Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Développement

université virtuelle Burkina Faso

CITADEL AI Summer School 2022

Regression

19-25.09.22

Rodrique Kafando, PhD
rodrique.kafando@citadel.bf

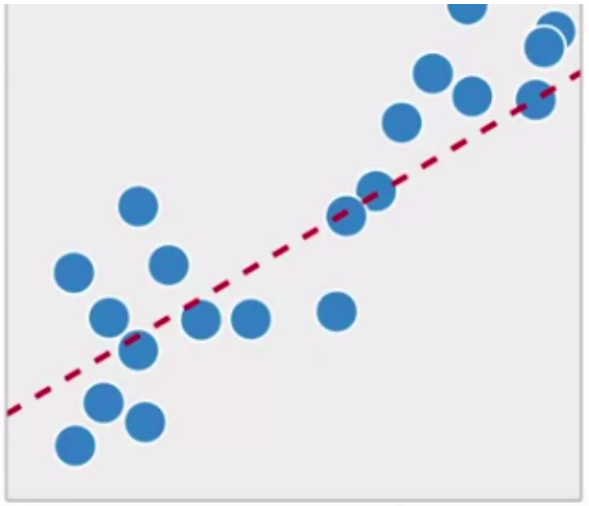
Supervised Learning : Regression

X : independente/explicative variable Y : dependente variable

[8]:

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232

Continuous Values



Regression→ Processus qui consiste à prédire des classes à **valeurs continues**

Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Développement

Supervised Learning : Regression

Simple Regression

Linear Regression
Non-linear Regression

→ Prédire **CO2emission** en se basant uniquement sur **EngineSize**

Multiple Regression

Linear Regression
Non-linear Regression

→ Prédire **CO2emission** en se basant sur **EngineSize** et **Cylinders**

Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Développement

Supervised Learning : Regression

Some Algorithms

* Ordinal regression
* Poisson regression
* Fast forest quantile regression
* Linear, Polynomial, Lasso, Stepwise, Ridge regression
* Bayesian linear regression
* Neural network regression
* Decision forest regression
* Boosted decision tree regression
* K-nearest neighbors

Centre d'Excellence Interdisciplinaire en
Intelligence Artificielle pour le Développement

Supervised Learning : Regression

[23]:

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	???

Centre d'Excellence Interdisciplinaire en
Intelligence Artificielle pour le Développement

Supervised Learning : Regression

Simple Linear Regression

Variable independante (x) : **EngineSize**
Variable dependante (y) : **CO2emission**

Multiple Linear Regression

Variables independantes (x) : **EngineSize, Cylenders , and more...**
Variable dependante (y) : **CO2emission**

Supervised Learning : Regression

$x_1 = 5.4$ independent variable

$y = 250$ actual Co2 emission of x1

$\hat{y} = \theta_0 + \theta_1 x_1$

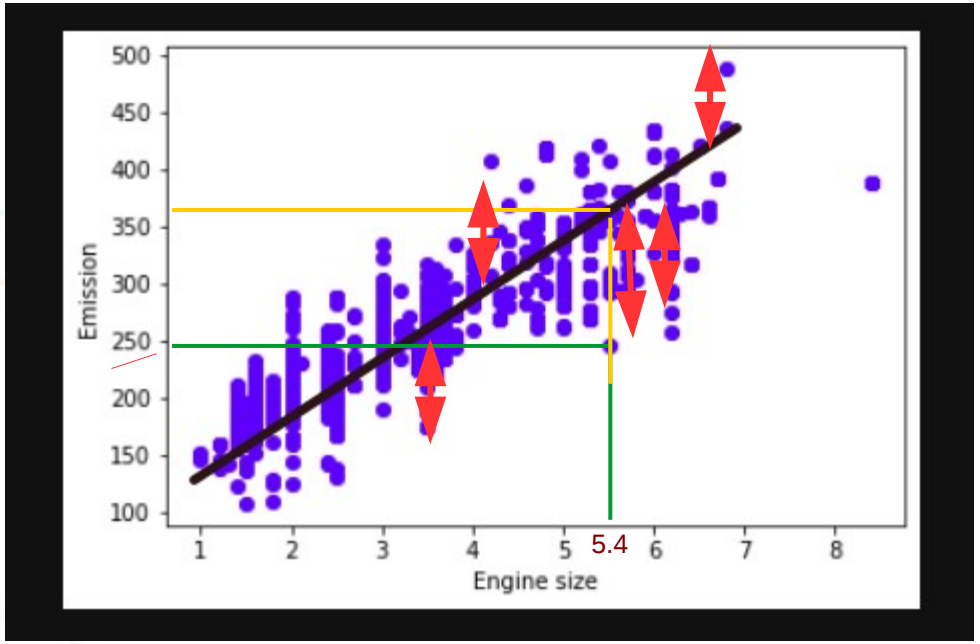
$\hat{y} = 340$ the predicted emission of x1

Error = $y - \hat{y}$

$= 250 - 340$

$= -90$

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



Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Développement

Regression – Model Evaluation

- Train and Test on the Same Dataset
- Train/Test Split

Regression Evaluation Metrics

Regression – Model Evaluation

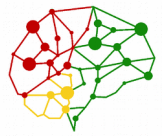
[23]:

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Train

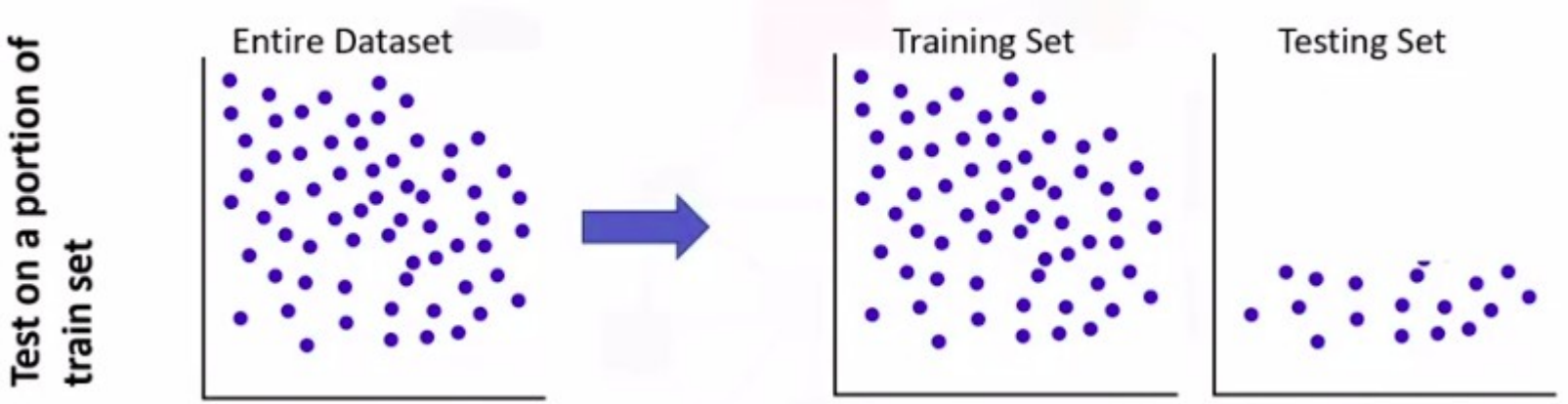
Test

Actual Values

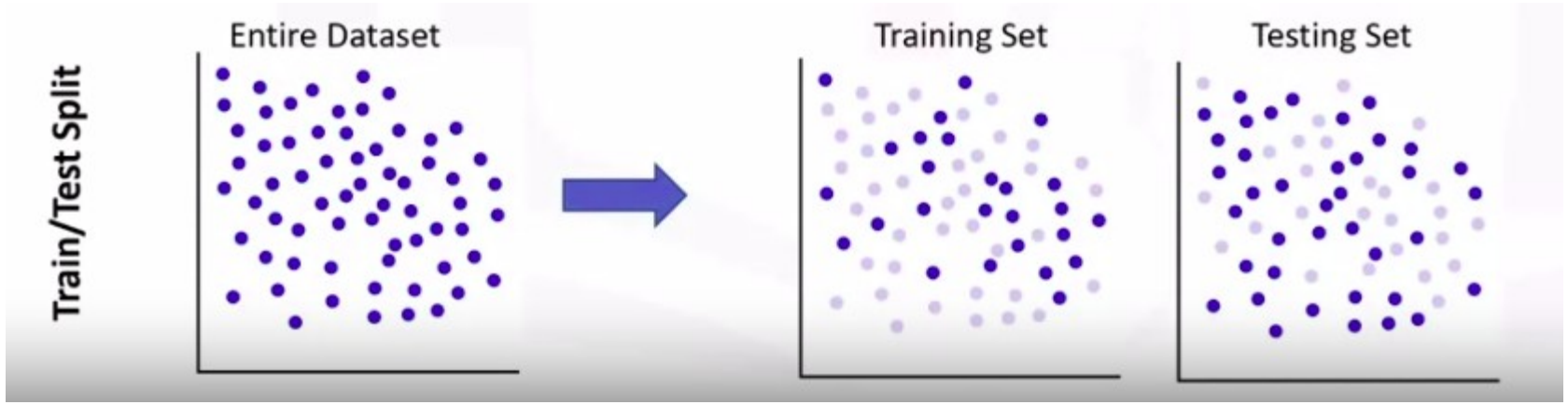


CITADEL
Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Développement

Regression – Model Evaluation



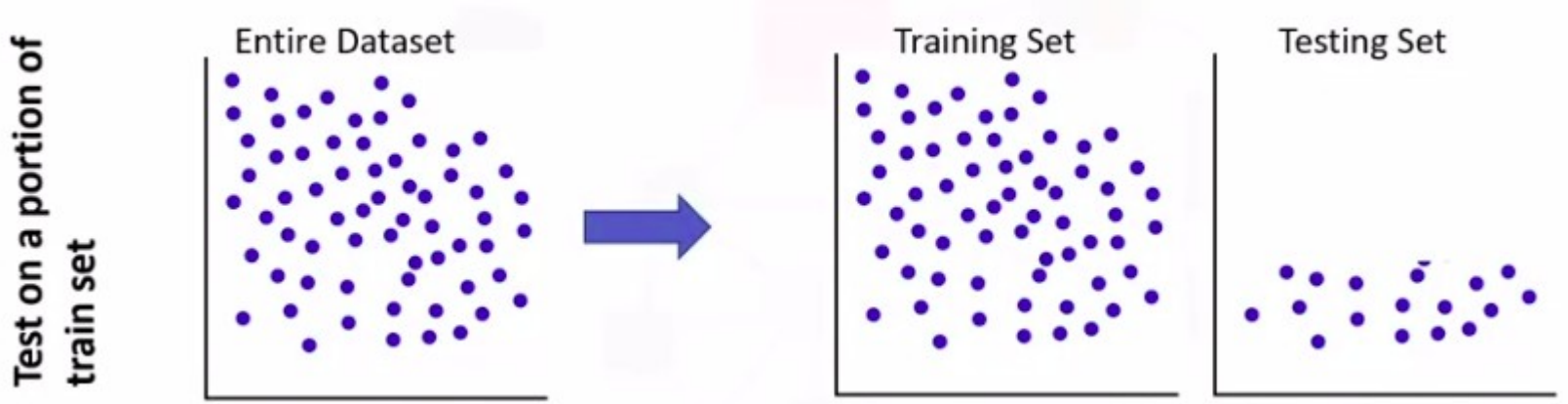
- Test-set is a portion of the train-set
- High "training accuracy"
- Low "out-of-sample accuracy"



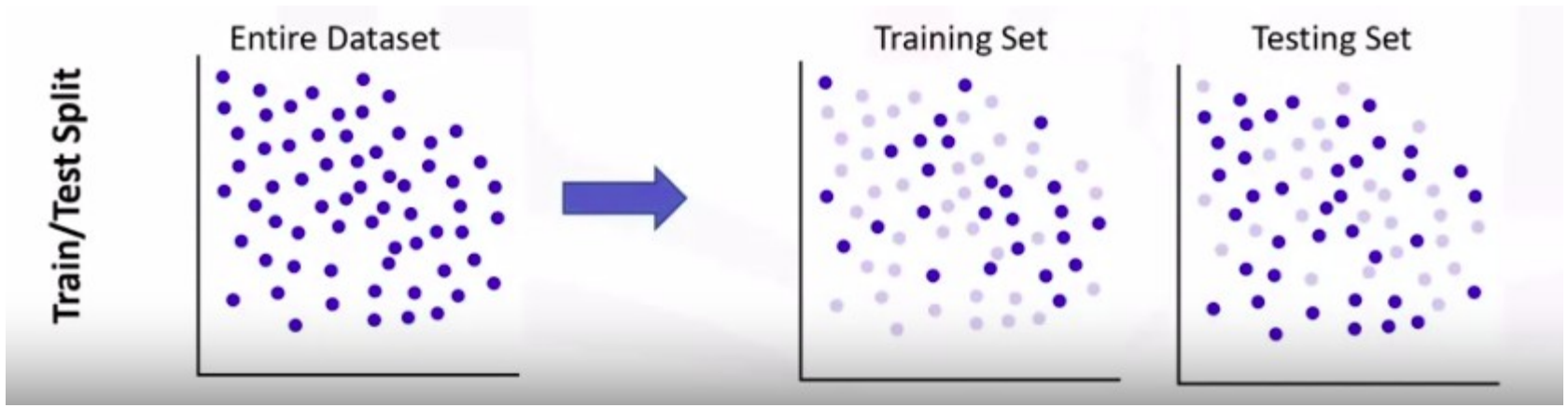
- Mutually exclusive
- More accurate evaluation on out-of-sample accuracy

CITADEL
Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Développement

Regression – Model Evaluation



- Test-set is a portion of the train-set
- High "training accuracy"
- Low "out-of-sample accuracy"

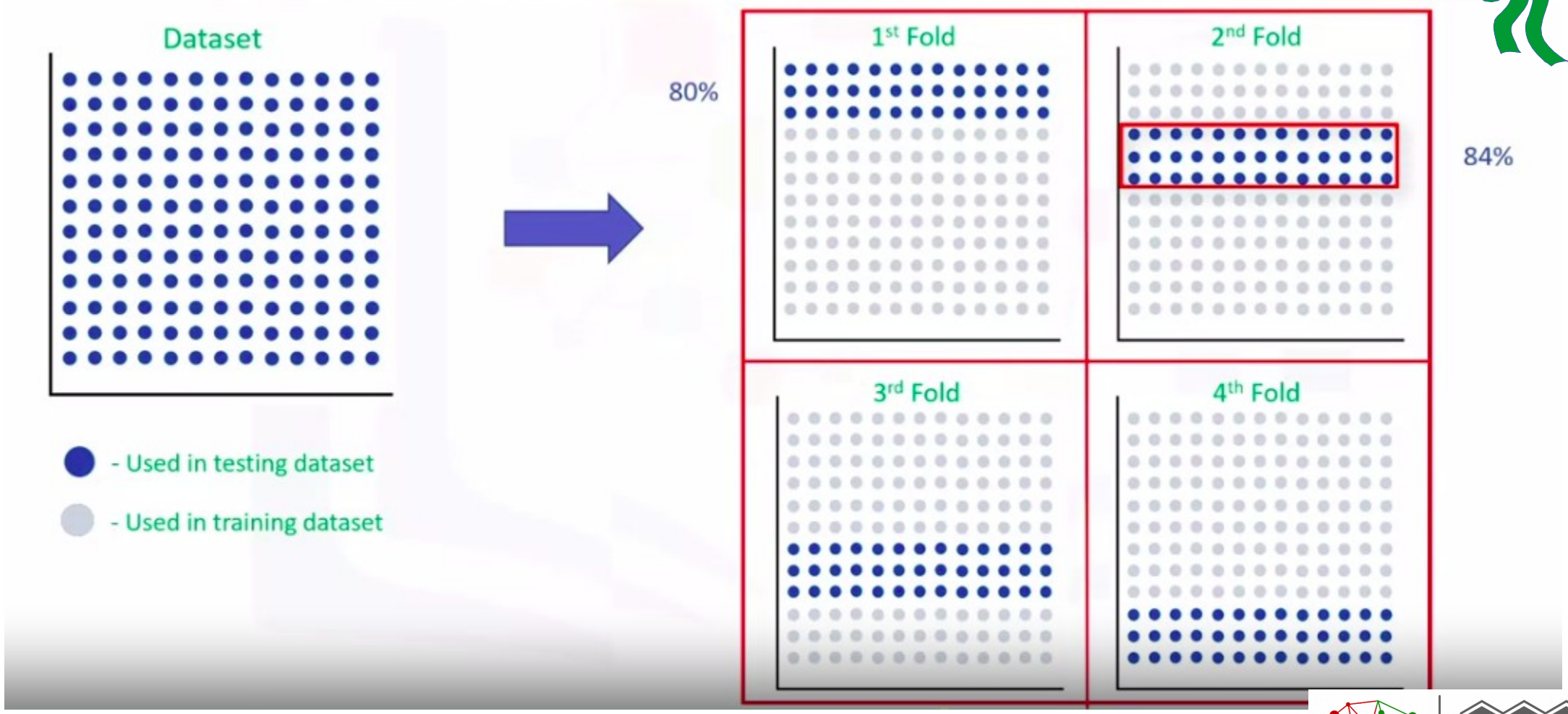


- Mutually exclusive
- More accurate evaluation on out-of-sample accuracy
- Highly dependent on which datasets the data is trained and tested

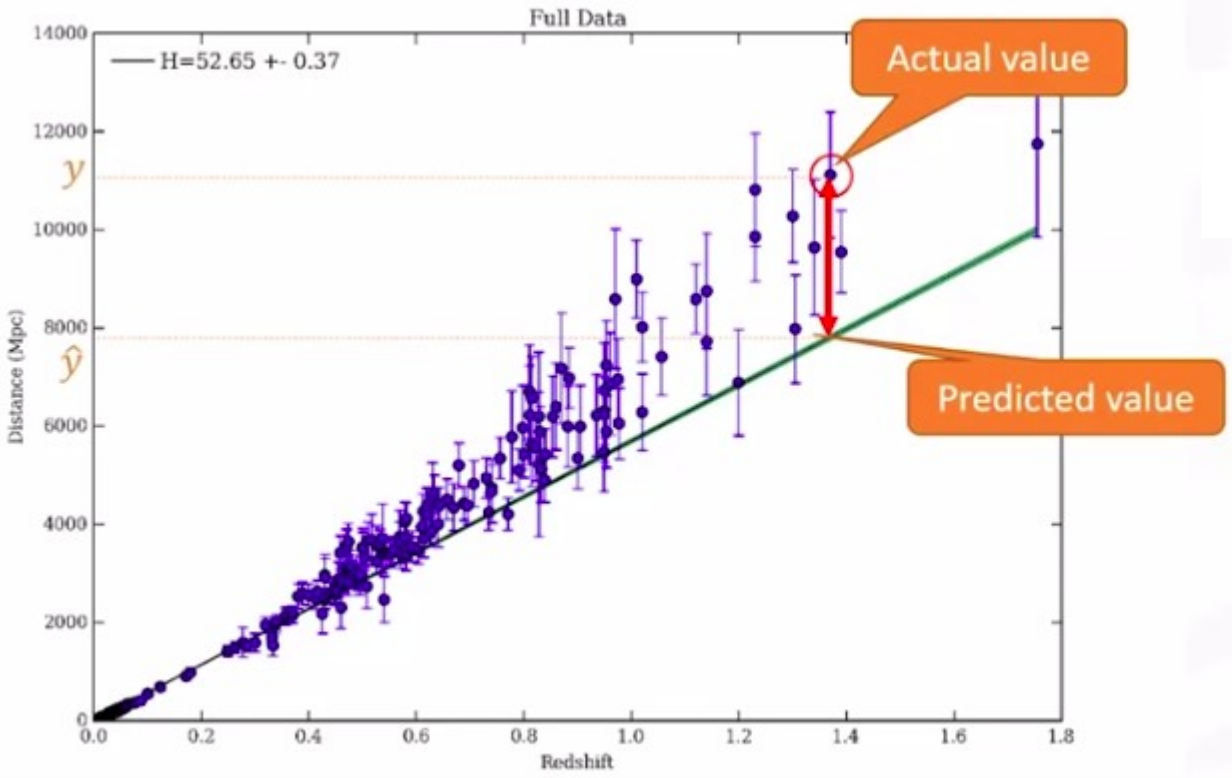
Forte Dependance

CITADEL
Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Développement

Regression – Model Evaluation



Regression – Evaluation Metrics



Error: measure of how far the data is from the fitted regression line.

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

$$R^2 = 1 - RSE$$

Inclusion