

EPISEN – ING3. SI

Machine Learning



Abdallah EL HIDALI

Tech Lead Sita For Aircraft
abdallah.el-hidali@sit.aero

EPISEN

2024/2025

VIII. Les métriques d'évaluation des modèles

Introduction

- En machine learning, l'objectif est de **résoudre des problèmes à partir de données**.
- **Plusieurs outils sont disponibles**, comme les modèles de régression, de classification et de boosting ...
- **Pour évaluer** lequel de ces modèles répond le mieux à nos besoins, il existe des méthodes spécifiques d'évaluation de performance.
- Ces outils permettent de comparer l'efficacité des modèles.
- Ils aident à **choisir le modèle qui offre les meilleures performances** pour résoudre notre problème.



PROBLEM



TOOLS

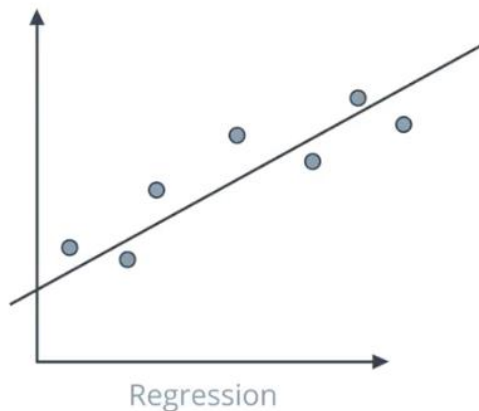


MEASUREMENT
TOOLS

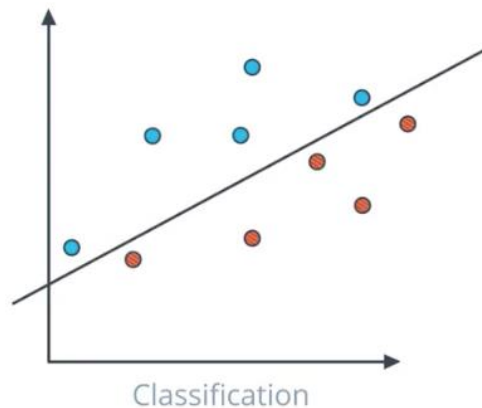
Test des modèles

REGRESSION AND CLASSIFICATION

Regression returns a numeric value



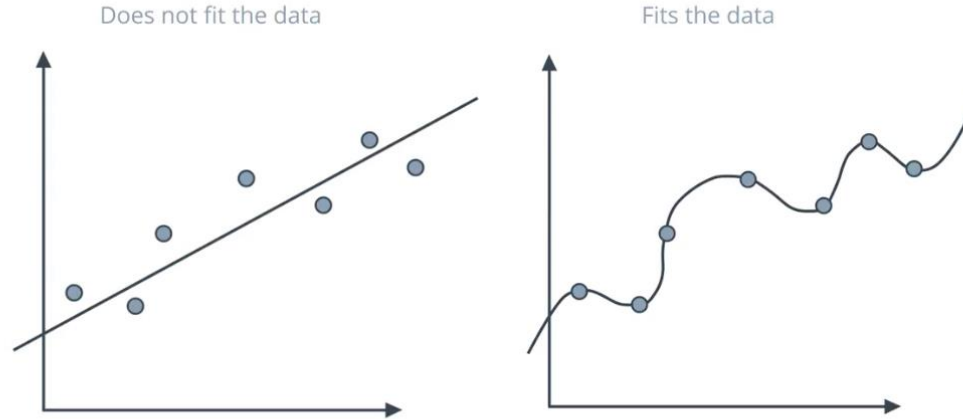
Classification returns a state



- Jusqu'à présent, nous avons exploré les modèles de régression et de classification.
- Les modèles de régression servent à prédire des valeurs numériques
- Les modèles de classification permettent de prédire des catégories ou états, tels que oui ou non, 0 ou 1, chat ou chien, etc.
- Comment mesurer la performance de nos modèles ?

Test des modèles

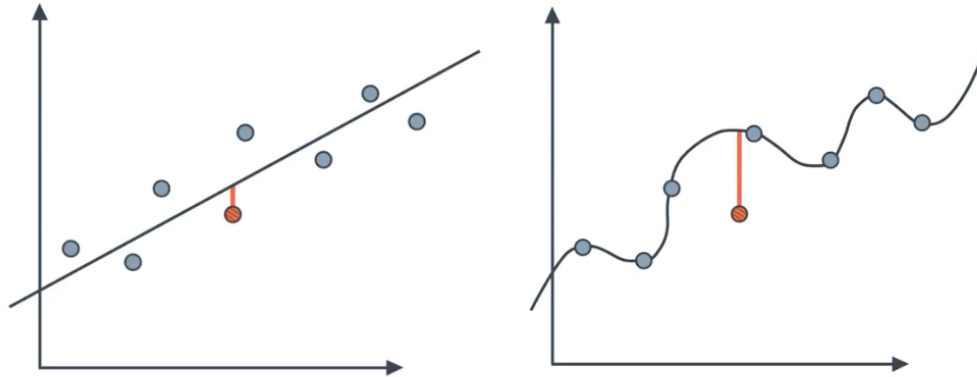
WHICH MODEL IS BETTER?



- Dans cet exemple, nous faisons face à un problème de régression.
- Nous choisissons deux modèles pour représenter la relation entre les données :
 - Le premier modèle est une régression linéaire.
 - Le deuxième est une courbe.
- Lequel des modèles est meilleur

Test des modèles

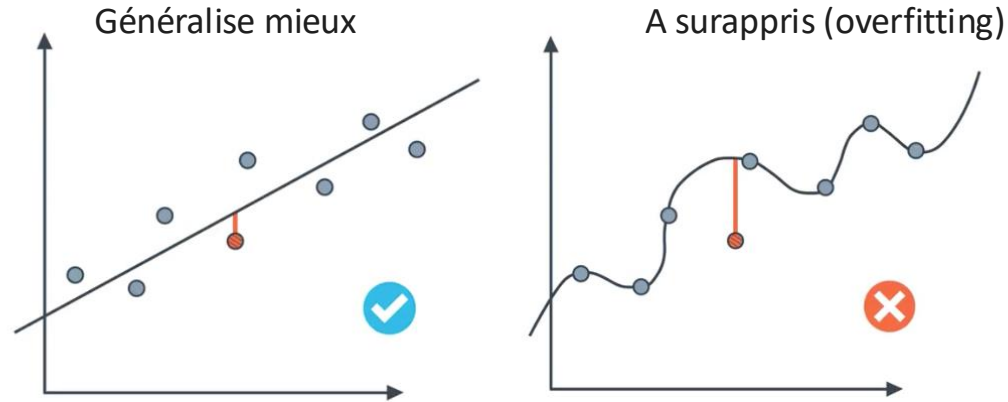
WHICH MODEL IS BETTER?



- Prenons un nouveau point (rouge) pour évaluer l'erreur de ces deux modèles.
- Il semble que le modèle de régression présente moins d'erreur que la courbe.

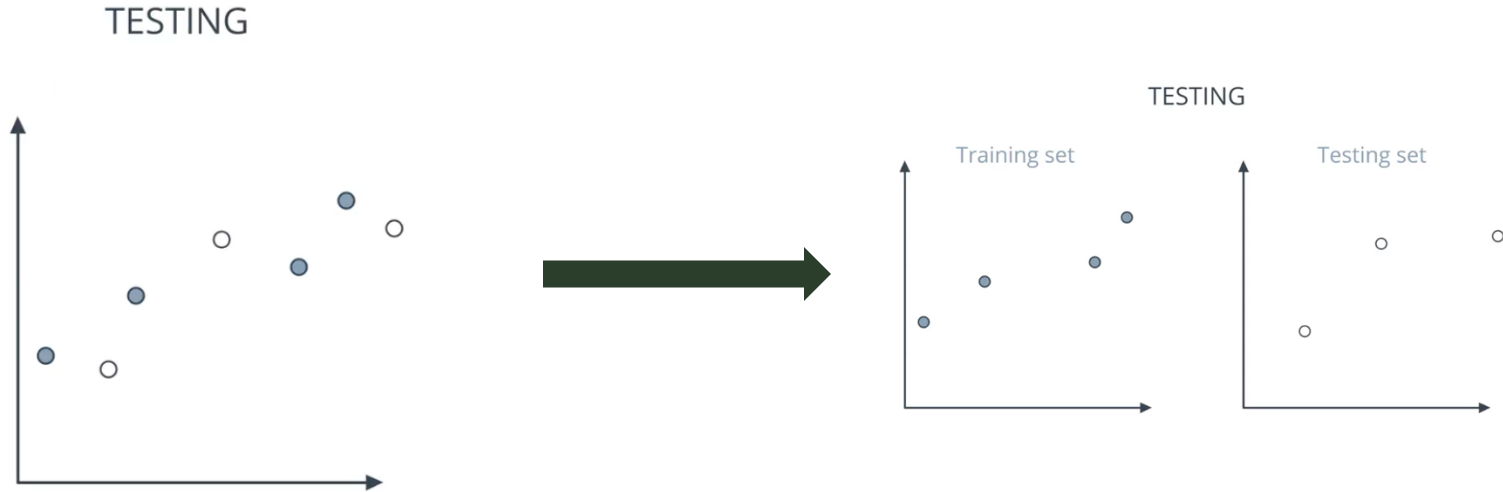
Test des modèles

WHICH MODEL IS BETTER?



Nous aborderons plus tard dans le cours la notion de surapprentissage (overfitting).
La question maintenant est : comment trouver un modèle qui généralise mieux ?

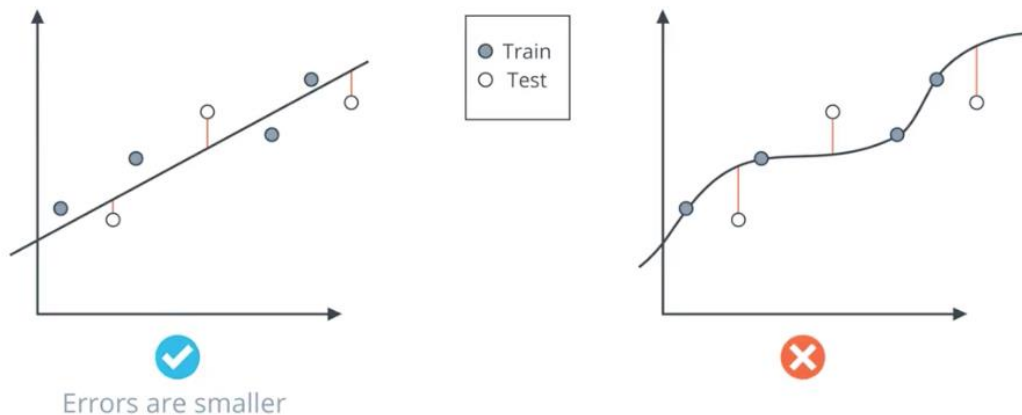
Test des modèles



- On introduit le concept de test.
- On divise les données en deux ensembles : le **training set** et le **testing set**.
- On utilise le **training set** pour entraîner notre modèle.
- On évalue ensuite le modèle sur le **testing set**.

Test des modèles

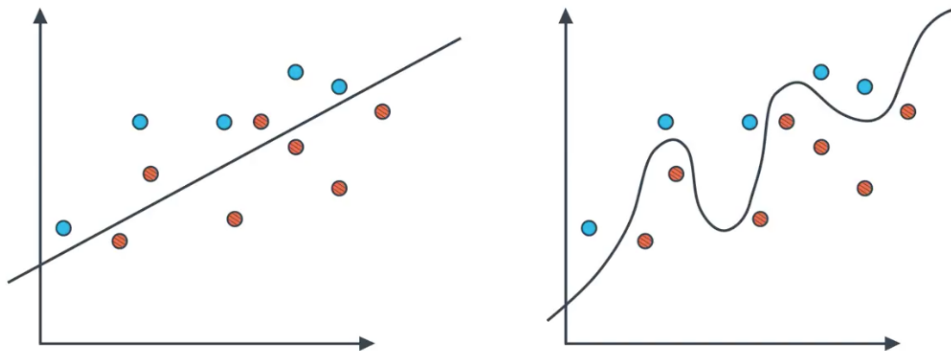
TESTING



- Dans cet exemple, le modèle de régression fait le moins d'erreur sur les données de test.
- On conclut donc que c'est le meilleur modèle.

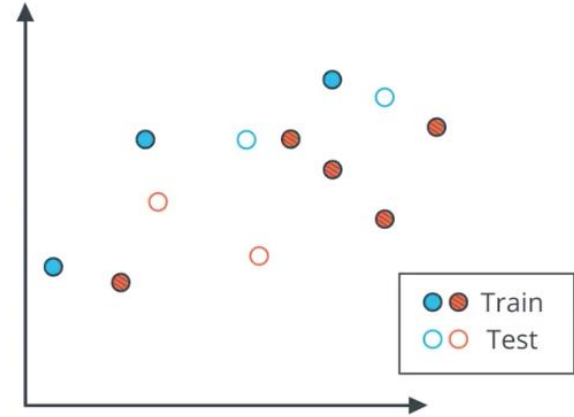
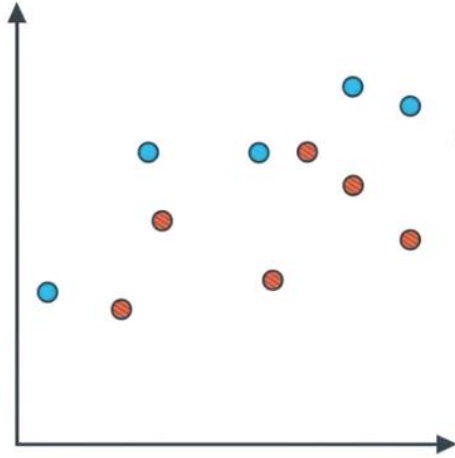
Test des modèles

WHICH MODEL IS BETTER?

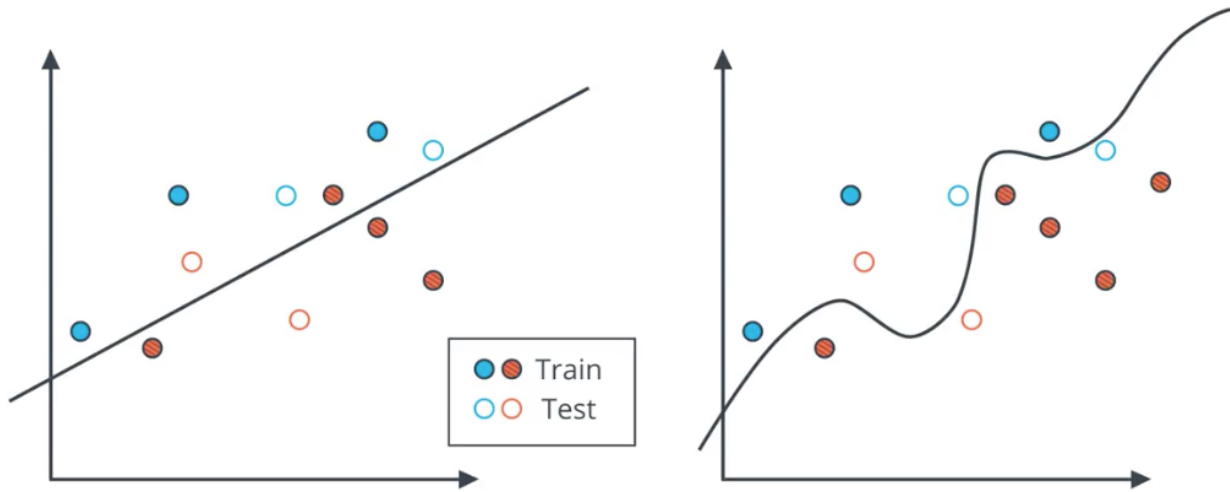


Le même raisonnement s'applique aux modèles de classification.

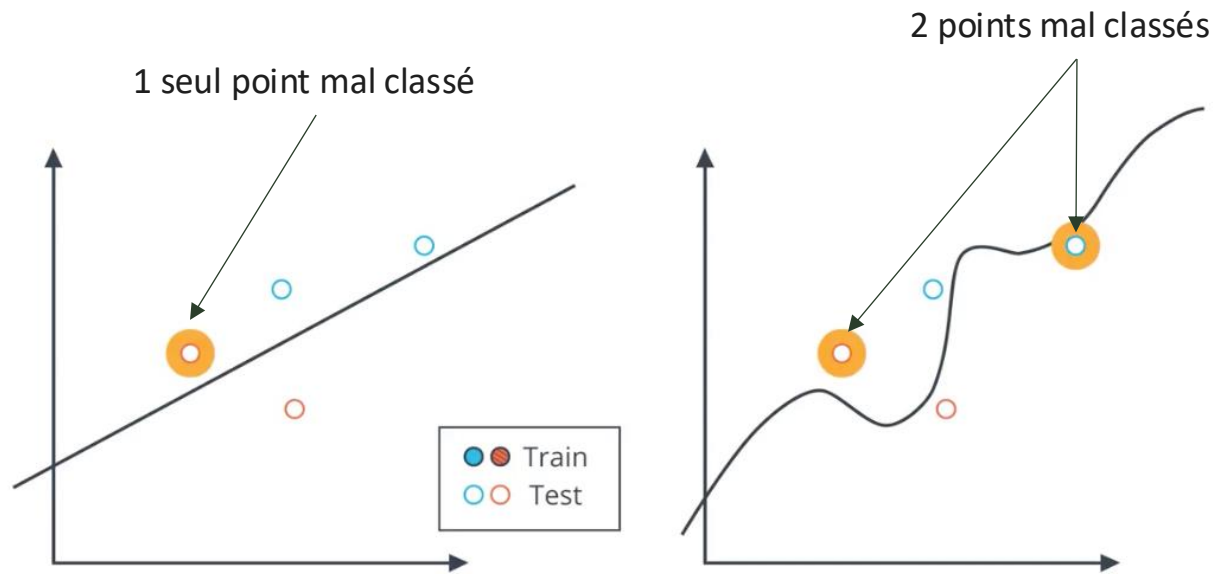
Test des modèles



Test des modèles



Test des modèles

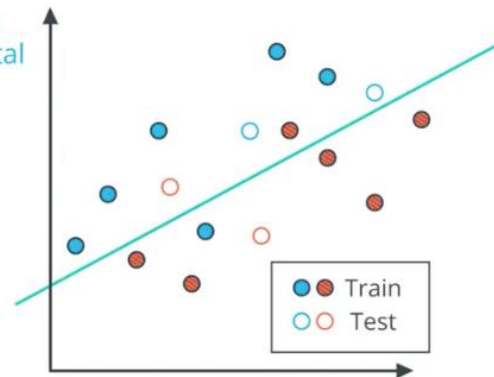
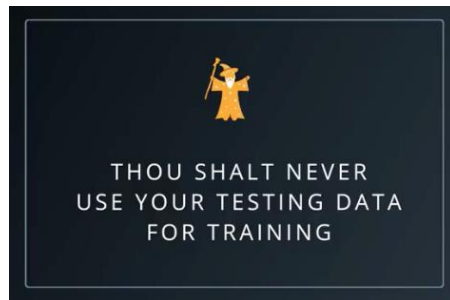


Test des modèles

TESTING IN SKLEARN

```
from sklearn.model_selection import train_test_split  
  
X_train, X_test, y_train, y_test =  
train_test_split(X,  
                y,  
                test_size = 0.25)
```

4 test
16 total



Exercise: <https://github.com/elhidali/EPISEN-2024/>

La Matrice de Confusion:

Comment le modèle se comporte-t-il ?

Nous allons étudier deux modèles



MEDICAL MODEL



HEALTHY



SICK



SPAM CLASSIFIER MODEL



NOT SPAM







SPAM

La Matrice de Confusion:



MEDICAL MODEL

	Diagnosed Sick	Diagnosed Healthy
Sick	 True Positive	 False Negative
Healthy	 False Positive	 True Negative

La Matrice de Confusion:

○ CONFUSION MATRIX



10, 000 PATIENTS

PATIENTS





DIAGNOSIS

	Diagnosed Sick	Diagnosed Healthy
Sick	1000 True Positives	200 False Negatives
Healthy	800 False Positives	8000 True Negatives

La Matrice de Confusion:



SPAM CLASSIFIER MODEL

	Sent to Spam Folder	Sent to Inbox
Spam	 True Positive	 False Negative
Not Spam	 False Positive	 True Negative

La Matrice de Confusion:

○ CONFUSION MATRIX



1000 EMAILS

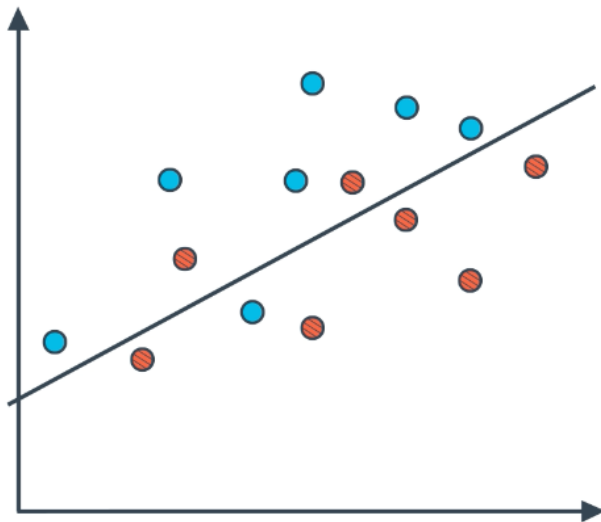
EMAIL

SPAM

	SPAM	
	Spam Folder	Inbox
Spam	100 True Positives	170 False Negatives
Not Spam	30 False Positives	700 True Negatives

La Matrice de Confusion:

○ CONFUSION MATRIX

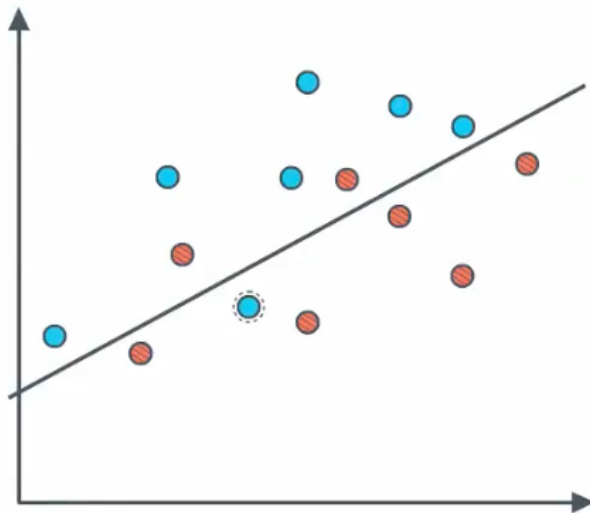


	Guessed Positive	Guessed Negative
Positive	? True Positives	? False Negatives
Negative	? False Positives	? True Negatives

Dans cet exemple, les points bleus sont étiquetés comme positifs et les points rouges comme négatifs. De plus, les points au-dessus de la ligne sont prédits (supposés) comme positifs, tandis que les points en dessous de la ligne sont prédits comme négatifs.

La Matrice de Confusion:

○ CONFUSION MATRIX



	Guessed Positive	Guessed Negative
Positive	6 True Positives	1 False Negatives
Negative	2 False Positives	5 True Negatives

L'exactitude (Accuracy)


L'exactitude (accuracy) répond à la question, parmi tous les patients, combien ont été bien classés ?

	Diagnosed sick	Diagnosed healthy
Sick	1,000	200
Healthy	800	8,000

$$\text{Accuracy} = \frac{1,000 + 8,000}{10,000} = 90\%$$

```
from sklearn.metrics import accuracy_score  
accuracy_score(y_true, y_pred)
```

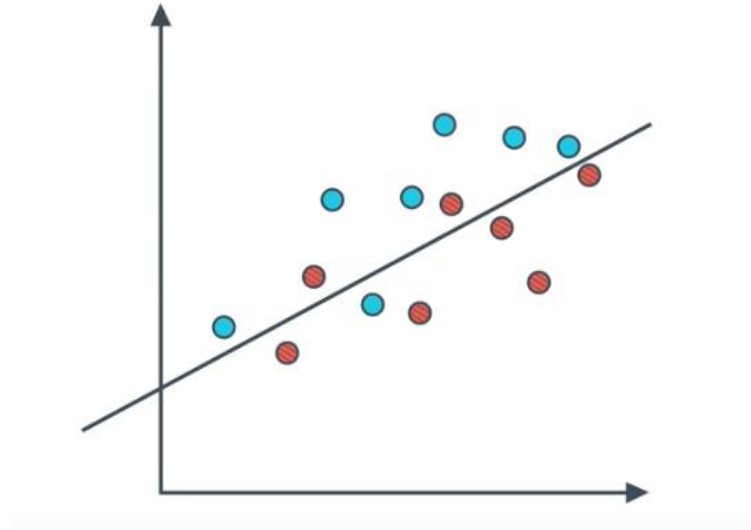
L' exactitude (Accuracy)

	Spam folder	Inbox
Spam	100	170
Not spam	30	700

$$\text{Accuracy} = \frac{100 + 700}{1,000} = 80\%$$

L'exactitude (Accuracy)

Quelle est l'exactitude (accuracy) du modèle suivant



La Précision (Accuracy)

Quelle est l'exactitude (accuracy) du modèle suivant

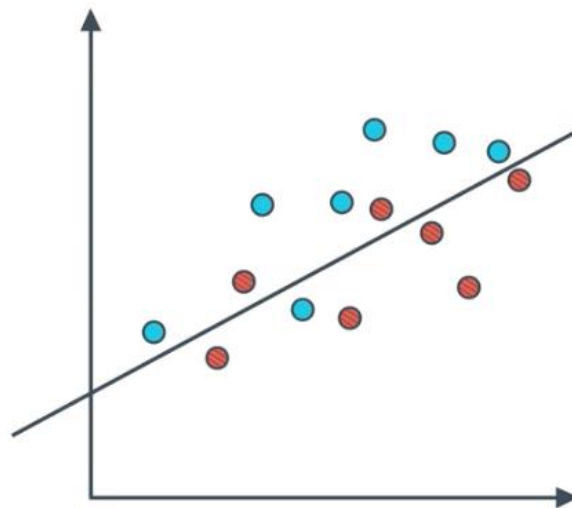
	Guessed Positive	Guessed Negative
Positive	6 True Positives	1 False Negatives
Negative	2 False Positives	5 True Negatives

$$\text{Accuracy} = \frac{\text{Correctly classified points}}{\text{All points}}$$

$$= \frac{11}{11 + 3}$$

$$= \frac{11}{14}$$

$$= 78.57\%$$



Les limites de la métrique « accuracy »

◦ CREDIT CARD FRAUD



284,335



472

Peut-on proposer un modèle avec une exactitude $> 99\%$?

Les limites de la métrique exactitude (accuracy)

◦ CREDIT CARD FRAUD






MODEL: ALL TRANSACTIONS ARE GOOD.

$$\text{ACCURACY} = \frac{284,335}{284,887} = 99.83\%$$




Bien que ce modèle affiche une grande exactitude, il n'identifie aucune transaction frauduleuse.

Les faux négatifs et les faux positifs

Dans le contexte médical, qu'est-ce qui est plus problématique : un faux positif ou un faux négatif ?




Diagnosis			
Patients		DIAGNOSED SICK	DIAGNOSED HEALTHY
	SICK		 FALSE NEGATIVE
	HEALTHY	 FALSE POSITIVE	

Dans le cas d'un détecteur de spam, lequel est plus problématique : un faux positif ou un faux négatif ?

Folder			
Emails		SENT TO SPAM	SENT TO INBOX
	SPAM		 FALSE NEGATIVE
	NOT SPAM	 FALSE POSITIVE	




Les faux négatifs et les faux positifs

Dans le contexte médical, qu'est-ce qui est plus problématique : un faux positif ou un faux négatif ?

Diagnosis			
Patients		DIAGNOSED SICK	DIAGNOSED HEALTHY
	SICK		 FALSE NEGATIVE
	HEALTHY	 FALSE POSITIVE	

Les faux négatifs

Dans le cas d'un détecteur de spam, lequel est plus problématique : un faux positif ou un faux négatif ?

Folder			
Emails		SENT TO SPAM	SENT TO INBOX
	SPAM		 FALSE NEGATIVE
	NOT SPAM	 FALSE POSITIVE	

Les faux positifs

Précision (Precision) et Rappel (Recall)

- SOLUTION: FALSE POSITIVES AND NEGATIVES



Medical Model

FALSE POSITIVES OK

FALSE NEGATIVES NOT OK

OK IF NOT ALL ARE SICK
FIND ALL THE SICK PEOPLE

HIGH RECALL



Spam Detector

FALSE POSITIVES NOT OK



FALSE NEGATIVES OK

DON'T NECESSARILY NEED
TO FIND ALL THE SPAM
BETTER BE SPAM

HIGH PRECISION

La Précision

○ PRECISION

		DIAGNOSIS	
PATIENTS		Diagnosed Sick	Diagnosed Healthy
	Sick	1000	200 
	Healthy	800	9000

↖
Total des patients diagnostiqués positifs

$$\text{PRECISION} = \frac{1,000}{1,000 + 800} = 55.6\%$$

La Précision

- PRECISION

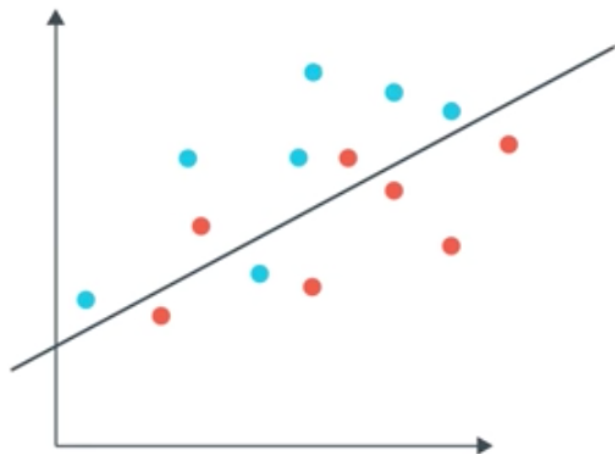
FOLDER

EMAIL		Sent to Spam Folder	Sent to Inbox
	Spam	100	170
	Not Spam	30 	700

$$\text{PRECISION} = \frac{100}{100 + 30} = 76.9\%$$

La Précision

PRECISION



OUT OF THE POINTS WE HAVE
PREDICTED TO BE POSITIVE,
HOW MANY ARE CORRECT?


Dans cet exemple, les points bleus sont étiquetés positifs, et les points rouges sont étiquetés négatifs. De plus, les points au-dessus de la ligne sont prédits comme positifs, et les points en dessous de la ligne sont prédits comme négatifs.

Quelle est la précision de ce modèle ?

	Guessed Positive	Guessed Negative
Positive	6 True Positives	1 False Negatives
Negative	2 False Positives	5 True Negatives

Le rappel

◦ RECALL



PATIENTS	DIAGNOSIS	
		
	Diagnosed Sick	Diagnosed Healthy
Sick	1000	200 
Healthy	800	8000

OUT OF THE SICK PATIENTS,
HOW MANY DID WE CORRECTLY
DIAGNOSE AS SICK?

$$\text{RECALL} = \frac{1,000}{1,000 + 200} = 83.3\%$$

Le rappel

◦ RECALL

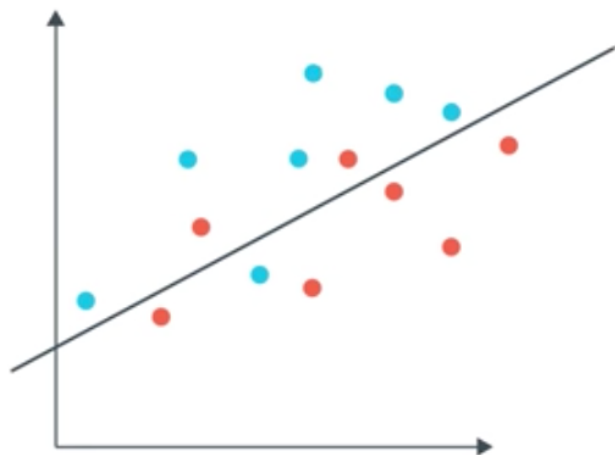
		FOLDER	
		 Sent to Spam Folder	Sent to Inbox
EMAIL	Spam	100	170
	Not Spam	30 	700

OUT OF ALL THE SPAM E-MAILS,
HOW MANY WERE CORRECTLY
SENT TO THE SPAM FOLDER?

$$\text{Recall} = \frac{100}{100 + 170} = 37\%$$

Le rappel

◦ PRECISION



OUT OF THE POINTS WE HAVE
PREDICTED TO BE POSITIVE,
HOW MANY ARE CORRECT?

Dans cet exemple, les points bleus sont étiquetés positifs, et les points rouges sont étiquetés négatifs. De plus, les points au-dessus de la ligne sont prédits comme positifs, et les points en dessous de la ligne sont prédits comme négatifs.

Quelle est le rappel de ce modèle ?

	Guessed Positive	Guessed Negative
Positive	6 True Positives	1 False Negatives
Negative	2 False Positives	5 True Negatives

Limite de la précision et du rappel

◦ PRECISION AND RECALL



MEDICAL MODEL
PRECISION: 55.7%
RECALL: 83.3%

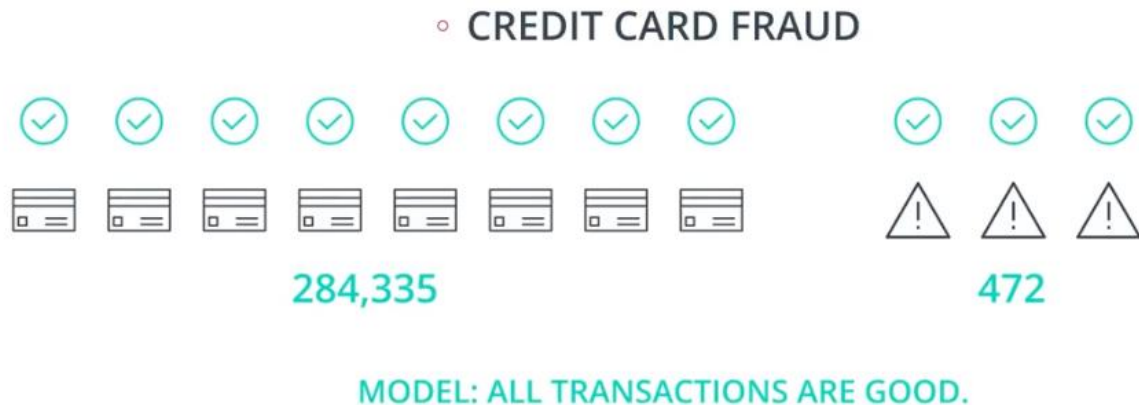
ONE SCORE?



SPAM DETECTOR
PRECISION: 76.9%
RECALL: 37%

Peut-on avoir une seule métrique qui combine la précision et le rappel en même temps ?

La moyenne entre la précision et le rappel



PRECISION = 100%

AVERAGE = 50%

RECALL = 0%

**La moyenne ne traduit pas le fait que
le modèle n'est pas bon en rappel**

La moyenne entre la précision et le rappel

◦ CREDIT CARD FRAUD



MODEL: ALL TRANSACTIONS ARE FRAUDULENT.

$$\text{PRECISION} = 472/284,807 = 0.16\%$$

$$\text{RECALL} = 472/472 = 100\%$$

$$\text{AVERAGE} = 50.08\%$$

**La moyenne ne traduit pas le fait que
le modèle n'est pas bon en précision**

La moyenne harmonique / F1 score

◦ HARMONIC MEAN

	Y		
ARITHMETIC MEAN =	$\frac{x+y}{2}$	PRECISION = 1 RECALL = 0 AVERAGE = 0.5 HARMONIC MEAN = 0	PRECISION = 0.2 RECALL = 0.8 AVERAGE = 0.5 HARMONIC MEAN = 0.32
HARMONIC MEAN =	$\frac{2xy}{x+y}$		
	X	ARITHMETIC MEAN (PRECISION, RECALL)	
		F ₁ SCORE = HARMONIC MEAN (PRECISION, RECALL)	

$$F1 \text{ Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 Score

◦ CREDIT CARD FRAUD



MODEL: ALL TRANSACTIONS ARE GOOD.

PRECISION = 100%

F_1 SCORE = 0

RECALL = 0%

F-beta Score

- QUIZ: F_{β} SCORE

$$F_1 \text{ SCORE} = 2 \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F_{\beta} \text{ SCORE} = (1 + \beta^2) \beta^2 \frac{\text{Precision} * \text{Recall}}{2 * \text{Precision} + \text{Recall}}$$



PRECISION

$F_{0.5}$ SCORE

F_1 SCORE

F_2 SCORE



RECALL

F-beta Score

◦ QUIZ: F_{β} SCORE



PRECISION

$F_{0.5}$ SCORE

F_1 SCORE

F_2 SCORE



RECALL

- Dans la détection des fraudes, on peut chercher à maximiser le rappel en choisissant une valeur élevée pour β .
- Cela risque d'augmenter le nombre de faux positifs et d'envoyer de nombreuses notifications aux clients pour des transactions erronément signalées comme frauduleuses.
- Un β plus faible favorisera la précision, réduisant ainsi le risque de manquer une transaction frauduleuse. Le choix du β n'est pas une science exacte et nécessite plusieurs itérations.

Récap

Predicted	Actual	
	Spam (Positive)	Not Spam (Negative)
	Spam (Positive)	Not Spam (Negative)
Spam (Positive)	True Positive (TP)	False Positive (FP)
Not Spam (Negative)	False Negative (FN)	True Negative (TN)

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Predicted	Actual	
	Spam (Positive)	Not Spam (Negative)
	Spam (Positive)	Not Spam (Negative)
Spam (Positive)	True Positive (TP)	False Positive (FP)
Not Spam (Negative)	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \frac{\text{True Positives}}{\text{TP} + \text{FP}}$$

Predicted	Actual	
	Spam (Positive)	Not Spam (Negative)
	Spam (Positive)	Not Spam (Negative)
Spam (Positive)	True Positive (TP)	False Positive (FP)
Not Spam (Negative)	False Negative (FN)	True Negative (TN)

$$\text{Recall} = \frac{\text{True Positives}}{\text{TP} + \text{FN}}$$

F-Beta Score

$$F_{\beta} = (1 + \beta^2) \frac{\text{Precision} * \text{Recall}}{(\beta^2 * \text{Precision}) + \text{Recall}}$$

Les métriques de régression

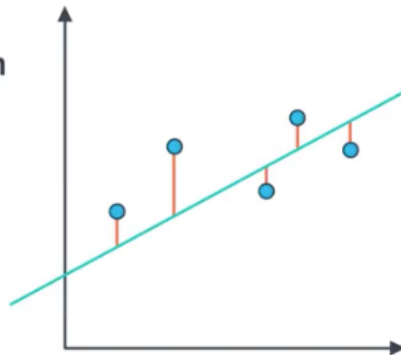
- MEAN ABSOLUTE ERROR IN SKLEARN

```
from sklearn.metrics import mean_absolute_error
from sklearn.linear_model import LinearRegression

classifier = LinearRegression()
classifier.fit(X,y)

guesses = classifier.predict(X)

error = mean_absolute_error(y, guesses)
```



Les métriques de régression

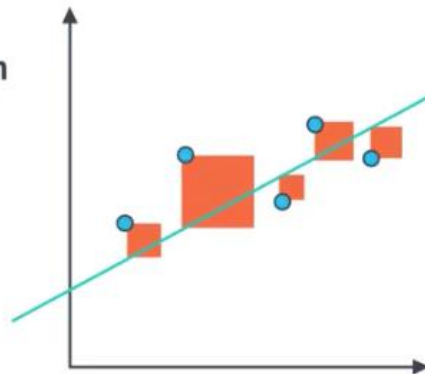
- MEAN SQUARED ERROR IN SKLEARN

```
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LinearRegression

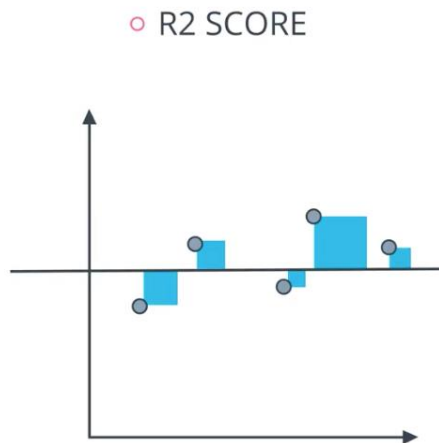
classifier = LinearRegression()
classifier.fit(X,y)

guesses = classifier.predict(X)

error = mean_squared_error(y, guesses)
```



Les métriques de régression

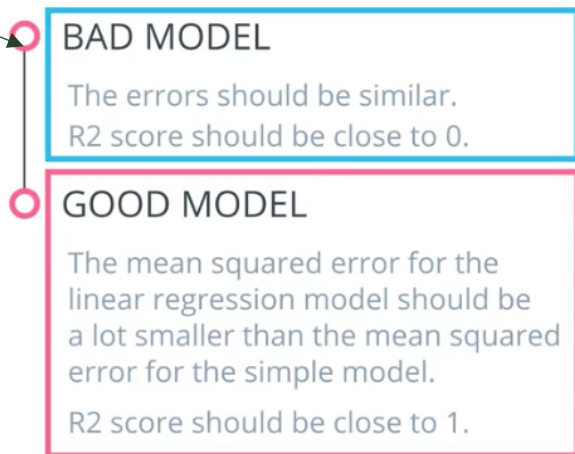


- Le score R^2 permet de comparer l'erreur du modèle à celle du modèle le plus simple, qui est basé sur la moyenne.

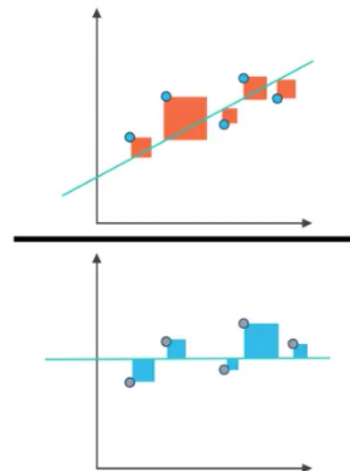
Les métriques de régression

La performance du modèle est similaire à celle du modèle le plus simple, qui est basé sur la moyenne.

○ R2 SCORE



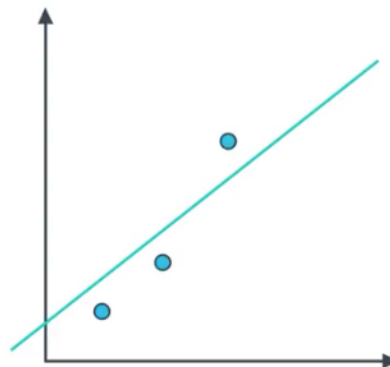
$$R^2 = 1 -$$



Les métriques de régression

○ R2 SCORE IN SKLEARN

```
from sklearn.metrics import r2_score  
  
y_true = [1, 2, 4]  
y_pred = [1.3, 2.5, 3.7]  
  
r2_score(y_true, y_pred)
```



Exercice: <https://github.com/elhidali/EPISEN-2024/>