REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN - TIARET

MEMOIRE

Présenté à :

FACULTÉ DES MATHEMATIQUES ET DE l'INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Génie Logiciel

Par :

MERMIT Ahmed
KAIS Oussama

Sur le thème

La proposition d'une approche de l'extraction d'informations basée sur l'algorithme K-Means

Soutenu publiquement le 26 / 06 / 2022 à Tiaret devant le jury composé de :

Mr BENDAOUD Mebarek	Pr	Université Ibn Khaldoun	Président
Mr DAOUD Mohamed Amine	MAA	Université Ibn Khaldoun	Encadrant
Mr DJAFRI Laouni	MCA	Université Ibn Khaldoun	Examinateur

2021-2022

Remerciements

Tout d'abord, nous remercions **ALLAH** qui nous aide et nous donne la patience et le courage durant ces années d‘étude.

Nous souhaitons d‘adresser nos remerciements les plus sincères aux personnes qui nous ont apporté leur aide et qui ont contribué à l‘élaboration de ce mémoire.

Ces remerciements vont au corps professoral et administratif de département d‘informatique de l‘université d‘Ibn Khaldoun de Tiaret pour la richesse et la qualité de leurs enseignements.

Ensuite nous tenons à remercier notre encadreur

Mr DAOUD Mohamed Amine

Pour l‘orientation, la confiance, la patience qui ont constitué un apport considérable sans lequel ce travail n‘aurait pas pu être mené. Qu‘il trouve dans ce travail un hommage vivant à sa haute personnalité.

Nous tenons aussi à remercier les membres du jury qui ont accepté d‘examiner notre mémoire.

Enfin, nous adressons nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours soutenue et encouragé au cours de la réalisation de ce mémoire.

Merci à tous.

Dédicaces

Je dédie cet humble travail :

À mon cher père et à ma chère mère Que Dieu les protège et leurs offre la chance et le bonheur.

À mes Frères qui je souhaite un avenir radieux plein de réussite

À toute ma famille.

À mes Amis qui me sont chers

Je remercie également tous mes professeurs et surtout mon encadreur

Mr Daoud Mohamed Amine

En un mot à tous les gens qui contribué ma réussite de près ou de loin.

Puisse Dieu vous donne santé, bonheur, courage et surtout réussite

(KAIS Oussama & MERMIT Ahmed)

TABLE DES MATIERES

Liste des abréviations

IA : Intelligence Artificielle.

EI: Extraction d'information.

MUC: Message Understanding Conference.

NER: Named Entity Recognition.

POS: Part of Speech Tagging.

RC : Résolution de Coréférence.

ER : Extraction de Relations.

EE : Extraction des Evénements.

ML : Machine Learning.

IDS : Intrusion Détection Système.

H-IDS : Host Intrusion Détection Système.

N-IDS : Network Intrusion Détection Système.

Liste des figures

Liste des tableaux

Abstract

Currently, the number of researches in the field of intrusion detection systems has increased considerably due to the importance of this field, and the results of these researches are often presented in the form of scientific papers, reports etc. containing evaluations of their work.

The problem is that the community has not thought to take advantage of the results obtained from published articles and scientific reports in the field of intrusion detection systems and to build a database for this field, to classify and categorize the field.

Our objective is to extract information on the documents of the domain of intrusion detection systems and build the dataset and then study this dataset by machine learning techniques.

Keywords: Information Extraction, Machine Learning, Intrusion Detection System, NER, POS Tagging, K-means.

Résumé

Actuellement, le nombre de recherches dans le domaine des systèmes de détection d'intrusion a considérablement augmenté en raison de l'importance de ce domaine, et les résultats de ces recherches se présentent souvent sous la forme de articles scientifique, rapport etc. contenant des évaluations de leur travail.

Le problème pose est que la communauté n'a pas pensé à tirer parti des résultats obtenus à partir d'articles publiés et de rapports scientifiques de domaine des systèmes de détection d'intrusion et à constituer une base de données pour ce domaine, afin de classifier et catégoriser le domaine.

Notre objectif consiste à faire l'extraction d'information sur le document de domaine des systèmes de détection d‘intrusions et construire le dataset ensuite étudier cet dataset par des techniques de la machine Learning.

Mots clés : Extraction d'information, Apprentissage automatique, Système de détection d'intrusion, NER, POS Tagging, K-means.

الملخص

في الوقت الحاضر، زاد عدد الأبحاث في مجال أنظمة الكشف عن التسلل بشكل كبير بسبب أهمية هذا المجال، وغالبا ما تكون نتائج هذا البحث في شكل مقالات علمية وتقارير وما إلى ذلك، وتحتوي على تقييمات لعملهم.

المشكلة أن المجتمع لم يفكر في الاستفادة من النتائج التي تم الحصول عليها من المقالات المنشورة والتقارير العلمية في مجال أنظمة كشف التسلل وبناء قاعدة بيانات لهذا المجال. لتصنيف المجال.

هدفنا هو استخراج المعلومات حول وثائق المجال لأنظمة الكشف عن التسلل وبناء مجموعة البيانات ثم دراسة مجموعة البيانات هذه باستخدام تقنيات التعلم الآلي.

الكلمات المفتاحية : استخراج المعلومات ، التعلم الآلي ، نظام كشف التسلل ، التعرف على الكيان المحدد، وضع علامات على جزء من الكلام، الخوارزمية التصنيفية .

INTRODUCTION GENERALE

INTRODUCTION GENERALE

En raison de la croissance rapide d'Internet, la quantité d'informations accessibles augmente à un rythme incroyable. La plus grande partie de ceci est sous forme textuelle surtout pour la communauté des chercheurs par la publication des articles scientifiques, des rapports etc. Le traitement manuel de ces types de documents demande un effort énorme. L'extraction d'informations est un domaine permettant d'extraire des informations utiles à partir de textes en langage naturel et de fournir une sortie lisible par machine. Elle est basée sur l'analyse de documents textuelle afin de rassembler et de structurer des informations précises.

Il existe plusieurs applications qui nécessitent l'extraction d'informations comme la surveillance des entreprises (extraction d'informations à partir des nouvelles les concernant). Dans notre cas, on s'intéresse au domaine des systèmes de détection d'intrusion, vus aux nombres croissants des articles scientifiques dans des revues et des conférences, la communauté n'a pas pensé à la capitalisation des résultats obtenus lors de la proposition de leurs solutions. Cela le transfert des données de domaine IDS (Intrusion détection Intrusion) à partir des articles scientifiques vers des repository (base de données spécialisée) et la construction de bases de connaissances pour la communauté de la détection (par exemple, la collecte à partir de la littérature scientifique), afin de classifier et catégoriser le domaine.

L'objectif de cette mémoire est de capitaliser le domaine des IDS par la proposition d'une approche d'extraction d'information à partir des articles scientifiques en se basant sur les mesures de performance des IDS. Cela a été examiné par les méthodes d'apprentissage automatique.
La structure de mémoire :

Afin de répondre à cet objectif, ce mémoire est structuré de la façon suivante :

- Le premier chapitre est un chapitre descriptif sur l'extraction d'information, par une ensemble définitions, ainsi que son évolution, l'architecture et les taches d'extraction d'information.
- Le deuxième chapitre présente les différentes techniques d'apprentissage automatique dont leur fonctionnement et ses algorithmes.
- Le troisième chapitre est consacré à notre cas d'étude, c'est le domaine des systèmes de détection d'intrusion, il présente la classification des systèmes de détection d'intrusion.
- Le dernier chapitre est consacré à présenter notre approche proposée et l'implémentation des résultats de notre travail.

CHAPITRE 1 :
Extraction d'informations

CHAPITRE 1 : EXTRACTION D'INFORMATION

1.1 Introduction

L'extraction d'information est désormais un sujet de recherche important dans le domaine du Traitement Automatique des Langues Naturelles (TALN). Elle connaît ces dernières années un intérêt grandissant car elle répond à un besoin devenu incontournable dans la société de l'information.[1]

1.2 Traitement du langage naturel

Le traitement du langage naturel est un domaine de l'intelligence artificielle (IA) qui donne aux machines la capacité de lire, de comprendre et de tirer un sens des langues humaines. C'est un composant important dans une large gamme d'applications logicielles que nous utilisons dans notre vie quotidienne.[2]

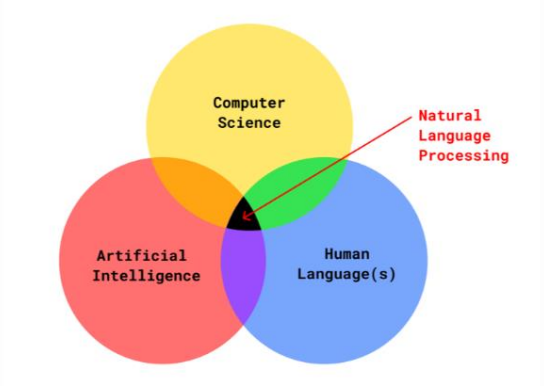

Figure 1 Le traitement du langage naturel

1.3 Définitions de L'extraction d'information

Il existe de nombreuses définitions pour extraction d'informations notamment :

- L'extraction d'information est un processus d'analyse de données non structurées et d'extraction d'informations essentielles dans des formats de données plus modifiables et structurés.[3]

- L'Extraction d'Information ou EI (en anglais, Information Extraction ou IE) désigne une technologie récente qui vise à extraire et à structurer automatiquement un ensemble d'informations précises apparaissant dans un ou plusieurs documents textuels écrits en langue naturelle. Ces informations sont

destinées à créer ou alimenter un entrepôt de données (appelé aussi banque de données)[4].

- L'extraction d'information est le processus automatique, qui permet d'extraire des informations pertinentes et précises à partir de documents non structurés en langage naturel et permet leur sauvegarde sous une forme structurée du type formulaire ou base de données[5] .

1.4 Evaluation de l'extraction d'information

Les conférences sur la compréhension des messages ou les concours de compréhension des messages (MUC) ont joué un rôle important dans le développement de l'extraction d'informations en tant que domaine d'étude. Cette conférence a été initiée par le Naval Ocean Systems Centre (NOSC) des États-Unis et a été parrainée par la Defence Advanced Research Project Agency (DARPA). Les MUC ont eu lieu sept fois entre 1987 et 1998. Bien que l'événement soit appelé "conférence", il peut être décrit avec d'autres mots comme "compétition" entre les groupes de recherche en extraction d'information ou "évaluation" des performances de leurs systèmes [6].

Les deux premières MUCs (1987-1989) se sont concentrées sur l'analyse automatique de messages militaires contenant des informations textuelles sur les batailles navales où le template à remplir contenait 10 champs (attributs). La MUC-3 (Lehnert, Cardie, Fisher, Riloff, & Williams, 1991) s'est intéressée à l'extraction à partir d'articles de presse des informations concernant des activités terroristes, des fondations internationales, des évènements de succession de gestion d'entreprises et des lancements de missiles et de véhicules spatiaux. Les structures des templates à remplir s'est complexifiée au fil du temps. À partir de MUC-5, la structure imbriquée des templates et l'EI multilingue ont été introduites. Les dernières MUCs ont défini plusieurs sous-tâches d'EI dans le but de faciliter l'évaluation et l'identification de sous-composants d'EI utilisables immédiatement. Les tâches d'EI génériques définies dans MUC-7 (1998) ont fourni progressivement des informations de haut niveau sur les textes.[7]

1.5 Tâches d'extraction d'information

Selon la complexité des phrases, l'extraction d'informations peut être classée en quatre catégories :

- **Reconnaissance des entités nommées:**

La reconnaissance des entités nommées (NER) - également appelée identification des entités ou extraction des entités - est une technique de traitement du langage naturel (NLP) qui identifie automatiquement les entités nommées dans un texte et les classes dans des catégories prédéfinies. Les entités peuvent être des noms de personnes, d'organisations, de lieux, d'heures, de quantités, de valeurs monétaires, de pourcentages, etc.

Avec la reconnaissance des entités nommées, vous pouvez extraire des informations clés pour comprendre le sujet d'un texte ou simplement l'utiliser pour recueillir des informations importantes à stocker dans une base de données.[8]

Exemple :

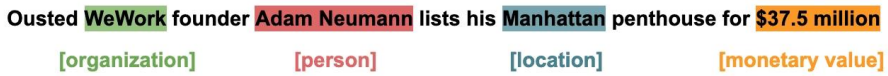

Figure 2 Exemple de NER

- **Résolution de coréférence :**

La résolution de coréférence (RC) consiste à trouver toutes les expressions linguistiques (appelées mentions) dans un texte donné qui font référence à la même entité du monde réel. Après avoir trouvé et regroupé ces mentions, nous pouvons les résoudre en remplaçant, comme indiqué ci-dessus, les pronoms par des phrases nominales.

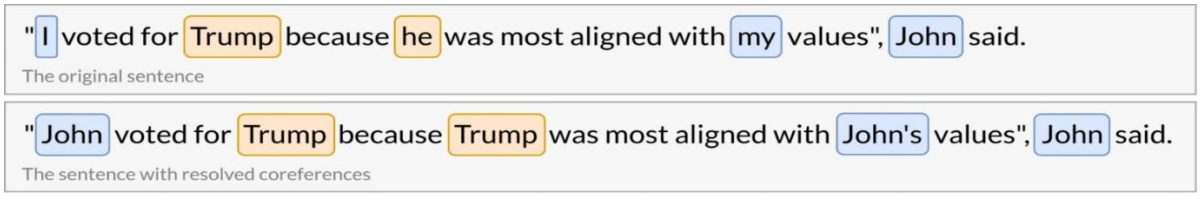

Figure 3 Exemple de RC

- **Extraction de relations :**

La tâche d'extraction des relations sémantiques entre les entités dans le texte est appelée Relation Extraction (RE). Alors que Named Entity Recognition (NER) consiste à identifier des entités dans le texte, RE consiste à trouver les relations entre les entités. Étant donné un texte non structuré, NER et RE nous aident à obtenir des représentations structurées utiles. Les deux tâches font partie de la discipline de l'extraction d'information (IE)[9]

- **Extraction d'événements :**

L'extraction d'événements (EE) constitue une tâche difficile dont le but est d'identifier rapidement les événements et leurs entités dans un grand nombre de documents. Un événement est décrit par un ensemble de participants (c'est-à-dire des attributs ou des rôles) dont les valeurs sont des extraits de texte. L'EE implique d'identifier les instances de types d'événements spécifiés dans le texte et les arguments qui leur sont associés. Chaque événement est représenté par une expression, une phrase ou un segment de texte, le déclencheur de l'événement (le plus souvent des verbes simples ou des verbes à particule, mais aussi des noms, des noms à particule, des pronoms et des adverbes), qui évoque cet événement. Après la détection et la classification des déclencheurs, il faut trouver les arguments de l'événement. Les arguments de l'événement sont des mentions d'entités ou des expressions temporelles qui sont impliquées dans un événement (en tant que participants).[10]

1.6 Application pour l'extraction d'informations

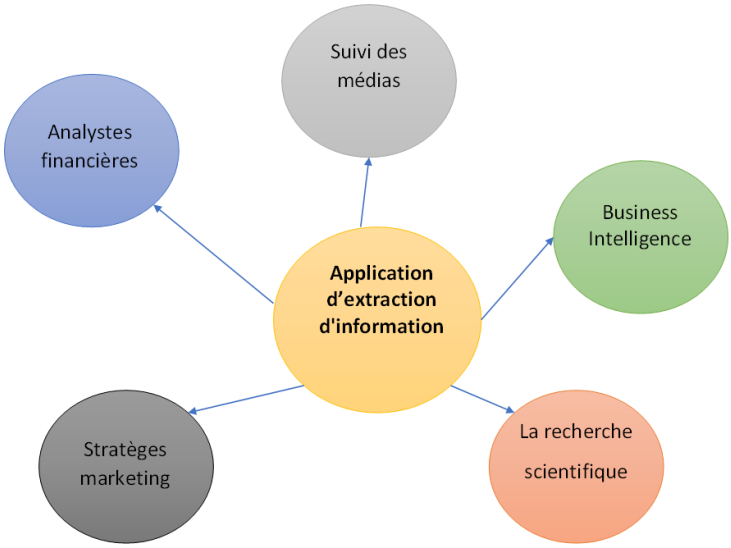

Figure 4 Application pour l'extraction d'informations

1.7 Pipeline général du processus d'extraction d'informations

Les étapes suivantes sont souvent impliquées dans l'extraction d'informations structurées à partir de textes non structurés [11]:

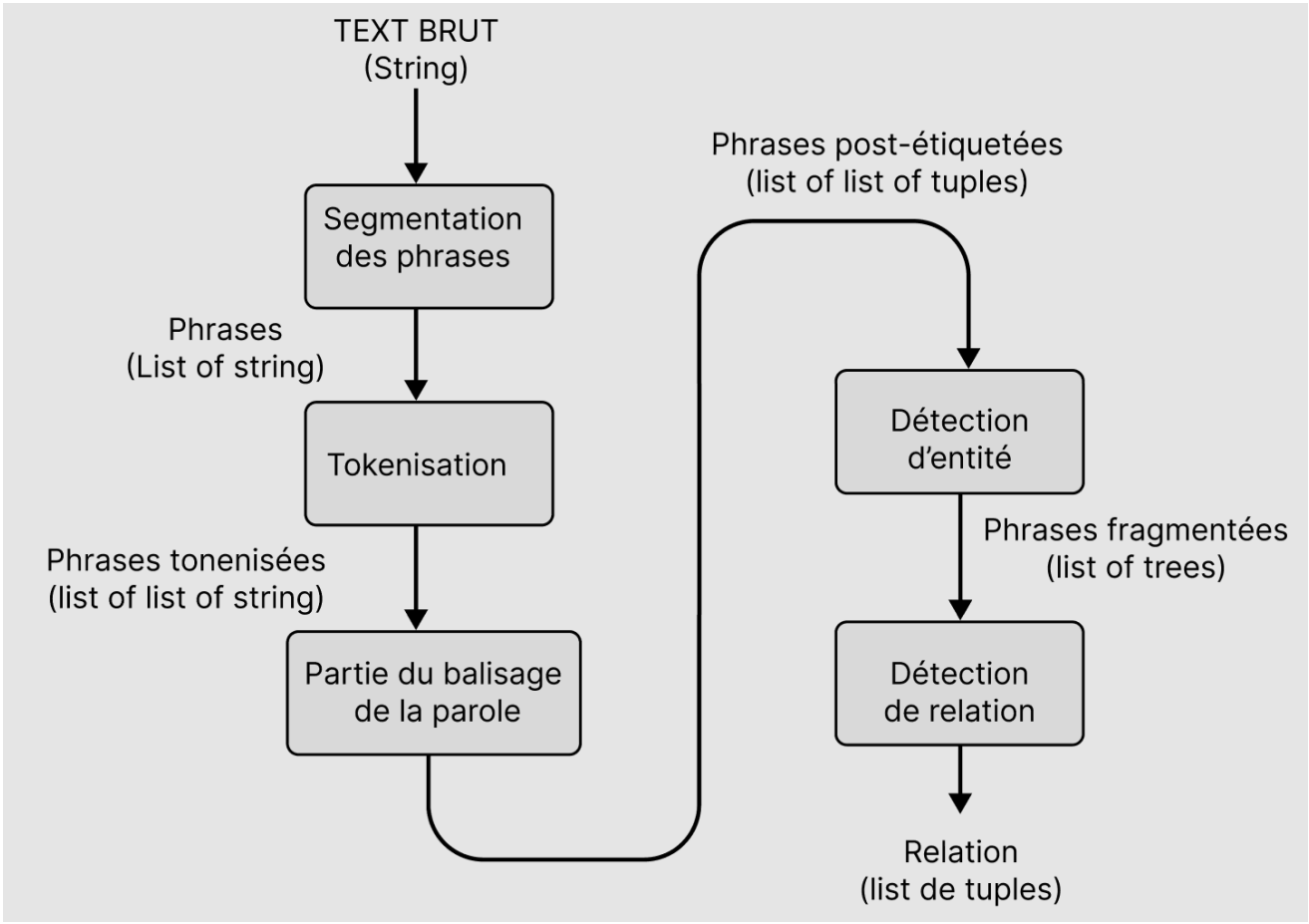

Figure 5 Processus d'extraction d'informations

1. Traitement initial : La première étape consiste à décomposer un texte en fragments tels que des zones, des phrases, des segments et des tokens. Cette fonction peut être assurée par des tokenizers, des zoners de texte, des segmenters et des splitters, entre autres composants. Au cours de l'étape initiale de traitement, l'étiquetage des parties du discours, l'identification des unités phrastiques (noms ou verbes) et l'étiquetage des parties du discours sont généralement les tâches suivantes.

2. Identification correcte des noms : L'une des étapes les plus importantes de la chaîne d'extraction d'informations est l'identification de diverses classes de noms propres, tels que les noms de personnes ou d'organisations, les dates, les montants monétaires, les lieux, les adresses, etc. Ils peuvent être trouvés dans pratiquement n'importe quel type de texte et sont largement utilisés dans le processus d'extraction. Les expressions régulières, qui sont un ensemble de motifs, sont utilisées pour reconnaître ces noms.

3. Analyse syntaxique : L'analyse syntaxique des phrases dans les textes est effectuée à cette étape. Après avoir reconnu les entités fondamentales à l'étape précédente, les phrases sont traitées pour trouver les groupes de noms qui entourent certaines de ces entités et les groupes de verbes. Lors de l'étape de comparaison de motifs, les groupes de noms et de verbes sont utilisés comme sections sur lesquelles commencer à travailler.

4. Extraction des événements et des relations : Cette étape établit des relations entre les idées extraites. Ceci est accompli en développant et en mettant en œuvre des règles d'extraction qui décrivent divers modèles. Le texte est comparé à certains modèles, et si une correspondance est découverte, l'élément de texte est étiqueté et récupéré ultérieurement.

5. Résolution des anaphores : La résolution de coréférence est utilisée pour identifier toutes les façons dont l'entité est nommée dans le texte. L'étape où l'on décide si les phrases nominales se rapportent ou non à la même entité est appelée résolution de coréférence ou d'anaphore.

6. Génération de résultats de sortie : Cette étape consiste à convertir les structures recueillies au cours des processus précédents en modèles de sortie qui suivent le format défini par l'utilisateur. Elle peut comprendre une variété de processus de normalisation.

1.8 Défis de l'extraction d'information

Le domaine de l'extraction d'informations contient un certain nombre de défis, notamment :

- La difficulté de développer un programme qui peut gérer différents types de texte
- L'extraction d'information prend beaucoup de temps et des ressources
- La transformation des données non structurées en un format structuré pour une meilleure représentation est la grande question.

1.9 Architecture d'un système d'EI

La tâche d'extraction est réalisée grâce au remplissage de formulaires prédéfinis (Template). Ces formulaires, dits formulaires d'extraction, sont définis dans le but de représenter la connaissance à rechercher par une structure déterminée a priori. Ils décrivent un ensemble d'entités, les relations entre celles-ci et les événements impliquant ces entités[4]. La structure de cette architecture comme sur le schéma dans la figure 7 ci-dessous, texte brut représente l'ensemble des documents utilisés dans l'EI, le programme représente un ensemble de processus utiliser pour l'EI, la sortie sans des champs remplis.

- Par exemple, un formulaire concernant des accidents de la route devra spécifier des champs comme « Lieu de l'accident », « Nombre de victimes », « Identité des victimes » ou encore « Cause de l'accident ». Les informations extraites par un système d'Extraction d'Information peuvent être consultées par des utilisateurs humains (par exemple via la génération de rapports d'événements).
  Un exemple d'extraction de faits de guerre à partir d'articles de journaux est présenté dans l'exemple 1.1.[12]

Exemple 1.1 (Extraction d'Information sur un extrait du journal *Libération*)

Texte :

Libération - lundi 27 octobre 2003 - Bagdad envoyé spécial -- Réveil agité hier à 6 heures du matin pour Paul Wolfowitz, le numéro 2 du Pentagone, qui passait la nuit à Bagdad dans l'hôtel Al-Rashid, transformé en bunker par les forces d'occupation américaines. Au moins six roquettes Katioucha, tirées depuis une remorque stationnée à 400 mètres de là, ont atteint la façade de l'hôtel de luxe. L'attaque a tué un soldat américain et blessé quinze autres personnes, en majorité des Américains.

Formulaire rempli :

Événement	:	attaque
Nature	:	tir de roquettes
Date	:	27 octobre 2003
Lieu	:	hôtel Al-Rashid, Bagdad, Irak
Cible	:	Paul Wolfowitz
Victimes	:	**Mort** : un soldat américain
		Blessés : quinze personnes, en majorité des Américains

Figure 6 Exemple sur l'EI

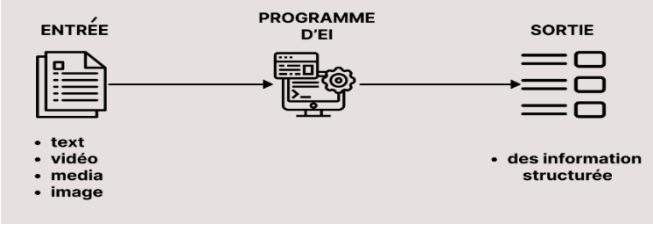

Figure 7 Architecture d'un système d'EI

1.10 La relation entre extraction d'information et apprentissage automatique

De nombreuses techniques sont utilisées dans le domaine de l'extraction d'informations, dont les plus importantes sont les algorithmes d'apprentissage automatique.

ML est un sous-domaine essentiel de l'intelligence artificielle qui conçoit et améliore la capacité de comportement humain intelligent en élargissant les connaissances et en résolvant des problèmes jamais rencontrés auparavant dans la majorité des systèmes IE. Le réseau de neurone et la machine à vecteurs de support (SVM) et la technologie de forêt aléatoire sont déjà utilisées dans IE, et elles ont un bon potentiel pour de nouvelles utilisations supplémentaires dans la recherche et les applications IE.

Divers outils ML qui correspondent à différents types de travail IE, qui sont classés en trois modes : ceux qui incluent l'exploration corporelle supervisée, semi-supervisée et non supervisée.

1.11 Conclusion

Nous avons apporté une description sur l'extraction d'informations par la présentation les principales tâches d'EI, l'architecture d'EI. Nous avons parlé des défis de ce domaine en intéressant à l'extraction à textes à partir des fichiers.

Le chapitre suivant présente en détail les algorithmes de Machine Learning.

CHAPITRE 2 :
Apprentissage automatique

2.1 Introduction

L'apprentissage automatique est un domaine qui a vu son utilisation croître avec les progrès technologiques. Toutefois, son utilisation reste conditionnée par une bonne maîtrise de ses fondements théoriques. C'est à ces fondements que cette section est consacrée.

2.2 Définition de L'intelligence artificielle

L'intelligence artificielle est l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine, ce qui permet de faire des logiciels ou bien des algorithmes qui rend l'ordinateur ou des machines dotées de programmes capables de performances similaires à l'intelligence humaine, ou même, amplifiées par la technologie. Ces machines sont en mesure de:

• Raisonner
• Traiter de grandes quantités de données
• Discerner des modèles indétectables par l'œil d'un humain
• Comprendre et analyser ces modèles
• Interagir avec l'Homme
• Apprendre progressivement
• Améliorer continuellement ses performances.[13]

2.3 Définition de l'apprentissage automatique

Le terme apprentissage automatique a été inventé par le pionnier américain des jeux informatiques et de l'intelligence artificielle Arthur chez IBM en 1959 en tant que projet scientifique.

L'apprentissage automatique est la programmation des ordinateurs pour optimiser une performance critère utilisant des données d'exemple ou une expérience passée. Nous avons un modèle défini jusqu'à certains paramètres et l'apprentissage est l'exécution d'un programme informatique pour optimiser les paramètres du modèle en utilisant les données de formation ou expérience passée. Le modèle peut être prédictif pour faire

des prédictions dans la futur, ou descriptif pour acquérir des connaissances des données, ou les deux.[14]

L'objectif visé est de rendre la machine ou l'ordinateur capable d'apporter des solutions à des problèmes compliqués, par le traitement d'une quantité astronomique d'informations. Cela offre ainsi une possibilité d'analyser et de mettre en évidence les corrélations qui existent entre deux ou plusieurs situations données, et de prédire leurs différentes implications.[15]

2.4 Définition de Classification

La classification est un processus de catégorisation d'un ensemble donné de données en classes. Elle peut être effectuée sur des données structurées ou non structurées. Le processus commence par prédire la classe de points de données donnés. Les classes sont souvent appelées cible, étiquette ou catégories .[16]

2.5 Définition de Régression

L'analyse de régression est un concept fondamental dans le domaine de l'apprentissage automatique . Il relève de l'apprentissage supervisé dans lequel l'algorithme est formé à la fois avec des caractéristiques d'entrée et des étiquettes de sortie. Il aide à établir une relation entre les variables en estimant comment une variable affecte l'autre. [17]

2.6 Définition de Regroupement

Le regroupement consiste à diviser la population ou les points de données en plusieurs groupes, de sorte que les points de données des mêmes groupes soient plus similaires aux autres points de données du même groupe qu'à ceux des autres groupes. En termes simples, l'objectif est de séparer les groupes ayant des traits similaires et de les affecter en clusters.[18]

2.7 Application d'apprentissage automatique

Des milliers d'applications qui utilisent l'apprentissage automatique, notamment Facebook et Netflix, Google traduction, et d'autre applications médicales ...etc. {figure 8}

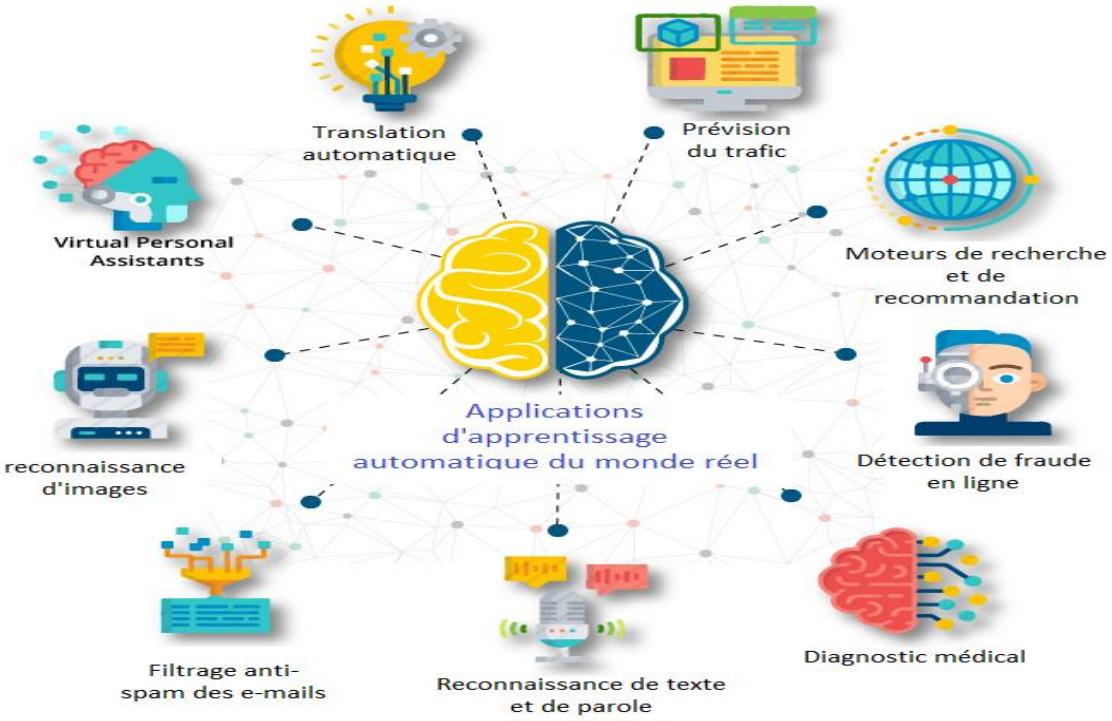

Figure 8 Une image montrant les applications les plus importantes de l'apprentissage automatique

2.8 Les types de l'apprentissage automatique

En général, Les principaux types d'algorithmes de l'apprentissage automatique sont utilisés aujourd'hui : l'apprentissage supervisé et l'apprentissage non supervisé et l'apprentissage automatique par renforcement.

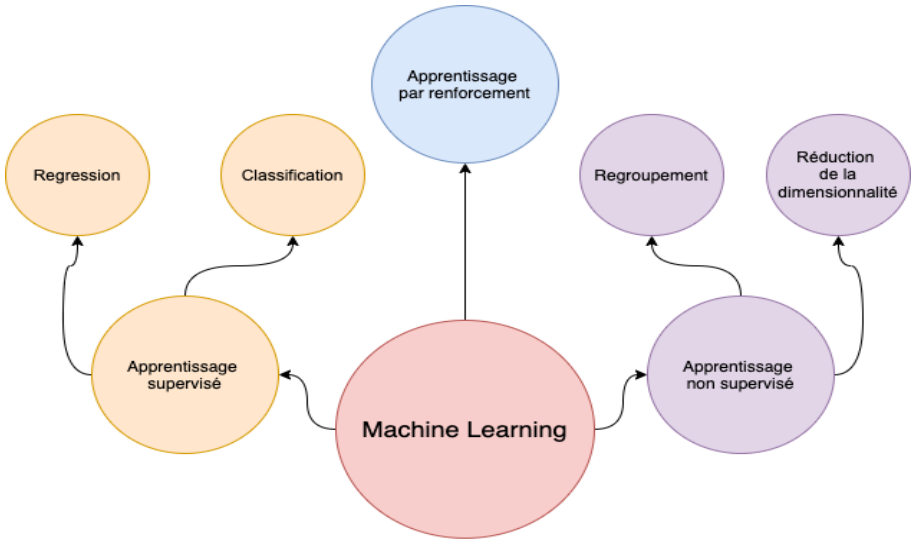

Figure 9 Les types d'apprentissage automatique

2.8.1 Apprentissage supervisé

L'apprentissage supervisé consiste à entraîner un modèle en lui fournissant la réponse (label). Cette réponse permet de superviser l'apprentissage du modèle en lui disant à quel point il est loin de la bonne réponse. Dans un apprentissage supervisé, nous avons un X (variable indépendante) et un Y (variable dépendante) lors de l'entraînement. Cette catégorie se divise en 2 sous catégories principales soit la classification et la régression.[19]

- Ce type d'apprentissage contient des algorithmes suivants :
  - Régression linéaire
  - Régression logistique
  - Arbres de décision
  - K-NN
  - Machines à vecteurs de support
  - Random Forest

2.8.1.1 Régression linéaire

La régression linéaire se classe parmi les méthodes d'analyses multi variées qui traitent des données quantitatives. C'est une méthode d'investigation sur données d'observations, ou d'expérimentations, où

l'objectif principal est de rechercher une liaison linéaire entre une variable Y quantitative et une ou plusieurs variables X également quantitatives.[20]

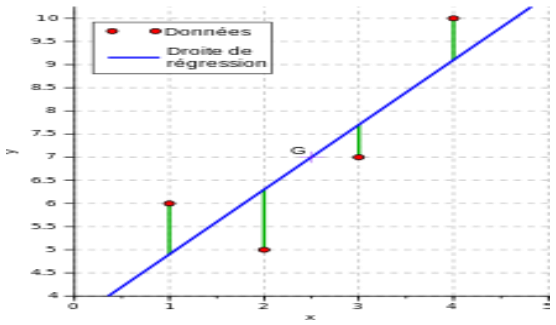

Figure 10 Régression linéaire

2.8.1.2 La régression logistique

La régression logistique est utilisée pour traiter un problème de classification. Elle donne le résultat binomial car elle donne la probabilité qu'un événement se produise ou non (en termes de 0 et 1) en fonction des valeurs des variables d'entrée. Par exemple, prédire si une tumeur est maligne ou bénigne ou si un e-mail est classé comme spam ou non sont les cas qui peuvent être considérés comme des résultats binomiaux. Résultat de la régression logistique.[21]

2.8.1.3 Les arbres de décision

Les arbres de décision sont une des techniques les plus populaires de l'apprentissage automatique et de la fouille de données. L'apprentissage par arbre de décision se situe dans le cadre de l'apprentissage supervisé, où la classe de chaque objet dans la base est donnée. Le but est de construire un modèle à partir d'un ensemble d'exemples associés aux classes pour trouver une description pour chaque classe à partir des propriétés communes entre les exemples. Une fois ce modèle construit, on peut extraire un ensemble de règles de classement. Ce modèle ou les règles extraites sont ensuite utilisés pour classer de nouveaux objets dont la classe est inconnue. Le classement se fait en parcourant un chemin depuis la racine jusqu'à à une feuille. La classe renvoyée est celle qui est la plus fréquente parmi les exemples de la feuille.[22] voir l'exemple ci-dessous [23]:

A	B	A AND B
F	F	F
F	T	F
T	F	F
T	T	T

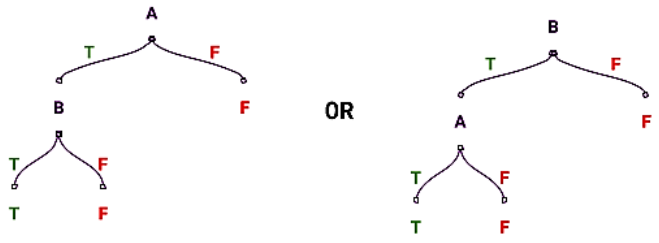

Figure 11 Example of Decision tree for an AND operation.

2.8.1.4 Machines à vecteurs de support

Machines à vecteurs de support (SVM) est un algorithme d'apprentissage automatique supervisé qui peut être utilisé pour les problèmes de classification ou de régression. Cependant, il est surtout utilisé pour les problèmes de classification.

Dans l'algorithme SVM, nous représentons chaque élément de données comme un point dans un espace à n-dimensions (où n est le nombre de caractéristiques que vous avez), la valeur de chaque caractéristique étant la valeur d'une coordonnée particulière. Ensuite, nous effectuons la classification en trouvant l'hyperplan qui différencie très bien les deux classes (regardez l'image ci-dessous).[24]

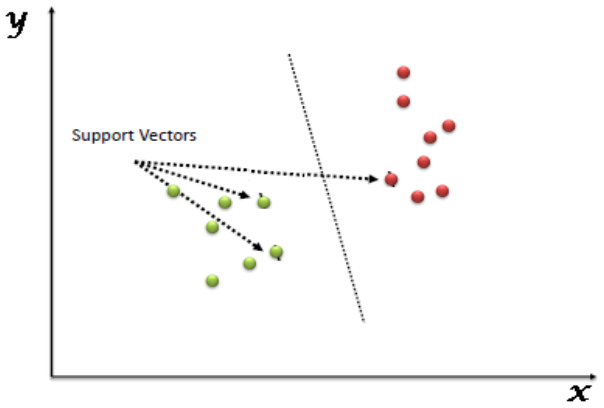

Figure 12 Machines à vecteurs de support modèle

Les vecteurs de support sont simplement les coordonnées d'une observation individuelle. Le classifieur SVM est une frontière qui sépare le mieux les deux classes (hyper-plan/ligne). [24]

2.8.1.5 K-Neirest Neighbors

L'algorithme K-NN figure parmi les plus simples algorithmes d'apprentissage artificiel. Dans un contexte de classification d'une nouvelle observation **x**, l'idée fondatrice simple est de faire voter les plus proches voisins de cette observation. La classe de **x** est déterminée en fonction de la classe majoritaire parmi les k plus proches voisins de l'observation **x**.

La méthode K-NN est donc une méthode à base de voisinage, non-paramétrique ; Ceci signifiant que l'algorithme permet de faire une classification sans faire d'hypothèse sur la fonction $y=f(x_{1, x2}, \ldots x_p)$ qui relie la variable dépendante aux variables indépendantes [25], regardez l'image ci-dessous [26].

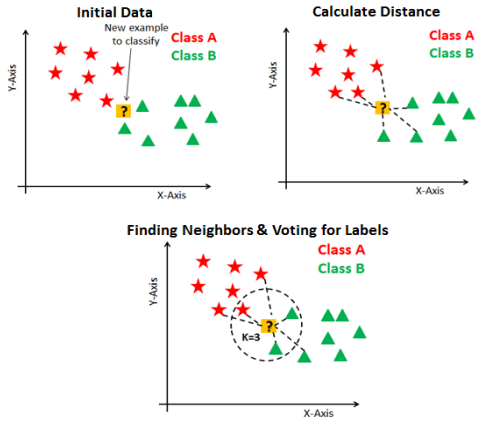

Figure 13 K-nearest neighbors

2.8.1.6 Random Forest

Random Forest est une méthode d'ensemble qui consiste en de nombreux arbres de décision dont les résultats sont agrégés pour donner une prédiction. Il y a deux paramètres dans une forêt aléatoire, l'un est L qui contrôle la taille de l'ensemble et l'autre est mtry qui représente le nombre de caractéristiques différentes sélectionnées aléatoirement à chaque nœud. Chaque arbre de la forêt est construit en utilisant l'algorithme suivant :

1. Construire un ensemble d'entraînement Bootstrap à partir de tous les échantillons d'entraînement pour générer chacun des arbres de décision de la forêt. Nous échantillonnons généralement le même nombre

d'échantillons d'entraînement avec remplacement.
2. À chaque nœud de l'arbre, sélectionnez aléatoirement mtry différentes caractéristiques à partir de l'ensemble de formation pour calculer la meilleure division à ce nœud.
3. Chaque arbre sera entièrement développé sans aucun élagage.
Ensuite, pour un échantillon de test, chaque arbre de décision de la forêt prédit son étiquette de classe. Enfin, l'ensemble produira l'étiquette de classe ayant reçu le plus de votes.[27]

2.8.2 Apprentissage non supervisé

L'apprentissage non supervisé consiste à entraîner un modèle à trouver les caractéristiques et extraire les relations entre les données. Dans un problème non supervisé, nous n'avons pas la réponse exacte que le modèle devrait trouver, nous avons seulement des données entrantes et donc le modèle est en quelque sorte « laissé à lui-même». En apprentissage non supervisé, nous avons seulement un X (variable indépendante) lors l'entraînement. L'apprentissage non supervisé se divise lui aussi en 2 principales catégories soit le regroupement (clustering) et la réduction de la dimensionnalité (dimensionnalité réduction).[19]

- ce type d'apprentissage contient l'algorithme suivant :
  - K-means

2.8.2.1 K-means

K-means est l'un des algorithmes d'apprentissage non supervisé les plus simples qui résolvent le problème de clustering bien connu. La procédure suit une manière simple et facile de classer un ensemble de données donné à travers un certain nombre de clusters. L'idée principale est de définir k centres, un pour chaque cluster. Ces centres doivent être placés de manière astucieuse car un emplacement différent entraîne un résultat différent. Ainsi, le meilleur choix est de les placer le plus loin possible les uns des autres.[28]

L'algorithme du kmeans est un algorithme itératif qui minimise la somme des distances entre chaque individu et le centroïde du cluster ; c'est la variabilité intra cluster.[29]
Le principe est :

1. Attribuer un cluster à chaque objet (ou sujet, ou point), de façon aléatoire.

2. Calculer le centroïde de chaque cluster (c'est-à-dire le vecteur des moyennes des différentes variables).
3. Pour chaque objet (ou sujet ou point) calculer sa distance euclidienne avec les centroides de chacun des clusters.
4. Attribuer à l'objet le cluster le plus proche de lui.
5. Calculer la somme de la variabilité intra-cluster.
6. Recommencer les étapes 2 à 5, jusqu'à atteindre un équilibre, on parle de convergence : plus aucun changement de clusters, ou stabilisation de la somme de la variabilité intra-cluster.

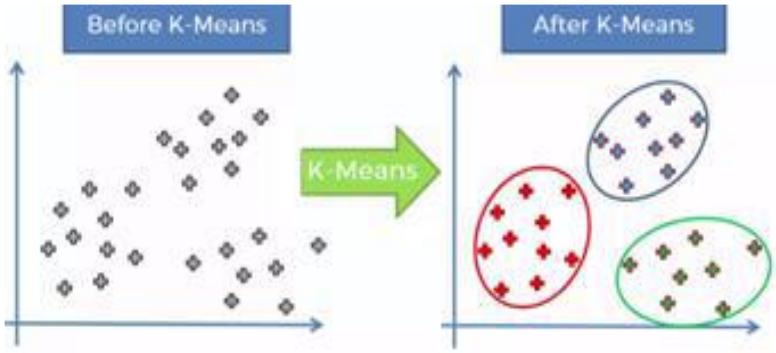

Figure 14 K-Means Clustering

2.8.3 Apprentissage par renforcement :

L'apprentissage par renforcement consiste en un agent qui interagit avec son environnement.

Voici une illustration qui montre le fonctionnement de l'apprentissage par renforcement. [19]

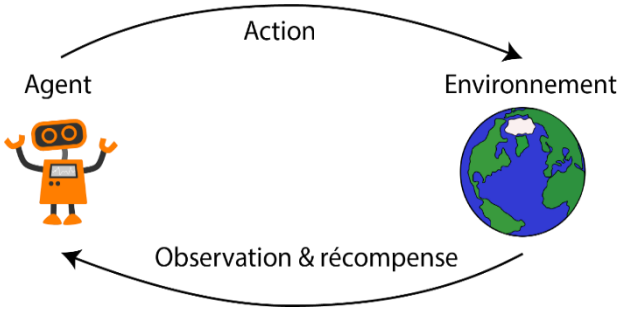

Figure 15 Apprentissage par renforcement

2.9 L'apprentissage supervisé contre non- supervisé

Paramètres	Technique d'apprentissage automatique supervisé	Technique d'apprentissage automatique non supervisée
Processus	Dans un modèle d'apprentissage supervisé, des variables d'entrée et de sortie seront données.	Dans le modèle d'apprentissage non supervisé, seules les données d'entrée seront données
Des données d'entrée	Les algorithmes sont formés à l'aide de données étiquetées.	Les algorithmes sont utilisés contre des données qui ne sont pas étiquetées
Algorithmes utilisés	Prise en charge de la machine vectorielle, du réseau neuronal, de la régression linéaire et logistique, de la forêt aléatoire et des arbres de classification.	Les algorithmes non supervisés peuvent être divisés en différentes catégories : comme les algorithmes de cluster, les K-means, le clustering hiérarchique, etc.
Complexité informatique	L'apprentissage supervisé est une méthode plus simple.	L'apprentissage non supervisé est complexe sur le plan informatique
Utilisation des données	Le modèle d'apprentissage supervisé utilise des données d'apprentissage pour apprendre un lien entre l'entrée et les sorties.	L'apprentissage non supervisé n'utilise pas de données de sortie.
Exactitude des résultats	Méthode très précise et fiable	Méthode moins précise et moins fiable.
Apprentissage en temps réel	La méthode d'apprentissage se déroule hors ligne.	La méthode d'apprentissage se déroule en temps réel.
Nombre de cours	Le nombre de classes est connu.	Le nombre de classes n'est pas connu.
Inconvénient principal	Classifier les métadonnées peut être un véritable défi en apprentissage	Vous ne pouvez pas obtenir d'informations précises concernant

	supervisé.	le tri des données, et la sortie en tant que données utilisées dans l'apprentissage non supervisé est étiquetée et inconnue.

Tableau 1 L'apprentissage supervisé VS non- supervisé

2.10 Conclusion

A cette partie, nous avons introduit les concepts de base de l'apprentissage automatique avec ses types et ses applications. Et nous avons détaillé l'apprentissage automatique avec ses types

Dans le chapitre suivant nous allons présenter le domaine d'étude sur lequel, on va travailler dans la partie d'implémentation, en étudiant le contexte.

CHAPITRE 3 : Système de détection d'intrusion

CHAPITRE 3 : SYSTEME DE DETECTION D'INTRUSION

3.1 Introduction

Les systèmes informatiques fonctionnent dans des environnements hautement dynamiques et distribués. Cependant, ils exigent des mécanismes de protection pour contrer les attaques, pour empêcher la violation intentionnelle ou involontaire aux politiques de sécurité, qui ont pour objectif d'assurer la disponibilité des services, la confidentialité et l'intégrité des données et des échanges.[30]
Dans ce chapitre, nous présentons un des mécanismes qui assure la sécurité informatique, c'est IDS.

3.2 Sécurité informatique

3.2.1 Définitions

La sécurité des SI est un concept qui couvre l'ensemble des méthodes, techniques et outils mis en œuvre pour la protection des ressources d'un système d'information contre les sinistres, les erreurs et les malveillances dans le but de rendre leur probabilité et/ou leur conséquences compatibles avec les exigences de sécurité dégagées.[31]

La sécurité informatique c'est l'ensemble des moyens mis en œuvre pour réduire la vulnérabilité d'un système contre les menaces accidentelles ou intentionnelles.[32]

3.2.2 Les critères de la sécurité informatique

La sécurité informatique vise généralement cinq principaux objectifs [33]:

- **La confidentialité**

La confidentialité consiste à rendre l'information inintelligible à d'autres personnes que les seuls acteurs de la transaction.

- **L'intégrité**

L'intégrité des données consiste à déterminer que les données n'ont pas été altérées durant la communication (de manière fortuite ou intentionnelle).

- **La disponibilité**

L'objectif de disponibilité est de garantir l'accès à un service ou à des ressources.

- **La non-répudiation**

La non-répudiation de l'information est la garantie qu'aucun des correspondants ne pourra nier la transaction.

- **L'authentification**

L'authentification consiste à assurer l'identité d'un utilisateur, c'est-à-dire de garantir à chacun des correspondants que son partenaire est bien celui qu'il croit être. Un contrôle d'accès peut permettre (par exemple par le moyen d'un mot de passe qui devra être crypté) l'accès à des ressources uniquement aux personnes autorisées.

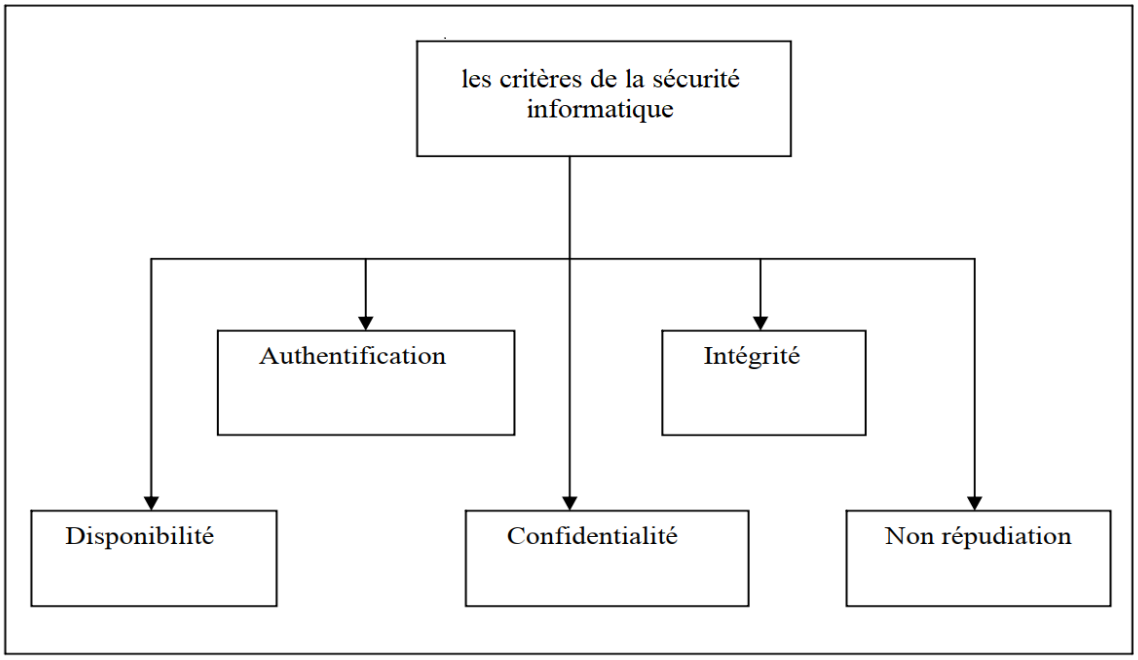

Figure 16 Les critères de la sécurité informatique

3.2.3 Quelques mécanismes pour la sécurité informatique

3.2.3.1 Les pare-feu

Un pare-feu est un appareil de protection du réseau qui surveille le trafic entrant et sortant et décide d'autoriser ou de bloquer une partie de ce trafic en fonction d'un ensemble de règles de sécurité prédéfinies.[34]

3.2.3.2 Le chiffrement

Le chiffrement est un procédé de cryptographie qui consiste à protéger des données qui sont alors incompréhensibles pour celui qui ne dispose pas de la clef du chiffrement. Le chiffrement des données a pour objectif de garantir la confidentialité des données stockées sur des systèmes informatiques (SI) ou en transit. Les données sont chiffrées à l'aide d'un algorithme et d'un jeu de clefs de chiffrement.[35]

3.2.3.3 Journalisation

En termes simples, un journal système est une collection d'enregistrements individuels qui représentent une activité spécifique, des événements, des conditions d'erreur, des défauts ou un état général sur un système d'information ou un réseau. Ces entrées de journal contiennent des données critiques qui aident les administrateurs du système et de la sécurité à comprendre ce qui se passe sur un système d'information.[36]

3.2.3.4 L'antivirus

L'antivirus est un type de logiciel utilisé pour prévenir, analyser, détecter et supprimer les virus d'un ordinateur. Une fois installés, la plupart des logiciels antivirus fonctionnent automatiquement en arrière-plan pour assurer une protection en temps réel contre les attaques de virus.[37]

3.2.3.5 Le système de détection d'intrusion

Un système de détection d'intrusion (IDS) est une application logicielle ou un appareil matériel qui surveille le trafic circulant sur les réseaux et à travers les systèmes pour rechercher des activités suspectes et des menaces connues, et envoie des alertes lorsqu'il trouve de tels éléments.[38]

3.3 Système de détection d'intrusion

3.3.1 Définition

Un système de détection d'intrusion (IDS) est un système qui surveille le trafic réseau à la recherche d'activités suspectes et émet des alertes lorsqu'une telle activité est découverte. Il s'agit d'une application logicielle qui scanne un réseau ou un système à la recherche d'une activité nuisible ou d'une violation de politique. Toute activité malveillante ou violation est normalement signalée soit à un administrateur, soit collectée de manière centralisée à l'aide d'un système de gestion des informations et des événements de sécurité .[39]

Un système de détection d'intrusions (ou IDS : Intrusion Detection System) est un logiciel ou un matériel qui automatise la surveillance et l'analyse des événements se trouvant dans un système ou dans un réseau.[40]

3.3.2 Historique

Le concept de système de détection d'intrusions a été introduit en 1980 par James Anderson [41]. Mais le sujet n'a pas eu beaucoup de succès. Il a fallu attendre la publication d'un modèle de détection d'intrusions par Denning en 1987 [42] pour marquer réellement le départ du domaine. La recherche dans le domaine s'est ensuite développée, le nombre de prototypes s'est énormément accru. Le gouvernement des Etats-Unis a investi des millions de dollars dans ce type de recherches dans le but d'accroître la sécurité de ses machines.[43]

3.3.3 Terminologies et concepts de base

Attaque : Une Attaque est une tentative malveillante et délibérée par un individu ou une organisation de violer le système d'information d'un autre individu ou organisation.[44]

Intrusion : événement ou combinaison d'événements permettant d'avoir indûment accès (sans autorisation) à un système et ses ressources.[45]

Vulnérabilité : défaut ou faiblesse dans la conception d'un système, son implémentation, fonctionnement ou administration et qui pourrait être exploité pour violer la politique de sécurité.[46]

Menace : danger potentiel pouvant exploiter une vulnérabilité pour violer la politique de sécurité causant éventuellement des dégâts. Une menace peut être réelle ou non fondée[46]

Détection d'intrusion : La détection d'intrusions est le processus de suivi des événements survenus dans un système ou réseau informatique, afin d'analyser les signes d'intrusions, vus comme des tentatives de compromettre la confidentialité, l'intégrité, la disponibilité, ou de contourner les mécanismes de sécurité d'un ordinateur ou d'un réseau.[30]

3.3.4 Architecture de l'IDS

Nous décrivons les composants qui constituent classiquement un système de détection d'intrusion. La Figure ci-dessous illustre les interactions entre ces composants. [47]

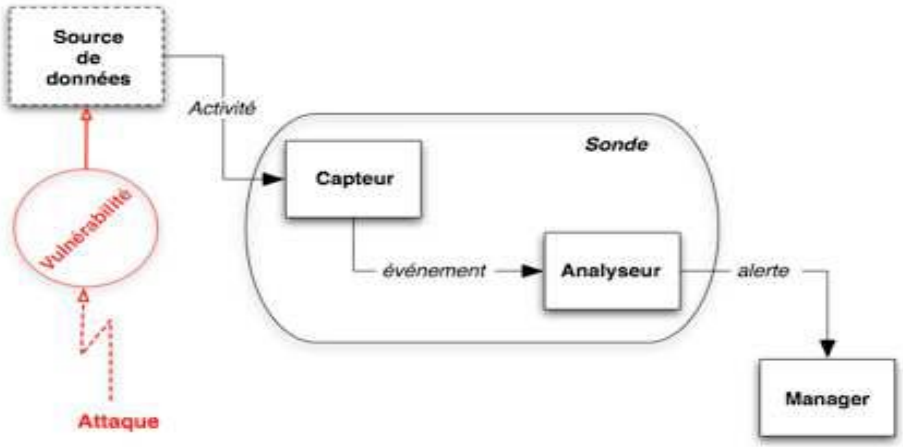

Figure 17 Architecture de l'IDS.

- **Source de données :** dispositif générant de l'information sur les activités des entités du système d'information.

- **Capteur :** génère des événements en filtrant et formatant les données brutes provenant d'une source de données.

- **Evénement :** message formaté et renvoyé par un capteur. C'est l'unité élémentaire utilisée pour représenter une étape d'un scénario d'attaques connu.

- **Analyseur :** c'est un outil logiciel qui met en œuvre l'approche choisie pour la détection (comportementale ou par scénarios), il génère des alertes lorsqu'il détecte une intrusion.

- **Sonde :** un ou des capteurs couplés avec un analyseur.

- **Alerte :** message formaté émis par un analyseur s'il trouve des activités intrusives dans une source de données.

- **Manager : est** responsable de la présentation des alertes a l'operateur (fonction de console de management).

3.3.5 Classification des systèmes de détection d'intrusion

Il existe plusieurs critères qu'on peut utiliser pour classifier les différents systèmes de détection d'intrusion, dont les principaux sont résumés dans Figure ci-dessous [48]:

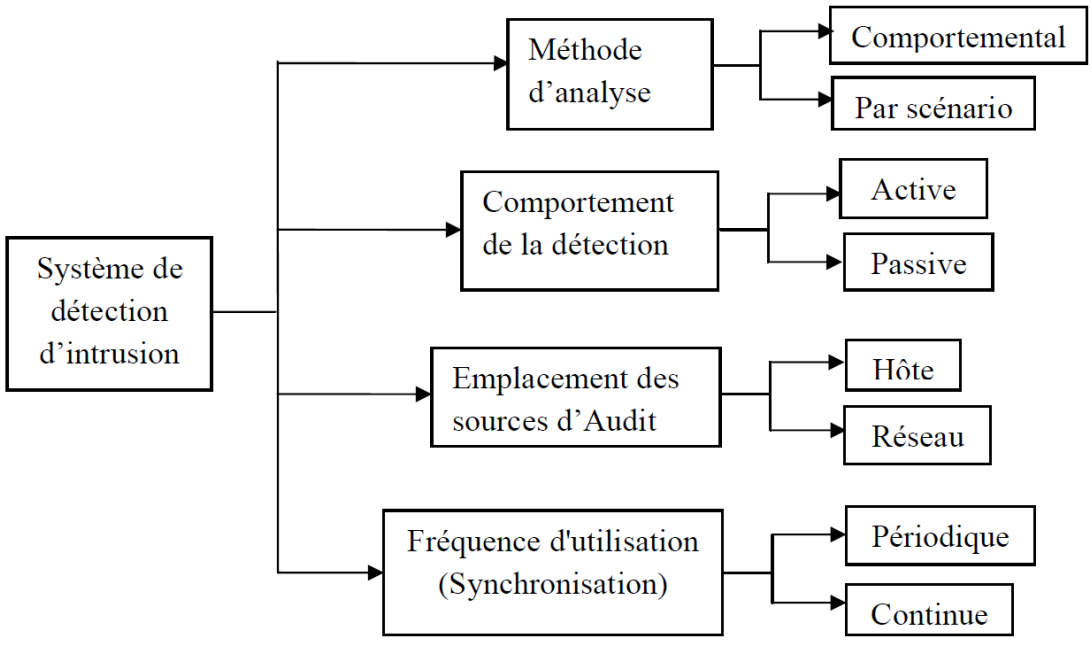

Figure 18 Taxonomie des systèmes de détection d'intrusion.

3.3.5.1 La méthode d’analyse

3.3.5.1.1 La détection par comportementale

L’approche comportementale, c'est-à-dire qu'on va chercher à savoir si un utilisateur a eu un comportement déviant par rapport à ses habitudes. Ceci signifierait qu'il essaye d'effectuer des opérations qu'il n'a pas l'habitude de faire. On peut en déduire, soit que c'est quelqu'un d'autre qui a pris sa place, soit que lui-même essaye d'attaquer le système en abusant de ses droits. Dans les deux cas, il y a intrusion.[49]

3.3.5.1.2 La détection par scénario

Les attaques connues sont répertoriées et les actions indispensables de cette attaque forment sa signature. On compare ensuite les actions effectuées sur le système avec ces signatures d'attaques. Si on retrouve une signature d'attaque dans les actions d'un utilisateur, on peut en déduire qu'il tente d'attaquer le système par cette méthode.[49]

3.3.5.2 Comportement de la détection

Un IDS réagit après avoir détecté une attaque, et la réponse peut être passive ou active. Une réponse passive consiste principalement à enregistrer et à notifier le personnel, tandis qu'une réponse active modifie également l'environnement pour bloquer l'attaque :[50]

3.3.5.2.1 IDS passif

Un IDS passif enregistre l'attaque et peut également déclencher une alerte pour prévenir quelqu'un. La plupart des IDS sont passifs par défaut. La notification peut prendre plusieurs formes, notamment un courriel, un message texte, une fenêtre contextuelle ou une notification sur un moniteur central.

3.3.5.2.2 IDS actif

Un IDS actif enregistre et notifie le personnel comme le fait un IDS passif, mais il peut également modifier l'environnement pour déjouer ou bloquer l'attaque.

3.3.5.3 Les types d'IDS

3.3.5.3.1 IDS réseau

3.3.5.3.1.1 Définition

Un système de détection d'intrusions réseau (N-IDS), permet de capturer et d'écouter ce qui transite dans le réseau. Il pose des capteurs à des endroits stratégiques, -pour contrôler un grand nombre de paquets et pouvoir surveiller plusieurs hôtes à la fois-, en gardant toujours un fonctionnement furtif (mode espion) vis-à-vis des attaquants. Les paquets qui circulent contiennent une grande quantité de trafic que le N-IDS doit analyser, pour déterminer les paquets suspects. Dans ce cas, des alertes seront émises à une console sécurisée, située sur un réseau isolé qui relie les capteurs et la console [51].

La figure suivante montre l'architecture de N-IDS [52] :

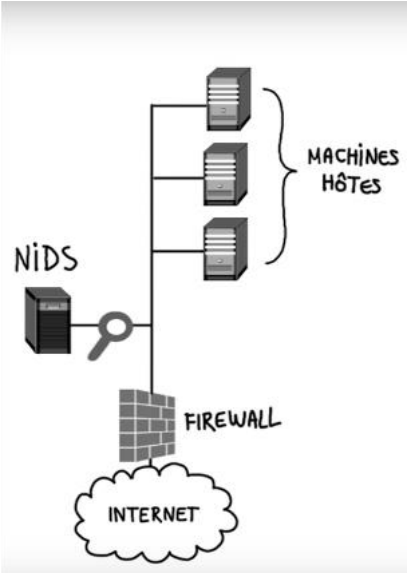

Figure 19 Architecture de N-IDS.

3.3.5.3.1.2 Les avantages et les inconvénients de N-IDS :[52]

Les avantages de N-IDS	Les Inconvénients de N-IDS
-Les capteurs peuvent être bien sécurisés puisqu'ils se contentent d'observer le trafic. -Détecter plus facilement les scans grâce aux signatures. -Filtrage de trafic. -assurer la sécurité contre les attaques puisqu'il est invisible.	-La probabilité de faux négatifs (attaques non détectées) est élevée et il est difficile de contrôler le réseau entier. -Ils doivent principalement fonctionner de manière cryptée d'où une complication de l'analyse des paquets. -A l'opposé des IDS basés sur l'hôte, ils ne voient pas les impacts d'une attaque.

Tableau 2 Les avantages et les inconvénients de N-IDS.

3.3.5.3.2 IDS Host

3.3.5.3.2.1 Définition

Un H-IDS est un système de sécurité qui analyse exclusivement les informations concernant l'hôte où il est installé. Contrairement au N-IDS, le H-IDS n'a pas à contrôler le trafic qui transite dans le réseau, mais analyse et inspecte les journaux de « log » du système en question, l'activité sur les ports réseaux ainsi que les paquets réseaux entrant et sortant de l'hôte concerné. De ce fait, ce système peut observer les activités se déroulant sur l'hôte (même dans des environnements cryptés) avec précision.[51]

La figure suivante montre l'architecture de H-IDS [53] :

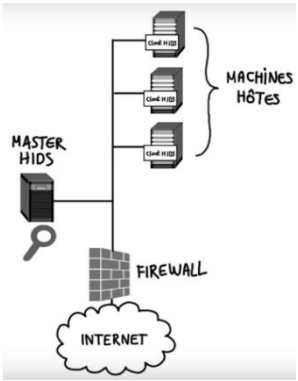

Figure 20 Architecture de H-IDS.

3.3.5.3.2.2 Les avantages et les inconvénients de H-IDS :[52]

Les avantages de H-IDS	Les Inconvénients de H-IDS
-Découvrir plus facilement un Cheval de Troie puisque les informations et les possibilités sont très étendues. -Détecter des attaques impossibles à détecter avec des IDS réseau puisque le trafic est souvent crypté. -Observer les activités sur l'hôte avec précision.	-Ils ont moins de facilité à détecter les scans. -Ils sont plus vulnérables aux attaques de type DOS. -Ils consomment beaucoup de ressources CPU.

Tableau 3 Les avantages et les inconvénients de H-IDS

3.3.5.4 La fréquence d'utilisation

La fréquence d'utilisation d'un système de détection d'intrusions peut exister selon deux formes [51]:

3.3.5.4.1 Analyse périodique

Dans ce type d'IDS, l'analyse se fait périodiquement sur les traces d'audit, intrusions et anomalies. Cela peut s'appliquer dans les environnements peu sensibles, en se contentant d'une analyse journalière.

3.3.5.4.2 Analyse continue

La majorité des IDS récents effectuent leur analyse des traces d'audit ou des paquets réseau continuellement et sans interruption, car la sensibilité des informations (données confidentielles…etc.) à protéger devient cruciale. Toutefois, cette analyse du système est très couteuse en temps de calcul.

3.3.6 Forces et limites des IDS

Bien que les systèmes de détection d'intrusion soient un complément précieux à l'infrastructure de sécurité d'une organisation, il y a des choses qu'ils font bien, et d'autres qu'ils ne font pas bien.[54]

3.3.6.1 Points forts des systèmes de détection d'intrusion

Les systèmes de détection d'intrusion remplissent bien les fonctions suivantes :

- Surveillance et analyse des événements système et des comportements des utilisateurs.
- Test des états de sécurité des configurations du système.
- établir une base de référence pour l'état de sécurité d'un système, puis suivre les modifications apportées à cette base de référence.
- Reconnaître les modèles d'événements système qui correspondent à des attaques connues.
- Reconnaître les modèles d'activité qui s'écartent statistiquement de l'activité normale.
- Gérer les mécanismes d'audit et de journalisation des systèmes d'exploitation et les données qu'ils génèrent.
- Alerter le personnel concerné par des moyens appropriés lorsque des attaques sont détectées.
- Mesurer l'application des politiques de sécurité encodées dans le moteur d'analyse.
- Fournir des politiques de sécurité de l'information par défaut.
- Permettre à des non experts en sécurité d'effectuer d'importantes fonctions de surveillance de la sécurité.

3.3.6.2 Limites des systèmes de détection d'intrusion

Les systèmes de détection d'intrusion ne peuvent pas remplir les fonctions suivantes :

- Compenser les mécanismes de sécurité faibles ou manquants dans l'infrastructure de protection. Ces mécanismes comprennent les pare-feu, l'identification et l'authentification, le cryptage des liens, les mécanismes de contrôle d'accès, ainsi que la détection et l'éradication des virus.
- Détecter, signaler et répondre instantanément à une attaque, en cas de forte charge de réseau ou de traitement.
- Détecter les attaques nouvellement publiées ou les variantes d'attaques existantes.
- Réponse efficace aux attaques lancées par des attaquants sophistiqués.
- Enquêter automatiquement sur les attaques sans intervention humaine.
- Résister aux attaques qui visent à les mettre en échec ou à les contourner.
- Compenser les problèmes de fidélité des sources d'information.

- Traiter efficacement les réseaux commutés.

3.3.7 Les mesures de performance du système de détection d'intrusion

Une matrice de confusion est un tableau qui est utilisé pour décrire les performances d'un modèle de classification sur un ensemble de données de test dont on connaît les vraies valeurs.[55]

	Classe prédite		
classe réelle		Classe = Oui	Classe = Non
	Classe = Oui	Vrai Positive	Vrai Négative
	Classe = Non	Faux Positive	Faux Négative

Tableau 4 Matrice de confusion

- **Vrais Positifs (VP) :** Il s'agit des valeurs positives prédites avec précision, ce qui signifie que la valeur réelle de la classe est oui et que la valeur prédite de la classe est également oui.
- **Vrais négatifs (VN) :** Ce sont les valeurs négatives correctement prédites, ce qui signifie que la valeur réelle de la classe est non et que la valeur attendue de la classe est également non.
- **Faux positifs (FP) :** Lorsque la classe est négative et que l'on s'attend à ce que la classe soit positive. Par exemple, si la classe réelle dit que ce passager n'a pas survécu, mais que la classe prédite vous dit que ce passager survivrait.
- **Faux négatifs (FN) :** Si la classe réelle est oui, mais qu'aucune classe n'est prévue. Par exemple, si la valeur réelle de la classe indique que ce passager a survécu et que la classe prévue vous dit que le passager va mourir.

- **Accuracy :** L'accuracy est la mesure la plus courante de la performance. Il s'agit simplement d'une proportion d'observations correctement prévues par rapport au total des observations. Si nous sommes très précis, on pourrait penser que notre modèle est le meilleur. Oui, la fiabilité est une excellente mesure, mais uniquement lorsque vous disposez d'ensembles de données symétriques présentant pratiquement les mêmes valeurs de faux positifs et de faux négatifs.

$$\text{Accuracy} = VP+VN/VP+FP+FN+VN$$

- **Précision :** La précision est le rapport entre les observations positives correctement prédites et le total des observations positives prédites.

Précision = VP/VP+FP

- **Rappel(Recall) :** Le rappel est le rapport entre les observations positives correctement prédites et l'ensemble des observations de la classe réelle - oui.

Rappel = VP/VP+FN

- **F1 Score :** Le score F1 est la moyenne pondérée de Précision et Rappel. Par conséquent, ce score tient compte à la fois des faux positifs et des faux négatifs. Intuitivement, il n'est pas aussi facile à comprendre que la précision, mais en général, le F1 est plus utile que la précision, en particulier si vous avez une distribution inégale des classes. La précision fonctionne mieux si les faux positifs et les faux négatifs ont le même coût. Si le coût des faux positifs et des faux négatifs est très différent, il est préférable d'examiner à la fois la précision et le rappel.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

3.4 Conclusion

Le but d'un système de détection d'intrusion (IDS) est de protéger la confidentialité, l'intégrité et la disponibilité d'un système. Ils sont conçus pour détecter des problèmes. Dans ce chapitre, nous avons donné une vision globale sur le système de détection d'intrusion en présentant ses types, ses méthodes. Par la suite nous allons passer au chapitre de l'implémentation.

Chapitre 4 : L'approche proposée et l'implémentation

CHAPITRE 4 : L'APPROCHE PROPOSEE ET L'IMPLEMENTATION

4.1 Introduction

Dans ce chapitre, nous présentons notre approche basée sur l'extraction d'information des mesures de performance (Accuracy, Recall, Precesion, F1-score) pour le domaine des IDS, ainsi une implémentation. Un ensemble d'outils de développement sera présente afin d'atteindre notre objectif.

4.2 L'environnement matériel

On a utilisé une machine, configurée comme suit :

- Machine PC ASUS TUF 15.
- Processeur : Intel(R) Core (TM) i5-10300H CPU @ 2.50GHz 2.50 GHz.
- Mémoire vive : 12 GO.
- Disque dur : 512 GO.
- Type de système : Windows 11.

4.3 Les outils de développement

4.3.1 Python

Le langage Python est un langage de programmation open source multi-plateformes et orienté objet. Grâce à des bibliothèques spécialisées, Python s'utilise pour de nombreuses situations comme le développement logiciel, l'analyse de données, ou la gestion d'infrastructures.[56]

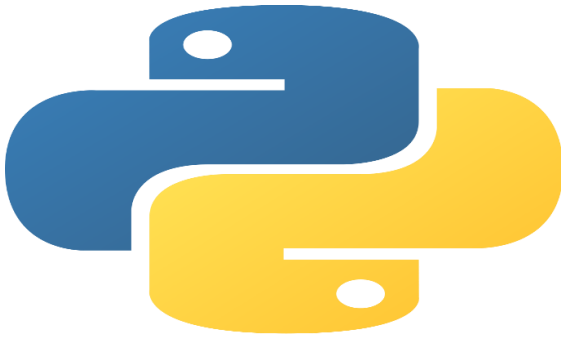

Figure 21 Logo python.

4.3.2 Jupyter

JupyterLab est le dernier environnement de développement interactif basé sur le Web pour les blocs-notes, le code et les données. Son interface flexible permet aux utilisateurs de configurer et d'organiser des flux de travail en science des données, en informatique scientifique, en journalisme informatique et en apprentissage automatique. Une conception modulaire invite les extensions à étendre et enrichit les fonctionnalités.[57]

Figure 22 Logo jupyter.

4.4 Les bibliothèques a utilisé

Nous avons besoin d'un ensemble de bibliothèques pour réaliser le projet

4.4.1 Pdfminer

PDFMiner est un outil d'extraction de texte pour les documents PDF.

4.4.2 Csv

Est un outil qui permet de manipuler les fichier csv

4.4.3 Spacy

Spacy est une bibliothèque open source gratuite pour le traitement avancé du langage naturel (NLP) en Python.[58]

4.4.4 Pandas

pandas est un outil d'analyse et de manipulation de données open source rapide, puissant, flexible et facile à utiliser, construit sur le langage de programmation Python .[59]

4.4.5 Tabula

Tabula est un outil d'extraction de tables à partir des documents.

4.4.6 Numpy

NumPy est le package fondamental pour le calcul scientifique en Python. Il s'agit d'une bibliothèque Python qui fournit un objet tableau multidimensionnel, divers objets dérivés (tels que des tableaux masqués et des matrices) et un assortiment de routines pour des opérations rapides sur des tableaux, y compris mathématiques, logiques, manipulation de forme, tri, sélection, E/S , transformées de Fourier discrètes, algèbre linéaire de base, opérations statistiques de base, simulation aléatoire et bien plus encore.[60]

4.4.7 Sklearn

Est une bibliothèque d'apprentissage automatique open source qui prend en charge l'apprentissage supervisé et non supervisé. Il fournit également divers outils pour l'ajustement du modèle, le prétraitement des données, la sélection du modèle, l'évaluation du modèle et de nombreux autres utilitaires.[61]

4.4.8 Matplotlib

Matplotlib est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python.[62]

4.4.9 Seaborn

Seaborn est une bibliothèque pour créer des graphiques statistiques en Python. Il s'appuie sur matplotlib et s'intègre étroitement aux structures de données pandas .[63]

4.5 L'approche proposée

CHAPITRE 4 : L'APPROCHE PROPOSEE ET L'IMPLEMENTATION

Dans cette partie nous présentons notre approche proposée.

Cette approche est divisée en deux parties :

- La première partie présente l'extraction d'information qui contient deux méthodes, l'extraction à partir du texte et l'extraction à partir des tableaux.

- Pour l'extraction du texte suit les processus suivants :

  - ✓ Extraire le texte du document.
  - ✓ Pré-traitement de texte (mettre le texte en minuscule) et désignation des zones des parties non important (abstract et les références).
  - ✓ Division du texte en phrases.
  - ✓ Prenez que les phrases qui contiennent les informations que nous voulons extraire.
  - ✓ Application de la tâche NER (Named Entity recognition) et POS (part of speech tagging) sur les phrases pour trouver les mots-clés et leurs valeurs.
  - ✓ Stocker le résultat sur un fichier CSV (data set)

- Pour la deuxième méthode (l'extraction du tables) contient les processus suivants :

  - ✓ Extraire tous les tableaux du document.
  - ✓ Stockez chaque table dans un fichier csv pour faciliter sa manipulation.
  - ✓ Choisissez les lignes ou les colonnes qui contiennent les informations nécessaires.
  - ✓ Stocker le résultat sur un fichier CSV (data set).

- La deuxième partie présente la création d'un modèle d'apprentissage automatique qui contient les processus se suivant :

  - ✓ Chargé le data set.
  - ✓ Créer un modèle k-means en spécifiant le nombre de clusters.
  - ✓ Évaluation du modèle (déterminer le score).

CHAPITRE 4 : L'APPROCHE PROPOSEE ET L'IMPLEMENTATION

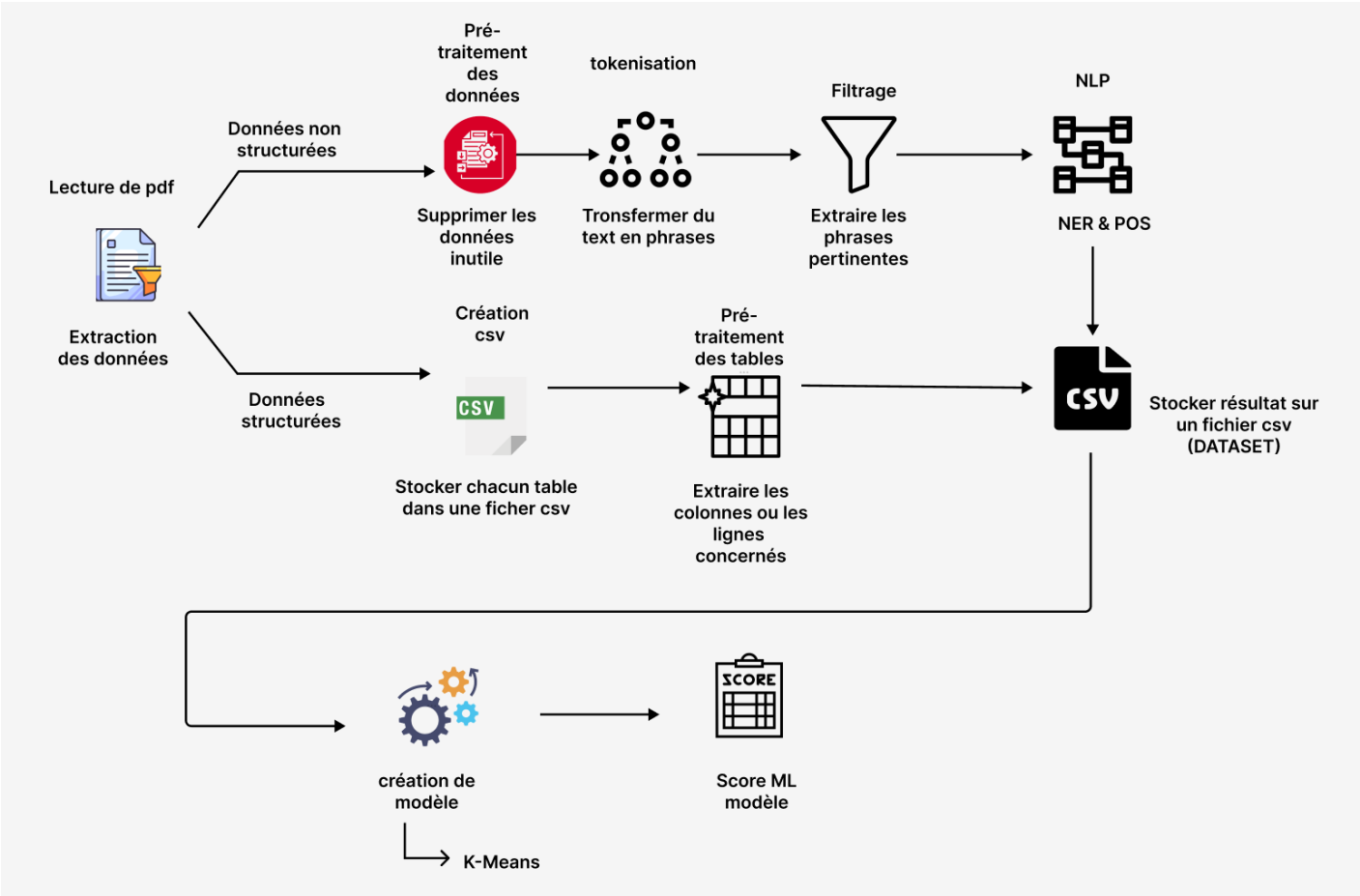

Figure 23 Approche proposée.

4.6 L'implémentation

4.6.1 Partie 1 : Extraction d'information

L'objectif de cette partie extraire ensemble des informations liées au domaine d'IDS de document et les stocker dans ficher csv comme une data set pour utiliser ultérieurement.

4.6.1.1 Importation des bibliothèque requises

Nous devons importer PDFminer, Tabula, csv, Spacy et string ; PDFminer pour l'extraction du texte et Tabula pour l'extraction des tables, CSV pour manipuler les fichier csv, spacy pour traitement des langages naturels, string pour traiter les chaines de caractère.

```
from pdfminer.high_level import extract_text
import tabula , csv , string
import spacy
```

Figure 24 Importation des bibliothèque requises.

4.6.1.2 Les fonctions utilisées

- La fonction « is_float(x) » retourne vrais si la valeur x de type « float » sinon retourne faux

```
def is_float(s) :
    try:
        float(s)
        return True
    except ValueError :
        return False
```

Figure 25 Fonction is_float.

- La fonction « data_filtering (doc) » fait la suppression des parties avant l'introduction et les références du document entrée, et supprimé les espaces entre les paragraphes.

```
def data_filtering(doc):
    doc = doc.replace("\n\n\n" , " ").replace("\n\n" , " ").replace("\n" , " ")
    doc = doc.lower()

    # delete before introduction
    if doc.find("introduction") != -1:
        doc = str(doc.split("introduction")[1:])
        doc = doc[2:-2]

    # delete references
    if doc.find("references") != -1:
        doc = str(doc.split("references")[:-1])
        doc = doc[2:-2]

    return doc
```

Figure 26 Fonction data_filtering.

- La fonction « add_zero () » ajouter des '0' a une table, dans notre cas, nous devons ajouter un '0' si nous ne trouvons pas la valeur des mots-clés (accuracy, recall, precesion, f1-score).

```
def add_zero(table , length , m) :
    while (length < m):
        table.append('0')
        length += 1
    return table
```

Figure 27 Fonction add_zero.

- La fonction « find_data(list_val) » pour éviter la répétition d'insertion des données.

```
def find_data(list_val):

    filename = 'D:\\DataSet_IDS.csv'
    with open(filename, 'r', newline="") as file:
        csvreader = csv.reader(file)
        trouve = False
        for row in csvreader:
            if row == list_val :
                trouve = True
    return trouve
```

Figure 28 Fonction find_data.

- La fonction « write_data_csv () » pour insérer les résultats dans le fichier csv.

```
def write_data_csv(path_file_csv , pdf_name , accuracy , recall , precision , f1_score , maxi) :

    with open(path_file_csv, 'a', newline="") as file:
        csvwriter = csv.writer(file)

        if maxi > 0:
            for i in range(maxi):
                data = []
                data.append(pdf_name)
                data.append(accuracy[i])
                data.append(recall[i])
                data.append(precision[i])
                data.append(f1_score[i])
                if find_data(data) == False:
                    print(data)
                    csvwriter.writerow(data)
            print('\n')
            file.close()

        else  :
            data = [pdf_name , '0','0','0','0']
            if find_data(data) == False:
                csvwriter.writerow(data)
                print(data)
        file.close()
```

Figure 29 Fonction write_data_csv.

- La fonction « structured_data () » : le traitement des tables commencé par cette fonction ;
  - Nous avons crié une loupe pour lire chacun tableau.
  - Nous vérifions si nos données se trouvent sur les colonnes ou les lignes du tableau.

```
def structured_data(path_doc) :
    add_row = 0
    tables = tabula.read_pdf(path_doc ,pages="all")
    table_number =1

    for table in tables :
        table.to_csv("d:\\table_trt\\file_csv{}.csv".format(table_number) ,index=False)
        with open('d:\\table_trt\\file_csv{}.csv'.format(table_number) ) as csv_file:
            try :
                csv_reader = csv.reader(csv_file, delimiter=',')

                id_accuracy = -1
                id_recall = -1
                id_precision = -1
                id_f1_score = -1

                table_colon = []
                table_ligne = []

                liste1 , liste2 , liste3 , liste4 = [] , [] ,[] , []
                for row in csv_reader:
                    i = 0
                    for item in row :
                        item = item.translate(str.maketrans('' , '' , string.punctuation)).lower().replace(' ' , '')

                        # get row accuracy
                        if item.strip().lower().find('accuracy')!=-1 :
                            id_accuracy = i
                            table_ligne.append(row)

                        # get column accuracy
                        if id_accuracy == i :
                             liste1.append(row[id_accuracy])


                        # get row recall
                        if item.strip().lower().find('recall') !=-1 :
                            id_recall = i
                            table_ligne.append(row)

                        # get column recall
```

Figure 30 Fonction des données structuré.

  - Après avoir trouvé les informations, nous les stockons dans des listes
  - Ensuite, nous les stockons dans un fichier csv.

- La fonction « unstructured_data ()» le traitement des textes commencé par cette fonction ;
  - Extrait de texte et faire une filtration par la fonction « data_filtering (doc) ».
  - Transformer le texte à des phrases.

```
def unstructured_data(path_doc):
    add_row = 0
    # fetch files (PDF).
    doc = data_filtering(extract_text(path_doc , 'rb'))



    nlp = spacy.load("en_core_web_lg")
    doc = nlp(doc)
    nlp.add_pipe("merge_entities")
    nlp.add_pipe("merge_noun_chunks")

    # step_1 : sentence
    sentences = list(doc.sents)

    # step_2 : tokenization
    list_token = []
    for i in range(len(sentences)) :
        list_sentence = []

        for token in sentences[i]:
            if(token.text != " "):
                list_sentence.append(token)
        list_token.append(list_sentence)
```

Figure 31 Fonction des données non structuré 1.

- Choisir seulement les phrases qui contient les mots clés

```
    # step_3 : choose useful sentences
    sent_new= []
    i = 0
    for sent in list_token:
        stop = False
        j=0
        while j<len(sent) and stop == False :
            if (sent[j].text == 'accuracy') or (sent[j].text == 'precision') or
            (sent[j].text == 'recall') or (sent[j].text == 'f1-score') or
            (sent[j].text == 'f-measure') or  (sent[j].text == 'f1') :
                sent_new.append(sentences[i])
                stop = True
            else:
                j =j+1
        i = i+1
```

Figure 32 Fonction des données non structuré 2

- Faire une boucle sur les phrases une par une, si trouve dans une phrase une entité de type pourcentage en cherchant les mots qui dépendant à cette entité, Ensuite, stockez les mots et leurs valeurs dans le tableau.

```
# pour pourcentage
if  token.ent_type_ == "PERCENT":
    if token.pos_ == 'NUM': val = token
    if token.pos_ == 'NOUN' :

        # We have an attribute and direct object, so check for subject
        if token.dep_ in ("attr", "dobj"):
            subj = [w for w in token.head.lefts if w.dep_ == "nsubj"]
            if subj:
                print(subj[0], "-->" , val , token)
                data.append(subj[0].text)
                data1.append(val.text)


        # We have a prepositional object with a preposition
        if token.dep_ == "pobj" and token.head.dep_ == "prep":
            print(token.head.head, "-->" , val , token)

            data.append(token.head.head.text)
            data1.append(val.text)


        if token.dep_ == "compound":
            print(token.head, "-->" , val , token)

            data.append(token.head.text)
            data1.append(val.text)


        if token.dep_ == "nmod" and token.head.dep_ == "conj":
            print(token.head.head, "-->" , val , token)
            data.append(token.head.head.text)
            data1.append(val.text)
```

Figure 33 Fonction des données non structuré 3

- Faire une boucle sur les phrases une par une, si on trouve dans une phrase un nombre entre 0 et 1 en cherchant les mots qui dépendant à ce nombre, Ensuite, stockez les mots et leur nombre dans une tableau.

```
# traiter les valeurs entre o et 1
if  token.ent_type_ == "CARDINAL":
    try:
        val =float(token.text)
        if token.pos_ == 'NUM' and val > 0 and val < 1 :

            # We have an attribute and direct object, so check for subject
            if token.dep_ in ("attr", "dobj"):
                subj = [w for w in token.head.lefts if w.dep_ == "nsubj"]
                if subj:
                    print(subj[0], "-->" , token)
                    data.append(subj[0].text)
                    data1.append(float(token.text)*100)


            # We have a prepositional object with a preposition
            if token.dep_ == "pobj" and token.head.dep_ == "prep":
                print(token.head.head, "-->" , token)

                data.append(token.head.head.text)
                data1.append(float(token.text)*100)


            if token.dep_ == "compound":
                print(token.head, "-->" , token)

                data.append(token.head.text)
                data1.append(float(token.text)*100)


            if token.dep_ == "nmod" and token.head.dep_ == "conj":
                print(token.head.head, "-->" , token)
                data.append(token.head.head.text)
                data1.append(float(token.text)*100)
    except:
        continue
```

Figure 34 Fonction des données non structuré 4

- Nous définissons quatre tableaux, chaque tableau représentant un mot-clé (accuracy, recall, precesion, f1-score).

- Stocker les valeurs dans leurs tableaux respectifs puis stocker ses tableaux sur le fichier csv.

```
        # result : accuracy , recall , precision , f1 score
    accuracy = []
    recall = []
    precision = []
    f1_score = []


    filename = 'D:\\DataSet_IDS.csv'
    a = path_doc.split(('\\'))[-1]
    pdf_name = a.split(('.pdf'))[0]

    import re

    if data != []:
        for i in range(len(data)) :
            if data[i].find("accuracy") != -1 and is_float(str(data1[i]))==True :
                accuracy.append(str(data1[i]))

            if data[i].find("recall") != -1 and is_float(str(data1[i]))==True :
                recall.append(str(data1[i]))

            if data[i].find("precision") != -1 and is_float(str(data1[i]))==True :
                precision.append(str(data1[i]))

            if data[i].find('f1 score') != -1 and data[i].find('measure') != -1 and is_float(str(data1[i]))==True :
                f1_score.append(str(data1[i]))



        maxi = max(max(len(accuracy) , len(recall)) , max(len(precision) , len(f1_score)))
        accuracy = add_zero(accuracy, len(accuracy) , maxi)
        recall = add_zero(recall, len(recall) , maxi)
        precision = add_zero(precision, len(precision) , maxi)
        f1_score = add_zero(f1_score, len(f1_score) , maxi)

        write_data_csv(filename , pdf_name , accuracy , recall , precision , f1_score , maxi)
        add_row += 1

    return add_row
```

Figure 35 Fonction des données non structuré 5

4.6.2 Partie 2 : Machine Learning

4.6.2.1 Importation des bibliothèque requises

Nous devrons importer Numpy et Pandas, Numpy et une bibliothèque qui contient des fonctions mathématiques utilisées pour le calcul scientifique tandis que pandas est utilisé pour importer et gérer les ensembles de données.

```
# importation des bibliothéque pertinente
from sklearn.model_selection import train_test_split
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.cluster import KMeans
```

Figure 36 Importation des bibliothèque requises

4.6.2.2 Chargement des données

```
# chargement des données
dataset = 'D:\\DataSet_IDS.csv'
data = pd.read_csv(dataset)
x = data.iloc[: ,[1,2,3,4]]
```

Figure 37 Chargement des données

4.6.2.3 Tracer les données

- On utiliser la fonction « pairplot () » de la bibliothèque « seaborn » :

```
# tracer les données
import seaborn as sns

sns.set_theme(style="ticks")
sns.pairplot(x)
```

Figure 38 Tracer les données 1

- Distribution des données :

La figure suivante montre la distribution des données de notre dataset, on remarque que les données sont bien distribuées en terme la recall et la précision.

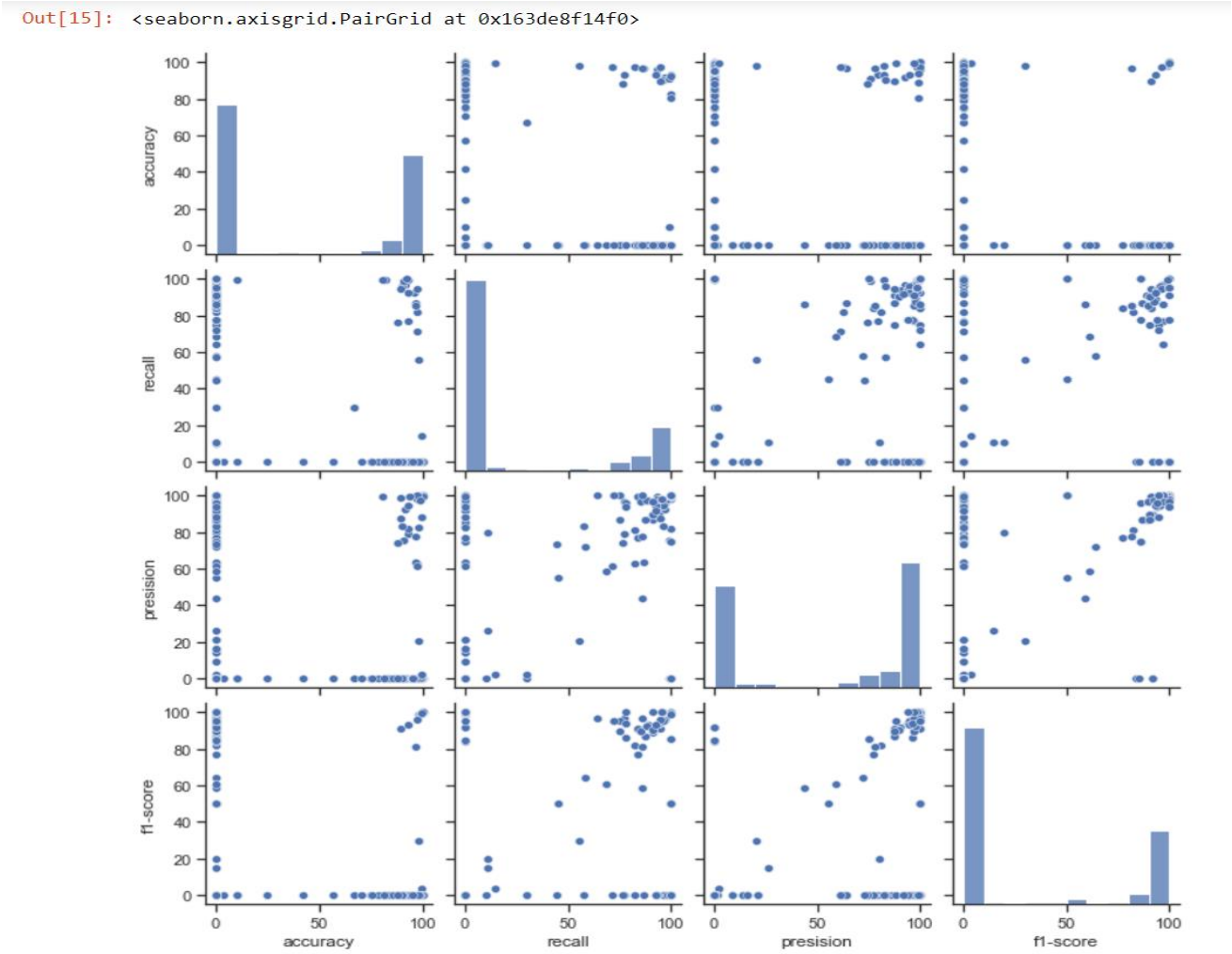

Figure 39 Tracer les données 2

4.6.2.4 Chercher le nombre de clusters

- On utilise la méthode Elbow pour déterminer nombre minimum cluster.

```
# Using the elbow method to find  the optimal number of clusters

from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
    y_kmeans=KMeans(n_clusters=i, init='k-means++', max_iter= 300, n_init= 10, random_state= 0)
    y_kmeans.fit(x)
    wcss.append(y_kmeans.inertia_)
plt.plot(range(1, 11),wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters K')
plt.ylabel('Average Within-Cluster distance to Centroid (WCSS)')
plt.show()
```

Figure 41 La méthode d'Elbow 1

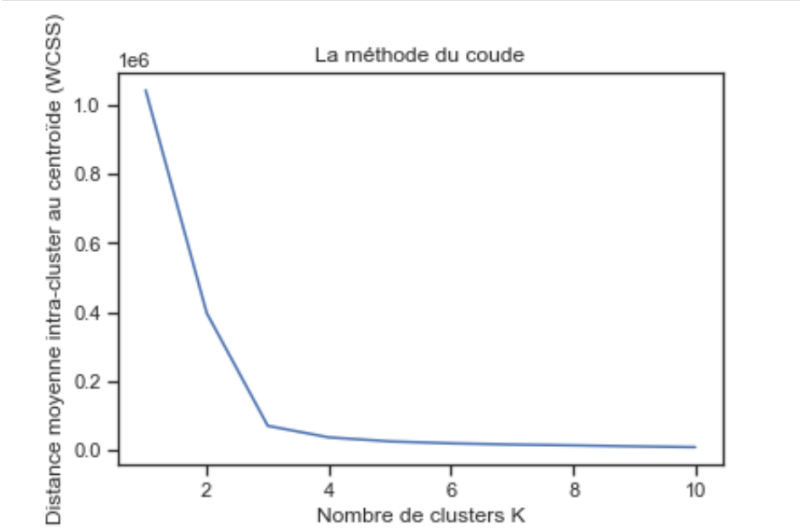

Figure 40 La méthode d'Elbow 2

- On remarque sur ce graphique, la forme d'un bras où le point le plus haut représente l'épaule et le point où K vaut 10 représente l'autre extrémité : la main. Le nombre optimal de clusters est le point représentant le coude. Ici le coude peut être représenté par K valant 3.

- Le résultat : selon le graphe de Elbow le nombre des clusters est 3.

4.6.2.5 Création de modèles

- Kmeans=KMeans () : créer une instance de classe KMeans avec les paramètres suivants :
  - n_clusters=3 : le nombre de cluster est 3.
  - Init='k-means++' : le type d'initialisation est 'k-means++'.
  - Max_iter =300 : le nombre maximal d'itérations de l'algorithme k-means pour une seule exécution est 300.

- Fit_predect () : calculer les centres de cluster et prédisez l'indice du cluster pour chaque échantillon.

```
# Applying k-means to the dataset

kmeans=KMeans(n_clusters = 3 , init='k-means++', max_iter= 300)
y_kmeans=kmeans.fit_predict(x)
```

Figure 42 Création de modèles

4.6.2.6 Déterminer le score du modèle

- Nous calculons le score du model par la fonction score :

```
a = kmeans.score(x)

print('model score : %s' % a )
```

Figure 43 Déterminer le score

- Le tableau ci-dessous montre la valeur du score en termes du nombre de clusters (k):

  On remarque que le modèle à trois clusters a donné un meilleur score que deux clusters.

Nombre de clusters (K)	K=2	K=3
Score de modèle	-397035.05	-69942.53

Tableau 5 Comparaison score de modèle en termes de k

- La distribution des données une fois utilise le k-means :

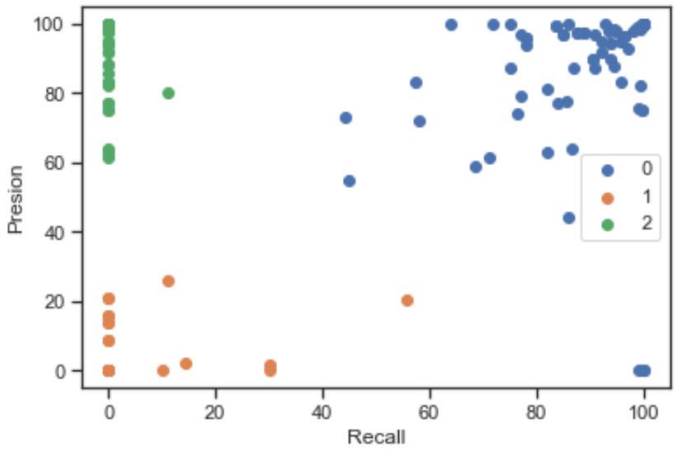

Figure 44 La distribution des données une fois utilise le k-means

- La distribution des données en utilisant k-means en remarque que si la valeur de recall et de précision est élevée forment cluster bleu, si la valeur de recall et de précision est faible forment cluster vert, et si la valeur précision sont élevé et la valeur de recall sont faible forment cluster jaune.

4.7 Conclusion

Dans ce chapitre, nous avons d'abord présenté les différents outils et bibliothèques que nous avons utilisés pour implémenter notre approche.

Dans un premier temps, nous présentons notre approche proposée pour l'extraction des mesures de performances à partir des fichiers PDF, afin de les capitaliser ces informations dans un seul fichier csv. Après une implémentation de notre approche a été présenté et examiné.

Dans notre travail, nous extrayons les informations des documents des systèmes de détection d'intrusion par les technologies NER et POS Tagging, ces informations représente les valeurs des variables de mesure de performance, et stockez-les dans un seul fichier qui représente le dataset, Ensuite, nous avons étudié le dataset à l'aide de l'algorithme k-means Grâce à laquelle nous avons obtenu trois classes pour notre ensemble de données.

Conclusion générale

5 CONCLUSION GENERALE

Ce modeste travail à l'opportunité d'approfondir nos connaissances dans les trois domaines ; Extraction d'information, Machine Learning et systèmes de détection d'intrusion.

Il vise à capitaliser les mesures de performance des IDS par l'extractions des informations du ce domaine étudié, afin de faciliter de classification des systèmes de détection d'intrusion. Tout d'abord, nous avons présenté la technologie de l'extraction d'information, en expliquant ses différents concepts, tâches et architecture. Après cette brève introduction sur l'extraction d'information, nous avons présenté les différentes techniques utilisées dans la Machine Learning et leurs algorithmes. Ensuite, le domaine d'études est étudié afin de se basant sur les mesures de performance des systèmes de détection d'intrusion.

6 BIBLIOGRAPHIE

[1] S. Zenasni, « Extraction d'information spatiale à partir de données textuelles non-standards », p. 164.
[4] F. Even, « Extraction d'Information et modélisation de connaissances à partir de Notes de Communication Orale », p. 253.
[5] A. Abdelmadjid et A. Hamid, « Vers un système d'extraction d'informations pour les textes de la presse arabophone en ligne ArIExtract », p. 10, 2009.
[7] S. Bannour, « Apprentissage interactif de règles d'extraction d'information textuelle », p. 197.
[12] F. Even, « Extraction d'Information et modélisation de connaissances à partir de Notes de Communication Orale », p. 253.
[14] « [Adaptive Computation and Machine Learning series] Ethem Alpaydin - Introduction to Machine Learning (2014, The MIT Press) - libgen.lc.pdf ».
[20] J. Confais et M. L. Guen, « Premiers pas en régression linéaire avec SAS® », p. 146.
[21] S. Ray, « A Quick Review of Machine Learning Algorithms », p. 5.
[22] L. Hawarah, « Une approche probabiliste pour le classement d'objets incomplètement connus dans un arbre de décision », p. 183.
[25] E. Mathieu-Dupas, « Algorithme des k plus proches voisins pondérés et application en diagnostic », p. 8.
[27] L. Zhang, Y. Ren, et P. N. Suganthan, « Instance based random forest with rotated feature space », in *2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL)*, Singapore, Singapore, avr. 2013, p. 31-35. doi: 10.1109/CIEL.2013.6613137.
[28] B. Mahesh, « Machine Learning Algorithms - A Review », vol. 9, n° 1, p. 6, 2018.
[30] K. A. Scarfone et P. M. Mell, « Guide to Intrusion Detection and Prevention Systems (IDPS) », National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 800-94, 2007. doi: 10.6028/NIST.SP.800-94.
[31] « Conception et développement dun outil daudit de Conformité à la PSSI (Politique de Sécurité du Système dInformation) issue de la Norme Internationale ISO 27002.pdf ».
[40] M. T. V. Tay, « LE SYSTÈME DE DÉTECTION DES INTRUSIONS », p. 49.
[41] J. P. Anderson, « Computer Security Threat Monitoring and Surveillance », p. 56.
[42] D. E. Denning, « An Intrusion-Detection Model », *IEEE Trans. Softw. Eng.*, vol. SE-13, n° 2, p. 222-232, févr. 1987, doi: 10.1109/TSE.1987.232894.
[43] « Gunadiz, Safia magister.pdf ».
[45] R. Shirey, « Internet Security Glossary », Internet Engineering Task Force, Request for Comments RFC 2828, 2000. doi: 10.17487/RFC2828.
[46] R. W. Shirey, « Internet Security Glossary, Version 2 », Internet Engineering Task Force, Request for Comments RFC 4949, 2007. doi: 10.17487/RFC4949.
[47] G. Hiet, « Détection d'intrusions paramétrée par la politique de sécurité grâce au contrôle collaboratif des flux d'informations au sein du système d'exploitation et des applications : mise en œuvre sous Linux pour les programmes Java », p. 161.
[48] H. Debar, M. Dacier, et A. Wespi, « A revised taxonomy for intrusion-detection systems », *Ann. Télécommunications*, vol. 55, n° 7-8, p. 361-378, juill. 2000, doi: 10.1007/BF02994844.
[51] K. Tabia et S. Benferhat, « Modèles graphiques et approches comportementales pour la détection d'intrusions ». France, 2008.
[52] M. Belkhatmi et M. Benamara, « Mise en place d'un systéme de détection et de prévention d'intusion », p. 75.

BIBLIOGRAPHIE

[54] « Special Publication 800-31 », p. 51.

[56] « Snapshot ». Consulté le: 25 mai 2022. [En ligne]. Disponible sur: https://www.futura-sciences.com/tech/definitions/informatique-python-19349/?fbclid=IwAR3V7HfHsPY2zAp7kZ0I_aYkTVycFPyqz71u63_LhmZlRuwIWOfSAACZnto

7 WEBOGRAPHIE

[2] « Information Extraction with Natural Language Processing. » https://www.linkedin.com/pulse/information-extraction-natural-language-processing-shubham-shankar (consulté le 28 février 2022).

[3] « What is Information Extraction? - A Detailed Guide », AI & Machine Learning Blog, 19 juillet 2021. https://nanonets.com/blog/information-extraction/ (consulté le 6 février 2022)..

[6] « Ipalakova-Madina.pdf ». Consulté le: 26 février 2022. [En ligne]. Disponible sur: https://studentnet.cs.manchester.ac.uk/resources/library/thesis_abstracts/BkgdReportsMSc10/Ipalakova-Madina.pdf

[8] « Named Entity Recognition: Concept, Tools and Tutorial », MonkeyLearn Blog, 30 mars 2020. https://monkeylearn.com/blog/named-entity-recognition/ (consulté le 27 février 2022).

[9] arvindpdmn, « Relation Extraction », Devopedia, 6 février 2020. https://devopedia.org/relation-extraction (consulté le 27 février 2022).

[10] leopardpan, « Introduction of Event Extraction ». http://www.caojiarun.com/2019/11/Introduction-of-Event-Extraction/ (consulté le 27 février 2022).

[11] « Information Extraction with Natural Language Processing. » https://www.linkedin.com/pulse/information-extraction-natural-language-processing-shubham-shankar (consulté le 23 février 2022).

[13] « Tout savoir sur l'intelligence artificielle », Microsoft experiences, 9 février 2018. https://experiences.microsoft.fr/articles/intelligence-artificielle/comprendre-utiliser-intelligence-artificielle/ (consulté le 20 février 2022).

[15] « Machine learning definition », Digital Insiders, 14 novembre 2016. https://digitalinsiders.feelandclic.com/machine-learning-definition (consulté le 22 février 2022).

[16] « Classification In Machine Learning | Classification Algorithms », Edureka, 4 décembre 2019. https://www.edureka.co/blog/classification-in-machine-learning/ (consulté le 31 mai 2022).

[17] « Understanding Regression In Machine Learning | Built In ». https://builtin.com/data-science/regression-machine-learning (consulté le 31 mai 2022).

[18] « Clustering | Types Of Clustering | Clustering Applications », Analytics Vidhya, 3 novembre 2016. https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/ (consulté le 31 mai 2022).

[19] « Les types de Machine Learning ». https://fr.linkedin.com/pulse/les-types-de-machine-learning-william-simetin-grenon (consulté le 6 février 2022).

[23] « Decision Tree Tutorials & Notes | Machine Learning », HackerEarth. https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/?fbclid=IwAR1WlsA_M4SyiEiK0tPY1YAq021U7K4wtCGTcEnOEqMT0zbLt-6nSF9xQLE (consulté le 22 février 2022).

[24] « SVM | Support Vector Machine Algorithm in Machine Learning », Analytics Vidhya, 12 septembre 2017. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (consulté le 13 février 2022).

[26] P. Dhote, « K-NN(K-Nearest Neighbors) », Medium, 5 août 2020. https://pradeep-dhote9.medium.com/k-nn-k-nearest-neighbors-3f34c60d5f2e (consulté le 22 février 2022).

[29] « La méthode des kmeans - DellaData », 6 mai 2020. https://delladata.fr/kmeans/ (consulté le 27 juin 2022).

WEBOGRAPHIE

[32] « III.1 La sécurité de réseau informatique ». https://www.institut-numerique.org/iii1-la-securite-de-reseau-informatique-525681a39077f (consulté le 1 mars 2022).

[33] « Introduction à la sécurité informatique ». https://web.maths.unsw.edu.au/~lafaye/CCM/secu/secuintro.htm?fbclid=IwAR1ozZJNwlR_cBrc419CWDbcCmTAlpiw0nAicelBEZdaYnN5m3l97so9HvI (consulté le 2 mars 2022).

[34] « Qu'est-ce qu'un pare-feu ? », Cisco. https://www.cisco.com/c/fr_fr/products/security/firewalls/what-is-a-firewall.html (consulté le 8 mars 2022).

[35] « Qu'est-ce que le chiffrement ? », Oracle France. https://www.oracle.com/fr/cloud/chiffrement-donnees-clef.html (consulté le 8 mars 2022).

[36] « Network/Security Logging & Monitoring: Challenges & Best Practices », Linford & Company LLP, 18 septembre 2019. https://linfordco.com/blog/logging-and-monitoring/ (consulté le 8 mars 2022).

[37] « What is Antivirus - Definition, Meaning & Explanation », Verizon Fios. https://www.verizon.com/info/definitions/antivirus/ (consulté le 8 mars 2022).

[38] « What is an intrusion detection system? How an IDS spots threats | CSO Online ». https://www.csoonline.com/article/3255632/what-is-an-intrusion-detection-system-how-an-ids-spots-threats.html?fbclid=IwAR1a8JJ1GKN5x9IPwNRqLr04wEDUExfKXlKKDYmTLOus5zISVUXuQQOvc0Q (consulté le 8 mars 2022).

[39] « Intrusion Detection System (IDS) », GeeksforGeeks, 8 avril 2019. https://www.geeksforgeeks.org/intrusion-detection-system-ids/ (consulté le 2 mars 2022).

[44] « What Is a Cyberattack? - Most Common Types - Cisco ». https://www.cisco.com/c/en/us/products/security/common-cyberattacks.html?fbclid=IwAR0G9d1i6JU9YxaQSM4O9OFIQyWEnCUYDU8Y1qXFhPtRIpm9Nu34ytH3LQI (consulté le 2 mars 2022).

[49] « Memoire Online - Mise en place d'un crypto systeme pour la sécurité des donnée et la détection d'intrusion dans un supermarché - landry Ndjate », Memoire Online. https://www.memoireonline.com/01/16/9388/m_Mise-en-place-dun-crypto-systeme-pour-la-securite-des-donnee-et-la-detection-dintrusion-da14.html (consulté le 7 mars 2022).

[50] « Active VS Passive IDS Responses », Get Certified Get Ahead, 18 septembre 2017. https://blogs.getcertifiedgetahead.com/active-vs-passive-ids-responses/ (consulté le 7 mars 2022).

[53] T. Ory, « L'intrusion des réseaux, qu'est ce qu'un IDS/IPS ? », Thomas Ory, 10 juillet 2020. https://thomasory.com//r%C3%A9seau/s%C3%A9curit%C3%A9/intrusion/What-is-an-IPS-and-IDS/ (consulté le 7 mars 2022).

[55] « Exactitude, précision, rappel et score F1 : interprétation des mesures de performance | LinkedIn ». https://www.linkedin.com/pulse/accuracy-precision-recall-f1-score-interpretation-mukul-choudhary/?fbclid=IwAR011629A6LfYRLzsntTAWn6QwN-ApT-JrygKaEiwp-wje3tImLaX9A66k8 (consulté le 6 juin 2022).

[56] « Snapshot ». Consulté le: 25 mai 2022. [En ligne]. Disponible sur: https://www.futura-sciences.com/tech/definitions/informatique-python-19349/?fbclid=IwAR3V7HfHsPY2zAp7kZ0I_aYkTVycFPyqz71u63_LhmZlRuwIWOfSAACZnto

[57] « Snapshot ». Consulté le: 25 mai 2022. [En ligne]. Disponible sur: https://jupyter.org/

[58] « spaCy 101: Everything you need to know · spaCy Usage Documentation », spaCy 101: Everything you need to know. https://spacy.io/usage/spacy-101 (consulté le 25 mai 2022).

[59] « pandas - Bibliothèque d'analyse de données Python ». https://pandas.pydata.org/ (consulté le 25 mai 2022).

WEBOGRAPHIE

[60] « Documentation NumPy — Manuel NumPy v1.22 ». https://numpy.org/doc/stable/ (consulté le 25 mai 2022).

[61] « Getting Started », scikit-learn. https://scikit-learn/stable/getting_started.html (consulté le 25 mai 2022).

[62] « Matplotlib documentation — Matplotlib 3.5.2 documentation ». https://matplotlib.org/stable/index.html (consulté le 25 mai 2022).

[63] « An introduction to seaborn — seaborn 0.11.2 documentation ». https://seaborn.pydata.org/introduction.html (consulté le 25 mai 2022).