Math behind Decision Tree Algorithm

MLMath.io Feb 20, 2019 · 9 min read

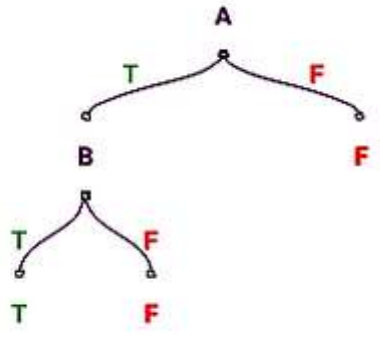
Decision tree algorithm is one of the most popular machine learning algorithm. It is a supervised machine learning algorithm, used for both classification and regression task. It is a model that uses set of rules to classify something.

This is the PART I of Decision Tree Tutorial.

Link For PART II DECISION TREE TUTORIAL

Lets see decision tree with this simple example, It is normal “AND’ operation problem, where ‘A’, ‘B’ are features and “A and B” are corresponding labels.

A	B	A AND B
F	F	F
F	T	F
T	F	F
T	T	T



Source Hackerearth

If A=F then result=F

If A=T and B=T, then result=T

If A=T and B=F, then result = F

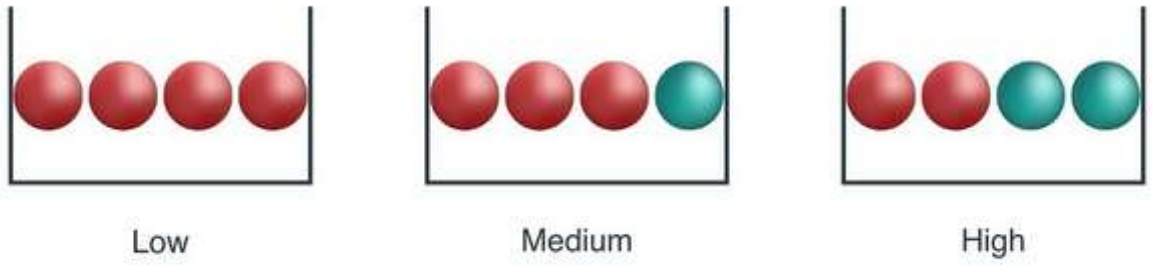
This is a an example of binary classifier. It classify “And” operation is ‘False’ or ‘True’.

This story is about understanding the full mathematical concept of Decision tree algorithm. So,lets break it!!!

Before diving deeper into the concept lets understand about impurity, types of measures of impurity. it will help us to understand the algorithm better.

Impurity

Let's understand Impurity with the following toy example,



source Medium

From above image, a ball is randomly drawn from each bowl. So how much information you needed to accurately tells the color of ball. So, left bowl needed less information as all of the ball is red colored, central bowl needed more information than left bowl to tell it accurately, and right bowl needed maximum information as both number of both color ball are same.

As information is measure of purity, so we can say that left bowl is pure node, middle is less impure and right is more impure.

> ***So, how can we measure impurity in sample??***

There are couple of impurity measures are there, but in this story we will talk about only two such measure,

1. Entropy
2. Gini index/ Gini impurity

Entropy

Entropy is amount of information is needed to accurately describe the some sample. So if sample is homogeneous, means all the element are similar than Entropy is 0, else if sample is equally divided than entropy is maximum 1.

So, left bowl has lowest entropy, middle bowl has more entropy and right bowl has highest entropy.

Mathematically it is written as ,

$$Entropy = -\sum_{i=1}^{n} p_i * \log(p_i)$$

Gini index / Gini impurity

Gini index is measure of inequality in sample. It has value between 0 and 1. Gini index of value 0 means sample are perfectly homogeneous and all element are similar, whereas, Gini index of value 1 means maximal inequality among elements. It is sum of the square of the probabilities of each class. It is illustrated as,

$$Gini\,index = 1 - \sum_{i=1}^{n} p_i^2$$

i is number of classes

> ***So what is the importance of impurity measure in decision tree??***

Impurity measures the homogeneity in the data sample. If the sample is homogeneous then sample are from same class.

Decision Tree Algorithm

Decision tree algorithm is a tree where nodes represents features(attributes), branch represents decision(rule) and leaf nodes represents outcomes(discrete and continuous).

So how decision tree algorithm are built actually??

There are Various algorithm that are used to generate decision tree from data, some are as following,

1. Classification and regression tree CART
2. ID 3
3. CHAID
4. ID 4.5

In this tutorial we will only talk about CART and next tutorial will explain concept of ID3 algorithm. These are mostly used in the industry.

> So lets start with a generating a classification tree with the help of CART algorithm,

CART

1. It is used for generating both classification tree and regression tree.
2. It uses Gini index as metric/cost function to evaluate split in feature selection in case of classification tree.

3. It is used for binary classification.
4. It use least square as a metric to select features in case of Regression tree.

> *Lets start with generating classification tree.*

Lets start with weather data set, which is quite famous in explaining decision tree algorithm,where target is to predict play or not(Yes or No) based on weather condition.

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

From data, outlook, temperature, humidity, wind are the features of data.

So, lets start building tree,

Outlook

Outlook is a nominal feature. it can take three value, sunny, overcast and rain. Lets summarize the final decision for outlook features,

Outlook	Yes	No	# Instances
sunny	2	3	5
overcast	4	0	4
rainfall	3	2	5

Gini index (outlook=sunny)= $1-(2/5)^2-(3/5)^2 = 1- 0.16–0.36 = 0.48$

Gini index(outlook=overcast)= $1- (4/4)^2-(0/4)^2 = 1- 1- 0 = 0$

Gini index(outlook=rainfall)= $1- (3/5)^2 -(2/5)^2 = 1- 0.36- 0.16 = 0.48$

Now , we will calculate the weighted sum of Gini index for outlook features,

Gini(outlook) = (5/14)*0.48 + (4/14) *0 + (5/14)*0.48 = 0.342

Temperature

Similarly, temperature is also a nominal feature, it can take three values, hot,cold and mild. lets summarize the final decision of temperature feature,

Temperature	Yes	No	# Instances
hot	2	2	4
cool	3	1	4
mild	4	2	6

Gini(temperature=hot) = $1-(2/4)^2-(2/4)^2 = 0.5$

Gini(temperature=cool) = $1-(3/4)^2-(1/4)^2 = 0.375$

Gini(temperature=mild) = $1-(4/6)^2-(2/6)^2 = 0.445$

Now, the weighted sum of Gini index for temperature features can be calculated as,

Gini(temperature)= (4/14) *0.5 + (4/14) *0.375 + (6/14) *0.445 =0.439

Humidity

Humidity	Yes	No	# Instances
high	3	4	7
Normal	6	1	7

Humidity is a binary class feature , it can take two value high and normal.

Gini(humidity=high) = $1-(3/7)^2-(4/7)^2 = 0.489$

Gini(humidity=normal) = $1-(6/7)^2-(1/7)^2 = 0.244$

Now, the weighted sum of Gini index for humidity features can be calculated as,

Gini(humidity) = (7/14) *0.489 + (7/14) *0.244=0.367

Wind

wind	Yes	No	# Instances
weak	6	2	8
strong	3	3	6

wind is a binary class feature , it can take two value weak and strong.

Gini(wind=weak)= $1-(6/8)^2-(2/8)^2 = 0.375$

Gini(wind=strong)= $1-(3/6)^2-(3/6)^2= 0.5$

Now, the weighted sum of Gini index for wind features can be calculated as,

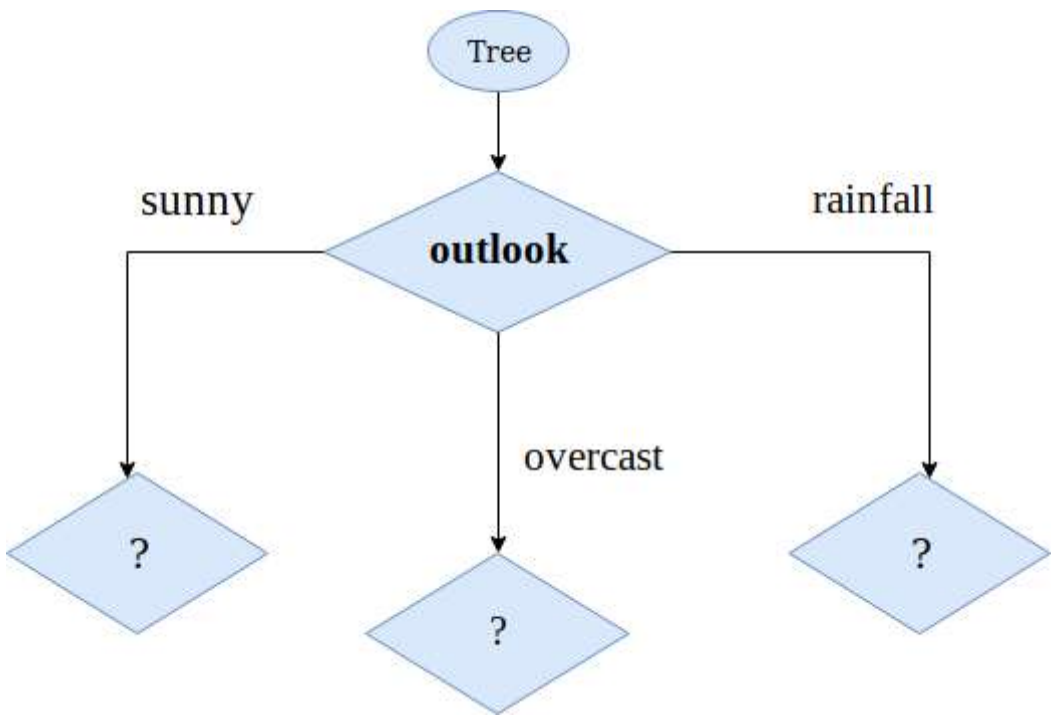
Gini(wind) = $(8/14) *0.375 + (6/14) *0.5=0.428$

Decision for root node

So,the final decision of all the features,

Features	Gini Index
outlook	0.342
temperature	0.439
humidity	0.367
wind	0.428

From table, you can seen that Gini index for outlook feature is lowest. So we get our root node.



Lets calculate the Gini index on sub data set for outlook feature, as you can seen we have three sub data section, sunny, overcast, and rainfall of outlook feature. we will use same method as above to find next split.

> ***So , lets focus on sub data on sunny outlook feature. we need to find the Gini index for temperature, humidity and wind feature respectively.***

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

Gini index for temperature on sunny outlook

Temperature	Yes	No	# Instances
hot	0	2	2
cool	1	1	1
mild	1	1	2

Gini(outlook=sunny & temperature=hot) = $1-(0/2)^2-(2/2)^2 = 0$

Gini(outlook=sunny & temperature=cool) = $1-(1/1)^2-(0/1)^2 = 0$

Gini(outlook=sunny & temperature=mild) = $1-(1/2)^2-(1/2)^2 = 0.5$

Now, the weighted sum of Gini index for temperature on sunny outlook features can be calculated as,

Gini(outlook=sunny & temperature)= (2/5) *0 + (1/5) *0+ (2/5) *0.5 =0.2

Gini Index for humidity on sunny outlook

Humidity	Yes	No	# Instances
high	0	3	3

Normal	2	0	2

Gini(outlook=sunny & humidity=high) = $1-(0/3)^2-(3/3)^2 = 0$

Gini(outlook=sunny & humidity=normal) = $1-(2/2)^2-(0/2)^2 = 0$

Now, the weighted sum of Gini index for humidity on sunny outlook features can be calculated as,

Gini(outlook = sunny & humidity) = (3/5) *0 + (2/5) *0=0

Gini Index for wind on sunny outlook

wind	Yes	No	# Instances
weak	1	2	3
strong	1	1	2

Gini(outlook=sunny & wind=weak) = $1-(1/3)^2-(2/3)^2 = 0.44$

Gini(outlook=sunny & wind=strong) = $1-(1/2)^2-(1/2)^2 = 0.5$

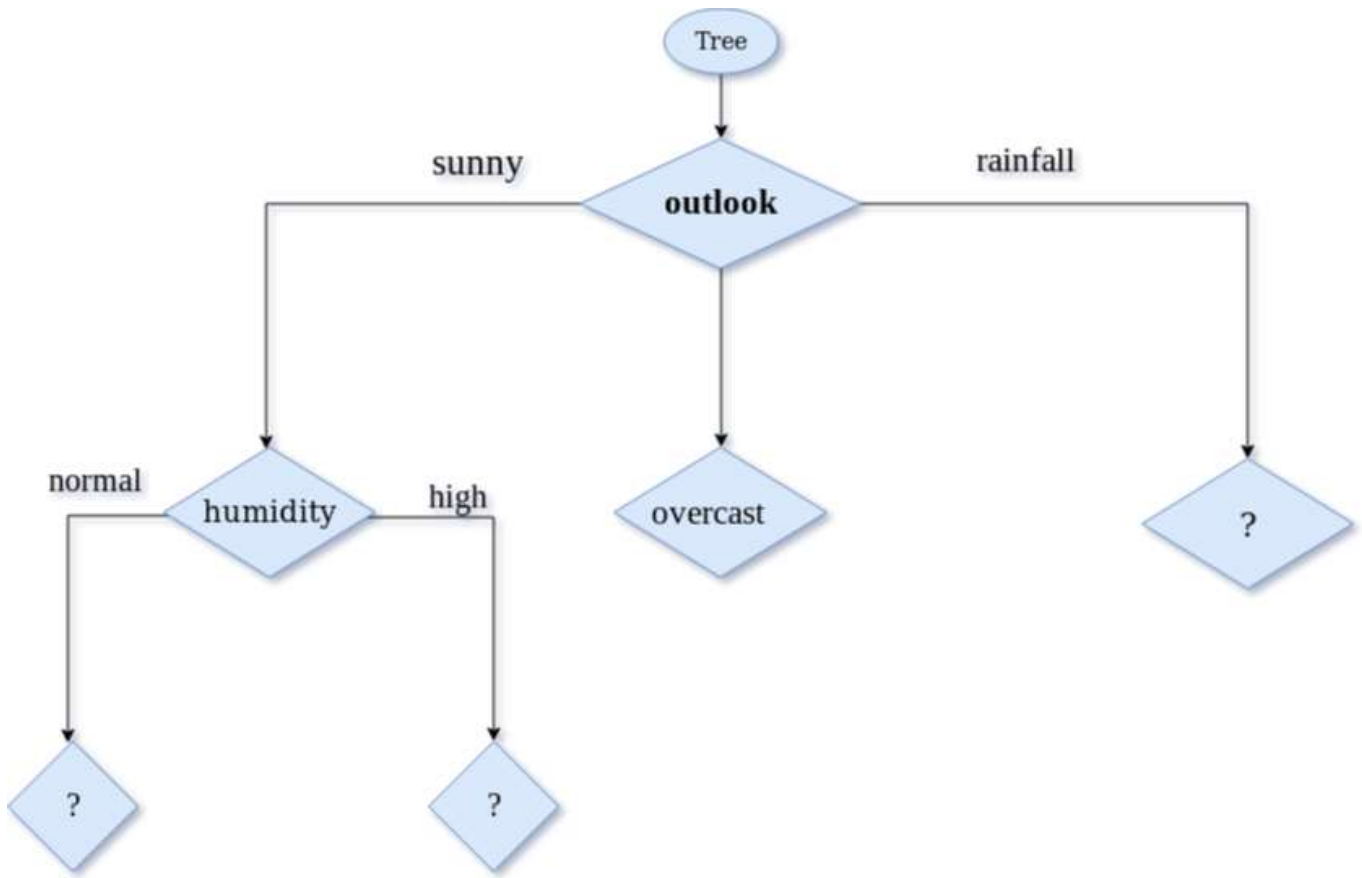
Now, the weighted sum of Gini index for wind on sunny outlook features can be calculated as,

Gini(outlook = sunny and wind) = (3/5) *0.44 + (2/5) *0.5=0.266+0.2= 0.466

Decision on sunny outlook factor

Features	Gini Index
temperature	0.2
humidity	0
wind	0.466

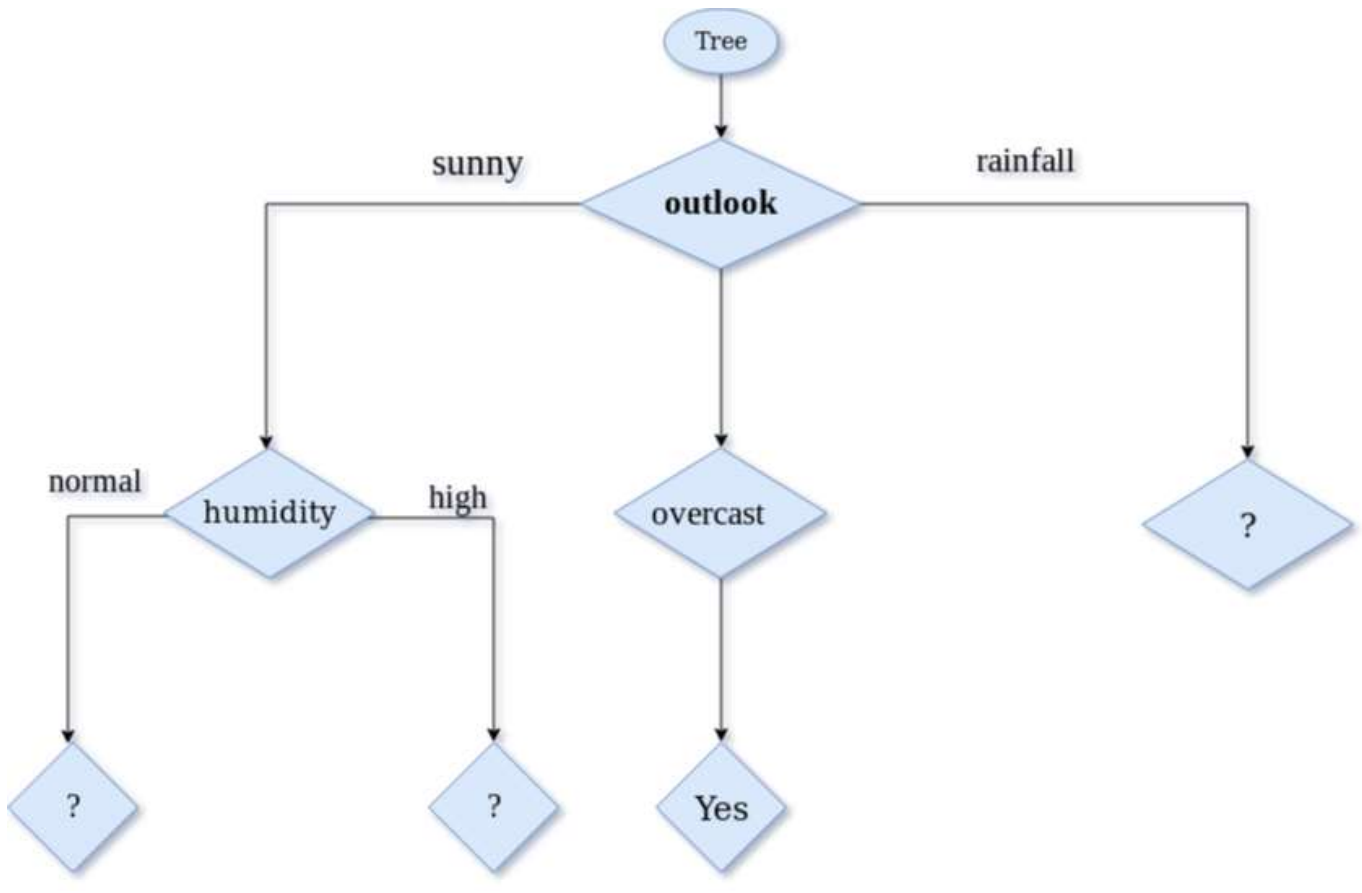
we have calculated the Gini index of all the features when the outlook is sunny. You can infer that humidity has lowest value. so next node will be humidity.



> *Now,Lets focus on sub data for overcast outlook feature.*

Day	outlook	temperature	humidity	wind	decision
3	overcast	hot	high	weak	Yes
7	overcast	cool	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes

As, you can see from the above table all the decision for overcast outlook feature is always ‘Yes’. Then Gini index for each feature is 0, means it is a leaf nodes.

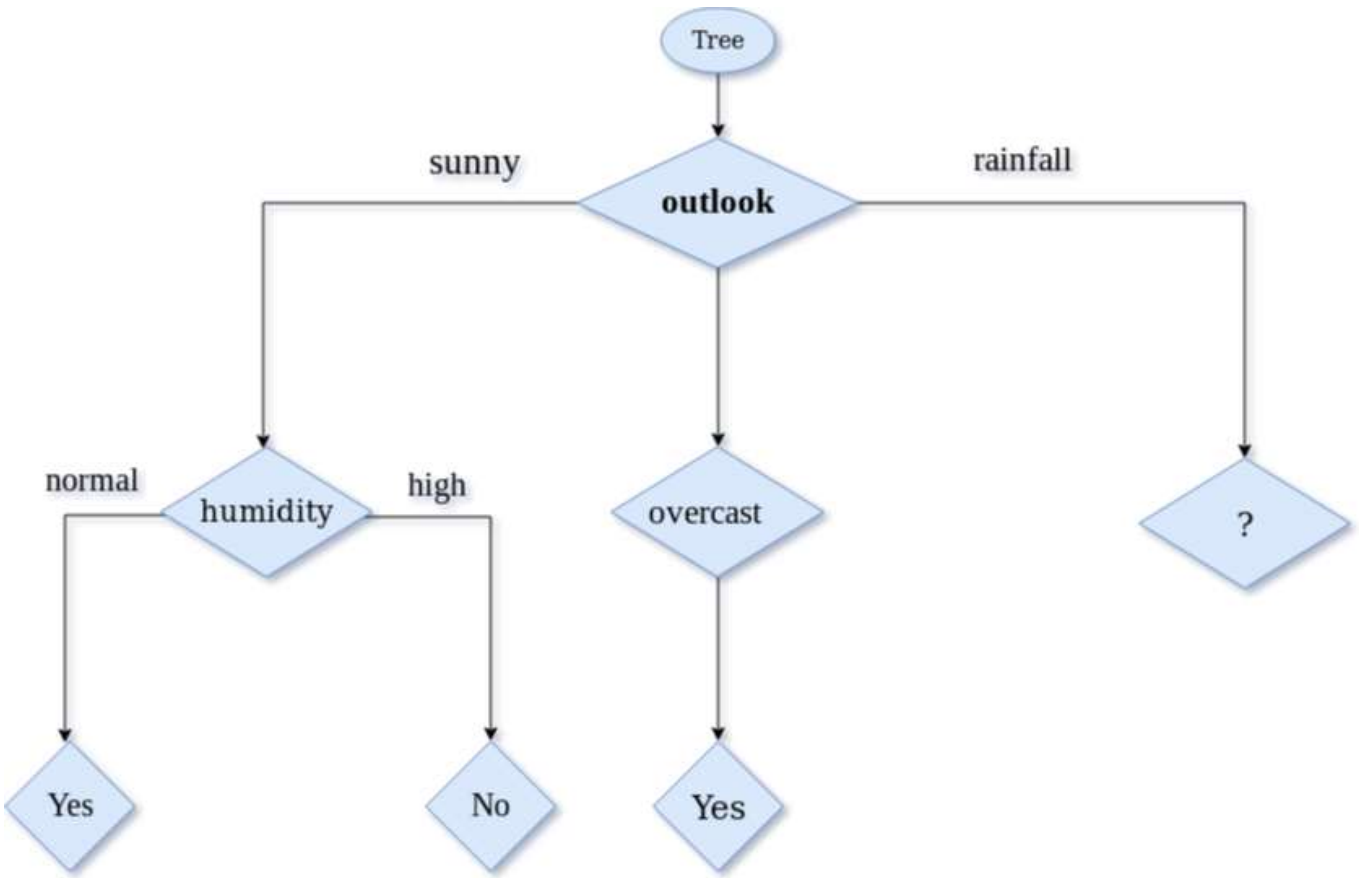


> *Now,Lets focus on sub data for high and normal humidity feature.*

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No

Day	outlook	temperature	humidity	wind	decision
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

From the given two table, the decision is always ‘No’ when humidity is ‘high’ and decision is always ‘Yes’ when humidity is ‘normal’. So we got leaf node. now decision tree can be viewed as,



> *Now,Lets focus on sub data for rainfall outlook feature. we need to find the Gini index for temperature,humidity and wind feature respectively.*

Day	outlook	temperature	humidity	wind	Decision
4	rain	mild	high	weak	Yes

5	rain	cool	normal	weak	Yes
6	rain	cool	normal	strong	No
10	rain	mild	normal	weak	Yes
14	rain	mild	high	strong	No

Gini index for temperature for rainfall outlook

temperature	Yes	No	# Instances
cool	1	1	2
mild	2	1	3

Gini(outlook=rainfall and temp.=Cool) = 1 — (1/2)2 — (1/2)2 = 0.5

Gini(outlook=rainfall and temp.=Mild) = 1 — (2/3)2 — (1/3)2 = 0.444

Gini(outlook=rainfall and temp.) = (2/5)*0.5 + (3/5)*0.444 = 0.466

Gini index for humidity for rainfall outlook

humidity	Yes	No	# Instances
high	1	1	2
normal	2	1	3

Gini(outlook=rainfall and humidity=high) = 1 — (1/2)2 — (1/2)2 = 0.5

Gini(outlook=rainfall and humidity=normal) = 1 — (2/3)2 — (1/3)2 = 0.444

Gini(Outlook=rainfall and humidity) = (2/5)*(0.5 + (3/5)*0.444 = 0.466

Gini index for wind for rainfall outlook feature

wind	Yes	No	# Instances

weak	3	0	3
strong	0	2	2

Gini(outlook=rainfall and wind=weak) = 1 — (3/3)2 — (0/3)2 = 0

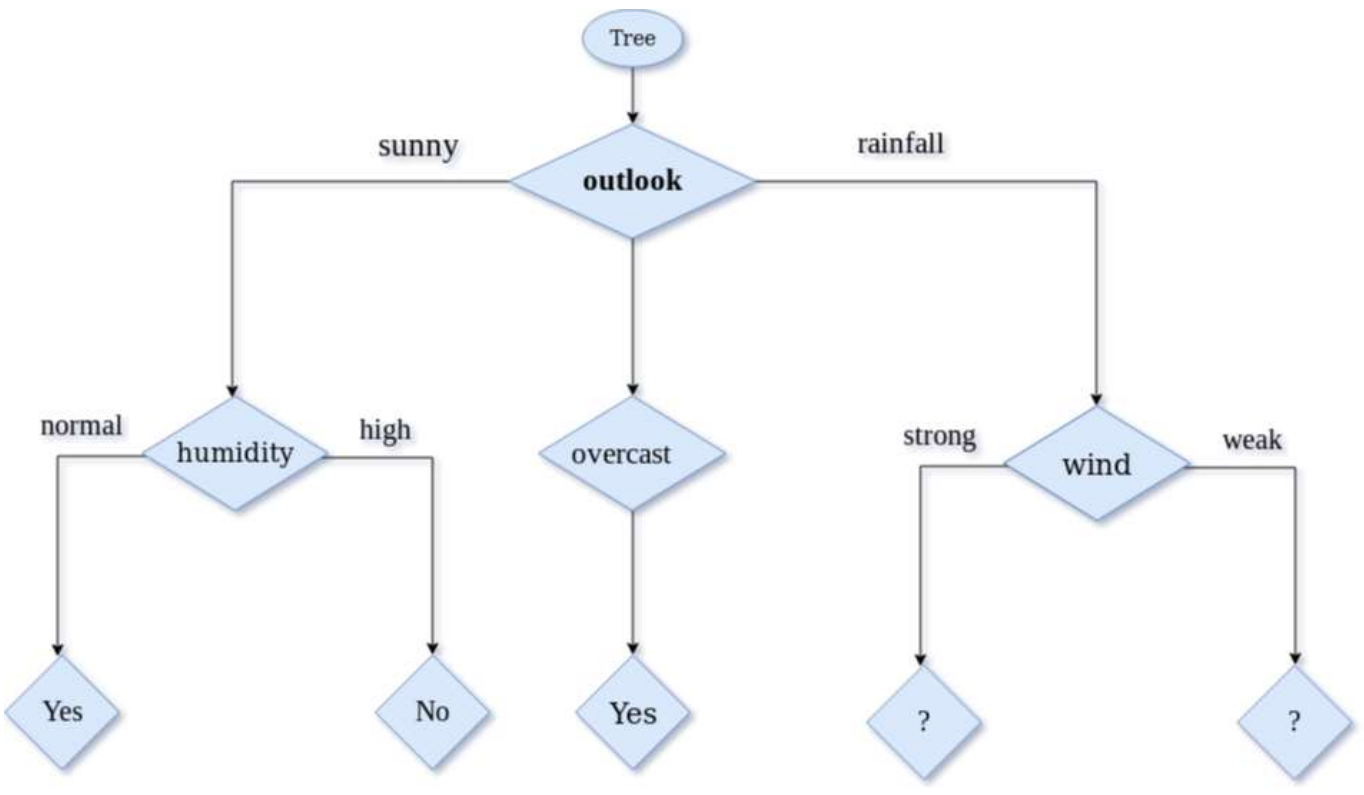
Gini(outlook=rainfall and wind=strong) = 1 — (0/2)2 — (2/2)2 = 0

Gini(outlook=rainfall and wind) = (3/5)*0 + (2/5)*0 = 0

Decision on rainfall outlook factor

Features	Gini Index
temperature	0.466
humidity	0.466
wind	0

we have calculated the Gini index of all the features when the outlook is rainfall. You can infer that wind has lowest value. so next node will be wind.



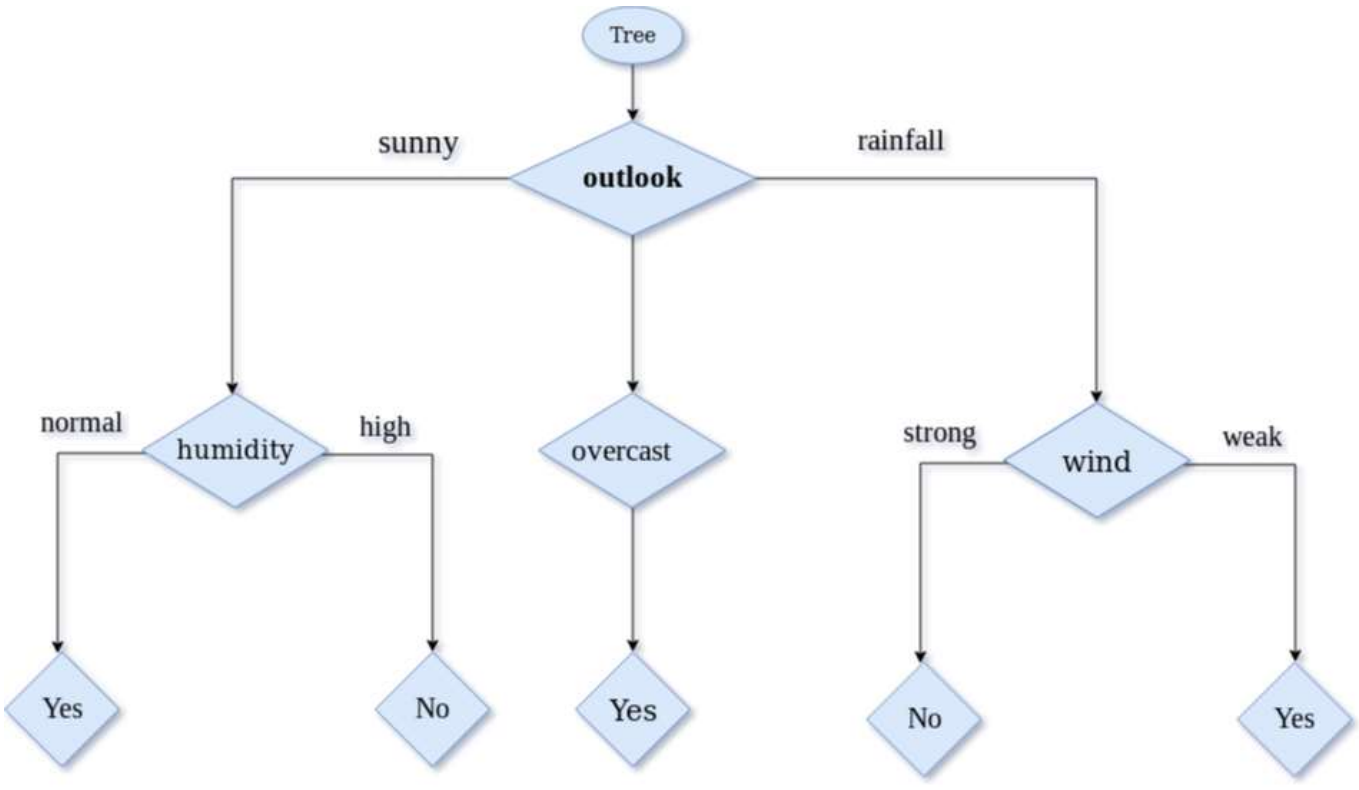
> *Now,Lets focus on sub data strong and weak for wind rainfall feature.*

Day	outlook	temperature	humidity	wind	decision
6	rainfall	cool	normal	strong	No

14	rainfall	mild	high	strong	No

Day	outlook	temperature	humidity	wind	decision
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes

From the above two table, the decision is always ‘No’ when wind is ‘strong’ and decision is always ‘Yes’ when wind is ‘weak’. So we got leaf node.



So, we have explained the generation of decision tree in step by step manner.

In the next tutorial we will generate tree with ID3 algorithm

That’s it. Thank you,

Machine Learning Decision Tree Decision Tree Classifier Data Science Gini Index