"Decision Tree using C4.5 Algorithm"

Note :⇒ ID3 favour attributes with large no. of values

C4.5 Principle :⇒

It is a Decision Tree Classifier which can be employed to generate a decision, based on a certain sample of data.

Iterative Dichotomiser 3(decision tree using Information Gain

↳ ID3 algorithm that we have seen earlier is biased towards multivalued attributes. C4.5 use Gain Ratio.

$$\text{Gain Ratio}: \quad \text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}$$

$$\text{SplitInfo}: \quad \text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

Attribute with maximum Gain Ratio is selected as "Splitting Attribute"

C4.5 Algorithm :⇒

1. For each attribute 'A', find the normalized information gain ratio from splitting on 'A'.
2. Let A_Best be the attribute with the highest normalized information Gain.
3. Create a decision node that splits on 'A_Best'

   Repeat these procedures on the sublists obtained by splitting on A_Best.
4. Add these nodes as children of node.

P.T.O

"Dataset"

Work Period	Income	Married	Bank	Buy House
Short	High	No	High	No
Short	High	No	High	No
Average	Medium	Yes	High	Yes
Long	Low	Yes	High	Yes
Long	Low	Yes	Low	Yes
Long	Low	No	High	No
Average	Low	No	Low	Yes
Short	Medium	Yes	Low	No
Short	Low	No	Low	Yes
Long	Medium	Yes	High	Yes
Short	Low	No	Low	Yes
Average	Medium	Yes	Low	No
Average	Medium	Yes	High	Yes
Long	High	Yes	Low	No
Average	High	No	Low	No

We have to first compute the information Gain.

$$\text{Information Gain} = -\sum_{j=1}^{v} P_j \cdot \log_2(P_j)$$

$$\text{Information Gain}_{(\text{Bank})}(D) = -\frac{7}{15}\log_2\left(\frac{7}{15}\right) - \frac{8}{15}\log_2\left(\frac{8}{15}\right)$$

$$= 0.9967$$

Now we need to compute information gain for each attribute :>

"Work-Period Attribute"

	Yes	No	Total
Short	2	3	5
Average	3	2	5
Long	3	2	5
			"15"

$$\text{Info}_{\text{WorkedPeriod}}(D) = \frac{5}{15} I(3,2) + \frac{5}{15} I(2,2) + \frac{5}{15} I(2,3)$$

$$= \frac{15}{15} I(3,2)$$

$$= -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.4421 + 0.5287$$

$$= 0.9708$$

Similarly let's try to compute Info for other attributes :>

$$\text{Info}_{\text{Income}}(D) = 0.5838$$

$$\text{Info}_{\text{married}}(D) = 0.2923$$

$$\text{Info}_{\text{Rank}}(D) = 0.9910$$

$$\text{Gain}_{\text{Income}}(D) = \text{Info}(D) - \text{Info}_{\text{Income}}(D)$$

$$= 0.9967 - 0.5838$$

$$= 0.4129$$

$$\text{Gain}_{\text{Work Period}}(D) = 0.0259$$

$$\text{Gain}_{\text{married}}(D) = 0.1044$$

$$\text{Gain}_{\text{Rank}}(D) = 0.0037$$

Now, we need to compute its SplitInfo for each attribute :>

$$\text{SplitInfo}_{\text{Income}}(D) = -\frac{4}{15}\log_2\left(\frac{4}{15}\right) - \frac{5}{15}\log_2\left(\frac{5}{15}\right) - \frac{6}{15}\log_2\left(\frac{6}{15}\right)$$

$$\text{SplitInfo}_{\text{Income}}(D) = 0.5085 + 0.5283 + 0.5287$$

$$= 1.5655$$

Similarly,

$$\left.\begin{aligned} \text{SplitInfo}_{\text{Work Period}}(D) &= 1.5849 \\ \text{SplitInfo}_{\text{Married}}(D) &= 0.9967 \\ \text{SplitInfo}_{\text{Rank}}(D) &= 0.9967 \end{aligned}\right\}$$

Now Compute Gain ratio for each attribute :→

$$\text{Gain Ratio} = \frac{\text{Info. Gain}}{\text{Split Ratio.}}$$

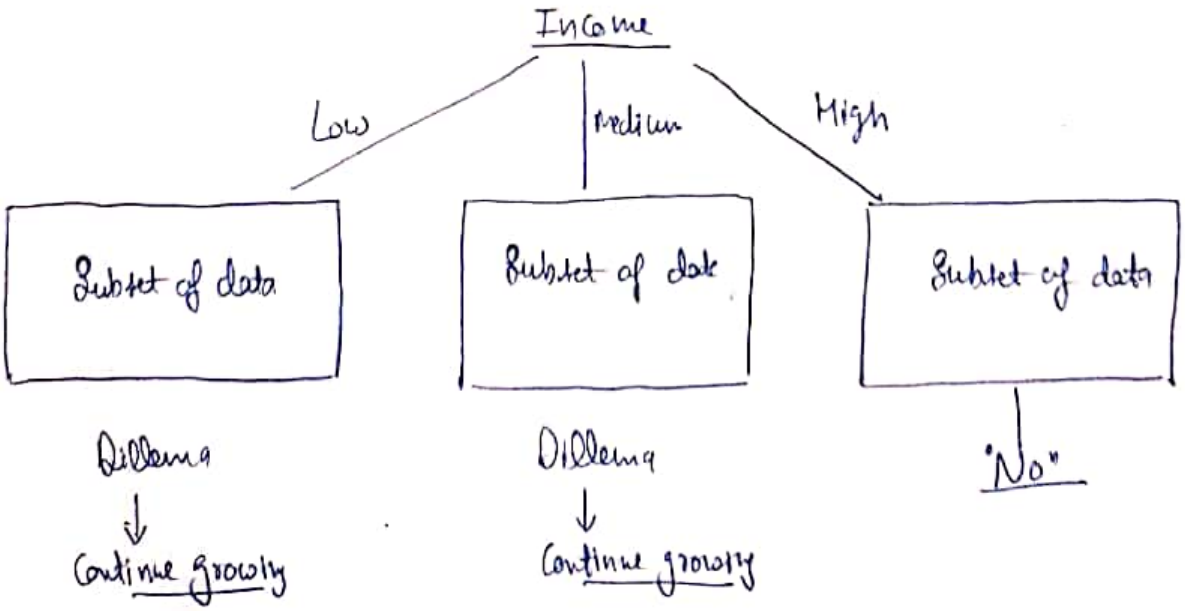
$$\text{Gain Ratio (Work Period)} = \frac{0.0259}{1.5849} \Big/ \frac{0.0259}{1.5849} = 0.0163$$

$$\text{Gain Ratio (Income)} = \frac{0.4129}{1.5655} = 0.9637$$

Gain Ratio (Married) = 0.1047 & Gain Ratio (Rank) = 0.0037

Since Gain Ratio of income is highest

∴ Income will become the 'Root'



CART (Classification & Regression Trees)

↳ "Another method to create Decision Trees"

Here we use Gini Index instead of "Entropy" & Binary split

Gini-Index, it also known as Gini Impurity

↓

It calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.

If all the elements are linked with a single class then it can be called pure.

Gini index varies b/w values 0 & 1 :→

↳ Where "0" expresses the purity of classification.

i.e. All the elements belongs to a specified class or only one class exists there.

↳ "1" indicates the random distribution of elements across various classes.

↳ The value of "0.5" of the Gini index shows an equal distribution of elements over some classes.

Let's now Construct a Decision Tree using "Gini Index"

P.T.O

'Dataset'

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

O/p variable

While constructing a Decision tree, we will be selecting an attribute which has highest information & we will make it as a root node.

Here the attribute which is having the minimum 'Gini index' will be the attribute that is having the maximum information. So we will select it as a root node.

Let's first compute the 'Gini index' of entire dataset.

In the dataset, we have four possible output variables :-

"Cinema, Tennis, Stay in & Shopping"

In total we have 6 instances where Cinema is present, 2 instances where Tennis is present, 1 instance of Stay in and 1 instance, where Shopping is present.

Now let's compute the Gini of entire dataset :-

$$\text{Gini}(S) = 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)^2\right]$$

$$= 0.58.$$

Now we try to compute the Gini of every attribute.

Computation of Gini index for Money Attribute.

It has two possible values (Rich & Poor)

[7 instances] [3 instances]

for "Money = Poor", there are three instances with "Cinema"

$$\text{Gini(s) Money=Poor} = 1 - \left[\left(\frac{3}{3}\right)^2\right] = \boxed{0}$$

For "Money = Rich" there are two instances with "Tennis" 3 instances with "Cinema" & 1 instance with "Play in" & "Shopping" each

$$\text{Gini(s) Money=Rich} = 1 - \left[\left(\frac{2}{7}\right)^2 + \left(\frac{3}{7}\right)^2 + \left(\frac{1}{7}\right)^2 + \left(\frac{1}{7}\right)^2\right]$$

$$= \boxed{0.694}$$

Now, we need to compute weighted average for the "money attribute".

$$\text{Weighted Average (Money)} = 0 * \left(\frac{3}{10}\right) + 0.694 * \frac{7}{10} = \boxed{0.486}$$

(0 → Gini value; $\frac{3}{10}$ → Proportion)

Similarly we will compute the "Gini" for remaining attributes {Parents & Weather}.

Computation of Gini Index for Parents Attribute.

It has two possible values of Yes (5 instances) & No (5 instances)

for Parents = Yes, there are 5 instances all with Cinema

$$\therefore \text{Gini(s) Parents=Yes} = 1 - \left[\left(\frac{5}{5}\right)^2\right] = 0$$

For Parents = NO, there are two instances with "Tennis" one instance with "Stay in" "Shopping" & "Cinema" each.

∴ Gini(S) Parents = NO

$$= 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = \boxed{0.72}$$

Now lets compute Weighted Average (Parents)

$$= 0 * \left(\frac{5}{10}\right) + 0.72 * \left(\frac{5}{10}\right) = \boxed{0.36}$$

Computation of Gini index for Weather Attribute.

It has three possible values of Sunny (3 instances), Rainy (3 instances) & Windy (4 instances).

for Weather = Sunny, there are 2 instances with 'Cinema' & 1 with 'Tennis'

$$\text{Gini(S)}(\text{Weather} = \text{Sunny}) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = \boxed{0.444}$$

for Weather = Rainy, there are 2 instances with "Cinema" & 1 with 'Stay in'

$$\text{Gini(S)}(\text{Weather} = \text{Rainy}) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = \boxed{0.444}$$

for Weather = Windy, there are 3 instances with "Cinema" & 1 with "Shopping"

$$\text{Gini(S)}(\text{Weather} = \text{Windy}) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = \boxed{0.375}$$

Now Weighted Average (weather)

$$= 0.444 * \left(\frac{3}{10}\right) + 0.444 * \left(\frac{3}{10}\right) + 0.375 * \left(\frac{4}{10}\right)$$

$$= \boxed{0.416}$$

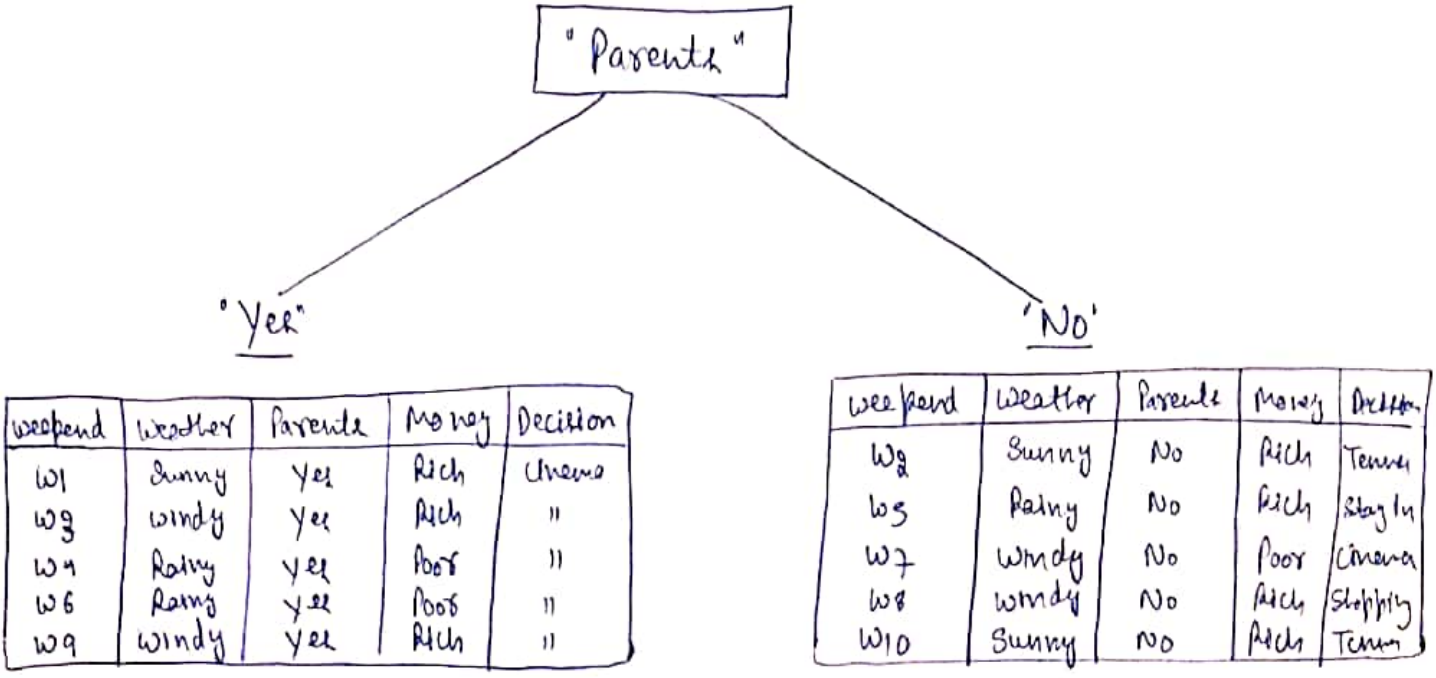
We have computed Gini index for all the three attributes :-

- For Weather – Gini index = 0.416
- For Parents – Gini index = 0.36
- For Money – Gini index = 0.486

Now out of these three, Gini index for Parents is minimum which means, it is having highest information.

So we select Parents as root node. (as it has smallest Gini index)

Now we have to divide the data, based on possible values of Parents [Yes, No]



weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	windy	Yes	Rich	"
W4	Rainy	Yes	Poor	"
W6	Rainy	Yes	Poor	"
W9	windy	Yes	Rich	"

weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	windy	No	Poor	Cinema
W8	windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Whenever Parents = 'Yes', we have a decision as 'Cinema' always.

No need to do computation any further from the end

P.T.O

When our Parents = No, we do not have a specific decision to take we have four decision [Tennis, Stay in Cinema, Shopping]

We will grow our tree using the subset.

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay in
W7	windy	No	Poor	Cinema
W8	windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Computation of Gini Index for Parents = No | Weather Attribute.

We have two instances of Sunny.

for Parents = No | Weather = Sunny, there are two examples/instances with "Tennis"

$$\text{Gini}(S) = 1 - \left[\frac{2}{2}\right]^2 = \boxed{0}$$

We have one instance of Rainy

for Parents = No | Weather = Rainy, there is 1 example with "Stay in".

$$\text{Gini}(S) = 1 - \left[\left(\frac{1}{1}\right)^2\right] = \boxed{0}$$

We have two instances of Windy

for Parents = No | Weather = Windy, there is 1 example with "Cinema" & 1 example with "Shopping"

$$\text{Gini}(S) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = \boxed{0.5}$$

Weighted Average (Parents = No | Weather) $= 0 * \left(\frac{2}{5}\right) + 0 * \left(\frac{1}{5}\right) + 0.5 * \left(\frac{2}{5}\right) = \boxed{0.2}$

P.T.O

Computation of Gini Index for Parents = No | Money Attribute)

We have four "Rich instances"

For Parents = No | Money = Rich, there is 1 example with "stay in" & "Shopping" each & 2 examples of "Tennis".

$\therefore$ Gini(S) = $1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$ = 0.625.

We have 1 "Poor instance"

For Parents = No | Money = Poor, there is 1 example with "Cinema".

$\therefore$ Gini(S) = $1 - \left[\left(\frac{1}{1}\right)^2\right] = 0$

$\therefore$ Weighted Average (Parents = No | Money) =

$0.625 * \left(\frac{4}{5}\right) + 0 * \left(\frac{1}{5}\right)$ = 0.5

for Parents = No | weather → Gini Index = 0.2

for Parents = No | Money → Gini Index = 0.5

Now weather is selected as it has the smallest "Gini index"

Now for weather we have three categories ∴ Sunny, Rainy & windy ∴ we will get three branches

Now subset of data when Parent = No & weather = Sunny we have all the instances as Tennis.

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W10	Sunny	No	Rich	Tennis

No dilemma

Now for Parents = No & weather = Rainy, we have all the instances as 'Stay in'.

Weekend	Weather	Parents	Money	Decision
W5	Rainy	No	Rich	Stay In

No dilemma

Now for Parents = No & weather = Windy, we need to split.

Weekend	Weather	Parents	Money	Decision
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping

We need to split

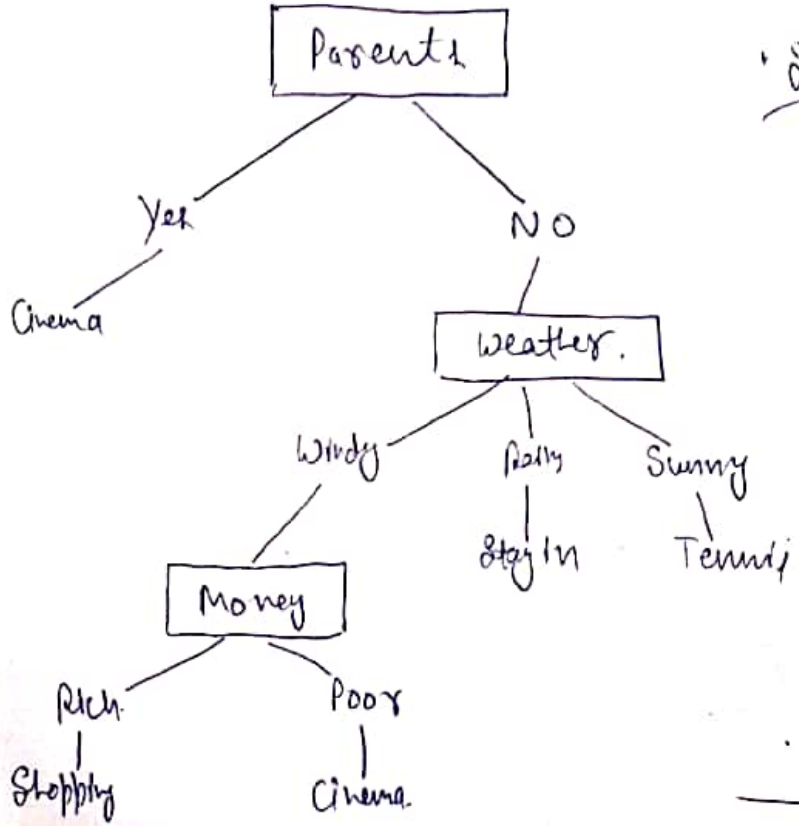
Now we need to consider remaining attribute.

When Parents = No, Weather = Windy, money = Poor, Decision is 'Cinema'

&

When Parents = No, Weather = Windy, Money = Rich, Decision is 'Shopping'.

∴ Decision Tree looks like :→



'Odd' 3pm' graph CLIO