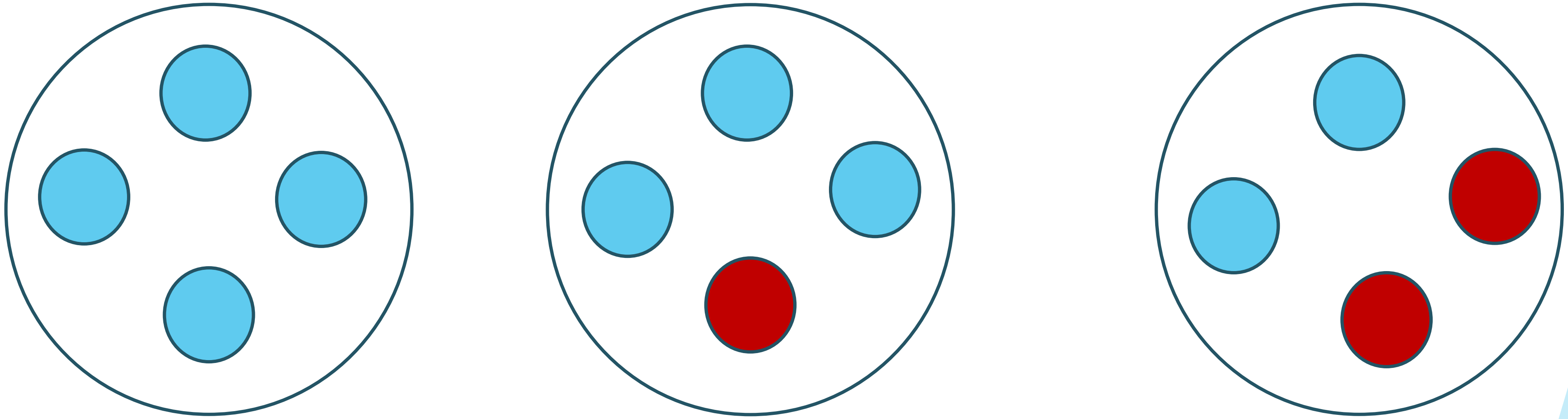
Decision tree

Entropy

$$\text{Entropy(s)} = \sum_{i=1}^{c} - p_i \, log_2^{p_i}$$



Gain

$$Gain(S, A) = Entropy(S) - \sum_{v \in values} \frac{|S_v|}{|s|} \text{Entropy}\,(\,S_v)$$

	Age	income	student	credit_rating	buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

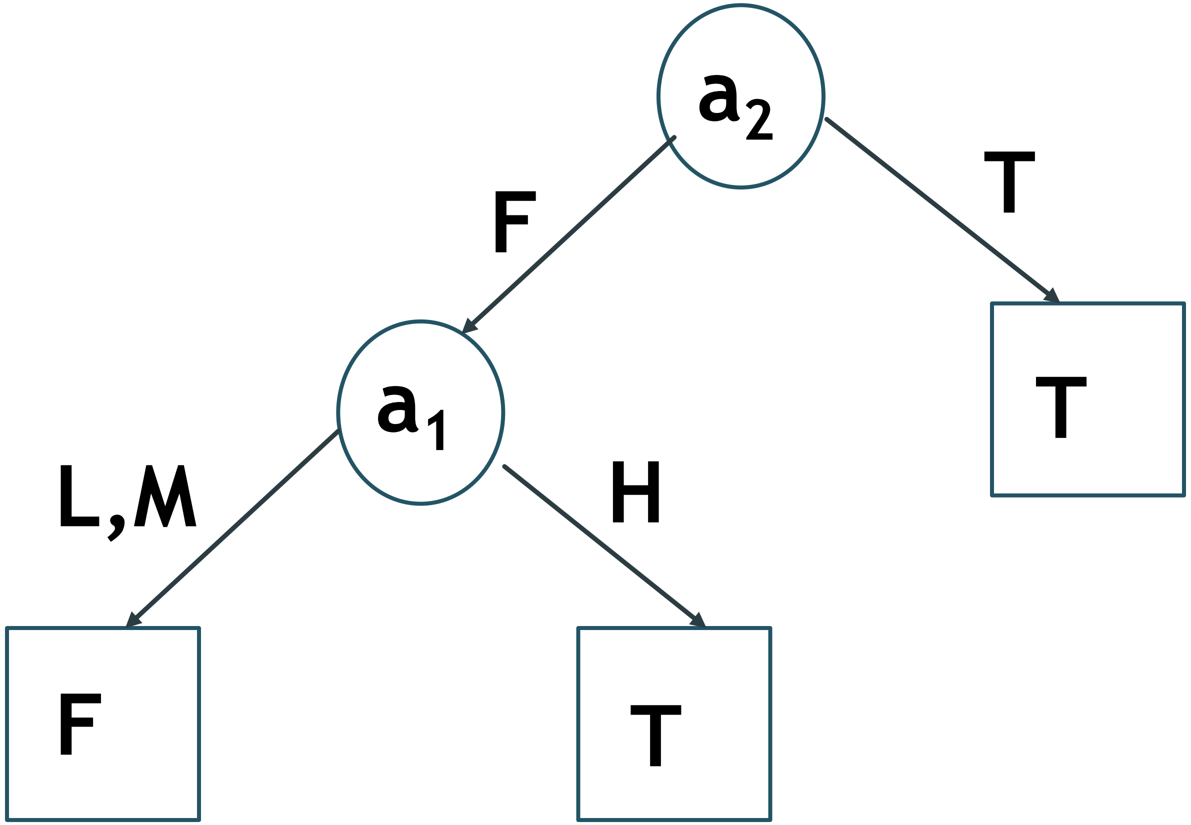
بر اساس جدول زیر و بر مبنای معیار آنتروپی بهترین ویژگی برای قرار گرفتن در ریشه درخت تصمیم کدام است؟

	a1	a2	class
1	Low	F	FALSE
2	low	M	TRUE
3	medium	F	FALSE
4	medium	M	TRUE
5	high	F	TRUE
6	high	M	TRUE

	a1	a2	class
1	Low	F	FALSE
2	low	M	TRUE
3	medium	F	FALSE
4	medium	M	TRUE
5	high	F	TRUE
6	high	M	TRUE

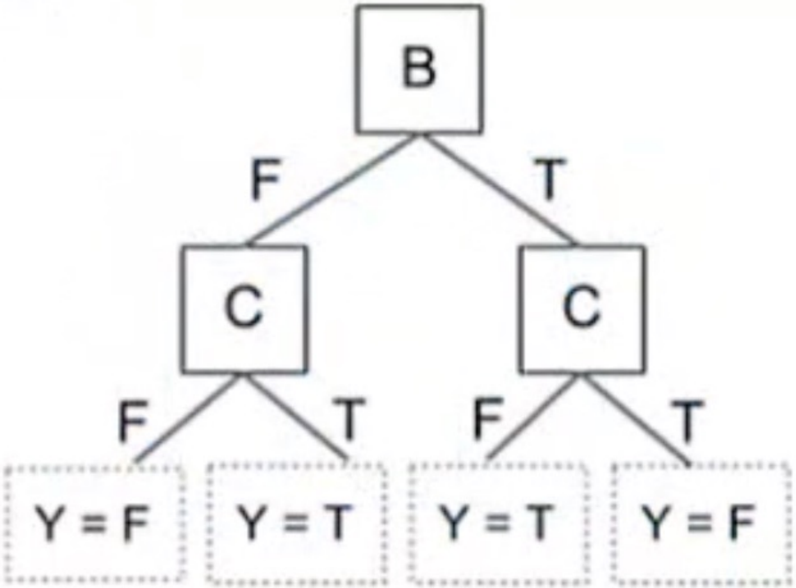
feature	a1	a2
gain	0.25	0.78

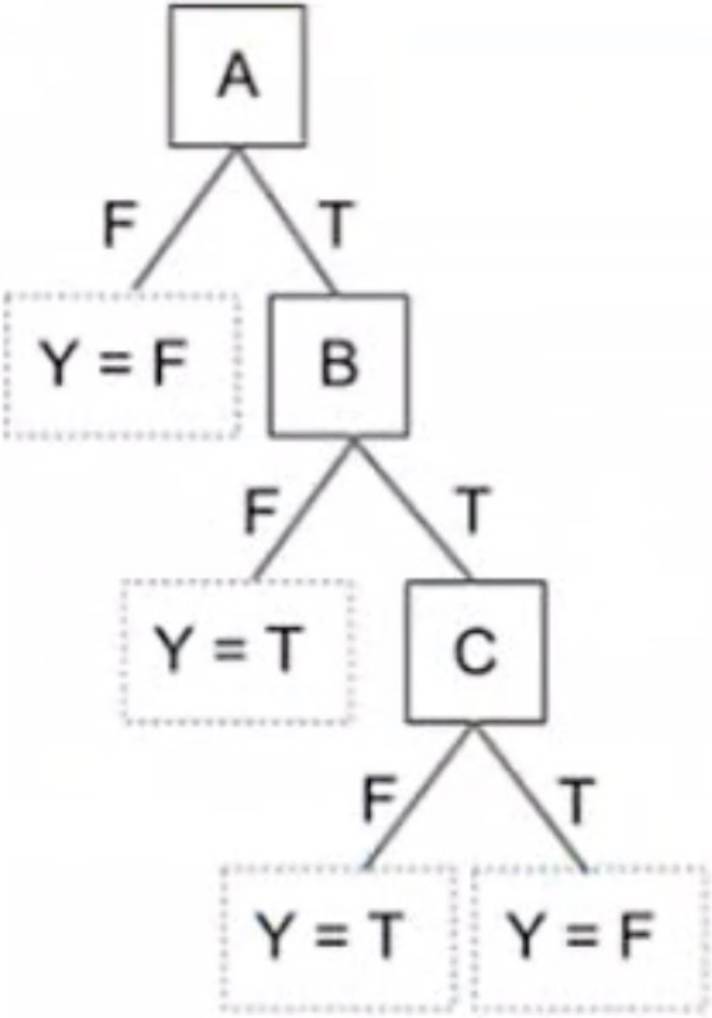
درخت تصمیم مثال به صورت زیر است؟



Decision Tree is a greedy algorithm

A	B	C	Y
F	F	F	F
T	F	T	T
T	T	F	T
T	T	T	F





ID3 (Examples,Target_Attribute,Attributes)

Create a root node for the tree

If all examples are positive, return the single-node tree Root, with label = +

If all examples are negative, return the single-node tree Root, with label = -

If number of predicting attributes is empty then

    return Root, with label = most common value of the target attribute in the examples

else

    A = The Attribute that best classifies examples.

    Testing attribute for Root = A.

    for each possible value, v_i, of A

        Add a new tree branch below Root, corresponding to the test A = v_i .

        Let Examples(v_i) be the subset of examples that have the value for A

        if Examples(v_i) is empty then

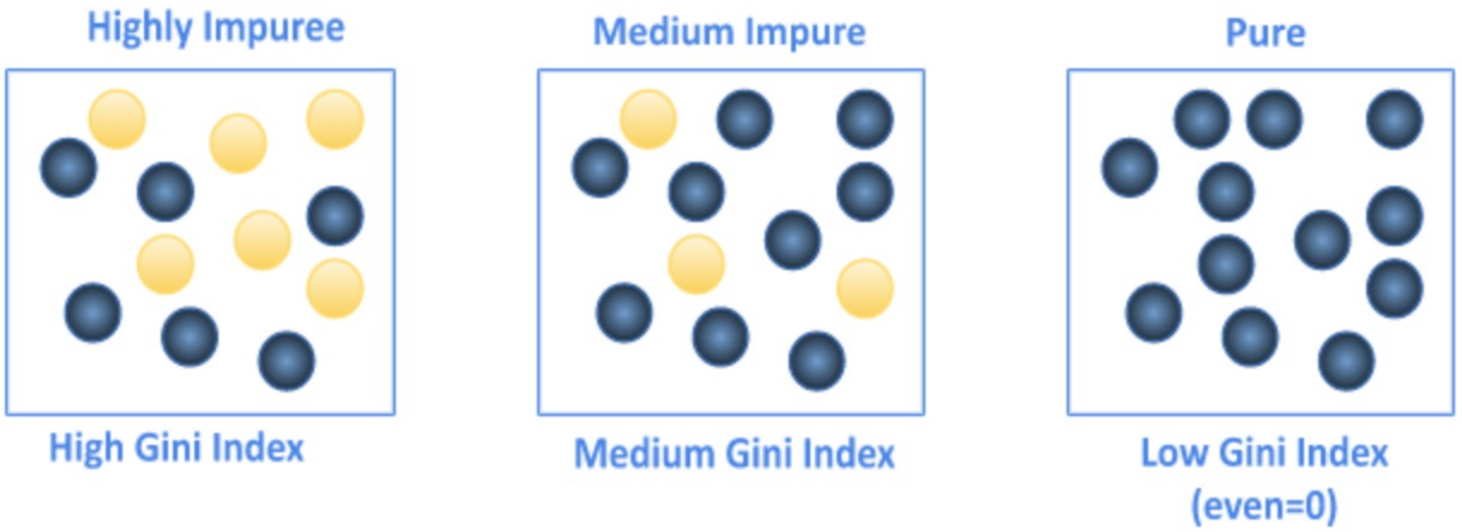
            below this new branch add a leaf node with label = most common target value in the examples

        else below this new branch add subtree **ID3 (Examples(v_i),Target_Attribute,Attributes – {A})**

return Root

Gini

شاخص یا ضریب جینی بصورت کلی نابرابری را در میان مقادیر مختلف یک متغیر اندازه‌گیری می‌کند. هرچه این شاخص بالاتر باشد، داده‌ها پراکنده‌تر هستند.



Gini mathematical formula

The computation of the Gini index is as follows:

$$Gini\ (t) = 1 - \sum_{i=1}^{c} p_i^2$$

Where c is number of classes

Savings	Assets	Salary ($1000)	Credit Note
High	High	20	Good
Low	High	25	Bad
High	Low	30	Good
Low	Low	35	Bad
Low	High	40	Good
Low	Low	50	Bad
High	Low	90	Good

Gini(credit Note) = 1 $-p_{Good}^2 - p_{Bad}^2$

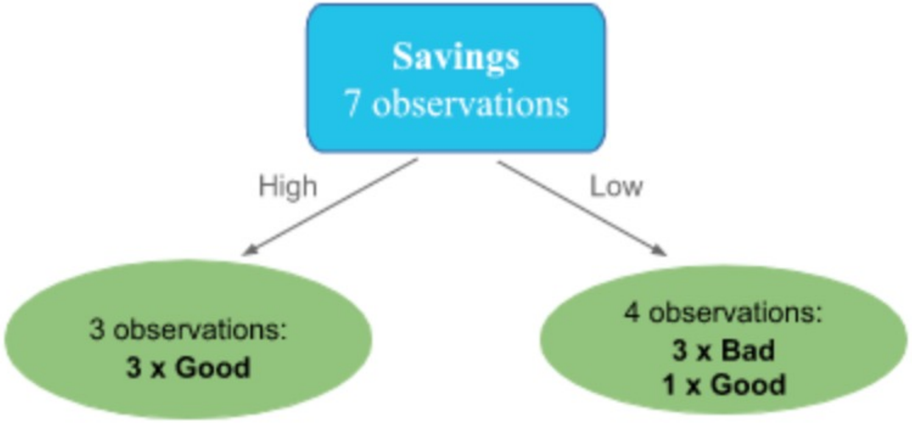
Gini(credit Note) = 1 - $(\frac{4}{7})^2 - (\frac{3}{7})^2$ = 0.49

Gini index for each feature

(weighted average value of the Gini index)

$$\text{Gini index (A)} = \sum_{i=1}^{c} \frac{n_i}{n} \, Gini(i)$$

Where A is an attrinbute



	p	p^2
Good	1 (3/3)	1
Bad		
Gini Index	0	

weight 0,43 (3/7)

	p	p^2
Good	0,25 (1/4)	0,0625
Bad	0,75 (3/4)	0,563
Gini Index	0,375	

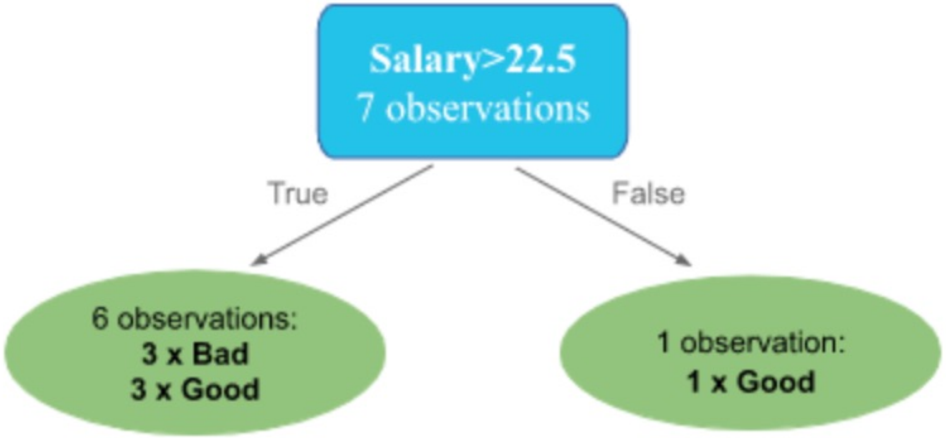
1-0.0625-0.563

weight 0,57 (4/7)

weighted average Gini Index=
3/7 * 0 + 4/7 * 0.375
0,214

Savings	Assets	Salary ($1000)	Credit Note
High	High	20	Good
Low	High	25	Bad
High	Low	30	Good
Low	Low	35	Bad
Low	High	40	Good
Low	Low	50	Bad
High	Low	90	Good

Savings	Assets	Salary ($1000)	Credit Note
High	High	20	Good
Low	High	25	Bad
High	Low	30	Good
Low	Low	35	Bad
Low	High	40	Good
Low	Low	50	Bad
High	Low	90	Good



	p	p^2
Good	0,5 *(3/6)*	0,25
Bad	0,5 *(3/6)*	0,25
Gini Index		0,5
		1-0.25-0.25

weight 0,86 *(6/7)*

	p	p^2
Good	1 *(1/1)*	1
Bad		
Gini Index		0

weight 0,14 *(1/7)*

weighted average Gini Index=
6/7 * 1 + 1/7 * 0
0,429

Savings	Assets	Salary ($1000)	Credit Note
High	High	20	Good
Low	High	25	Bad
High	Low	30	Good
Low	Low	35	Bad
Low	High	40	Good
Low	Low	50	Bad
High	Low	90	Good

feature	Saving	Assets	Salary
Gini index	0.241	0.467	0.429

The lowest is the value of the Gini Index, the better is the feature selected to split the node. Thus leading to more pure subsets for the branches.

So, we select saving to split the node

Gini vs Entropy

Entropy(s) = $\sum_{i=1}^{c} - p_i \, log_2^{p_i}$

$$Gini\ (t) = 1 - \sum_{i=1}^{c} p_i^2$$