الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة سعيدةد. مولاي الطاهر

كلية التكنولوجيا

قسم: الإعلام الآلي

Mémoire de Master

Spécialité : réseau informatique et system réparti

Thème

Système de détection d'intrusion à base D'apprentissage automatique

Présenté par :

Dris chaima

Abdelhakim Halima

Dirigé par :

Benyahia Kadda

Année universitaire 2022-2023

June 22, 2023

Remerciement

Avant tout, le grand et le vrai merci revient à Allah qui nous a donné la force, la foi et la vie pour accomplir cette tâche, qui semblait d'abord une mission difficile Au terme de ces travaux, nous tiens à remercier notre directeur de thèse **Mr.Benyahia Kadda** qui a dirigé ce travail, pour sa sagesse, son expérience, ses conseils et encouragements, sa patience, de nous faire confiance et laissé la liberté nécessaire à l'accomplissement de notre travail. Merci Cheikh. Nous tiens également à remercier les membres du Jury. Nos plus vifs remerciements s'adressent aussi à tout le cadre professoral et administratif de Faculté de technologie d'université de DR.Moulay Taher , Saida. Et bien sure nous gardons une place toute particulière à nos parents, nos frères et nos sœurs , nos amis qui sont toujours à nos côtés. Nos remerciements vont enfin à tous ceux qui ont Contribué directement ou indirectement au développement de ces travaux

Dédicaces

À mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études, A mes chères sœurs pour leurs encouragements permanents, et leur soutien moral, A toute ma famille pour leur soutien tout au long de mon parcours universitaire, Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fuit de votre soutien infaillible, Merci d'être toujours là pour moi.

Abedelhakimi Halima

Dédicaces

À mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études, à mes sœurs et frères pour leurs encouragements constants et leur soutien moral et matériel, à tous les membres de ma famille pour leur soutien tout au long de mon parcours universitaire, J'espère que ce travail sera l'exaucement de vos soi-disant souhaits, et je remercie également mon amie Amina Nouari pour son soutien à mon égard. Merci à Dieu en tout cas.

Dris CHaima

résumé

Le modèle d'IDS d'apprentissage automatique est l'une des techniques de ce type les plus populaires, qui peut détecter, déterminer et prédire les attaques avant les résultats. Dans le présent travail, nous couplons cinq algorithmes d'apprentissage automatique avec trois algorithmes de sélection d'attributs sur le jeu de données CICIDS-2017 afin de sélectionner le modèle qui améliore mieux la détection d'intrusion.

Mots-clés sécurité, détection d'intrusion, d'apprentissage automatique, dataset cicids2017

Abstract

The machine learning IDS model is one of the most popular such techniques, which can detect, determine and predict attacks before results. In the present work, we couple five machine learning algorithms with three attribute selection algorithms on the CICIDS-2017 dataset to select the model that better improves intrusion detection.

Keywords security, intrusion detection, machine learning, dataset cicids2017

ملخص

للتعلم الآلي أحد أكثر هذه التقنيات شيوعًا، حيث يمكنه اكتشاف الهجمات وتحديدها IDS يعد نموذج والتنبؤ بها قبل ظهور النتائج. في العمل الحالي، قمنا بربط خمس خوارزميات للتعلم الآلي مع ثلاث لتحديد النموذج الذي يحسن اكتشاف CICIDS-2017 خوارزميات لاختيار السمات في مجموعة بيانات التسلل بشكل أفضل.

الكلمات الرئيسية الأمن، كشف التسلل، التعلم الآلي، اختيار العناصر

Contents

List of Figures

Introduction générale

Internet a révolutionné la société et devient rapidement un besoin de la vie quotidienne. Alors que l'expansion d'Internet continue de donner des découvertes révolutionnaires et des avantages qui changent la vie de la société, elle permet également aux ennemis de se livrer à des comportements nuisibles dans ce domaine numérique. Ces ennemis utilisentl'Internet pour exploiter la connectivité de la société afin de réaliser un objectif malveillant. Ces objectifs peuvent inclure le vol de propriété intellectuelle, le déni de service, l'interruption des activités, le vol d'informations personnellement identifiables ou d'informations de carte de crédit, la fraude financière, les demandes de rançon, la destruction de biens physiques et d'autres activités malveillantes.

Un système de détection d'intrusion (IDS) examine toutes les activités réseau entrantes et sortantes, à la recherche de modèles suspects qui pourraient indiquer une attaque réseau ou système ou une attaque par quelqu'un tentant de s'introduire ou de compromettre un système. Les algorithmes d'apprentissage automatique ont obtenu de bons résultats en matière de détection d'intrusion. Dans ce mémoire nous appliquons cinq algorithmes de classification (Random Forest (RF), Decisiontree , LogisticRegression (LR) , Naive Bayes et Stochastic Gradient Descent) avec trois algorithmes de sélection des attributs (Anova,IPCA et CHI2)sur le jeu de données CICIDS-2017 pour choisir quel est la bonne combinaison qui convient pour augmenter les performances de détection des intrusions.

Notre mémoire s'organise en trois chapitres : Le premier chapitre présente une description détaillé des notions autour de la sécurité informatique, Dans, le deuxième chapitre nous présentons les systèmes de détections d'intrusion , Les expérimentations et la discussion des résultats font l'objet du troisième chapitre.

Chapitre 01

1 sécurité informatique

Introduction

En raison de plusieurs facteurs notamment l'ouverture des systèmes d'information sur Internet, l'évolution de la technologie et des moyens de communication ainsi que la transmission de données à travers les réseaux, des risques d'accès et de manipulation des données par des personnes non autorisées d'une façon accidentelle ou bien intentionnelle sont apparus. Donc la mise en place d'une politique de sécurité autour de ces systèmes est devenue une nécessité incontournable.
Le système de détecté d'intrusion est l'une des techniques utilisées pour garantir un contrôle permanent des attaques ainsi que la détection de toute violation de cette politique, c'estàdire toute intrusion. Dans ce premier chapitre nous introduisons les principales notions de base de la sécurité informatique y compris sa définition, ses objectifs, les problèmes et les attaques informatiques et aussi iles mécanismes permettant d'amélioreras écurité.Ensuite, nous présentons les systèmes de détection d'intrusions, leur définition, architecture, classification..., et nous terminons par les limites des systèmes de détection d'intrusions actuels.

1.1 Sécurité informatique :

1.2 Définition La sécurité informatique :

La protection des informations contre les menaces potentielles est connue sous le nom de sécurité de l'information, qui vise à promouvoir la continuité des activités, à minimiser les risques commerciaux et à optimiser les retours sur investissement et les opportunités commerciales [1].

1.3 Objectifs de la sécurité :

Lors de la résolution de problèmes de sécurité, on vise à atteindre certains objectifs dont les principaux sont [2] :

• **L'authentification :** :
C'est un mécanisme qui peut vérifier l'identité des participants à la communication en distinguant les nœuds légitimes des intrus. En effet, lors de la communication entre deux nœuds légitimes, tout nœud malveillant peut altérer les paquets de données afin de les modifier ou d'injecter d'autres paquets supplémentaires. Par conséquent, le destinataire doit s'assurer que les données utilisées proviennent de sources légitimes. Par conséquent, l'authentification est essentielle pour vérifier l'identité de l'expéditeur des données et pour authentifier les données lors de leur passage sur le réseau.[2]

•**L'intégrité :** C'est un service qui vérifie que les données n'ont pas été altérées ou corrompues de quelque manière que ce soit, intentionnellement ou

Figure 1: Objectifs de la sécurité informatique. [3]

accidentellement, et conserve un format qui permet de les utiliser lors du traitement, du stockage ou de la transmission.[2]

•La confidentialité :
Les seuls nœuds autorisés doivent être en mesure d'accéder aux données. La confidentialité garantit que les données d'un nœud ne sont accessibles ou révélées qu'au destinataire . Ce service fait référence à la capacité du réseau à garantir que ses services fonctionnent correctement en garantissant aux parties communicantes la présence et l'utilisation de l'information au moment souhaité. La disponibilité reste difficile à garantir, car les nœuds peuvent agir comme serveurs. En effet, un nœud peut ne pas utiliser des informations pour ne pas épuiser ses ressources d'énergie, de mémoire et de calcul, ce qui entraîne un comportement inapproprié. [2]

• La disponibilité :
Ce service fait référence à la capacité du réseau à garantir que ses services fonctionnent correctement en garantissant aux parties communicantes la présence et l'utilisation de 'information au moment souhaité. Les contraintes qui pèsent sur ces réseaux rendent cette propriété difficile à garantir dans les RCSF. En effet, un nœud peut ne pas utiliser des informations pour éviter l'épuisement de ses ressources d'énergie, de mémoire et de calcul, ce qui entraîne un mauvais comportement. [2]

• Le contrôle d'accès :
Ce service consiste à empêcher des éléments externes d'accéder au réseau et à attribuer des droits d'accès aux participants légitimes afin de distinguer les messages provenant des sources internes du réseau de ceux externes. [2]

• Non-répudiation :
Ce service crée, maintient, rend disponible et valide un élément de preuve concernant un événement ou une action revendiquée afin de résoudre des litiges sur la réalisation ou non de l'événement ou de l'action. C'est donc un mécanisme destiné à garantir l'impossibilité que la source ou la destination nient avoir reçu ou émis un message.[2]

1.4 Soucis de la sécurité informatique :

Les vulnérabilités, les menaces et les attaques sont trois problèmes de sécurité informatique . [4]

a) Les vulnérabilités : Les failles ou les défauts dans les spécifications, la conception, l'implémentation ou la configuration des systèmes informatiques peuvent entraîner une intrusion. [4].

b) Les menaces : Une menace est la possibilité qu'une propriété de sécurité soit violée intentionnellement ou accidentellement en exploitant une ou plusieurs vulnérabilités. [4].

c) Les attaques :

Une attaque est une action malveillante qui tente d'exploiter une faiblesse dans un système et de violer un ou plusieurs règles de sécurité. [4].

1.5 Classification des attaques informatiques :

Une attaque peut être classée selon son objectif, son point d'initiation ou la façon d'ardes ¬ sera la victime désirée.[5]

a) Selon l'objectif d'attaque : On trouve deux types d'attaques principaux : passives et actives.

• **Les attaques passives :** Une attaque passive tente d'apprendre ou d'utiliser l'information du système, mais n'affecte pas les ressources du système. C'est relativement difficile à détecter, mais plus facile à prévenir .

• **Les attaques actives :** Une attaque active cherche à modifier les ressources du système ou à perturber leur fonctionnement.

b) Selon le point d'initiation: On distingue deux types d'attaques pour ce critère de classification : attaques de l'intérieur et attaques de l'extérieur.

• **Les attaques de l'intérieur :**
Des utilisateurs légitimes d'un système lorsqu'ils agirent de manière illégale.

• **Les attaques de l'extérieur :** Venant de l'extérieur, généralement via Internet, en utilisant des méthodes telles que l'usurpation d'identité

c) Selon la façon d'adresser la victime : L existe deux façons pour adresser la victime soit d'une manière directe ou bien indirecte.

• **Les attaques directes :** Dans ce type d'attaque, l'attaquant adresse directement la victime sans passer par un intermédiaire.

• **Les attaques indirectes :** Dans ce type d'attaque, l'adversaire envoie ses paquets à une entité intermédiaire, qui les retourne à la victime.

d) Selon la façon d'adresser la victime :
Il existe deux façons pour adresser la victime soit d'une manière directe ou bien indirecte.

• **Les attaques directes :**
dans ce type d'attaque, l'intrus adresse ses paquets directement à la victime sans passer par un intermédiaire.

• **Les attaques indirectes :**
dans ce type d'attaque, l'adversaire envoie ses paquets vers une entité intermédiaire qui à son tour les retransmet vers la victime.

1.6 Buts des attaques :

Il existe plusieurs objectifs pour les attaques informatiques. [6]

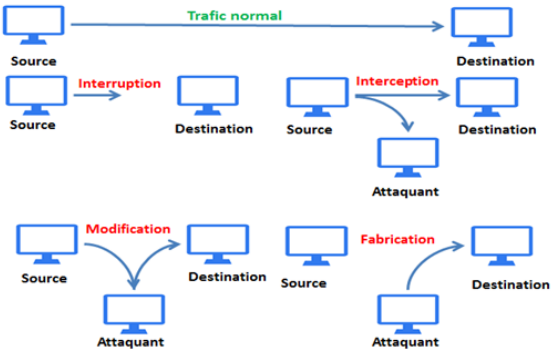

Figure 2: This frog was uploaded via the file-tree menu[16].

• **Interruption :**

•Un atout du système est détruit ou devient indisponible ou inutilisable.

• C'est une attaque portée à la disponibilité.

• Interception : Une tierce partie non autorisée obtient un accès à un atout. C'est une attaque portée à la confidentialité.

• Modification : vise l'intégrité des informations.

• Fabrication : C'est une attaque portée à l'authenticité.

1.7 Terminologie de la sécurité informatique :

L'ensemble des termes utilisés dans le domaine de la sécurité informatique peut se résumer ainsi .[7]

1) Vulnérabilité : Une vulnérabilité ou une faille est une faiblesse dans un système ou un logiciel qui permet à un attaquant de porter atteinte à la sécurité d'un système ou d'un logiciel.

2) Menace : Ces actions peuvent nuire à un système informatique. Les risques peuvent être causés par une variété d'actes et d'origines différentes.

3) Risque : Un risque désigne la probabilité d'un événement dommageable ainsi que les coûts qui en découlent. Le risque dépend également des montants des valeurs à protéger.

4) Attaque : Une attaque est l'utilisation d'une faille dans un système informatique à des fins qui ne sont pas connues par l'exploitant et qui peuvent être préjudiciables. Et parmi les diverses attaques actuelle, nous pouvons citer le sopor d'IP, le sniffions, les virus, les Ver, les attaques Dos et Man in the middle.

1.8 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 . [8]

1.8.1 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.

1.8.2 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.

1.8.3 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 [9]

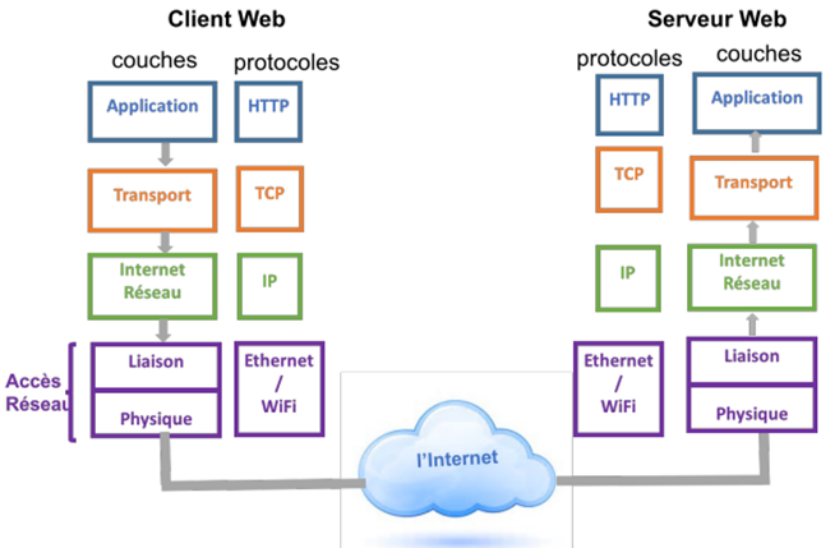

Figure 3: La représentation en couches des protocoles de sécurité[9] .

• 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 :[10.11]

— 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.[10]
Il existe aussi les algorithmes de hachage qui nous permettent d'obtenir une signature numérique à partir des données comme :

• IP Sec : il est créé pour faire face aux problèmes d'authentification et de confidentialité du protocole IP. IP Sec opère au niveau IP et il encapsule nativement tous ces protocoles (TCP, UDP, ICMP, etc.). IP Sec offre des services de contrôle d'accès, d'intégrité, d'authentification, de confidentialité de plus il fait face aux attaques de type paquets replay [10.11].

• 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 [10.11].

• SSH (Secure Shell) : il opère au niveau application et permet d'obtenir un interprète des commandes (Shell) à distance d'une manière sécurisé [10.11].

1.8.4 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 [10.11] :

• 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.[11]

• 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...

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 différentes classifications, il est donc nécessaire d'assurer la protection des systèmes 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 100C'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 02

2 Détection d'intrusion

Introduction

De nombreuses organisations passent aujourd'hui à la Cloud. Leur traitement beaucoup plus simple grâce an un accès à distance. Mais il est également confronté à plusieurs nouvelles menaces d'attaques. Ainsi, en raison de la nature dispersée, les risques d'intrusion sont more élevés avec l'aptitude des nouvelles attaques. Il existe une variété de méthodes d'analyse de sécurité, dont les systèmes de détection d'intrusion sont essentiels pour le Cloud. En utilise la détection des intrusions pour lutter contre les attaques malveillantes dans les réseaux virtuels du Cloud.

2.1 Définition d'un système de détection :

Les intrusions dans les systèmes informatiques peuvent être définies comme toutes les activités qui contreviennent à la Policy de sécurité du système. Les intrus ou attaquant essayant de obtenir un accès non autorisé aux informations, de causer des dommages ou de se livrer à, autres activités malveillantes. La détection d'intrusion le processus de surveillance des événements survenant dans un réseau ou sur un système informatique et de recherche de signes de menaces imminentes de violation des normes de sécurité or des politiques . Un système de détection d'intrusion un ensemble de composants matériels et logiciels conçus pour automatiser le processus de détection d'intrusion. Ce sort de logiciel de cyber sécurité lié à la sécurité du réseau comme un pare-feu. Il est considéré comme une deuxième protection pour détecter diverses activités malveillantes qui ne peuvent pas être détectées un pare-feu classique. Il surveille constamment les événements qui se produisent dans un système et détermine if ces événements sont des signes, attaque qui entravent ; utilisation légitime du système. [12] IDS signifie Intrusion Détection System. Il s'agit d'un équipement permettant de surveiller l'activité d'un réseau ou d'un hôte donné, afin de détecter toute tentative d'intrusion et éventuellement de réagir à cette tentative. [12]. Un appareil ou une application de détection ; intrusion (IDS) alerte ; administrateur en cas de faille de sécurité, de violation de règles or ; autres problèmes qui pourraient compromettre son réseau informatique. [12] Ils vérifient ; intégrité des fichiers en analysant les configurations des systèmes et leurs vulnérabilités. Ils sont capables de reconnaître des stratégies ; attaque traditionnelles. Pour ce faire, ils surveillent les comportements inhabituels et suivent les utilisateurs qui violent les règles.

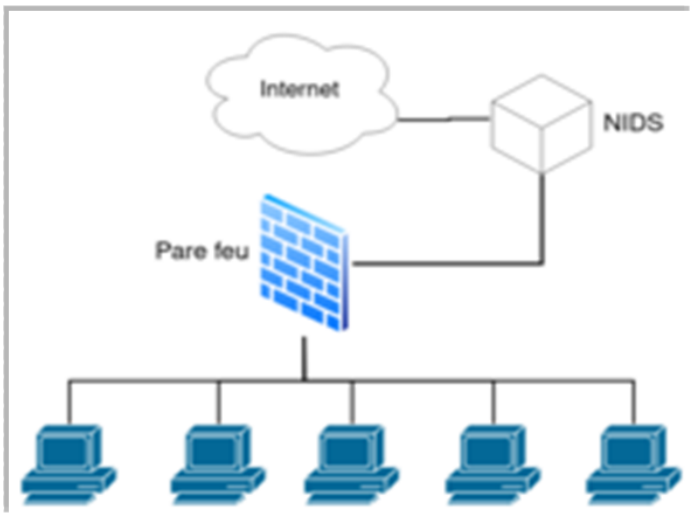

Figure 4: Système de détection d'intrusions[12]

2.2 Architecture d'un IDS:

Le groupe de travail IDWG (Intrusion Détections Worhing Group) de l'IETF a proposé un modèle fonctionnel d'IDS constitué de trois composants de base. La figure 2 illustre les Relationship entre ces trois composants. Un capteur (ou senseur) est chargé de collecter des informations sur l'évolution de l'état du système et de fournir une séquence des événements qui traduit cette évolution. Un analyseur détermine if un sous-ensemble des événements produits par un capteur est caractéristique d'une activité malveillante. Un gestionnaire recueillit les alertes générées par le capteur, les organise et les présente a opérateur. En cas d'échec, le gestionnaire est responsable de la réaction à adopter. Nous abordons ensuite chacun de ces trois éléments des actions intrusives dans une source de données. [13]

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

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

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

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

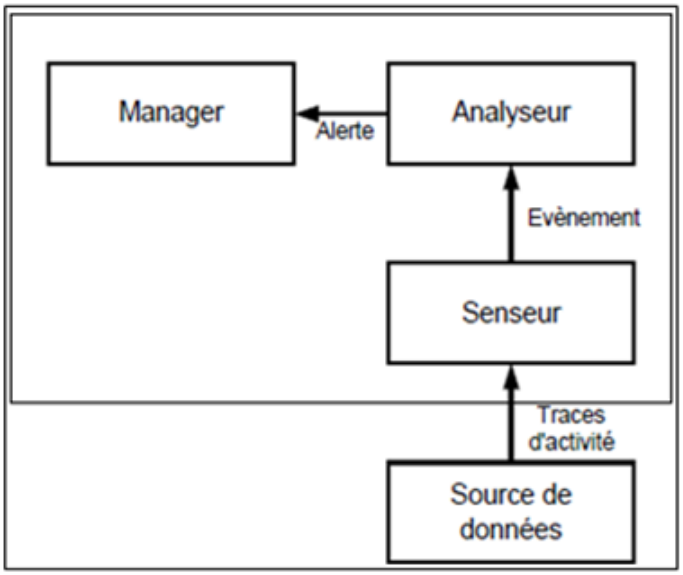

Figure 5: Modèle d'architecture IDS d'IDWG [13]

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

*Alerte : message formaté émis par un analyseur s'il trouve des activités intrusives dans une source de données. [13]

Il existe deux grandes catégories d'IDS, les plus connues sont les détections par signatures (reconnaissance de programme malveillant) et les détections par anomalies (détecter les écarts par rapport à un modèle représentant les bons comportements, cela est souvent associé a de l'apprentissage automatique) (figure 3).

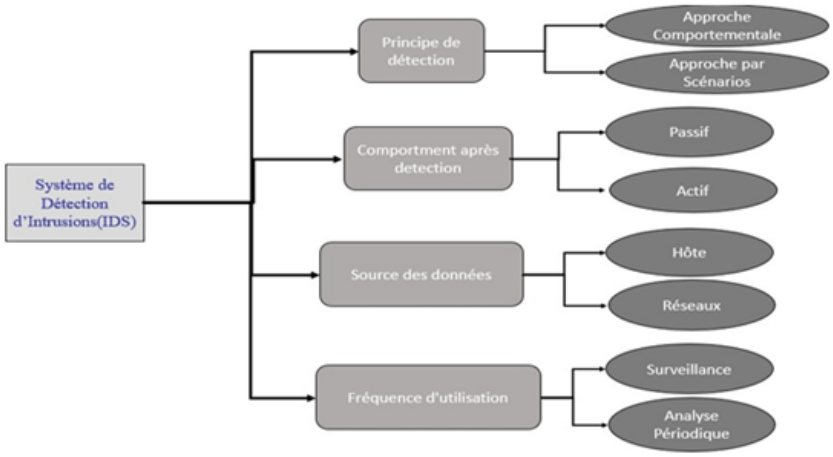

Figure 6: Taxonomie des systèmes de détection d'intrusions. [13]

• La méthode de détection utilisée (principe).
• Le comportement après détection
• La source des données à analyser.
• La fréquence de l'analyse

2.3 Les IDS hiérarchiques :

Les IDS hiérarchiques ont été proposées pour les réseaux sans fil multicouches. Cette architecture utilise le même principe que les IDS réparties en clusters pour la création des groupes avec chef de groupe, seulement les différents chefs de groupes ne coopèrent pas directement, mais forment un autre groupe d'un niveau supérieur avec un chef de groupe qui agit comme une station de base. Une station de base centrale effectue cette tâche de manière hiérarchique jusqu'à atteindre la couche la plus haute.
La détection d'intrusions dans les réseaux de capteurs ad hoc utilise fréquemment cette architecture. Dans la figure 4 nous présentons l'architecture hiérarchique proposée par [14]

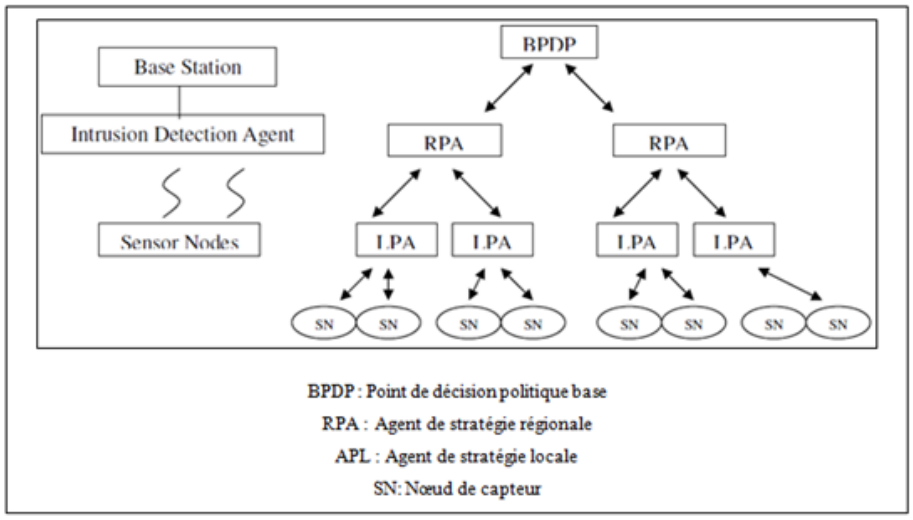

Figure 7: Architecture d'IDS Hiérarchique[14]

2.4 Les méthodes d'analyses des systèmes de détections d'intrusion (principe):

Les systèmes de détection d'intrusion courants utilisent généralement deux techniques de détection d'intrusion. Lorsque le système de détection d'intrusion utilise des informations sur le comportement normal des systèmes qu'il surveille, on le qualifie de comportement (détection par comportement). Lorsque le système de détection d'intrusion use des informations sur les attaques, on qualifie (détection par signature).

a) Détection par signature (scénario) :

La détection par signature adopte la politique suivante ce ne est pas dangereux, alors est normal Il donc essentiel de disposer d'une base de toutes les attaques connues.

La détection par signature (également appelée détection de mauvaise utilisation) analyse les informations recueillies et les compare à une base de données de signatures d'attaques connues (i.e., qui ont déjà été documentées), et toute activité associée est considérée comme une attaque (avec différents niveaux de sévérité). [15]

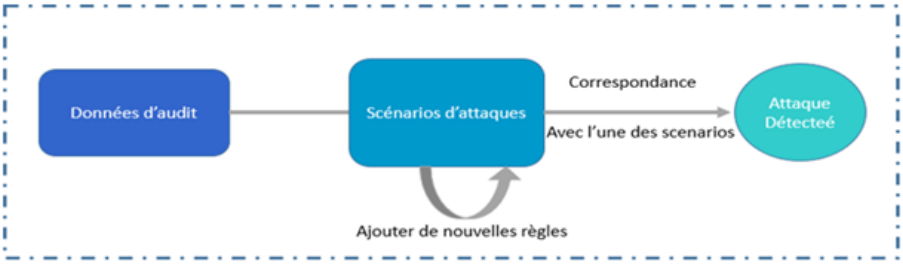

Figure 8: Fonctionnement d'un IDS par l'approche basée connaissance[15]

a) Détection par comportement :
La détection par comportement consiste à considérer comme hostile tout ce qui n'est pas normal, au sens où on cherchera plutôt à bien définir ce qui est un comportement normal sur le système pour pouvoir y opposer toute déviation, que l'on considérera comme étant une attaque : si ce n'est pas normal, alors c'est dangereux .

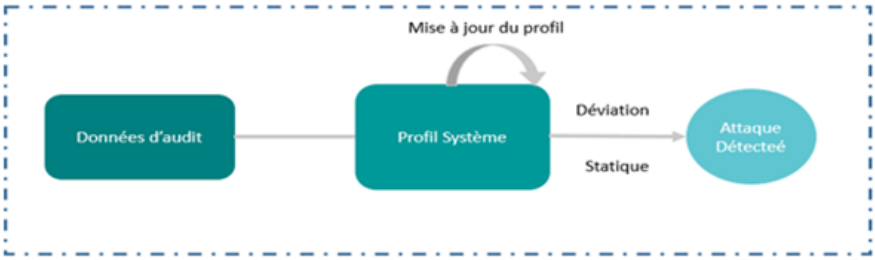

Figure 9: Fonctionnement d'un IDS par l'approche comportementale. [15]

2.5 Comportement après détection :

Nous pouvons également faire une distinction entre les IDS en se basant sur le type de réaction lorsqu'une attaque est détectée.

a. Passive :

En général, la majorité des systèmes de détection d'intrusions ne répondent qu'à l'intrusion de manière passive. En d'autres termes, lorsque d'une attaque est détectée, ils déclenchent une alarme et informent administrateur système par e-mail, message dans une console, voire même par SMS. L'opérateur est responsable de prendre les mesures nécessaires.

b. Active:

En plus de la notice à opérateur, les systèmes de détection d'intrusions peuvent prendre automatiquement des mesures pour stopper l'attaque en cours. Ils peuvent, par exemple, couper les connexions suspectes ou même reconfigurer le pare-feu pour qu'il refuse tout ce qui provient du site incriminé en cas l'attaque externe. Cependant, il semble que ce type de fonctionnalité automatique présente un risque, car il peut entraîner des interruptions de service causées par l'intrusion d'alarme. Par exemple, un attaquant peut tromper l'identification des intrus en utilisant des adresses de réseau locales qui seront alors considérées comme la source d'attaque par l'identification des intrus. Il spréférable offrir une réaction facultative an un opérateur humain.[16]

2.6 Source des données à analyser :

Les sources de données à analysions font partie des caractéristiques essentielles des systèmes de détection d'intrusions. Les journaux du système exploitation, les journaux des appas, les informations du réseau ou les alertes d'autres systèmes ; alarme peuvent fournir ces données. [16]

2.7 Fréquence de l'analyse :

Une autre caractéristique des systèmes de détection d'intrusions est leur fréquence d'utilisation, dans ce cas nous distinguons deux (2) types :

1. IDS online (continue) :

Ce sont des agents de détection d'intrusion qui analysent continuellement or en permanence les paquets réseau ou les fichiers d'audit afin de détecter une attaque au moment où elle est produite, c'est une détection en temps réel. Ce type d'identification des intrusions nécessite beaucoup de ressources systèmes cause il doit analysions à la volée tout ce qui se passe sur le système, ce qui le rend inadéquat en situations de ressources importantes telles comme les serveurs de messagerie.[17]

2. IDS offline (périodique) :

Ce type d'IDS fait l'analyse dans des durées périodiques afin de détecter des traces d'attaques au but de modéliser des signatures d'attaques pour la base du système, l'avantage de ce type est qu'il ne consomme pas beaucoup de ressources système.

2.7.1 Types des IDS :

1) IDS réseau :

L'IDS réseau (Network IDS ou NIDS) est sur un réseau isolé et ne voit qu'une copie du trafic, c'est-à-dire des paquets qui circulent. Le NIDS peut lever des alertes et ordonner des actions pour bloquer un flux un danger est détectée. Le NIDS est installé sur un réseau isolé et analysés la copie du trafic entre ses points d'entrées et les terminaux du réseau à surveiller. Il convient de noter qu'il est complètement passif et ne peut pas interagir avec le réseau surveillé.

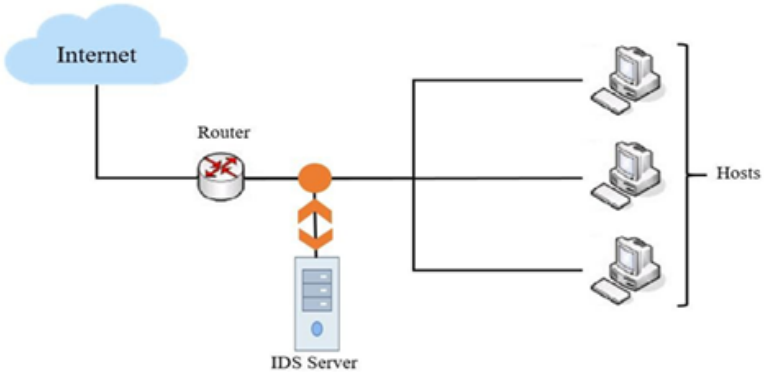

Figure 10: Exemple d'une architecture d'un NIDS[17]

2) IDS hôte :

Il y a ensuite les IDS hôte (Host IDS ou HIDS) ou IDS système. Les HIDS (Host Intrusion Détection System), surveillent l'état de la sécurité des hôtes selon différents critères :

- Activité de la machine (comme par exemple le nombre et listes de processus, le nombre d'utilisateurs, ressources consommées, etc.).

- Le second critère de surveillance est l'activité de l'utilisateur sur la machine : horaires et d'urée des connexions, commandes utilisées, programmes activés [18]

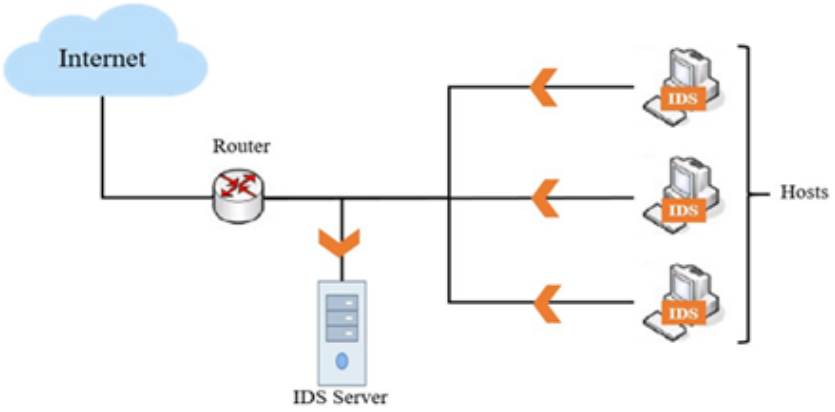

Figure 11: Exemple d'une architecture d'un HIDS [18]

3) IDS hybride :

Les IDS hybrides sont, généralement, utilisés dans un environnement décentralisé, ils permettent de réunir les informations de diverses sondes placées sur le réseau, et agissent comme NIDS et/ou HIDS suivant leurs emplacements.

Avantages d'IDS hybride :
Moins de faux positifs. Possibilité de réaction sur les analyseurs.

Inconvénients d'IDS hybride :
Taux élevé de faux positifs. [19]

2.7.2 Le modèle de base d'un système de détection d'intrusion :

Le système de détection d'intrusion se compose de plusieurs outils, ou chaque outil a sa propre tâche, dont l'objectif général est la détection d'intrusion au premier temps, puis d'informer l'opérateur ou le personnel informatique de la possibilité d'une intrusion dans le réseau. L'IDWG (Intrusion Detection Working Group) de l'IETF a proposé un modèle général pour la structure d'un système de détection d'intrusion qui englobe et standardise la structure. La figure suivante représente en détail les différentes parties de ce système. [20]

L'administrateur : est chargé de mettre en place la Policy de sécurité de l'organisation et de déployer et de configurer les différents éléments d'intrusion. Pour répondre aux besoin d'un système d'information, il prend en charge la déclaration prédéfinie des activités autorisées à se dérouler sur le réseau ou sur des hôtes particuliers. [20]

La source de données : Il existe de nombreux kinds de données provenant de diverses sources, telles que le réseau, le système, les applications et les alertes. Le système d'identification des intrusions ne limite pas les sources de données utilisées, mais il utilise les capteurs appropriés pour analyser les informations provenant de ces sources pour détecter les activités non autorisées or non désirées. [21]

Le capteur et l'analyseur : sont des composants importants du système. Au début, le capteur accède aux données brutes et collecte toutes les informations sur les activités survenant, puis les transfère à l'analyseur sous la forme d'événements. Ensuite, il va analyser ces événements pour identifier les activités non autorisées or indésirables ou les événements qui pourraient intéresser l'administrateur de sécurité. Le capteur et l'analyseur sont intégrés dans la plupart des IDS actuels. [21]

Le gestionnaire : De plus, c'est un élément crucial et le moyen par lequel l'opérateur gère les différentes parties du système. La configuration du capteur, la configuration de l'analyseur, la gestion des notifications d'événements, la consolidation des données et la gestion des rapports sont généralement des fonctions de gestionnaires.

La réponse : c'est les actions prises en réponse an un événement. elle peut être effectuée automatiquement par une entité dans l'architecture de l'identification des intrusions, tout comme elle peut être initiée par un humain. L'envoi d'une notice à l'opérateur est une réponse très courante. La journalisation de l'activité, l'enregistrement des données brutes (à partir de la source de données) qui ont caractérisé l'événement, l'arrêt du réseau ou de l'utilisateur ou de la session de l'application, la modification des contrôles d'accès réseau ou système sont d'autres réponses. [22]

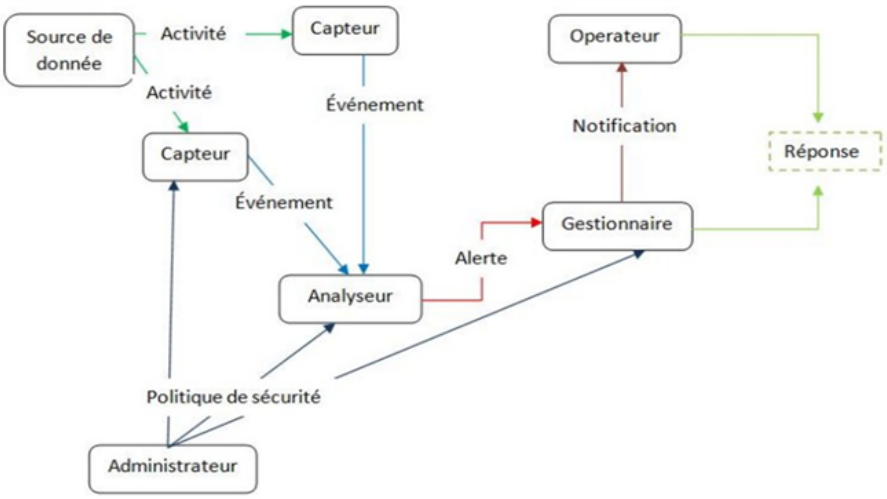

Figure 12: Un modèle fonctionnel du Système de détection d'intrusion proposé par l'IDWG [23]

2.7.3 Taxonomie des IDS :

Il existe différents kinds de technologie des systèmes de détection d'intrusions, chacun caractérisé par une variété d'approches de l'architecture du système, de l'environnement de déploiement, de la surveillance et des stratégies de détection. La littérature a proposé de nombreuses taxonomies d'identification des intrusions (IDS). Liao et al. Ont proposé une nouvelle perspective de ces taxonomies et ont utilisé une variété de critères basés sur des termes largement acceptés. Comme le montre la figure 13 quatre principaux critères de catégorisation l'illustrent :

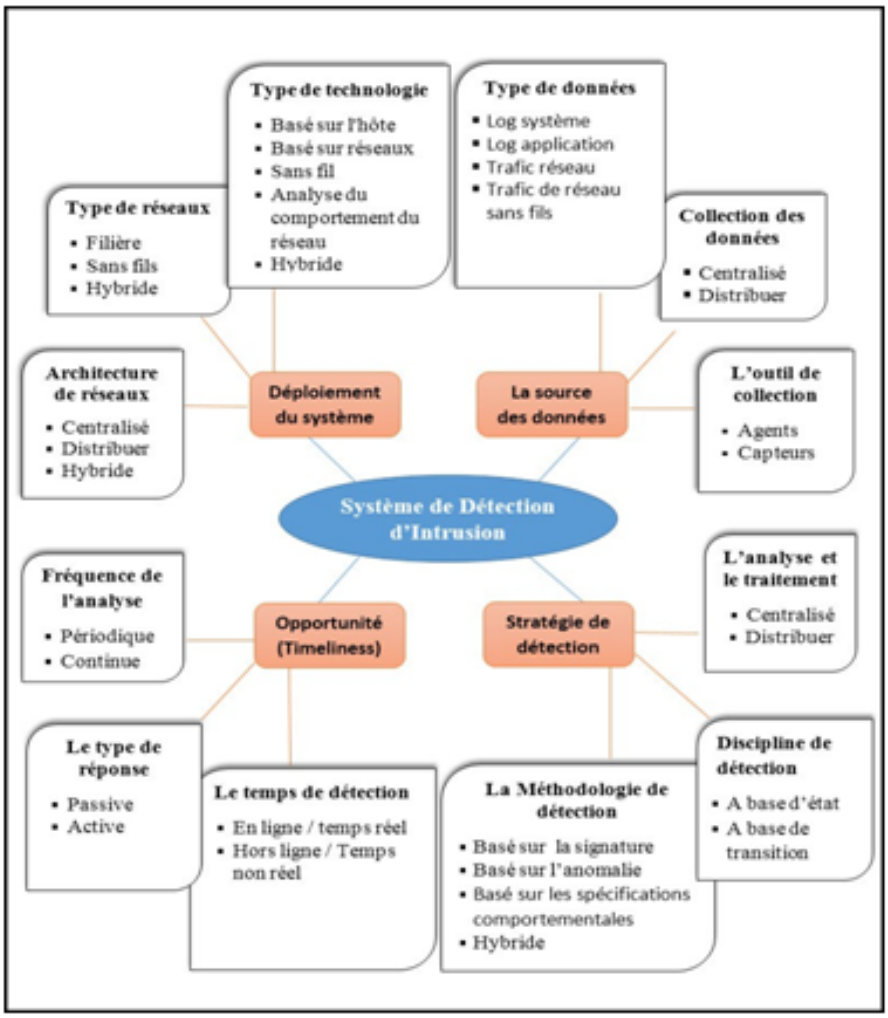

Figure 13: Taxonomie des systèmes de détection d'intrusion proposée par Liao et al. [23]

Conclusion

Dans ce chapitre, nous avons découvert que les IDS sont de plus en plus fiables, c'est pourquoi ils sont souvent inclus dans les solutions de sécurité contemporaines. Ils sont favorisés en raison de leurs avantages par rapport aux autres outils de sécurité. De plus, cela nous an appris que ces derniers sont essentiels aux fournisseurs de Cloud pour garantir leur sécurité Cloud. Le chapitre suivant présentera les différentes méthodes d'apprentissage automatique. Le second critère de surveillance est l'activité de l'utilisateur sur la machine horaires et durée des connexions, commandes utilisées, programmes activés, etc.

Chapitre 03

3 Expérimentations et résultas

Introduction

Dans ce chapitre, nous décrivons les procédures suivies pendant la création du modèle de détection d'intrusion incluant l'ensemble de données (Dataset) utilisé, scénario expérimental, résultats et explications. Différents algorithmes d'apprentissage automatique ont été utilisés sur l'ensemble de données d'intrusion (CICIDS2017).La comparaison entre les résultats obtenu nous a permet de sélectionner un meilleur algorithme pour notre modèle.

3.1 La Méthodologie :

Ce travail vise à proposer un modèle hybride de détection d'intrusion dans lequel l'ensemble de données collectées sera analysé à l'aide d'algorithmes d'apprentissage automatique. L'accent sera mis sur le prétraitement des données, la sélection des attributs, la normalisation, les algorithmes d'apprentissage automatique et l'ensemble de données CICIDS-2017. Les

performances de cinq algorithmes ML différents seront évaluées sur la base de mesures prédéfinies. Ensuite, en combinant chaque algorithme avec trois algorithmes de sélection d'attributs, un nouvel algorithme hybride est généré . Pour chaque classificateur(Random Forest (RF), Decisiontree ,

LogisticRegression (LR) , Naive Bayes et Stochastic Gradient Descent) , 3 algorithmes hybrides ont été construits en combinant les trois algorithmes de sélection d'attributs (Anova,IPCA et CHI2). L'ensemble de données utilisé est CICIDS-2017 utilisé. Toute l'expérimentation de ce module a été menée sous python (Jupyter notebook).

La figure 01 montre la méthodologieproposée que nous utilisions :

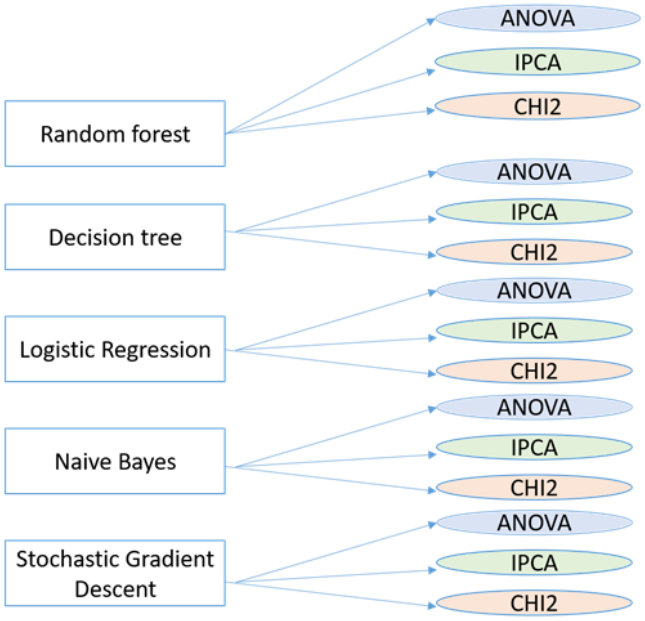

Figure 14: couplageclassificateur/sélection d'attributs

3.2 Pré-traitement:

Le prétraitement des données consiste à préparer les données d'une manière ou d'un format particulier, qui convient à la mise en œuvre effective des algorithmes d'apprentissage automatique souhaités. Le prétraitement des données est l'une des étapes majeures de l'apprentissage automatique. Dans le prétraitement, parfois, les données ne sont pas toujours présentes à un seul endroit, elles doivent donc être rassemblées à différents endroits et finalement converties en un format unique et approprié, suivies d'un nettoyage des données et d'une normalisation/discrétisation, avant d'y appliquer un algorithme d'apprentissage automatique différent. [24]

3.3 Normalisation:

Lorsqu'il s'agit d'attributs à plusieurs échelles, la normalisation est généralement essentielle ; sinon, l'efficacité d'un attribut significatif ou tout aussi important peut être diluée en raison des valeurs d'autres attributs à plus grande échelle. Simplement, lorsque de nombreuses caractéristiques existent mais que leurs valeurs sont à différentes échelles, un modèle de données insuffisant peutsurviennent lors de l'exécution de procédures d'apprentissage

automatique. De ce fait, elles sont normalisées afin de mettre toutes les qualités sur la même échelle.[25]

3.4 La sélection des attributs :

Après le prétraitement des données, l'ensemble de données propre est fourni comme entrée au processus de sélection des attributs. Ici, les attributsredondantssontdétachés de l'ensemble de données. Celaaméliorel'efficacité du modèled'apprentissageautomatique. Dans la prédiction, l'étape de sélection des attributs seconcentre sur la sélection du sous-ensemble, qui est un attributextrêmement unique. Dans la sélection des attributs, les attributspertinentes et uniquessontsélectionnées pour la construction du modèle. L'ensemble de données CICIDS-2017 contient 79 entités. Endehors de cela, après la miseenœuvre des algorithmes de sélection des attributs, moins d'attributs contribuent de manièresignificative au processus de prise de décision pour classer les attaques. C'est la raison pour laquelle la sélection des attributsestimplémentée sur le jeu de données, de sortequ'ellepeutréduire la taille du jeu de données avec les attributshautementcorrélées. Dansnotre travail nous avonsutilisé trois algorithmes de sélectiond'attributs :

1-Anova:

L'ANOVA, ou analyse de la variance, est une méthode statistique utilisée pour comparer les moyennes de trois groupes ou plus dans le but de déterminer s'il existe des différences significatives entre ces groupes. L'ANOVA permet de déterminer si les différences observées entre les moyennes des groupes sont dues à des variations réelles des groupes ou simplement à des variations aléatoires.[26]

2-Incremental PCA :

incrémentale (Incremental PCA en anglais) est une variante de l'analyse en composantes principales (ACP) qui permet de traiter efficacement de grands ensembles de données en les divisant en sous-ensembles plus petits. Contrairement à l'ACP traditionnelle qui nécessite de charger l'ensemble complet de données en mémoire, l'ACP incrémentale traite les données par lots (ou sous-ensembles) successifs, réduisant ainsi les exigences de mémoire.[27]

3-ChI2:

ChI2 (Chi-carré) est une mesure statistique utilisée pour évaluer la dépendance ou l'indépendance entre deux variables catégorielles. Il est principalement utilisé dans le cadre des tests d'indépendance ou des tests d'adéquation.[28]

3.5 Les algorithmes de classifications :

Cinq algorithmes d'apprentissage automatique différents ont été choisis pour cette étude :

1. Random Forest:

C'est un algorithme d'apprentissage automatique flexible et facile à apprendre. Cet algorithme appartient à la famille des algorithmes * d'ensemble *. Ces algorithmes génèrent de nombreux modèles connus sous le nom d'apprenants faibles. De plus, ces apprenants de la semaine sont combinés pour la prise de décision. Cette méthode augmente la précision de la prédiction. Cet algorithme suit le concept de construction d'une "forêt" d'"arbres" de décision aléatoire. Ces arbres sont utilisés pour la classification d'une nouvelle instance.

Chaque arbre est généré par un sous-ensemble de variables aléatoires à partir des variables prédictives du candidat. Un sous-ensemble aléatoire de données, en revanche, est généré à l'aide de bootstrap. Il est également possible d'utiliser cette approche algorithmique pour estimer la pertinence des variables. Cette approche algorithmique peut être utilisée pour la classification ainsi que pour la régression et utilise des hyperparamètres similaires à l'arbre de décision ou au classificateur d'ensachage. Dans la forêt aléatoire, un sous-ensemble aléatoire de caractéristiques fournit des résultats plus précis sur de grands ensembles de données [29]. En plus de cela, un grand nombre d'arbres aléatoires peuvent être créés en définissant un seuil aléatoire pour les caractéristiques globales plutôt qu'en explorant le seuil le plus précis. Le problème de sur-ajustement peut également être résolu en utilisant cette approche algorithmique.

2 . Decision-Tree:

Unarbre de décisionest un outild'aide à la décision, qui utilise le modèled'arbre pour la décision et leurs résultats possibles, y comprisl'utilité, les conséquences et le coût des ressources. C'est unefaçond'afficher un algorithme qui ne comprend que des instructions de contrôleconditionnelles. Ilmontre des règles, oùtous les domainesaffichent un test sur les attributs, la branchespécifie le rendement pour effectuer le test et quitte les classes d'affichage. Pour commencer à prendreunedécision, les décideurspeuventchoisir la meilleure alternative et passer de la racine à l'épisodeillustrant le classement du roman enfonction des informations les plus complètes. L'arbre de décisionestlargementutilisé par de nombreuxscientifiquesdans le domaine[29].

3.Naive Bayes:

L'algorithme Naive Bayes estbasé sur l'application deThéorème de Bayes. Dansce classificateur probabi liste, la présomptionest que les fonctions/prédicteurs sont indépendants, ce qui signifie que l'existence d'un attributspécifiquen'affecte pas l'autre. C'estpourquoi on l'appelle naïf.

Naive Bayes peutêtreutilisé pour la classification enutilisant des approximations basées sur la densité. Ilclasse les données, sur la base des information sprécédentes, ce qui estrapide et simple à appliquer. Commetoutes les fonctionnalitéssontindépendantes les unes des autres, ilprésenteégalement de nombreuxinconvénients.

Il existe trois types d'algorithmes Naive Bayes :

(i) Bernoulli Naïve Bayes :

- Si les caractéristiques de sortie sont de type booléenou de type binaire.

(ii) Gaussian Naïve Bayes :

- Danscetteméthode, les caractéristiquesdépendantesprennentunevaleur continue et suivent la distribution gaussienne.

(iii) Naïve Bayes multi-nominal :

- Il a le type d'informationnumérique et les donnéesnécessaires avec des caractéristiquesdiscrètes.[29]

4. Logistic Regression (LR) :

Dans la régressionlogistique, initialement, un ensemble de attributspondéréessontextraites de l'entrée. Ensuite, les caractéristiquesextraiteset les journauxsontcombinés de manièrelinéaire. Celaimplique que toutes les caractéristiquessontmultipliées par unpoids et ensuiteadditionnées. LR estunesorte de régression. Ce modèle de régressioninsère des donnéesdansunelogistiquefonction pour générer des prédictions sur la survenance d'un événement. Cetteapprocheutilise de nombreuses variables prévisionnelles, tout commed'autres types d'analyse de régression [30]. Ces variables peuven têtre catégoriellesou numériques[30].

5. Stochastic Gradient Descent:

La descente de gradient stochastique (SGD) estuneapproche simple maistrèsefficace pour ajuster les classificateurs et les régresseurslinéaires sous des fonctions de perteconvexestelles que les machines à vecteurs de support (linéaires) et la régressionlogistique. Mêmesi SGD existedepuislongtempsdans la communauté de l'apprentissageautomatique, il a récemmentreçuune attention considérabledans le contexte de l'apprentissage à grandeéchelle.

SGD a été appliqué avec succès à des problèmesd'apprentissageautomatique à grandeéchelleetclairseméssouventrencontrésdans la classification de texte et le traitement du langage naturel. Étantdonné que les donnéessontrares, les classificateurs de ce module s'adaptentfacilement aux problèmes avec plus de 10'5 exemples de formation et plus de 10'5 fonctionnalités.[31]

3.6 Matérielsetoutilslogiciels :

3.6.1 Matériels :

Nous avons utiliséun pc dont les caractéristiques sont :

CPU	I7
RAM	8 GB
HARD DISK	500 GB

Table01 :Spécificationsmatérielles .

3.6.2 Outils logiciels

Systèmed'exploitation : Windows 10

Langage de programmation :

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)

Figure 15: Logo de langage python

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.[32]

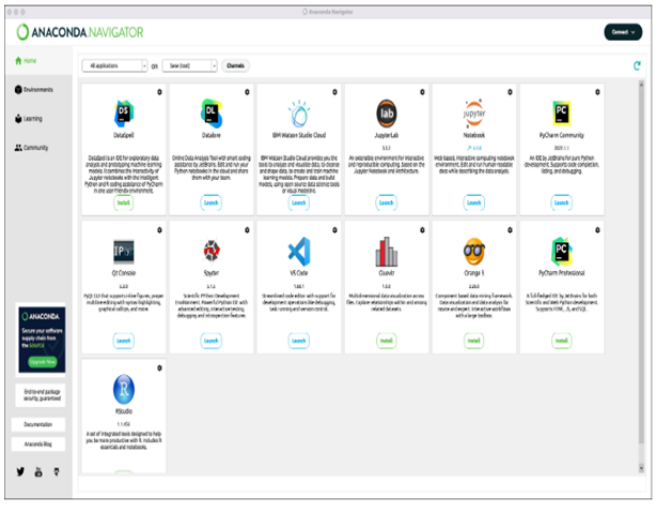

Figure 16: Interface Anaconda

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-deréaliserdescalepinsounotebooksquisontutilisésensciencedesdonnéespour 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.[33]

Figure 17: Interface Jupyter

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

1 . 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

2 . 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.

3 . 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.

3.8 Descriptions of CICIDS2017:

L'Institut canadien pour la cybersécurité a publié CICIDS-2017 en 2017 en tant qu'ensemble de données d'évaluation de la détection des intrusions. Il s'agit d'un ensemble de données réaliste avec un trafic bénin (normal) et malveillant (attaques fréquentes les plus récentes) collecté et enregistré dans des fichiers PCAP. Les données sont identiques aux données réelles réelles (PCAP).[34]

De plus, cet ensemble de données comprend également des flux de données, qui sont étiquetés par horodatage, adresse IP source, adresse IP de destination, port source, port de destination, protocoles utilisés et types d'agressions (fichiers CSV).

Il a été créé à l'aide de la technique du profil B, ce qui implique que l'ensemble de données a été créé à l'aide d'interactions humaines dans le réseau. CICIDS-2017 a été créé à partir des données de 25 utilisateurs qui ont utilisé divers protocoles tels que HTTP

3.8.1 Les Attaques dans CICIDS-2017

Plus de 2,8 millions de flux sont inclus dans l'ensemble de données CICIDS 2017. Cet ensemble de données tente également de couvrir une gamme variée et actualisée d'agressions que l'on peut trouver dans les réseaux d'aujourd'hui. Cet ensemble de données comprend huit attaques qui reflètent plusieurs catégories d'attaques : Web, force brute, DoS, DDoS, infiltration, heartbleed, botnet et portscan.

Classes	Number of samples
BENIGN	2273097
DoS	380699
PortScan	158930
BruteForce	1385
WebAttack	2180
Bot	1966
Infiltration	36

Table02 La distribution des classes dans CICIDS2017

	Flow Duration	Total Fwd Packets	Total Backward Packets	Total Length of Fwd Packets	Total Length of Bwd Packets	Fwd Packet Length Max	Fwd Packet Length Min	Fwd Packet Length Mean	Fwd Packet Length Std	Active Min	Idle Mean	Idle Std	Idle Max	Idle Min	Label
0	3	2	0	12	0	6	6	6.0	0.00000	0	0.0	0.0	0	0	BENIGN
1	109	1	1	6	6	6	6	6.0	0.00000	0	0.0	0.0	0	0	BENIGN
2	52	1	1	6	6	6	6	6.0	0.00000	0	0.0	0.0	0	0	BENIGN
3	34	1	1	6	6	6	6	6.0	0.00000	0	0.0	0.0	0	0	BENIGN
4	3	2	0	12	0	6	6	6.0	0.00000	0	0.0	0.0	0	0	BENIGN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2830738	32215	4	2	112	152	28	28	28.0	0.00000	0	0.0	0.0	0	0	BENIGN
2830739	324	2	2	84	362	42	42	42.0	0.00000	0	0.0	0.0	0	0	BENIGN
2830740	82	2	1	31	6	31	0	15.5	21.92031	0	0.0	0.0	0	0	BENIGN
2830741	1048635	6	2	192	256	32	32	32.0	0.00000	0	0.0	0.0	0	0	BENIGN
2830742	94939	4	2	188	226	47	47	47.0	0.00000	0	0.0	0.0	0	0	BENIGN

2830743 rows × 78 columns

Figure 18: Aperçu de l'ensemble de données CICIDS2017.

3.9 Evaluation des performances:

Nous évaluon snos algorithmes à l'aide de quatremétriques qui sontl'exactitude, la précision, le rappel et le score f1.Oùchacune de cesmesuresestextraite de la matrice de confusion comme suit. Si nous avons un tableau 04 commecelle ci-dessous,

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Table3 .1 Exemple de la matrice de confusion.

Accuracy= (TP + TN)/ (TP+TN+FP+FN)

Precision = TP/ (TP + FP) (4.10)

Recall= TP/(TP+FN) (4.11)

F- Measure= 2*((Precision * Recall)/ (Precision + Recall))

3.10 Discussion des résultats:

La partie expérimentations a été réalisée selon deux volets :

-Nous évaluons les algorithmes avec jeu de données avec deux classes (binaires).

-Nous évaluons les algorithmes avec jeu de données multi-classes.

3.11 . Classification binaire :

Dans cette première phase notre data set est normalisé de façon que le jeu de données contient seulement deux classes : Normal pour la classe BENIGN et anormal pour les autres classes. Le tableau 3.2 et la figure 19 montrent la répartition de notre jeu de données après modification :

normal	497938
anormal	320231
	818 169

Table3.2 : Répartition des classes binaires.

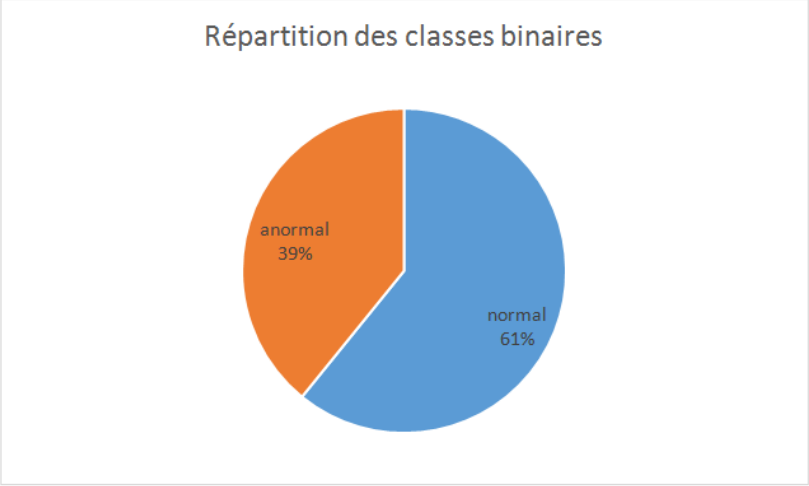

Figure 19: : Répartition des classes binaires.

3.12 Tests avant la sélection des attributs : :

Comme première phase de test nous avons appliqué les cinq algorithmes sur notre jeu de données binaires, les résultats obtenus sont inscrits dans le tableau 3.3 :

	Accuracy	Recall	Precision	F1-SCORE
RF	99,95	0.99	0.99	0.99
GNB	66,23	0.44	0.98	0.61
DT	98,37	0.98	0.99	0.98
LR	97,54	0.96	0.99	0.98
SGD	96,71	0.95	0.98	0.97

3.13 Tests après la sélection des attributs :

Dans la deuxième phase de test , nous avons appliqué 3 algorithmes de sélection des attributs et nous appliquons les algorithmes de machine learning sur l'ensemble de données obtenus. Le tableau 3.4 englobe les résultats obtenus, et les figures 16-17 montrent les résultats après chaque test.

Anova:

	Accuracy	Recall	Precision	F1-SCORE
RF	99.32	0.99	0.99	0.99
GNB	85.30	0.90	0.85	0.88
DT	97.12	0.99	0.95	0.97
LR	90.27	0.98	0.86	0.92
SGD	90.06	0.99	0.86	0.92

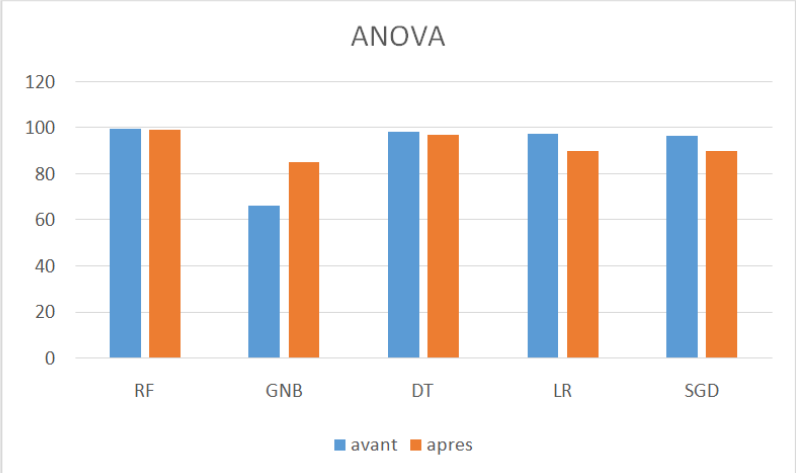

Figure 20: résultats d'applications d'Anova

Ipca:

	Accuracy	Recall	Precision	F1-SCORE
RF	93.42	0.99	0.90	0.94
GNB	69.61	0.54	0.92	0.68
DT	93.80	0.95	0.94	0.94
LR	94.41	0.93	0.97	0.95
SGD	94.11	0.92	0.97	0.95

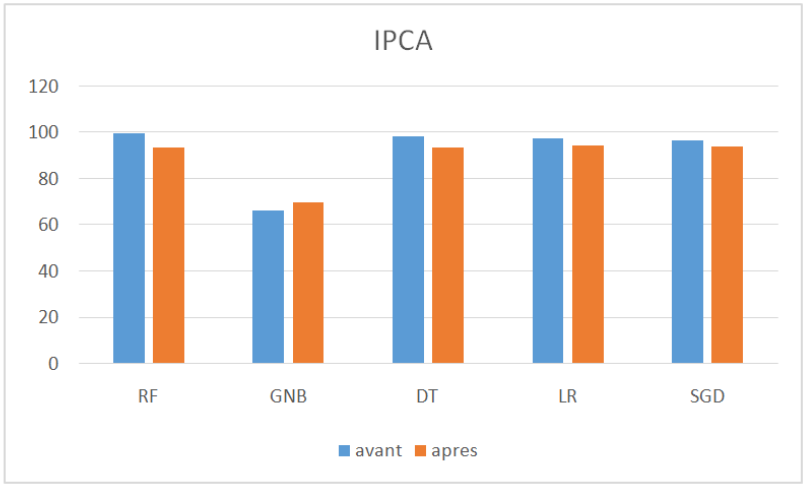

Figure 21: résultats d'applications d'IPCA

Chi2:

	Accuracy	Recall	Precision	F1-SCORE
RF	99,95	0.99	0.99	0.99
GNB	66,23	0.90	0.85	0.8
DT	98,37	0.99	0.95	0.97
LR	97,54	0.97	0.86	0.91
SGD	96,71	0.99	0.86	0.92

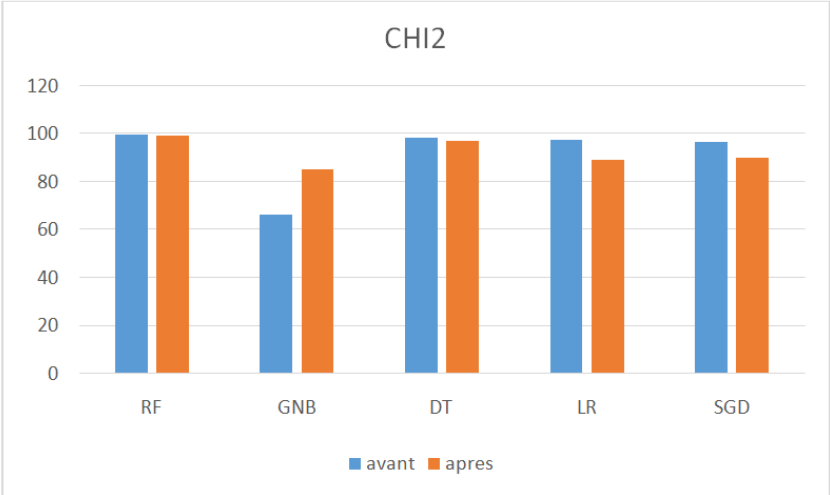

Figure 22: résultats d'applications de CHI2

Les résultats

Les résultats montrent que RandomForest appliqué avec CHI2 sur le jeu de données avec classification binaire , donne de meilleurs résultats.

3.14 CLASSIFICATION Multiclasses :

Dans cette phase d'évaluation, nous appliquons les cinq algorithmes sur le jeu de données multi-classes , dans une première phase sur le jeu de données entier , et dans la deuxième phase après l'application des algorithmes de sélection des attributs.

BENIGN	497898
DoS Hulk	171509
DDoS	128007
DoSGoldenEye	10279
DoSslowloris	5289
DoSSlowhttptest	5176
Heartbleed	11
	818169

Table3.5 :la répartition des attaques dans le jeu de données

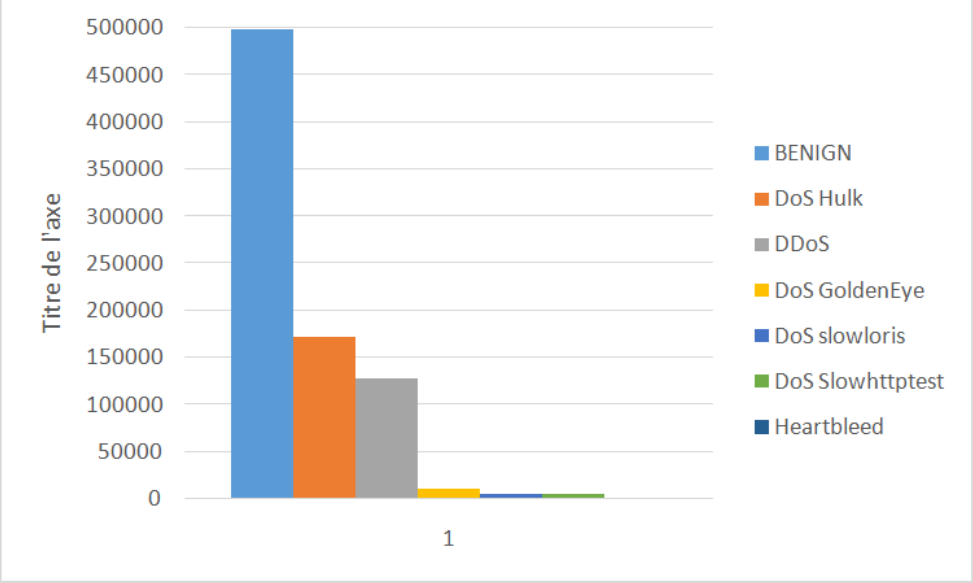

Figure 23: : la répartition des attaques dans le jeu de données

3.15 avant la sélection des attributs:

Nous avons appliqué l'ensemble des algorithmes sur le jeu de données multi-classes complet , les résultats obtenus sont montré dans le tableau 3.6 el la figure 24

	Accuracy	Recall	Precision	F1-SCORE
RF	99,94	0.99	0.99	0.99
GNB	90,59	0.91	0.65	0.70
DT	96,13	0.55	0.64	0.59
LR	97,72	0.78	0.81	0.79
SGD	95,98	0.69	0.79	0.73

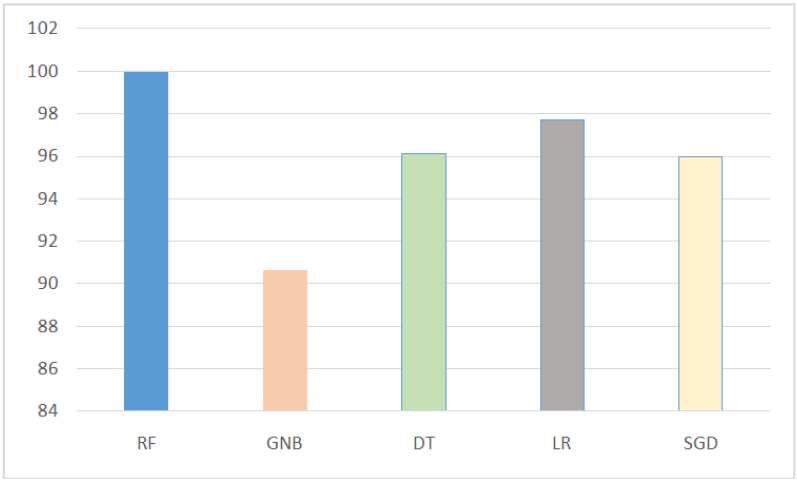

Figure 24: Accuracy après les tests sur le jeu de données complet.

Nous remarquons que Randomforest donne des meilleurs résultats avec une accuracy qui égale à 99.94 et GaussianNaive Bayes a donné 90.59 comme accuracy.

3.16 Tests après la sélection des attributs:

Dans cette deuxième phase, nous avons appliqué trois algorithmes de sélection d'attribut sur le jeu de données multi-classes , puis nous avons appliqué les cinq algorithmes de classifications. Les tableau 3.7 -3.9 et la figure 25-27 , montrent les résultats obtenus :

Anova

	Accuracy	Recall	Precision	F1-SCORE
RF	99,95	0.57	0.82	0.64
GNB	66,23	0.35	0.36	0.32
DT	98,37	0.48	0.55	0.48
LR	97,54	0.22	0.24	0.22
SGD	96,71	0.26	0.32	0.26

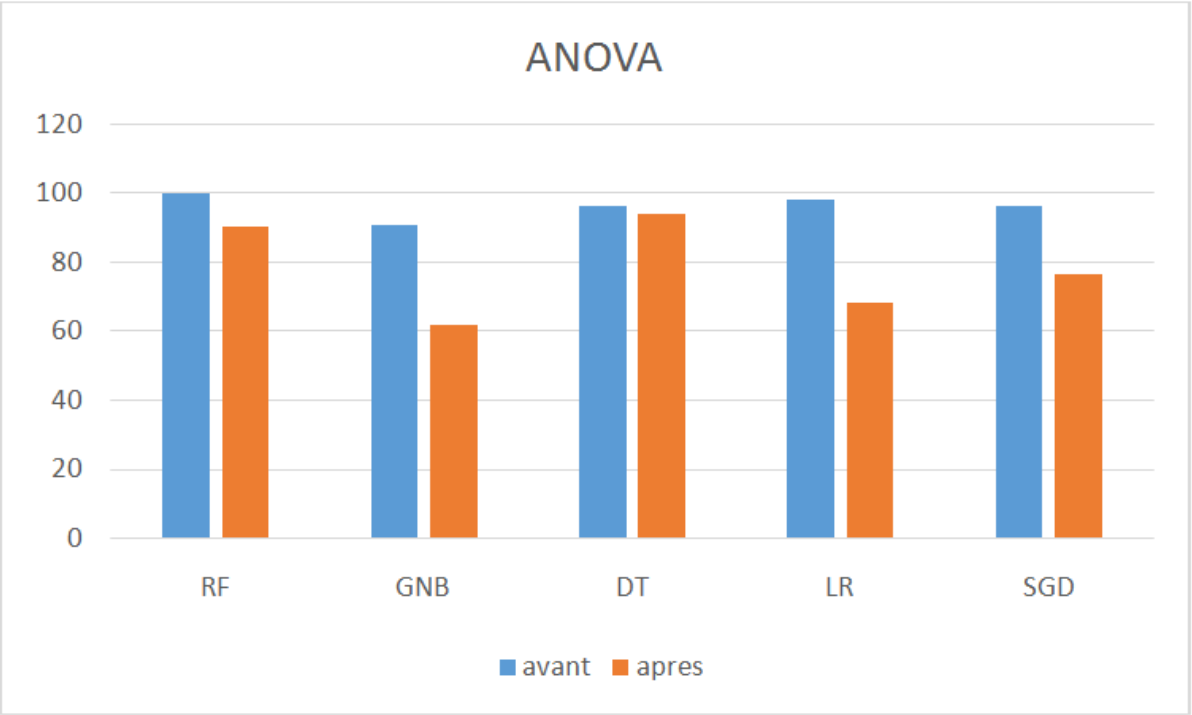

Figure 25: résultats d'applications d'ANOVA

Nous remarquons que Anova couplé avec DecisionTree donnent de meilleurs résultats 93.74 d'accuracy.

IPCA:

	Accuracy	Recall	Precision	F1-SCORE
RF	93.23	0.67	0.91	0.74
GNB	86.99	0.78	0.62	0.65
DT	88.34	0.39	0.36	0.37
LR	91.44	0.52	0.51	0.50
SGD	91.24	0.45	0.48	0.45

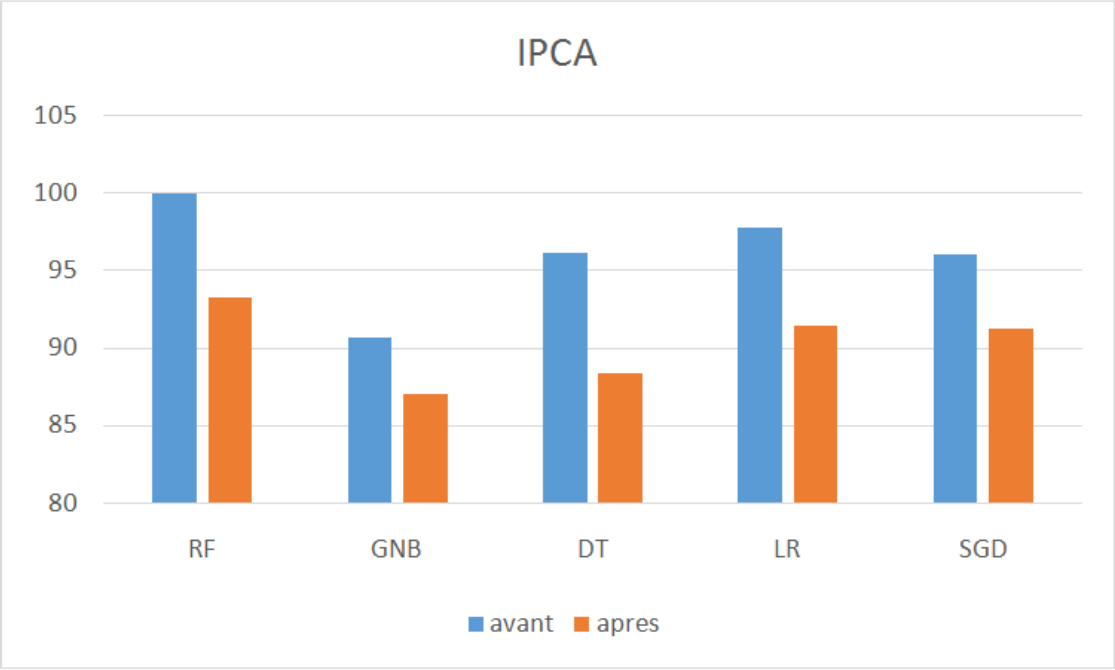

Figure 26: : accuracyapres application de IPCA

Selon les résultats montrés sur le graphe 3. , nous remarquons que IPCA couplé avec Random Forest donnent de meilleurs résultats 93.23 d'accuracy.

CHI2:

	Accuracy	Recall	Precision	F1-SCORE
RF	99,95	0.59	0.97	0.68
GNB	66,23	0.29	0.20	0.22
DT	98,37	0.42	0.40	0.41
LR	97,54	0.24	0.22	0.23
SGD	96,71	0.27	0.27	0.25

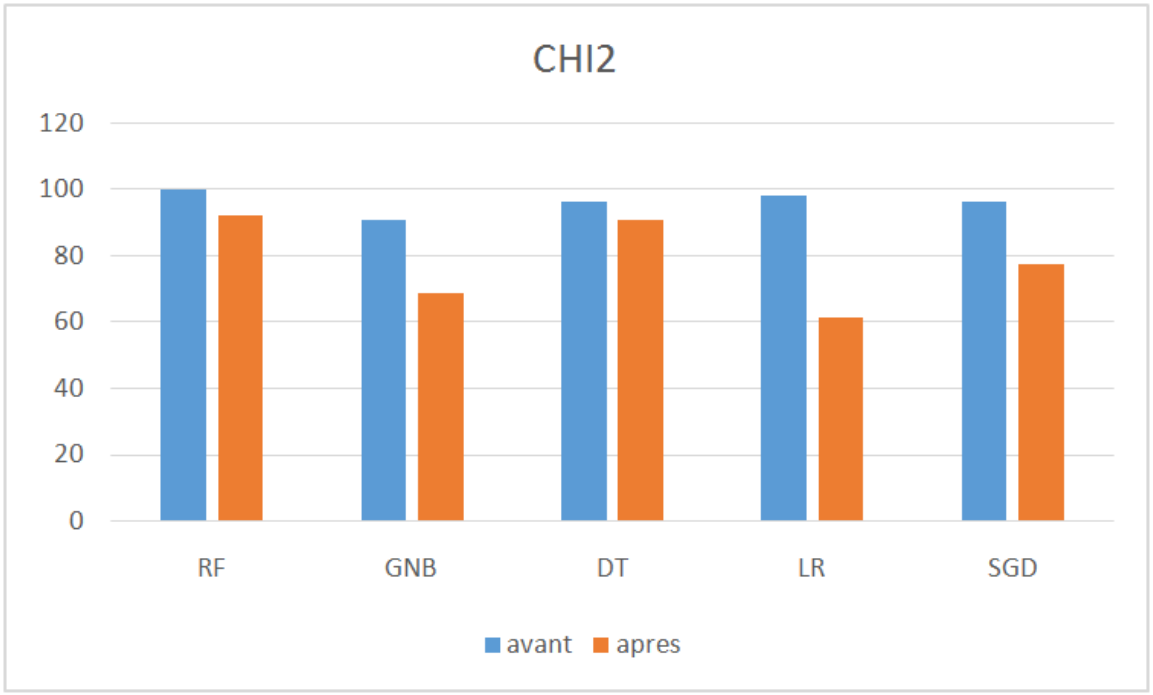

Figure 27: résultats d'applications de CHI2

Selonles résultatsmontrés sur le graphe3.11 , nous remarquons que CHI couplé avec Random-Forest donnent de meilleursrésultats 93.23 d'accuracy.

Conclusion

Notre étudesert à choisir la bonne combinaison entre l'algorithme de sélection d'attributs et le classificateur, nous avonsutilisé 3 classificateurAnova ,Ipca et CHI2. Et cinq algorithmes de classification qui sont : Random Forest , Gaussian Naive Bayes , Decision Tree , Logistic Regressionet Stochastic Gradient Descentappliqué sur le jeu de donnée CICIDS 2017.

Dans la classification binaire, Les résultatsmontrentque Random forest appliqué avec CHI2 sur, donnent de meilleursrésultats. Dans la classification multi-classe, Les résultatsontmontré que decision- tree avec Anova ontdonnéles meilleurs résultatsselonl'accuracyobtenue.

Doncconcevoir un modèle qui utilisecesdeuxalgorithmesselon le cas(binaireou multi-classes) contribuedans la détectiond'intrusion et donne de meilleursrésultats.

Conclusion générale

En raison de la popularité d'Internet et des réseaux locaux, les incidents d'intrusion dans les systèmes informatiques se multiplient. La propagation rapide des réseaux informatiques a modifié les possibilités de sécurité des réseaux. Cela a créé le besoin d'un système capable de détecter non seulement les menaces sur le réseau, mais également de s'appuyer sur des systèmes de prévention des intrusions. La détection de tels dangers fournit non seulement des informations sur l'évaluation des dommages, mais aide également à prévenir de futures attaques.

Les fonctions du système de détection d'intrusion sont réunies pour analyser toutes les violations de sécurité possibles et pour analyser les informations provenant de différentes zones au sein d'un ordinateur ou d'un réseau. Au cours des dix dernières années, la détection des intrusions et d'autres technologies de sécurité telles que la cryptographie, l'authentification et les pare-feu ont rapidement pris de l'importance.

De nombreux algorithmes d'apprentissage automatique ont été proposé, nous avons appliquécinq algorithmessur le jeu de données CICIDS-2017 après l'utilisation de trois algorithmes de sélection d'attribut pour la détection d'intrusion.

Nous optons pour la combinaison algorithmes randomforest et CHI2la classification binaire et decision- tree avec Anova pour la classification multi-classes.

Bibliographie

[1]: Renaud Tabary: tabary@enseirb.fr 2008-2009

[2]: Mr M. Daoui -Mme R. Aoudjit -Mme Ben Hocini-Mme M.Belkadi Système de détection d'intrusion hybride et hiérarchique pour les réseaux de capteur sans fil(2011-2012)

[3]: https://protectam.fr/iso-27001/clause-6-2-objectifs-securite-information-et-planification/

[4]: Cherfi Sarra Détection d'intrusions via des réseaux de neurones optimisés par des métaheuristiques(2019/2020)

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

[6]: Université Abou Bakr Belkaid Faculté des sciences Département d'Informatique Master 1 Réseaux et systèmes distribués 2019-2020

[7]: Mr. Fares KHELOUFI Mr. Yacine IKHLEF Proposition de solution de sécurité pour le Réseau local de l'hôpital d'Amizour (2015 - 2016) Proposition de solution de sécurité pour le Réseau local de l'hôpital d'Amizour Présenté par Mme Labraoui N.

[8]: Université Abou Bakr Belkaid Faculté des sciences Département d'Informatique Master 1 Réseaux et systèmes distribués 2019-2020

[9]:Cédric LIORENS. Tableaux de bord de la sécurité réseau. Editions Eyrolles, 2011.isbn : 2-212- 11973-9

[10]: Guillaume DESGEORGE. La sécurité des réseaux. 2000

[11]: Ahmed Chaouki LOKBANI. Le problème de sécurité par le Data Mining .Thèse de doct. Université Djillali Liabes - Sidi Bel Abbes, 2017

[12]: définition Hamouda Djallel : Un système de détection d'intrusion pour la cybersécurité Octobre 2020

[13]: définition BOUROUBA Hadjer et CHAOUCHE Ouidad : Optimisation des IDS Cloud Computing par les techniques de machines Learning 2019/2020.

[14]: Un appareil ou une application de détection ; intrusion (IDS) alerte : BOUROUBA Hadjer et CHAOUCHE Ouidad : Optimisation des IDS Cloud Computing par les techniques de machines Learning 2019/2020.

[**15**]**:** L'architecture IDS BENYETTOU Lahouari : DETECTION D'INTRUSIONS DANS LES RESEAUX AD HOC 06 juillet 2011 .

[**16**]**:** Alerte BOUROUBA Hadjer CHAOUCHE Ouidad :Optimisation des IDS du Cloud Computing par les techniques de machines Learning 07 / 09/ 2020 .

[**17**]**:** Les IDS hiérarchiques BENYETTOU Lahouari :DETECTION D'INTRUSIONS DANS LES RESEAUX AD HOC 06 juillet 2011

[**18**]**:** Détection par signature BOUROUBA Hadjer et CHAOUCHE Ouidad : Optimisation des IDS du Cloud Computing par les techniques de machines Learning 2019/2020

. [**19**]**:** active:BENYETTOU Lahouari :DETECTION D'INTRUSIONS DANS LES RESEAUX AD HOC 06 juillet 2011

[**20**]**:** Source des données à analyser BOUROUBA Hadjer et CHAOUCHE Ouidad : Optimisation des IDS du Cloud Computing par les techniques de machines Learning 2019/2020

[**21**]**:** IDS online (continue) : BENYETTOU Lahouari Octobre 2020

[**22**]**:** IDS hôte :BENYETTOU Lahouari :DETECTION D'INTRUSIONS DANS LES RESEAUX AD HOC 06 juillet 2011 .

[**23**]**:** Inconvénients d'IDS hybride Hamouda Djallel :Un système de détection d'intrusion pour la cybersécurité Octobre 2020

[**24**]**:** BERRI Nacira 06 juillet 2019 Utilisation des techniques de Deep Learning pour l'extraction des concepts à partir des documents textuels mémoire master Systèmes d'information, Optimisation et décision.

[**25**]**:** helenclu 18/03/2023 Principes de base de la normalisation des bases de données.

[26] : Julie Colas 06 April, 2020 Comprendre l'analyse de la variance (ANOVA) et le test F

[27] : https://scikit-learn.org/stable/

[28] : https://spss.espaceweb.usherbrooke.ca/test-de-chi-2/

[29] :HERIMANITRA RANAIVOSON JUIN 2019 CLASSIFICATION DE LA SÉVÉRITÉ DES BOGUES PAR L'UTILISATION DE MÉTRIQUES TIRÉES DE L'HISTORIQUE GIT MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES.

[30] : Alice Charguéraud Le taux de transformation en automobile : comparaison de différentes Mémoire présentéle : pour l'obtention du diplôme de Statisticien Mention Actuariat et l'admission à l'Institut des Actuaires méthodes d'apprentissage.