الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العلي و البحث العلمي
Ministry of Higher Education and Scientific Research.

Nº Ref: ……

Université

Abd Elhafid Boussouf

Mila

Institut des sciences et technologies Département de Mathématique et Informatique

Mémoire de fin d'études
en vue de l'obtention du diplôme de Master en Informatique
Spécialité :
Science et Technologies de l'Information et de la Communication (STIC)

Détection des attaques DDOS dans le Cloud Computing

Présentée par: -Fethi Amira
-Litim Yousra

Soutenue devant le jury composé de:

Mme Bouchemal Nardjes	**MCA Université A. Boussouf, Mila**	**Président**
Mme Boufaghes Hamida	**MAA Université A. Boussouf, Mila**	**Examinateur**
Mme Hedjaz Sabrine	**MCA Université A. Boussouf, Mila**	**Encadreur**

Année universitaire :2020/2021

Résumé

Le cloud computing est une technologie florissante et adoptée par de nombreuses entreprises. Cependant, cette évolution a rendu le système du cloud computing hautement vulnérable à plusieurs menaces de sécurité. Le cloud Computing l'un des techniques les plus évolué il prend de plus en plus sa place dans le monde car il fournit à un grand nombre d'utilisateurs des services performants et stockage de données évolutif.

Cependant cette évolution a rendu le système de cloud computing avec une sécurité très vulnérable surtout qu'il y a les menaces des attaques ddos. L'attaque DDoS est l'une des techniques de piratage les plus puissantes sur cloud. L'arme de base que le pirate utilise ces types d'attaques essentiellement pour consommer les ressources du cloud.

Ce travail effectué dans le cadre de ce mémoire vient répondre aux problèmes des attaques ddos dans le cloud computing et a pour objectif d'implémenter une solution intelligente de détection des attaques ddos via les différentes modèles du machine learning.

Nous avons pris l'initiative de présenter les différents concepts du cloud computing, puis on s'est concentrée sur les problèmes et les solutions de ces attaques. Dans ce contexte on a utilisé la dernière data set ddos pour faire un systéme de détection des attaques ddos a base machine learning où on à utilisé plusieurs algorithmes tell que l'arbre de décision, l'algorithme des K plus proches voisins, régression Logistique, gaussienne NB etc.

Abstract

Cloud computing is a flourishing technology adopted by many companies. However, this development has made the cloud computing system highly vulnerable to several security threats. Cloud Computing, one of the most advanced techniques, is increasingly taking its place in the world because it provides a large number of users with high-performance services and scalable data storage.

However this evolution has made the cloud computing system with very vulnerable security especially as there are threats from ddos attacks. DDoS attack is one of the most powerful hacking techniques in the cloud. The basic weapon that the hacker uses these types of attacks is basically to consume cloud resources.

This work carried out within the framework of this dissertation answers the problems of ddos attacks in cloud computing and aims to implement an intelligent solution for detecting ddos attacks via the different models of machine learning.

We took the initiative to present the different concepts of cloud computing, then focus on the problems and solutions of these attacks. In this context, we used the latest ddos data set to make a machine learning-based ddos detection system or to use several algorithms such as the decision tree, the K nearest neighbors algorithm, Logistic regression, Gaussian NB etc. ,and the results sound good compared to other articles.

Remerciements

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui nous voudrons témoigner toute ma gratitude.

Tout d'abord, nous tenons à exprimer notre gratitude à la directrice de ce mémoire, Mme Sabrine Hedjaz, pour sa patience, sa disponibilité et, surtout, ses conseils avisés qui ont contribué à notre réflexion. Nous adressons nos sincères remerciements à tous les professeurs, conférenciers, et toutes les personnes qui ont orienté leurs réflexions, écrits, conseils et critiques et ont accepté de nous rencontrer et de répondre à nos questions au cours de nos recherches, en particulier Mme Nerdjes Bouchmal et Mme Hamida Boufagas. Nous remercions nos chers parents qui ont toujours été là pour nous. Nous remercions mes sœurs et frères pour leurs encouragements.

Dedicace 1

je tiens a dédie ce modeste travail :

- A l'amour de ma vie, mon adorable mère Naziha, pour son amour, ses encouragement et ses sacrifices.
- A mon cher père Abdelaziz qui m'a supporté, m'a dirigé et m'a accordé sa confiance.
- A mes frères Mohamed Aymen et Taki eddine que dieu les protèges.
- A la joie de ma vie, ma petite sœur Youssra, tu connais ta place dans mon cœur..je t'aime
- A ma chére grande mère, que Dieu lui donne une longue et joyeuse vie.
- A tous mes amies en particulier Marwa et Hadjer.
- A mon chère binôme yousra.
- On tient à remercier tout particulièrement notre encadrante Dr.Hadjez Sabrine pour nous avoir suivis et conseillés tout au long de la réalisation de ce mémoire.

Amira

Dedicace 2

Avec l'expression de ma reconnaissance, je dédie ce modeste travail à ceux qui, quels que soient les termes embrassés, je n'arriverais jamais à leur exprimer mon amour sincère.

- A la femme qui a souffert sans me laisser souffrir, qui n'a jamais dit non à mes exigences et qui n'a épargné aucun effort pour me rendre heureuse ; Quoi que je fasse ou que je dise, je ne saurai point te remercier comme il se doit. Ton affection me couvre, ta bienveillance me guide et ta présence à mes côtés a toujours été ma source de force pour affronter les différents obstacles : mon adorable mère Assia.
- A l'homme, mon précieux offre du dieu, vous avez été à mes côtés pour me soutenir et m'encourager. Que ce travail reflète ma gratitude et mon amour pour toi : mon cher père Ammar.
- A ma chère sœur Lamis qui n'a pas cessée de me conseiller, encourager et soutenir tout au long de mes études. Que Dieu les protège et leurs offre la chance et le bonheur.
- A mes adorables frères mouhcen zinedine et mouhamed abdelhay, qui sait toujours comment procurer la joie et le bonheur pour toute la famille.
- A mes grands-parents, Que Dieu leur donne une longue et joyeuse vie. Merci pour leurs amours et leurs encouragements.
- A tous ceux que j'aime et ceux qui m'aiment.
- Sans oublier mon binôme Amira pour son soutien moral, sa patience et sa compréhension tout au long de ce projet.

Yousra

Table des matières

Table des figures

Liste des tableaux

Liste des symboles et abréviations

DDoS : Déni de service distribué.
DoS : Déni de service.
MSSQL : MSSQL signifie Microsoft SQL.
SSDP : protocole de découverte de services simple.
NTP : protocole de temps réseau.
TFTP : protocole de transfert de fichiers trivial.
DNS : nom de domaine du service.
LDAP : protocole d'accès aux annuaires léger.
NetBIOS : système d'entrée/sortie de base du réseau.
SNMP : protocole de gestion de réseau simple.
SYN : une attaque par inondation SYN est également appelée attaque semi-ouverte.
UDP : protocole de datagramme utilisateur.
UDP-Lag : cela peut être un problème grave lorsque le serveur nécessite un court délai.

Introduction Géneral

La technologie du cloud computing est devenue plus populaire aujourd'hui, car il fournit à un grand nombre d'utilisateurs des services hautes performances et des supports de stockage de données évolutifs. Pour que le Cloud soit utilisable,il faut être correctement connecté à Internet car La disponibilité du Cloud est un facteur primordial de la confiance qu'on peut avoir en un fournisseur donné.

Malheureusement, l'un des problèmes de sécurité critiques qui menacent le cloud computing est les attaques DDos qui sont devenues l'une des menaces puissantes et dangereuses pour le cloud computing , il peut utiliser des centaines de milliers d'hôtes compromis situés dans différents réseaux dans le monde pour attaquer la cible dans le cloud en utilisant l'usurpation d'adresse IP, le nombre d'attaques DDoS signalés a considérablement augmenté. La plus grande attaque signalée par un fournisseur de services en 2017 était de 600 Gbit/s. ce qui rend la détection de la source de l'attaque complexe et difficile. Il refuse également aux utilisateurs légitimes l'accès aux services fourni par le cloud ou arrête une organisation pour une longue période de temps. En 2016, pendant deux heures une attaque DDoS a été menée sur Amazon, Twitter et Spotify, ce qui a entraîné d'énormes pertes financières a cause de l'interruptions de service [1].

Par conséquent, les mécanismes de défense traditionnel sont devenue inefficace contre ce genre d'attaque c'est pour ça la plupart des études ont recours à des solutions intelligents comme data mining, deep learning et machine learning etc., dans ce contexte on va présenté une solution intelligente de détection des attaques ddos via les différentes modèles du machine learning, car c'est l'un des piliers les plus courants de l'IA, Elle est capable d'analyser des données complexes non structurées et fournit une solution aux tâches avec des règles difficiles à définir. Afin d'obtenir des bons résultats nous suivons les étapes de développement d'un projet machine learning.

Ce mémoire compose de cinq chapitres organisés de la manière suivante : Le premier chapitre définit un concept général du cloud computing, le deuxième est dédié à ce qui concerne les attaques ddos et les modes de ses attaques ainsi que les types des attaques ddos et concentre sur les problèmes et les solutions des attaques ddos dans le cloud computing, le troisième chapitre décrit l'importance de machine learning et les algorithmes de classifications et le quatrième chapitre présente la partie la plus importante de notre projet qui présente la conception et et la réalisation du notre projet, enfin le dernier chapitre consacré pour la présentations des résultats finale que nous avons obtenus dans notre projet ainsi que la comparaison avec les résultats d'une autre étude.

Chapitre 1

Cloud Computing

1.1 Introduction

Ces dernières années, le cloud computing est devenu un nouveau modèle de gestion et d'utilisation des systèmes informatiques. cette technologie permettant de délocaliser les données et les applications sur des infrastructures dématérialisées accessibles depuis internet. Dans ce chapitre, nous allons expliquer cette nouvelle technologie.

1.2 Historique du Cloud Computing

Avant l'idée du cloud computing a pris naissance, les informaticiens utilisaient déjà des services de Cloud computing comme le webmail2, le stockage de données en ligne (photos, vidéos, etc.) et même le partage d'informations sur les réseaux sociaux dans les années 90, un autre concept avait déjà préparé le terrain au Cloud computing. Il s'agit de L'ASP (Application Service Provider) qui permet aux clients de louer l'accès au logiciel installé sur le serveur distant du fournisseur de services sans avoir à installer le logiciel sur leurs propres ordinateurs.le cloud computing fournit ici le concept de flexibilité, et de nouveaux utilisateurs et de nouveaux services peuvent être ajoutés en cliquant sur la souris.

Le concept du cloud computing a été initié par Amazon en 2002, ce dernier avait alors investi dans un parc informatique afin de pallier les surcharges des serveurs dédiés au commerce en ligne constatées durant les fêtes de fin d'année. Mais une fois que les fêtes de fin d'année sont passées les ressources informatiques d'Amazon restait sont peu utilisées. Amazon a eu l'idée de louer ses capacités informatiques le reste de l'année à des clients pour qu'ils stockent les données et qu'ils utilisent les serveurs.Ces services étaient accessibles via Internet et avec une adaptation en temps réel de la capacité de traitement, le tout facturé à la consommation.
Cependant, ce n'est qu'en 2006 que Amazon comprit qu'un nouveau mode de consommation de l'informatique et d'internet faisait son apparition .
le Cloud computing est enfin apparu avec les différents progrès technologiques réalisés durant ces 50 dernières années, tant que sur le plan matériel, logiciel et conceptuel, qu'aux avancées des mécanismes de sécurité à l'élaboration de réseaux standardisés comme Internet, et à l'expérience dans l'édition et la gestion de logiciels, services, infrastructures et stockage de données .[2]

1.3 Définition du Cloud Computing

Le cloud computing est une infrastructure dans laquelle la puissance de calcul et le stockage sont gérés par des serveurs distants auxquels les utilisateurs se connectent au serveur distant via une liaison internet sécurisée. L'ordinateur de bureau ou portable, le téléphone mobile, la tablette tactile et autres objets connectés deviennent des points d'accès pour exécuter des applications ou consulter des données qui sont hébergées sur les serveurs. Le cloud se caractérise également par sa souplesse qui permet aux fournisseurs d'ajuster automatiquement la capacité de stockage et la puissance de calcul en fonction des besoins des utilisateurs.[3]

Selon la définition du National Institute of Standards and Technology (NIST), le Cloud computing est l'accés via un réseau de télécommunications, à la demande et d'une manière ubiquitaire a des ressources informatiques partagées et configurables (par exemple : réseaux, serveurs, stockage, applications et services). Il s'agit donc d'une délocalisation de l'infrastructure informatique.[4]

1.4 Les différents modèle de service du Cloud Computing

Le cloud computing offre trois modèles de services qui sont : IaaS, PaaS et SaaS : [5]

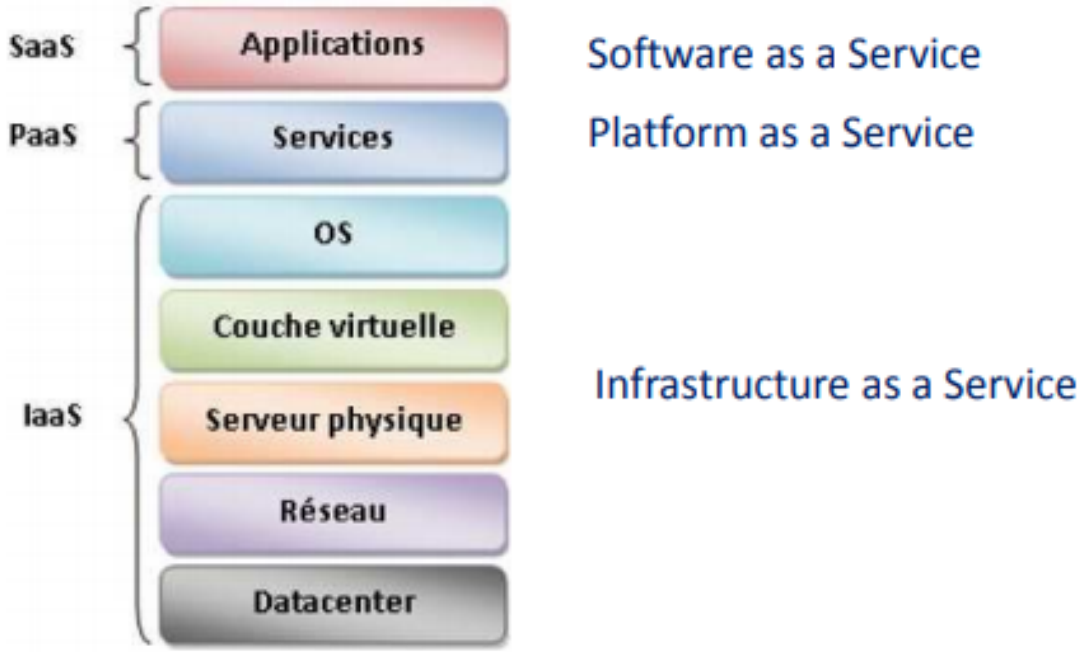

FIGURE 1.1 – modèles de services

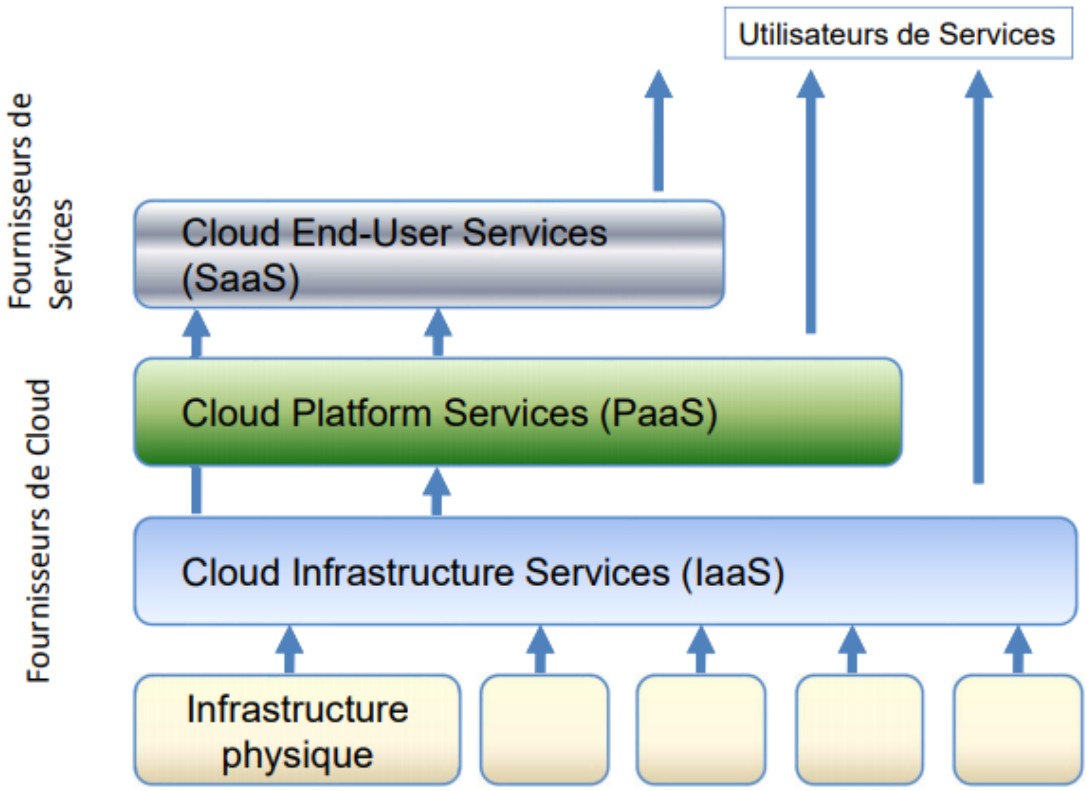

FIGURE 1.2 – Laboratoire d'infrastructure automatisé

1.4.1 SaaS (Software as Service)

Ceci est le plus haut niveau de la couche de la pile du cloud directement utilisé (consommé) par l'utilisateur final.
Le software as a service (SaaS) est accessible aux entreprises et il est facturé au nombre d'utilisateurs. L'entreprise loue les applications du fournisseur de services, il n'est plus besoin d'acheter un logiciel.
Ces applications sont accessibles via différentes interfaces, navigateurs web, clients légers, etc.,parmi les fournisseurs SaaS existants : salesForce Officz Live, la figure ci-dessous présente 'Platform as a Service'.[5]

FIGURE 1.3 – SaaS (Software as Service)

1.4.2 PaaS (Platform as a Service)

La platform as a Service (PaaS) est un service cloud qui présente une plateforme partagées sur laquelle des développeurs ou éditeurs de logiciels peuvent déployer ses propres applications.
L'entreprise loue un environnement middleware avec une infrastructure masquée. Parmi les fournisseurs PaaS existants : Google, AppEngine, Microsoft Azure, la figure 1.4 ci-dessous présente 'Platform as a Service', la figure ci-dessous présente 'Platform as a Service'.[5]

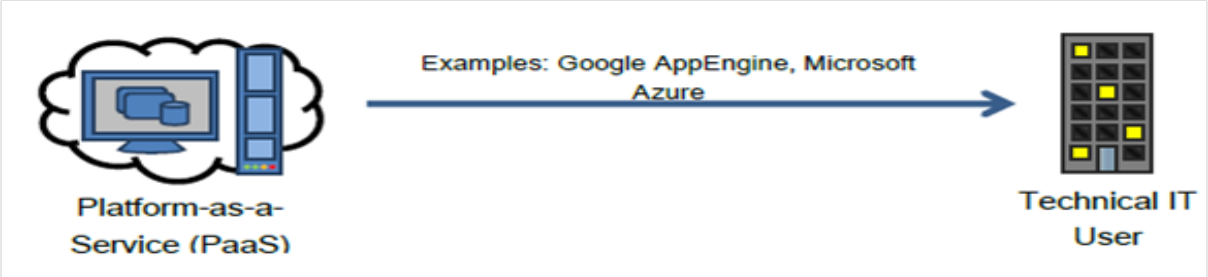

FIGURE 1.4 – PaaS (Platform as a Service)

1.4.3 IaaS (Infrastructure as a Service)

Ceci est la couche de base du modèle de la pile du cloud.
L'infrastructure as à Service (IaaS) est la mise à disposition par l'internet d'infrastructures facilement modifiables et hautement disponible.
L'entreprise loue ainsi des capacités de traitement, de stockage et autres infrastructures qu'elle peut structurer et gérer de façon autonome coté logiciel dès le système d'exploitation. Parmi les fournisseurs IaaS existants : Amazon EC2 et s3, la figure ci-dessous présente 'Infrastructure as à Service'.[5]

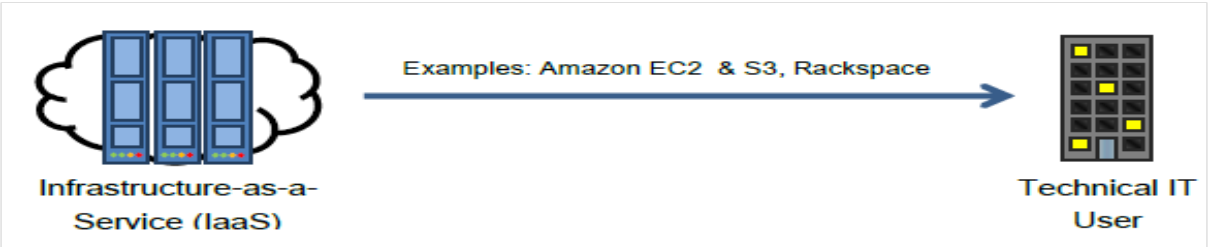

FIGURE 1.5 – IaaS (Infrastructure as a Service)

Cette figure présente une vue d'ensemble des zones de contrôle entre le client et le fournisseur en fonction du service offert sur le cloud :

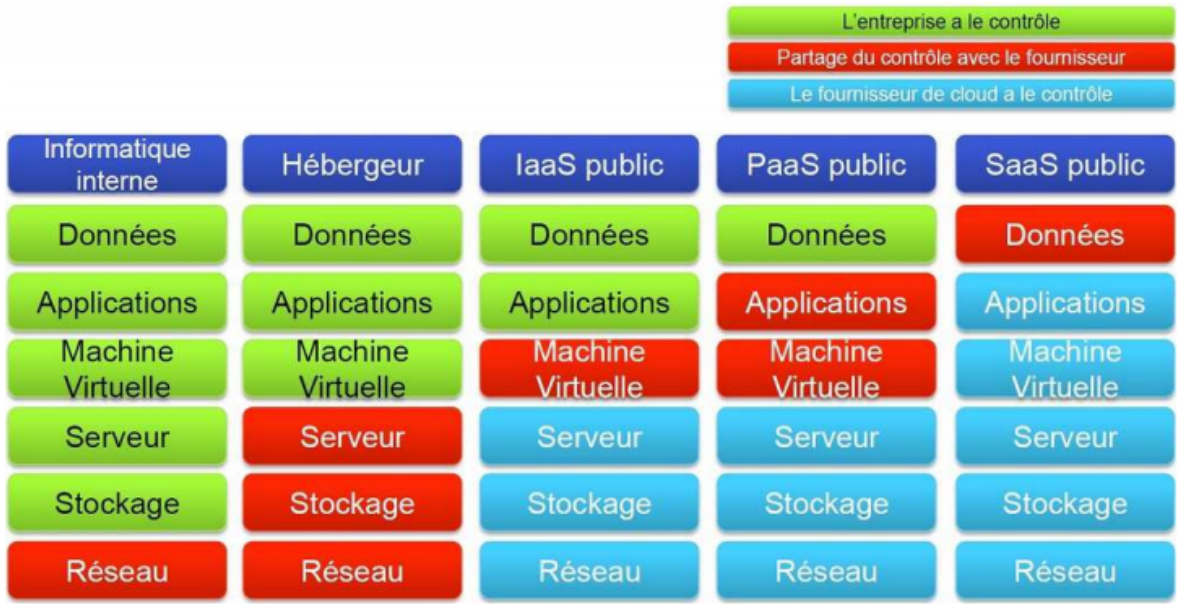

FIGURE 1.6 – Résponsabilité du cloud computing dans l'entreprise

Les services du cloud computing offrent plusieurs avantages aux clients en économisant du temps et de l'argent, néanmoins des inconvénients sont soulevés qui limite leur utilisations par les clients. Les avantages et inconvénients des services du cloud sont présentés dans le tableau suivant : [6]

- **Avantages et Inconvénients des services :**

Types	Avantages.	Inconvénients .
Saas	- Pas d'installation . – Plus de licence . – Migration .	– Logiciel limité. – Besoin de sécurité . – Dépendance des prestataires. – Services dédiés.
Paas	- Ne nécessite pas d'infrastructure. – Pas d'installation . – Environnement hétérogène .	– Limitation des langages . – Pas de personnalisation dans la configuration des machines virtuelles.
Iaas	– Administration . – Personnalisation . – Flexibilité d'utilisation .	– Besoin de sécurité. – Besoin d'un administrateur Système .

TABLE 1.1 – Avantages et Inconvénients des services de cloud computing [6].

1.5 Caractéristiques du Cloud Computing

En étudiant divers services disponibles sur le cloud, un ensemble de caractéristiques commun peut être mis en évidence : [7]

1.5.1 Accès instantané

Pour utiliser un service Cloud, il suffit généralement de se créer un compte qui sera directement utilisable une fois le mode de paiement est validé. Le service cloud est accessible lorsque l'utilisateur le souhaite.[7]

1.5.2 Self service

Les capacités de calcul et les ressources sont misent à disposition des clients , au besoin sans la nécessité de l'intervention du prestataire du services. [7]

1.5.3 Élasticité

Dans le cloud l'utilisateur(client)choisit les ressources à consommer ainsi que la quantité. Celle-ci peut varier à tout moment en fonction des besoins.[7]

1.5.4 Paiement à la carte

Les clients du cloud ont une facture qui ne regroupe le listing des services consommés. Tout ce qui touche à la maintenance, au personnel ou aux infrastructures matérielles n'apparaît pas dans la facture. la facturation se fait généralement à court terme, par exemple mensuellement.[7]

1.5.5 Service mesuré

Possibilité de surveiller, contrôler et mesurer l'utilisation des ressources l'argument en faveur de la valeur, de la flexibilité et de la qualité apportée par ces fonctionnalités est si convaincant que la transition vers le cloud computing va définir le paysage technologique des administrations au cours de la prochaine décennie. Toutefois, la transition vers le cloud computing présente des difficultés pour certaines administrations. Bon nombre d'entre elles sont découragées par la complexité des questions liées à la sécurité des données, aux rôles et aux modèles commerciaux.[7]

1.6 Architecture de référence du cloud computing

L'architecture globale du cloud computing comporte essentiellement :

Acteur	Définition .
Cloud Consumer	Une personne ou une organisation qui entretient une relation d'affaires avec et utilise le service de fournisseurs de cloud.
Cloud Provider	Une personne, une organisation ou une entité chargée de rendre un service disponible pour les parties intéressées.
Cloud Auditor	Une partie qui peut mener une évaluation indépendante des services cloud, les opérations du système d'information, les performances et la sécurité de la mise en œuvre du cloud.
Cloud Broker	Une entité qui gère l'utilisation,les performances et la livraison du cloud services et négocie les relations entre les fournisseurs de cloud et consommateurs cloud.
Cloud Carrier	Un intermédiaire qui assure la connectivité et le transport du cloud services des fournisseurs de cloud aux consommateurs de cloud.

TABLE 1.2 – Acteurs du Cloud Computing.

1.6.1 Cloud Consumer

Le consommateur du cloud est la principale partie prenante du service de cloud computing. Un consommateur de cloud représente une personne ou une organisation qui entretient une relation commerciale avec et utilise le service depuis un fournisseur de cloud.[8]

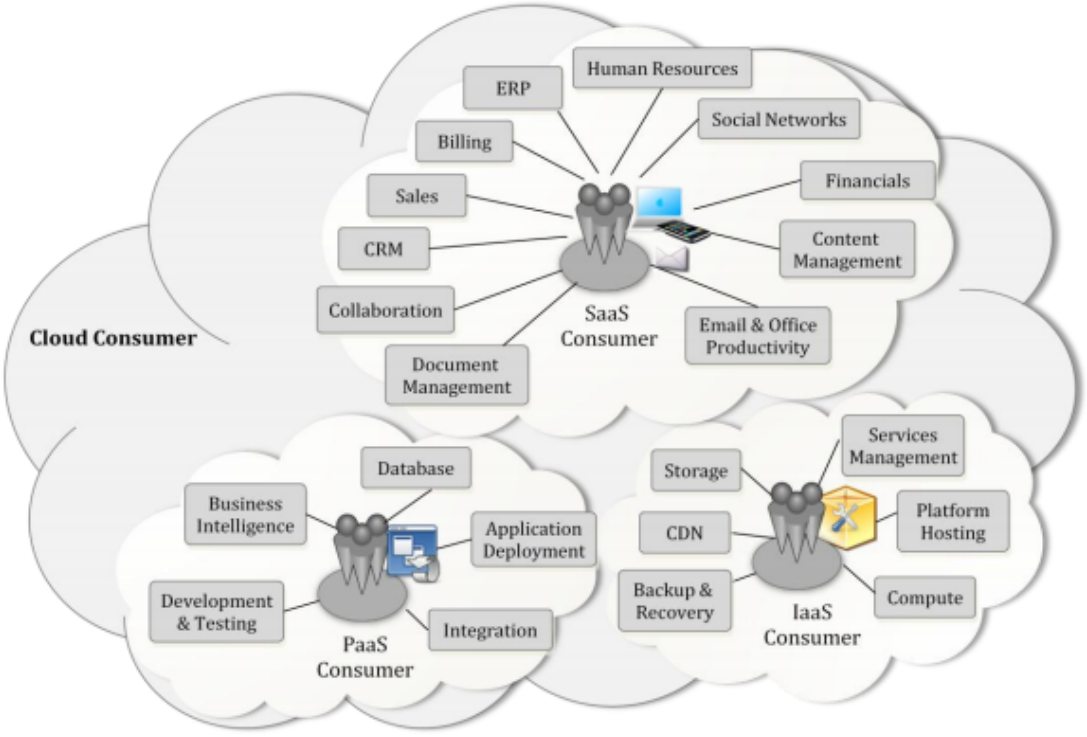

FIGURE 1.7 – Exemples de services disponibles pour un Cloud Consumer

1.6.2 Cloud Provider

Un fournisseur de cloud est une personne, une organisation, c'est l'entité responsable de la mise à disposition d'un service parties intéressées.
Un fournisseur de cloud acquiert et gère l'infrastructure informatique requise pour fournir les services, exécuter le logiciel cloud qui fournit les services et prendre des dispositions pour fournir les services cloud aux consommateurs cloud via un accès réseau.[8]

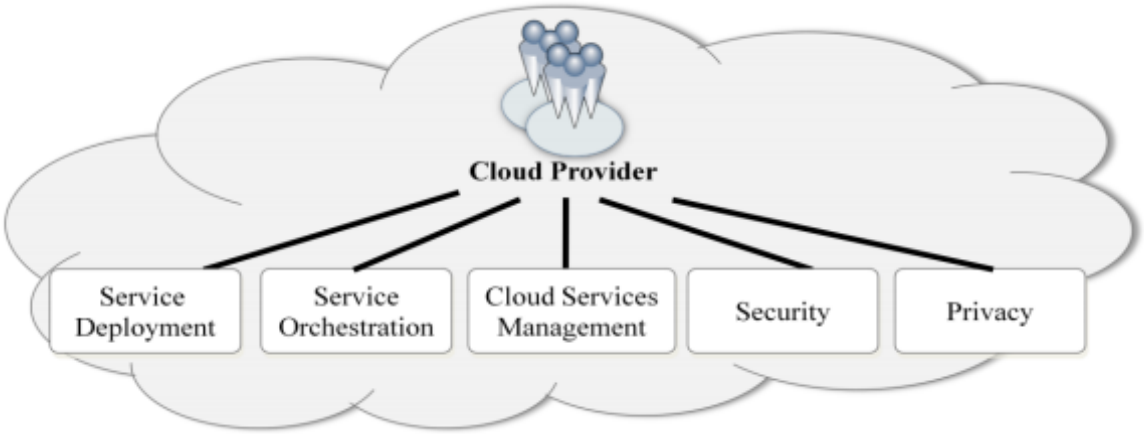

FIGURE 1.8 – Cloud Provider - Activités principales

1.6.3 Cloud Auditor

C'est la partie qui peut procéder à une évaluation indépendante des services fournis par le cloud computing, des opérations des systémes d'informations, de la vices fournis par le cloud computing, de la performance et de la sécurité de l'implémentation du cloud.[9]

1.6.4 Cloud Broker

C'est l'entité qui gère l'utilisation, la performance et la prestation de services de cloud computing, et qui négocie les relations entre les fournisseurs de services cloud et les client du cloud.[10]

1.6.5 Cloud Carrier

C'est l'entité intermédiaire qui fournit la connectivité et le transport des services cloud entre consommateurs de cloud et fournisseurs de cloud.
Les opérateurs de cloud offrent un accès aux consommateurs via le réseau, télécommunications et autres dispositifs d'accès. Par exemple, les consommateurs du cloud peuvent obtenir des services cloud via des appareils d'accès au réseau, tels que des ordinateurs, des ordinateurs portables, des téléphones mobiles, des appareils Internet mobiles (MID), etc.[8]

1.7 Les modèle de déploiement de Cloud Computing

Pour le grand public, le cloud computing fait référence globalement à internet, pour les entreprises il n'est pas le cas. Pour cela différents modèles de déploiement du cloud existent :[13]

1.7.1 Cloud privé

Les services et ressources cloud peuvent être utilisés pour une seule organisation (client / entreprise). Il peut être géré par l'organisation elle-même (cloud privé interne) ou par tiers (cloud privé externe). Dans ce dernier cas, l'infrastructure est entièrement spécifique à l'entreprise et est accessible via un réseau sécurisé tel qu'un VPN (Virtuel Private Network).[13]

1.7.2 Cloud public

Les services et ressources sont accessibles via internet et gérés par des prestataires de services externes (fournisseurs). Ces ressources et services sont partagés et utilisés par plusieurs clients sur demande. Ces services peuvent être gratuits ou payants.[13]

1.7.3 Cloud hybride

Il s'agit d'une combinaison de deux ou plusieurs clouds (cloud public et cloud privé) pour collaborer, communiquer et partager des applications et des données entre eux.[13]

1.7.4 Cloud communautaire

L'infrastructure cloud communautaire est partagée par plusieurs organisations ayant des intérêts communs (ex : exigences de sécurité, conformité ...). Comme un cloud privé, il peut être géré par l'organisation elle-même ou par un tiers.[13]

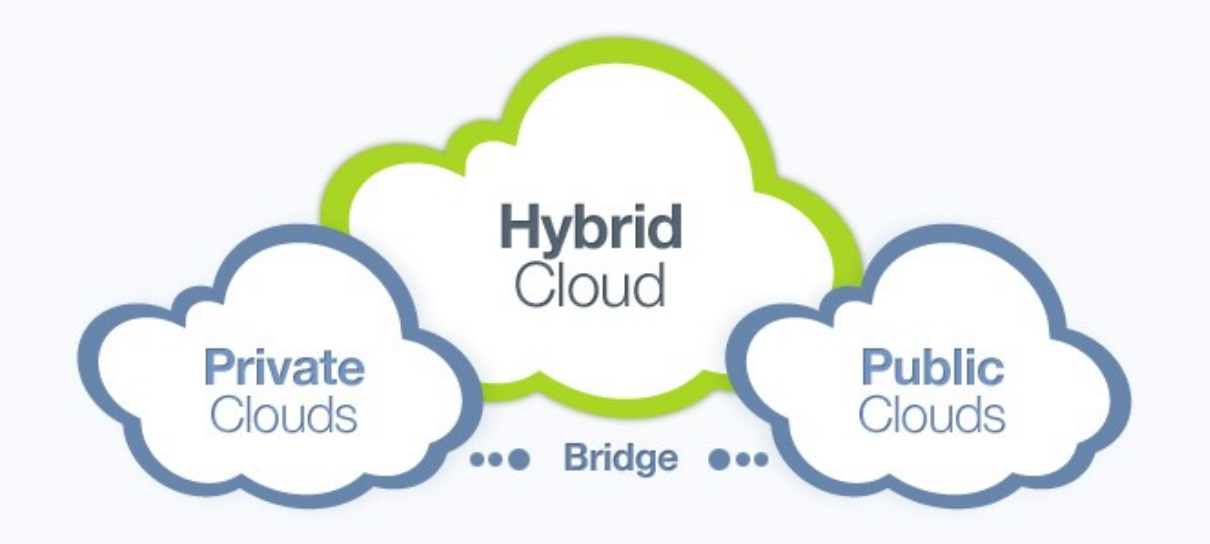

FIGURE 1.9 – Les type de cloud computing

1.8 Avantage du Cloud Computing

Il existe plusieurs avantages du cloud computing : [14]

- **Un démarrage rapide :** Le cloud computing permet de tester le plan business rapidement, à couts réduits et avec facilité.
- **L'agilité pour l'entreprise :** Résolution des problèmes de gestion informatique simplement sans avoir à s'engager à long terme.
- **Un développement plus rapide des produits :** Réduisons le temps de recherche pour les développeurs sur le paramétrage des applications.
- **Pas de dépenses de capital :** Plus besoin des locaux pour élargir vos infrastructures informatiques.

1.9 Limites du Cloud Computing :

Le cloud computing présente les inconvénients (limites) suivantes : [14]

- **La bande passante :** Besoin d'une bande passante gigantesque, et les coûts seraient tellement importants qu'il est plus avantageux d'acheter le stockage nous-mêmes plutôt que de le loué.
- **Les performances des applications peuvent être amoindries :** Un Cloud public n'améliorera définitivement pas les performances des applications.
- **La fiabilité du Cloud :** Un grand risque lorsqu'on met une application qui donne des avantages compétitifs ou qui contient des informations clients dans le Cloud.
- **Taille de l'entreprise :** Si votre entreprise est grande alors vos ressources sont grandes, ce qui inclut une grande consommation du Cloud. vous trouverez peut être plus d'intert à mettre au point votre propre Cloud plutôt que d'en utiliser un externalisé.
- **Perte de la maîtrise de son informatique (confiée à un ou des tiers) :** Du fait que l'on ne peut pas toujours exporter les données d'un service cloud, la réversibilité (ou les coûts de sortie associés) n'est pas toujours prise en compte dans le cadre du projet.[15]
- **Problèmes de sécurisation de ses données informatiques :**
  L'utilisation des réseaux publics, dans le cas du cloud public, entraîne des risques liés à la sécurité du cloud. En effet, la connexion entre les postes et les serveurs applicatifs passe par le réseau internet, et expose à des risques supplémentaires de cyberattaques, et de violation de confidentialité. Le risque existe pour les particuliers, mais aussi pour les grandes et moyennes entreprises, qui ont depuis longtemps protégé leurs serveurs et leurs applications des attaques venues de l'extérieur grâce à des réseaux internes cloisonnés.[15]

1.10 Problèmes de sécurité dans le cloud computing

Le cloud computing apporte de nombreux problèmes de sécurité, qui peuvent exposer les données et les rendre plus vulnérables aux attaques. Par exemple, lors de la transmission, les données sont plus facilement exposées car le pare-feu ne couvre plus les données. De même, le cloud est conçu pour connecter plusieurs personnes, ce qui le rend

plus vulnérable aux attaques, car celui qui dit « multi-utilisateurs » est « multi-accès ». Plus les utilisateurs (et les appareils) accèdent au cloud, plus le risque d'intrusion de cybercriminels est grand.[16]

Violation de données Si vous ne gérez pas la sécurité des données dans le cloud, vous pouvez subir une violation de données. Si votre service cloud (ou l'appareil qui y est connecté) fuit, quelqu'un a accédé à des données sensibles. S'il s'agit d'un cybercriminel, il peut le propager. Si les données stockées sont transmises électroniquement ou physiquement, cela est considéré comme une fuite. Étant donné que le cloud ne dépend pas du matériel, les cybercriminels peuvent divulguer des données cloud sur Internet ou stocker des informations, puis les diffuser. Dans le cloud computing, les fuites de données, également appelées attaques « à basse vitesse », sont un risque très courant.
Les violations de données sont généralement liées aux données personnelles de santé ou d'identité, aux secrets commerciaux et aux droits de propriété intellectuelle. Ce type d'information nécessite le plus haut niveau de sécurité.[16]

Données perdues Un autre risque courant du stockage en nuage est la perte de données. Ici, les données ne sont pas volées ou divulguées, mais complètement effacées. Cela peut être dû à des attaques de pirates, des virus ou des défaillances du système. Si les données n'ont pas été sauvegardées auparavant, cela peut causer des problèmes et ne fera que souligner la valeur de la protection des services cloud. Cependant, si les cybercriminels ciblent des données spécifiques, ils peuvent également attaquer la sauvegarde.
La perte de données peut nuire à l'entreprise car certaines informations peuvent être difficiles, voire impossibles à récupérer. Essayer de les récupérer peut nécessiter beaucoup de temps, d'argent et de ressources. Bien qu'il soit parfois possible de recréer des données ou de copier des données à partir de formats papier, la capacité de travail est toujours gravement perturbée.[16]

Cryptojacking Le cryptojacking est une menace qui se cache à l'intérieur de votre appareil pour utiliser ses ressources et exploiter la crypto-monnaie. Il peut cibler les systèmes de sécurité défaillants : leur infrastructure cloud devient fragile et les attaquants peuvent contrôler le réseau cloud pour pirater les navigateurs Web et compromettre les points de terminaison. Tous ces éléments sont inconnus des utilisateurs.
L'extraction de crypto-monnaie est légale, mais parce qu'elle peut utiliser beaucoup de ressources, les cybercriminels préfèrent utiliser l'équipement d'autres personnes. Par conséquent, les victimes de cryptojacking peuvent subir une augmentation des factures d'électricité, une durée de vie réduite de la batterie ou des processus plus lents. L'extraction de crypto-monnaie peut être une entreprise rentable, mais avant d'y arriver, vous devez d'abord dépenser beaucoup d'argent sur les ressources que vous utilisez.[16]

1.11 Conclusion

Dans ce chapitre, nous avons fourni une base théorique sur le cloud computing, en présentant ses différents services (IaaS, PaaS, SaaS) et leurs caractéristiques, types, avantages et inconvénients.
L'ouverture du cloud et sa simplicité d'utilisation, sa résilience expose les réseaux informatiques aux attaques et aux dangers de la communauté des hackers et c'est devenu le principal obstacle à la progression du Cloud services informatiques. Le deuxième chapitre identifiera le problème des attaques ddos en cloud computing et comment les éliminer.

Chapitre 2

Les attaques DDOS

2.1 Introduction

La technologie du cloud computing est devenue plus populaire, et concurrente sérieuse des systèmes informatiques traditionnels [17], car il facilite l'accès aux données via les différentes techniques d'accès au réseau.
Tant que le cloud computing permet aux utilisateurs d'accéder à des services et des ressources via internet en raison de leur nature multi-locataires, cela permet à plusieurs machines virtuelles (VM) appartenant à différents clients de partager la même infrastructure physique, c'est ce qui l'a rendu encore plus largement exposées à divers types de menaces de sécurité, qui peuvent conduire à des mauvaises qualité de service.
L'une de ces attaques, qui a été une attaque très visible est l'attaque par déni de service (DoS).
Traditionnellement, l'attaque Dos est une attaque qui vise à rendre une application informatique incapable de répondre aux requêtes de ses utilisateurs, les auteurs du [18] disent que le déni de service distribué (DDoS) est une attaque DoS émise depuis plusieurs origines distinctes. Ce type d'attaque est extrêmement complexe à bloquer. De plus, les méthodes traditionnelles de la détection et l'atténuation sont tout simplement insuffisantes par rapport aux différentes méthodes d'attaques DDoS désormais utilisées par les pirates.[19]

2.2 C'est quoi une attaque DDoS ?

L'attaque par déni de service distribué, ou DDoS (en anglais Distributed Denial of Service), vise à perturber ou paralyser totalement le fonctionnement d'un serveur informatique en le bombardant à outrance de requêtes erronées.
L'objectif peut être d'affecter les services en ligne ou le réseau de l'entreprise en saturant l'une des ressources du système : bande passante, espace de stockage, puissance de traitement de la base de données, calculs du processeur, RAM, l'ouverture d'un grand nombre de nouvelles sessions TCP dans un intervalle du temps très court,ou encore d'un nombre trop important de traitements concurrents effectués par une base de données.[20]
Le DoS distribué appelé DDoS est la méthode d'attaque DoS la plus populaire car il est capable de causer des effets plus graves facilement et rapidement.
voici la figure 2.1 qui montre un exemple simple d'une attaque ddos.

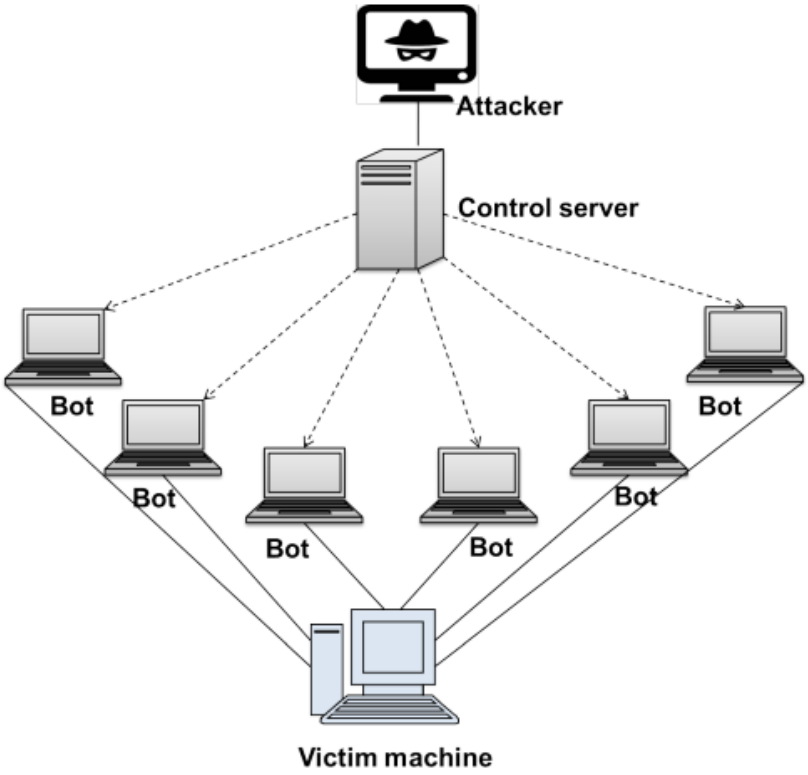

FIGURE 2.1 – Exemple d'attaque DDoS

2.2.1 L'attaque DDOS dans le cloud computing

Les résultats montrent que les attaques DDoS sont plus compliquées que les attaques DOS car les attaques DDoS se compose de trois éléments de base :
1) chef ou attaquant.
2) hôte (zombie)
3) La machine victime.
Dans la première étape de l'attaque, l'attaquant principal (chef) a trouvé de nombreux hôtes (Zombie) et implémenter un logiciel dessus pour communiquer avec lui à l'avenir (Temps d'attaque). L'étape principale de l'attaque se produit par l'hôte conformément à l'ordre du chef. Les attaques impliquant de nombreuses sources différentes sont beaucoup plus difficiles que les attaques, il ne contient qu'une seule source.
Dans les systèmes de cloud computing, la situation est plus compliquée, car dans le cloud computing il existe un serveur distribué (Figure 2.2 (ci-dessous)), ce qui pose un problème plus important. Virtualisation, l'utilisation de la multi-location et des ressources partagées sont quelques-unes des raisons des attaques DDOS par rapport à l'ancien réseau, les problèmes du système de cloud computing sont plus graves et plus difficiles (Il n'y a pas d'environnement virtualisé). La figure 2.2 montre l'attaque DOS et DDOS dans l'ancien système et l'attaque DDOS dans le cloud computing systèmes.
Les attaques DDoS et leur caractérisation deviennent complètement différentes lorsqu'il est appliqué au contexte du nuage[21]. Ces attaques ont été très réussies où les attaquants exploitent fonctions du nuage (mise à l'échelle automatique, comptabilité au fur et à mesure et multi-locataires). Comme il a été discuté dans[21], ces caractéristiques donnent

l'avantage d'exécuter plus de VM de différents propriétaires de VM sur un seul serveur physique, et permet un nuage d'utiliser des ressources sans les acheter physiquement. Comme mentionné par Gubta et al [21], lorsqu'une VM est infectée par code malveillant et si cette VM lance une attaque DDos sur un hôte physique, puis il peut causer des problèmes à d'autres VM sur le même hôte. En outre avec le même ensemble de fonctionnalités, un lourd l'utilisation des ressources sur le serveur victime peut être transformée dans un problème grave.

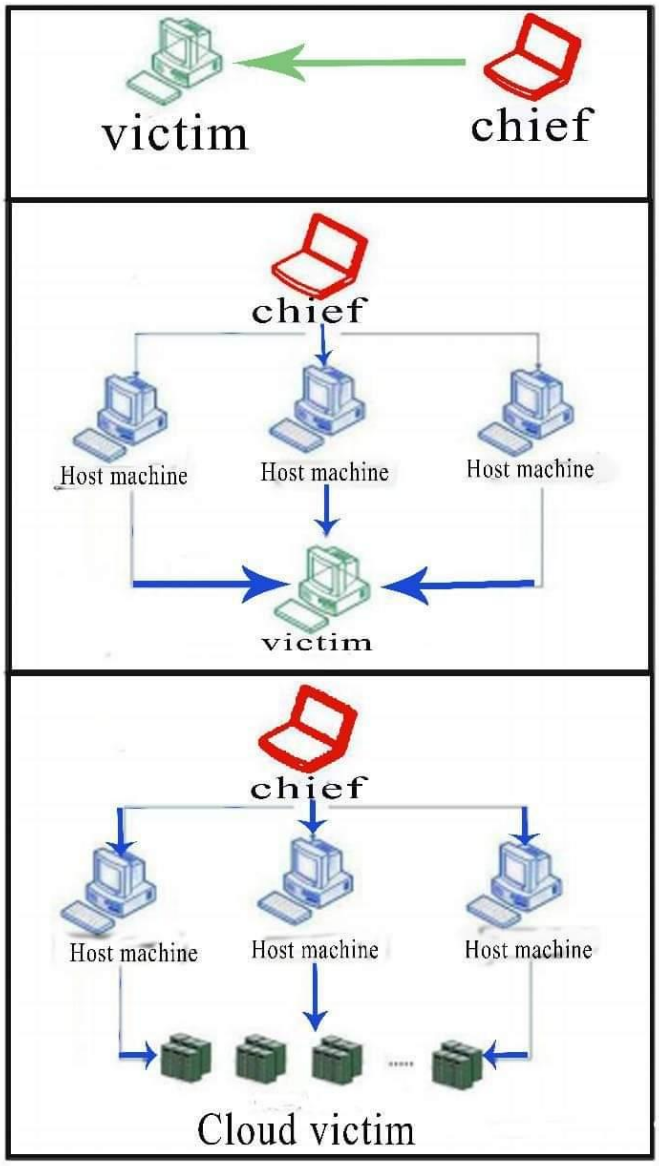

FIGURE 2.2 – Attaque DoS (ci-dessus) et attaque DDoS (au milieu) dans l'ancien système et attaque DDoS dans le cloud système informatique (ci-dessous)

2.3 Les Modes des attaques DDoS

Les attaques DoS ont un large éventail d'attaques. Voici les différentes catégories principales de schémas d'attaque :

2.3.1 Consommez des ressources limitées

Les machines informatiques ont besoin de diverses ressources pour fonctionner efficace comme mémoire, bande passante puissance de traitement, etc., et si ces ressources consommés par des facteurs externes, puis des programmes informatiques pour les raisons suivantes, le besoin réel de ces ressources peut être plus lent l'accès est restreint. Voici les sous-catégories de ces types d'attaques,

- **Connectivité réseau :**
  Dans ce type d'attaque, l'attaquant empêche les utilisateurs légitimes d'accéder au service des manières suivantes :
  Un exemple d'une telle attaque est Attaque "SYN Flood". Dans cette attaque, l'attaquant envoyez le paquet de données au serveur, demandez-en un nouveau 'connectez-vous' avec lui. Mais l'attaquant n'envoie pas 'confirmez' le paquet de données pour confirmer la connexion.
  La connexion semi-ouverte reste vivante jusqu'à ce que la minuterie expire. Quand l'attaquant est perdu un serveur victime avec un grand nombre de tels paquets, le serveur doit conserver la structure des données dans sa mémoire tous ces paquets jusqu'à ce que la minuterie expire. Par conséquent, le serveur est toujours occupé et ne peut pas être servi autres demandes légitimes pendant cette période.
- **Utiliser les ressources des clients contre eux-mêmes :**
  L'attaquant peut attaquer en faisant deux victimes machines pour communiquer entre elles sans cesse. Un exemple de ce type d'attaque est décrit dans [22] . Il utilise des paquets UDP falsifiés pour connecter le service d'écho d'une machine à une autre machine dans le réseau de la victime, ce qui fait que deux machines consomment tout leur réseau bande passante entre eux.
- **Consommation de bande passante**
  L'attaquant peut rediriger un grand nombre de paquets vers le réseau de la victime, consommant tout son bande passante entrante. Ces paquets peuvent être quoi que ce soit, généralement utilisés sont les paquets ICMP Echo. Pour rendre cette attaque efficace, un attaquant peut utiliser plusieurs machines pour envoyer un grand nombre de paquets vers la victime.
- **Consommation d'autres ressources**
  De nombreux attaquants peuvent tenter de consommer des ressources autres que la bande passante du réseau, comme la puissance de traitement ou la mémoire.
  Le serveur peut stocker également temporairement des informations dans ses données structures et un attaquant peut exploiter cette mécanisme en demandant au serveur d'allouer informations par diverses demandes à la fois. Serveur le stockage regorge d'informations relatives aux attaquants, rendant le système indisponible pour desservir d'autres demandes.
  L'attaquant peut également demander en générant divers processus pour le CPU qui doivent attendre pour être exécuté. Les exemples sont les bombardements par e-mail, générant intentionnellement des erreurs qui doivent être consigné.

2.3.2 Destruction ou modification des informations de configuration

Si certaines informations de configuration sont modifiées ou retiré de la machine, alors on ne peut pas utiliser la machine correctement et il se peut que la machine ne puisse pas

servir les autres utilisateurs qui s'y sont connectés.
L'attaquant tente de s'introduire dans la machine en modifiant sa configuration et rendre le système inaccessible aux utilisateurs généraux. Par exemple, si un attaquant est capable de changer de routage informations de table d'un routeur mal sécurisé, puis d'autres les utilisateurs peuvent ne pas être en mesure d'atteindre la machine connectée.[22]

2.3.3 Destruction physique et altération des composants du réseau

Dans ce type d'attaque, l'attaquant peut violer physiquement l'environnement de la victime. Ensuite, l'attaquant peut détruire ou modifier les composants du réseau

2.4 Les types des attaques DDOS

Différents types d'attaques DDoS ciblent différents composants de la connexion réseau. Afin de comprendre le fonctionnement des différentes attaques DDoS, il est nécessaire de comprendre comment les connexions réseau sont établies. La connexion réseau sur internet se compose de nombreux composants indépendants appelés « couches », chaque couche du modèle est conçue dans un but différent.

Le modèle OSI (illustré ci-dessous) est un cadre conceptuel pour décrire les connexions réseau dans sept couches différentes :

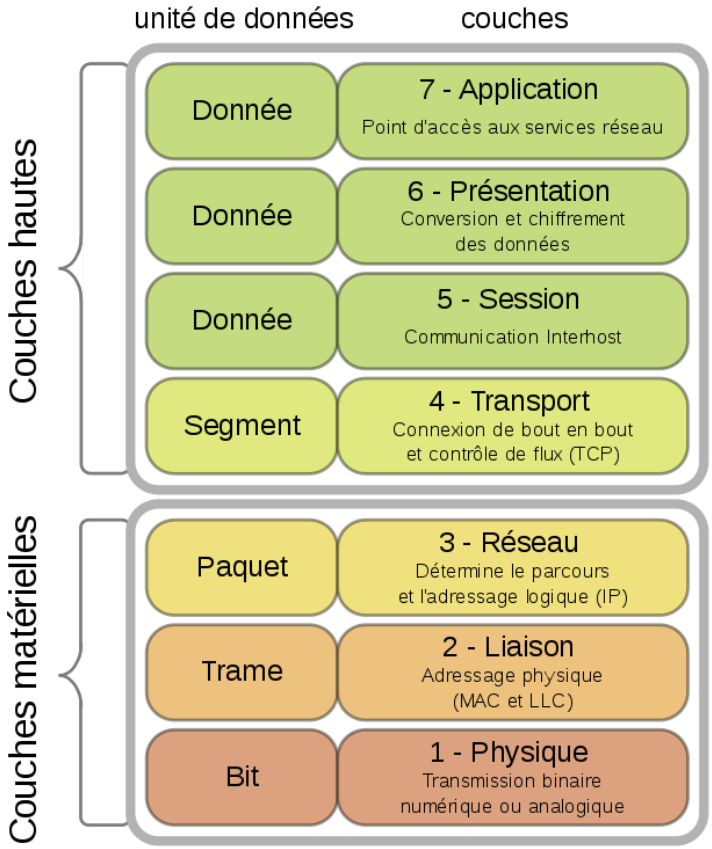

FIGURE 2.3 – Le modèle OSI

Les attaques DDoS sont classées dans trois grandes catégories en fonction de leur cible [15] :

2.4.1 Attaques basées sur le volume

Il s'agit du type d'attaque DDoS le plus courant, les attaquants essaient de consommer toute la bande passante du réseau en envoyant un nombre volumineux de paquets à un serveur victime afin que cela devienne très difficile pour utilisateurs légitimes d'accéder au serveur en raison d'un blocage à routeurs d'interface.
L'ampleur de ces types d'attaques est mesuré en bits par seconde. Exemple : Amplification DNS ,UDP Flood, ICMP Flood, autres inondations de paquets usurpés.[23]

2.4.1.1 L'attaque Amplification DNS :

Il s'agit d'une attaque DDos (déni de service distribué) volumétrique, fondée sur la réflexion, lors de laquelle un pirate exploite la fonctionnalité de résolveurs DNS ouverte pour surcharger un serveur ou un réseau cible avec une quantité de trafic amplifié, afin de rendre inaccessibles le serveur et son infrastructure environnante.[24].

L'attaque par amplification DNS peut être divisée en quatre étapes :

1- Le pirate utilise un point de terminaison compromis pour envoyer des paquets UDP à un récurseur DNS à l'aide d'adresses IP usurpées. L'adresse usurpée contenue dans les paquets renvoie vers l'adresse IP réelle de la victime.

2- Chacun des paquets UDP envoie une requête à un résolveur DNS, en transmettant souvent un argument du type « ANY », afin de recevoir la réponse la plus volumineuse possible.

3- Après avoir reçu ces requêtes, le résolveur DNS (qui essaie d'effectuer son travail en répondant) envoie une réponse volumineuse à l'adresse IP usurpée.

4- L'adresse IP de la cible reçoit la réponse et l'infrastructure de réseau environnante se retrouve submergée sous un flot de trafic, provoquant ainsi un déni de service.[24]

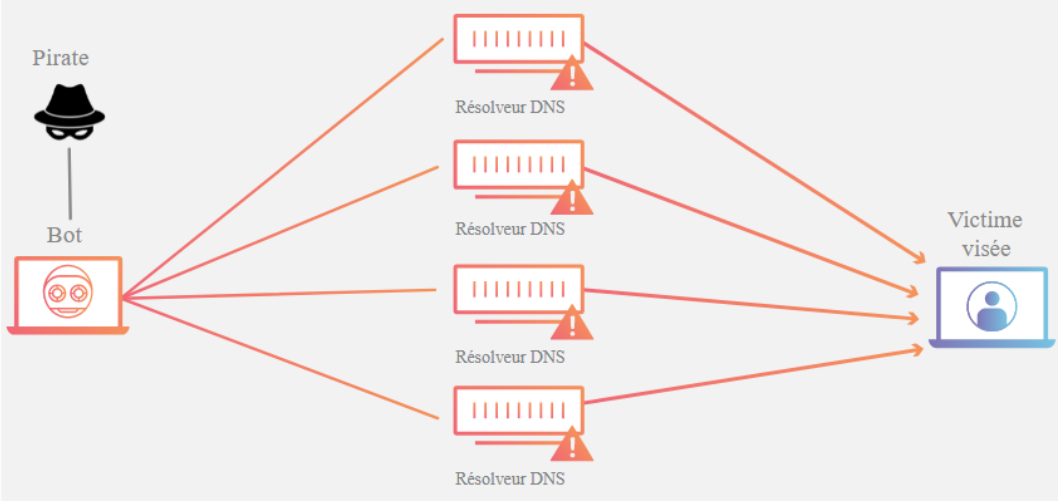

FIGURE 2.4 – Exemple d'amplification

2.4.1.2 UDP Flood :

UDP Flood est un type d'attaque par déni de service dans laquelle un grand nombre de paquets UDP (User Datagram Protocol) sont envoyés à un serveur ciblé dans le but d'écraser la capacité de cet appareil à traiter et à répondre.
Le pare-feu protégeant le serveur ciblé peut également s'épuiser à la suite d'une inondation UDP, entraînant un déni de service du trafic légitime. [24]

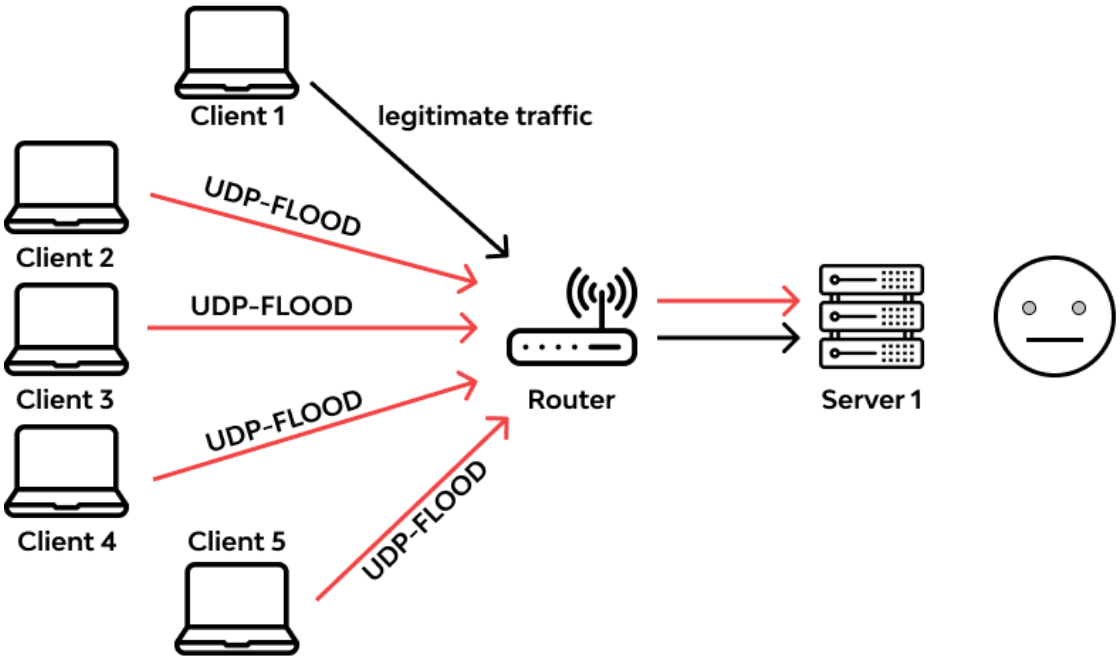

FIGURE 2.5 – Exemple d'attaques udp flood [25]

2.4.1.3 L'attaque Peer-To-Peer (P2P)

De nos jours, les systèmes de partage de fichiers P2P ont été la cible d'attaques massives qui exploitent des bogues dans les serveurs P2P pour exécuter un DDoS attaque pour que l'attaquant P2P utilise des clients P2P connectés au P2P des hubs de partage de fichiers au lieu d'utiliser un botnet pour attaquer l'hôte victime[25].

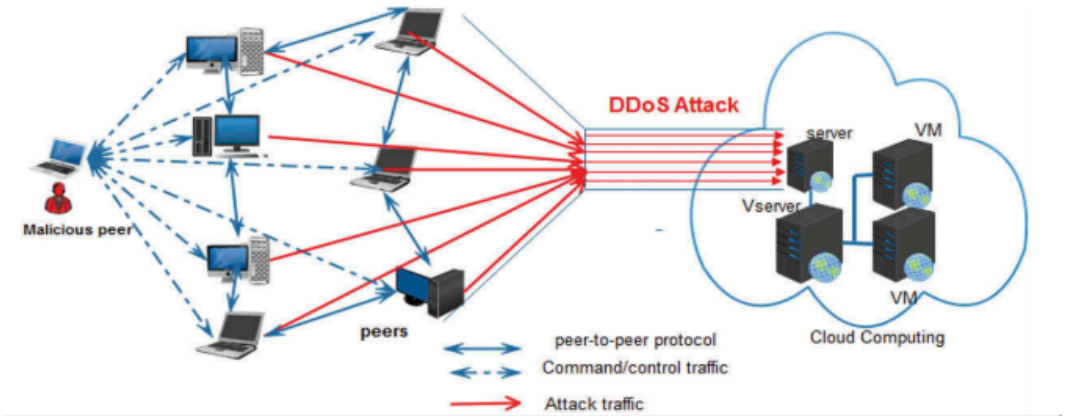

FIGURE 2.6 – L'architecture du mécanisme d'attaque P2P [25]

2.4.2 Attaques de protocole

Les attaques de protocole consomment les ressources de la victime machine comme la mémoire, la capacité de traitement, etc.. peut créer de longues files d'attente ou d'autres structures de données sur les périphériques réseau intermédiaires tels que les pare-feu, les routeurs ou équilibreurs de charge, etc. Ampleur du protocole les attaques sont mesurées en paquets par seconde. Exemple :SYN Flood, attaque de paquets fragmentés.

Exemple d'attaque de protocole : [24]

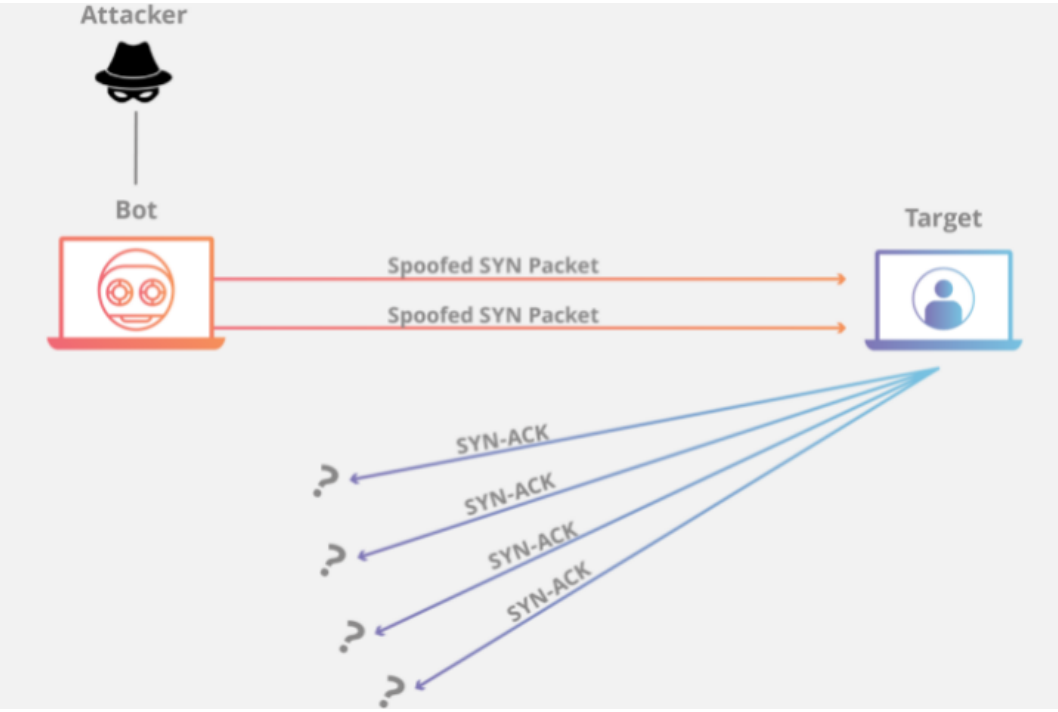

FIGURE 2.7 – Exemple d'attaque de protocole [24]

2.4.2.1 L'attaque SYN flood

Une attaque SYN flood est un type d'attaque par déni de service (DDoS) qui vise à rendre un serveur indisponible pour le trafic légitime en consommant toutes les ressources serveur disponibles. En envoyant à plusieurs reprises des paquets de demande de connexion initiale (SYN), le pirate est en mesure de submerger tous les ports disponibles sur une machine serveur ciblée, ce qui oblige l'appareil ciblé à répondre lentement au trafic légitime, ou l'empêche totalement de répondre.
l'attaques SYN flood fonctionne en utilisant le processus d'établissement de liaison des connexions TCP.
Généralement,une connexion TCP a trois processus indépendants pour établir une connexion.
1- Tout d'abord,le client envoie un paquet SYN au serveur pour établir une connexion.
2- Le serveur répond alors à ce paquet initial avec un paquet SYN/ACK pour confirmer la réception de la communication.
3- Enfin,le client renvoie un paquet ACK pour confirmer la réception du paquet depuis le serveur.
Après avoir terminé cette séquence d'envoi et de réception de paquets de données, la connexion TCP est ouverte et les données peuvent être envoyées et reçues.[24]

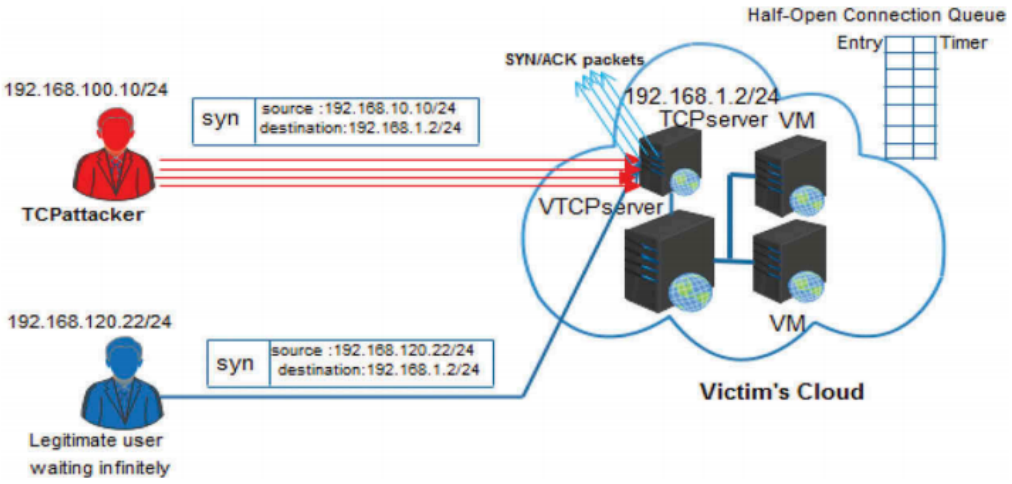

FIGURE 2.8 – Les détails du mécanisme d'attaque SYN Flood [25]

2.4.2.2 L'attaque Ping of death

Une attaque Ping of death (PoD) est une attaque par déni de service (DoS), dans laquelle l'attaquant vise à perturber une machine ciblée en envoyant un paquet plus grand que la taille maximale autorisée, provoquant le blocage ou le crash de la machine cible.
Le ping original de l'attaque mortelle est moins courant aujourd'hui. Une attaque connexe connue sous le nom d'attaque par inondation ICMP est plus répandue.[24]

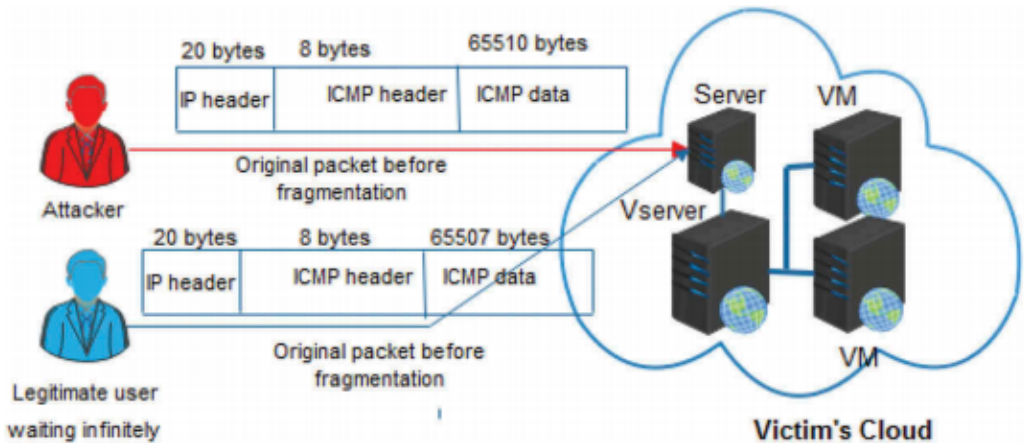

FIGURE 2.9 – L'architecture d'attaque Ping of death [25]

2.4.3 Attaques d'application

Considérées comme étant les types d'attaques DDoS les plus sérieuses et sophistiquées, ces attaques ciblent les applications web en exploitant les vulnérabilités qu'elles comportent. Aussi appelées « Attaques couche 7 », les attaques d'application fonctionnent toujours de la même manière mais demandent beaucoup moins de force car elles ciblent les faiblesses au sein des serveurs ciblés.
Beaucoup moins de trafic web est nécessaire pour monopoliser les processus et protocoles sur ces points faibles, ce qui rend aussi l'attaque bien plus difficile à détecter en raison

du faible volume de trafic généré qui semble légitime [23] .

Exemple d'attaque de la couche d'application : [24]

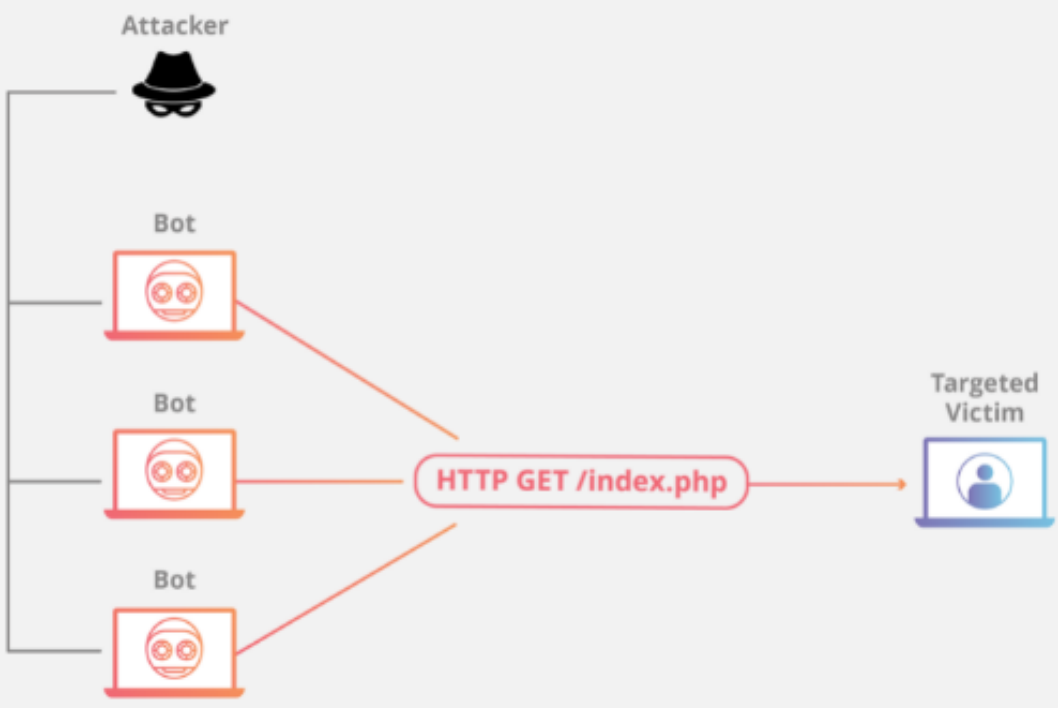

FIGURE 2.10 – Exemple d'attaque de la couche d'application [24]

2.4.3.1 L'attaque Back

Pour lancer une attaque réussie sous cette forme, l'attaquant continue l'envoie d'un grand nombre de requêtes contenant de nombreuses barres obliques ("////// . ..///") au serveur Web Apache. Chaque demande a besoin plus de temps de traitement qui ralentit officiellement le Web Apache serveur.[25].

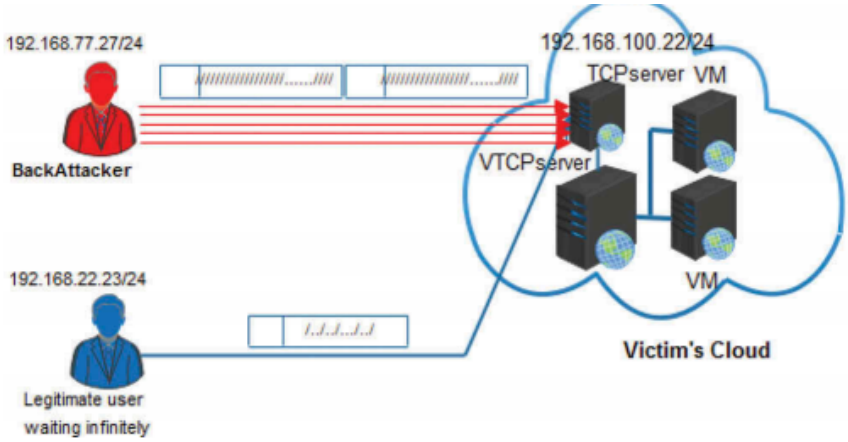

FIGURE 2.11 – l'architecture d'attaque back [25]

2.4.3.2 L'attaque Slowloris

Slowloris est une attaque de la couche applicative dans laquelle le cybercriminel envoie périodiquement des requêtes HTTP GET incomplètes mais légitimes au serveur web ciblé. Il maintient ainsi les connexions ouvertes et dévore lentement et méticuleusement les prises de connexion du serveur web, brouillant ainsi toutes les autres demandes légitimes. [26]

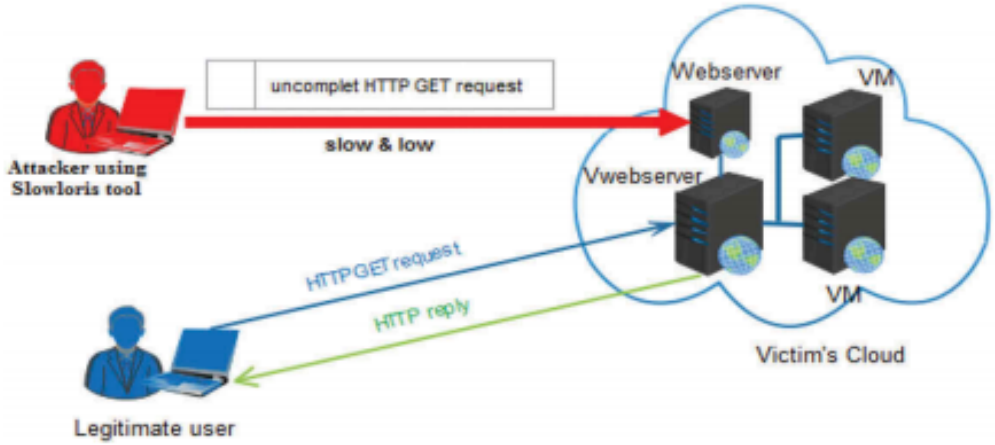

FIGURE 2.12 – Les détails du mécanisme d'attaque de Slowloris. [25]

2.5 Quelques vecteurs d'attaque

2.5.1 Les botnets :

Les attaques DDoS peuvent être lancées à partir du réseau de la machine infectée Cela s'appelle un botnet.
De nombreux outils disponibles en ligne peuvent tirer parti des botnets [27].
Dans ces dernières années, des services DDoS en ligne (communément appelés bootstrap ou stress) ont vu le jour [28]. Le prix de ces services permet utilisé par des individus et permet aux utilisateurs de lancer des attaques opposez-vous au but qu'il a choisi. En outre, certains programmes de démarrage fournissent également des services de test gratuit pendant quelques minutes.
La diversité des outils et services permet lancer une attaque par déni de service distribué peut aider à augmenter beaucoup de ces attaques.

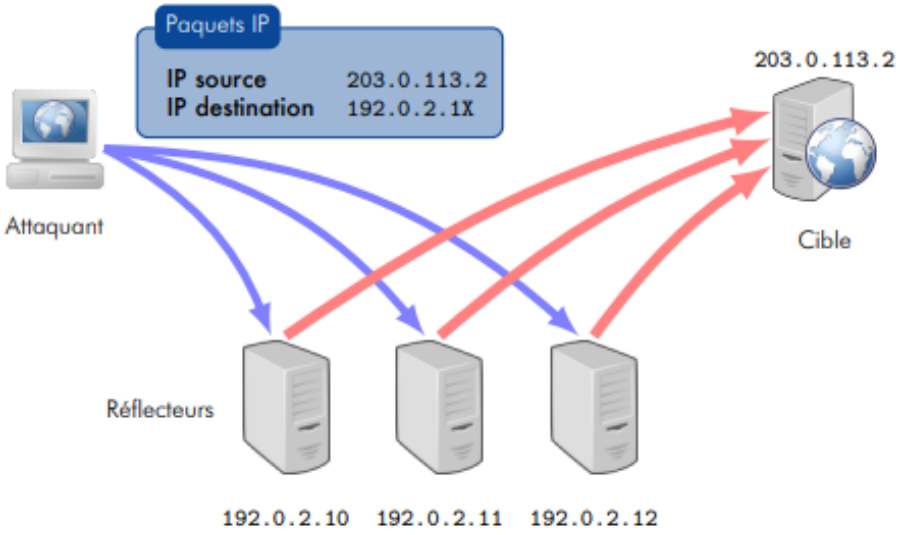

FIGURE 2.13 – Principe d'une attaque par réflexion

2.5.2 Les attaques basées sur la réflexion

Certaines attaques utilisent des ordinateurs accessibles sur internet et répondent demandes de n'importe quelle source : ce sont des réflecteurs. L'attaque de La réflexion consiste à envoyer des paquets à ces réflecteurs en utilisant les adresses IP suivantes : La victime est l'adresse IP source : c'est ce qu'on appelle l'usurpation d'adresse IP. Cette réaction de ces réflecteurs à la victime peut induire un trafic non sollicité de la destination de ce dernier.
Ce débit peut être suffisamment important pour saturer lien réseau de la victime, qui conduit à un déni de service. L'attaquant interroge le serveur en usurpant l'adresse IP ses victimes (203.0.113.2). Par conséquent, le serveur envoie la réponse à la demande faite par l'agresseur à la victime. Si le déni de service est effectif, Le trafic causé par ces réponses dépasse la bande passante du réseau disposition de la victime.
Les attaques par réflexion impliquent généralement des protocoles basés sur des protocoles transmission UDP. En fait, le protocole UDP laisse la tâche à la couche application identifié la source afin que l'usurpation d'adresse IP puisse être effectuée. Au fait,UDP Il n'est pas nécessaire d'établir une session avant d'envoyer des données (contrairement à TCP). Cette fonctionnalité permet à l'attaquant d'utiliser UDP pour demander un service via un seul paquet de données et générer un la réponse du réflecteur.
Les attaques par réflexion ne se limitent pas aux protocoles de transmission UDP. Par exemple, vous pouvez envoyer des paquets TCP SYN en usurpant l'adresse IP de la victime génère un paquet SYN-ACK en réponse à la cible.
Enfin, il convient de noter que les attaques par réflexion peuvent être lancées à partir de botnets.

2.5.3 Les attaques basées sur l'amplification :

Le but d'une attaque par lots est d'épuiser la bande passante réseau disponible pour rendre un ou plusieurs services inaccessibles.

Ce type d'attaque est généralement réalisé en utilisant les propriétés de certains protocoles pour maximiser le trafic généré. De plus,le but des attaques de volume est de produire très grand nombre de paquets de données par seconde pour saturer les ressources de traitement le but.
La réponse générée par certains protocoles est beaucoup plus grande que réclamer. Le nombre de paquets provoqués par la réponse peut également être plus grand dépassement du nombre de paquets requis pour envoyer la demande. L'amplification produit grâce à ces protocoles, il peut être utilisé pour mener des attaques de volume.
Dans la plupart des cas, les attaques volumétriques utilisent la réflexion et l'amplification. De nombreux protocoles peuvent être utilisés pour implémenter ce genre d'attaque. Parmi eux, on peut notamment citer DNS (Domain Name System [29]), NTP (Network Time Protocol [30]), SNMP (Simple Network Management) Protocole [31]), SSDP (Simple Service Discovery Protocol) [32] ou CHARGEN (Protocole Générateur de Caractères [33]).
Il est important de noter qu'une entité peut être victime d'une attaque de volume utilisez le protocole, même s'il n'a aucun service actif exposé sur internet Basé sur le même accord.

2.5.4 Les attaques ciblant des applications

Certaines attaques visent à épuiser les capacités de traitement d'une cible. Par exemple, un attaquant peut chercher à atteindre la limite du nombre de connexions concurrentes qu'un serveur web peut traiter. Dans ce cas, l'attaquant envoie en permanence un grand nombre de requêtes HTTP GET ou POST au serveur ciblé. Il est également possible d'envoyer des requêtes partielles, puis de transmettre la suite de ces requêtes à intervalles réguliers, dans le but de maintenir les connexions ouvertes le plus longtemps possible et d'éviter la fermeture des connexions au-delà d'un délai fixé [34].
D'autres types d'attaques applicatives cherchent à épuiser les ressources de calcul d'un serveur en initiant un grand nombre de sessions TLS, ou encore à tirer parti de faiblesses dans la conception d'une application web [35] .

2.6 Qui peut être visé

Toute entité dont l'activité dépend de l'infrastructure réseau connectée à Internet peut être la cible d'attaques DDoS. Les motivations et les objectifs des assaillants sont variés, allant des revendications idéologiques à la vengeance en passant par l'extorsion.
En outre, il semble que certaines attaques aient été menées pour détourner l'attention et dissimuler d'autres activités illégales, telles que des transactions bancaires frauduleuses [36] [37].
Bien que de nombreuses entités soient affectées par cette menace, certains types d'activités sont plus vulnérables aux cibles DDoS. Parmi eux, on peut citer notamment le e-commerce, les institutions financières, le gouvernement et même les structures d'hébergement informatique.
Dans ce cas, il est particulièrement important de prévoir des solutions de protection adaptées dès le début d'un projet de mise en œuvre d'un système d'information et d'une infrastructure réseau.
Les attaques DDoS sont très courantes aujourd'hui. Par exemple, les opérateurs français ont observé plus d'un millier d'attaques chaque jour en 2014.

En outre, certains rapports publics indiquent que le nombre de ces attaques est en augmentation [38].
En plus de l'augmentation du nombre, l'ampleur des attaques a également augmenté de manière significative ces dernières années [39] [40] .

2.7 Les trois phases de gestion d'une crise DDoS

La plupart des article comme [41] quand ils résolvent les solutions des attaques ddos ils classent les phases de gestion d'une crise DDoS en trois phases : avant l'attaque, pendant l'attaque et après l'attaques, le schéma ci-dessous montre les détails de chaque phase :

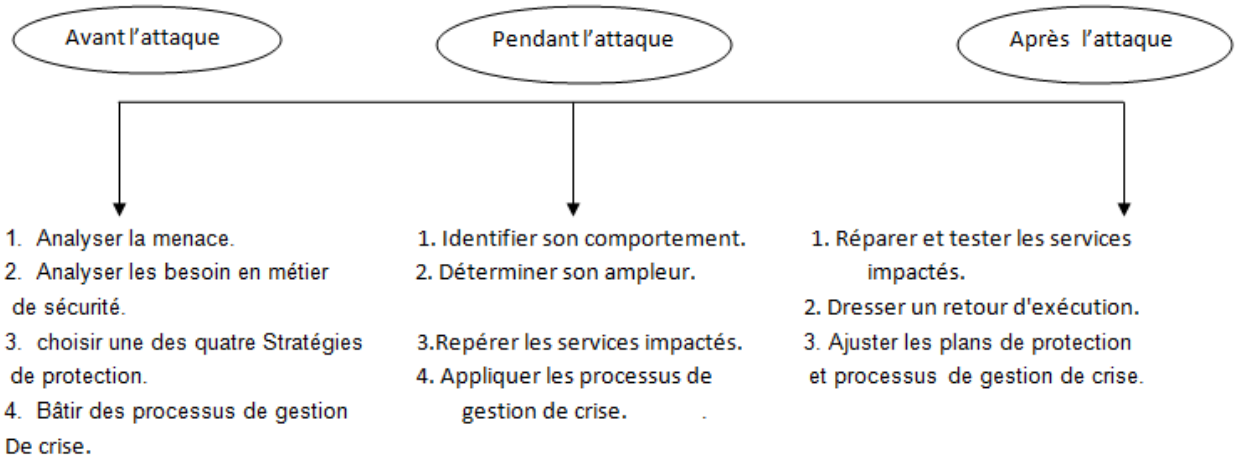

FIGURE 2.14 – Les trois phases de gestion d'une crise ddos [41]

2.8 Les solutions des attaques DDOS dans le Cloud Computing Environnement :

Les attaques DDoS sont un problème difficile à résoudre. Premièrement, il n'y a pas de caractéristiques communes flux DDoS pouvant être utilisés pour leur détection.
De plus, la nature distribuée des attaques DDoS les rend extrêmement difficiles à combattre ou à remonter.
De plus, les outils automatisés qui rendre possible le déploiement d'une attaque DDoS peut être facilement téléchargé.
Les attaquants peuvent également utiliser usurpation d'adresse IP afin de cacher leur véritable identité, et cela rend la traçabilité des attaques DDoS encore plus difficile.
Enfin, il n'y a pas de niveau de sécurité suffisant sur toutes les machines d'internet, alors qu'il existe des failles de sécurité persistantes dans les hôtes Internet.

La plupart des œuvres classées les mécanismes de défense des attaques DDos en trois parties [42], La figure 2.15 illustre les méthodes de défense des attaques DDoS dans le cloud computing.

- **préventions des attaques :**
  La prévention des attaques dans le cloud est une étape pro-active dans le filtrage ou l'abandon des demandes des attaquants suspects avant d'affecter le serveur [43]. Cette prévention est appropriée système pour protéger les données et les services contre les attaques DDoS à divers endroits[16].
  Le système est déployé pour fonctionner comme des outils de surveillance qui collectent des informations sur l'hôte ou le réseau contre les attaques DDoS et également suivre l'exécution du système de prévention [44]
- **détections des attaques :**
  La détection des attaques comprend l'analysent des systèmes en cours d'exécution pour identifier les sources malveillantes qui conduire à des attaques DDoS ou identifier des paquets malveillants dans le trafic réseau [16], bien que la technique de prévention des attaques est utilisée pour se défendre contre les attaques DDoS. malgré cela, la détection des attaques est l'une des techniques les plus importantes pour se défendre contre les attaques DDoS que d'autres techniques défensives [45]. La prévention des attaques a un problème d'utilisabilité,car il s'applique à tous les utilisateurs. Cela entraîne une surcharge du serveur supplémentaire lorsque la prévention est appliquée aux utilisateurs légitimes [43]. Par conséquent, les mécanismes de détection des attaques jouent un rôle important dans la détection des attaques DDoS [46].
- **mitigations des attaques :**
  Au moment de l'attaque, un serveur victime peut continuer à servir les requêtes de l'utilisateur en en utilisant des techniques d'atténuation 'mitigation' des attaques [43].
  Depuis ces attaques, les ressources du serveur victime ne sont pas entièrement consommées au début de l'attaque L'atténuation des attaques DDoS complète la défense en évaluer la force de l'attaque et sélectionner la bonne réponse au bon moment.
  De plus, un système d'intervention traite des contre-mesures appropriées adaptées aux types d'attaque [16]. L'atténuation des attaques comprend des stratégies, telles que des techniques de traçage et filtrage des paquets [44].
  D'autre part, plusieurs éléments, tels que les serveurs,routeur, commutateur, protocoles et applications sont impliqués dans une architecture de sécurité [47]
  Selon d'autres études [48], [49] et [50], la deuxième classification divise les solution des défenses contre les attaques ddos selon l'emplacement déploiement :
- En fonction de l'endroit où le mécanisme de défense est appliqué (Près de source de l'attaque, à proximité de la destination de l'attaque, à routeurs intermédiaires, hybrides).
- Basé sur le moment où le mécanisme de défense est appliqué (avant attaque, après attaque, pendant attaque)

2.8.1 Mécanismes de détection :

1. **Mécanismes de détection basés sur les signatures :**
   La détection basée sur la signature implique un ensemble de signatures d'attaque de différents modèles d'attaques connues.
   Les signatures des modèles d'attaque seront stockées dans une base de données.

Tout le trafic entrant est comparé aux signatures existantes qui sont stockées dans la base de données pour détecter les modèles de trafic d'attaque .
Ces mécanismes consistent à détecter avec précision attaques connues si la base de données est toujours mise à jour. En raison de la facilité de reconfiguration, les règles sont nécessaires pour mettre à jour les signatures dans la base de données pour les attaques inconnues.
Cependant, les attaques inconnues ne peuvent pas être détectées si les signatures ne sont pas mises à jour. Comme un résultat, toute variation de la signature d'attaque stockée connue ou des attaques obscures entraînera un fort taux de faux négatifs.De plus, la mise à jour des signatures pour chaque attaque n'est pas faisable à mesure que les attaques changent.[46]

2. **Mécanismes de détection basés sur les anomalies :**
   La détection d'anomalies implique l'identification de modèles dans les données qui sont incompatibles avec comportement attendu. Ces modèles qui ne sont pas compatibles sont appelés anomalies modèles .
   Ces mécanismes visent à identifier des événements qui semblent anormaux dans le comportement normal du système. Normalement, le comportement du trafic est collecté sur une certaine période de temps pour comparer avec le trafic réseau entrant .
   De nos jours, des mécanismes de détection d'anomalies sont utilisés pour détecter les attaques DDoS sur le cloud informatique .
3. **Détection hybride :**
   La combinaison de deux ou plusieurs mécanismes est utilisée par la détection hybride. Tel les mécanismes de détection hybride comprennent l'utilisation à la fois des mécanismes - signature et détection basée sur les anomalies pour augmenter le taux de détection.
   Les forces et les limites de ces mécanismes de détection dépendent de l'algorithme utilisé par une telle détection mécanismes .
   Ces mécanismes de détection sont capables de classer les règles avec précision, en raison de l'avantage combiné de plusieurs techniques. Cependant, les performances et le coût de calcul de ces mécanismes de détection dépendent du nombre de techniques de détection combinées.
4. **Détection de trace de source et d'usurpation :**
   La technique de traçage est utile pour trouver la source à partir de laquelle les attaques DDoS sont généré. Ces attaques ont tendance à utiliser des adresses frauduleuses pour lancer des attaques DDoS. Une attaque par réflecteur est un exemple de cette attaque [2]. Le traçage de la source et l'usurpation sont extrêmement important pour les mécanismes de détection. Cependant, ce processus nécessite un service les fournisseurs prennent en charge et plusieurs composants réseau, tels que les serveurs et les routeurs de périphérie. De plus, il n'est pas facile de concevoir un système de défense en cloud computing contre botnets à grande échelle avec des adresses IP falsifiées, car ce mécanisme de détection nécessite beaucoup d'efforts entre les prestataires de services.
5. **Filtrage basé sur le nombre :**
   Les paramètres de ressource réseau sont utilisés pour indiquer le début de l'attaque dans le filtrage basé sur le nombre et sont ensuite utilisés pour détecter une occurrence d'attaque telle que le nombre de sauts, nombre de connexions et nombre de

requêtes pour une seule source dans une unité de temps.
La principale force de ces mécanismes de détection peut être facilement déployée, ces mécanismes permettre aux administrateurs de surveiller rapidement la situation. Cependant, il faut une base de données pour être continuellement mis à jour et avoir également une variété d'applications hétérogènes dans différents systèmes qui causent des problèmes lors de l'utilisation de ces méthodes, et cela nécessite également un étalonnage du taux de fausses alarmes.
En outre, il souffre également d'usurpation d'adresse IP qui conduit à des problèmes d'intégrité et de précision .

6. **Détection BotCloud :**
   Un attaquant peut exploiter les fonctionnalités de cloud computing pour créer des robots dans le cloud au lieu infecter les machines des utilisateurs . Par conséquent, le cloud devient une plate-forme d'utilisateurs pour lancer des attaques. Ces robots sont appelés BotCloud .
   Détection de BotCloud mécanismes visent à détecter ou trouver une attaque ciblant la machine virtuelle dans le cloud. La principale force de ces mécanismes est leur emplacement de déploiement à la fin du fournisseur de services cloud. Cependant, ces mécanismes ne peuvent pas détecter tous les types d'attaques.
   De plus, seul le bord du cloud à l'origine de l'attaque peut fonctionner avec de tels mécanismes de détection. Si les fournisseurs de services cloud ne prennent pas en charge ces détections, ces attaques pourraient devenir importantes, en raison de l'utilisation de ressources de cloud computing lourdes.[46]
7. **Utilisation des ressources (consommation des ressources) :**
   L'environnement informatique en Cloud computing utilise un serveur virtualisé appelé hyperviseur exécutant une plate-forme de système d'exploitation virtuel pour exécuter une machine invitée, telle qu'une machine virtuelle ou serveur.
   L'hyperviseur a la capacité de gérer les ressources disponibles dans chaque invité machine.
   Cette méthode utilise des fonctionnalités de trafic entrant telles que le débit, l'utilisation du processeur, et l'utilisation de la mémoire en tant que métrique de détection d'attaque ddos.
   Lorsque la machine virtuelle ou le serveur atteint la limite spécifiée d'utilisation des ressources, cela indique le soupçon q'une attaque s'est produite.
   En conséquence, la consommation de diverses machines virtuelles ou serveurs ressources est capable de fournir des informations vitales sur l'attente ou l'occurrence de l'attaque DDoS .[46]

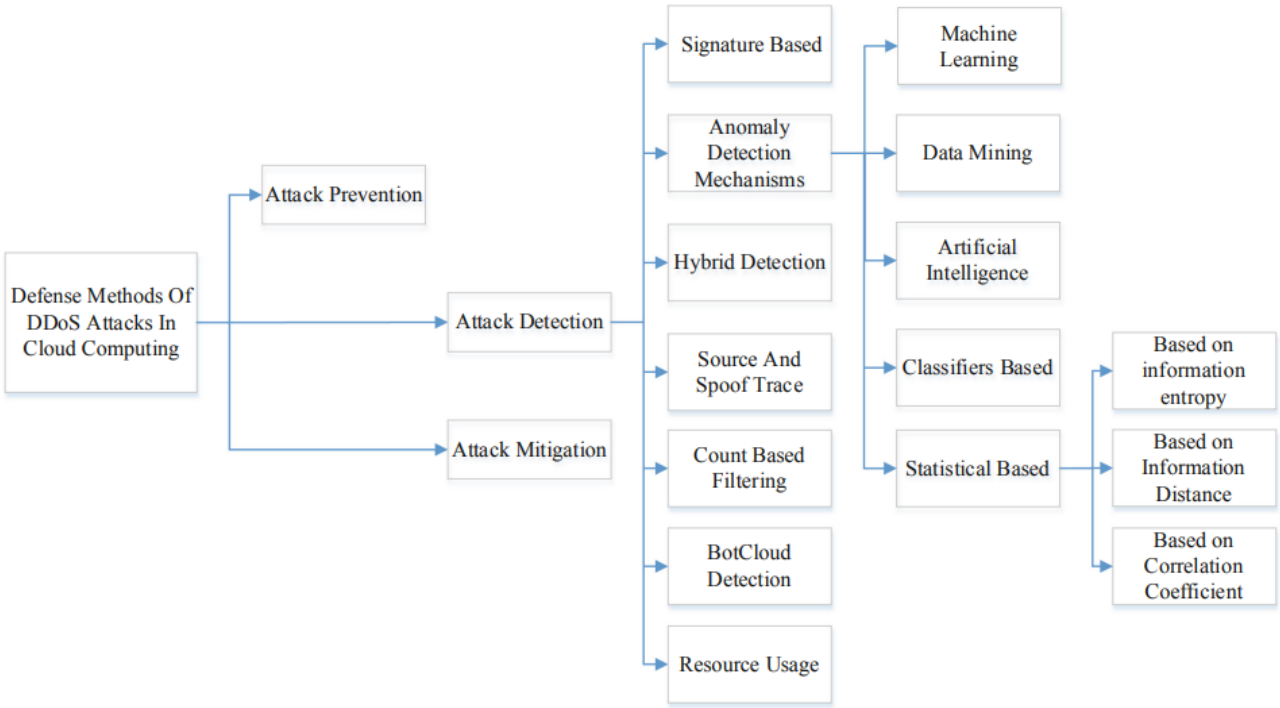

FIGURE 2.15 – Méthodes de défense des attaques DDoS dans le cloud computing [46]

2.8.2 Les classes de mécanismes de détection d'anomalies

Les mécanismes de détection d'anomalies sont classés en cinq classes selon la classification de [46] en fonction des algorithmes utilisés dans le processus de détection, à savoir l'apprentissage automatique, exploration de données, intelligence artificielle, classificateurs, et statistique.
La figure suivante présente une comparaison de la détection d'anomalies classes basées sur les algorithmes utilisés dans le processus de détection ainsi que leurs forces et limites.[46]

classes	Avantages.	limites .
Basé sur l'apprentissage automatique	• Détection des modèles d'attaque DDoS à haute efficacité. • cela peut changer leur stratégie d'exécution tout au long de la détection	• Des frais généraux élevés entraînent baisse des performances.
Basé sur Exploration de données	• Augmenter la vitesse de détection processus. • Faible coût de calcul. • Il peut gérer une énorme base de données.	• Inefficace en cas de forte volume du réseau entrant circulation. • Valeurs manquantes de l'ensemble de données influencera la détection processus.
Basé sur l'Intelligence Artificiel	• Classer les comportements comme normaux ou Intrusif. • Détecter efficacement les attaques DDoS.	• La précision de détection dépend de le profil de formation. • Problèmes d'évolutivité
Basé sur Classificateur	• Le taux de détection dépend du seuil Les paramètres. • Taux d'adoption élevé pour mettre à jour la détection stratégies.	• détection d'attaques inconnues a besoin d'une formation adéquate. • La consommation de ressources est élevée.
Statistique Basé	• Apprentissage du comportement attendu de observations sans connaissance préalable des activités normales. • Détection précise des programmes malveillants Activité.	• Difficulté de fixer une valeur optimale seuil. • Besoin de justifier les hypothèses qui sont nécessaires pour déterminer le taux de classement.

TABLE 2.1 – Comparaison des classes de mécanismes de détection d'anomalies.

2.9 Comment se protéger contre une attaque DDoS

- **Le choix de l'hébergeur**
  Qui peut apporter ses services pour lutter contre différents types et attaques techniques. Ainsi, OVH propose plusieurs technologies pour protéger les sites Internet, comme l'atténuation. Il filtre le trafic « illégal » représenté par les botnets.[51]
- **Des solutions comme Akamaï ou Cloudfare**
  Distribuent vos données sur plusieurs serveurs à travers le monde pour permettre un accès permanent. Mais pour les sites sensibles ou très fréquentés, ils reposent également sur le principe de la décentralisation. D'autres solutions DDoS géolocalisent efficacement les utilisateurs pour bloquer le trafic en provenance de certains pays ou de certaine région.[51]
- **La mise en place d'un site miroir**
  Une copie certifiée conforme de votre site Web sur un autre domaine (par exemple, l'achat d'un domaine à partir de .com ou .net en même temps). Ce processus doit bien entendu être complété en amont. Plusieurs plugins WordPress peuvent automatiser ce processus, y compris wp-mirror.[51]
- **Préparer une version allégée du site**
  Certains sites ont mis en place des versions allégées dans certains grands groupes de personnes (résultats électoraux, événements majeurs) afin de charger des structures plus légères, réduisant ainsi la bande passante. Mais cette version n'est généralement pas jolie.[51]
- **Mettre en place et préparer une "whitelist"**
  Afin de restreindre l'accès, quand nécessaire, aux seuls techniciens et administrateurs du site, par exemple pour des opérations de maintenance.[51]
- **Installer des systèmes d'alerte efficaces et automatiques** Afin d'effectuer les premières interventions automatiquement. Cela implique une bonne connaissance

du réseau par les administrateurs, qui doivent également savoir identifier et autoriser le trafic légitime [51] .

2.10 Conclusion

Bien que les attaques DDoS se multiplient dans le cloud computing. Ce chapitre présente brièvement les attaques DDoS dans le cloud computing, puis la classification des attaques, types et diverses contre mesures pour atténuer les attaques DDoS. Cette enquête accorde DDoS technologie de détection, de prévention et de tolérance.

Chapitre 3

Machine learning

3.1 Introduction

Machine learning, ou plus habituellement, l'intelligence artificielle, où l'apprentissage automatique est un sous-domaine, évoque la science-fiction pour la plupart des gens.
Cependant, l'apprentissage automatique n'est pas un rêve d'avenir mais cela fait déjà partie de la vie quotidienne pour être honnête, il est debout depuis des décennies comme les filtres anti-spam dans les années 80 ou encore les jeux vidéo depuis le début des années 2000.
Le machine learning est partout aujourd'hui il à changer complètement le monde, grâce à la voiture autonome de Google, la reconnaissance vocale de Siri, facebook même le magasin indépendant d'Amazon.

3.2 Définition de L'intelligence artificielle :

Selon le Larousse, l'intelligence artificielle se définirait comme étant «l'ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence». Ce serait, de ce fait, des ordinateurs ou des machines dotées de programmes capables de performances similaires à l'intelligence humaine, ou même, amplifiées par la technologie.
Ces machines sont en mesure de :

- Raisonner.
- Traiter de grandes quantités de données.
- Discerner des modèles indétectables par l'œil d'un humain.
- Comprendre et analyser ces modèles.
- Interagir avec l'Homme.
- Apprendre progressivement.
- Améliorer continuellement ses performances [52].

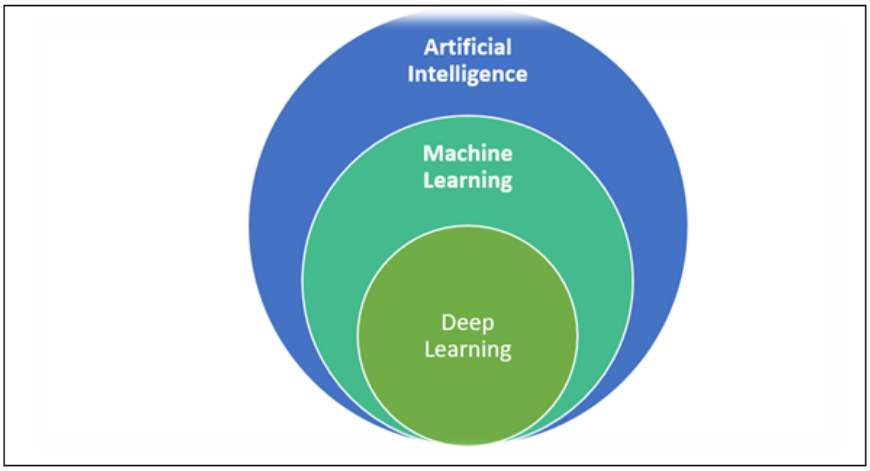

FIGURE 3.1 – Intelligence artificial, Machine learning, Deep Learning

3.3 Définition de Machine Learning

Le Machine Learning, aussi appelé apprentissage automatique en français, est une forme d'intelligence artificielle permettant aux ordinateurs d'apprendre sans avoir été programmés explicitement à cet effet. Cette technologie permet de développer des programmes informatiques pouvant changer en cas d'exposition à de nouvelles données. Découvrez la définition, le fonctionnement et les secteurs d'applications du Machine Learning.
Le machine Learning est une méthode d'analyse de données permettant d'automatiser le développement de modèle analytique. Par le biais d'algorithmes capables d'apprendre de manière itérative,le machine learning permet aux ordinateurs de découvrir des insights cachées sans être programmés pour savoir où les chercher [53].

3.3.1 Cycle de développement du Machine Learning

le domaine d'application du machine learning est trés varié :
La prédiction de valeurs financières, intrusion dans le domaine de la sécurité informatique, le moteur de recherche l'influençable par le profil de l'utilisateur, la détection de vols de machine, l'implémentions d'un anti-virus et la cryptanalyse. Le cycle de vie d'une implémentation de machine learning est la suivante :

1. Obtention et nettoyage des données.
2. Réalisation du modéle.
3. phase d'apprentisage.
4. phase de validation.
5. phase d'execution.

Le développement d'un modèle de Machine Learning ne se fait pas du premier coup. On commence souvent avec une première idée de modèle simple et rapide à développer, puis on analyse si on a une variance ou un biais et on tente une nouvelle idée pour corriger les problèmes rencontrés, etc., La figure 3.2 résume le cycle de développement d'un projet de Machine Learning.

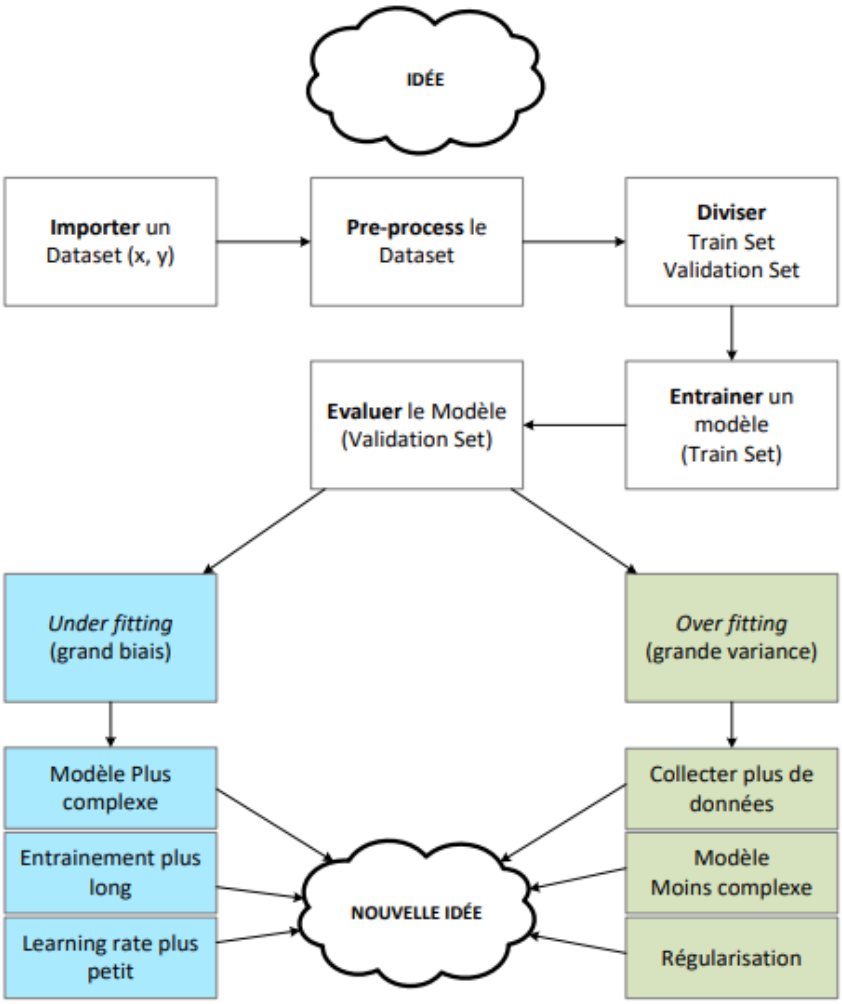

FIGURE 3.2 – Cycle de développement du Machine Learning

3.4 Types de systèmes d'apprentissage

Il existe plusieurs types de système d'apprentissage et cela varie en fonction du type de problème que l'on se pose. Il est alors utile de les classer en différentes catégories.
Les systèmes de machine learning peuvent être classés en fonction de l'importance et de la nature de la supervision qu'ils requièrent durant la phase d'entraînement. On distingue alors quatre grandes catégories : l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage semi-supervisé et l'apprentissage avec renforcement.

3.4.1 Apprentissage supervisé

L'apprentissage supervisé est une catégorie d'apprentissage automatique dans laquelle les algorithmes apprennent la variable d'entrée (X) en tant que superviseur ou enseignant pour prédire la variable de sortie (Y).
Cette catégorie est appelée apprentissage supervisé car l'algorithme peut apprendre de l'ensemble de données d'entraînement étiqueté,tandis que arrêter lorsque l'algorithme atteint un processus d'apprentissage niveau de performance approprié. Le plus couramment

utilisé les algorithmes d'apprentissage supervisé incluent des vecteurs de support Machine (SVM), régression linéaire, régression logistique, naïf Algorithme bayésien et k voisin le plus proche (KNN).
• Un exemple d'utilisation de l'apprentissage supervisé est le filtre anti-spam,l'apprentissage s'effectue à l'aide de nombreux exemples d'é-mails qu'on a étiqueté spam ou normal, a partir de cela,le filtre doit alors être capable de classer de nouveaux e-mails.
• Un autre exemple consiste à prédire le prix d'une voiture à partir des valeurs d'un certain nombre d'attributs ou variables qu'on appelle caractéristiques d'une observation ou features en anglais. Ces variables peuvent être le kilométrage, l'âge, la marque, etc Ils sont également appelés variables explicatives ou prédictives. L'entraînement se fait alors à partir de ces variables et des étiquettes. La catégorie de la variable prédite Y fait décliner. [54] [55] [56]

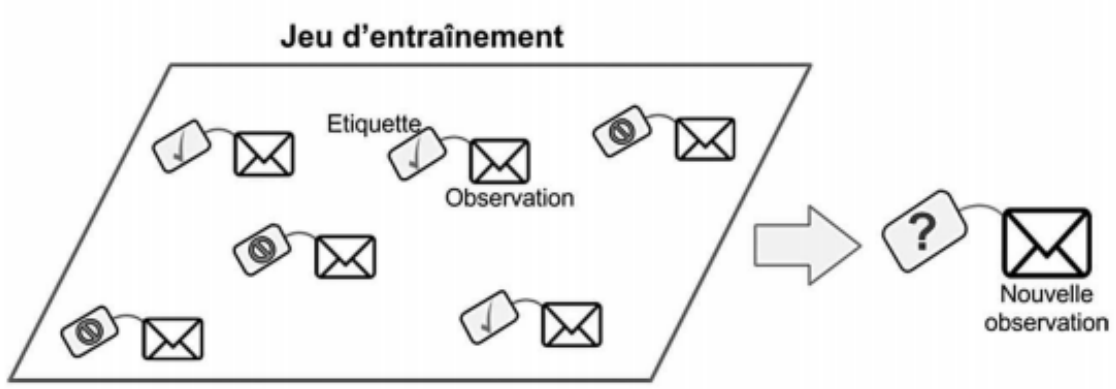

FIGURE 3.3 – Un jeu d'entraînement étiqueté pour un apprentissage supervisé

Voici quelques-uns des plus importants algorithmes d'apprentissage supervisé :
- K plus proches voisins.
- Régression linéaire.
- Régression logistique.
- Machines à vecteurs de support.
- Arbres de décision et forêts aléatoires.
- Réseaux neuronaux.

Il existe deux types d'apprentissage supervisé :

1. **Classification :**
   Un problème de classification survient lorsque la variable de sortie est une catégorie, telle que «rouge», «bleu» ou «maladie» et «pas de maladie». Exemples :
   - En finance et dans le secteur bancaire pour la détection de la fraude par carte de crédit (fraude, pas fraude).
   - Détection de courrier électronique indésirable (spam, pas spam).
   - Dans le domaine du marketing utilisé pour l'analyse du sentiment de texte (heureux, pas heureux).
   - En médecine, pour prédire si un patient a une maladie particulière ou non.

2. **Régression :**
   Un problème de régression se pose lorsque la variable de sortie est une valeur réelle, telle que «dollars» ou «poids».Exemples :[57]
   - Prédire le prix de l'immobilier.
   - Prédire le cours de bourse.

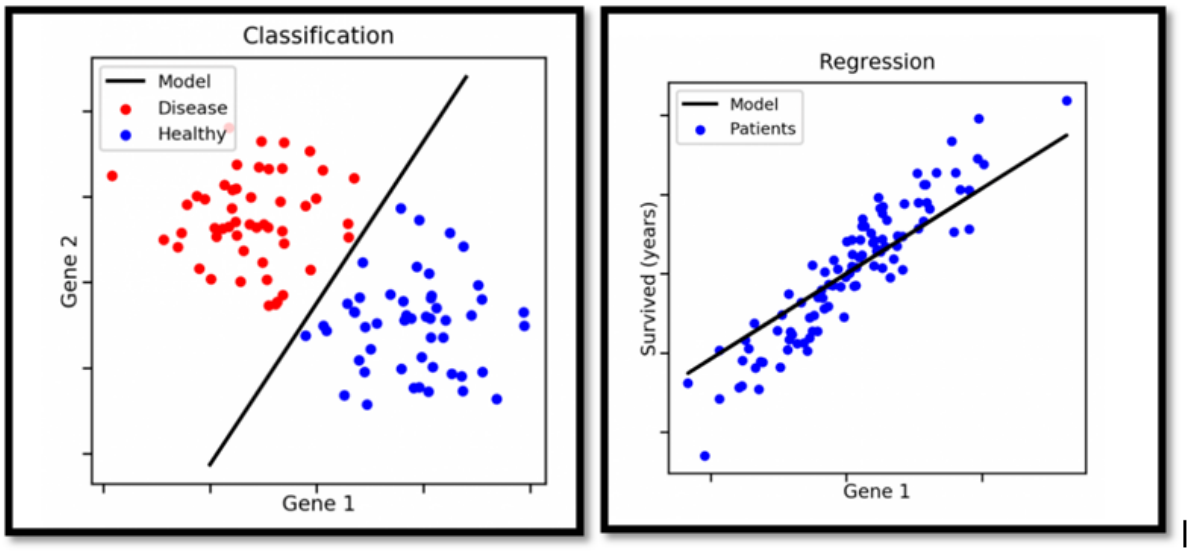

FIGURE 3.4 – La Classification et la Régression [57]

3.4.1.1 Problème Linéaire et Non-Linéaire :

Les méthodes de classification supervisée peuvent être basées sur :
- **des hypothèses probabilistes** (cas du classifieur naïf bayésien).
- **des notions de proximité** (exemple, k plus proches voisins).
- **des recherches dans des espaces d'hypothèses** (exemple, arbres de décisions).

Selon le problème, il faut pouvoir choisir un classificateur approprié, c'est-à-dire Il sera possible de séparer au maximum les données d'entraînement.
Si les exemples de classes différentes sont, on dit qu'un problème est **linéairement séparable**
Peut être complètement séparé par un hyperplan (appelé hyperplan de séparation ou séparateur), Ce type de problème est résolu par un classificateur assez simple dont le but est de trouver Équation de l'hyperplan séparant.
Mais, le problème peut également être non séparable de manière linéaire comme illustré dans la figure 3.3.
Dans ce cas, il faut utiliser d'autres types de classifieurs, souvent plus longs à paramétrer, mais qui obtiennent des résultats plus précis. [58]

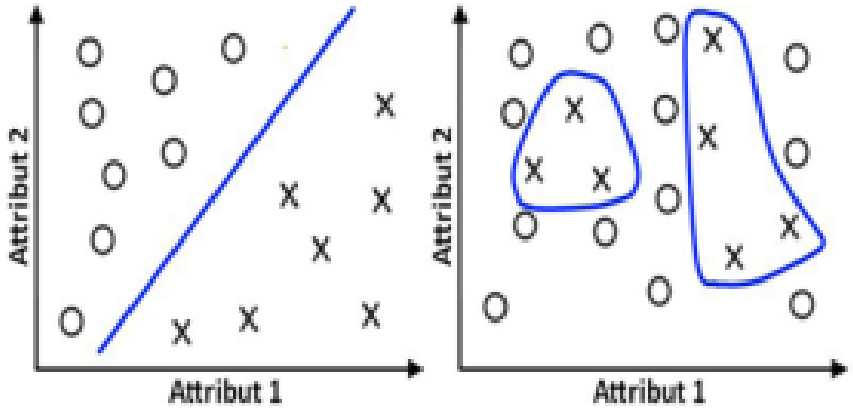

FIGURE 3.5 – A Gauche : Problème linéairement séparable (Frontière linéaire). A Droite : Problème non linéairement séparable

3.4.2 Apprentissage non supervisé :

Il vise à concevoir un modèle structurant l'information. La différence ici est que les comportements (ou catégories ou encore les classes) des données d'apprentissage ne sont pas connus, c'est ce que l'on cherche à trouver [59].
Les différents genres des approches d'apprentissage non supervisé peut être divisé en deux groupes :.

- **Clustering :** les points de données sont regroupés dans différents groupes. Les points de données avec des caractéristiques ou des propriétés similaires seront être regroupés dans le même groupe, tandis que les points de données qui ne partager des caractéristiques similaires seront regroupés dans différents groupes
- **Association :** Une technique de ML basée sur des règles conçue pour découvrir la relation ou l'association entre un grand ensemble d'éléments.

3.4.3 Apprentissage semi-supervisé :

L'apprentissage semi-supervisé vise à résoudre des problèmes avec relativement peu de données marquées et grandes quantités de données non marquées.
Apprentissage semi-supervisé simplifié par rapport à la formation, le temps d'étiqueter de grandes quantités de données surveillance.
Il a été démontré que l'utilisation de données non étiquetées, combinées à des données améliore considérablement la qualité de l'apprentissage. Ce gars l'objectif du processus d'apprentissage est de classer certaines données non étiquetées en utilisant les méthodes suivantes une collection d'informations marquées.
Un exemple illustrant l'utilisation d'un apprentissage semi-supervisé est le service d'héber gement d'image : Google Photos. Une fois avoir téléchargé des photos de famille sur ce service, le système arrive à reconnaître qu'une personne A apparaît sur telle ou telle photos et qu'une personne B sur telle autres.
Cela est dû à la partie non supervisé de l'algorithme. Une fois que vous aviez identifié

ces personnes, juste une étiquette par personne, le système sera capable de nommer les personnes figurant sur chaque photo, ce qui est utile pour des recherches ultérieures [59].

3.4.4 Apprentissage avec renforcement

L'apprentissage par renforcement est très différent des types d'apprentissage vus jusqu'ici. Il consiste à apprendre, à partir d'expériences successives, ce qu'il convient de faire de façon à trouver la meilleure solution.
Le système d'apprentissage, que l'on appelle ici « agent », interagit avec l'environnement, en sélectionnant et accomplissant des actions afin de trouver la solution optimale et obtenir en retour des récompenses.
L'agent essaie plusieurs solutions, on parle d'«exploration », observe la réaction de l'environnement et adapte son comportement, c'est-à-dire les variables pour trouver la meilleure stratégie. Pour ce type d'apprentissage, les données d'entraînement proviennent directement de l'environnement [59].

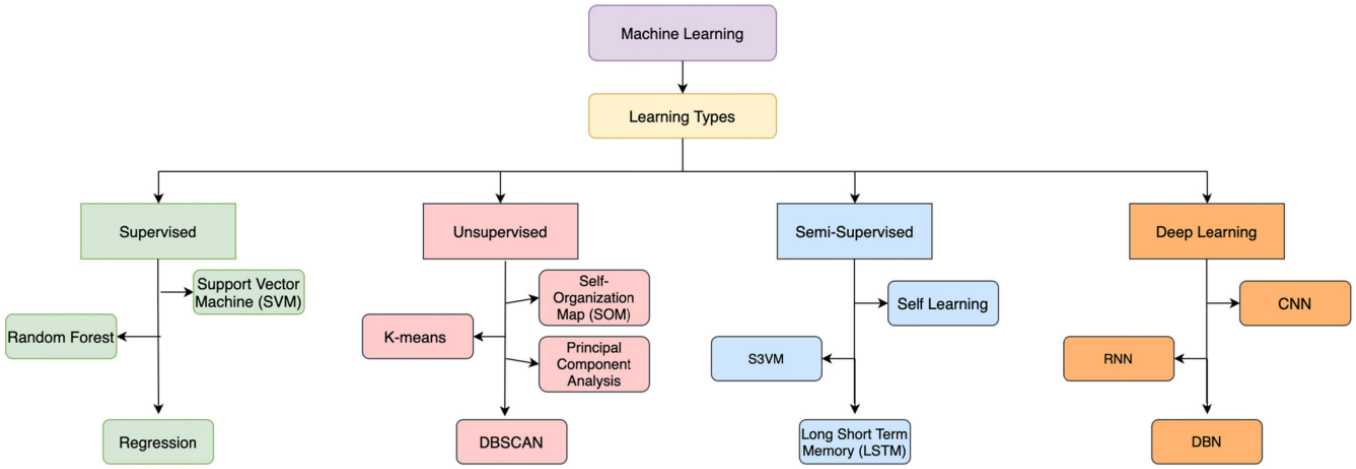

FIGURE 3.6 – Les techniques de ML

3.5 Algorithmes de Classification :

La classification est une méthode d'exploration de données utilisée pour attribuer des instances de données à l'une des quelques catégories. De nombreux algorithmes de classification ont été développés pour se surpasser.
Ils utilisent toues des techniques mathématiques pour fonctionner, telles que les arbres de décision, la programmation linéaire et les réseaux de neurones. Ces techniques analysent les données disponibles de diverses manières pour faire des prédictions. Cependant, certains sont plus couramment utilisées que d'autres :

3.5.1 Naïve bayes :

Naive Bayes Classifier est un algorithme populaire en Machine Learning. C'est un algorithme du Supervised Learning utilisé pour la classification. Il est particulièrement utile pour les problématiques de classification de texte. Un exemple d'utilisation du Naive

Bayes est celui du filtre anti-spam [58].
Bayes théorème :

$$P(H/X) = \frac{P(X/H.P(H)}{P(H)}$$

X :Données avec des classes qui n'ont pas connu
H : L'hypothèse des données est une classe spécifique
P(H|X) : La probabilité de l'hypothèse H est basée sur la condition X (probabilité a priori)
P (H) : Probabilité de l'hypothèse H (probabilité a priori)
P (X|H) : Probabilité X basée sur la condition de l'hypothèse H (la probabilité que l'évenement X se réalise sachant que l'évenement H s'est déjà réalisé).
P (X) : Probabilité X.
Où C et X sont deux événements (par exemple, la probabilité que le train arrivera à l'heure étant donné que le temps est pluvieux).
De tels classificateurs naïfs de Bayes utilisent la probabilité théorie pour trouver la classification la plus probable d'un instance (non classée) .
L'algorithme effectue positivement avec des données catégorielles mais mal si on a données numériques dans l'ensemble d'apprentissage (the training set).

3.5.1.1 Avantages et limitations du Naive Bayes Classifier

. **Avantages :**

- le Naive Bayes Classifier est très rapide pour la classification :
  en effet les calculs de probabilités ne sont pas très coûteux.
- La classification est possible même avec un petit jeu de données.

Inconvénients :

- l'algorithme Naive Bayes Classifier suppose l'indépendance des variables : C'est une hypothèse forte et qui est violée dans la majorité des cas réels[45] cependant, il est très sensible à leur corrélation. [59].

3.5.2 Decision tree classifier

Cette technique divise le problème de classification en sous-problèmes. Il crée un arbre de décision qui est utilisé pour développer un modèle utilisé aux fins de la classification. Effectuer une classification est relativement simple. Il suffit de répondre aux questions en descendant dans l'arbre de décision.
Les arbres de décision sont des modèles non-paramétriques et trouvent des règles en général assez puissantes. Ils peuvent traiter des très grands dataset et ils peuvent aussi utiliser des prédicateurs mixtes (catégoriques et nombres).
Les variables redondantes sont éliminées, parce que l'arbre ne les prend même pas en compte
A chaque noeud, on doit donc trouver deux choses : quelle variable 'feature' utiliser et quel est le point de séparation des deux zones, de sorte à minimiser l'erreur quadratique pour la régression et l'impureté pour la classification. On fait ainsi grandir l'arbre de la

manière la plus grande possible.[60]

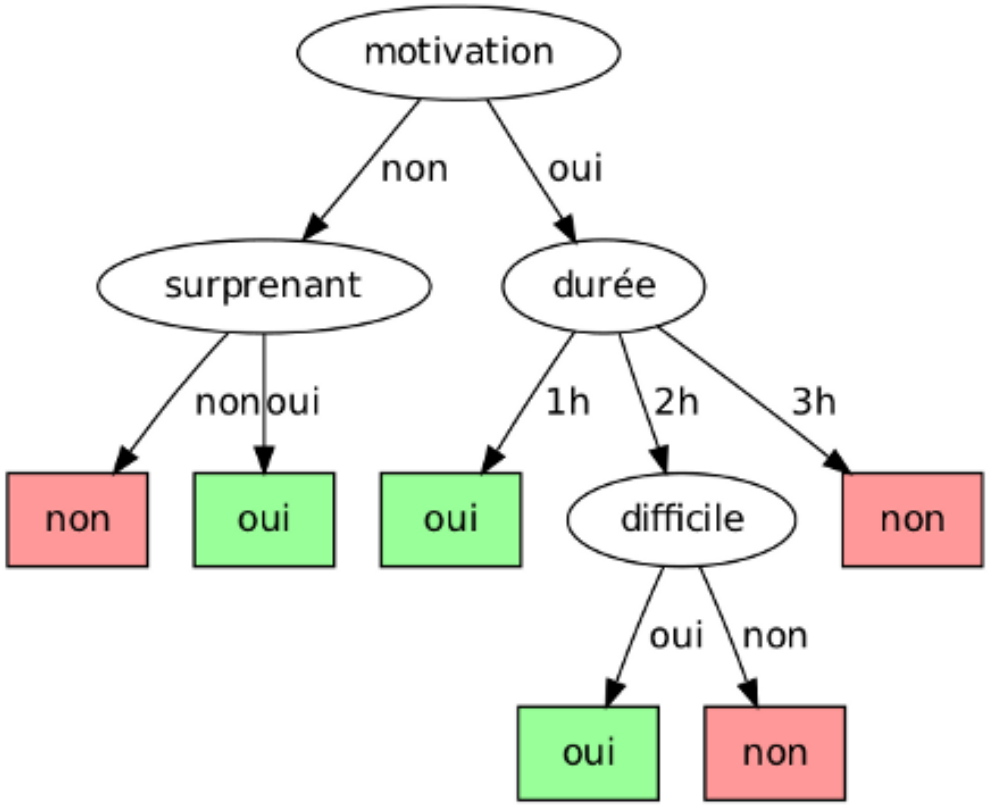

FIGURE 3.7 – Decision tree classifier

3.5.2.1 Avantages et limitations du Decision tree classifier

Avantages :

- Ils sont faciles à comprendre.
- Ils permettent de sélectionner l'option la plus appropriée parmi plusieurs.
- Temps d'exécution raisonnable.
- Il est facile de les associer à d'autres outils de prise de décision.[73]

Inconvénients :

- L'apprentissage de l'arbre de décision optimal est NP- complet concernant plusieurs aspects de l'optimalité.
- L'apprentissage par arbre de décision peut amener des arbres de décision très complexes, qui généralisent mal l'ensemble d'apprentissage .[74]

3.5.3 L'algorithme K Nearst Neighbors (K-NN) :

L'algorithme des K plus proches voisins ou K-nearest neighbors (kNN) est un algorithme de Machine Learning qui appartient à la classe des algorithmes d'apprentissage

supervisé simple et facile à mettre en œuvre qui peut être utilisé pour résoudre les problèmes de classification et de régression [61] . Pour effectuer une prédiction, l'algorithme K-NN va se baser sur le jeu de données en entier. En effet, pour une observation, qui ne fait pas parti du jeu de données, qu'on souhaite prédire, l'algorithme va chercher les K instances du jeu de données les plus proches de notre observation.
Ensuite pour ces K voisins, l'algorithme se basera sur leurs variables de sortie (output variable) y pour calculer la valeur de la variable y de l'observation qu'on souhaite prédire[62].
En général, pour définir la distance entre deux objets x et y, la formule de distance euclidienne est utilisée dans l'équation suivante :

$$D(x, y) = \sqrt{\sum_{j=1}^{n}(x_j - y_j)^2}$$

par ailleurs :

- Si K-NN est utilisé pour la régression, c'est la moyenne (ou la médiane) des variables y des K plus proches observations qui servira pour la prédiction.
- Si K-NN est utilisé pour la classification, c'est le mode des variables y des K plus proches observations qui servira pour la prédiction.

3.5.3.1 Avantages et limitations du K-NN

Avantages :

- L'algorithme est **simple et facile à mettre en œuvre**.
- Il n'est pas nécessaire de créer un modèle, de régler plusieurs paramètres ou de formuler des hypothèses supplémentaires.
- L'algorithme est polyvalent. Il peut être utilisé pour **la classification** ou **la régression**.

Inconvénients :

- L'algorithme devient beaucoup plus lent à mesure que le nombre d'observation et de variables indépendantes augmente.

3.5.4 Support vecteur Machine(Svm)

Les machines à vecteurs de support (SVM) sont un ensemble de méthodes d'apprentissage supervisé utilisées pour la classification, la régression et la détection des valeurs aberrantes [63]
En effet, l'idée principale des SVM est de reconsidérer le problème dans un espace de dimension supérieure, éventuellement de dimension infinie.
Dans ce nouvel espace, il est alors probable qu'il existe un hyperplan séparateur linéaire. Si c'est le cas, les SVM cherchent parmi l'infinité des hyperplans séparateurs celui qui maximise la marge entre les classes [64].

-

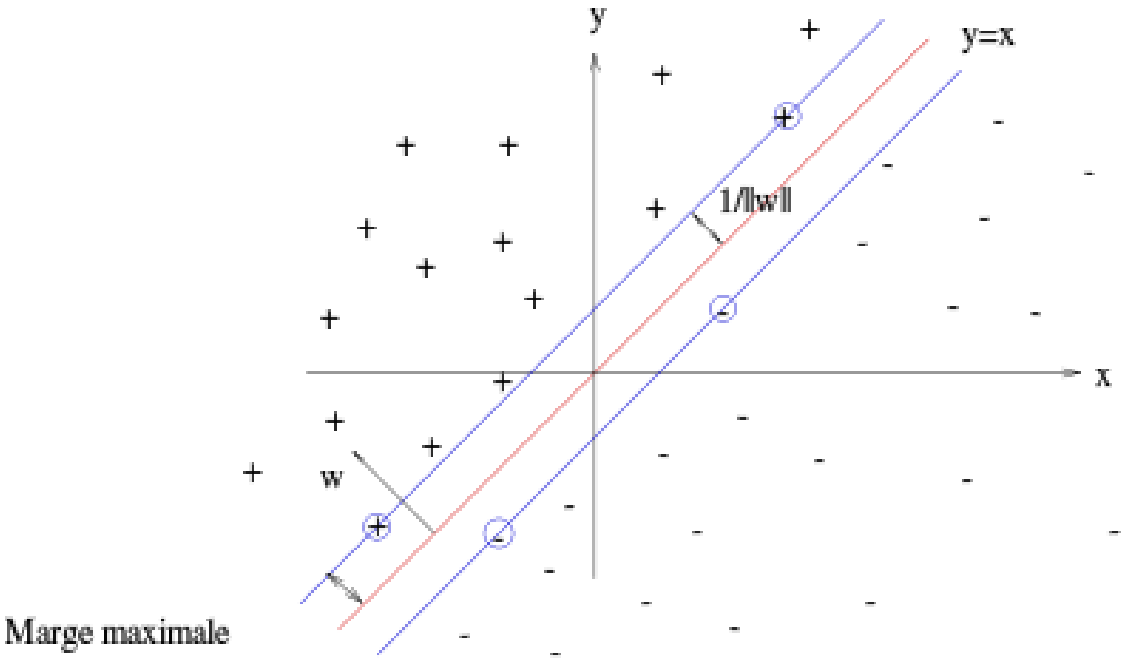

FIGURE 3.8 – Principe du Séparateur à Vaste Marge (SVM)

3.5.4.1 Avantages et limitations du SVM

Avantage :

- Efficace dans les espaces de grande dimension.
- Toujours efficace dans les cas où le nombre de dimensions est supérieur au nombre d'échantillons.
- Utilise un sous-ensemble de points d'apprentissage dans la fonction de décision (appelés vecteurs de support), il est donc également efficace en mémoire.
- différentes fonctions du noyau peuvent être spécifiées pour la fonction de décision. Des noyaux communs sont fournis, mais il est également possible de spécifier des noyaux personnalisés.

Inconvénients :

- Si le nombre de caractéristiques est bien supérieur au nombre d'échantillons, évitez le sur-ajustement dans le choix des fonctions du noyau et le terme de régularisation est crucial.
- Les SVM ne fournissent pas directement d'estimations de probabilité, celles-ci sont calculées à l'aide d'une validation croisée quintuple coûteuse.[65]

3.5.5 Random Forest Classifier

Random Forest Classifier est un algorithme d'apprentissage supervisé et composé de plusieurs arbres de décision. En faisant la moyenne de l'impact de plusieurs arbres de décision, les forêts aléatoires tendent à améliorer la prédiction.[66]
Il fonctionne en quatre étapes :

1-Sélectionnez des échantillons aléatoires à partir d'un ensemble de données donné.

2- Construire un arbre de décision pour chaque échantillon et obtenez un résultat de prédiction de chaque arbre de décision.

3- Effectuez un vote pour chaque résultat prévu.

4- Sélectionnez le résultat de prédiction avec le plus votes comme prédiction finale.
-

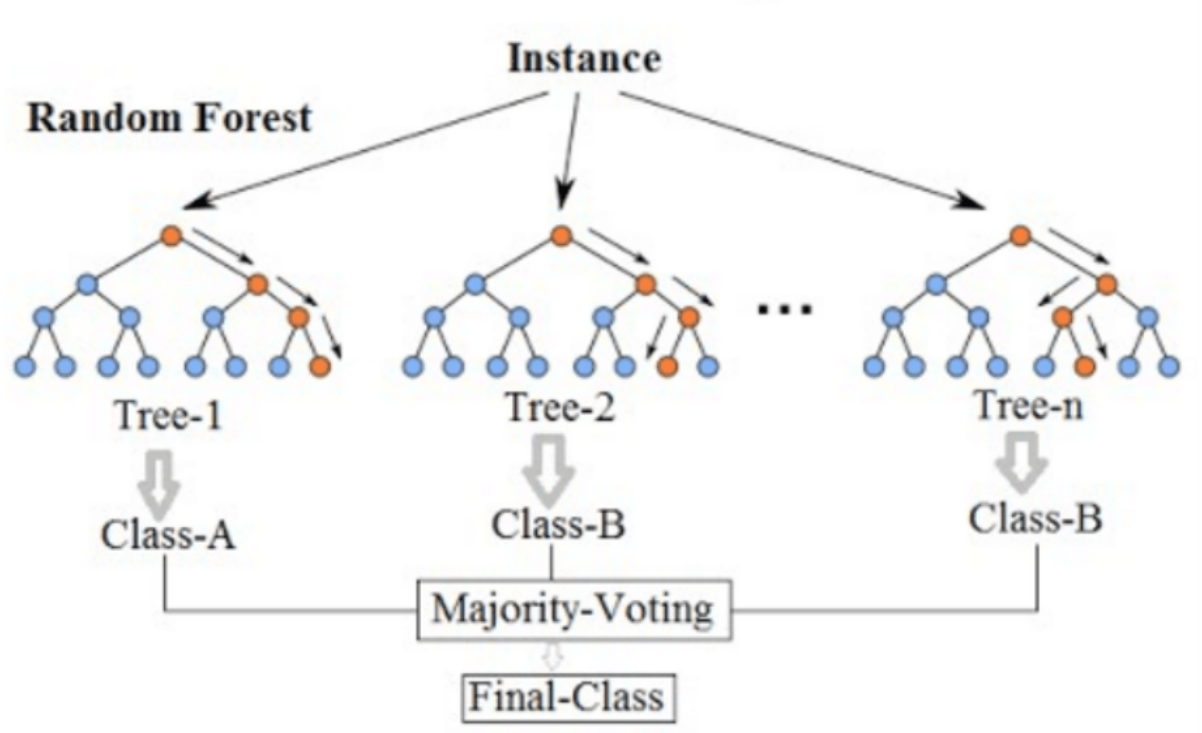

FIGURE 3.9 – Random Forest Classifier

3.5.5.1 Avantages et limitations du Random Forest Classifier

Avantages :

- une méthode très précise et robuste en raison de nombre d'arbres de décision participant au processus.
- L'algorithme peut être utilisé dans les problèmes de classification et de régression.

Inconvénients :

- Le modèle est difficile à interpréter par rapport à un arbre de décision, où l'on peut facilement faire une décision en suivant le chemin dans l'arbre.[66]

3.5.6 Réseau de neurones

Le nom "réseau de neurones" vient du fait que Développé à l'origine comme modèle de cerveau Humanité. Chaque unité du modèle représente un neurone et les connexions

entre unité (les liens, ou connexions, qui apparaissent dans la Figure 3.8) représentent les synapses.
Le terme réseau neuronal englobe une grande classe de modèles. Nous décrivons ici le réseau de neurones le plus utilisé, souvent appelé "réseau à couche cachée unique avec apprentissage par rétropropagation" ou "perceptron simple couche" [67]. **Un neurone** est une cellule du système nerveux spécialisée dans la communication et le traitement d'informations.[68]

-

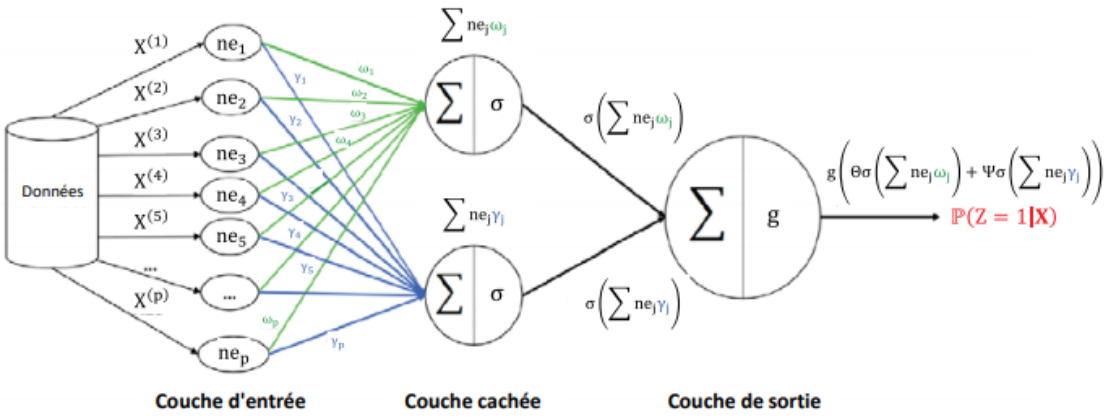

FIGURE 3.10 – Illustration d'un réseau de neurones

3.5.6.1 Avantages et limitations du Réseau de neurones

Avantage :

- Capacité de représenter n'importe quelle fonction, linéaire ou pas, simple ou complexe
- Résistance au bruit ou au manque de fiabilité des données

Inconvénients :

- L'absence de méthode systématique permettant de définir la meilleure topologie du réseau et le nombre de neurones à placer dans la (ou les) couche(s) cachée(s).
- Le choix des valeurs initiales des poids du réseau et le réglage du pas d'apprentissage,qui jouent un rôle important dans la vitesse de convergence.[67]

3.6 Métriques utilisés :

3.6.1 Cross validation :

On parle en générale de validation croisée à K blocs (ou K-fold cross validation) pour désigner une technique d'évaluation d'un algorithme de Machine Learning. Cela consiste à découper le dataset en K sous-ensemble (ou K folds) puis prendre un des K sous-ensemble comme dataset de validation (validation set) et les K-1 restants comme dataset d'entrainement (training set). On répète l'opération sur toutes les combinaisons possibles. On obtient K mesures de performance dont la moyenne représente la performance de l'algorithme.[69]

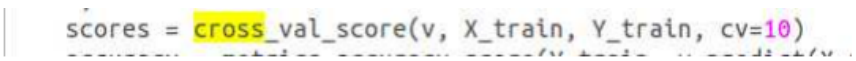

FIGURE 3.11 – Méthode de cross-validation

3.6.2 Accuracy :

Dans la classification multi-étiquette, cette fonction calcule la précision des sous-ensembles : l'ensemble d'étiquettes prédites pour un échantillon doit correspondre exactement a l'ensemble d'étiquettes correspondant dans y-true.

$$Accuracy = \frac{vraiPositif + vraiNégatif}{total}$$

3.6.3 Confusion matrix :

Une Confusion matrix (matrice de confusion) ou tableau de contingence est un outil permettant de mesurer les performances d'un modèle de Machine Learning en vérifiant notamment à quelle fréquence ses prédictions sont exactes par rapport à la réalité dans des problèmes de classification.[71]

ainsi, dans la classification binaire, le nombre de vrai négatifs sont C1, 0, les vrai positifs C1, 1, et les faux positifs C0, 1.

3.6.4 Classification reporte :

sklearn.metrics.classification-report :résumé textuel de la précision, rappel, score f1, pour chaque classe. Dictionnaire retourné si output-dict est true. Le dictionnaire a la structure suivante :

- précision :
  La précision est le rapport entre les vrais positifs et le total des vrais positifs et des faux positifs. La précision regarde pour voir combien de junk positifs ont été jetés dans le mélange. S'il n'y a pas de mauvais positifs (ces fps), alors le modèle a eu la précision 100%. Plus il y a de fps qui entrent dans le mélange, plus la précision sera laide. Mathématiquement :[70]

$$précision = \frac{tp}{tp + fp}$$

- Recall :
  Le rappel est la mesure de notre modèle identifiant correctement les vrais positifs. Ainsi, pour tous les patients qui souffrent réellement d'une maladie cardiaque, le rappel nous indique combien nous avons correctement identifiés comme ayant une maladie cardiaque. Mathématiquement :[70]

$$Recall = \frac{tp}{tp + fn}$$

- F1 Score :
  Le score F1 est la moyenne harmonique de la précision et du rappel'Recall' :[70]

$$F1Score = \frac{2(p * r)}{p + r}$$

3.6.5 balanced accuracy :

La précision équilibrée 'balanced accuracy' est une mesure que l'on peut utiliser pour évaluer la qualité d'un classificateur binaire. Il est particulièrement utile lorsque les classes sont déséquilibrées, i.e. une des deux classes apparaît beaucoup plus souvent que l'autre. Cela se produit souvent dans de nombreux contextes tels que la détection d'anomalies et la présence d'une maladie.Mathématiquement :[72]

$$balanced_accuracy = \frac{sensitivity + pecificity}{2}$$

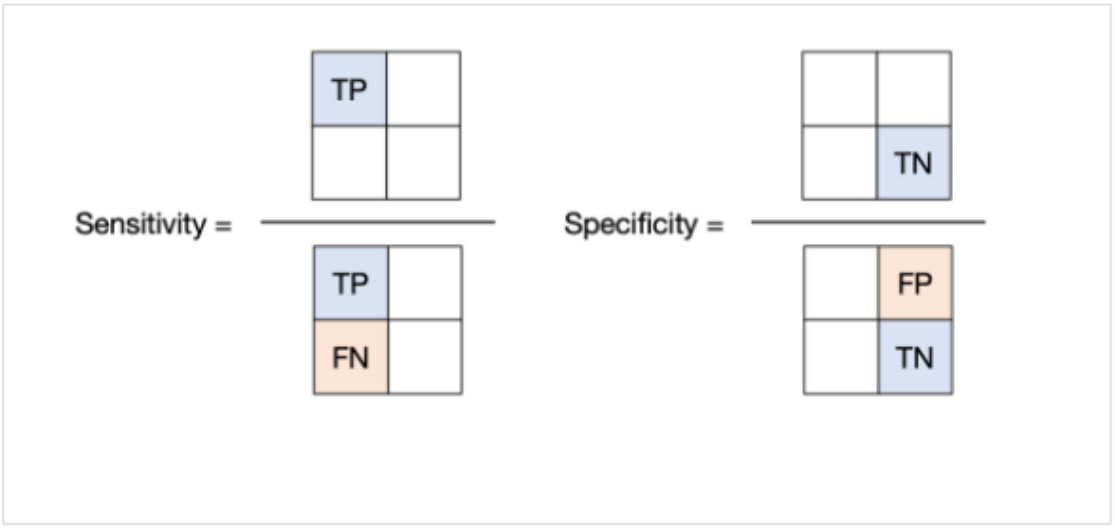

FIGURE 3.12 – Balanced accuracy

3.7 Conclusion

Ce chapitre nous a permis d'avoir une idée globale sur l'intelligence artificielle, l'importance de machine learning, le types de systèmes d'apprentissage et les algorithmes de classification avec leurs avantages et inconvénients. Le prochaine chapitre va présenté la conception et la réalisation de notre projet.

Chapitre 4

Conception et Réalisation

4.1 Introduction

Comme nous l'avons dit précédemment dans le chapitre trois que la machine learning est partout, donc dans ce chapitre, nous avons suivi les étapes d'un projet d'apprentissage automatique dans le but de détecter les attaques ddos dans un environnement cloud et d'obtenir des bons résultats pour cette découverte.

4.2 Conception de notre solutions :

Comme notre solution propos de la détection des attaques DDOS basé sur le machine learning, nous devons tout d'abord passer par les étapes de machine learning (nettoyage, prétraitement des données, normalisation, apprentissage, validation et exécution), le fonctionnement du systéme de détection d'attaque proposé est illustré dans un organigramme présenté dans la figure suivante :

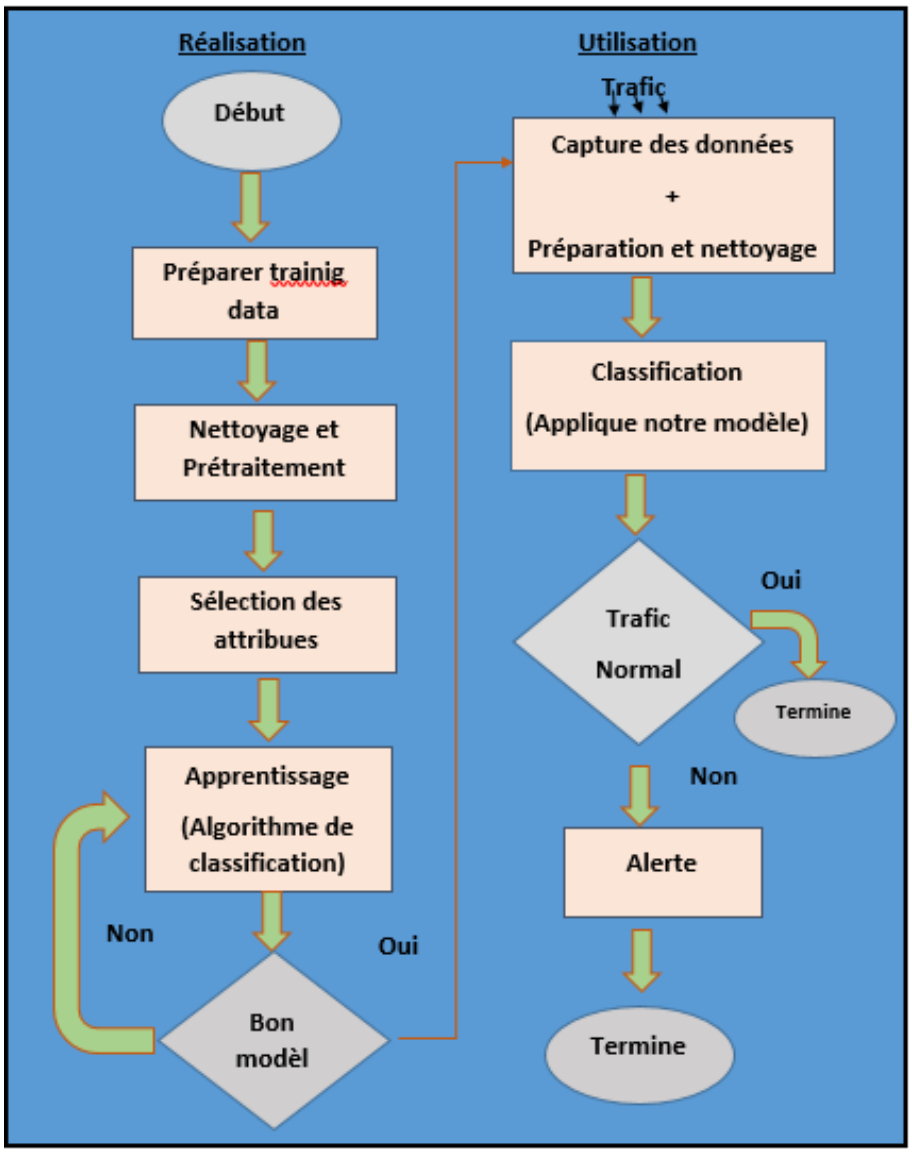

FIGURE 4.1 – Les étapes de machine learning

4.3 Solution dans un environnement cloud :

La figure 4.2 présente le système de détection des attaques ddos propose dans un environnement cloud.

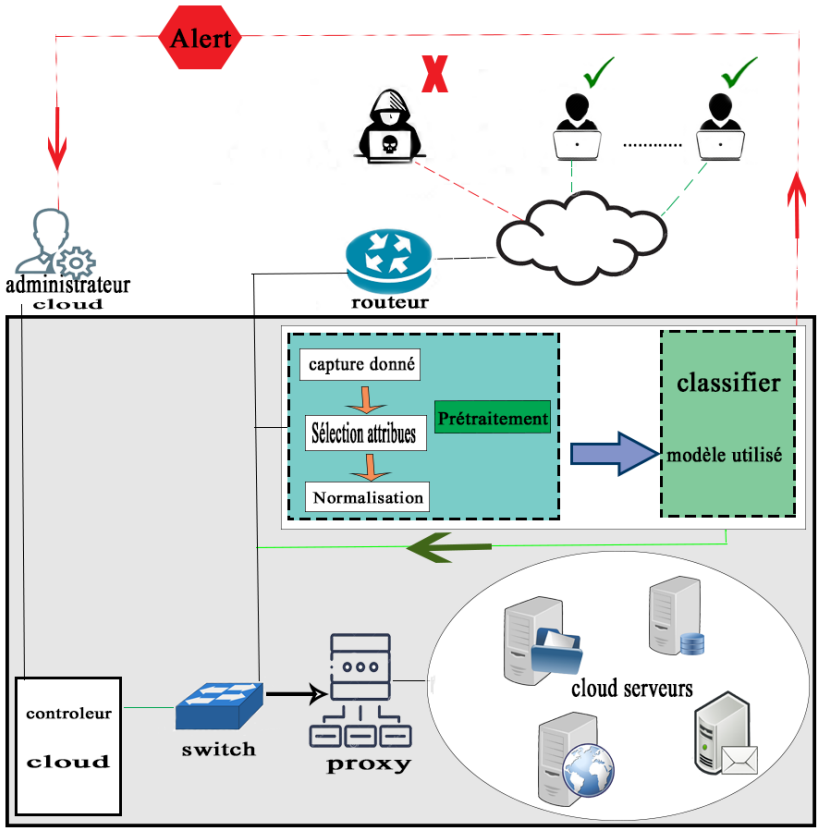

FIGURE 4.2 – Système de détection des attaques ddos propose dans un environnement cloud

4.4 Choix du data set :

Dans cette étude, l'ensemble de données utilisé a été obtenu à partir du CICDDoS2019. CIC-DDoS 2019 est la dernière data set synthétiques de haute qualité. Cet ensemble de données est un projet conjoint de l'établissement de la communication et sécurité canadien(CSE) et l'institut canadien de cyber sécurité (CIC).
Cette data set contient seulement des attaques ddos et bénin, elle a 84 variables qui sont plus détaillé dans la figure 4.3
Elle a été collecté dans deux jours pour le training et l'évaluation des tests. L'ensemble d'entraînement a été capturé le 12 janvier 2019, et contient 12 types d'attaques DDoS différents, chaque type d'attaque séparé dans un fichier PCAP. Les types d'attaque dans la journée de training comprend UDP, SNMP, NetBIOS, LDAP, TFTP, NTP, SYN, WebDDoS, MSSQL, UDP-Lag, DNS et SSDP DDoS based attack. Les données de test ont été créées le 11 mars 2019 et contiennent 7 types d'attaque DDoS SYN, MSSQL, UDP-Lag, LDAP, UDP, PortScan et NetBIOS. Ces captures est représentent dans la figure suivante :

Days	Attacks	Attack Time
First Day	PortMap	9:43 - 9:51
	NetBIOS	10:00 - 10:09
	LDAP	10:21 - 10:30
	MSSQL	10:33 - 10:42
	UDP	10:53 - 11:03
	UDP-Lag	11:14 - 11:24
	SYN	11:28 - 17:35
Second Day	NTP	10:35 - 10:45
	DNS	10:52 - 11:05
	LDAP	11:22 - 11:32
	MSSQL	11:36 - 11:45
	NetBIOS	11:50 - 12:00
	SNMP	12:12 - 12:23
	SSDP	12:27 - 12:37
	UDP	12:45 - 13:09
	UDP-Lag	13:11 - 13:15
	WebDDoS	13:18 - 13:29
	SYN	13:29 - 13:34
	TFTP	13:35 - 17:15

FIGURE 4.3 – les types d'attaques ddos dans dataset 2019

Les attributs de notre Base de Donnée :

Le tableau suivant donne la description des attributs de fichier CICDDoS 2019 avec l'explication de chaque attribut :

attributs	Description .
Source Port	Numéro du port source
Cloud Destination Port	Numéro du port destination
Protocol	Numéro du protocol utilisé
Flow Packets/s	débit de paquets de flux qui est le nombre de paquets transféré par seconde.
Flow IAT Mean	Le temps moyen entre les deux flux
Flow IAT Std	Ecart type temps deux flux
Flow IAT Max	Temps maximum entre deux flux
Flow IAT Min	Temps minimum entre deux flux
Flow Duration	Durée du debit
Total Fwd Packets	Nombre total de paquets dans le sens direct
Total Backward Packets	Nombre total de paquets dans le sens inverse

Fwd IAT Total	le Temps total entre deux paquets envoyés dans le transfert direction
Fwd IAT Meant	Le temps moyen entre deux paquets envoyés dans le sens avant.
Fwd IAT Std	le Temps d'écart type entre deux paquets envoyés dans le sens avant
Bwd IAT Total	le Temps total entre deux paquets envoyés dans le sens inverse
Bwd IAT Mean	Le temps moyen entre deux paquets envoyés dans le sens arrière
Bwd IAT Std	le Temps d'écart type entre deux paquets envoyés dans le sens arrière.
Bwd IAT Max	le Temps maximum entre deux paquets envoyés dans le sens arrière
Bwd IAT Mine	le Temps minimum entre deux paquets envoyés dans le sens arrièr
Fwd PSH Flags	le Nombre de fois où l'indicateur PSH a été défini dans les paquets se déplaçant dans le sens direct (0 pour UDP)
Bwd PSH Flags	Nombre de fois où l'indicateur PSH a été défini dans les paquets se déplaçant dans le sens arrière (0 pour UDP).
Fwd URG Flags	le Nombre de fois où l'indicateur PSH a été défini dans les paquets se déplaçant dans le sens arrière (0 pour UDP)
Bwd URG Flags	Nombre de fois où le drapeau URG a été défini dans les paquets se déplaçant dans le sens direct (0 pour UDP)
Fwd Header Length	le Nombre total d'octets utilisés pour les en-têtes vers l'avant.
Bwd Header Length	le Nombre total d'octets utilisés pour les en-têtes vers l'avant
Fwd Packets/s	le Nombre de paquets de transfert par seconde
Bwd Packets/s	le Nombre de paquets en arrière par seconde
Min Packet Length	Longueur minimale d'un flux
Max Packet Length	La longueur maximale d'un flux
Packet Length Mean	Longueur moyenne d'un flux
Packet Length Std	Longueur d'écart type d'un flux.
Packet Length Variance	Temps d'inter-arrivée minimum du paquet
FIN Flag Count	Nombre de paquets avec FIN
SYN Flag Count	Nombre de paquets avec SYN
RST Flag Count	Nombre de paquets avec RST
PSH Flag Count	le Nombre de paquets avec PUSH
ACK Flag Count	Nombre de paquets avec ACK
URG Flag Count	Nombre de paquets avec URG
CWE Flag Count	le Nombre de paquets avec CWE
ECE Flag Count	Nombre de paquets avec ECE
Down/Up Ratio	Taux de téléchargement et de téléchargement
Average Packet Size	La taille moyenne des paquets
Avg Fwd Segment Size	la Taille moyenne observée dans le sens direct
Avg Bwd Segment Sizee	la Taille moyenne observée dans le sens arrièr
Fwd Header Length.1	nombre total d'octets utilisée pour les en-tête vers l'avant
Fwd Avg Bytes/Bulk	Le nombre moyen d'octets débit en bloc dans le sens direct
Fwd Avg Packets/Bulk	Le taux de masse moyen du nombre de paquets dans le sens avant
Fwd Avg Bulk Rate	Le nombre moyen de taux de gros dans le contrat à terme direction
Bwd Avg Bytes/Bulk	Le taux de masse moyen du nombre d'octets dans le sens inverse
Bwd Avg Packets/Bulk	Le taux de masse moyen du nombre de paquets dans le sens arrière
Bwd Avg Bulk Rate	Le nombre moyen de taux de vrac dans le vers l'arrière direction
Subflow Fwd Packets	Le nombre moyen de paquets dans un sous-flux dans le sens avant
Subflow Fwd Bytes	Le nombre moyen d'octets dans un sous-flux dans le sens avant
Subflow Bwd Packets	Le nombre moyen de paquets dans un sous-flux dans le sens arrière
Idle Max	La durée maximale d'inactivité d'un flux avant de devenir actif
Idle Std	Déviation standard du temps pendant laquelle un flux était inactif avant de devenir actif
Idle Mean	Pendant ce temps un flux était inactif avant de devenir actif
Active Max	La durée maximale d'activité d'un flux avant de devenir inactif
Active Std	Écart type de temps pendant lequel un flux était actif avant devenir inactif

Label	0 ou 1 c'est-a-dire BENIGN Ou ddos
Subflow Bwd Bytes	Le nombre moyen d'octets dans un sous-flux dans le sens arrière
Active Min	La durée minimale d'activité d'un flux avant de devenir inactif
Active Mean	Le temps moyen pendant lequel un flux était actif avant de devenir inactif
min-seg-size-forward	Taille de segment minimale observée dans le sens direct
act-data-pkt-fwd	Le nombre de paquets avec au moins 1 octet de TCP charge utile de données dans le sens direct
Init-Win-bytes-backward	Le nombre d'octets envoyés dans la fenêtre initiale dans la direction arrière
Total Length of Fwd Packets	La taille totale des paquets dans le sens direct
Total Length of Bwd Packets	La taille totale des paquets dans le sens arrière
Fwd Packet Length Max	la Taille maximale des paquets dans le sens direct
Fwd Packet Length Min	La taille minimale des paquets dans le sens direct
Fwd Packet Length Mean	La taille moyenne des paquets dans le sens direct
Fwd Packet Length Std	la Taille de l'écart type des paquets dans le transfert direction
Bwd Packet Length Max	la Taille maximale des paquets dans le sens arrière
Bwd Packet Length Min	La taille minimale des paquets dans le sens arrière.
Bwd Packet Length Min	La taille minimale des paquets dans le sens arrière
Bwd Packet Length Mean e	la Taille moyenne des paquets dans le sens inverse
Bwd Packet Length Std	la Taille de l'écart type des paquets dans le sens inverse direction
Flow Bytes/s	débit d'octets qui est le nombre de paquets transférés par seconde

TABLE 4.1 – la description de chaque attribut.

4.5 Implémentions :

Voici les étapes que nous avons suivie pour faire l'implémentions de notre projet :

4.5.1 L'analyse exploratoire des données

C'est l'étape pour comprendre les variables pour définir la stratégie de modélisation. il se divise on deux partie :

1. **Analyse du forme**
   - Identification de la variable Target : Label (prédire si une attaque ddos ou bénin).
   - Nombre de lignes et de colonnes :

```
[8]  df.shape

     (140365, 84)
```

FIGURE 4.4 – Le nombre de ligne et de colonne

   - Types de variables : on a deux types de variables :
     qualitative = 45 et quantitative = 39.

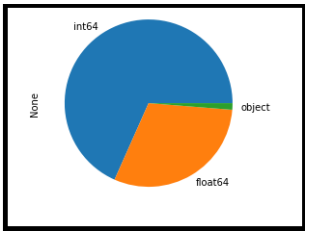

FIGURE 4.5 – Les types de variables

- Identification des valeurs manquantes : On peut voir si on a des valeurs manquantes à travers la requête suivante :

```
(df.isna().sum()/df.shape[0]).sort_values(ascending= True)
```

FIGURE 4.6 – Identification des valeurs manquantes

2. **Analyse du Fond :**
   - **La visualisation de la Target** :
     - 0.77% de ddos
     - 0.23% de bénin

     On peut visualiser ça dans la figure suivante :

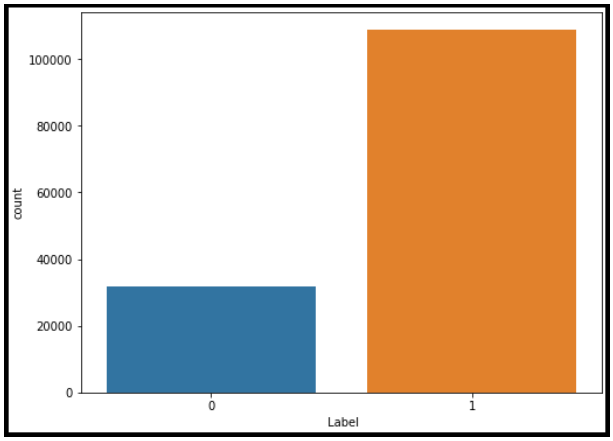

FIGURE 4.7 – visualisation de la target

   - **Visualisation des relations d'attributs :**
     Pour mieux comprendre les données et voire quelque statistique, il est souvent utile de les visualiser. On peut représenter les données dans un espace avec la bibliothèques seaborn. Voici un exemple de représentation graphique pour la relation de la target avec quelques variables :

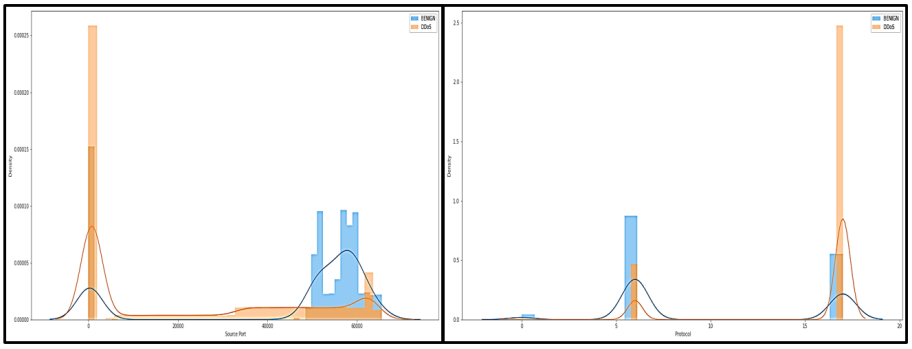

FIGURE 4.8 – la relation Target-port source/la relation de la terget-protocole

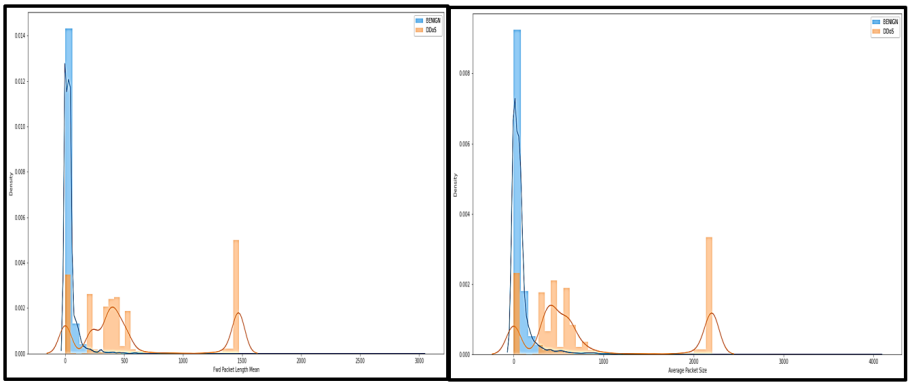

FIGURE 4.9 – La relation de la Target-FWD packet length mean/La relation de la Target-average packet size

4.5.2 Prétraitement :

Le prétraitement c'est l'étape qui consiste a préparé nos données avant de les fournir à la machine pour son apprentissage, notre objectif et de :

1. **Mètre les données dans un format propice au machine learning :**
   - **Préparation du data set :**
     Nous avons formé la base de données accumulant 5% de chaque fichier puis nous l'avons enregistrer dans un fichier csv.
     cette 'data set' l'ensemble du training est : 140365 et l'ensemble du test est : 83624.
     Les figures suivantes présente notre étape de préparation du data set :

```
#prenez 5% de dataset
df1partie1=df1.sample(frac=0.05).reset_index(drop=True)
#afficher le nombre de lignes et de colonnes
print(df1partie1.shape)
```

FIGURE 4.10 – préparation du data set

```
7]:  #concatener les trois types dattack exploitation
     ddos=pd.concat([dat1, dat2, dat3, dat4, dat5, dat6, dat7, dat8, dat9, dat10, dat11], ignore_index=True)
```

FIGURE 4.11 – préparation du data set

```
3]:  #sauvgarder la data set en format csv
     ddos.to_csv ('DDOS.csv', encoding = 'utf-8', index = False)
```

FIGURE 4.12 – préparation du data set

- **Encodage :**
  L'encodage consiste à convertir les données catégorielles (qualitatives) par une valeur numérique comprise entre 0 et le nombre de classes moins 1 (n classes-1.). Si la valeur de la variable catégorielle contient 5 classes distinctes, nous utilisons (0, 1, 2, 3 et 4).
  - Il existe cinq transformers d'encodage : Label Encoder, LabelBinarized, MultilabelBinarized, OnHotEncoder, OrdinalEncoder.
  Ces cinq transformers permet d'effectuer deux types d'encodage : -Encodage ordinal.
  -Encodage One-Hote.
  La figure suivante présente notre étape d'encodage :

```
Entrée [38]:  #creer un dictionnaire
              code={'Syn':1,
                   'UDP-lag':1,
                   'DrDoS_UDP':1,
                   'WebDDoS':1,
                   'UDP':1,
                   'UDPLag':1,
                   'MSSQL':1,
                   'DrDoS_MSSQL':1,
                   'TFTP':1,
                   'DrDoS_DNS':1,
                   'DrDoS_SNMP':1,
                   'DrDoS_SSDP':1,
                   'DrDoS_LDAP':1,
                   'DrDoS_NTP':1,
                   'DrDoS_NetBIOS':1,
                   'BENIGN':0}

Entrée [39]:  #encodage train set
              for col in dfd.select_dtypes('object'):
                  dfd[col]=dfd[col].map(code)


              #afficher le target %
              print(dfd[' Label'].value_counts())
              dfd.head()
```

FIGURE 4.13 – Encodage

- **Nettoyage :**
  (a) Élimination des nan, inf : c'est-à-dire éliminé tout les valeurs manquantes et tout valeurs infinis en peut voir ça dans la figure suivante :

```
df3=df3[~df3.isin([np.nan, np.inf, -np.inf]).any(1)]

df3=df3.dropna()
```

FIGURE 4.14 – Code élimination des nan,inf

  (b) Élimination des valeurs ou la variance égale à zéro :
  nous procédons au calcul de la variance et puis l'élimination des valeurs où la variance égale à zéro. Après le calcul de la variance on a Quinze

valeurs dont la variance est zéro et voici un exemple de quelque valeur avec le code d'élimination.

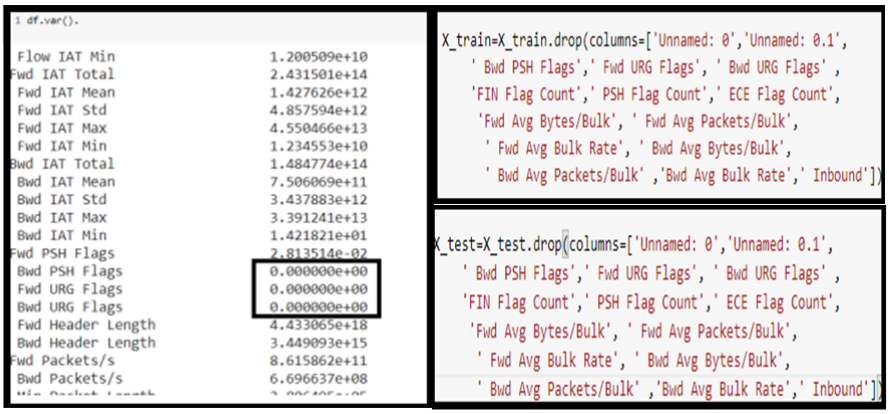

FIGURE 4.15 – Code de la variance

Après l'élimination des valeurs ou la variance est égale a zéro on obtient 68 variables 'features' au lieu de 83.

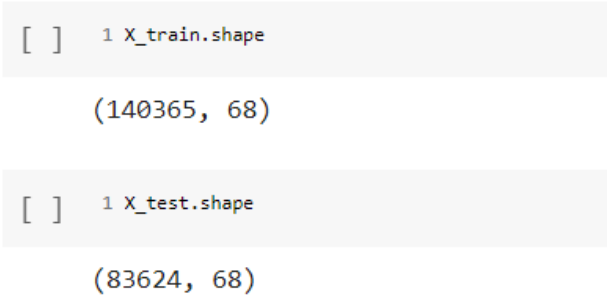

FIGURE 4.16 – Le nombres d'attributs

(c) Élimination Duplicate :
pour supprimer les lignes qui sont répétées La figure suivante représente le code d'élimination du Duplicate :

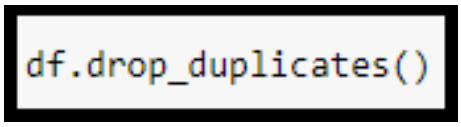

FIGURE 4.17 – Code élimination duplicate

2. Améliorer la performance de notre modèle.
   - **Normalisation des donnés et feature scaling :**
     La plupart du temps, en machine Learning, les bases de données proviennent avec des ordres de grandeurs différents.Cette différence d'échelle peut conduire à des performances moindres.
     Pour palier à cela, des traitements préparatoires sur les données existent. nous

appliquerons un procédé qui s'appelle **:feature scalling** [76]

Les différentes techniques de Feature Scaling :
Normalisation :
il existe beaucoup de technique de normalisation :

(a) **normalisation min-max :**
La normalisation **Min Max** conserve la distribution de scores originale à un facteur d'échelle près et transforme tous les scores dans l'intervalle [0,1].[77]

$$Xscaled = \frac{X - Xmin}{Xmax - Xmin}$$

(b) **standarisation :**
transforme nos données a ce que chaque variable est une moyenne égale à zéro et un écart-type égale à 1.

$$Xscaled = \frac{X - Ux}{\delta x}$$

X : chaque variable.
Ux : moyenne initial de notre variable.

La figure 4.17 présente la partie du notre code :

```
scaler= StandardScaler()

X_train=  pd.DataFrame(scaler.fit_transform(X_train),columns=X_train.columns)

X_test=  pd.DataFrame(scaler.fit_transform(X_test),columns=X_test.columns)
```

FIGURE 4.18 – code de normalisation

(c) **Le transformeur Robuste skyler :**
Soustrait nos données a la médiane de chaque variable et on divise par l'inter quartile de nos donnée.

$$Xscaled = \frac{X - Médiane}{IQR}$$

4.5.3 Modélisation

on a quatre etapes :
1- sélectionner un estimateur et préciser les paramètres :
Model= LinearRegression()
2- Entraîner le modèle(x,y) :
model.fit(x,y)
3- évalué le modèle :
Model.score(x,y)
4- utilisé le modéle :
Model.predict(x,y).

- **Aprentisage de l'algorithme par les données :**

Nous allos enfin pouvoir appliquer nos algorithme de classification, pour chaque algorithme nous allons montrer les différentes fonctions utilisés :

```
11
12 from sklearn.linear_model import LogisticRegression
13 from sklearn.neighbors import KNeighborsClassifier
14 from sklearn.tree import DecisionTreeClassifier
15 from sklearn.metrics import accuracy_score
16 from sklearn.naive_bayes import GaussianNB
17 from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
18 from sklearn.svm import  SVC
19 from sklearn.pipeline import make_pipeline
20 from sklearn.feature_selection import SelectKBest,f_classif
21 from sklearn.decomposition import  PCA
22
```

FIGURE 4.19 – importation des algorithmes

Nous avons utiliser six modèles de classification :
l'arbre de décision, l'algorithme des K plus proches voisins, les machines à vecteurs de support, le classificateur de forêt aléatoire, régression Logistique, gaussienne NB.

figure 4.19 représente les estimateurs Les algorithmes que nous avons utilisés.

```
DC = DecisionTreeClassifier(random_state=0,splitter='random',max_depth=12,criterion='entropy')

RF = RandomForestClassifier(n_estimators=100,random_state=0,max_depth=5)

svm = SVC(random_state=0)

knn = KNeighborsClassifier(n_neighbors=2)

lgr = LogisticRegression(solver='liblinear',max_iter=5)

NB = GaussianNB(var_smoothing=0.001)
```

FIGURE 4.20 – Déclaration des algorithmes

4.6 Conclusions

Dans ce chapitre nous avons donner un aperçu de notre solution et présenté notre choix de data set, et pour bien comprendre notre travail nous avons fourni les étapes que nous avons suivies dans notre projet.

Chapitre 5

Résultat et discution

5.1 Introduction

Dans ce cinquième chapitre, nous présenterons les résultats de chaque algorithme que nous avons obtenus, ainsi que les différentes méthodes de sélection des attributs. Enfin Nous faisions une comparaison avec les résultats d'une autre étude. [78]

5.2 Résultat des Algorithmes de machine learning

À travers le chapitre 04, nous avons appliquer six algorithmes de classification (l'arbre de décision, l'algorithme des K plus proches voisins, les machines à vecteurs de support, le classificateur de forêt aléatoire, régression Logistique, gaussienne NB).
Les tableaux suivants montrent les résultats de chacun d'eux pour obtenir le meilleur résultat.

Les tableaux suivants présentent les paramètres utilisée pour obtenir les meilleurs résultats :

1. **Application de l'algorithme «l'arbre de décision»**
   - Résultat des testes :

Algorithme	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
DT	Par défauts	0.72	0.99	1.00	0.15	0.84	0.25	0.734
	random_state=0 splitter='random' max_depth=10 criterion='entropy'	0.87	0.74	0.89	0.69	0.88	0.72	0.830

FIGURE 5.1 – L'algorithme « l'arbre de décision »

   - Meilleur résultat :

Algorithmes	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
DT	random_state=0 splitter='random' max_depth=12 criterion='entropy'	0.87	0.77	0.91	0.71	0.89	0.74	0.84

FIGURE 5.2 – Milleur résultat de l'algorithme « l'arbre de décision »

2. **Application de l'algorithme «le classificateur de forêt aléatoire»**
   - Résultat des testes :

Algorithme	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
RF	Par défauts	0.88	0.47	0.59	0.82	0.70	0.60	0.660
	random_state=0 max_depth=4 n_estimators=100	0.86	1.00	1.00	0.64	0.92	0.78	0.887

FIGURE 5.3 – L'algorithme « le classificateur de forêt aléatoire »

   - Meilleur résultat :

Algorithme	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
RF	random_state=0 max_depth=5 n_estimators=100	0.92	0.98	0.99	0.82	0.96	0.89	0.94

FIGURE 5.4 – L'algorithme « le classificateur de forêt aléatoire »

3. **Application de l'algorithme «Les machines à vecteurs de support»**
   - Résultat des testes :

Algorithme	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
SVM	Par défauts	1.00	0.98	0.99	0.99	0.99	0.99	0.991
	random_state=0	1.00	0.98	0.99	0.99	0.99	0.99	0.991

FIGURE 5.5 – L'algorithme « Les machines à vecteurs de support »

- Meilleur résultat :

Algorithme	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
SVM	random_state=0	1.00	0.98	0.99	0.99	0.99	0.99	0.99

FIGURE 5.6 – Meilleur résultat de l'algorithme « Les machines à vecteurs de support »

4. **Application de l'algorithme «L'algorithme des K plus proches voisins»**
   - Résultat des testes :

Algorithme	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
KNN	Par défauts	0.99	1.00	1.00	0.98	0.99	0.99	0.992
	n_neighbors=3	0.99	1.00	1.00	0.98	0.99	0.99	0.991

FIGURE 5.7 – L'algorithme des K plus proches voisins

- Meilleur résultat :

Algorithme	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
KNN	n_neighbors=2	0.99	0.99	1.00	0.98	0.99	0.99	0.99

FIGURE 5.8 – Meilleur résultat de l'algorithme «L'algorithme des K plus proches voisins»

5. **Application de l'algorithme « Régression Logistique » :**
   - Résultat des testes :

Algorithme	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
LR	Par défauts	0.94	1.00	1.00	0.86	0.97	0.92	0.955
	solver=' liblinear' max_iter=10	0.97	1.00	1.00	0.92	0.98	0.96	0.975

FIGURE 5.9 – L'algorithme « Régression Logistique » :

- Meilleur résultat :

Algorithme	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
LR	solver='liblinear' max_iter=5	0.98	1.00	1.00	0.96	0.99	0.98	0.97

FIGURE 5.10 – Meilleur résultat de l'algorithme « Régression Logistique » :

6. **Application de l'algorithme « Gaussienne NB » :**
   - Résultat des testes :

Algorithme	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
NB	Par défauts	0.00	0.31	0.00	1.00	0.00	0.47	0.310
	var_smoothing=0.01	0.99	0.92	0.96	0.97	0.97	0.95	0.965

FIGURE 5.11 – L'algorithme « Gaussienne NB » :

- Meilleur résultat :

Algorithmes	Paramètre	Précision		Recall		F1-score		Accuracy
		Attack	Benign	Attack	Benign	Attack	Benign	
NB	var_smoothing=0.001	0.99	0.92	0.96	0.98	0.98	0.95	0.97

FIGURE 5.12 – Meilleur résultat de l'algorithme « Gaussienne NB » :

5.3 Sélection des attributs :

La sélection des variables est un processus qui consiste à chercher dans l'ensemble des variables explicatives disponibles un ensemble optimal des caractéristiques les plus importantes à un système donné. Ceci est dans le but de mener à bien la tâche pour laquelle il a été conçu [85]. on a appliqué deux méthodes de la sélection : Analyse en composantes principales, univariate sélection.

1. Analyse en Composantes principales (PCA) : L'analyse en composantes principales (ACP ou PCA en anglais pour principal component analysis), est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorrélées les unes des autres. Ces nouvelles

variables sont nommées « composantes principales », ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante.
On a appliquer la méthode pca avec 20 feature. La figure ci-dessous présente une partie du code :

```
pca=PCA(n_components=20)
```

```
pca.fit(X_train)
pca.fit(X_test)
```

```
PCA(copy=True, iterated_power='auto', n_components=20, random_state=None,
    svd_solver='auto', tol=0.0, whiten=False)
```

```
x_pca=pca.transform(X_train)
x_pcaTest=pca.transform(X_test)
```

```
X_train.shape
```

```
(140365, 68)
```

```
x_pca.shape
```

```
(140365, 20)
```

FIGURE 5.13 – code de la méthode pca

Voici le résultat de chaque algorithmes dans les figures suivantes :

Algorithmes	Précision		Recall		F1-score		Accuracy
	Attack	Benign	Attack	Benign	Attack	Benign	
DT	0.98	0.98	0.99	0.95	0.98	0.96	0.98
RF	0.97	0.97	0.99	0.93	0.98	0.95	0.97
SVM	0.99	0.98	0.99	0.99	0.99	0.98	0.99
KNN	0.98	0.99	1.00	0.95	0.99	0.97	0.98
LR	0.97	1.00	1.00	0.94	0.99	0.97	0.98
NB	0.84	0.87	0.96	0.59	0.90	0.70	0.85
	0.85	0.87	0.96	0.61	0.90	0.72	0.85

FIGURE 5.14 – la méthode pca avec 20 feature

- Remarque :
  - Amélioration dans le résultat Random Forest Classifier et décision tree classifier .
  - Le résultat SVM restes stables (une petite différence dans balanced accuracy).
  - le résultat montre une légère différence : accuracy(0,992 -> 0,983) F1score(0,994 -> 0,987) balanced accuracy(0,974 ->0,989).
  - Une légère différence aussi dans le résultat de logistique Régression.
  - Une différance dans le résultat NB :accuracy (97 -> 85).
  - Donc on a essayé d'amélioré le résultat de l'algorithme NB avec les paramètre par défaut le résultat dans le tableau ci-dessous :

```
evaluation2(NBpca)

train score
0.8891390303850675
test score
0.8520281258968717
[[15912 10065]
 [ 2309 55338]]
              precision    recall  f1-score   support

           0       0.87      0.61      0.72     25977
           1       0.85      0.96      0.90     57647

    accuracy                           0.85     83624
   macro avg       0.86      0.79      0.81     83624
weighted avg       0.85      0.85      0.84     83624

f1 score
0.8994392523364487
accuracy
0.8520281258968717
balanced accuracy
0.7862438707261852
```

FIGURE 5.15 – Le résultat de l'algorithme NB

2. Univariate Selection (SelectKBest) :
   La méthode 'SelectKBest' sélectionne les caractéristiques en fonction du k score le plus élevé. En modifiant le paramètre 'score func', nous pouvons appliquer la méthode à la fois aux données de classification et de régression.
   La sélection des meilleures fonctionnalités est un processus important lorsque nous préparons un grand ensemble de données pour la formation.[85]
   On a applique la méthode SelectKBest avec 20 attributs. La figure ci-dessous présente une partie du code :

```
preprocessor =make_pipeline(SelectKBest(f_classif, k=20))#prendre 20 meilleur variable )
```

```
Dc= make_pipeline(preprocessor,DecisionTreeClassifier(random_state=0,splitter='random',max_depth=12,criterion='entropy'))
RandomForest = make_pipeline(preprocessor,RandomForestClassifier(n_estimators=100,random_state=0,max_depth=5))
SVM= make_pipeline(preprocessor, SVC(random_state=0))
KNN= make_pipeline(preprocessor, KNeighborsClassifier(n_neighbors=2))
LGR= make_pipeline(preprocessor, LogisticRegression(solver='liblinear',max_iter=5))
nb= make_pipeline(preprocessor, GaussianNB(var_smoothing=0.001))
```

FIGURE 5.16 – Code de la méthode SelectKBest

Le résultat de chaque algorithmes dans le tableau ci-dessous :

Algorithmes	Précision		Recall		F1-score		Accuracy
	Attack	Benign	Attack	Benign	Attack	Benign	
DT	0.97	0.64	0.88	0.49	0.83	0.56	0.76
RF	0.92	0.98	0.99	0.80	0.95	0.88	0.93
SVM	0.93	0.81	0.91	0.84	0.92	0.83	0.89
KNN	0.99	1.00	1.00	0.98	1.00	0.99	0.99
LR	0.96	1.00	1.00	0.91	0.98	0.95	0.97
NB	0.98	0.80	0.89	0.96	0.93	0.87	0.91

FIGURE 5.17 – La méthode SelectKBest

5.4 Comparaison entre les algorithmes :

Algorithmes	Précision		Recall		F1-score		Accuracy
	Attack	Benign	Attack	Benign	Attack	Benign	
DT	0.87	0.77	0.91	0.71	0.89	0.74	0.84
RF	0.92	0.98	0.99	0.82	0.96	0.89	0.94
SVM	1.00	0.98	0.99	0.99	0.99	0.99	0.99
KNN	0.99	0.99	1.00	0.98	0.99	0.99	0.99
LR	0.98	1.00	1.00	0.96	0.99	0.98	0.97
NB	0.99	0.92	0.96	0.98	0.98	0.95	0.97

FIGURE 5.18 – Comparaison entre les algorithmes

5.5 Comparaison entre les Résultats de la sélection des attributs :

	PCA				SelectKBest			
Algorithmes	Précision	Recall	F1-score	accuracy	Précision	Recall	F1-score	accuracy
DecisionTreeClassifier	0.97	0.98	0.98	0.98	0.79	0.87	0.83	0.76
RandomForestClassifier	0.96	0.98	0.97	0.97	0.91	0.99	0.95	0.93
SVM	0.99	0.99	0.99	0.99	0.92	0.91	0.91	0.89
KNeighborsClassifier	0.97	0.99	0.98	0.98	0.99	0.99	0.99	0.99
LogisticRegression	0.97	0.99	0.98	0.98	0.96	0.99	0.97	0.97
GaussianNB	0.83	0.96	0.89	0.85	0.97	0.89	0.93	0.91

FIGURE 5.19 – Comparaison entre les Résultats de Pca et Kbest

Déscusion :

- On remarque que la méthode pca est améliorer le résultat de l'algorithme arbre de décision par rapport à la première résultat(avant d'appliquer les méthodes de sélection des attributs), d'autre part en remarque que la méthode selectkbest est diminué le résultat (accuracy 76% et précision 79%).

- Les les Résultats de l'algorithme randomForest est augmenter lorsque on a appliquer la méthode pca (accuracy, f1 score 97% et précision 96%) par contre après l'application de la méthode selectkbest le résultat presque reste stable(accuracy 93%, précision 91%, donc la méthode pca augmente le résultat de l'algorithme randomeForest.

- La méthode PCA diminuer le résultat d'algorithme GausianNb par rapport au résultat avant l'application de cette methode (précision de 83%, accuracy de 85%.

- Lorsque on applique la méthode selectkbest le résultat de gaussienNB reste presque stable(accuracy 91% , f1scor 93% , recall 89% et précision de 97% , Mais en peut remarquer que la meilleure résultat pour ce algorithme est avant l'application des méthodes de la sélection.

- En remarque que les deux l'algorithme Knn et svm ont des bons résultats avant l'application des méthodes de la sélection avec une accuracy, f1score, recall, precision de 99%, d'autre part la méthode PCA garde le meme résultat de svm mais elle a diminuer le résultat de Knn avec f1scor, accuracy de 98% et pression 97% par contre lorsque on applique la méthode select kbest on a les meme résultats de knn et les résultats de svm diminue avec f1scor, recall 91%, precision 92% et accuracy 89%.
  La figure suivante présente la comparaison entre avant et après l'application des méthodes de selection :

	PCA				SelectKBest				Résultat avant			
Algorithmes	Précision	Recall	F1-score	accuracy	Précision	Recall	F1-score	accuracy	Précision	Recall	F1-score	accuracy
DecisionTreeClassifier	0.97	0.98	0.98	0.98	0.79	0.87	0.83	0.76	0.87	0.90	0.88	0.84
RandomForestClassifier	0.96	0.98	0.97	0.97	0.91	0.99	0.95	0.93	0.92	0.99	0.95	0.94
SVM	0.99	0.99	0.99	0.99	0.92	0.91	0.91	0.89	0.99	0.99	0.99	0.99
KNeighborsClassifier	0.97	0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
LogisticRegression	0.97	0.99	0.98	0.98	0.96	0.99	0.97	0.97	0.98	0.99	0.99	0.98
GaussianNB	0.83	0.96	0.89	0.85	0.97	0.89	0.93	0.91	0.98	0.96	0.97	0.97

FIGURE 5.20 – La comparaison entre avant et après l'application des méthodes de sélection

5.6 Comparaison entre nos résultats et les résultats d'un autre article [78] :

Nous avons comparé nos résultats finale avec les résultats de l'article :«DDoS datasets : Use of machine learning to analyse intrusion détection performance».[78]
D'après les premières observations, on a des bon résultats en comparaison avec le [78]
• le fonctionnement d'algorithme neive bayes de [78] n'est pas bien en comparaison avec notre résultat (accuracy 45%, precision et recall de 66% et 54% successivement.

Au contraire notre modèle a une très bon résultat avec un accuracy et F1-score de 97%, et un recall de 96% et précision de 98%.
• Les résultats de nos modèles (SVM et KNeighbors) sont bons par rapport à [78].
Les deux tableaux suivants montrent les différents résultats obtenu :

	Nos résultats				les résultats de l'article			
Algorithmes	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**	**Accuracy**
DecisionTreeClassifier	0.87	0.90	0.88	0.84	0.99	0.99	0.99	0.99
RandomForestClassifier	0.92	0.99	0.95	0.94	0.99	0.99	0.99	0.99
Svm	0.99	0.99	0.99	0.99	0.86	0.87	0.85	0.86
KNeighborsClassifier	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98
LogistiqueRegression	0.98	0.99	0.99	0.98	0.99	0.98	0.99	0.98
GaissianNB	0.98	0.96	0.97	0.97	0.66	0.54	0.38	0.45

FIGURE 5.21 – Comparaison entre nos résultats et les résultats d'un autre article [78]

5.7 Les outils de développement

dans cette partie nous allons présenter les principaux outils utilisés pour la mise en place de notre projet :

1. Python :
   Le langage Python est un langage de programmation open source multi-plateformes et orienté objet. Grâce à des bibliothèques spécialisées, Python s'utilise pour de nombreuses situations comme le développement logiciel, l'analyse de données, ou la gestion d'infrastructures. Il n'est donc pas, comme le langage HTML par exemple, uniquement dédié à la programmation web.[79]

2. sklearn :
   Nous avons utilisé Sklearn qui est une bibliothèque d'apprentissages statique en python ,contient toutes les fonctions de l'état de l'art du Machine Learning. On y trouve les algorithmes les plus importants ainsi que diverses fonctions de pré-processing.[80] Les fonctionnalités fournies par scikit-learn incluent :
   - Régression, (la régression linéaire et logistique).
   - Classification, (les voisins les plus proches).
   - Clustering, (K-Means et K-Means++).
   - Sélection du modèle.
   - Pré-traitement, (la normalisation Min-Max).[81]

elle set construite a base de :

-Pandas :
Pandas est une excellente bibliothèque pour importer vos tableaux Excel (etautres formats) dans Python dans le but de tirer des statistiques et de charger votre Dataset dans Sklearn.[80]

-Matplotlib :
Matplotlib est la bibliothèque qui permet de visualiser nos Datasets, nos fonctions, nos résultats sous forme de graphes, courbes et nuages de points.[80]

-Seaborn :
Seaborn est une bibliothèque permettant de créer des graphiques statistiques en Python. Elle est basée sur Matplotlib, et s'intègre avec les structures Pandas.[82]

3. google Colaboratory :
   Colaboratory, souvent raccourci en "Colab", est un produit de Google Research. Colab permet à n'importe qui d'écrire et d'exécuter le code Python de son choix par le biais du navigateur. C'est un environnement particulièrement adapté au machine learning, à l'analyse de données et à l'éducation.
   En termes plus techniques, Colab est un service hébergé de notebooks Jupyter qui ne nécessite aucune configuration et permet d'accéder gratuitement à des ressources informatiques, dont des GPU.[83]

4. Overleaf :
   est une plateforme en ligne gratuite permettant d'éditer du texte en LATEX sans aucun téléchargement d'application. En outre, elle offre la possibilité de rédiger des documents de manière collaborative, de proposer ses documents directement à différents éditeurs (IEEE Journal, Springer, etc.) ou plateformes d'archives ouvertes (**arXiv, engrxiv**, etc.) pour une éventuelle publication.
   Cette plateforme est très compatible avec différents supports tels que tablettes et smartphones.[84]

5.8 Conclusion

Dans ce chapitre nous avons présenté un résumé de notre travail avec les résultats que nous avons obtenue.

Conclusion Géneral

Au cours de la dernière décennie, le cloud a été la cible de la majorité des attaques DDoS car il est devenu un fournisseur important de services pour les organisations dans les pays développés ainsi que pays en développement en utilisant uniquement Internet et votre ordinateur. Par conséquent, le but général de ces attaques est de interrompre précisément la disponibilité des services pendant une longue période de temps. Cela cause de sérieux dommages financiers aux organisations. Alors, il est très important pour comprendre le problème de DoS et DDoS attaques dans l'environnement de cloud computing.

Dans ce projet nous avons présentés les différents concepts du cloud computing ainsi que les attaques DDoS dans son environnement où sont devenues l'un des problèmes de sécurité critiques qui menacent le cloud. Nous avons appris quelques concepts de Machine learning ainsi que nous avons implémenté une solution intelligente de détection des attaques ddos via les différentes modèles du machine learning en utilisant le langage Python et nous avons obtenus des bons résultats dans la plupart des algorithmes en comparaison avec d'autre étude.

Concernant les travaux futurs nous espérons réaliser une simulation d'une attaque réel dans le cloud computing.

Bibliographie

[1] Behal, S., Kumar, K. : Detection of DDoS attacks and flash events using novel information theory metrics. Comput. Netw. 116, 96–110 (2017)

[2] https ://www.futura-sciences.com/tech/definitions/informatique-cloud-computing-11573/

[3] P. Mell, T. Grance, The NIST Definition of Cloud Computing, Recommendation of NIST. Special Publication 800-145, 2011. http ://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[4] J-F Pépin, S. Bouteiller, A-S. Boissard, J. Watrinel, Fondamentaux du Cloud computing- le point de vue des grandes entreprise. Réseau de Grandes Entreprises (CIGREF). Mars 2013. http ://www.eurocloud.fr/doc/cigref2.pdf.(Derniére consultation avril 2016)

[5] P.P. Codo. Conception d'Une Solution de Cloud Computing Privé Basée sur un Algorithme de Supervision Distribué : Application aux Services IAAS. Ecole Polytechnique d'Abomey-Calavi (EPAC), 2012.

[6] H. Saouli. Découverte de services web via le Cloudcomputing à base d'agents mobiles. Thése de doctorat. Université Mohamed Khider de Biskra, 2015.

[7] A.A.Y. Elwessabi. Une approche basée agent mobile pour le cloud computing. Mémoire de magister. Université HADJ LAKHDAR-BATNA, 2014.

[8] NIST SP 800-146, "NIST Cloud Computing Synopsis and Recommendations", http ://csrc.nist.gov/publications/drafts/800-146/Draft-NIST-SP800-146.pdf

[9] Federal Cloud Computing Strategy,http ://www.cio.gov/documents/Federal-Cloud-ComputingStrategy.pdf

[10] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger and D, NIST SP 500-292,NIST Cloud Computing Reference Architecture, Leaf, 2011.

[11] DDoS Attack Types and Mitigation Methods, IncapsulaInc, 2014, http ://www.incapsula.com/ddos/ddos-attacks/. Accessed on Sept 2014

[12] Shameli-Sendi, A., Pourzandi, M., Fekih-Ahmed, M., Cheriet, M. : Taxonomy of Distributed Denial of Service mitigation approaches for cloud computing. J. Netw. Comput. Appl. 58,165–179 (2015)

[13] D. Yuanshun, X. Yanping, Z.Gewei. Self-healing and Hybrid Diagnosis in Cloud Computing. Proceedings CloudCom of 1st International Conference on CloudComputing. Beijing, China, pp. 45-56, 2009.

[14] http ://www.renaudvenet.com/cloud-computing-avantages-et-inconvenients-2011-01-26.html(Dernièere consultation mai 2016)

[15] https ://images.cigref.fr/Publication/2012-2013-Fondamentaux-Cloud-Computing-Point-de-vue-grandes-entreprises.pdf

[16] https ://blog.avast.com/fr/data-security-issues-in-cloud-computing

[17] Somani, G., Gaur, M.S., Sanghi, D., Conti, M., Buyya, R. : DDoS attacks in cloud computing : issues, taxonomy, and future directions. Comput. Commun. 107, 30–48 (2017)

[18] https ://www.altospam.com/glossaire/deni-de-service.php

[19] D.Hubert, W.Ping, O.S.Raed, and R.Andrew, "A Software-Defined Networking (SDN) Approach to Mitigating DDoS Attacks," Springer International Publishing AG, 141-145 (2018).

[20] https ://www.futura-sciences.com/tech/definitions/internet-deni-service-2433/

[21] G. Somani, M. S. Gaur, D. Sanghi, M.Conti, and R. Buyya, "DDoS Attacks in Cloud Computing : Issues, Taxonomy, and Future Directions," in Computer Communications 107 (2017) 30–48, Elsevier B.V (2017).

[22] CERT Advisory, CA-1996-01, Carnegie Mellon University,2014,http ://www.cert.org/historical/advisories/CA-1996-01.cfm. Accessed on Sept 2014

[23] https ://softwarelab.org/fr/attaque-ddos/

[24] https ://www.cloudflare.com/fr-fr/learning/ddos/what-is-a-ddos-attack/

[25] https ://doi.org/10.1080/07366981.2018.1453101

[26] https ://www.cloudprotector.com/fr/attaque-ddos

[27] Curt Wilson, Arbor Networks ASERT, « Attack of the Shuriken : Many Hands, Many Weapons ». <http ://www.arbornetworks.com/asert/2012/02/ddos-tools/>, fév. 2012.

[28] Brian Krebs, « DDoS Services Advertise Openly, Take PayPal ». <http ://krebsonsecurity.com/2013/05/ddos-services-advertise-openlytake-paypal/>, mai 2013.

[29] P. Mockapetris, « Domain names - concepts and facilities ». RFC 1034 (INTERNET STANDARD), nov. 1987. Updated by RFCs 1101, 1183, 1348, 1876, 1982,2065, 2181, 2308, 2535, 4033, 4034, 4035, 4343, 4035, 4592, 5936

[30] D. Mills, J. Martin, J. Burbank et W. Kasch, « Network Time Protocol Version 4 : Protocol and Algorithms Specification ». RFC 5905 (Proposed Standard), juin 2010.

[31] R. Presuhn, « Version 2 of the Protocol Operations for the Simple Network Management Protocol (SNMP) ». RFC 3416 (INTERNET STANDARD), déc. 2002.

[32] Contributing Members of the UPnP Forum, « UPnP™Device Architecture 1.1 ».<http ://www.upnp.org/specs/arch/UPnP-arch-DeviceArchitecturev1.1.pdf>, oct. 2008.

[33] J. Postel, « Character Generator Protocol ». RFC 864 (INTERNET STANDARD),mai 1983.

[34] NTP development team, « NTP Software Downloads ». <http ://www.ntp.org/downloads.html>.

[35] Wong Onn Chee, Tom Brennan, «H.....t.....t....p.......p....o....s....t ». <https ://www.owasp.org/images/4/43/Layer7DDOS.pdf>, nov. 2010.

[36] Bill Brenner, Akamai, « DDoS Attacks Used As Cover For Other Crimes ». <https ://blogs.akamai.com/2013/08/DDoS-Attacks-Used-As-CoverFor-Other-Crimes.html>, août 2013.

[37] FBI, FS-ISAC, IC3, « Fraud Alert - Cyber Criminals Targeting Financial Institution Employee Credentials to Conduct Wire Transfer Fraud ». <http ://www.ic3.gov/media/2012/ FraudAlertFinancialInstitutionEmployeeCredentialsTargeted.pdf>,sept. 2012.

[38] Prolexic Security Engineering et Response Team, « Prolexic Quarterly Global DDoS Attack Report Q4 2012 ». <http ://www.stateoftheinternet.com/ resources-web-security-2012-q4-global-ddos-attack-report.html>,jan. 2013.

[39] Prolexic Security Engineering et Response Team, « Prolexic Quarterly Global DDoS Attack Report Q4 2013 ». <http ://www.stateoftheinternet.com/ resources-web-security-2013-q4-global-ddos-attack-report.html>, jan. 2014.

[40] Prolexic Security Engineering et Research Team, « Q4 2014 State of the Internet– Security Report».<http ://www.stateoftheinternet.com/resources-websecurity-2014-q4-internet-security-report.html>, jan. 2015.

[41] https ://www.nystek.com/cache-sites/nystek-editions/IMG/pdf/risques-attaques-ddosv1r0.pdf.

[42] G. Somani, M. S. Gaur, D. Sanghi, M.Conti, and R. Buyya, "DDoS Attacks in Cloud Computing : Issues, Taxonomy, and Future Directions," in Computer Communications 107 (2017) 30–48, Elsevier B.V (2017).

[43] Somani, G., Gaur, M.S., Sanghi, D., Conti, M., Buyya, R. : DDoS attacks in cloud computing : issues, taxonomy, and future directions. Comput. Commun. 107, 30–48 (2017).

[44] Gupta, B.B., Badve, O.P. : Taxonomy of DoS and DDoS attacks and desirable defense mechanism in a Cloud computing environment. Neural Comput. Appl. 28(12), 3655–3682

[45] Tao, Y., Yu, S. : DDoS attack detection at local area networks using information theoretical metrics. In : 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, Melbourne, VIC, Australia, pp. 233–240. IEEE (2013).

[46] Detection Mechanisms of DDoS Attack in Cloud Computing Environment : A Survey.

[47] Bonguet, A., Bellaiche, M. : A survey of denial-of-service and distributed denial of service attacks and defenses in cloud computing. Future Internet 9(3), 43 (2017).

[48] O. Osanaiye, K. Raymond, and M. Dlodlo, "DDoS Resilience in Cloud :Review and Conceptual Cloud DDoS Mitigation Framework," Journal of Network and Computer Applications (2016).

[49] B. B. Gupta and O. P. Badve, "Taxonomy of DoS and DDoS attacks and desirable defense mechanism in a Cloud computing environment,"in The Natural Computing Applications Forum, Springer (2016).

[50] N. Agrawal and S. Tapaswi, "Defense schemes for variants of distributed denial-of-service (DDoS) attacks in cloud computing : A survey," Information Security Journal, 1-13 (2017).

[51] https ://www.1min30.com/developpement-web/5-conseils-pour-vous-proteger-des-attaques-ddos-19858.

[52] Microsoft Experiences, Tout savoir sur l'Intelligence Artificielle, (consulté le 09/02/2019), disponible

sur :https ://experiences.microsoft.fr/business/intelligenceartificielle-ia-business/comprendre-utiliser-intelligence-artificielle/

[53] https ://www.lebigdata.fr/machine-learning-et-big-data.

[54] A. Mercier, « L'information face à l'intelligence artificielle : promesses et dangers »https ://larevuedesmedias.ina.fr/linformation-face-lintelligence-artificielle-promesses-et-dangers, 05 février 2018.

[55] A. Géron, « Machine Learning avec Scikit-Learn », Dunod : Paris, 2017.

[56] M. Taffar, « Initiation à l'apprentissage automatique », Cours Master, Ment InformatiqueFaculté des Sciences Exactes et de l'Informatique.

[57] https ://analyticsinsights.io/apprentissage-supervise-vs-non-supervise/

[58] https ://mrmint.fr/naive-bayes-classifier

[59] H. B. Barlow. Unsupervised Learning. Neural Computation, vol. 1, no. 3, pages 295-311,1989.

[60] http ://cedric.cnam.fr/vertigo/cours/ml2/coursArbresDecision.html.

[61] https ://datascientest.com/knn.

[62] https ://mrmint.fr/introduction-k-nearest-neighbors.

[63] https ://runebook.dev/fr/docs/scikit-learn/modules/svm.

[64] https ://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificie lle/1501879-machine-a-vecteurs-de-support-svm-definition-et-cas-d-usage/

[65] https ://scikit-learn.org/stable/modules/svm.html

[66] https ://scikit-learn.org/stable/modules/svm.html https ://datascientest.com/random-forest-definition

[67] https ://hal.archives-ouvertes.fr/hal-02063544/document.

[68] https ://www.futura-sciences.com/sante/definitions/biologie-neurone-209/

[69] https ://datascientest.com/glossary/validation-croisee-cross-validation.

[70] https ://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/

[71] https ://www.lebigdata.fr/confusion-matrix-definition

[72] https ://statisticaloddsandends.wordpress.com/2020/01/23/what-is-balanced-accuracy.

[73] https ://www.lucidchart.com/pages/fr/arbre-de-decision

[74] https ://fr.slideshare.net/mariemchaaben/les-arbres-de-dcisions https ://www.memoireonline.com/04/12/5750/mIdentification-et-commande-des-systemes-non-lineaires21.html

[75] A. W. Omar, B. Jamal, O. Hadi, and M. Azzam, "Optimal Load Distribution for the Detection of VM-based DDoS Attacks in the Cloud," in IEEE transactions on services computing, 1-14 (2017)

[76] https ://mrmint.fr/data-preprocessing-feature-scaling-python

[77] https ://mrmint.fr/data-preprocessing-feature-scaling-python http ://thesis.univ-biskra.dz/944/7/Chap

[78] University of New Brunswick, "DDoS Evaluation Dataset (CICDDoS2019)," unb.ca, 2019. [Online].Available :https ://www.unb.ca/cic/datasets/ddos-2019.html.

[79] https ://www.futura-sciences.com/tech/definitions/informatique-python-19349/

[80] Apprendre le ML en une semaine.pdf

[81] https ://www.codecademy.com/articles/scikit-learn

[82] https ://datascientest.com/seaborn

[83] https ://research.google.com/colaboratory/faq.html

[84] https ://paris-sorbonne.libguides.com/c.php ?g=497641p=4637541 : :text=Overleaf

[85] https ://stats.stackexchange.com/questions/253086/selectkbest-feature-selection-python-scikit-learn : :text=SelectKBest