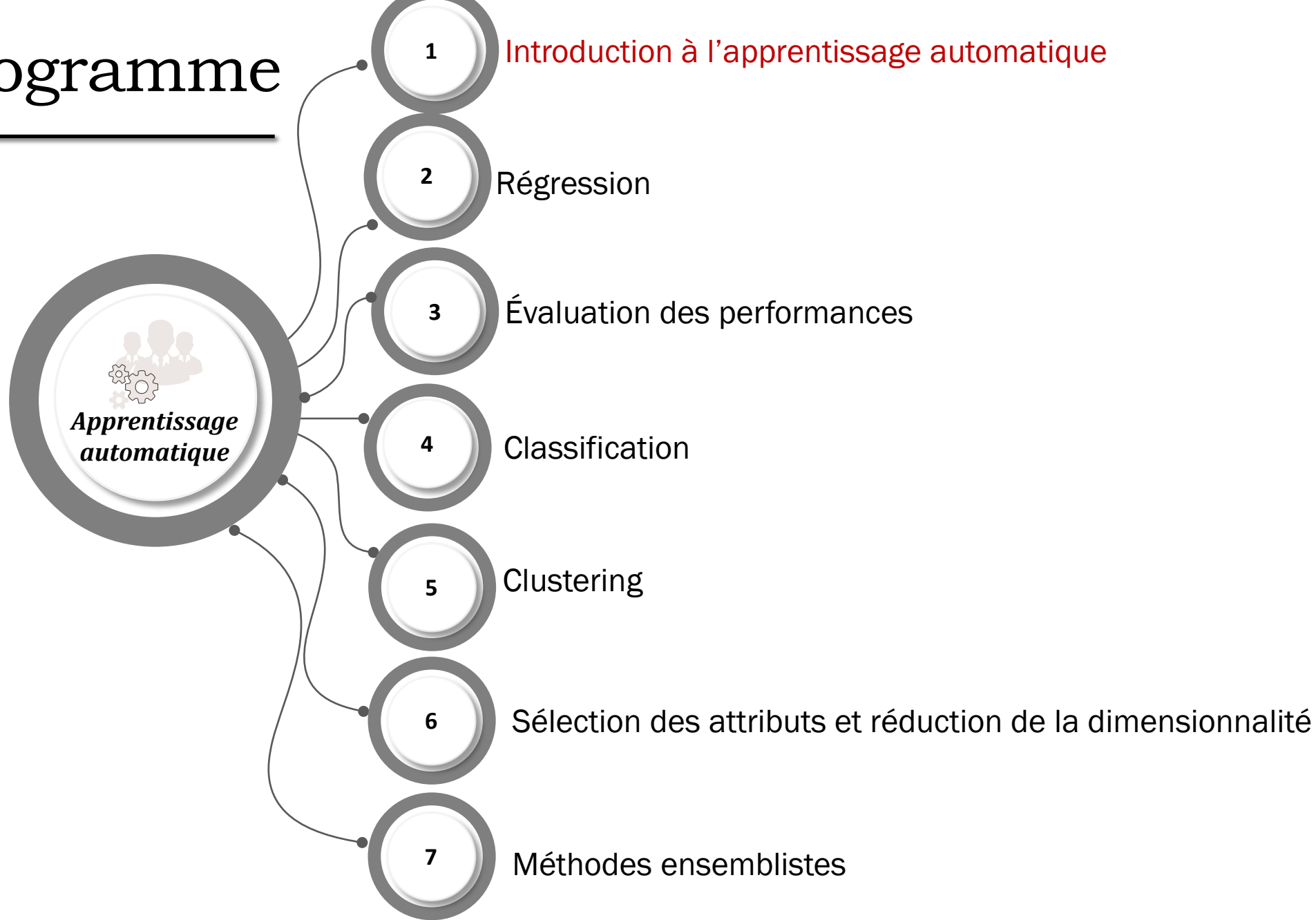


Apprentissage Automatique

*Intelligence Artificielle et
Sciences de Données
(IASD)*

DR N. DIF

Programme



Objectifs

du

cours

- Comprendre le processus de l'apprentissage automatique
- Distinguer entre les techniques d'apprentissage supervisé et non supervisé.

- Maîtrise du processus de fonctionnement des algorithmes d'apprentissage automatique et des méthodes d'évaluation.

- Comprendre le processus et l'utilité des techniques de sélection des attributs et réduction de la dimensionnalité

- Comprendre le processus des méthodes ensemblistes, leur utilité et leurs avantages par rapport aux méthodes classiques

Références

Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal, Data Mining Practical Machine Learning Tools and Techniques: Fourth Edition. Morgan Kaufmann, 2017.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Albon, C. (2018). Machine learning with python cookbook: Practical solutions from preprocessing to deep learning. "O'Reilly Media, Inc".

S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach (Pearson Series in Artificial Intelligence). 4th Edition, 2021.

JavaTpoint. (2011-2021). Machine Learning Tutorial. <https://www.javatpoint.com/machine-learning>.

ÉVALUATION

Moyenne = ?

Note TD = assiduité + préparation + test

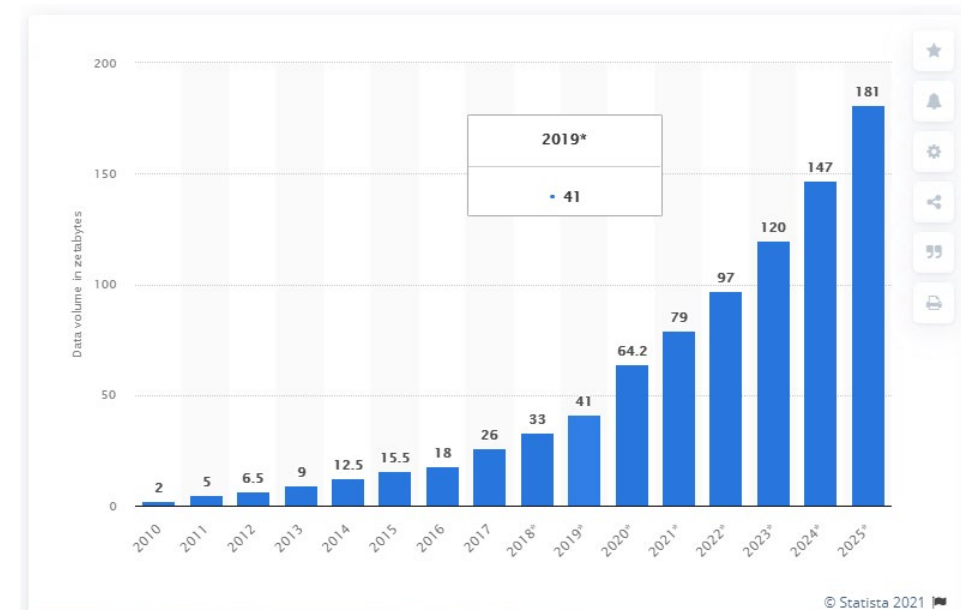
Note TP = assiduité + préparation + mini projet

1. INTRODUCTION

INTRODUCTION

- Développement de l'internet, informatisation des systèmes, la capacité élevée des équipements.
- Toute transaction est enregistrée sur la toile.
- **Conséquence** : Augmentation exponentielle de la quantité de données (Figure 1).

FIGURE 1. VOLUME DE DONNÉES/INFORMATIONS CRÉÉES, CAPTURÉES, COPIÉES ET CONSOMMÉES DANS LE MONDE DE 2010 À 2025 (EN ZETTAOCTETS) [1].



1. INTRODUCTION



PROBLÉMATIQUE

- La croissance rapide de données limite notre capacité de compréhension.
- La difficulté d'analyse et d'extraction des informations utiles à l'œil nu.



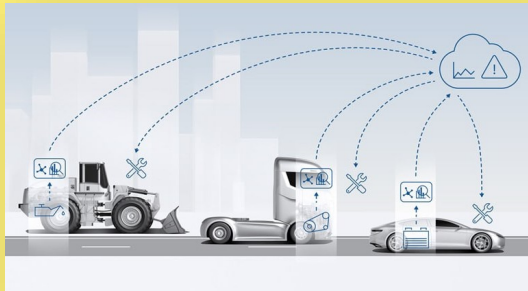
MOTIVATION

- Donner plus de sens aux données.
- Extraire les informations utiles.
- Créer des modèles ou des théories afin de réaliser des prédictions qui peuvent nous aider dans l'avenir à travers **les techniques d'apprentissage automatique.**

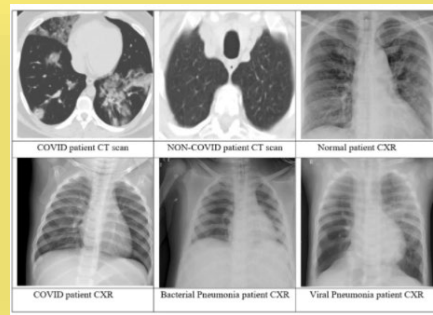
1. INTRODUCTION

Les informations collectées et les modèles de prédiction peuvent nous servir à quoi ?

Maintenance prédictive
des Machines



Détection du covid-19 à
partir des images X-ray
et CT-scans.



Détection des vols dans les
lieux publics



Suggestion de produits les
plus fréquemment achetés
dans un supermarché



2. DATA MINING (FOUILLE DE DONNÉES) : DÉFINITION ET DOMAINES D'APPLICATIONS

2.1. La différence entre une donnée, une information, et une connaissance

01



Donnée

Un élément **brut**, qui n'a pas été encore interprété, mis en contexte.

02



Information

Une donnée qui est transformée et classée dans une forme intelligible (structurée, organisée, traitée, et présentée dans son contexte), qui peut être utilisée dans le processus de prise de décision (**une donnée interprétée**).

03



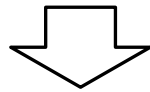
Connaissance

Signifie la familiarité et la conscience d'une **personne** de façons de prendre des décisions rassemblées par le biais de l'apprentissage, de la perception ou de la découverte. La combinaison d'informations, d'expérience et d'intuition mène à des connaissances.

2. DATA MINING (FOUILLE DE DONNÉES) : DÉFINITION ET DOMAINES D'APPLICATIONS

2.2. La définition de la fouille de donnée

- L'analyse de données depuis différentes perspectives et le fait de transformer ces données en informations utiles, en établissant des relations entre les données ou en repérant des patterns.
- Le processus de découverte des nouvelles connaissances en examinant de larges quantités de données (stockées dans des entrepôts, souvent appelés mégabase)

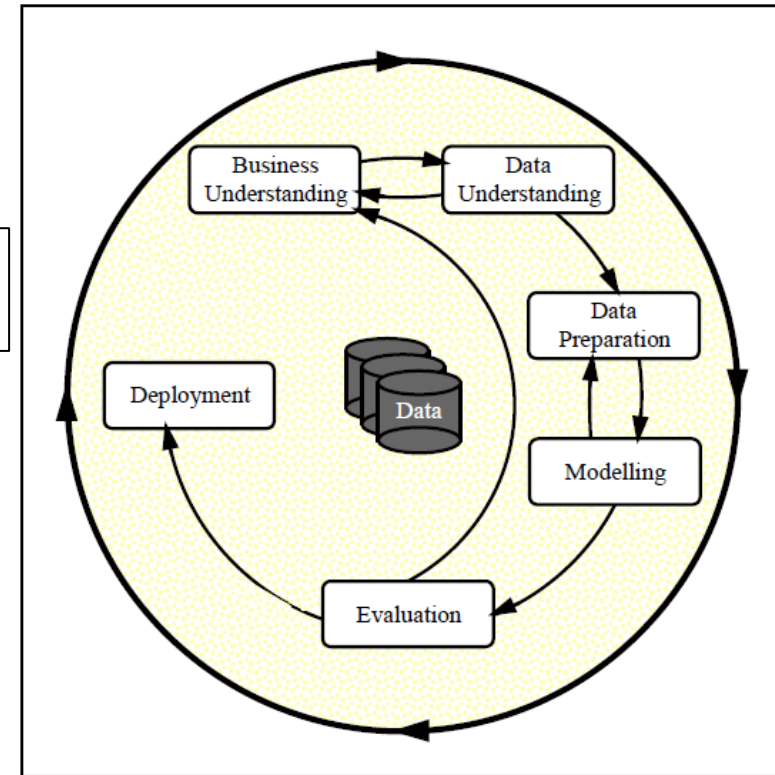


Elle permet d'extraire la connaissance à partir **des exemples (dataset, base d'apprentissage)** et à base des **algorithmes d'apprentissage automatique**

2. DATA MINING (FOUILLE DE DONNÉES) : DÉFINITION ET PRINCIPES

2.3. Le processus de la fouille de données (Crisp-DM)

Le processus Crisp-DM (Cross-Industry Standard Process for Data Mining) [2] est composé de 6 phases principales.



2.3. Le processus

de la fouille de

données (Crisp-DM)

1. Définir et comprendre le problème

- Comprendre les objectives du problème et le convertir en un problème de data mining.

2. Collecte des données

- Collecter les données nécessaires pour effectuer l'apprentissage.

3. Prétraitement

- Convertir les données collectées et **non structurées** à des **données structurées**.
- **Préparer les données à algorithme d'apprentissage** (discrétisation, traitement des valeurs manquantes et aberrantes, extraction d'attributs).
- **Améliorer la performance de l'algorithme** (augmentation de données, sélection des attributs).

But

➤ **Structurée** : un format prédéfini en lignes et en colonnes (Fichier Excel).

➤ **Semi-structurée** : une certaine structure de base est présente, mais le contenu lui-même n'est pas structuré (les courriers électroniques).

➤ **Non structurée** : Le type de fichier est connu, mais le contenu est complètement indépendant en soi. Ils ne sont pas situés dans des bases de données (images, audio, vidéo, texte).

4. Apprentissage

- Exploitation des **données prétraitées précédemment** et des **algorithmes d'apprentissage automatique** afin de générer un modèle (**pattern**) qui permet d'effectuer la **prédiction**, à ce niveau intervient le **processus d'apprentissage automatique**

5. Évaluation :

- Évaluer le modèle sur des données non exploitables durant l'apprentissage, pour éviter le problème de surapprentissage.
- Mesurer la performance du modèle, afin de déployer le modèle le plus robuste et fiable.

6. Déploiement

- Exploiter le modèle généré afin d'effectuer la prédiction

2. DATA MINING (FOUILLE DE DONNÉES) : DÉFINITION ET PRINCIPES

2.4. La structure de la base d'apprentissage

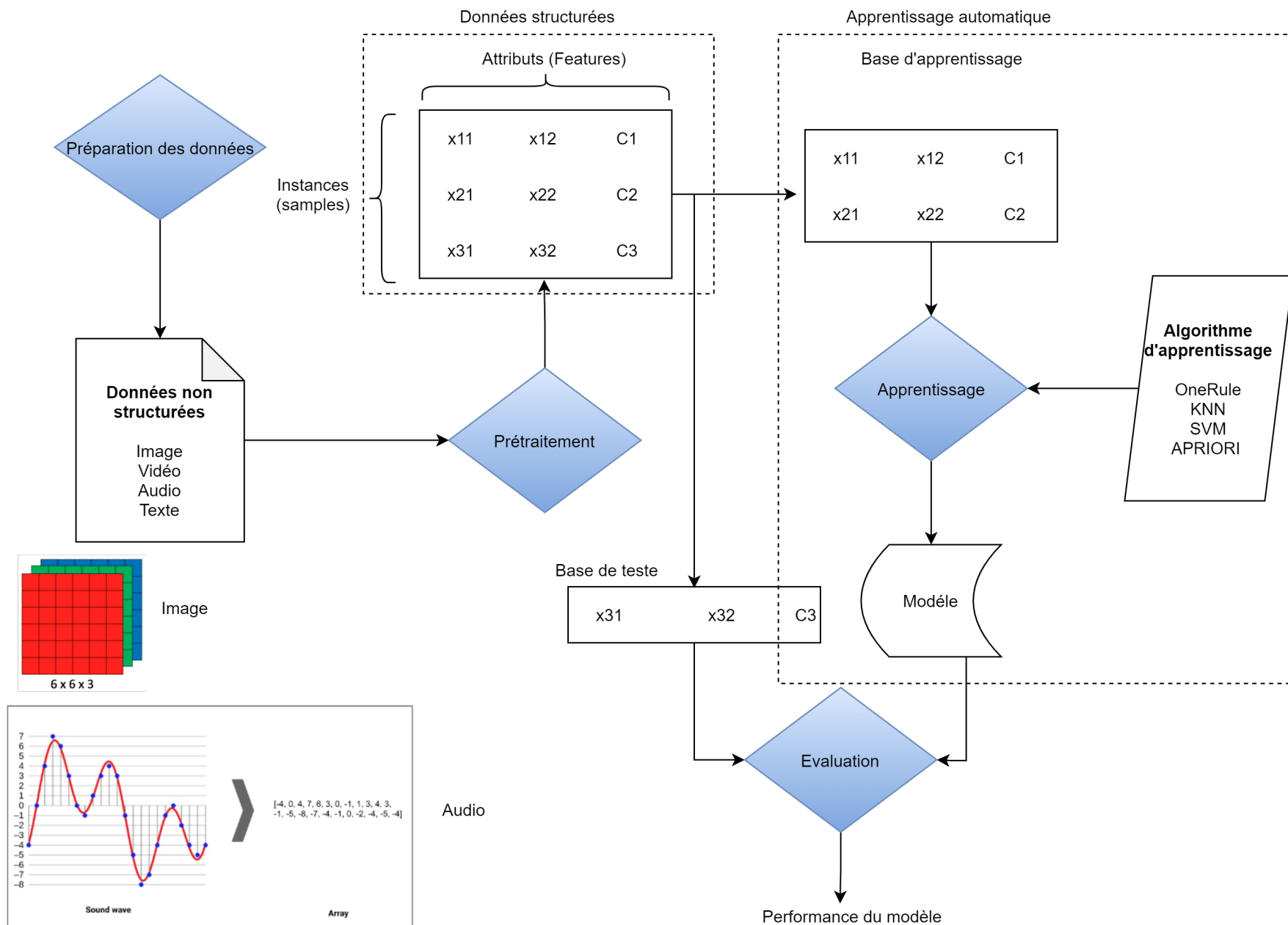
- La base d'apprentissage (dataset, benchmark) est composée d'un nombre d'attributs et d'instances.
- Les instances sont les exemples.
- Les attributs sont les caractéristiques de chaque exemple.
- Le choix des caractéristiques et des instances est très important et influence la qualité de l'apprentissage.
- Les valeurs des attributs sont soit de type **nominal symbolique** ou de type **numérique**.
- Pour plus de dataset : <https://www.kaggle.com/datasets>

Attributs				Classe
Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

3. APPRENTISSAGE AUTOMATIQUE

3.1. Définition

- L'apprentissage automatique est une application d'intelligence artificielle (IA).
- Il permet aux systèmes d'apprendre et de s'améliorer automatiquement à partir de l'expérience elle-même.
- Un algorithme d'apprentissage automatique prend en entrée une dataset, et il génère en sortie un modèle (prédiction).
- La qualité du modèle généré dépend essentiellement de la qualité de la dataset et la pertinence de l'algorithme d'apprentissage sur cette dataset.



3. APPRENTISSAGE AUTOMATIQUE

3.2. Le schéma d'apprentissage automatique

3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage automatique

Il existe 3 types d'apprentissages :

01



Apprentissage supervisé

Les instances sont classifiées au préalable (un attribut de type classe est présent dans la base, comme PlayTennis).

02



Apprentissage semi supervisé

Quelques instances sont classifiées, et d'autres non.

03



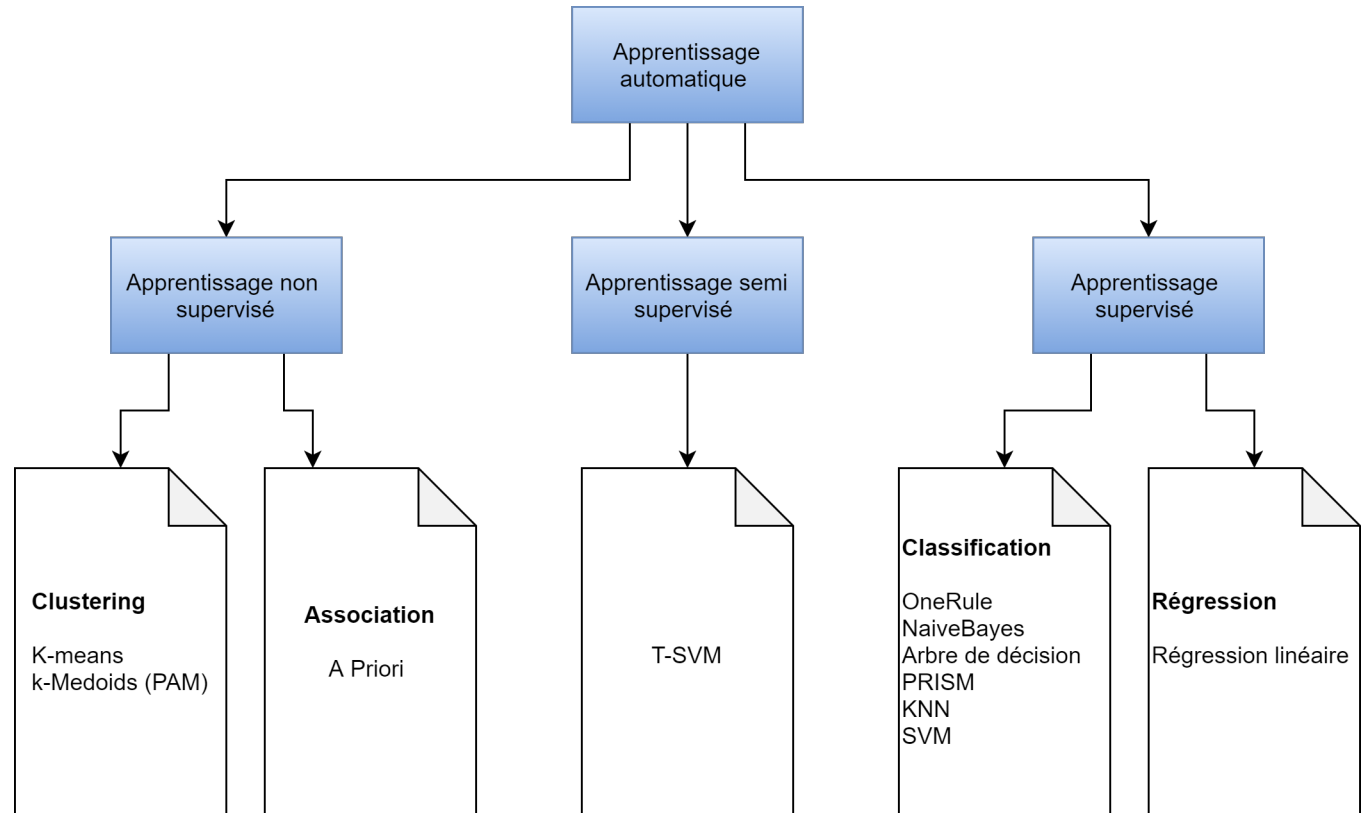
Apprentissage non supervisé

Toutes les instances ne sont pas classifiées.

3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage automatique

➤ Avant de commencer n'importe quel apprentissage, il faut avoir une idée sur le type de problème à résoudre.



3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage profond

Apprentissage supervisé	Apprentissage non supervisé
Le but des modèles est de trouver la fonction de mappage pour mapper la variable d'entrée (X) avec la variable de sortie ou classe (Y). $Y = f(X)$	Le but d'apprentissage non supervisé est de trouver en autonomie des modèles à partir des données.
Les données sont classifiées dans la base d'apprentissage	Les données sont classifiées dans la base d'apprentissage
Prédire la sortie (ou la classe).	Retrouve les patterns cachés à partir des données.
Nécessite une supervision pour construire le modèle.	Ne nécessite pas une supervision pour construire le modèle.
Le modèle d'apprentissage supervisé est généralement caractérisé par des résultats performants.	Le modèle d'apprentissage non supervisé produit des résultats moins performants.
N'est pas vraiment proche à l'intelligence humaine, car dans ce cas, le modèle est entraîné à partir des données pré classifiées.	Plus intuitive et proche à l'intelligence humaine, car il apprend de la même manière qu'un enfant apprend les choses de la routine quotidienne par ses expériences.
Il peut être catégorisé à des problèmes de classification et de régression.	Il peut être catégorisé à des problèmes de clustering et d'associations.

3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage automatique

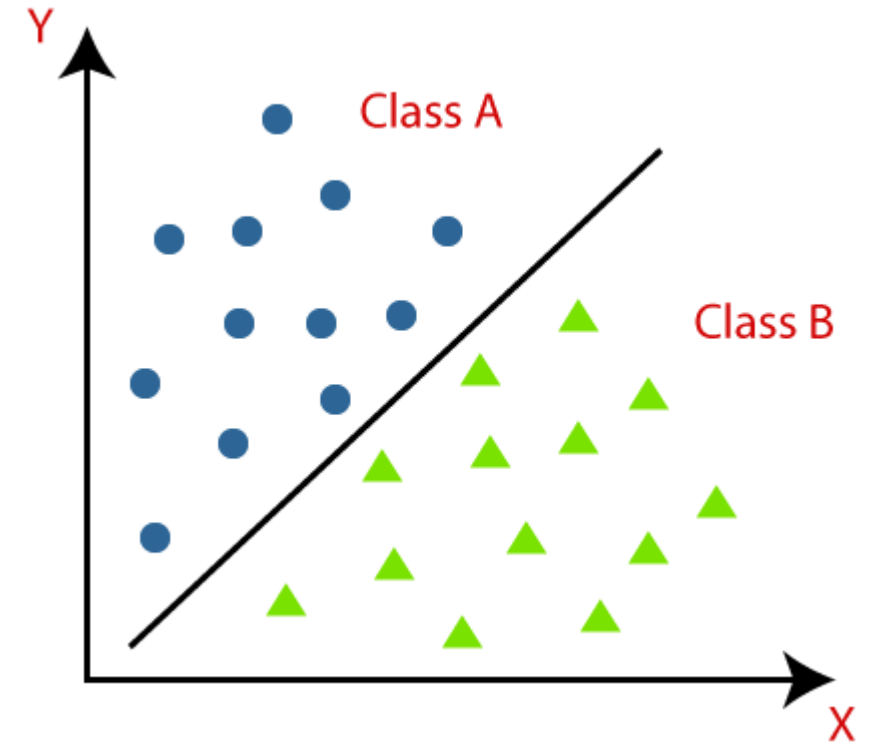
3.3.1. Apprentissage supervisé

3.3.1.1. Classification

- Approche supervisée.
- Prédire la catégorie des nouvelles instances.
- Chaque observation est associée à une classe de type discret : $Y \in \{C1, C2, \dots, Cn\}$, n est le nombre de catégorie ou classe (ex : Oui ou Non).

Algorithmes
paresseux

Algorithmes qui génèrent un modèle
(eager learners)



3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage automatique

3.3.1. Apprentissage supervisé

3.3.1.1. Classification

Écoute social et analyse des sentiments appliquée aux dialectes arabes et multilingues [3]

L'analyse des sentiments est une application de l'apprentissage automatique du traitement du langage naturel.

Elle permet d'analyser les commentaires et extraire l'opinion ou le sentiment que comportent ces derniers.

Collecte des données à partir de tweeter (une écoute sociale sur twitter à l'aide des hashtags , ou mot clé, par exemple sur un produit précis)

The screenshot shows a web interface with the title "écoute twitter". Below the title, there is a subtitle: "ici, vous pouvez écrire n'importe quel mot (ex: votre nom de marque) et obtenir les derniers tweets, hashtags et mentions à ce sujet". The interface contains two input fields: the first is labeled "Mot Q" and has the placeholder text "le mot recherché"; the second is labeled "Nombre de tweets récents" and has the placeholder text "combien de tweets?". Below these fields is a button labeled "Envoyer".

3. APPRENTISSAGE AUTOMATIQUE


3.3. Les topologies des techniques d'apprentissage automatique

3.3.1. Apprentissage supervisé

3.3.1.1. Classification

Écoute social et analyse des sentiments appliquée aux dialectes arabes et multilingues [3]
Etape de prédiction

Chargez votre fichier ici
le fichier doit être au format csv



Chargez un fichier

☐ LR ☐ DCT
☒ KNN ☐ RF
☐ MNB ☐ Voting
☐ SVM ☐ DL

Envoyer

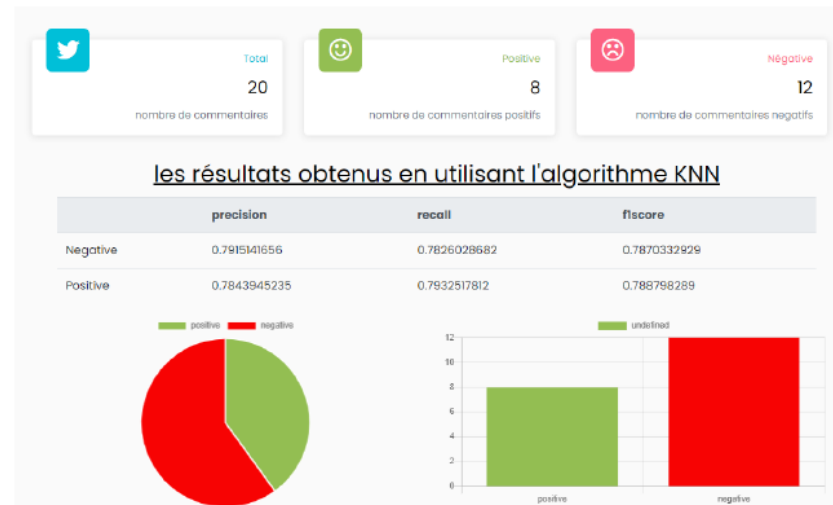


Table des résultats

id	tweet	class
0	لطفه براهن تاخ تسجب	Positive
1	فرحتكم ببيت الحمد لله	Positive
2	مادي شغفه وقليه ادب	Negative
3	لكيا فور دايه يزاف	Positive
4	أشرف حاك لعمه الأشفاق ورانه ممبر	Positive
5	بانه لخر والهناريه	Negative

« 1 2 3 4 »

3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage automatique

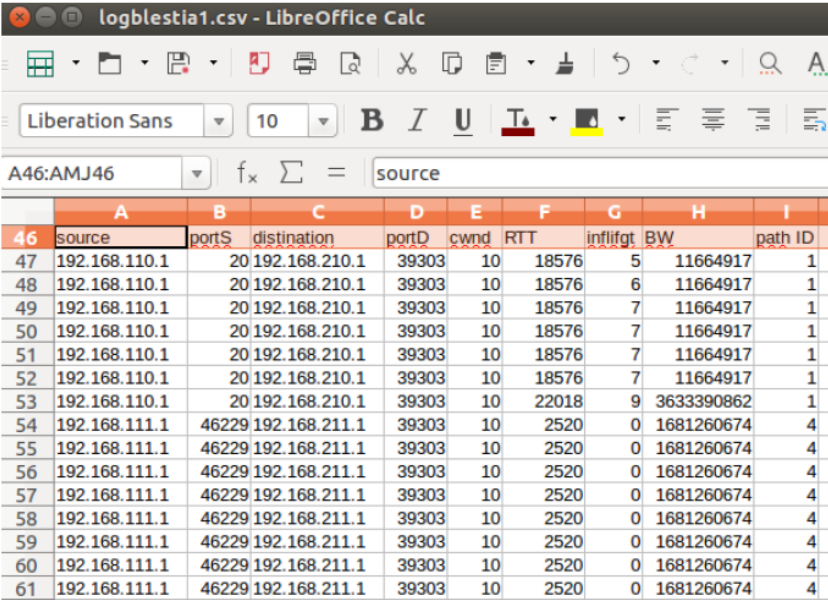
3.3.1. Apprentissage supervisé

3.3.1.1. Classification

Elaboration et intégration d'une stratégie d'ordonnancement du protocole MTCP basée sur le Machine Learning [4]

Concevoir et intégrer une nouvelle stratégie heuristique d'ordonnancement MPTCP de noyau Linux, basée sur le Machine Learning.

	Champ	Signification
1	source	l'adresse IP de la source
2	portS	le port source
3	destination	l'adresse IP de la destination
4	portD	le port destination
5	cwnd	la fenêtre de congestion
6	RTT	le temps Round Trip Time
7	inflight	le nombre des paquets envoyés mais non encore acquittés
8	BW	la bande passante
9	path ID	ID de chemin défini la sortie de modèle, il est attribué au pair (@source ;port source – @destination ;port destination)



	A	B	C	D	E	F	G	H	I
46	source	portS	destination	portD	cwnd	RTT	inflight	BW	path ID
47	192.168.110.1	20	192.168.210.1	39303	10	18576	5	11664917	1
48	192.168.110.1	20	192.168.210.1	39303	10	18576	6	11664917	1
49	192.168.110.1	20	192.168.210.1	39303	10	18576	7	11664917	1
50	192.168.110.1	20	192.168.210.1	39303	10	18576	7	11664917	1
51	192.168.110.1	20	192.168.210.1	39303	10	18576	7	11664917	1
52	192.168.110.1	20	192.168.210.1	39303	10	18576	7	11664917	1
53	192.168.110.1	20	192.168.210.1	39303	10	22018	9	3633390862	1
54	192.168.111.1	46229	192.168.211.1	39303	10	2520	0	1681260674	4
55	192.168.111.1	46229	192.168.211.1	39303	10	2520	0	1681260674	4
56	192.168.111.1	46229	192.168.211.1	39303	10	2520	0	1681260674	4
57	192.168.111.1	46229	192.168.211.1	39303	10	2520	0	1681260674	4
58	192.168.111.1	46229	192.168.211.1	39303	10	2520	0	1681260674	4
59	192.168.111.1	46229	192.168.211.1	39303	10	2520	0	1681260674	4
60	192.168.111.1	46229	192.168.211.1	39303	10	2520	0	1681260674	4
61	192.168.111.1	46229	192.168.211.1	39303	10	2520	0	1681260674	4

Les adresses IP source et destination qui sont des champs de types texte (encodage à l'aide de *One-Hot-Encoding*).

Les algorithmes de classification utilisés : Arbre de décision, Forêt aléatoire, KNN, SVM, Régression logistique, SVM.

3. APPRENTISSAGE AUTOMATIQUE

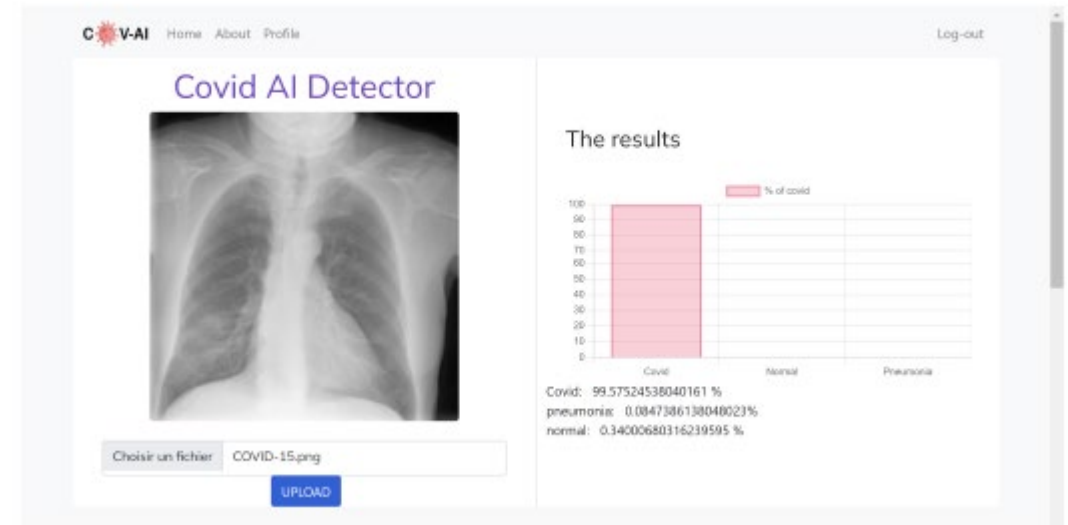
3.3. Les topologies des techniques d'apprentissage automatique

3.3.1. Apprentissage supervisé

3.3.1.1. Classification

Detection of COVID-19 from Chest X-Ray Images using Deep learning [5]

L'utilisation des techniques d'apprentissage automatique pour réduire la charge sur les professionnelles de santé et d'éviter leur subjectivité. Ces méthodes sont utilisées pour la détection du covid-19 à partir des images X-rays.



3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage automatique

3.3.1. Apprentissage supervisé

3.3.1.2. Régression

- Approche supervisée.
- Le résultat à prédire est une valeur numérique $Y \in [a, b]$.
- Exemple : le calcul de la température, l'estimation du prix d'une voiture occasion, l'estimation du salaire.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	Price
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

La base d'apprentissage (boston house prices) pour la prédiction des prix des maisons.

3. APPRENTISSAGE AUTOMATIQUE

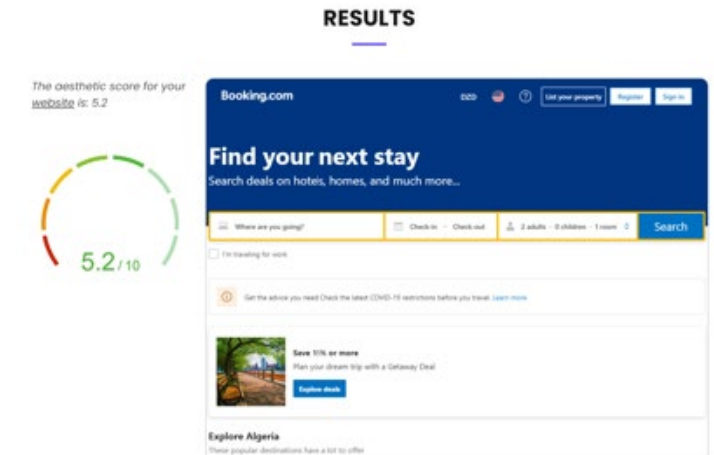
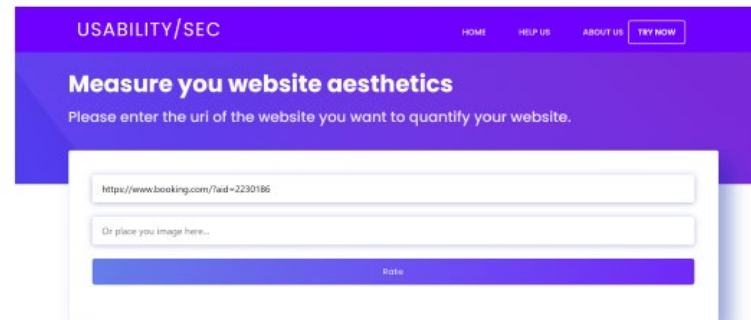
3.3. Les topologies des techniques d'apprentissage automatique

3.3.1. Apprentissage supervisé

3.3.1.2. Régression

Quantifying The Aesthetics of Graphic Interfaces With Deep Learning [6]

Cette méthode utilise une capture d'écran d'un site web comme entrée, puis détermine s'il s'agit d'une jolie interface sur une échelle de 1 à 9 en fonction des évaluations des utilisateurs.



3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage automatique

3.3.2. Apprentissage non supervisé

3.3.2.1. Association

- L'association est une technique d'apprentissage non supervisé.
- Trouver les relations ou les associations entre les variables (ou attributs) de la base d'apprentissage.
- Basée sur les règles d'association (Déclaration si et sinon : si A alors B) afin d'extraire ces relations.



L'analyse du panier de marché :

- Une technique utilisée par les grands supermarchés pour découvrir les associations entre les articles.
- Rechercher les combinaisons d'éléments qui se produisent fréquemment ensemble dans les transactions (bon d'achat).



3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage profond

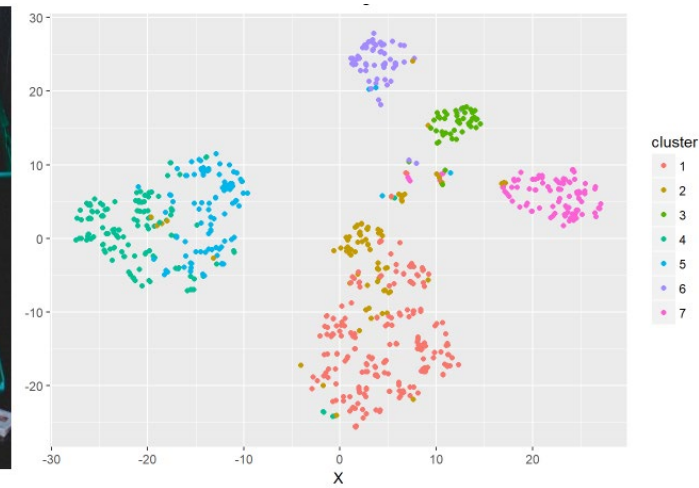
3.3.2. Apprentissage non supervisé

3.3.2.2. Clustering

- Technique d'apprentissage non supervisé.
- Regrouper les données dans différents clusters. Ce regroupement est basé sur une fonction de similarité.
- Maximiser la variance intercluster et minimiser la variance intracluster. L'association est une technique d'apprentissage non supervisé.



Exemples : la segmentation des affaires dans un supermarché, l'analyse des réseaux sociaux (regroupement en communauté), segmentation des images.



3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage profond

3.3.2. Apprentissage non supervisé

3.3.2.2. Clustering

Predicting COVID-19 from chest X-ray images using fine tuning and transfer learning [7]

L'utilisation des techniques de clustering au lieu des méthodes de classification pour la détection du covid-19.



3. APPRENTISSAGE AUTOMATIQUE

3.3. Comment choisir l'algorithme d'apprentissage approprié

- Avec un attribut dominant. (OneRule)
- Tous les attributs contribuent indépendamment et équitablement à la classification. (Naïve Bayes)
- Avec peu d'attributs représentatifs qui peuvent être représentés sous forme d'un arbre de décision. (Arbre de décision)
- Basé sur quelques règles de décision qui peuvent attribuer les instances aux différentes classes. (Prisme)
- Contenant des dépendances entre ses attributs. (Les algorithmes de règles d'association)
- Contenant des dépendances linéaires entre ses attributs numériques. (Régression linéaire)
- Avec des classes qui peuvent être liées à une certaine distance entre les instances. (KNN)
- Avec des instances qui peuvent être regroupées dans différents groupes. (Clustering)

Impossible d'analyser les bases d'apprentissage de grands volumes de données à l'œil nu !

Références

- [1] Arne Holst, Jun 7. Amount of data created, consumed, and stored 2010-2025, 2021
- [2] Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1). London, UK: Springer-Verlag.
- [3] Kouadri A. R., Alahoum N. Écoute social et analyse des sentiments appliquée aux dialectes arabes et multilingues. Thesis (2020/2021), Ecole supérieure en informatique 8 Mai 1945 Sidi Bel Abbés.
- [4] ARBAOUI M., MENNI A. S. Elaboration et intégration d'une stratégie d'ordonnancement du protocole MTCP basée sur le Machine Learning. Thesis (2020/2021), Ecole supérieure en informatique 8 Mai 1945 Sidi Bel Abbés.
- [5] ARIOUI A., ZEBLAH I. Detection of COVID-19 from Chest X-Ray Images using Deep learning . Thesis (2021/2022), Ecole supérieure en informatique 8 Mai 1945 Sidi Bel Abbés.
- [6] LAMRI M. C.. Quantifying The Aesthetics of Graphic Interfaces With Deep Learning. Thesis (2021/2022), Ecole supérieure en informatique 8 Mai 1945 Sidi Bel Abbés.
- [7] Dermache M. D., Boularaoui M. A. Predicting COVID-19 from chest X-ray images using fine tuning and transfer learning. Thesis (2021/2022), Ecole supérieure en informatique 8 Mai 1945 Sidi Bel Abbés.