الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة سعيدةد. مولاي الطاهر

كلية التكنولوجيا

قسم: الإعلام الآلي


# Mémoire de Master

## Spécialité : Sécurité Informatique et Cryptographie

## Thème

# Application d'une méta-heuristique pour la détection d'intrusion: les algorithmes génétiques(AG)

**Présenté par :**

**OUDAYA Ahmed Serradj Eddine.**

**MOHAMMEDI Mohamed Nadir.**

**Dirigé par :**

DR LOKBANI Ahmed Chaouki.

Promotion 2021 - 2022

# REMERCIEMENT

Tout d’abord, nous tenons à remercier le Dieu de nous avoir donné la force et le courage pour pouvoir réaliser ce travail.

Nous adressons le grand remerciement à notre encadreur Dr LOKBANI Ahmed Chaouki et Dr HAMOU Reda pour leurs précieux conseils et remarques qui nous ont permet d’accomplir ce travail.

A nos collègues qui nous ont poussé dans les moments difficiles.

On remercie tous les enseignants du département informatique qui ont participé à cette formation de cinq ans.

Finalement, nous tenons à exprimer notre profonde gratitude à nos familles qui nous ont toujours soutenues pour réaliser ce travail.

# DÉDICACE

Nous dédions ce travail à nos chers parent qui n'ont jamais cessé à nous encourager grâce à leur amour, leurs sacrifices et prières.

Et à nos chers frères et sœurs pour leurs encouragement et leur soutien moral.

A nos amis et collègues pour leur compagnie et les bons moments passés ensemble.

Merci d'être là pour nous.

Que Dieu vous garde.

# Table des matières

**CHAPITRE2 : Les systèmes de détection d'intrusion.**

**CHAPITRE3 : Les algorithmes génétiques**



**CHAPITRE 4 : Implémentation et discussion des résultats.**

# Table des figures

## Liste des Tableaux :

# INTRODUCTION GÉNÉRALE

# Introduction générale :

Dans l'époque actuelle, ou le monde informatique est en progression très rapide, toutes les communications et les échanges commerciales, financières, militaires se passent par internet, donc des informations très sensibles circulent et risquent d'être accessibles aux gens non-autorisés vu que tous type de personnes utilisent l'internet et parmi eux se cachent les voleurs.

Malheureusement, les moyens de sécurité comme le pare-feu et les antivirus ne sont pas efficaces tout le temps, surtout quand il s'agit de nouvelles attaques qui sont pas connues par le système se sécurité.

C'est pour cela, les systèmes de détection d'intrusions(IDS) sont des nouvelles solutions qui cherchent a renforcer la sécurité informatique  et protéger les données que ce soit à l'intérieur ou à l'extérieur du système.

## Problématique

Les anciens moyens de sécurité informatique ont prouvé leur inefficacité contre les attaques non-connues, les systèmes de détection d'intrusion offre une solution pour résoudre ce problème, dans notre travail, nous cherchons à améliorer les performances des IDS en basant sur les algorithmes génétiques.

## Organisation du mémoire

Dans notre projet, nous essayons de réaliser un système de détection d'intrusion basé sur une méthode méta-heuristique ( les algorithmes génétiques) et renforcé par les algorithmes du data mining pour avoir plus de résultats valides et moins des fausses alertes.

Dans le premier chapitre de de ce travail, nous allons parler de la sécurité informatique, les différents objectifs, les attaques les plus fréquentes dans le temps actuel et quelques moyens de sécurité.

Puis dans le deuxième chapitre, nous allons parler d'un de ces moyens qui est le système de détection d'intrusion, ses types, méthodologies et emplacement.

Nous allons utilisé une méthode meta-heuristique, donc nous parlerons sur les algorithmes génétiques et la technique du datamining dans le chapitre trois.

Pour finaliser, nous allons expérimenter les résultats obtenus et conclure en les discutant dans le dernier chapitre.

# CHAPITRE 1

# LA SÉCURITÉ INFORMATIQUE

## 1-Introduction :

De nos jours, avec l'ouverture des systèmes d'information à l'extérieur, la progression des moyens de communication et le développement de technologies de transmission de données, l'internet est devenu le moyen utilisé pour l'échange des informations personnelles, financières et d'autres qui doivent être secrètes . Donc il est incontournable d'établir une politique de sécurité pour assurer le bon transfert de ces données et protéger le flux de transmission de l'accès non-autorisé.

## 2-La sécurité informatique:

# 2-1-Définition:

La sécurité informatique est l'ensemble des moyens physiques, logiques et administratifs mis en place dans un système dans le but de minimiser les vulnérabilités et protéger ses données des menaces intentionnelles ou accidentelles, afin de réaliser les principaux objectifs : la confidentialité, l'intégrité et la disponibilité des données.

# 2-2-Objectifs de la sécurité :

La sécurité informatique vise à réaliser les éléments de la triade (CIA) qui désigne les principaux objectifs de la sécurité des systèmes d'information :

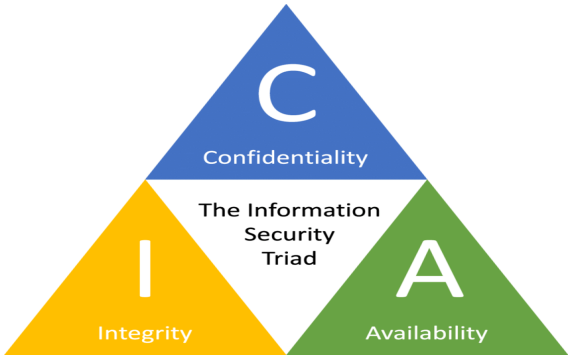



**FIGURE 1-1 : La triade CIA**

- **La confidentialité :** assure que l'information soit protégée contre toute divulgation accidentelle ou malveillante aux parties non autorisées.
- **L'intégrité** : assure que l'information et les systèmes soient protégés contre toute modification ou destruction accidentelle ou malveillante.
- **La disponibilité** : assure que l'information et les systèmes soient accessibles et utilisables par les parties autorisées aux moments où elles en ont besoin.

A coté de ces caractéristiques de bases nous rencontrons également les composantes suivantes :

- **Authentification :** assure l'identification d'un individu, d'une entité mais également l'origine de l'information ou encore d'une opération effectuée sur celle-ci.
- **Non-répudiation :** assure-le fait qu'une personne ou entité ne puisse nier avoir effectué une activité. Le destinataire ne pourra nier avoir reçu l'information, et l'expéditeur de la source de l'information ne peut nier avoir envoyé l'information. [1]

### 2-2-1 La politique de sécurité informatique:

Une politique de sécurité informatique est une stratégie visant à maximiser la sécurité informatique d'une entreprise. Elle est matérialisée dans un document qui reprend l'ensemble des enjeux, objectifs, analyses, actions et procédures faisant parti de cette stratégie. [2]

### 2-2-2 Faille de sécurité :

Une faille de sécurité est une faiblesse dans un produit, qui pourrait permettre à un attaquant de compromettre l'intégrité, la disponibilité ou la confidentialité de celui-ci. [3]

### 2-2-3 Les programmes malveillants :

Un logiciel malveillant (ou malware) est un type de programme conçu pour endommager ou nuire à un ordinateur, serveur, client, réseau informatique, une / infrastructure sans que les utilisateurs finaux ne s'en aperçoivent. [4]

### 2-2-4 Les risques et menaces de la messagerie électronique:

La sécurité des e-mails, et la protection de la messagerie électronique revêtent une importance capitale, vu la place qu'occupe l'e-mail dans notre communication quotidienne et la réalisation de notre travail. Exploitation des données personnelles, pertes de documents confidentiels et sensibles, cyber-harcèlement… sont tous des risques auxquels vous ou votre entreprise êtes exposés à chaque fois que vous cliquez sur un courriel entrant. Ils surviennent à partir de virus, de spam, de malwares et d'autres cybers attaques à travers votre boite e-mail. [5]

### 2-2-5 Les risques et menaces sur le réseau

- Divulgation des données.
- Atteinte à l'intégrité des données.
- Destruction des données.
- Perte de service.
- Usurpation d'identité. [6]

# 2-3) La classification des attaques informatiques :

Une attaque informatique est une action malveillante qui vise à exploiter une faille d'un système dans le but de causer des dommages, elles peuvent être classées selon les critères suivants :

## 2-3-1) Selon l'objectif :

On classe les attaques en deux catégories :

- **Les attaques passives :** sont les attaques par lesquelles l'attaquant se livre à une écoute illicite, ne surveillant que la transmission ou la collecte d'informations. L'espion n'apporte aucune modification aux données ni au système.
- **Les attaques actives :** sont les attaques dans lesquelles l'attaquant tente de modifier les informations ou crée un faux message, Elle nécessite généralement plus d'effort et une implication souvent plus dangereuse. Lorsque le pirate informatique tente d'attaquer, la victime en prend conscience. [7]

## 2-3-2) Selon la source de l'attaque:

Cela veut dire selon le point d'initiation de l'attaque :

- **Les attaques internes :** venant des employés et des utilisateurs légitimes du système qui dépassent la zone d'utilisation autorisée.
- **Les attaques externes :** venant de l'extérieur du système (internet).

## 2-3-3) Selon la technique utilisée:

Les hackers utilisent plusieurs techniques d'attaques. Ces attaques peuvent être regroupées en trois familles différentes : Les attaques directes, les attaques indirectes par rebond et les attaques indirectes par réponses.

- a) les attaques directes : C'est la plus simple des attaques. Le hacker attaque directement sa victime à partir de son ordinateur.

- b) les attaques indirectes par rebond : Cette attaque est très prisée des hackers. En effet, le rebond à deux avantages :
  - Masquer l'identité du hacker.
  - Eventuellement, utiliser les ressources de l'ordinateur intermédiaire car il est plus puissant pour attaquer.

- c) les attaques indirectes par réponses : Cette attaque est un dérivé de l'attaque par rebond. Elle offre les mêmes avantages, du point de vue du hacker. Mais au lieu d'envoyer une attaque à l'ordinateur intermédiaire pour qu'il la répercute, l'attaquant va lui envoyer une requête. [8]

# 2-4) Buts des attaques informatiques:

- **Interruption :** touche la disponibilité des données.
- **Interception :** touche la confidentialité des données.
- **Modification :** touche l'intégrité des données.
- **Fabrication :** touche l'authenticité des données.
-

# 2-5) Motivations des attaques :

Les motivations des attaques peuvent être de différentes sortes :

- Obtenir un accès au système.
- Voler des informations, tels que des secrets industriels ou des propriétés intellectuelles.
- Glaner des informations personnelles sur un utilisateur.
- Récupérer des données bancaires.
- S'informer sur l'organisation (entreprise de l'utilisateur, etc.).
- Troubler le bon fonctionnement d'un service.
- Utiliser le système de l'utilisateur comme « rebond » pour une autre attaque.
- Utiliser les ressources du système de l'utilisateur, notamment lorsque le réseau sur lequel il est situé possède une bande passante élevée. [9]

## 2-6) Les attaques informatiques les plus fréquentes :

**a) <u>Le déni de service :</u>**

Les attaques de ce type ont pour but de saturer un routeur ou un serveur afin de le "crasher" ou en préambule d'une attaque massive. Ces types d'attaque sont très faciles à mettre en place et très difficile à empêcher. [11]

**b) <u>Le phishing :</u>**

C'est est une tentative de récupérer des informations sensibles (généralement des informations financières comme les détails de cartes de crédit, mot de passe...), en envoyant des e-mails non sollicités avec des URL truqués. [11]

**c) <u>Le pharming :</u>**

C'est une autre attaque réseau visant à rediriger le trafic d'un site web vers un autre sit web. [11]

La redirection de la demande de l'utilisateur se fait grâce à une manipulation du protocole DNS, qui est chargé de convertir le nom textuel de l'hôte (adresse URL) en

une adresse IP numérique. Ce processus de conversion offre aux criminels deux angles d'attaque pour détourner le processus.

**d) <u>Attaque de l'homme au milieu (MitM)</u> :**

Il s'agit d'une technique de piratage consistant à intercepter des échanges cryptés entre deux personnes ou deux ordinateurs pour en décoder le contenu. Le hacker doit donc réceptionner les messages des deux parties et répondre à chacune se faisant passer pour l'autre. [12]

# 2-7) Comment protéger notre système informatique :

## 2-7-1) Définition d'antivirus :

Un antivirus est un programme qui a pour finalité de protéger la machine ou l'appareil sur lequel il est installé contre les logiciels malveillants. La détection d'un logiciel malveillant peut reposer sur trois méthodes [13]:

- La reconnaissance d'un code déjà connu (appelé signature) et mémorisé dans une base de données.
- Analyse du comportement d'un logiciel (méthode heuristique).
- La reconnaissance d'un code typique d'un virus.

## 2-7-2) Définition d'un antispyware :

L'Anti-spyware est un type de logiciel conçu pour détecter et supprimer les programmes espions indésirables. Les logiciels espions sont des types de logiciels malveillants installés sur un ordinateur à l'insu de l'utilisateur afin de collecter des informations à leur sujet. Cela peut poser un risque pour la sécurité de l'utilisateur, mais le plus souvent, les logiciels espions dégradent les performances du système en absorbant la puissance de traitement, en installant des logiciels supplémentaires ou en redirigeant l'activité du navigateur des utilisateurs.[14]

## 2-7-3) L'anti-spam :

Un logiciel anti-spam vise à empêcher les e-mails indésirables et malveillants d'atteindre les messageries professionnelles. L'objectif principal d'un système anti spam est de filtrer les emails avant même qu'ils arrivent dans votre boite email pour éviter les emails commerciaux/spam. [15]

### 2-7-4) Le pare-feu :

Il s'agit d'un système, logiciel ou matériel placé entre deux réseaux ou plus, dont le rôle est de filtrer le trafic réseau se présentant a ses interfaces.il tente d'isoler ces réseaux de façon à les protéger les uns des autres. [16]

### 2-7-5) Système de détection d'intrusion :

Un système de détection d'intrusions (« Intrusion Détection Systems » ou IDS) est un appareil ou une application qui alerte l'administrateur en cas de faille de sécurité, de violation de règles ou d'autres problèmes susceptibles de compromettre son réseau informatique. [17]

## 3) Conclusion :

Dans ce chapitre, nous avons abordé quelques notions de base de la sécurité informatique, on parlé de les risques, les attaques informatiques et leurs classifications, leurs buts et motivations, puis les attaques les plus courantes dans le temps actuel.

Enfin, nous avons défini les moyens de sécurité contre ces attaques pour protéger notre système vu que notre projet vise à concevoir un système de détection pour réduire les menaces, augmenter le taux de sécurité et préserver nos information.

Dans le chapitre suivant, on va détaillé un des moyens de sécurité qui est le système de détection d'intrusion avec ces différents types.

# CHAPITRE2

# LES SYSTÈMES DE DÉTECTION D'INTRUSION.

# 1) Introduction :

La sécurité des systèmes informatiques devient un objectif principal à réaliser, pour cela de nombreux moyens sont proposés comme des solutions pour augmenter le taux de sécurité des données informatiques.

Les moyens traditionnels comme le pare-feu et les antivirus étaient inefficaces contre les attaques non connues. L'apparition des **Systèmes de détection d'intrusion** est une solution probable pour essayer de résoudre ce problème.

Dans ce chapitre, nous allons parlé des systèmes de détection d'intrusion, leurs types, méthodologie, emplacement, fonctionnement, avantages et inconvènients.

# 2) Système de détection d'intrusion :

## 1) Historique :

Le premier modèle de détection d'intrusion est développé en 1984 par Dorothy Denning et Peter Neuman, qui s'appuie sur des règles d'approche comportementale. Ce système fut appelé IDES (Intrusion Detection Expert System). Très rapidement IDES a évolué pour donner naissance en 1993 à NIDES (Next generation Intrusion Detection Expert System). En 1988, le projet Haystack de Crosby Marks (à la demande de l'armée de l'air américaine) aboutie à un IDS qui fonctionne selon des signatures et non plus des règles de modélisation. Ce projet évoluera par la suite vers DIDS (Distributed Intrusion Detection System) qui analysera serveurs et machines clientes. L'approche comportementale allait commencer à laisser la place à une approche dite par scénario. En 1994, une autre société ISS (Internet Security System) est fondée par Thomas Noonan et Christopher Klauss, ISS (Internet Security System). En 1997, ISS lance RealSecure, disponible pour NT4 (premier IDS destiné à l'environnement Windows).Ensuite une multitude de fusions et d'acquisitions se sont produites.

## 2) Architecture d'un IDS :

Les IDS ont trois composants nécessaires, qui sont :

- Le capteur : Son rôle est de récolter les informations qui circulent autour de lui dans un message appelé évènement.
- L'analyseur : il analyse les évènements donnés par le capteur et envoie une alerte en cas de détection d'une activité malveillante.

- Le manager : il collecte les informations d'analyseur et il fait l'isolement de l'attaque ou la suppression définitive.

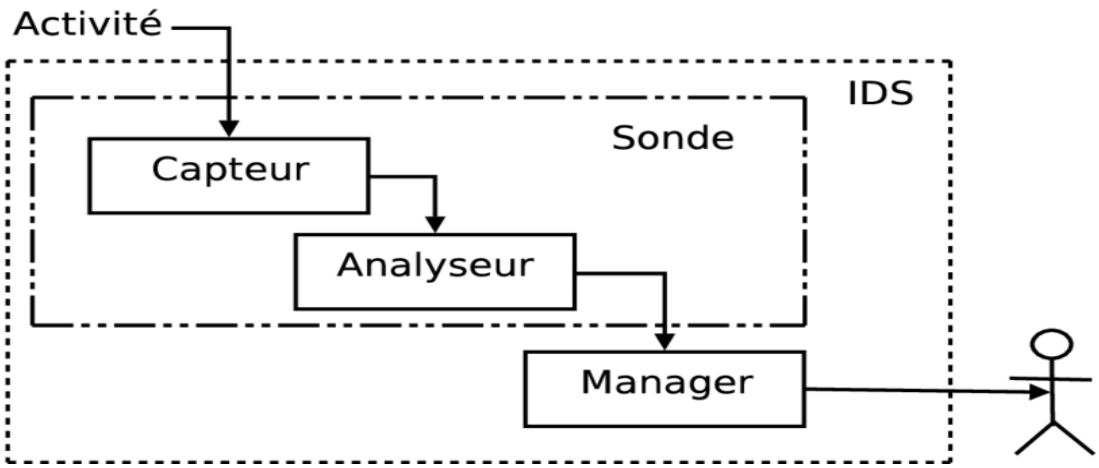



**FIGURE 2-1: Les composants d'un IDS**

## 3) Fonctionnement de l'IDS :

Il existe 3 modes de fonctionnement globaux d'un système de détection d'intrusion (IDS) :

### 3-1) Analyse basée sur l'anomalie:

Les IDS basés sur les anomalies fonctionnent généralement en prenant une base de référence du trafic et de l'activité normale qui se déroule sur le réseau. Ils peuvent mesurer l'état actuel du trafic sur le réseau par rapport à cette ligne de base afin de détecter des modèles qui ne sont pas présents dans le trafic normal. Ces méthodes peuvent fonctionner très bien lorsque nous cherchons à détecter de nouvelles attaques ou des attaques qui ont été délibérément assemblées pour éviter les IDS [18]. Ils ont été principalement introduits pour détecter les attaques inconnues, en partie en raison du développement rapide des logiciels malveillants. L'approche de base consiste à utiliser l'apprentissage automatique pour créer un modèle d'activité digne de confiance, puis à comparer le nouveau comportement à ce modèle [19].

### 3-2) Analyse basée sur les signatures :

Les IDS basés sur les signatures fonctionnent de manière très similaire à la plupart des systèmes antivirus. Ils maintiennent une base de données des signatures qui pourraient signaler un type particulier d'attaque et comparent le trafic entrant à ces signatures. En général, cette méthode fonctionne bien, sauf lorsque nous rencontrons une attaque nouvelle ou spécifiquement conçue pour ne pas correspondre aux signatures d'attaques existantes. L'un des principaux inconvénients de cette méthode est que de nombreux systèmes basés sur les signatures s'appuient uniquement sur leur base de données de

signatures pour détecter les attaques. Si nous n'avons pas de signature pour l'attaque, nous ne pouvons pas la voir du tout. En outre, l'attaquant qui manipule le trafic peut avoir accès aux mêmes outils IDS que nous utilisons et peut être en mesure de tester l'attaque contre eux afin de contourner spécifiquement nos mesures de sécurité. [20]

## 3-3) Systèmes adaptatifs :

Les systèmes adaptatifs commencent par des règles généralisées pour l'environnement, puis apprennent ou s'adaptent à des conditions locales qui seraient autrement inhabituelles. Après la période d'apprentissage initiale, le système comprend comment les gens interagissent avec l'environnement, puis avertit les opérateurs des activités inhabituelles. [21]
Voici une image qui résume le fonctionnement des IDS :

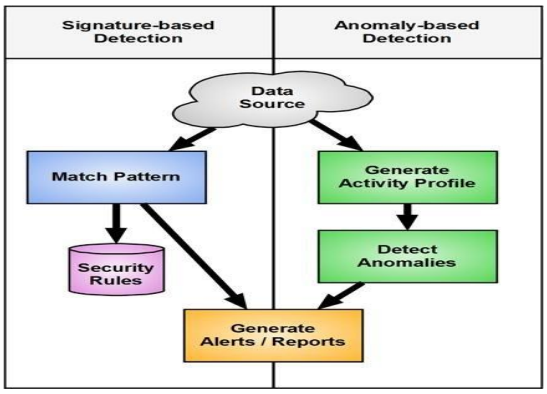



**FIGURE 2-2 : Le fonctionnement d'un IDS.**

## 4)Les types des IDS :

Il existe plusieurs types d'IDS :

## 4-1) Les IDS réseaux(NIDS) :

Ils sont connus aussi sous le nom de Network-based IDS. Les IDS réseaux analysent et interprètent les paquets circulant sur un réseau (ou un segment du réseau) afin de repérer les paquets à contenus malicieux. Le paquet est analysé sur toutes ses couches (réseau, transport, application). Par dissection des paquets et la connaissance des protocoles, les

NIDS sont capables de détecter des paquets malveillants conçus pour outrepasser un pare-feu. [22]

Ils effectuent une analyse du trafic passant sur l'ensemble du sous - réseau et fait correspondre le trafic transmis sur les sous-réseaux à la bibliothèque des attaques connues. Une fois qu'une attaque est identifiée ou qu'un comportement anormal est détecté, l'alerte peut être envoyée à l'administrateur. Un exemple de NIDS serait de l'installer sur le sous-réseau où se trouvent les pare-feu afin de voir si quelqu'un essaie de s'introduire dans le

pare-feu. Idéalement, il faudrait analyser tout le trafic entrant et sortant, mais cela pourrait créer un goulot d'étranglement qui réduirait la vitesse globale du réseau. OPNET et NetSim sont des outils couramment utilisés pour simuler des systèmes de détection d'intrusion réseau. Les systèmes NID sont également capables de comparer les signatures de paquets similaires pour lier et supprimer les paquets détectés nuisibles qui ont une signature correspondant aux enregistrements dans le NIDS. Lorsque nous classons la conception du NIDS en fonction de la propriété d'interactivité du système, il existe deux types : les NIDS en ligne et hors ligne, souvent appelés mode en ligne et mode tap, respectivement. Le NIDS en ligne traite le réseau en temps réel. Il analyse les paquets Ethernet et applique quelques règles, pour décider s'il s'agit d'une attaque ou non. Le NIDS hors ligne traite les données stockées et les transmet à certains processus pour décider s'il s'agit d'une attaque ou non.

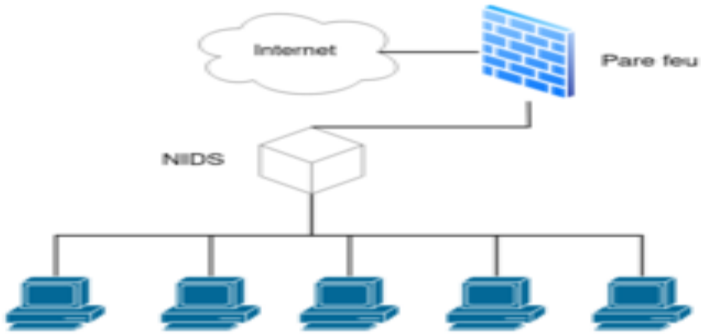



**FIGURE 2-3 : Emplacement d'un NIDS**

## 4-1-1) Les avantages d'un NIDS :

-Le NIDS peut être placé dans un réseau qui contient un grand nombre de machines.
-Le NIDS peut être invisible pour les attaquants.
-Placé sans interférer l'opération normale d'un réseau.

## 4-1-2) Les inconvénients d'un NIDS :

-Probabilité élevée de faux négatifs (une vraie tentative non détectée).
- Traitement difficile dans le cas de haut trafic du réseau.
-Il ne peut pas analyser les informations chiffrées.

## 4-1-3) Exemple de NIDS :

**Snort** : Snort est un système de détection d'intrusion (IDS) et un système de prévention d'intrusion (IPS) open-source qui fournit une analyse du trafic réseau en

temps réel et un enregistrement des paquets de données. Il utilise un langage basé sur des règles qui combine des méthodes d'inspection des anomalies, des protocoles et des signatures pour détecter les activités potentiellement malveillantes. L'utilisation du Snort permet de détecter les attaques de type DOS.[23]

**BRO** : Ce logiciel est programmé de façon totalement différente de Snort, il s'appuie sur les mêmes bases théoriques (filtrage par motif, formatage aux normes RFC, etc.), mais il intègre un atout majeur : l'analyse de flux réseau. Cette analyse permet de concevoir une cartographie du réseau et d'en générer un modèle. Ce modèle est comparé en temps réel au flux de données et toute déviance lève une alerte.[24]

## 4-2) Les IDS hôtes(HIDS) :

Les systèmes de détection d'intrusion basés sur l'hôte, aussi appelés Host-based IDS, analysent exclusivement les activités concernant l'hôte sur lequel ils sont installés (serveur, poste client, pare-feu, etc.), recherchant des activités suspectes. La détection peut se faire en utilisant les logs d'audit de sécurité, les logs système, le trafic réseau de l'hôte, les processus en cours d'exécutions, les accès aux fichiers, les changements de configurations des applications, etc. Le plus souvent les HIDS sont déployés sur les hôtes critiques comme les serveurs contenant des informations de sensibilités élevées et les serveurs publiquement accessibles.

### 4-2-1) Exemples de HIDS :

**AIDE :** AIDE (Advanced Intrusion Detection Environment) est un vérificateur d'intégrité des fichiers et des répertoires. Il crée une base de données à partir des règles d'expression régulière qu'il trouve dans le(s) fichier(s) de configuration. Une fois cette base de données initialisée, elle peut être utilisée pour vérifier l'intégrité des fichiers**.** [25]

**DarkSpy :** DarkSpy est un système de détection d'intrusion à usage individuel. [26]

### 4-2-2) Les avantages d'un HIDS :

-Possibilité d'analyser les informations chiffrées.
-L'analyse des événements locaux non-détectés par le NIDS.
-Possibilité de détecter des attaques qui touchent l'intégrité du système.

### 4-2-3) Les inconvénients d'un HIDS :

-Configuration pour chaque machine surveillée.
-Le HIDS peut analyser que les paquets de ses machines.
-Il peut être dépassé par des attaques de déni de service.

### 4-3) Les IDS hybrides :

La nouvelle tendance en matière de détection d'intrusion est de combiner les NIDS et les HIDS pour concevoir des IDS hybrides. Les systèmes hybrides de détection d'intrusion sont flexibles et augmentent le niveau de sécurité. Ils combinent plusieurs localisations des systèmes IDS et recherchent si bien les attaques visant des éléments particuliers que celles visant l'ensemble du système. Un exemple d'IDS hybride est l'ISS RealSecure. [21]

### 4-4) Wireless IDS :

il surveille le trafic du réseau sans fil et analyse ses protocoles afin d'identifier toute activité suspecte impliquant les protocoles eux-mêmes. Il est le plus souvent déployé à porter du réseau sans fil d'une organisation pour le surveiller, mais il peut aussi être déployé dans des endroits où des réseaux sans fil non autorisés se produire. [27]

### 4-5) APHIDS (Agent-Based Programmable Hybrid Intrusion Detection System) :

Ce type de système de détection d'intrusion se base sur des agents autonomes réactifs, capables de communiquer avec d'autres systèmes, ou de se déplacer d'hôte en hôte (on parle alors d'agents mobiles), permettant ainsi de réduire l'impact réseau du système de détection d'intrusion pour sa collecte de données. [28]

## 5) Emplacement d'un IDS :

Il existe plusieurs endroits stratégiques où il convient de placer un IDS.

Le schéma suivant illustre un réseau local ainsi que les trois positions que peut y prendre un IDS :

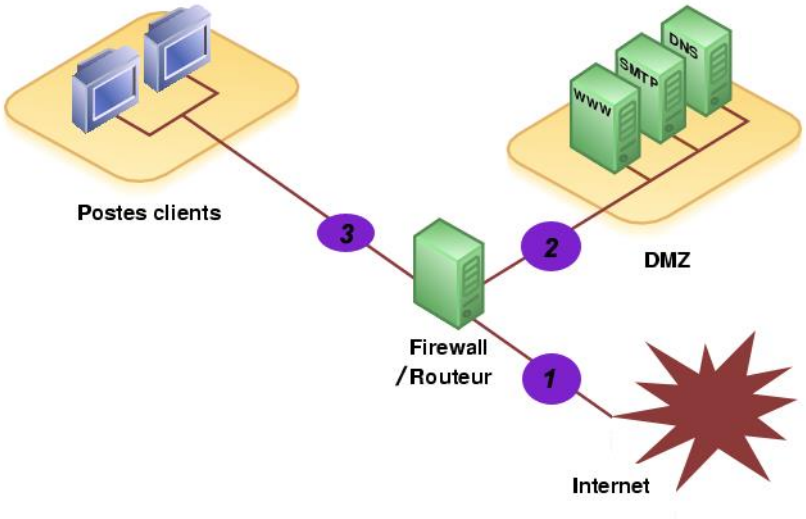



**FIGURE 2-4 : L'emplacement d'un IDS**

- **Position (1):** Sur cette position, l'IDS va pouvoir détecter l'ensemble des attaques frontales, provenant de l'extérieur, en amont du firewall. Ainsi, beaucoup (trop) d'alertes seront remontées ce qui rendra les logs difficilement consultables.

  Si un IDS est placé au-delà du pare-feu d'un réseau, son objectif principal serait de se défendre contre le bruit provenant d'Internet mais, plus important encore, de se défendre contre les attaques courantes, telles que les analyses de ports et le mappeur de réseau. Un IDS dans cette position surveillerait les couches 4 à 7 du modèle OSI et serait basé sur la signature. C'est une pratique très utile, car plutôt que d'afficher les violations réelles dans le réseau qui ont réussi à traverser le pare-feu, les tentatives de violation seront affichées, ce qui réduit le nombre de faux positifs. L'IDS dans cette position aide également à réduire le temps nécessaire pour découvrir des attaques réussies contre un réseau.[29]

- **Position (2):** Si l'IDS est placé sur la DMZ, il détectera les attaques qui n'ont pas été filtrées par le firewall et qui relèvent d'un certain niveau de compétence. Les logs seront ici plus clairs à consulter puisque les attaques bénignes ne seront pas recensées.

- **Position (3):** L'IDS peut ici rendre compte des attaques internes, provenant du réseau local de l'entreprise. Il peut être judicieux d'en placer un à cet endroit étant donné le fait que 80% des attaques proviennent de l'intérieur. De plus, si des Trojans ont contaminé le parc informatique (navigation peu méfiante sur internet) ils pourront êtres ici facilement identifiés pour être ensuite éradiqués. [EMPIDS], Ignorer la sécurité au sein d'un réseau peut causer de nombreux problèmes, cela permettra soit aux utilisateurs d'engendrer des risques de sécurité, soit à un attaquant qui s'est déjà introduit dans le réseau de se déplacer librement. La sécurité intense de l'intranet rend difficile, même pour les pirates au sein du réseau, de manœuvrer et d'élever leurs privilèges. [30]

## 6) Les critères de choix :

Les systèmes de détection d'intrusion sont devenus indispensables lors de la mise en place d'une infrastructure de sécurité opérationnelle. Ils s'intègrent donc toujours dans un contexte et dans une architecture imposants des contraintes très diverses. Certains critères (toujours en accord avec le contexte de l'étude) peuvent être dégagés:

**Fiabilité :** Les alertes générées doivent être justifiées et aucune intrusion ne doit pouvoir lui échapper.

**Réactivité :** Un IDS doit être capable de détecter les nouveaux types d'attaques le plus rapidement possible ; pour cela il doit rester constamment à jour. Des capacités de mise à jour automatique sont pour ainsi dire indispensables.

**Facilité de mise en œuvre et adaptabilité :** Un IDS doit être facile à mettre en œuvre et surtout s'adapter au contexte dans lequel il doit opérer ; il est inutile d'avoir un IDS émettant des alertes en moins de 10 secondes si les ressources nécessaires à une réaction ne sont pas disponibles pour agir dans les mêmes contraintes de temps.

**Performance :** la mise en place d'un IDS ne doit en aucun cas affecter les performances des systèmes surveillés. De plus, il faut toujours avoir la certitude que l'IDS a la capacité de traiter toute l'information à sa disposition (par exemple un IDS réseau doit être capable de traiter l'ensemble du flux pouvant se présenter à un instant donné sans jamais supprimer de paquets) car dans le cas contraire il devient trivial de masquer les attaques en augmentant la quantité d'information. [31]

## 7) Comparaison d'un IDS avec un pare-feu :

Bien qu'ils soient tous deux liés à la sécurité du réseau, un IDS diffère d'un pare-feu en ce qu'un pare-feu de réseau traditionnel (distinct d'un pare-feu de nouvelle génération ) utilise un ensemble de règles statiques pour autoriser ou refuser les connexions réseau. Il empêche implicitement les intrusions, en supposant qu'un ensemble de règles approprié a été défini. Essentiellement, les pare-feu limitent l'accès entre les réseaux pour empêcher les intrusions et ne signalent pas une attaque de l'intérieur du réseau. Un IDS décrit une intrusion suspectée une fois qu'elle a eu lieu et signale une alarme. Un IDS surveille également les attaques qui proviennent de l'intérieur d'un système. Ceci est traditionnellement réalisé en examinant les communications réseau, en identifiant les heuristiques et les modèles (souvent appelés signatures) des attaques informatiques courantes et en prenant des mesures pour alerter les opérateurs. Un système qui met fin aux connexions s'appelle un système de prévention d'intrusion et effectue un contrôle d'accès comme un pare-feu de couche d'application . [32]

## 8) L'avantage d'un IDS :

Naturellement, le principal avantage d'un IDS est d'identifier les menaces pour la sécurité de vos réseaux. Ils constituent un système d'alerte précoce, conçu pour éviter que les attaques malveillantes ne se propagent au sein du réseau et ne causent davantage de dommages. Si vous optez pour un système actif, il peut également contribuer à neutraliser le fil jusqu'à ce que vos administrateurs puissent régler le problème.

En plus d'identifier (et potentiellement de neutraliser) les menaces de sécurité, les systèmes de détection d'intrusion à Phoenix enregistrent également les attaques. Les enregistrements détaillés des attaques malveillantes aident les administrateurs à identifier les faiblesses, à résoudre les problèmes et à surveiller les attaques futures.

Les journaux détaillés sont également utiles si vous devez prouver que votre réseau est conforme aux réglementations industrielles. Vous pouvez utiliser les journaux pour montrer comment vous traitez les problèmes de sécurité et prouver que votre réseau a été correctement sécurisé. Ils permettent également d'observer plus facilement l'activité sur l'ensemble du réseau.

Enfin, les IDS facilitent l'amélioration de vos alertes de sécurité et de votre réponse, sur la base des données qui circulent sur le réseau, des dispositifs ciblés et de la façon dont la réponse de sécurité précédente a traité la menace. [33]

## 9) L'inconvénient d'un IDS :

-L'inconvénient du système de détection d'intrusion est qu'il ne peut pas détecter la source de l'attaque et en cas d'attaque, il verrouille simplement l'ensemble du réseau.

- Le bruit et la détection des mauvais paquets peuvent augmenter le nombre de fausses alertes.

-Un problème de fiabilité, l'intrus peut modifier les programmes du système et les règles de détection.

## 3) Conclusion :

Les systèmes de détection d'intrusion sont des outils très utiles pour renforcer la sécurité de notre système, Ils utilisent deux méthodes générales : la détection par anomalie et l'autre par signatures. Notre travail consiste à utiliser une méthode méta-heuristique à la base des algorithmes génétiques pour réaliser cet objectif. Ce qu'on va détailler dans le prochain chapitre.

# CHAPITRE3

# LES ALGORITHMES GÉNÉTIQUES.

## 1) Introduction :

La classification de données est une technique très utilisé dans le domaine informatique tel que le machine learning, reconnaissance des formes et des images pour faciliter la prise de décision.

Dans la détection d'intrusion, la classification est utilisée pour classifier les connexion du trafic en deux classes générales : une connexion normale et une classe anormale.

Dans ce chapitre, nous parlons des algorithmes méta-heuristiques en générale, puis on détaille l'un de ses algorithmes qui est l'algorithmes génétique et ses différentes étapes.

## 2) Les algorithmes méta-heuristiques :

Une métaheuristique est un algorithme d'optimisation visant à résoudre des problèmes d'optimisation difficiles (souvent issus des domaines de la recherche opérationnelle, de l'ingénierie ou de l'intelligence artificielle) pour lesquels on ne connaît pas de méthode classique plus efficace.

Les métaheuristiques sont généralement des algorithmes stochastiques itératifs, qui progressent vers un optimum global, c'est-à-dire l'extremum global d'une fonction, par échantillonnage d'une fonction objectif. Elles se comportent comme des algorithmes de recherche, tentant d'apprendre les caractéristiques d'un problème afin d'en trouver une approximation de la meilleure solution (d'une manière proche des algorithmes d'approximation).

Il existe un grand nombre de métaheuristiques différentes, allant de la simple recherche locale à des algorithmes complexes de recherche globale. Ces méthodes utilisent cependant un haut niveau d'abstraction, leur permettant d'être adaptées à une large gamme de problèmes différents. [34]

Les métaheuristiqes sont inspirées de la nature tels que les algorithmes de colonies de fourmis, les algorithmes génétiques…

## 3) Présentation des algorithmes génétiques :

Les algorithmes génétiques sont des algorithmes d'optimisation s'appuyant sur des techniques dérivées de la génétique et de l'évolution naturelle : croisements, mutations, sélection, etc. Les algorithmes génétiques ont déjà une histoire relativement ancienne puisque les premiers travaux de John Holland sur les systèmes adaptatifs remontent à 1962 [35].

Pour résumer, Lerman et Ngouenet (1995) distinguent 4 principaux points qui font la différence fondamentale entre ces algorithmes et les autres méthodes :
1. Les algorithmes génétiques utilisent un codage des paramètres, et non les paramètres eux mêmes.
2. Les algorithmes génétiques travaillent sur une population de points, au lieu d'un point unique.
3. Les algorithmes génétiques n'utilisent que les valeurs de la fonction étudiée, pas sa dérivée, ou une autre connaissance auxiliaire.
4. Les algorithmes génétiques utilisent des règles de transition probabilistes, et non déterministes.

## 3-1) Terminologie :

**Chromosome:** En biologie, il est défini comme le porteur de l'information génétique nécessaire à la construction et au fonctionnement d'un organisme. Dans le cadre des AG, il correspond à un élément représentant une solution possible d'un problème donné.

**Gène :** En biologie, il représente une partie du chromosome, chaque chromosome est constitué d'un certain nombre de gènes. Pour un AG, chaque chromosome est divisé en un ensemble d'unités le constituant appelé gènes.

**Génotype:** Dans les systèmes naturels, l'ensemble du "matériel" génétique est appelé le génotype. Dans les AG, l'ensemble des chaines est appelé structure.

**Phénotype:** Dans les systèmes naturels, l'organisme formé par l'interaction de l'ensemble du matériel génétique avec son environnement est appelé phénotype. Dans les AG, les structures décodées forment un ensemble de paramètres donnés, ou solutions ou bien points de l'espace des solutions.

**Allèle:** Dans les systèmes naturels, l'allèle est une composante du gène. Les allèles sont les différentes valeurs que peuvent prendre les gènes. Dans les AG, l'allèle est également appelé valeur caractéristique.

**Locus:** Le locus est la position d'un gène dans le chromosome.
**Individu:** En biologie un individu est une forme qui est le produit de l'activité des gènes. Pour un AG, il est réduit à un chromosome et on l'appelle donc chromosome ou individu pour désigner un même objet.

**Population:** Dans un système naturel, une population est simplement un ensemble d'individus. Par analogie, elle se définit comme l'ensemble des chromosomes. Elle est aussi appelée une génération.

**Parents:** Dans un système naturel, les individus peuvent se reproduire en créant de nouveaux individus formant une nouvelle génération afin d'assurer la continuité de la vie.[36]

Un algorithme génétique recherche le ou les extrema d'une fonction définie sur un espace de données. Pour l'utiliser, on doit disposer des cinq éléments suivants:

1. Un principe de codage de l'élément de population. Cette étape associe à chacun des points de l'espace d'état une structure de données. Elle se place généralement après une phase de modélisation mathématique du problème traité. La qualité du codage des données conditionne le succès des algorithmes génétiques. Le codage binaires ont été très utilisés à l'origine. Les codages réels sont désormais largement utilisés, notamment dans les domaines applicatifs pour l'optimisation de problèmes à variables réelles.

2. Un mécanisme de génération de la population initiale. Ce mécanisme doit être capable de produire une population d'individus non homogène qui servira de base pour les générations futures. Le choix de la population initiale est important car il peut rendre plus ou moins rapide la convergence vers l'optimum global. Dans le cas où l'on ne connaît rien du problème à résoudre, il est essentiel que la population initiale soit répartie sur tout le domaine de recherche.

3. Une fonction à optimiser. Celle-ci retourne une valeur de Â $^{+}$ appelée *fitness* ou fonction d'évaluation de l'individu.

4. Des opérateurs permettant de diversifier la population au cours des générations et d'explorer l'espace d'état. L'opérateur de croisement recompose les gènes d'individus existant dans la population, l'opérateur de mutation a pour but de garantir l'exploration de l'espace d'états.

5. Des paramètres de dimensionnement : La taille de la population, le nombre total de générations ou critère d'arrêt, les probabilités d'application des opérateurs de croisement et de mutation. [37]

## 3-2) Description :

### 3-2-1) Le Codage :

Historiquement le codage utilisé par les algorithmes génétiques était représenté sous forme de chaînes de bits contenant toute l'information nécessaire à la description d'un point dans l'espace d'état. Ce type de codage a pour intérêt de permettre de créer des opérateurs de croisement et de mutation simples. C'est également en utilisant ce type de codage que les premiers résultats de convergence théorique ont été obtenus.

Cependant, ce type de codage n'est pas toujours bon comme le montrent les deux exemples suivants :

- deux éléments voisins en terme de distance de Hamming ne codent pas nécessairement deux éléments proches dans l'espace de recherche. Cet inconvénient peut être évité en utilisant un codage de Gray.
- Pour des problèmes d'optimisation dans des espaces de grande dimension, le codage binaire peut rapidement devenir mauvais. Généralement, chaque variable est représentée par une partie de la chaîne de bits et la structure du problème n'est pas bien reflétée, l'ordre des variables ayant une importance dans la structure du chromosome alors qu'il n'en a pas forcément dans la structure du problème. [38]

### 3-2-1-1) Les types de codage :

- **Codage binaire :** Méthodes d'encodage les plus courantes. Les chromosomes sont des strings de 1 et de 0 et chaque position dans le chromosome représente une caractéristique particulière du problème. [39]
- **Codage de gray :** Le code de Gray, également appelé code Gray ou code binaire réfléchi, est un type de codage binaire permettant de ne modifier qu'un seul bit à la fois quand un nombre est augmenté d'une unité. Cette propriété est importante pour plusieurs applications. [WIKI]

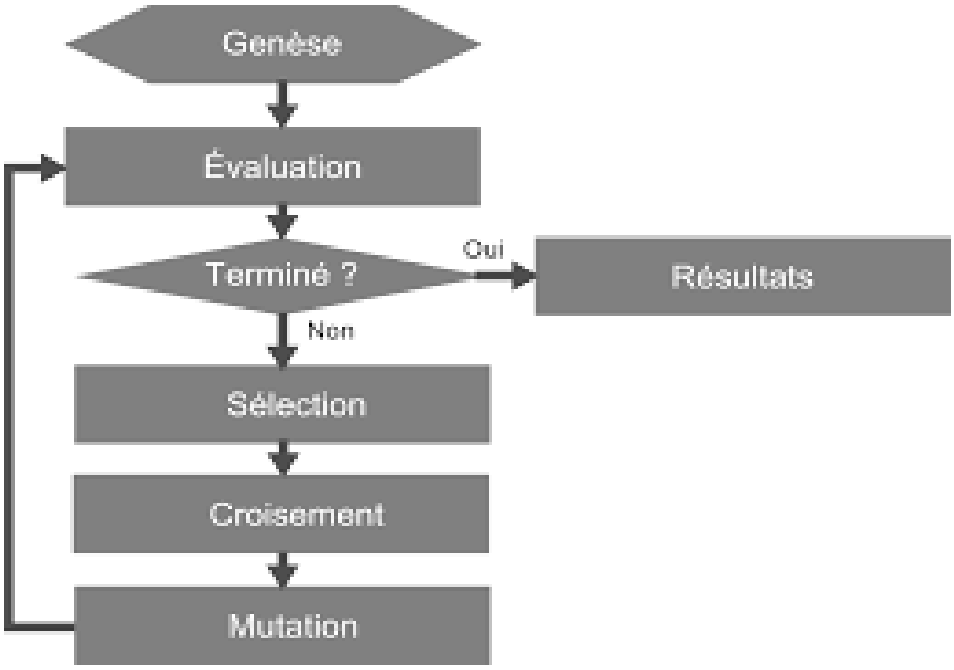
## 3-2-2) Génération de la population initiale:

Le choix de la population initiale d'individus conditionne fortement la rapidité de l'algorithme. Une connaissance des solutions de bonne qualité comme point d'initialisation permet à l'algorithme de converger plus rapidement vers l'optimum ou du moins s'y rapprocher.

Les individus sont alors générés dans un sous-domaine particulièrement proche de ces solutions de départ. Dans le cas où l'on ne dispose que peu d'informations sur le problème à résoudre, il est naturel de générer aléatoirement des individus, mais il est essentiel que la population initiale soit répartie sur tout le domaine de recherche. Tout en veillant à ce que les individus produit respectent les contraintes. Dans l'hypothèse où la gestion des contraintes ne peut se faire directement, les contraintes sont généralement incluses dans le critère à optimiser sous forme de pénalités. La diversité de la population doit être entretenue aux cours des générations afin d'explorer le plus largement possible l'espace de recherche. C'est le rôle des opérateurs de croisement et de mutation. [39]

### 3-2-3) La fonction de fitness :

La fonction de fitness détermine l'aptitude d'un individu (sa capacité à rivaliser avec d'autres individus). Elle donne un score de fitness à chaque individu. La probabilité qu'un individu soit sélectionné pour la reproduction est basée sur son score de fitness. [40]




**FIGRURE 3-1 : les étapes de AG**

### 3-2-4) Les opérateurs de l'AG :

### 3-2-4-1) Le Crossover :

Le Crossover est un opérateur génétique utilisé pour faire varier la programmation d'un ou plusieurs chromosomes d'une génération à l'autre. Deux chaînes sont choisies au hasard pour se croiser afin de produire une descendance supérieure. La méthode choisie dépend de la méthode de codage. [41]

### 3-2-4-1-1) Les types de crossover :

**One point crossover :**

Dans cet opérateur, un point de combinaison est sélectionné pour les chromosomes des deux parents. Les sections chromosomiques après ces points de combinaison sont échangées l'une avec l'autre, ce qui donne naissance à deux nouvelles descendances [42]

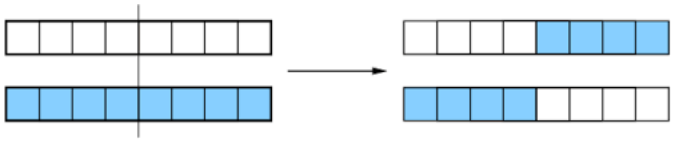


**FIGURE 3-2 : Crossover a un point**

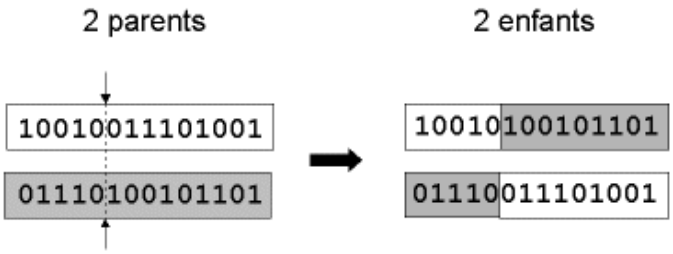



**FIGURE 3-3 : Crossover a un point binaire**

**Corssover à N points :**

Le croisement à N points a été mis en œuvre pour la première fois par De Jong en 1975. Il comporte de nombreux sites de croisement mais la règle utilisée est la même que celle utilisée dans le croisement à un point.[42]

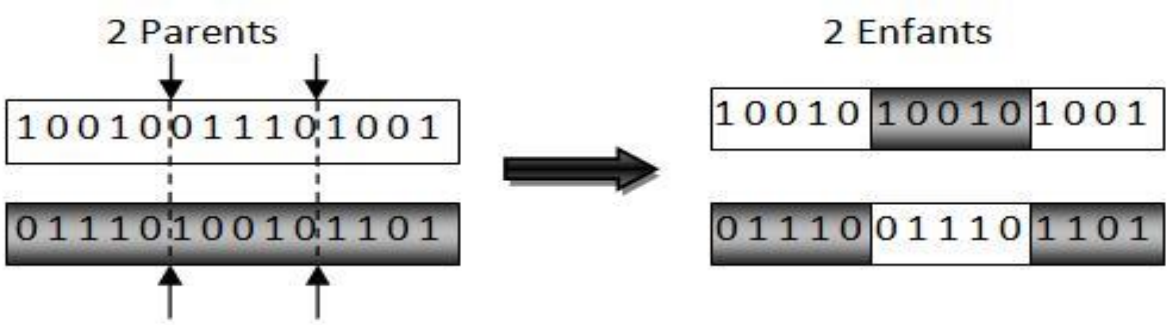



**FIGURE 3-4: Crossover à 2 points**

**Uniform Crossover :**

Le crossover uniforme permet de combiner uniformément les bits des deux parents. Il effectue cette opération de permutation des bits dans les parents à inclure dans la progéniture en choisissant un nombre réel aléatoire uniforme u (entre 0 et 1)

Le croisement uniforme sélectionne les deux parents pour le croisement. Il crée deux descendants de n gènes sélectionnés uniformément parmi les deux parents de manière uniforme. Le nombre réel aléatoire décide si le premier enfant sélectionne le ième gène du premier ou second parent. [42]

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Parent 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| Parent 2 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| Mask | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| Offspring 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Offspring 2 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |

**FIGURE 3-5 : Uniform crossover**

### 3-2-4-1-2 Problème avec le Crossover :

Selon le codage, de simples croisements peuvent avoir de fortes chances de produire une progéniture illégale.

Par exemple, dans TSP avec un simple codage binaire ou de chemin, la plupart des descendants seront illégaux car toutes les villes ne seront pas dans la progéniture et certaines villes y seront plus d'une fois.[43]

### 3-2-4-2) La Mutation :

La mutation consiste à altérer un gène dans un chromosome selon un facteur de mutation. Ce facteur est la probabilité qu'une mutation soit effectuée sur un individu. Cet opérateur est l'application du *principe de variation* de la théorie de Darwin et permet, par la même occasion, d'éviter une convergence prématurée de l'algorithme vers un extremum local. [44]

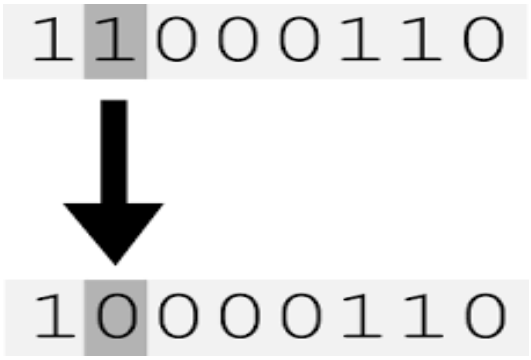



**FIGURE 3-6 : La Mutation**

### 3-2-4-3) La Sélection :

La sélection consiste à choisir les individus les mieux adaptés afin d'avoir une population de solution la plus proche de converger vers l'optimum global. Cet opérateur est l'application du *principe d'adaptation* de la théorie de Darwin. Il existe plusieurs techniques de sélection :

**Sélection par roulette (Roulette Wheel sélection) :**

Pour chaque individu, la probabilité d'être sélectionné est proportionnelle à son adaptation au problème. Le principe de Roulette Wheel sélection est celui de la roue de la fortune biaisée. Cette roue est une roue de la fortune classique sur laquelle on associe à chaque individu un

segment dont la longueur est proportionnelle à sa fitness. On effectue ensuite un tirage aléatoire utilisé dans les roulettes de casinos avec une structure linéaire. Avec ce système, les grands segments, c'est-à-dire les bons individus, seront plus souvent adressés que les petits. Sélection par tournoi :

La sélection par tournoi consiste à sélectionner n individus au hasard et à prendre le meilleur parmi ces n individus. On organise autant de tournois qu'il y a d'individus à repêcher. Le nombre n permet de donner plus ou moins de chance aux individus peu adaptés. Avec un nombre élevé de participants, un individu faible sera presque toujours sûr de perdre. Le nombre d'individus par tournoi détermine les paramètres d'exploration (n petit) et d'exploitation (n grand) du bassin génétique.

**Sélection par rang :**

La sélection par rang trie d'abord la population par fitness. Ensuite, chaque chromosome se voit associé un rang en fonction de sa position. Le plus mauvais chromosome aura le rang, le suivant, et ainsi de suite jusqu'au meilleur chromosome qui aura le rang (pour une population de chromosomes). La sélection par rang d'un chromosome est la même que par roulette, mais les proportions sont en relation avec le rang plutôt qu'avec la valeur de 6 l'évaluation, c'est à dire les individus choisis sont ceux qui possèdent les meilleurs scores d'adaptation (meilleur rang), le hasard n'entre donc pas dans ce mode de sélection.

**Sélection « steady-state » :**

L'idée principale est qu'une grande partie de la population puisse survivre à la prochaine génération. A chaque génération sont sélectionnés quelques chromosomes (parmi ceux qui ont le meilleur coût) pour créer des chromosomes fils. Ensuite les chromosomes les plus mauvais sont retirés et remplacés par les nouveaux. Le reste de la population survie à la nouvelle génération.

**Elitisme :**

A la création d'une nouvelle population, il y a de grandes chances que les meilleurs chromosomes soient perdus après les opérations d'hybridation et de mutation. Pour éviter cela, on utilise la méthode d'élitisme. Elle consiste à copier un ou plusieurs des meilleurs chromosomes dans la nouvelle génération. Ensuite, on génère le reste de la population selon l'algorithme de reproduction usuel. Cette méthode améliore considérablement les algorithmes génétiques, car elle permet de ne pas perdre les meilleures solutions.

**Sélection uniforme:**

La sélection se fait aléatoirement, uniformément et sans intervention de la valeur d'adaptation. Chaque individu a donc une probabilité 1/P d'être sélectionné, où P est le nombre total d'individus dans la population [45]

## 4) L'utilisation des AG dans la détection d'intrusion :

L'objectif de l'application de l'AG est de générer des règles qui ne correspondent qu'aux connexions anormales. Ces règles sont testées sur les connexions historiques et sont utilisées pour filtrer les nouvelles connexions afin de trouver le trafic réseau suspect.

Dans cette implémentation, le trafic réseau utilisé pour GA est un ensemble de données pré-classifié qui différencie les connexions réseau normales des connexions anormales.

En démarrant l'AG avec un petit ensemble de règles générées de manière aléatoire, nous pouvons générer un ensemble de données plus important qui contient des règles pour les IDS. Ces règles sont des solutions "suffisamment bonnes" pour l'AG et peuvent être utilisées pour filtrer le nouveau trafic réseau.[46]

## 5) Les avantages des AG :

Les algorithmes génétiques présentent de nombreux avantages par rapport aux algorithmes d'optimisation traditionnels. Les deux plus notables sont : la capacité de traiter des problèmes complexes et le parallélisme. Les algorithmes génétiques peuvent traiter différents types d'optimisation, que la fonction objectif (fitness) soit stationnaire ou non stationnaire (change avec le temps), linéaire ou non linéaire, continue ou discontinue, ou avec un bruit aléatoire. Comme les multiples descendants d'une population agissent comme des agents indépendants, la population (ou tout sous-groupe) peut explorer l'espace de recherche dans plusieurs directions simultanément. Cette caractéristique en fait un outil idéal pour paralléliser les algorithmes d'implémentation. Différents paramètres et même différents groupes de chaînes codées peuvent être manipulés en même temps.

Cependant, les algorithmes génétiques présentent également certains inconvénients. La formulation de la fonction de fitness, l'utilisation de la taille de la population, le choix des paramètres importants tels que le taux de mutation et de croisement, et les critères de sélection de la nouvelle population doivent être effectués avec soin. Tout choix inapproprié rendra difficile la convergence de l'algorithme ou produira simplement des résultats sans signification. Malgré ces inconvénients, les algorithmes génétiques restent l'un des algorithmes d'optimisation les plus utilisés dans l'optimisation non linéaire moderne. [47]

# 6) Le Data Mining :

## 6-1)Présentation générale :

Le Data Mining est un sujet qui dépasse aujourd'hui le cercle restreint de la communauté scientifique pour susciter un vif intérêt dans le monde des affaires. La littérature spécialisée et la presse ont pris le relais de cet intérêt et proposent pléthore de définitions générales du Data Mining :

• « l'extraction d'informations originales, auparavant inconnues, potentiellement utiles à partir de données » .

• « la découverte de nouvelles corrélations, tendances et modèles par le tamisage d'un large volume de
données » ;

• « un processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données » ;

• D'autres, plus poétiques, parlent de « torturer l'information disponible jusqu'à ce qu'elle avoue ».[48]

## 6-2) Définition:

Data Mining est une collection de techniques pour la découverte automatisée et efficace de modèles précédemment inconnus, , utiles et compréhensibles dans de grandes bases de données. Les modèles doivent être exploitables afin de pouvoir être utilisés dans le processus de décision d'une entreprise.[49]

## 6-3) Pourquoi le Data Mining?

Le data mining est né dans la nécessité de valoriser les immenses bases de données d'entreprises et conduit à analyser et prévoir les comportements individuels des consommateurs.[50]

## 6-4) Définition de la classification :

La classification est un processus technique qui utilise des algorithmes pour analyser des données à partir de plusieurs perspectives et d'extraire des modèles significatifs qui peuvent être utilisés pour prédire le comportement futur des utilisateurs.[51]

## 6-5) L'apprentissage supervisé :

L'exploration supervisée des données, comme son nom l'indique, fait référence aux algorithmes d'apprentissage utilisés pour la classification et la prédiction. L'algorithme supervisé apprend à partir des données d'apprentissage qui sont étiquetées et la tâche est contrôlée par l'ingénieur des connaissances et le concepteur du système. Avec des données supervisées, nous devons avoir des entrées connues correspondant à des sorties connues, déterminées par des experts du domaine.[52]

## 6-6) L'apprentissage non-supervisé :

Contrairement à la technique supervisée, l'exploration de données non supervisée n'a pas de fonction objective prédéterminée et ne prédit pas de valeur cible.Les techniques non supervisées sont celles où il n'y a pas de variable de résultat à prédire ou à classer. Par conséquent, il n'y a pas d'apprentissage à partir de cas où une telle variable de résultat est connue.[52]

## 6-7) L'algorithme KNN :

L'algorithme de classification KNN oui l'algorithme des K-voisins les plus proches, est l'un des algorithmes de classification les plus utilisés dans le domaine de l'intelligence artificielle. Son idée de base est la suivante : lors de la saisie de nouvelles données de catégorie inconnue à classer, la catégorie des données à classer doit être déterminée en fonction de la catégorie des autres échantillons.Il est basé sur une fonction de distance qui calcule la différence ou la similarité entre deux instances. [53]

A partir de ces K petites distances, l'instance est assignée à la classe majoritaire.

La distance Euclidienne :

$$d_{Eucl}(p,q) = \\ = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2} = \\ = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

**FIGURE 3-7 : La distance Euclidienne.**

P et q : vecteurs.

n : nombre des attributs.

La distance de Manhattan :

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

**FIGURE 3-8 : la distance de Manhattan.**

## 6-8)L'algorithme de naÏve bayes :

Naïve Bayes est un algorithme d'apprentissage automatique probabiliste basé sur le théorème de Bayes qui calcule les probabilités conditionnelles:

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

**FIGURE 3-9 : Théorème de Bayes**

Le terme P(A/B) se lit : **la probabilité que l'événement A se réalise sachant que l'événement B s'est déjà réalisé.**

## 6-9) L'algorithme K Means :

Le K-Means est un algorithme d'apprentissage non supervisé, qui regroupe l'ensemble de données non étiquetées en différents clusters.

Il nous permet de regrouper les données en différents groupes et constitue un moyen pratique de découvrir les catégories de groupes dans l'ensemble de données non étiquetées sans avoir besoin d'entraînement.

Il s'agit d'un algorithme basé sur les centroïdes, où chaque cluster est associé à un centroïde. L'objectif principal de cet algorithme est de minimiser la somme des distances entre le point de données et les clusters correspondants.[54]

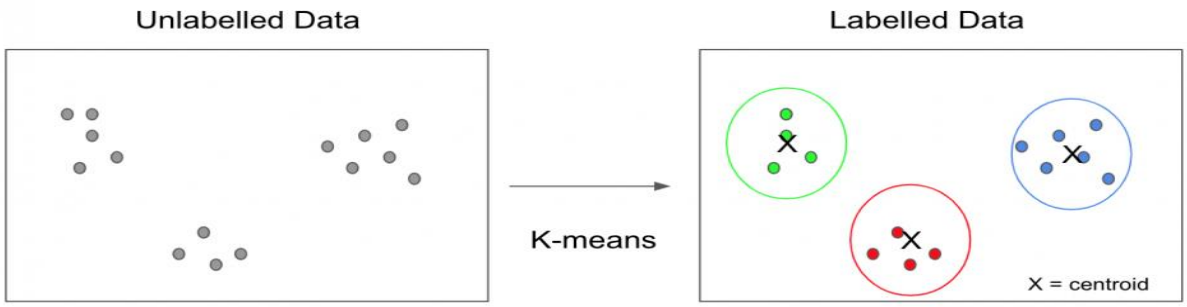



**FIGURE 3-10 : L'algorithme de K Means.**

## 7) Conclusion :

Dans ce chapitre, nous avons présenté les algorithmes méta-heuristiques, puis nous avons détaillé les algorithmes génétiques qui concernent notre travail.

On a commencé par des notions de base, puis les étapes de l'algorithme génétique.

Enfin, nous avons parlé de l'utilisation des AG pour la détection d'intrusion, et les algorithmes du Data Mining et les deux types de classification, ce que nous allons implémenté et illustré dans le prochain chapitre.

# CHAPITRE4

# IMPLÉMENTATION ET DISCUSSION DES RÉSULTATS.

## 1-Introduction :

La détection d'intrusion est un problème np-complet, les méthodes métaheuristiques sont très efficaces pour résoudre ce type de problèmes.

L'une de ces méthodes est les algorithmes génétiques qu'on a décrits dans le chapitre précédent.

Cette méthode nous donne la possibilité de classifier les données en deux classes : « Normal » et « Intrusion ».

Dans ce chapitre nous allons parler sur les étapes de développement de notre application, en commençant par le data set utilisé, puis la technique pour choisir les meilleurs attributs, enfin l'utilisation des algorithmes génétiques et les algorithmes de data mining suivi par la discussion des résultats.

## 2-Outils de réalisation :

### 2-1) Langage de programmation python :

Python est un langage de programmation interprété multi-paradigme. Il favorise la programmation impérative structurée, et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ; il est ainsi similaire à Perl, Ruby, Scheme, Smalltalk et Tcl.[55]

Ce langage supporte les méthodes Data Mining et facilite la classification.

Nous avons utilisé les bibliothèques suivantes :

Matplotlib: pour représentation des données sous forme graphiques.
Numpy : pour la structuration des données.
Pandas : pour la manipulation et l'analyse des données.
SciKit-Learn : destinée à l'apprentissage automatique.
Streamlit : pour l'interface graphique.

### 2-2) Google Colab :

Google Colab ou Colaboratory est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plateforme permet d'entraîner des modèles de Machine Learning directement dans le

cloud. Sans donc avoir besoin d'installer quoi que ce soit sur notre ordinateur à l'exception d'un navigateur.[56]

## 3) Présntation du dataset (KDDCUP99):

Le programme d'évaluation de la détection d'intrusion du DARPA de 1998 a été préparé et géré par les Lincoln Labs du MIT. L'objectif était de recenser et d'évaluer la recherche en matière de détection des intrusions. Un ensemble standard de données à vérifier, qui comprend une grande variété d'intrusions simulées dans un environnement de réseau militaire, a été fourni. Le concours de détection d'intrusion KDD 1999 utilise une version de cet ensemble de données.

Les Lincoln Labs ont mis en place un environnement permettant d'acquérir neuf semaines de données brutes de TCP pour un réseau local (LAN) simulant un LAN typique de l'armée de l'air américaine. Ils ont exploité le réseau local comme s'il s'agissait d'un véritable environnement de l'armée de l'air, mais l'ont soumis à de multiples attaques.

Les données brutes d'entraînement étaient constituées d'environ quatre gigaoctets de données TCP binaires compressées provenant de sept semaines de trafic réseau. Elles ont été transformées en environ cinq millions d'enregistrements de connexion. De même, les deux semaines de données de test ont donné environ deux millions d'enregistrements de connexion.

Une connexion est une séquence de paquets TCP commençant et se terminant à des moments bien définis, entre lesquels les données circulent vers et depuis une adresse IP source vers une adresse IP cible selon un protocole bien défini. Chaque connexion est étiquetée soit comme normale, soit comme une attaque, avec exactement un type d'attaque spécifique. Chaque enregistrement d'une connexion est constitué d'environ 100 octets. [57]

```
#Reading the dataset
import pandas as pd
df = pd.read_csv("/content/drive/MyDrive/bin_data.csv")
df.head()
```

| | duration | protocol_type | service | flag | src_bytes | dst_bytes | land | wrong_fragment | urgent | hot | ... | dst_hos |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | tcp | http | SF | 181 | 5450 | 0 | 0 | 0 | 0 | ... | |
| 1 | 0 | tcp | http | SF | 239 | 486 | 0 | 0 | 0 | 0 | ... | |
| 2 | 0 | tcp | http | SF | 235 | 1337 | 0 | 0 | 0 | 0 | ... | |
| 3 | 0 | tcp | http | SF | 219 | 1337 | 0 | 0 | 0 | 0 | ... | |
| 4 | 0 | tcp | http | SF | 217 | 2032 | 0 | 0 | 0 | 0 | ... | |

5 rows × 42 columns

**Figure 4-1: les cinq premières lignes du dataset.**

Le KDD99 contient 41 attributs et le dernier pour la classe de connexion, la figure suivante affiche l'ensemble des attributs :

| Nr | Features | |
|---|---|---|
| | Name | Description |
| 1 | duration | duration of connection in seconds |
| 2 | protocol_type | connection protocol (tcp, udp, icmp) |
| 3 | service | dst port mapped to service (e.g. http, ftp, ..) |
| 4 | flag | normal or error status flag of connection |
| 5 | src_bytes | number of data bytes from src to dst |
| 6 | dst_bytes | bytes from dst to src |
| 7 | land | 1 if connection is from/to the same host/port; else 0 |
| 8 | wrong_fragment | number of 'wrong' fragments (values 0,1,3) |
| 9 | urgent | number of urgent packets |
| 10 | hot | number of 'hot' indicators (bro-ids feature) |
| 11 | num_failed_logins | number of failed login attempts |
| 12 | logged_in | 1 if successfully logged in; else 0 |
| 13 | num_compromised | number of 'compromised' conditions |
| 14 | root_shell | 1 if root shell is obtained; else 0 |
| 15 | su_attempted | 1 if 'su root' command attempted; else 0 |
| 16 | num_root | number of 'root' accesses |
| 17 | num_file_creations | number of file creation operations |
| 18 | num_shells | number of shell prompts |
| 19 | num_access_files | number of operations on access control files |
| 20 | num_outbound_cmds | number of outbound commands in an ftp session |
| 21 | is_hot_login | 1 if login belongs to 'hot' list (e.g. root, adm); else 0 |
| 22 | is_guest_login | 1 if login is 'guest' login (e.g. guest, anonymous); else 0 |
| 23 | count | number of connections to same host as current connection in past two seconds |
| 24 | srv_count | number of connections to same service as current connection in past two seconds |
| 25 | serror_rate | % of connections that have 'SYN' errors |
| 26 | srv_serror_rate | % of connections that have 'SYN' errors |
| 27 | rerror_rate | % of connections that have 'REJ' errors |
| 28 | srv_rerror_rate | % of connections that have 'REJ' errors |
| 29 | same_srv_rate | % of connections to the same service |
| 30 | diff_srv_rate | % of connections to different services |
| 31 | srv_diff_host_rate | % of connections to different hosts |
| 32 | dst_host_count | count of connections having same dst host |
| 33 | dst_host_srv_count | count of connections having same dst host and using same service |
| 34 | dst_host_same_srv_rate | % of connections having same dst port and using same service |
| 35 | dst_host_diff_srv_rate | % of different services on current host |
| 36 | dst_host_same_src_port_rate | % of connections to current host having same src port |
| 37 | dst_host_srv_diff_host_rate | % of connections to same service coming from diff. hosts |
| 38 | dst_host_serror_rate | % of connections to current host that have an S0 error |
| 39 | dst_host_srv_serror_rate | % of connections to current host and specified service that have an S0 error |
| 40 | dst_host_rerror_rate | % of connections to current host that have an RST error |
| 41 | dst_host_srv_rerror_rate | % of connections to the current host and specified service that have an RST error |
| 42 | connection_type | |

**Figure 4-2 : Les attributs du dataset KDDCUP99**

Ce dernier attribut  représente à la base 23 classes entre attaques et connexions normales,

Dans notre travail, on va utiliser 2 classes :

Une classe « NonIntrusion » qui représente les connexions normales et une classe «Intrusion »

qui représente toutes les attaques.

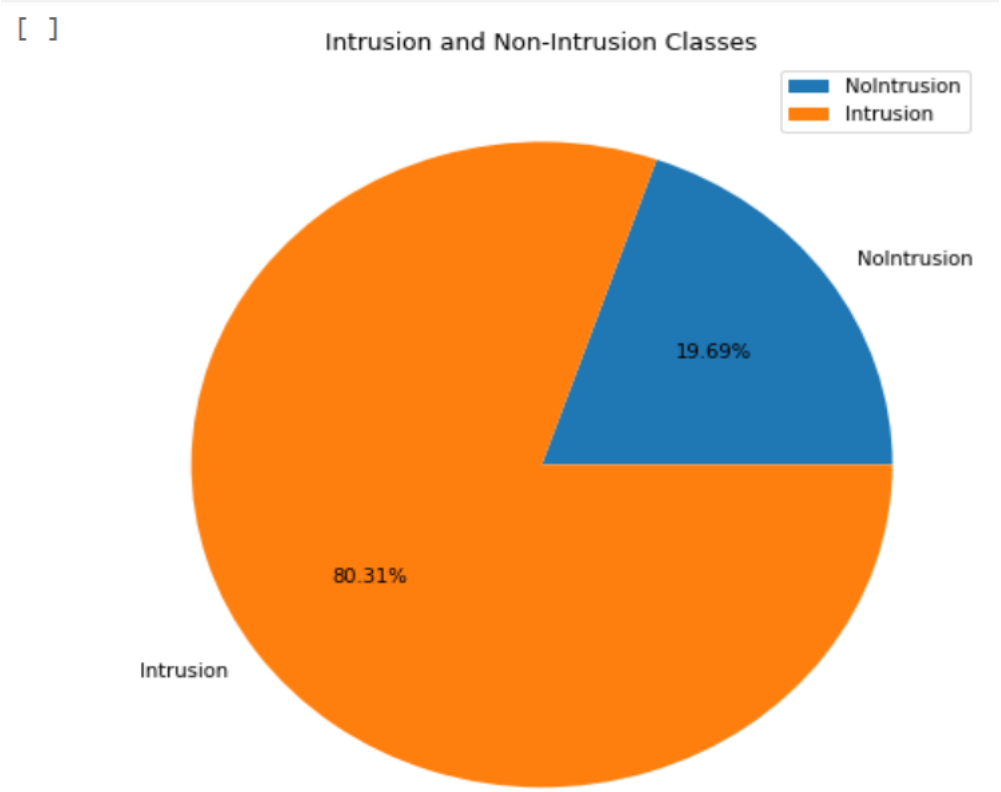



**Figure 4-3 : Intrusion and Non-Intrusion Classes**

## 3-1) Numérisation des valeurs des attributs :

Ce dataset contient 3 attributs qui ont des valeurs catégoriques « protocol_type , service, flag ». Les individus de la population initiale de l'algorithme génétique doivent etre de meme type (binaire). Pour cela on va numériser ces valeurs :

| Les valeurs catégoriques de l'attribut « type_protocol | Conversion |
|---|---|
| Icmp | 0 |
| tcp | 1 |
| udp | 2 |

**Tableau 4-1 : Conversion alphabétique simple de l'attribut « protocol_type »**

Et on fait le même travail pour les 2 attributs :

```
# flag feature mapping
fmap = {'SF':0, 'S0':1, 'REJ':2, 'RSTR':3, 'RSTO':4, 'SH':5, 'S1':6, 'S2':7, 'RSTOS0':8, 'S3':9, 'OTH':10}
DataFtr['flag'] = DataFtr['flag'].map(fmap)

#Service
Smap = {'http':0, 'smtp':1, 'finger':2, 'domain_u':3, 'auth':4, 'telnet':5, 'ftp':6,
       'eco_i':7, 'ntp_u':8, 'ecr_i':9, 'other':10, 'private':11, 'pop_3':12, 'ftp_data':13,
       'rje':14, 'time':15, 'mtp':16, 'link':17, 'remote_job':18, 'gopher':19, 'ssh':20,
       'name':21, 'whois':22, 'domain':23, 'login':24, 'imap4':25, 'daytime':26, 'ctf':27,
       'nntp':28, 'shell':29, 'IRC':30, 'nnsp':31, 'http_443':32, 'exec':33, 'printer':34,
       'efs':35, 'courier':36, 'uucp':37, 'klogin':38, 'kshell':39, 'echo':40, 'discard':41,
       'systat':42, 'supdup':43, 'iso_tsap':44, 'hostnames':45, 'csnet_ns':46, 'pop_2':47,
       'sunrpc':48, 'uucp_path':49, 'netbios_ns':50, 'netbios_ssn':51, 'netbios_dgm':52,
       'sql_net':51, 'vmnet':52, 'bgp':53, 'Z39_50':54, 'ldap':55, 'netstat':56, 'urh_i':57,
       'X11':58, 'urp_i':59, 'pm_dump':60, 'tftp_u':61, 'tim_i':62, 'red_i':63}
```

**Figure 4-4 : mapping des valeurs des attributs «flag » et « service » en python**

## 3-2) La sélection des attributs :

Chaque ligne du dataset contient 41 attributs, ce que rendre la classification difficile. Donc l'étape de sélection des attributs est très importante dans notre travail pour réduire le cout et le temps d'exécution.

Il existe plusieurs techniques pour sélectionner les attributs, dans notre travail nous avons utilisé la technique CHI-deux (Khi-deux).

### 3-2-1) La méthode CHI-deux :

Cette méthode permet de sélectionner les termes en tenant compte de leurs fréquences dans chaque classe, avec l'idée d'extraire les meilleurs termes qui caractérisent une classe par rapport à l'autre [58].

La formule est la suivante :

```
from scipy.stats import chi2
chi_square=sum([(o-e)**2./e for o,e in zip(Observed_Values,Expected_Values)])
chi_square_statistic=chi_square[0]+chi_square[1]
```

**Figure 4-5 : CHI-deux.**

o : l'effectif théorique.

e : l'effectif observé.

Après le classement des attributs du meilleur au pire, nous avons utilisé la technique SelectK-best pour choisir les K meilleurs attributs, dont K est le nombre d'attributs (au choix).

La valeur de CHI2 est un paramètre de la méthode K-Best.

```
FtChi2 = SelectKBest(score_func= chi2 ,k=5)
```

Voici le résultat des meilleurs attributs:

| Numéro | Attribut | CHI2 |
|---|---|---|
| 1 | Duration | 7.183210e+07 |
| 2 | protocol_type | 9.498235e+04 |
| 3 | service | 4.630333e+05 |
| 5 | src_bytes | 1.397802e+08 |
| 6 | dst_bytes | 8.829356e+08 |

**TABLEAU 4-2 : Choix des attributs.**

## 4) Apprentissage de l'algorithme par les données :

Après la préparation des données, nous pouvons appliquer les algorithmes de classification.

### 4-1) L'application de l'algorithme génétique :

L'entrée de l'AG est un ensemble d'individus de type binaire constitués de suite de 1 et de 0 appelée la population initiale.

Le dernier bit représente la classe comme suite :

0 : pour la classe « Non-Intrusion ».

1 : pour la classe « Intrusion ».

### 4-1-1) La première étape « Sélection » :

Nous avons choisi la sélection par tournoi :

```
# tournament selection
def selection(pop, scores, k=3):
  # first random selection
  selection_ix = randint(len(pop))
  for ix in randint(0, len(pop), k-1):
    # check if better (e.g. perform a tournament)
    if scores[ix] < scores[selection_ix]:
      selection_ix = ix
  return pop[selection_ix]
```

**Figure 4-6 : L'étape sélection d'AG.**

### 4-1-2) La deuxième étape « Crossover » :

Consiste à faire croisement entre deux individus comme suite :

```
# crossover two parents to create two children
def crossover(p1, p2, r_cross):
  # children are copies of parents by default
  c1, c2 = p1.copy(), p2.copy()
  # check for recombination
  if rand() < r_cross:
    # select crossover point that is not on the end of the string
    pt = randint(1, len(p1)-2)
    # perform crossover
    c1 = p1[:pt] + p2[pt:]
    c2 = p2[:pt] + p1[pt:]
  return [c1, c2]
```

**Figure 4-7 :L'étape crossover d'AG.**

p1 ,p2 :les parents.

r_cross : le facteur de croisement.

c1, c2 : les enfants.

pt : point de croisement.

### 4-1-3)La troisième étape «mutation » :

Consiste à choisir un bit au hasard et le changer.

```
# mutation operator
def mutation(bitstring, r_mut):
  for i in range(len(bitstring)):
    # check for a mutation
    if rand() < r_mut:
      # flip the bit
      bitstring[i] = 1 - bitstring[i]
```

**Figure 4-8 : l'étape de mutation d'AG.**

On fait appel à ces méthodes, le résultat de l'AG est un sac des individus optimales binaires, on fait la conversion vers le décimal pour appliquer les algortihmes du Data Mining .

Exemple :

```
[4095, 3, 127, 1073741823, 8388607]
1
```

**Figure 4-9 : résultat optimal de l'AG.**

### 4-1-4) la fonction de fitness :

La fonction de fitness que nous avons choisit pour l'algorithme génétique est la moyenne de chaque instance, elle égale à la somme des valeurs des attributs divisée sur leur nombre. Chaque instance va être classée par rapport à sa valeur de fitness, et on prend les meilleures instances.

## 4-2)Classification à la base des algorithmes de Data Mining :

Pour la classification supervisée, nous avons utilisé l'algorithme K plus proche voisins et le naive bayes, et le k-means pour la classification non-supervisée.

| | | Actual Values | |
|---|---|---|---|
| | | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | TP | FP |
| | Negative (0) | FN | TN |

**Figure 4-10 : Matrice de confusion.**

TP **:** une attaque détectée correctement.

FP : une connexion normale détectée comme intrusion.

FN : une connexion normale correctement détectée.

TN : une intrusion détectée comme connexion normale.

### 4-2-1) Les mesures d'évaluation :

**La précision :** désigne la probabilité qu'une prédiction soit correcte.

**Précision= (TP\TP+FP)*100%**

**Le rappel (Recall) :** désigne le rapport entre le nombre d'intrusion correctement détecté et le nombre total d'intrusion.

**Rappel= (TP\FN+TP)*100%**

**Accuracy :** désigne le rapport entre les détections correctes et le nombre total de détection.

**Accuracy= (TP+TN\ FP+FN+TP+TN)*100%**

**L'entropie:** Désigne quantité d'information délivrée par une source d'information.

**Entropie= -précision*log(précision)**

**F-score (F-mesure):** Il combine le Recall et l'Accuracy comme suite:

**F-score= (Précsion*Rappel)\(Précision+Rappel)**

**La courbe ROC** (Receiver Operating Characteristic) : Est une mesure de performance de classificateur binaire, elle représente la sensibilité en fonction de 1.

**AUC** (Area Under the Curve : Aide à choisir quel modèle performe le mieux, plus le AUC est élevé, plus le modèle performe mieux.

### 4-2-2)Discussion des résultats :

Nous avons voir deux cas :

Le premier cas donne les résultats des performances des algorithmes du datamining appliqués sur le KDDCUP (70% du dataset pour l'apprentissage et 30% pour le test).

Dans le deuxième cas, les algorithmes du datamining ont comme base d'apprentissage 70% du dataset et un sac des instances optimales obtenues par l'algorithme génétique pour le test.

L'objectif de cette étape est de voir si l'algorithme génétique améliore les performances des algorithmes du datamining et augmente les mesures d'évaluation.

## 4-2-2-1) Les résultats des algorithmes du Data Mining: (Avant l'algorithme génétique)

### 4-2-2-1-1) L'algorithme KNN :

L'application du KNN nous a donné les statistiques présentées dans le tableau suivant :

| Les mesures d'évaluation | Accuracy | Entropie | Précision | Rappel | F-score |
|---|---|---|---|---|---|
| Les valeurs(%) | 99,9 | 0,09 | 99,9 | 99,7 | 99,94 |

**TABLEAU 4-3 : Tableau représente les mesures d'évaluation du KNN.**

**Le ROC du KNN**

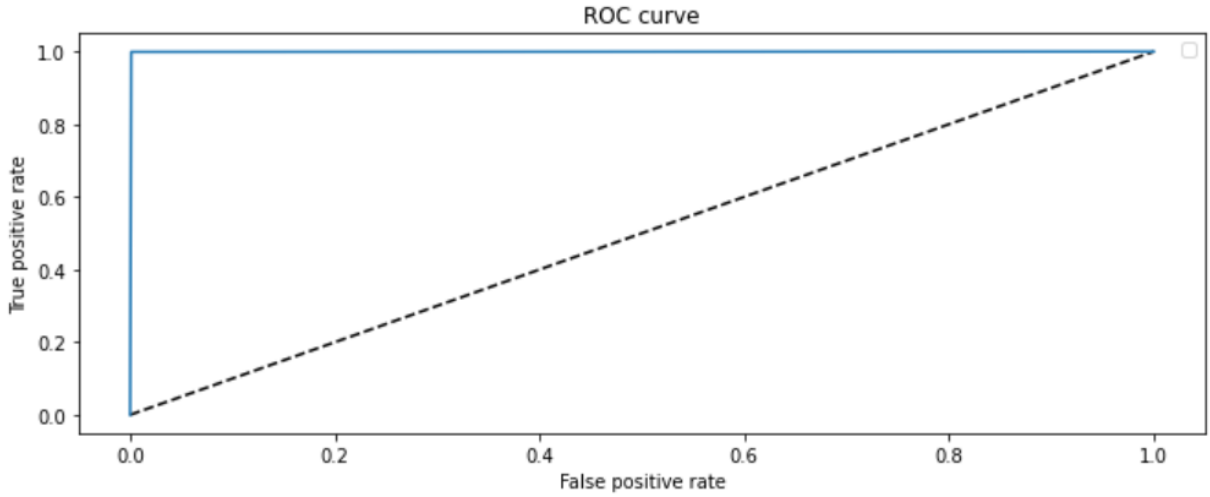



**Figure 4-11 : Le ROC du KNN.**

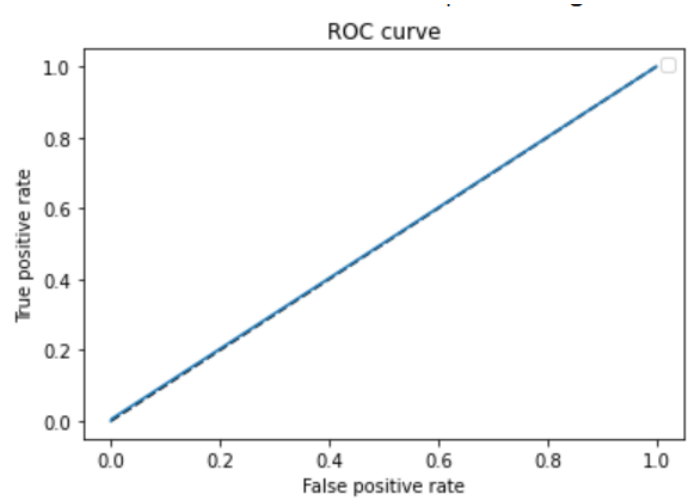
### 4-2-2-1-2)L'algorithme Naive Bayes :

L'algorithme NB donne les résultats suivants :

| Les mesures d'évaluation | Accuracy | Entropie | Précision | Rappel | F-score |
|---|---|---|---|---|---|
| Les valeurs(%) | 20,06 | 2,8 | 0,5 | 95,16 | 1,08 |

**TABLEAU 4-4 : Tableau représente les mesures d'évaluation du NB.**

**Le ROC de NB :**




```
----------------
----------------
array([0.       , 0.0054581, 1.       ])
```

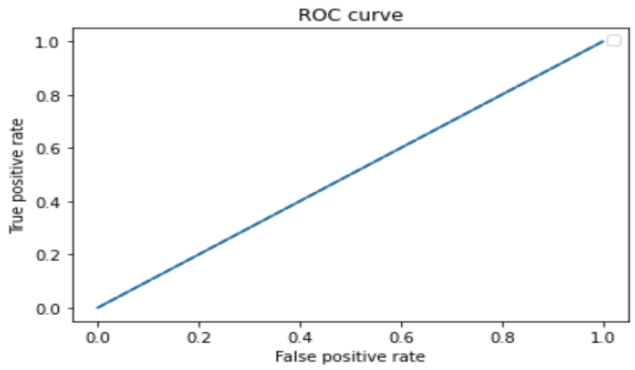
**Figure 4-12: Le ROC du NB.**

### 4-2-2-1-3) L'algorithme K-Means :

| Les mesures d'évaluatin | Accuracy | Entropie | Précision | Rappel | F-score |
|---|---|---|---|---|---|
| Les valeurs(%) | 19,64 | 31,96 | 19,64 | 19,64 | 19,64 |

**TABLEAU 4-5 : Tableau représente les mesures d'évaluation du K-Means.**

**ROC du K-Means:**




```
--------------------------
--------------------------
array([0.00000000e+00, 8.39708117e-06, 4.19854059e-05, 1.76338705e-04,
       1.00000000e+00])
```

**Figure 4-13 : Le ROC du K-Means.**

## 4-2-2-1-4) Comparaison entre les algorithmes :

Nous avons comparé entre les algorithmes, en se basant sur les deux mesures suivantes :

L'accuracy et l'entropie.

### 4-2-2-1-4-1) Avec l'accuracy :

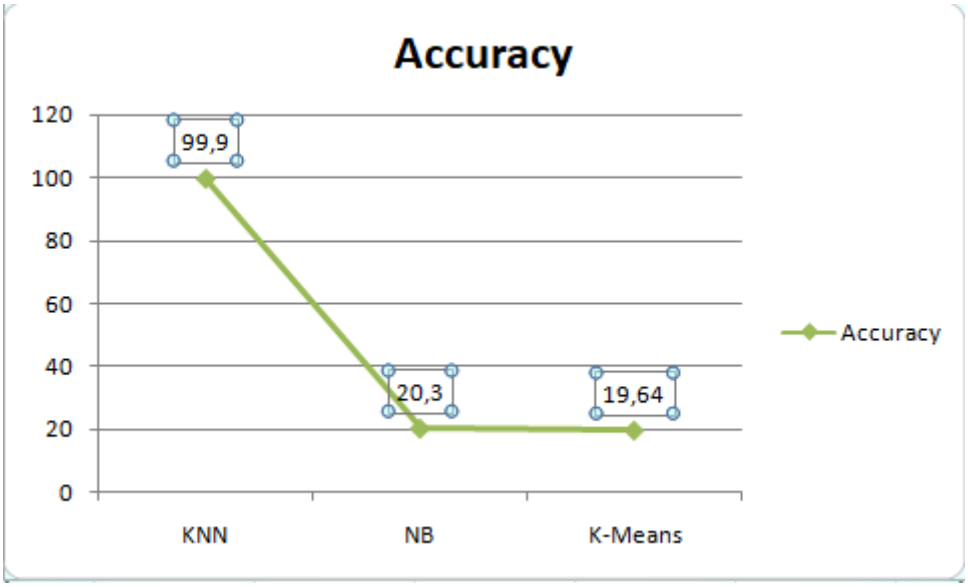



**FIGURE 4-14 : Comparaison avec l'Accuracy**

On remarque que la valeur de l'accuracy de l'algorithme KNN est très élevée, par contre elle est faible dans le NB et le K-Means. On déduit que le KNN a l'avantage dans cette mesure.

### 4-2-2-1-4-2) Avec l'Entropie :

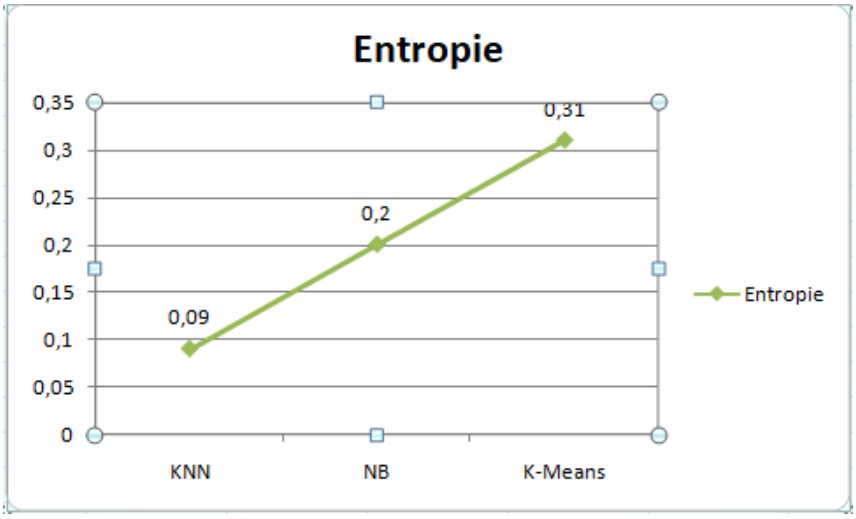



**FIGURE 4-15 : Comparaison avec l'Entropie.**

L'entropie du KNN est trop basse et elle converge vers le zero, l'Entropie de NB et K means est acceptable.

La quantité d'information perdue est petite avec le KNN par rapport aux autres algorithmes donc le KNN est meilleur.

En final, le KNN est meilleur que les autres algorithmes si on prend ces mesures d'évaluation.

## 4-2-2-2) Les résultats des algorithmes du Data Mining: ( Aprés l'algorithme génétique)

### 4-2-2-2-1) L'algorithme KNN :

| Les mesures d'évaluatin | Accuracy | Entropie | Précision | Rappel | F-score |
|---|---|---|---|---|---|
| Les valeurs(%) | 68,75 | 0,9 | 99,9 | 99,9 | 81,477 |

**TABLEAU 4-6 : Le KNN après l'utilisation de l'AG.**

### 4-2-2-2-2) L'algorithme NB :

| Les mesures d'évaluatin | Accuracy | Entropie | Précision | Rappel | F-score |
|---|---|---|---|---|---|
| Les valeurs(%) | 68,74 | 1,96 | 68,745 | 68,745 | 81,477 |

**TABLEAU 4-7 : Le NB après l'utilisation de l'AG.**

### 4-2-2-2-3) L'algorithme K-Means :

| Les mesures d'évaluatin | Accuracy | Entropie | Précision | Rappel | F-score |
|---|---|---|---|---|---|
| Les valeurs(%) | 60,05 | 30 | 60,54 | 60,54 | 60,54 |

**TABLEAU 4-8 : Le K-Means après l'utilisation de l'AG.**

### 4-2-2-3-) Comparaison entre les deux cas :

Dans cette étape, nous avons comparé les résultats avant et après l'application de l'algorithme génétique pour savoir s'il a un effet sur les mesures d'évaluation :

**Avec l'Entropie :**

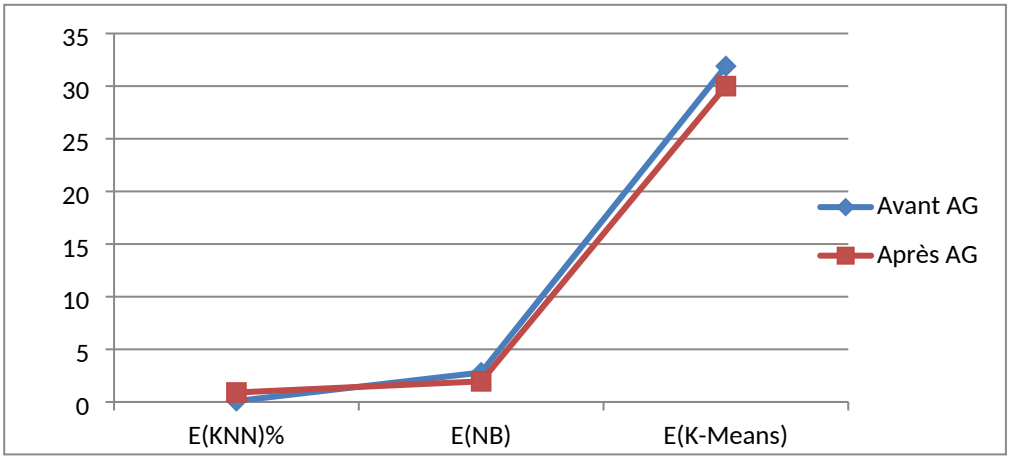



**FIGURE 4-16 : Comparaison avant et après l'AG par rapport à l'entropie.**

On remarque que les valeurs de  cette mesure sont très proches dans l'algorithme KNN,

On a obtenu moins d'informations perdues dans le cas du NB et K-Means.

**Avec l'Accuracy :**

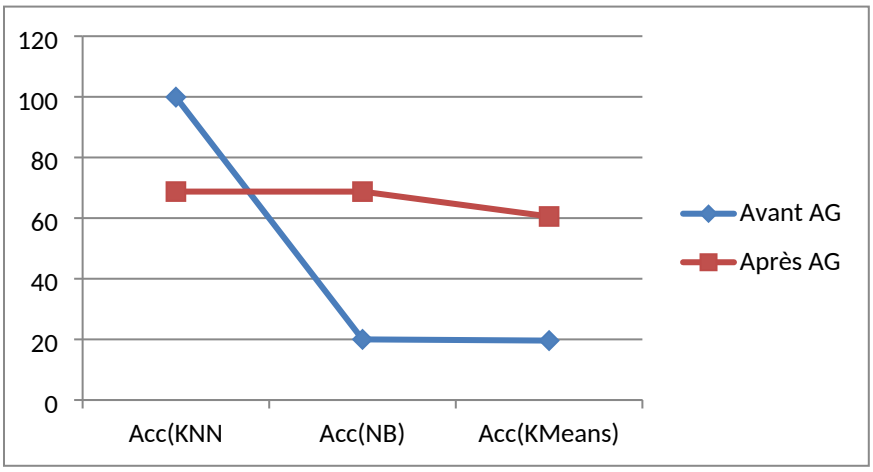



**FIGURE 4-17 : Comparaison avant et après l'AG par rapport à l'Accuracy.**

Les performances du KNN, la valeur de l'Accuracy est un peu moins que le premier cas, mais elle reste acceptable. L'Accuracy dans les NB et K-means est beaucoup mieux que le premier cas, on remarque une progression dans les performances de ces deux algorithmes.

## 5) Présentation de l'application :

Voici la page d'accueil de notre application :

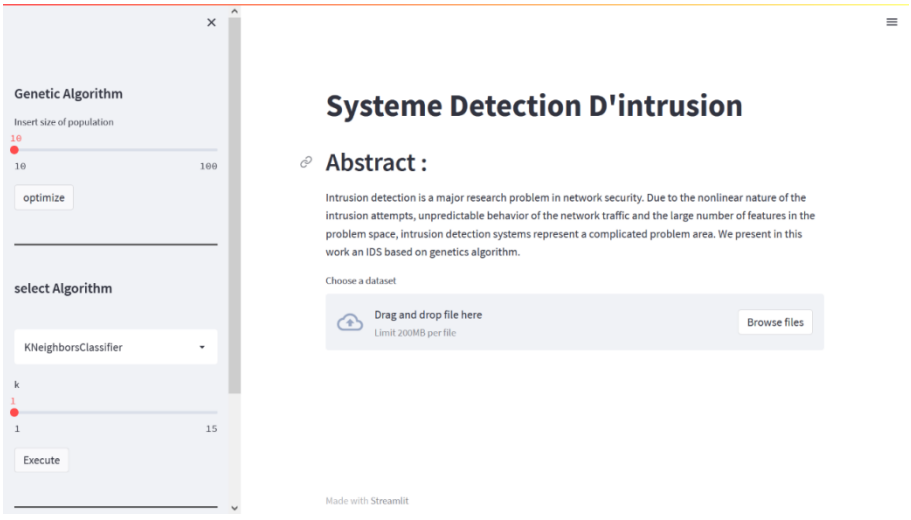



**FIGURE 4-18 : Page D'accueil.**

La page d'accueil présente l'entrée de notre application, au milieu de la page on trouve un résumé sur les IDS en général et à gauche on trouve les algorithmes utilisés.

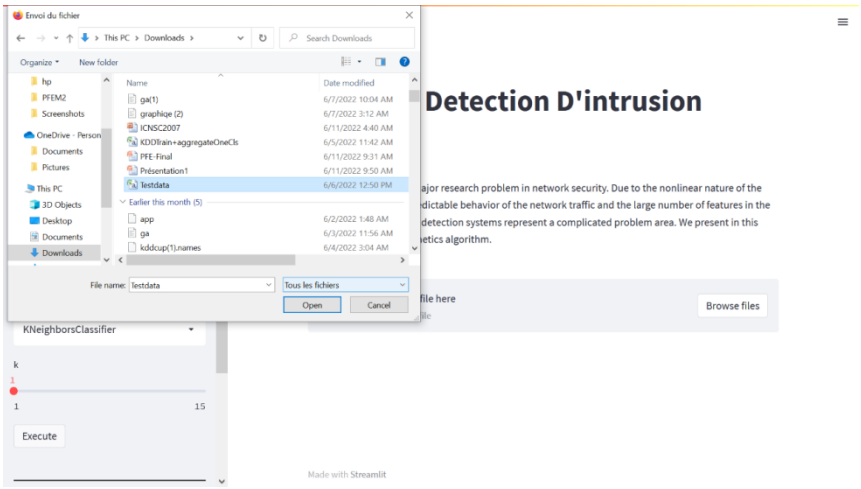



**FIGURE 4-19 : Charger le Dataset.**

Notre application donne la main aux utilisateurs de charger le dataset dans lequel elle fait l'apprentissage. Cette opération est faite en cliquant sur le bouton « Browse files ».

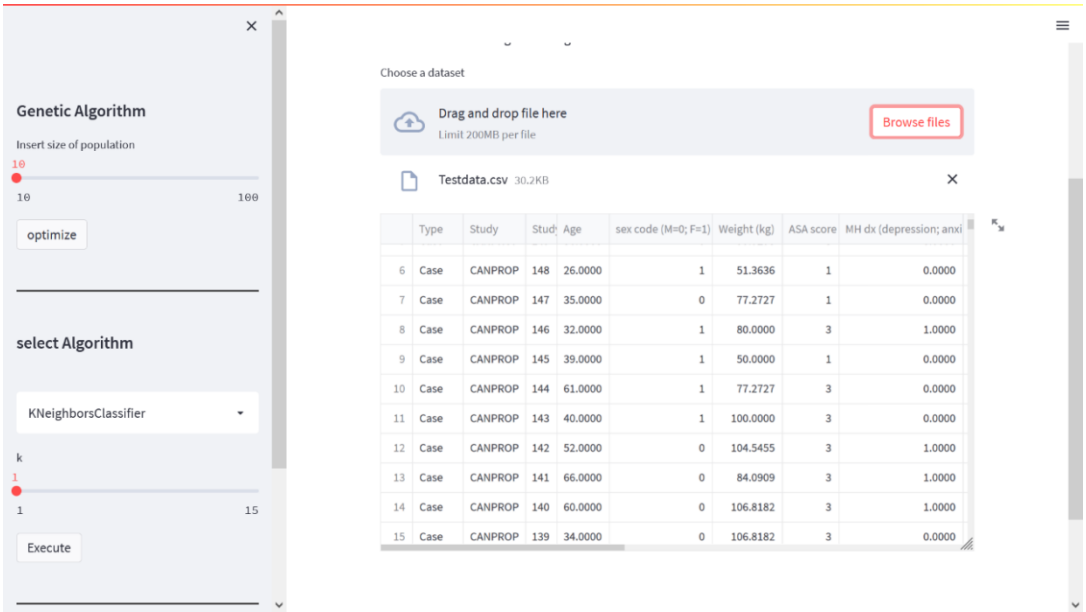



**FIGURE 4-20 : Exemple de chargement de dataset.**

Après le choix du fichier, il s'affiche au milieu de la table comme il est illustré dans la (FIGURE 4-20)

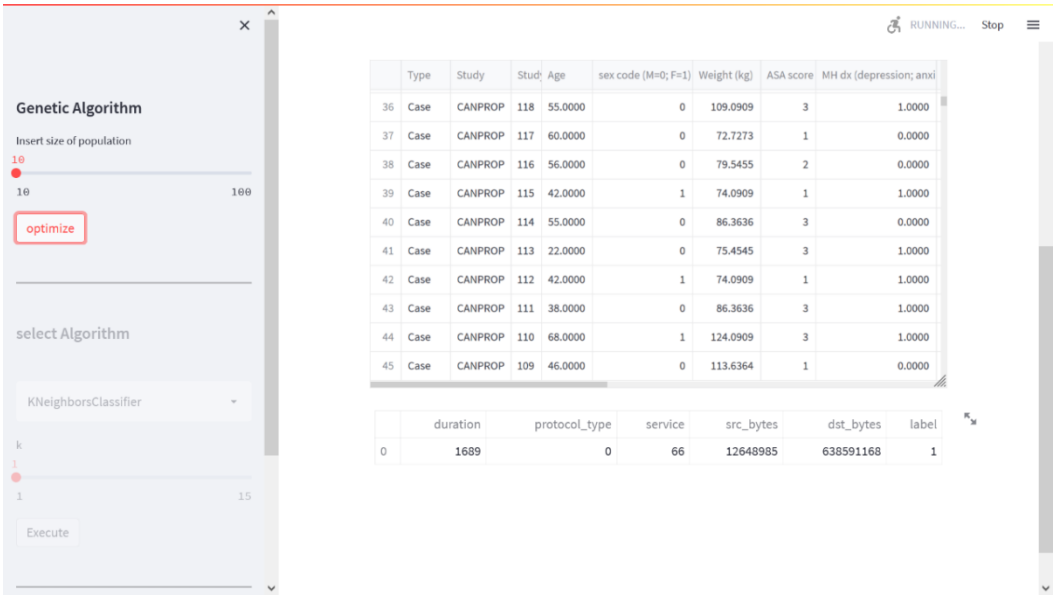



**FIGURE 4-21 : L'exécution de l'algorithme génétique.**

Dans le haut de la partie gauche de la page, on trouve une glissière (slider) pour déterminer le nombre des individus de la population initiale pour l'algorithme génétique.

En cliquant sur ce bouton, l'algorithme génétique s'applique sur le dataset chargé.

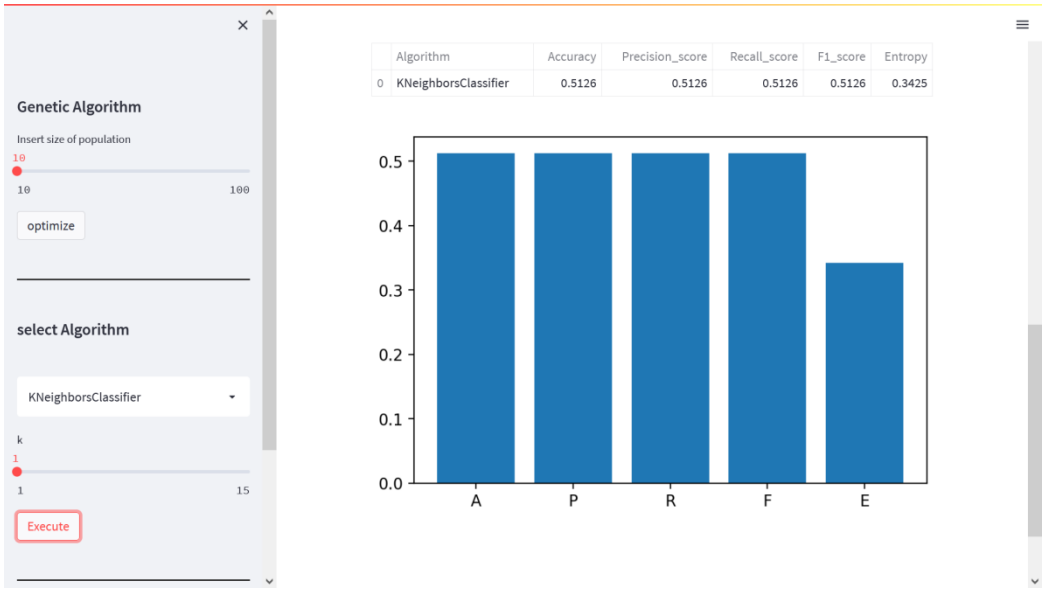



**FIGURE 4-22 : L'affichage de l'exécution du KNN.**

Dans le bas du coté gauche, on trouve les algorithmes du datamining, on choisit l'algorithme et la valeur des paramètres choisit en glissant le Slider. Pour lancer l'algorithme, on clique sur le bouton « Exécuter ». Au milieu de la plage s'affiche un tableau qui contient les mesures d'évaluation, et au dessous se trouve l'histogramme de ces mesures.

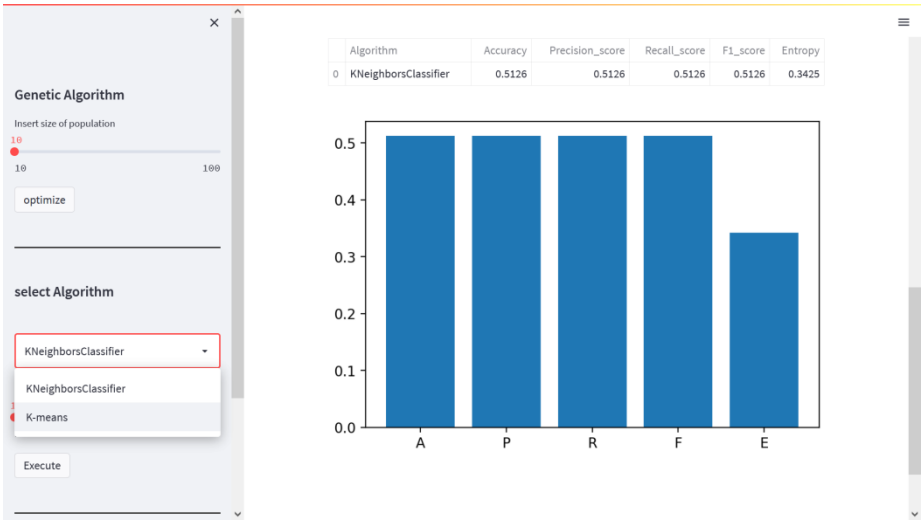



**FIGURE 4-23 : Changement de l'algorithme.**

Pour changer l'algorithme, on clique sur la flèche.
Dans le bas du coté gauche se trouve l'algorithme naive bayes :

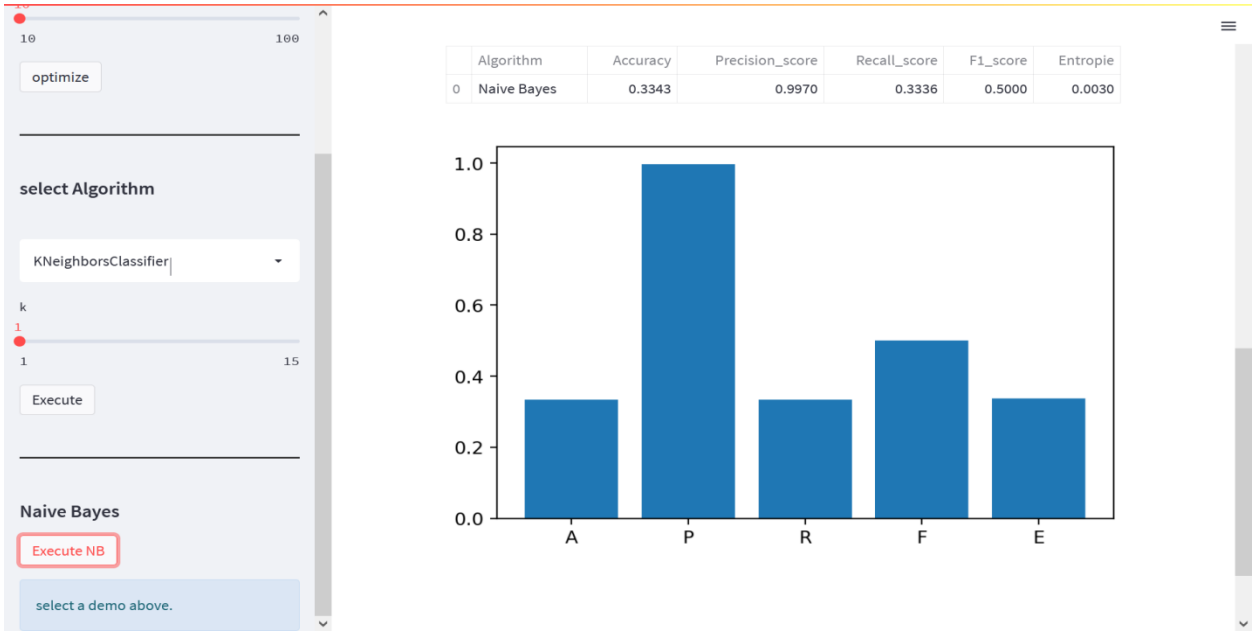



**FIGURE 4-24 : L'algorithme NB.**

## 5) Conclusion :

Dans ce dernier chapitre, nous avons présenté le dataset KDDCUP99 qui a été choisit comme une base de données d'apprentissage de notre modèle de détection qui se base sur les algorithmes génétiques.

En plus, nous avons parlé des outils utilisés pour réaliser ce travail et les différentes phases de développement, puis nous avons passé à la présentation des résultats obtenus pour chaque algorithme appliqué.

D'après ce qu'on a vu, les algorithmes génétiques ont prouvé leur efficacité dans le modèle de détection, Ils permettent d'augmenter le taux de réussite des algorithmes du datamining après l'amélioration de la population et les données test du dataset dans deux algorithmes différents, et la valeur dans le dernier algorithme était très acceptable. La perte d'information est beaucoup moins avec les algorithmes génétiques.

Enfin, on peut dire que les algorithmes génétiques ont amélioré les performances des algorithmes du Data Mining.

# CONCLUSION GÉNÉRALE

## Conclusion générale :

Le développement informatique facilite énormément notre quotidien, les nouvelles technologies nous ont donné beaucoup d'avantages, mais en autre coté, on doit affronter quelques défis, la sécurité informatique est l'un de ces défis.

La sécurité informatique est un sujet sensible et un objectif principal pour toutes les entreprises, la mise en œuvre des moyens de sécurité devient une obligation pour protéger les données contre les différentes attaques informatiques.

Contrairement aux outils de sécurité traditionnels, Les systèmes de détection d'intrusion ont comme avantage la possibilité de détecter des nouvelles attaques.

Dans ce travail, nous avons proposé un système de détection d'intrusion basé sur une metaheuristique (les algorithmes génétiques).

Dans la première partie de ce travail, nous avons parlé de la sécurité informatique, des notions de base, les différentes attaques, et les moyens de protection.

Dans la deuxième partie, nous avons parlé des systèmes de détection d'intrusion l'un de nouveaux moyens de sécurité, les méthodologies, les types et d'autres détails.

Puis, nous avons présenté les algorithmes génétiques avec leurs différentes étapes.

Dans la dernière partie, nous avons fait l'évaluation de notre travail en utilisant les mesures et les algorithmes du Data Mining pour renforcer notre modèle.

Ce travail, nous a permis de mieux comprendre les algorithmes génétiques et les méthodes du data mining. Cette expérience nous pousse à continuer la recherche dans notre domaine de sécurité informatique.

[1]- Livre sécurité informatique Risques, Stratégies et solutions Echec au cyber-roi deuxième édition Didier Godart (p18).

[2] https://www.ivision.fr/mettre-en-place-une-politique-de-securite-informatique-les-bonnes-pratiques/ visité le 07/03/2022

[3] Site officiel de Microsoft, https://www.microsoft.com/fr-FR/msrc/definition-of-a-security-vulnerability visité le 07/03/2022

[4] https://www.cyberark.com/fr/what-is/malware/ visité le 07/03/2022

[5] https://www.unsimpleclic.com/les-risques-de-la-messagerie-electronique-et-comment-sen-proteger/ visité le 09/03/2022

[6] : https://www.unsimpleclic.com/les-risques-de-la-messagerie-electronique-et-comment-sen-proteger/ visité le 09/03/2022

[7] https://fr.gadget-info.com/difference-between-active visité le 09/03/2022

[8] https://www.securiteinfo.com/attaques/hacking/typesattaques.shtml visité le 10/02/2022

[9] https://web.maths.unsw.edu.au/~lafaye/CCM/attaques/attaques.html visité le 10/02/2022

[10] Jean-Olivier Gerphagnon∗, Marcelo Portes de Albuquerque† & M´arcio Portes de Albuquerque‡ Centro Brasileiro de Pesquisas F´ısicas – CBPF/CNPq Coordena¸c˜ao de Atividade T´ecnicas – CAT
Rua Dr. Xavier Sigaud 150 22290-180 Rio de Janeiro – RJ – Brazil

[11] Ali Sadiqui, sécurité des réseaux informatiques(p10).

[12] https://www.oodrive.com/fr/blog/securite/top-10-differents-types-cyberattaques/ visité le 14/03/2022

[13] https://www.futura-sciences.com/tech/definitions/informatique-antivirus-10999/ visité le 14/03/2022

[14] https://fr.theastrologypage.com/anti-spyware visité le 25/03/2022

[15 ] https://www.capterra.fr/directory/30996/anti-spam/software visité le 25/02/2022

[16] Mikaeil PIRIO, Linux Red Hat Fedora TCP/IP, les services réseaux,Chapitre 11 P418

[17] https://www.lemagit.fr/definition/Systeme-de-detection-dintrusions visité le 02/04/2022

[18] : Jason Andress, in The Basics of Information Security (Second Edition), 2014

[19] Rowayda, A. Sadek; M Sami, Soliman; Hagar, S Elsayed (novembre 2013). "Système de détection d'intrusion d'anomalie efficace basé sur un réseau de neurones avec indicateur variable et réduction approximative". Journal international des problèmes informatiques (IJCSI)

[20] : Jason Andress, in The Basics of Information Security, 2011

[21] : https://iotindustriel.com/glossaire-iiot/systeme-de-detection-dintrusion-ids/ visité le 06/04/2022

[22] :Marouane HACHIMI, Détection intelligente de brouillage dans les réseaux 5G,MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE AVEC MEMOIRE EN GÉNIE ÉLECTRIQUEM. Sc.(page 40-43)

[23] https://www.fortinet.com/resources/cyberglossary/snort visité le 06/04/2022

[24] http://www.bro-ids.org/people.html visité le 06/04/2022

[25] https://aide.github.io/ visité le 06/04/2022

[26] Unionpédia, https://fr.unionpedia.org/i/DarkSpy visité le 07/04/2022

[27] International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 8, August 2012) visité le 09/04/2022

[28] Boukhlouf Djemaa et Kazar Okba, « Intrusion Detection System: Hybrid Approach based Mobile Agent », International Conference on Education and e-Learning Innovations, 2012

[29] : http://www-igm.univ-mlv.fr/~dr/XPOSE2004/IDS/IDSSnort.html visité le 11/04/2022

[30] A b c Richardson, Stephen (2020-02-24). "Placement IDS - Sécurité CCIE" . Expert certifié Cisco . Récupéré le 2020-06-26 .

[31] Lehmann Guillaume ; http://lehmann.free.fr/RapportMain/node10.html visité le 19/04/2022

[32] : Vacca, John R. (2013-08-26). Sécurité du réseau et du système

[33] https://www.ssmi-controls.com/2020/12/benefits-of-intrusion-detection-systems

[34] Wikipédia sous licence CC-BY-SA 3.0. visité le 03/05/2022

[35] : Article Jean-Marc Alliot, Nicolas Durand March 14, 2005

[36]:MEHIDID Fadila, « les algorithmes génétiques » Mémoire de fin d'étude pour l'obtention du diplôme de Master de Mathématiques, IBN BADIS MOSTAGANEM 2013.

[37] : Jean-Marc Alliot, Nicolas Durand »les algorithmes génétiques

[38] : : Jean-Marc Alliot, Nicolas Durand »les algorithmes génétiques : http://pom.tls.cena.fr/GA/FAG/ag.html#AGREF1 visité le 05/05/2022

[39] : MEHIDID Fadila, « les algorithmes génétiques » Mémoire de fin d'étude pour l'obtention du diplôme de Master de Mathématiques, IBN BADIS MOSTAGANEM 2013.

[40] : https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3 visité le 08/05/2022

[41] : https://fr.acervolima.com/crossover-dans-l-algorithme-genetique/ visité le 10/05/2022

[42] : Alireza Shafiee, ... Ali Abbas, in Computer Aided Chemical Engineering, 2016

[43] : Article écrit par Avik_Dutta et traduit par Acervo Lima de Crossover in Genetic Algorithm.

[44] : http://igm.univ-mlv.fr/~dr/XPOSE2013/tleroux_genetic_algorithm/fonctionnement.html visité le 14/05/2022

[45] : BELBACHIR Assia DEAU Raphaël LENNE Renaud SNOUSSI Jihene La Programmation Génétique

[46] : Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS 39762

[47] : Xin-She Yang, in Nature-Inspired Optimization Algorithms (Second Edition), 2021

[48] : Mémo technique, Le Data Mining ; softsomputing.com

[49] : introduction to datamining with case studies Third edition; GK GUPTA Adjunct professor of Computer science Lonash University Clayton, Australia.

[50] : Data Mining et statistique décisionnelle, l'intelligence dans dans les bases de données , Stéphane TUFFERY Universités De Rennes.

[51] : Data mining classification algorithms: An overview, Saeed Ngmaldin Bardab 1, *, Tarig Mohamed Ahmed 2, 3, Tarig Abdalkarim Abdalfadil Mohammed 1

[52] : Khillar, S. (2020, October 22). Difference Between Data Mining Supervised and Unsupervised. Difference Between Similar Terms and Objects. http://www.differencebetween.net/technology/difference-between-data-mining-supervised-and-unsupervised/. visité le 05/06/2022

[53] : Intelligent Systems with Applications Volume 14, May 2022, 200073

[54] : Java T Point ; https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning visité le 07/06/2022

## Résumé

Devant le développement de l'informatique, l'évolution technologique et l'utilisation de l'internet dans l'échange des données et le transfert d'information, la sécurité informatique devient une nécessité et un objectif à réaliser pour les entreprises.
Et vu que de nouveaux types d'attaques sont apparus, de nouveau moyens de sécurité doivent être mise en œuvre pour essayer de protéger ces échanges, l'un de ces moyens est le système de détection d'intrusion.Les méthodes méta-heuristiques et plus précisément les algorithmes génétiques ont prouvé leur efficacité contre ce genre de problèmes.
Dans ce travail, nous essayons de réaliser un système de détection d'intrusion à la base des algorithmes génétiques en utilisant le dataset KDDCUP99 et les algorithmes du Data Mining pour faire l'évaluation de notre projet.



Mot clé: CIA- Déni de service- Algorithme génétique- KPPV- Data Mining- IDS- NIDS- sécurité- Classification.



## Abstract

In front of the development of data processing, the technological evolution and the use of the Internet in the exchange of the data and the transfer of information, computer security becomes a necessity and an objective to be carried out for the companies. And since new types of attacks have appeared, new security tools should be implemented to try to protect these exchanges. Meta-heuristic methods and more precisely genetic algorithms have proven their efficiency against this kind of problems.

In this work, we try to realize an intrusion detection system based on genetic algorithms using the KDDCUP99 dataset and data mining algorithms to evaluate our project.



Key Words : CIA- Denial-of-Service- genetic algorithm- KNN- Data Mining- IDS- NIDS- security- Classification .

**الملخص**

مع تطور تكنولوجيا المعلومات والتطور التكنولوجي واستخدام الإنترنت في تبادل البيانات ونقل المعلومات، أصبح أمن تكنولوجيا المعلومات ضرورة وهدفا يتعين على الشركات تحقيقهما

ومع ظهور أنواع جديدة من الهجمات، وجب تنفيذ وسائل أمنية جديدة لمحاولة حماية هذه التبادلات، وإحدى هذه الوسائل هي نظام الكشف عن التسلل. أثبتت الطرق الاستدلالية الفوقية والخوارزميات الجينية بدقة أكبر فعاليتها ضد هذه الأنواع من المشاكل..

في هذا العمل، نحاول إنشاء نظام للكشف عن التسلل يعتمد على الخوارزميات الجينية باستخدام مجموعة البيانات KDDCUP99. وخوارزميات تعدين البيانات لتقييم مشروعنا.

كلمات مفتاحية: الأمن-خوارزميات-الكشف عن التسلل-التصنيف