

# Majeure Machine Learning

Data  
Preparation

# Contenu



- Qu'est-ce que le nettoyage des données
- Pourquoi nettoyer les données
- Pourquoi le feature engineering
- Pourquoi la Cross Validation (CV)
- Pratique de jupyter / Dataiku

# Ce que vous devrez savoir faire



- Définir les termes techniques
- Comprendre l'enjeu de la data preparation
- Comprendre l'impact de la connaissance métier
- Expliquer l'intérêt de la CV
- Avoir une bonne connaissance de jupyter / Dataiku

# Mise en Pratique

# Approche technique



# Approche intuitive



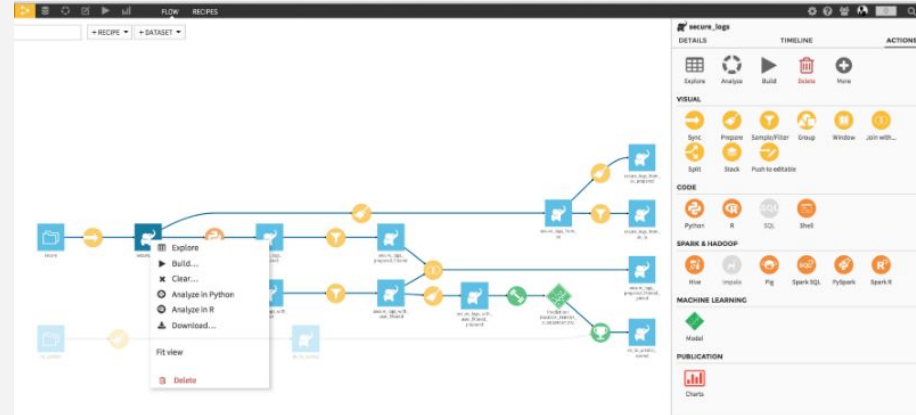


data  
iku

# Dataiku DSS

Dataiku dss est une plateforme d'analyse de données permettant d'effectuer toutes les étapes d'un projet de Machine Learning sans développement (au clique)

# Dataiku dss Studio



# Data Preparation

Data Cleaning | Feature engineering

# Vocabulaire

Data Cleaning | Data filtering | Data enrichissement | Feature engineering  
| Data preparation | Data discretization | Data snooping | Data  
Visualization | Data exploration | Data preprocessing | Data mining | Data  
scientist | Data analyst | Data engineer | Data transformation



# Pourquoi le Data Cleaning ?

Apprendre à partir d'une base de données contenant des valeurs fausses ne peut engendrer que des prédictions fausses.

L'objectif du data cleaning est :

De n'avoir que des  
données  
exploitables



De n'avoir que des  
données exactes



De n'avoir aucune  
ligne incomplète



De conserver un  
maximum de  
données



# Data Cleaning

Ensemble des techniques qui permettent de transformer, enrichir, modifier les données afin d'optimiser les performances du modèle d'apprentissage.

Prénom	Nom	âge	Ville	Profession	...	Nationalité
Yann	Lecun	53	New York	Chercheur	...	Français
Hugo	Larochelle	? 39	<del>Canada</del>	Chercheur	...	Canadien
Andrew	Ng	Etats-Unis	↔ 42	Chercheur	...	Anglais/Chinois

Data Cleaning

Data enrichissement

# Pourquoi le feature engineering ?

Augmenter la performance de son modèle en bénéficiant du savoir d'un expert métier

# Définition du feature engineering

Le feature engineering est le processus consistant à extraire, sélectionner et créer des caractéristiques pertinentes, informatives et distinctives à partir de données, puis de les exploiter au travers d'un algorithme apprenant.

# Exemple de feature engineering - Date

date	Date - semaine	Date - vacances	Date - weekend	Date - Jour férié
01/03/2018	9	oui	non	non
11 / 11 / 2018	45	non	oui	oui

# Exemple de feature engineering - Age

profession	âge	Majeur ?
étudiant	17	non
étudiant	19	oui
étudiant	20	oui

# Exemple de feature engineering - one hot encoding

profession	Étudiant ?	Chercheur ?
étudiant	1	0
chercheur	0	1
Etudiant / chercheur	1	1

# Exemple de feature engineering - normalisation

âge	Âge - normalisation
17	$(17-17)/(84-17) = 0$
19	$(19-17)/(84-17) = 0.029$
45	$(45-17)/(84-17) = 0.417$
84	$(84-17)/(84-17) = 1$
20	$(20-17)/(84-17) = 0.044$

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$



# Exemple de feature engineering - standardisation

âge	Âge - normalisation
17	$(17-37)/(84-17) = \mathbf{-0.298}$
19	$(19-37)/(84-17) = \mathbf{-0.0268}$
45	$(45-37)/(84-17) = \mathbf{0.119}$
84	$(84-37)/(84-17) = \mathbf{0.701}$
20	$(20-37)/(84-17) = \mathbf{-0.253}$

$$x_{new} = \frac{x - x_{mean}}{x_{max} - x_{min}}$$

# Normalisation

$[0,1]$

Inconvénients : efface les outliers

Efficace pour :

- Image processing
- Neural networks

# Standardisation

$[-1,1]$

Efficace pour :

- PCA
- Classification
- Algorithmes ensemblistes
- etc.

# Exemple de feature engineering - Golden feature

Une entreprise de sport souhaite prédire le nombre d'articles vendus par catégorie pour la semaine suivante. Un réapprovisionnement de 50 articles par catégorie est effectué chaque semaine. Le data scientist dispose des données de ventes des 2 dernières années par semaine.

Semaine	Catégorie	Nbr d'articles en stock	Articles vendus
42	Vélo	232	$S_{41} - s_{42} + 50 = 30$
43	Vélo	167	$232 - 167 + 50 = 115$

inventaire
non
oui

# Pourquoi la Cross Validation ?

La CV permet de valider un modèle basé sur un ensemble de modèles apprenant chacun sur une segmentation de données différentes. Cette approche permet de ne pas être pénalisé lorsque les données sont hétérogènes

Augmenter la  
perception de la  
robustesse du  
modèle



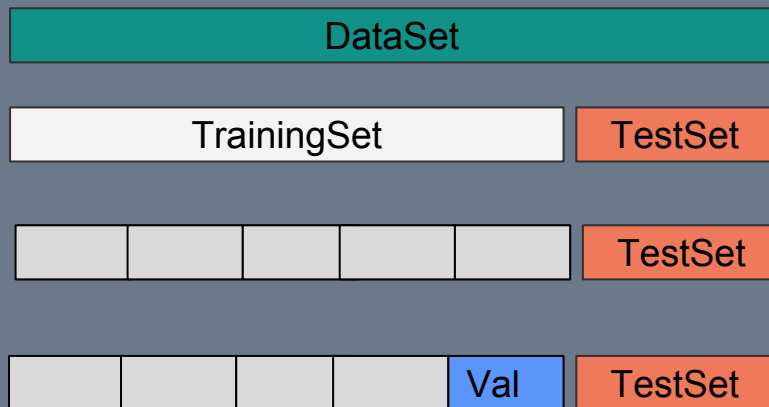
Tester sur  
l'ensemble des  
données



Identifier plus  
facilement si le  
modèle a Overfitté



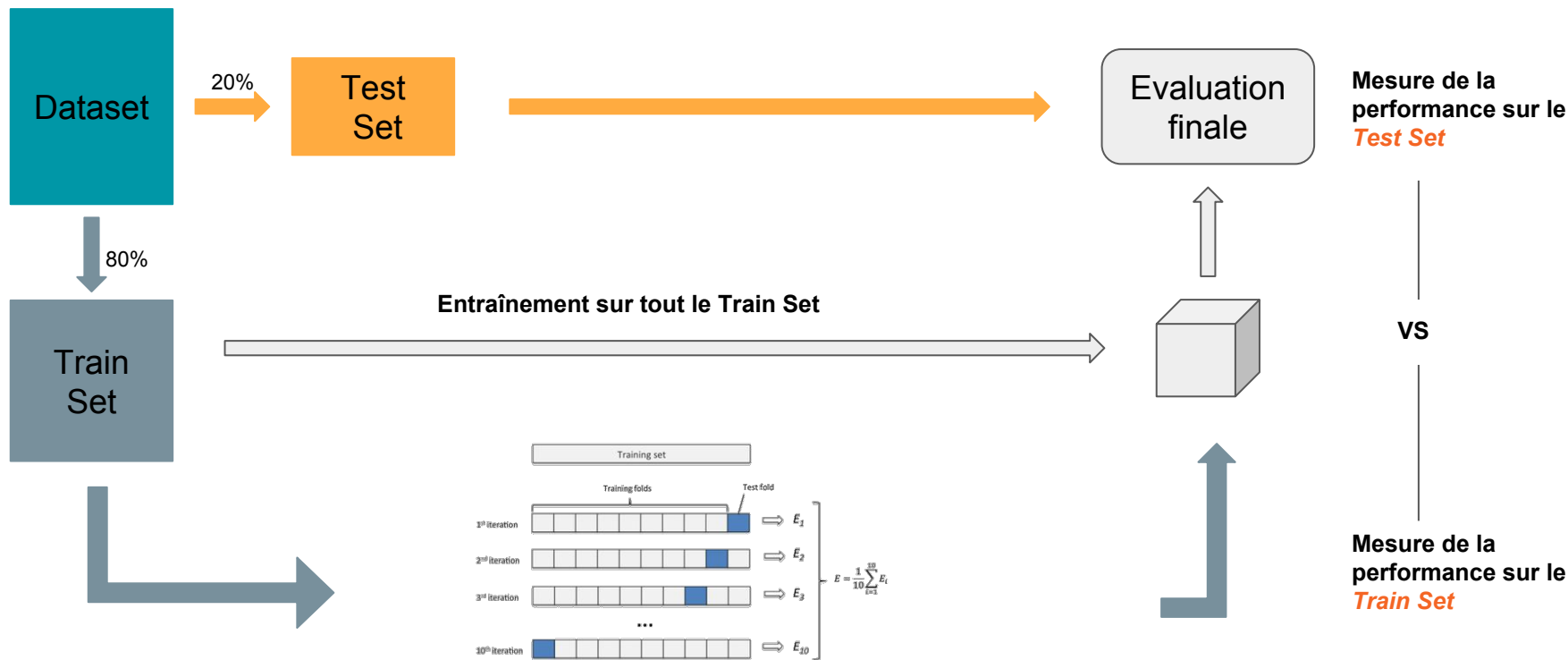
# Définition la Cross Validation 1



# Définition la Cross Validation 2



# Différence entre la performance sur le Train set et le Test set



# Mesure de la performance

# Indicateurs de Classification

## Accuracy

$$\frac{nb\_elements\_bien\_predits}{nb\_total\_elements}$$

## F1score

$$\begin{aligned} precision &= \frac{TP}{TP + FP} \\ recall &= \frac{TP}{TP + FN} \\ F1 &= \frac{2 \times precision \times recall}{precision + recall} \end{aligned}$$

## Confusion Matrix

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)



# Indicateurs de Classification

$\hat{y}$	y
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	0
1	0

Accuracy

$$\frac{8}{10} = 0.8 \Rightarrow 80\%$$

F1score

Précision :  $\frac{8}{8+2} = 0.8$

Sensibilité (Recall) :  $\frac{8}{8+0} = 1$

Confusion Matrix

		Prédit	
		P	N
Vrai	P	8	0
	N	2	0

$$\text{F1score} : \frac{2 \cdot 0.8 \cdot 1}{0.8 + 1} = \frac{1.6}{1.8} \approx 0.88 \Rightarrow 88\%$$

# Calcul de distance des erreurs

Intuition : A quel point la prédiction est loin de la vérité

$$RSS = \sum_{i=1}^n (f(x_i) - y_i)^2$$

Somme des carrés des résidus

$$MSE = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$$

Normalisation avec n (moyenne)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2}$$

On se ramène à l'unité de y avec l'ajout de la racine

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(f(x_i) + 1) - \log(y_i + 1))^2}$$

On rajoute des logs pour ne pas pénaliser les grandes erreurs

# R<sup>2</sup> - Mesure de la corrélation

Intuition : indice de confiance

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

Valeur à prédire

Valeur prédite

Moyenne des mesures à prédire

Mesure de la corrélation entre la prédiction et la valeur réelle (doit être au plus proche de 1)

$$\in ] - \infty, 1]$$



Fin du chapitre 2.2