

THE USE OF MACHINE LEARNING AND DEEP LEARNING TO ENHANCE THE SECURITY IN THE IOT APPLICATIONS

MASTER 1 2ÈME ANNÉE EEEA

Rapport de stage

Réalisé par :
Yacoubi MOHAMED REDA

Encadré par :
M.Saadane RACHID

Table des matières

Chapitre 1 : Contexte général du projet	3
1 Introduction	4
2 Problématique	4
3 Description du déroulement du stage	4
4 Les outils utilisés	5
4.1 Scikit Learn	5
4.2 Pandas	5
5 Conclusion	6
Chapitre 2 : Familiarisation avec quelques procédés IoT	7
1 Introduction	8
2 Quelques définitions	8
2.1 L'Internet des objets (IoT)	8
2.2 Le cloud computing	9
2.3 Big Data (macro données)	10
3 Les standards de communication et protocoles de l'Iot	10
3.1 Les réseaux courte portée	10
3.2 Les réseaux longue portée	12
4 La sécurité IoT	15
4.1 Qu'est-ce que la sécurité des IoT ?	15
4.2 Les plus grands défis liés à la sécurité de l'IoT	16
4.3 Solutions pour sécuriser un objet connecté	17
5 L'Internet des Objets dans l'entreprise	17
5.1 Avantages de l'IoT dans les entreprises	17
5.2 Scénarios IoT dans des industries clés	18
Chapitre 3 : L'intelligence artificielle - Généralités	19
1 Introduction	20
2 L'intelligence artificielle	20
3 Le Machine Learning (ML)	20
3.1 Définition	20
3.2 Fonctionnement	21
3.3 Méthodes	21
3.4 Le machine learning au service de la cybersécurité	23

4 Les réseaux de neurones	24
4.1 Définition	24
5 Le Deep Learning (DL)	24
5.1 Définition	24
5.2 Fonctionnement	25
6 La cybersécurité	26
6.1 Définition	26
6.2 Domaines	26
6.3 Les cybermenaces	27
Chapitre 4 : Etude de cas	28
1 Introduction	29
2 Jeu de données	29
3 Sélection des variables	30
3.1 Définition	30
3.2 Résultats	32
4 La technique SMOTE	32
4.1 Définition	32
4.2 Résultats	33
5 La classification	34
5.1 Les réseaux bayésiens	34
5.1.1 L'apprentissage des classifieurs de réseaux bayésiens	38
5.1.2 Classifieur Naive Bayes (NB)	39
5.1.3 Classifieur de réseau bayésien naïf augmenté par arbre (TAN) .	40
5.1.4 Classifieur Chow-liu	41
5.2 Le classifieur Random Forest	42
5.3 Résultats	42
5.3.3 Classifieur NB	44
5.3.2 Classifieur TAN	44
5.3.2 Classifieur CHOW	45
5.3.4 Classifieur Random Forest	45
6 Conclusion	46
Conclusion	47
Bibliographie	48

Table des figures

1	Scikit-learn	5
2	Pandas	6
3	Objectifs de la sécurité IoT	16
4	Modèles d'apprentissage automatique	22
5	Réseau de neurones et deep learning	24
6	Graphe d'un réseau bayésien	25
7	Algorithmee de sélcction de vaiables	31
8	Comparaison Sans/ Avec SMOTE	33
9	Nombre des échantillons initiale du jeu de données	33
10	Nombre d'échantillons des classes après application de SMOTE	34
11	Graphe d'un réseau bayésien	35
12	Relation entre noeud	35
13	table de probabilités conditionnelles	36
14	Graphe de dépendance entre les noeuds	37
15	Graphe de dépendance entre les noeuds	37
16	Graphe de dépendance entre les noeuds	38
17	classifieur Naïve Bayes	39
18	Structure classifieur Naïve Bayes	40
19	Structure classifieur TAN	40
20	Structure classifieur Chow-liu	41
21	L'algorithme Random Forest et ses paramètres	42
22	Matrice de confusion	43
23	Precision et Recall	43

Remerciements

Avant d'entamer ce travail, j'adresse tout d'abord mes profondes gratitude au corps professoral et administratif de l'École nationale supérieure d'électrotechnique, d'électronique, d'informatique, d'hydraulique et des télécommunications (**ENSEEIHT**), pour leur formation et leur encadrement durant toute l'année.

Je tiens à remercier tout particulièrement mon encadrant Monsieur **SAADANE Rachid** pour m'avoir proposé un bon sujet dont j'ai beaucoup appris, ainsi pour m'avoir permis de m'épanouir dans ce stage par l'autonomie et la liberté dont j'ai pu disposer pour réaliser mes objectifs aussi que pour sa disponibilité malgré ses responsabilités nombreuses.

Enfin, je présente mes remerciements à tous ceux et celles qui ont contribué de près ou de loin à l'élaboration de ce travail.

Mohamed Reda YACOUBI

Abstract

The Internet of Things (IoT) connects millions of computing devices and paves the way for future technology. Therefore, the increasing reliance on the IoT requires a focus on its privacy and security issues. Implementing security through encryption, authentication, access control and communication security is the need of the hour. These needs can be better met with the use of machine learning (ML) and deep learning (DL) that can help achieve secure intelligent systems.

In our study we started by making a preliminary study in the topic of IoT, machine learning and deep learning systems. We then examine the security and privacy issues in IoT systems.

Thus, in the practical part of our work, we will make a case study to visualize the role of machine learning to improve the security and privacy property of an IoT system(**Embedded intrusion detection systems using feature selection and classification**).



Chapitre 1

Contexte général du projet

1 Introduction

L'Internet des objets (IoT) connecte des millions d'appareils informatiques et ouvre la voie à une technologie future. Par conséquent, l'augmentation de la dépendance à l'IoT nécessite de se concentrer sur ses problèmes de confidentialité et de sécurité. La mise en œuvre de la sécurité par le cryptage, l'authentification, le contrôle d'accès et la sécurité des communications est le besoin de l'heure. Ces besoins peuvent être mieux satisfaits avec l'utilisation de l'apprentissage automatique (ML) et de l'apprentissage en profondeur (DL) qui peuvent aider à réaliser des systèmes intelligents sécurisés.

2 Problématique

Au service de l'IoT, l'Intelligence Artificielle ainsi que ses sous-catégories peuvent apporter une meilleure surveillance et une sécurisation des réseaux. En combinant leurs forces, l'intelligence artificielle et l'internet des objets offrent plusieurs possibilités aux applications.

Séparément, l'intelligence artificielle et l'internet des objets sont des technologies qui permettent d'améliorer notre quotidien d'une manière révolutionnaire. Ensemble, ils offrent encore plus de puissance aux applications pour vivre dans un environnement technologique de pointe. L'IA peut rendre les infrastructures IoT plus intelligentes et elle permet de toujours garder le contrôle sur les réseaux.

L'IA est donc la solution de cybersécurité idéale pour les entreprises qui cherchent à prospérer aujourd'hui. Les professionnels de la sécurité ont besoin d'un soutien solide de la part de machines intelligentes et de technologies avancées telles que l'IA pour travailler avec succès et protéger les organisations contre les cyberattaques.

3 Description du déroulement du stage

Le stage s'est déroulé en 3 phases :

- **La première phase :** Comprendre l'Internet des objets et ses défis en matière de sécurité.
- **La deuxième phase :** Comprendre l'intelligence artificielle, ses 3 concepts incontournables(Machine Learning, Deep Learning et les réseaux de neurons) et son utilisation dans le domaine de la cybersécurité .
- **La troisième phase :** Faire une étude de cas afin de bien visualiser le rôle de l'apprentissage automatique pour améliorer la propriété de sécurité et de confidentialité d'un système IoT(Systèmes de détection d'intrusion embarqués).

4 Les outils utilisés

4.1 Scikit Learn

Dans ce projet j'ai utilisé la bibliothèque SKlearn, Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs, notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria.

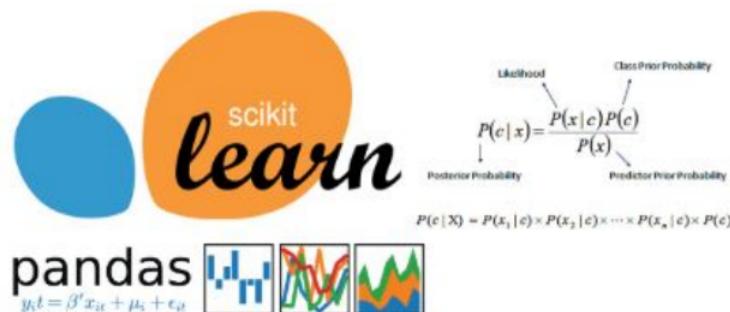


FIGURE 1 – Scikit-learn

Parmi les outils qui offrent cette bibliothèque :

- - **Classification** : Identifier la catégorie à laquelle appartient un objet.
- - **Regression** : Prévision d'un attribut à valeur continue associé à un objet.
- - **Clustering** : Regroupement automatique d'objets similaires en ensembles.
- - **Dimensionality reduction** : Réduire le nombre de variables aléatoires à prendre en compte.
- - **Model selection** : Comparer, valider et choisir des paramètres et des modèles.
- - **Preprocessing** : Extraction et normalisation des caractéristiques.

4.2 Pandas

La bibliothèque logicielle open-source **Pandas** est spécifiquement conçue pour la manipulation et l'analyse de données en langage **Python**. Elle est à la fois performante, flexible et simple d'utilisation.

Grâce à Pandas, le langage Python permet enfin de charger, d'aligner, de manipuler ou encore de fusionner des données. Les performances sont particulièrement impressionnantes quand le code source back-end est écrit en **C** ou en **Python**.

Le nom « **Panda** » est en fait la contraction du terme « **Panel Data** » , désignant les ensembles de données incluant des observations sur de multiples périodes temporelles. Cette bibliothèque a été créée comme un outil de haut niveau pour l'analyse en Python.

Les créateurs de Pandas comptent faire évoluer cette bibliothèque pour qu'elle devienne l'outil d'analyse et de manipulation de données open-source le plus puissant et flexible dans n'importe quel langage de programmation.

Outre l'analyse de données, Pandas est très utilisé pour le « **Data Wrangling** ». Ce terme englobe les méthodes permettant de transformer les données non structurées afin de les rendre exploitables.

De manière générale, Pandas excelle aussi pour traiter les données structurées sous forme de tableaux, de matrices ou de séries temporelles. Il est également compatible avec d'autres bibliothèques Python.

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```



FIGURE 2 – Pandas

5 Conclusion

Dans ce chapitre introductif, il était question de présenter le contexte général du projet, en commençant par une présentation générale du projet à étudier, pour ensuite passer à une présentation des objectifs et finir par une simple présentation des outils utilisés. Les deux chapitres suivants sont consacrés à la description de l'état d'art et de la présentation des différents concepts en relation avec la problématique.



Chapitre 2 :

Familiarisation avec quelques procédés IoT

1 Introduction

De plus en plus, le monde est régi par des systèmes informatiques et des objets en réseau. Cet internet des objets (IoT) est présent dans la quasi-totalité des aspects de notre vie. Connectés à Internet, ces systèmes sont confrontés à un fort risque de menace.

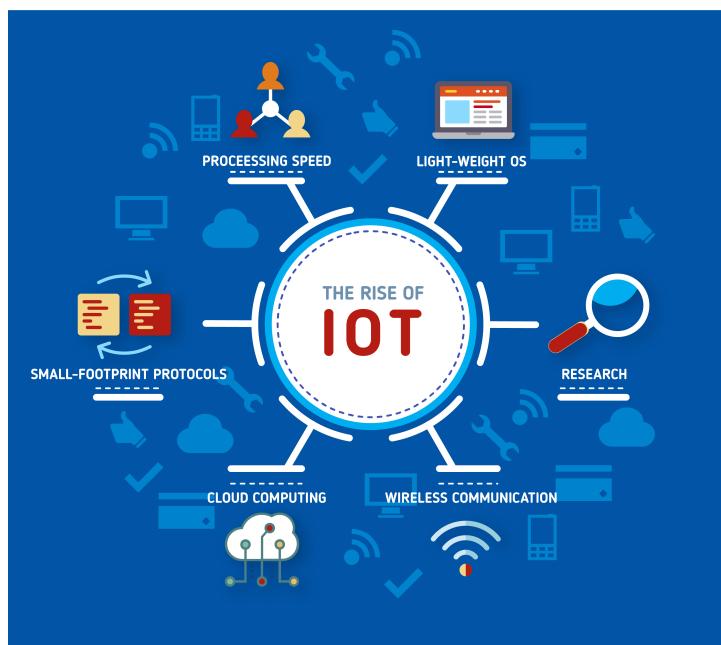
2 Quelques définitions

2.1 L'Internet des objets (IoT)

L'Internet des objets (IoT) décrit le réseau de terminaux physiques, les « objets », qui intègrent des capteurs, des softwares et d'autres technologies en vue de se connecter à d'autres terminaux et systèmes sur Internet et d'échanger des données avec eux.

L'IoT est donc un réseau de réseaux qui permet, via des systèmes d'identification électronique normalisés et unifiés, et des dispositifs mobiles sans fil, d'identifier directement et sans ambiguïté des entités numériques et des objets physiques et ainsi de pouvoir récupérer, stocker, transférer et traiter, sans discontinuité entre les mondes physiques et virtuels, les données s'y rattachant.

Tout objet qui peut se connecter à un réseau ouvert sur Internet est potentiellement un objet connecté. C'est dans ces usages qu'il trouve son utilité, usages définis par les programmes embarqués, les algorithmes, ou par des solutions déportées sur des serveurs (dans le cloud) qui reçoivent des informations venant des objets, des capteurs par exemple, les stockent, les analysent, les traitent, et automatisent éventuellement des actions qui sont renvoyées vers les objets.



2.3 Big Data (macro données)

Ce concept couvre les infrastructures, les technologies et les services qui apportent des solutions au traitement de grands ensembles de données qui, en raison de leur quantité, dépassent les capacités des outils des logiciels habituels. L'analyse de ces données permet également de configurer les situations futures susceptibles d'influencer la prise de décision.



3 Les standards de communication et protocoles de l'Iot

3.1 Les réseaux courte portée

➤ Wifi



À l'heure actuelle, la 802.11n s'impose comme la norme Wi-Fi la plus utilisée dans le contexte privé et professionnel. Cette norme offre un débit élevé, de l'ordre de centaines de mégabit par seconde, idéal pour les transferts de fichiers, mais peut-être trop énergivore pour la plupart des applications IIoT. RS propose une série de kits de développement RF conçus pour la création d'applications Wi-Fi.

- ◆ **Norme :** basée sur 802.11n (actuellement la norme la plus utilisée pour un usage privé)
- ◆ **Fréquences :** bandes de 2,4 GHz et 5 GHz
- ◆ **Portée :** environ 50m
- ◆ **Vitesses de transmission :** 600 Mbit/s maximum, mais les vitesses habituelles sont plus proches de 150 Mbit/s, en fonction de la fréquence de canal utilisée et du nombre d'antennes (la dernière norme 802.11-ac devrait permettre des vitesses pouvant atteindre 500 Mbit/s à 1 Gbit/s)

quelques bits par seconde à 100 Kbit/s sur la même liaison simple ; par ailleurs il peut résister jusqu'à 10 à 15 ans sur le terrain.

- ◆ **Norme :** Neul
- ◆ **Fréquences :** 900 MHz (ISM), 458 MHz (UK), 470-790 MHz (White Space)
- ◆ **Portée :** 10 km
- ◆ **Vitesses de transmission :** de quelques bit/s à 100 Kbit/s
- **SIGFOX**



Autre technologie à longue portée, Sigfox s'insère entre les technologies Wi-Fi et cellulaire en termes de portée. Elle utilise les bandes ISM, qui peuvent être utilisées gratuitement sans licences, pour transmettre des données sur un spectre très étroit, à partir et à destination d'objets connectés. La portée du Wi-Fi est trop courte, tandis que celle de la technologie cellulaire est trop coûteuse et énergivore. Ce protocole peut aussi offrir une durée de veille type de 20 ans avec une batterie 2,5 Ah, contre 0,2 an seulement pour la communication cellulaire.

Déjà activé pour des dizaines de milliers d'objets connectés, le réseau est en cours de déploiement dans les grandes villes d'Europe, dont dix au Royaume-Uni. Associant robustesse, efficacité et évolutivité, le réseau permet de connecter des millions de circuits alimentés par batterie sur des zones de plusieurs kilomètres carrés. Cette distance est idéale pour différentes applications M2M, qui devraient comprendre les objets suivants : compteurs intelligents, moniteurs de patients, dispositifs de sécurité, éclairage de rue et capteurs d'environnement.

- ◆ **Norme :** Sigfox
- ◆ **Fréquences :** 900 MHz
- ◆ **Portée :** 30-50km (environnements ruraux), 3-10km (environnements urbains)
- ◆ **Vitesses de transmission :** 10-1 000 bit/s
- **NB-IOT (Narrow Band IOT)**



Le NB-IoT (Narrow Band Internet of Things) est une solution standardisée se basant sur l'infrastructure de la 4G. Elle utilise une bande étroite de 180 KHz et assure une pénétration à l'intérieur des bâtiments ou en sous-sol, ce qui lui permet de répondre aux besoins de parcs importants d'appareils fixes nécessitant un faible volume de données, comme la télé-relève de compteurs d'eau ou de gaz. Vodafone et Huawei comptent parmi les principaux promoteurs du NB-IoT. En France, SFR a été le premier à lancer une offre NB-IoT, qui s'appuiera sur près de 18 000 antennes 800 MHz du réseau 4G. Selon la GSMA, 68 réseaux NB-IoT sont opérés dans le monde.

Le débit de transmission de données du NB-IoT se limite à 150 Kbits/s et le déploiement de ce standard dépend des opérateurs mobiles et de leurs capacités à faire évoluer leurs stations de base.

2 goals of IoT security

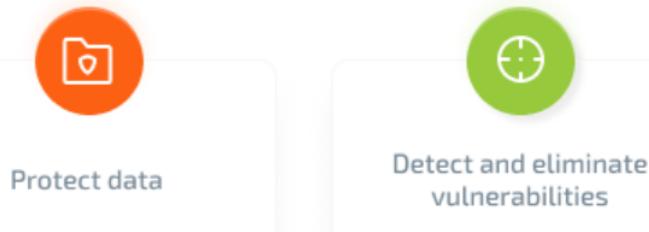


FIGURE 3 – Objectifs de la sécurité IoT

4.2 Les plus grands défis liés à la sécurité de l'IoT

Pour comprendre comment sécuriser les systèmes IoT, il est essentiel d'explorer d'abord les risques potentiels de cybersécurité. Voici une liste des défis les plus courant de la sécurité des IoT :

- **Faible puissance de traitement** : La plupart des applications IoT nécessitent extrêmement peu de données. Cela réduit les coûts et augmente la durée de vie de la batterie, mais cela rend les mises à jour OTA (Over-the-Air) difficiles et empêche l'appareil d'utiliser des outils de cybersécurité tels que les pare-feu, les antivirus et le cryptage de bout en bout.
- **Actifs traditionnels** : Si une application n'était pas établi avec la connectivité cloud à l'esprit, il est le plus susceptible d'être sensible aux cyberattaques contemporaines. Ces actifs plus anciens, par exemple, peuvent ne pas adhérer aux normes de cryptage de fichiers plus actuelles.
- **Manque de cryptage** : L'absence de cryptage de fichiers sur les transmissions régulières est l'un des problèmes de sécurité IoT les plus importants. De nombreux appareils IoT ne sécurisent pas les informations qu'ils transfèrent, ce qui implique que si quelqu'un s'introduit dans le réseau, il peut capturer les mots de passe et autres détails délicats envoyés vers et depuis le gadget.
- **Communications non sécurisées** : La plupart des mécanismes de sécurité existants ont été initialement conçus pour les ordinateurs de bureau et sont difficiles à mettre en œuvre sur des appareils IoT à ressources limitées. C'est pourquoi les mesures de sécurité traditionnelles ne sont pas aussi efficaces lorsqu'il s'agit de protéger la communication des appareils IoT.
- **Manque de sensibilisation des utilisateurs** : Du fait que l'IoT est une toute nouvelle innovation, de nombreuses personnes ne connaissent pas ses principes et ses fonctionnalités. En conséquence, les fabricants, les consommateurs et les entreprises peuvent présenter des dangers de sécurité importants dans les gadgets IoT.

4.3 Solutions pour sécuriser un objet connecté

Avec la prise de conscience des menaces et des conséquences que le monde de l’IoT peut engendrer, la notion de sécurité de l’Internet des Objets s’impose de plus en plus comme un prérequis à son développement.

Des mesures et normes de sécurité sont peu à peu imposées pour cadrer la conception et l’utilisation de ces objets connectés afin de diminuer les principales vulnérabilités des appareils et les risques de piratage.

Suivre un processus de développement logiciel s’appuyant sur les quatre étapes suivantes permet d’améliorer très nettement la sécurité, la fiabilité et la qualité des produits connecté :

- **Conception guidée par la Sécurité (Secure by Design)** : pour les objets connectés, la sécurité doit absolument guider toutes les phases de développement.
- **Evaluation des menaces au niveau Système** : pour les objets connectés faisant partie d’une infrastructure IoT plus vaste, une évaluation et une analyse des menaces et des vecteurs d’attaques au niveau système global est essentielle. Cette étape permet ainsi d’intégrer les besoins et les réponses pertinentes apportées dans le suivi des exigences du projet.
- **Automatisation des tests et analyses** : la sécurité ajoute des tâches supplémentaires aux équipes de développement. L’automatisation de celles-ci et l’utilisation de techniques spécifiques comme l’analyse statique avancée de code sont des éléments clés pour atteindre les objectifs, en réduisant la dépendance aux interprétations ou attentions des développeurs. De la même façon, une analyse statique spécifique doit être réalisée en cas d’utilisation de codes open source.
- **Analyse des binaires pour les code tierces-parties** : le recours à des logiciels tiers et notamment des librairies binaires s’accentue. S’assurer de la qualité et la sécurité de ses briques logicielles, même si on ne dispose pas des sources, est primordial.

5 L’Internet des Objets dans l’entreprise

5.1 Avantages de l’IoT dans les entreprises

L’Internet des Objets (IoT) a le potentiel de transformer les entreprises, en changeant en profondeur la manière dont les organisations recueillent des données et des informations en réunissant les principales tendances techniques et commerciales liées à la mobilité, l’automatisation et l’analyse de données. L’IoT consiste à la mise en réseau d’objets physiques qui, via l’utilisation de capteurs, d’actionneurs et d’autres dispositifs, peuvent collecter et transmettre des informations sur l’activité en temps réel au sein du réseau. L’ensemble des données issues de ces équipements peuvent ensuite être analysées par l’organisation pour :

- Optimiser les produits et les processus, en réduisant les coûts d'exploitation, en augmentant la productivité et le développement des nouveaux produits et services.
- En savoir plus sur les attentes et les préférences de la clientèle, permettant aux entreprises d'offrir des produits et services mieux personnalisés.
- Rendre les entreprises plus intelligentes et plus efficaces, en surveillant de manière proactive l'infrastructure critique et en permettant de créer des processus plus efficaces.
- Améliorer les expériences des utilisateurs, en offrant des produits et services améliorés pour permettre à une entreprise gérant des données de se différencier de la compétition.

5.2 Scénarios IoT dans des industries clés

Les solutions IoT promettent de rendre les organisations plus intelligentes et d'avoir plus de réussite dans ce qu'elles réalisent. Ces améliorations peuvent être particulièrement remarquables dans certains domaines :

- **Santé** : L'IoT a le potentiel de redéfinir comment les personnes, la technologie et les équipements interagissent et se connectent les uns aux autres dans l'environnement des soins de santé, aidant à promouvoir de meilleurs soins, à réduire les coûts et à améliorer les résultats.
- **Éducation** : L'IoT change à la fois les expériences d'apprentissage et d'enseignement dans les salles de classe connectées, améliorant la façon dont les écoles et les campus peuvent surveiller le fonctionnement et la sécurité à la fois dans l'éducation primaire et secondaire.
- **Gouvernement** : L'IoT fournit aux agences gouvernementales la possibilité d'assurer des services de qualité supérieure, de rationaliser les processus, de réduire les coûts et de trouver des façons innovantes d'apporter de la valeur ajoutée aux citoyens.
- **Transport** : L'IoT est au cœur de la stratégie globale de transformation du transport de façon à assurer une plus grande sécurité, des voyages plus efficaces, une maintenance améliorée des véhicules et des avions et une gestion stratégique du trafic.



Chapitre 3

L'intelligence artificielle - Généralités

1 Introduction

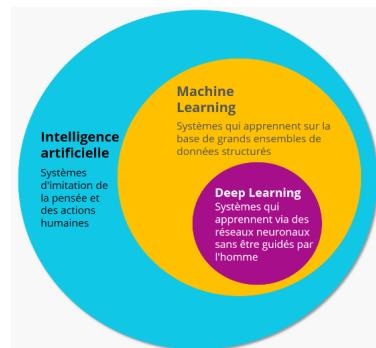
Aujourd’hui, l’analyse de données représente un facteur clé dans la prise de décision des entreprises. Ces données nécessitent d’être pré-traitées et analysées en utilisant l’intelligence artificielle grâce à des méthodes de Machine Learning comme le Deep Learning. L’IA est ainsi devenue une réalité aux multiples applications quotidiennes dont le nombre ne fera qu’augmenter dans les années à venir.

2 L’intelligence artificielle

L’**intelligence artificielle (IA)** est à la fois la théorie et le développement concret de machines, systèmes et logiciels qui imitent l’intelligence humaine pour accomplir des tâches très évoluées.

L’IA est basée sur une démarche d’apprentissage afin de reproduire une partie de l’intelligence humaine à travers une application, un système ou un processus. La reconnaissance de la parole, la perception visuelle et la traduction linguistique sont des exemples de systèmes d’intelligence artificielle.

Le **machine learning** et le **deep learning** sont des sous-ensembles de l’intelligence artificielle.



3 Le Machine Learning (ML)

3.1 Définition

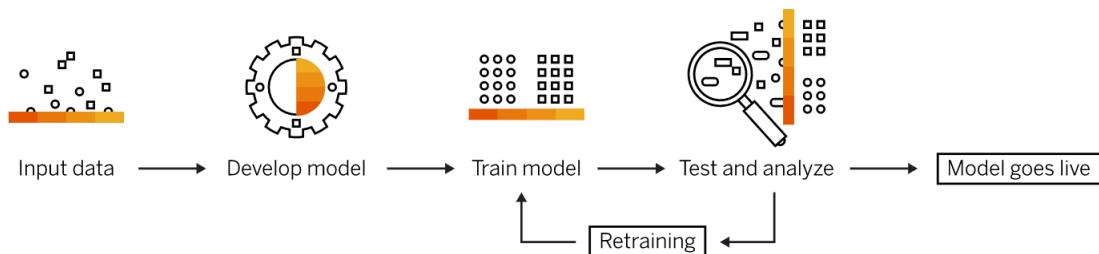
Le **Machine Learning** a été défini par son pionnier Arthur Samuel en 1959 comme le « champ d’étude qui donne aux ordinateurs la capacité d’apprendre sans être explicitement programmés à apprendre ».

En effet, c’est une approche fondée sur des analyses statistiques permettant aux ordinateurs d’améliorer leurs performances à partir de données, et à résoudre des tâches sans être explicitement programmées pour celles-ci. En fonction de la présence ou non des cibles, les apprentissages peuvent être classifiés en plusieurs types : supervisé, semi-supervisé, non-supervisé ou par renforcement.

3.2 Fonctionnement

Un algorithme d'apprentissage automatique est composé de trois parties principales :

1. **Un processus de décision** : en général, les algorithmes d'apprentissage automatique sont utilisés pour faire une prédiction ou une classification. Sur la base de certaines données d'entrée, qui peuvent être étiquetées ou non, l'algorithme produira une estimation d'un modèle dans les données.
2. **Une fonction d'erreur** : une fonction d'erreur évalue la prédiction du modèle. S'il existe des exemples connus, une fonction d'erreur peut effectuer une comparaison pour évaluer la précision du modèle.
3. **Un processus d'optimisation du modèle** : si le modèle peut mieux s'adapter aux points de données de l'ensemble d'apprentissage, les poids sont ajustés pour réduire l'écart entre l'exemple connu et l'estimation du modèle. L'algorithme répétera ce processus « d'évaluation et d'optimisation », mettant à jour les pondérations de manière autonome jusqu'à ce qu'un seuil de précision soit atteint.



3.3 Méthodes

Les modèles d'apprentissage automatique se répartissent en quatre catégories principales :

1. **L'apprentissage supervisé** : est défini par son utilisation d'ensembles de données étiquetés pour former des algorithmes afin de classer les données ou de prédire les résultats avec précision. Au fur et à mesure que les données d'entrée sont introduites dans le modèle, le modèle ajuste ses pondérations jusqu'à ce qu'il ait été ajusté de manière appropriée. Cela se produit dans le cadre du processus de validation croisée pour s'assurer que le modèle évite le surajustement ou le sous-ajustement. L'apprentissage supervisé aide les organisations à résoudre une variété de problèmes du monde réel à grande échelle, tels que la classification du spam dans un dossier distinct de votre boîte de réception. Certaines méthodes utilisées dans l'apprentissage supervisé comprennent les réseaux de neurones, les baies naïves, la régression linéaire, la régression logistique, la forêt aléatoire et la machine à vecteurs de support (SVM).

2. **L'apprentissage non supervisé :** utilise des algorithmes d'apprentissage automatique pour analyser et regrouper des ensembles de données non étiquetés. Ces algorithmes découvrent des modèles ou des regroupements de données cachés sans intervention humaine. La capacité de cette méthode à découvrir les similitudes et les différences dans les informations la rend idéale pour l'analyse exploratoire des données, les stratégies de vente croisée, la segmentation de la clientèle et la reconnaissance d'images et de modèles. Il est également utilisé pour réduire le nombre de caractéristiques dans un modèle grâce au processus de réduction de la dimensionnalité. L'analyse en composantes principales (ACP) et la décomposition en valeurs singulières (SVD) sont deux approches courantes pour cela. D'autres algorithmes utilisés dans l'apprentissage non supervisé comprennent les réseaux de neurones, le clustering k-means et les méthodes de clustering probabiliste.
3. **L'apprentissage semi-supervisé :** offre un juste milieu entre l'apprentissage supervisé et non supervisé. Pendant la formation, il utilise un ensemble de données étiquetées plus petit pour guider la classification et l'extraction de caractéristiques à partir d'un ensemble de données plus grand et non étiqueté. L'apprentissage semi-supervisé peut résoudre le problème du manque de données étiquetées pour un algorithme d'apprentissage supervisé. Cela aide également s'il est trop coûteux d'étiqueter suffisamment de données.
4. **L'apprentissage par renforcement :** est un modèle d'apprentissage automatique similaire à l'apprentissage supervisé, mais l'algorithme n'est pas formé à l'aide d'échantillons de données. Ce modèle apprend au fur et à mesure en utilisant des essais et des erreurs. Une séquence de résultats positifs sera renforcée pour développer la meilleure recommandation ou politique pour un problème donné.

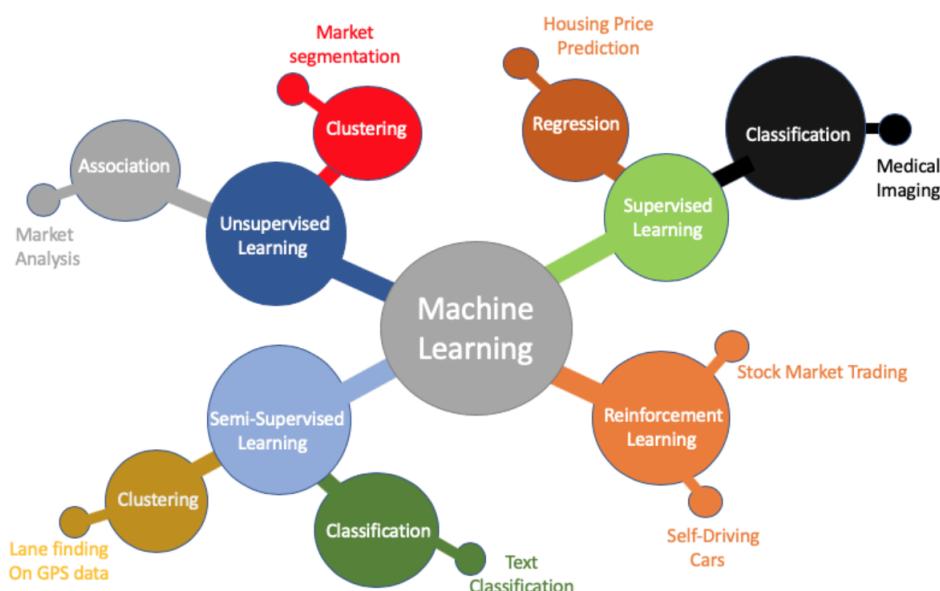


FIGURE 4 – Modèles d'apprentissage automatique

3.4 Le machine learning au service de la cybersécurité

L'apprentissage automatique est conçu pour se rapprocher des processus de l'esprit humain et permet aux ordinateurs d'analyser des informations, de prendre des décisions et d'apprendre des expériences passées.

Dans le domaine de la cybersécurité, les algorithmes d'apprentissage automatique aident les équipes de sécurité à gagner du temps en identifiant automatiquement les incidents et les menaces de sécurité, en les analysant et même en y répondant automatiquement dans certains cas. L'apprentissage automatique est intégré à de nombreux outils de sécurité modernes. Il remplace progressivement les anciennes méthodes d'inférence, telles que les règles définies manuellement et les corrélations statistiques.

Les algorithmes d'apprentissage automatique se présentent sous de nombreuses formes, mais la plupart d'entre eux effectuent l'une des trois tâches suivantes :

Machine Learning Task	How it Works	Cybersecurity Example
Regression	Cet algorithme identifie les corrélations entre différents ensembles de données et comprend comment et dans quelle mesure ils sont liés les uns aux autres.	La régression peut être utilisée pour prédire le prochain appel système d'un processus du système d'exploitation et le comparer à l'appel réel pour identifier les anomalies.
Classification	Habituellement effectué par des algorithmes d'apprentissage supervisé, qui "s'entraînent" sur un ensemble de données d'observations précédentes, et essaient d'appliquer ce qu'ils apprennent à de nouvelles données invisibles. Implique de prendre des artefacts, qui peuvent être du contenu textuel ou multimédia, et de les classer dans l'une de plusieurs étiquettes	La classification peut être utilisée pour classer un fichier binaire dans des catégories telles que les logiciels légitimes, les logiciels espions, les logiciels publicitaires et les logiciels de rançon.
Clustering	Généralement effectué par des algorithmes d'apprentissage non supervisés, qui travaillent directement sur de nouvelles données sans tenir compte des exemples précédents. Le regroupement consiste à identifier les points communs entre les artefacts et à les regrouper en fonction de leurs caractéristiques communes.	Le clustering peut être utilisé pour analyser les sessions de trafic et identifier les groupes de sessions pouvant provenir de la même source, afin d'identifier les attaques DDoS.

4 Les réseaux de neurones

4.1 Définition

Les réseaux de neurones artificiels sont une application du machine learning s'inspirant du cerveau humain. Lorsque le neurone artificiel reçoit une ou plusieurs données en entrée, elle calcule grâce à une fonction de combinaison une somme pondérée grâce aux variables d'entrée. De la même façon qu'il y a plus ou moins de connexions synaptiques entre certains neurones, le calcul sera pondéré par un poids de connexion. Une fonction d'activation (décris grâce à un seuil) permettra ensuite de définir une donnée en sortie.

Une phase d'apprentissage est bien sûr nécessaire afin que les poids et le seuil soit correctement configuré pour la prédiction.

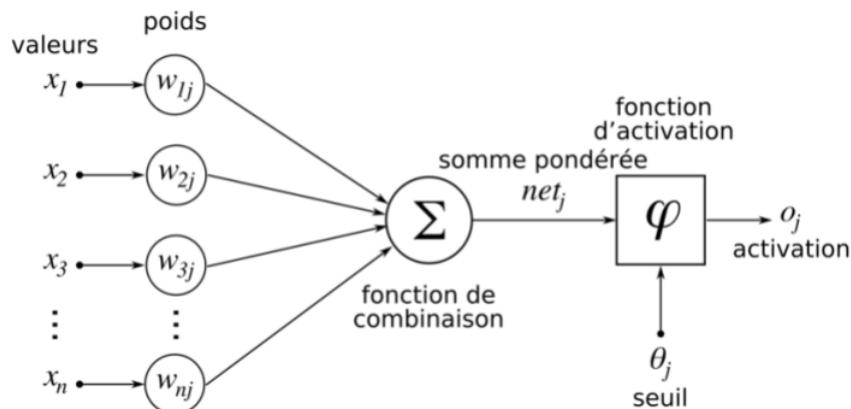


FIGURE 5 – Réseau de neurones et deep learning

Un neurone ou une couche de neurones ne suffisant pas à traiter les problèmes complexes, nous pouvons les coupler pour un apprentissage profond.

5 Le Deep Learning (DL)

5.1 Définition

Le **Deep learning**, sous-catégorie du Machine Learning, est une méthode d'apprentissage automatique qui s'inspire du fonctionnement du système nerveux des êtres vivants.

Les algorithmes du deep learning traitent l'information reçue de façon similaire à ce que feraient nos réseaux de neurones en réponse aux signaux nerveux qui leur sont destinés. En fonction du type et de la fréquence des messages reçus, certains réseaux de neurones vont se développer quantitativement et qualitativement alors que d'autres vont régresser.

5.2 Fonctionnement

Les réseaux de neurones d'apprentissage en profondeur(DL), ou réseaux de neurones artificiels, tentent d'imiter le cerveau humain grâce à une combinaison d'entrées de données, de pondérations et de biais. Ces éléments fonctionnent ensemble pour reconnaître, classer et décrire avec précision les objets dans les données.

Les réseaux de neurones profonds se composent de plusieurs couches de nœuds interconnectés, chacun s'appuyant sur la couche précédente pour affiner et optimiser la prédiction ou la catégorisation. Cette progression des calculs à travers le réseau est appelée propagation vers l'avant. Les couches d'entrée et de sortie d'un réseau neuronal profond sont appelées couches visibles. La couche d'entrée est l'endroit où le modèle d'apprentissage en profondeur ingère les données pour le traitement, et la couche de sortie est l'endroit où la prédiction ou la classification finale est effectuée.

Un autre processus appelé rétropropagation utilise des algorithmes, comme la descente de gradient, pour calculer les erreurs dans les prédictions, puis ajuste les poids et les biais de la fonction en reculant à travers les couches dans le but d'entraîner le modèle. La propagation directe et la rétropropagation permettent à un réseau de neurones de faire des prédictions et de corriger les erreurs.

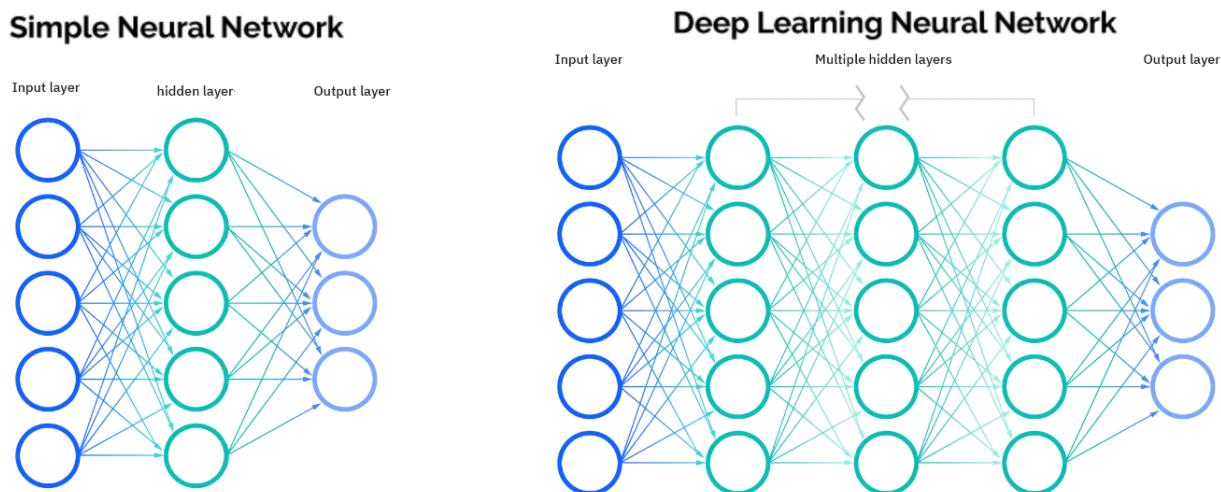


FIGURE 6 – Graphe d'un réseau bayésien

6 La cybersécurité

6.1 Définition

La cybersécurité, également connue sous le nom de sécurité des technologies de l'information (IT), est la pratique consistant à défendre les actifs numériques, y compris les réseaux, les systèmes, les ordinateurs et les données, contre les cyberattaques [2].

Bien que toute organisation ou tout individu puisse être la cible d'une cyberattaque, la cybersécurité est particulièrement importante pour les organisations qui travaillent avec des données ou des informations sensibles.

Afin de se protéger et de se défendre contre les attaques numériques, les organisations doivent développer et déployer une stratégie de sécurité complète qui comprend à la fois des mesures préventives, ainsi que des capacités de détection et de réponse rapides.

6.2 Domaines

Une stratégie de cybersécurité robuste comporte des couches de protection contre les cyberattaques qui tentent d'accéder à des données, de les modifier ou de les détruire, d'extorquer de l'argent aux utilisateurs ou à l'organisation, ou de perturber le fonctionnement de l'entreprise. Les contre-mesures doivent couvrir :

- ❖ **Sécurité de réseau :** les mesures de sécurité visant à protéger un réseau informatique contre les intrus, y compris les connexions filaires et sans fil (Wi-Fi).
- ❖ **Sécurité des applications :** Processus qui permettent de protéger les applications fonctionnant sur site et dans le cloud. La sécurité doit être intégrée aux applications dès leur conception, en tenant compte de la manière dont les données sont traitées, de l'authentification des utilisateurs, etc.
- ❖ **Sécurité du cloud :** Plus précisément, une véritable informatique confidentielle qui chiffre les données du cloud au repos (dans le stockage), en mouvement (lorsqu'elles se déplacent vers, à partir et dans le cloud) et en cours d'utilisation (pendant le traitement) pour le respect de la confidentialité des informations des clients, des exigences commerciales et des normes de conformité aux réglementations.
- ❖ **Sécurité des informations :** veille à garantir l'intégrité et la confidentialité des données, qu'elles soient stockées ou en transit.
- ❖ **Formation des utilisateurs finaux :** Sensibilisation de l'ensemble de l'organisation à la sécurité, afin de renforcer la sécurité des terminaux.
- ❖ **Planification de la reprise après sinistre/de la continuité des opérations :** outils et procédures permettant de réagir à des événements imprévus, tels que des catastrophes naturelles, des pannes de courant ou des incidents de cybersécurité.

Chapitre 4 : Etude de cas

Systèmes de détection d'intrusion embarqués

1 Introduction

L’industrie des systèmes aériens sans pilote (UAV) est devenue un immense terrain de jeu technologique dans le monde entier. Leur utilisation généralisée rend les UAV très populaires auprès des secteurs public et privé, par exemple : les forces armées, l’agriculture, les forces de l’ordre, les agences météorologiques et les services médicaux. Les fabricants de drones investissent massivement dans des systèmes intelligents de haute technologie allant des drones militaires de la taille d’un avion aux micro-drones.

Certes, l’utilisation dans presque tous les aspects de l’activité humaine a accru le besoin d’évolution des UAVs, mais elle a également accru les risques de sécurité. Un agresseur malveillant peut attaquer les drones via diverses méthodes pour accéder aux informations que le drone collecte ou que le drone contient.

En effet, il est nécessaire de développer un système de détection de ces intrusions en faisant une étude des potentielles attaques. Pour se faire, il faut développer un classifieur performant permettant de détecter les attaques.

2 Jeu de données

L’ensemble de données CICIDS2017 contient des attaques courantes bénignes de pointe, similaires à la vérité terrain (PCAP). Il inclut également les résultats d’analyse du trafic réseau à l’aide de CICFlowMeter qui marque les flux en fonction de l’horodatage, de l’adresse IP source et de destination, du port source et de destination, du protocole et de l’attaque (c’est un fichier CSV). Les définitions des fonctions d’extraction sont également disponibles [10].

La période de saisie des données a commencé à 9 h, le lundi 3 juillet 2017 et s’est terminée à 17 h, le vendredi 7 juillet 2017, pour un total de 5 jours. Les attaques mises en œuvre incluent Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet et DDoS. Ils ont été exécutés le matin et l’après-midi du mardi, mercredi, jeudi et vendredi.

Le jeu de données contient 3,1 millions d’enregistrements de flux et construit le comportement abstrait de 25 utilisateurs basé sur les protocoles HTTP, HTTPS, FTP, SSH et e-mail en se basant sur 78 variables. Il comprend les types d’attaque suivants : **Benign**, **DoS Hulk**, **Port Scan**, **DDoS**, **DoS**, **FTP Patator**, **SSH Patator**, **DoS**, etc. **Slow Loris**, **DoS Slow HTTP Test**, **Botnet**, **Web Attack** : **Brute Force**, **Web Attack** : **Infiltration**, **Attaque Web** : **SQL Injection** et **Heartbleed**.

Étant donné que ce jeu de données est conçu pour attaquer des serveurs, il faut l'adapter à l'environnement du drone. Par conséquence, trois attaques seront sélectionnées sur onze Attaque plus flux normal :

- **BENIGN** : Flux normal.
- **DOS** : Une attaque par déni de service (DoS) est une tactique pour surcharger une machine ou un réseau afin de le rendre indisponible. Les attaquants y parviennent en envoyant plus de trafic que la cible ne peut en gérer, ce qui entraîne son échec, ce qui l'empêche de fournir des services à ses utilisateurs normaux. Les exemples de cibles peuvent inclure les e-mails, les services bancaires en ligne, les sites Web ou tout autre service reposant sur un réseau ou un ordinateur ciblé.
Il existe différents types d'attaques DoS telles que l'épuisement des ressources et les attaques par inondation. Les attaques par épuisement des ressources obligent l'infrastructure ciblée à utiliser toutes ses ressources de mémoire ou de stockages disponibles, ralentissant les performances du service ou l'arrêtant complètement. Les attaques par inondation envoient un nombre écrasant de paquets qui dépassent la capacité du serveur.
- **DDOS** : Une attaque DDoS se produit lorsque plusieurs systèmes orchestrent une attaque DoS synchronisée vers une seule cible. La différence essentielle est qu'au lieu d'être attaquée depuis un seul endroit, la cible est attaquée depuis plusieurs endroits à la fois. La répartition des hôtes qui définit un DDoS offre à l'attaquant de multiples avantages :
 - Il peut tirer parti du plus grand volume de machine pour exécuter une attaque sérieusement perturbatrice.
 - Le lieu de l'attaque est difficile à détecter en raison de la répartition aléatoire des systèmes d'attaque (souvent dans le monde entier).
 - Il est plus difficile d'arrêter plusieurs machines qu'une seule.
 - La véritable partie attaquante est très difficile à identifier, car elle est déguisée derrière de nombreux systèmes.
- **Infiltration** : une attaque qui consiste à entrer et/ou endommager le système d'un utilisateur.

3 Sélection des variables

3.1 Définition

Certains problèmes de modélisation prédictive comportent un grand nombre de variables qui peuvent ralentir le développement et la formation des modèles et nécessitent une grande quantité de mémoire système. De plus, les performances de certains modèles peuvent se dégrader lors de l'inclusion de variables d'entrée qui ne sont pas pertinentes pour la variable cible.

3.2 Résultats

L'algorithme de sélection de variables a été appliqué en utilisant deux types de classifiants : Gaussian et Random Forest sur la partie entraînement du jeu de données (75% du jeu de données).

Jeu de données	Gaussian	Random Forest
Nombre de variables	36	52
Temps d'entraînement	16h	48h
Nombre de variables communes	27	29

4 La technique SMOTE

4.1 Définition

Dans une classification déséquilibrée, il y a trop peu d'exemples de la classe minoritaire pour qu'un modèle apprenne efficacement la frontière de décision.

Une façon de résoudre ce problème est de sur échantillonner les exemples de la classe minoritaire. Ceci peut être réalisé en dupliquant simplement des exemples de la classe minoritaire dans l'ensemble de données d'apprentissage avant d'ajuster un modèle. Cela peut équilibrer la distribution des classes mais ne fournit aucune information supplémentaire au modèle.

Une amélioration de la duplication d'exemples de la classe minoritaire consiste à synthétiser de nouveaux exemples de la classe minoritaire. Il s'agit d'un type d'augmentation de données pour les données tabulaires et peut être très efficace.

SMOTE se base sur le principe de l'algorithme d'apprentissage unsupervisée K-NN Neighborhood, il fonctionne en sélectionnant des exemples qui sont proches dans l'espace des caractéristiques, en traçant une ligne entre les exemples dans l'espace des caractéristiques et en dessinant un nouvel échantillon à un point le long de cette ligne [7].

Plus précisément, un exemple aléatoire de la classe minoritaire est d'abord choisi. Ensuite, on trouve k des voisins les plus proches de cet exemple (généralement $k=5$). Un voisin choisi au hasard est sélectionné et un exemple synthétique est créé à un point choisi au hasard entre les deux exemples dans l'espace des caractéristiques.

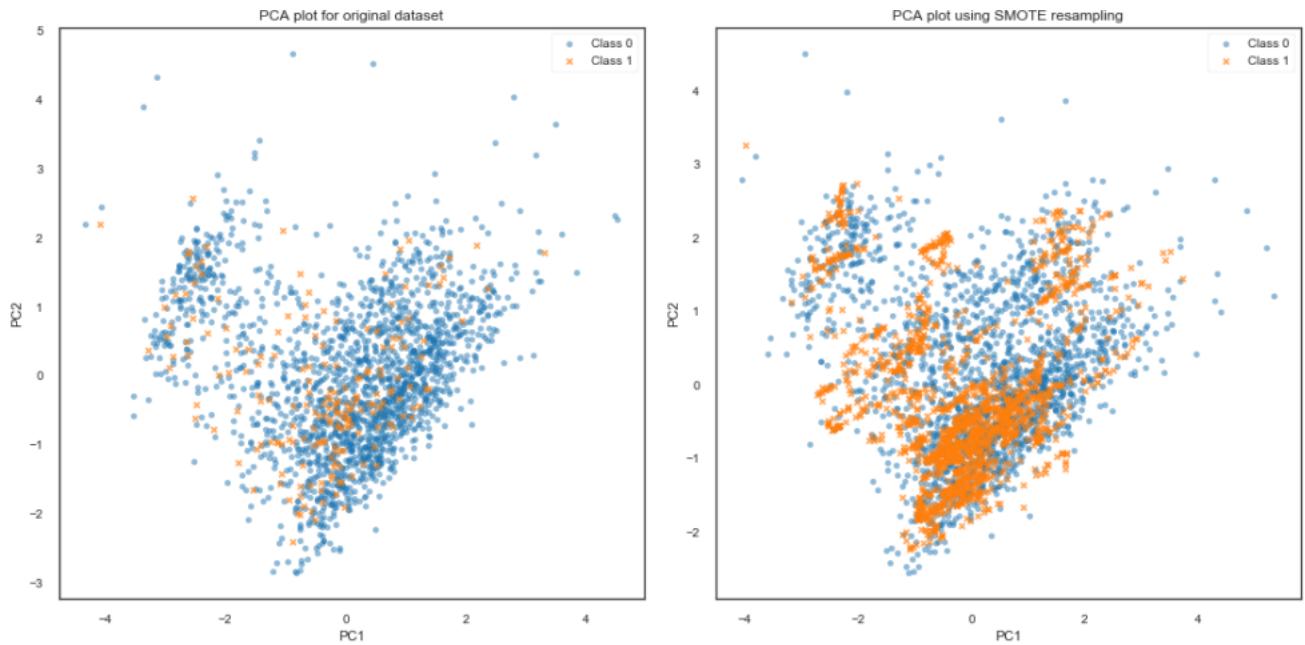


FIGURE 8 – Comparaison Sans/ Avec SMOTE

4.2 Résultats

Après avoir réduit le nombre de variable du jeu de donné en utilisant la sélection des variables, on a appliqué la technique SMOTE afin d'augmenté le nombre d'échantillons de la classe minoritaire Infiltration.

Le SMOTE a été utilisé pour augmenter le nombre des échantillons aux trois valeurs suivantes : 100000, 200000 et 300000 et il a été appliqué sur la partie entrainement du jeu de donnée (75% du jeu de données)

Le tableau suivant représente le nombre des échantillons initial ainsi que leurs pourcentages dans le jeu de données :

Label	Nombre d'échantillons	Répartition
BENIGN	1703163	85,62%
DDOS	96139	4,83%
DOS	189779	9,54%
Infiltration	25	0,0012%
Total	1989106	100%

FIGURE 9 – Nombre des échantillons initiale du jeu de données

- la causalité joue un rôle important (des événements en causent d'autres)
- notre compréhension de la causalité des événements est incomplète (on doit recourir aux probabilités)

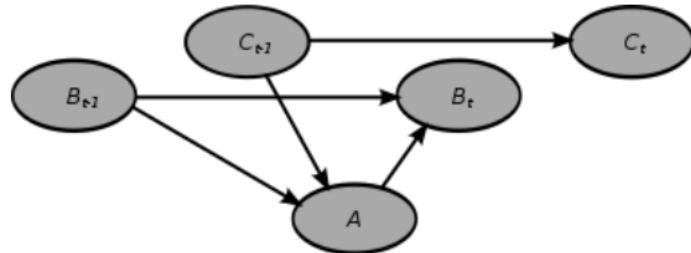


FIGURE 11 – Graphe d'un réseau bayésien

Un RB est un graphe :

- Orienté : chaque noeud il' a un successeur et un prédécesseur
- Acyclique : est un graphe ne contenant aucun cycle
- Les nœuds sont des variables aléatoires
- Les arcs représentent :
 - Des dépendances (par exemple des causalités) probabilistes entre les variables
 - Des distributions de probabilités conditionnelles (locales) pour chaque variable étant donnés ses parents

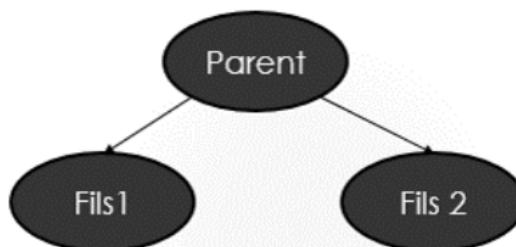


FIGURE 12 – Relation entre noeud

➤ Le calcul de probabilité dans un réseau bayésien

Au debut on commence par définir des notations :

- Une **table de probabilités conditionnelles** (TPC) donne la probabilité pour chaque valeur du nœud étant donnés les combinaisons des valeurs des parents du nœud (c'est l'équivalent d'une distribution).

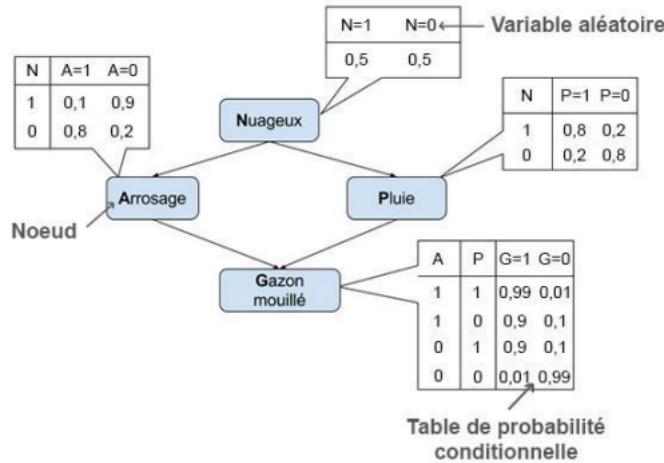


FIGURE 13 – table de probabilités conditionnelles

- S'il y a un arc d'un nœud X vers un nœud Y, cela signifie que la variable X influence la variable Y.
 - X est appelé le parent de Y.
 - Parents(X) est l'ensemble des parents de X.
- Si X n'a pas de parents, sa distribution de probabilités est dite **inconditionnelle ou a priori**.
- Si X a des parents, sa distribution de probabilités est dite **conditionnelle**.
- Si X est une variable observée, on dit que c'est une évidence.

Probabilités conjointes :

Un RB est une façon compacte de représenter des probabilités conjointes.

Par définition, la probabilité conjointe de X₁ et X₂ est donnée par la distribution P(X₁,X₂), pour une valeur donnée de X₁ et X₂ .

La probabilité conditionnelle de X₁ sachant X₂ est notée P(X₁| X₂) :

$$P(X_1, X_2) = P(X_1|X_2)P(X_2)$$

En d'autres mots, la distribution conjointe des variables d'un RB est définie comme étant le produit des distributions conditionnelles.

Probabilités conditionnelles :

Après définition de la probabilité conjointes et sa relation avec la probabilité conditionnelle, On peut alors calculer toute probabilité conditionnelle.

Parmi les avantages d'un RB est qu'il est très simple à identifier les indépendances conditionnelles, ceci pour but de réduire les calculs à faire.

Indépendance conditionnelle dans un RB :

L'indépendance est une notion probabiliste qualifiant de manière intuitive des événements aléatoires n'ayant aucune influence l'un sur l'autre. Il s'agit d'une notion très importante en statistique et en théorie des probabilités. Par exemple, la valeur d'un premier lancer de dé n'a aucune influence sur la valeur du second lancer.

Il existe trois types de définir une indépendance :

① 1er cas :

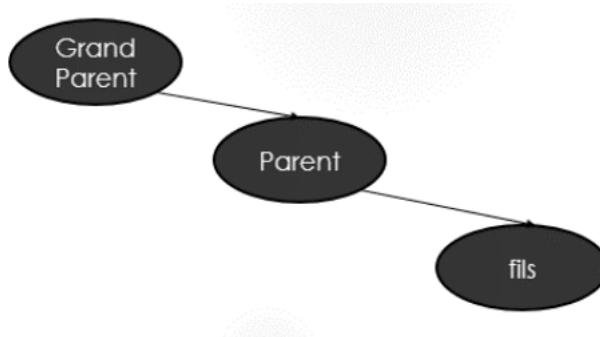


FIGURE 14 – Graphe de dépendance entre les noeuds

La règle : Relation entre **grand-parent** et **enfant** étant donné parent : sont indépendants si **parent** observé.

Le **fils** et le **grand-parent** sont dépendants a priori vu la loi du réseau bayesien, mais ils sont indépendants étant donné **parent**. Donc l'équation mathématique s'écrit sous forme :

$$P(F/GP, P) = P(F/P)$$

D'autre mot, si P est connu, GP n'intervient pas dans le calcul, alors connaître P bloque le chemin entre GP et F.

② 2ème cas :

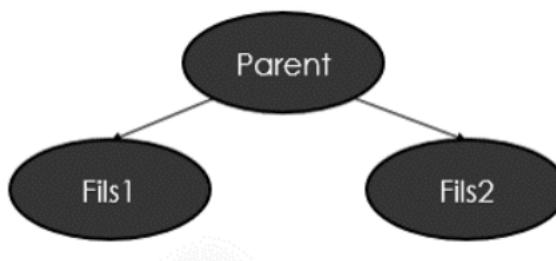


FIGURE 15 – Graphe de dépendance entre les noeuds

La relation entre deux enfants étant donné parent : sont indépendants si parent observé. Le **Fils1** et **Fils2** sont dépendants a priori, mais ils sont indépendants étant donné **parent**. Donc l'équation mathématique s'écrit sous forme :

$$P(F1/P, F2) = P(F1/P)$$

Pour résumer, Si **parent** est connu, **Fils2** n'intervient pas dans le calcul

③ 3ème cas :

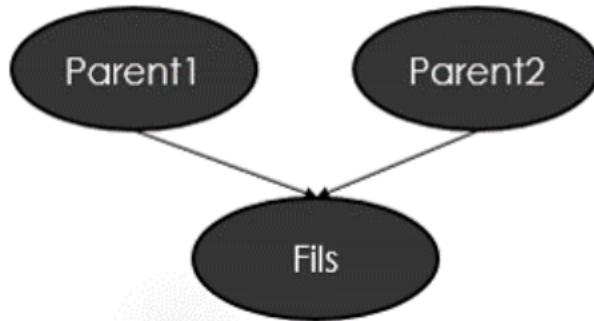


FIGURE 16 – Graphe de dépendance entre les noeuds

Relation entre deux parents étant donné enfant : sont indépendants si enfant non-observé. Le **Parent1** et **Parent2** sont dépendants a priori, mais ils sont indépendants étant donné **Fils**. $P(C|A, S)$ n'est pas simplifiable, parce que $P(A|C, S)$ n'est pas simplifiable. Pour simplifier ne pas connaître le fils « bloque » le chemin entre le Parent1 et Parent2.

Maintenant on vas présenter un outil qui simplifie la détection des indépendance entre les noeuds.

* Indépendance conditionnelle dans un RB : D-séparation :

D-séparation : critère général pour décider si un noeud X est indépendant d'un noeud Y, étant donnés d'autres noeuds $Z = Z_1, \dots, Z_m$. la règle pour que X soit indépendant de Y sachant Z si tous les chemins non-dirigés entre X et Y sont bloqués par Z, c'est-à-dire un chemin est bloqué s'il contient au moins un noeud N qui satisfait une ou l'autre des conditions suivantes :



ou

ou

où $N \in Z_1, \dots, Z_m$

où $N \in Z_1, \dots, Z_m$, ni aucun des descendants de N.

5.1.1 L'apprentissage des classifieurs de réseaux bayésiens

En pratique, les paramètres BN sont pour la plupart inconnus car les paramètres d'apprentissage sont appris à partir d'ensembles de données appelés problèmes d'apprentissage BN. Les problèmes d'apprentissage sont énoncés comme ayant une donnée d'entraînement et une information préalable (par exemple, une connaissance experte et une relation causale). Le BN a évalué la disposition du réseau de graphes et les paramètres de la distribution de probabilité conjointe dans le BN. Les structures graphiques du BN et l'apprentissage des paramètres sont d'une importance majeure.

Cependant, il y a deux manières de considérer un BN comme une approche d'apprentissage ; la première est l'apprentissage de l'arrangement des variables qui inclut la distribution conjointe des variables qui correspondent le mieux aux données et conduit à des algorithmes d'apprentissage basés sur la notation.

La seconde est l'arrangement BN qui inclut la relation d'indépendance conditionnelle entre les nœuds représentés dans les nœuds DAG selon le concept de séparation. Apprendre les arrangements de structure implique d'identifier les relations d'indépendance conditionnelle entre les attributs. Certains tests statistiques tels que le test du chi carré, le coefficient de corrélation (covariance) entre les attributs et le test d'information mutuelle ont été utilisés pour déterminer les relations indépendantes conditionnelles entre les nœuds d'attributs.

5.1.2 Classifieur Naïve Bayes (NB)

» Définition

En statistique, les classificateurs de Bayes naïfs sont une famille de "classificateurs probabilistes" simples basés sur l'application du théorème de Bayes avec des hypothèses fortes (naïves) d'indépendance entre les caractéristiques. Ils font partie des modèles de réseaux bayésiens les plus simples, mais couplés à l'estimation de la densité du noyau, ils peuvent atteindre des niveaux de précision plus élevés.

Dans la littérature statistique et informatique, les modèles de Bayes naïfs sont connus sous divers noms, dont Bayes simple et Bayes indépendant. Tous ces noms font référence à l'utilisation du théorème de Bayes dans la règle de décision du classificateur, mais le modèle Bayes naïf n'est pas (forcément) une méthode bayésienne.

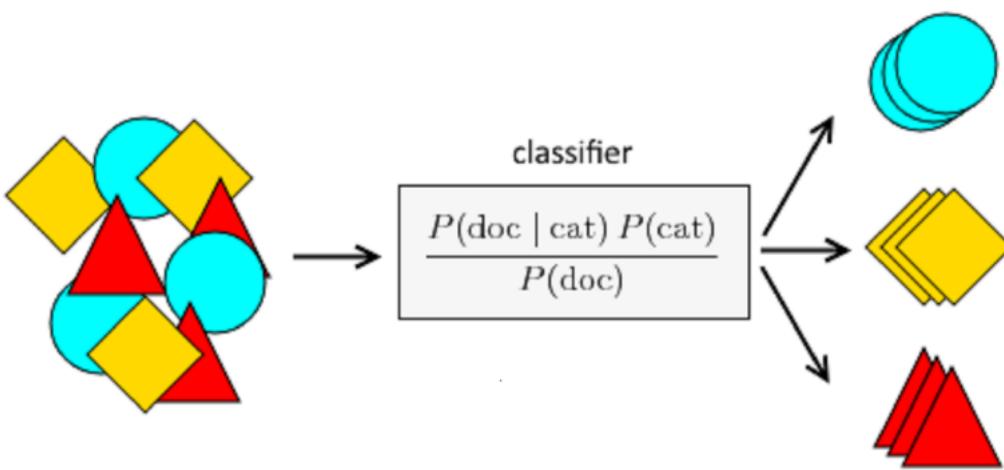


FIGURE 17 – classifieur Naïve Bayes

Dans le domaine informatique, Naïve Bayes est connu comme un algorithme d'apprentissage automatique supervisé basé sur le théorème de Bayes qui est utilisé pour résoudre les problèmes de classification en adoptant une approche probabiliste. Il part du principe que les variables prédictives d'un modèle d'apprentissage automatique sont indépendantes les unes des autres.

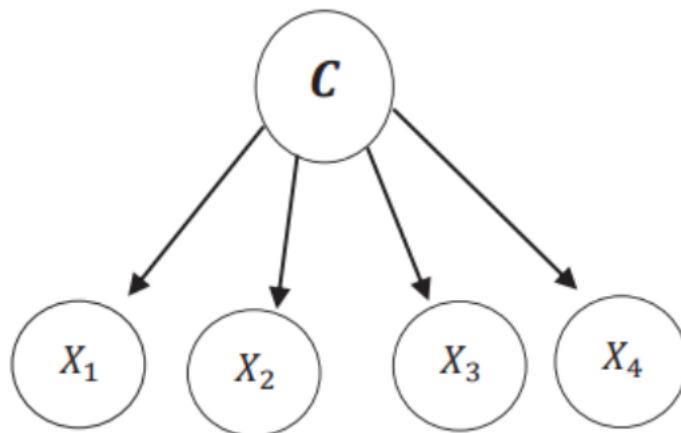


FIGURE 18 – Structure classifieur Naïve Bayes

5.1.3 Classifieur de réseau bayésien naïf augmenté par arbre (TAN)

Le modèle Naive Bayes présenté dans la section précédente, force des hypothèses d’indépendance incorrectes selon lesquelles, étant donné l’étiquette de classe, les attributs sont indépendants les uns des autres. Mais dans le monde réel, les attributs de n’importe quel système sont le plus souvent corrélés et le cas du modèle Bayes naïf se produit rarement. En dépit de ces hypothèses d’indépendance incorrectes, le modèle de Naive Bayes semble fonctionner assez bien. Ainsi, si le modèle prend également en compte les corrélations entre les attributs, la précision de la classification peut être améliorée.

Un modèle Naive Bayes à arbre augmenté (TAN) impose une structure arborescente au modèle Naive Bayes, en limitant l’interaction entre les variables à un seul niveau.

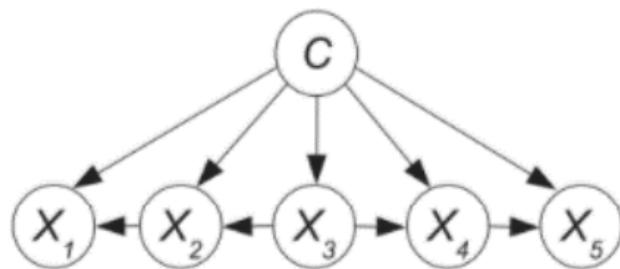


FIGURE 19 – Structure classifieur TAN

5.1.4 Classifieur Chow-liu

En théorie des probabilités et en statistique, l'arbre de Chow-Liu est une méthode efficace pour construire une approximation de produit de second ordre d'une distribution de probabilité conjointe. Les objectifs d'une telle décomposition, comme avec de tels réseaux bayésiens en général, peuvent être soit la compression des données, soit l'inférence.

La méthode de Chow-Liu décrit une distribution de probabilité conjointe $P(X_1, X_2, \dots, X_n)$ comme un produit de distributions conditionnelle et marginale du second ordre. Par exemple, la distribution à six dimensions $P(X_1, X_2, X_3, X_4, X_5, X_6)$ peut être approché comme :

$$P(X_1, X_2, X_3, X_4, X_5, X_6) = P(X_6/X_5)P(X_5/X_2)P(X_4/X_2)P(X_3/X_2)P(X_2/X_1)P(X_1)$$

Où chaque nouveau terme dans le produit introduit une seule nouvelle variable, et le produit peut être représenté comme un arbre de dépendance de premier ordre, comme le montre la figure ci-dessous.

L'algorithme de Chow-Liu détermine les probabilités conditionnelles à utiliser dans l'approximation du produit. En général, à moins qu'il n'y ait pas d'interactions de troisième ordre ou d'ordre supérieur, l'approximation de Chow-Liu est en effet une approximation et ne peut pas capturer la structure complète de la distribution d'origine.

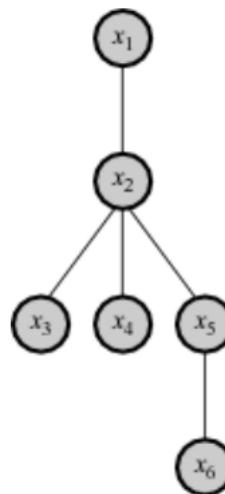


FIGURE 20 – Structure classifieur Chow-liu

5.2 Le classifieur Random Forest

Un random forest [11] est constitué d'un ensemble d'arbres de décision indépendantes. Chaque arbre dispose d'une vision parcellaire du problème du fait d'un double tirage aléatoire :

- un tirage aléatoire avec remplacement sur les observations (les lignes de la base de données). Ce processus s'appelle le tree bagging.
- un tirage aléatoire sur les variables (les colonnes de la base de données). Ce processus s'appelle le feature sampling.

A la fin, tous ces arbres de décisions indépendants sont assemblés. La prédiction faite par le random forest pour des données inconnues est alors la moyenne (ou le vote, dans le cas d'un problème de classification) de tous les arbres.

Random Forest Classifier

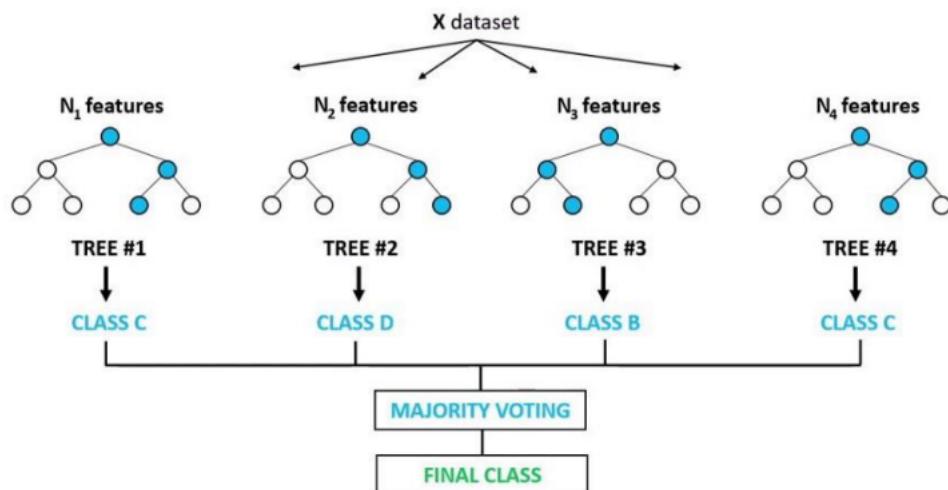


FIGURE 21 – L'algorithme Random Forest et ses paramètres

5.3 Résultats

Évaluation du modèle :

Avant de commencer Voici quelques termes courants à connaître :

Vrais positifs (TP) : Prévus positifs et qui sont effectivement positifs.

Faux positifs (FP) : Prévu positif et qui est en fait négatif.

Vrais négatifs (TN) : Prévision négative et résultats réellement négatifs.

Faux négatifs (FN) : Prévus négatifs et qui sont en fait positifs.

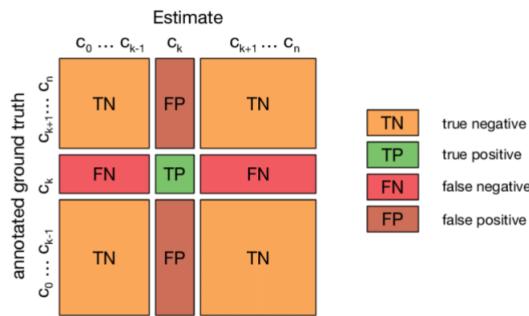


FIGURE 22 – Matrice de confusion

- **Accuracy** : métrique la plus couramment utilisée pour juger un modèle et n'est en fait pas un indicateur clair de la performance. Le pire se produit lorsque les classes sont déséquilibrées, comme dans notre projet Une "Accuracy" acceptable or au niveau des classes il y a une mauvaise "precision" (Infiltration). Lors d'évaluation on va considérer de plus "Precision" et "Recall".

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

- **Precision** : Pourcentage d'instances positives sur le total des instances positives prédites. Ici, le dénominateur est la prédiction du modèle faite comme positive à partir de l'ensemble des données données. Il s'agit de déterminer "dans quelle mesure le modèle a raison quand il dit qu'il a raison".

$$Precision = \frac{tp}{tp + fp}$$

- **Recall** : Pourcentage d'instances positives sur le total des instances positives réelles. Le dénominateur est donc ici le nombre réel d'instances positives présentes dans l'ensemble de données. Il s'agit de déterminer "combien de bonnes instances supplémentaires le modèle a manqué lorsqu'il a montré les bonnes instances".

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

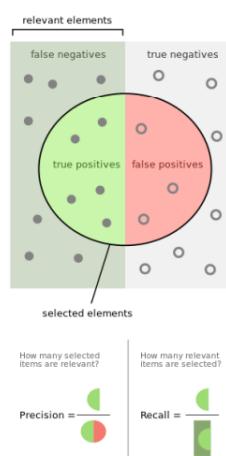


FIGURE 23 – Precision et Recall

Résultats :

La classification est appliquée sur les deux parties entraînement et test du jeu de données.

La classification du jeu de données est appliquée après sélection des variables (en utilisant l'algorithme Random Forest).

Les résultats de la classification en comparant les 3 valeurs d'infiltration : 100000, 200000 et 300000 avec les résultats précédents ainsi que le temps d'entraînement sont affichés pour chaque classifieur.

5.3.1 Classifieur NB

Nombre d'échantillon pour infiltration	Precision			Recall			f1-Score		
	100000	200000	300000	100000	200000	300000	100000	200000	300000
BENIGN	94%	94%	94%	95%	95%	94%	94%	94%	94%
DDOS	54%	54%	54%	58%	58%	58%	56%	56%	56%
DOS	72%	72%	72%	54%	54%	54%	62%	62%	62%
Infiltration	0%	0%	0%	45%	45%	45%	0%	0%	0%

Nombre d'échantillon pour infiltration	Accuracy			Temps d'entraînement		
	100000	200000	300000	100000	200000	300000
Total	89%	89%	89%	7,25 mins	8 mins	9 mins

On remarque que le problème principal est que le modèle ne détecte pas bien la classe Infiltration et la détecte en étant une classe Benign, ce qui influence sur les autres classes, c'est à dire beaucoup de faux positifs pour les autres classes, comme cela, lors du calcul des "precision" et "recall" ils seront basse. Ainsi pour réduire ce problème nous avons décidé de changer la structure du modèle (TAN, Chow-Liu) parce qu'on n'aura pas négliger les dépendances entre les variables, ce qui nous permettra d'avoir beaucoup d'information lors de l'apprentissage.

5.3.2 Classifieur TAN

Nombre d'échantillon pour infiltration	Precision			Recall			f1-Score		
	100000	200000	300000	100000	200000	300000	100000	200000	300000
BENIGN	94%	94%	94%	99%	99%	99%	97%	96%	96%
DDOS	54%	60%	60%	27%	26%	26%	36%	37%	37%
DOS	75%	75%	75%	53%	56%	56%	62%	64%	64%
Infiltration	0%	0%	0%	45%	45%	64%	0%	0%	0%

On remarque que le problème persiste, on n'arrive pas à détecter la classe Infiltration.

Sans sélection de variables :

Nombre d'échantillon pour infiltration	Precision			Recall			f1-Score		
	100000	200000	300000	100000	200000	300000	100000	200000	300000
BENIGN	99%	99%	100%	100%	100%	100%	99%	99%	100%
DDOS	100%	100%	100%	89%	84%	100%	94%	91%	100%
DOS	96%	98%	96%	99%	99%	98%	98%	97%	97%
Infiltration	100%	100%	100%	64%	73%	73%	78%	84%	84%

Nombre d'échantillon pour infiltration	Accuracy			Temps d'entraînement		
	100000	200000	300000	100000	200000	300000
Total	99%	99%	99%	14 s	15 s	17 s

On remarque depuis la comparaison entre les structures que le modèle Random Forest sans utiliser l'algorithme de sélection de variables est bien meilleur que les autres modèles par rapport à tous les classes dans le jeu de données. En plus nous avons pu encombrer le problème de faux positives pour la classe Infiltration ce qui reflète les bons valeurs "precision" et "recall".

On peut déduire que nos variables sont déjà bien sélectionnées.

6 Conclusion

Il a été constaté d'après la comparaison des performances des différents classificateurs que l'application de la technique SMOTE avec un nombre d'échantillons de la classe Infiltration : 200000 et une classification Random Forest sans faire de sélection de variables demeure le modèle le plus performant.

Conclusion

A l’issue de ce projet réalisé au Laboratoire LAGES-EHTP, j’ai eu l’opportunité de travailler sur un sujet d’actualité, sur le thème de la cybersécurité et de l’intelligence artificielle. Ce projet s’est révélé très enrichissant. C’était une occasion tout d’abord de se familiariser avec la technologie IoT, le domaine de l’intelligence artificielle, de la machine learning et une occasion aussi de maîtriser le langage Python.

Bibliographie

- ① <https://www.crowdstrike.com/cybersecurity-101/internet-of-things-iot-security/>
2. <https://www.ibm.com/fr-fr/topics/cybersecurity>
3. <https://www.mdpi.com/journal/mathematics>
4. <https://connect.ed-diamond.com/>
5. <https://www.udemy.com/course/intelligence-artificielle-az/>
6. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*
7. Darío Ramos-López and Ana D Maldonado. Cost-sensitive variable selection for multi-class imbalanced datasets using bayesian networks. *Mathematics*,
8. Sara Zermani. Implémentation sur SoC des réseaux Bayésiens pour l'état de santé et la décision dans le cadre de missions de véhicules autonomes.
9. <https://www.unb.ca/cic/datasets/ids-2017.html>
10. <https://www.scirp.org/journal/paperinformation.aspx?paperid=104256>