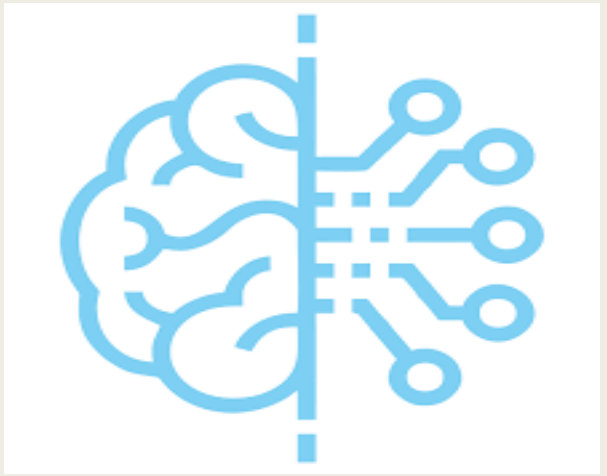
DATA PREPROCESSING

Using python

The Machine Learning Steps:

Step1: Data Pre-processing

- Import the dataset and libraries
- Clean the data
- Split the data into training and testing sets



Step2: Modeling

- Create the model
- Train the model
- Make predictions

Step3: Evaluation

- Calculate the performance
- Make a decision

Data Preprocessing (a.k.a. Data Wrangling)

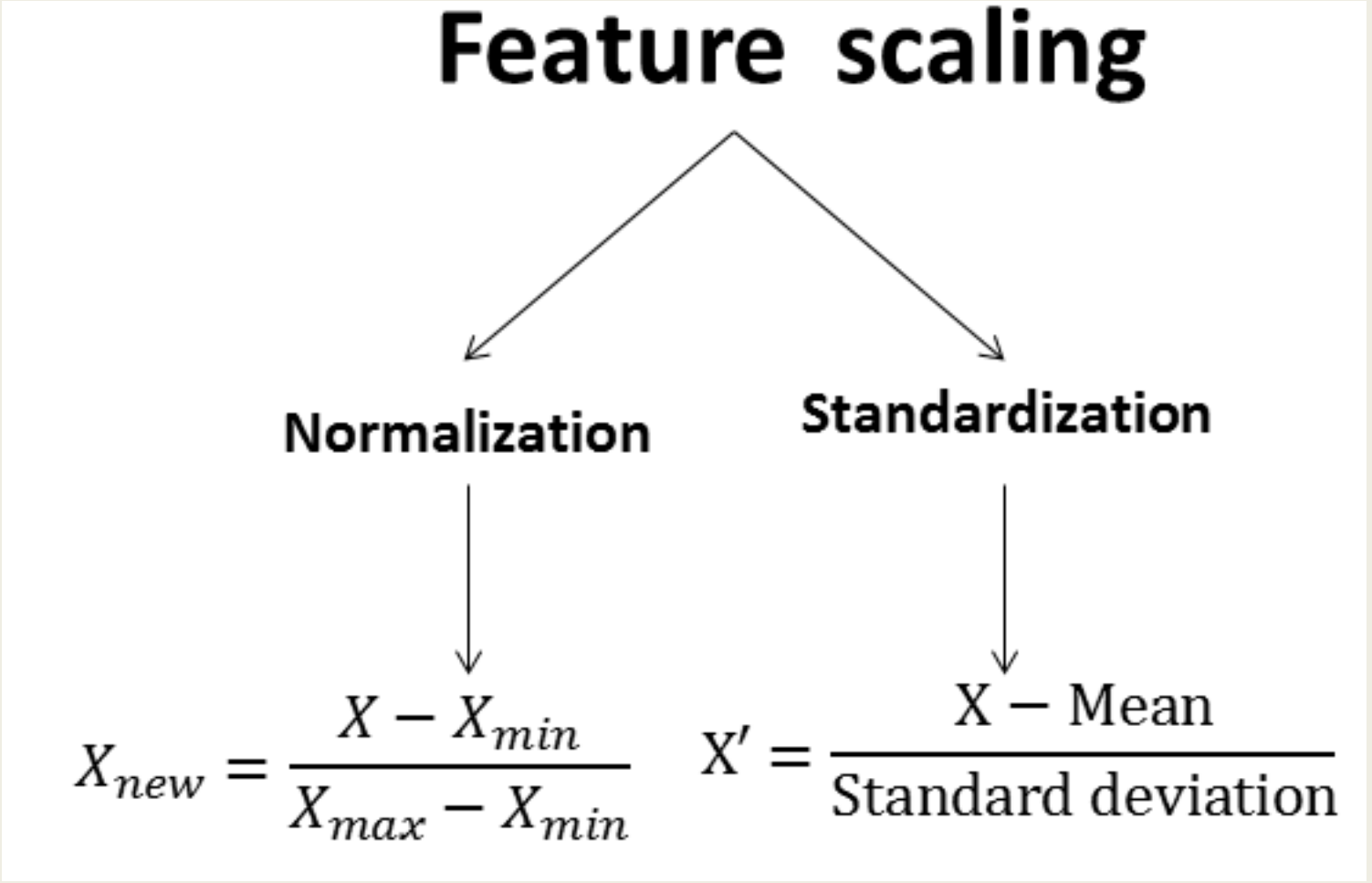
Clean the data
- Missing values
- Mis-formatted data
- Outliers

Transform the data
- Scale
- Turn categorical data into metric data
- Turn metric data into categorical data

Exploratory data analysis
- Visualization
- Statistical analysis

Feature Scaling

house size (square feet)	House age (years)	House price ($ dollar)
2,467	10	800,000
1,050	5	500,000
3,000	25	1,500,000
7,000	2	2,000,000
5,500	35	2,500,000
1,900	8	900,000



‘Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution. Normalization scales in a range of [0,1] or [-1,1]. Standardization is not bounded by range(mostly will be between [-3,+3]’.

Feature Scaling example

1,500 2

?

2,000 6

3,000 8

Feature Scaling example

size?

AGE

1,500 2

500 4

2,000 6

1000 2

3,000 8

Feature Scaling example

Normalization

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

1,500 - 1,500/(3,000-1,500) = 0

2-2/(8-2) = 0

2,000 - 1,500/(3,000-1,500) = 0.33

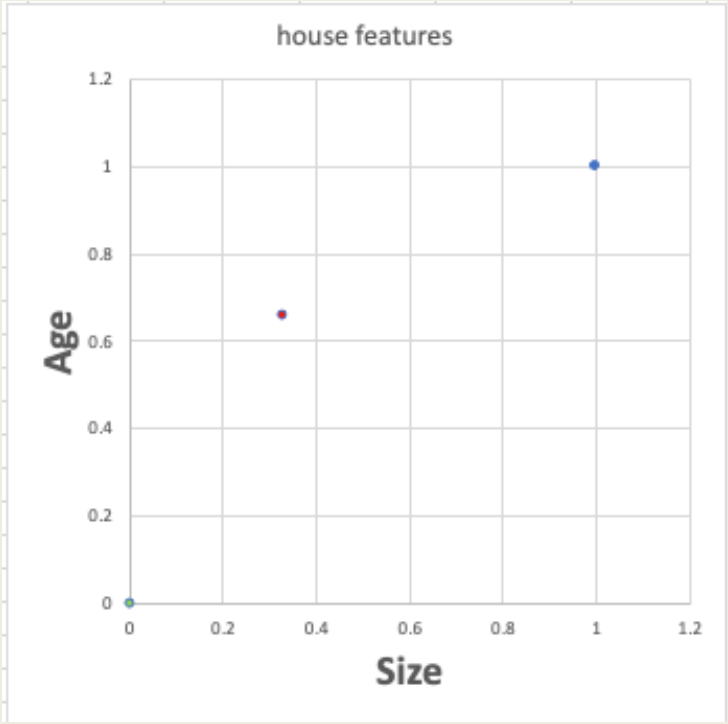
6-2/(8-2) = 0.66

3000 - 1,500/(3,000-1,500) = 1

8-2/(8-2) = 1

Feature scaling example continued

	size	age
green house	0	0
red house	0.33	0.66
blue house	1	1



$$Distance = \sqrt{(0-0.33)^2 + (0-0.66)^2} = 0.737$$

$$Distance = \sqrt{(0.33-1)^2 + (0.66-1)^2} = 0.751$$

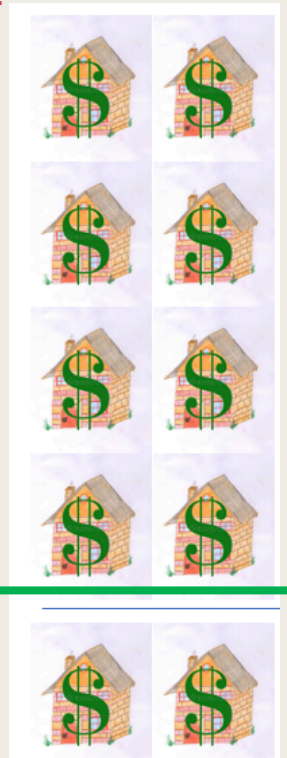
Conclusion: the red house is more similar to the green house than the blue house, since they're closer in distance.

Training and Testing sets

Based on

Training
80%

Testing
20%



1) Use the training set to train the model, e.g., multiple linear regression model:

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2$$

2) Use this testing set to test the model that created in step 1

3) Evaluate the performance of the model. Compare what the Model predicted for the testing Set, to the actual values of the Testing set.

Predicted values vs Actual values