

An illustration of a philosopher with a long white beard, wearing a purple and white robe, standing on a path and pointing upwards. He is positioned next to a large tree on the left. The background features rolling green hills, a blue sky with birds, and a smaller tree on the right.

Knowledge Discovery and Data Mining process

Nathan Hoche

Definition:

Knowledge Discovery se réfère au processus global de découverte de connaissances utiles à partir de données.

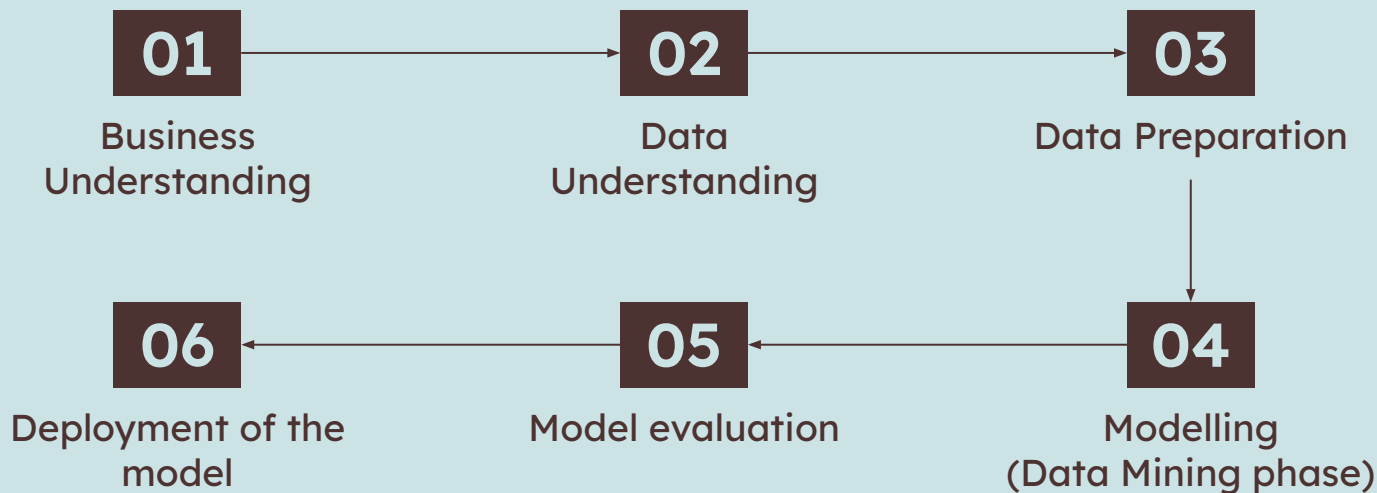
Comprend l'exploration des données, mais aussi plusieurs étapes de préparation des données et de validation des connaissances.

L'exploration de données est l'étape centrale du processus de KDD et se réfère à l'application d'algorithmes pour extraire des modèles des données.



CRISP-DM Process Model

Cross-Industry Standard Process for Data Mining



CRISP-DM Process Model

Cross-Industry Standard Process for Data Mining

01

Business Understanding

Comprendre les objectifs et les exigences de l'entreprise pour les convertir en un problème de data mining

CRISP-DM Process Model

Cross-Industry Standard Process for Data Mining

02

Data Understanding

Comprendre les données à extraire, prendre en compte les questions de qualité des données

CRISP-DM Process Model

Cross-Industry Standard Process for Data Mining

03

Data Preparation

Nettoyage des données, sélection des attributs, transformation des données, ...

En fonction du ou des algorithmes d'exploration de données à utiliser durant l'étape (4)

CRISP-DM Process Model

Cross-Industry Standard Process for Data Mining

04

Modelling

1. Choisir un ou plusieurs algorithmes d'exploration de données à appliquer aux données
2. Ajuster ses (leurs) paramètres
3. construire un modèle de données

CRISP-DM Process Model

Cross-Industry Standard Process for Data Mining

05

Model Evaluation

Interpréter et valider les modèles d'un point de vue commercial

CRISP-DM Process Model

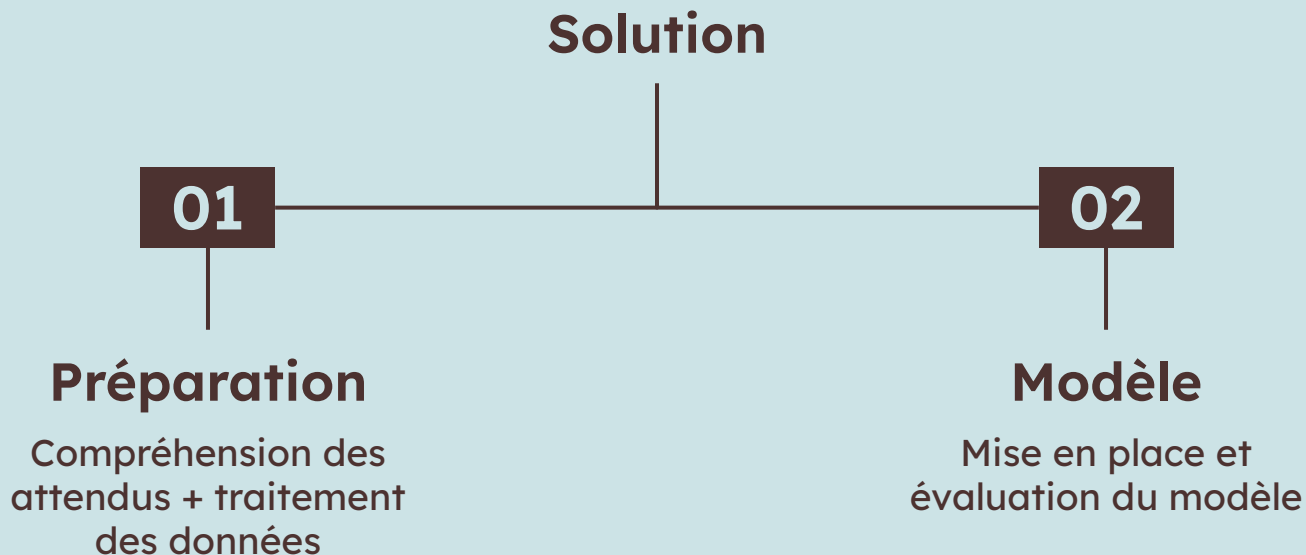
Cross-Industry Standard Process for Data Mining

06

Deployment of the model

Déployer les résultats de l'exploration de données sous la forme d'un plan, surveiller et maintenir le déploiement du plan.

Résumé

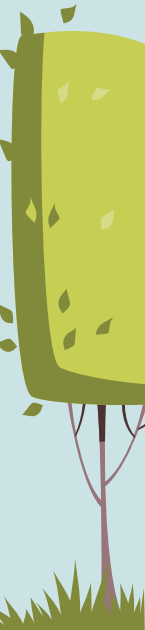


Mais c'est quoi
traiter les données?



Précédemment:


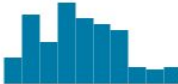
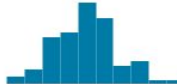

- Nettoyage des données
- Sélection des attributs
- Transformation des données (ex: PCA)





Dataset d'exemple: Iris Dataset

Objectif: Identifier l'espèce d'une plante avec la taille des pétales/sépales

# Id	# SepalLengthCm	# SepalWidthCm	# PetalLengthCm	# PetalWidthCm	Species
SPL-SPW-PTL-PTW(CM)	Length of the sepal (in cm)	Width of the sepal (in cm)	Length of the petal (in cm)	Width of the petal (in cm)	Species name
				3 unique values	
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5.0	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3.0	1.4	0.1	Iris-setosa
14	4.3	3.0	1.1	0.1	Iris-setosa
15	5.8	4.0	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa





1- Nettoyage des données

Qu'est-ce que je dois changer?



# Id	# SepalLengthCm	# SepalWidthCm	# PetalLengthCm	# PetalWidthCm	Species
SPL-SPW-PTL-PTW(CM)	Length of the sepal (in cm)	Width of the sepal (in cm)	Length of the petal (in cm)	Width of the petal (in cm)	Species name
					3 unique values
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5.0	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3.0	1.4	0.1	Iris-setosa
14	4.3	3.0	1.1	0.1	Iris-setosa
15	5.8	4.0	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa

4 grands problèmes :

Noisy data	Erreurs dans les valeurs d'attributs ou de classes
Incompleteness	Valeurs manquantes pour certains attributs / les attributs ne sont pas propices au problème
Inconsistency	Données contradictoires (même attribut prédicateur, mais classes différentes), incohérences dans la valeur de l'attribut (ex : km vs miles)
Redundant data	Les caractéristiques en double, un attribut peut être dérivé d'un autre

Raison :

Noisy data

- un équipement de collecte ou de transmission des données défectueux
- Erreur humaine lors de la saisie des données dans la base de données
- Défaut d'enregistrement d'une modification des données
- La valeur de l'attribut de classe peut dépendre de l'appréciation humaine.

Raison :

Incompleteness

- Données non saisies en raison d'un malentendu ou d'un manque de temps
- la valeur d'un attribut n'est pas applicable dans certaines situations/enregistrements
- la valeur de l'attribut n'est pas connue

Inconsistency

- Attributs pertinents en cas de désordre
- Incohérence dans les attributs
- Dataset mal choisi

Raison :

Redundant data

- Même attribut, mais les noms sont différents dans des bases de données différentes
- un attribut peut être dérivé d'un autre (ex : date de naissance et âge)
- attribut différent mais fortement lié (ex : "âge" et "retraité")

Solutions :

Noisy data

Binning

Incompleteness

- Supprimer des données les enregistrements comportant des valeurs manquantes
- Remplacer la valeur manquante par une valeur par défaut
- Prédire la valeur manquante sur la base des valeurs de l'autre attribut

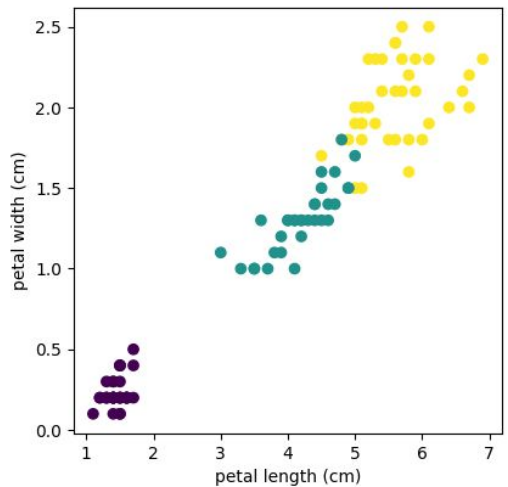
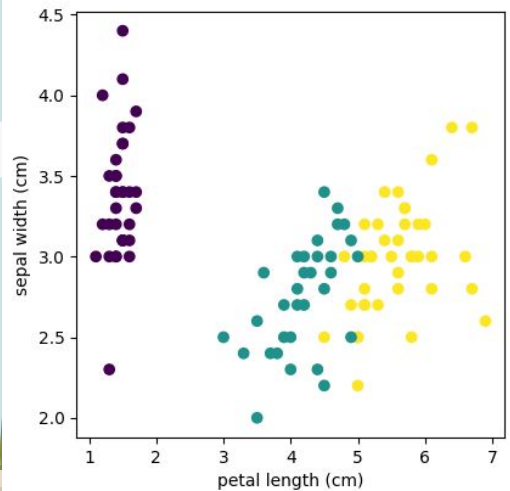
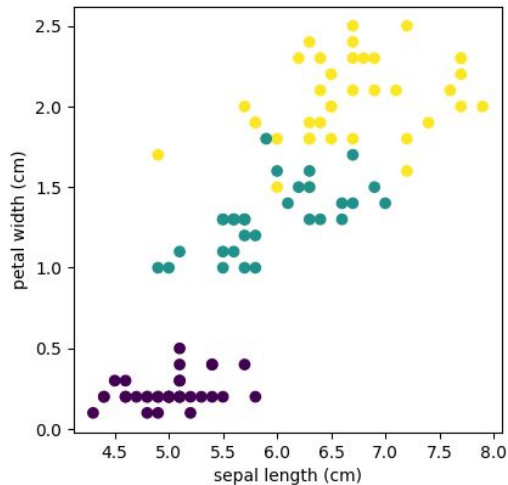
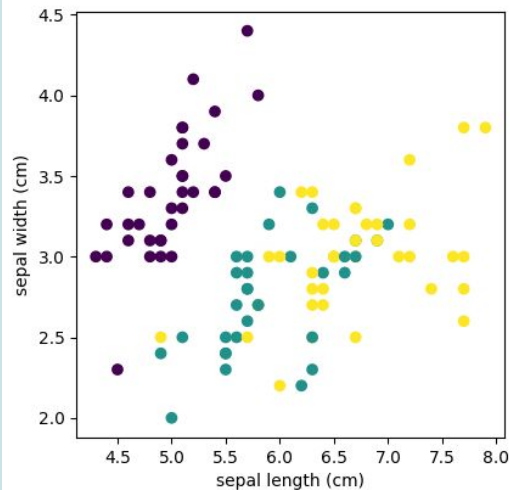
Solutions :

Inconsistency

- Création d'un entrepôt de données intégrant des données provenant de différents services dans un format cohérent
- Changement de dataset

Redundant data

Sampling



2- Selection des données

Quelles features garder?



Conclusion

Le traitement des données est une part importante de la mise en place de modèle et ne doit pas être négligé !





**Avez-vous des
questions?**