Classification I

DigitalSkola
Uncover The World Of Digital Skills With Us

Profile

linkedin.com/in/romansyasetyo/

DigitalSkola
Uncover The World Of Digital Skills With Us

Table of Content

What will We Learn Today?

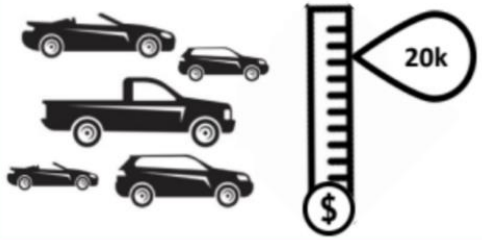
1. What is Classification
2. Examples of Classification Projects
3. K-Nearest Neighbors
4. Decision Trees
5. Random Forest
6. Evaluation Metrics

Hands on using Python

DigitalSkola
Uncover The World Of Digital Skills With Us

REGRESSION

PREDICT VALUE



CLASSIFICATION

PREDICT CLASS

CLUSTERING

GROUP OBJECT

DigitalSkola
Uncover The World Of Digital Skills With Us

What is Classification?

Classification Model

A model that will predict the class **labels/categories** for the new data

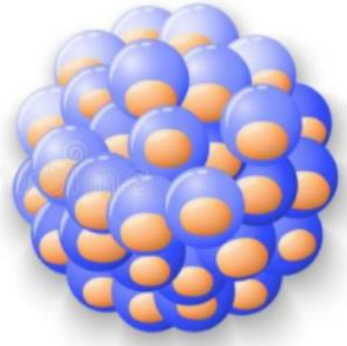
DigitalSkola
Uncover The World Of Digital Skills With Us

Examples of Classification Projects

DigitalSkola
Uncover The World Of Digital Skills With Us

Input	Machine Learning Model	Output
Banyaknya Roda (2,3,4,6,8)	**if banyaknya roda >2 :** Mobil **else:** Motor	Mobil atau Motor

Fraud Detection

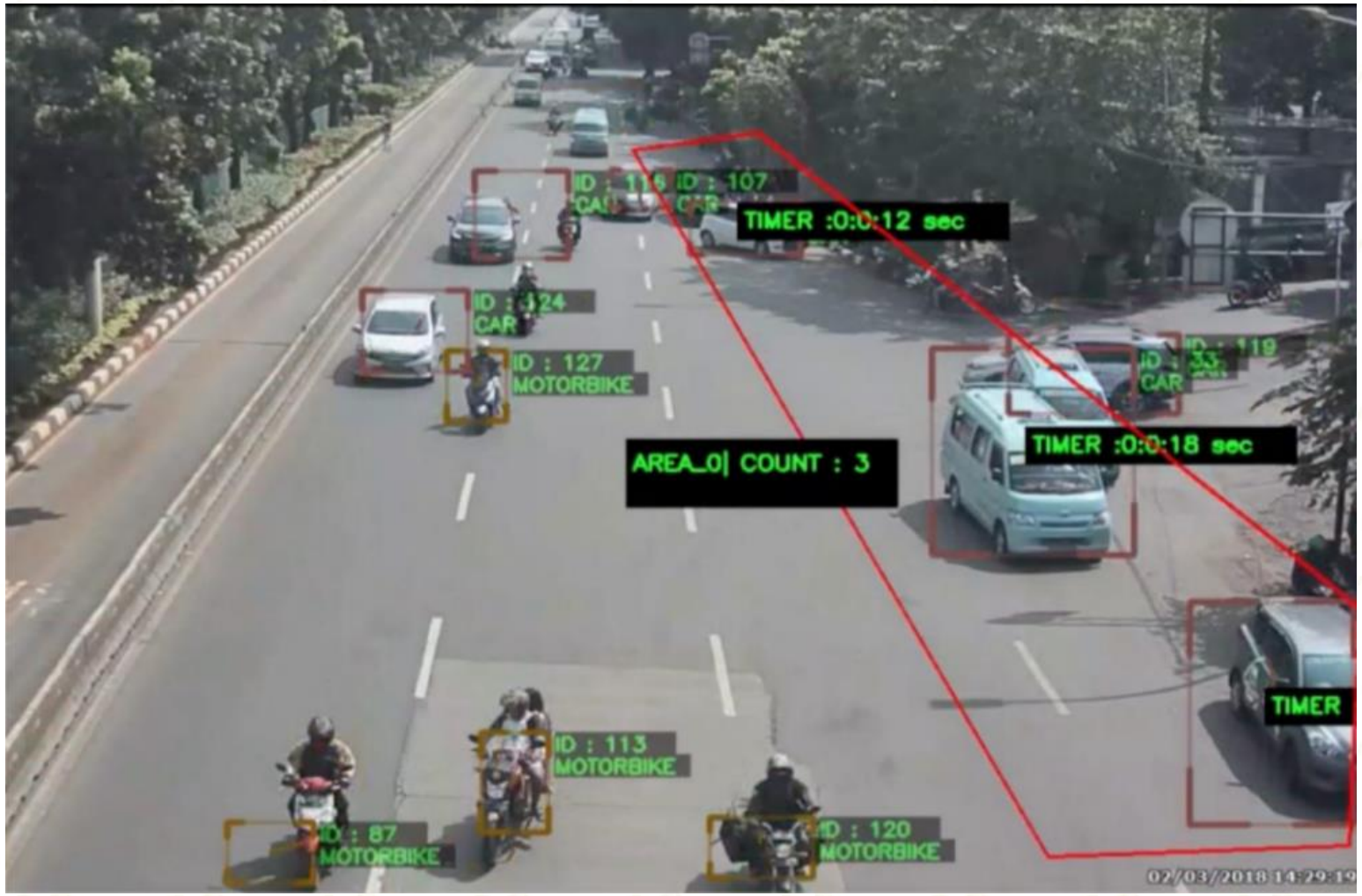


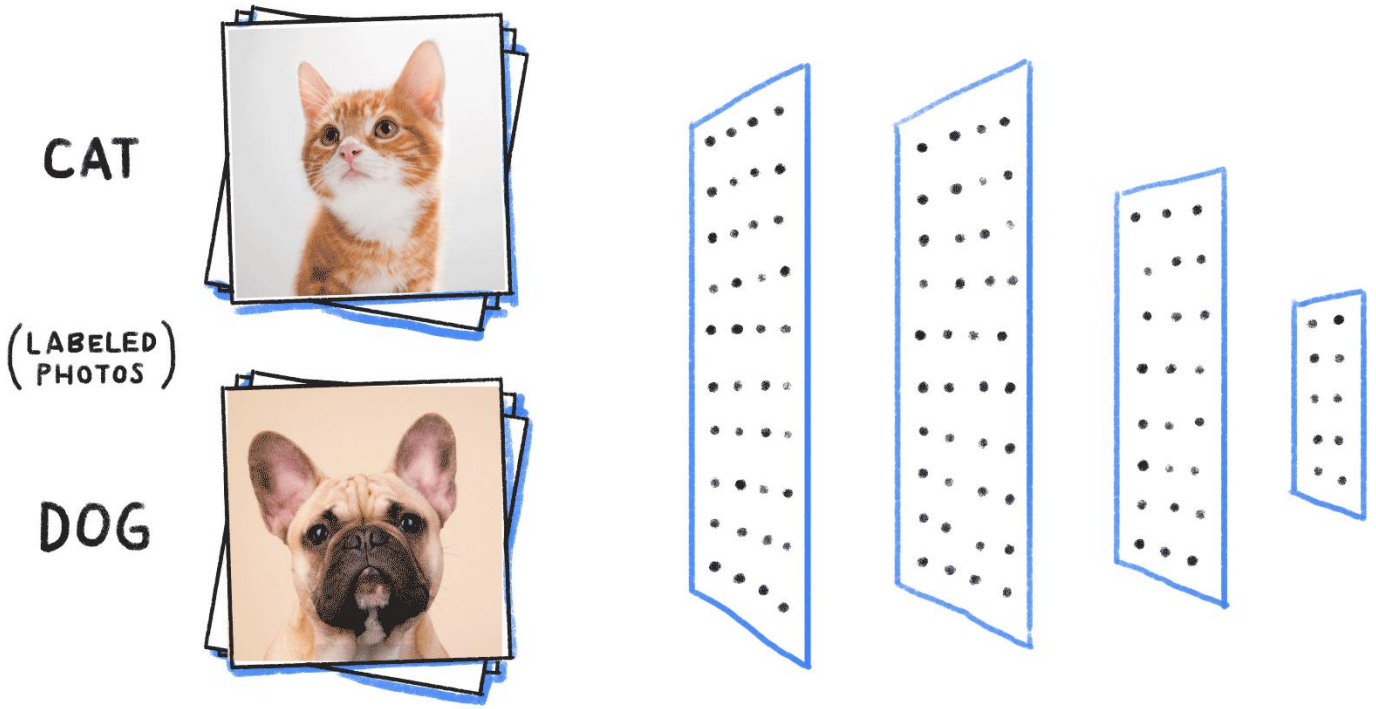
Cancer Cell Classification

Credit Scoring

Spam Detector

Churn Analysis



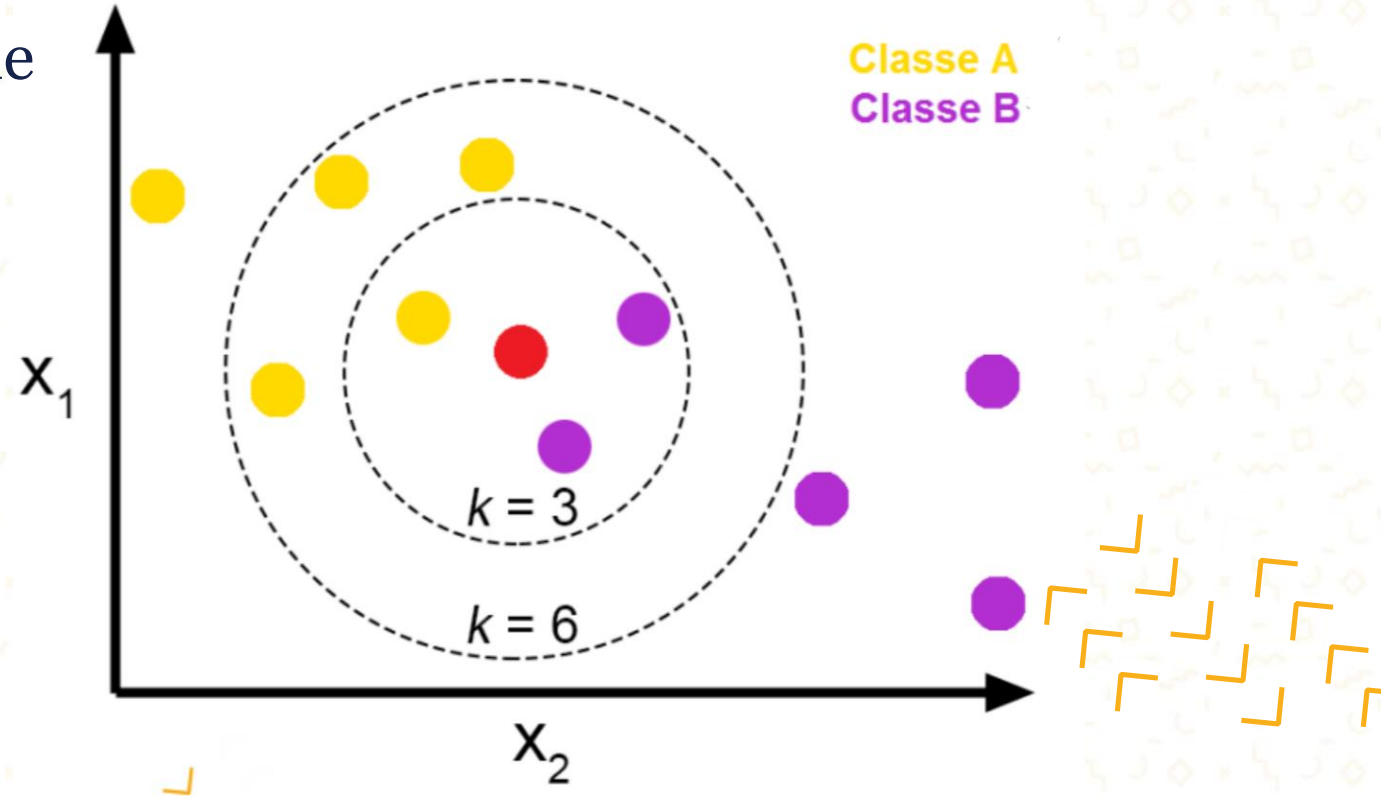


DigitalSkola
Uncover The World Of Digital Skills With Us

Classification : KNN

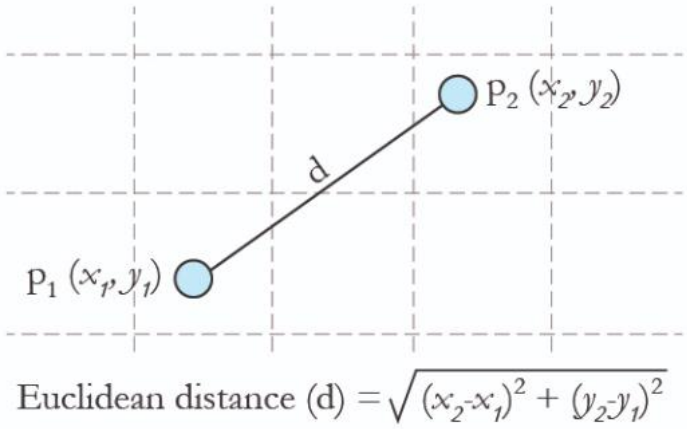
K-Nearest Neighbors

- K-NN is an algorithm that define class of an entity based on the closest neighbors
- It uses distance algorithm to define the class



Distance

- Closest neighbor is identified based on the distance
- There are several method to measure distance. The popular one is Euclidean.



Illustration

(x1, y1) = (1, 2)

(x2, y2) = (5, 4)

Euclidean distance

$= \sqrt{(5-1)^2 + (4-2)^2}$

$= \sqrt{(4)^2 + (2)^2}$

$= \sqrt{20}$

$= 4.47$

Issue with Distance

- Given X is Area with unit of hectare
- Given Y is Corn Production with unit of kg.

Illustration

(x1, y1) = (3, 2000)

(x2, y2) = (5, 4000)

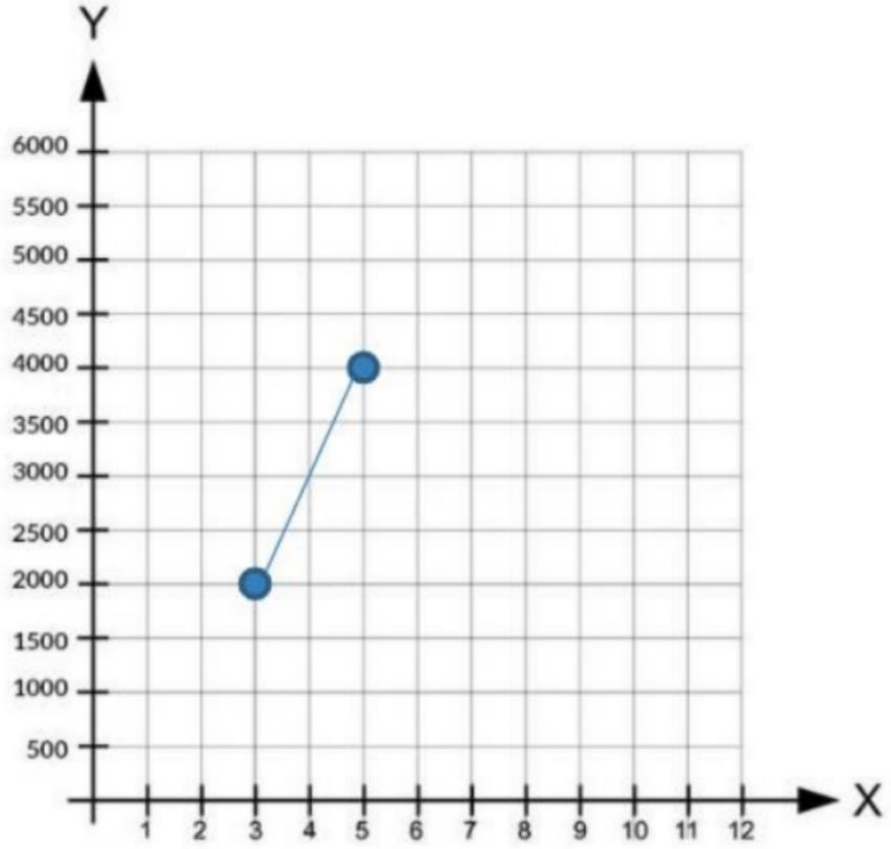
Euclidean distance

$$= \sqrt{(5-3)^2 + (4000-2000)^2}$$

$$= \sqrt{(2)^2 + (2000)^2}$$

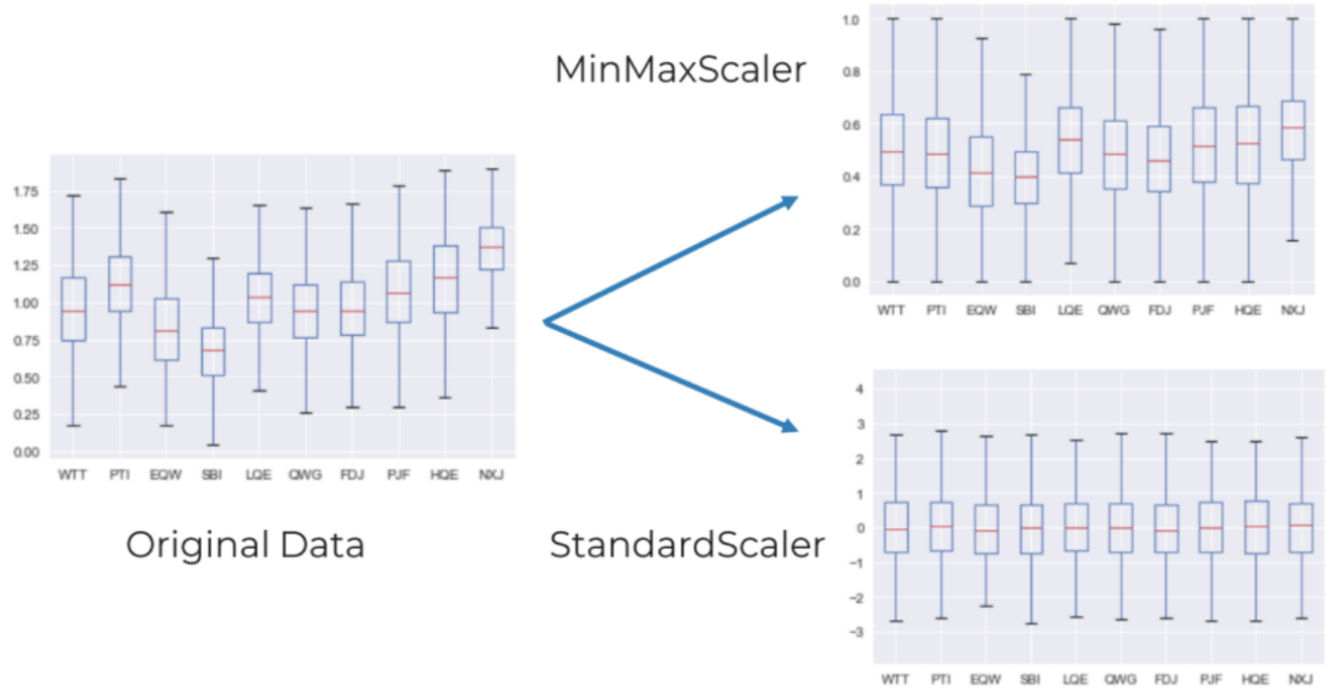
$$= \sqrt{4\,000\,004}$$

$$= 2000$$



Issue with Distance : Solution

Scaling / Normalization



Simple Illustration

Misalnya ada sebuah rumah yang berada tepat di tengah perbatasan antara Kota Bandung dan Kabupaten Bandung, sehingga pemerintah kesulitan untuk menentukan apakah rumah tersebut termasuk kedalam wilayah Kota Bandung atau Kabupaten Bandung.

Rumah	Lat	Long	Lokasi
A	11	26	Kota
B	15	29	Kota
C	19	28	Kota
D	18	30	Kota
E	16	26	Kota
F	23	25	Kabupaten
G	25	22	Kabupaten
H	21	24	Kabupaten
I	23	25	Kabupaten
J	29	24	Kabupaten
X	19	25	?

Simple Illustration

Step 1 : Pilih jumlah K

Step 2 : Hitung jarak dan urutkan dari jarak terkecil

Step 3 : Dari K neighbor yang paling dekat, terlihat bahwa 1 rumah berada di Kabupaten sedangkan 2 rumah berada di Kota

Step 4 : Kesimpulannya, rumah X berada di Kota

X	19	25

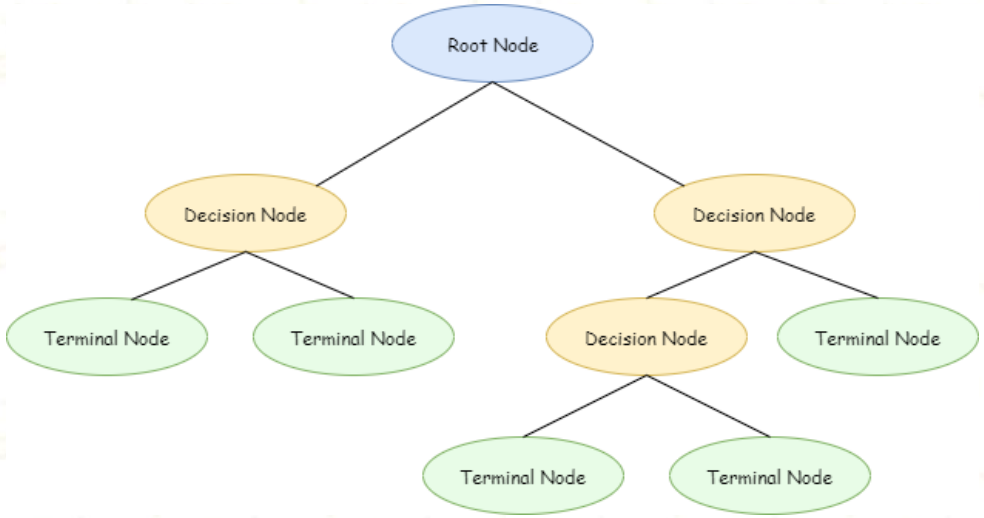
Rumah	Lat	Long	Jarak Terhadap Rumah X
H	21	24	2.24
C	19	28	3.00
E	16	26	3.16
F	23	25	4.00
I	23	25	4.00
D	18	30	5.10
B	15	29	5.66
G	25	22	6.71
A	11	26	8.06
J	29	24	10.05

DigitalSkola
Uncover The World Of Digital Skills With Us

Classification : Decision Tree

Decision Tree

- Decision tree learning is a method commonly used in machine learning.
- Basically the algorithm divide a condition into two choices. And it happens until can't be divided.



Contoh Kasus Pola Prediksi: Decision Tree

DigitalSkola
Uncover The World Of Digital Skills With Us

STEP 1

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

Gender	Age	App
F	15	
F	25	
M	32	
F	40	
M	12	
M	14	

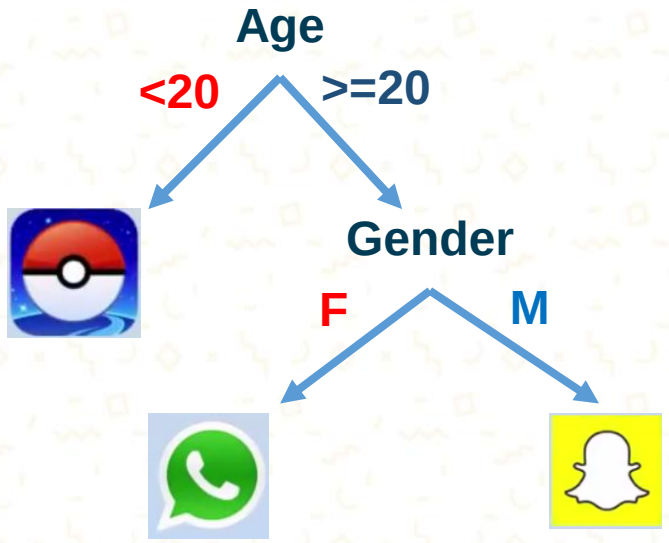
STEP 2

Gender	Age	App
F	25	
M	32	
F	40	

Test Subject

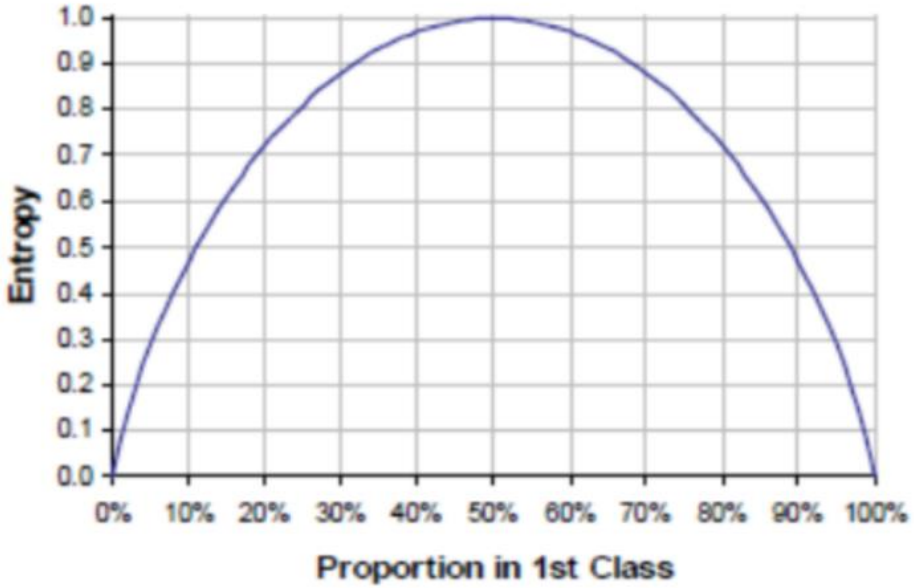
Antonio

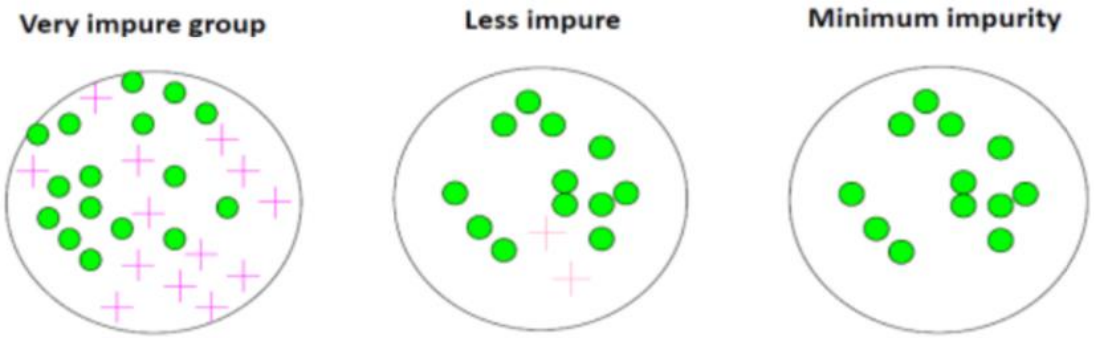
Male
20 years old



Decision Tree : Entropy

- Entropy is measure of Heterogeneity/Impurity.
- Entropy is 0 if the outcome is certain. Entropy is maximum if we have no knowledge of the system (outcome is equally possible)
- Entropy is used to calculate Information Gain.
- Information Gain is measure of Homogeneity.





Decision Tree : Information Gain

- Entropy

Given a dataset D, contains YES and NO

P(YES) = p, and P(NO) = 1-p

Entropy of D, E(D)

$$E(D) = -p\,log_2(p) - (1-p)log_2(1-p)$$

- Information Gain

Dataset D split into D1, D2, ..., Dk based on variable V

Entropy of each Di can be calculated as E(Di)

Information Gain IG(D,V)

$$IG(D,V) = E(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} E(D_i)$$

Decision Tree : Basic Idea

Perform 3 steps for every single Node and its splitting result

- Step-1

  Find best splitter on each variable

- Step-2

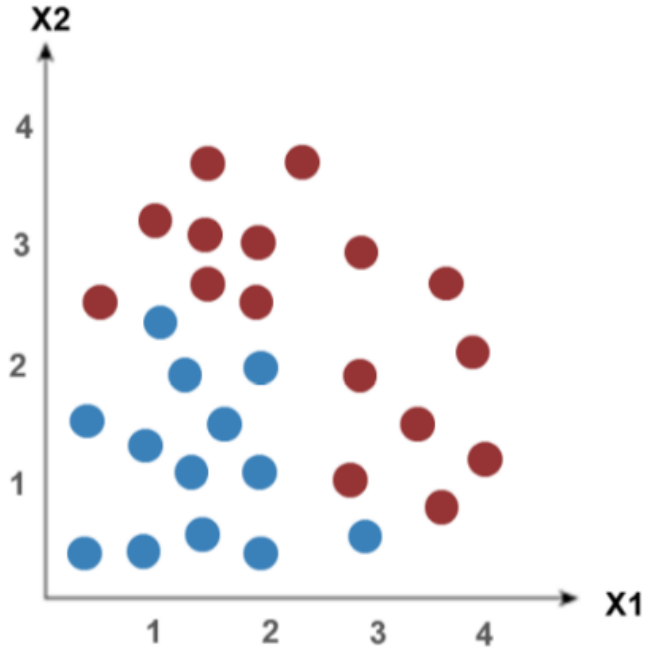
  Select best variable for splitting

- Step-3

  Perform splitting based on result on Step-2.

  Check if the splitting should stop.

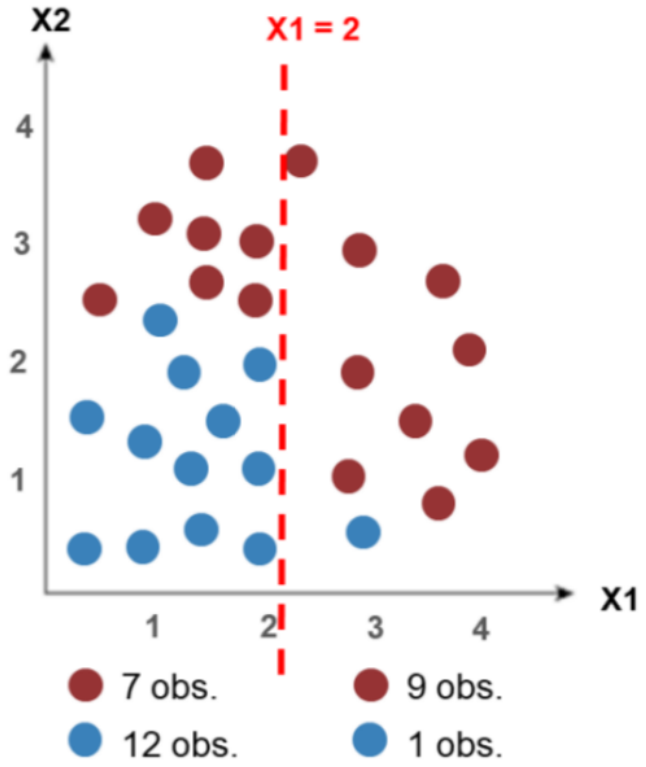
Decision Tree : Basic Idea

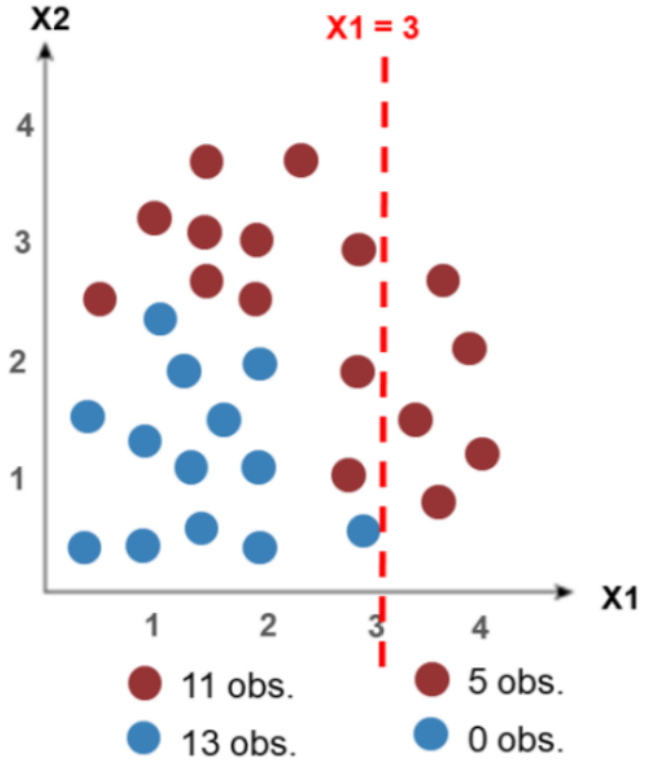


- 16 obs.
- 13 obs.

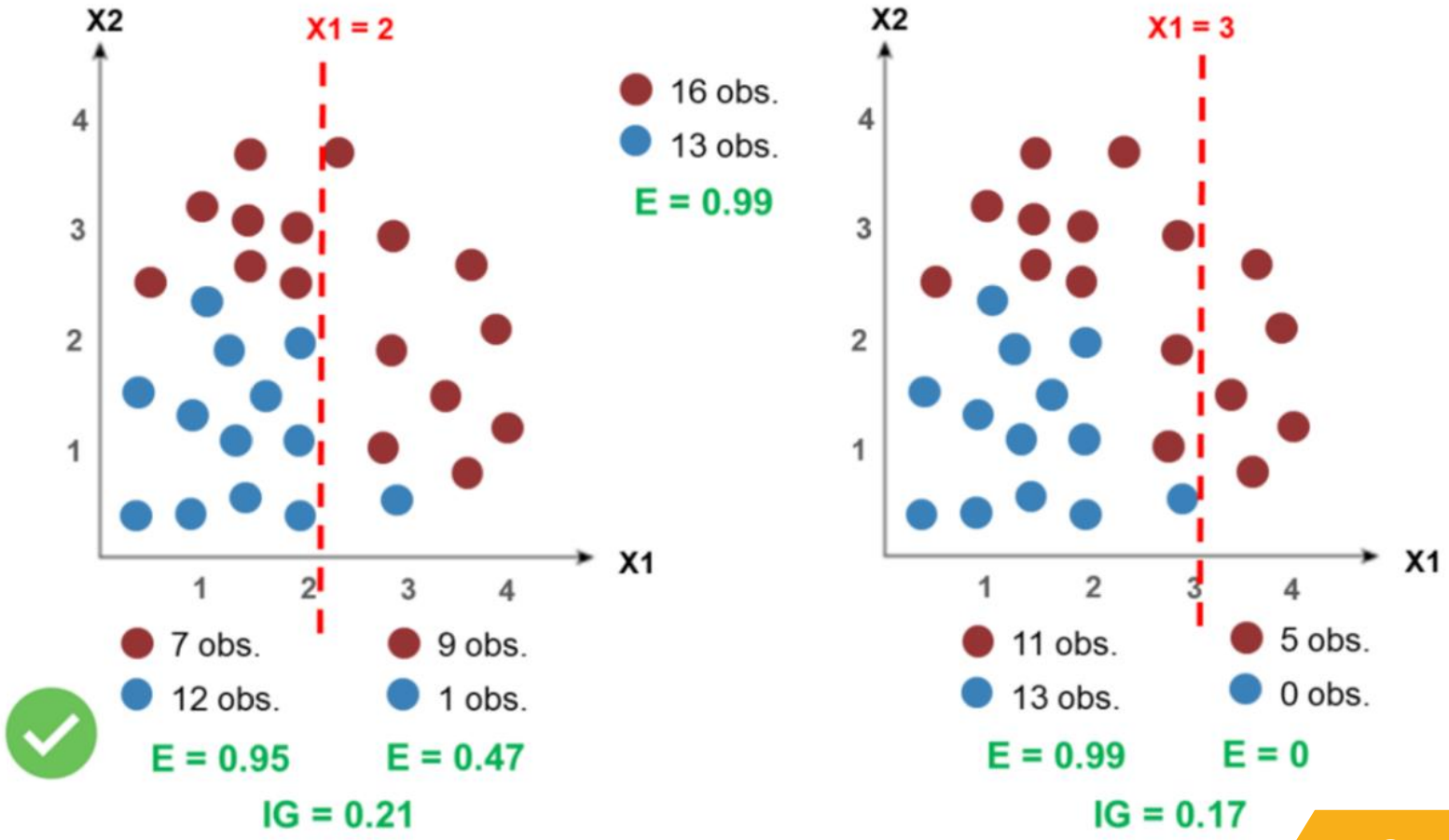
- Find the best splitter between ● & ●
- Best splitter is the one results most homogenous element in each class

Decision Tree : Basic Idea

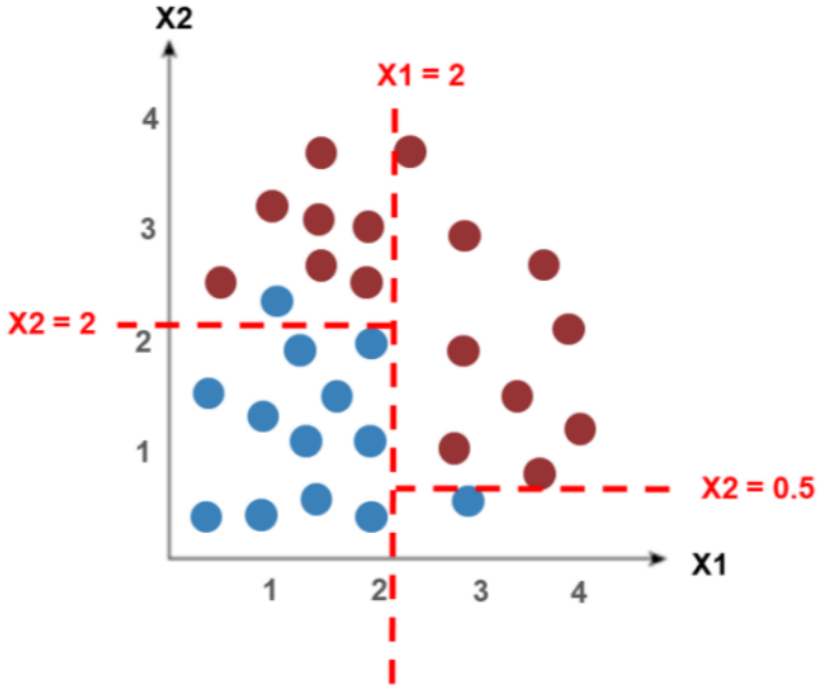


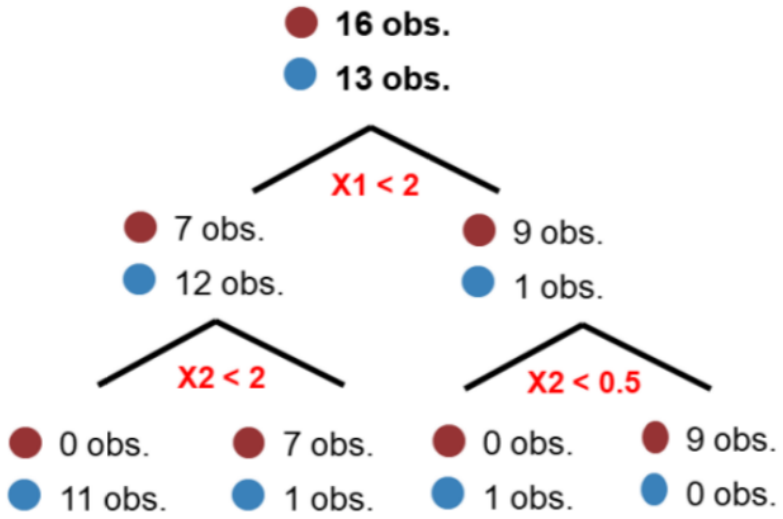


Decision Tree : Basic Idea



Decision Tree : Basic Idea



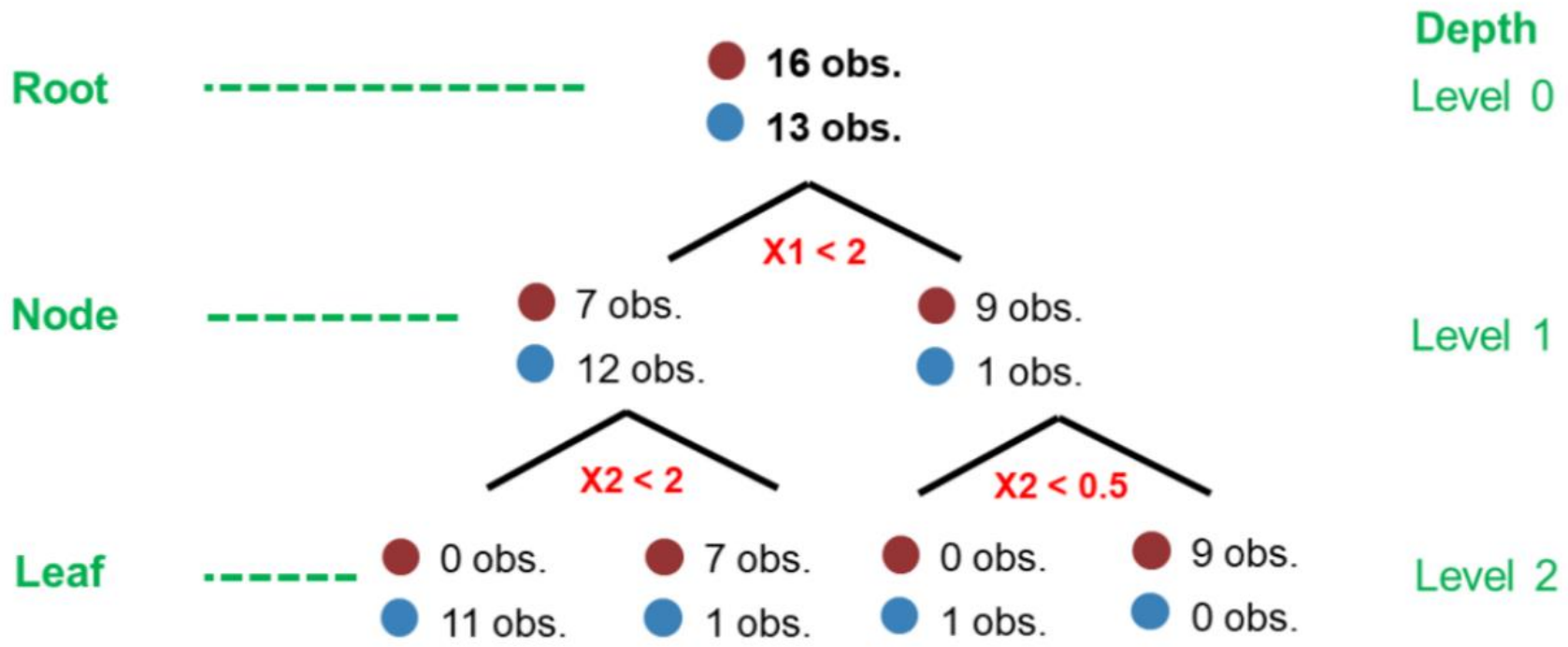


Continue splitting on each group.

At X2 = 2 for class X1 < 2

At X2 = 0.5 for class X1 > 2

Decision Tree : Basic Idea

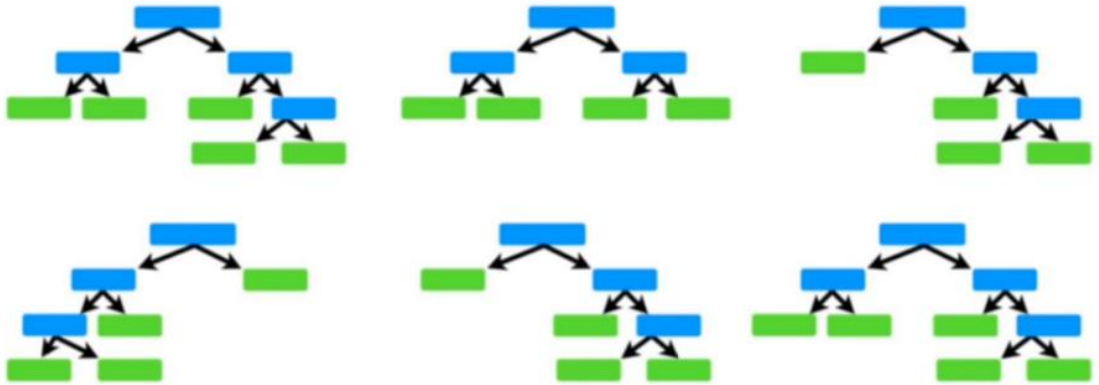


DigitalSkola
Uncover The World Of Digital Skills With Us

Classification : Random Forest

Random Forest

- Random Forests are made out of decision trees.
- Decision trees inside random forest are composed of different bootstrapped data
- Random Forest combine simplicity of decision trees with flexibility resulting in a vast improvement in accuracy

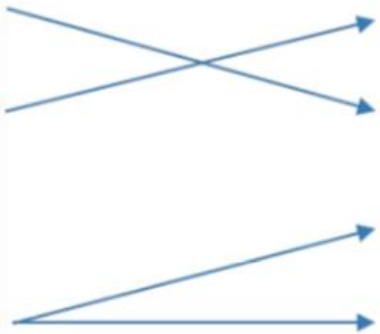


Random Forest : Bootstrap

- Bootstrap is used to estimate parameter by using sampling with replacement.

Original Dataset

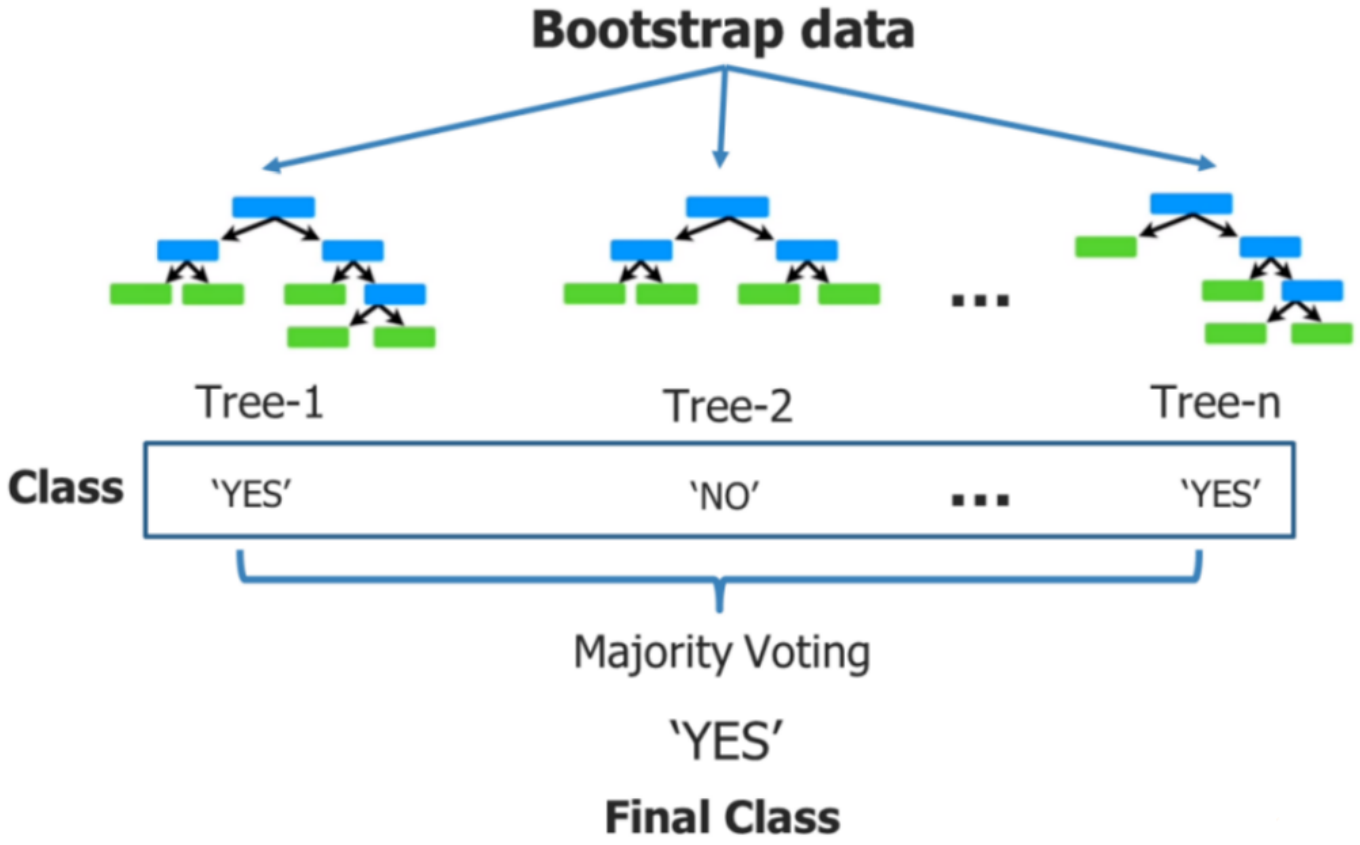
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes



Bootstrapped Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Random Forest



DigitalSkola
Uncover The World Of Digital Skills With Us

Evaluation Metrics

Confusion Matrix

N x N matrix where N is the number of classes being predicted.

Some definitions:

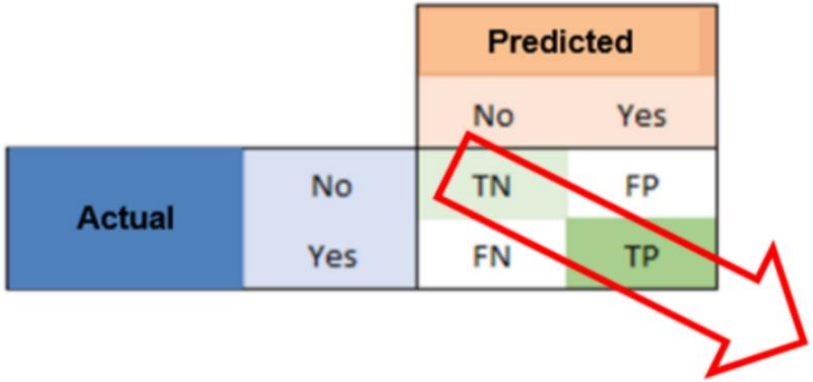
- Accuracy
- Precision (Positive Predictive Value)
- Sensitivity (Recall)
- Specificity

		Predicted	
		No	Yes
Actual	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

Accuracy

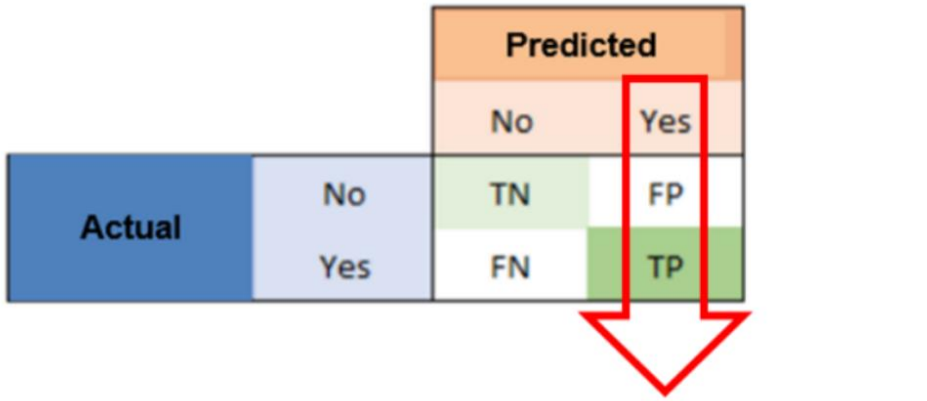
Proportion of total number of correct prediction



Accuracy (Correct Rate) = (TP + TN) / N

Precision

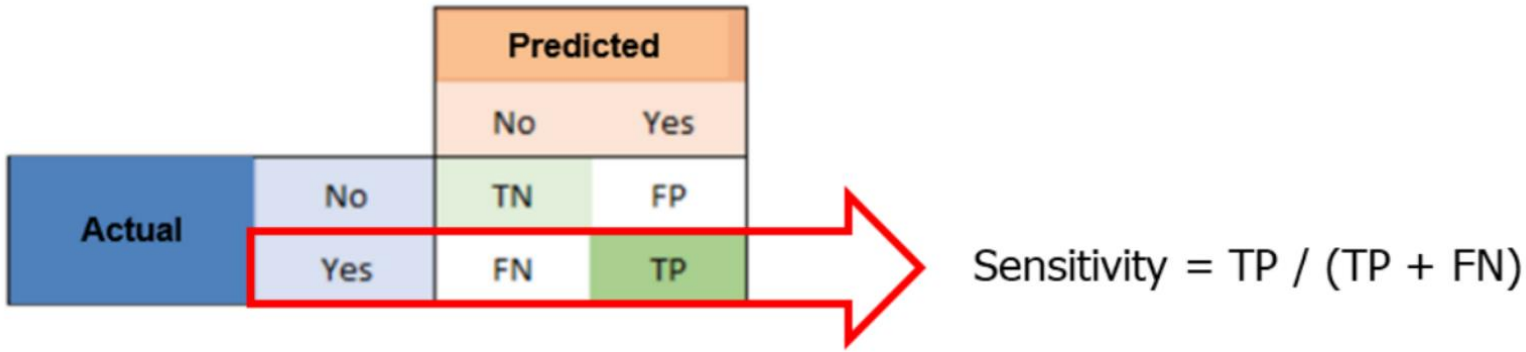
From all predicted “Yes”, how many actually “Yes”



Precision = TP / (TP + FP)

Recall

From all actual “Yes”, how many predicted as “Yes”



DigitalSkola

Uncover The World Of Digital Skills With Us

Thank You