ID3 Algorithm →

* ~~Inductive~~ Iterative Dichotomiser 3 Algorithm

Entropy → Measures of disorder in a System

* In a particular Node, all the Examples are ~~positive~~ or all are negative. that mean all the Examples belong to same class. then it is a Homogenous Set of Example and Entropy is low.

* How ever if we have two classes all the Examples half belong to one class and half belong to another class then Entropy is highest.

Information Gain → When we decide which attribute to split on we will use the principal of Information Gain.

* if all the Example have same target classification. then information Gain is high.
* if suppose 90% belong to one class then information gain is quiet high.
* 50% belong to one class and 50% another class then information Gain is low.

Gain is measure of how much we can reduce uncertainty. if the Example belong to the same class there is no uncertainty.

if the Example Spread among the classes almost uniformly there is high uncertainty.

What an Entropy does?

Entropy controls how a Decision tree decides to split the data. It actually effects how a Decision tree draws its boundaries.

What is information gain and, Why it is Matter in Decision Tree?

* measures how much "information" a feature gives us about the class.

Matter:-

* Information gain is the main key that is used by Decision Tree Algorithms to Construct a Decision tree.
* Decision Tree algorithm will always tries to Maximize information gain.
* An attribute with highest information Gain will tested / split first.

* ID3 is one of the first decision tree algorithm.
* only categorical attributes supported by ID3.

* C4-5 (Successor of ID3)

Improved Version of ID3

$$\text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{Split Info}(A)}$$

$$\text{Split Info}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|}$$

In attributes No of outcome ↑ split Info ↑

No of outcome ↓ split Info ↓

for Example

age	income	Student	credit_rating	buys_Computer
Youth	high	No	fair	No
Youth	high	No	Excellent	No
Middle-aged	high	No	fair	yes
Senior	Medium	No	fair	yes
Senior	low	yes	fair	yes
Senior	low	yes	Excellent	No
Middle-aged	low	yes	Excellent	yes
Youth	Medium	No	fair	No
Youth	low	yes	fair	yes
Senior	Medium	yes	fair	yes
Youth	Medium	yes	Excellent	yes
Middle-aged	Medium	No	Excellent	yes
Middle-aged	high	yes	fair	yes
Senior	Medium	No	Excellent	No

buys_ Computer	
yes	NO
9	5

Entropy

$$\text{info}(D) = (I\,9,5) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$\Rightarrow 0.9403$$

(i) Calculate Entropy of class attribute :-

(ii) Calculate Gain Ration of all other attributes.

		class		
		Yes	NO	
age	Youth	2	3	5
	Middle-aged	4	0	4
	Senior	3	2	5
				14

$$\text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{Split Info}(A)}$$

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

Info of Dataset — Info of Attribute

$$-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$\text{Info}_A(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} \times \text{Info}(D_j)$$

$$\text{Info}_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2)$$

$$\Rightarrow \frac{5}{14} \times 0.971 + \frac{4}{14} * 0 + \frac{5}{14} * 0.971$$

$$\Rightarrow 0.3467 + 0 + 0.3467$$

$$\Rightarrow 0.6934$$

$$\text{Gain}(age) \Rightarrow \text{Info}(D) - \text{Info}_{age}(D)$$

$$\Rightarrow 0.9403 - 0.6934$$

$$\Rightarrow 0.2469$$

$$\text{Split Info}_{age}(D) = -\frac{5}{14} * \log_2\left(\frac{5}{14}\right) - \frac{4}{14} * \log_2\left(\frac{4}{14}\right) - \frac{5}{14} * \log_2\left(\frac{5}{14}\right)$$

$$\Rightarrow 1.5774$$

Gain Ration (age) = $\frac{0.246}{1.5774}$

$\Rightarrow 0.1559$

		class		
		yes	No	
Income	low	3	1	4
	Medium	4	2	6
	high	2	2	4
				14

$\text{Info}_{\text{income}}(D) = \frac{4}{14} I(3,1) + \frac{6}{14} I(4,2) + \frac{4}{14} I(2,2)$

$\Rightarrow \frac{4}{14} * 0.8113 + \frac{6}{14} * 0.9183 + \frac{4}{14} * 1$

$\Rightarrow 0.2318 + 0.3935 + 0.2857$

$\Rightarrow 0.911$

Gain (income) = $0.9403 - 0.911$

$\Rightarrow 0.0293$

Split Info (income) $(D) \Rightarrow -\frac{4}{14} * \log\left(\frac{4}{14}\right) - \frac{6}{14} * \log_2\left(\frac{6}{14}\right) -$

$\frac{4}{14} * \log_2\left(\frac{4}{14}\right)$

$\Rightarrow 1.5566$

Gain Ration (income) = $\frac{0.0293}{1.5566}$

$\Rightarrow 0.0188$

		class		
		yes	NO	
student	yes	6	1	7
	NO	3	4	7
				14

$\text{Info student}(D) = \frac{7}{14} I(6,1) + \frac{7}{14} I(3,4)$

$\Rightarrow \frac{7}{14} * 0.5917 + \frac{7}{14} * 0.9852$

$\Rightarrow 0.2958 + 0.4926$

$\Rightarrow 0.7884$

Gain (Student) $\Rightarrow$ 0.9403 − 0.7884

$\Rightarrow$ 0.1519

$\text{Split info}_{\text{student}}^{(D)} = -\frac{7}{14} * \log_2\left(\frac{7}{14}\right) - \frac{7}{14} * \log_2\left(\frac{7}{14}\right)$

$\Rightarrow 1$

Gain Ratio (Student) $\Rightarrow \frac{0.1519}{1} \Rightarrow 0.1519$

		class		
		Yes	No	
Credit rating	Fair	6	2	8
	Excellent	3	3	6
				14

$\text{Info}_{\text{credit rating}}^{(D)} = \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3)$

$\Rightarrow \frac{8}{14} * 0.8113 + \frac{6}{14} * 1$

$\Rightarrow 0.4636 + 0.4286$

$\Rightarrow 0.8922$

Gain (Credit rating) $\Rightarrow 0.9403 - 0.8922$

$\Rightarrow 0.0481$

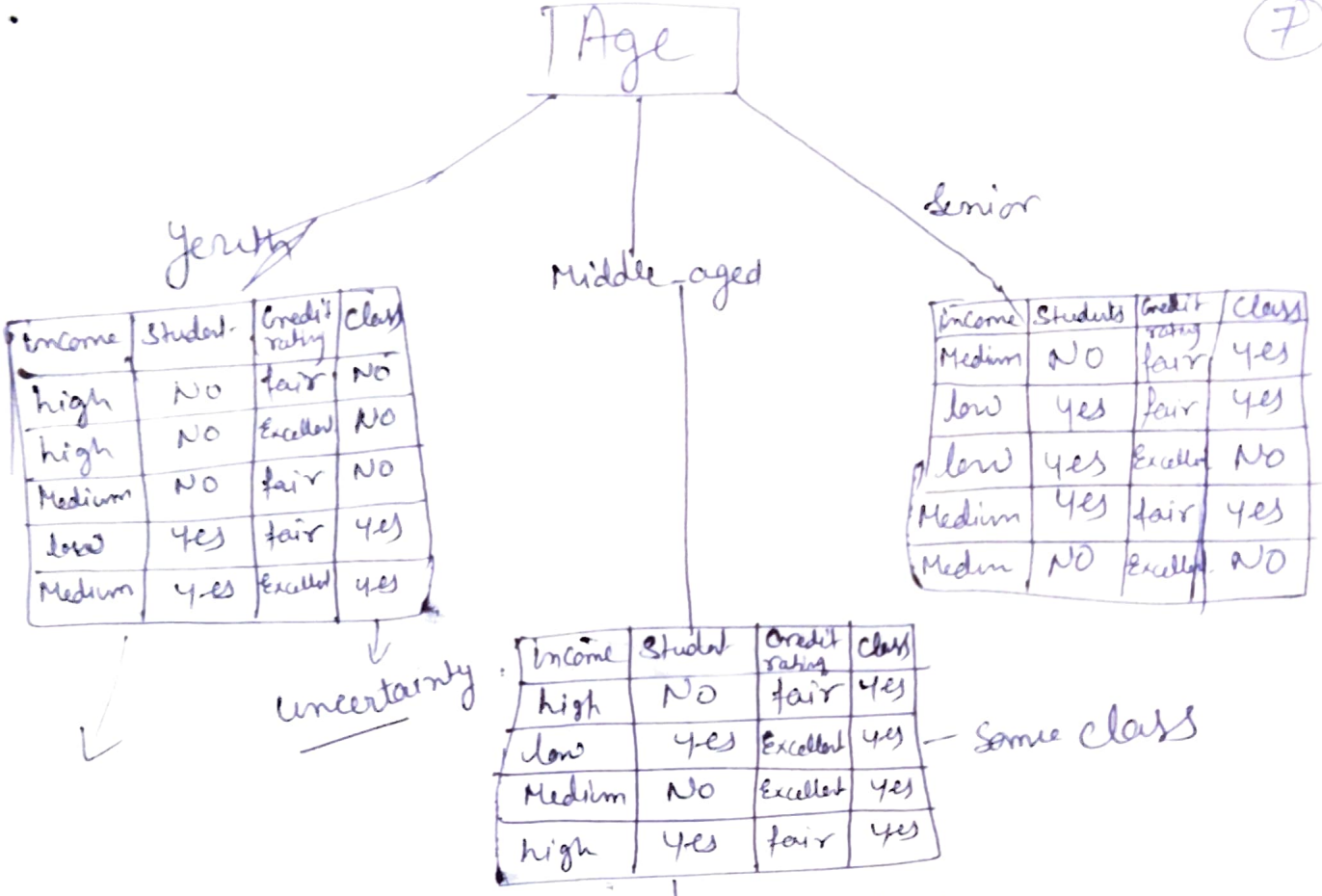
$\text{Split Info}_{\text{credit-rating}}^{(D)} \Rightarrow -\frac{8}{14} * \log_2\left(\frac{8}{14}\right) - \frac{6}{14} \log_2^{\left(\frac{6}{14}\right)}$

$\Rightarrow 0.9852$.

Gain Ratio (Credit_rating) $\Rightarrow \frac{0.0481}{0.9852}$

$\Rightarrow 0.0488$

* As the Gain Ratio of "age" is highest.
* Age is the best attributes & become the Root Node.



Income	Student	Credit rating	Class
high	NO	fair	NO
high	NO	Excellent	NO
Medium	NO	fair	NO
low	yes	fair	yes
Medium	yes	Excellent	yes

Income	Student	Credit rating	Class
Medium	NO	fair	yes
low	yes	fair	yes
low	yes	Excellent	NO
Medium	yes	fair	yes
Medium	NO	Excellent	NO

Income	Student	Credit rating	Class
high	NO	fair	yes
low	yes	Excellent	yes
Medium	NO	Excellent	yes
high	yes	fair	yes

↓ replace it by yes

buys Computer

yes	NO
2	3

$$info(D) = I(2,3) \Rightarrow -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$\Rightarrow 0.971$$

Calculate Gain Ratio of all other attributes

Income		Class		
		yes	NO	
	low	1	0	1
	Medium	1	1	2
	high	0	2	2
				5

$$Info_{income}(D) = \frac{1}{5}I(1,0) + \frac{2}{5}I(1,1) + \frac{2}{5}I(0,2)$$

$$\Rightarrow \frac{1}{5} * 0 + \frac{2}{5} * 1 + \frac{2}{5} \times 0$$

$$\Rightarrow 0 + 0.4 + 0$$

$$\Rightarrow 0.4$$

$$Gain(income) = 0.971 - 0.4$$

$$\Rightarrow 0.571$$

Split Info$_{income}(D) = -\frac{1}{5} * \log_2\left(\frac{1}{5}\right) - \frac{2}{5} * \log_2\left(\frac{2}{5}\right) - \frac{2}{5} * \log_2\left(\frac{2}{5}\right)$

$\Rightarrow 1.5219$

Gain Ratio (Income) $\Rightarrow \frac{0.571}{1.5219}$

$\Rightarrow 0.3751$

		class		
		Yes	NO	
Credit rating	fair	1	2	3
	Excellent	1	1	2
				5

Info$_{\text{Credit rating}}(D) \Rightarrow \frac{3}{5} I(1,2) + \frac{2}{5} I(1,1)$

$\Rightarrow \frac{3}{5} * 0.9183 + \frac{2}{5} * 1$

$\Rightarrow 0.3443 + 0.4$

$\Rightarrow 0.7443$

Gain(credit rating) $\Rightarrow 0.971 - 0.7443$

$\Rightarrow 0.2267$

Split Info$_{(\text{credit rating})} = -\frac{3}{5} * \log_2\left(\frac{3}{5}\right) - \frac{2}{5} * \log_2\left(\frac{2}{5}\right)$

$\Rightarrow 0.9709$

Gain Ratio$_{(\text{Credit Rating})} \Rightarrow \frac{0.2267}{0.9709}$

$\Rightarrow 0.2335$

		class		
		Yes	NO	
Student	Yes	2	0	2
	NO	0	3	3
				5

Info$_{\text{Studn}}(D) = \frac{2}{3} I(2,0) + \frac{3}{5} I(0,3)$

$\Rightarrow \frac{2}{5} * 0 + \frac{3}{5} * 0$

$\Rightarrow 0$

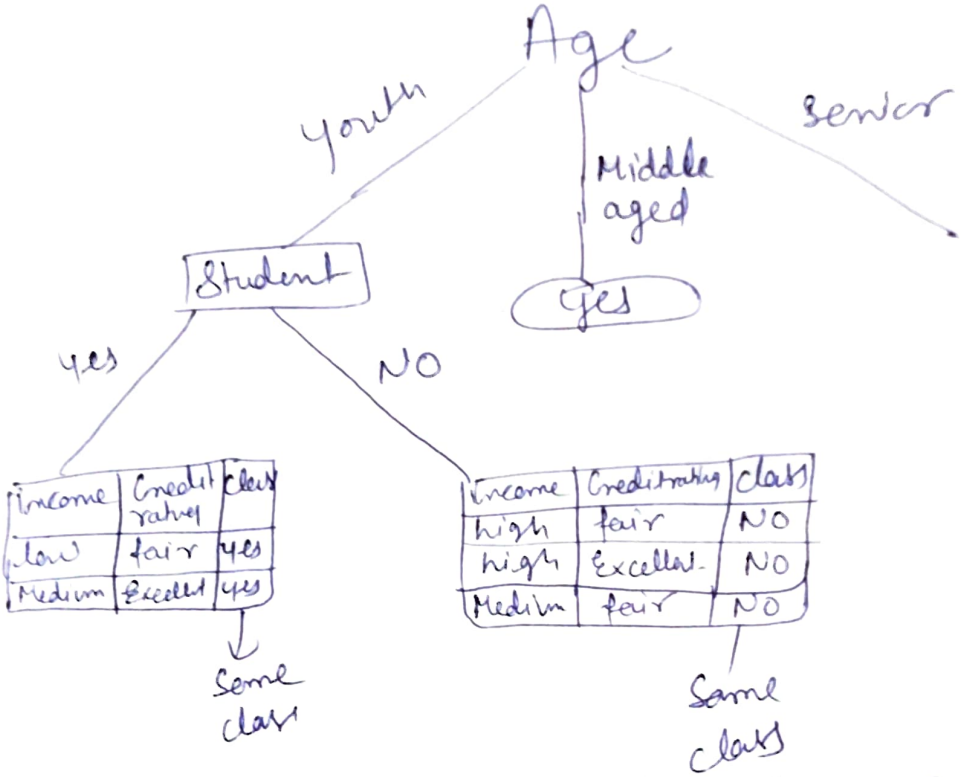
Gain(Studn) $\Rightarrow 0.971 - 0 \Rightarrow 0.971$

Split Info $_{Student} \Rightarrow -\frac{2}{5} * \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \log_2\left(\frac{3}{5}\right)$

$\Rightarrow 0.9709$

Gain Ratio $_{(Student)} \Rightarrow \frac{0.971}{0.9709}$

$\Rightarrow 1$



Income	Credit rating	class
low	fair	yes
Medium	Excellent	yes

Income	Credit rating	class
high	fair	NO
high	Excellent	NO
Medium	fair	NO

Same we will do for Senior

buys Computer

yes NO

3 2

$Info(D) \Rightarrow I(3,2) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$

$\Rightarrow 0.971$

		class yes	class NO	
income	low	1	1	2
	Medium	2	1	3
	high	0	0	0
				5

$Info_{income}(D) = \frac{2}{5} I(1,1) + \frac{3}{5} I(2,1)$

$\Rightarrow \frac{2}{5} * 1 + \frac{3}{5} * 0.9183$

$\Rightarrow 0.4 + 0.551$

$\Rightarrow 0.951$

$Gain_{(income)} \Rightarrow 0.971 - 0.951 \Rightarrow 0.02$

$Split\ Info_{income}(D) \Rightarrow -\frac{2}{5} log_2(\frac{2}{5}) - \frac{3}{5} log_2(\frac{3}{5})$

$\Rightarrow 0.9709$

$Gain\ Ratio(income) \Rightarrow \frac{0.02}{0.9709}$

$\Rightarrow 0.0205$

		Class		
		Yes	NO	
Credit rating	fair	3	0	3
	Excellent	0	2	2
				5

$Info_{Credit\ rating}(D) = \frac{3}{5} I(3,0) + \frac{2}{5} I(0,2)$

$\Rightarrow \frac{3}{5} * 0 + \frac{2}{5} * 0$

$\Rightarrow 0$

$Gain_{(Credit\ rating)} \Rightarrow 0.971 - 0 \Rightarrow 0.971$

$Split\ Info_{(Credit\ rating)}(D) \Rightarrow -\frac{3}{5} log_2(\frac{3}{5}) - \frac{2}{5} log_2 \frac{2}{5}$

$\Rightarrow 0.9709$

$Gain\ Ratio_{(Credit\ rating)} \Rightarrow \frac{0.971}{0.9709} \Rightarrow 01$

		Class		
		Yes	NO	
Student	Yes	2	1	3
	NO	1	1	2
				5

$Info_{student}(D) \Rightarrow \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1)$

$\Rightarrow \frac{3}{5} * 0.9183 + \frac{2}{5} * 1$

$\Rightarrow 0.551 + 0.4$

$\Rightarrow 0.951$

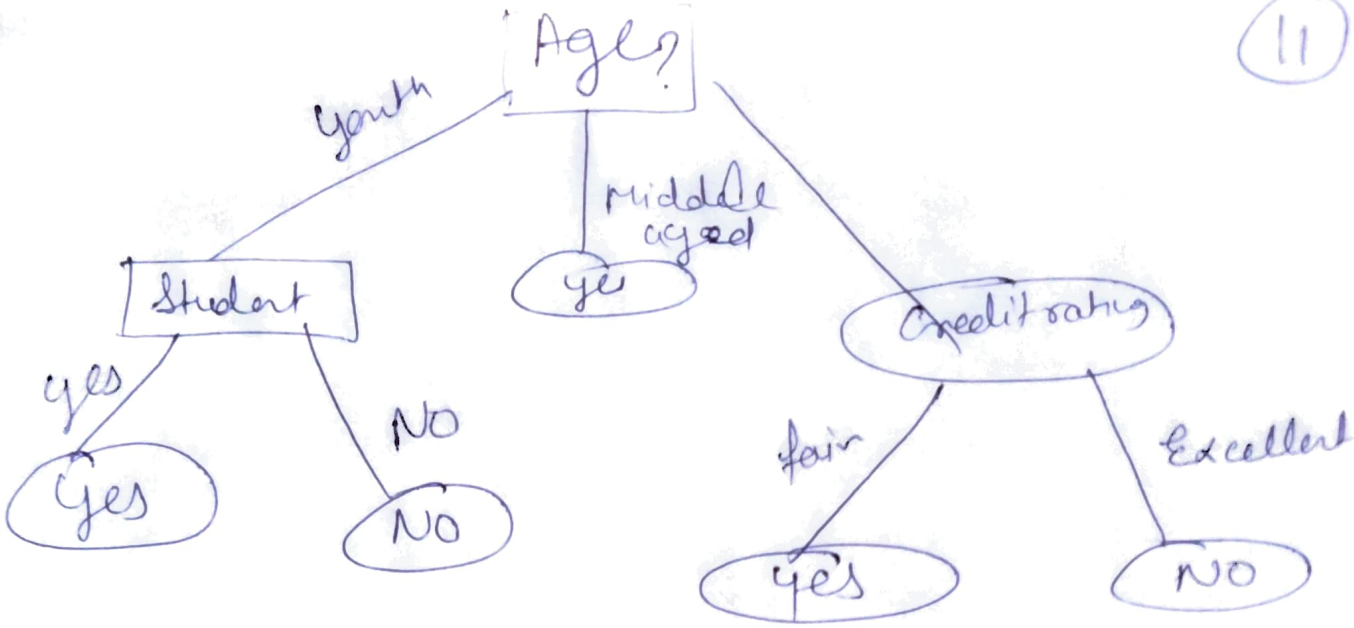
$Gain_{(Student)} \Rightarrow 0.971 - 0.951$

$\Rightarrow 0.02$

$Split\ Info_{stude} \Rightarrow -\frac{3}{5} log_2(3/5) - \frac{2}{5} * log_2(\frac{2}{5})$

$\Rightarrow 0.9702$

$Gain\ Ratio = \frac{0.02}{0.9702} = 0.02$


Age?
Youth
Middle aged
Student
Credit rating
yes
NO
fair
Excellent
yes
NO
yes
NO