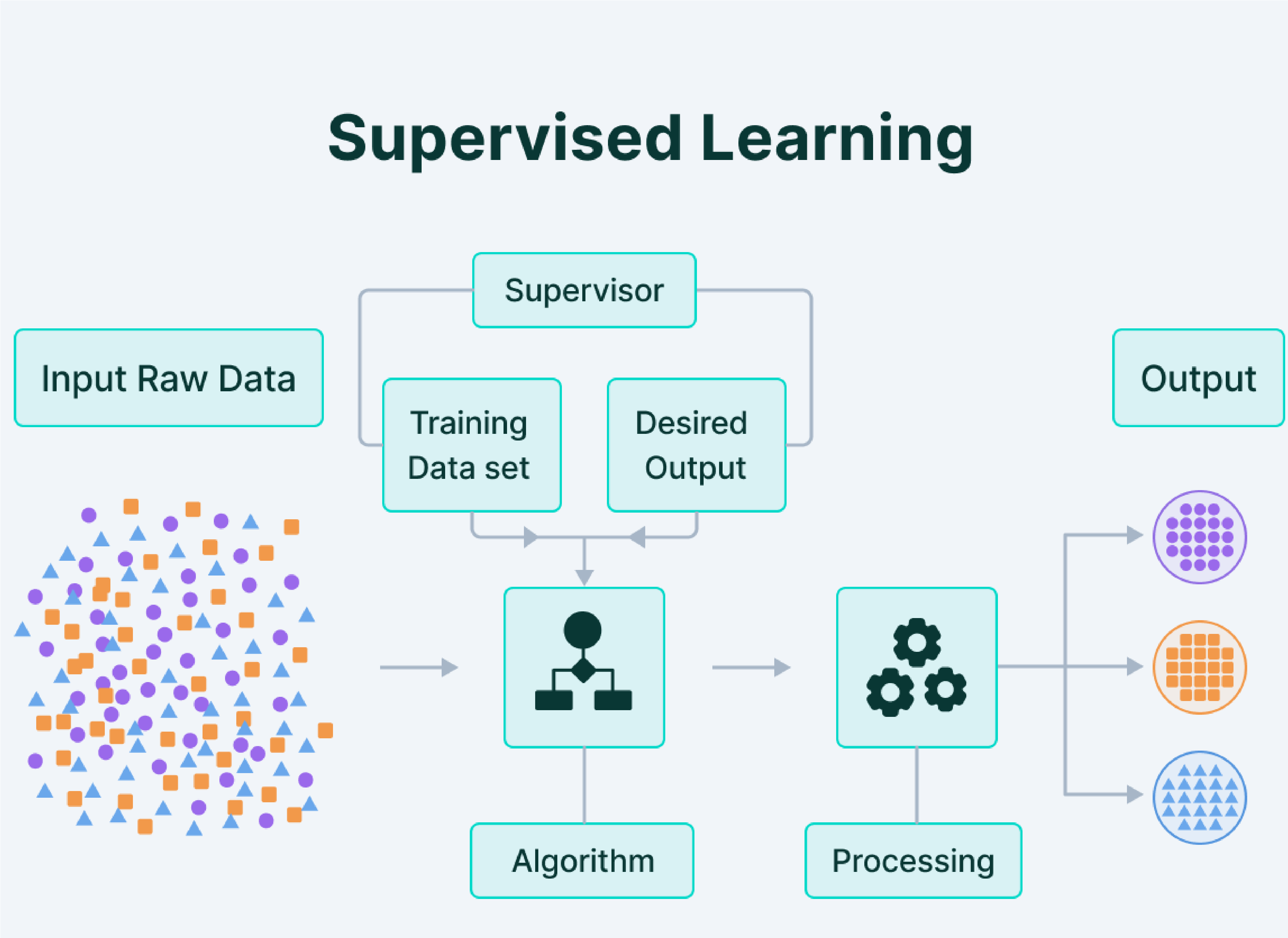
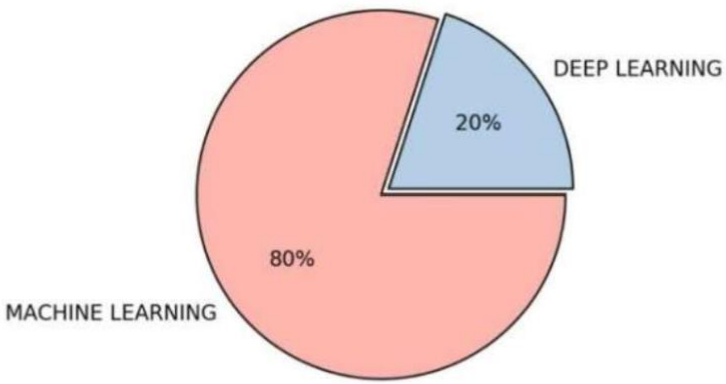
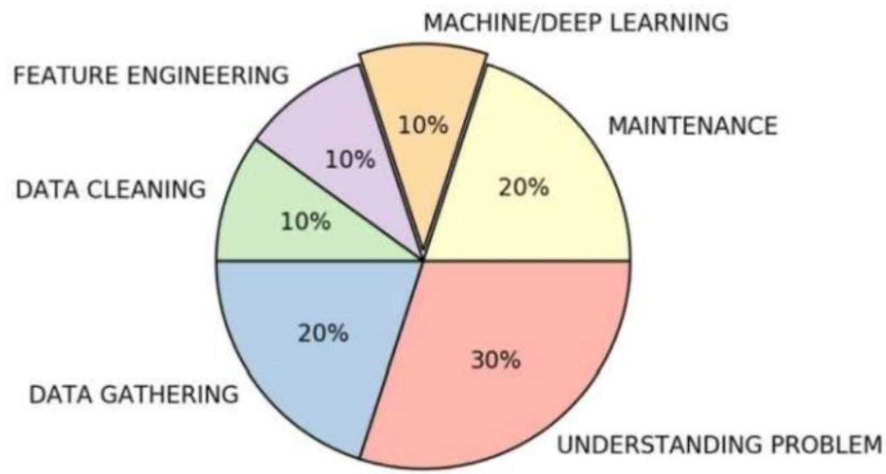

Supervised Learning
Supervisor
Input Raw Data
Training Data set
Desired Output
Output
Algorithm
Processing

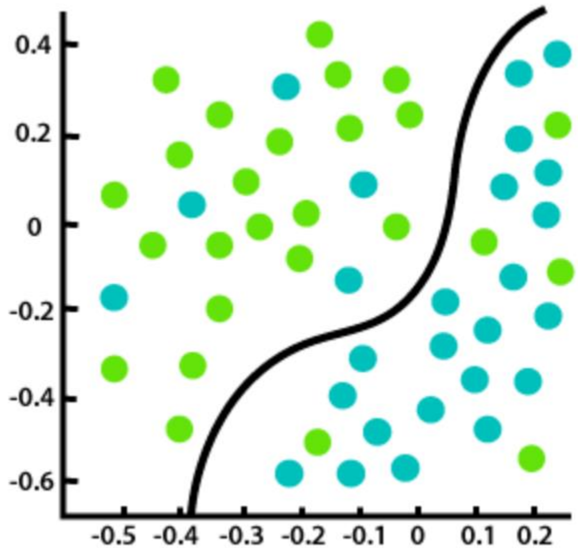
EXPECTATION



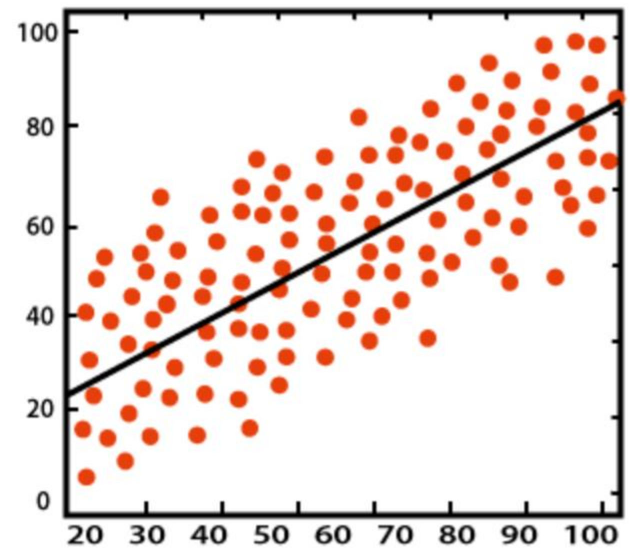
REALITY



SUPERVISIONATO

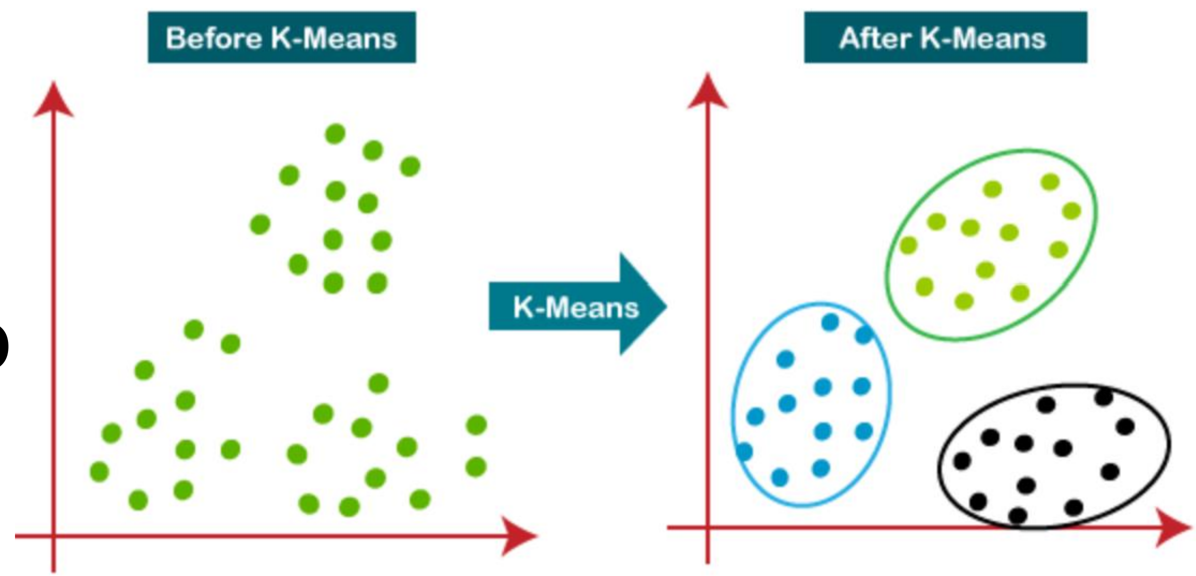


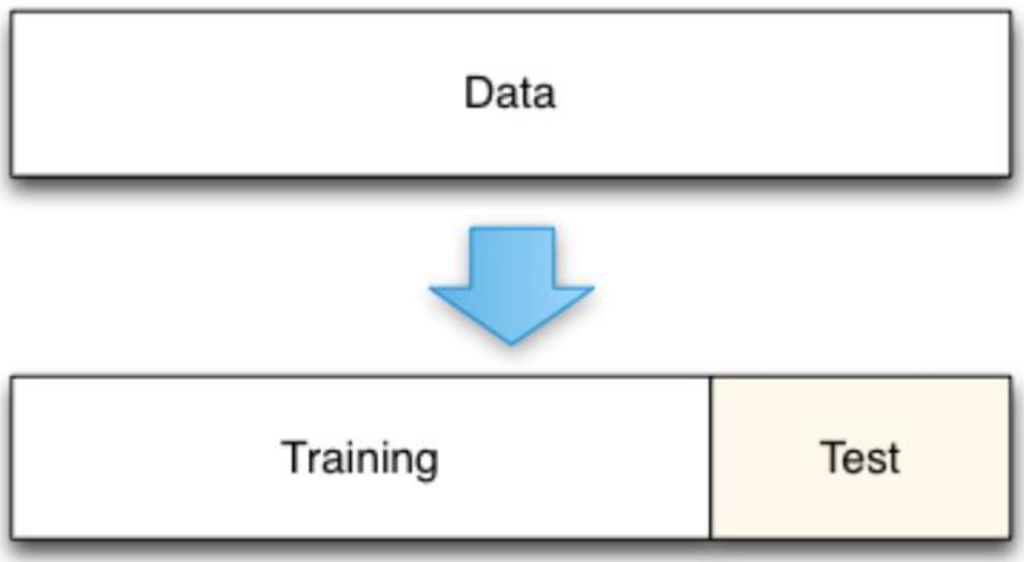
Classification



Regression

NON SUPERVISIONATO

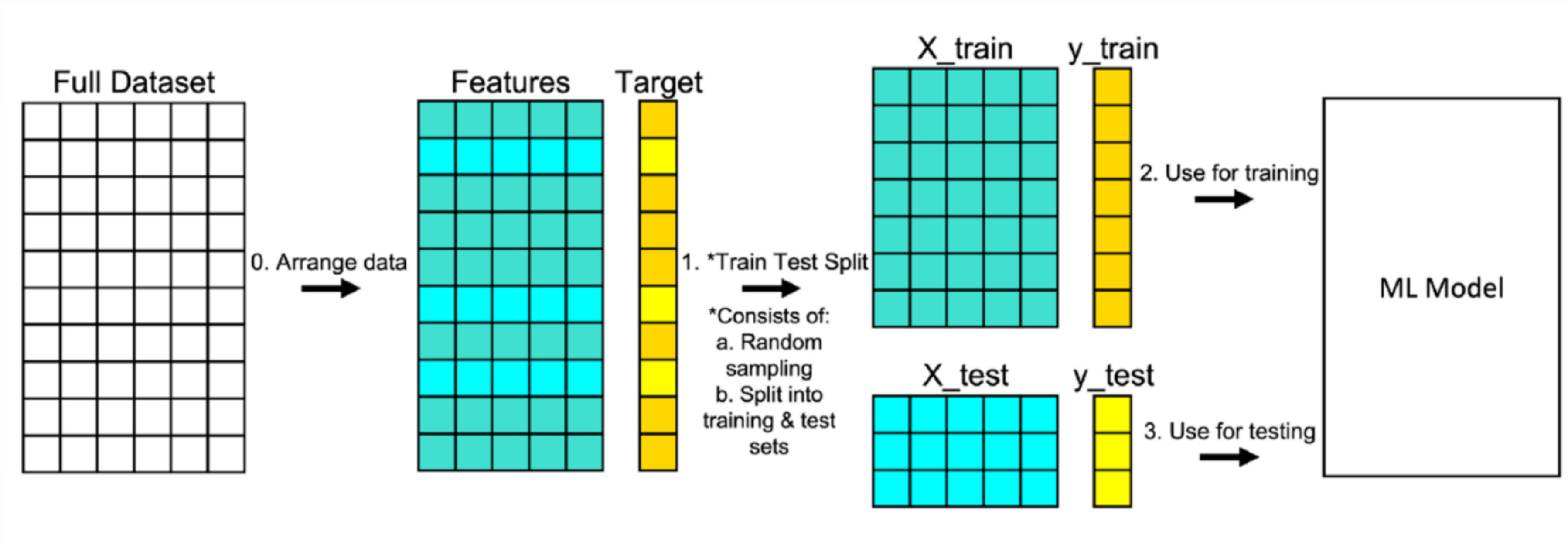


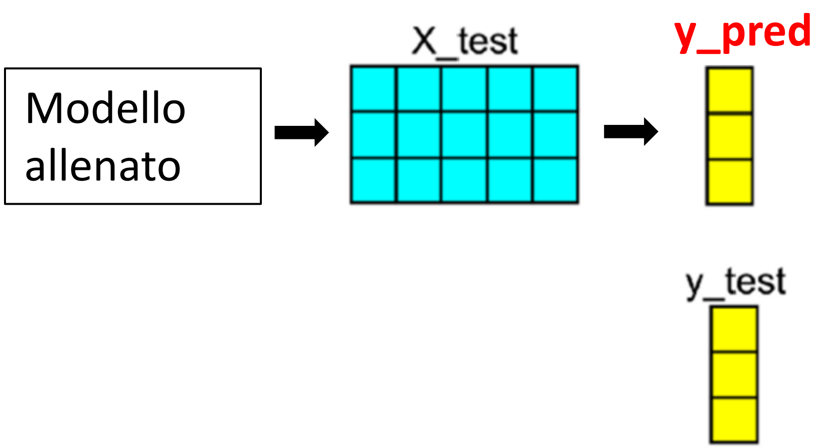

Data
Training
Test


Testing set
Training set

INPUT = FEATURES

OUTPUT = TARGET

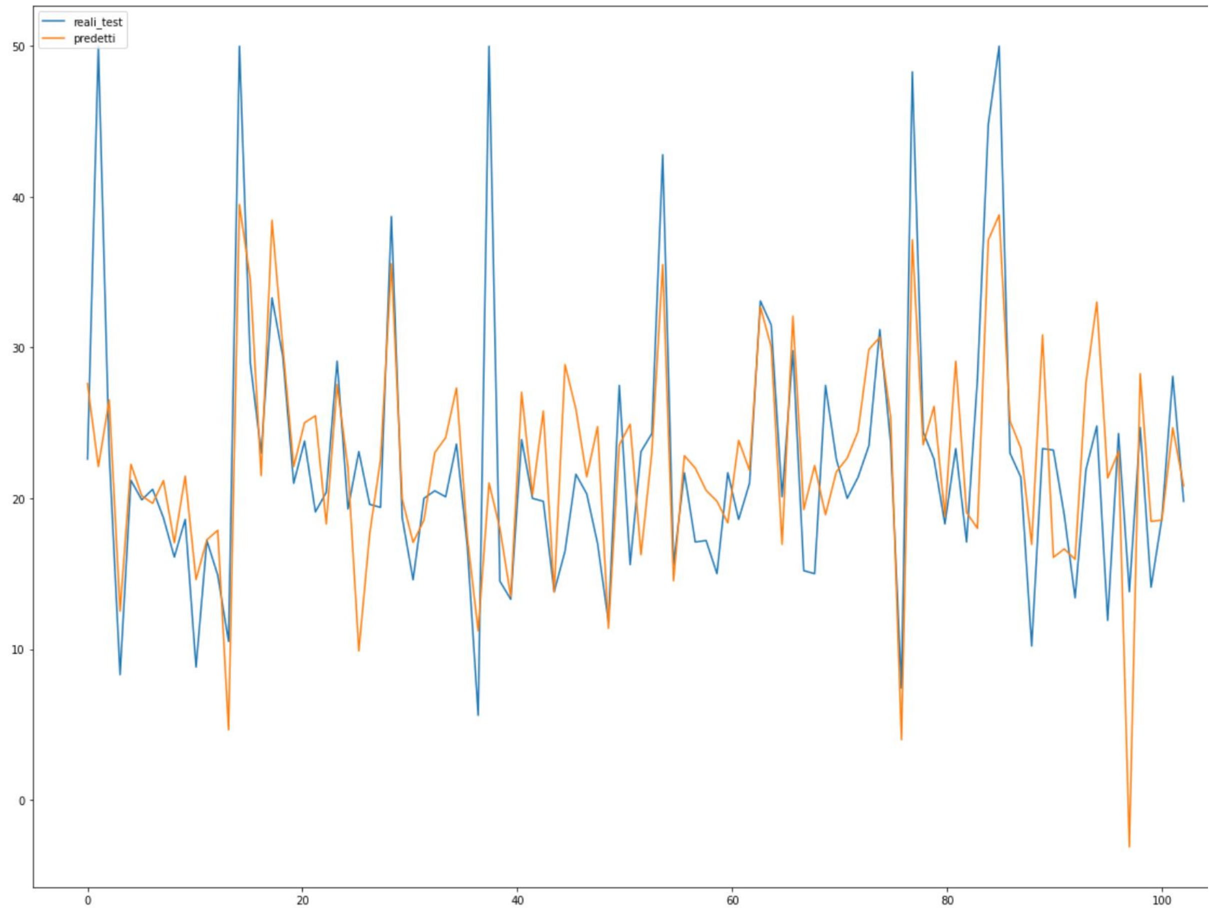


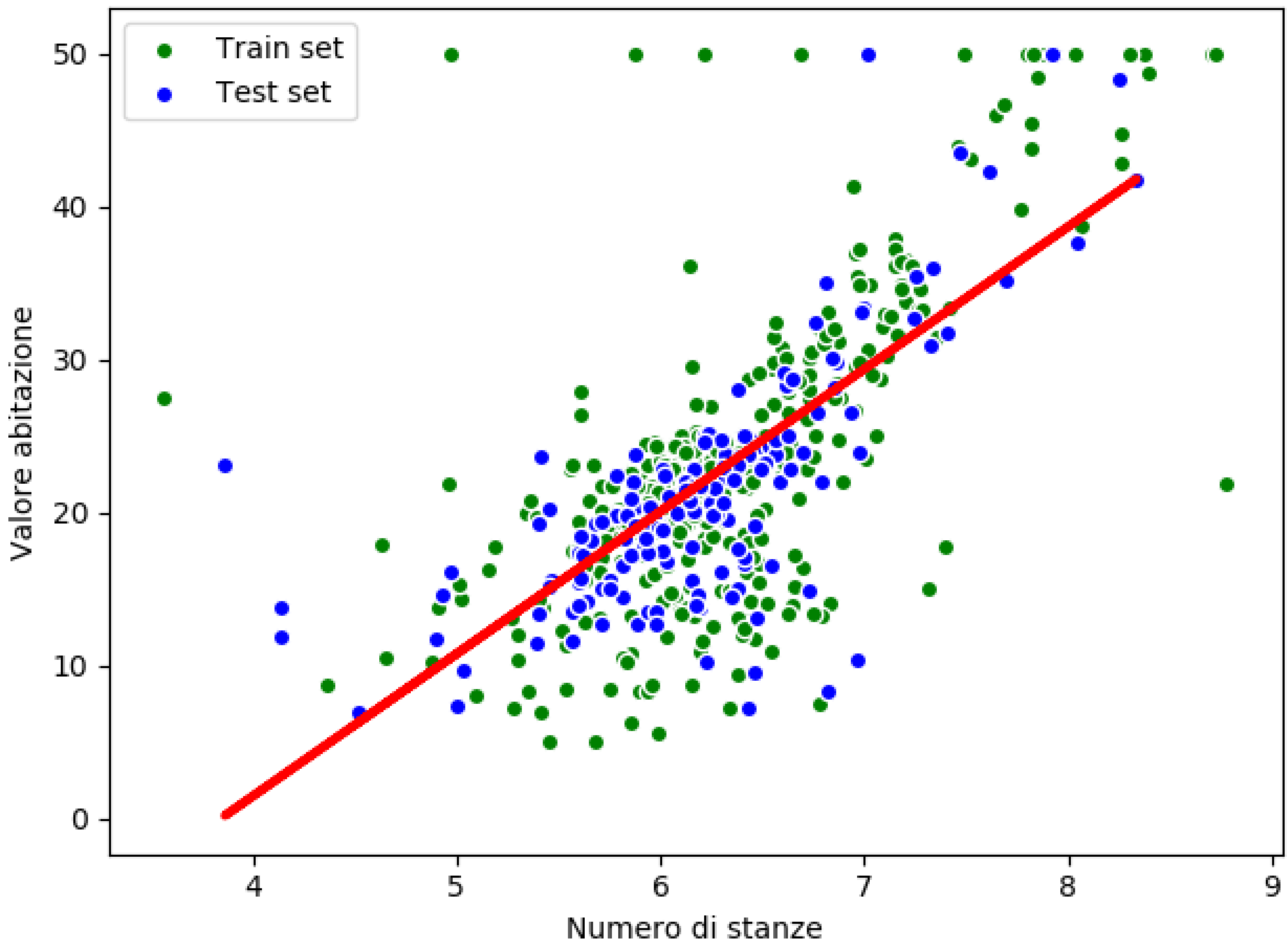


Confronto (y_pred y_test) per valutare errori

	predetti	reali_test
0	27.609031	22.6
1	22.099034	50.0
2	26.529255	23.0
3	12.507986	8.3
4	22.254879	21.2
...	...	...
97	28.271228	24.7
98	18.467419	14.1
99	18.558070	18.7
100	24.681964	28.1
101	20.826879	19.8

	predetti	reali_test
0	27.609031	22.6
1	22.099034	50.0
2	26.529255	23.0
3	12.507986	8.3
4	22.254879	21.2
...	...	...
97	28.271228	24.7
98	18.467419	14.1
99	18.558070	18.7
100	24.681964	28.1
101	20.826879	19.8




Train set
Test set
Valore abitazione
Numero di stanze
0
10
20
30
40
50
4
5
6
7
8
9

Introducendo opportune assunzioni si ottiene il modello di regressione lineare semplice.

Assunzione 1:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$ **per ogni osservazione i=1,...n**

Assunzione 2:

Le ε_i sono variabili casuali indipendenti con valore atteso $E(\varepsilon_i) = 0$ e varianza costante $V(\varepsilon_i) = \sigma^2$ per ogni i=1,...,n

Assunzione 3:

I valori x_i della variabile esplicativa X sono noti senza errore

MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?

The objective of Linear Regression is to find a line that minimizes the prediction error of all the data points.

- The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Where,

$\hat{y}$ – *predicted value of y*
$\bar{y}$ – *mean value of y*

- Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

- Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.

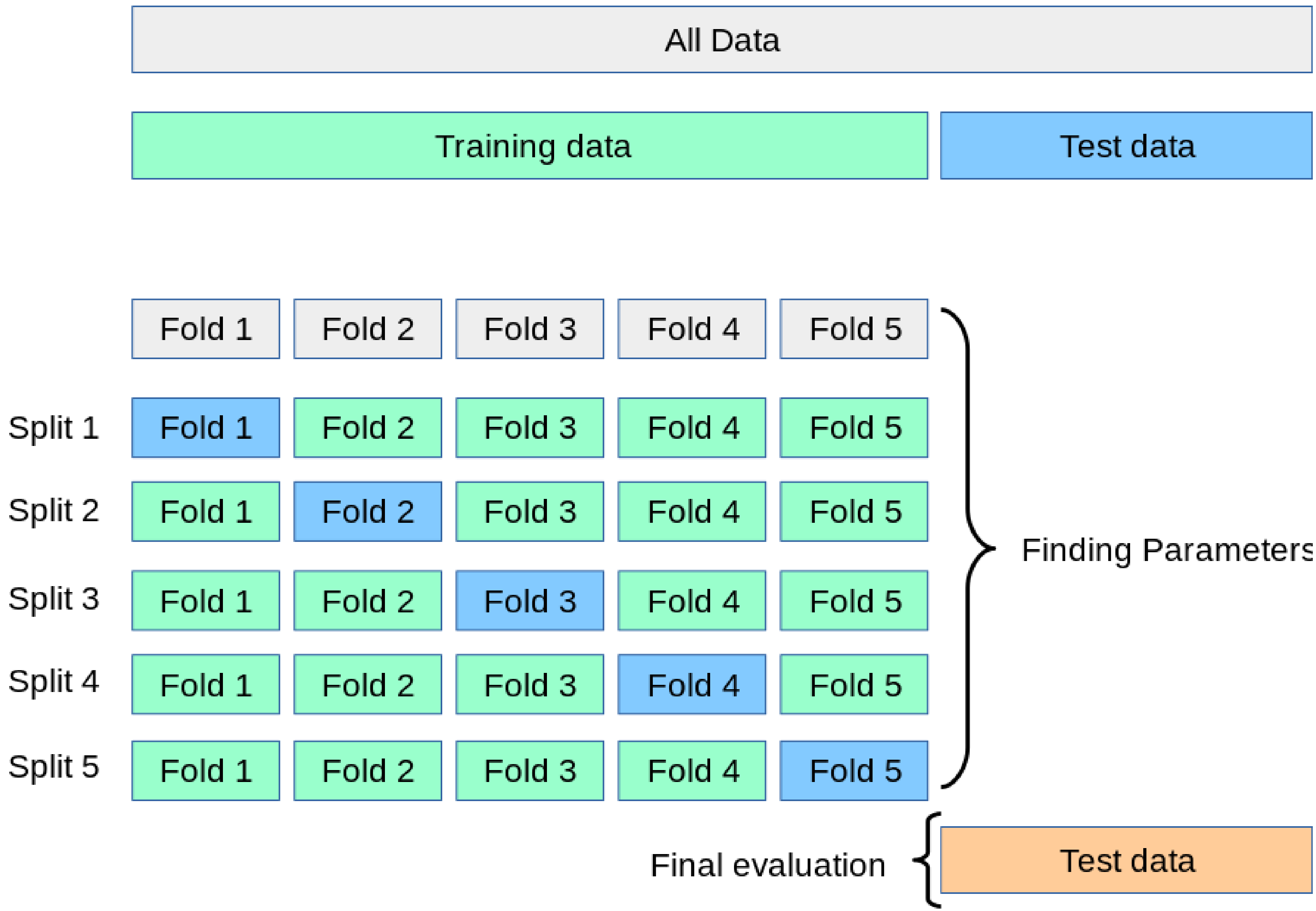
$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

- The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2}$$

- Adjusted R squared is a modified version of R square, and it is adjusted for the number of independent variables in the model, and it will always be less than or equal to R^2.In the formula below **n** is the number of observations in the data and **k** is the number of the independent variables in the data.

$$R_{adj}^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$


All Data
Training data
Test data
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Split 1
Split 2
Split 3
Split 4
Split 5
Finding Parameters
Final evaluation
Test data