Résumé

Les développements récents des technologies de l'information ont entraîné des changements continus dans tous les domaines, la transmission de l'information par les réseaux et les ordinateurs devenant l'un des signes indispensables de l'ère numérique. Faciliter les exigences modernes en matière de durée de vie en réduisant le volume d'affaires et en développant des méthodes de production, de stockage et de distribution des problèmes d'information et des risques qui menacent et affectent la sécurité de l'économie de l'information de l'entreprise, Cela exige une vigilance afin d'assurer la sécurité nécessaire de ses informations. Économie de l'information, en particulier dans la mesure où elle vit dans l'économie de l'information du détenteur d'information contrôlé, De nouvelles approches appelées **IDS** ont émergé. Leur but est d'analyser le mouvement des requêtes et de détecter les comportements nuisibles.

Notre objectif dans ce projet est de traiter un modèle ou un algorithme d'apprentissage automatique pour détection d'intrusion, ce qui donne de bons résultats en sécurité informatique, qui rivalisent avec d'autres algorithmes d'apprentissage automatique.

Mots clés : Sécurité Informatique, système de détection d'intrusion, Data Mining, Data Science, Sélection des attributs, KDDCup99.

Abstract

Dvelopments in information technology have led to continuous changes in all areas, with the transmission of information through networks and computers becoming one of the indispensable signs of the digital age. To facilitate modern life requirements by reducing business volume and developing methods of producing, storing and distributing information information problems and risks that threaten and affect the security of the company's information economy, This requires vigilance in order to ensure the necessary security of its information. information economy ", particularly as it lives in the information economy of the controlled information holder, New approaches called IDS have emerged. Their goal is to analyze the movement of requests and detect harmful behavior.

Our goal in this project is to process a machine learning model or algorithm for intrusion detection, which gives good results in computer security, which rival other machine learning algorithms.

Keywords : Security, detection intrusion system, Data Mining, Data Science, features selection, KDDCup99.

ملخص

قد أدت التطورات الأخيرة في تكنولوجيا المعلومات إلى تغييرات مستمرة في جميع المجالات، حيث أصبح نقل المعلومات عن طريق الشبكات والحواسيب أحد العلامات التي لا غنى عنها للعصر الرقمي. لتسهيل متطلبات الحياة الحديثة من خلال تقليل حجم الأعمال وتطوير طرق إنتاج وتخزين وتوزيع المعلومات، وقد أدى ذلك إلى زيادة مشاكل المعلومات والمخاطر التي تهدد وتؤثر على أمن اقتصاد المعلومات في الشركة ، الأمر الذي يتطلب اليقظة من أجل ضمان الأمن اللازم لمعلوماتها، لا سيما وأنها تعيش في اقتصاد المعلومات لحائز المعلومات الخاضعة للسيطرة، وقد ظهرت نُهج جديدة تسمى نظم الكشف عن التسلل (IDS). هدفهم هو تحليل حركة الطلبات واكتشاف السلوك الضار.

هدفنا في هذا المشروع هو معالجة نموذج أو خوارزمية التعلم الآلي لاكتشاف التسلل، والتي تعطي نتائج جيدة في أمان الكمبيوتر، والتي تتنافس مع خوارزميات التعلم الآلي الأخرى.

الكلمات الرئيسية : الأمن، كشف التسلل، تنقيب البيانات، علم البيانات، إختيار العناصر.

Table des matières

Résumé **i**

Table des figures **viii**

Liste des tableaux **x**

Introduction générale **1**
Introduciton . . . 2
Problématique . . . 2
Organisation du mémoire . . . 3

1 La sécurité informatique **4**
1.1 Introduction . . . 5
1.2 La sécurité informatique . . . 5
1.2.1 Définitions de la sécurité informatique . . . 5
1.2.2 Objectifs de la sécurité . . . 5
1.2.3 Attaques de sécurité . . . 7
1.3 Classification des attaques . . . 7
1.3.1 La première classification . . . 7
1.3.2 La deuxième classification . . . 7
1.3.3 La troisième classification . . . 8
1.3.4 La quatrième classification . . . 8
1.4 Exemples des attaques informatiques . . . 8
1.4.1 L'attaque de déni de service distribué (DDoS) . . . 8
1.4.2 L'attaque man-in-the-middle . . . 9
1.4.3 Les attaques virales . . . 10
1.5 Les techniques de sécurité . . . 10
1.5.1 La sécurisation des accès réseau . . . 11
1.5.2 Superviser les connexions réseau . . . 11
1.5.3 Assurer la confidentialité des connexions . . . 11
1.5.4 La protection des équipements réseau . . . 13
1.6 Conclusion . . . 14

2 Détection d'intrusion **15**
2.1 Introduction . . . 16
2.2 Intrusion . . . 16
2.3 Protection . . . 16
2.4 Système de détection d'intrusion . . . 16
2.4.1 Définition d'un IDS . . . 17
2.4.2 Évaluation des IDS . . . 18
2.4.3 Les Caractéristiques d'un système de détection d'intrusions . . . 18
2.5 Classification des systèmes de détection d'intrusion . . . 19
2.6 Les méthodes de détection d'intrusion . . . 19
2.6.1 L'approche comportementale . . . 20
2.6.1.1 Avantages . . . 20
2.6.1.2 Inconvénients . . . 20
2.6.2 L'approche par scénario . . . 21
2.6.2.1 Avantages . . . 21
2.6.2.2 Inconvénients . . . 21
2.7 Les différents types d'IDS . . . 22
2.7.1 IDS basé sur le hôte . . . 22
2.7.1.1 Les avantages . . . 22
2.7.1.2 Les inconvénients . . . 22
2.7.2 IDS basé sur le réseau . . . 23
2.7.2.1 Les avantages . . . 23
2.7.2.2 Les inconvénients . . . 23
2.8 Conclusion . . . 25

3 Machine Learning **26**
3.1 Machine Learning . . . 27
3.1.1 Comment Machine Learning fonctionne-t-il ? . . . 27
3.1.2 Caractéristiques de Machine Learning . . . 28
3.1.3 Différences entre Machine learning algorithmes et les algorithmes traditionnels basés sur des règles . . . 28
3.1.3.1 Algorithmes fondés sur des règles . . . 28
3.1.3.2 Machine Leaning . . . 28
3.1.4 Machine Learning Classification . . . 28
3.2 Data Science . . . 30
3.2.1 Qu'est-ce que data science ? . . . 30
3.2.2 Quelle est la différence entre data science, l'intelligence artificielle et le machine learning ? . . . 30
3.2.3 Théorie et méthodes analytiques avancées pour la classification supervisé . . . 31
3.2.3.1 Arbres de décision(Decision Trees) . . . 31
3.2.3.1.1 Les avantages . . . 33
3.2.3.1.2 Les inconvénients . . . 33
3.2.3.2 Naive Bayes . . . 33
3.2.3.2.1 Les avantages . . . 34
3.2.3.2.2 Les inconvénients . . . 34

3.2.3.3 K plus proches voisins (KPPV) . . . 34
3.2.3.3.1 Avantages et inconvénients de (KPPV) . . . 35
3.2.4 Théorie et méthodes analytiques avancées pour la classification non supervisé . . . 35
3.2.4.1 K-means . . . 35
3.2.4.1.1 Critiques du K moyennes . . . 36
3.2.5 Composantes de Data Science . . . 37
3.3 Data Mining . . . 38
3.3.1 Qu'est-ce que le data mining ? . . . 38
3.3.2 Comment ça marche Data Minnig ? . . . 38
3.3.2.1 Étape 1 : Définition du problème . . . 39
3.3.2.2 Étape 2 : Collecte des données . . . 39
3.3.2.3 Étape 3 : Construire le modèle d'analyse . . . 39
3.3.2.4 Étape 4 : Étude des résultats . . . 39
3.3.2.5 Étape 5 : Formalisation et diffusion . . . 39
3.3.3 Données (data) . . . 39
3.3.4 Fouille (Mining) . . . 39
3.3.5 Concepts de bases . . . 40
3.3.6 La différence entre donnée, information et connaissance . . . 40
3.3.7 Les instances . . . 40
3.3.8 Modèle . . . 41
3.3.9 Étiquettes (classe) . . . 41
3.3.10 Données structurées . . . 41
3.3.11 Les données non-structurées . . . 41
3.4 La différence entre Data Science, et Data Mining . . . 42
3.5 Conclusion . . . 43

4 Expérimentation et résultats **44**
4.1 Introduction . . . 45
4.2 DATA SETS . . . 45
4.3 Description de Dataset KDDCup99 . . . 46
4.3.1 Le Contenue de KDDCup99 . . . 46
4.4 Présentation des outils utilisés . . . 49
4.4.1 Langage de programmation . . . 49
4.4.2 Bibliothèques utilisées . . . 50
4.5 Expérimentations et discussions . . . 51
4.5.1 Présentation de KDDCup99 . . . 51
4.5.2 Data Cleaning . . . 51
4.5.3 Features selection (sélection des attributs) . . . 52
4.5.3.1 L'information Mutuelle . . . 53
4.5.3.2 Chi-square (Chi2) . . . 53
4.5.4 Split data set . . . 54
4.5.4.1 Base d'apprentissage VS base de test . . . 54
4.5.5 Classification algorithme . . . 54
4.5.5.1 Les mesures de performances utilisées . . . 55
4.5.5.1.1 Matrice de confusion : . . . 55

4.5.5.1.2 Précision et Rappel 55
4.5.5.1.3 Accuracy 56
4.5.5.1.4 TP_rate et FP_rate 57
4.5.5.1.5 AUC - ROC Curve 57
4.5.5.1.6 F-mesures et entropie 58
4.5.5.2 L'implémentions 59
4.5.5.3 Naive Bayes 60
4.5.5.3.1 La discussion des résultats : 60
4.5.5.4 KPPV 62
4.5.5.4.1 La discussion des résultats : 62
4.5.5.5 Arbre de décision 64
4.5.5.5.1 La discussion des résultats : 65
4.5.5.6 K-means 66
4.5.5.6.1 La discussion des résultats : 67
4.6 Comparaison entre algorithmes 68
4.7 Remarque 70
4.7.1 Pour les attributs sélectionnés 70
4.7.1.1 Pour L'information Mutuelle : 70
4.7.1.2 Pour Chi-square(chi2) : 70
4.8 Conclusion 71

Conclusion Générale **72**

Future Work **73**

Table des figures

1.1 L'attaque Ddos 9
1.2 L'attaque man-in-the-middle 10
1.3 La représentation en couches des protocoles de sécurité 12

2.1 Modèle simplifié d'un système de détection d'intrusions. 17
2.2 Les critères de classification des IDSs 19

3.1 Comment Machine Learning fonctionne-t-il ? 27
3.2 Différences entre Machine learning algorithmes et les algorithmes traditionnels basés sur des règles 28
3.3 Apprentissage supervisé 29
3.4 Apprentissage non supervisé 29
3.5 La différence entre data science, l'intelligence artificielle et le machine learning 30
3.6 Exemple d'arbre de décision. 31
3.7 Modèle KPPV 35
3.8 Modèle de K-means 36
3.9 Composantes de Data Science 37
3.10 Processus Data Mining 38

4.1 Interface Anaconda 49
4.2 Interface Jupyter 50
4.3 Notre approche proposé 51
4.4 Pourcentage de classes 52
4.5 Information mutuelle 53
4.6 AUC ROC Curve 58
4.7 Visualisation graphique d'attributs sélectionnés par Naive Bayes 60
4.8 Visualisation par courbe d'attributs sélectionnés par Naive Bayes 60
4.9 TP_rate et FP_rate de Naïve Bayes 61
4.10 Visualisation graphique d'attributs sélectionnés par KPPV 62
4.11 Visualisation par courbe d'attributs sélectionnés par KPPV 62
4.12 TP_rate et FP_rate de KPPV 63
4.13 Visualisation graphique d'attributs sélectionnés par Arbre de décision . . 64
4.14 Visualisation par courbe d'attributs sélectionnés par Arbre de décision . 64

4.15 TP_rate et FP_rate de KPPV 65
4.16 Visualisation graphique d'attributs sélectionnés par K-means 66
4.17 Visualisation par courbe d'attributs sélectionnés par K-means 66
4.18 TP_rate et FP_rate de K-means 67
4.19 Visualisation par courbe de la comparaison des résultats de l'entropie avec 2 attributs sélectionnés 68
4.20 Visualisation par courbe de la comparaison des résultats de l'entropie avec 6 attributs sélectionnés 69
4.21 Visualisation par courbe de la comparaison des résultats de l'entropie avec 41 attributs sélectionnés 70

Liste des tableaux

3.1 Statistique montre la quantité des données générer dans le web 40
3.2 La différence entre Data Science, et Data Mining 43

4.1 Une vue d'ensemble complète des data sets. 46
4.2 Caractéristiques de base des connexions TCP individuelles. 47
4.3 Fonctionnalités de contenu au sein d'une connexion suggérée par la connaissance du domaine. 48
4.4 Caractéristiques de circulation calculées à l'aide d'une fenêtre de temps de deux secondes. 48
4.5 Matrice de confusion 55

Introduction générale

Introduction

Nous vivons dans ce siècle, le soi-disant âge de l'information, et le premier et dernier crédit pour la diffusion très facile de l'information revient à l'ordinateur. L'ordinateur est devenu l'une des choses indispensables dans tous les endroits de la maison. En outre, la personne est devenue très dépendante de l'ordinateur pour faciliter toutes les opérations et activités, et pour cette raison, l'ordinateur est devenu l'épine dorsale de la vie. Ce développement phénoménal s'accompagne naturellement de l'accroissement du nombre d'utilisateurs. Qui ne sont pas forcément pleins de bonnes intentions vis-à-vis de ces systèmes informatiques. Ils peuvent exploiter les vulnérabilités des réseaux et les systèmes pour essayer d'accéder à des informations sensibles dans le but de les lire, les modifier ou les détruire, portant atteinte au bon fonctionnement du système.

Il existe un ensemble d'objectifs spécifiques adoptés par la science de la sécurité de l'information, et les plus importants de ces objectifs sont les suivants : vise à protéger l'accès et la manipulation des données et les ressources d'un système par des mécanismes d'authentification, d'autorisation, de contrôle d'accès, etc.

Les attaques les plus ré- centres profitent des failles de sécurité des services ou systèmes informatiques qui sont plus vulnérables. Pour pallier ce problème, des nouvelles approches appelées système de détection d'intrusion (IDS) surveille le trafic réseau, surveille les activités suspectes et alerte l'administrateur système ou réseau.

Problématique

Les systèmes et réseaux d'information sont constamment exposés à des attaques résultant d'une mauvaise utilisation du système informatique par des utilisateurs externes comme des utilisateurs internes. L'importance de la sécurité des systèmes informatiques stimule divers angles de recherche visant à fournir de nouvelles solutions prometteuses qui ne peuvent être garanties par les méthodes traditionnelles. Les systèmes de détection d'intrusion font partie de ces solutions qui permettent de détecter une utilisation non autorisée.

Cependant, les systèmes et réseaux à protéger sont de plus en plus complexes et volumineux, et la nature des interventions actuelles et futures nous incite à développer des outils de défense automatique. Les systèmes de détection d'intrusion sont largement répandus de nos jours pour la sécurité de systèmes informatiques. Diverses techniques d'‘apprentissage automatique ont été utilisées pour développer les IDS. Avec l'apparition des techniques d'apprentissage profond et ses succès dans plusieurs domaines, nous avons élaboré une approche basée sur la machine Learning.

Organisation du mémoire

Les chapitres qui composent cette thèse sont organisés en quatres chapitres :

1. **La sécurité informatique :** Nous présentons dans ce chapitre les notions générales de la sécurité informatique en débutant par donner des définitions de da sécurité informatique et les points essentiels que touche la notion de sécurité tel que : les réseaux, les systèmes et applications.

2. **Détection d'intrusion :** nous présentons dans ce chapitre la littérature nous offre en essayant de ce focaliser sur la notion d'intrusion, la détection d'intrusion et enfin les SDI en passant par les types d'IDS.

3. **Machine learning :** Nous avons présenté dans ce chapitre les concepts de base de Machine learning .A la fin du chapitre nous la notion du Data Mining et Data Science. Et la différence entre eux.

4. **Expérimentation et résultats :** Dans ce chapitre nous avons présenté le corpus KDDCup99, qui est utilisé pour l'évaluation des anomalies et la détection d'intrusion. Enfin de ce chapitre nous avons argumenté les résultats fournis par notre algorithme.

Chapitre 1

La sécurité informatique

1.1 Introduction

Avec le développement des réseaux de communication, l'internet est devenu une chose très important dans la vie des gens, La croissance explosive d'utilisateurs d'Internet a motivé l'expansion rapide de commerce électronique et d'autres services en ligne. Malheureusement derrière la convenance et l'efficacité de ces services, les risques et les chances d'intrusions malveillantes sont aussi augmentés. La sécurité des systèmes informatiques est devenue un défi majeur dont l'objectif est d'assurer la disponibilité des services, la confidentialité et l'intégrité des données et des échanges.

De nombreux mécanismes ont été développés pour assurer la sécurité des systèmes informatiques, particulièrement pour détecter les intrusions dont le but principal est de construire un système sécurisé en déterminant et éliminant les vulnérabilités de sécurité.

Citons par exemple l'authentification qui consiste à prouver l'identité des utilisateurs, le contrôle d'accès qui consiste à définir les droits d'accès accordés aux utilisateurs sur les données et les pare-feux qui filtrent l'accès aux services du système informatique vis-à-vis de l'extérieur.

1.2 La sécurité informatique

1.2.1 Définitions de la sécurité informatique

Les diverses définitions de la sécurité informatique que relate la littérature spécifient que c'est l'ensemble des techniques et outils collaborations permettant de garantir trois objectifs essentiels de la sécurité (confidentialité, intégrité, et disponibilité) , ces outils peuvent êtres, organisationnels, matériels, logiciels, ou juridiques dont le but est de protéger les informations et les systèmes d'information contre l'accès, l'utilisation malveillante ou non autorisée, la modification, la divulgation, et la destruction des données et connaissances.[1]

1.2.2 Objectifs de la sécurité

L'objectif de la sécurité est d'assurer les cinq principes clés suivants :

- **La confidentialité :** est définie par l'organisation internationale de normalisation (ISO) comme « le fait de s'assurer que l'information n'est seulement accessible qu'à ceux dont l'accès est autorisé », elle consiste à préserver la révélation non autorisée d'information sensible. La révélation pourrait être intentionnelle comme les attaques qui visent de casser le chiffrement des données et lire les informations, ou involontaire dû au manque de vigilance ou de l'incompétence des individus qui manient les informations.[2]

- **La disponibilité :** assure la pérennité du service opportun aux utilisateurs autorisés qui ont un accès non interrompu aux informations dans le système et le réseau.[2]

- **L'intégrité :** est la propriété d'une information de ne pas être altérée. Donc le système informatique doit :[2]
  - → Empêcher une modification par une personne non autorisée ou une modification incorrecte par une personne autorisée.
  - → Faire en sorte qu'aucun utilisateur ne puisse empêcher une modification légitime de l'information. En plus, il faut se prémunir contre les fautes affectant l'intégrité des données, en intégrants dans le système des mécanismes permettant de détecter les modifications des informations d'une part et de contrôler l'accès à ces dernières d'autre part (en gérant les droits d'accès des programmes et utilisateurs). De plus, une validation en amont peut également être réalisée pour prévenir les fautes accidentelles.

- **L'authentification :** est tout simplement le contrôle d'accès, ce service signifie que celle-ci les personnes autorisées peuvent accéder aux informations, des simples moyens comme la gestion des mots de passe permet de garantir les services d'authentification.[2]

- **La non répudiation :** Cette propriété garantie qu'un sujet ayant réalisé une action dans le système ne puisse nier l'avoir réalisée. Assurer la non répudiation d'une transmission signifie assurer que les extrémités d'une transmission (émetteur et récepteur) sont bien les seules personnes autorisées à envoyer ou réceptionner les informations sans aucune remise en cause, qui est d'ailleurs généralement garanti par le moyen d'un fichier électronique appelé certificat numérique qui assure l'identité de l'émetteur et le récepteur. Les certificats eux mêmes sont protégés par le moyen des signatures des utilisateurs, dans certains cas la signature d'un utilisateur est son identité. L'importance de ces principes diffère selon le contexte de l'application,[2] par exemple :
  - → la confidentialité est la plus importante dans le cadre d'une transmission des messages secrets entre deux agences de sécurité nationale ou internationale, si quelqu'un arrive à décrypter le message transmis la sécurité sera compromise et l'information sera divulguée.
  - → Par contre la disponibilité est la plus importante pour les sites de e-commerce, la non-disponibilité est catastrophique pour des sites comme amazone et eBay.[2]

De façon générale, la sécurité informatique peut être définie par l'ensemble des moyens matériels, logiciels, et humains mis en oeuvre pour minimiser les vulnérabilités d'un système d'informations, et le protéger contre les menaces accidentelles ou intentionnelles, provenant de l'intérieur ou de l'extérieur de l'entreprise. Du point de vue organisationnel, nous pouvons découper le domaine de la sécurité informatique de la façon suivante :

- **La sécurité logicielle :** Gère la sécurité au niveau logiciel du système d'informations (par exemple : l'intégration des protections logicielles comme les antivirus).
- **La sécurité du personnel :** Comprend la formation et la sensibilisation des personnes utilisant ou travaillant avec le système d'informations.

- **La sécurité physique :** Regroupe la politique d'accès aux bâtiments, la politique d'accès aux matériels informatiques et les règles de sécurité pour la protection des équipements réseaux.
- **La sécurité procédurale :** Définit les procédures et les règles d'utilisation du système d'informations.
- **La sécurité réseau :** S'occupe de l'architecture physique et logique du réseau, la politique d'accès aux différents services, la gestion des flux d'informations sur les réseaux et surtout les points de contrôle et de surveillance du réseau.
- **La veille technologique :** Souvent oubliée permet d'évaluer la sécurité au cours du temps afin de maintenir un niveau suffisant de protection du système d'informations.

1.2.3 Attaques de sécurité

Une attaque peut être définie comme toute action ou ensemble d'actions qui peut porter atteinte à la sécurité des informations d'un système ou réseau informatique. Etant donné le nombre important d'attaques possibles, nous allons d'abord commencer par les classer puis nous en présenterons quelques-unes.

1.3 Classification des attaques

Vu le nombre important des attaques possibles, elles peuvent être classées selon différentes classifications :[3]

1.3.1 La première classification

Se compose de deux types :

- **Les attaques passives :** ce type d'attaque vise à l'obtention d'accès pour pénétrer dans le système sans compromettre ces ressources.
- **Les attaques actives :**dont le résultat de cette attaque est un changement non autorisé d'état des ressources de système.

1.3.2 La deuxième classification

Se compose de deux types :

- **Les attaques internes :** ce type d'attaque est causé : Soit par les utilisateurs autorisés du système qui essayent d'utiliser des privilèges complémentaires dont ils n'ont pas le droit.
- **Les attaques externes :** ce type d'attaque est causé par des utilisateurs externes qui essayent d'accéder à des informations ou des ressources d'une manière illégitime et non autorisée.

1.3.3 La troisième classification

Selon cette classification, les attaques de cette catégorie peuvent porter atteinte à :

- **La confidentialité** des informations en brisant les règles privées.
- **L'intégrité** en altérant les données.
- **l'authenticité** des données.
- **La disponibilité** en rendant un système ou un réseau informatique indisponible. On parle alors d'attaque de **déni de service.**

1.3.4 La quatrième classification

Les attaques de sécurité peuvent également être classées en termes :

- **d'attaques réseaux :** leur but principal est d'empêcher les utilisateurs d'utiliser une connexion réseau, de rendre indisponible une machine ou un service et desurveiller le trafic réseau dans le but de l'analyser et d'en récupérer des informations pertinentes.
- **d'attaques systèmes :**ce sont des attaques qui portent atteinte au système, comme par exemple effacer des fichiers critiques (tel que le fichier "password") ou modifier la page web d'un site dans le but de le discréditer ou tout simplement le ridiculiser.

1.4 Exemples des attaques informatiques

1.4.1 L'attaque de déni de service distribué (DDoS)

Elle représente la version distribuée de l'attaque de déni de service. Le but de cette variante de l'attaque DoS est que la victime n'arrive pas à isoler les attaquants vu le nombre important des machines utilisées pour réaliser cette attaque. Pour réaliser cette attaque, il faut premièrement pénétrer par diverses méthodes des systèmes dits "handlers" et agents. Où l'attaquant contrôle un ensemble de systèmes "handlers" qui contrôlent eux-mêmes un ensemble de systèmes agents. Le hacker lance l'attaque en ordonnant les systèmes "handlers", qui eux-mêmes ordonnent les agents.[4]

Le schéma suivant illustre le fonctionnement de l'attaque DDoS (Voir La figure 1.1)

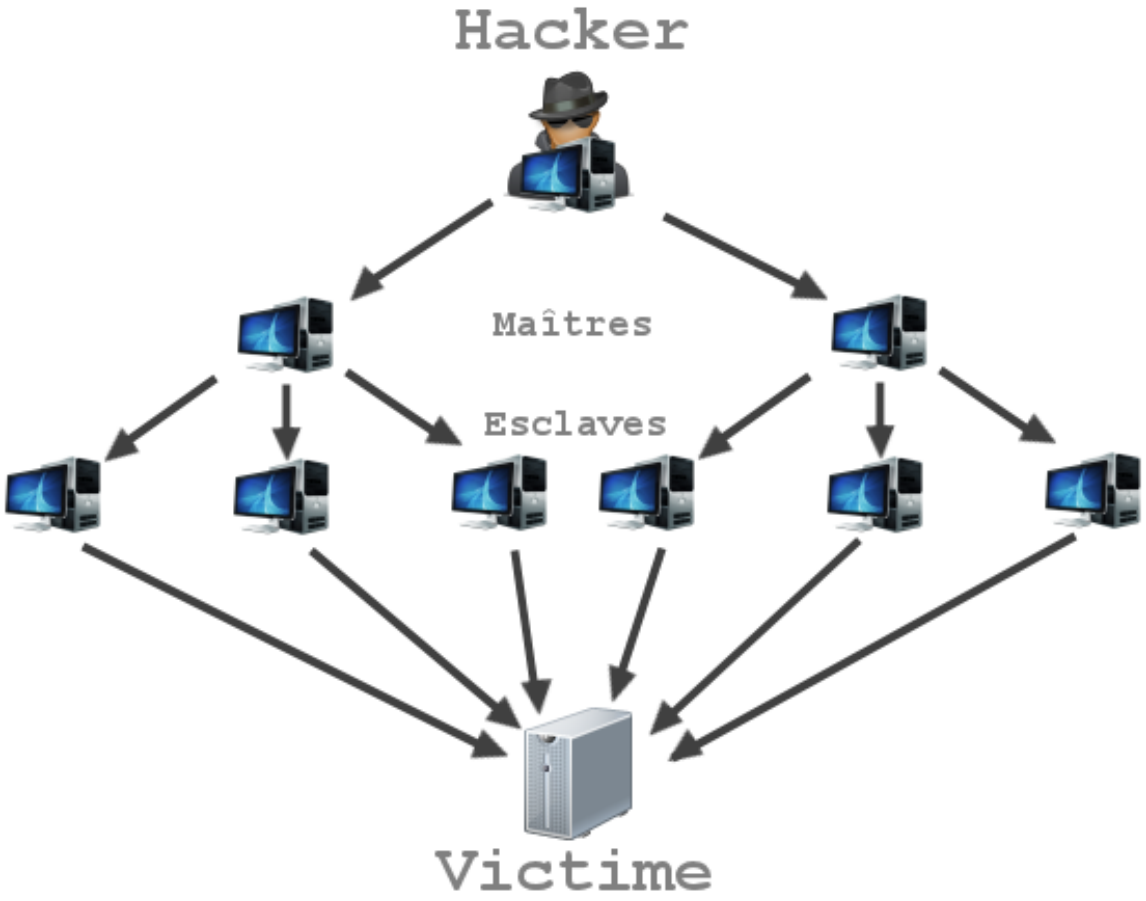


Figure 1.1 – L'attaque Ddos [5]

1.4.2 L'attaque man-in-the-middle

L'attaquant s'introduit entre deux systèmes sans que l'un d'entre eux aperçoive l'existence d'un troisième système qui fait passer les échanges réseau. Pour réussir une telle attaque, il faut que la machine de l'attaquant soit physiquement entre les deux machines victimes ou que l'attaquant arrive à modifier le routage réseau afin que sa machine devienne un des points de passage.[4]

Le schéma suivant illustre le fonctionnement de l'attaque man-in-the middle (Voir la figure 1.2

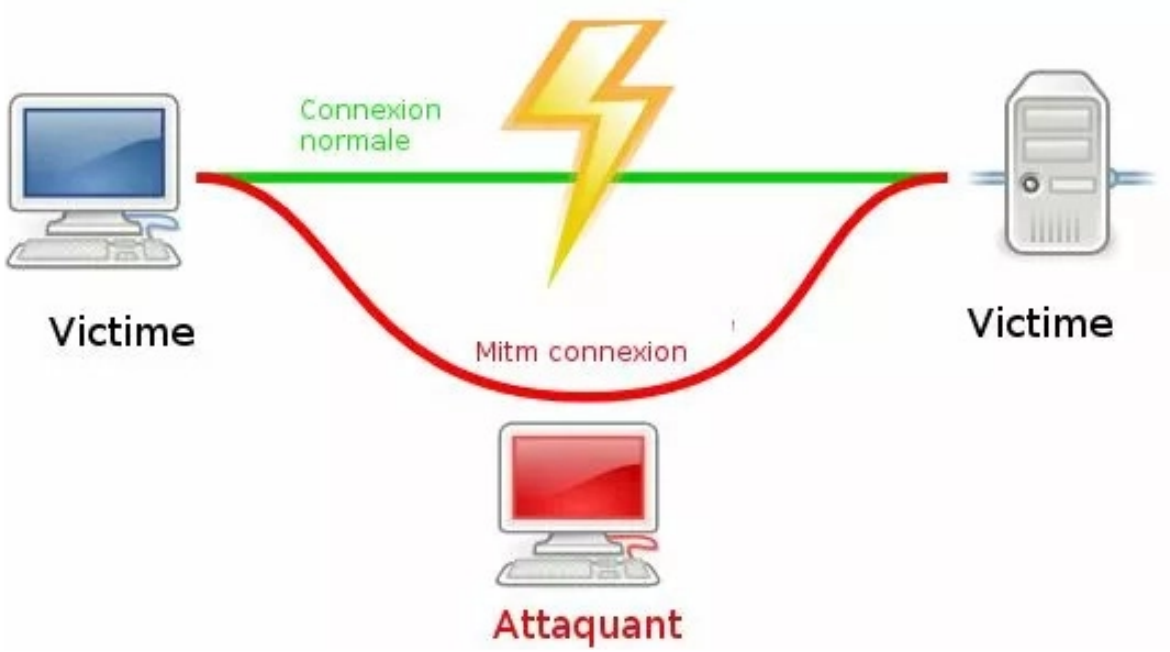


Figure 1.2 – L'attaque man-in-the-middle [5]

1.4.3 Les attaques virales

Il existe principalement quatre types de menaces distinctes, mais il arrive fréquemment qu'une attaque s'appuie sur plusieurs mécanismes :

- **Les virus :** Se reproduisent en infectant le corps de programmes hôtes0
- **Les chevaux de Troie (trojan horses) :** Exécutent des tâches malignes en se dissimulant au sein d'une coquille applicative d'aspect inoffensif0
- **Les vers (worms) :** Profitent des réseaux informatiques pour se propager, par courrier électronique par exemple
- **Les accès cachés (backdoors) :** Permettent à un utilisateur externe de prendre le contrôle d'une application par des moyens détournés.

1.5 Les techniques de sécurité

La mise en œuvre d'une politique de sécurité consiste à déployer les différents moyens et dispositifs visant la sécurisation du système d'information ainsi que l'application des règles définit dans la politique de sécurité adoptée. Ce qui signifie, faire le bon choix de l'ensemble des mécanismes et des techniques les plus simples possible permettant de protéger les ressources d'une manière très efficace avec un faible coût. Il existe différentes techniques utilisées contre les attaques informatiques, ces techniques sont classées en cinq catégories qui sont : [4]

1.5.1 La sécurisation des accès réseau

La maîtrise du flux réseau à l'aide des pare-feux assure un niveau de confidentialité des données grâce aux protocoles de sécurité tel que l'IPSec, ce qui permet la sécurisation des accès réseau.

1.5.2 Superviser les connexions réseau

La vérification du trafic réseau consiste à ne laisser passer que les connexions autorisées. Cela est possible par :

- la création d'un périmètre de sécurité
- limiter le nombre de points d'accès pour rendre la gestion de la sécurité plus facile.
- Et disposer de trace des systèmes en cas d'incident de sécurité. Nous citons certains dispositifs de contrôle et de filtrage de connexion.
- Le pare-feu : C'est le système qui permet de mettre en œuvre la politique du filtrage au sien de l'organisation, selon plusieurs principes de filtrage .
  - le filtrage des paquets au niveau réseau (IP, etc.).
  - le filtrage à mémoire des paquets de manière dynamique.
  - la passerelle de niveau transport filtrant les paquets en gérant le concept de session.
  - la passerelle de niveau applicatif filtrant les paquets du niveau applicatif.
- Contrôle de l'accès réseau : C'est un nouveau concept développé par Cisco, et ayant pour but le contrôle des accès les plus près à leurs sources où il permet de vérifier un certain nombre de points de sécurité avant d'autoriser un système à se connecter au réseau local.

1.5.3 Assurer la confidentialité des connexions

La confidentialité des données est assurée au sien d'un réseau informatique par l'utilisation du chiffrement, par un cryptage des données avant leur envoi et un décryptage à leur réception. Le schéma suivant (Voir Figure 1.4) montre ce le chiffrement dans l'architecture de communication TCP/IP. [4]

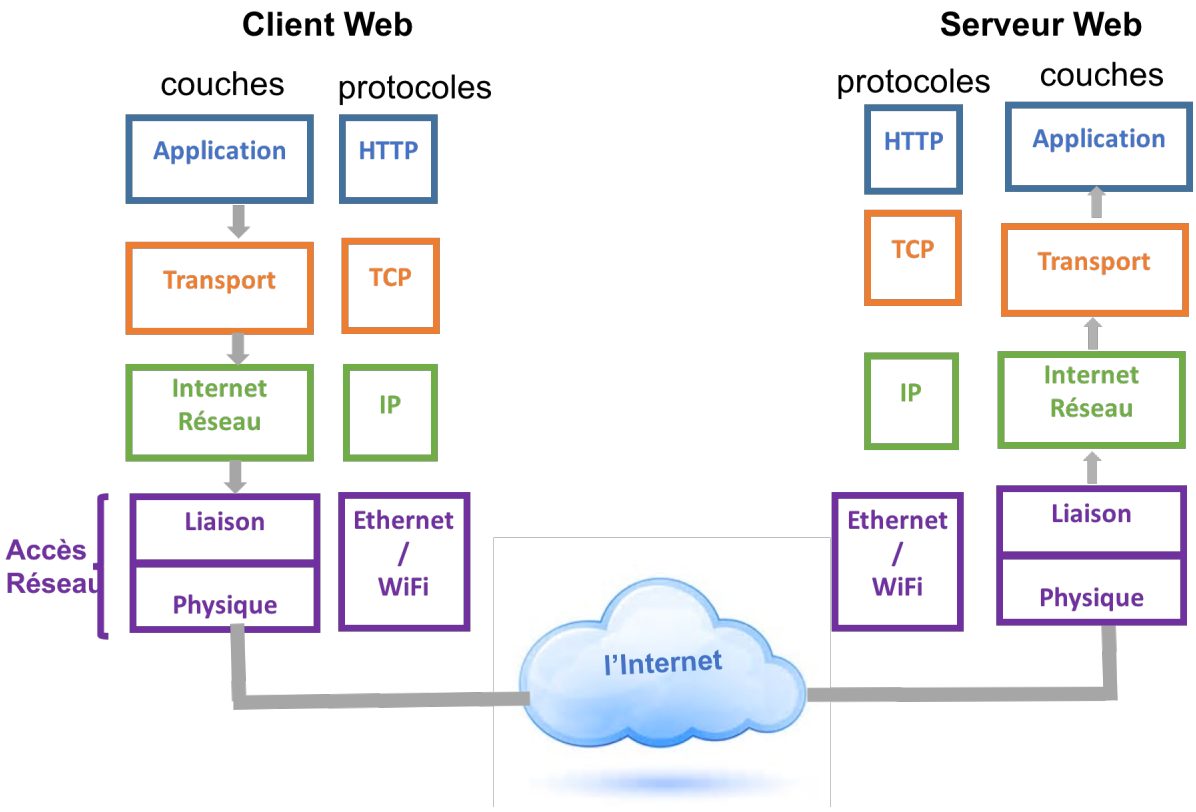


Figure 1.3 – La représentation en couches des protocoles de sécurité

- Les algorithmes cryptographiques : le chiffrement-déchiffrement des données est effectué par des algorithmes cryptographiques qui reposent sur des problèmes mathématiques difficiles à résoudre. Il existe deux grandes catégories d'algorithme de cryptographie :[6, 7]
  - les algorithmes cryptographiques à clé secrète ou symétrique qui se basent sur une même clé qui chiffre et déchiffre. Cette clé est partagée par les deux communicants
  - Les algorithmes cryptographiques à clé publique ou asymétrique qui se basent sur une clé publique de chiffrement et une clé secrète de déchiffrement.[4]

Il existe aussi les algorithmes de hachage qui nous permettent d'obtenir une signature numérique à partir des données comme :

- IPSec : il est créé pour faire face aux problèmes d'authentification et de confidentialité du protocole IP. IPSec opère au niveau IP et il encapsule nativement tous ces protocoles (TCP, UDP, ICMP, etc.). IPSec offre des services de contrôle d'accès, d'intégrité, d'authentification, de confidentialité de plus il fait face aux attaques de type paquets replay [6, 7].
- SSL (Secure Sockets Layer) : opère au-dessus de la couche TCP et offre aux navigateurs internet la possibilité d'établir des sessions authentifiées et chiffrées. Le protocole SSL a été standardisé par le groupe de travail TLS (Transport Layer Security) formé au sein de l'IETF [6, 7].

- SSH (Secure Shell) : il opère au niveau application et permet d'obtenir un interprète des commandes (Shell) à distance d'une manière sécurisé [6, 7].

1.5.4 La protection des équipements réseau

Protéger un réseau informatique c'est assurer la protection des équipements qui le composent et qui recouvre les trois domaines suivants [6, 7] :

- **La protection physique :** c'est la sécurité physique des équipements face aux menaces physiques externes comme le feu, l'inondation, le survoltage, l'accès illégal à la salle informatique. . . etc.
- **La protection du système d'exploitation :** c'est la sécurité des systèmes d'exploitation contre les faiblesses de sécurité ou les bugs.
- **La protection logique :** la mise en œuvre d'une politique de sécurité passe par une configuration de l'équipement réseau.

  La sécurité des équipements réseau nous permet de se protéger contre les attaques suivantes.[7]
- Les attaques par déni de service visant à exploiter des faiblesses de configuration.
- Les attaques permettant d'obtenir un accès non autorisé à un équipement réseau suite à des faiblesses de configuration
- Les attaques exploitant un bug référencé du système d'exploitation Cisco, Microsoft, RedHat...

1.6 Conclusion

Nous avons présenté dans ce chapitre une introduction à la sécurité informatique et on a définie les différents propriétés de la sécurité avec les services, d'autre part on a définie les attaques avec les différents classification , il est donc nécessaire d'assurer la protection des système informatique, afin de lutter contre les menaces qui pèsent sur l'intégrité, la confidentialité et la disponibilité des ressources.

La malveillance informatique est souvent à l'origine de ces menaces, qu'il s'agisse de vol d'information ou de sabotage, n'importe qui pouvant s'improviser pirate informatique avec des outils adaptés. Beaucoup de compétences sont nécessaires pour assurer une sécurité optimale, mais il est impossible de garantir la sécurité de l'information à 100%. Malgré tout, il existe des moyens efficaces pour faire face à ces agressions.

C'est pour cela qu'il est utile de bien savoir gérer les ressources disponibles et comprendre les risques liés à la sécurité informatique, pour pouvoir construire une politique de sécurité adaptée aux besoins de la structure à protéger. La mise en place d'un dispositif de sécurité efficace ne doit cependant jamais dispenser d'une veille régulière au bon fonctionnement du système.

Chapitre 2

Détection d'intrusion

2.1 Introduction

L'ordinateur se faisant chaque jour plus présent dans nos vies quotidiennes, la question de la sécurité informatique prend elle aussi de l'importance. Avec l'essor d'internet et du « tout connecté », nous dépendons de plus en plus de la fiabilité de nos appareils, sans même parfois nous en rendre compte. alors grâce a ces différent problème sécurité des systèmes de détection d'intrusions ont été apparu.

2.2 Intrusion

Une intrusion est toute utilisation d'un système informatique à des fins autres que celles prévues, généralement dues à l'acquisition de privilèges de façon illégitime. L'intrus est généralement vu comme une personne étrangère au système informatique qui a réussi à en prendre le contrôle, mais les statistiques montrent que les utilisations abusives (du détournement de ressources à l'espionnage industriel) proviennent le plus fréquemment de personnes internes ayant déjà un accès au système.[8]

Une intrusion dans un système informatique est aussi définie par Heady et al. comme :[9]

«N'importe quel ensemble d'actions essayant de compromettre l'intégrité, la confidentialité ou l'accessibilité d'une ressource».

En dépit de différentes formes d'intrusions, elles peuvent être regroupées dans deux classes :[10]

- **Les intrusions connues :** Ces intrusions sont des attaques bien définies qui généralement exploitent des failles connues du système cible.
- **Les intrusions inconnues ou anomalies :** Ces intrusions sont considérées comme des déviations du profil normal d'un système. Elles sont détectées des qu'il est observé un comportement anormal du système.

2.3 Protection

Nous venons de voir qu'il existe un grand nombre d'attaques connues. Une première idée est de vouloir lutter contre celles déjà existantes. Pour cela, on utilise un IDS qui surveille l'état de l'ensemble du système à protéger. Il existe plusieurs manières pour différencier les types d'IDS. Ainsi, certains s'implémentent de manière software alors que d'autres le sont en hardware. Une autre différenciation se fait sur ce que l'IDS regarde. Il peut être HIDS (Host-based Intrusion Detection System), NIDS (Network-based Intrusion Detection System) ou un mélange des deux types. [11]

2.4 Système de détection d'intrusion

La détection d'intrusions c'est l'analyse des informations collectées par les mécanismes d'audit de sécurité, à la recherche d'éventuelles attaques sur les systèmes informatiques.

Les méthodes de détection d'intrusion diffèrent sur la manière d'analyser le journal d'audits.[8]

2.4.1 Définition d'un IDS

Un système de détection d'intrusion (IDS) est tout équipement, méthode ou ressource nous permettant de surveiller un réseau ou un hôte donné afin de prévoir ou identifier toute action suspecte et non autorisée et éventuellement réagir à cette action. Les systèmes de détection d'intrusion actuels détectent les activités réseau qui peuvent être une intrusion ou non et non pas l'intrusion. La détection d'intrusion est précisément une partie d'un système de protection total installé autour d'un système ou appareil. Il n'est pas une mesure de protection autonome [12]

Modèle simplifié d'un IDS

Selon Hervé Debar, on a simplifié un IDS en un détecteur qui analyse les informations en provenance du système surveillé (Voir Figure 2.1).[13]

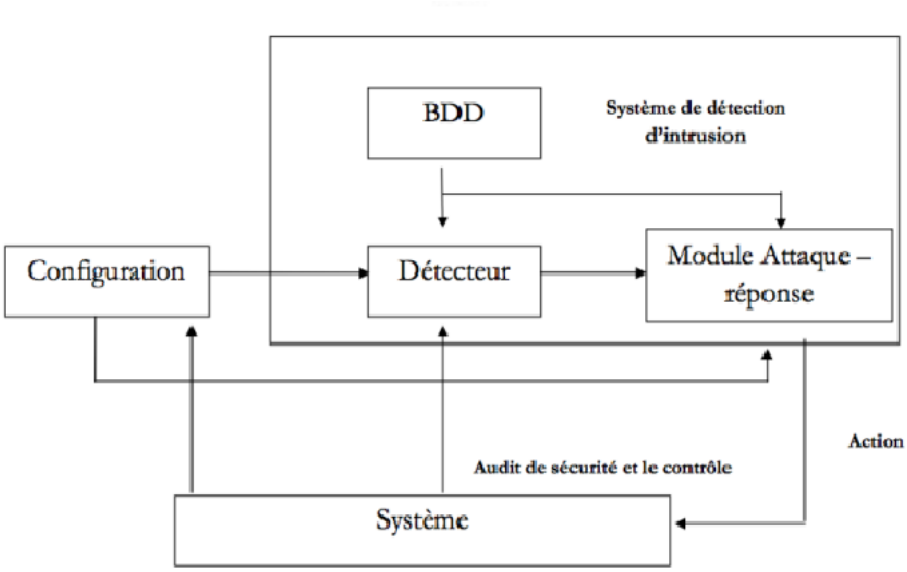


Figure 2.1 – Modèle simplifié d'un système de détection d'intrusions.[14]

Un détecteur, qui est l'élément principal de cette architecture, analyse trois types d'informations :

1. Les informations relatives aux techniques utilisées dans la détection (Base de donnée de signatures).
2. Les informations relatives à la configuration du système déterminant son état actuel.
3. Les informations relatives à l'audit décrivant les évènements survenus dans le système.

2.4.2 Évaluation des IDS

Les systèmes de détection d'intrusions actuels tendent à garantir les cinq propriétés suivantes :[15]

1. **Exactitude :** elle se traduit par une détection parfaite des attaques avec un risque minimal de faux positifs.
2. **Performance :** une détection rapide des intrusions avec une analyse approfondie des événements est indispensable pour mener une détection efficace en temps réel.
3. **complétude :** une détection exhaustive des attaques connues et inconnues.
4. **Tolérance aux fautes :** les systèmes de détection d'intrusions doivent résister aux attaques ainsi qu'à leurs conséquences.
5. **Rapidité :** une analyse rapide des données permet d'entreprendre instantanément les contre mesures nécessaires pour stopper l'attaque et protéger les ressources du réseau et du système de détection d'intrusions.

2.4.3 Les Caractéristiques d'un système de détection d'intrusions

Tout IDS doit présenter les caractéristiques suivantes [7] :

- Pouvoir effectuer une surveillance permanente et émettre une alarme en cas d'intrusion.
- Fournir beaucoup d'information pour pouvoir réparer le système et la responsabilité de l'intrus.
- S'adapter aux différentes plates formes et architectures réseaux par sa modulation et configuration.
- Assurer sa propre défense en supportant qu'une partie ou la totalité du système soit hors service.
- Avoir un taux de faux positifs faible.
- Avoir une réponse automatique en cas d'attaques coordonnées ou distribuées.
- Être en mesure de repérer les premiers événements de corruption pour réparer correctement le système.
- Ne pas créer de vulnérabilités supplémentaires.

Les systèmes de détection d'intrusion offrent beaucoup d'avantages comme [7] :

— Beaucoup plus efficace qu'une détection manuelle des intrusions.

— La prédiction des intrusions par l'utilisation d'une base de connaissance plus grande.

— La capacité de traiter un large volume de données.

— La réaction de l'IDS par une alerte en temps réel réduit le dommage important des attaques.

— Des mesures de contre-attaque automatique sont prises comme la fermeture des sessions, désactivation des comptes utilisateur...

2.5 Classification des systèmes de détection d'intrusion

Les Systèmes de détection d'intrusion existants peuvent être classifiés d'après plusieurs critères. Nous citons dans la figure suivante (Voir Figure 2.2) les cinq critères de classification des IDS introduits par Hervé Debar, Marc Dacier et Andreas Wespi. [13]

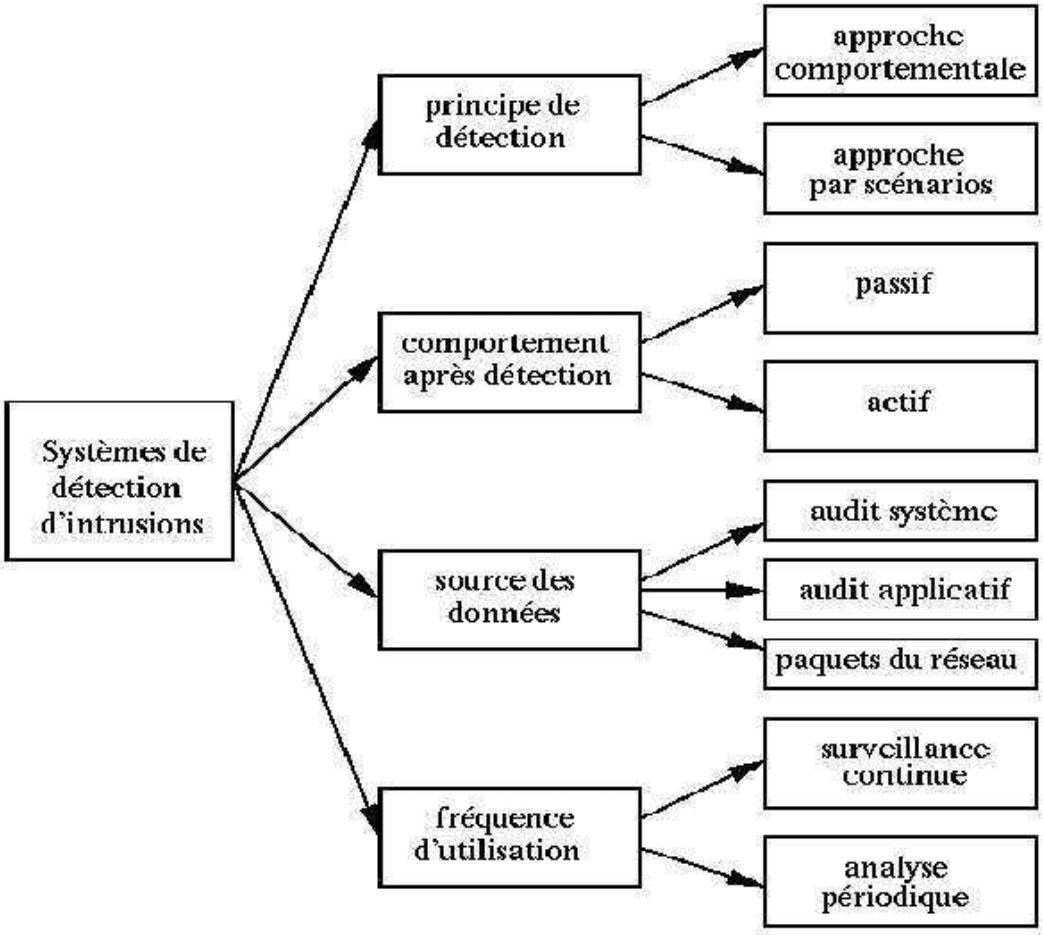


Figure 2.2 – Les critères de classification des IDSs[16]

2.6 Les méthodes de détection d'intrusion

A ce jour deux méthodes de détection sont proposées, la première est appelée approche Comportementale (anomaly detection) qui consiste à créer un modèle basé sur le comportement normal du système et toute déviation par rapport à celui-ci est considérée comme suspecte. La seconde est appelée approche par scénario (misuse detection ou knowledge based detection) qui consiste à utiliser des connaissances accumulées sur les attaques puis on en tire des scénarios d'attaques et on recherche dans les traces d'audit leur éventuelle survenue [7].

2.6.1 L'approche comportementale

Cette approche a été proposée par Anderson [17] puis développée par Denning [18], et se base sur l'hypothèse que l'exploitation d'une faille du système nécessite une utilisation anormale du système, donc un comportement inhabituel de l'utilisateur. Elle repose sur l'observation du système et toute déviation par rapport au comportement normal prévu du système est considérer comme intrusion.

Cette approche consiste, dans une première phase, à définir un modèle de comportement du système, des utilisateurs des applications, etc. qui sera considéré comme « normal ». Dans une seconde phase, l'activité actuelle du système est confrontée avec le modèle établi dans la phase une par l'IDS. En cas où une déviation est détectée, une alerte sera déclenchée.

En plus, cette approche considère comme intrusion, tout comportement qui n'est pas précédemment enregistré. Par conséquent, la précision reste son plus grand souci [17, 18].

2.6.1.1 Avantages

- La détection de nouvelles formes d'attaques exploitant de nouvelles formes de vulnérabilités non connues auparavant.
- Elle est moins dépendante du système d'exploitation par rapport à l'approche par scénario.
- Elle détecte les attaques d'abus de privilège qui n'exploite aucune vulnérabilité.

2.6.1.2 Inconvénients

- Un taux de fausses alarmes très élevé parce que l'ensemble du périmètre du comportement d'un système d'information ne peut pas être complètement couvert pendant la phase d'apprentissage. En outre, le comportement peut changer au fil du temps. Ce qui nous oblige à refaire l'apprentissage du comportement normal, ce qui cause soit l'indisponibilité temporaire du système de détection d'intrusion ou des fausses alarmes supplémentaires.
- De plus le système d'information peut subir des attaques au moment d'apprentissage. Par conséquent, le profil de comportement normal contiendra des comportements intrusifs qui ne seront pas détectés comme anormale.
- n cas de profondes modifications de l'environnement du système cible, le modèle statistique déclenche un flot d'alarmes ininterrompu, du moins pendant la phase de transition du système.
- Il est difficile d'affirmer que les observations faites sont des activités à prohiber.

Les systèmes de détection d'intrusion basés sur cette approche sont implémentés à l'aide de différentes techniques tel que : les systèmes experts, la méthode statistique, les réseaux de neurone [13].

2.6.2 L'approche par scénario

Elle vise à détecter des signes d'attaques connues selon une BDD d'attaques connues (les connaissances accumulées sur des attaques spécifiques et les vulnérabilités du système). Le système de détection d'intrusion contient les informations sur les vulnérabilités et cherche toute tentative de les exploiter. L'IDS confronte le comportement observé du système à la base de données d'attaques connues, Si ce comportement correspond à l'une des signatures de la base une alerte est déclenchée. En d'autres termes, toute action qui n'est pas explicitement reconnue comme une attaque est considérée comme acceptable. Par conséquent, la précision des systèmes de détection d'intrusion basée sur l'approche par scénario est bonne. Cependant, cette précision dépend toujours de la mise à jour des connaissances sur les attaques qui doit être régulière [7, 13].

2.6.2.1 Avantages

- le taux très faible de fausse alarme.
- l'analyse contextuelle très détaillée.

2.6.2.2 Inconvénients

- Les bases de signatures sont difficiles à construire.
- La difficulté de mettre à jour la base des attaques connues avec les nouvelles vulnérabilités.
- La difficulté de mettre à jour la base des attaques connues avec les nouvelles vulnérabilités.
- Un IDS est lié à son environnement par le fait que la connaissance des attaques est très liée au système d'exploitation, la version, la plateforme et l'application.
- Les attaques par abus de privilèges (interne) sont difficiles à détectées, car il n'y a aucune vulnérabilité exploitée par l'attaquant.
- Il n'y a pas de détection d'attaques non connues[7].

Actuellement, de nombreux systèmes de détection d'intrusion dites par scénario ont été implémentés à l'aide de multiples techniques tel que : les systèmes experts, l'analyse de la signature, les réseaux de pétri, l'analyse de l'état transition[7].

Les systèmes de détection d'intrusion commerciaux actuels se penchent vers l'approche par scénario pour les raisons suivantes [7] :

- L'approche par scénario est plus facile que l'approche comportementale dans l'implémentation.
- le taux élevé de fausse alarme pour l'approche comportementale la rend inapproprié pour des IDSs commerciaux.
- La vitesse de traitement des audits est un facteur très important c'est la raison pour laquelle les signatures sont utilisées à la place des règles.

2.7 Les différents types d'IDS

En raison de la variété des attaques menées par les pirates, la détection d'intrusion doit être effectuée à plusieurs niveaux. Il existe donc différents types d'IDS :

- IDS basé sur le hôte
- IDS basé sur le réseau

2.7.1 IDS basé sur le hôte

Un IDS basé sur l'hôte analyse plusieurs domaines pour déterminer le mauvais usage (activité malveillante ou abusive à l'intérieur du réseau) ou des intrusions (brèches de l'extérieur). Les IDS basés sur l'hôte consultent plusieurs types de fichiers journaux (noyau, système, serveur, réseau, pare-feu et autres) et comparent les journaux à une base de données interne de signatures courantes d'attaques connues. Les IDS UNIX et Linux basés sur l'hôte utilisent énormément la commande syslog et sa capacité à séparer les évènements enregistrés selon leur sévérité (par exemple, des messages d'imprimante mineurs contre des avertissements du noyau majeurs). La commande syslog est disponible lors de l'installation du paquetage sysklogd, inclus avec Red Hat Enterprise Linux. Ce paquetage offre une journalisation du système et la capture de messages du noyau. Les IDS basés sur l'hôte filtrent les journaux (qui, dans le cas de journaux d'évènements de noyau ou de réseau, peuvent être très commentés), les analysent, marquent à nouveau les messages avec leur propre système d'évaluation de sévérité et les rassemblent dans leur propre journal spécialisé pour être analysé par les administrateurs[19].

Les IDS basés sur l'hôte peuvent également vérifier l'intégrité de données de fichiers et d'exécutables importants. Ils vérifient une base de données de fichiers confidentiels (et tout fichier ajouté par l'administrateur) et créent une somme de contrôle de chaque fichier avec un utilitaire d'analyse de fichiers messages comme la commande md5sum (algorithme 128-bit) ou la commande sha1sum (algorithme 160-bit). Les IDS basés sur l'hôte sauvegardent alors les sommes dans un fichier en texte clair et, de temps en temps, comparent les sommes de contrôle de fichiers avec les valeurs dans le fichier texte. Si l'une des sommes ne correspond pas, alors les IDS avertissent l'administrateur par courrier électronique ou pager[19].

2.7.1.1 Les avantages

➢ La capacité de contrôler les activités locales des utilisateurs avec précision. Capable de déterminer si une tentative d'attaque est couronnée de succès.

➢ L'IDS base sur le hôte fonctionne sur les traces d'audit des système d'exploitation ce qui lui permet de détecter certains types d'attaques (ex : Cheval de Troie).

2.7.1.2 Les inconvénients

➢ La vulnérabilité aux attaques du type déni de service puisque l'IDS peut résider dans l'hôte cible par les attaques.

- ➢ La difficulté de déploiement et de gestion, surtout lorsque le nombre d'hôtes qui ont besoin de protection est large.
- ➢ Ces systèmes sont incapables de détecter des attaques contre de multiples cibles dans le réseau

2.7.2 IDS basé sur le réseau

Les systèmes de détection d'intrusions basés sur le réseau fonctionnent différemment des IDS basés sur l'hôte. La philosophie de conception d'un IDS basé sur le réseau est de scanner les paquets réseau au niveau de l'hôte ou du routeur, analysant les informations de paquets et enregistrant tous les paquets suspects dans un fichier journal spécial avec des informations détaillées. Selon ces paquets suspects, un IDS basé sur le réseau peut scanner sa propre base de données de signatures d'attaques réseau connues et assigner un niveau de sévérité pour chaque paquet. Si les niveaux de sévérité sont assez élevés, un message électronique d'avertissement ou un appel de pager est envoyé aux membres de l'équipe de sécurité afin qu'ils puissent étudier plus en profondeur la nature de l'anomalie [19].

Les IDS basés sur le réseau sont devenus populaires avec l'internet grandissant en taille et trafic. Les IDS qui peuvent scanner les quantités volumineuses d'activités réseau et marquer avec succès les transmissions suspectes, sont accueillis dans le domaine de la sécurité. Les protocoles TCP/IP n'étant peu sûrs de nature, il est devenu impératif de développer des scanneurs, des renifleurs et d'autres outils d'analyse de réseau et de détection pour éviter les brèches de sécurité provenant d'activités réseau malveillantes comme [19] :

- L'usurpation d'identité
- Les attaques par déni de service
- La corruption de cache arp
- La corruption de noms DNS
- Les attaques man-in-the-middle

2.7.2.1 Les avantages

- ➢ L'IDS base réseau est capable de contrôler un grand nombre d'hôte avec un petit cout de déploiement.
- ➢ Il n'influence pas sur les performances des entités surveillées.
- ➢ L'IDS base réseau est capable d'identifier les attaques de /a multiples hôtes.
- ➢ L'IDS base réseau assure une grande sécurité contre les attaques parce qu'il est invisible aux attaquants.

2.7.2.2 Les inconvénients

- ➢ L'IDS base réseau ne peut pas fonctionner dans des environnements crêpes.
- ➢ Ce type d'IDS ne permet pas d'assurer si une tentative d'attaque est couronnée de succès.

➢ L'évaluation et la comparaison des systèmes de détection d'intrusions est un problème en soi de par la diversité des sources de données possibles et la représentativité des données utilisées lors des tests notamment. Une. Les systèmes de détection d'intrusions sont évalues traditionnellement suivant deux critères :

— **La fiabilité de l'IDS :** toute intrusion doit effectivement donner lieu à une alerte. Une intrusion non signalée constitue une défaillance de l'IDS, appelée faux négatif. La fiabilité d'un système de détection d'intrusions est liée à son taux de faux négatifs (c'est a-dire le pourcentage d'intrusions non-détectées), qui doit être le plus bas possible.

— **La pertinence des alertes :** toute alerte doit correspondre à une intrusion effective. Toute « fausse alerte » (appelée également faux positif) diminue la pertinence de l'IDS. Un bon IDS doit présenter un nombre de faux positifs aussi bas que possible.

2.8 Conclusion

Nous avons présenté dans ce chapitre une étude des systèmes de détection d'intrusions.

Ils nous est paru évident que ces systèmes sont à présent indispensables aux entreprises afin d'assurer leur sécurité. La plupart des systèmes de détection d'intrusions sont construits dans une architecture hiérarchique dont une grande quantité de données transférée à travers le réseau peut résulter une congestion de réseau et sont susceptibles d'être attaqués.

Pour offrir un système de détection d'intrusions pour les réseaux actuels mais aussi pour la nouvelle génération, nous proposons une nouvelle génération de systèmes de détection d'intrusions fondés sur des architectures utilisant des méthodes heuristique distribuées et basés sur des modèles informatique.

Chapitre 3

Machine Learning

3.1 Machine Learning

"L'apprentissage automatique permet à une machine d'apprendre automatiquement des données, d'améliorer les performances des expériences et de prédire les choses sans être explicitement programmée."

Arthur Samuel

Selon Arthur Samuel, l'apprentissage automatique est défini comme le domaine d'études qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmé. Arthur Samuel était célèbre pour son programme de jeu de dames. Machine learning (ML) est utilisé pour enseigner aux machines comment gérer les données plus efficacement. Parfois, après avoir vu les données, on ne peut pas interpréter l'information extraite des données. Dans ce cas, on applique l'apprentissage automatique. Avec l'abondance des ensembles de données disponibles, la demande pour l'apprentissage automatique est en hausse. De nombreuses industries appliquent l'apprentissage automatique pour extraire des données pertinentes. Le but de l'apprentissage automatique est d'apprendre des données. De nombreuses études ont été réalisées sur la façon de faire apprendre les machines par elles-mêmes sans être explicitement programmées. Beaucoup de mathématiciens et de programmeurs appliquent plusieurs approches pour trouver la solution de ce problème qui ont d'énormes ensembles de données [20].

3.1.1 Comment Machine Learning fonctionne-t-il ?

Un système d'apprentissage automatique apprend des données historiques, construit les modèles de prédiction, et chaque fois qu'il reçoit de nouvelles données, prédit la sortie pour elle. L'exactitude de la sortie prévue dépend de la quantité de données, car la quantité énorme de données aide à construire un meilleur modèle qui prédit la sortie plus précisément.

Supposons que nous ayons un problème complexe, où nous avons besoin d'effectuer des prédictions, donc au lieu d'écrire un code pour cela, nous avons juste besoin d'alimenter les données en algorithmes génériques, et avec l'aide de ces algorithmes, la machine construit la logique selon les données et de prédire la sortie. L'apprentissage automatique a changé notre façon de penser le problème. Le schéma ci-dessous explique le fonctionnement de l'algorithme d'apprentissage automatique :[21]

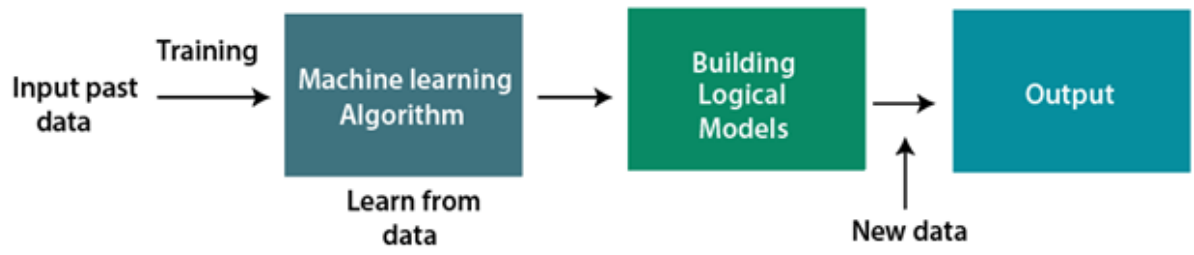


Figure 3.1 – Comment Machine Learning fonctionne-t-il ? [21]

3.1.2 Caractéristiques de Machine Learning

- Machine learning(apprentissage automatique) utilise des données pour détecter divers modèles dans un ensemble de données donné.
- Il peut apprendre des données passées et s'améliorer automatiquement.
- C'est une technologie axée sur les données.
- Apprentissage automatique ressemble beaucoup à l'exploration de données, car elle traite également de la quantité énorme de données.

3.1.3 Différences entre Machine learning algorithmes et les algorithmes traditionnels basés sur des règles

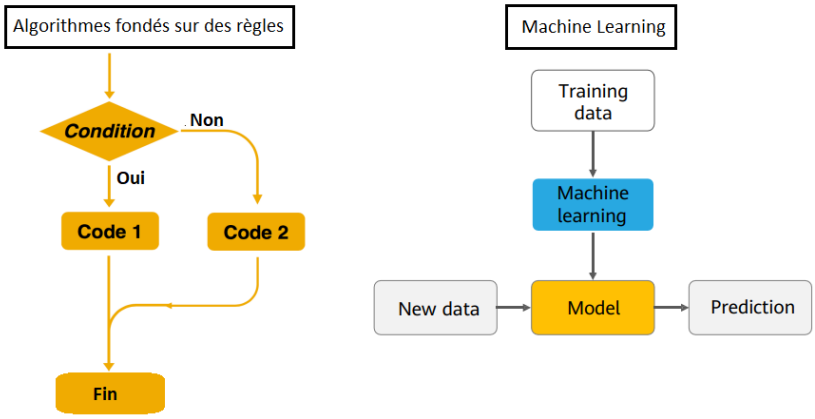


Figure 3.2 – Différences entre Machine learning algorithmes et les algorithmes traditionnels basés sur des règles

3.1.3.1 Algorithmes fondés sur des règles

- La programmation explicite est utilisée pour résoudre les problèmes.
- Les règles peuvent être spécifiées manuellement

3.1.3.2 Machine Leaning

- Des échantillons sont utilisés pour la formation.
- Les règles de prise de décision sont complexes ou difficiles à décrire.
- Les règles sont automatiquement apprises par les machines.

3.1.4 Machine Learning Classification

- **Apprentissage supervisé :** Obtenir un modèle optimal avec le rendement requis par la formation et l'apprentissage en fonction des échantillons de catégories connues. Ensuite, utilisez le modèle pour mappage toutes les entrées aux sorties et vérifiez la sortie dans le but de classer les données inconnues.[22]

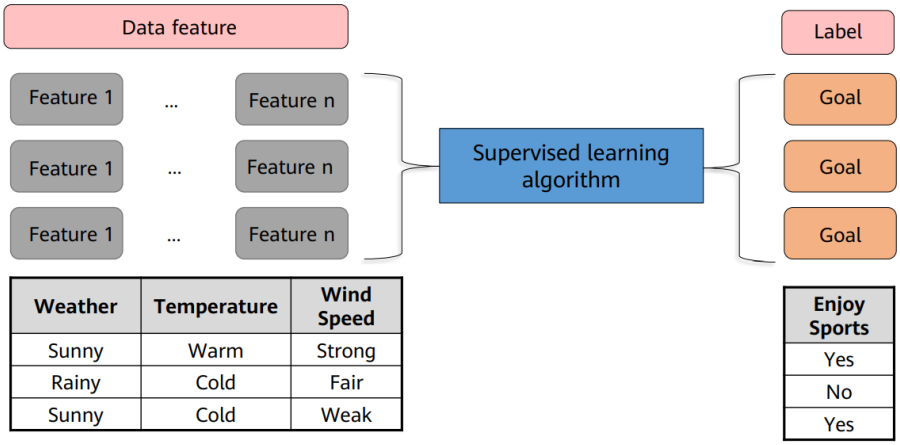


Figure 3.3 – Apprentissage supervisé [22]

- **Apprentissage non supervisé :** Pour les échantillons non étiquetés, les algorithmes d'apprentissage modélisent directement les ensembles de données d'entrée. Le regroupement est une forme courante d'apprentissage non supervisé. Il suffit de réunir des échantillons très semblables, de calculer la similarité entre les nouveaux échantillons et les échantillons existants, et de les classer par similarité.[22]

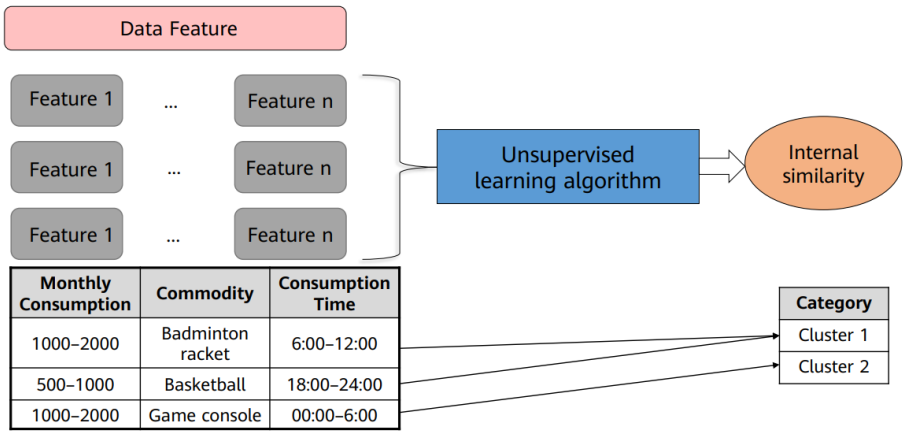


Figure 3.4 – Apprentissage non supervisé [22]

- **Apprentissage semi-supervisé :** Dans une tâche, un modèle d'apprentissage automatique qui utilise automatiquement une grande quantité de données non étiquetées pour faciliter l'apprentissage direct d'une petite quantité de données étiquetées [22]
- **Renforcement de l'apprentissage :** C'est un domaine de l'apprentissage automatique concerné par la façon dont les agents devraient prendre des mesures dans un environnement pour maximiser une certaine notion de récompense cumulative. La différence entre l'apprentissage par renforcement et l'apprentissage supervisé est le signal de l'enseignant. Le signal de renforcement fourni par l'environnement dans l'apprentissage du renforcement est utilisé pour évaluer l'action (signal scalaire) plu-

tôt que de dire au système d'apprentissage comment effectuer les actions correctes [22]

3.2 Data Science

3.2.1 Qu'est-ce que data science ?

Data Science est un domaine interdisciplinaire qui se concentre sur l'extraction des connaissances à partir d'ensembles de données qui sont généralement énormes. Le domaine englobe l'analyse, la préparation des données aux fins d'analyse et la présentation des constatations pour éclairer les décisions de haut niveau d'une organisation. À ce titre, il intègre des compétences en informatique, en mathématiques, en statistique, en visualisation de l'information, en graphisme et en affaires[23]

3.2.2 Quelle est la différence entre data science, l'intelligence artificielle et le machine learning ?

Pour mieux comprendre la data science, et comment vous pouvez l'exploiter, il est tout aussi important de connaître d'autres notions liées à ce domaine, telles que l'intelligence artificielle (IA) et le machine learning[24].

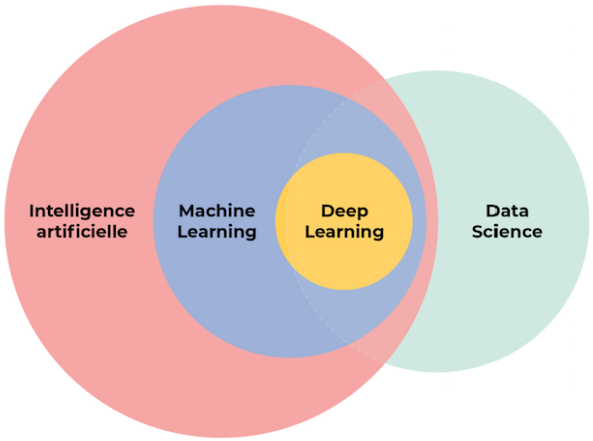


Figure 3.5 – La différence entre data science, l'intelligence artificielle et le machine learning

Voici leurs définitions :

- **L'IA, intelligence artificielle**, permet à un ordinateur d'imiter le comportement humain d'une manière ou d'une autre.
- **La data science** est un sous-ensemble de l'IA, qui désigne les domaines interconnectés des statistiques, des méthodes scientifiques et de l'analyse des données. Tous ces éléments sont utilisés pour extraire du sens et des perspectives des données.
- **Le machine learning**, autre sous-ensemble de l'IA, comprend des techniques qui permettent aux ordinateurs de comprendre les choses à partir des données et de fournir des applications d'IA. Et pour faire bonne mesure, voici une autre définition.

- **Le deep learning**, sous-ensemble du machine learning, permet aux ordinateurs de résoudre des problèmes plus complexes.

3.2.3 Théorie et méthodes analytiques avancées pour la classification supervisé

3.2.3.1 Arbres de décision(Decision Trees)

Decision Tree (également appelé arbre de décision) utilise une arborescence pour spécifier des séquences de décisions et conséquences. Commentaires reçus x = $\{x_1, x_2, ...x_n\}$, le but est de prédire une réponse ou une variable de sortie Y. Chaque membre de l'ensemble x = $\{x_1, x_2, ...x_n\}$ est appelé une variable d'entrée. La prédiction peut être réalisée en construisant un arbre de décision avec des points de test et des branches. À chaque point d'essai, il est décidé de choisir une branche spécifique et de descendre l'arbre. Finalement, un point final est atteint, et une prédiction peut être faite. Chaque point de test dans un arbre de décision implique de tester une variable (ou un attribut) d'entrée particulière, et chaque branche représente la décision prise. En raison de sa flexibilité et de sa visualisation facile, les arbres de décision sont couramment déployés dans des applications d'exploration de données à des fins de classification.[25]

Aerçu d'un arbre de décision

La figure 3.6 montre un exemple d'utilisation d'un arbre de décision pour prédire si les clients achèteront un produit. Le terme branche désigne le résultat d'une décision et est visualisé comme une ligne reliant deux nœuds. Si un la décision est numérique, la branche "supérieure à" est généralement placée à droite, et la branche "inférieure à" est placée à gauche. Selon la nature de la variable, l'une des branches peut devoir inclure une composante "égale à".

Les noeuds internes sont les points de décision ou de test. Chaque noeud interne se réfère à une variable d'entrée ou un attribut. Le noeud interne supérieur est appelé racine. L'arbre de décision dans la figure 7-1 est un arbre binaire en ce sens que chaque noeud interne n'a pas plus de deux branches. La ramification d'un nœud est appelée un split.[25]

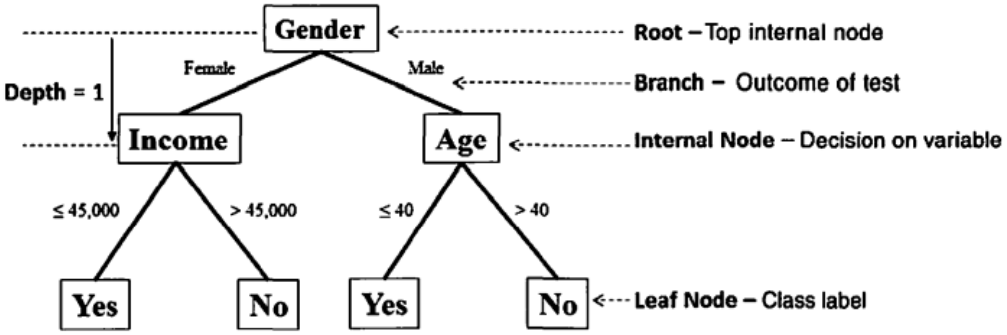


Figure 3.6 – Exemple d'arbre de décision[25]

Parfois, les arbres de décision peuvent avoir plus de deux branches provenant d'un nœud. Par exemple, si une variable d'entrée Météo est catégorique et a trois choix-Ensoleillé, Pluvieux, et Neigeux-le noeud correspondant Météo dans l'arbre de décision peut avoir trois branches étiquetées comme Ensoleillé, Pluvieux, et Neigeux, respectivement.

- La profondeur(**depth**) d'un nœud est le nombre minimum d'étapes nécessaires pour atteindre le nœud à partir de la racine. Dans la figure 3.6, par exemple, les nœuds Revenu et Âge ont une profondeur de un, et les quatre nœuds au bas de l'arbre ont une profondeur de deux. [25]
- Les nœuds foliaires(**Leaf nodes**) sont à la fin des dernières branches de l'arbre. Ils représentent les étiquettes de classe, le résultat de toutes les décisions antérieures. Le chemin de la racine à un nœud de feuille contient une série de décisions prises à différents nœuds internes.[25]

Algorithmes de l'arbre de décision

Il existe de multiples algorithmes pour implémenter les arbres de décision, et les méthodes de construction des arbres varient selon les algorithmes. Certains algorithmes populaires comprennent ID3, C4.5. et CART.

ID3 Algorithme

ID3 (ou Iterative Oichotomiser 3) est l'un des premiers algorithmes d'arbre de décision, et il a été développé par John Ross Quinlan. Que A soit un ensemble de variables d'entrée catégoriques, P soit la variable de sortie (ou la classe prédite), et T soit l'ensemble de formation.[25]

C4.5

L'algorithme C4.5 algorithme introduit un certain nombre d'améliorations par rapport à l'algorithme original ID3. L'algorithme C4.5 peut gérer les données manquantes. Si les enregistrements de formation contiennent des valeurs d'attribut inconnues, le C4.5 évalue le gain pour un attribut en ne considérant que les enregistrements où l'attribut est défini.

Les attributs cal et continu sont supportés par C4.5. Les valeurs d'une variable continue sont triées et partitionnées. Pour les enregistrements correspondants de chaque partition, le gain est calculé, et la partition qui maximise le gain est choisie pour le split suivant.[25]

CART

CART (ou Classification et arbres de régression) est souvent utilisé comme acronyme générique pour l'arbre de décision, bien qu'il s'agisse d'une implémentation spécifique. Semblable à C4.5, CART peut gérer les attributs continus. Alors que C4.5 utilise des critères fondés sur l'entropie pour classer les tests, CART utilise l'indice de diversité de Gini défini dans l'équation 3.1.

$$Gini_x = 1 - \sum_{\forall x \in X} P(x)^2 \tag{3.1}$$

Alors que C4.5 utilise des règles d'arrêt, CART construit une séquence de soustractions, utilise la validation croisée pour estimer le coût de la mauvaise classification de chaque soustraction, et choisir celui dont le coût est le plus bas.[25]

3.2.3.1.1 Les avantages

- Facile à comprendre
- Règles faciles à générer
- Il y a des hyper-paramètres presque nuls à régler
- Les modèles complexes d'arbre de décision peuvent être considérablement simplifiés par leurs visualisations

3.2.3.1.2 Les inconvénients

- Risque de débordement.
- Ne fonctionne pas facilement avec les données non numériques.
- Faible précision de prédiction pour un ensemble de données en comparaison avec d'autres algorithmes.
- Quand il y a beaucoup d'étiquettes de classe,les calculs peuvent être complexes.

3.2.3.2 Naive Bayes

Naive Bayes est un algorithme de classification intéressé à sélectionner la meilleure hypothèse h données d en supposant qu'il n'y a pas interaction entre les caractéristiques.

Un classificateur naive bayes suppose que la présence ou l'absence d'une caractéristique particulière d'une classe n'est pas liée à la présence ou à l'absence d'autres caractéristiques. Par exemple, un objet peut être classé en fonction de ses attributs tels que la forme, la couleur et le poids. Une classification raisonnable pour un objet sphérique, jaune, et moins de 60 grammes de poids peut être une balle de tennis. Même si ces caractéristiques dépendent les unes des autres ou de l'existence des autres caractéristiques, un classificateur na'ive Bayes considère toutes ces propriétés pour contribuer indépendamment à la probabilité que l'objet est une balle de tennis.[25]

théorème de bayes

La représentation est basée sur le théorème de Bayes :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.2}$$

Préparation des données

- Remplacer les entrées numériques par des entrées catégoriques (regroupement) ou quasi gaussiennes (supprimer les valeurs aberrantes, transformation log & boxcox)
- D'autres distributions peuvent être utilisées à la place de Gaussien
- La transformation logarithmique des probabilités peut éviter le débordement
- Les probabilités peuvent être mises à jour à mesure que les données deviennent disponibles

3.2.3.2.1 Les avantages

- Simple, facile et rapide
- Non sensible aux caractéristiques non pertinentes
- Fonctionne très bien dans la pratique
- Besoin de moins de données sur la formation
- Pour la classification multi-classes et binaire
- Fonctionne avec des données continues et discrètes

3.2.3.2.2 Les inconvénients

- Accepte que chaque fonction soit indépendante. Ce n'est pas toujours la vérité.

3.2.3.3 K plus proches voisins (KPPV)

L'algorithme des K plus proches voisins(KPPV) ou **K-nearest neighbors (kNN)** un algorithme de Machine Learning qui appartient à la classe des algorithmes d'apprentissage supervisé simple et facile à mettre en œuvre qui peut être utilisé pour résoudre les problèmes de classification et de régression. Dans cet article, nous allons revenir sur la définition de cet algorithme, son fonctionnement ainsi qu'une application directe en programmation. [26]

L'intuition derrière l'algorithme des K plus proches voisins est l'une des plus simples de tous les algorithmes de Machine Learning supervisé :

- **Étape 1 :** Sélectionnez le nombre K de voisins

- **Étape 2 :** Calculez la distance Manhattan 3.3 et Manhattan 3.4

$$\sum_{i=1}^{n} |x_i - y_i| \tag{3.3}$$

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.4}$$

- **Étape 3 :**Prenez les K voisins les plus proches selon la distance calculée.
- **Étape 4 :** Parmi ces K voisins, comptez le nombre de points appartenant à chaque catégorie
- **Étape 5 :** Parmi ces K voisins, comptez le nombre de points appartenant à chaque catégorie
- **Étape 6 :**Notre modèle est prêt :

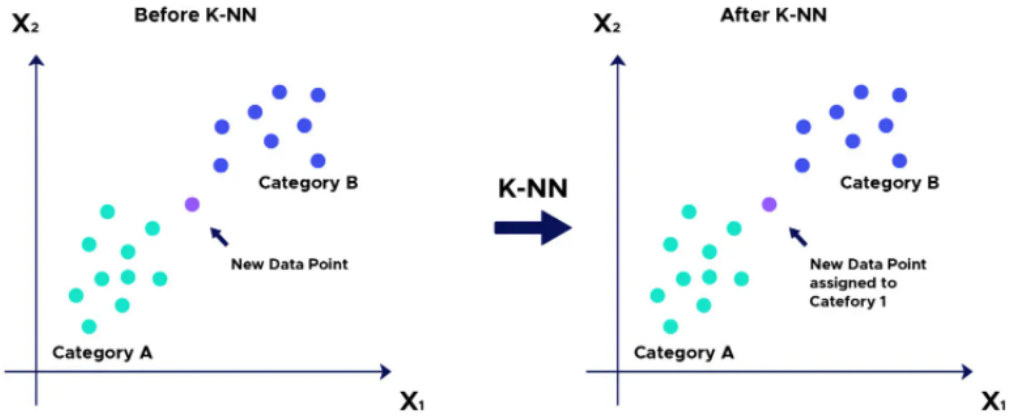


Figure 3.7 – Modèle KPPV[26]

3.2.3.3.1 Avantages et inconvénients de (KPPV)

1. Elle est lourd puisqu'elle nécessite beaucoup de temps de calcul et d'espace mémoire surtout si la base d'apprentissage est importante.
2. Problème dans le choix de la distance et la valeur de K.
3. Problème si les valeurs des attributs ne sont pas uniformes.
4. Elle donne des résultats efficaces.

3.2.4 Théorie et méthodes analytiques avancées pour la classification non supervisé

3.2.4.1 K-means

C'est l'un des algorithmes de clustering les plus répandus. Il permet d'analyser un jeu de données caractérisées par un ensemble de descripteurs, afin de regrouper les données

"similaires" en groupes (ou clusters).[27]

La similarité entre deux données peut être inférée grâce à la "distance" séparant leurs descripteurs ; ainsi deux données très similaires sont deux données dont les descripteurs sont très proches. Cette définition permet de formuler le problème de partitionnement des données comme la recherche de K "données prototypes", autour desquelles peuvent être regroupées les autres données.[27]

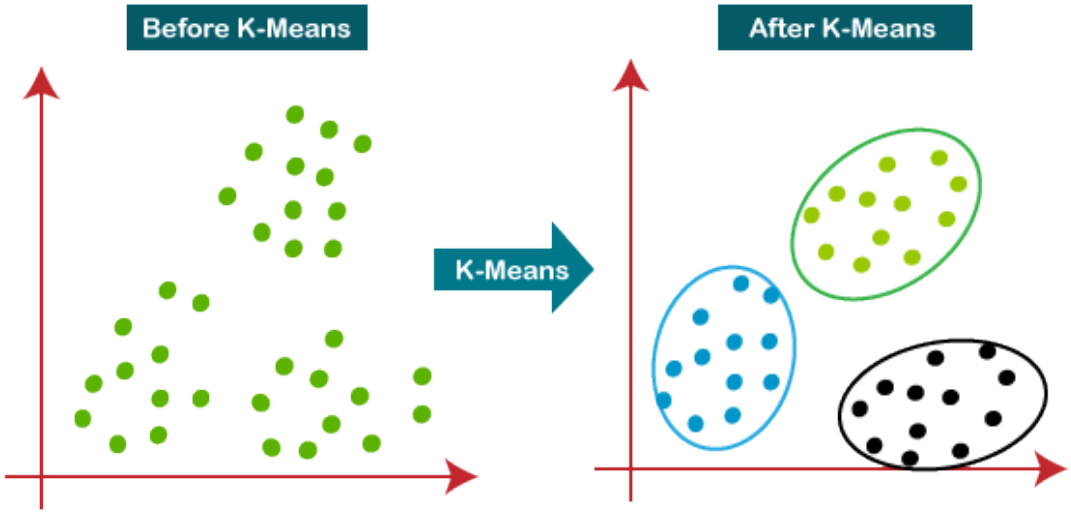


Figure 3.8 – Modèle de K-Means[28]

Comment fonctionne l'algorithme K-Means ?

Le fonctionnement de l'algorithme K-Means est expliqué dans les étapes ci-dessous :[28]

- **Étape 1 :**Sélectionnez le nombre K pour décider du nombre de clusters.
- **Étape 2 :**Sélectionnez des points K aléatoires ou des centroïdes. (Il peut s'agir d'autres points de l'ensemble de données(Datasets) d'entrée).
- **Étape 3 :**Assignez chaque point de données à leur centroïde le plus proche, qui formera les clusters K clusters.
- **Étape 4 :**Calculer la variance et placer un nouveau centroïde de chaque cluster.
- **Étape 5 :**Répétez la troisième étape, ce qui signifie réaffecter chaque point de données au nouveau centroïde le plus proche de chaque cluster.
- **Étape 6 :**Si une réaffectation se produit, passez à l'étape 4, sinon allez à TERMINER.
- **Étape 7 :**Le modèle est prêt.

3.2.4.1.1 Critiques du K moyennes

1. Facile à comprendre et à métre en œuvre
2. Il est applicable aux données de grandes tailles et aussi à tous types de données.
3. Le nombre de classe doit être fixer au départ

4. Le résultat dépend de tirage initial des centres des classes
5. Les clusters sont construits par rapport à des objets inexistants (les milieux).
6. Le regroupement final obtenu dépend de la mesure utilisée pour calculer la distance entre les objets et les barycentres des clusters.

3.2.5 Composantes de Data Science

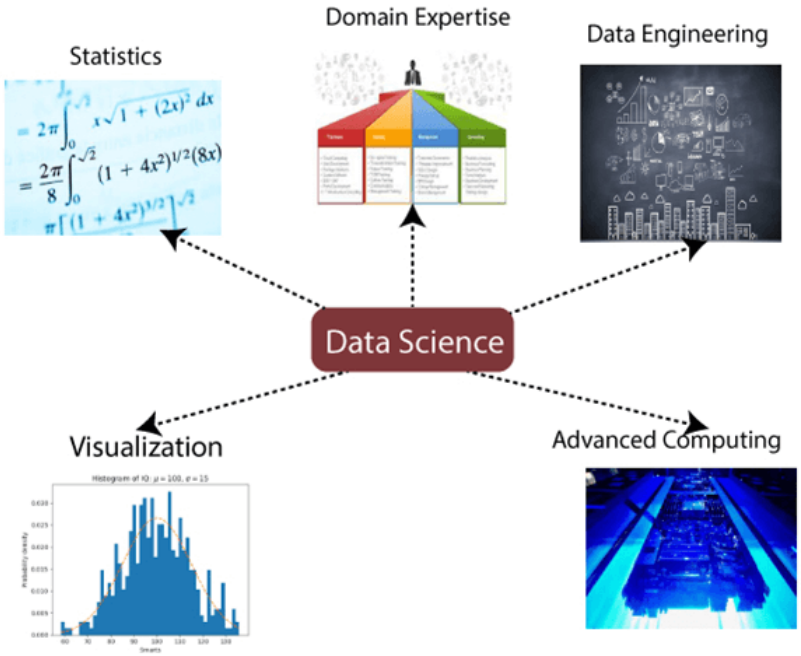


Figure 3.9 – Composantes de Data Science[29]

Les principales composantes de Data Science sont les suivantes :[29]

- **Statistics** ou Statistiques : Les statistiques sont l'une des composantes les plus importantes de la science des données. Les statistiques sont un moyen de recueillir et d'analyser les données numériques en grande quantité et d'en tirer des renseignements utiles.
- **Domain Expertise** ou Expertise du domaine : En science des données, l'expertise du domaine lie la science des données. L'expertise du domaine désigne les connaissances ou les compétences spécialisées d'un domaine particulier. En science des données, nous avons besoin d'experts dans divers domaines.
- **Data engineering** ou Ingénierie des données : L'ingénierie des données fait partie de la science des données, qui comprend l'acquisition, le stockage, la récupération et la transformation des données. L'ingénierie des données comprend également des métadonnées (données sur les données) aux données.
- **Visualization** ou Visualisation : La visualisation de données consiste à représenter des données dans un contexte visuel afin que les gens puissent facilement comprendre l'importance des données. La visualisation des données facilite l'accès à l'énorme quantité de données dans les visuels.

- **Advanced computing** ou Informatique avancée : Le gros du travail de la science des données est l'informatique avancée. L'informatique avancée implique la conception, l'écriture, le débogage et la maintenance du code source des programmes informatiques.

3.3 Data Mining

Dans le monde digital d'aujourd'hui il y'a trop de données, d'information, d'application et de services Mais moins de connaissances. Pour éclaircir le contexte général dans lequel ce cours est inscrit et d'expliciter la problématique au cœur du data mining, il semble opportun de décomposer le sujet de ce cours en deux interdépendances :[30]

3.3.1 Qu'est-ce que le data mining ?

Le data mining est l'exploration et l'analyse de données dans le but de découvrir des modèles ou des règles de nature significative. Il est classé comme une discipline dans le domaine de la data science. Les techniques de data mining servent à créer des modèles de machine learning (ML) qui activent des applications d'intelligence artificielle (IA). Les algorithmes des moteurs de recherche et les systèmes de recommandation sont des exemples de data machine learning dans l'intelligence artificielle.[31]

3.3.2 Comment ça marche Data Minnig ?

Le principe : une démarche (simplifiée et didactique) en 5 temps majeurs. (Voir La figure 3.10)

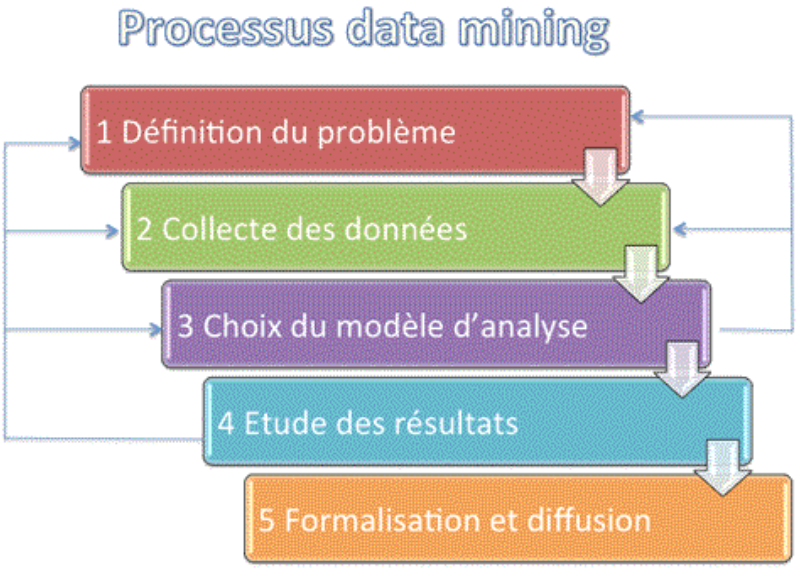


Figure 3.10 – Processus Data Mining [32]

3.3.2.1 Étape 1 : Définition du problème

Quel est le but de l'analyse, que recherche-t-on ? Quels sont les objectifs ? Comment traduire le problème en une question pouvant servir de sujet d'enquête pour cet outil d'analyse bien spécifique ? À ce sujet, se souvenir que l'on travaille à partir des données existantes. La question doit être ciblée selon les données disponibles.[32]

3.3.2.2 Étape 2 : Collecte des données

Une phase absolument essentielle. On n'analyse que des données utilisables, c'est-à-dire "propres" et consolidées. On n'hésitera pas à extraire de l'analyse les données de qualité douteuse. Bien souvent, les données méritent d'être retravaillées. S'assurer au final que la quantité de données soit suffisante pour éviter de fausser les résultats. Cette phase de collecte nécessite le plus grand soin. Voir en seconde partie de l'article un cas concret de projet Data mining où la qualité de la collecte laisse un peu à désirer...[32]

3.3.2.3 Étape 3 : Construire le modèle d'analyse

Ne pas hésiter à valider vos choix d'analyse sur plusieurs jeux d'essais en variant les échantillons. Une première évaluation peut nous conduire à reprendre les points 1 ou 2.[32]

3.3.2.4 Étape 4 : Étude des résultats

Il est temps d'exploiter les résultats. Pour affiner l'analyse, on n'hésitera pas à reprendre les points 1, 2 ou 3 si les résultats s'avéraient insatisfaisants. C'est-à-dire qu'ils ne seraient pas en phase avec les objectifs fixés au temps 1.[32]

3.3.2.5 Étape 5 : Formalisation et diffusion

Les résultats sont formalisés pour être diffuser. Ils ne seront utiles qu'une fois devenus une connaissance partagée. C'est bien là l'aboutissement de la démarche. C'est aussi là que réside la difficulté d'interprétation et de généralisation.[32]

3.3.3 Données (data)

Ces dernières années avec l'arrivée du WWW, le développement des moyens de communications et Les outils automatiques de collection de données ont résulté à la génération d'une masse de données électronique énormes stockées dans des bases de données, des entrepôts de données, des serveurs et d'autres référentiels de stockage comme le montre le tableau suivant :

3.3.4 Fouille (Mining)

En parallèle avec le point précédent, le monde numérique d'aujourd'hui peut être photographié comme un « océan » alimenté par les recherches sur Google, les réseaux sociaux, les blogs et les sites Web commerciaux. Cet océan a conduit à un regain d'intérêt

Google	20 milliards de pages web
Facebook	1 milliards d'utilisateur
Entreprise	Walmart a 20 millions de transactions / jour et une base de données de 10 téraoctets. Blockbuster compte plus de 36 millions de clients résidentiels
Twitter	377 millions de comptes
Bing	1 millions de requéte par jour

Table 3.1 – Statistique montre la quantité des données générer dans le web [30]

et enthousiasme pour la technologie du data mining. L'un des points forts de cette technologie est sa capacité d'épauler les utilisateurs à trouver l'aiguille caché (connaissance) dans une botte de foin de données numériques. [30]

3.3.5 Concepts de bases

3.3.6 La différence entre donnée, information et connaissance

Ces trois concepts sont différents techniquement mais nous allons essayer de les différencier par la suite : [30]

- **Les données :**Appelés faits qui peuvent être des images, textes ou nombres brute du monde réel qui n'ont pas encore étaient interpréter et qui seraient les préceptes qu'un système de machine obtient à-propos du monde. Elle pourrait être structuré ou non structuré, bruyant ou propre, pertinente ou non pertinente, et ainsi de suite. C'est la matière première dans la production d'informations.[33] peuvent être des images, textes..etc, exemple Age = 20.

- **Information :**L'information est une donnée à laquelle un sens et une interprétation ont été donnés après manipulation. Elle est une représentation structurée des données pertinentes. (par exemple, les données peuvent être un flux binaire, alors que l'information est la chaîne ASCII représentée par le flux binaire). Une information permet à un responsable de prendre une décision sur une action à mener.[34] Exemple Reda a 20 ans.

- **Connaissance :** La connaissance acquis par l'expérience ou l'apprentissage peut être vue comme une information comprise, assimilée et utilisée pour aboutir à une action.[34] Donc l'exemple Reda est adulte.

3.3.7 Les instances

Appelées aussi des caractéristiques qui représentent des variables d'entrées et qui peuvent être dans la base d'apprentissage (préétiqueté) ou dans la base de test (étiquette inconnue). Par exemple pour un problème de filtrage de spam, les instances peuvent inclure les éléments suivants :

- Les mots dans le corps de l'e-mail.
- Adresse de l'expéditeur.
- Heure à laquelle l'e-mail a été envoyé.
- E-mail contient l'expression "Une astuce étrange".

3.3.8 Modèle

C'est la connaissance extraite à partir des données pour faire de la prédiction et qu'est créer via un processus d'apprentissage.[33]

3.3.9 Étiquettes (classe)

Une étiquette est le résultat de la prédiction comme la classe spam ou ham pour le problème de filtrage de spam.[35]

3.3.10 Données structurées

Les données structurées résident généralement dans des bases de données relationnelles. Les champs des BDD stockent des données comme les numéros de téléphone, des numéros de sécurité sociale, des codes postaux, les chaînes de texte de longueur variable comme les noms.[35]

3.3.11 Les données non-structurées

Les données non structurées n'ont pas une structure définit via un modèle de données ou un schéma prédéfini. Elles peuvent être textuelles ou non textuelles et généré par l'homme ou la machine. Elles peuvent également être stockées dans une base de données non relationnelle comme NoSQL.[33] Par exemple :

- Fichiers texte : traitement de texte, tableurs, présentations, courriel, journaux
- Email : Email a une structure interne grâce à ses métadonnées, et nous l'appelons parfois semi-structuré. Cependant, son champ de message n'est pas structuré et les outils d'analyse traditionnels ne peuvent pas l'analyser. Médias sociaux : Données de Facebook, Twitter, LinkedIn.
- Site Web : YouTube, Instagram, sites de partage de photos.
- Données mobiles : messages texte, emplacements.
- Communications : chat, messagerie instantanée, enregistrements téléphoniques, logiciels de collaboration.
- Médias : MP3, photos numériques, fichiers audio et vidéo.
- Applications métier : documents MS Office, applications de productivité.
- Imagerie satellitaire : données météorologiques, formes terrestres, mouvements militaires..
- Données scientifiques : exploration pétrolière et gazière, exploration spatiale, imagerie sismique, données atmosphériques.

- Surveillance numérique : photos et vidéos de surveillance.
- Données de capteur : Trafic, météo, capteurs océanographiques.

3.4 La différence entre Data Science, et Data Mining

Data Sciene

Data Science : Science des données est un domaine ou un domaine qui comprend et implique de travailler avec une énorme quantité de données et les utilise pour construire des modèles analytiques prédictifs. Il s'agit de creuser, capturer (construire le modèle) analyser (valider le modèle) et utiliser les données (déployer le meilleur modèle). C'est une intersection de données et de calcul. C'est un mélange du domaine de l'informatique, des affaires et des statistiques ensemble.[36]

Data Mining

Data Mining : L'extraction de données est une technique permettant d'extraire de l'information et des connaissances importantes et vitales d'un vaste ensemble de données. Il tire la perspicacité en extrayant soigneusement, examinant, et traitant les données énormes pour trouver le modèle et les co-relations qui peuvent être importantes pour l'entreprise. Il est analogue à l'exploitation aurifère où l'or est extrait des roches et des sables.[36]

Voici un tableau (3.2) des différences entre la science des données(Data Science) et l'exploration des données(Data Mining) :

Data Science	**Data Minnig**
Data Science est un domaine	Data Mining est une technique
Il s'agit de la collecte, du traitement, de l'analyse et de l'utilisation des données dans diverses opérations. C'est plus conceptuel.	Il s'agit d'extraire les informations vitales et précieuses des données
C'est un domaine d'études tout comme l'informatique, les statistiques appliquées ou les mathématiques appliquées.	C'est une technique qui fait partie de Extraction de connaissances à partir de données (ECD)
L'objectif est de créer des produits dominants en données pour une entreprise.	Le but est de rendre les données plus vitales et utilisables, c.-à-d. en extrayant seulement les renseignements importants.

Table 3.2 – La différence entre Data Science, et Data Mining [36]

3.5 Conclusion

Nous avons présenté dans ce chapitre une introduction général sur le Machine Learning, d'autre part on a définie Data Science pour les algorithme qui nous allons utiliser à notre l'expérimentation et résultats, et aussi on a défini le Data Mining général. Et enfin la différence entre Data Science et Data Mining.

Chapitre 4

Expérimentation et résultats

4.1 Introduction

Dans ce chapitre, nous décrivons les procédures suivies pendant la création du modèle de détection d'intrusion incluant l'ensemble de données(Data set) utilisé, scénario expérimental, résultats et explications. Différents algorithmes d'apprentissage automatique ont été utilisés sur l'ensemble de données d'intrusion (KDDCup99)..La comparaison entre les résultats obtenu nous a permet de sélectionner un meilleur algorithme pour notre modèle.

4.2 DATA SETS

Data Sets (Jeu de données) existant permettant de détecter les intrusions. (Voir Table 4.1) fournis une vue d'ensemble complète des data sets(Jeu de données) et des classes d'attaque qui s'y trouvent.

- **KDDCup99 :**Est l'un des data sets les plus populaires et les plus utilisés de la communauté IDS. Il a plus de cinq millions d'enregistrements pour la formation et deux millions de disques pour les tests. Chaque enregistrement est classé comme normal ou voie de fait et contient 41 traits ou attributs différents. Déni de service (DoS), Probing, Remote to Local (R2L) et User to Root sont quatre principaux types d'agressions (U2R).[37]
- **Kyoto 2006 :**Ce data set a été développé en utilisant des enregistrements de trafic réseau recueillis par l'Université de Kyoto grâce à l'utilisation de pots de miel, de capteurs de darknet, de serveurs de messagerie, de robots Web et d'autres mesures de sécurité réseau. 134 De 2006 à 2015, l'ensemble de données le plus récent comprend des données sur le trafic. Chaque enregistrement possède 24 attributs statistiques, dont 14 sont tirés de la KDDCup99 data set tandis que les dix autres sont facultatifs.[38]
- **NSL-KDD :**Il s'agit d'une version révisée et améliorée du KDDCup99 data set qui répond à de nombreuses préoccupations clés. Comme décrit dans KDDCup99, ce data set possède 41 fonctionnalités et les attaques sont classées en quatre types.[39]
- **UNSW-NB15 :**Cet ensemble de données est créé par l'Australian Center for Cyber Security.[40] Il contient environ deux millions d'enregistrements avec un total de 49 fonctionnalités, qui sont extraites à l'aide de Bro-IDS, des outils Argus, et quelques rithms algo nouvellement développés. Ce jeu de données contient les types d'attaques nommés Worms, Shellcode, Reconnaissance, Ports Scans, Generic, Backdoor, DoS, Exploits et Fuzzers.[41]
- **CIC-IDS2017 :**Ce data set est créé par l'Institut canadien de la cybersécurité (CIC) en 2017. Il contient les flux normaux et les attaques mises à jour dans le monde réel. Le trafic réseau est analysé par CICFlowMeter en utilisant les informations basées sur les horodatages, la source et les adresses IP de destination, les protocoles et les attaques. 136 De plus, CICIDS2017 comprend des scénarios d'attaque courants comme Brute Force Attack, HeartBleed Attack, Botnet, Denial of Service (DoS) Attack, Distributed DoS (DDoS) Attack, Web Attack et Infiltration Attack.[42]

- **CSE-CIC-IDS2018 :**Ce data set a été créé conjointement par le Centre de la sécurité des télécommunications (CST) et CIC en 2018. Les profils utilisateurs contenant la représentation abstraite des différents événements sont créés.[43]

Data sets	Année	Types d'attaques	Attaques
KDDCup99[37]	1998	4	DoS, Probing, R2L,U2R
Kyoto 2006[38]	2006	2	Known Attacks, Unknown Attacks
NSL-KDD[39]	2009	4	DoS, Probing, R2L, U2R
UNSW-NB15[40]	2015	9	Backdoors, DoS, Exploits, Fuzzers, Generic, Port Scans, Reconnaissance, Shellcode, Worms
CIC-IDS2017[42]	2015	7	Brute Force, HeartBleed, Botnet, DoS, DDoS, Web, Infiltration
CSE-CIC-IDS201[43]	2015	7	HeartBleed, DoS, Botnet, DDoS, Brute Force, Infiltration

Table 4.1 – Une vue d'ensemble complète des data sets.

4.3 Description de Dataset KDDCup99

Donc nous allons choisir le data set kddCup99, Il s'agit du data set utilisé pour le troisième concours international d'outils de découverte de connaissances et d'exploration de données, qui a eu lieu en même temps que KDDCup99 The Fifth International Conference on Knowledge Discovery and Data Mining. La tâche de la concurrence consistait à construire un détecteur d'intrusion de réseau, un modèle prédicatif capable de distinguer les «mauvaises» connexions, appelées intrusions ou attaques, des «bonnes» connexions normales. Cette base de données contient un ensemble standard de données à vérifier, qui comprend une grande variété d'intrusions simulées dans un environnement de réseau militaire.[44]

4.3.1 Le Contenue de KDDCup99

Les logiciels de détection des intrusions dans le réseau protègent un réseau informatique contre les utilisateurs non autorisés, y compris peut-être les initiés. La tâche d'apprentissage du détecteur d'intrusion consiste à créer un modèle prédicatif (c.-à-d. un classificateur) capable de distinguer les «mauvaises» connexions, appelées intrusions ou attaques, des «bonnes» connexions normales.[45]

Une connexion est une séquence de paquets TCP commençant et se terminant à des moments bien définis, entre lesquels les données circulent vers et depuis une adresse IP source vers une adresse IP cible sous un protocole bien défini. Chaque connexion est

étiquetée comme normale, ou comme une attaque, avec exactement un type d'attaque spécifique. Chaque enregistrement de connexion se compose d'environ 100 octets.[45]

Le programme d'évaluation de la détection d'intrusion de la DARPA de 1998 a été préparé et géré par le MIT Lincoln Labs. L'objectif était d'étudier et d'évaluer la recherche en détection d'intrusion. Un ensemble standard de données à vérifier, qui comprend une grande variété d'intrusions simulées dans un environnement de réseau militaire, a été fournis. Le concours de détection d'intrusion KDD de 1999 utilise une version de ce data set.[46]

Les attaques se divisent en quatre grandes catégories

- **DoS :** déni de service, p.ex. syn flood.
- **R2L :** accès non autorisé depuis une machine distante, p.ex. deviner le mot de passe.
- **U2R :** accès non autorisé aux privilèges de super utilisateur local (root), par exemple, diverses attaques de «dépassement de tampon».
- **Probing :**surveillance et autres sondes, p. ex., balayage des ports.

Il est important de noter que les données de la base de test incluant des types d'attaques spécifiques qui ne figure pas dans le training set. Certains experts d'intrusion croient que la plupart des nouvelles attaques sont des variantes d'attaques connues. Le training set contient 24 types d'attaque, avec 14 types supplémentaires dans la base de test. Le nom et description détaillée des types d'attaque sont répertoriés dans le travail de chercheur Lippmann.[47]

Nom d'attributs	Description	Types
`duration`	longueur (nombre de secondes) de la connexion	continu
`protocol_type`	type de protocole, par exemple :tcp, udp,etc..	discret
`service`	service réseau sur la destination parexemple :http, telnet, etc..	discret
`src_bytes`	nombre d'octets de données de source à destination	continu
`dst_bytes`	nombre d'octets de données de destination à la source	continu
`flag`	état normal ou erreur de la connexion	discret
`land`	1 si la connexion est depuis/vers le même hôte/port ; 0 sinon	discret
`wrong_fragment`	nombre de " mauvais " fragments	continu
`urgent`	nombre de paquets urgents	continu

Table 4.2 – Caractéristiques de base des connexions TCP individuelles.

Nom d'attributs	Description	Types
hot	nombre d'indicateurs " chauds "	continu
num_failed_logins	nombre de tentatives de connexion qui ont échoué	continu
logged_in	1 si correctement connecté ; 0 sinon	discret
num_compromised	number of "compromised" conditions	continu
root_shell	1 si le shell root est obtenue ; 0 sinon	discret
su_attempted	1 si la commande " su root" tenté ; 0 sinon	discret
num_root	nombre d'accès " route "	continu
num_file_creations	nombre d'opérations de création de fichier	continu
num_shells	nombre d'invites de shells	continu
num_access_files	nombre d'opérations sur les fichiers de contrôle d'accès	continu
num_outbound_cmds	nombre de commandes sortants dans une session ftp	continu
is_hot_login	1 si la connexion appartient à la liste " chaude " ; 0 sinon	discret
is_guest_login	1 si la connexion est une connexion " guest" ; 0 sinon	discret

Table 4.3 – Fonctionnalités de contenu au sein d'une connexion suggérée par la connaissance du domaine.

Nom d'attributs	Description	Types
count	nombre de connexions vers le même hôte que la connexion en cours dans les deux dernières secondes Remarque : Les fonctionnalités suivantes se refer aux connexions des méme hotes	continu
serror_rate	% de connexions qui contiennent des erreurs " SYN"	continu
rerror_rate		continu
same_srv_rate	% de connexions qui contiennent des erreurs " REJ"	continu
diff_srv_rate	% des connexions au même service	continu
srv_count	nombre de connexions pour le même service que la connexion en cours dans les deux dernières secondes Remarque : Les fonctionnalités suivantes se réfère aux connections du même service.	continu
srv_serror_rate	% de connexions qui contiennent des erreurs " SYN	continu
srv_rerror_rate	% de connexions qui contiennent des erreurs " REJ"	continu
srv_diff_host_rate	% de connexions sur différents hôtes	continu

Table 4.4 – Caractéristiques de circulation calculées à l'aide d'une fenêtre de temps de deux secondes.

Donc la KDDCup99 est un corpus benchmark pour la détection d'intrusion , caractérisé de 42 attributs de différents services pour la détection d'intrusion, et l'attribut numéro 42 est la classe de la connexion.

4.4 Présentation des outils utilisés

Le matériel joue un rôle essentiel dans les performances du modèle, Nous avons utilisé un ordinateur bureau et portable sous Windows 10, Système d'exploitation 64 bits avec 8Go de RAM. Le processeur est core i5 avec une vitesse d'horloge de 2,6GHz.

4.4.1 Langage de programmation

Python : Python est un langage de programmation interprété, Développé en 1989. Il est utilisé pour de nombreuses applications différentes. Il est utilisé par des développeurs de logiciels professionnels dans des endroits tels que Google, la NASA..., Ainsi python est le langage le plus utilisé dans le domaine d'apprentissage automatique. Ses principales caractéristiques sont :

- ➢ «open-source» : son utilisation est gratuite et les fichiers sources sont disponibles et modifiables, Simple et très lisible.
- ➢ Doté d'une bibliothèque de base très fournie
- ➢ Importante quantité de bibliothèques disponibles : pour le calcul scientifique, les statistiques, les bases de données, la visualisation.
- ➢ Grande portabilité : indépendant vis à vis du système d'exploitation (linux, Windows, MacOs)

Anaconda Anaconda est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique (traitement de données à grande échelle, analyse prédictive, calcul scientifique), qui vise à simplifier la gestion des paquets et de déploiement. Les versions de paquetages sont gérées par le système de gestion de paquets conda. La distribution Anaconda est utilisée par plus de 6 millions d'utilisateurs et comprend plus de 250 paquets populaires en science des données adaptés pour Windows, Linux et MacOs.[48]

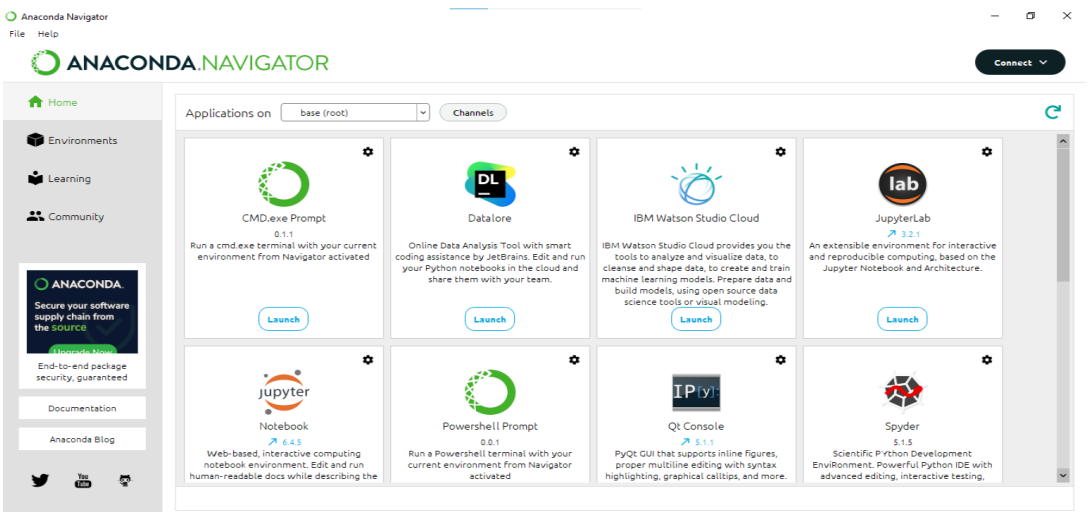


Figure 4.1 – Interface Anaconda

Jupyter Jupyter est une application web utilisée pour programmer, initialement développés pour les langages de programmation Julia , Python et R (d'où le nom Jupyter), et supporte près de 40 langages. Jupyter est une évolution du projet IPython. Jupyter permet de réaliser des calepins ou notebooks qui sont utilisés en science des données pour explorer et analyser des données. La cellule est l'élément de base d'un notebook jupyter. Elle peut contenir du texte formaté au format markdown ou du code informatique qui pourra être exécuté.[49]

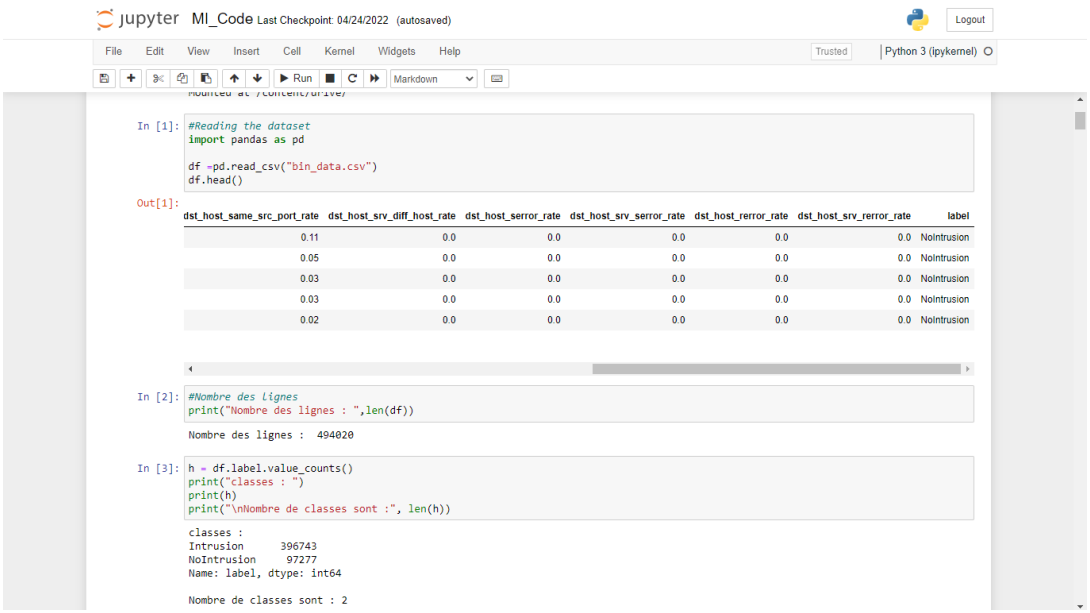


Figure 4.2 – Interface Jupyter

4.4.2 Bibliothèques utilisées

Pour traiter l'ensemble de données et mettre en œuvre l'apprentissage automatique, nous avons utilisé de nombreuses bibliothèques python.

- **Sklearn :** Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle comprend des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python, notamment NumPy et SciPy. La bibliothèque Sklearn est principalement utilisée pour créer la matrice de confusion, pour diviser un ensemble, pour effectuer le prétraitement des données et pour la procédure d'ingénierie des fonctionnalités
- **Matplotlib :** la bibliothèque Matplotlib est utilisée pour visualiser les données sous format graphique. Cette bibliothèque prend en charge le graphique à barres, le nuage de points et de nombreux autres graphiques qui aident à comprendre et analyser clairement les résultats obtenus.
- **Pandas :** la bibliothèque Pandas prend en charge l'analyse des données. Nous utilisons la bibliothèque pandas pour importer l'ensemble de données au format de fichier .CSV et pour manipuler les données.

4.5 Expérimentations et discussions

Nous pouvons commencer à appliquer des techniques d'apprentissage automatique pour la classification dans un ensemble de données qui détecte les intrusions.

Nous décrivons les procédures suivies pendant la création du modèle de détection d'intrusion, scénario expérimental, résultats et explications. Différents algorithmes d'apprentissage automatique ont été utilisés sur l'ensemble de données KddCup99. La comparaison entre les résultats obtenus nous a permet de sélectionner un meilleur algorithme pour notre modèle.

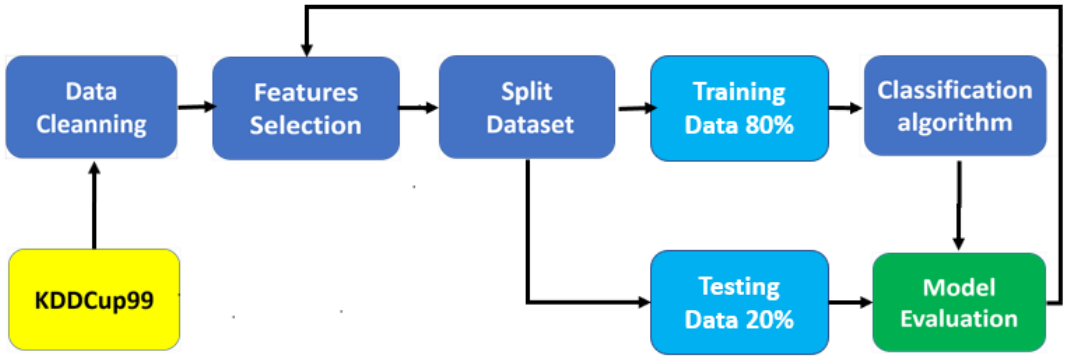


Figure 4.3 – Notre approche proposé

4.5.1 Présentation de KDDCup99

La base de données utilisée est un fichier CSV :

Nous avons 494020 instances et 41 attributs (Voir les tableaux ci-dessus : Table 4.2, Table 4.3, Table 4.4)

4.5.2 Data Cleaning

Nettoyage des données (Data cleaning), dans ce processus, nous allons faire la classe binaire, sa signification faire deux classes INTRUSION et NOINTRUSION, Dans KddCup99 nous avons **23** classes, qui sont (smurf, neptune, **normal**, back, satan, ipsweep, portsweep, warezclient, teardrop, pod, nmap, guess_passwd, buffer_overflow, land, warezmaster, imap, rootkit, loadmodule, ftp_write, multihop, phf, perl, spy) donc la classe normale est égale à NoIntruision, et d'autres classes est INTRUSION.

Pour faire ça en python, on va mettre le code comme ça :

```
1 bin_data = pd.DataFrame(df.label.map(lambda x:'NoIntrusion' if
  ↪  x=='normal' else 'Intrusion'))
2
3 # df = l'ancien data set avec 23 classes
4
```

```
h = bin_data.label.value_counts()
print("classes : ")
print(h)
print("\nNombre de classes sont :", len(h))
# Result ↓
# classes :
# Intrusion       396743
# NoIntrusion      97277
#Nombre de classes son : 2
```

Donc on voit qu'il y a des classes Intrusion c'est plus que NoIntrsuion classe, en pourcentage voir la figure (4.4)

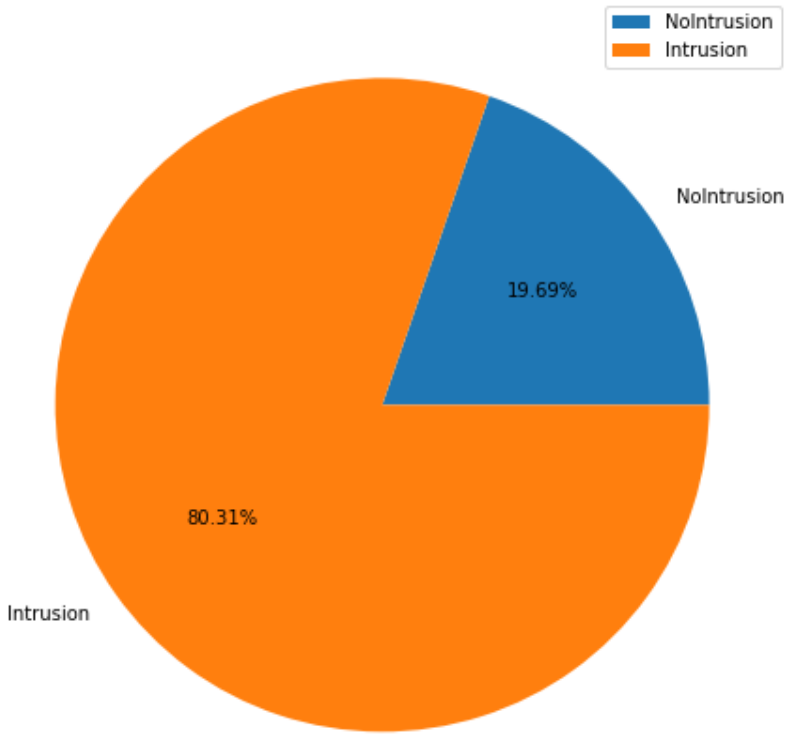


Figure 4.4 – Pourcentage de classes

Après cela, nous passerons au nettoyage des données, donc nous vérifierons s'il manque des valeurs dans les échantillons.

Nous l'avons vérifié, et nous n'avons aucune valeur manquante.

4.5.3 Features selection (sélection des attributs)

Dans ce processus, nous allons utiliser 2 méthodes pour choisir des attributs parfaites, les méthodes que nous allons utiliser, sont **information mutuelle** et **Chi-square(chi2)**.

4.5.3.1 L'information Mutuelle

L'information mutuelle (IM) [50] entre deux variables aléatoires est une valeur non négative, qui mesure la dépendance entre les variables. Elle est égale à zéro si et seulement si deux variables aléatoires sont indépendantes, et des valeurs plus élevées signifient une plus grande dépendance.

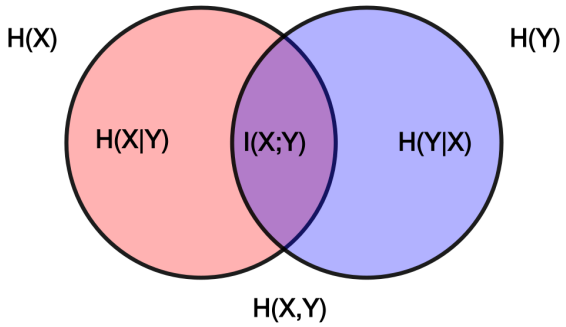


Figure 4.5 – Information mutuelle

$I(X;Y) = H(X)–H(X|Y)$ Où I(X ; Y) est l'information mutuelle pour X et Y, H(X) est l'entropie pour X et H(X | Y) est l'entropie conditionnelle pour X donnée Y. Le résultat a les unités de bits.

Le principe de l'information mutuelle est basé sur le nombre d'occurrence d'un mot dans une certaine catégorie. L'information mutuelle du mot et de la classe est élevée si la fréquence d'apparition du mot dans la classe (catégorie) est élevée et vice versa l'information mutuelle du mot et de la classe est faible si la fréquence du mot apparaît en dehors de la classe (et plus une classe va apparaître sans le mot) puis une moyenne des scores du mot jumelé à chacune des classes est calculée. La faiblesse de cette mesure (information mutuelle) est sa grande influence des fréquences des mots car un mot rare risque d'être avantagé vu sa faible probabilité conditionnelle.[51]

$$IM(T_i, R) = log\frac{P(T_i, R)}{P(T_i)P(R)} \tag{4.1}$$

Pour la réduction du vocabulaire, toutes ces mesures nécessitent un choix d'une valeur de seuil pour décider à partir de quelle valeur supprimer ou garder un terme. Le choix de ce seuil est très important pour la qualité de résultat d'une classification (il est souvent empirique). Des études ont prouvés que pour la sélection d'attribut, les méthodes, de gain d'information et la mesure statistique du x^2 , donnent les meilleurs résultats pour différents classificateurs [104]. Toujours dans ce contexte de réduction de vocabulaire, un avantage intéressant est la réduction du risque de sur apprentissage ("overfitting"). [51]

4.5.3.2 Chi-square (Chi2)

Ou La statistique du X^2, Le principe de cette mesure statistique est son évaluation du manque d'indépendance entre un mot et une classe. Les caractéristiques de cette mesure sont presque identiques à celle de l'information mutuelle puisqu'elle utilise la concurrence

mot/classe sauf qu'elle est soumise à une normalisation et peu pratique pour les mots rares[51]. Le principe de la réduction est comme suit :

Étant donnée la matrice N_{ij} des occurrences du mots i dans le texte j.

- Calculer les fréquences f_{ij} correspondants par $f_{ij} = \frac{N_{ij}}{N}$
- Calculer la contribution de (ij) à la statistique du x^2

$$x_{ij}^2 = \frac{N_{ij} \frac{N_i * N_j}{N}}{\frac{N_i * N_j}{N}} = N * \frac{(f_{ij} - f_i * f_j)^2}{f_i * f_j} \tag{4.2}$$

- Trier le tableau de x^2 par ordre décroissant.
- Enfin, déterminer la liste des k premiers mots pour chaque texte pour normalisation.

4.5.4 Split data set

Dans ce processus, nous allons diviser l'ensemble de données à 80% en base d'apprentissage et à 20% en base de test.

4.5.4.1 Base d'apprentissage VS base de test

Les données d'apprentissage comprennent un ensemble d'exemples contenant des instances d'entrées ayant des sorties désirées. Dans l'apprentissage supervisé pour le traitement d'images, par exemple, un système doit y'avoir des images étiquetées de véhicules dans des catégories telles que les voitures et les camions. Après une quantité suffisante d'observations, le système devrait être capable de distinguer et de catégoriser les nouvelles images non étiquetées dans ces deux catégories. En d'autres termes un algorithme d'apprentissage supervisé doit y'avoir un data set pré étiqueté afin d'étiqueté des nouvelle données (la base de teste)[30].

```
1 #Splitting data
2
3 X_train, X_test, y_train, y_test = train_test_split(Xsp, ysp,
  ↪  test_size=0.2, random_state=44, shuffle =True)
4
5 #test_size = 0.2 ( c'est à dire 20% base de tese)
```

Donc 20% base de test et 80% base d'apprenti sage

4.5.5 Classification algorithme

Dans ce procédure nous allons utiliser 3 algorithmes de classification supervisé (Naive Bayes, Arbres de décision(Decision Trees), K plus proches voisins (KPPV)) Et un seul algorithme de classification non supervisé : K-means

		Prédit	
		Intrusion	NoIntrusion
Actuel	Intrusion	VP	FN
	NoIntrusion	FP	VN

Table 4.5 – Matrice de confusion

4.5.5.1 Les mesures de performances utilisées

L'évaluation des performances du modèle d'apprentissage automatique est effectuée en générant une matrice de confusion pour chaque algorithme de machine Learning à gagner aperçu du type d'erreur commise par l'apprentissage automatique modèle qui nous aide à comprendre les autres métriques telles que précision qui en découlent. Nous avons dérivé l'exactitude, la précision, le rappel et le score F1 pour évaluer les performances du modèle.

4.5.5.1.1 Matrice de confusion : Une matrice de confusion, aussi appelée matrice d'erreur est une matrice N x N utilisée pour évaluer les performances d'un modèle de classification, où N est le nombre de classes cibles. La matrice compare les valeurs cibles réelles avec celles prédites par le modèle d'apprentissage automatique. Cela nous donne une vision globale de la performance de notre modèle de classification et des types d'erreurs qu'il commet. Elle comporte 4 valeurs essentielles :

- **Vrai Positif (VP) :** Le nombre de connexion attribués à une catégorie convenablement(document attribués a leur vrai catégorie).

  Dans notre cas connexion attribués **Intrusion** par le modèle et leurs vraies catégories **Intrusion** dans le corpus.

- **Faux Positif (FP) :** Le nombre de documents attribués à une catégorie inconcevablement,(Documents attribués à des mauvaises catégories).

  Dans notre cas connexion attribués **NoIntrusion** par le modèle et leurs vraies catégories **Intrusion** dans le corpus.

- **Faux Négatif (FN) :** Le nombre de documents inconcevablement non attribués,(Qui auraient dû être attribués à une catégorie mais qui ne l'ont pas été).

  Dans notre cas connexion attribués **Intrusion** par le modèle et leurs vraies catégories **NoIntrusion** dans le corpus.

- Vrai Négatif(VN) : Le nombre de documents non attribués à une catégorie convenablement,(Qui n'ont pas à être attribués à une catégorie, et ne l'ont pas été).

  Dans notre cas connexion attribués **NoIntrusion** par le modèle et leurs vraies catégories **NoIntrusion** dans le corpus.

4.5.5.1.2 Précision et Rappel Certains principes d'évaluation sont utilisés de manière courante dans les différents domaine . Les performances en termes de classification sont généralement mesurées à partir de deux indicateurs traditionnellement utilisés c'est les mesures de rappel et précision. Initialement elles ont été conçues pour les systèmes de

recherche d'information, mais par la suite la communauté de classification de textes les a adoptées.

Formellement, pour chaque classe C_i, on calcule deux probabilités qui peuvent être estimées à partir de la matrice de contingence correspondante, ainsi ces deux mesures peuvent être définies de la manière suivante :[52]

- **Le rappel(Recall) :** étant la proportion de documents correctement classés dans par le système par rapport à tous les documents de la classe C_i

$$Rappel(C_i) = \frac{Nombre\ de\ documents\ bien\ classés\ dans\ C_i}{Nombre\ de\ documents\ de\ la\ classe\ C_i} \tag{4.3}$$

$$R_i = \frac{VP}{VP + FN} \tag{4.4}$$

  Le rappel mesure la capacité d'un système de classification à détecter les documents correctement classés. Cependant, un système de classification qui considérerait tous les documents comme pertinents obtiendrait un rappel de 100%. Un rappel fort ou faible n'est pas suffisant pour évaluer les performances d'un système. Pour cela, on définit la **précision**

- **La précision :** est la proportion de documents correctement classés parmi ceux classés par le système dans C_i [52]

$$Précision(C_i) = \frac{Nombre\ de\ documents\ bien\ classés\ dans\ C_i}{Nombre\ de\ documents\ classé\ dans\ C_i} \tag{4.5}$$

$$P_i = \frac{VP}{VP + FP} \tag{4.6}$$

  La précision mesure la capacité d'un système de classification à ne pas classer un document dans une classe, un document qui ne l'est pas. Comme elle peut aussi être interprétée par la probabilité conditionnelle qu'un document choisi aléatoirement dans la classe soit bien classé par le classifieur.

  Ces deux indicateurs pris l'un indépendamment de l'autre ne permettent d'évaluer qu'une facette du système de classification : la qualité ou la quantité.

4.5.5.1.3 Accuracy Accuracy est le rapport entre le nombre de classes correctement prédites et le nombre total de prédictions. Il est présenté en pourcentage. La précision est analysée lorsque les vrais positifs (VP) et les vrais négatifs (VN) sont cruciaux. [53]

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.7}$$

4.5.5.1.4 TP_rate et FP_rate : Une courbe ROC (receiver operating characteristic) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs.

- **TP_rate :** (True Positive) Taux de vrais positifs : (TVP) est l'équivalent du rappel. Il est donc défini comme suit : 4.8

$$TP\_rate = \frac{VP}{VP + FP} \tag{4.8}$$

- **FP_rate :** (False Positive) Taux de faux positifs : (TFP) est défini comme suit : 4.9

$$FP\_rate = \frac{FP}{VP + FP} \tag{4.9}$$

4.5.5.1.5 AUC - ROC Curve AUC - Courbe ROC - La courbe ROC est une mesure du rendement pour les problèmes de classification à divers paramètres de seuil. ROC est une courbe de probabilité et AUC représente le degré ou la mesure de la séparabilité. Il indique dans quelle mesure le modèle est capable de distinguer les classes. Plus le AUC est élevé, mieux le modèle prédit 0 classe comme 0 et 1 classe comme 1. Par analogie, plus le AUC est élevé, mieux le modèle permet de distinguer les patients atteints de la maladie de ceux qui ne sont atteints d'aucune maladie.

La courbe ROC est tracée avec TPR par rapport au FPR où TPR est sur l'axe des y et FPR sur l'axe des x.[54]

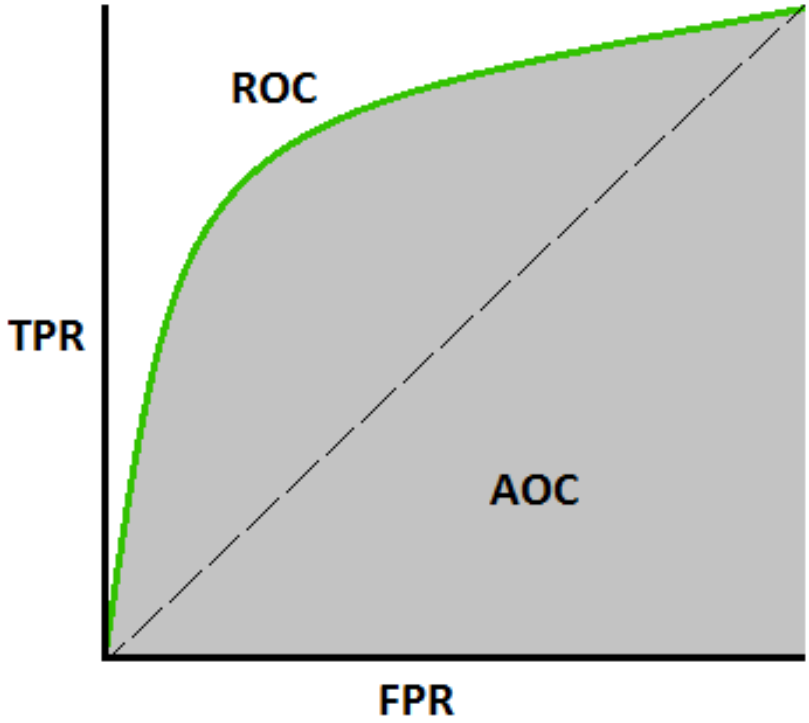


Figure 4.6 – AUC ROC Curve

4.5.5.1.6 F-mesures et entropie Observés conjointement, les indicateurs les plus célèbres à savoir le rappel et la précision, sont une estimation courante de la performance d'un système de classification.

Cependant plusieurs mesures ont été développées afin de synthétiser cette double information. Nous ne retiendrons ici la mesure F_β décrite dans (Van Rijsbergen,1979)[52].

La F-mesure est la mesure de synthèse communément adoptée depuis les années 80 pour évaluer les algorithmes de classification de données textuelles à partir de la précision et du rappel.

Elle est employée indifféremment pour la classification (Non supervisé) ou la catégorisation (Supervisé), pour la problématique de recherche d'information ou de classification. Elle permet donc, de combiner, selon un paramètre !, rappel et précision.[52]

On définit la mesure F_β comme la moyenne harmonique entre le rappel et la précision :

$$F_\beta = \frac{(\beta^2 + 1) * précision * rappel}{\beta^2 * précision * rappel} \tag{4.10}$$

Pour utiliser cette mesure, il est donc nécessaire de fixer préalablement un seuil de décision pour le classement, et de calculer la valeur de F_β pour ce seuil. Le paramètre ! permet de choisir l'importance relative que l'on souhaite donner à chaque quantité. On choisit en général de donner la même importance aux deux critères, donc habituellement, la valeur de β est fixée à 1 et la mesure est ainsi notée : [52]

$$F = \frac{2 * précision * rappel}{précision + rappel} \tag{4.11}$$

Entropie : c'est la perte d'information , elle se calcule par la formule : 4.12

$$E = -log(précision) \tag{4.12}$$

4.5.5.2 L'implémentions

Après avoir split KDDcup99, et les méthodes appliquées chi2 (Chi-square) & IM(l'information mutuelle) (pour la sélection des attributs), nous allons appliquer les méthodes de sélection d'attributs de CHi2(Chi-square) et la méthode de l'information mutuelle.

Nous avons essayé avec toutes les attributs, aavec les attributs sélectionnées, (Multiplier 2 : exemple 2 attributs sélectionnés 4 attributs sélectionnés 6..40) et (Multiplier 3 : 3 6 9 ..39), (Multiplier 5 : 5 10 15...40) Et tous les attributs 41. avec les 2 méthodes de CHi2 (Chi-square) et la méthode de l'information mutuelle.

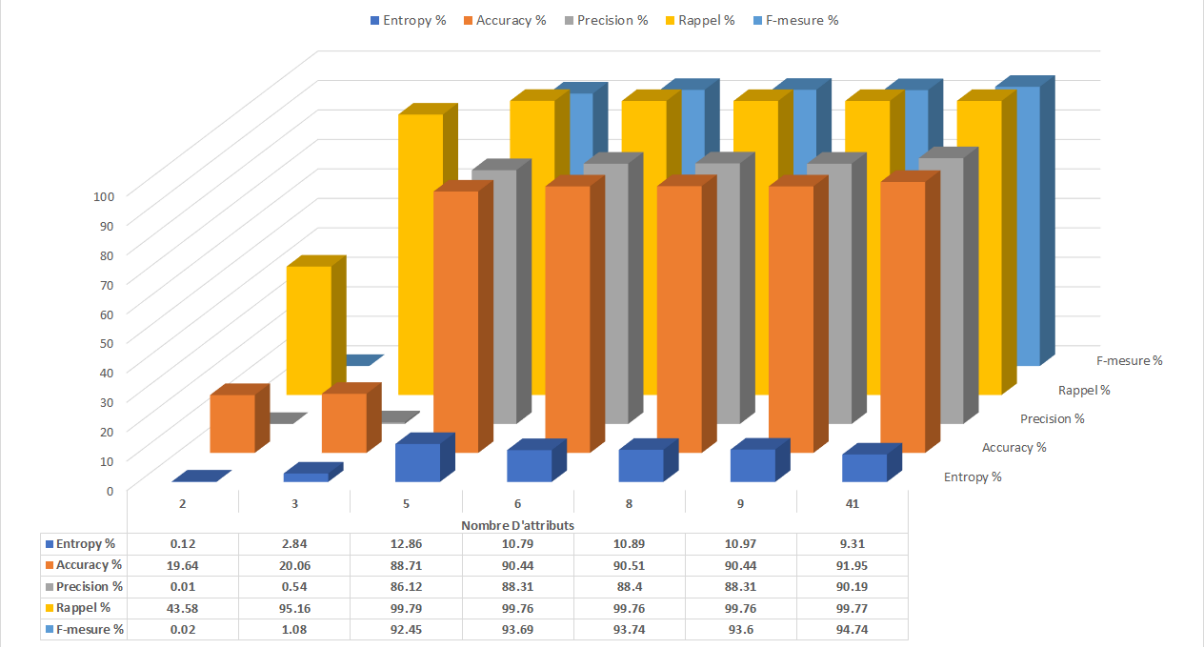
Donc Nous avons appliqué, les différents algorithmes de machine Learning sur le data set« KDDCup99 » et nous avons obtenus les résultats suivants :

Nous avons sélectionné, des attributs comme celui-ci 2 3 5 6 8 9 41, parce que dans ces attributs , entropie il a changé, voir les résultats dans les algorithmes ci-dessous

Dans notre travaille on a prendre la base d'apprentissage de KDDCup99 avec un pourcentage de 80% (395216 instances) , et on a prendre la base de test 20% (98804 instances).

4.5.5.3 Naive Bayes

Avec les performances : Entropie, Accuracy, Précision, Rappel, F-mesure



	2	3	5	6	8	9	41
Entropy %	0.12	2.84	12.86	10.79	10.89	10.97	9.31
Accuracy %	19.64	20.06	88.71	90.44	90.51	90.44	91.95
Precision %	0.01	0.54	86.12	88.31	88.4	88.31	90.19
Rappel %	43.58	95.16	99.79	99.76	99.76	99.76	99.77
F-mesure %	0.02	1.08	92.45	93.69	93.74	93.6	94.74

Figure 4.7 – Visualisation graphique d'attributs sélectionnés par Naive Bayes

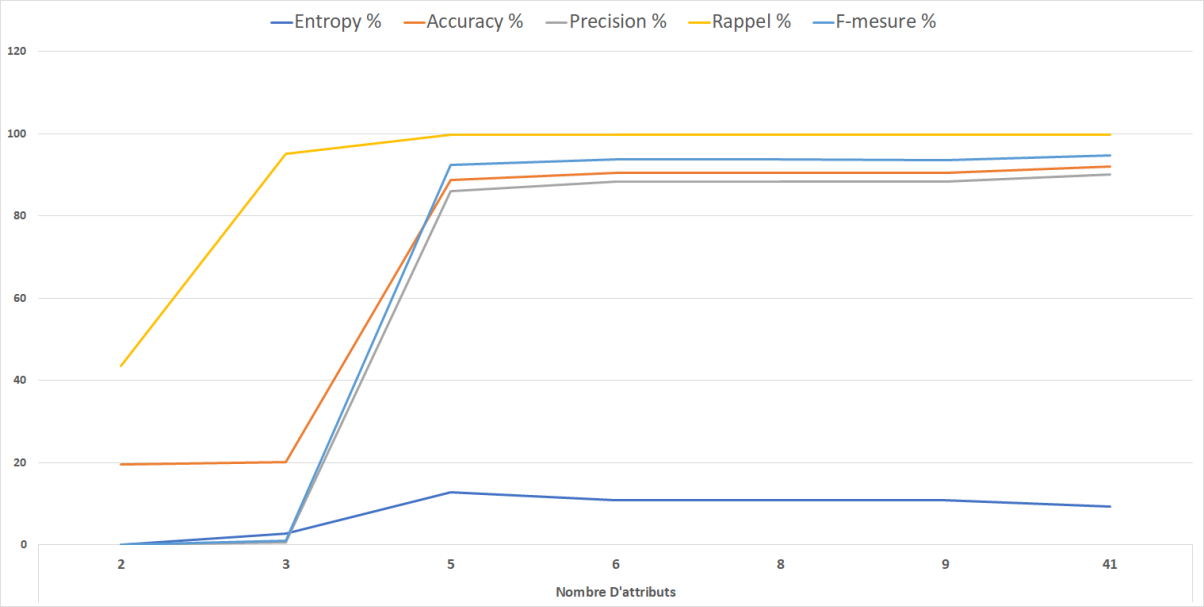


Figure 4.8 – Visualisation par courbe d'attributs sélectionnés par Naive Bayes

4.5.5.3.1 La discussion des résultats : a partir des figures (Figure 4.7, Figure 4.8) on conclue que :

- Pour le **Accuracy**, on voie que accuracy s'augmente progressivement de 2 colonnes à 3 attributs, et à partir de 5 attributs elle commence a se stabiliser avec un pourcentage de 90% un peut prêt.
- Pour La **précision**, on voie qu'il commence a diminué progressivement de 2 à 3 colonnes , et à partir de 5 attributs elle commence a se stabiliser avec un pourcentage de 89% un peut prêt.
- Pour Le **Rappel**, à partir de 3 attributs elle commence a se stabiliser avec un pourcentage de 98% un peut prêt.
- Pour Le **F-mesure**, on voie que accuracy s'augmente progressivement de 2 colonnes à 3 attributs, et à partir de 5 attributs elle commence a se stabiliser avec un pourcentage de 93% un peut prêt.
- Pour **Entropie**, avec l'addition a chaque fois le nombre d'attributs de la kDDCup99 on voie qu'elle commence a diminué progressivement de 2 à 3 colonnes , et à partir de 5 colonnes elle commence a se stabiliser avec un pourcentage de 10% un peut prêt , donc la perte d'informations se diminue progressivement.

Avec performances de TP_rate et FP_rate :

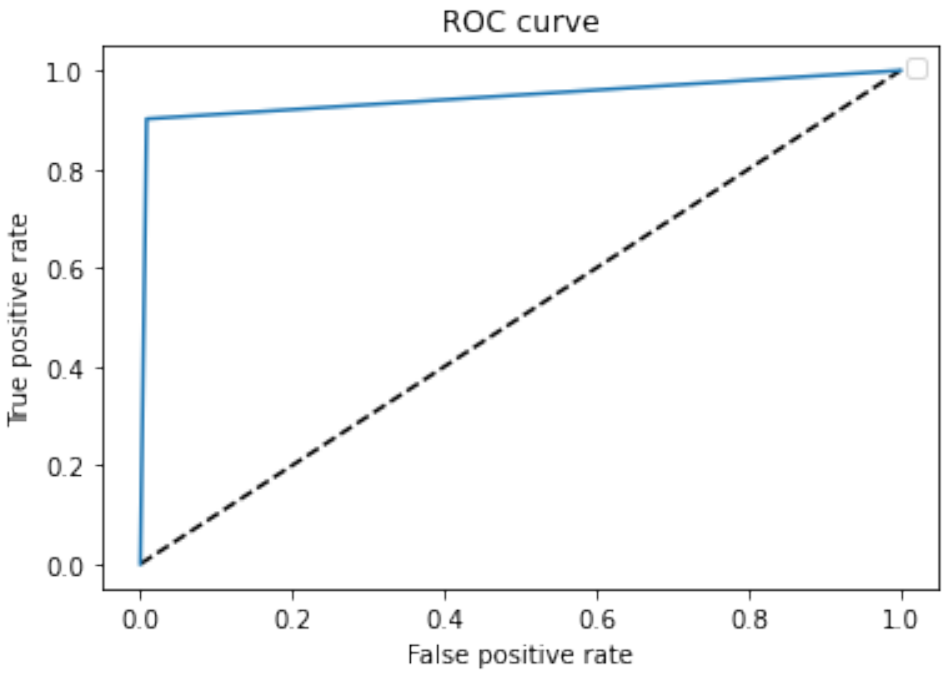


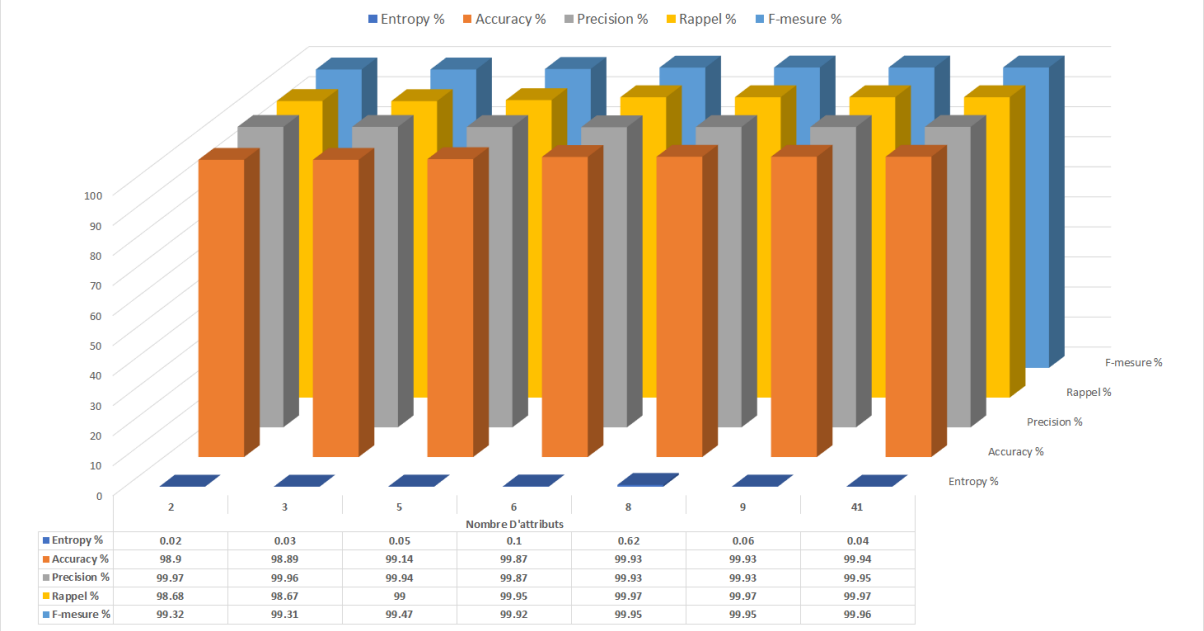
Figure 4.9 – TP_rate et FP_rate de Naïve Bayes

- TP_rate = 90.16%
- FP_rate = 0.82%

Nous voyons que cela nous a donné le meilleur résultats avec `TP_rate` et `FP_rate`, donc c'est la signification qui peut être détecter Intrusion et Nointrusion.

4.5.5.4 KPPV

Avec les performances : Entropie, Accuracy, Précision, Rappel, F-mesure



	2	3	5	6	8	9	41
Entropy %	0.02	0.03	0.05	0.1	0.62	0.06	0.04
Accuracy %	98.9	98.89	99.14	99.87	99.93	99.93	99.94
Precision %	99.97	99.96	99.94	99.87	99.93	99.93	99.95
Rappel %	98.68	98.67	99	99.95	99.97	99.97	99.97
F-mesure %	99.32	99.31	99.47	99.92	99.95	99.95	99.96

Figure 4.10 – Visualisation graphique d'attributs sélectionnés par KPPV

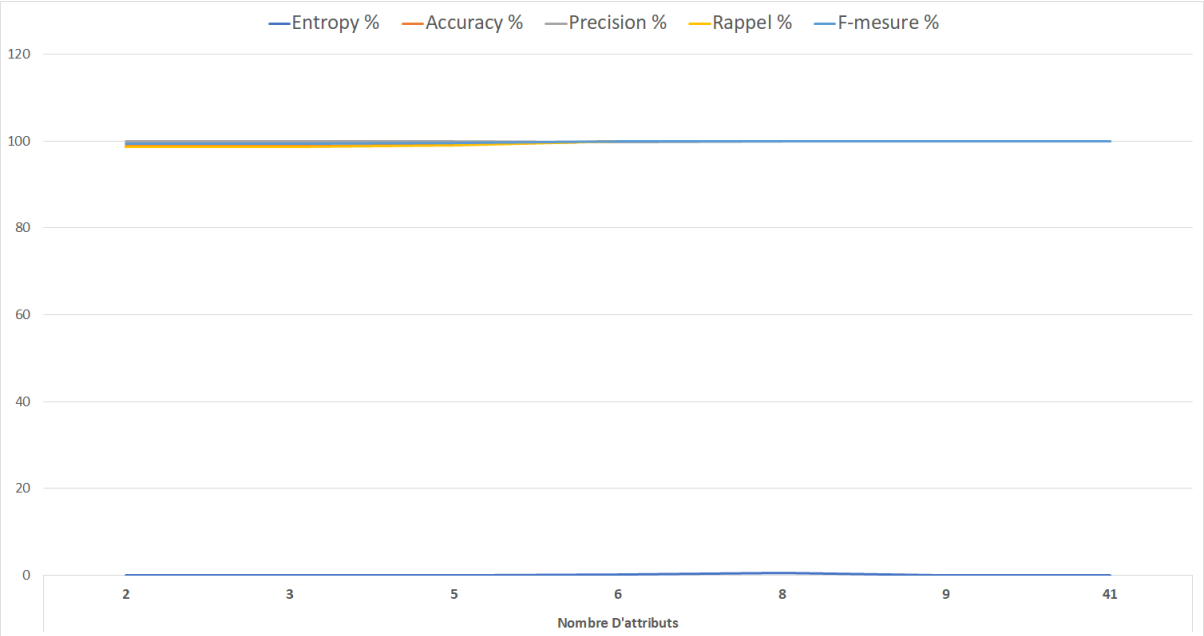


Figure 4.11 – Visualisation par courbe d'attributs sélectionnés par KPPV

4.5.5.4.1 La discussion des résultats : a partir des figures (Figure 4.10, Figure 4.11) on conclue que :

- Pour le **Accuracy**, elle est stable de 2 à 41 attributs avec un pourcentage de 98% un peut prêt.
- Pour La **précision**, elle est stable de 2 à 41 attributs avec un pourcentage de 99% un peut prêt.
- Pour Le **Rappel**, elle est stable de 2 à 41 attributs avec un pourcentage de 98% un peut prêt.
- Pour Le **F-mesure**, elle est stable de 2 à 41 attributs avec un pourcentage de 99% un peut prêt.
- Pour **Entropie**, Augmentation progressive de 2 à 6.avec un pourcentage de 0.08%, Et diminution progressive de 8 à 41. avec un pourcentage de 0.04% un peut prêt.

Avec performances de TP_rate et FP_rate :

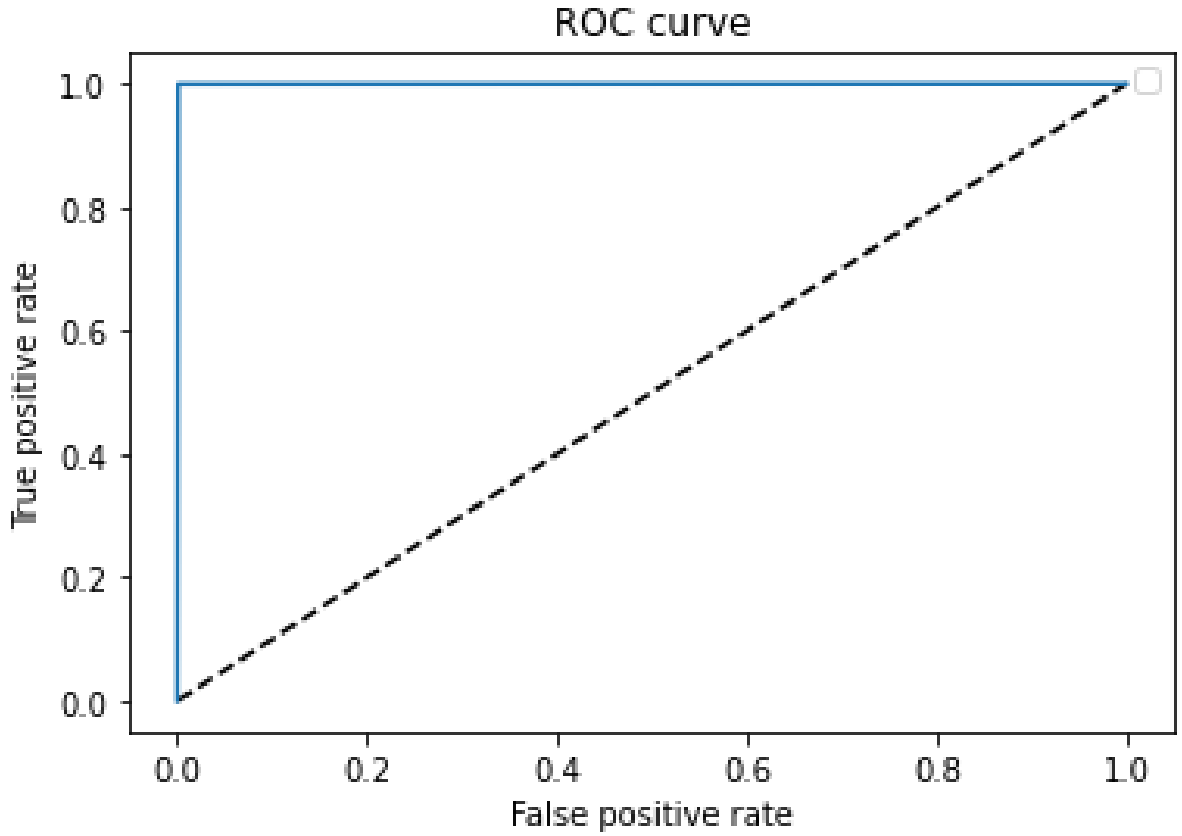


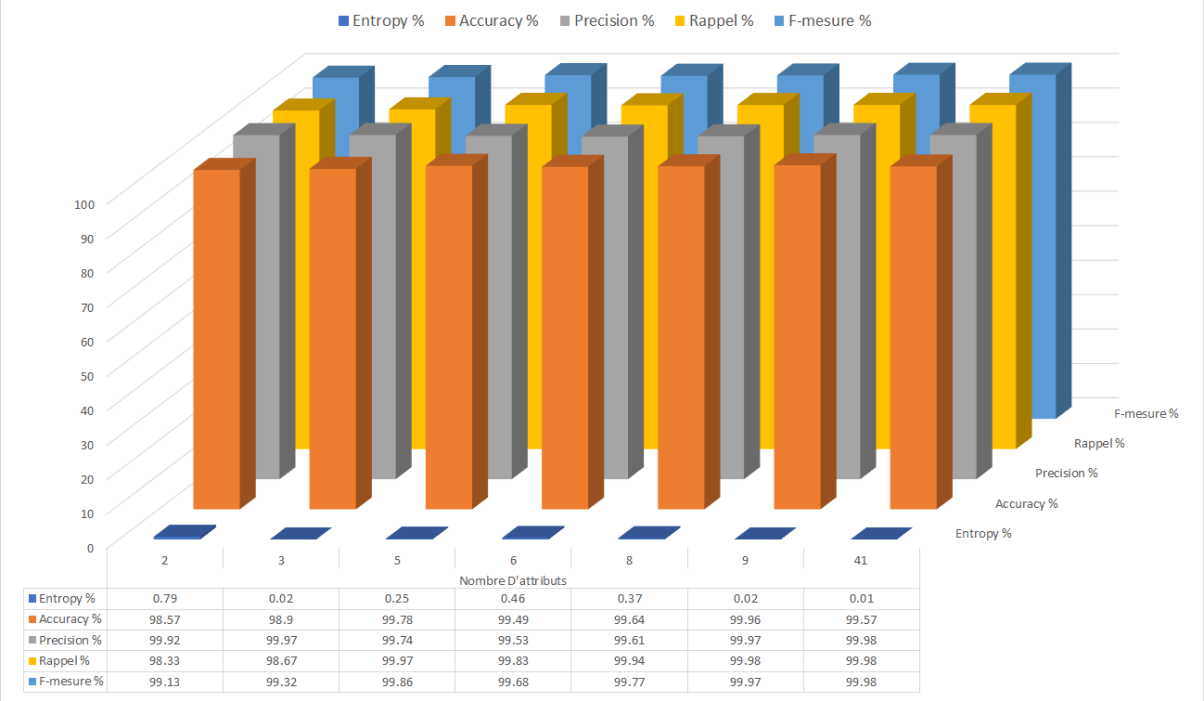
Figure 4.12 – TP_rate et FP_rate de KPPV

- TP_rate = 99.95%
- FP_rate = 0.1%

Nous voyons que cela nous a donné le meilleur résultats avec `TP_rate` et `FP_rate`, donc c'est la signification qui peut être détecter Intrusion et Nointrusion.

4.5.5.5 Arbre de décision

Avec les performances : Entropie, Accuracy, Précision, Rappel, F-mesure



	2	3	5	6	8	9	41
Entropy %	0.79	0.02	0.25	0.46	0.37	0.02	0.01
Accuracy %	98.57	98.9	99.78	99.49	99.64	99.96	99.57
Precision %	99.92	99.97	99.74	99.53	99.61	99.97	99.98
Rappel %	98.33	98.67	99.97	99.83	99.94	99.98	99.98
F-mesure %	99.13	99.32	99.86	99.68	99.77	99.97	99.98

Figure 4.13 – Visualisation graphique d'attributs sélectionnés par Arbre de décision

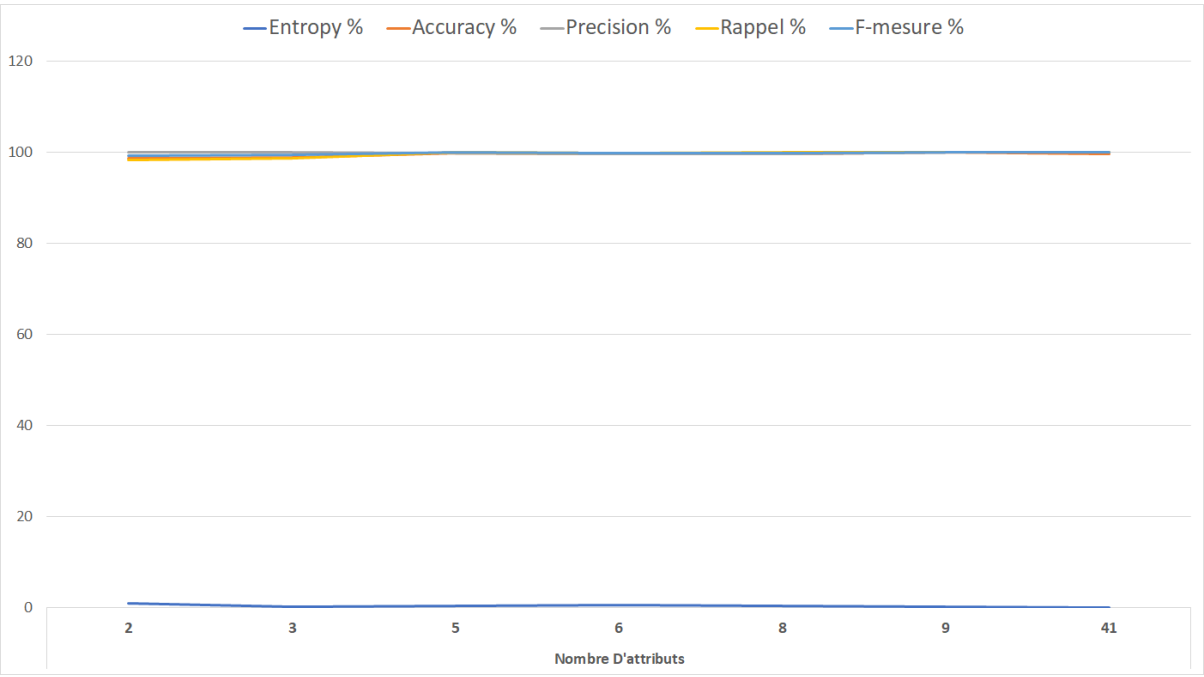


Figure 4.14 – Visualisation par courbe d'attributs sélectionnés par Arbre de décision

4.5.5.5.1 La discussion des résultats : a partir des figures (Figure 4.13, Figure 4.14) on conclue que :

- Pour le **Accuracy**, Augmentation progressive de 2 à 41 attributs avec un pourcentage de 98% un peut prêt.
- Pour La **précision**, Augmentation progressive de 2 à 41 attributs avec un pourcentage de 99% un peut prêt.
- Pour Le **Rappel**, Augmentation progressive de 2 à 41 attributs avec un pourcentage de 98% un peut prêt.
- Pour Le **F-mesure**, Augmentation progressive de 2 à 41 attributs avec un pourcentage de 99% un peut prêt.
- Pour **Entropie**, Augmentation progressive de 2 à 5.avec un pourcentage de 0.08%, Et diminution progressive de 6 à 41. avec un pourcentage de 0.04% un peut prêt.

Avec performances de TP_rate et FP_rate :

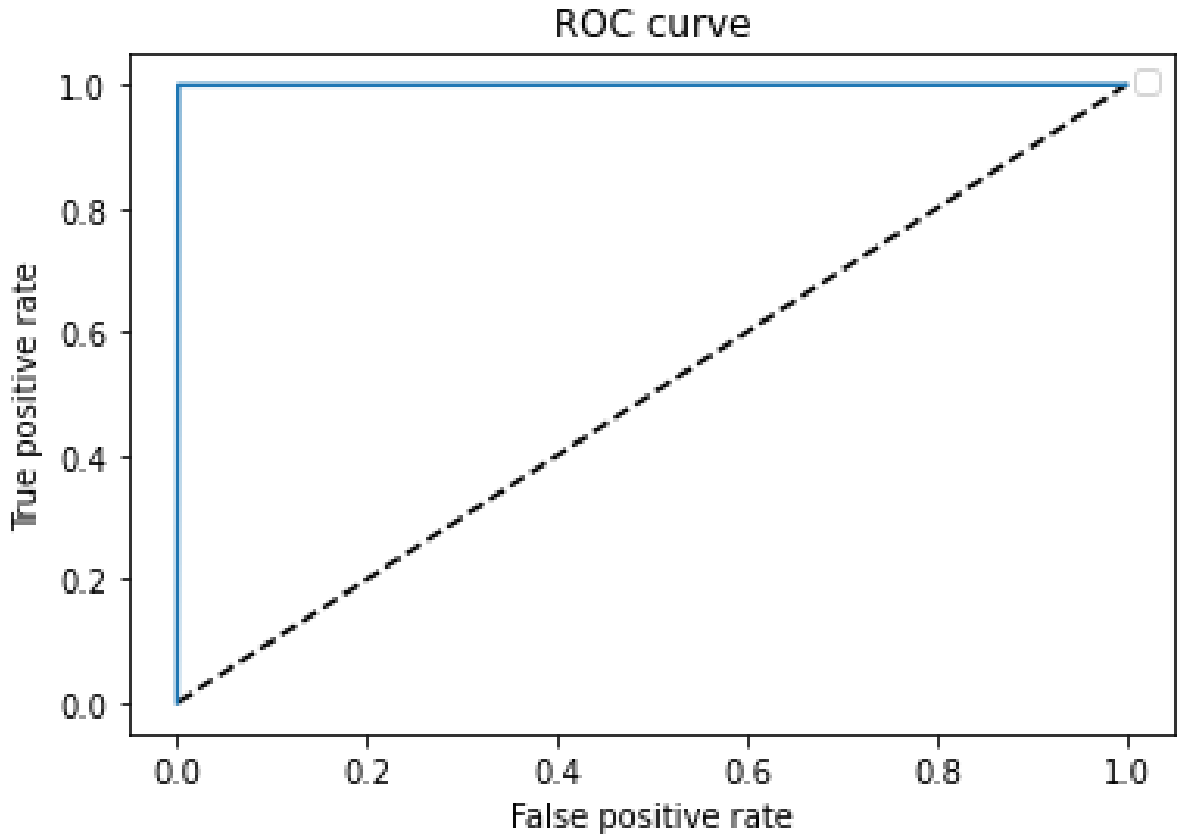


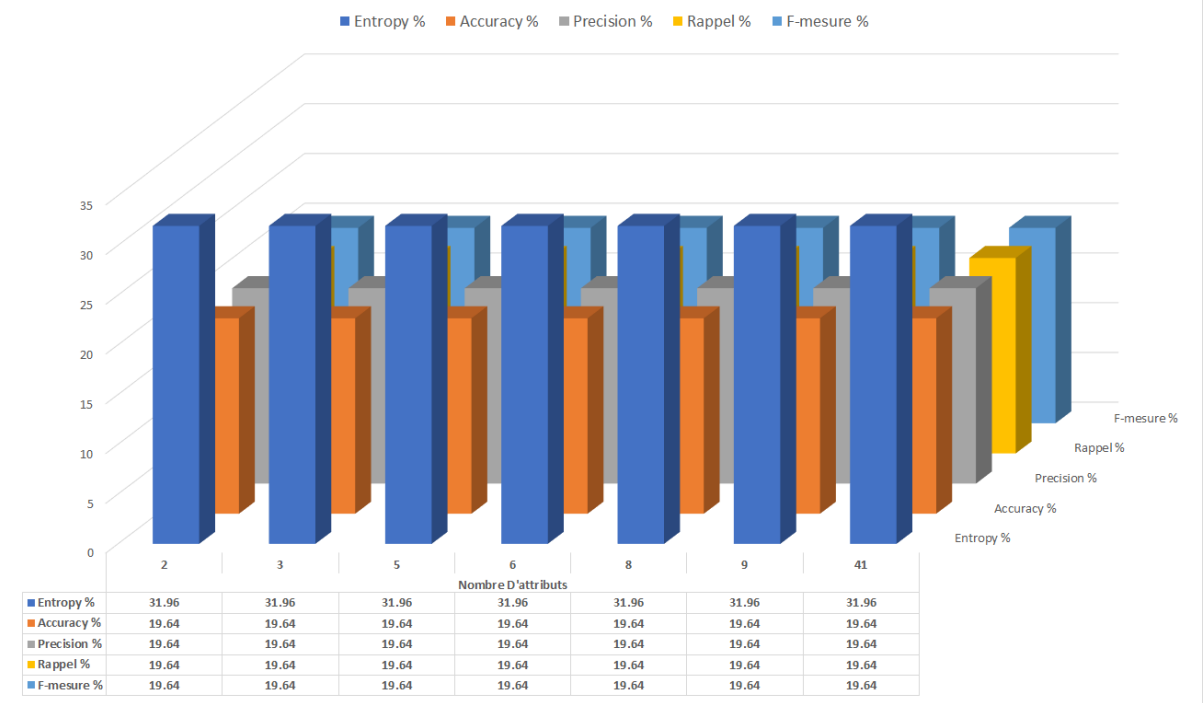
Figure 4.15 – TP_rate et FP_rate de KPPV

- TP_rate = 99.98%
- FP_rate = 0.02%

Nous voyons que cela nous a donné le meilleur résultats avec `TP_rate` et `FP_rate`, donc c'est la signification qui peut être détecter Intrusion et Nointrusion.

4.5.5.6 K-means

Avec les performances : Entropie, Accuracy, Précision, Rappel, F-mesure



	2	3	5	6	8	9	41
Entropy %	31.96	31.96	31.96	31.96	31.96	31.96	31.96
Accuracy %	19.64	19.64	19.64	19.64	19.64	19.64	19.64
Precision %	19.64	19.64	19.64	19.64	19.64	19.64	19.64
Rappel %	19.64	19.64	19.64	19.64	19.64	19.64	19.64
F-mesure %	19.64	19.64	19.64	19.64	19.64	19.64	19.64

Figure 4.16 – Visualisation graphique d'attributs sélectionnés par K-means

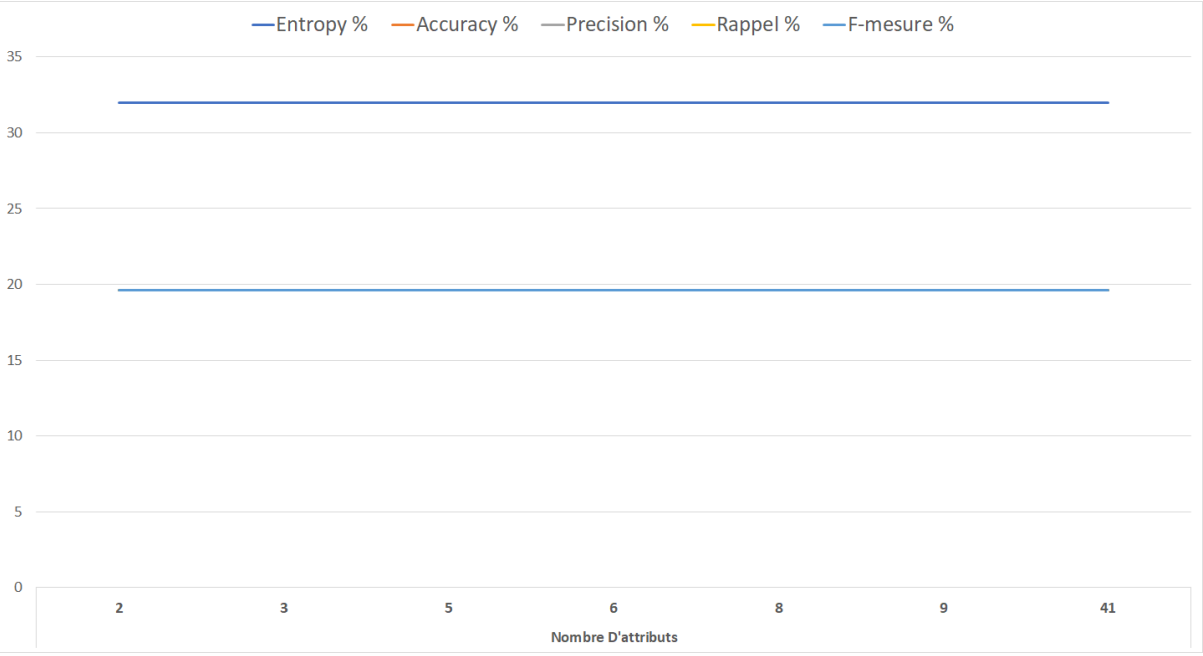


Figure 4.17 – Visualisation par courbe d'attributs sélectionnés par K-means

4.5.5.6.1 La discussion des résultats : a partir des figures (Figure 4.16, Figure 4.17) on conclue que :

- Pour le **Accuracy**, elle est stable de 2 à 41 attributs avec un pourcentage de 98% un peut prêt.
- Pour La **précision**, elle est stable de 2 à 41 attributs avec un pourcentage de 99% un peut prêt.
- Pour Le **Rappel**, elle est stable de 2 à 41 attributs avec un pourcentage de 98% un peut prêt.
- Pour Le **F-mesure**, elle est stable de 2 à 41 attributs avec un pourcentage de 99% un peut prêt.
- Pour **Entropie**, elle est stable de 2 à 41 attributs avec un pourcentage de 31.96% un peut prêt. **perte d'information**

Avec performances de TP_rate et FP_rate :

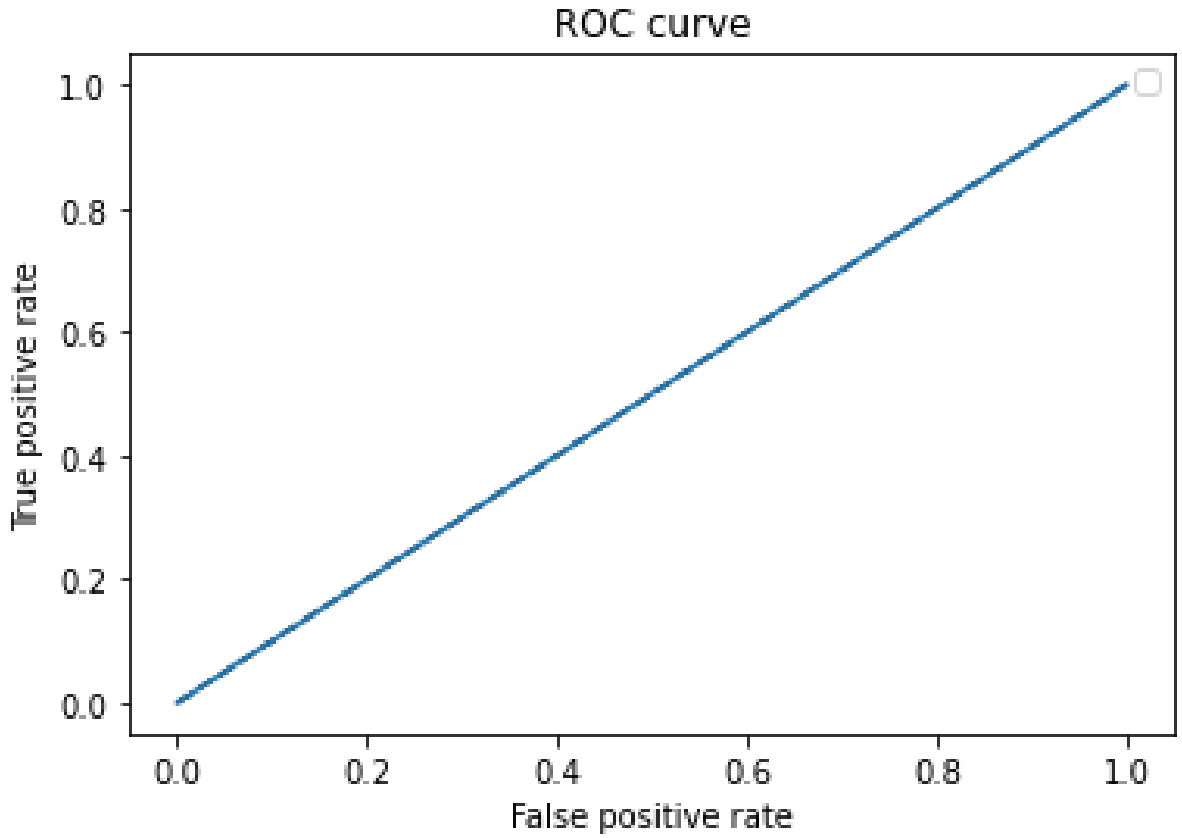


Figure 4.18 – TP_rate et FP_rate de K-means

- TP_rate = 0.0004%
- FP_rate = 0%

On voit que ça nous a donné, les pires résultats avec `TP_rate` et `FP_rate` , c'est à dire que k-means ne pas pouvoir détecter INTRUSION et NoINTRUSION.

4.6 Comparaison entre algorithmes

On comparera avec la performance de l'entropie : avec le nombre d'attributs sélectionnés

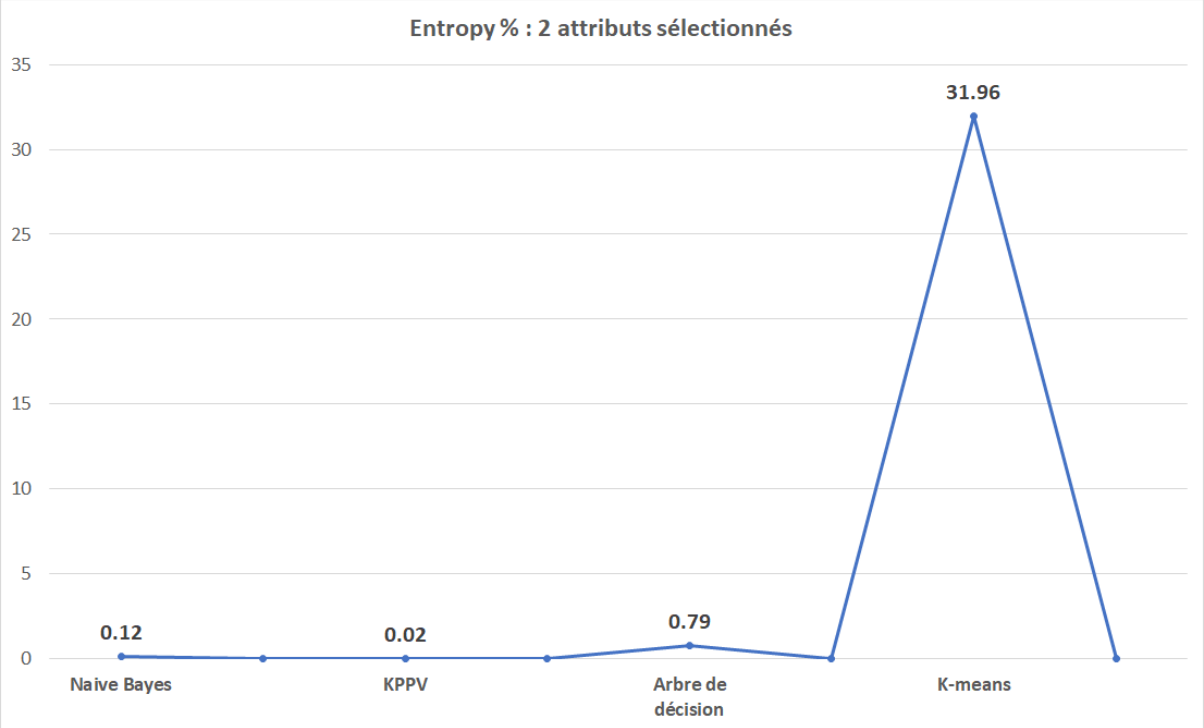


Figure 4.19 – Visualisation par courbe de la comparaison des résultats de l'entropie avec 2 attributs sélectionnés

Comme on a vu dans la figure précédente (Figure 4.19), l'algorithme de **KPPV** nous a donné dans le cas de 2 attributs sélectionnés de la KDDCup99 une entropie inférieur que les autres algorithmes.

Mais aussi les autres algorithmes comme Naive bayes et arbre de décision, ils nous ont donné de bons résultats avec l'entropie de 2 attributs sélectionnés, mais le meilleur modèle est l'entropie la plus petit. Dans ce cas, le meilleur modèle est **KPPV**

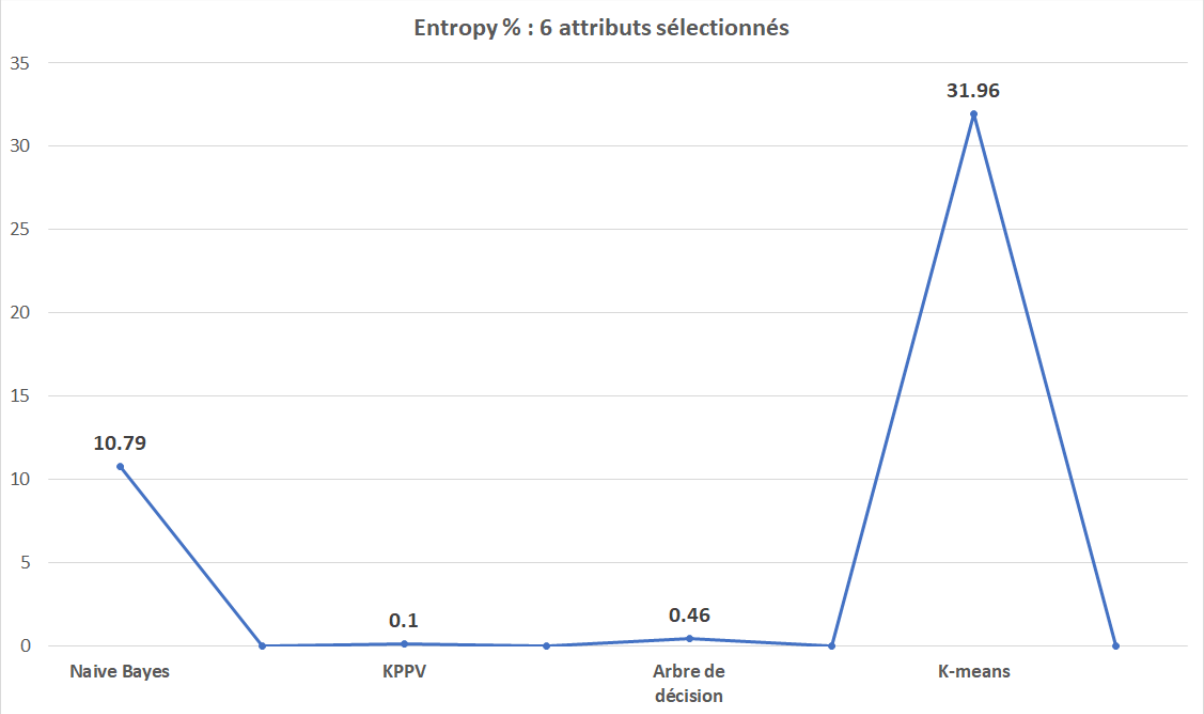


Figure 4.20 – Visualisation par courbe de la comparaison des résultats de l'entropie avec 6 attributs sélectionnés

Comme on a vu dans la figure précédente(4.20), l'algorithme de **KPPV** nous a donné dans le cas de 2 attributs sélectionnés de la KddCup99 une entropie inférieur que les autres algorithmes.Avec entropie de 0.1%

Mais aussi l'algorithme de l'arbre de décision, ils nous ont donné de bons résultats avec l'entropie de 2 attributs sélectionnés, mais le meilleur modèle est l'entropie la plus petit. Dans ce cas, le meilleur modèle est **KPPV**

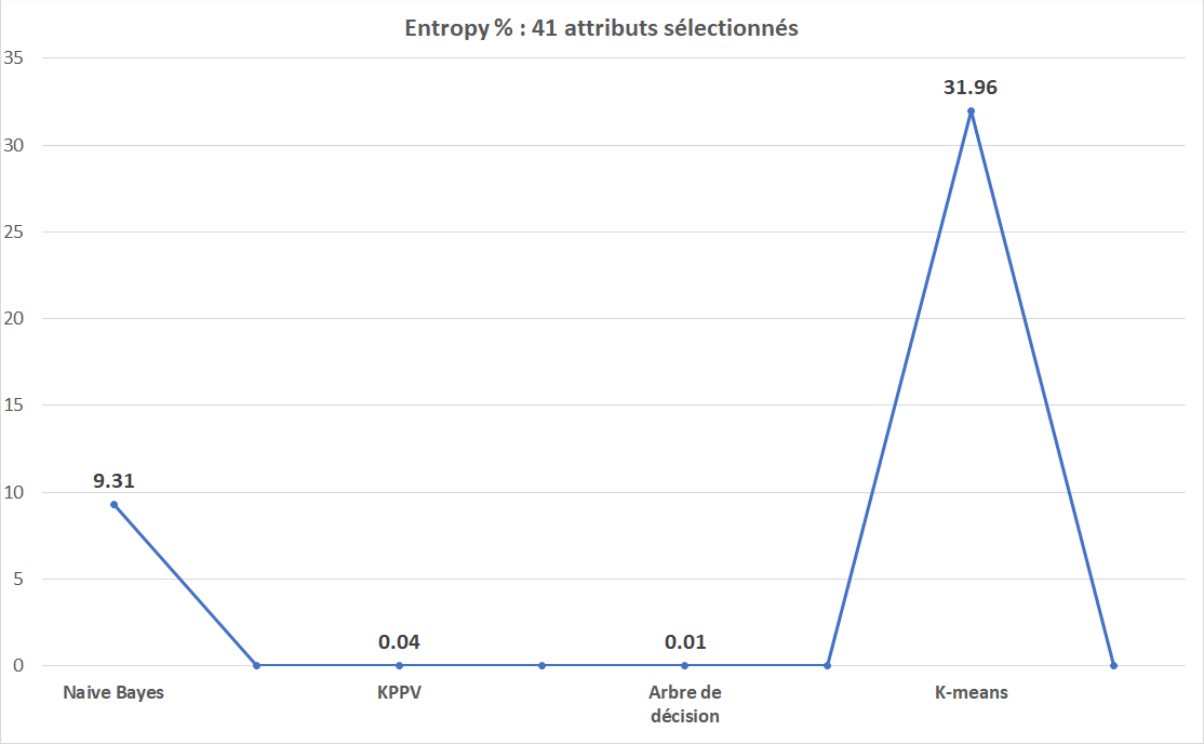


Figure 4.21 – Visualisation par courbe de la comparaison des résultats de l'entropie avec 41 attributs sélectionnés

Dans le cas de tous les attributs sélectionnés (41 attributs), nous avons vu des Naive bayes son entropie est élevée, aussi k-means, mais l'algorithme KPPV et l'arbre de décision, leur entropies est très petit,

Avec une petite différence entre arbre de décision et naive bayes en entropie, mais dans le reste du nombre d'attributs sélectionnés, nous avons donné de bons résultats **KPPV**.

4.7 Remarque

4.7.1 Pour les attributs sélectionnés

4.7.1.1 Pour L'information Mutuelle :

- **Pour 2 attributs sélectionnés :** `src_bytes, count.`
- **Pour 6 attributs sélectionnés :** `src_bytes, count ,service, dst_bytes, dst_host_same_src_port_rate, srv_count.`
- **Pour 41 attributs sélectionnés :** Tous les attributs de KDDCup 99 sélectionnés.

4.7.1.2 Pour Chi-square(chi2) :

- **Pour 2 attributs sélectionnés :** `dst_bytes, src_bytes.`

- **Pour 6 attributs sélectionnés :** `dst_bytes, src_bytes, duration, count, dst_host_count, service`
- **Pour 41 attributs sélectionnés :** Tous les attributs de KDDCup 99 sélectionnés.

4.8 Conclusion

Dans ce chapitre, nous avons présenté notre phase d'expérimentations et résultats, nous avons appliqué l'ensemble de 3 algorithmes d'apprentissage supervisé (Naive Bayes, KPPV : K plus Proche Voisins, et Arbre de décision) et un algorithme d'apprentissage non supervisé (K-means) sur un ensemble de données 'KDDCup99', et nous avons calculé les performances des mesures qui nous permettent de choisir le meilleur algorithme.(Le meilleur dans notre phase d'expérimentations et résultats est KPPV)

.

Conclusion Générale

Dans ce travail, on a proposé une approche pour la détection des intrusions en utilisant l'ensemble de données (Data Set)« KDDCup99 », ça veut dire, définir s'il y a INTRUSION ou NoIntrusion. testant différents algorithmes de Machine Learning et en sélectionnant les algorithmes les plus optimaux pour notre problématique.

Afin d'aboutir à ces résultats, on doit lire et étudier plusieurs publications et articles pour voir ce qui se fait de mieux en la détection et pour pouvoir concevoir notre propre approche.

D'après les expériences vues dans les articles on remarque que c'est important de tester plusieurs algorithmes pour savoir qui est le mieux adapté pour une certaine problématique, parce qu'il n'y a pas une règle générale qui dit qu'un algorithme est meilleur qu'un autre

Grâce à l'avancée de l'intelligence artificielle, comme le Machine Learning ça nous a permet de traiter des problématiques plus complexes et d'avoir de bons résultats ce qui est très important dans le domaine de la sécurité informatique, L'intelligence artificielle offre de nombreux avantages pour explorer la sécurité informatique. Cela promet d'innover dans le système d'informatique pour un avenir meilleur.

Cette étude nous a permis de mieux comprendre les applications de machine Learning dans le détection d'intrusion. C'était également une leçon importante sur les conversions de données et l'analyse des algorithmes.

Notre objectif dans ce Travail est d'aborder un modèle ou un algorithme de Machine Learning pour la détection d'intrusion, et qui donnes des bonnes résultats dans la sécurité informatique, et qui se concurrent avec les autres algorithmes de machine learning, avec des longues recherche et des bonnes orientations de notre encadreur.

Donc, l'expérimentation que nous avons initiée et par ses résultats satisfaisants montre le bien fait qui nous a guidés à choisir des algorithmes de Machine Learning.

Future work

La contribution présentée dans le présent document est une méthode de détection des intrusions fondée sur la combinaison de méthodes de sélection des attributs et de méthodes de classification. Le système conçu est capable de détecter un grand nombre d'intrusions tout en maintenant le taux de faux positifs raisonnablement bas, obtenant de meilleures performances que le gagnant KDDCup99. Plus précisément, les algorithmes de Naïve Bayes et KPPV et Arbre de décision obtiennent le meilleur résultat. En outre, alors que les travaux précédents avaient déjà réduit le nombre nécessaire d'attributs, nous avons réalisé une amélioration des performances générales en utilisant le même nombre d'attributs.

Lorsqu'il s'agit de la version multiclasse de l'ensemble de données KDDCup99, la reconnaissance des modèles et les algorithmes d'apprentissage automatique obtiennent généralement de mauvaises performances, car ils ne parviennent généralement pas à détecter la plupart des types d'attaques les moins représentés. Grâce aux bons résultats obtenus dans ce travail, les travaux futurs consisteront à élargir la méthode proposée pour traiter les problèmes multiclasses et à l'appliquer à la version multiclasses de l'ensemble de données KDDCup99.

En outre, des méthodes dynamiques d'initialisation des nœuds seront envisagées afin de pouvoir traiter plus efficacement les problèmes de chevauchement des classes.

Bibliographie

[1] Arnold JOHNSON. *Guide for Security-Focused Configuration Management of Information Systems : The National Institute of Standards and Technology Special Publication 800-128*. 2012 (page 5).

[2] E COLE. *Propagation of Sound in Porous Media*. second. Chichester : John Wiley et Sons, 2009, p. 72-89 (pages 5, 6).

[3] Piotr Dorosz Przemysiam Kazienko. *Intrusion Detection Systems (IDS) Part I - (network intrusion ; attack symptoms ; IDS tasks ; and IDS architecture*. 2004 (page 7).

[4] Cédric LIORENS. *Tableaux de bord de la sécurité réseau*. Editions Eyrolles, 2011. ISBN : 2-212- 11973-9 (pages 8-12).

[5] *Attaque Man in the Middle (MITM)*. https://www.malekal.com/man-in-the-middle/. Consulté à : 14-05-2022 (pages 9, 10).

[6] Guillaume DESGEORGE. *La sécurité des réseaux*. 2000 (pages 12, 13).

[7] Ahmed Chaouki LOKBANI. « Le problème de sécurité par le Data Mining ». Thèse de doct. Université Djillali Liabes - Sidi Bel Abbes, 2017 (pages 12, 13, 18, 19, 21).

[8] Philipe Biondi. « Architecture expérimentale pour la détection d'intrusions dans un système informatique ». In : Avril-Septembre 2001 (pages 16, 17).

[9] Salima Hassas. « ystèmes Complexes à base de Multi-Agents Situés » ,Université Claude Bernard-Lyon 1 ». In : 2003 (page 16).

[10] Nicolas Nobelis. « Un modèle de Case-Based Reasoning pour la détection d'intrusion ». In : Université nice SOPHIA ANTIPOLIS,Rapport de stage DEA RSD/ESSI3 SAR, 2002 (page 16).

[11] Liran LERMAN. « Les systèmes de détection d'intrusion basés sur du machine learning ». In : Université LIBRE de BRUXELLES (page 16).

[12] Jim MELLANDER Carl F ENDORF Eugene SCHULTZ. « Intrusion detection & prevention. McGrawHill Osborne Media ». In : 2004 (page 17).

[13] M.Dacier et A.Wespi H.Debar. *A revised taxonomy for intrusion detection systems*. Annales des télécommunications, July-August 2000 (pages 17, 19-21).

[14] *Introduction et Initiation à la sécurité informatique*. SecuriteInfo.com. Consulté à : 18-01-2022 (page 17).

[15] Ghenima BOURKACHE. « Un IDS réparti basé sur une société d'agents intelligents ». Mém. de mast. 2006 (page 18).

[16] Farah JEMILI. *Système de Détection et de Prévision d'Intrusions : À base de réseaux d'inférence incertaine et imprécise dans une architecture multi-agent*. Déc. 2013. ISBN : 978-3-8416-2121-4 (page 19).

[17] James P ANDERSON. *Computer security threat monitoring and surveillance*. Rapp. tech. Fort Washington, Pennsylvania, 1980 (page 20).

[18] Dorothy E DENNING. *An intrusion-detection model*. IEEE Transactions on software engineering 2, 1987, p. 222-232 (page 20).

[19] ABIZA IMAD. « Les systèmes de détections d'intrusion basés sur machine learning ». Mém. de mast. 2018 (pages 22, 23).

[20] Batta MAHESH. « Machine Learning Algorithms -A Review ». In : (jan. 2019). DOI : 10.21275/ART20203995 (page 27).

[21] *What is Machine Learning ?)* https://www.javatpoint.com/machine-learning. Consulté à : 12-05-2022 (page 27).

[22] *Classification of Machine Learning)*. https://forum.huawei.com/enterprise/en/classification-of-machine-learning/thread/700327-895. Consulté à : 12-05-2022 (pages 28-30).

[23] *What is Data Science*. https://www.geeksforgeeks.org/what-is-data-science/. Consulté à : 11-05-2022 (page 30).

[24] *Qu'est-ce que la data science ?* https://www.oracle.com/dz/data-science/what-is-data-science/. Consulté à : 11-05-2022 (page 30).

[25] Sons JOHN WILEY. *Data Science , Big Data Analytics Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, 2015 (pages 31-33).

[26] *Qu'est ce que l'algorithme KNN ?* https://datascientest.com/knn. Consulté à : 12-05-2022 (pages 34, 35).

[27] *K-MEANS (OU K-MOYENNES)*. https://dataanalyticspost.com/Lexique/k-means-ou-k-moyennes/. Consulté à : 12-05-2022 (page 36).

[28] *K-Means Clustering Algorithm)*. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning. Consulté à : 12-05-2022 (page 36).

[29] *Data Science Tutorial)*. https://www.javatpoint.com/data-science. Consulté à : 12-05-2022 (page 37).

[30] Hadj Ahmed BOUARARA. « Fouille de données (Data Mining) ». In : *Polycopié de cours - USMT*. 2019 (pages 38, 40, 54).

[31] *Qu'est-ce que le data mining ?* https://www.tibco.com/fr/reference-center/what-is-data-mining. Consulté à : 18-05-2022 (page 38).

[32] *Data Mining, explorer les données du Data Warehouse*. https://www.piloter.org/business-intelligence/datamining.htm. Consulté à : 18-05-2022 (pages 38, 39).

[33] M. & Meira Zaki. « Data mining and analysis : fundamental concepts and algorithms ». In : *Cambridge University Press*. 2014 (pages 40, 41).

[34] T. G Dietterich. « Ensemble methods in machine learning. In International workshop on multiple classifier systems ». In : *Berlin, Heidelberg*. 2000, June (page 40).

[35] X Dua S & Du. « Data mining and machine learning in cybersecurity ». In : *Auerbach Publications*. 2016 (page 41).

[36] *Difference Between Data Science and Data Mining*. https://www.geeksforgeeks.org/difference-between-data-science-and-data-mining/. Consulté à : 18-05-2022 (pages 42, 43).

[37] S. Hettich et S. D. Bay. *The UCI KDD Archive [http ://kdd.ics.uci.edu]*. Irvine, CA : University of California, Department of Information et Computer Science, 1999 (pages 45, 46).

[38] J. Song et al. « Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation ». en. In : *Paper presented at : Proceedings of the 1st Workshop on Building Analysis Datasets and Gathering Experience Return for Security*. Salzburg Austria, 2011 (pages 45, 46).

[39] M. Tavallaee et al. « A detailed analysis of the KDD CUP 99 data set ». en. In : *Paper presented at : Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications*. Ottawa, ON, Canada : IEEE, 2009 (pages 45, 46).

[40] N. Moustafa et J. Slay. « UNSW-NB15 : a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set ». en. In : *Paper presented at : Proceedings of the Military Communications and Information Systems Conference (MilCIS*. Canberra, ACT, Australia : IEEE, 2015 (pages 45, 46).

[41] N. Moustafa et J. Slay. « The evaluation of network anomaly detection systems : statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set ». en. In : *Inf Sec J A Global Perspect* 25.1-3 (2016), p. 18-31. doi : 10.1080/19393555.2015.1125974 (page 45).

[42] Razan Abdulhammed et al. « Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection ». In : *Electronics* 8.3 (2019). issn : 2079-9292. doi : 10.3390/electronics8030322. url : https://www.mdpi.com/2079-9292/8/3/322 (pages 45, 46).

[43] Gozde Karatas, Onder Demir et Ozgur Koray Sahingoz. « Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset ». In : *IEEE Access* 8 (2020), p. 32150-32162. doi : 10.1109/ACCESS.2020.2973219 (page 46).

[44] *KDD Cup 1999 Data Abstract*. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. Consulté à : 12-02-2022 (page 46).

[45] *INTRUSION DETECTOR LEARNING*. http://kdd.ics.uci.edu/databases/kddcup99/task.html. Consulté à : 12-02-2022 (pages 46, 47).

[46] Salvatore STOLFO et al. « Cost-Based modeling and evaluation for data mining with application to fraud and intrusion detection : Results from the JAM project ». In : (sept. 1999) (page 47).

[47] Richard LIPPMANN et al. « Evaluating intrusion detection systems : the 1998 DARPA off-line intrusion detection evaluation ». In : t. 2. Fév. 2000, 12-26 vol.2. ISBN : 0-7695-0490-6. DOI : 10.1109/DISCEX.2000.821506 (page 47).

[48] *Anaconda*. https://www.anaconda.com/ (page 49).

[49] *Jupyter Notebook*. https://jupyter.org/ (page 50).

[50] Alexander KRASKOV, Harald STÖGBAUER et Peter GRASSBERGER. « Estimating mutual information ». In : *Phys. Rev. E* 69 (6 juin 2004), p. 066138. DOI : 10.1103/PhysRevE.69.066138. URL : https://link.aps.org/doi/10.1103/PhysRevE.69.066138 (page 53).

[51] Reda Mohamed HAMOU. « Reduction de dimention et similarité ». In : *Polycopié de cours - USMT*. 2019 (pages 53, 54).

[52] MATALLAH HOCINE. « Classification Automatique de Textes Approche Orientée Agent ». Mém. de mast. 2011 (pages 56, 58).

[53] RAHMANI Sihem & DAHMANI RADJAA. « La sécurité dans un IOT : Machine Learning pour la détection des botnets dans l'IOT ». Mém. de mast. 2021 (page 56).

[54] *What is the AUC - ROC Curve ?* https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. Consulté à : 12-05-2022 (page 57).

ملخص

قد أدت التطورات الأخيرة في تكنولوجيا المعلومات إلى تغييرات مستمرة في جميع المجالات، حيث أصبح نقل المعلومات عن طريق الشبكات والحواسيب أحد العلامات التي لا غنى عنها للعصر الرقمي. لتسهيل متطلبات الحياة الحديثة من خلال تقليل حجم الأعمال وتطوير طرق إنتاج وتخزين وتوزيع المعلومات، وقد أدى ذلك إلى زيادة مشاكل المعلومات والمخاطر التي تهدد وتؤثر على أمن اقتصاد المعلومات في الشركة ، الأمر الذي يتطلب اليقظة من أجل ضمان الأمن اللازم لمعلوماتها، لا سيما وأنها تعيش في اقتصاد المعلومات لحائز المعلومات الخاضعة للسيطرة، وقد ظهرت نُهج جديدة تسمى نظم الكشف عن التسلل (IDS) هدفهم هو تحليل حركة الطلبات واكتشاف السلوك الضار.

هدفنا في هذا المشروع هو معالجة نموذج أو خوارزمية التعلم الآلي لاكتشاف التسلل، والتي تعطي نتائج جيدة في أمان الكمبيوتر، والتي تتنافس مع خوارزميات التعلم الآلي الأخرى.

الكلمات الرئيسية :لأمن، كشف التسلل، تنقيب البيانات، علم البيانات، إختيار العناصر.

Abstract

Developments in information technology have led to continuous changes in all areas, with the transmission of information through networks and computers becoming one of the indispensable signs of the digital age. To facilitate modern life requirements by reducing business volume and developing methods of producing, storing and distributing information information problems and risks that threaten and affect the security of the company's information economy, this requires vigilance in order to ensure the necessary security of its information. information economy ", particularly as it lives in the information economy of the controlled information holder, New approaches called IDS have emerged. Their goal is to analyze the movement of requests and detect harmful behavior.

Our goal in this project is to process a machine learning model or algorithm for intrusion detection, which gives good results in computer security, which rival other machine learning algorithms.

Keywords: Security, detection intrusion system, Data Mining, Data Science, features selection, KDDCup99.

Résumé

Les développements récents des technologies de l'information ont entraîné des changements continus dans tous les domaines, la transmission de l'information par les réseaux et les ordinateurs devenant l'un des signes indispensables de l'ère numérique. Faciliter les exigences modernes en matière de durée de vie en réduisant le volume d'affaires et en développant des méthodes de production, de stockage et de distribution des problèmes d'information et des risques qui menacent et affectent la sécurité de l'économie de l'information de l'entreprise, Cela exige une vigilance afin d'assurer la sécurité nécessaire de ses informations. Économie de l'information, en particulier dans la mesure où elle vit dans l'économie de l'information du détenteur d'information contrôlé, De nouvelles approches appelées IDS ont émergé. Leur but est d'analyser le mouvement des requêtes et de détecter les comportements nuisibles.

Notre objectif dans ce projet est de traiter un modèle ou un algorithme d'apprentis sage automatique pour détection d'intrusion, ce qui donne de bons résultats en sécurité informatique, qui rivalisent avec d'autres algorithmes d'apprentissage automatique.

Mots clés : Sécurité Informatique, système de détection d'intrusion, Data Mining, Data Science, Sélection des attributs, KDDCup99.