node.

Note :-

The Node for which the Gini Impurity is least is Selected as the Root Node to Split. //

Support, 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 Impurity is an Approach which Works with Categorical data Not Continuous Data.



*) If we have to predict that in Class 11 there is a Female, is she going to stay in Hostel or Not.

So, by above Tree Diagram we can Predict that yes

Note

If a New Student in Class 11 is male, then Whether he is going to Stay in Hostel or Not ?

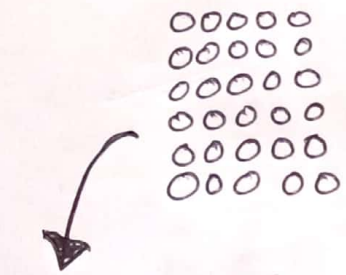
⇒ Yes, because Most of Time Male Students are Staying in Hostel of Class 11.

✱) Entropy :-

⇒ Entropy is the measure of Randomness in data.

In Other Words,

It gives the Impurity present in the dataset.



Low Entropy

High Entropy

$$\text{Entropy}(E) = \sum_{i=1}^{n} P \cdot \log_2(P)$$

P.T.O

Information Gain :-

⇒ Information Gain Calculates the decrease in Entropy after Splitting a node.

It is Different Between Entropies before and After the Split.

$$Gain\ (T, X) = E(T) - E(T, X)$$

Here,

E = Entropy

T = Parent Node before Split

X = Split Node From T

The More the Information Gain, the More Entropy is removed.

Based on Dataset of Table-1 we Calculate Entropy and based on Entropy we try to Calculate Information Gain.

P.T.O

Explanation :—

⇒ We have 14 Records and 2 Class in our dataset. Out of 14 Records, How many instances gives Yes as the output.

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

*) Calculating Entropy for YES and NO.

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

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

$$= 0.9852 //$$

$$\therefore E(L) = 0.9852 //$$

Here, we are able to Calculate Entropy of Label Column.

P.T.O

Calculating Entropy for Class Column and Gender Column

$$E(8) = -P(y)\cdot\log_2 P(y) - P(N)\cdot\log_2 P(N)$$

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

$$= \boxed{0.9182}$$

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

$$= \boxed{0.9182}$$

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

$$= \boxed{0.811}$$

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

$$= \boxed{0.811}$$

P.T.O

*) Information Gain from Class Column :—

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

$$+$$

$$\frac{\text{Total Record of Class 9}}{\text{Total No. of Records}} \left[\text{Entropy of Class 9}\right]$$

$$+$$

Now,

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

$$I(\text{class}) = 0.8574$$

(Total Information Gain of Class)

Now,

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

Where, $E_{\text{Before}} = E(\text{label Column})$

$E_{\text{After}} = E(\text{label Column})$

$\therefore IG = 0.9852 - 0.8572$

$= 0.1278$ //

P.T.O

Note :- 0.1278 will be the Total Information Gain means this is the Total Different Between the Entropy of Label Column and Entropy of Class Columns.

✱) Information Gain for Gender Column :—

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

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

$$= 0.9544.$$

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

$$= 1 \text{ //}$$

$$\therefore \text{Entropy (Gender)} = \frac{8}{14}\times 0.9544 + \frac{6}{14}\times 1$$

$$= 0.9739 \text{ //}$$

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

$$\therefore \text{For Gender Column} = E(L) - E(G)$$

$$= 0.9852 - 0.9739$$

$$= 0.01127$$

V.V.I

$\therefore$ Information Gain (Gender) = 0.01127
Information Gain (Class) = 0.1278 //

Case 01 :-

Feature can be Categorial
Outcome can be Categorial
} Classification Problem.

Case 02 :-

Feature can be Continuous
Outcome can be Categorial
} Classification Problem.

Case 03 :-

Feature can be Continuous
Outcome can be Continuous
} Regression Problem

✱) Example :-

X_1	X_2	X_3	X_4
1.1	7.5	2.5	A
2.2	8.8	5.5	B
3	9.2	6	A
3.6	5.1	6.7	A
5	5.4	7	B
5.8	2	8.9	B
8	1	9.1	A

Here :-

Feature ⇒ Continuous

Outcome ⇒ Categorial

P.T.O

*) How to Select the node in this Case?

⇒ The Concept here is we have to create a threshold,

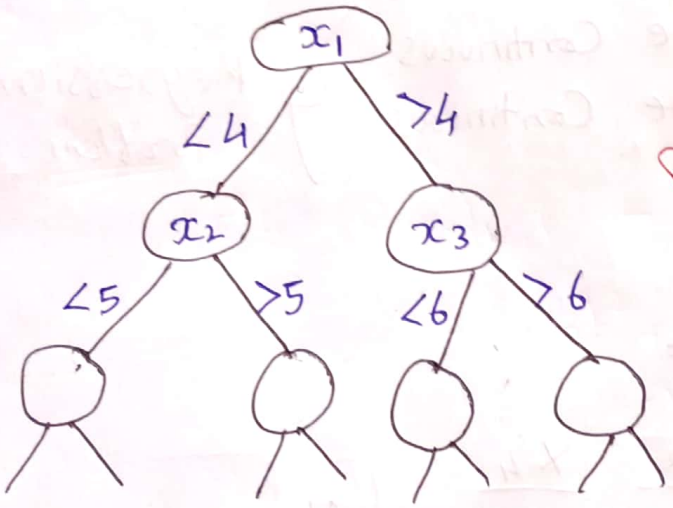
Support:-

We take the threshold for $x_1 = 4$

(≈ average)

Similarly,

For $x_2 = 5$, $x_3 = 6$



Here,

We will Divide until we get the Leaf Node.

P.T.O

✱) Another Dataset :-

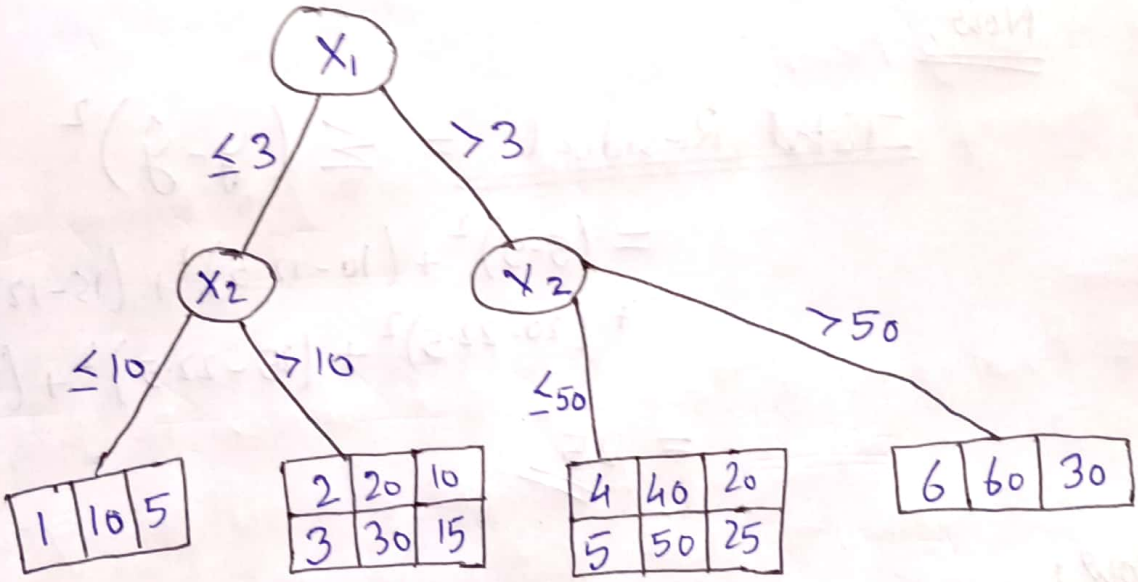
X_1	X_2	Y
1	10	5
2	20	10
3	30	15
4	40	20
5	50	25
6	60	30

✱) This is Regression Problem:-

⇒ Support the Information Gain X_1 is high.

So, X_1 will become root node.

X_1 threshold = 3
X_2 threshold = 30 ,,



Now, Calculate $y - \hat{y}$

Here :-
y : Actual Value
$\hat{y}$: Predicted Value.

P.T.O

For | 1 | 10 | 5 | :- $y = 5$, $\hat{y} = 5$

For | 2 | 20 | 10 |
| 3 | 30 | 15 | :- $y = 10$, $\hat{y} = \left(\frac{10+15}{2}\right) = 12.5$

$y = 15$, $\hat{y} = 12.5$

For | 4 | 40 | 20 |
| 5 | 50 | 25 | :- $y = 20$, $\hat{y} = \left(\frac{20+25}{2}\right) = 22.5$

$y = 25$, $\hat{y} = 22.5$

For | 6 | 60 | 30 | :- $y = 30$, $\hat{y} = 30$

Now,

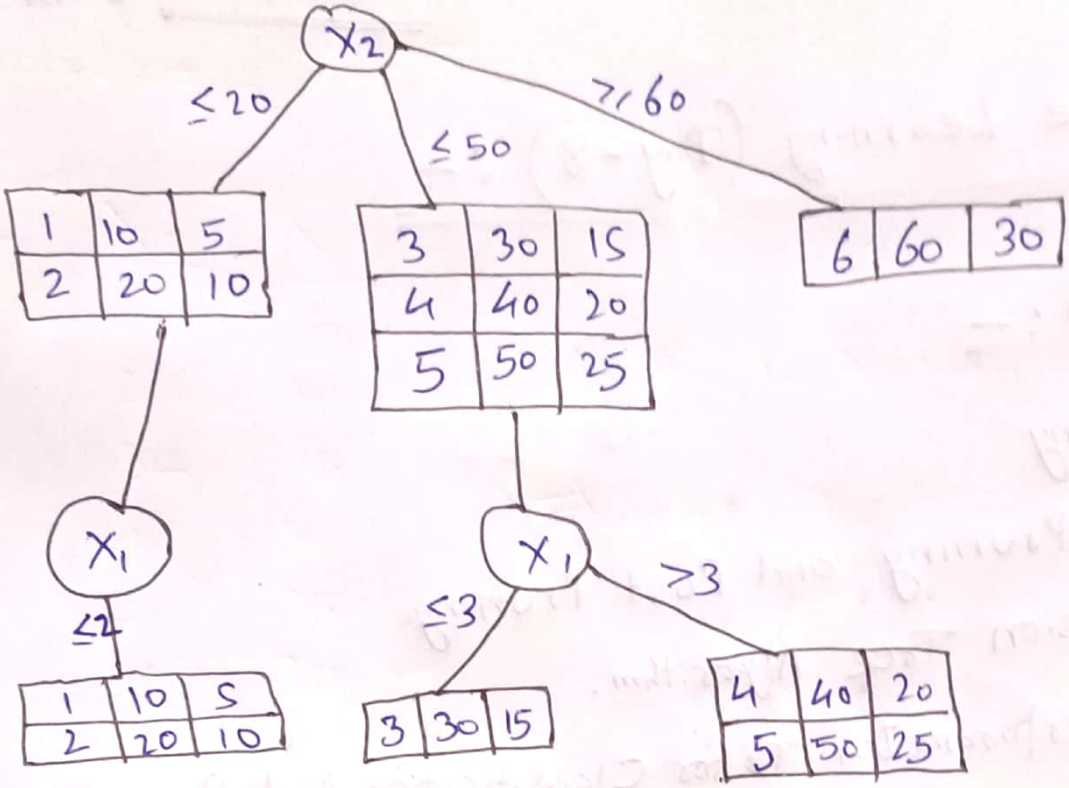
$$\text{Total Residual} = \sum (y - \hat{y})^2$$

$$= (5-5)^2 + (10-12.5)^2 + (15-12.5)^2 + (20-22.5)^2 + (25-22.5)^2 + (30-30)^2$$

$$= 25$$

Now,

Support the Information Gain of x_2 is high. and thus X_2 will become the Root node.



Now,

Test the Tree $(y-\hat{y})^2$

For

1	10	5
2	20	10

$y=5,\ \hat{y} = \left(\frac{5+10}{2}\right) = 7.5$ //

$y=10,\ \hat{y} = 7.5$ //

For

3	30	15

$y=15,\ \hat{y} = 15$

For

4	40	20
5	50	25

$y=20,\ \hat{y} = 22.5$

$y=25,\ \hat{y} = 22.5$

★) Total Residuals $\Rightarrow (5-7.5)^2 + (10-7.5)^2 + (15-15)^2 + (20-22.5)^2 + (25-22.5)^2$

$\Rightarrow 6.25 + 6.25 + 6.25 + 6.25$

$\Rightarrow 25$ //

Note:-

Take the Situation which has maximum Residual.