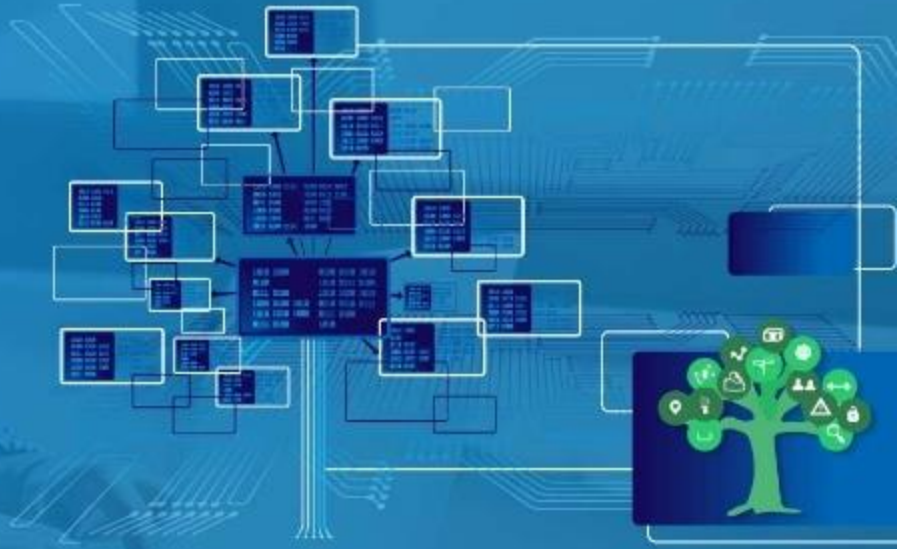
Decision Tree
Algorithm



edureka!

https://www.edureka.co/machine-learning-certification-training

Agenda for Today's Session

- What is Classification?
- Types of Classification
- Classification Use case
- What is Decision Tree?
- Terminologies associated to a Decision Tree
- Visualizing a Decision Tree
- Writing a Decision Tree Classifier form Scratch in Python using CART Algorithm


What is Classification?

edureka!

What is Classification?

"Classification is the process of dividing the data into different categories or groups by adding lab

- **Note:** It adds the data point to a particular labelled group on the basis of some condition"

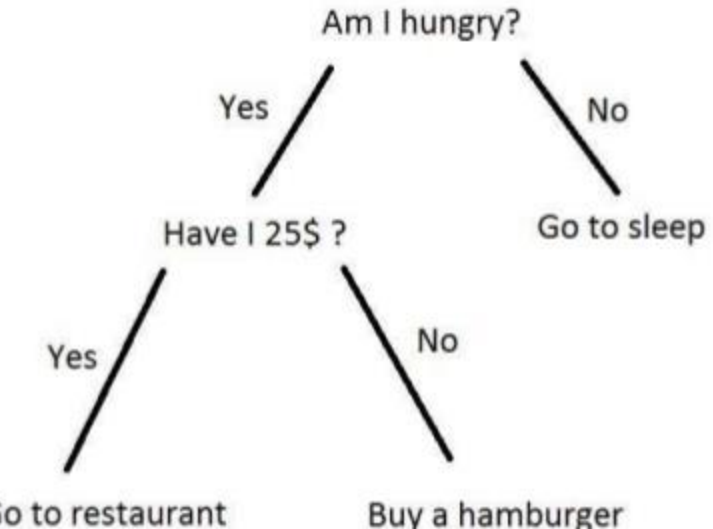
edureka!

Types of Classification

- Decision Tree
- Random Forest
- Naïve Bayes
- KNN

Decision Tree

- Graphical representation of all the possible solutions to a deci
- Decisions are based on some conditions
- Decision made can be easily explained



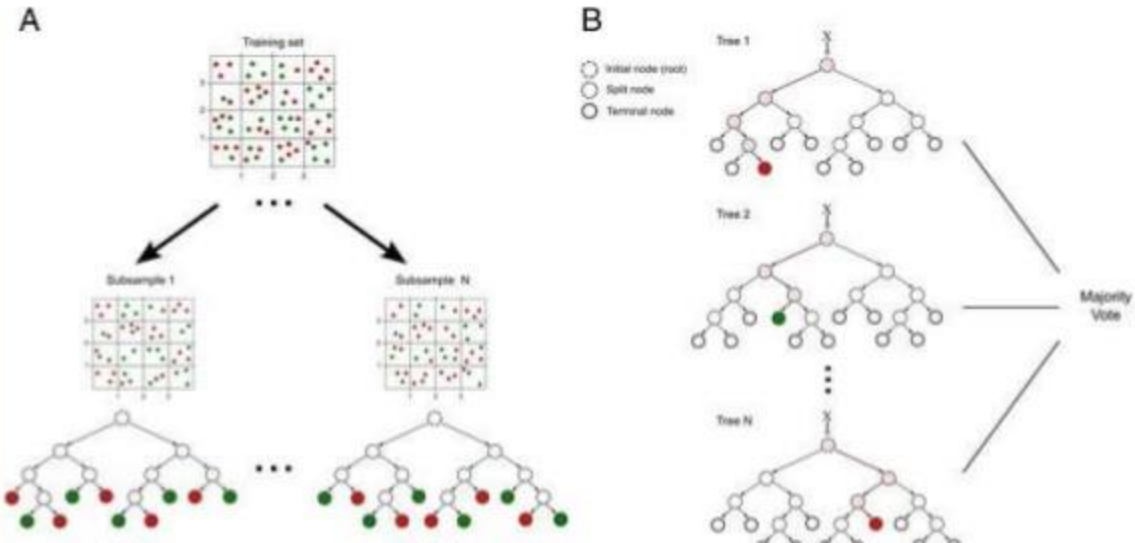
edureka!

Types of Classification

- Decision Tree
- Random Forest
- Naïve Bayes
- KNN

Random Forest

- Builds multiple decision trees and merges them together
- More accurate and stable prediction
- Random decision forests correct for decision trees' habit of overfitting to their training set
- Trained with the “bagging” method



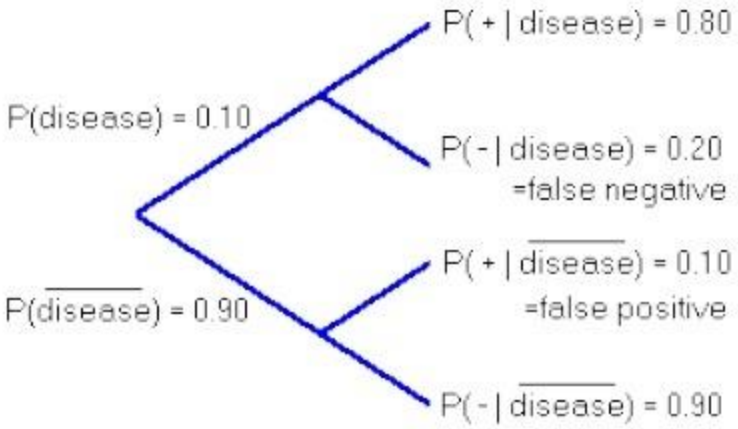
edureka!

Types of Classification

- Decision Tree
- Random Forest
- Naïve Bayes
- KNN

Naïve Bayes

- Classification technique based on Bayes' Theorem
- Assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature



edureka!

Types of Classification

- Decision Tree
- Random Forest
- Naïve Bayes
- KNN

K-Nearest Neighbors

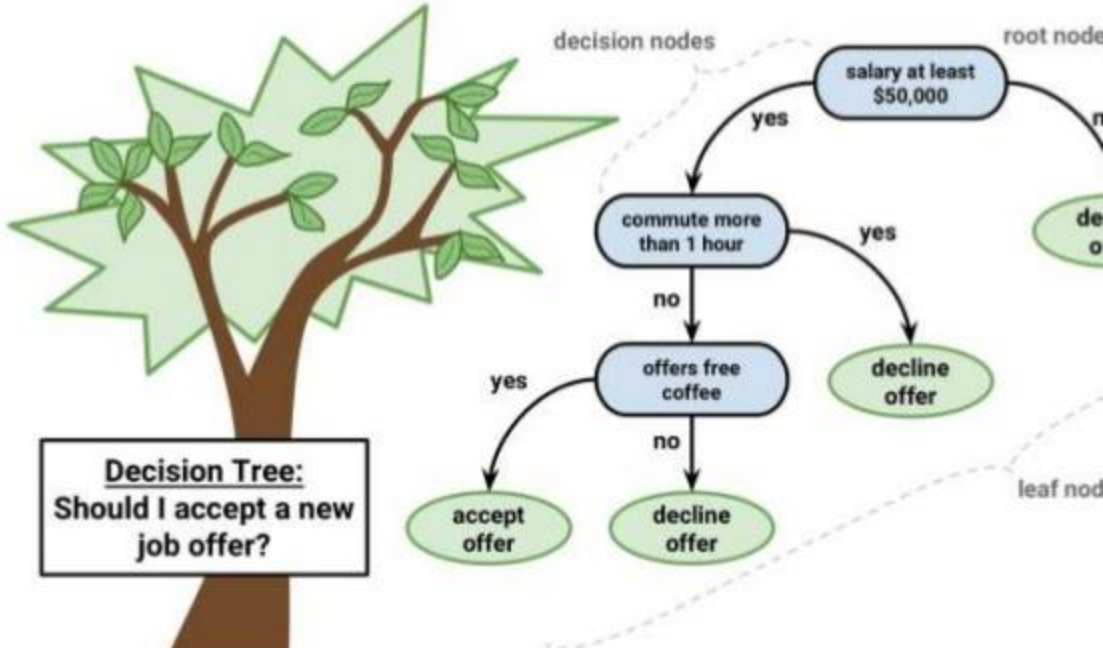
- Stores all the available cases and classifies new cases based on a similarity measure
- The "K" is KNN algorithm is the nearest neighbors we wi to take vote from.

What is Decision Tree?

edureka!

What is Decision Tree?

"A decision tree is a graphical representation of the possible solutions to a decision based on cert conditions"



Machine Leaning Training Using Python

Understanding a Decision Tree

edureka!

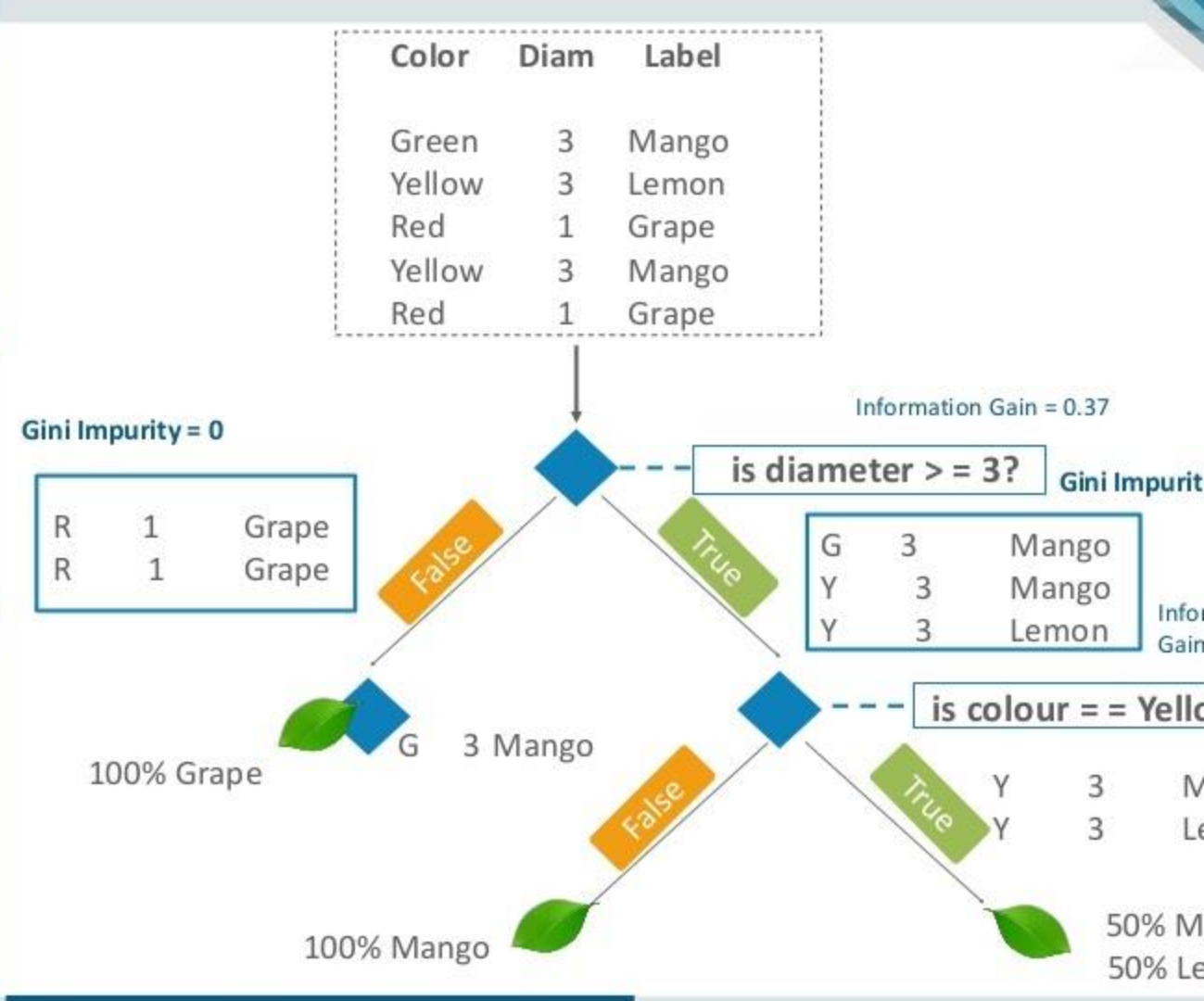
Dataset

This is how our dataset looks like!

Colour	Diameter	Label
Green	3	Mango
Yellow	3	Mango
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

edureka!

Decision Tree



edureka!

What is Decision Tree?

Green	3	Mango
Yellow	3	Lemon
Yellow	3	Mango

Is the colour green?

Is the diameter >=3

Is the colour yellow

TRUE

False

Machine Leaning Training Using Python

Decision Tree Terminologies

Decision Tree Terminology

Pruning

Opposite of Splitting, basically removing unwanted branches from the tree

Branch/SubTree

Formed by splitting the tree/node

Parent/Child Node

Root node is the parent node and all the other nodes branched from it is known as child node

Splitting

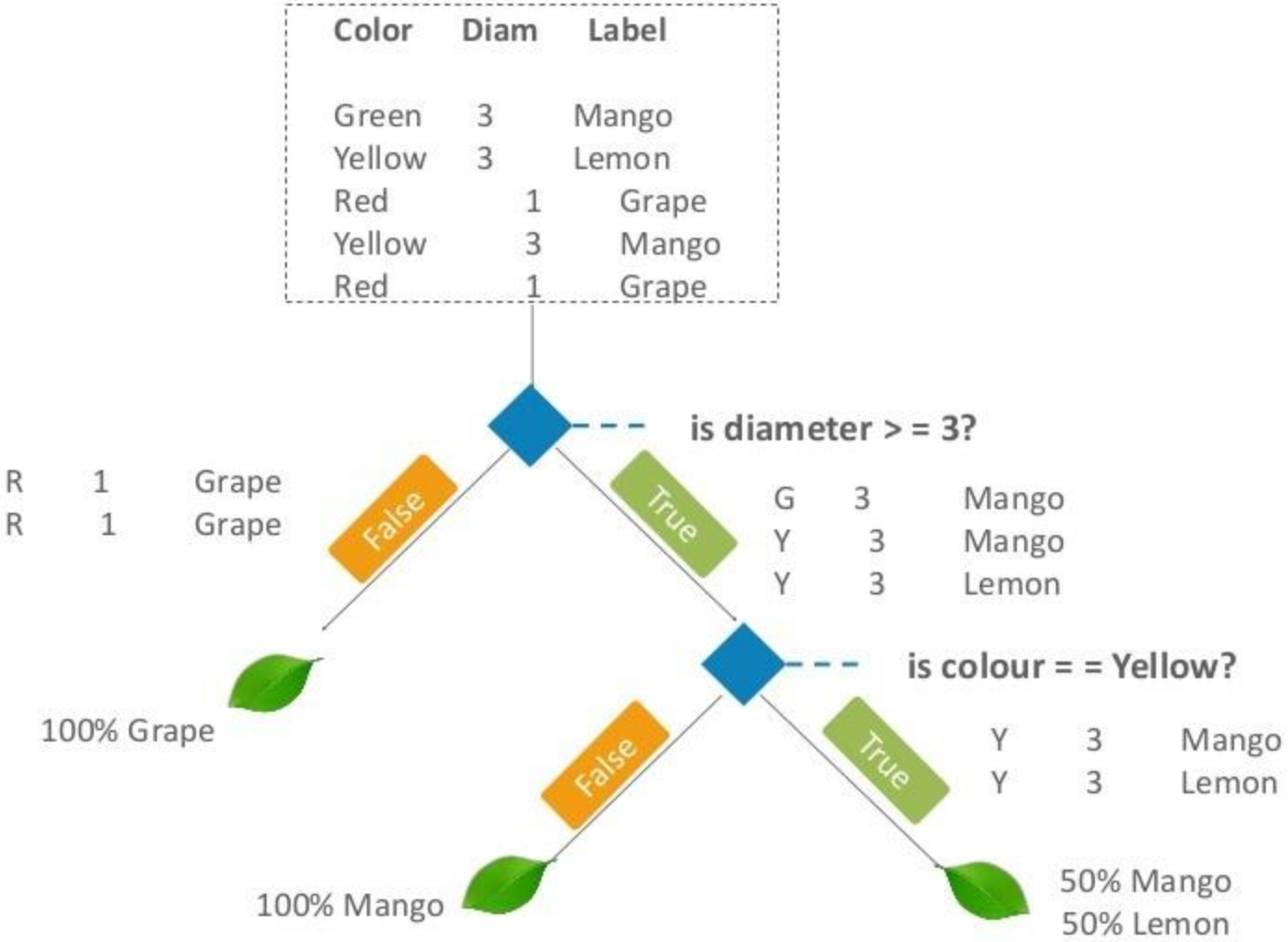
Splitting is dividing the root node/su node into different parts on the basi of some condition.

Root Node

It represents the entire population or sample and this further gets divided into two or more homogenous sets.

Leaf Node

Node cannot be further segregated into further nodes


Color Diam Label
Green 3 Mango
Yellow 3 Lemon
Red 1 Grape
Yellow 3 Mango
Red 1 Grape
is diameter > = 3?
False
True
R 1 Grape
R 1 Grape
G 3 Mango
Y 3 Mango
Y 3 Lemon
100% Grape
is colour = = Yellow?
False
True
Y 3 Mango
Y 3 Lemon
100% Mango
50% Mango
50% Lemon

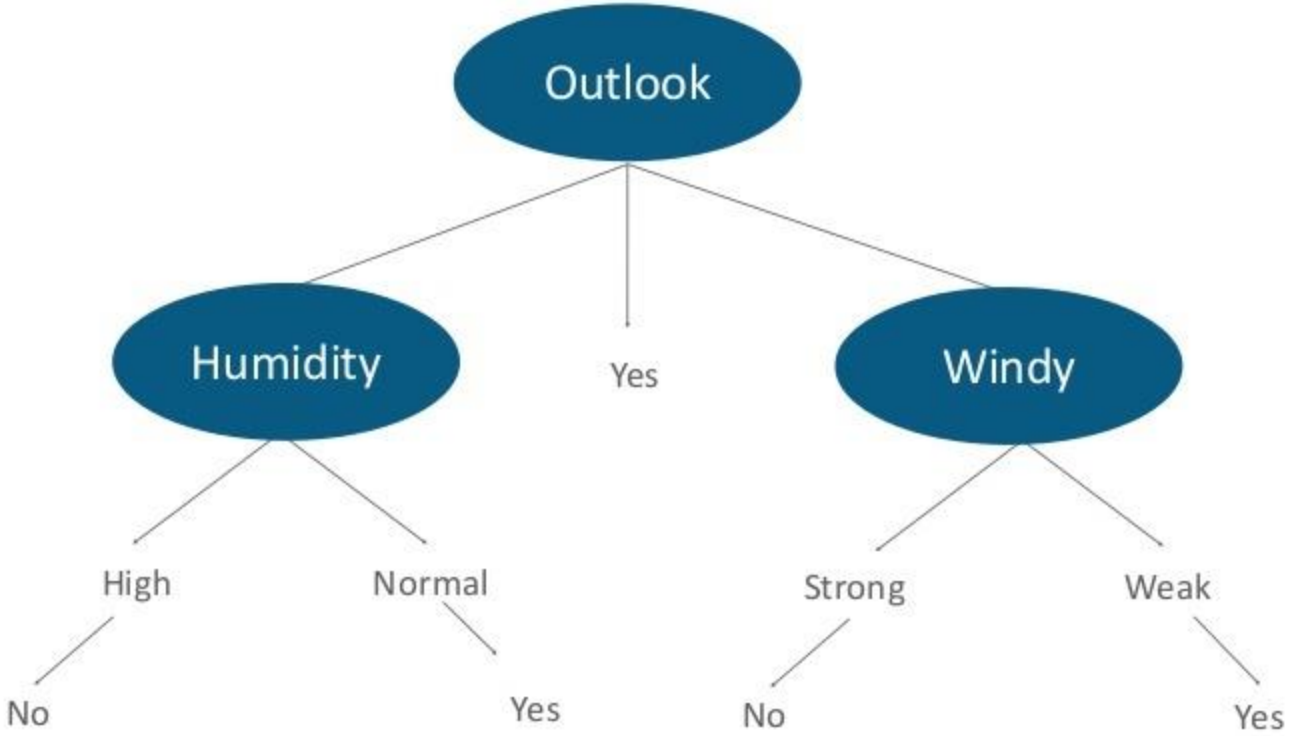
Machine Leaning Training Using Python

CART Algorithm

Let's First Visualize the Decision Tree

Which Question to ask and When?

Let's First Visualize the Decision Tree



Learn about Decision Tree

Which one among them should you pick first?

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Learn about Decision Tree

Answer: Determine the attribute that best classifies the training data

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Learn about Decision Tree

But How do we choose the best attribute?

Or

How does a tree decide where to split?

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

How Does A Tree Decide Where To Split?

Gini Index

The measure of impurity (or purity) used in building decision tree in CART is Gini Index

Chi Square

It is an algorithm to find out the statistical significance between the differences between sub-nodes and parent node

Information Gain

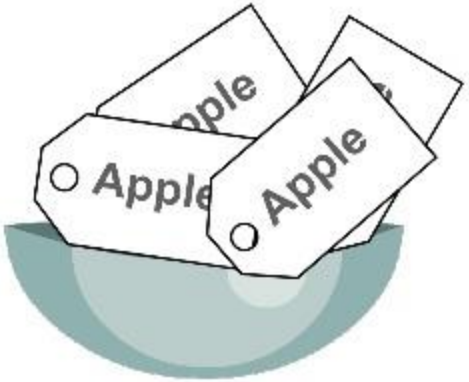
The information gain is the decrease in entropy after a dataset is split on the basis of an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain

Reduction in Variance

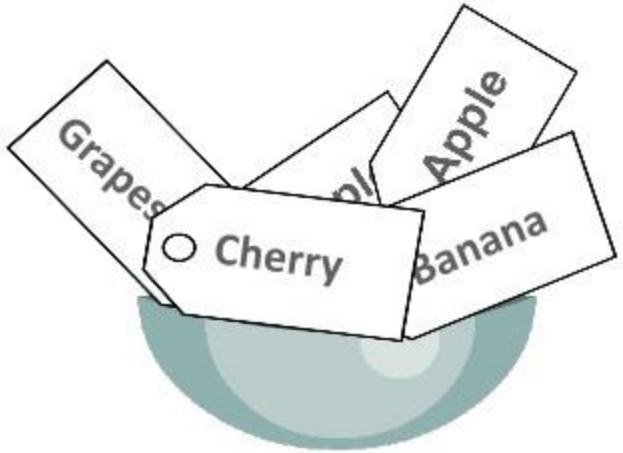
Reduction in variance is an algorithm used for continuous target variables (regression problems). The split with lower variance is selected as the criteria to split the population

Let's First Understand What is Impurity

Impurity = 0



Let's First Understand What is Impurity

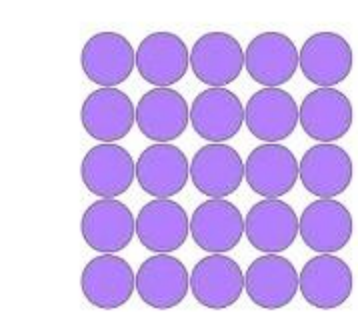


Impurity $\neq 0$

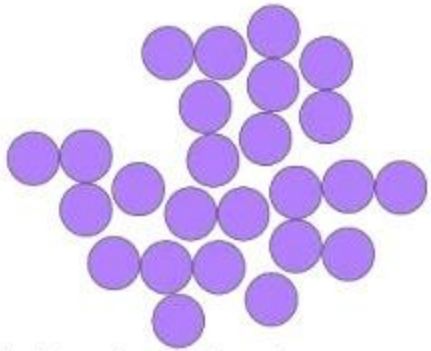
edureka!

What is Entropy?

- Defines randomness in the data
- **Entropy** is just a metric which measures the impurity or
- The first step to solve the problem of a decision tree

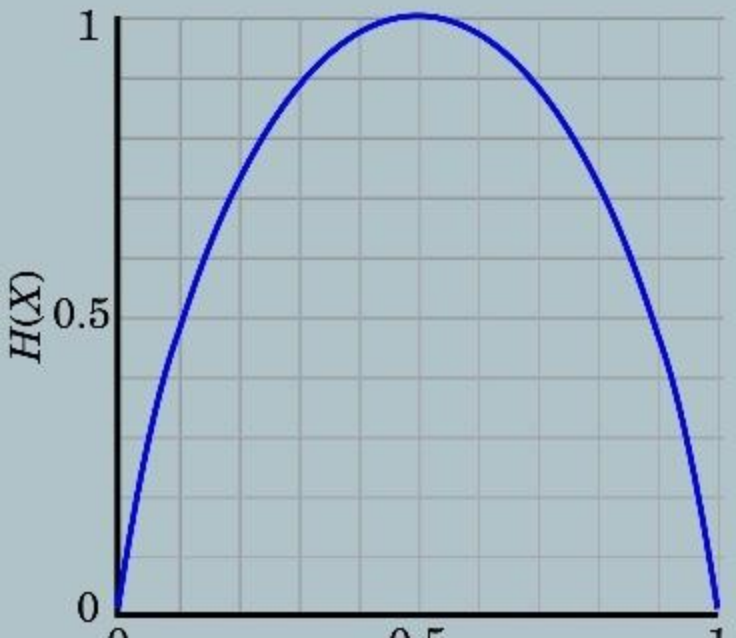


Low Entropy



High Entropy

What is Entropy?



Entropy(s) =- P(yes) $\log_2$ P(yes) – P(no) $\log_2$ P(no)

Where,

- S is the total sample space,
- P(yes) is probability of yes

If number of *yes* = number of *no ie P(S) = 0.5*

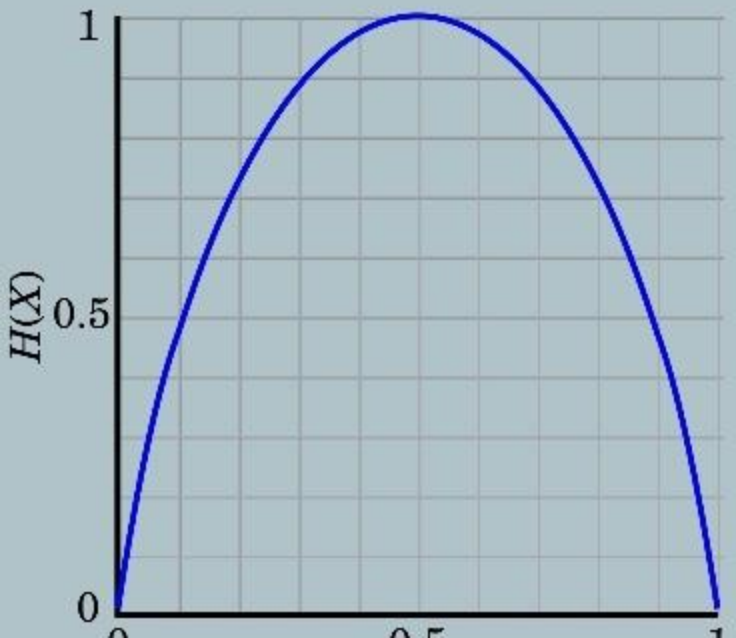
⇒ *Entropy(s) = 1*

If it contains all yes or all no ie P(S) = 1 or 0

⇒ *Entropy(s) = 0*

edureka!

What is Entropy?



$E(S) = -P(Yes) \log_2 P(Yes)$

When P(Yes) =P(No) = 0.5 ie YES + NO = Total Sample(S)

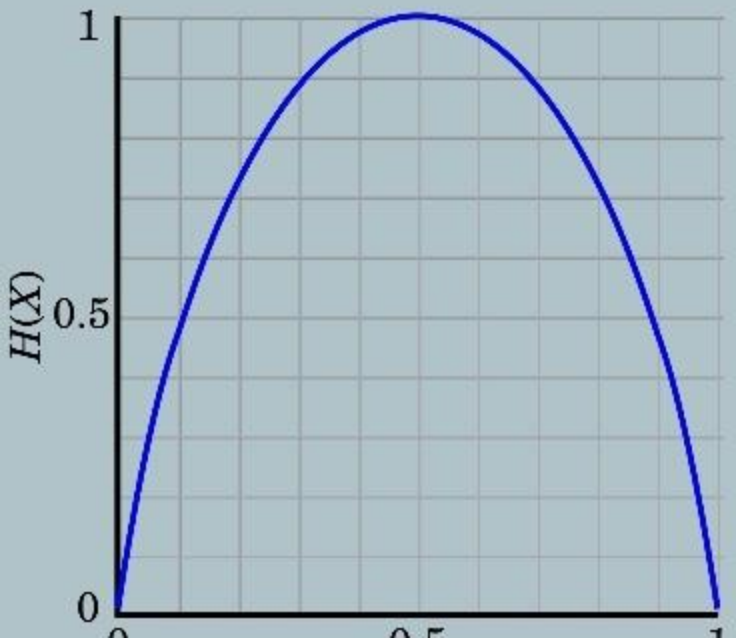
$E(S) = 0.5 \log_2 0.5 - 0.5 \log_2 0.5$

$E(S) = 0.5(\log_2 0.5 - \log_2 0.5)$

$E(S) = 1$

edureka!

What is Entropy?



$E(S) = -P(Yes) \log_2 P(Yes)$

When P(Yes) = 1 ie YES = Total Sample(S)

$E(S) = 1 \log_2 1$

$E(S) = 0$

$E(S) = -P(No) \log_2 P(No)$

When P(No) = 1 ie No = Total Sample(S)

$E(S) = 1 \log_2 1$

$E(S) = 0$

edureka!

What is Information Gain?

- Measures the reduction in entropy
- Decides which attribute should be selected as the decision node

If S is our total collection,

Information Gain = Entropy(S) – [(Weighted Avg) x Entropy(each feature

Let's Build Our Decision Tree

Step 1: Compute the entropy for the Data set

Out of 14 instances we have 9 YES and 5 NO

So we have the formula,

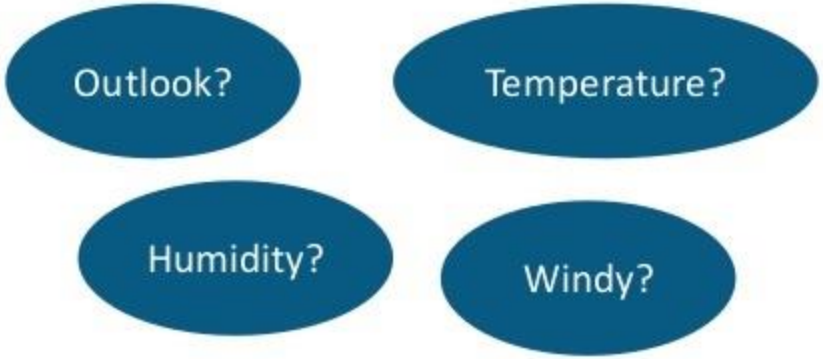
$E(S) = -P(Yes) \log_2 P(Yes) - P(No) \log_2 P(No)$

$E(S) = -(9/14)* \log_2 9/14 - (5/14)* \log_2 5/14$

$E(S) = 0.41+0.53 = 0.94$

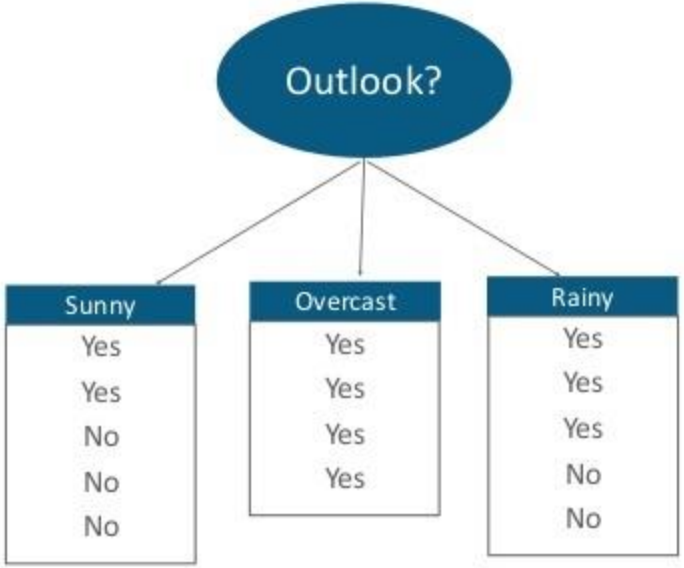
	outlook	temp.	humidity	windy	play
D1	sunny	hot	high	false	no
D2	sunny	hot	high	true	no
D3	overcast	hot	high	false	yes
D4	rainy	mild	high	false	yes
D5	rainy	cool	normal	false	yes
D6	rainy	cool	normal	true	no
D7	overcast	cool	normal	true	yes
D8	sunny	mild	high	false	no
D9	sunny	cool	normal	false	yes
D10	rainy	mild	normal	false	yes
D11	sunny	mild	normal	true	yes
D12	overcast	mild	high	true	yes
D13	overcast	hot	normal	false	yes
D14	rainy	mild	high	true	no

Which Node To Select As Root Node?



outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Which Node To Select As Root Node: Outlook



outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Which Node To Select As Root Node: Outlook

E(Outlook = Sunny) = -2/5 $\log_2 2/5 - 3/5 \log_2 3/5$ *= 0.971*

E(Outlook = Overcast) = -1 $\log_2 1 - 0 \log_2 0$ *= 0*

E(Outlook = Sunny) = -3/5 $\log_2 3/5 - 2/5 \log_2 2/5$ *= 0.971*

Information from outlook,

I(Outlook) = 5/14 x 0.971 + 4/14 x 0 + 5/14 x 0.971 = 0.693

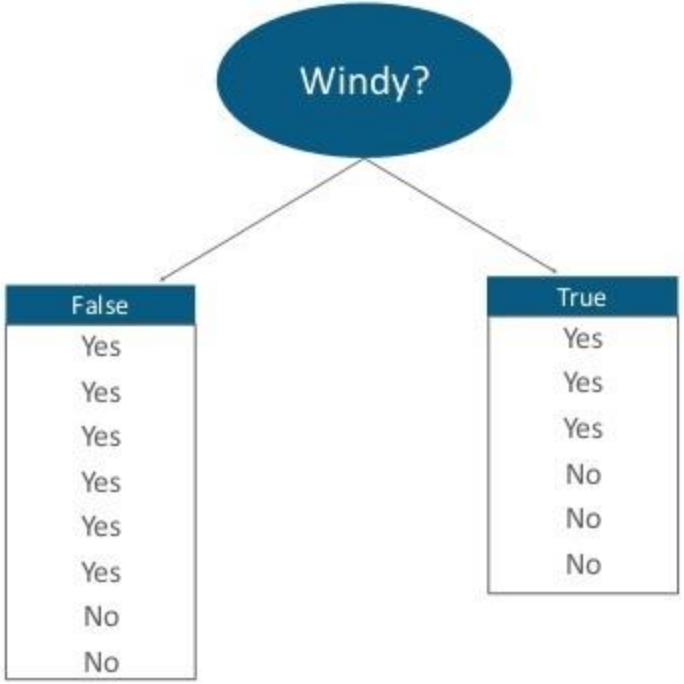
Information gained from outlook,

Gain(Outlook) = E(S) – I(Outlook)

0.94 – 0.693 = 0.247

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Which Node To Select As Root Node: Outlook



outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Which Node To Select As Root Node: Windy

E(Windy = True) = 1

E(Windy = False) = 0.811

Information from windy,

I(Windy) = 8/14 x 0.811+ 6/14 x 1 = 0.892

Information gained from outlook,

Gain(Windy) = E(S) – I(Windy)

0.94 – 0.892 = 0.048

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Similarly We Calculated For Rest Two

Which Node To Select As Root Node

Outlook:
Info 0.693
Gain: 0.940-0.693 0.247

Temperature:
Info 0.911
Gain: 0.940-0.911 0.029

Humidity:
Info 0.788
Gain: 0.940-0.788 0.152

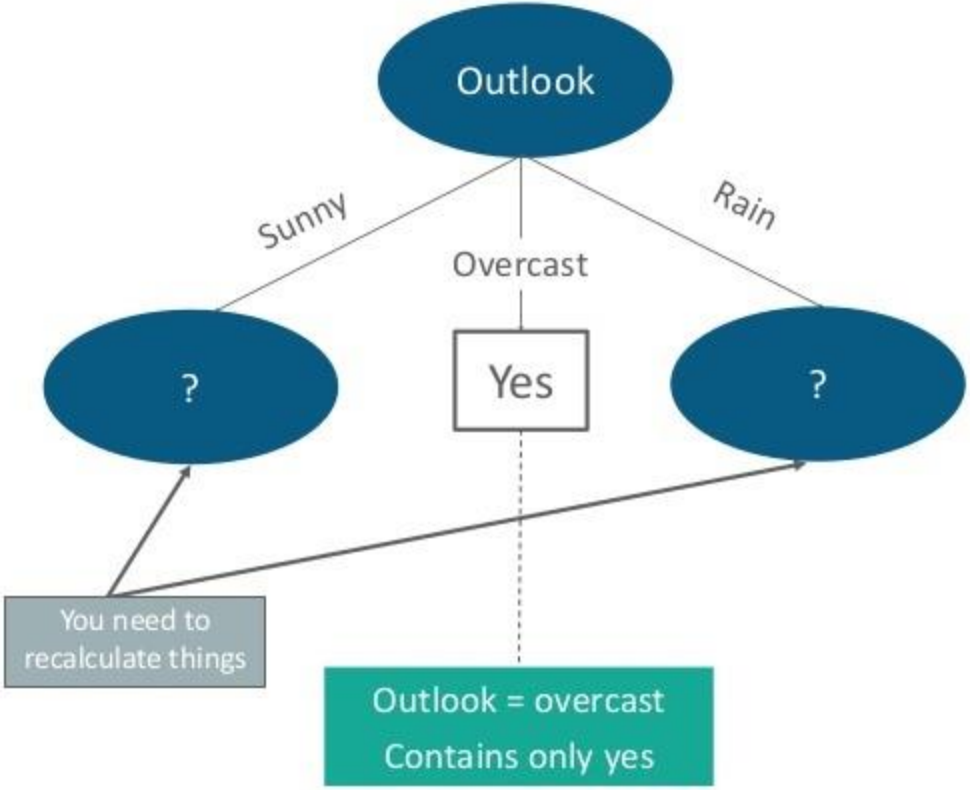
Windy:
Info 0.892
Gain: 0.940-0.982 0.048

Since Max gain = 0.247,

Outlook is our ROOT Node

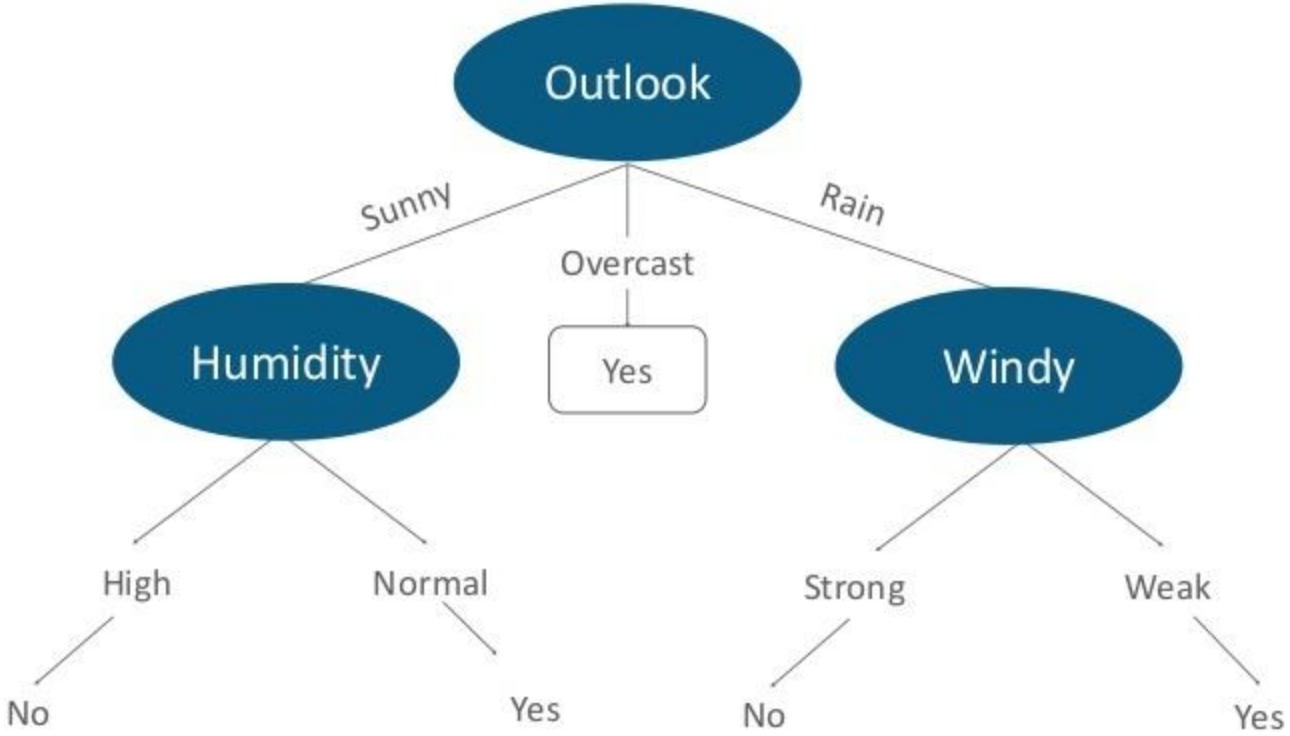
outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Which Node To Select Further?



outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

This Is How Your Complete Tree Will Look Like



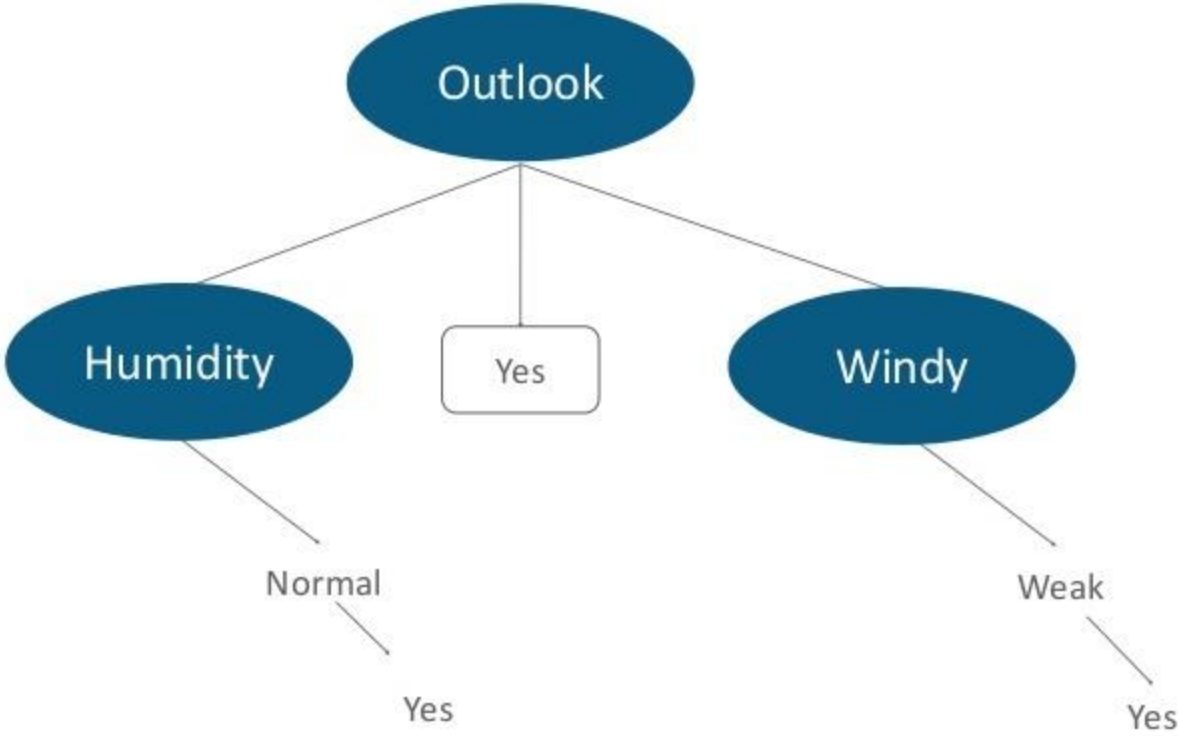
What Should I Do To Play - Pruning

edureka!

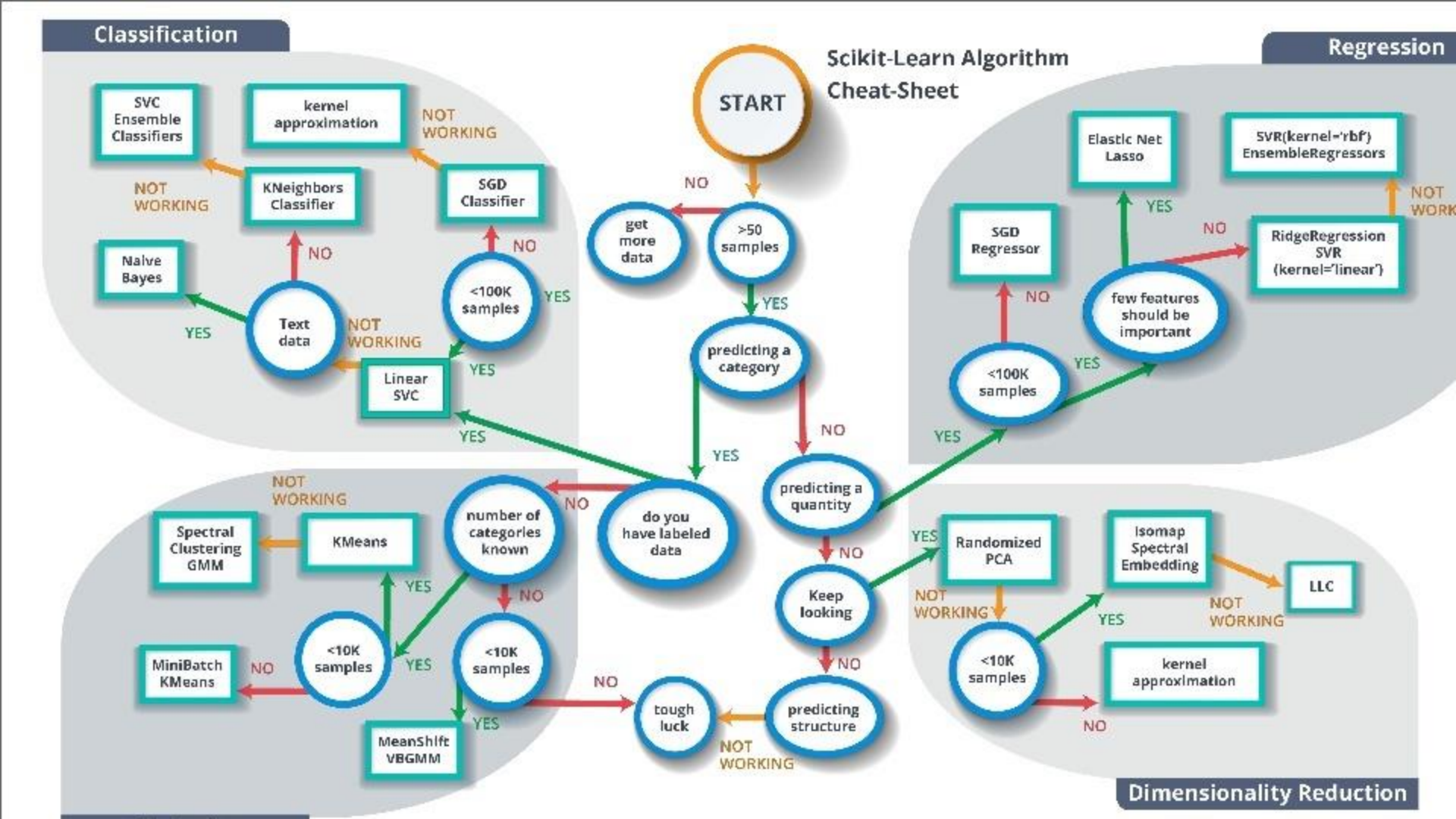
What is Pruning?

"A decision tree is a graphical representation of the possible solutions to a decision based on cert conditions"

Pruning: Reducing The Complexity



Are tree based models better than linear models?


Scikit-Learn Algorithm Cheat-Sheet
START
Classification
Regression
Dimensionality Reduction
NO
get more data
>50 samples
YES
predicting a category
predicting a quantity
do you have labeled data
NO
YES
NO
YES
SVC Ensemble Classifiers
kernel approximation
NOT WORKING
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
NO
Naive Bayes
Text data
<100K samples
YES
YES
NOT WORKING
YES
Linear SVC
YES
Elastic Net Lasso
SVR(kernel='rbf') EnsembleRegressors
YES
NOT WORK
SGD Regressor
NO
RidgeRegression SVR (kernel='linear')
NO
few features should be important
<100K samples
YES
YES
NOT WORKING
Spectral Clustering GMM
KMeans
number of categories known
NO
YES
YES
NO
MiniBatch KMeans
NO
<10K samples
YES
<10K samples
NO
MeanShift VBGMM
YES
tough luck
NOT WORKING
predicting structure
NO
Keep looking
NO
YES
Randomized PCA
Isomap Spectral Embedding
LLC
NOT WORKING
NOT WORKING
YES
<10K samples
kernel approximation
NO

edureka!