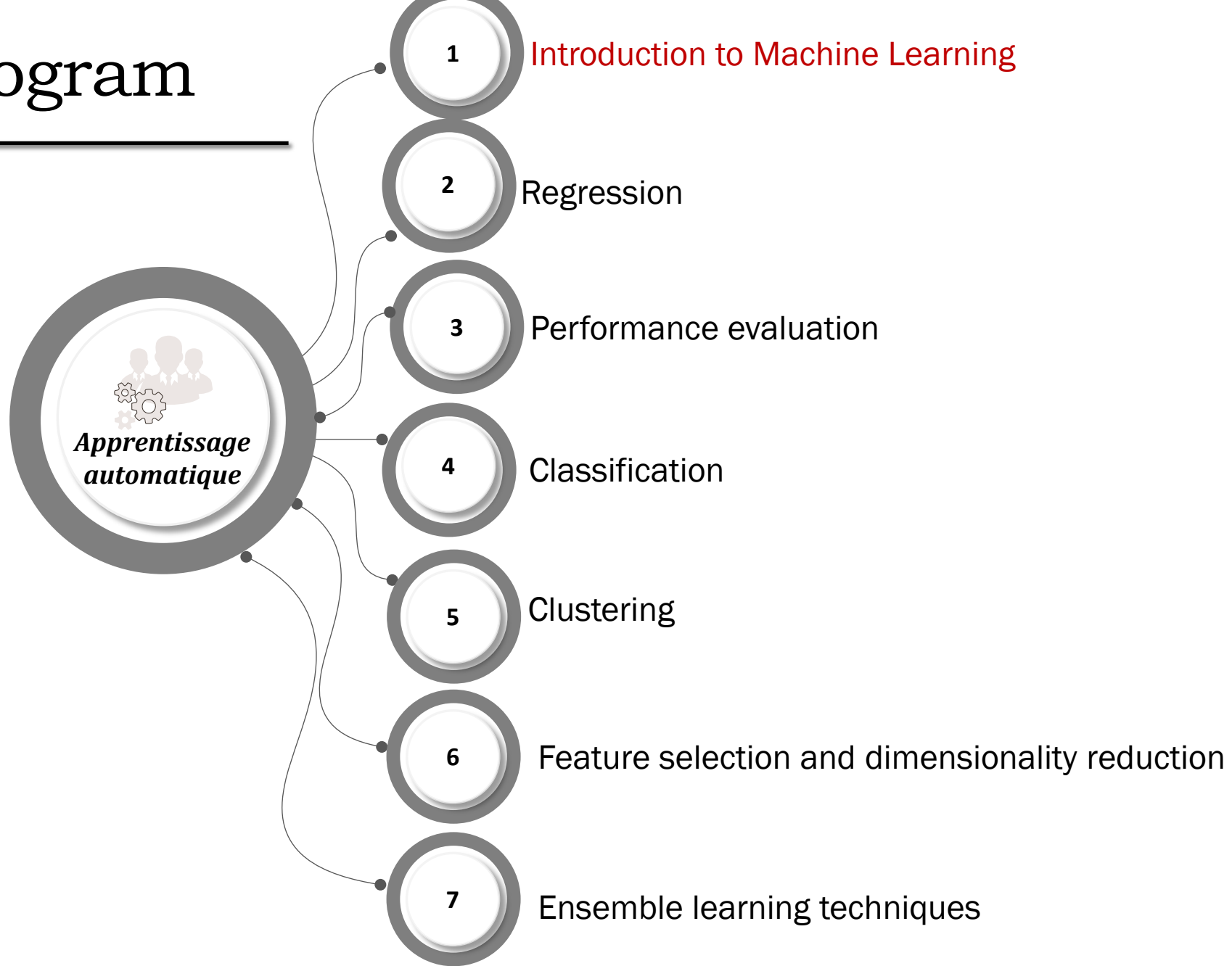


Machine Learning

*Intelligence Artificielle et
Sciences de Données
(IASD)*

DR N. DIF

Program



Objectifs

du

cours

- Understand the Machine Learning Process.
- Distinguish Between Supervised and Unsupervised Learning Techniques.

- Understand the process of machine learning algorithms and evaluation method.

- Understand the process and the advantages of feature selection and dimensionality reduction techniques

- Understanding the process of ensemble learning techniques, their benefits, and their advantages compared to classical methods

References

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>.

Evaluation

Average = ?

Tutorials = Punctuality + Preparation + Test

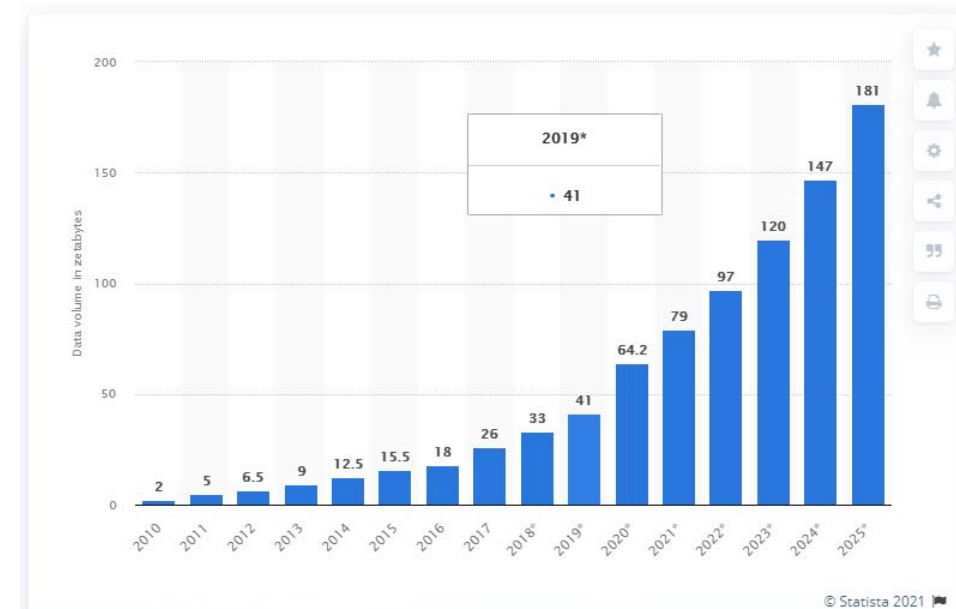
Practical Works = Punctuality + Preparation + Mini Project

1. INTRODUCTION

INTRODUCTION

- Internet growth, computerization of systems, High-capacity devices.
- Every transaction is recorded.
- **Consequence** : The amount of data has augmented exponentially (Figure 1).

FIGURE 1. Volume Of Data/Information Created, Captured, Copied, And Consumed From 2010 To 2025 (In Zettabytes) [1].



1. INTRODUCTION



PROBLEMATIC

- The rapid growth of data limits our capacity of understanding.
- The difficulty of analyzing and extracting useful information manually.



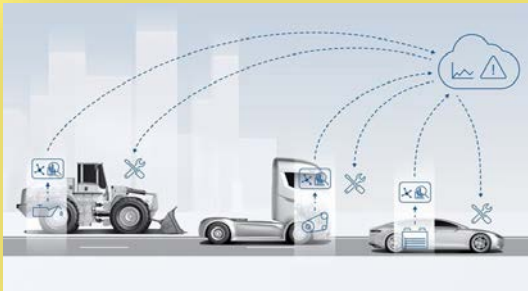
MOTIVATION

- Adds more meaning to data.
- Extract useful information.
- Use machine learning techniques to create models in order to make predictions that can assist us in the future.

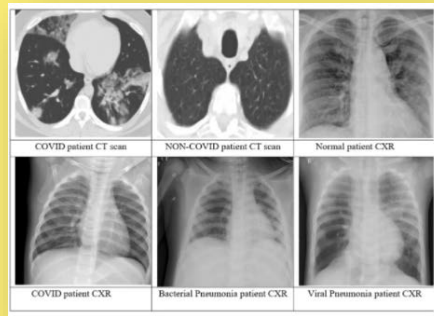
1. INTRODUCTION

Application domains

Predictive maintenance systems for machines.



Detecting Covid-19 from X-ray images or CT-scans.



Theft detection in public places.



Product recommendation systems for Supermarket.



2. DATA MINING : DEFINITION AND APPLICATION DOMAINS

2.1. The difference between data, information, and knowledge

01

Data

Raw data, which has not yet been interpreted or contextualized .

02

Information

Data that is transformed and organized into an intelligible form (structured, organized, processed, and presented within its context), which can be used in the decision-making process (**interpreted data**)

03

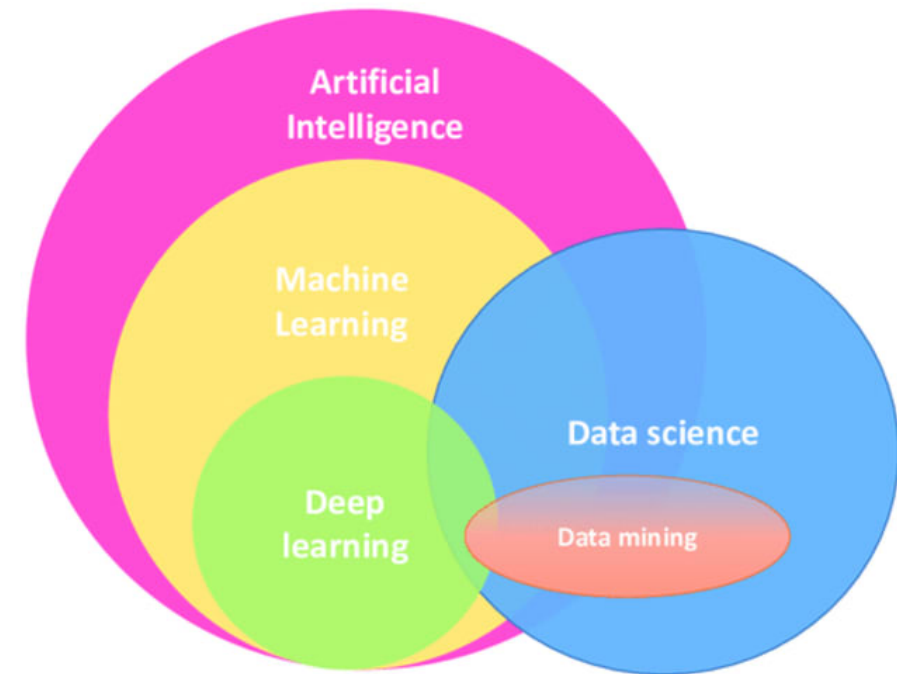
Knowledge

Signifies a person's familiarity of making decisions gathered through learning, perception, or discovery. The combination of information, experience, and intuition leads to knowledge

2. DATA MINING : DEFINITION AND APPLICATION DOMAINS

2.2. Data mining definition

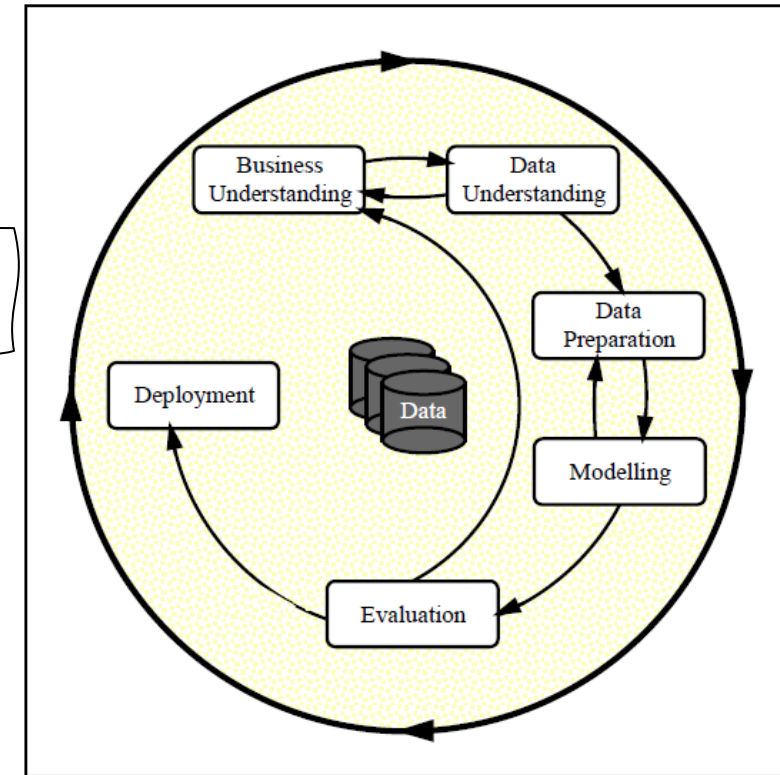
Data mining is the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.



2. DATA MINING : DEFINITION AND APPLICATION DOMAINS

2.3. *The Crisp-DM process in data mining*

The process Crisp-DM (Cross-Industry Standard Process for Data Mining) [2] consists of 6 main steps .



2.3. Data mining

process

(Crisp-DM)



1. Business Understanding

- Understand the problem and convert it into a data mining problem.

2. Data Understanding

- Collect data for the learning process

3. Preprocess

- Convert collected and unstructured data into structured data
- Prepare the data for the machine learning algorithms.
- Enhance the algorithm's performance.

4. Modelling

- Use the previously preprocessed data and machine learning algorithms to generate a model for making predictions; at this stage the machine learning process is used.

5. Evaluation :

- Evaluate the model on unseen data during training to avoid overfitting problems
- Compute the model's performance to deploy the most robust and reliable (fiable) model.

6. Deployment

- Exploit the generated model in an interface to make predictions.

2. DATA MINING : DEFINITION AND APPLICATION DOMAINS

2.4. Preprocess

The main purpose of preprocessing techniques is to :

- Convert unstructured data into structured data.
 - Prepare the data for the machine learning algorithm (discretization, handling missing and outlier values, attribute extraction).
 - Improve the algorithm's performance (data augmentation, feature selection).
-

Structured: A predefined format in rows and columns (Excel file).

Semi-structured: Some basic structure is present, but the content itself is not structured (emails).

Unstructured: They are not located in databases (images, audio, video, text).

2. DATA MINING : DEFINITION AND APPLICATION DOMAINS

2.5. The dataset's structure and characteristics

- The training dataset (dataset, benchmark) is composed of a number of features and samples.
- Samples are the examples.
- Features are the characteristics of each sample.
- The choice of features and samples is very important and influences the learning process and the performance of the generated model.
- The values of features can be nominal type or numeric.
- For more datasets, visit:

<https://www.kaggle.com/datasets>

features				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

2. DATA MINING : DEFINITION AND APPLICATION DOMAINS

2.5. The dataset's structure and characteristics

Relevance: The dataset should represent the problem to be solved by including relevant and representative examples from the real world.

Diversity, Variability, and Complexity: Diversifying the dataset is important to prevent overfitting problems and to ensure generalization. It should include a large number of variable scenarios that differ in complexity.

Quality: A good-quality dataset is essential for an efficient predictive model. It should be cleaned of outliers, errors, and missing values.

Size: The size of the dataset should be sufficient to guarantee an efficient learning process. In general, high-quality models are trained on large datasets.

Consistency: Consistency guarantees that the ML model can learn patterns effectively. Diverse formats, scales, or interpretations across features can introduce errors, uncertainty, and biases.

2. DATA MINING : DEFINITION AND APPLICATION DOMAINS

2.5. The dataset's structure and characteristics

Balance: The distribution of feature values or examples over labels should be balanced. Unbalanced datasets risk bias toward the majority class.

Labeling: High-quality labeling is important in the machine-learning process. Accurate labeling helps to generate high-quality predictive models.

Timeliness: The dataset should accurately reflect the current state of the ML problem. It ensures that the model can make predictions based on recent examples.

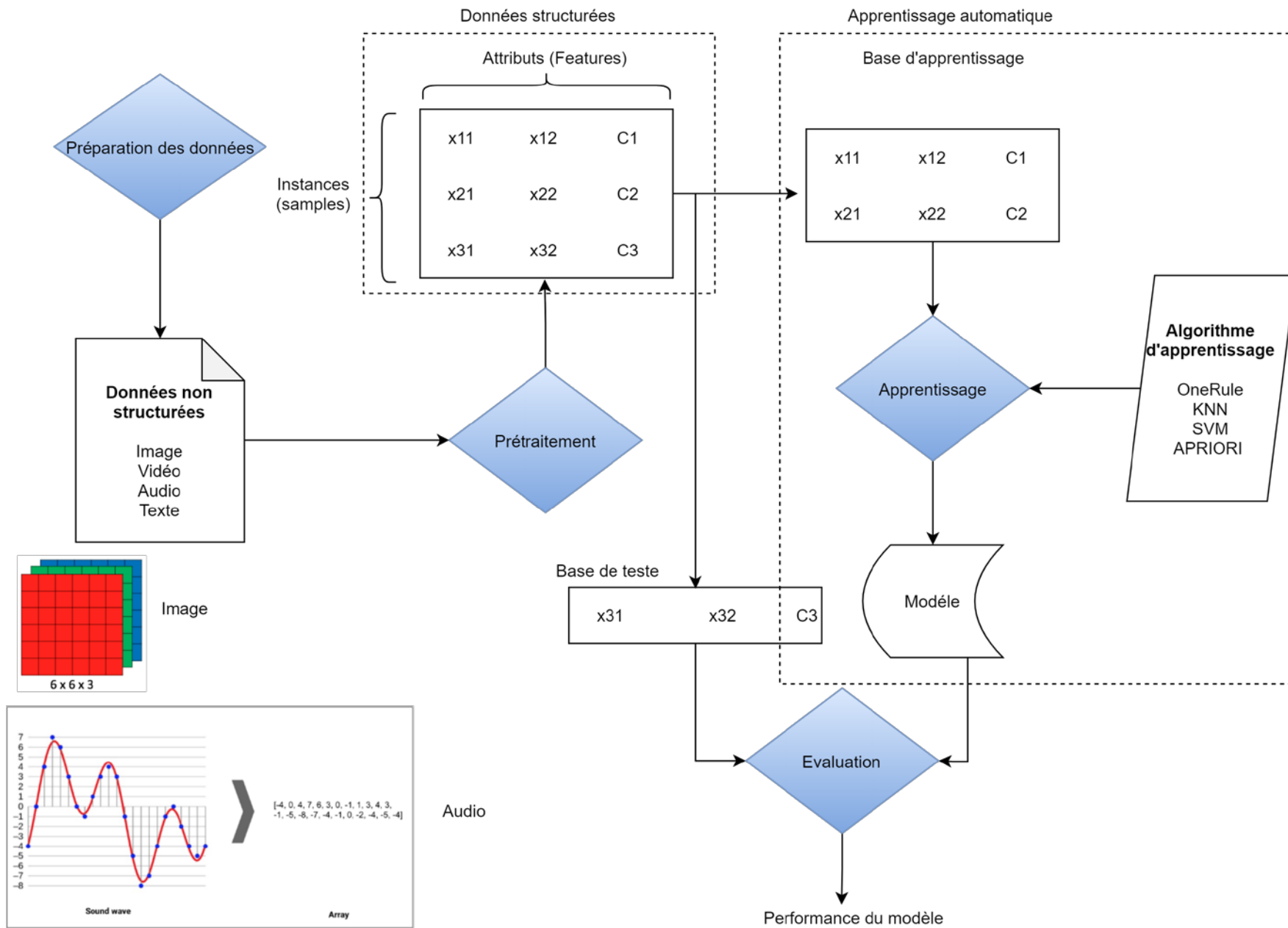
3. MACHINE LEARNING

3.1. Definition

- Machine learning is a subfield of artificial intelligence (AI).
- It enables systems to learn and improve automatically from their experience.
- A machine learning algorithm takes a dataset as input, and it produces a model as an output.
- The quality of the generated model depends primarily on the quality of the dataset and the suitability of the machine learning algorithm for learning on this dataset.

3. MACHINE LEARNING

3.2. The machine learning scheme



3. MACHINE LEARNING

3.3. *Types of machine learning*

01



Supervised learning

The samples are classified in the dataset (a feature class is present, such as PlayTennis).

02



Semi-supervised learning

Only a subset of samples is classified.

03



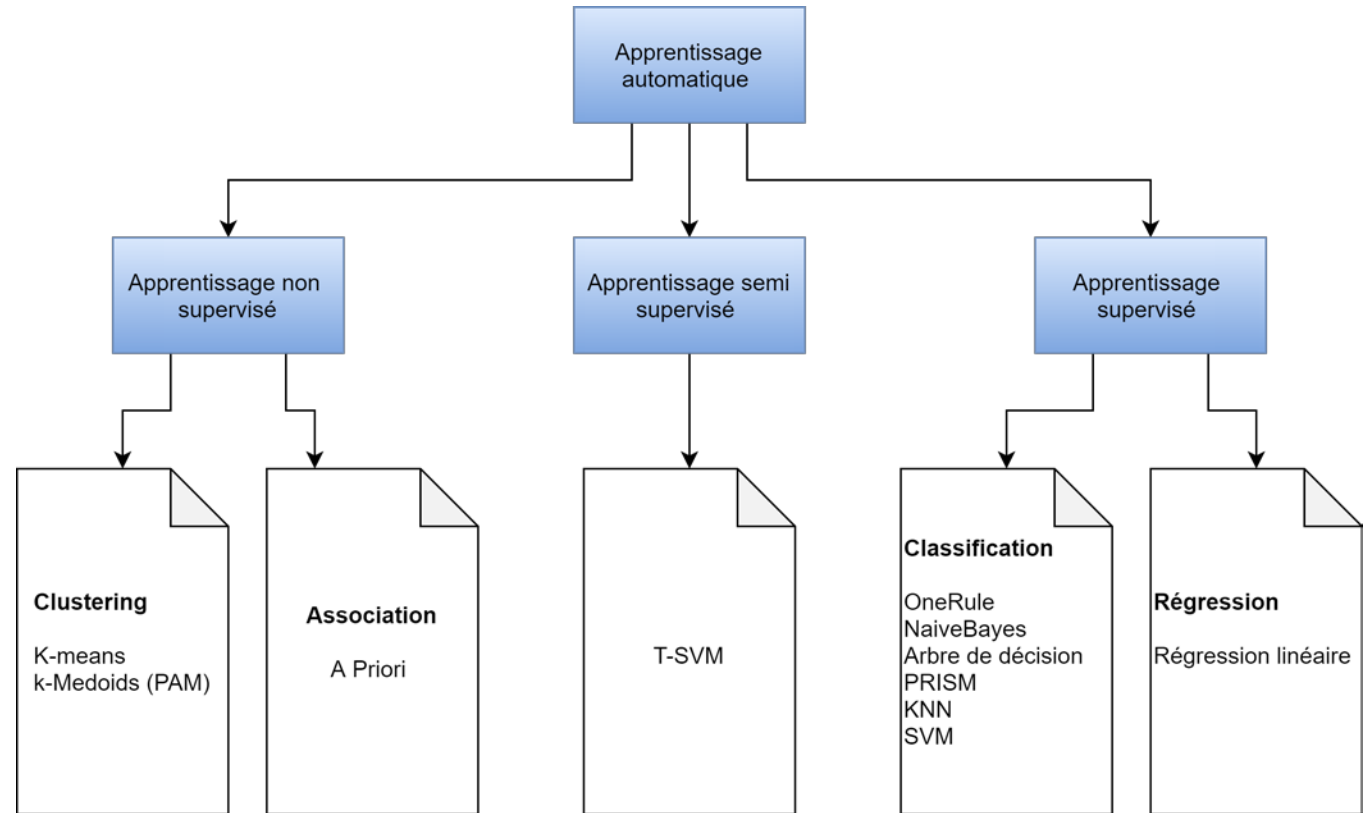
Unsupervised learning

All samples are unclassified.

3. MACHINE LEARNING

3.3. Types of machine learning

➤ Before starting any learning process, we should have an idea of the type of problem to solve.



3. MACHINE LEARNING

3.3. *Types of machine learning*

Supervised learning	Unsupervised learning
The main purpose of models is to find the mapping function to map the input variable (X) to the output variable or class (Y). $Y = f(X)$	Find models from data without supervision
Samples are classified in the dataset.	Samples are not classified in the dataset.
Supervised learning models are typically characterized by accurate results.	Unsupervised learning models are characterized by less accurate results.
It's not very close to human intelligence because the model is trained from pre-classified data.	More intuitive and closer to human intelligence, as it learns in the same way a child learns things from daily routine through its experiences.
It can be categorized into classification and regression problems.	It can be categorized into clustering and d'association problems.

3. MACHINE LEARNING

3.3. Types of machine learning

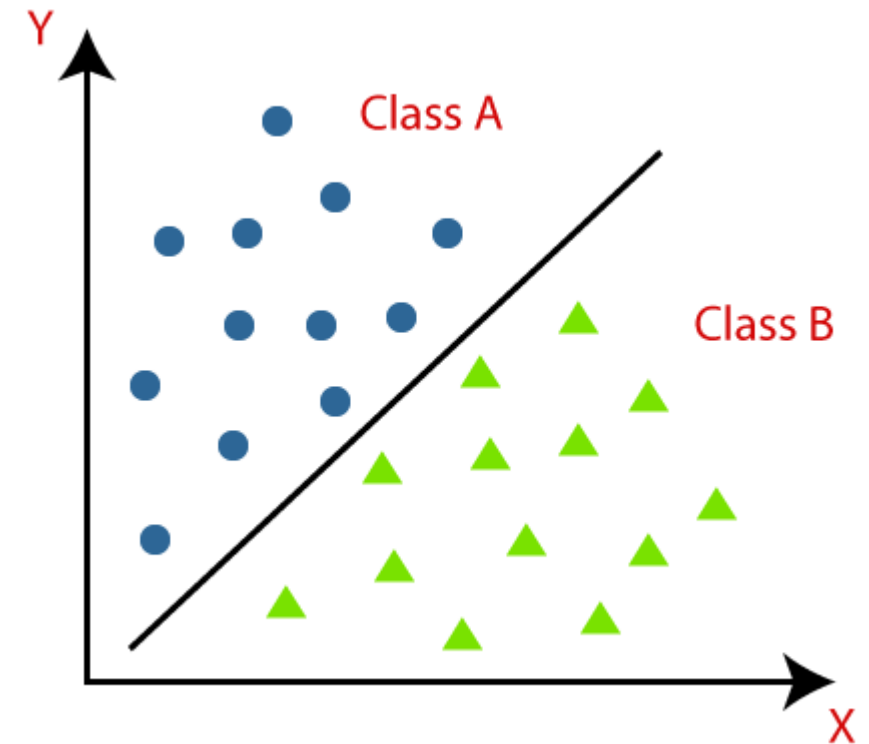
3.3.1. Supervised learning

3.3.1.1. Classification

- Supervised approach.
- Predict the class of new samples.
- Each observation is associated with a discrete class.: $Y \in \{C1, C2, \dots, Cn\}$, n is the number of classes (ex : Yes or No).

Lazy learners

Eager learners



3. MACHINE LEARNING

3.3. Types of machine learning

3.3.1. Supervised learning

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 for 'écoute twitter'. It includes a title, a subtitle explaining the purpose, a search input field, a field for the number of tweets, and a submit button.

écoute twitter

ici, vous pouvez écrire n'importe quel mot (ex: votre nom de marque)
et obtenir les derniers tweets, hashtags et mentions à ce sujet

Mot Q

le mot recherché

Nombre de tweets récents

combien de tweets?

Envoyer

3. MACHINE LEARNING


3.3. Types of machine learning

3.3.1. Supervised learning

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



☐ LR

☒ KNN

☐ MNB

☐ SVM

☐ DCT

☐ RF

☐ Voting

☐ DL

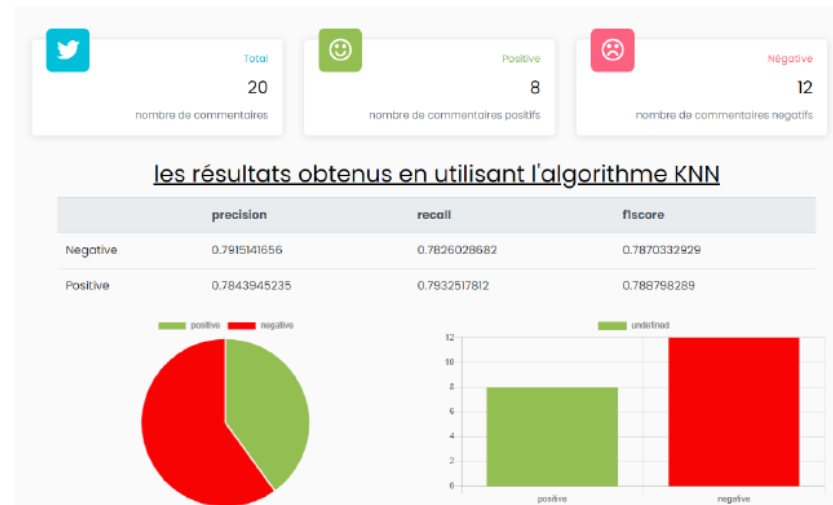


Table des résultats

id	tweet	class
0	لعله براهن تاخ تسجب	Positive
1	فرحتكم ببيت الحمد لله	Positive
2	مادي شعله وقليه ادب	Negative
3	لكيا فور دايه يزاف	Positive
4	أشرف حاك لعمه الاشفاق وراسه مسمر	Positive
5	بانه لخرنا والفانزيه	Negative

«

1

2

3

4

»

3. MACHINE LEARNING

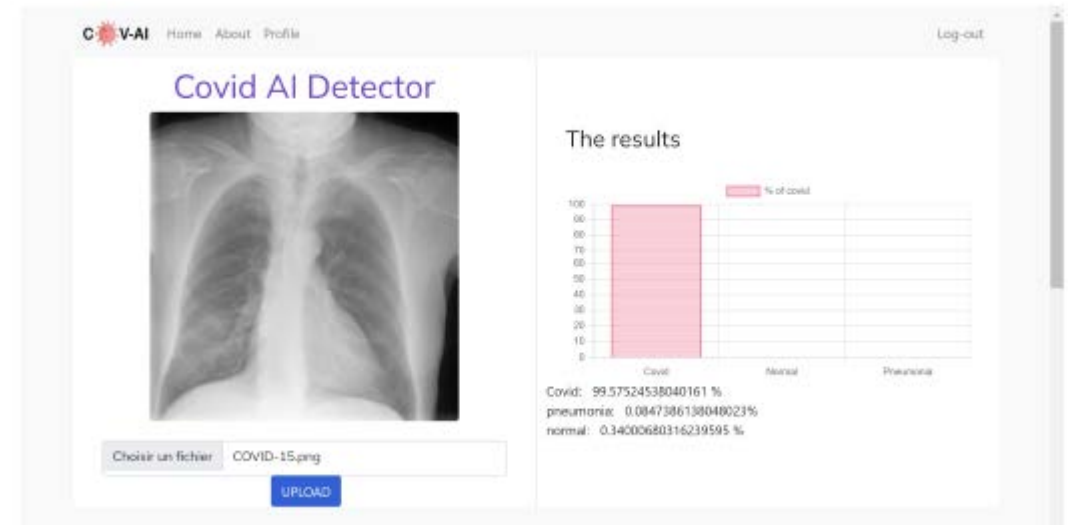
3.3. Types of machine learning

3.3.1. Supervised learning

3.3.1.1. Classification

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

The exploitation of machine learning techniques to reduce the workload on healthcare professionals and to avoid their subjective decisions. These methods are used for the detection of COVID-19 from X-ray images.



3. MACHINE LEARNING

3.3. Types of machine learning

3.3.1. Supervised learning

3.3.1.2. Regression

- A supervised approach.
- The predicted class is numeric $Y \in [a, b]$.
- Example : estimating the temperature, the price of a used car, the salary...

	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. MACHINE LEARNING

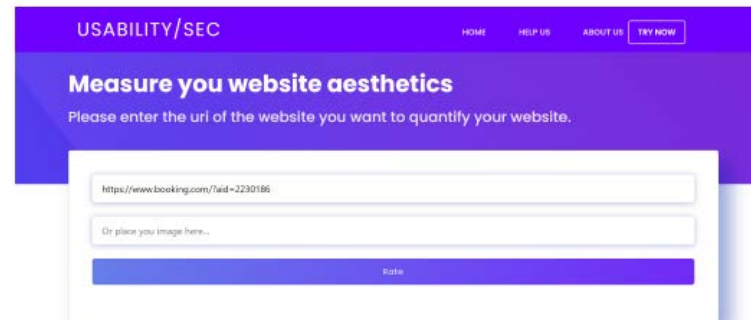
3.3. Types of machine learning

3.3.1. Supervised learning

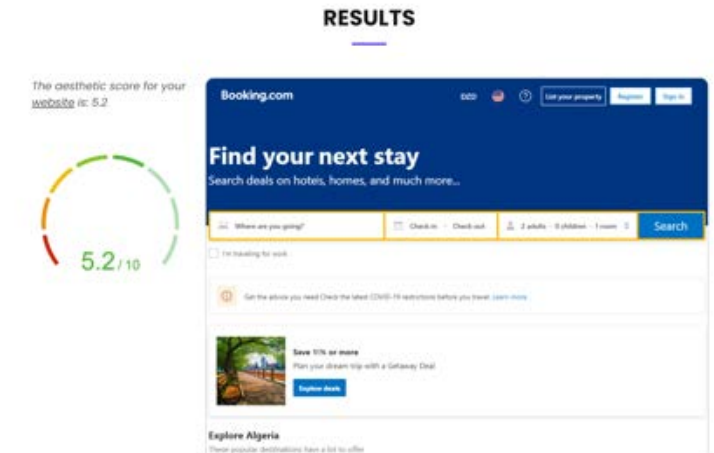
3.3.1.2. Regression

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

This method uses a screenshot of a website as input, then determines whether it is a visually appealing (attirante) interface on a scale of 1 to 9 based on user ratings.



The screenshot shows a web interface for 'USABILITY/SEC'. It has a purple header with navigation links: HOME, HELP US, ABOUT US, and TRY NOW. The main heading is 'Measure your website aesthetics' with a subtext 'Please enter the url of the website you want to quantify your website.' Below this is a text input field containing 'https://www.booking.com/?aid=2230186'. There is also a placeholder for an image with the text 'Or place your image here...'. At the bottom is a large blue button labeled 'Rate'.



3. MACHINE LEARNING

3.3. Types of machine learning

3.3.2. Unsupervised learning

3.3.2.1. Association

- Unsupervised learning technique.
- Find the relationships or associations between the variables (or features) in the training dataset.
- Based on association rules (If-Then statements: If A, then B) in order to extract these relationships.



Market Basket Analysis :

- A technique used by huge supermarkets to discover associations between items.
- Search for combinations of elements that frequently occur together in transactions (purchase receipts).



3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage automatique

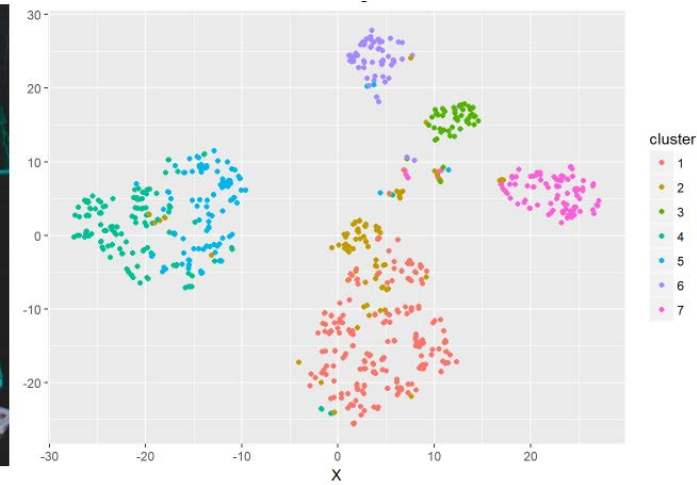
3.3.2. Apprentissage non supervisé

3.3.2.1. Clustering

- Unsupervised technique.
- Group data into different clusters. This clustering is based on a similarity function.
- Maximize inter-cluster variance and minimize intra-cluster variance.



Examples : products segmentation in a supermarket, social network analysis (community clustering), image segmentation.



3. APPRENTISSAGE AUTOMATIQUE

3.3. Les topologies des techniques d'apprentissage automatique

3.3.2. Apprentissage non supervisé

3.3.2.1. Clustering

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

The exploitation of clustering techniques instead of classification methods for COVID-19 detection.



3. MACHINE LEARNING

3.3. How to choose the machine learning algorithm

- With a dominant feature. (OneRule)
- All features contribute independently and equally to classification. (Naïve Bayes)
- With a few representative features that can be represented as a decision tree. (Decision tree)
- Based on a few decision rules that can assign instances to different classes. (Prism)
- Contains dependencies between its attributes. (Association rule algorithms)
- Contains linear dependencies between its numerical attributes. (Linear regression)
- With classes that can be related to a distance between instances. (KNN)
- With instances that can be grouped into different clusters. (Clustering)

It's impossible to analyze large volumes of training data manually with eye!

References

- [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.