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

UNIVERSITE IBN KHALDOUN - TIARET

MEMOIRE

Présenté à :

FACULTÉ MATHEMATIQUES ET INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Réseaux et Télécommunication

Par :

DAHMANI Mostefa
OUARDANI Kamel

Sur le thème

Proposition d'un outil d'assistance pour la construction des systèmes de détection d'intrusion

Soutenu publiquement le 19/ 11/ 2020 à Tiaret devant le jury composé de :

Mr. DAOUD Mohamed Amine	Université Ibn Khaldoun	Encadreur
Dr. OUARED AEK	Université Ibn Khaldoun	Président
Mlle. HAMDANI ABDIA	Université Ibn Khaldoun	Examinatrice

2019-2020

REMERCIEMENTS

Louange à notre seigneur « الله » de nous avoir donné le savoir et l'opportunité de pouvoir poursuivre nos études, et qui nous a dotés de la merveilleuse faculté de raisonnement. Louange à notre créateur qui nous a incités à acquérir le savoir c'est à lui que nous adressons toutes nos gratitudes en premier lieu.

Nous tenons à adresser nos sincères remerciements à notre promoteur Mr. DAOUD Mohamed Amine pour son soutien,
ses conseils, sa présence et ses encouragements et les outils
qu'il nous a donnés, afin de mener à bien notre travail.
Nos profonds remerciements vont à l'ensemble des enseignants ayant contribué à notre formation.
Nos vifs remerciements vont à l'ensemble des membres du département informatique, tout particulièrement le Pr. DAHMANI.Y et Mr.SI ABDELHADI.A.

Nous tenons aussi à adresser nos vifs remerciements aux membres du jury, pour avoir accepté de juger notre travail.

Enfin, nous tenons à remercier
tous ceux qui ont participé de près
ou de loin à l'élaboration du présent travail.

DEDICACES

C'est avec l'aide et la grâce de DIEU que j'ai achevé cet humble travail que je dédie aux personnes à qui je tiens vraiment en reconnaissance de tout le soutien qu'ils m'ont apporté durant les moments difficiles :

À mes très chers parents.

À ma grand-mère, pour tout ce qu'elle a fait pour moi.

À ma femme.

À ma fille : LINA.

À mes frères.

À mes sœurs.

À mon binôme : Kamel et sa famille

À toute la promotion.

À la mémoire de nos martyres d'hier et d'aujourd'hui.

Mostefa

DEDICACES

A l'aide de DIEU tout puissant, qui trace le chemin de ma vie, j'ai pu arriver à réaliser ce modeste travail que je dédie :

A Mes très chers parents ;

Ma chère femme

A Mes frères et toute ma famille OUARDANI;

A Mes cousines, mon petit prince ahmed ;

la famille SAMI spécialement maya rihab ; raghed ;

la famille DAHMANI .BENMHEL.

MAMOUNI.BELAIDE

.SAFRANI.AICHOUBA.HANIFI.ABED....

A Mes amis, mes stagiaires au INSFEP de tissemsilet ;

Dont le mérite, les sacrifices et les qualités humaines m'ont permis de vivre ce jour.

OUARDANI KAMAL

Contenu

Listes des figures

Liste des tableaux

Liste des abréviations

IAAS : Infrastructure as a Service ;

PAAS: Platform as a Service;

SAAS: Software as a Service;

IDS : systèmes de détection d'intrusion ;

IA : intelligence artificielle ;

ML: Machine Learning ;

SVM : Machines à vecteurs de support.

Résumé :
Le problème posé par les utilisateurs de Cloud Computing c'est bien la sécurité de leurs données stockées, le système de détection des intrusions revient à jouer ce rôle. Le problème formulé peut être considéré comme un problème de classification, dont l'objectif est d'avoir une bonne optimisation dans lequel la fonction objective est de maximiser le taux de détection par type. Notre objectif consiste à faciliter le choix de la manière de sécuriser un Cloud Computing à l'aide d'un système de détection d'intrusions par la proposition d'un modèle pour avoir le meilleur résultat, et l'adaptabilité par des techniques de la machine Learning. Nous avons proposé un modèle IDS qui nous permet d'améliorer le taux de détection par rapport aux travaux précédents.
Le modèle proposé est une combinaison d'un algorithme d'apprentissage automatique supervisé et non supervisé afin de détecter les attaques. On s'attend à ce que ce système proposé qui utilise une combinaison du K-means et de SVM propose aux clients des Cloud Computing une vue sur les types des systèmes de détection d'intrusion existants.
Mots-clés:
Cloud, Systèmes de Détection d'Intrusion, Apprentissage Automatique, K-moyennes, Machines à vecteurs de support.

Abstract:
The problem posed by Cloud Computing users is the security of their stored data; the intrusion detection system comes down to playing this role. The formulated problem can be considered as a classification problem, the objective of which is to have a good optimization in which the objective function is to maximize the detection rate by type. Our objective is to facilitate the choice of the way to secure a Cloud Computing using an intrusion detection system by proposing a model to have the best result, and adaptability by techniques of the Machine Learning. We proposed an IDS model that allows us to improve the detection rate compared to previous work.
The proposed model is a combination of a supervised and unsupervised machine learning algorithm in order to detect attacks. The proposed system which uses a combination of K-means and SVM is expected to provide cloud computing customers with a view of the types of existing intrusion detection systems.

Keywords: Cloud, Intrusion Detection Systems, Machine Learning, K-means, Support vector machines.

ملخص:
المشكلة التي يطرحها مستخدمو الحوسبة السحابية هي أمن بياناتهم المخزنة ، ونظام كشف التطفل يلعب هذا الدور. يمكن اعتبار المشكلة المصاغة على أنها مشكلة تصنيف ، والهدف منها هو الحصول على تحسين جيد حيث تكون الوظيفة الموضوعية هي زيادة معدل الاكتشاف حسب النوع. هدفنا هو تسهيل اختيار طريقة تأمين الحوسبة السحابية باستخدام نظام كشف التطفل من خلال اقتراح نموذج للحصول على أفضل نتيجة ، والقدرة على التكيف من خلال تقنيات التعلم الالي. اقترحنا نموذجا IDS الذي يسمح لنا بتحسين معدل الكشف مقارنة بالأعمال السابقة.
النموذج المقترح عبارة عن مزيج من خوارزمية تعلم آلي خاضعة للإشراف وغير خاضعة للإشراف من أجل اكتشاف الهجمات. من المتوقع أن يوفر النظام المقترح الذي يستخدم مزيجًا من K-mean و SVM لعملاء الحوسبة السحابية عرضًا لأنواع أنظمة اكتشاف التسلل الحالية.
الكلمات الرئيسية: سحابة، أنظمة كشف التسلل ، التعلم الالي ، K-mean، SVM.

INTRODUCTION GÉNÉRALE

INTRODUCTION GÉNÉRALE

Contexte du travail

Vu que la technologie de l'Internet se développe d'une manière exponentielle depuis sa création, une nouvelle forme de TIC (Technologies de l'information et de la communication) a fait son apparition pour accroitre la productivité des entreprises et pour répondre à l'évolution des systèmes d'information en termes de ressources et d'espace, dont on trouve du Cloud Computing.

Au cours des dernières années, le Cloud Computing a pris un pas considérable comme un nouveau paradigme de l'informatique pour l'approvisionnement de divers services. Avec le Cloud Computing, les applications distribuées peuvent regrouper des services et des ressources informatiques évolutives à la demande. En dépit d'un nombre considérable de recherches sur le traitement de divers problèmes de Cloud Computing, la sécurité du service dans le Cloud est considérée comme un sujet très intéressant et important.

En effet, dans le contexte du Cloud Computing, nous devons revoir les défis de la sécurité du Cloud pour plusieurs facteurs.

La politique de sécurité définie, elle convient de la mettre en œuvre au sein du Cloud (la prévention des attaques et leur détection) en appliquant un contrôle a priori sur les actions effectuées au sein du système, s'assure que les utilisateurs ne peuvent pas enfreindre la politique.

Le déploiement, l'accessibilité à l'information et la disponibilité des services expose le Cloud à des activités malveillantes et à des attaques. Ce qui rend nécessaire la détection des failles de sécurité quand elles se produisent en utilisant des mécanismes de détection d'intrusion (IDS) et de tels mécanismes possèdent leurs propres limitations, qui peuvent porter sur des aspects théoriques des modèles sous-jacents ou sur leur implémentation.

Objectifs

Afin de qualifier un IDS, on s'intéresse à sa fiabilité, et à sa pertinence, autrement dit on s'intéresse en premier lieu à sa capacité à émettre une alerte pour toute violation de la politique de sécurité, en deuxième lieu à sa capacité aussi à n'emmètre une alerte. Les systèmes de détection d'intrusion (IDS) ont été largement utilisés pour détecter les comportements malveillants dans les communications réseau et les hôtes. Face à de nouveaux scénarios d'application dans le Cloud Computing, les approches IDS posent plusieurs problèmes puisque l'opérateur de l'IDS doit être l'utilisateur et non l'administrateur de l'infrastructure Cloud. L'extensibilité, la gestion efficace et la compatibilité avec le contexte basé sur la virtualisation doivent être introduites dans de nombreuses implémentations IDS existantes. De plus, les fournisseurs de Cloud doivent permettre de déployer et configurer IDS pour l'utilisateur.

Notre travail s'articule autour des IDS dans le Cloud Computing dont il consiste à faciliter le choix de la manière de sécuriser un Cloud Computing à l'aide d'un système de détection d'intrusions par la proposition d'un modèle pour avoir le meilleur résultat.

Problématique

Étant donné que le Cloud est un cas particulier des systèmes distribués, ces derniers couvrent plusieurs aspects tels que la sécurité, alors la sécurité des Cloud devient une tache très compliqué.

La problématique traiter est : Comment ce fait pour choisir un algorithme de sécurité et de prévention adéquat dans l'environnement Cloud ?

Structure de mémoire

Ce présent mémoire est structuré en trois chapitres :

Le premier chapitre est un chapitre descriptif sur le Cloud Computing, dont lequel nous avons les définitions essentielles du Cloud Computing à partir de différentes perspectives et des caractéristiques importantes sont présentées y compris son historique d'évolution. La conception de l'architecture du Cloud a été discutée, tel que les modèles de services et les modèles de déploiement, ainsi que les technologies clés derrière lui.

Le deuxième chapitre est consacré à présenter les systèmes de détection d'intrusion, le mode de fonctionnement, la classification des IDS et enfin la méthode de détection d'une intrusion.

Le troisième chapitre présent les différentes techniques la machine Learning et leurs fonctionnements et utilisation.

Le dernier chapitre est consacré présenté notre approche proposé et l'implémentation puis discutions des résultats de notre travail.

CHAPITRE I :

LE CLOUD COMPUTING

Chapitre I : LE CLOUD COMPUTING

I.1. Introduction

Le Cloud Computing a récemment émergé comme un paradigme pour la gestion et la prestation de services sur Internet. L'accroissement du Cloud Computing est en train de changer rapidement la vision des technologies de l'information. Il promet d'offrir des services fiables à travers les data-centers basés sur les technologies de calcul et de stockage virtuelles Les utilisateurs seront en mesure d'accéder à des applications et des données d'un nuage partout dans le monde à travers le modèle financier pay-as you-go [5].

Un nombre considérable d'étude tournent autour de la sécurité des Cloud [6]. Avec une telle progression rapide du Cloud Computing et sa présence dans la plupart des affaires des entreprises et les domaines de recherche scientifique, il devient crucial de comprendre tous les aspects de cette technologie, à savoir : ses concepts, ses architectures, ses modèles . . . etc. L'objectif de ce chapitre est de fournir une vue sur le Cloud Computing.

Le Cloud Computing est apparu pour certains comme une révolution et pour d'autres comme un simple terme de marketing, la question qui se pose c'est : d'où est-on parti pour arriver à cette informatique dans les nuages ? De quoi le Cloud Computing est-il constitué ? Quelles sont ses différentes architectures ?

I.2. Historique

La notion de Cloud fait référence à un nuage, tel qu'on a l'habitude de l'utiliser dans des schémas techniques lorsque l'on veut représenter Internet. Un réseau comme Internet est constitué d'une multitude de systèmes fournissant des services et des informations [7]. En fait le Cloud Computing n'est pas nouveau, il est exploité depuis les années 2000, les changements qui ont permis l'apparition du Cloud Computing sont nombreux [8, 9] :

— Internet fait son apparition dans les années 1960, mais sa diffusion dans le monde de l'entreprise n'a un réel impact que dans les années 1990.

— Depuis les années 70, il y'avait les mainframes qui offrent des services en local. Puis il invente la notion de < service bureau > pour qualifier une entreprise louant des lignes téléphoniques, répondeurs, services informatiques, etc. IBM lui-même était un < service bureau > en proposant la notion de < on-demande >.

— Pendant les années 80, il y'avait l'apparition du modèle client /serveur et dans lequel le client bénéficie des services offerts par le serveur.

— Le World Wide Web est né en 1991 et, en 1993, le navigateur Internet qui rend le <www> célèbre est créé par les américains dans sa première version. Aussi, les universitaires commencent à penser sérieusement à la façon de connecter ces machines, afin d'agréger les différentes ressources informatiques non exploitées (stockage et puissance de calcul). C'est alors que l'idée du réseau Grid a commencé à prendre forme.

— Lorsque la vitesse et la fiabilité des connexions s'est améliorée, un nouveau type d'entreprise est apparue : les fournisseurs d'applications en ligne (Application Service Provider, ASP). Le principe est que l'ASP achetait le matériel informatique et exécutait continuellement l'application moyennant le paiement d'un abonnement mensuel pour le client souhaitant accéder à Internet.

— La virtualisation a été la première pierre vers l'architecture du Cloud Computing. En effet, elle permet une gestion optimisée des ressources matérielles dans le but de pouvoir y exécuter plusieurs systèmes < virtuels > sur une seule ressource physique et fournir une couche supplémentaire d'abstraction du matériel. Les premiers travaux peuvent être attribués à IBM (les années 60, développement du premier hyperviseur).

— L'apparition de SOA (Service Orienté Architecture) a contribué à la naissance du Cloud Computing. En effet, les services de Cloud Computing sont exposés comme des services Web.

— Le Web 2.0 est une technologie émergente décrivant les tendances novatrices pour utiliser la technologie World Wide Web et la conception de sites Web qui vise à améliorer la créativité, le partage de l'information, la collaboration et la fonctionnalité du Web.

Ainsi le Cloud Computing s'est popularisé pendant ces dernières années. Il consiste à réorganiser l'utilisation des ressources informatique. La Figure I.1 représente l'évolution des systèmes informatiques depuis les années 1960 jusqu'à aujourd'hui.

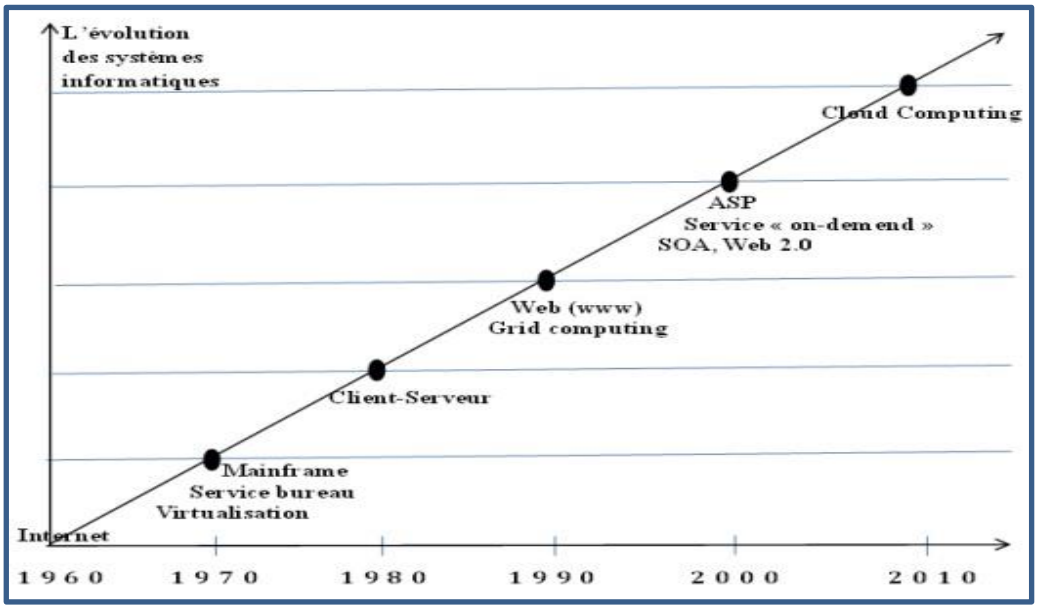


Figure 1. Historique de l'évolution des systèmes informatiques.

I.3. Définitions du Cloud Computing

Plusieurs définitions ont été proposées dans la littérature pour le concept Cloud Computing, ce dernier est encore en train de changer et ces définitions montrent comment le Cloud est vu aujourd'hui. Le Cloud Computing a été inventé comme un terme générique pour décrire une catégorie de services informatiques à la demande, initialement offerts par les fournisseurs commerciaux. Il représente un modèle sur lequel une infrastructure informatique est considérée comme un " Nuage ", à partir de laquelle les entreprises et les individus accèdent aux applications à partir de n'importe où dans le monde à la demande [10]. Il, offre le calcul, le stockage, et des applications " comme un service ". Il existe de nombreuses définitions du Cloud Computing, mais ils semblent tous se concentrer uniquement sur certains aspects de la technologie informatique [11].

Définition 1: *Une première définition du Cloud Computing a été proposé selon Lizhe.W et Gregor.L [12] comme suit : Un Cloud Computing est un ensemble de services permis par un réseau, fournissant une Qualité de Service (QoS) évolutive et garantie, normalement personnalisés, des plateformes informatiques peu coûteux à la demande, qui pourraient être accessibles d'une manière simple.*

Définition 2: *Selon le NIST (National Institute of Standards and Technology) [13] : Le Cloud Computing est un modèle qui permit, un accès pratique par un réseau de télécommunications à la demande, à des ressources informatiques partagés et configurables (réseaux, serveurs, stockage, applications et services) qui peuvent être provisionnés rapidement et libérés avec un effort minimale de gestion, et une interaction minimale du prestataire de services.*

Définition 3: *Buyya et al. [14] ont ajouté que pour atteindre le marché commercial, il est nécessaire de renforcer le rôle de la SLA (Service Level Agreement) entre les fournisseurs de services et les consommateurs de ces services. McFedries [15] décrit le data center (conçue comme une énorme collection de grappes) comme unité de base de l'offre Cloud, avec énormes quantités de puissance de calcul et des ressources de stockage. Le rôle de la virtualisation dans le Cloud est également souligné comme un élément clé [14]. En outre, le Cloud a été définis, comme du hardware et du software virtualisés, plus l'approvisionnement et les technologies de surveillance [16].*

Définition 4: *Une autre définition dit que le Cloud Computing est basée sur l'utilisation des ressources informatiques distribuées qui sont facilement allouées, migrées et éventuellement réattribuées sur la demande de l'utilisateur. À ce titre, le Cloud repose largement sur l'utilisation des technologies de virtualisation, il est capable d'offrir une quantité presque illimitée de ressources informatiques virtuelles. La virtualisation contrôle l'accès à des ressources physiques d'une manière transparente, il est possible d'offrir des ressources de calcul avec plein contrôle, de manière que les utilisateurs finaux peuvent les configurer comme administrateurs, sans aucune restriction [17].*

Définition 5: *Nous pouvons synthétiser, comme une définition globale, le Cloud Computing est un concept désignant des nouvelles pratiques et services, avec la mise en commun des ressources numériques et matérielles, des plateformes de développement, des serveurs de calcul et de stockage, des applications et des services distants, par l'intermédiaire d'un réseau, généralement Internet. Ces ressources se caractérisent par une flexibilité immédiate, une reconfiguration dynamique (scalabilité), une possibilité de paiement à la demande et une virtualisation des systèmes, afin d'accorder ainsi une forte abstraction des services du point de vue utilisateur.*

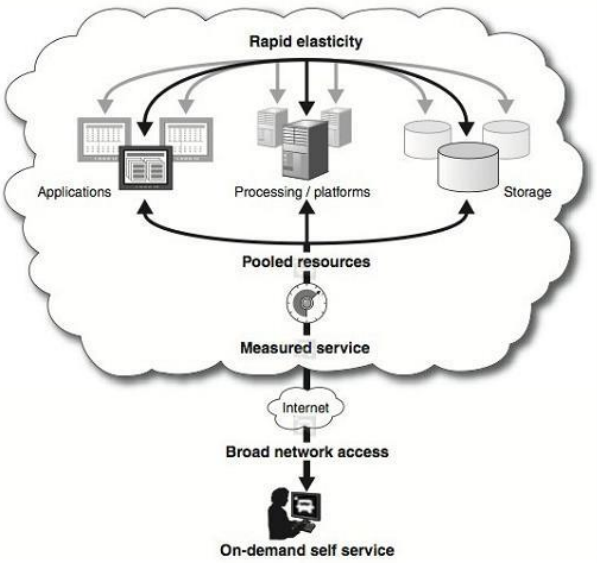


Figure 2. Le Cloud Computing [1].

I.4. Caractéristiques du Cloud

Le Cloud Computing offre d'énormes quantités d'énergie en termes de calcul et de stockage, tout en offrant une évolutivité et une élasticité améliorée. En outre, avec une efficacité et une économie évolutive, les services de Cloud Computing deviennent non seulement une solution moins coûteuse, mais beaucoup plus pour construire et déployer des services informatiques [18].

Le Cloud Computing se distingue des autres paradigmes informatiques par les aspects suivants [12, 19] :

— Haute disponibilité et fiabilité : La disponibilité des serveurs est élevée et plus fiable que les risques d'échec de l'infrastructure qui sont minimes.

— Evolutivité : L'évolutivité et la flexibilité sont les caractéristiques les plus importantes qui entraînent l'émergence du Cloud Computing. Les services et les plateformes informatiques offerts par le Cloud Computing pourraient être étendues à travers diverses préoccupations ; telles que les emplacements géographiques, la performance du matériel, et des configurations logicielles. Les plateformes informatiques doivent être souples pour s'adapter aux différents besoins d'un nombre potentiellement élevé d'utilisateurs.

— Service à la demande : Le Cloud Computing fournit des ressources et des services pour les utilisateurs à la demande. Les utilisateurs peuvent personnaliser et dépersonnaliser leurs environnements informatiques plus tard, par exemple, l'installation du logiciel, la configuration du réseau, les utilisateurs possèdent généralement des privilèges administratifs.

— Offre QoS garantie : Les environnements informatiques fournis par le Cloud Computing peuvent garantir la qualité de service pour les utilisateurs, par exemple, la performance du matériel comme la vitesse du CPU, bande passante d'E/S et la taille de la mémoire. Le Cloud Computing exprime le QoS par le traitement de Service Level Agreement (SLA) avec les utilisateurs.

— Services de Mode Pay-Per-Use : Les SLAs entre le fournisseur et l'utilisateur doivent être définies lors de l'offre de services en mode pay per use. Ceci peut être basé sur la complexité des services offerts. Application Programming Interfaces (API) peuvent être

offerts aux utilisateurs afin qu'ils puissent accéder à des services sur le Cloud en utilisant ces APIs.

— Multi-Partage : Avec le Cloud, il travaille dans un mode distribué et partagé, plusieurs utilisateurs et applications peuvent travailler plus efficacement avec des réductions de coûts en partageant une infrastructure commune.

— Système autonome : Le Cloud Computing est un système autonome et il est géré de manière transparente pour les utilisateurs. Le matériel, le logiciel et les données à l'intérieur de Cloud peuvent être reconfigurés automatiquement, orchestrées et consolidés pour présenter une image de plateforme unique, finalement rendu aux utilisateurs.

I.5. Architecture du Cloud

En fait, il n'y a pas une architecture fixe pour le Cloud Computing. De nombreux organisations et chercheurs ont définit l'architecture de Cloud Computing. L'ensemble du système peut être divisé en une pile de noyau de gestion. Dans la pile de noyau, il y a trois couches :

La couche ressources, la couche plateforme et la couche application [20]. La couche ressources : est la couche de l'infrastructure qui est composée de l'informatique physique et virtualisée, des ressources de stockage et de réseaux.

La couche plateforme : est la partie la plus complexe qui pourrait être divisée en plusieurs sous-couches ; par exemple. Un framework informatique gère la transaction d'envoi et/ou la planification des tâches. Une sous-couche de stockage offre un stockage illimité et la capacité de mise en cache.

La couche application : elle prenne en charge la même logique d'application générale comme avant dotée d'un système soit à la demande ou une gestion flexible. Basé sur les ressources sous-jacentes et des composants, l'application pourrait soutenir des opérations et des grandes gestions et distribuées d'énorme volume de données. Toutes les couches fournissent un service externe via un service Web ou d'autres interfaces ouvertes. Une des architectures du Cloud Computing est représentée sur la Figure 3 [20].

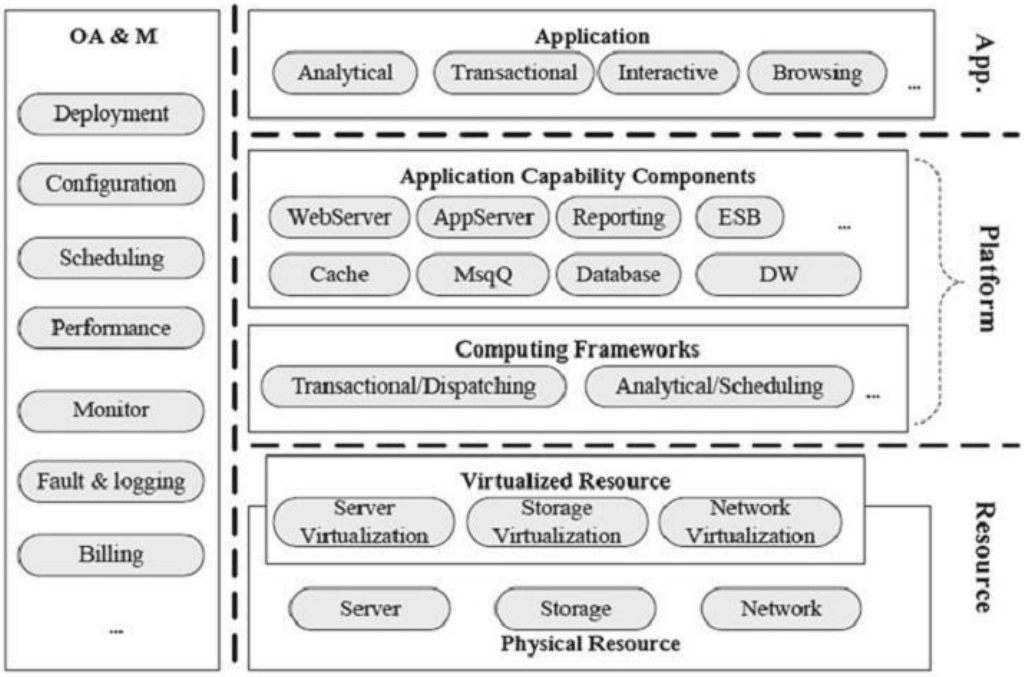


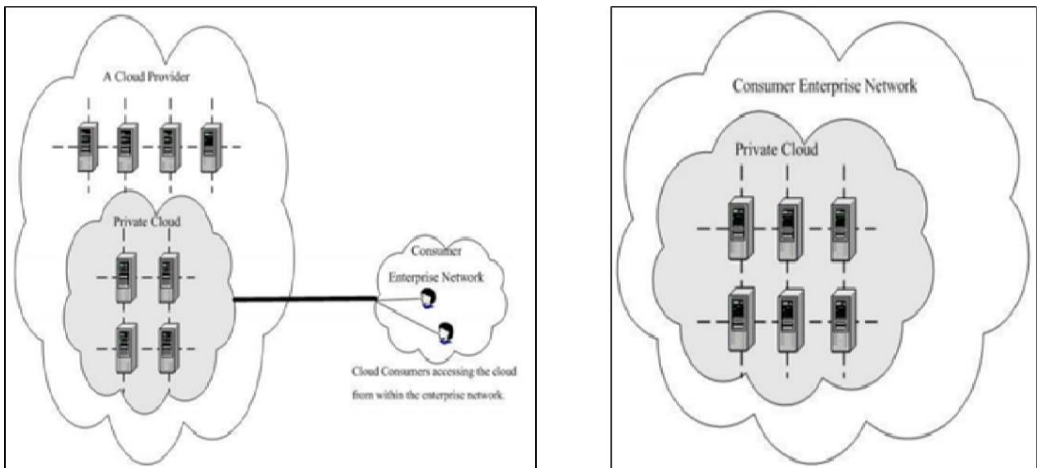
Figure 3. Architecture du Cloud.

I.6. Modèles de déploiement

Ce n'est pas évident d'avoir une simple description du Cloud et une claire explication de ces modèles de déploiement proposés. Généralement, on distingue quatre modèles de déploiement, un Cloud public, Cloud privé, un Cloud communautaire et Cloud hybride. Les différences sont basées sur la façon dont l'exclusivité des ressources informatiques est faite à un consommateur de Cloud [21]. Le Cloud public se réfère aux services fournis à des parties externes. Les entreprises construisent et exploitent le Cloud privé pour eux-mêmes. Le Cloud communautaire pour la communauté spécifique. Les Clouds Hybrides partagent des ressources entre le reste des autres Clouds par un réseau sécurisé [20].

I.6.1. Cloud Privé

L'infrastructure Cloud est utilisée uniquement pour une organisation, c.-à-d. l'exclusivité de l'accès et de l'usage de l'infrastructure et les ressources informatiques. Il peut être géré par l'organisation ou par un tiers et peut exister sur site ou bien externalisé. La figure 4.a et la figure 4.b présentes un Cloud privé externalisé et un Cloud sur site respectivement [21].



(a) Cloud privé externalisé. (b) Cloud privé sur site.

Figure 4. Le Cloud privé.

I.6.2. Cloud communautaire

L'infrastructure Cloud est partagée par plusieurs organisations et supporte une communauté spécifique qui a des préoccupations communes (par exemple, une mission, des exigences de sécurité, de la politique et des considérations de conformité). Il peut être géré par les organisations ou un tiers et peut exister sur site ou hors site (externalisé). La Figure 5 montre un Cloud communautaire, d'un conté, un ensemble d'organisations qui sont responsables à la fourniture des services du Cloud de l'autre côté ensemble d'organisations consommatrices de ces derniers [21].

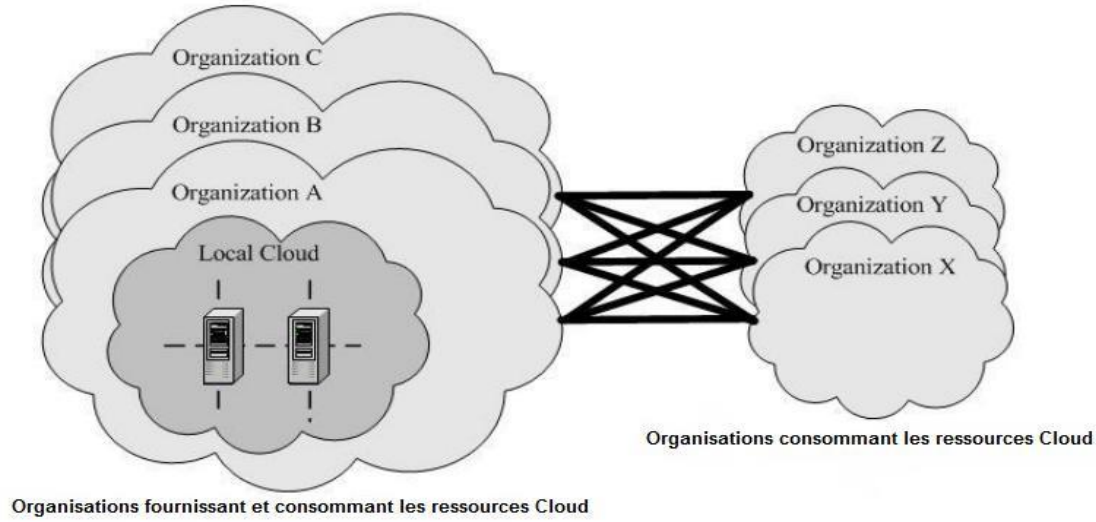


Figure 5. Le Cloud communautaire.

I.6.3. Cloud public

L'infrastructure Cloud est mise à la disposition de l'infrastructure et les ressources informatiques du grand public ou un grand groupe divers de l'industrie et elle est possédée par une organisation vendant des services Cloud. La Figure 6 présente une vue simple d'un Cloud public et ses clients [21].

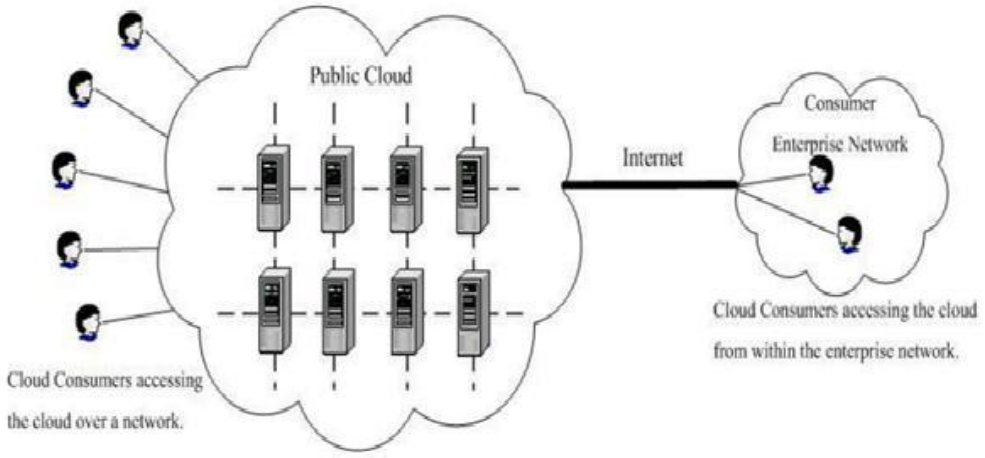


Figure 6. Le Cloud public.

I.6.4. Cloud hybride

L'infrastructure Cloud est une composition de deux ou plusieurs Clouds (privés, communautaires ou publics) qui restent des entités uniques, mais sont liés entre eux par la technologie standard ou propriétaire qui permet la portabilité des données et des applications. La Figure 7 présente une vue simple d'un Cloud hybride qui pourrait être construit avec un ensemble de Clouds dans les cinq variantes modèles de déploiement [21].

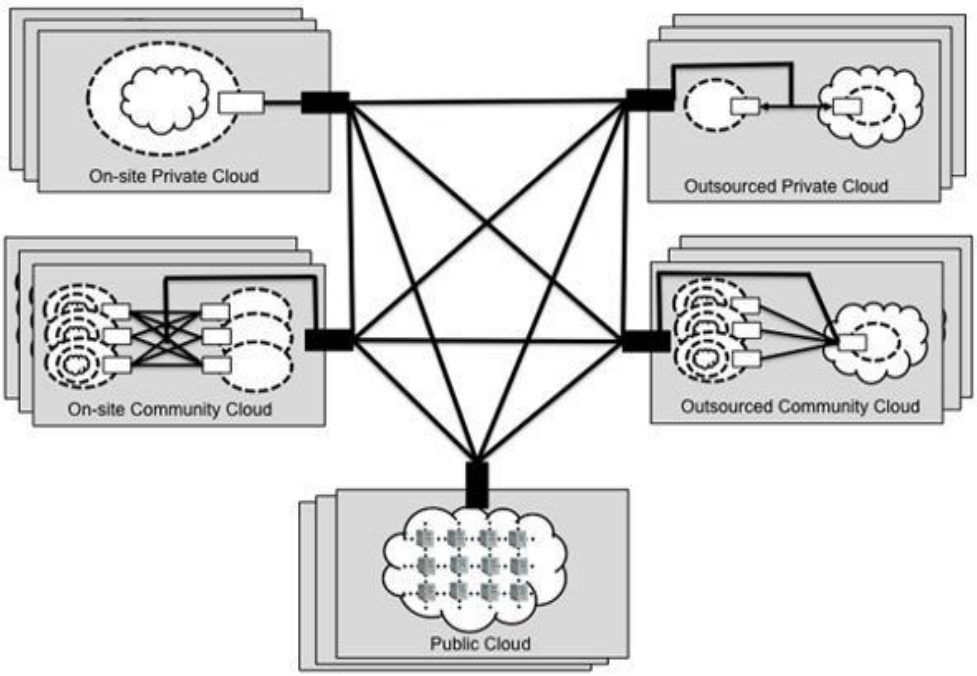


Figure 7. Le Cloud hybrid.

I.7. Modèles de service

De plus en plus, le Cloud Computing gagne de la popularité avec l'introduction de la nouvelle catégorie des services XaaS, progressivement il prend la place de nombreux types de ressources de calcul et de stockage utilisées aujourd'hui [22]. Le Cloud Computing fournit des services, qui sont mis à la disposition, sur l'abonnement dans un modèle pay-as-you-go aux consommateurs tel que : Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) et Software-as-a-Service (SaaS) [23, 24]. La Figure 8 ci-dessous montre les trois modèles de service du Cloud Computing [25].

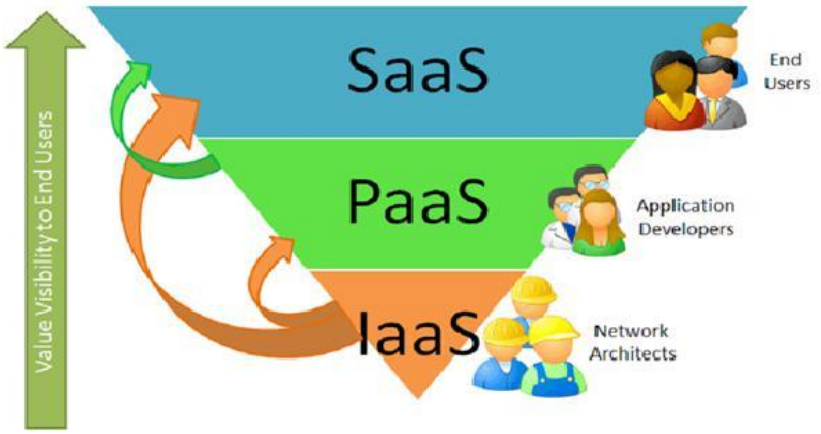


Figure 8. Les modèles de service du Cloud Computing.

I.7.1. Infrastructure-as-a-Service (IaaS)

Egalement appelé Hardware-as-a-Service (HaaS) a été inventé éventuellement en 2006. Comme le résultat des progrès rapides de la virtualisation matérielle, une automatisation informatique et utilisation du comptage et la tarification, les utilisateurs peuvent acheter du matériel informatique, ou même un Datacenter entier, comme un service d'abonnement pay-as-you-go [12]. Les solutions IaaS ou Haas offrent des infrastructures basées sur les ressources physiques ou virtuelles comme un produit aux clients. Ces ressources sont conformes aux exigences de l'utilisateur final en termes de mémoire, type de CPU et la puissance et la capacité de stockage. Les utilisateurs sont facturés sur une base pay-per-use. Ils doivent mettre en place leurs applications au-dessus de ces ressources qui sont gérées et hébergées dans des datacentres disposées par le fournisseur. Amazon EC2, Eucalyptus, Salesforce et Microsoft sont des exemples de Clouds qui proposent Infrastructure as a Service [8].

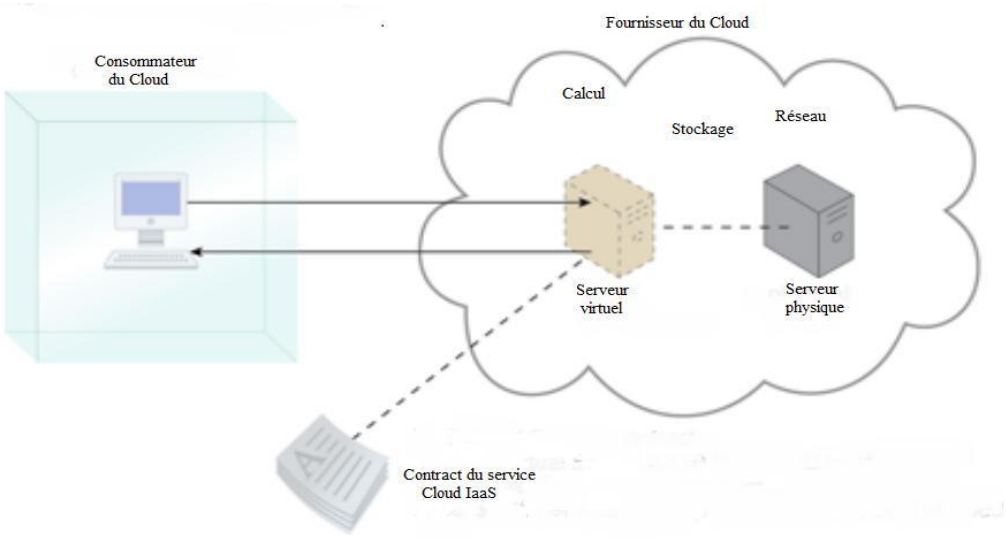


Figure 9. Infrastructure as a Service [1].

I.7.2. Platform-as-a-Service (PaaS)

Il s'agit des plateformes du Cloud, regroupant principalement les serveurs mutualisés et leurs systèmes d'exploitation. Les PaaS fournissent une application ou plateforme de développement dans lequel les utilisateurs peuvent créer leurs propres applications, qui s'exécuteront sur le Cloud [8]. Plus précisément, ils fournissent un Framework d'application et un ensemble d'API qui peuvent être utilisés par les développeurs d'applications dans le Cloud. Les solutions PaaS intègrent souvent une infrastructure au-dessus de ce que les applications seront exécutées. Ceci est le cas de Google AppEngine et Microsoft Azure [1, 24].

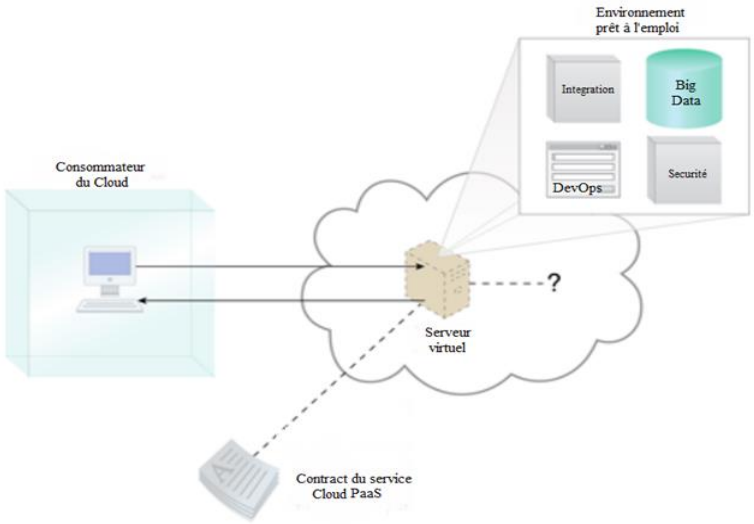


Figure 10. Plateforme as a Service [1].

I.7.3. Software-as-a-Service (SaaS)

Le Software ou l'application est hébergé en tant que service et est fournie aux clients à travers l'Internet. Ce modèle (SaaS) élimine le besoin d'installer et d'exécuter l'application sur les ordinateurs locaux des clients. Donc SaaS allège la charge au client et l'utilisation reste transparente pour les utilisateurs, qui ne se soucient ni de la plateforme, ni du matériel, qui sont mutualisés avec d'autre entreprises ce qui réduit le cout de l'achat, de la mise à jour et de la maintenance des logiciels [8, 24]. Les solutions SaaS sont à l'extrémité supérieure de la pile du Cloud Computing et elles fournissent aux utilisateurs finaux un service intégré comprenant le hardware, les plateformes de développement, et les applications. Les utilisateurs ne sont pas autorisés à personnaliser le service, mais ils auront un accès à une application spécifique hébergée dans le Cloud. Un exemple de service SaaS est les services fournis par Google pour l'automatisation de bureau, comme Google Mail, Google Documents et Google Calendar, qui sont livrés gratuitement aux utilisateurs de l'Internet et accusés pour des services de qualité professionnelle. Des exemples de solutions commerciales sont Dropbox, SalesForce.com et Clarizen.com, qui fournissent des services de CRM (Customer Relationship Management) et des gestions de projets en ligne, respectivement [1, 24].

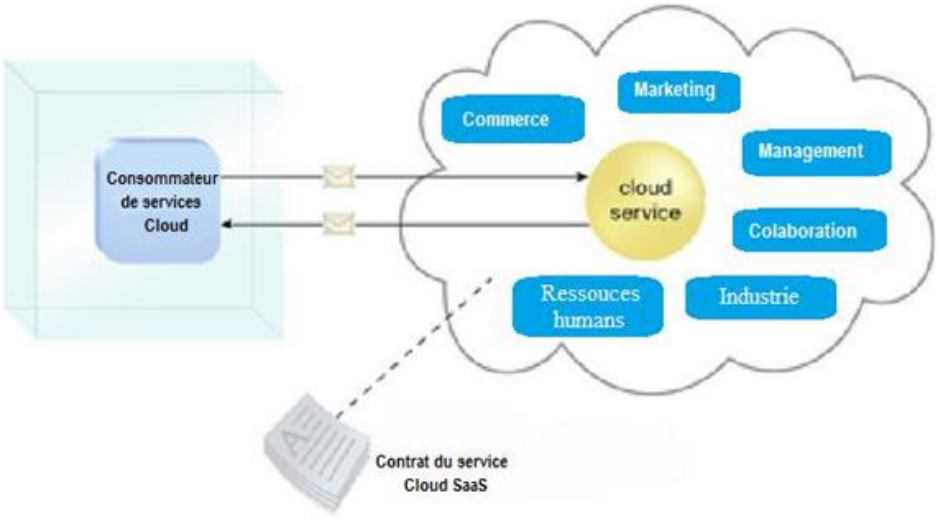


Figure 11. Software as a Service [1].

I.8. Avantages et inconvénients du Cloud Computing

Malgré l'avancé technique et technologique, mais il y'a toujours des avantages et des inconvénients pour chaque technologie émergente. Pour cela nous signons quelques avantages et inconvénients pour le Cloud Computing [26].

I.8.1. Avantages

Coût efficace : Les clients sont facturés pour seulement ce qu'ils utilisent et ils dépensent moins sur l'infrastructure où le coût global est réduit.

Améliorer l'accessibilité : Les clients pourront accéder à leurs applications et données de partout via Internet.

Sauvegarde et stockage : Les clients peuvent stocker une grande quantité de données et une copie est créée pour la sauvegarde par le serveur hôte.

Evolutivité et performances : Cette technologie est flexible pour répondre aux besoins immédiats des clients.

Minimiser les licences de nouveaux logiciels : Les clients ne doivent pas acheter un ensemble de logiciels ou de licences pour chaque système. Au lieu de cela, le client pourrait payer une redevance mesurée à une société de Cloud Computing.

I.8.2. Inconvénients

Questions techniques : Nécessité d'une vitesse élevée, fiable, fonctionnement haut débit de la connexion Internet pendant toute la durée de travail.

Forte dépendance : Il y a une plus grande dépendance sur les fournisseurs des services rendant les clients plus vulnérables.

Sécurité : Il y a toujours une sécurité potentielle et des risques de la confidentialité notamment en raison de la présence des données sensibles sur le Cloud.

Contrôle limité : Étant donné que les clients qui achètent les services souvent ont besoin de faire des compromis et d'adopter des solutions qui précisément ne répondent pas à leurs besoins.

I.9. Conclusion

Dans ce chapitre nous avons présenté un état de l'art sur le Cloud Computing, les concepts de base, notamment son historique d'évolution, son architecture, ses caractéristiques, ses modèles de services, les modèles de déploiements et enfin les avantages et les inconvénients. La technologie Cloud est la partie la plus récente de l'informatique distribuée. Il donne des promesses fortes, comme la haute disponibilité, l'allocation dynamique des ressources, le paiement que pour les ressources utilisées et l'offre des services. Le chapitre suivant sera consacré pour les Systèmes de Détection d'Intrusion.

CHAPITRE II :

SYSTEME DE DETECTION D‘INTRUSION

Chapitre II : SYSTEME DE DETECTION D'INTRUSION

II.1 Introduction

Le Cloud Computing, comme tout système informatique réparti, est continuellement exposé à de nombreuses menaces. Ainsi, sa sécurité est aujourd'hui une préoccupation très important par les fournisseurs et des utilisateurs du Cloud. Pour se prémunir des attaques, des mécanismes de sécurité sont déployés afin de protéger les données hébergée et partagées dans les infrastructures virtuelles. Les pare-feu sont responsables du filtrage de paquets afin de contrôler l'accès réseau. Néanmoins, le pare-feu ne procure pas une sécurité complète contre les attaques, car ils s'intéressent à l'attaque elle-même sans se soucie des attaquants. Les systèmes de détection d'intrusions viennent pour compléter le travail du pare-feu, en s'intéressant au comportement de l'attaquant.

Les systèmes de détection d'intrusions (IDS) détectent les attaques survenant sur le réseau et les systèmes. L'objectif des administrateurs de sécurité est de prévenir et de détecter les attaques sans perturber le bon fonctionnement du Cloud.

Nous consacrons ce deuxième chapitre à la sécurité, aux mécanismes de sécurité, ainsi qu'aux systèmes de détection d'intrusions.

II.2 Sécurité et mécanismes de sécurité

II.2.1 Définition de la sécurité

La sécurité informatique est l'utilisation de la technologie, des politiques et de l'éducation des personnes pour assurer la confidentialité, l'intégrité et l'accessibilité des données durant leur stockage, leur traitement et leur transmission. La protection des données doit dépendre du système à protéger [27].

II.2.2 Objectifs de la sécurité

La sécurité d'un système informatique a pour mission la protection des informations et des ressources contre toute dévaluation, modification ou destruction.

Les objectifs pris en confédération dans la sécurité sont les suivant [27] :

1. ***La confidentialité***

Permet de garder les informations secrètes de tous sauf des personnes autorisées à les consulter.

2. ***L'authentification***

Permet la confirmation de l'identité d'une entité avant de lui donner l'accès à une ressource.

3. ***L'intégrité des informations***

Permet d'assurer que les informations n'ont pas été altérées par des personnes qui ne sont pas autorisées.

4. ***La disponibilité***

Permet de garantir l'accès à un service ou une donnée.

5. ***Non répudiation***

Permet d'empêcher le démenti d'engagement ou de l'action précédente.

II.2.3 Mécanismes de défense

Il existe plusieurs mécanismes ou technologies de défense pour faire face aux attaques, nous allons dans ce qui suit citer les principales technologies [28] :

1. ***Authentification***

Permet de vérifier la véracité des utilisateurs, du réseau et des documents.

2. ***Cryptographie***

Permet la confidentialité des informations et la signature électronique.

3. ***Contrôle d'accès***

Permet de vérifie les droits d'accès d'un acteur aux données.

4. ***Antivirus***

C'est un logiciel censé de protéger l'ordinateur contre les logiciels (ou fichiers potentiellement exécutables) néfastes. L'antivirus ne protège pas contre un intrus qui emploie un logiciel légitime, ni contre un utilisateur légitime qui accède à une ressource alors qu'il n'est pas autorisé à le faire .

5. ***Le pare-feu***

C'est un élément (logiciel ou matériel) du réseau informatique contrôlant les communications qui traversent le réseau. Il a pour fonction de faire respecter la politique de sécurité du réseau.

6. ***Journalisation (logs)***

Permet l'enregistrement des activités de chaque acteur et de constater que des attaques ont eu lieu, de les analyser et défaire en sorte qu'elles ne se reproduisent pas plus tard.

7. ***Analyse des vulnérabilités (Security audit)***

Permet l'identification des points de vulnérabilité du système qui ne détecte pas les attaques ayant déjà eu lieu, ou lorsqu'elles auront lieu.

8. ***Détection d'intrusion***

Permet de détection des comportements anormaux d'un utilisateur ou des attaques connues.

II.3 Les systèmes de détection d'intrusion (IDS)

Le concept de système de détection d'intrusions a été introduit en 1980 par James Anderson, mais le sujet n'a pas eu beaucoup de succès. Il a fallu attendre la publication d'un modèle de détection d'intrusions par Denning [29] en 1987 pour marquer réellement le départ du domaine. En 1988, il existait au moins trois prototypes (IDS). La recherche dans le domaine s'est ensuite développée, le nombre de prototypes s'est énormément accru. Le gouvernement des Etats-Unis a investi des millions de dollars dans ce type de recherches dans le but d'accroitre la sécurité de ses machines .

II.3.1 Intrusion

Nous appellerons intrusion 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 proviennent le plus fréquemment de personnes internes ayant déjà un accès au système [29].

II.3.2 Détection d'intrusion

Les techniques de détection d'intrusion tentent de faire la différence entre une utilisation normale du système et une tentative d'intrusion et donnent l'alerte. Typiquement, les données d'audit du système sont parcourues à la recherche de signatures connues d'intrusions ou de comportements anormaux. La détection peut être faite en temps réel, dans ce cas, le programme (IDS) peut donner l'alerte, auquel le personnel qualifié pourra tenter de remédier à l'intrusion, en coupant la connexion ou en remontant la piste[29] .

II.3.3 Systèmes de détection d'intrusion

Les IDSs sont des outils permettant de détecter les attaques intrusions des systèmes sur lequel ils sont placés .

Il existe deux types d'IDS : les IDS à base de nœuds et les IDS réseaux [30] :

- ✓ ***Les IDSs à base de nœud (Host Intrusion Détection System-HIDS)***

Analysent le fonctionnement et l'état des machines sur lesquels ils sont installés afin de détecter les attaques en se basant sur des démons. L'intégrité des systèmes est alors vérifiée périodiquement et des alertes peuvent être levées.

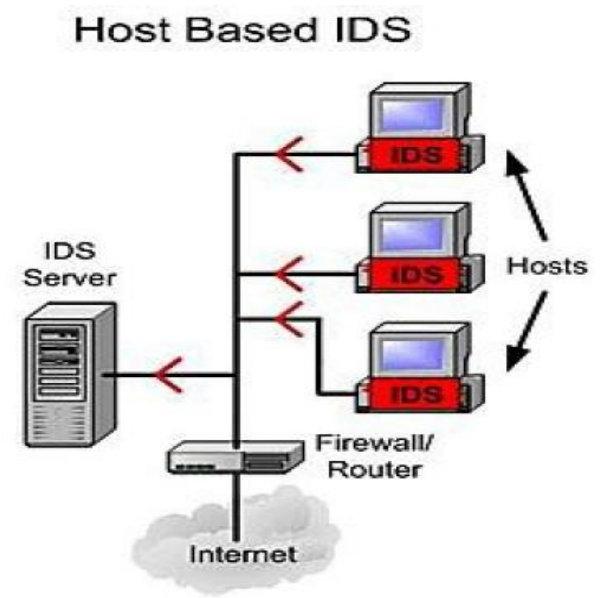


Figure 12. Les IDS à base de nœud [30].

- ✓ ***Les IDSs réseaux (Network Intrusion Detection System-NIDS)***

Analysent en temps réel le trafic qu'ils aspirent à l'aide d'une sonde (carte réseau en mode promiscues). Ensuite, les paquets sont décortiqués puis analysés. En cas, de détection d'intrusion, des alertes peuvent être envoyées.

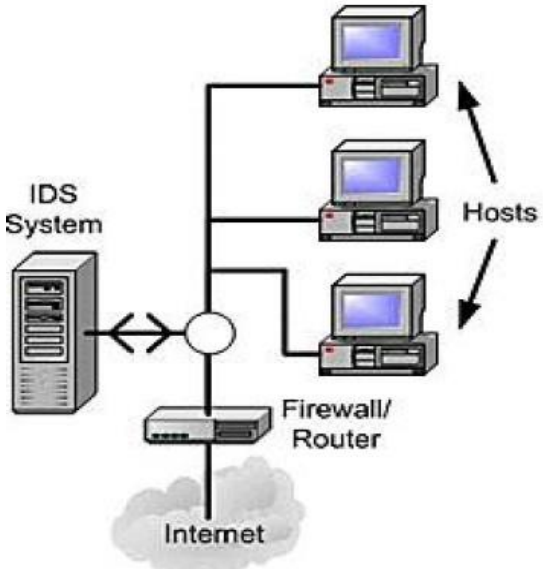


Figure 13. Les IDS réseaux [30].

Les deux types d'IDS possèdent des avantages ainsi que des inconvénients, ces derniers sont illustrer dans le tableau suivant :

Types	Avantages	Inconvénients
IDSs HIDS	Permet de constater l'impact d'une attaque et pout donc mieux réagir. Observation des activités sur l'hôte avec précision.	Ils ont moins de facilité à détecter les scans. Ils sont plus vulnérables aux attaques de type Dos. L'analyse des traces d'audit du système est très contraignante en raison de la taille de ces dernières. Consommation de ressources CPU.
IDSs NIDS	Les capteurs peuvent être bien sécurisés puisqu'ils se "contentent" d'observer le trafic. Détection facile des scans grâce aux signatures. Peut filtrer le trafic.	La probabilité de faux négatifs (attaques non détectées comme telles) est élevée et il est difficile de contrôler le réseau entier. Ils doivent principalement fonctionner de manière cryptée d'où une complication de l'analyse des paquets. A l'opposé des HIDS, ils ne voient pas les impacts d'une attaque.

Tableau 1. Les avantages et inconvénients des deux types d'IDSs [30].

II.4 Les critères de classifications des IDSs

Les IDSs peuvent être classifiés en cinq (5) critères de classification qui ont été introduits par Debar et al. dans la figure II.3 résume les critères de classification des IDSs [31].

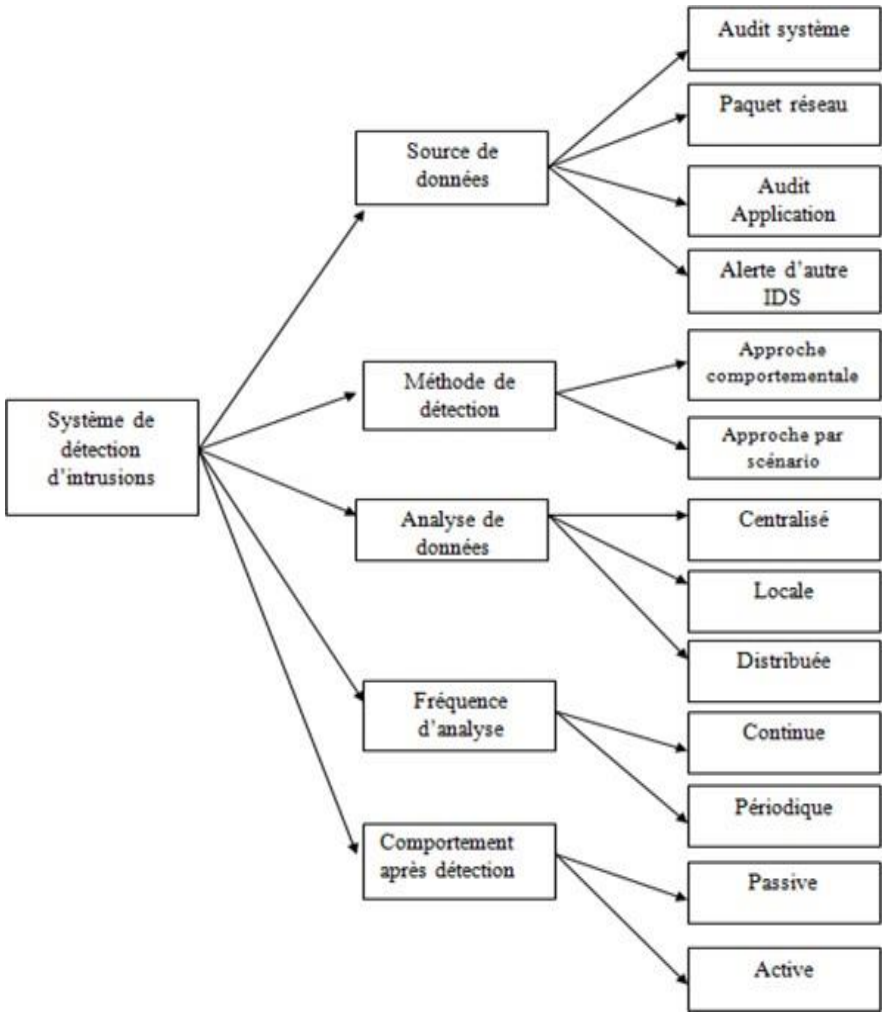


Figure 14. Les critères de classification des IDSs [32].

II.4.1 Sources de données

Les sources des données constituent les informations qu'elles fournissent pour analyser le système pour d'éventuelles intrusions. Il existe quatre sources de données qui sont: l'audit système, le trafic réseau, l'audit applicatif et les alertes d'autres IDSs [32] :

Audit système

Permet d'enregistrer les actions effectuées sur le système en exploitant l'outil d'administration système. Cette source de données est très pertinente, du fait que la totalité des attaques provoquent une interaction avec le système. Les données d'audit système sont le seul moyen pour recueillir des informations sur les activités des utilisateurs d'une machine donnée.

Paquet réseau

Représente les paquets récupérés du réseau. La plupart des accès aux ordinateurs se font via les réseaux informatiques, alors que la capture des paquets avant qu'ils entrent au serveur est le moyen le plus efficace pour les contrôler. L'analyse des

paquets représente le moyen le plus efficace pour détecter les attaques de type déni de service (DOS).

Audit application

C'est une source d'information de haut niveau, qui représente les services Web tel que le FTP (File Transfert Protocol) et le HTTP (Hyper Text Transfert Protocol). Avec la grande augmentation de l'utilisation des serveurs d'application, les fichiers logs des applications sont devenus une source d'information pour les systèmes de détection d'intrusions.

Alerte d'autre IDSs

C'est une source d'information basée IDS. Ce sont les alertes remontées par des analyseurs provenant d'un IDS. Chaque alerte synthétise déjà un ou plusieurs événements intéressants du point de vue de la sécurité. La corrélation de ces alertes conduit parfois à la détection d'une intrusion complexe de plus haut niveau.

II.4.2 Méthode de détection

Il existe deux méthodes de détection, la première consiste à utiliser des connaissances accumulées sur les attaques puis les exploiter afin de prouver l'existence d'autres attaques. Le second consiste à créer un modèle basé sur le comportement habituel du système et surveiller toute déviation de ce comportement. La première méthode est appelée approche par scénario et la seconde est l'approche comportementale [32] :

L'approche par scénario

Elle se base sur des attaques connues préalablement. Cela nécessite une connaissance à priori des attaques à détecter. Elle se base sur les connaissances accumulées sur les attaques spécifiques et les vulnérabilités du système. Le système de détection d'intrusion contient les informations et cherche toute tentative de les exploiter. Si l'IDS détecte une tentative, une alarme est déclenchée. Par conséquent, la précision des systèmes de détection d'intrusions 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.

L'approche comportementale

Les techniques de détection d'intrusion basée sur l'approche comportementale supposent que l'intrusion peut être détectée par l'observation de la déviation par rapport au comportement normal ou prévu du système ou des utilisateurs. Au début le modèle du comportement normal est extrait à partir des informations de référence recueillies par divers moyens, puis le système de détection d'intrusions compare ce modèle avec l'activité

actuelle, si une déviation est détectée, une alerte est déclenchée. D'une manière générale, on peut dire que cette approche considère tout comportement non enregistré précédemment, comme intrusion. Par conséquent, cette approche peut être complète, mais la précision reste son plus grand souci.

II.4.3 Analyse de données

Les données peuvent être analysée en centralisé, en locale ou bien distribué [32] :

1. ***Analyse centralisé***

Consiste à centraliser les alertes et le contrôle au sein d'une seule machine (administrateur réseau).

2. ***Analyse locale***

L'analyse se fait localement, c'est-à-dire sur chaque hôte séparément des autres. Donc, on lance l'analyse sur un hôte seulement, ce qui fait perdre du temps. De plus, il n'y a pas de surveillance entière du réseau en même temps.

3. ***Analyse distribué***

L'analyse se fait de manière distribuée sur le réseau. L'analyse est lancée sur toutes les machines au même temps. Ce qui fait gagner du temps dans l'analyse et permet la surveillance entière du réseau et évite donc d'avoir des attaques au moment de l'analyse.

II.4.4 Fréquence d'analyse

Il existe deux façons dont les systèmes de détection d'intrusions effectuent leurs analyses qui sont : **analyse continue et analyse périodique** [32] :

La surveillance continue

Une analyse continue et en temps réel par l'acquisition d'informations sur les mesures prises sur l'environnement et analyse ce cliché a la recherche des logiciels vulnérables, des erreurs de configuration, etc.

L'analyse périodique

Cette analyse revient à surveiller et à analyser le système de manière périodique, c'est-à-dire, à chaque période de temps qui est choisie par l'administrateur (observation périodique dans le temps).

II.4.5 Comportement après détection

Le comportement de l'IDS après une détection d'intrusion peut être passive, active ou bien les deux au même temps [32] :

Action passive

L'IDS informe directement l'utilisateur ou le manager qu'une intrusion est détectée en déclenchant une alerte.

Action active

L'IDS en cas d'une attaque l'IDS non seulement il informe le manager ou l'utilisateur mais aussi-il réagi contre cette attaque en coupant le courant par exemple.

Action passive et active

L'IDS peut informer l'utilisateur ou bien le manager qu'une intrusion est détectée et il réagit directement.

II.4.6 Les limites actuelles de la détection d'intrusions

Il existe des limites spécifiques pour la détection d'abus ainsi que des limites pour la détection d'anomalies qui sont les suivant [33] :

1. Limites spécifiques à la détection d'abus

Les principaux défis actuels de cette technique sont les suivants :

- Base de signature d'attaques délicate à construire ;
- Seules les attaques contenues dans la base sont détectées ;
- Nécessite la mise à jour de la base de signature d'attaques comme les Antivirus ;
- Incapables de détecter certains types d'attaque. Ils sont eux-mêmes vulnérables aux attaques ;
- Problème des faux négatifs qui sont le fait que les nouvelles attaques passent l'IDS sans être détectées.

2. Limites spécifiques à la détection d'anomalie

Cette technique comporte elle-aussi de nombreux problèmes complexes à résoudre ; voici les plus couramment évoqués :

- Choix délicat des mesures à retenir pour un système cible donné ;
- Pour un utilisateur au comportement erratique, toute activité est " normale " ;
- En cas de profonde modification de l'environnement du système cible, déclenchement d'un flot ininterrompu d'alarmes (faux positifs).
- Utilisateur pouvant changer lentement de comportement dans le but d'habituer le système à un comportement intrusif (faux négatifs).

II.5 Les IDS dans le Cloud Computing

Au cours des dernières années, plusieurs travaux de recherche ont été faits dans le but de proposer des solutions d'IDS dans l'environnement du Cloud Computing. Nous allons citer dans cette partie les travaux les plus connus [34] :

1. Architecture IDS de l'environnement Cloud Computing

Une interaction entre les services est offerte dans [35], l'IDS fournit des services et des services de stockage pris en charge par chaque nœud de l'environnement Cloud. Le système de service IDS est composé en deux éléments : l'analyseur et le système d'alerte. Les données provenant de diverses ressources sont capturé par le vérificateur (auditeur) de l'événement. Le système de service IDS reçoit des données de l'auditeur de l'événement. Ces données sont utilisées pour détecter les intrusions en utilisant une technique basée sur la basée du comportement. Dans cette approche, le réseau neuronal artificiel (ANN) est utilisé pour détecter les attaques inconnues. Le système d'alerte informe les autres nœuds lorsqu'une attaque ou intrusion est détectée.

2. Approche basé sur les statistiques et la théorie des probabilités

L'approche CP (Co-variante Probability) est basée sur les statistiques et la théorie des probabilités [36]. Les statistiques se basent sur des ensembles mutuellement exclusifs et ils sont utilisés pour décomposer cet ensemble. La détection d'intrusion est construite en utilisant les sous-ensembles générés à partir des espaces de l'échantillon considéré.

3. IDS associe à une machine virtuelle

Cette approche se compose de deux éléments [37] : Unité de gestion IDS et capteur IDS. Unité de gestion IDS se compose d'un rassembleur d'événement, d'une base de données d'événements, du composant d'analyse et d'un contrôleur à distance. Le capteur IDS identifie les comportements malveillants. Le rassembleur d'événements collecte les comportements et les stocke dans la base de données d'événements. Le composant d'analyse accède à la base de données d'événements et d'analyser les événements qui sont configurés par les utilisateurs. Le contrôleur d'IDS gère l'IDS-VM (machines virtuelles) et peut communiquer avec l'IDS-VM et le capteur IDS.

4. Approche basé sur agent mobile

Pour détecter l'intrusion dans des applications Cloud Dastjerdi et al. [38] ont proposé une méthode évolutive, flexible et rentable en utilisant les agents mobiles. Cette méthode est utilisée pour protéger les machines virtuelles qui se trouvent à l'extérieur de l'organisation. L'ensemble des attaques sont collectées par l'agent mobile. Une analyse plus poussée et une vérification sont appliquées sur cette preuve.

II.6 Les Caractéristiques et les limites des différentes approches proposées

Dans le tableau suivant, une comparaison des différentes approches, qui ont été proposé, est présentée [34] :

IDS	Inconvénients
Architecture IDS de l'environnement Cloud Computing	Nécessite plus de temps de formation et d'exemples pour la précision de la détection ; Il ne peut pas détecter toutes intrusions internes en cours d'exécution sur les ma- chines virtuelles.
Approche basé sur les statistiques et la théorie des probabilités	Utilisé pour détecter tous les types d'attaques ; Limitation du temps de calcul.
IDS associe à une machine virtuelle	La VM peut être attaqué ; Méthode très complexe ; Ne peut pas détecter les attaques sur les machines virtuelles.
Approche basé sur agent mobile	Produit la charge du réseau avec une augmentation de VM attaché ; Fournit un IDS pour l'application Cloud indépendamment de leur emplacement.

Tableau 2. Les limites des différentes approches proposées [34].

II.7 Conclusion

De manière générale, l'efficacité d'un système de détection d'intrusion dépend de sa "configurabilité" (possibilité de définir et d'ajouter de nouvelles spécifications d'attaque), de sa robustesse (résistance aux défaillances) et de la faible quantité de faux positifs (fausses alertes) et de faux négatifs (attaques non détectées) qu'il génère. Dans ce chapitre, nous avons présenté la sécurité des Cloud, ses objectifs et les mécanismes de défense, les systèmes de détection d'intrusions ainsi que quelques travaux de recherche qui ont été proposées par les chercheurs.

La détection d'intrusion est devenue une industrie mature et une technologie éprouvée. Néanmoins, quelques voies restent cependant relativement inexplorées : les mécanismes de réponse aux attaques, les architectures pour les systèmes de détection d'intrusions distribués. Les standards d'interopérabilité entre différents systèmes de détection d'intrusion, et la recherche de nouveaux paradigmes pour effectuer la détection d'intrusion.

Dans le chapitre suivant, nous allons voir la machine Learning l'outil utilisé pour notre travail.

CHAPITRE III:

MACHINE LEARNING

Chapitre III: MACHINE LEARNING

III.1 Introduction

Lorsqu'on parle de « Machine Learning », que l'on traduit en français par « apprentissage automatique », il est présent depuis des décennies dans certaines applications spécialisées telles que la reconnaissance optique de caractères ou OCR. La première application ML ayant véritablement touché un large public, améliorant le quotidien de centaines de millions de personnes, s'est imposé dans les années 1990 : il s'agit du filtre anti-spam. S'il n'est pas vraiment présenté comme tel, il possède pourtant techniquement les caractéristiques d'un système d'apprentissage automatique (d'ailleurs, il a si bien appris que vous n'avez que rarement à signaler un e-mail comme indésirable). Il a été suivi par des centaines d'applications ML intégrées désormais discrètement dans des centaines de produits et fonctionnalités que vous utilisez régulièrement, depuis les recommandations jusqu'à la recherche vocale.

III.2 Origines de l'Apprentissage Automatique

La discipline de l'apprentissage automatique possède de riches fondements théoriques. On sait, désormais, répondre à des questions comme :

1. Combien d'exemples d'entrainement faut-il fournir à un programme d'apprentissage pour être certain qu'il apprenne avec une efficacité donnée ?

Etant donnée la variété d'apprentissages qu'on peut rencontrer, il est aisé de deviner que les fondements de cette discipline, en occurrence l'apprentissage automatique, proviennent de diverses sciences :

1. Des mathématiques pour l'informatique : algèbre linéaire, la probabilité, la logique, l'analyse élémentaire, …
2. La théorie statistique de l'estimation,
3. L'apprentissage Bayésien,
4. L'inférence grammaticale ou l'apprentissage par renforcement, et tant d'autres.

III.3 Définitions et Particularités

Parmi les principales caractéristiques et facultés adoptées par les modèles d'apprentissage, nous citons : l'entrainement, la généralisation, l'adaptation, l'amélioration, l'intelligibilité et la prédiction.

III.3.1. Apprentissage

L'apprentissage, ou *Learning* en anglais, c'est le processus de **construire un modèle général** à partir de **données** (observations) **particulières** du monde réel [35].

Ainsi, le but est double :

1. **Prédire** un comportement face à une nouvelle donnée.
2. **Approximer** une fonction ou une densité de probabilité.

Deux (02) branches d'apprentissage existent :

1. Apprentissage **symbolique**, issue de l'IA.
2. Apprentissage **numérique**, issue des statistiques.

Dans la pratique, le mot **entraînement** est souvent synonyme de : *Apprentissage*. Ainsi, en science cognitive, l'apprentissage est défini comme étant la **capacité d'améliorer les performances** au fur et à mesure de l'exercice d'une activité.

III.3.2. Adaptation

Elle peut être vue comme étant la disposition du modèle (algorithme ou système) à corriger son comportement ou à remanier sa réponse (ex., prédiction) par rapport à de nouvelles situations.

Pour les tâches de perception, en vision artificielle, on accumule les **bonnes** et **mauvaises expériences**, et à partir d'elles, on peut faire **évoluer les règles** pour mieux effectuer la tâche, c'est le phénomène d'**adaptation** ou d'**amélioration** [35].

III.3.3. Dilemme de l'Apprentissage : Précision Vs Généralisation

1. ***Précision*** : c'est **l'écart** entre une **valeur mesurée** ou **prédite** par le modèle et une **valeur réelle**.
2. ***Généralisation*** : c'est la **capacité de reconnaître de nouveaux exemples** jamais vus auparavant.

NB : Souvent, un *seuil de généralisation* est utilisé et est propre à chaque modèle pendant l'apprentissage.

III.3.4. Intelligibilité

C'est amélioré la compréhension des résultats d'apprentissage, afin que le modèle puisse fournir une connaissance claire et compréhensible, au sens interprétable.

Exemple, quand un expert extrait de la connaissance des bases de données, il *apprend une manière de les résumer* ou *de les formuler* (expliquer, expliciter de manière simple et précise).

D'un point de vue fouille de données, ça revient purement et simplement à contrôler l'intelligibilité (clarté) d'un modèle obtenu.

Actuellement, la **mesure d'intelligibilité** se réduit à vérifier que la connaissance produite est intelligible et que les **résultats** sont exprimés dans le **langage de l'utilisateur** et la **taille des modèles n'est pas excessive**.

III.4 Principes

Les algorithmes utilisés permettent, dans une certaine mesure, à un système piloté par ordinateur (un robot éventuellement), ou assisté par ordinateur, d'adapter ses analyses et ses comportements en réponse, en se fondant sur l'analyse de données empiriques provenant d'une base de données ou de capteurs.

La difficulté réside dans le fait que l'ensemble de tous les comportements possibles compte tenu de toutes les entrées possibles devient rapidement trop complexe à décrire. On confie donc à des programmes le soin d'ajuster un modèle pour simplifier cette complexité et de l'utiliser de manière opérationnelle. Idéalement, c'est-à-dire que la nature des données d'entrainement n'est pas connue

Ces programmes, selon leur degré de perfectionnement, intègrent éventuellement des capacités de traitement probabiliste des données, d'analyse de données issues de capteurs, de reconnaissance (reconnaissance vocale, de forme, d'écriture…).

III.5 Domaines de l'Apprentissage Automatique

Les principaux domaines d'applications de l'apprentissage automatique sont les fouilles de données et l'intelligence artificielle.

La fouille de données est le processus d'extraction de la connaissance : il consiste à sélectionner les données à étudier à partir de bases de données (hétérogènes ou homogènes), à épurer ces données et enfin à les utiliser en apprentissage pour construire un modèle.

Exemples :

1. Systèmes de vidéo surveillance pour la détection des intrus.
2. Logiciel biométrique de reconnaissance de visages et d'empreintes digitales.

III.6 Classification et régression

En analyse de données, la classification consiste à regrouper des ensembles d'exemples (souvent, de manière non-supervisée) en classes. Ces classes sont généralement organisées en une structure : clusters (groupes ou grappes).

III.6.1 Classification

C'est le processus de reconnaissance en intention (par leurs propriétés) des classes décrites en extension (par les valeurs de leurs descripteurs).

Si les valeurs à prédire sont des classes en petit nombre, on parle alors de classification.

III.6.2 Régression

Elle traite des exemples où les valeurs à prédire sont numériques.

III.7 Apprentissage automatique

III.7.1 Types de systèmes d'apprentissage automatique

Il existe tellement de types de systèmes d'apprentissage automatique différents qu'il est utile de les classer en grandes catégories :

- selon que l'apprentissage s'effectue ou non sous supervision humaine (apprentissage supervisé, non supervisé, semi-supervisé ou avec renforcement),
- selon que l'apprentissage s'effectue ou non progressivement, au fur et à mesure (apprentissage en ligne ou apprentissage groupé),
- selon qu'il se contente de comparer les nouvelles données à des données connues, ou qu'il détecte au contraire des éléments de structuration dans les données d'entraînement et construise un modèle prédictif à la façon d'un scientifique (apprentissage à partir d'observations ou apprentissage à partir d'un modèle).

Ces critères ne sont pas exclusifs, vous pouvez les combiner comme vous le souhaitez. Ainsi, un filtre anti-spam dernier cri peut apprendre au fur et à mesure en s'appuyant sur un modèle de réseau neuronal profond dont l'apprentissage s'effectue sur des exemples de messages indésirables ou non : ceci en fait un système supervisé d'apprentissage en ligne à partir d'un modèle. Examinons maintenant plus soigneusement chacun de ces critères.

III.7.2 Apprentissage supervisé / non supervisé

Les systèmes d'apprentissage automatique peuvent être classés en fonction de l'importance et de la nature de la supervision qu'ils requièrent durant la phase d'entrainement. Il existe quatre catégories majeures : l'apprentissage supervisé,

l’apprentissage non supervisé, l’apprentissage semi-supervisé et l’apprentissage avec renforcement.

III.7.2.1 L’apprentissage supervisé

La majorité des apprentissages automatiques utilisent un apprentissage supervisé. L’apprentissage supervisé consiste à des variables d’entrée (x) et une variable de sortie (Y). C’est un algorithme qui apprend une fonction de cartographie de l’entrée à la sortie Y = f (X) [35].

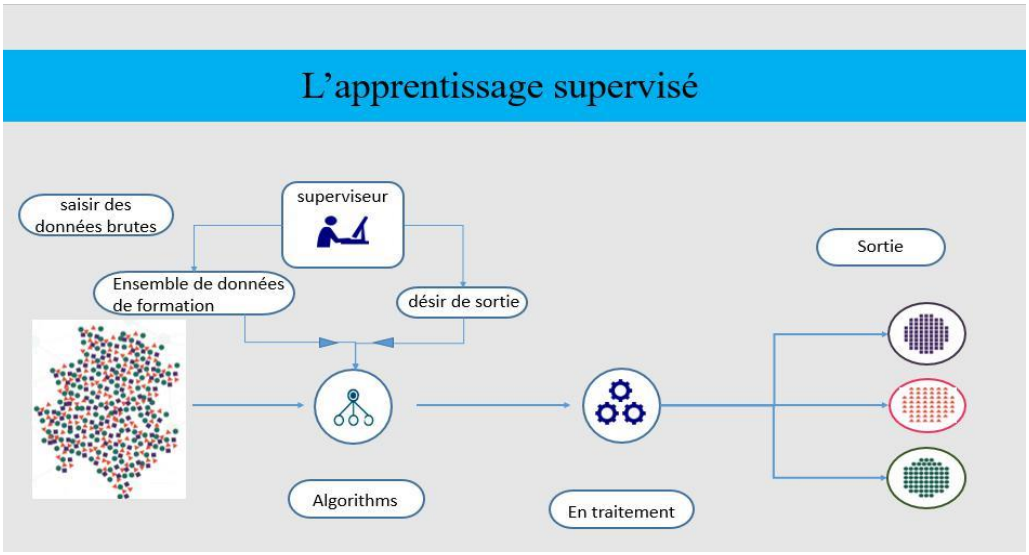


Figure 15. L’apprentissage Supervise

Dans l'apprentissage supervisé, l'ordinateur est fourni avec des exemples d'entrées qui sont étiquetés avec les sorties souhaitées. Le but de cette méthode est que l'algorithme puisse «apprendre» en comparant sa sortie réelle avec les sorties «enseignées» pour trouver des erreurs et modifier le modèle en conséquence. L'apprentissage supervisé utilise donc des modèles pour prédire les valeurs d'étiquettes sur des données non étiquetées supplémentaires [36].

Les algorithmes de régression peuvent être utilisés également en classification, et inversement. Par exemple, la régression logistique s’utilise couramment en classification, car elle peut fournir une valeur correspondant à la probabilité d’appartenance à une classe donnée (par exemple, 20 % de chances qu’un courriel soit du spam).

Voici quelques-uns des plus importants algorithmes d’apprentissage supervisé (présentés dans ce livre) :

- K plus proches voisins
- Régression linéaire
- Régression logistique
- Machines à vecteurs de support (SVM)

- Arbres de décision et forêts aléatoires
- Réseaux neuronaux.

Algorithme Support Vector Machine SVM

«Support Vector Machine» (SVM) est un algorithme d'apprentissage automatique supervisé qui peut être utilisé à la fois pour les défis de classification ou de régression. Cependant, il est principalement utilisé dans les problèmes de classification.

Les domaines d'applications SVM

Est une méthode de classification qui montre de bonnes performances dans la résolution de problèmes variés. Cette méthode a montré son efficacité dans de nombreux domaines d'applications tels que le traitement d'image, la catégorisation de textes ou le diagnostiques médicales et ce même sur des ensembles de données de très grandes dimensions.

La réalisation d'un programme d'apprentissage par SVM se ramène à résoudre un problème d'optimisation impliquant un système de résolution dans un espace de dimension conséquente. L'utilisation de ces programmes revient surtout à sélectionner une bonne famille de fonctions noyau et à régler les paramètres de ces fonctions. Ces choix sont le plus souvent faits par une technique de validation croisée, dans laquelle on estime la performance du système en la mesurant sur des exemples n'ayant pas été utilisés en cours d'apprentissage. L'idée est de chercher les paramètres permettant d'obtenir la performance maximale. Si la mise en œuvre d'un algorithme de SVM est en général peu coûteuse en temps, il faut cependant compter que la recherche des meilleurs paramètres peut requérir des phases de test assez longues [47].

Principe de fonctionnement général de SVM

Pour deux classes d'exemples donnés, le but de SVM est de trouver un classificateur qui va séparer les données et maximiser la distance entre ces deux classes. Avec SVM, ce classificateur est un classificateur linéaire appelé hyperplan. Dans le schéma qui suit, on détermine un hyperplan qui sépare les deux ensembles de points [47].

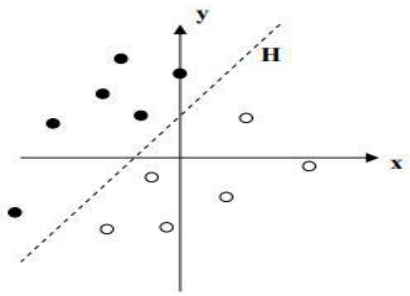


Figure 16. Exemple d'un hyperplan séparateur.

Les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan, sont appelés vecteurs de support.

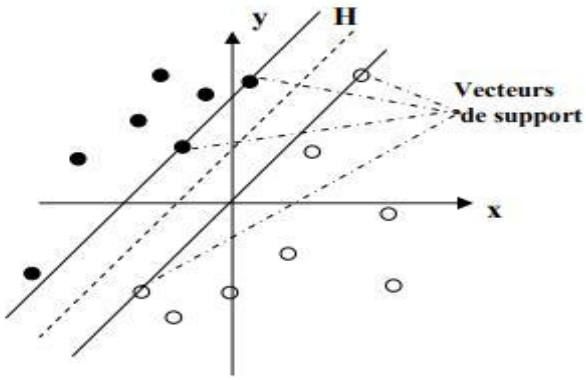


Figure 17. Exemple de vecteurs de support.

Il est évident qu'il existe une multitude d'hyperplan valide mais la propriété remarquable des SVM est que cet hyperplan doit être optimal. Nous allons donc en plus chercher parmi les hyperplans valides, celui qui passe « au milieu » des points des deux classes d'exemples. Intuitivement, cela revient à chercher l'hyperplan le « plus sûr » [48]. En effet, supposons qu'un exemple n'ait pas été décrit parfaitement, une petite variation ne modifiera pas sa classification si sa distance à l'hyperplan est grande. Formellement, cela revient à chercher un hyperplan dont la distance minimale aux exemples d'apprentissage est maximale. On appelle cette distance « marge » entre l'hyperplan et les exemples. L'hyperplan séparateur optimal est celui qui maximise la marge. Comme on cherche à maximiser cette marge, on parlera de séparateurs à vaste marge [48].

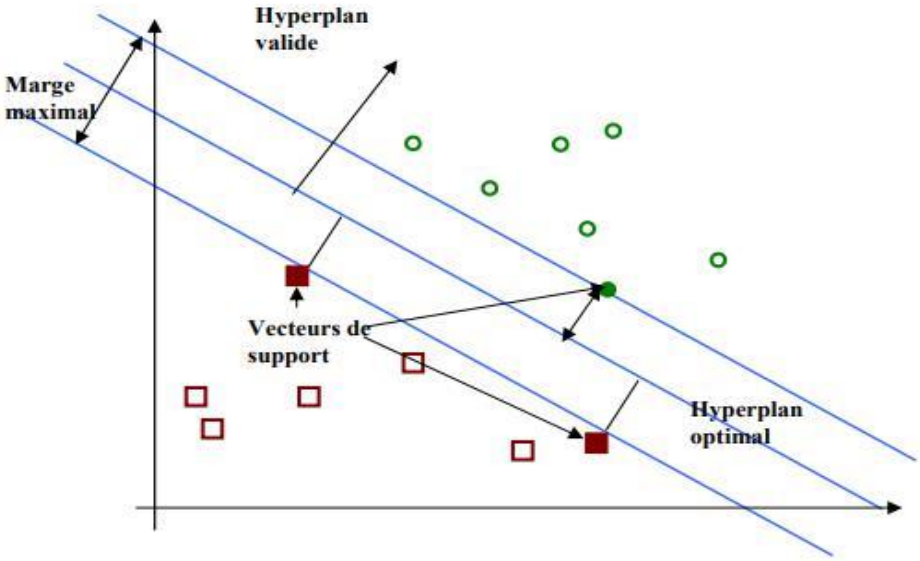


Figure 18. Exemple de marge maximal (hyperplan valide).

III.7.2.2 L'apprentissage non supervisé

L'apprentissage non supervisé consiste à ne disposer que de données d'entrée (X) et pas de variables de sortie correspondantes. L'objectif de l'apprentissage non supervisé est de modéliser la structure ou la distribution sous-jacente des données afin d'en apprendre davantage sur les données.

On appelle apprentissage non supervisé car, contrairement à l'apprentissage supervisé ci-dessus, il n'y a pas de réponse correcte ni d'enseignant. Les algorithmes sont laissés à leurs propres mécanismes pour découvrir et présenter la structure intéressante des données.

L'apprentissage non supervisé comprend deux catégories d'algorithmes :

Algorithmes de regroupement et d'association. [35]

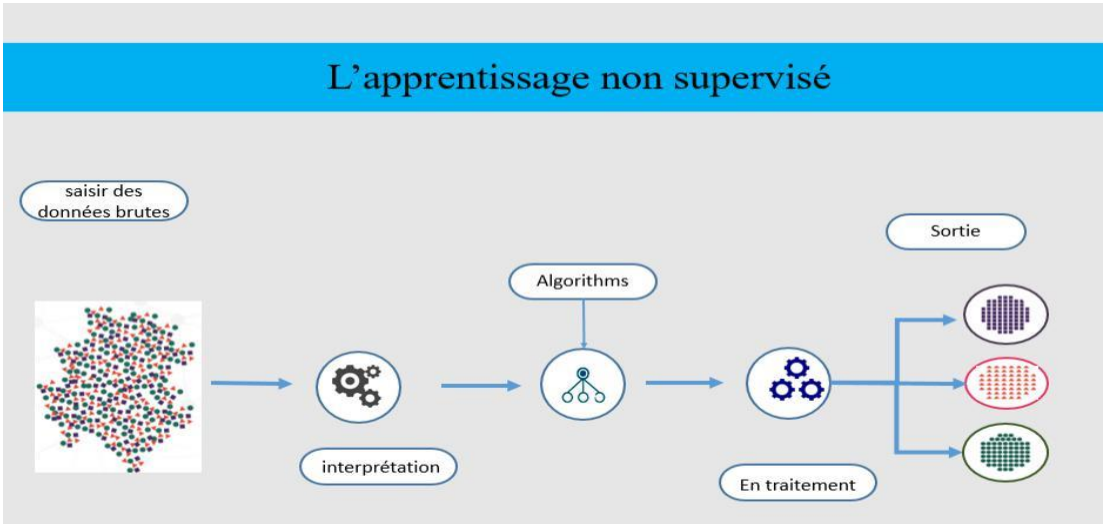


Figure 19. L'apprentissage Non Supervise

Dans l'apprentissage non supervisé, les données sont non étiquetées, de sorte que l'algorithme d'apprentissage trouve tout seul des points communs parmi ses données d'entrée. Les données non étiquetées étant plus abondantes que les données étiquetées, les méthodes d'apprentissage automatique qui facilitent l'apprentissage non supervisé sont particulièrement utiles.

L'objectif de l'apprentissage non supervisé peut être aussi simple que de découvrir des modèles cachés dans un ensemble de données, mais il peut aussi avoir un objectif d'apprentissage des caractéristiques, qui permet à la machine intelligente de découvrir automatiquement les représentations nécessaires pour classer les données brutes. [36]

Voici quelques-uns des plus importants algorithmes d'apprentissage non supervisés

- Partitionnement

— K-moyennes (K-means)

— Partitionnement hiérarchique — Maximum de vraisemblance

- Visualisation et réduction de dimension

— Analyse en composantes principales

— Analyse en composantes principales à noyaux

— Plongement localement linéaire (Locally-Linear Embedding ou LLE)

— Méthode t-SN/E (t-distributed Stochastic Neighbor Embedding)

Algorithme de k-means

K-means (k-moyennes) est un algorithme non supervisé de clustering, populaire en Machine Learning. Il permet de regrouper en K clusters distincts les observations du data set. Ainsi les données similaires se retrouveront dans un même cluster. Par ailleurs, une observation ne peut se retrouver que dans un cluster à la fois (exclusivité d'appartenance). Une même observation, ne pourra donc, appartenir à deux clusters différents.

Notion de similarité

Pour pouvoir regrouper un jeu de données en K cluster distincts, l'algorithme K-means a besoin d'un moyen de comparer le degré de similarité entre les différentes observations. Ainsi, deux données qui se ressemblent, auront une distance de dissimilarité réduite, alors que deux objets différents auront une distance de séparation plus grande.

Choisir K : le nombre de clusters

Choisir un nombre de cluster K n'est pas forcément intuitif. Spécialement quand le jeu de données est grand et qu'on n'ait pas un a priori ou des hypothèses sur les données. Un nombre K grand peut conduire à un partitionnement trop fragmenté des données. Ce qui empêchera de découvrir des patterns intéressants dans les données. Par contre, un nombre de clusters trop petit, conduira à avoir, potentiellement, des clusters trop généralistes contenant beaucoup de données. Dans ce cas, on n'aura pas de patterns "fins" à découvrir.

Pour un même jeu de données, il n'existe pas un unique clustering possible. La difficulté résidera donc à choisir un nombre de cluster K qui permettra de mettre en lumière des patterns intéressants entre les données. Malheureusement il n'existe pas de procédé automatisé pour trouver le bon nombre de clusters.

III.8 Avantages ET Inconvénients Des Algorithmes De ML

APPRENTISSAGE	ALGORITHMES		AVANTAGES	INCONVENIENTS
Supervisé	**Classification**	KNN	1. facile à implémente. 2. efficace. 3. L'algorithme est polyvalent	1. Calculer chaque fois la similarité entre les k. 2. grande capacité de stockage. 3. utilise de nombreuses données de références pour classifier les nouvelles entrées
		SVM	1. Leur capacité à manipuler de grandes quantités de données 2. Le faible nombre d'hyper paramètres. 3. Elles sont bien fondées théoriquement.	1. complexes pour la classification des corpus. 2. demande un temps énorme pendant les phases de test.
		Arbre de Décision	1. faciles à comprendre. 2. Ils permettent de sélectionner l'option la plus appropriée parmi plusieurs. 3. Il est facile de les associer à d'autres outils de prise de décision.	1. instables. 2. Certains concepts sont difficiles à exprimer à l'aide d'arbres de décision (comme XOR).
		Naïve Bayes	1. La facilité et la simplicité de leur implémentation. 2. Leur rapidité. 3. Les méthodes Naïve Bayes donnent de bons résultats.	1. faire le même travail de classification.
	REGRESSION	Linéaire	1. Simplicité d'interprétation. 2. facilité de calcul	Elle ne traite pas les valeurs manquantes de variables continues sensible aux valeurs hors norme de variables continue
Non Supervisé	**REDUCTION DES DIMENSIONS**	PCA	1. Simplicité Mathématique 2. Simplicité des résultats 3. Puissance 4. Flexibilité	1. l'ACP n'a pas réellement 2. s'applique simplement sur des cas précis 3.Perte d'information par l'emploi fréquent de la 1ère composante principale uniquement.
	CLUSTERING	K-means	1. Simple 2. Flexible 3. Efficace 4. Complexité temporelle.	1. Ensemble non optimal de clusters 2. Manque de cohérence 3. Limitation des calculs 4. Spécifiez les valeurs k

Tableau 3. Avantages ET Inconvénients Des Algorithmes De ML.

III.9 Conclusion

Le machine Learning est un outil très puissant qui permet d'effectuer de multiples actions comme classifier des données, faire apprendre à un programme à partir d'expérimentations ou encore de créer un programme évolutionnaire qui s'améliore sans cesse. Ainsi, même avec un échantillon peu fourni (le machine Learning nécessite habituellement des échantillons avec 50 spécimens) et des données influencées par la subjectivité de celui qui les mesure, le machine Learning reste relativement précis malgré quelques lacunes.

CHAPITRE IV :

APPROCHE PROPOSEE ET IMPLEMENTATION

Chapitre IV: APPROCHE PROPOSEE ET IMPLENMENTATION

IV.1. Introduction

Un projet de machine learning commence généralement avec un jeu de données et un problème à résoudre. Celui-ci se décrit par trois éléments, des données (X), une cible (Y) et une fonction d'erreur qui permette d'évaluer la distance entre la prédiction et la cible. Dans ce chapitre on va traiter notre dataset (X) par passage sur les étapes de la machine learning pour arriver à un target (Y).

IV.2. L'environnement de simulation

IV.2.1. L'environnement matériel

On a utilisé une machine, configurée comme suit :

- ✓ Machine pc HP probook ;
- ✓ Mémoire Vive : 4 Go ;
- ✓ Disque Dur : 320 Go ;
- ✓ Processeur : Intel (R) Core (TM) i3 2.10 GHz ;
- ✓ Type de système : Windows 7.

IV.2.2. Environnement Logiciel

Lors de notre travail, on a utilisé, les outils logiciels suivants:

- ✓ Langage de programmation Python ;
- ✓ Anaconda navigator (jupiter) ;
- ✓ Microsoft Excel.

a. Définition du langage Python en informatique

Python est le langage de programmation open source le plus employé par les informaticiens. Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages. [37]

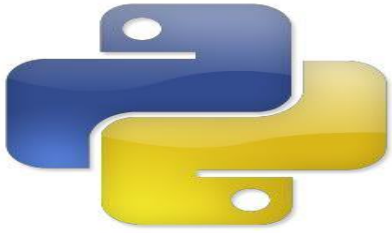


Figure 20. Logo Python.

b. Définition de l'anaconda

Anaconda est une plate-forme informatique scientifique et de traitement de donnée basée sur Python. Il a intégré de nombreuses bibliothèques tierces très utiles. [38]

Figure 21. Logo Anaconda.

Anaconda est donc un utilitaire dont on ne peut quasiment pas se passer lorsque l'on a un projet incluant du Python. [39]

c. Définition jupyter

Jupyter se présente comme un outil extrêmement simple à mettre en œuvre qui vous permettra de transformer vos Jupyter Notebooks en applications web ou en Dashboard quasiment automatiquement. [40]

Figure 22. Logo Jupytre.

d. Bibliothèques Supplémentaires

Afin d'atteindre les objectifs de ce projet, nous avons utilisé d'autres bibliothèques externes pour effectuer certaines tâches spécifiques. En plus de celles fournies par la bibliothèque standard de Python.

1. **Matplotlib :**

Est probablement l'un des packages Python les plus utilisés pour la représentation de graphiques en 2D. Il fournit aussi bien un moyen rapide de visualiser des données grâce au langage Python, que des illustrations de grande qualité dans divers formats [41]

2. Seaborn :

Est une librairie qui vient s'ajouter à Matplotlib, remplace certains réglages par défaut et fonctions, et lui ajoute de nouvelles fonctionnalités. Seaborn vient corriger trois défauts de Matplotlib.

Seaborn fournit une interface qui permet de pallier ces problèmes. Il utilise toujours Matplotlib "sous le capot", mais le fait en exposant des fonctions plus intuitives. Pour commencer à l'utiliser, rien de plus simple. [42]

3. Scikit-learn :

Est une bibliothèque développée en Python, un langage de programmation de haut niveau. Elle est dédiée à l'apprentissage statistique (machine Learning) et peut être utilisée comme middleware, notamment pour des tâches de prédiction. [43]

4. NumPy :

Est le package fondamental pour le calcul scientifique avec Python. Il contient entre autres :

- un puissant objet tableau N-dimensionnel ;
- Fonctions sophistiquées (diffusion) ;
- Outils d'intégration de code C / C ++ et Fortran.

Outre ses utilisations scientifiques évidentes, NumPy peut également être utilisé comme un conteneur multidimensionnel efficace de données génériques. Des types de données arbitraires peuvent être définis. Cela permet à NumPy de s'intégrer de manière transparente et rapide à une grande variété de bases de données.

5. Pandas :

Est une librairie Python qui a pour objectif de vous faciliter la vie en matière de manipulation de données. Les structures de données gérées par Pandas peuvent contenir tout type d'éléments à savoir (dans le jargon Pandas) des Séries et Data Frame et des Panel. Dans le cadre de nos expérimentations on utilisera plutôt les Data frame car ils offrent une vue bidimentionnelle des données (comme un tableau Excel), et c'est exactement ce que l'on va chercher à utiliser pour nos modèles. [44]

6. SciPy:

SciPy est un logiciel open source pour les mathématiques, les sciences et l'ingénierie. La bibliothèque SciPy dépend de NumPy, qui fournit une manipulation de tableau N dimensionnelle pratique et rapide.

La bibliothèque SciPy est conçue pour fonctionner avec les baies NumPy et fournit de nombreuses routines numériques conviviales et efficaces telles que des routines pour l'intégration et l'optimisation numériques. [45]

7. **SKLEARN :**

Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs2 notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria3. Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. [46]

IV.3. Approche proposée

Tout d'abord il faut imaginer un chemin entre les données initiales et la valeur à prédire. Notre proposition se traite en problème supervisé et non supervisé car on cherche le plus souvent à reproduire un processus humain. L'aspect non supervisé intervient sous la forme d'une étape intermédiaire (clustering, k-means), avant d'appliquer l'aspect supervisé sous la forme de classification (SVM).

Pour fournir des données plus appropriées pour le classificateur, l'ensemble de données est passé par un groupe d'opérations de prétraitement (nettoyage et normalisation des donnés) figure 21. Les étapes sont présentées comme suite :

- ✓ La première étape consiste à diviser l'ensemble de données en un ensemble d'entités et des étiquettes correspondantes, stocke les ensembles d'entités dans la variable X et la série d'étiquettes correspondantes dans la variable y.
- ✓ Nettoyage des données par élimination des espaces blancs, certaines des étiquettes multi-classes du jeu de données incluent des espaces blancs.
- ✓ Encodage des étiquettes multi-classes de l'ensemble de données sont fournies avec les noms de l'attaque, qui sont des valeurs de chaînes. Ainsi, il est important de coder ces valeurs en valeurs numériques, afin que le classificateur puisse apprendre le numéro de classe auquel appartient chaque type. Par l'application du k-means. La conversation numérique est fait avec pandas.to_numeric () est l'une des fonctions générales de Pandas qui est utilisée pour convertir l'argument en type numérique.
- ✓ Normalisation des données, les données numériques de l'ensemble de données sont de différentes plages, ce qui pose certains défis au classificateur pendant la formation pour compenser ces différences. Ainsi, il est important de normaliser les valeurs de chaque attribut, de sorte que la valeur minimale de chaque attribut soit nulle, tandis que le maximum est un. Cela fournit des valeurs plus homogènes au classificateur tout en maintenant la relativité entre les valeurs de chaque attribut.

- ✓ Avec la technique SVM on a normalisé les sous-groupes des k-means, telle qu'il transformera les données de telle sorte que leur distribution aura une valeur de 0 et 1 ou bien 2 qu'on va l'expliquer après.

Nous évaluons l'ensemble de dataset et nous comparons les résultats du système de détection d'intrusions d'anomalies basé sur SVM, voir la figure suivante qui représente

L'enchaînement de la combinaison des K-MEANS et SVM :

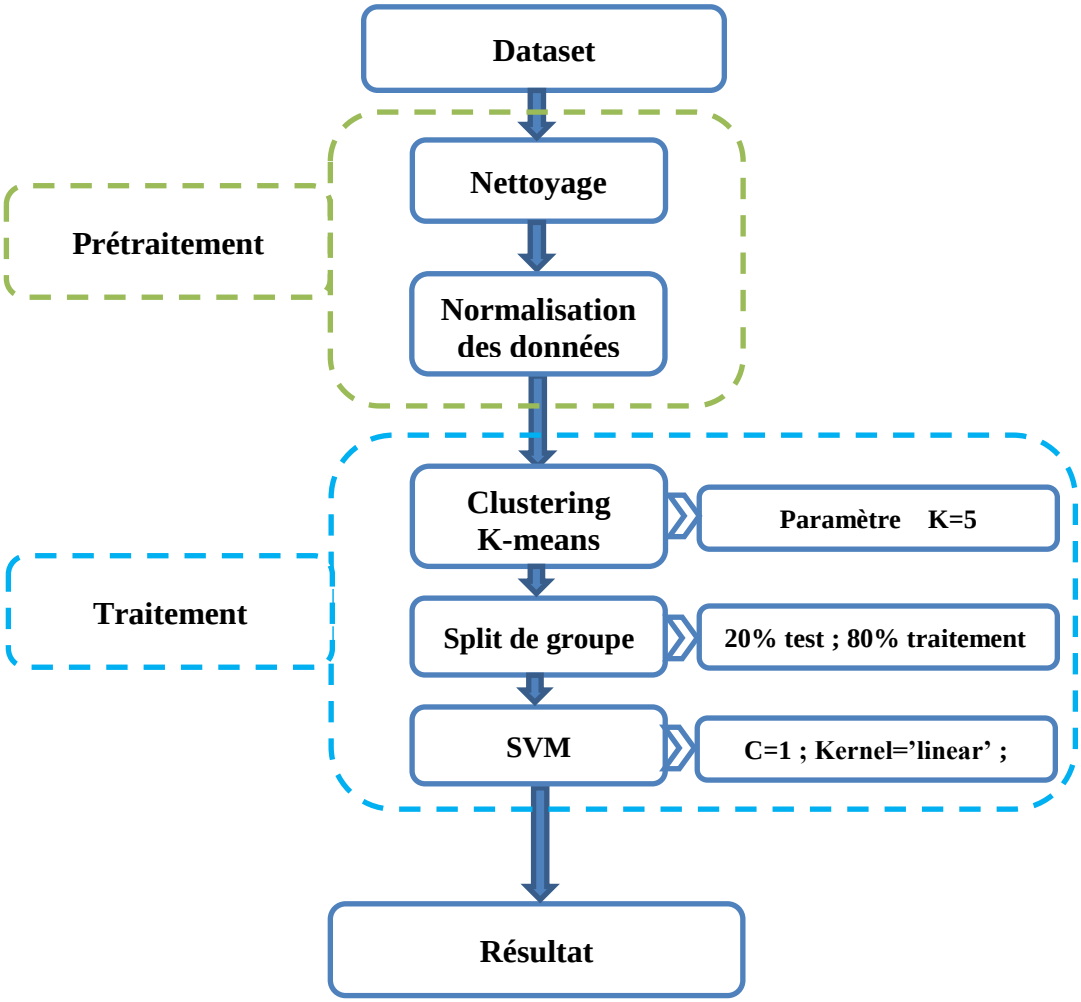


Figure 23. Model proposé.

IV.4. Implémentation

IV.4.1. Dataset

Notre dataset contient 54 lignes et 15 colonnes préparés au préalable par des chercheurs (mise à jour 2017-2019) qui ont utilisé des différentes technique pour définir les types d'IDS dans un model machine learning (supervisé et non supervisé).

Table	Author	Color	ShortTitle	Title	Year	Accuracy	Precision	Recall	F1_Score	Dataset	Reduction	Technique	nbre_of_feau	Time_training
1	Yulianto1	#FBEE23	ImprAda	Improving AdaBoost-based Intrusion Detection	2019	81,47	85,15	94,92	89,77	CIC IDS 2017	_	AdaBoost+EFS	25	_
2	Yulianto2	#FBEE23	ImprAda	Improving AdaBoost-based Intrusion Detection	2019	81,47	81,49	99,93	89,78	CIC IDS 2017	_	AdaBoost+EFS	25	_
3	Yulianto3	#FBEE23	ImprAda	Improving AdaBoost-based Intrusion Detection	2019	81,47	81,69	95,76	88,17	CIC IDS 2017	PCA	AdaBoost+SM	16	_
4	Yulianto4	#FBEE23	ImprAda	Improving AdaBoost-based Intrusion Detection	2019	81,83	81,83	100	90,01	CIC IDS 2017	PCA	AdaBoost	16	_
5	ImanAsh	#47DBCD	IMAN	IMAN	2018	77	77	84	77	CIC IDS 2017	_	AdaBoost	16	_
6	Hosseini1	#F3A0F2	hybrid	The hybrid technique for DDoS detection with s	2019	93,1	93,6	87,3	92,7	KNN-NSL	_	naïve Bayes	12	_
7	Hosseini2	#F3A0F3	hybrid	The hybrid technique for DDoS detection with s	2019	98,9	99,6	99,8	99,7	KNN-NSL	_	random fores	14	_
8	Hosseini3	#F3A0F4	hybrid	The hybrid technique for DDoS detection with s	2019	98,2	99,4	99,8	99,6	KNN-NSL	_	decision tree	10	_
9	Hosseini4	#F3A0F5	hybrid	The hybrid technique for DDoS detection with s	2019	96,1	93,4	91,8	94,9	KNN-NSL	_	MLP	20	_
10	Hosseini5	#F3A0F6	hybrid	The hybrid technique for DDoS detection with s	2019	97,7	99,8	99,8	99,8	KNN-NSL	_	K-NN	11	_
11	Zhoo1	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	96,9	94,7	99,5	97	KDDCup'99	CFS-BA	C4.5	12	0,36
12	Zhoo2	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	96,4	99,8	97,2	98,6	KDDCup'99	CFS-BA	RF	12	5,49
13	Zhoo3	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	97,5	99,8	99,8	99,8	KDDCup'99	CFS-BA	ForestPA	12	10,21
14	Zhoo4	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	97,6	99,8	99,8	99,8	KDDCup'99	CFS-BA	C4.5, FR, Fore	12	13,66
15	Zhoo5	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	95,7	99,7	92,9	96,2	KDDCup'99	CFS-BA	C4.5	41	2,53
16	Zhoo6	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	96,3	99,8	97,2	98,5	KDDCup'99	CFS-BA	RF	41	9,25
17	Zhoo7	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	94,9	99,7	94,9	95	KDDCup'99	CFS-BA	ForestPA	41	23,75
18	Zhoo8	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	95,2	99,7	95,3	97,3	KDDCup'99	CFS-BA	C4.5+FR+ Fore	41	30,43
19	Zhoo9	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	93,9	97	93,3	95,1	NSLKDD	CFS-BA	C4.5	41	1,56
20	Zhoo10	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	95,1	92,5	98,8	95,6	NSLKDD	CFS-BA	RF	41	10,64
21	Zhoo11	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	94,8	95,7	97	96,4	NSLKDD	CFS-BA	ForestPA	41	35,57
22	Zhoo12	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	95,2	99,7	95,3	97,4	NSLKDD	CFS-BA	C4.5+FR+ Fore	41	39,43
23	Zhoo13	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	98,7	99,1	98,8	98,9	NSLKDD	CFS-BA	C4.5	10	0,27
24	Zhoo14	#CE1317	Efficient	An Efficient Intrusion Detection System Based c	2019	98,8	99,1	98,8	98,9	NSLKDD	CFS-BA	RF	10	4,7

Figure 24. Apercu du dataset

Parmi les articles utilisés dans ce dataset on a :

L'ensemble de données CICIDS2017

Contient les attaques communes, qui ressemblent aux vraies données réelles .Il inclut également les résultats de l'analyse du trafic réseau à l'aide de CICFlowMeter avec des flux étiquetés basés sur l'horodatage, les IP source et de destination, les ports source et de destination, les protocoles et les attaques.

Le jeu de données CIDSID2017

Contient l'attaque la plus courante basée sur le rapport McAfee 2016 (Dos, DDos, Web based, Brute force, Infiltration, Heart-bleed, Bot et Scan) avec plus de 80 fonctionnalités extraites du trafic réseau généré.[49]

UNBS-NB. et KDDCup'99

Cet article a réalisé une analyse expérimentale des méthodes d'apprentissage automatique pour la détection d'attaques Botnet DDoS. L'évaluation est effectuée sur les UNBS-NB 15 et KDD99 qui sont des ensembles de données publicitaires bien connus pour la détection d'attaques Botnet DDoS. Les méthodes d'apprentissage automatique généralement Support Vector Machine (SVM)

DDoS Attack

Le déni de service distribué (DDoS) est une menace majeure parmi de nombreux problèmes de sécurité. Pour pallier ce problème, de nombreuses études ont été menées par des chercheurs, mais en raison de l'inefficacité de leurs techniques en termes de précision et de coût de calcul, proposer une méthode efficace pour détecter les attaques DDoS reste un sujet brûlant dans la recherche

Ces articles sont regroupés dans un tableau microsoft excel, par la suite dans notre programme on va importer et lire ce dataset dans un variable X sous forme d'une matrice dans un environnement de programme python comme suit :

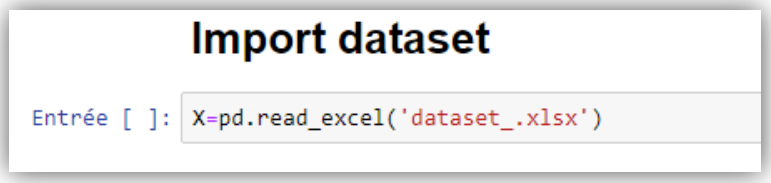


Figure 25. Importation dataset

Dans ce dataset il existe certains case vide remplacé par la suite par des '_' pour être normalisé.

IV.4.2. Indexation des types d'IDS

Dans ce stade là on a arrivé a commencé la manipulation sur le dataset, en commençant par une indexation des types d'IDS et les techniques

Indexing

Entrée [4]:

```
X['Type_IDS'] = X['Type_IDS'].replace(['NIDS'], '1')
X['Type_IDS'] = X['Type_IDS'].replace(['Hybrid'], '2')
X['Type_IDS'] = X['Type_IDS'].replace(['_'], '0')
X['Type_technique'] = X['Type_technique'].replace(['_'], '0')
X['Type_technique'] = X['Type_technique'].replace(['Classification'], '1')
X['Type_technique'] = X['Type_technique'].replace(['Clustring'], '2')
X['Type_Detection'] = X['Type_Detection'].replace(['Anomaly'], '2')
X['Type_Detection'] = X['Type_Detection'].replace(['Signature '], '1')
X['Type_Detection'] = X['Type_Detection'].replace(['Signature'], '1')
X['Type_Detection'] = X['Type_Detection'].replace(['_'], '0')
X['Reduction'] = X['Reduction'].replace(['_'], '0')
X['Time_training']  =X['Time_training'] .replace(['_'], '0')
X['Time_testing']  =X['Time_testing'] .replace(['_'], '0')
```

Figure 26. Indexation des types d'IDS.

IV.4.3. Traitement d'apprentissage (machine learning)

Pour arriver à des résultats satisfaites on mesure la performance du modèle sur la base de test.

Dans notre model on va passer par le clustering puis la classification

IV.5.3.1. Clustering (K-means)

Le clustering est un algorithme non supervisé il doit découvrir par lui-même la structure des données, à partir de l'échantillon d'apprentissage 'X' qu'on a, on applique le partitionnement en K-means par le choix des différentes valeurs de K et de calculer la variance des différents clusters pour arriver à la variance minimale qui définit le k idéal.

fonction clustermap,

Model K-means

```
Entrée [9]: modelkmeans= KMeans(n_clusters=5)
            kmeans_fit=modelkmeans.fit(features_X)
            kmeans_predict=modelkmeans.predict(features_X)
            modelkmeans.score(features_X)
Out[9]: -21457.1251641782
```

Figure 27. Fonction K-means.

On obtient la matrice de K-means suivante :

Matrix of K-means

```
Entrée [10]: sns.heatmap(features_X.corr())
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x497007a0b8>
```

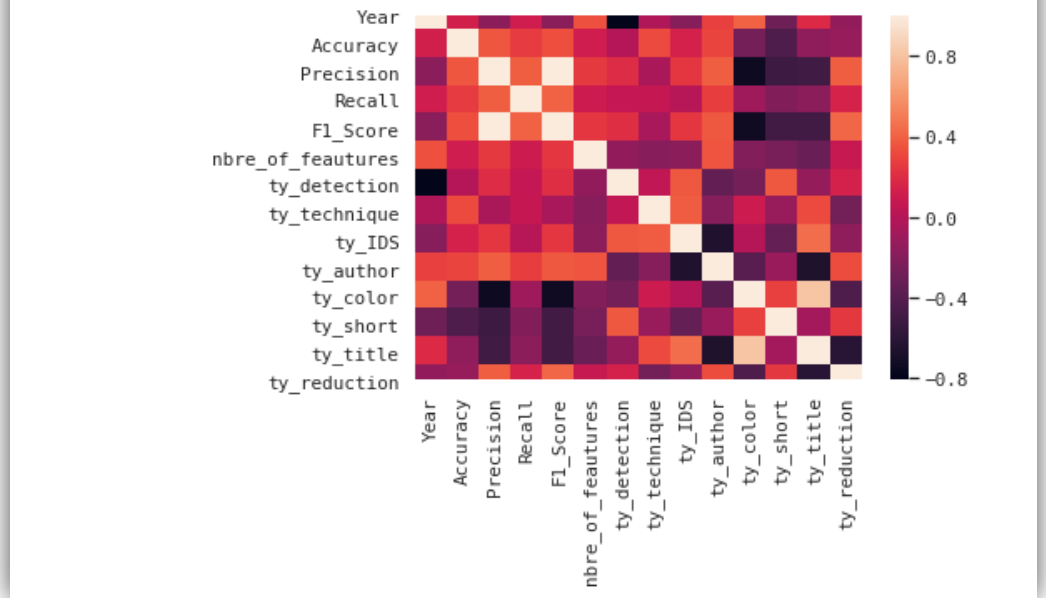


Figure 28. Matrice de K-means

Comparaison entre les vrais types d'IDS et les résultats du K-means :

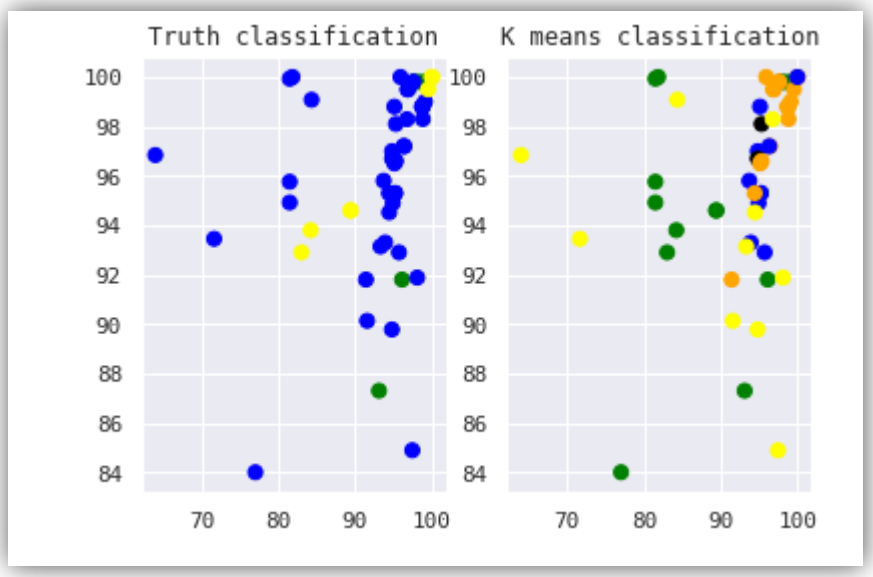


Figure 29. Comparaison entre les vrais types d'IDS et les résultats du K-means.

IV.5.3.2. Classification (SVM)

Par l'application de l'algorithme d'apprentissage automatique SVM sur les clusters (sous matrice) obtenus précédemment par le k-means (5 clusters) on suive les étapes suivantes :

Régularisation des paramètres (nettoyage des données):

SVM

```
Entrée [20]: C = 1.0  # SVM regularization parameter
Classifier = svm.SVC(kernel='linear', C=C, decision_function_shape='ovo', probability=True).fit(X_train, y_train)
#rbf_svc = svm.SVC(kernel='rbf', gamma=0.7, C=C).fit(X_train, y_train)
#poly_svc = svm.SVC(kernel='poly', degree=3, C=C).fit(X_train, y_train)
#lin_svc = svm.LinearSVC(C=C).fit(X_train, y_train)
SVMpredict=Classifier.predict(X_test)
Classifier.score(X_test, y_test)

Out[20]: 1.0
```

Figure 30. Fonction SVM.

Notre noyau va être linéaire, C'est une évaluation de à quel point nous voulons bien classer ou ajuster tout. Le domaine de ML est relativement nouveau et expérimental. Il existe de nombreux débats sur la valeur de C, ainsi que sur la façon de calculer la valeur de C. Nous allons nous en tenir à 1.0 pour l'instant, ce qui est un paramètre par défaut intéressant C'est égal à 1,0.

IV.4.4. Évaluation des résultats

Pour évaluer la qualité de sortie du classificateur on a utilisé :

La courbe précision- Accuracy/Recall montre le compromis entre précision et rappel pour différents seuils. Une zone élevée sous la courbe représente à la fois un rappel élevé et une haute précision, où une précision élevée est liée à un faible taux de faux positifs et un rappel élevé correspond à un faible taux de faux négatifs. Des scores élevés pour les deux montrent que le classificateur renvoie des résultats précis (haute précision), ainsi que la majorité de tous les résultats positifs (rappel élevé).

```
# Accuracy/Recall

Entrée [29]: from sklearn.metrics import auc, precision_recall_curve
from sklearn.multioutput import MultiOutputClassifier

y_pred2 = clf2.predict(X_test)
y_prob2 = clf2.predict_proba(X_test)
y_min2 = y_pred2.min()
import numpy
y_score2 = numpy.array( [y_prob2[i,p-y_min2] for i,p in enumerate(y_pred2)] )
y_score2[:5]

precision = dict()
recall = dict()
threshold = dict()
nb_obs = dict()

for i in clf2.classes_:
    precision[i], recall[i], threshold[i] = precision_recall_curve(y_test == i, y_score2)
    nb_obs[i] = (y_test == i).sum()

i = "all"
precision[i], recall[i], threshold[i] = precision_recall_curve(y_test == y_pred2, y_score2)
nb_obs[i] = (y_test == y_pred2).sum()
```

Figure 31. Fonction Accuracy/Recall

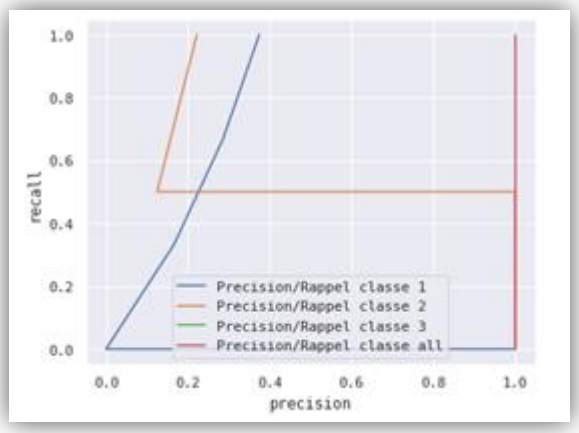


Figure 32. Graphes de précision.

Une fois on a appliqué notre modèle proposé sur le dataset, on a une certaine évaluation :

Procedure d'evaluation

```
Entrée [37]: def evaluation (model):
                 model.fit(X_train, y_train)
                 ypred=model.predict(X_test)

                 print(confusion_matrix(y_test, ypred))
                 print(classification_report(y_test, ypred))

                 N, train_score, val_score= learning_curve(model, X_train, y_train,
                                                           scoring='f1_weighted',
                                                           train_sizes=np.linspace(0.2, 1.0))
                 plt.figure(figsize=(12,8))
                 plt.plot(N, train_score.mean(axis=1), label='train score')
                 plt.plot(N, val_score.mean(axis=1), label='Validation score')
                 plt.legend()
```

Figure 33. Fonction d'évaluation.

Modelisation

```
Entrée [111]: evaluation(Classifier)
```

```
[[2 0 2]
 [0 3 0]
 [0 0 4]]
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         4
           1       1.00      1.00      1.00         3
           2       0.67      1.00      0.80         4

    accuracy                           0.82        11
   macro avg       0.89      0.83      0.82        11
weighted avg       0.88      0.82      0.81        11
```

Figure 34. Résultats d'évaluation.

En apprentissage automatique supervisé, la matrice de confusion est une matrice qui mesure la qualité d'un système de classification. Chaque ligne correspond à une classe réelle, chaque colonne correspond à une classe estimée. La cellule ligne L, colonne C contient le nombre d'éléments de la classe réelle L qui ont été estimés comme appartenant à la classe C1.

Un des intérêts de la matrice de confusion est qu'elle montre rapidement si un système de classification parvient à classifier correctement.

Comme montré dans la figure 35, la matrice de confusion obtenue a des résultats satisfaisants.

IV.5. Conclusion

Dans ce chapitre nous avons, en premier lieu, présenté les différents outils et langages que nous avons utilisés pour implémenter notre approche hybride de (K-means + SVM), et on a interprétée les résultats de la simulation.

Parmi les techniques de classification, le classifier SVM a atteint le taux de précision le plus élevé pour la détection et la classification de tous les types d'attaques dans les Cloud Computing.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

CONCLUSION GÉNÉRALE

Le Cloud Computing est une technologie qui permet aux clients ainsi qu'aux fournisseurs de services de bénéficier de capacités de traitement illimitées, avec des coûts d'utilisation et de déploiement très compétitifs. Les récents efforts visent à concevoir et à développer des services tout en se focalisant sur la définition des nouvelles méthodes, des politiques et des mécanismes efficaces. La sécurité contre les intrusions représente l'un des axes de recherches les plus importants dans ce domaine.

Ce mémoire a pour objectif de Proposer un outil qui va aider les chercheurs de trouver les solutions IDS pour le Cloud Computing, ainsi afin d'obtenir une base des métriques pour ces solution pour qu'on facilite la tâche de classification des IDS.

En premier lieu, nous avons traité la technologie du Cloud Computing, en clarifiant ses différents concepts, caractéristiques, modèles de services et de déploiements, afin de présenter les principaux défis du Cloud.

Après avoir introduit des généralités sur la sécurité des Cloud, nous avons présenté un état de l'art sur les IDS et les différentes approches utilisés dans la machine learning.

Enfin, nous avons traité notre dataset en passant par les étapes de traitement par machine learning.

Notre projet a été une opportunité pour approfondir nos connaissances dans le domaine de Machine Learning et d'apprendre ses différents modèles et leurs applications. Il est important pour nous de dire que l'un des avantages majeurs de ce travail est de familiariser avec la compréhension des articles et la maîtrise de plusieurs bibliothèques où nous avons vu et les exploiter pour la création des modèles.

L'approche proposée est de combiner des algorithmes d'apprentissage automatique supervisé et non supervisé afin de détecter les attaques. On s'attend à ce que le système proposé qui utilise une combinaison du K-means et de SVM propose aux clients des Cloud Computing une vue sur les types des systèmes de détection d'intrusion existants.

Comme perspective, on propose une réalisation d'un outil avec interface utilisateur (option génie logiciel) peut être développé avec différentes combinaisons d'algorithmes d'apprentissage automatique pour obtenir de meilleures performances peuvent et testé sur différents environnements Cloud.

RÉFÉRENCES BIBLIOGRAPHIQUES

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Di Marzo Serugendo. G. Cloud computing Architectures, services et risques.

[2] Zhou Xiangbing and Mao Fang. A semantics web service composition approach based on cloud computing. In Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pages 807–810. IEEE, 2012.

[3] Hassina Nacer and Djamil Aissani. Semantic web services : Standards, applications, challenges and solutions. Journal of Network and Computer Applications, 44 :134–151, 2014.

[4] Miranda Zhang, Rajiv Ranjan, Anna Haller, Dimitrios Georgakopoulos, Michael Menzel, and Surya Nepal. An ontology-based system for cloud infrastructure services' discovery. In Collaborative Computing : Networking, Applications and Worksharing (Collaborate-Com), 2012 8th International Conference on, pages 524–530. IEEE, 2012.

[5] A. Shawish and M. Salama. Cloud computing : Paradigms and technologies. Studies in Computational Intelligence, 2014.

[6] R DE LA ROSA-ROSERO. Découverte et sélection de services web pour une application mélusine. Master's thesis, l'Institut d'Informatique et de Mathématiques Appliquées de Grenoble, 2004.

[7] Le cloud computing : Réelle révolution ou simple évolution ? 2011.

[8] Allan Lefort Kuhn. Cloud computing. 2010.

[9] F. STEPHAN N. Couraud. CLOUD COMPUTING :Défnitions & Concepts,Enquête et Analyse des Tendances. 2010.

[10] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic. Cloud computing and emerging it platforms : Vision, hype, and reality for delivering computing as the 5th utility. Future Generation computer systems, 25(6) :599–616, 2009.

[11] Rajkumar Buyya, James Broberg, and Andrzej M Goscinski. Cloud computing : principles and paradigms, volume 87. John Wiley & Sons, 2010.

[12] Lizhe Wang, Jie Tao, Marcel Kunze, Alvaro Canales Castellanos, David Kramer, and Wolfgang Karl. Scientific cloud computing : Early definition and experience. In HPCC, volume 8, pages 825–830, 2008.

[13] Michael Hogan, Fang Liu, Annie Sokol, and Jin Tong. Nist cloud computing standards roadmap. NIST Special Publication, 35, 2011.

[14] Rajkumar Buyya, Chee Shin Yeo, and Srikumar Venugopal. Market-oriented cloud computing : Vision, hype, and reality for delivering it services as computing utilities. In High Performance Computing and Communications, 2008. HPCC'08. 10th IEEE International Conference on, pages 5–13. Ieee, 2008.

[15] Paul Mcfredries. Technically speaking : The cloud is the computer. Spectrum, IEEE, 45(8) :20–20, 2008.

[16] Jeremy Geelan et al. Twenty-one experts define cloud computing. Cloud Computing Journal, 2 :1–5, 2009.

[17] Rajkumar Buyya, Suraj Pandey, and Christian Vecchiola. Cloud Computing : First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. Proceedings, chapter Cloudbus Toolkit for Market Oriented Cloud Computing, pages 24–44. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[18] Justin M Grimes, Paul T Jaeger, and Jimmy Lin. Weathering the storm : The policy implications of cloud computing. in Proceedings of iConference, University of North Carolina (Chapel Hill, USA), 2009.

[19] Udeze Chidiebele. C Christiana. C. Okezie and Okafor Kennedy .C. Cloud computing : A cost effective approach to enterprise web application implementation : A case for cloud erp web model. Academic Research International, 03, July 2012.

[20] Ling Qian, Zhiguo Luo, Yujian Du, and Leitao Guo. Cloud computing : an overview. In Cloud computing, pages 626–631. Springer, 2009.

[21] Mao. J Bohn. R Messina. J Badger. L Leaf. D Liu. F, Tong. J. NIST Cloud Computing Reference Architecture. NIST, September 2011.

[22] Francesco Maria Aymerich, Gianni Fenu, and Simone Surcis. A real time financial system based on grid and cloud computing. In Proceedings of the 2009 ACM symposium on Applied Computing, pages 1219–1220. ACM, 2009.

[23] Rajkumar Buyya, Rajiv Ranjan, and Rodrigo N Calheiros. Intercloud : Utility oriented federation of cloud computing environments for scaling of application services. In Algorithms and architectures for parallel processing, pages 13–31. Springer, 2010.

[24] Rajkumar Buyya, Suraj Pandey, and Christian Vecchiola. Cloudbus toolkit for market oriented cloud computing. In Cloud Computing, pages 24–44. Springer, 2009.

[25] Hüsemann. S Chardonnens. T. Les enjeux du Cloud Computing en entreprise. Université de Fribourg, 2012.

[26] S. Karan and A. Kritika. Cloud computing. Engineering Studies and Technical Approach, 01, APril 2015.

[27] O.Markowitch, cours de ” cryptologie distribuée et protocoles ”, computer science department, université libre de Bruxelles, Belgium, 2009.

[28] W.Stallings, Network security Essentials, 2nd edition, prentice Hall, 2003.

[29] D.E Denning. ”An intrusion dection model” In :proccedings of the IEEE Transactions on software engineering, Septembre 2007 .

[30] H. GUILLAUME, Détection d’intrusions paramétrée par la politique de sécurité, soupélec, campus de rennes, équipe SSIR, 7 février 2005.

[31] H.Debar, M.Dacier, A.Wespi, a revised taxonomy for intrusion detection systems, annals des telecommunications, vol 55, 2000, NO 7-8,PP 361-378.

[32] Ludovic Mé. ” Détection des intrusions dans les systèmes d’information : la nécessaire prise en compte des caractéristique du système surveillé ”. Habilitation à diriger des recherches, université de Rennes 1 l’institut de formation Supérieure en informatique et en Communication de Rennes1, 2003.

[33] Nathalie Dagorn, ”Détection et prévention d’intrusion : présentation et limites”, Rapport de recherche, 2006. https ://hal.archives-ouvertes.fr/inria-00084202/document

[34] L. Sellami, D. Idoughi, PF. Tiako, An Intrusion Detection System Based on Nodes in Cloud Computing Environments. In : P. Ivànyi, B.H.V. Topping, (Editors), Proceedings of the Fourth International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering : Civil-Comp Press. Stirlingshire : UK. Paper 22, (2015).

[35] https://le-datascientist.fr/apprentissage-supervise-vs-non-supervise Accès le 31/03/2020 à 17:05.

[36] https://www.supinfo.com/articles/single/6041-machine-learning-introduction-apprentissage-automatique Accès le 29/03/2020 à 22 :28.

[37] https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/Accède le 09/05/2020 a 20 :26 .

[38] https://www.it-swarm.dev/fr/python/quel-est-le-lien-entre-anaconda-et-python/829887380/Accède le 09/05/2020 a 20 :26.

[39] https://www.yubigeek.com/developper-en-python-avec-anaconda/Accède le 09/05/2020 a 22 :29.

[40] https://www.stat4decision.com/fr/voila-dashboards-a-partir-de-vos-jupyter-notebooks/Accède le 09/05/2020 a 22 :30.

[41] https://python.developpez.com/tutoriels/graphique2d/matplotlib/Accède le 20/05/2020 a 22 :04.

[42] https://openclassrooms.com/fr/courses/4452741-decouvrez-les-librairies-python-pour-la-data-science/5559011-realisez-de-beaux-graphiques-avec-seaborn Accède le 20/05/2020 à 22:41.

[43] https://www.inria.fr/fr/lancement-de-linitiative-scikit-learn?fbclid=IwAR1r89W0NsQHyu7BN31qRQJq5YEUS0iORwj37i51Zj0ds35stAwHCL-8N8c Accède le 20/05/2020 a 22 :11.

[44] https://www.datacorner.fr/pandas_1/ Accède le 20/05/2020 a 22 :31.

[45] https://pypi.org/project/scipy// Accède le 20/05/2020 a 23 :18.

[46] https://www.geeksforgeeks.org/how-to-use-glob-function-to-find-files-recursively-in-python/ Accède le 24/05/2020 a 10 :51.

[47] Mohamadally Hasan,Fomani Boris :" SVM machine à vecteurs de support ou séparateur à vaste marge ". BD Web, ISTY3,Versailles St Quentin, France, janvier 2006.

[48] A. Cornuéjols :" Une nouvelle méthode d'apprentissage : Les SVM. Séparateurs à vaste marge". Université de Paris-Sud, Orsay, France, Juin 2002.

[49] https://www.unb.ca/cic/datasets/nsl.html IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB Accède le 17/03/2020 a 10 :51. »

ANNEXES

ANNEXES

Annexe A

Administrateur : Personne chargée de mettre en place la politique de sécurité, et par conséquent, de déployer et configurer les IDS.

Attaque : Synonyme d'intrusion. Exploit : terme utilisé pour désigner un programme d'attaque.

Analyseur : 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 à partir des événements remontés par les capteurs ou à partir d'alertes générées par d'autres analyseurs.

Opérateur : Personne chargée de l'utilisation du manager associé à l'IDS. Elle propose ou décide de la réaction à apporter en cas d'alerte. C'est parfois la même personne que l'administrateur.

Intrusion : Action (ou tentative d'action) qui a pour conséquence de compromettre l'intégrité, la confidentialité ou la disponibilité d'une ressource (violation de la politique de sécurité).

Signature : règle utilisée par certains analyseurs pour identifier parmi les activités surveillées celles qui sont caractéristiques d'une intrusion.

Détection d'intrusions : processus logiciel de recherche des intrusions qui s'appuie sur la surveillance des activités des entités dans les systèmes et les réseaux.

Approche comportementale : Ensemble des techniques utilisées par les IDS qui basent leur processus de détection sur l'hypothèse que toute déviation significative du comportement observé d'une entité par rapport à son modèle de comportement normal constitue une intrusion potentielle.

Système de détection d'intrusions : Ensemble complet composé de capteur(s), d'analyseur(s) et de manager(s).

Politique de sécurité : Spécification des règles à respecter dans le réseau d'une organisation, afin de garantir l'intégrité, la confidentialité et la disponibilité des ressources sensibles. Elle définit quelles activités sont autorisées et Les quelles sont interdites.

Faux positif : Alerte émise en présence d'une action légitime rapportée à tort comme étant une intrusion par un système de détection d'intrusions (fausse alerte).

Alerte : Message formaté qui décrit un événement relatif à une action qui compromet la sécurité d'un système ou d'un réseau. Les alertes sont produites par un analyseur.

Approche par scénarios : Ensemble des techniques utilisées par les IDS qui détectent les intrusions en recherchant dans les activités courantes celles qui sont caractéristiques de

scénarios d'attaques connus (comparaison avec une base de signatures d'attaques). Aussi appelée approche par signatures.

Capteur : Logiciel qui génère les événements enfutant et formatant les données brutes intéressantes provenant d'une unique source d'information (paquets du réseau, logs du système ou logs applicatifs).

Manager : Composant d'un IDS permettant à l'opérateur de gérer les autres composants. Ses fonctions comportent généralement la configuration des capteurs et analyseurs, la notification des alertes à l'opérateur et éventuellement la réaction.

Scénario : Suite des étapes d'une intrusion. Réaction : mesures passives ou actives qui peuvent être prises en réponse à la détection d'une attaque, pour la stopper ou pour corriger ses effets.

Sonde : Regroupement (logique ou fonctionnel) d'un capteur et d'un analyseur.

Faux négatif : Absence d'alerte en présence d'une action qui constitue bien une intrusion mais qui n'a pas été détectée comme telle par un système de détection d'intrusions.

Annexe B

Algorithme K-Means

Algorithme de Lloyd (1957), Forgy (1965), MacQueen (1967)

Algorithme particulièrement simple

Entrée : X (n obs., p variables), K #classes

Initialiser K centres de classes G_k

REPETER

Allocation. Affecter chaque individu à la classe dont le centre est le plus proche

Représentation. Recalculer les centres de classes à partir des individus rattachés

JUSQU'À Convergence

Sortie : Une partition des individus caractérisée par les K centres de classes Gk

Peut être K individus choisis au hasard. Ou encore, K moyennes calculées à partir d'une partition au hasard des individus en K groupes.

Variante MacQueen : remettre à jour les centres de classes à chaque individu traité. Accélère la convergence, mais le résultat dépend de l'ordre des individus.

Propriété fondamentale : l'inertie intra-classe diminue à chaque étape (nouvelles valeurs des barycentres conditionnels G_k)

Nombre d'itérations fixé
Ou aucun individu ne change de classe
Ou encore lorsque W ne diminue plus
Ou lorsque les G_k sont stables

Algorithme SVM

- Initialiser W aléatoirement
- Tant qu'il existe X tel que :

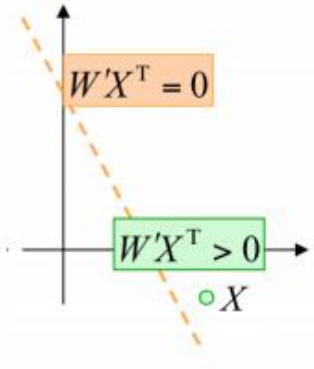
$$X \in C_1 \Rightarrow WX^T > 0 \text{ et } X \in C_2 \Rightarrow WX^T < 0 \text{ non satisfaite}$$

Faire :

$$W \leftarrow W + \lambda.\delta(X)X$$

λ petite constante

$$\begin{cases} X \in C_1 \Rightarrow \delta(X) = +1 \\ X \in C_2 \Rightarrow \delta(X) = -1 \end{cases}$$



$X \in C_1$ mais $WX < 0$

on cherche ΔW tel que $W'X^T=(W+\Delta W)X^T > 0$:

On a : $W'X^T=(W + \lambda.1.X)X^T = \left(WX^T + \lambda\|X\|^2\right) > WX^T$

$X \in C_2$ mais $WX^T > 0$

on cherche ΔW tel que $W'X^T=(W+\Delta W)X^T < 0$:

On a : $W'X^T=(W + \lambda.(-1).X)X^T = \left(WX - \lambda\|X\|^2\right) < WX^T$

Résumé :
Le problème posé par les utilisateurs de Cloud Computing c'est bien la sécurité de leurs données stockées, le système de détection des intrusions revient à jouer ce rôle. Le problème formulé peut être considéré comme un problème de classification, dont l'objectif est d'avoir une bonne optimisation dans lequel la fonction objective est de maximiser le taux de détection par type. Notre objectif consiste à faciliter le choix de la manière de sécuriser un Cloud Computing à l'aide d'un système de détection d'intrusions par la proposition d'un modèle pour avoir le meilleur résultat, et l'adaptabilité par des techniques de la machine Learning. Nous avons proposé un modèle IDS qui nous permet d'améliorer le taux de détection par rapport aux travaux précédents.
Le modèle proposé est une combinaison d'un algorithme d'apprentissage automatique supervisé et non supervisé afin de détecter les attaques. On s'attend à ce que ce système proposé qui utilise une combinaison du K-means et de SVM propose aux clients des Cloud Computing une vue sur les types des systèmes de détection d'intrusion existants.
Mots-clés:
Cloud, Systèmes de Détection d'Intrusion, Apprentissage Automatique, K-moyennes, Machines à vecteurs de support.

Abstract:
The problem posed by Cloud Computing users is the security of their stored data; the intrusion detection system comes down to playing this role. The formulated problem can be considered as a classification problem, the objective of which is to have a good optimization in which the objective function is to maximize the detection rate by type. Our objective is to facilitate the choice of the way to secure a Cloud Computing using an intrusion detection system by proposing a model to have the best result, and adaptability by techniques of the Machine Learning. We proposed an IDS model that allows us to improve the detection rate compared to previous work.
The proposed model is a combination of a supervised and unsupervised machine learning algorithm in order to detect attacks. The proposed system which uses a combination of K-means and SVM is expected to provide cloud computing customers with a view of the types of existing intrusion detection systems.

Keywords: Cloud, Intrusion Detection Systems, Machine Learning, K-means, Support vector machines.

ملخص:
المشكلة التي يطرحها مستخدمو الحوسبة السحابية هي أمن بياناتهم المخزنة ، ونظام كشف التطفل يلعب هذا الدور. يمكن اعتبار المشكلة المصاغة على أنها مشكلة تصنيف ، والهدف منها هو الحصول على تحسين جيد حيث تكون الوظيفة الموضوعية هي زيادة معدل الاكتشاف حسب النوع. هدفنا هو تسهيل اختيار طريقة تأمين الحوسبة السحابية باستخدام نظام كشف التطفل من خلال اقتراح نموذج للحصول على أفضل نتيجة ، والقدرة على التكيف من خلال تقنيات التعلم الالي. اقترحنا نموذجا IDS الذي يسمح لنا بتحسين معدل الكشف مقارنة بالأعمال السابقة.
النموذج المقترح عبارة عن مزيج من خوارزمية تعلم آلي خاضعة للإشراف وغير خاضعة للإشراف من أجل اكتشاف الهجمات. من المتوقع أن يوفر النظام المقترح الذي يستخدم مزيجًا من K-mean و SVM لعملاء الحوسبة السحابية عرضًا لأنواع أنظمة اكتشاف التسلل الحالية.
الكلمات الرئيسية: سحابة، أنظمة كشف التسلل ، التعلم الالي ، K-mean، SVM.