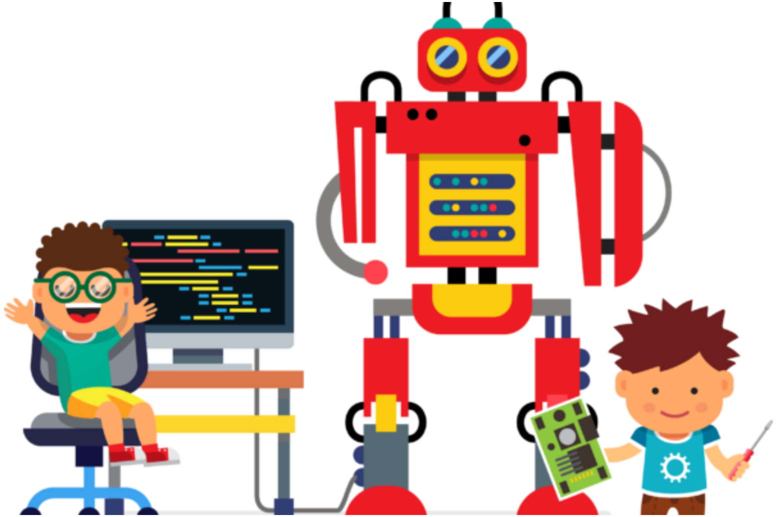
Machine Learning Beginner Course

2.Data Preprocessing

2. معالجة البيانات

دورة تعليم الالة للمبتدئين



Salam again!

In the last class we‘ve introduce ML and general python programming language, and to start doing ML, the first thing is to find and prepare the data, this what we are going to see today.

“Data will talk If you choose listen to it.”

1.

ML Workflow

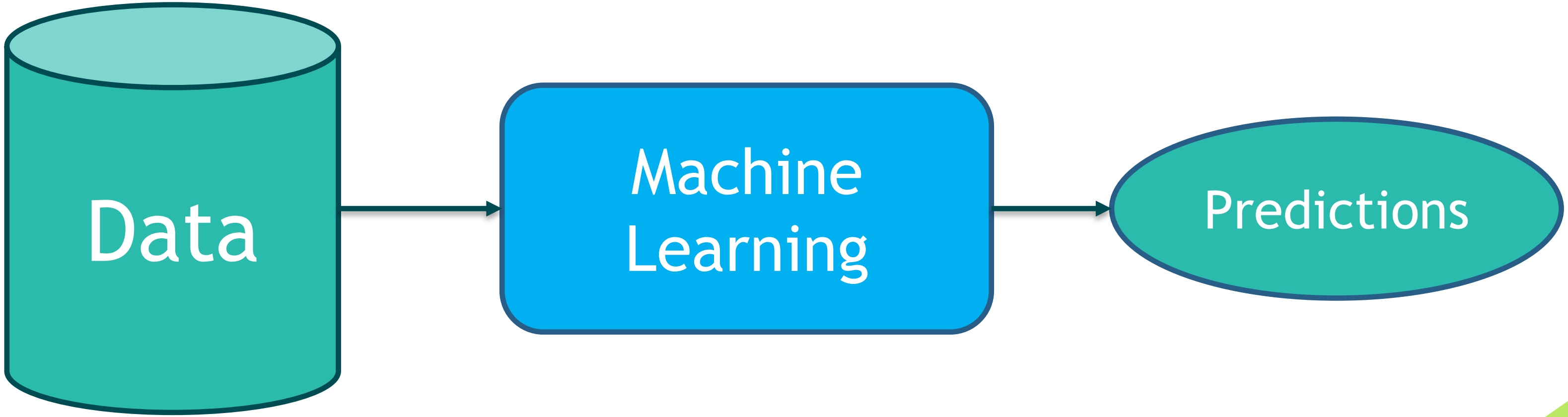
The cores Tasks In ML

Machine Learning Workflow

Each Domain Has it's own principles and Best practices, and machine learning is the same thing, it does include many tasks which we going to present in the following slides.

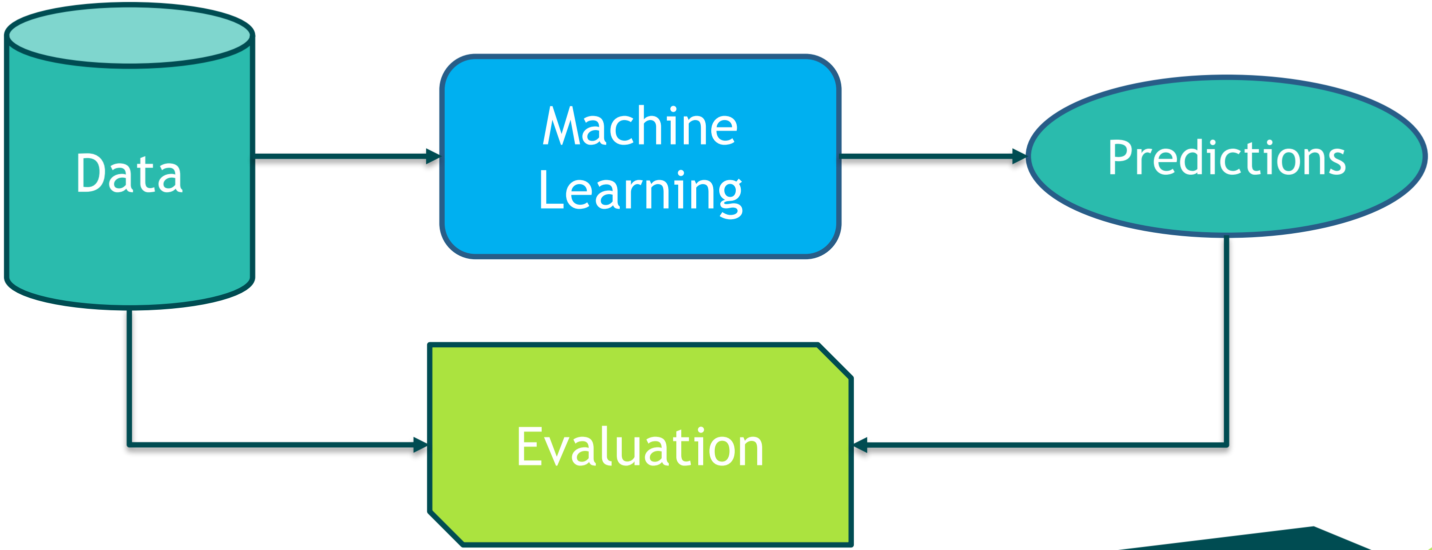
High Level View

At the first Place here is what we can say about ML Task

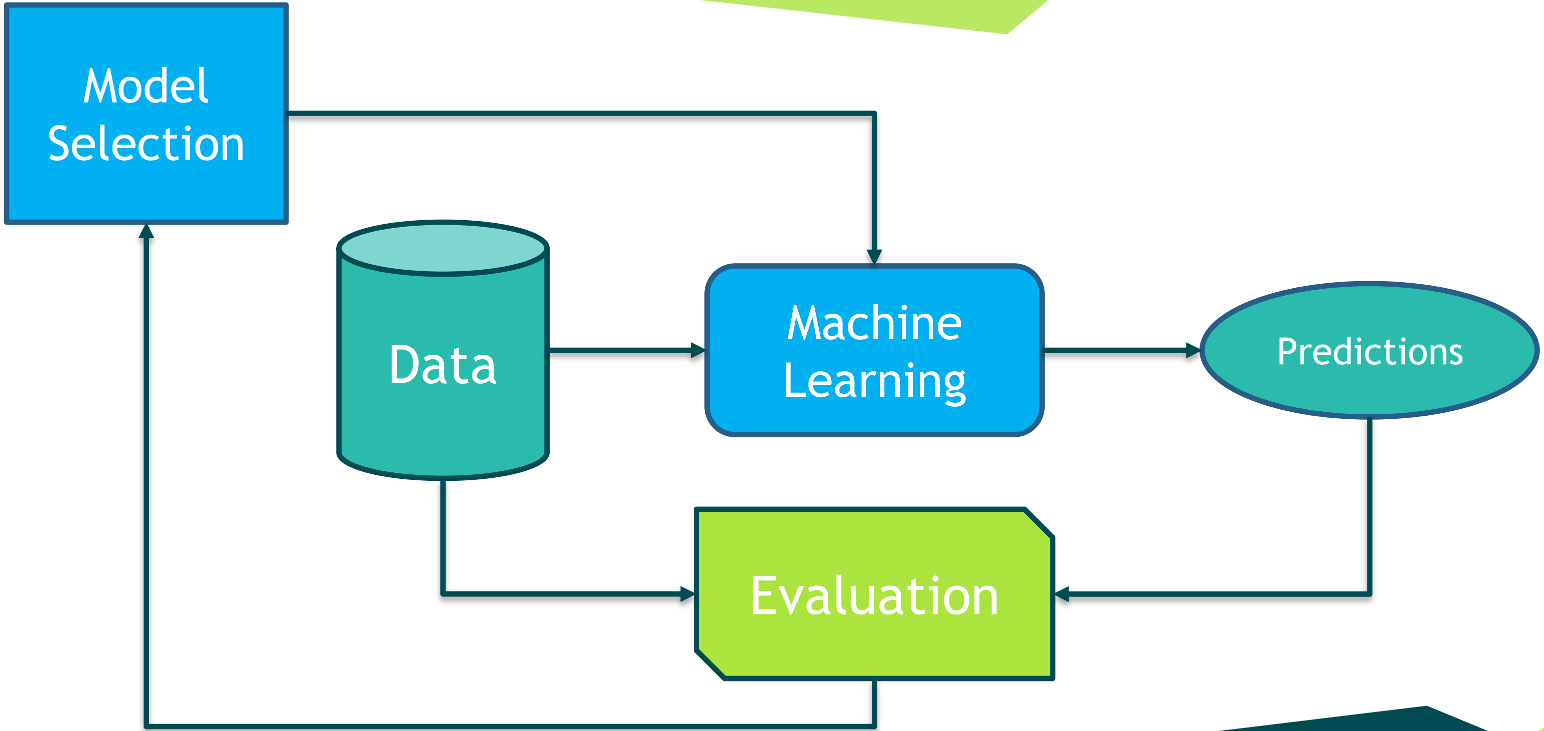


Adding Evaluation

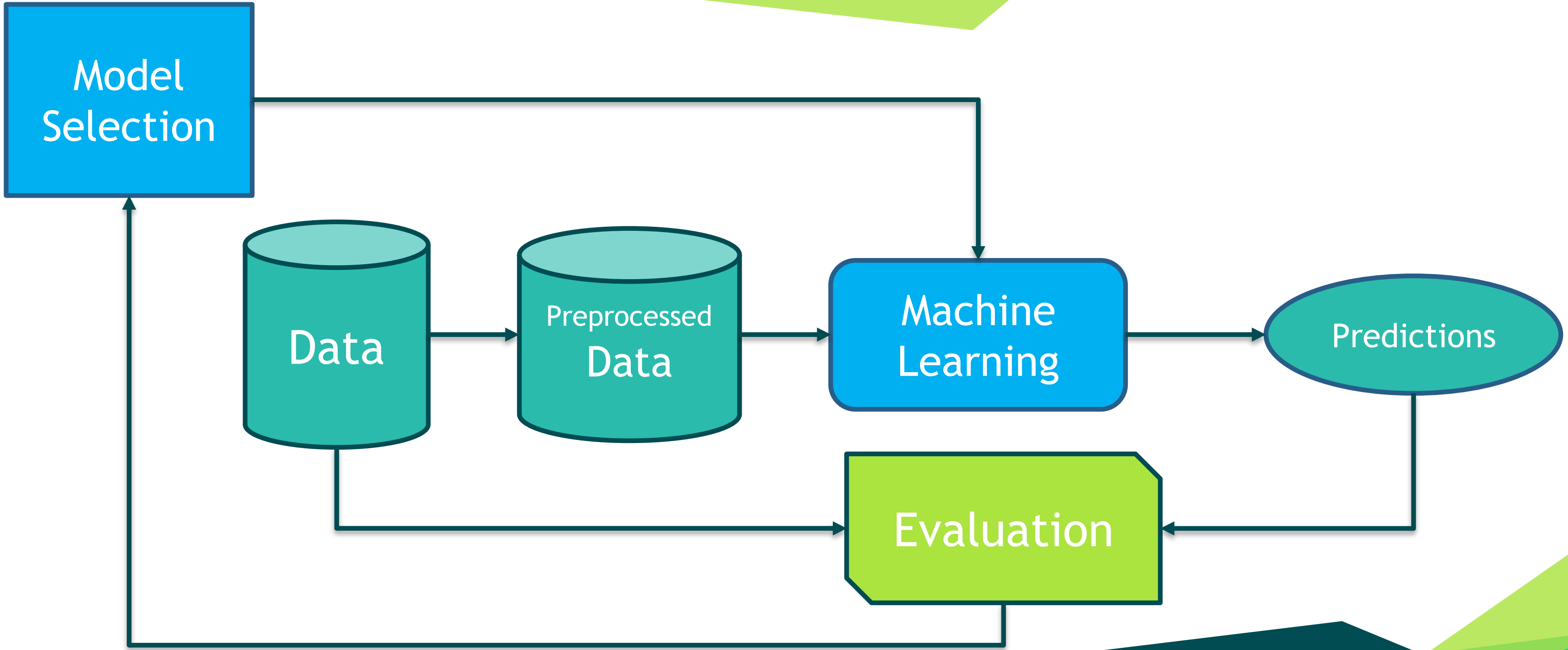
The missing thing in the last diagram was evaluation.



Model Selection



Data Preprocessing



2.

Get the data.

Where to get Some Data

Places

There a lot of places you can get fresh data to work with, here is the famous one

kaggle

https://www.kaggle.com/datasets

UCI

Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets.html

https://github.com/awesomedata/awesome-public-datasets

3.

Preprocessing.

Task in Preprocessing

“Data”

Features or Labels

Country	Age	Salary	Purchased
Algeria	44	72000	No
Tunis	27	48000	Yes
Morocco	30	54000	No
Tunis	38	61000	No
Morocco	40		Yes
Algeria	35	58000	Yes
Tunis		52000	No
Algeria	48	79000	Yes
Morocco	50	83000	No
Algeria	37	67000	Yes

Example

Target

Load Data

How to load data with Pandas?

Load Data

```
import pandas as pd
data = pd.read_csv('data.csv')
```

Of course there are other format of data so it cause you to read it with another function which is suitable for the data format.

Missing Data

How to deal with missing data?

How deal with this?

It will be represented as nan for Not a number

Country	Age	Salary	Purchased
Algeria	44	72000	No
Tunis	27	48000	Yes
Morocco	30	54000	No
Tunis	38	61000	No
Morocco	40		Yes
Algeria	35	58000	Yes
Tunis		52000	No
Algeria	48	79000	Yes
Morocco	50	83000	No
Algeria	37	67000	Yes

What to do ?

Just delete it :D

Country	Age	Salary	Purchased
Algeria	44	72000	No
Tunis	27	48000	Yes
Morocco	30	54000	No
Tunis	38	61000	No
Algeria	35	58000	Yes
Algeria	48	79000	Yes
Morocco	50	83000	No
Algeria	37	67000	Yes

But we can’t lose our data just like that, especially when nan is a lot

Replace With Statistics

Country	Age	Salary	Purchased
Algeria	44	72000	No
Tunis	27	48000	Yes
Morocco	30	54000	No
Tunis	38	61000	No
Morocco	40	60333.33	Yes
Algeria	35	58000	Yes
Tunis	37.333	52000	No
Algeria	48	79000	Yes
Morocco	50	83000	No
Algeria	37	67000	Yes

Replacing By the Mean can solve the problem.

Replace With Statistics

Country	Age	Salary	Purchased
Algeria	44	72000	No
Tunis	27	48000	Yes
Morocco	30	54000	No
Tunis	38	61000	No
Morocco	40	56000	Yes
Algeria	35	58000	Yes
Tunis	37.5	52000	No
Algeria	48	79000	Yes
Morocco	50	83000	No
Algeria	37	67000	Yes

Also Replacing By the Median we can solve the problem.

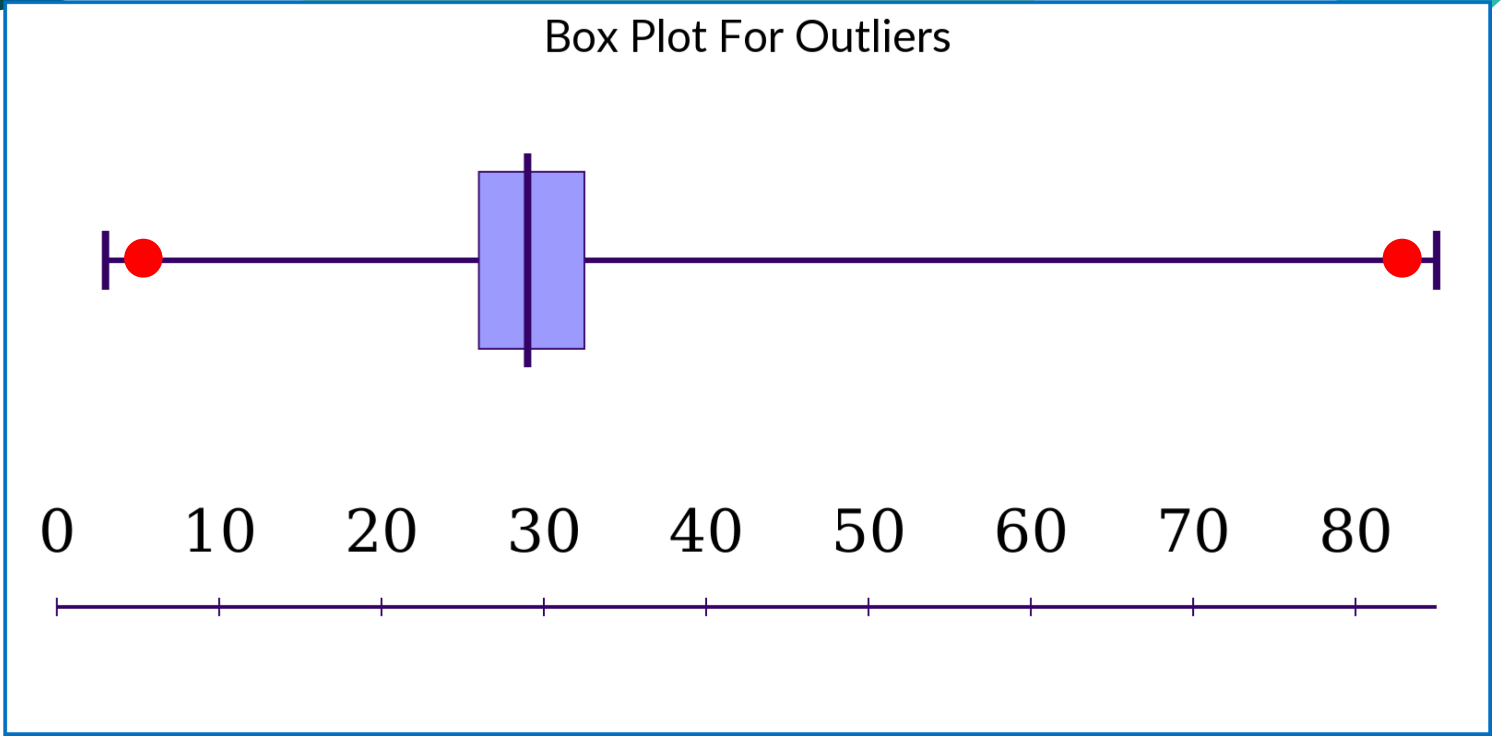
Be carful of outliers!!

Really they can cause huge problems !! In statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error.

Here is an Example:

25	29	3	32	85	33	27	29

Box Plot for outliers.



Bad Example

Salaries Data

Average is 5211.7

8945
2905
2862
2577
2183
5522
8294
9882
7929
1018

Salaries Data With Rabrab or Ali Hadad Included

8945
2905
2862
2577
2183
5522
8294
9882
7929
1018
1000000

Average now is 95647 !!

How to encode Categorical data?

Categorical Data

ML Algorithm Are Math

Country	Age	Salary	Purchased
Algeria	44	72000	No
Tunis	27	48000	Yes
Morocco	30	54000	No
Tunis	38	61000	No
Morocco	40	60333.33	Yes
Algeria	35	58000	Yes
Tunis	37.333	52000	No
Algeria	48	79000	Yes
Morocco	50	83000	No
Algeria	37	67000	Yes

ML Algorithm Expect Numbers and not String !

Encoding Binary Variable

Country	Age	Salary	Purchased
Algeria	44	72000	0
Tunis	27	48000	1
Morocco	30	54000	0
Tunis	38	61000	0
Morocco	40	60333.33	1
Algeria	35	58000	1
Tunis	37.333	52000	0
Algeria	48	79000	1
Morocco	50	83000	0
Algeria	37	67000	1

Just Put ones and zeros

✓

Yes/no Questions, Male Female and any variable that can be at most in two states

Encoding Multi-class Variables

Country	Age	Salary	Purchased
1	44	72000	0
2	27	48000	1
3	30	54000	0
2	38	61000	0
3	40	60333.33	1
1	35	58000	1
2	37.333	52000	0
1	48	79000	1
3	50	83000	0
1	37	67000	1

Just put a number corresponding to each label.

✖

But why?

Numbers are tricky, The integer values have a natural ordered relationship between each other and machine learning algorithms may be able to understand and harness this relationship. using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results

For example, ordinal variables like the "High Medium and Low" example would be a good example where a label encoding would be sufficient.

One hot Encoding

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction by adding a new binary variable for each unique value.

One hot Encoding Example

Color
RED
GREEN
BLUE
RED
GREEN

One Hot

Color Red	Color Green	Color Blue
1	0	0
0	1	0
0	0	1
1	0	0
0	1	0

Back to our data

Country	Age	Salary	Purchased
Algeria	44	72000	0
Tunis	27	48000	1
Morocco	30	54000	0
Tunis	38	61000	0
Morocco	40	60333.33	1
Algeria	35	58000	1
Tunis	37.333	52000	0
Algeria	48	79000	1
Morocco	50	83000	0
Algeria	37	67000	1

We need to add 3 columns to encode it.

Back to our data

Algeria	Tunis	Morocco	Age	Salary	Purchased
1	0	0	44	72000	0
0	1	0	27	48000	1
0	0	1	30	54000	0
0	1	0	38	61000	0
0	0	1	40	60333.3	1
1	0	0	35	58000	1
0	1	0	37.3	52000	0
1	0	0	48	79000	1
0	0	1	50	83000	0
1	0	0	37	67000	1

Here is the expected result !

What is this ?

Feature Scaling

Variables War

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations, this is a problem.

Example

# Room	Price
1	20000 DA
3	35000 DA
2	25000 DA
5	12130 DA
1	40021 DA
7	32000 DA
5	13222 DA

Example

# Room	Price
1	20000 DA
3	35000 DA
2	25000 DA
5	12130 DA
1	40021 DA
7	32000 DA
5	13222 DA

Ecludian Distance

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Here is a calculation: ✖

$\sqrt{(1 - 3)^2 + (20000 - 35000)^2}$ = 15000.001

Solution is to Scale

To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling. There are four common methods to perform Feature Scaling.

Standardization

Standardization replaces the values by their Z scores.

$$\boldsymbol{\grave{x} = \frac{x - \bar{x}}{\sigma}}$$

Mean Normalization

This distribution will have values between -1 and 1with μ=0.

$$\grave{x} = \frac{x \ - mean(x)}{\max(x) - \min(x)}$$

Min-Max Scaling

This scaling brings the value between 0 and 1.

$$\grave{x} = \frac{x\ - min(x)}{\mathrm{max}(x) - \mathrm{min}(x)}$$

Unit Vector

Scaling is done considering the whole feature vector to be of unit length.

$$\grave{\boldsymbol{x}} = \frac{\boldsymbol{x}}{||\boldsymbol{x}||}$$

When to Scale

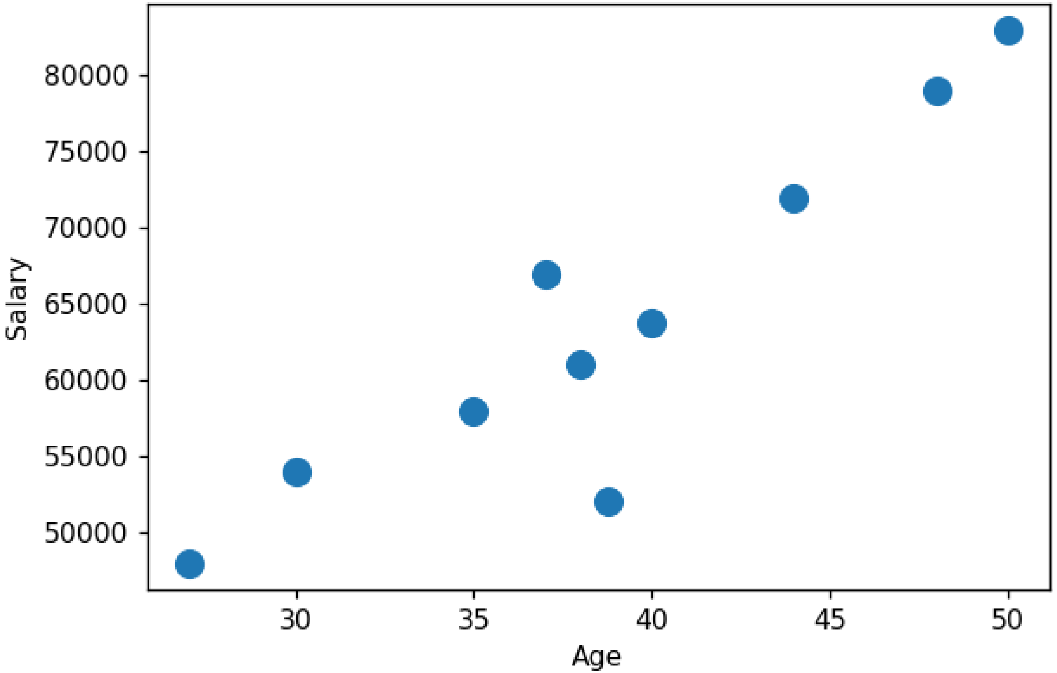
Rule of thumb I follow here is any algorithm that computes distance or assumes normality, scale your features !!!

Back to our data

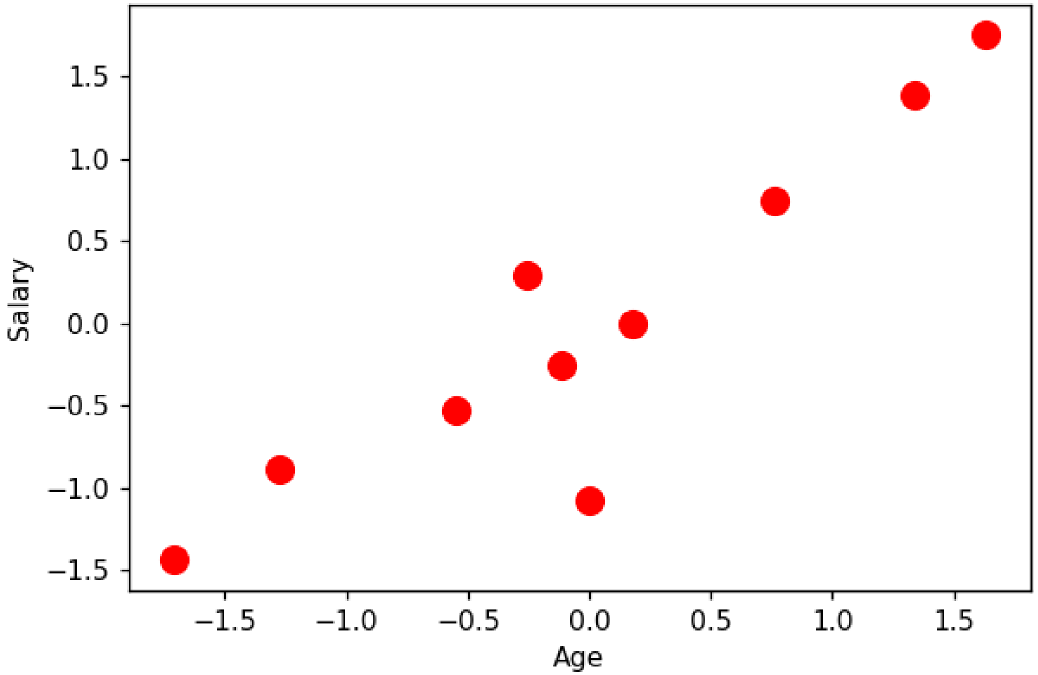
Algeria	Tunis	Morocco	Age	Salary	Purchased
1	0	0	0.758874	0.749473	0
0	1	0	-1.7115	-1.43818	1
0	0	1	-1.27555	-0.891265	0
0	1	0	-0.113024	-0.2532	0
0	0	1	0.177609	6.63219e-16	1
1	0	0	-0.54897	-0.526657	1
0	1	0	0	-1.07357	0
1	0	0	1.34014	1.38754	1
0	0	1	1.63077	1.75215	0
1	0	0	-0.25834	0.293712	1

Here is the expected result !

Before Scaling



After Scaling



How to evaluate model?

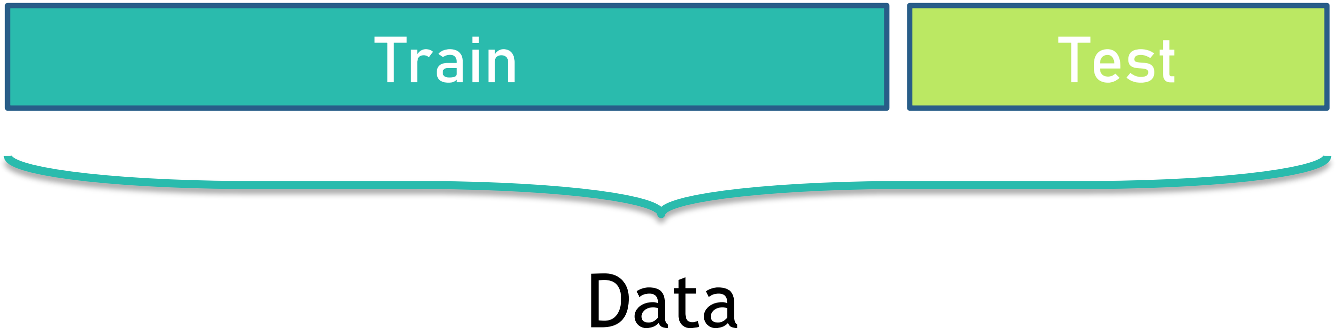
Train/Test Splitting

The problem

After we got our data, we will give it to ML algorithm , form which the algorithm will learn, but how we can test it and evaluate it, if we just input the data which he trained on we can risk of memorizing this data point and for that he will give good results, so we need a completely new data, an **unseen** data from which we are going to evaluate the resulting model.

Splitting

Simple solution is to split the data in two parts, one for training and the other for testing.



Train

Test

Data

Splitting Example

```
import pandas as pd
data = pd.read_csv('data.csv' )
x_values = data.iloc[:,:-1].values
y_values = data.iloc[:,-1].values

from sklearn.model_selection import train_test_split
x_tr,x_ts,y_tr,y_ts = train_test_split(x_values , y_values, test_size = 0.2)
```

Practical Time

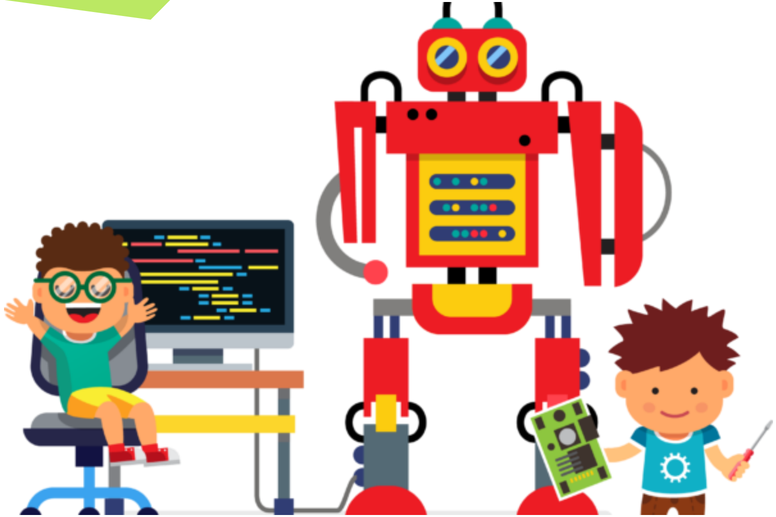
Open up your PC, launch your anaconda and let's do some data preprocessing.

]

Coding Interview

Dropbox

Spreadsheets often use this alphabetical encoding for its columns:
"A", "B", "C", ..., "AA", "AB", ..., "ZZ", "AAA", "AAB", ...
Given a column number, return its alphabetical column id.
For example, given 1, return "A". Given 27, return "AA"



شكرا لحضوركم

Thanks for Assisting!