


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Akli Mohand Oulhadj de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master

en infomatique

Spécialité : *Ingénierie des systèmes d'information et des logiciels*

Thème

---

apprentissage automatique Pour Le Controle d'Accès

Encadré par

— M. BOUDJELABA Hakim

Réalisé par

— Taleb Amel

— Ziane Asmaa

2023/2024

Remerciements

Nos tenons à exprimer toute nos reconnaissance à notre encadrant , Monsieur Hakim Boudjelaba. Nous le remercions de nous avoir encadré, orienté, aidé ,conseillé, et aussi pour sa disponibilité inconditionné .

Nos vifs remerciements iront aussi aux membres de jury qui nous ont fait l'honneur de juger notre modeste travail.

Nos remerciements sont destinés de même à, tous nos enseignants du département informatique de l'université Akli Mohand Oulhadj Bouira, qui ont contribué à notre formation.

Nous remercions tous nos amis (es) et camarades de promotion pour leurs aides, leurs remarques et leurs critiques qui nous ont fait évoluer.

Dédicaces

Ce modeste mémoire est dédié à :

Tous les membres de nos précieuses familles.

Nous exprimons notre gratitude envers tous nos professeurs pour leur soutien ininterrompu et leurs conseils précieux tout au long de notre parcours académique.

À tous ceux qui nous ont aidés, nous sommes extrêmement reconnaissants.

Asma , Amel

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ML	machine learning
Svm	support vector machine
KNN	k Nearest Neighbor
PA	permission d'accès
UA	utilisation d'accès
ACL	liste de contrôle d'accès
RBAC	Role-Based Access Control
ReBA	:contrôle d'accès basè sur les relations
ABAC	contrôle d'accès basé sur mes attributs
AC	accès contrôle

Introduction générale

Contexte

L'apprentissage machine (Machine Learning) s'est révélé extrêmement efficace pour résoudre des problèmes fastidieux dans de nombreux domaines. Comme d'autres domaines, celui du contrôle d'accès comporte de nombreuses questions pour lesquelles une solution manuelle pilotée par l'homme pourrait aboutir, au mieux, à une solution sous-optimale. Au fil du temps, les chercheurs en contrôle d'accès ont pris l'initiative d'utiliser la puissance du ML pour obtenir des solutions plus efficaces et ont proposé des solutions basées sur la ML pour concevoir des politiques de contrôle d'accès qui sont plus robustes que les approches traditionnelles. Ces dernières années ont également été marquées par l'utilisation de l'apprentissage automatique pour la vérification des politiques d'accès, ce qui pourrait améliorer la qualité des politiques de contrôle d'accès mise en œuvre et alléger les charges sous-jacentes en automatisant les processus de contrôle d'accès ont utilisons l'apprentissage automatique pour la prise de décision en matière de contrôle d'accès lorsqu'un modèle de ML entraîné décidera si une demande d'accès doit être acceptée ou refusée. Comme on peut le constater, l'effet de l'apprentissage automatique dans le contrôle d'accès s'avère positif et à d'énormes répercussions sur le contrôle d'accès, cela permet d'améliorer les performances du système d'information.

Problématique

Le contrôle d'accès manuel peut être une tâche longue et sujette aux erreurs. L'effort et le coût sont très élevés.

Plusieurs défis se posent dans la pratique actuelle du contrôle d'accès utilisant le machine learning :

— Nous observons que les méthodes appliquées par ML sont cas par cas, donc il n'existe pas de stratégie commune pour l'utilisation du ML dans le domaine du contrôle d'accès.
— Il est difficile de sélectionner le meilleur modèle pour un nouveau problème.
— L'approche d'apprentissage automatique proposée est-elle applicable à l'ensemble du domaine du contrôle d'accès ou convient-elle uniquement à un modèle particulier ?
— Quelles sont les méthodes de ML que l'approche respective utilise ?
— Quelles sont les données d'entrée et de sortie du modèle d'apprentissage automatique ? En particulier, le modèle d'apprentissage automatique formé peut-il prendre des décisions en matière de contrôle d'accès ?
— Quel type de données a été utilisé pour l'entraînement de l'algorithme ML ? Quel type de prétraitement des données est nécessaire pour une approche similaire ?
— Quelle est l'efficacité d'une méthode de ML pour résoudre le problème de contrôle d'accès souhaité ?

Objectif

L'objectif principal de notre travail est de proposer une méthode ou un modèle de contrôle d'accès puis le tester avec plusieurs méthodes de ML on ajoutant des amélioration possibles et arriver à de meilleure performance.

Organisation du mémoire

Notre mémoire comporte quatre chapitres :

— Dans le premier chapitre nous allons parler de l'authentification, ensuite nous allons présenter les principaux concepts du contrôle d'accès et les différentes techniques existantes ainsi que les différents modèles de contrôle d'accès existants, et une comparaison entre ces modèles.
— Le deuxième chapitre nous allons présenter c'est quoi l'intelligence artificielle, de la machine Learning et de ses principales tâches. Nous allons également explorer

le Deep Learning, ses types, ainsi que ses architectures, pour conclure par une comparaison.

— Le troisième chapitre explore les systèmes de contrôle d'accès basés sur la machine Learning, avec une attention particulière à une enquête sur l'application de l'apprentissage automatique dans le contrôle d'accès.

— Le dernier chapitre, est consacré à proposer une méthode ou un modèle de contrôle d'accès, que nous testerons avec différentes approches de machine Learning, en intégrant des améliorations potentielles pour atteindre une performance presque optimale.

— Nous clôturons ce mémoire par une conclusion générale pour résumer notre travail et évaluer les résultats obtenus, ainsi que quelques perspectives d'avenir pour améliorer la solution.

Chapitre 1

Authentification et contrôle d'accès

1.1 Introduction

Pour accéder à un système d'information, une identification et une authentification préalables sont nécessaires. L'utilisation de comptes partagés ou anonymes est interdite. Des mécanismes doivent être mis en place pour limiter l'accès aux services, données et privilèges en fonction du rôle de l'utilisateur au sein de l'organisation. Le contrôle d'accès est un système permettant de gérer l'entrée et la sortie des personnes, informations ou ressources dans un lieu donné. Il joue un rôle essentiel dans la sécurité globale d'une organisation en garantissant que seules les personnes autorisées puissent accéder aux zones sensibles. Ce contrôle peut s'appliquer aux locaux physiques et systèmes informatiques, assurant ainsi une protection complète des actifs de l'entreprise.

1.2 Authentification

1.2.1 Définition

L'authentification est le processus de vérification de l'identité d'un utilisateur ou d'un système. Elle est cruciale pour la sécurité des systèmes informatiques, des applications et des services en ligne. Il s'agit de vérifier la véracité des utilisateurs, du réseau et des documents.

L'authentification ne se limite pas à l'utilisation de mots de passe ; il existe diverses autres méthodes. Ces mécanismes peuvent être classés en trois catégories basées sur au moins un des critères suivants :

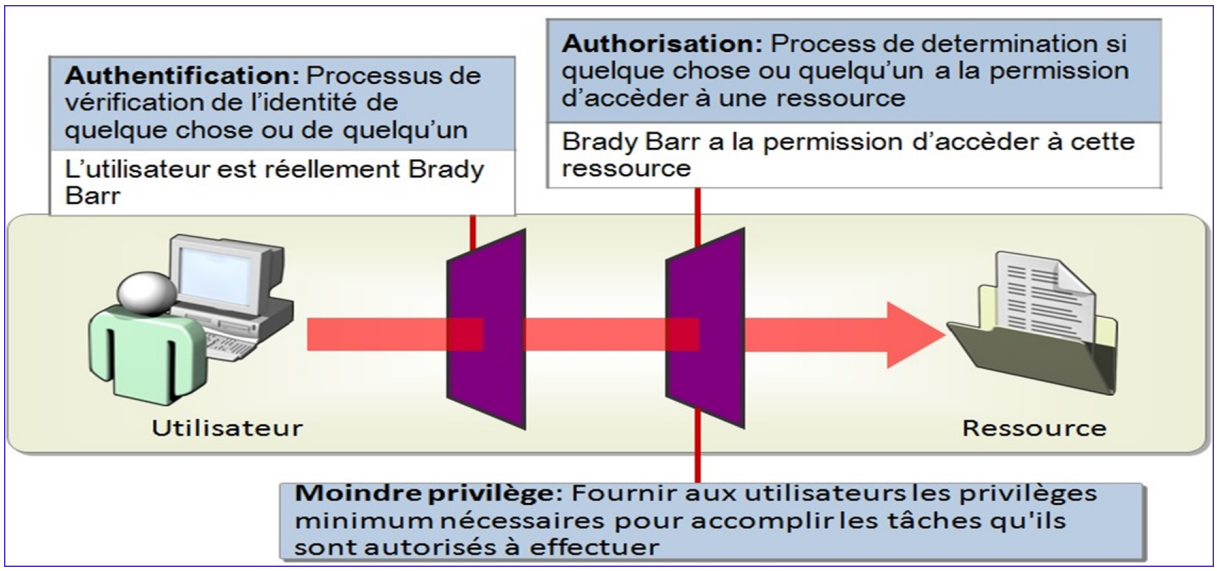


FIGURE 1.1 – Authentification / Autorisation

Quelque chose que l'on est :

— Techniques biométriques, comme la prise d'empreintes digitales, l'analyse rétinienne, l'analyse de la voix, la reconnaissance faciale, etc.

Quelque chose que l'on sait :

— Système de mot de passe traditionnel.

Quelque chose que l'on a :

— Mécanismes comme des listes de questions-réponses, des one-time pads (blocs à usage unique), des cartes à puces, etc.

1.2.2 Types d'Authentification

Voici les principaux types d'authentification utilisés [1] :

Authentification par Facteurs Uniques :

— **Mot de Passe**

Utilisation d'un mot de passe, généralement en combinaison avec un nom d'utilisateur.

Ex : Connexion à un compte de messagerie, réseaux sociaux.

— **Code PIN**

Utilisation d'un code numérique, souvent pour des appareils mobiles ou des cartes

de crédit.

Ex : Déverrouillage de smartphone, paiement par carte bancaire.

Authentification par Facteurs Multiples

— **Authentification à Deux Facteurs**

Deux méthodes, par exemple un mot de passe et un code envoyé par SMS.

Ex : Connexion à des services bancaires en ligne.

— **Authentification à Trois Facteurs**

Combinaison de trois méthodes d'authentification, par exemple, quelque chose que l'utilisateur connaît (mot de passe), quelque chose qu'il possède (carte) et quelque chose qu'il est (empreinte digitale).

Ex : Accès à des systèmes de haute sécurité.

Authentification par Connaissance

— **Questions de Sécurité**

Réponse à des questions prédéfinies basées sur des informations personnelles.

Ex : "Quel est le nom de jeune fille de votre mère ?"

Authentification Biométrique

— **Empreinte Digitale**

Vérification par empreinte digitale.

Ex : Déverrouillage de smartphones, terminaux de paiement.

— **Reconnaissance Faciale**

Identification via la structure faciale.

Ex : Déverrouillage d'appareils Apple Face ID.

— **Reconnaissance Vocale**

Utilisation de la voix pour l'identification.

Ex : Services bancaires par téléphone.

— **Authentification Basée sur la Localisation**

  — **Géolocalisation**

  Authentification basée sur l'emplacement géographique de l'utilisateur.

  Ex : Accès à des comptes bancaires uniquement depuis des lieux de confiance.

— **Authentification Basée sur le Comportement**

Analyse du comportement de l'utilisateur, comme la vitesse de frappe ou les habitudes de navigation.

Ex : Systèmes de détection de fraude dans les banques.

1.3 Contrôle D'accès

1.3.1 Définition

Le contrôle d'accès est un moyen de restreindre l'entrée ou l'utilisation de certaines zones ou ressources uniquement aux personnes autorisées. Il utilise des méthodes telles que l'identification (par exemple, un badge d'accès) ou l'authentification (par exemple, un mot de passe) pour vérifier l'identité d'un individu et décider s'il peut ou non accéder à un lieu ou à une information spécifique. L'une des exigences majeures du partage des données entre plusieurs utilisateurs est la protection de ces données contre des atteintes à la confidentialité, l'intégrité et la disponibilité. Chaque accès aux données doit être contrôlé et tous les accès non autorisés doivent être bloqués. Cela est appelé le contrôle d'accès [2].

1.3.2 Importance du contrôle d'accès

Le contrôle d'accès est crucial pour la sécurité d'une organisation. Il permet de prévenir les intrusions non autorisées, de protéger les biens et les informations confidentielles, d'éviter les vols et les actes de vandalisme, et de limiter les risques de perturbation des activités. En empêchant l'accès non autorisé, il garantit également la confidentialité des données sensibles et la conformité aux réglementations en matière de protection des renseignements personnels [3].

1.3.3 Objectifs du contrôle d'accès

Les principaux objectifs du contrôle d'accès sont d'assurer la sécurité physique et la sécurité des informations dans un environnement donné. Il vise à prévenir les incidents de sécurité en empêchant l'accès non autorisé, en réduisant les risques de vol, de sabotage ou de mauvaise utilisation des ressources. Il permet également de protéger la vie privée

des individus en contrôlant l'accès aux données personnelles et en veillant à ce qu'elles ne soient pas divulguées à des personnes non autorisées. Enfin, il contribue à maintenir l'intégrité des systèmes et des réseaux informatiques en limitant l'accès à ceux qui ont les permissions nécessaires [3].

1.3.4 Les méthodes du contrôle d'accès

Il existe plusieurs méthodes de contrôle d'accès, chacune adaptée à des besoins spécifiques en matière de sécurité, de commodité et de coût. Voici les méthodes les plus couramment utilisées :

— **Méthodes physiques**

Les méthodes physiques utilisent des équipements matériels pour restreindre l'entrée dans des zones spécifiques. Cela peut inclure des serrures à clé, des systèmes de verrouillage électronique et des portes de sécurité. Par exemple, les serrures à clé nécessitent une clé spécifique pour être déverrouillées, tandis que les dispositifs de verrouillage électronique utilisent des cartes d'accès ou des codes PIN. Les barrières et portes de sécurité peuvent également être utilisées pour contrôler physiquement le passage.

— **Méthode logique**

Le contrôle d'accès logique gère l'accès aux systèmes informatiques, aux réseaux et aux données électroniques. Son objectif est de restreindre l'accès aux utilisateurs autorisés tout en protégeant les informations sensibles. Des mesures telles que les mots de passe, les cartes à puce et les certificats numériques sont couramment utilisées.

— **Méthode biométrique**

Les systèmes biométriques vérifient l'identité d'une personne en fonction de caractéristiques physiologiques ou comportementales spécifiques. Les technologies populaires incluent la reconnaissance des empreintes digitales, la reconnaissance faciale, le balayage de l'iris et la reconnaissance vocale.

1.3.5 Politiques du contrôle d'accès

Elle établit les règles selon lesquelles il est nécessaire de réguler le contrôle d'accès. Les politiques sont un élément essentiel qui définit d'une part la structure du contrôle

d'accès et d'autre part les utilisateurs autorisés à réaliser de telles actions sur de telles ressources.[1].

Nous identifions trois concepts essentiels d'une politique de contrôle d'accès à partir de cette définition :

— **Sujet** : S est une entité active qui symbolise les utilisateurs d'un système. Par exemple, on peut citer une personne, une application, un processus et une adresse IP.

— **Objet** : Les données protégées dans un système sont représentées par une entité passive, notée O. Par exemple : un document, une ressource, une table de relation, un logiciel, une donnée.

— **Action** : Les données protégées dans un système sont représentées par une entité passive, notée O. Par exemple : un document, une ressource, une table de relation, un logiciel, une donnée.

Différentes mesures ont été suggérées afin de satisfaire aux diverses exigences de contrôle d'accès des applications :

Liste de contrôle d'accès (ACL) L'état du contrôle d'accès d'un système est représenté de manière basique par la liste de contrôle d'accès. La liste des autorisations ou des restrictions d'accès à une ressource spécifique, comme un fichier, un répertoire, un réseau ou un périphérique, est appelée ACL. Toutes les règles d'une liste de contrôle d'accès définissent les actions autorisées à réaliser par un utilisateur ou un groupe d'utilisateurs sur la ressource en question, telles que la lecture, l'écriture, l'exécution, etc. Dans les systèmes d'exploitation et les équipements réseau, les ACL sont employés afin de mettre en place la politique de contrôle d'accès établie par l'entreprise. [3].

Les cellules de la matrice contiennent une valeur notée *value[i,j]* telle que :

— Valeur *[i, j]* = droit d'accès j associé à l'utilisateur ou au groupe i

— Un droit d'accès est aussi appelé **permission**

— La valeur peut être discrète : 1 ou 0

  — 1 : permission est accordée

  — 0 : permission est refusée

— Par défaut, la permission n'est ni accordée ni refusée (NULL).

Il existe trois types de listes de contrôle d'accès :

— **Listes de contrôle d'accès gratuites** : Ces listes de contrôle d'accès, également

RESOURCE=R		ACCESS RIGHTS		
		AD1	AD2	AD3
USERS GROUPS	UG1	value	value	value
	UG2	value	value	value
	UG3	value	value	value

FIGURE 1.2 – principe d'autorisation

connues sous le nom de listes de contrôle d'accès basées sur l'utilisateur, sont gérées par le propriétaire de la ressource et permettent à chaque utilisateur de définir l'accès à ses propres ressources.

— **Listes de contrôle d'accès système** : également appelées listes de contrôle d'accès basées sur des groupes, ces listes de contrôle d'accès sont définies par le système d'exploitation ou l'administrateur système et appliquées à toutes les ressources du système.

— **Listes de contrôle d'accès étendues** : Ces listes de restriction d'accès offrent la possibilité de définir des règles plus complexes, comme autoriser le filtrage en fonction de critères tels que l'adresse IP source, le port source, etc.

En général, les listes de contrôle d'accès sont mises en place dans le système d'exploitation (comme Windows, Linux) afin de gérer l'accès aux fichiers et aux répertoires.
Les ACL peuvent être installées sur des équipements tels que les routeurs, les commutateurs et les pare-feu dans les réseaux informatiques afin de gérer le trafic réseau entrant et sortant.

Contrôle d'accès discrétionnaire (DAC) : Dans le modèle de contrôle d'accès discrétionnaire, chaque objet ou ressource du système possède un propriétaire (un sujet), qui a la capacité de déterminer les privilèges d'accès à cet objet. Le sujet exerce une totale maîtrise sur tous les objets qui lui sont propres ; il a la possibilité de modifier

les autorisations d'accès, de transférer des objets authentifiés ou d'accorder des accès à l'information à d'autres personnes. C'est ainsi qu'il est qualifié de discrétionnaire. Dans ce schéma, les droits sont directement octroyés à des sujets en fonction de leur identité.[3]

L'inconvénient de cette méthode réside dans le fait que, dans les grands systèmes, la question de savoir s'il faut accorder des autorisations à une ressource donnée à un seul utilisateur est fastidieuse et difficile. Il est également difficile de révocation de la permission lorsque l'utilisateur quitte l'entreprise ou change de poste, par exemple. Il est possible de copier l'information d'un objet à un autre, ce qui permet d'accéder à une copie même si le propriétaire initial ne donne pas accès à l'originale. Comme les propriétaires ont la possibilité de modifier facilement les politiques du DAC, les logiciels malveillants exécutés en leur nom pourront également modifier ces mêmes politiques, ce qui représente un désavantage pour les systèmes DAC. [4]

En synthèse : [5]

- Le système = l'ensemble (Objet + Sujet).
- Objet et Sujet sont énumérés.
- Autoriser (s, o) = {Θ, Lire, Ecrire, Exécuter…} /s ∈ Sujet, o ∈ Objet.
- Représentation :
  - Une matrice (n, m) tel que n=card (Sujet) ; m=card (Objet).
  - Matrice (i, j) = {Θ, Lire, Ecrire, Exécuter…}

Une matrice de contrôle d'accès consiste en une ligne pour chaque sujet et une colonne pour chaque objet du système. Les lignes et les colonnes se rencontrent pour décrire l'accès du sujet à l'objet (lecture, écriture, exécution, etc). On peut observer un exemple de matrice de contrôle d'accès dans la figure suivante :

Contrôle d'accès obligatoire MAC (Mandatory Access Control) : Dans le modèle MAC, la décision de politique est prise par l'administrateur en fonction des étiquettes de sécurité des sujets (autorisation) et des objets (classification). Les droits d'accès ne sont pas contrôlés, définis et modifiés par l'utilisateur. En général, on définit le contrôle d'accès discrétionnaire par opposition au contrôle d'accès obligatoire ou MAC, qui impose des

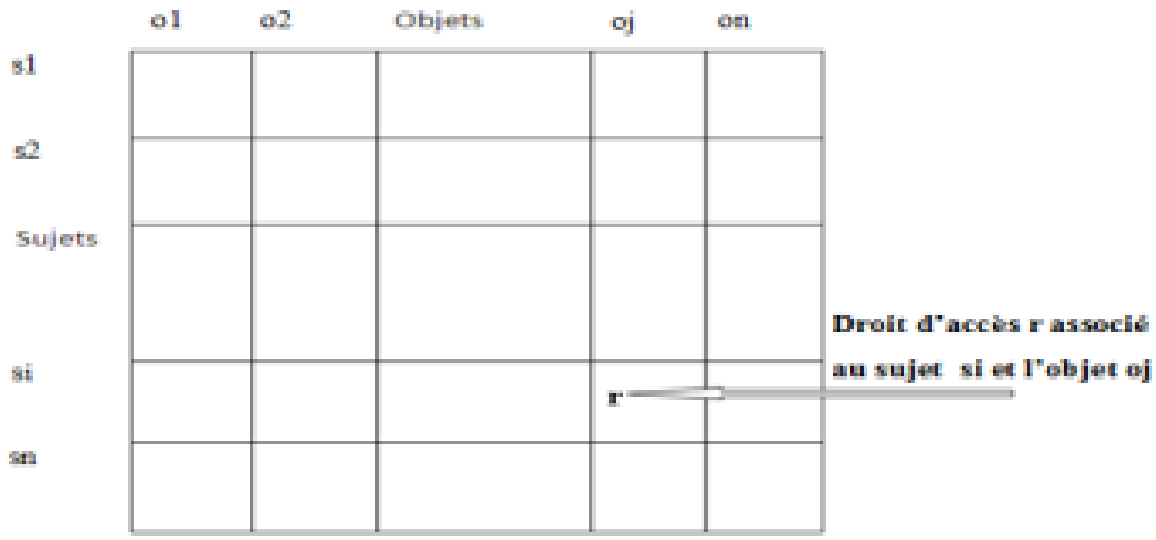


FIGURE 1.3 – exemple de matrice de contrôle d'accès

règles indispensables pour assurer l'atteinte des objectifs de sécurité établis. [3] ,[5].

Dans ce genre de contrôle d'accès, les personnes concernées ne peuvent pas avoir d'influence sur l'octroi des droits d'accès. Ce contrôle d'accès présente une rigidité supérieure à celle du contrôle d'accès discrétionnaire, mais il est tout de même plus sûr. MAC est employé lorsqu'il est prévu dans la politique de sécurité du système d'information que le propriétaire de l'objet en question ne doit pas prendre la décision de protection, c'est-à-dire que la décision de protection doit lui être imposée par le système. Optez pour une classification des rubriques et des objets en fonction des niveaux de sécurité préétablis utilisés lors de la prise de décision concernant l'accès. [6]

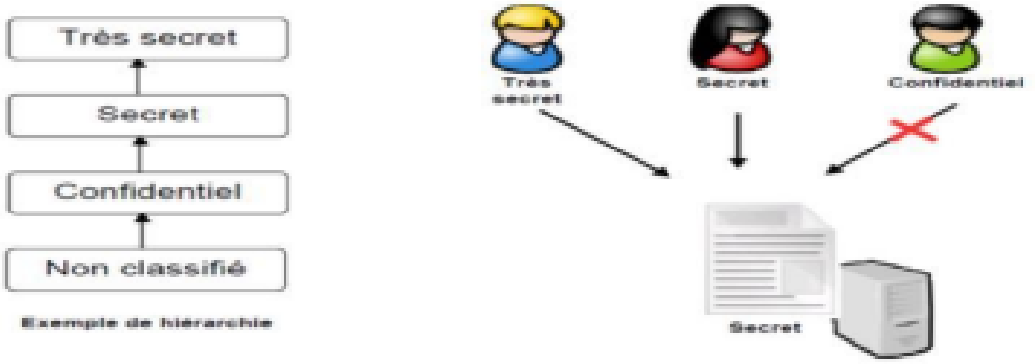


FIGURE 1.4 – Exemple de contrôle d'accès avec les niveaux de sécurité

Quatre niveaux de sécurité sont généralement considérés : Très Secret (TS), Secret

(S), Confidentiel (C), et non classifiés (U), avec l'ordre suivant : TS > S > C > U.[5].

Contrôle d'accès basé sur les rôles RBAC (Role-Based Access Control) Dans le RBAC, les droits d'accès de l'utilisateur sont conditionnés par son rôle au sein de l'organisation. Il est possible de considérer le contrôle d'accès basé sur les rôles comme une solution alternative au contrôle d'accès obligatoire (MAC) et au contrôle d'accès discrétionnaire (DAC). Ce modèle attribue les autorisations qui peuvent être représentées par paire (o, a) à un rôle spécifique (o O et a A) et non directement à l'identité comme dans le modèle précédent. Par la suite, il est possible d'attribuer des rubriques à des rôles qui découleraient habituellement d'une organisation. [7].

Le système RBAC permet également de séparer les tâches de manière statique et dynamique, ce qui simplifie les opérations comme l'ajout ou la suppression d'un sujet, tout en permettant de définir d'autres contraintes telles que des rôles, des autorisations et une cardinalité mutuellement exclusifs. En effet, les autorisations ne sont pas attribuées individuellement aux sujets et les sujets ne peuvent obtenir ces autorisations qu'à partir de leurs fonctions, ce qui rend RBAC un système « parfait » pour les entreprises où le personnel change fréquemment. [3] ,[7].

s

FIGURE 1.5 – Attribution des permissions en RBAC

Un rôle peut posséder différentes autorisations et une autorisation peut être liée à différents rôles. Plusieurs rôles peuvent être joués par un utilisateur et un rôle peut être attribué à plusieurs utilisateurs. Une permission P est accordée à un sujet si et seulement si ce sujet est assigné à un rôle qui possède cette permission. [7].

Contrôle d'accès basé sur les attributs (ABAC) Gestion d'accès basée sur les caractéristiques (ABAC). L'accès d'un utilisateur à une ressource est défini dans le cadre de l'ABAC en fonction des caractéristiques de l'utilisateur, des caractéristiques de la ressource, des conditions environnementales et d'un ensemble de politiques préétablies. Les combinaisons d'attributs de l'utilisateur, de la ressource ou de l'environnement sont définies par une politique ABAC, qui permet à l'utilisateur de travailler sur une ressource. Quand le système ABAC intercepte une demande d'accès, il compare la demande avec les politiques ABAC sous-jacentes, où sont définies toutes les conditions et restrictions. Si la demande d'accès est conforme à la politique, elle recevra l'accès désiré. L'élaboration de la politique d'accès repose sur les règles, ABAC est l'un des modèles de contrôle d'accès.[8]

Contrôle d'accès basé sur les relations (ReBAC) CReBAC : contrôle d'accès basé sur les relations. Le contrôle d'accès fondé sur les relations établit les choix d'accès en se basant sur les liens entre les sujets ou les objets. Les réseaux sociaux sont les principaux exemples de contrôle d'accès basé sur les relations. Sur Facebook, par exemple, l'utilisateur permet à ses amis ou à ses amis de visualiser ses photos ou vidéos. Cependant, les amis de ces amis sont incapables de consulter les photos. De cette manière, ReBAC permet l'accès lorsque l'utilisateur entretient une relation avec d'autres entités du système. Les avantages de ReBAC surpassent ceux des autres modèles de contrôle d'accès en accordant l'accès en fonction de relations multiples entre les entités et en prenant des décisions pour certaines entités plutôt que pour d'autres.[2]

1.4 Comparaison entre les différents modèles de contrôle d'accès

Le Tableau suivant présente une comparaison entre modèles de contrôle d'accès (RBAC, DAC et MAC ...) par rapport aux critères suivants : flexibilité, complexité, support de la collaboration, granularité et utilisation des informations contextuelles.

Critères	ACL (Liste de contrôle d'accès)	DAC (Contrôle d'Accès Discrétionnaire)	MAC (Contrôle d'Accès Obligatoire)	RBAC (Contrôle d'Accès Basé sur les Rôles)	LSE (Contrôle d'Accès Basé sur les Étiquettes)	ABAC (Contrôle d'Accès Basé sur les Attributs)	ReBAC (Contrôle d'Accès Basé sur les Règles)
Définition	Permet aux propriétaires de ressources de définir les autorisations pour les utilisateurs ou les groupes.	Les propriétaires des ressources déterminent les autorisations d'accès pour leurs ressources.	Basé sur des règles prédéfinies en fonction de la classification des ressources et des utilisateurs.	Les autorisations sont attribuées aux rôles, et les utilisateurs obtiennent ces autorisations par leur assignation à des rôles.	Les ressources et les utilisateurs sont étiquetés, et l'accès est basé sur la compatibilité entre les étiquettes.	Les autorisations d'accès sont basées sur les caractéristiques des utilisateurs, des ressources et du contexte de la demande d'accès.	Les autorisations sont déterminées par des règles de politique spécifiques définies par l'administrateur
Flexibilité	Variable selon la mise en œuvre spécifique.	Offre une grande flexibilité car les propriétaires ont un contrôle total sur les autorisations.	Moins flexible en raison de la rigidité des politiques de classification.	Peut être flexible en fonction de la granularité de la définition des rôles.	Variable selon la façon dont les étiquettes sont attribuées et utilisées.	Peut être flexible en fonction de la complexité des politiques d'attribution d'attributs.	Variable en fonction de la complexité des règles définies.
Gestion des autorisations	Généralement décentralisée, chaque propriétaire gère les autorisations pour ses propres ressources.	Décentralisée, les propriétaires de ressources définissent les autorisations.	Centralisée, basée sur des politiques définies au niveau du système.	Centralisée, les autorisations sont définies en fonction des rôles et attribuées aux utilisateurs.	Souvent décentralisée, les étiquettes sont associées aux ressources par leurs propriétaires.	Peut être centralisée ou décentralisée, les attributs sont définis et gérés selon les besoins.	Centralisée, les règles sont généralement définies au niveau de l'administration

Complexité de gestion	Peut devenir complexe dans les environnements avec de nombreuses ressources et utilisateurs.	Peut devenir complexe dans les environnements avec de nombreuses ressources et utilisateurs, car chaque propriétaire doit définir les autorisations.	Souvent complexe en raison de la nécessité de classer les ressources et les utilisateurs de manière cohérente.	Peut simplifier la gestion en attribuant des autorisations en fonction de rôles prédéfinis.	Peut devenir complexe en raison de la nécessité de définir et de maintenir des politiques d'étiquetage cohérentes.	Peut simplifier la gestion en basant les autorisations sur des attributs définis.	La complexité dépend de la complexité des règles de politique définies.
Granularité	Variable, peut être fine ou grossière selon la façon dont les listes de contrôle d'accès sont définies.	Variable, dépend de la façon dont les propriétaires définissent les autorisations pour leurs ressources.	Généralement fine, car les politiques de classification sont souvent spécifiques et précises.	Peut être fine ou grossière selon la granularité de la définition des rôles.	Généralement fine, car les étiquettes peuvent être spécifiques et détaillées.	Variable, peut être fine ou grossière en fonction de la façon dont les attributs sont définis et utilisés.	Variable, peut être fine ou grossière selon la complexité des règles définies.

1.5 Conclusion

La protection de l'information dans l'entreprise revêt une importance capitale, ce qui en fait une fonction essentielle. Le contrôle d'accès est l'un des éléments de la sécurité informatique, qui implique de donner des autorisations d'accès en fonction de contraintes et de règles déjà validées. Ce chapitre a abordé la définition, l'importance et l'objectif du contrôle d'accès. Nous avons également exposé quelques modèles de base de contrôle d'accès, puis nous avons comparé ces différents modèles en fonction de quelques critères spécifiques.

Chapitre 2

L'apprentissage automatique ML/DL

2.1 Introduction

L'apprentissage automatique, aussi connu sous le nom de machine learning, est un domaine de recherche en intelligence artificielle. De nos jours, l'intelligence artificielle et le machine Learning sont largement utilisés dans le domaine de la sécurité informatique. L'objectif de la machine Learning est de développer des systèmes intelligents capables de traiter d'énormes volumes de données, de les analyser et de prendre des décisions à une vitesse bien supérieure à celle d'un être humain. L'apprentissage artificiel se réfère à la capacité d'un système à acquérir et à intégrer de manière autonome des connaissances. Cette notion englobe toutes les méthodes permettant de construire un modèle de la réalité à partir de données, que ce soit en améliorant un modèle existant ou en créant un modèle entièrement nouveau .

2.2 Définition de l'apprentissage automatique

Dans le domaine de l'intelligence artificielle, le Machine Learning offre aux machines la possibilité d'analyser, de résoudre des problèmes et de mettre en place des solutions de manière autonome en utilisant intensivement des données et des algorithmes d'apprentissage[9]

En principe, le Machine Learning se base sur la conception d'algorithmes capables de recevoir des données en entrée et d'utiliser des analyses statistiques pour prédire des résultats, tout en ajustant ces prédictions au fur et à mesure que de nouvelles données sont dispo-

nibles. [10].

Pour rendre une machine véritablement intelligente, il ne suffit pas de l'ensemencer de connaissances, mais il faut aussi lui conférer la capacité d'apprendre à partir d'événements observés et d'utiliser cette expérience pour réagir de manière plus adaptée à des situations similaires à l'avenir[11].

Chaque méthode d'apprentissage automatique comporte deux étapes : la première consiste à choisir un modèle (par exemple, un réseau de neurones) et à adapter ses paramètres à partir de données d'entrée spécifiques, telles que des images de chats et de chiens pour un modèle de reconnaissance visuelle. La deuxième étape, l'inférence, exploite les connaissances acquises pour accomplir des tâches spécifiques, comme différencier des images de chats de celles de chiens[12].

2.3 Les principales tâches de l'apprentissage automatique

Plusieurs tâches peuvent être associées à l'apprentissage automatique, parmi elles nous citons :

2.3.1 La classification

La classification consiste à identifier et à différencier des classes de données ou des concepts en utilisant un modèle (ou une fonction). L'analyse d'un ensemble de données d'apprentissage (c'est-à-dire des objets de données dont les étiquettes de classe sont connus) permet de déterminer le modèle. On utilise le modèle pour anticiper la classe d'objets dont l'étiquette de classe est inconnue.[13]

Les algorithmes de classification

Le data mining propose une grande diversité de méthodes et d'algorithmes pour la recherche de données. Certains de ces algorithmes proviennent de différentes sources, telles que la statistique (régression), l'intelligence artificielle (réseaux de neurones, arbres de décision...), tandis que d'autres s'inspirent de la théorie de l'évolution (algorithmes génétiques...). Grâce à cette fusion de technologies, il devient plus facile de résoudre, de

comprendre, de modéliser et de prévoir les problèmes.[14]

L'algorithme k-Means (Macqueen, 1967) L'algorithme K-Means (K-moyennes) est le plus célèbre dans le domaine de l'apprentissage supervisé. C'est un algorithme de regroupement. Il va regrouper les données qui se ressemblent dans des "zones" (Cluster). Les données dans le même cluster présentent des similitudes. L'approche de K-Means implique l'affectation aléatoire de centres de clusters (connus sous le nom de centroïde), puis l'attribution de chaque point de nos données au centroïde le plus proche [29]. Ainsi, chaque groupe de la partition est caractérisé par ses objets et son centre. Par la suite, on redéfinit les centres en utilisant les objets qui ont été attribués aux différents clusters. Ensuite, les objets sont classés selon leur distance aux nouveaux centres, et ainsi de suite. Ainsi, l'algorithme se répète jusqu'à l'existence d'une convergence. La méthode K moyen est une méthode itérative qui tend à converger vers une solution, peu importe son point de départ. [14]

Le résultat est un groupe de clusters compacts et bien séparés, à condition que la valeur K du nombre de clusters soit correcte [15]

Le fonctionnement de l'algorithme K-Means peut être résumé comme suit :

1. On choisit K points distincts $c_1, \ldots, c_k$ au hasard parmi $\{x_1, \ldots, x_n\}$ et on considère que $c_1, \ldots, c_K$ sont les centres des clusters initiaux, avec $\{x_1, \ldots, x_n\}$ étant les objets (nos données).
2. On répète jusqu'à la « stabilisation » des c_k :
   (a) Assigner chaque x_i au cluster $C_k(i)$ tel que $dist(x_i, c_k(i))$ est minimum pour obtenir un ensemble de K classes.
   (b) Recalculer les centres c_k des clusters.

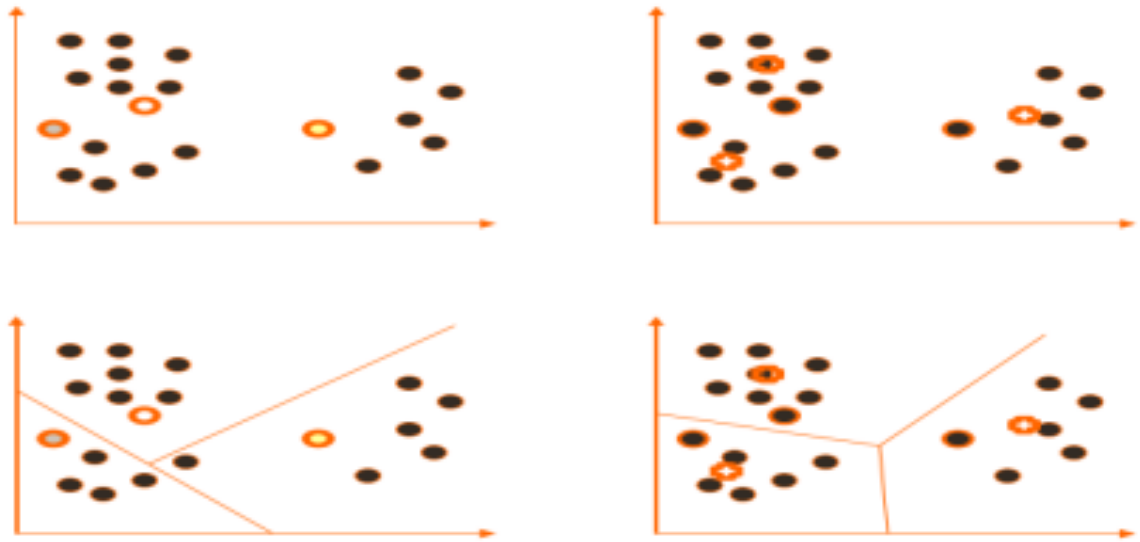


FIGURE 2.1 – Figure : Exemple sur l'algorithme K-Means

La Complexité Le nombre de calculs de distance réalisés ici est extrêmement facile car à chaque étape, nous calculerons la distance entre chaque point et chaque centre. [16]. La complexité de l'algorithme du K-moyen est de $O(1kn)$ [14]. Où l c'est le nombre d'itérations et k c'est le nombre de clusters avec $k < n$.

- **Les avantages :**
  - Cette méthode est simple et facile à mettre en œuvre.
  - • Elle est rapidement assimilée.
  - • Elle présente une faible complexité en matière de calcul.
  - • S'adapte à toutes les mesures standard.
  - • Ne tient pas compte de l'ordre des données.
- **Les inconvénients :**
  - • Il est nécessaire de préciser le nombre de classes à l'avance.
  - • Il n'est pas valable en présence d'attributs non numériques.
  - • • Le résultat final est grandement influencé par le choix des centroids initiaux.
  - • Il est incapable de repérer les groupes non-convexes.
  - • Réceptif au bruit et aux écarts.

l'algorithme SVM (Support Vector Machine) Un algorithme d'apprentissage automatique supervisé appelé machine à vecteurs de support peut être employé pour classifier et régresser. Les SVM ont une utilisation plus courante dans les contextes de classifi-

cation. Les SVM sont basés sur l'idée de déterminer un hyperplan qui permet de séparer au mieux un jeu de données en deux [17] de manière à ce que les données soient séparables de manière linéaire.

Le calcul de la fonction de classement séparant les classes est effectué en utilisant la méthode de construction de l'hyperplan optimal [18]

— Les vecteurs des différentes classes sont situés de part et d'autre de l'hyperplan.

— La marge maximale entre les vecteurs et l'hyperplan doit être la plus petite possible.

Elles se fondent donc sur deux concepts fondamentaux : la notion de marge maximale et la notion de fonction noyau. Si un séparateur linéaire est trouvé, c'est-à-dire qu'il y a un hyperplan séparateur, alors le problème est considéré comme linéairement séparable, sinon il n'est pas linéairement séparable et il n'y a pas d'hyperplan séparateur. [14].

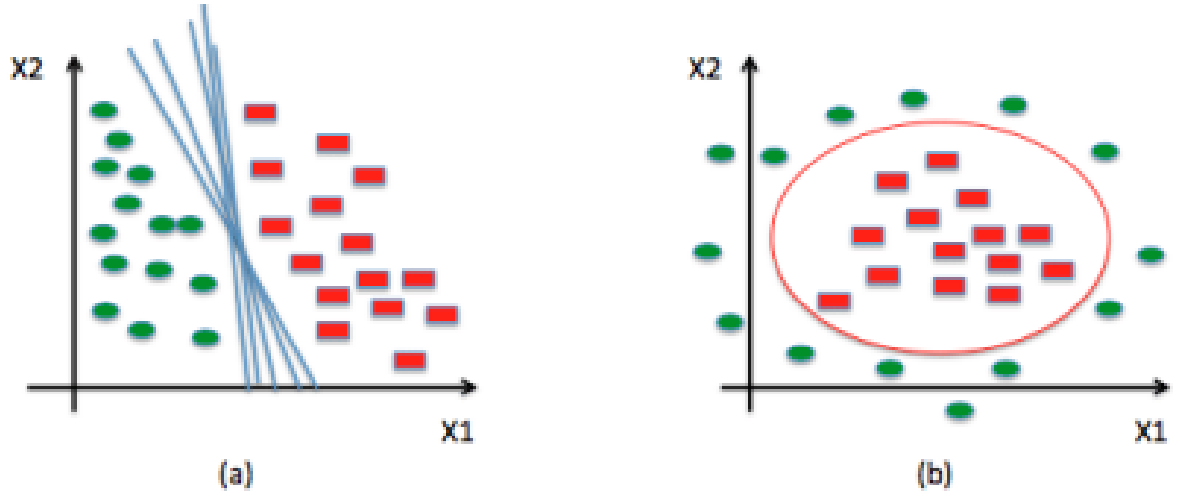


FIGURE 2.2 – Figure Problème de classification à deux classes avec une séparatrice linéaire et non linéaire

Pour résoudre le problème du non linéarité séparatrice, l'idée des SVM est d'augmenter la dimension d'espace de données. Dans ce cas, il est alors probable qu'il existe un séparateur linéaire. En effet, la chance de trouver un hyperplan séparateur augmente proportionnellement avec la dimension d'espace de données.

— **Avantages :** On peut citer parmi les bénéfices des SVM :
  — Elles possèdent une fondation théorique solide [14].
  — Elles peuvent avoir une meilleure performance car elles font appel à un ensemble plus restreint de points d'entraînement [17].

- On compare uniquement les exemples de test avec les supports vecteur (les points de données les plus proches de l'hyperplan) et non avec toutes les données.

— **Inconvénients :** On peut mentionner parmi les désavantages des SVM :

- Elles font appel à des fonctions mathématiques complexes pour la classification.
- Les jeux de données plus volumineux ne sont pas adaptés, car le temps d'apprentissage avec les SVM peut être élevé [17].
- La performance diminue sur les jeux de données qui incluent des bruits et de nombreux outils.

2.4 L'algorithme KNN

La méthode de classification K Nearest Neighbors (PPV Plus Proches Voisins) peut être utilisée pour des tâches d'estimation. La méthode PPV consiste à raisonner à partir de cas concrets. Elle commence par prendre des décisions en cherchant un ou plusieurs cas similaires déjà résolus en mémoire.

Il est nécessaire de trouver les k échantillons les plus proches de l'objet et de les attribuer à la classe la plus représentative parmi ces k échantillons (« Dis-moi qui sont tes amis, et je te dirai qui tu es »). La meilleure méthode consiste à chercher le cas le plus semblable et à prendre la même décision, on utilise le terme 1-NN. Si cette méthode permet d'obtenir des résultats satisfaisants sur des problèmes simples où les objets sont bien répartis en groupes denses de même classe, il est généralement nécessaire de prendre en compte un nombre de voisins plus élevé afin d'obtenir des résultats satisfaisants.

À la différence des autres méthodes de classification, il n'existe pas d'étape d'apprentissage qui consiste à construire un modèle à partir d'un échantillon d'information. Le modèle est constitué par l'échantillon d'apprentissage, qui est lié à une fonction de distance et à une fonction de choix de la classe en fonction des classes plus proches.[19].

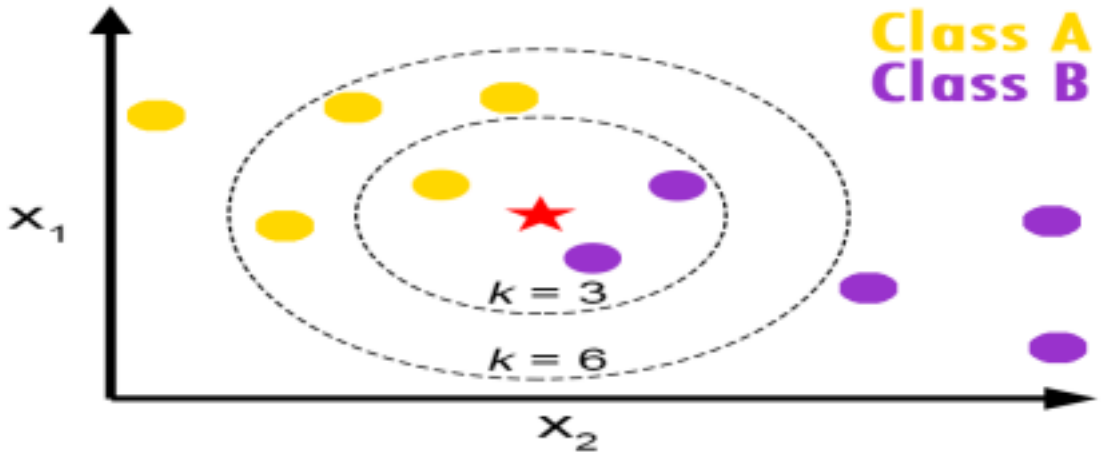


FIGURE 2.3 – Figure : Exemple d'algorithme KNN

Définition de la distance

Elle permet de mesurer le degré de différence entre deux vecteurs. Il existe plusieurs types de distance parmi lesquels on trouve :

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Où : x, y sont des vecteurs.

L'algorithme des k plus proches voisins [14]

Pour $i = 1$ à m faire :
Calculer la distance $d(X_i, x)$ fin pour
Construire l'ensemble I contenant les indices pour k plus petites distances $d(X_i, x)$ Retourner l'étiquette majoritaire pour $\{Y_i \mid i \in I\}$

Où : $\{X$: ensemble d'entraînement, Y : étiquettes de classe de X, x : individu inconnu$\}$

Choix de k :

- -Il est nécessaire que l'utilisateur détermine le paramètre k : k N.
- - Il est avantageux de choisir k impair afin d'éviter les votes égalitaires.
- - Le choix optimal de k est influencé par le jeu de données.
- -Il est nécessaire de tester différentes valeurs de k et de sélectionner celle qui offre le meilleur résultat.

Avantages des k plus proches voisins :

— La méthode des k plus proches voisins est simple à mettre en place et offre des résultats précis et clairs. Elle est efficace lorsque les données sont larges et incomplètes.

— Il est préférable de ne pas apprendre lors de l'introduction de nouvelles caractéristiques [2]. Gérer toutes sortes de données comportant un grand nombre d'attributs.

— L'algorithme KNN présente une résistance face aux données bruitées.

Inconvénients des k plus proches voisins :

— La vitesse de classification est limitée en raison du grand nombre de distances à calculer.

— Cette méthode nécessite une capacité de stockage importante pour le traitement des corpus.

— Il est nécessaire de déterminer le nombre de voisins les plus proches.

Random Forest (RF) L'approche de la Forêt Aléatoire (RF) est une méthode d'apprentissage supervisé qui peut être utilisée pour résoudre des problèmes de classification et de régression. Un ensemble d'arbres de décision est utilisé dans ce modèle afin de procéder à une reconnaissance de formes à grande échelle et multivariée. La classification par discrimination stochastique est basée sur l'approche du sous-espace aléatoire, en collaboration avec la méthode de RF.

. Toutefois, elle pose des difficultés lorsqu'il s'agit de gérer des attributs multiples à plusieurs dimensions. L'un des soucis fréquents. Les forêts aléatoires sont souvent utilisées pour des tâches de régression, ce qui entraîne un risque de sur-ajustement du modèle aux données d'entraînement. [20].

Naive Bayes La famille Naïve Bayes est composée de classificateurs probabilistes basés sur le théorème de Bayes. Il implique une autonomie conditionnelle des caractéristiques (variables), d'où l'appellation. Grâce à cette simplification, il est possible de calculer les probabilités de façon plus performante. [21].

Théorème de Bayes :

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

Où :

— $P(C \mid X)$ est la probabilité de la classe C donnée les caractéristiques X (probabilité a posteriori).
— $P(X \mid C)$ est la probabilité des caractéristiques X données la classe C (vraisemblance).
— $P(C)$ est la probabilité a priori de la classe C.
— $P(X)$ est la probabilité des caractéristiques X (évidence).

Application : Classification de textes, analyse de sentiments, filtrage de spam.

2.4.1 L'estimation

La pratique de l'estimation impliquera de combler une valeur manquante dans un champ spécifique en se basant sur les autres champs de l'enregistrement. Elle ressemble à la classification, mais la sortie est une variable numérique plutôt que verbale. [22].

2.4.2 La prédiction

La prédiction ressemble à la classification et à l'estimation, à la différence que pour la prédiction, les résultats sont anticipés. Il est donc possible d'utiliser toutes les méthodes et techniques employées pour la classification et l'estimation dans des circonstances appropriées pour la prédiction [23]. Effectivement, la prédiction implique de prédire la valeur à venir d'un attribut en se basant sur d'autres caractéristiques

2.5 L'association

Le concept d'association implique de trouver des liens captivants entre des variables dans de vastes bases de données. Par exemple, les individus qui font l'acquisition d'une maison neuve ont également tendance à faire l'acquisition de nouveaux meubles. Il apprend que des éléments peuvent co-apparaître dans une collection [10].

2.5.1 Le clustering

Le clustering est une méthode utilisée pour organiser des objets similaires dans une classe connue sous le nom de cluster. Le clustering ne consiste pas à classer, à estimer ou

à prédire la valeur des variables cibles, mais plutôt à séparer l'ensemble des données en sous-groupes homogènes [24].

2.5.2 La régression

Contrairement à la classification, la régression est une méthode utilisée pour représenter des fonctions à valeurs continues. Il sert à anticiper les valeurs de données numériques manquantes ou non disponibles plutôt que les étiquettes de classe (étiquettes cachées). [25].

2.6 Les types d'apprentissage automatique

Le Machine Learning, l'un des domaines majeurs de l'intelligence artificielle, inclut les techniques et les algorithmes qui permettent à un ordinateur d'acquérir des connaissances. On le relie à l'intelligence computationnelle, qui cherche à automatiser la création de modèles analytiques. Ce domaine offre aux ordinateurs la possibilité de créer des idées, de prendre des décisions et d'anticiper les résultats à venir [25].

Pour l'apprentissage automatique, il est essentiel d'avoir deux ensembles de données :

- **Base de connaissances pour l'entraînement :** C'est un ensemble de données utilisé pour entraîner l'algorithme d'apprentissage. Au cours de cette étape, les réglages du modèle sont adaptés en fonction des résultats obtenus.
- **Recueil de données destinées au :**Seulement utilisé pour évaluer les résultats du modèle sur des données non observées.

La théorie de l'apprentissage utilise les outils mathématiques de la théorie des probabilités et de l'information pour évaluer l'efficacité des différentes méthodes par rapport aux autres[26]. Les outils mathématiques de la théorie des probabilités et de l'information sont utilisés dans la théorie de l'apprentissage pour évaluer l'efficacité des différentes méthodes par rapport aux autres [26]. Trois types d'algorithmes d'apprentissage automatique peuvent être mentionnés : .

Apprentissage supervisé.

Apprentissage non supervisé.

Apprentissage par renforcement.

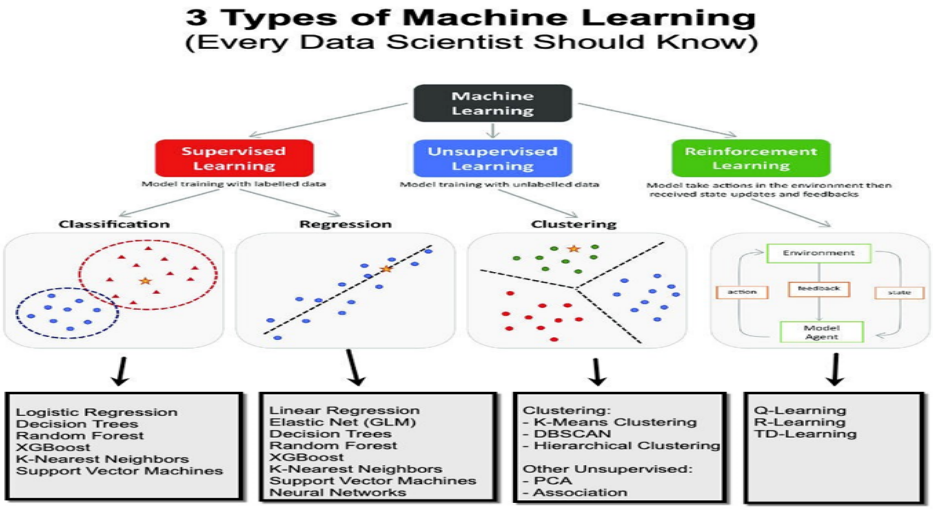


FIGURE 2.4 – Types d'apprentissage automatique

Apprentissage supervisé

L'apprentissage supervisé est l'une des tâches les plus simples et les plus connues en apprentissage automatique. Il repose sur un ensemble d'exemples pré-classifiés où la catégorie de chaque entrée est connue à l'avance. L'enjeu principal dans ce contexte est la capacité du système à généraliser : après avoir analysé un échantillon d'exemples, il doit produire un modèle qui fonctionne efficacement pour toutes les entrées possibles.[4]

L'ensemble de données d'entraînement est constitué d'objets accompagnés de leurs étiquettes de classe associées. Ces exemples étiquetés forment l'ensemble d'apprentissage.

Pour illustrer ce concept, prenons l'exemple suivant : un utilisateur reçoit quotidiennement de nombreux e-mails, certains provenant d'entreprises importantes tandis que d'autres sont des spams non sollicités. Dans un cadre d'apprentissage supervisé, un algorithme serait alimenté avec une grande quantité d'e-mails déjà étiquetés par l'utilisateur comme spam ou non spam. L'algorithme analyse ces données étiquetées pour prédire si chaque nouvel e-mail est un spam ou non.

Ainsi, l'algorithme examine chaque exemple, utilise les informations fournies par les exemples étiquetés pour faire des prédictions sur chaque e-mail, déterminant s'il appartient à la catégorie des spams ou non. [15]

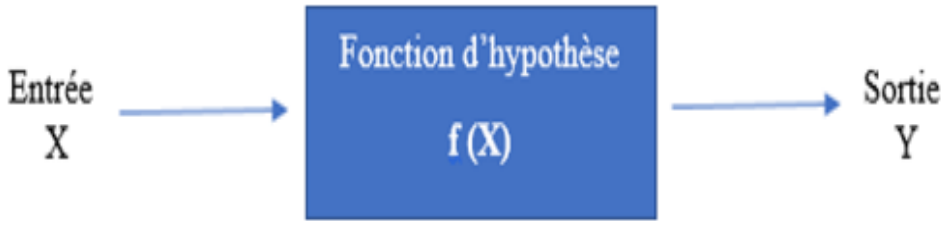


FIGURE 2.5 – Types d'apprentissage automatique

Les algorithmes d'apprentissage automatique supervisé les plus répandus dans la littérature sont les suivants :

- Arbre de prise de décision
- Les voisins les plus proches (KNN)
- Machine à support de vecteurs (SVM)[4] .

L'apprentissage non supervisé

L'apprentissage non supervisé (Unsupervised Learning) implique que les données d'entrée (X) ne sont pas accompagnées de variables de sortie correspondantes [10].
Une illustration un peu plus mathématique afin de clarifier la notion :
Seulement des observations brutes de variables aléatoires sont reçues : x1, x2, x3, x4,... et on souhaite identifier la relation avec des variables latentes structurelles : Xi → Yi. L'apprentissage non supervisé vise à représenter la structure ou la répartition sous-jacente des données pour en saisir davantage les caractéristiques. On l'appelle apprentissage non supervisé parce qu'il n'y a pas de réponse correcte ni d'enseignant, contrairement à l'apprentissage supervisé. Les mécanismes des algorithmes sont laissés à eux-mêmes pour trouver et exposer la structure captivante des données [10] .
L'apprentissage non supervisé vise à représenter la structure ou la répartition sous-jacente des données pour en saisir davantage les caractéristiques. On l'appelle apprentissage non supervisé parce qu'il n'y a pas de réponse correcte ni d'enseignant, contrairement à l'apprentissage supervisé. Les mécanismes des algorithmes sont laissés à eux-mêmes pour trouver et exposer la structure captivante des données [10] .
On peut classer les problèmes d'apprentissage non supervisés en problèmes de clustering

et d'association. Voici quelques exemples d'algorithmes non supervisés d'apprentissage automatique :

• K-means (K-moyen).

• Minimisation de la taille .

• Regroupement hiérarchique.

Apprentissage par renforcement

L'approche de l'IA appelée apprentissage par renforcement se concentre sur l'acquisition de connaissances par le système à travers ses interactions avec son environnement. En utilisant la méthode de renforcement de l'apprentissage, le système ajuste ses paramètres en fonction des réactions de l'environnement.
, qui donne ensuite un retour d'information concernant les choix effectués. À titre d'exemple, un système qui simule un joueur d'échecs qui exploite les résultats des étapes précédentes pour améliorer ses performances est un système qui utilise le renforcement pour apprendre. La recherche en cours sur l'apprentissage avec renforcement est très interdisciplinaire et implique des chercheurs qui se spécialisent dans les algorithmes génétiques, les réseaux de neurones, la psychologie et les méthodes de contrôle. [26]

2.7 7 Synthèse sur le Machine Learning

L'apprentissage automatique (ML) est une discipline clé de l'intelligence artificielle quipermet aux systèmes informatiques d'apprendre et de s'améliorer à partir de l'expérience sansêtre explicitement programmés. En utilisant des algorithmes qui détectent des modèles dansles données, la machine Learning transforme les données brutes en informations exploitablespour des tâches telles que la classification, la régression et la prédiction.

Des domaines tels que la vision par ordinateur (reconnaissance d'images), le traitement du langage naturel (analyse de sentiments, traduction automatique), les systèmes de recommandation (personnalisation du contenu), et bien d'autres encore sont couverts par les applications de la machine Learning. Le fait qu'il puisse s'adapter à de nouvelles données et automatiser des processus complexes en fait un outil indispensable pour l'innovation

et l'efficacité dans de multiples domaines [27].

2.8 L'apprentissage en profondeur

L'apprentissage profond repose sur l'idée des réseaux de neurones artificiels et est conçu pour gérer de grandes quantités de données en intégrant des couches au réseau. Un modèle d'apprentissage profond peut extraire des caractéristiques à partir des données brutes en utilisant différentes couches de traitement qui incluent des transformations linéaires et non linéaires [**?**].

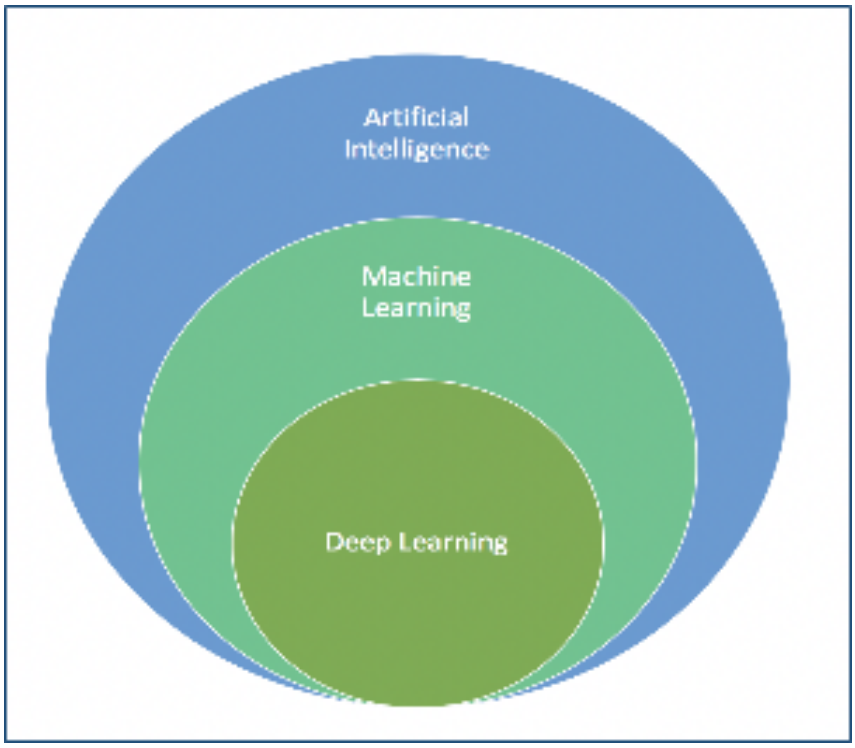


FIGURE 2.6 – Intelligence artificielle, machine learning et deep learning

2.8.1 Pourquoi le choix deep Learning

Premièrement, les divers algorithmes de Deep Learning n'ont émergé qu'après l'échec de l'apprentissage automatique qui cherche à résoudre une grande diversité de problèmes de l'intelligence artificielle (IA) [7] :

— Pour améliorer le développement des algorithmes traditionnels dans de telles tâches de l'IA [7].

— Pour développer une grande quantité de données telles que les big data.

— Pour s'adapter à n'importe quel type de problème.

— Pour extraire automatiquement les caractéristiques.

2.9 Reseau de noronnes

L'ANN, également connu sous le nom de réseau de neurones artificiels, a été introduit comme un modèle basique de traitement de l'information dans le cerveau humain. Par conséquent, la base d'un ANN consiste en un réseau de petits nœuds de calcul reliés entre eux par des liens dirigés et pondérés Les neurones sont symbolisés par les nœuds, tandis que les liens pondérés symbolisent la puissance des connexions synaptiques qui les relient les uns aux autres.

Le neurone a donc la capacité de représenter les potentiels des signaux synaptiques qui lui parviennent, et il transmet ensuite une information basée sur cette somme grâce à une fonction de transfert de préférence non linéaire.

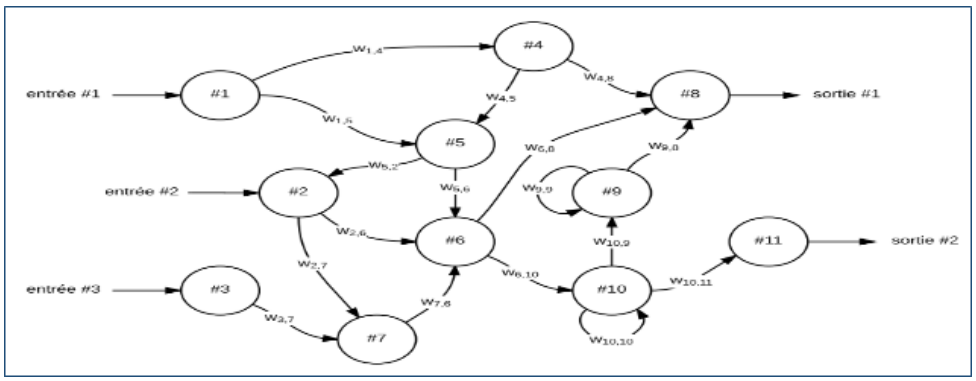


FIGURE 2.7 – Représentations d'un réseau de neurones artificiels (ANN)

En activant un ANN, on injecte des données au niveau de tous ou d'une partie des nœuds, puis on diffuse l'information en suivant les liens pondérés. Une fois que l'information est diffusée, on peut recueillir les niveaux d'activation de tous ou d'une partie des nœuds et les utiliser comme commande. La sortie est ensuite transmise aux autres nœuds de calcul.

2.10 Les architectures du Deeplearning

La structure d'un réseau dépend de l'algorithme d'apprentissage que vous avez l'intention d'utiliser. En général, nous pouvons identifier 3 classes de réseaux :

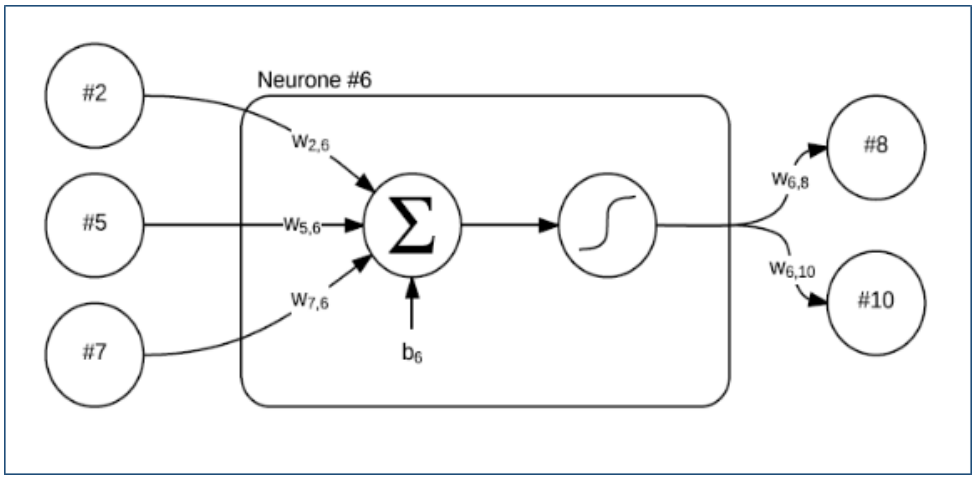


FIGURE 2.8 – Structure interne d'un nœud de calcul.

2.10.1 Réseaux à une couche (feed-forward)

Dans ce réseau en couches simple, on dispose de nœuds d'entrée et d'une couche de neurones (couche de sortie), le signal se propage dans le réseau de manière linéaire, à partir de la couche d'entrée et se terminant par la couche de sortie, sans connexions de retour et sans connexions transversales dans la couche de sortie.[4].

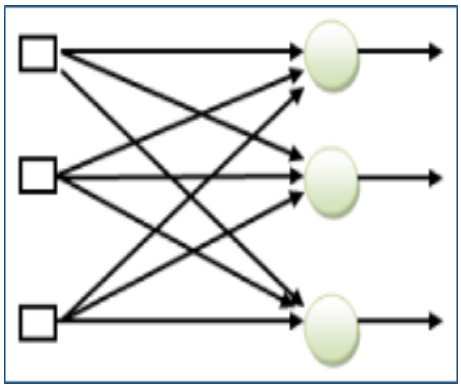


FIGURE 2.9 – Réseaux à une couche

2.10.2 Réseaux multi-couches (feed-forward)

La classe de réseaux feed-forward présente une différence par rapport à la précédente en raison de la présence d'une ou plusieurs couches de neurones cachés entre les couches d'entrée et de sortie. Les connexions entre chaque couche et la couche suivante sont linéaires, ce qui signifie que le signal se propage de manière linéaire, sans cycles et sans connexions transversales. Ce genre d'architecture permet au réseau d'avoir une vision

globale car il favorise les échanges entre les neurones.

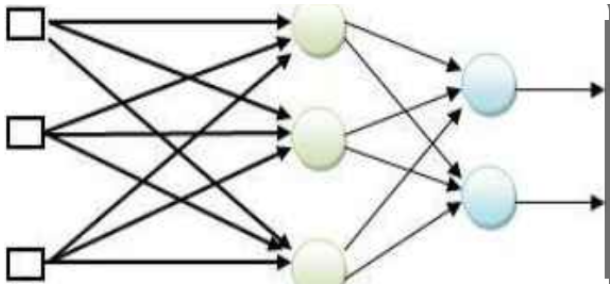


FIGURE 2.10 – Figure : Réseaux à une couche

Réseaux récurrents (feed-back)

Le réseau cyclique diffère du précédent en raison de sa périodicité. La présence de cycles a une influence significative sur la capacité d'apprentissage et les performances du réseau, notamment en dynamisant le système.

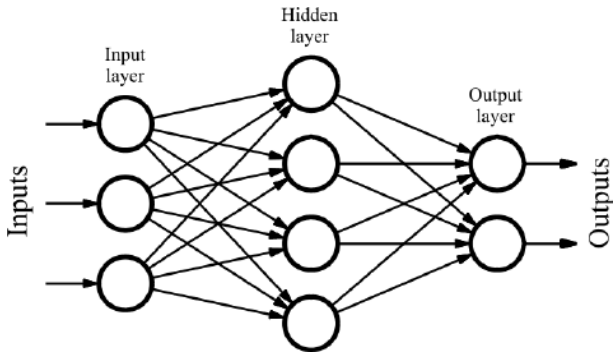


FIGURE 2.11 – Figure : Réseaux à une couche

2.11 Les types de Deep Learning

Le Deep Learning évolue rapidement, avec de nouvelles architectures, variantes ou algorithmes qui se présentent chaque semaine. Toutefois, les performances de ces algorithmes ne peuvent pas toujours être comparées car elles ne sont pas évaluées sur le même ensemble de données.

2.11.1 Perceptron multi-couches (MLP)

Nous commencerons donc par présenter le réseau de neurones le plus classique : le perceptron multi-couches (MultiLayer Perceptron en anglais ou MLP). Ce réseau de neurones acyclique (Feed-Forward Neural Network - FFNN) est structuré en couches. Ainsi, un perceptron multi-couches est composé d'une couche d'entrée, d'une ou plusieurs couches intermédiaires dites cachées et d'une couche de sortie. Chaque couche est connectée à tous les nœuds de la couche précédente. donne une représentation d'un perceptron multi-couches avec deux couches .

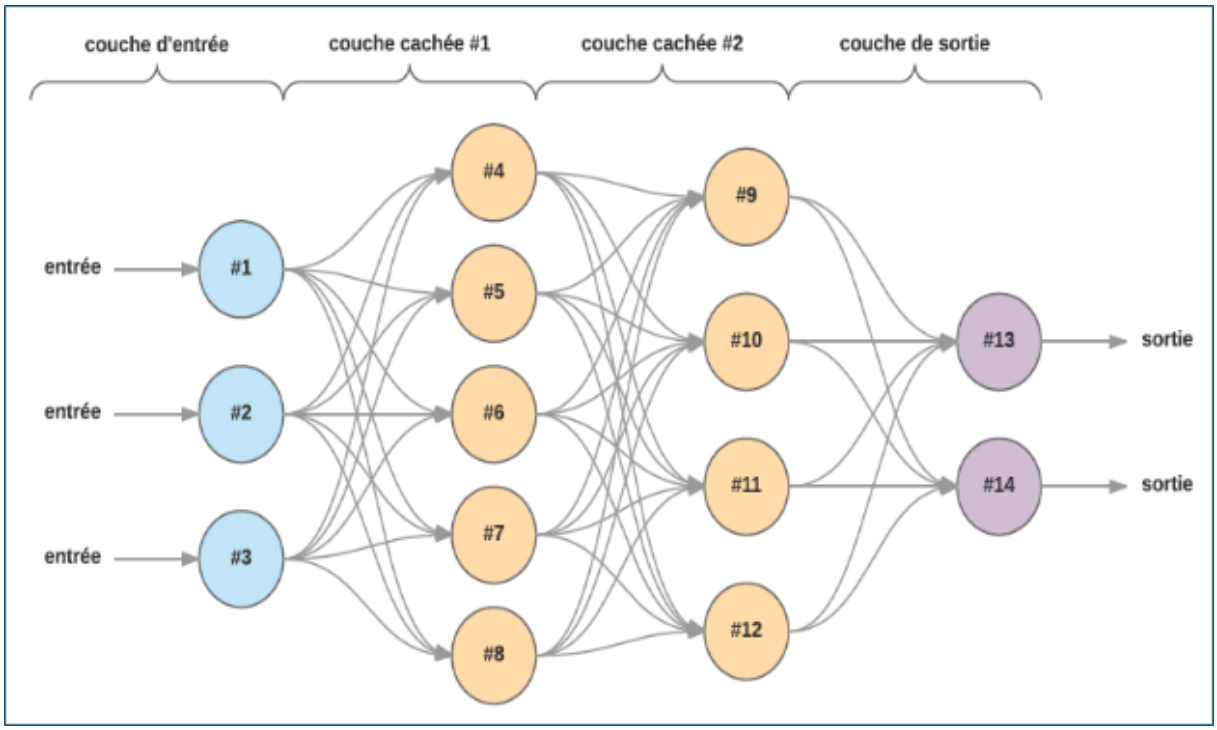


FIGURE 2.12 – Réseau de neurones dense, acyclique et structuré en couches communément appelé perceptron multi-couches (MLP)

2.11.2 Réseaux de neurones profonds

Les réseaux de neurones dits "profonds" (Deep Neural Networks en anglais) sont des perceptronsmulti-couches avec un nombre de couches supérieur à trois. Pendant longtemps, la méthode d'apprentissage de ce type de réseau de neurones acyclique ne permet pas la convergence vers un réseau de neurones puissant. Grâce à des progrès importants dans les techniques d'apprentissage et dans le choix de la fonction de transfert d'unités

linéaires rectifiées (ReLU), l'effet de la dilution du gradient dans les couches inférieures du réseau est réduit, ce qui permet d'utiliser davantage de réseaux de neurones[4].

2.11.3 Réseau de neurones récurrents

Le concept de RNN repose sur l'utilisation d'informations séquentielles. Au sein des réseaux de neurones classiques, nous considérons que toutes les entrées (et sorties) sont autonomes. Cependant, cela peut être une très mauvaise idée pour de nombreuses tâches. Pour prédire le prochain mot dans une phrase, il est nécessaire de connaître les mots précédents. On nomme récurrents les RNN, car ils effectuent la même tâche pour chaque élément d'une séquence, la sortie étant influencée par les calculs précédents[4]. Une autre perspective sur les RNN est qu'ils possèdent une "mémoire" qui retient les informations sur les calculs effectués jusqu'à présent. En théorie, les RNN ont la capacité d'utiliser des données dans des séquences infiniment longues, mais dans la réalité, on ne les limite qu'à observer quelques étapes en suivant. On l'emploie pour :

— La création de modèles linguistiques et la création de texte
— La traduction automatique
— La reconnaissance vocale
— La description des images

2.11.4 Les réseaux de neurones convolutifs

Le réseau neuronal convolutif (CNN) est un réseau spécialisé de neurones utilisé pour traiter les données avec une structure en forme de grille. Ces techniques ont été très efficaces dans les domaines de la reconnaissance et de la classification d'images et de vidéos. Les visages, les objets, les panneaux de circulation et l'auto-conduite des voitures ont été identifiés grâce à la CNN . Les CNN ont récemment démontré leur efficacité dans diverses tâches de traitement du langage naturel (comme la classification des phrases) . L'un des types de réseaux de neurones feed-forward est un réseau convolutif, qui comprend une ou plusieurs couches de neurones cachés entre les couches d'entrée et de sortie. La connexion entre chaque couche et la couche suivante est linéaire, sans cycles et sans interconnexions, ce qui permet une propagation du signal de manière linéaire. Ce genre d'organisation donne au réseau une vision globale car elle favorise les échanges entre les neurones [10].

2.12 Exemples d'application de Deep Learning

Le Deep Learning est employé dans différents domaines, allant de la conduite automatisée aux dispositifs médicaux. Grâce au deep Learning, il est désormais possible d'ajouter des sons à des films qui restent silencieux [21].

— Effectuer une traduction automatisée.

— Classer les objets en utilisant des photographies.

— Produire des écrits automatiques.

— Création d'une légende visuelle.

— Jeux automatiques.

2.13 Synthèse sur le Deep Learning

Le Deep Learning, une branche avancée du machine Learning, repose sur des réseaux deneurones artificiels profonds pour modéliser des données complexes. Sa force réside dans sacapacité à apprendre des représentations hiérarchiques et abstraites directement à partir desdonnées brutes, rendant possible des progrès significatifs dans des domaines comme la visionpar ordinateur, le traitement du langage naturel, et l'analyse de données complexes. Le DeepLearning a révolutionné la capacité des machines à percevoir et interpréter lesdonnées de manière autonome, surpassant souvent les méthodes traditionnelles en termes deprécision et d'efficacité pour les tâches complexes. Cependant, il pose aussi des défis entermes de besoins en calcul, interprétabilité des modèles, et gestion des données massives.

2.14 Conclusion

Dans ce chapitre, nous avons défini l'intelligence artificielle, Machine Learning et DeepLearning et la différence entre eux. Nous avons montré les déférents types d'apprentissage automatique, ensuite nous avons expliqué que les réseaux de neurones artificiels sont calculatoires et sont des systèmes inspirés par les processus biologiques qui se produisent dans le cerveau humain, et on a montré ces architectures et les déférents types des réseaux de neurones. Et enfin on a cité quelques exemples d'application de Deep Learning.

Chapitre 3

Apprentissage automatique dans le contrôle d'accès

3.1 Introduction

Les méthodes de preuve de modèle et de structure de données acceptent initialement la politique en cours de vérification comme irréprochable jusqu'à ce qu'elle soit testée par rapport à des cas ou des exigences spécifiés. Les défauts sont ensuite revendiqués si le modèle de politique ou la structure de données ne peut pas tenir. Par conséquent, afin de détecter toutes les failles possibles embarquées dans une politique, l'enjeu est de composer des cas de test ou des exigences capables de les découvrir. D'autres méthodes similaires, cependant, peuvent s'appuyer sur un oracle de test qui contient toutes les demandes de contrôle d'accès et les autorisations d'accès possibles en entrée pour vérifier si les autorisations des demandes entrent en conflit avec les autorisations attendues attribuées aux règles. Par conséquent, un temps de calcul ou des ressources excessifs sont nécessaires pour un grand nombre d'attributs de stratégie. Par exemple, étant donné n, 2n+1 demandes d'accès (1 pour les états d'autorisation) sont requises pour l'oracle de test. La plupart des politiques de contrôle d'accès actuelles peuvent facilement avoir des centaines d'attributs. Par conséquent, les oracles de test seront trop grands pour être effectués en pratique. En plus de la preuve du modèle et des données structurées, un système simulé est conçu pour simuler les règles de la politique de contrôle d'accès à des fins de vérification. Dans de tels systèmes, chaque règle de stratégie est représentée par des composants système simulés de sorte que les erreurs peuvent être détectées en déclenchant les fonctions système

qui provoquent des erreurs. avant d'ajouter des circuits de règles qui compliquent encore l'effort de détection. En d'autres termes, au lieu de vérifier en retraçant les

interrelations entre les règles une fois la stratégie terminée, il vérifie uniquement la nouvelle règle ajoutée par rapport aux règles précédentes « correctes ». Même si cette méthode ne nécessite pas de grands nombres complets[28].

3.2 Machine Learning pour la vérification du contrôle d'accès

L'apprentissage automatique a été utilisé pour le contrôle des appareils, l'analyse des systèmes et les prévisions commerciales. Cependant, l'utilisation du ML pour les tests logiciels n'en est qu'à ses débuts. En particulier, les applications de ML pour la preuve de modèle sont encore insaisissables et doivent encore être explorées. La classification ML permet de générer ou de prédire des classes cibles pour de nouvelles données d'entrée à l'aide d'exemples de données d'entraînement produites par l'exécution du système au lieu de données d'entrée complètes. Pour la vérification de la stratégie de contrôle d'accès, les données d'entraînement se voient attribuer les valeurs des attributs des règles de stratégie, et la cible de la classification est les autorisations d'accès (par exemple, accorder et refuser) affectées aux règles. Les données sont consommées par un algorithme de classification ML pour générer un modèle de classification. L'interrogation de l'exactitude des données d'entraînement par rapport au modèle permet de détecter les incohérences, en indiquant les erreurs détectées parmi les règles de stratégie. En outre, pour vérifier les règles de stratégie nouvelles ou mises à jour, elles doivent être contenues dans les données de test pour une analyse de précision supplémentaire des données de d'essai[28].

3.3 Intégration de ML dans les systèmes de CA

L'intégration du machine learning (ML) dans les politiques de contrôle d'accès (CA) révolutionne la gestion des accès dans les systèmes d'information modernes[3].

3.4 Une taxonomie et une enquête

est une enquête sur l'utilisation de l'apprentissage automatique (machine learning, ML) dans le contrôle d'accès. Elle est organisée en différentes branches, chacune représentant une catégorie ou un aspect spécifique du sujet. (figure ...) Voici une explication des principaux éléments[3]. :

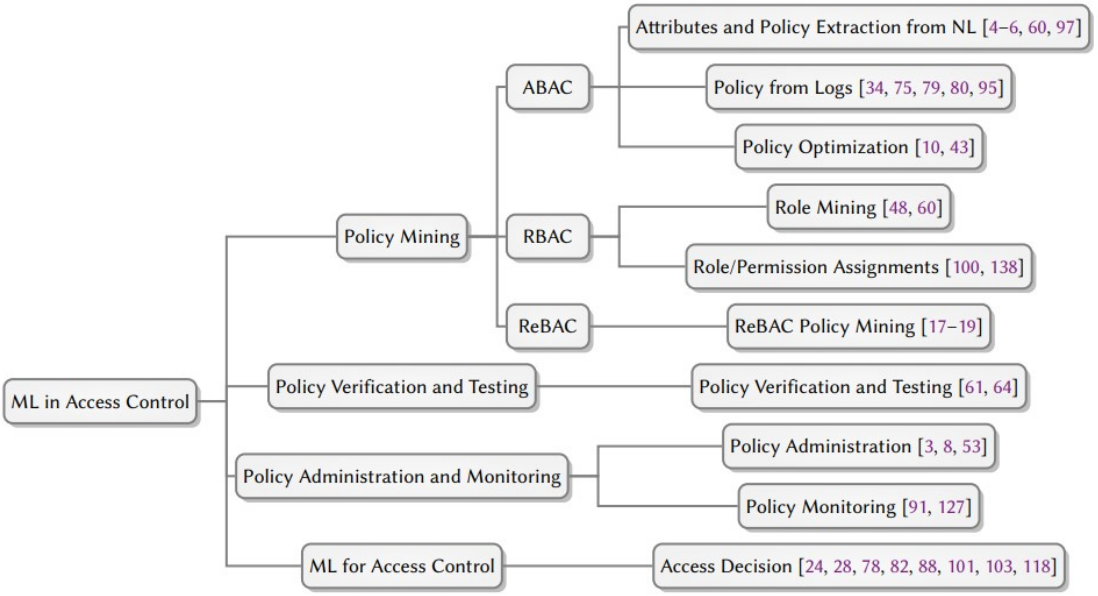


FIGURE 3.1 – taxonomie et enquête de M/L pour le contrôle d'acces

3.4.1 ML in Access Control (Apprentissage automatique dans le contrôle d'accès)

Policy Mining (Extraction de politiques)

C'est le sujet principal, et les différentes branches explorent comment l'apprentissage automatique peut être appliqué dans divers aspects du contrôle d'accès, Elle se divise en plusieurs sous-catégories[3] :

ABAC (Attribute-Based Access Control)

1-Extraction d'attributs et de politiques à partir du langage naturel : Utilise le ML pour extraire des attributs et des politiques d'accès depuis le texte en langage naturel

2. Politiques à partir de journaux : Génère des politiques basées sur l'analyse des journaux d'activité.

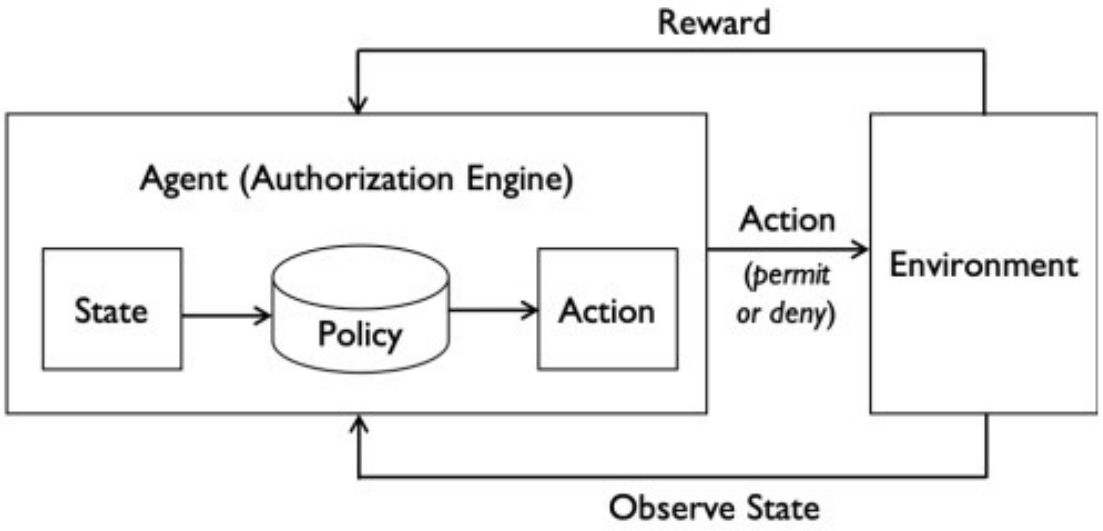


FIGURE 3.2 – Présentation de l'apprentissage des politiques ABAC basé sur l'apprentissage par renforcement

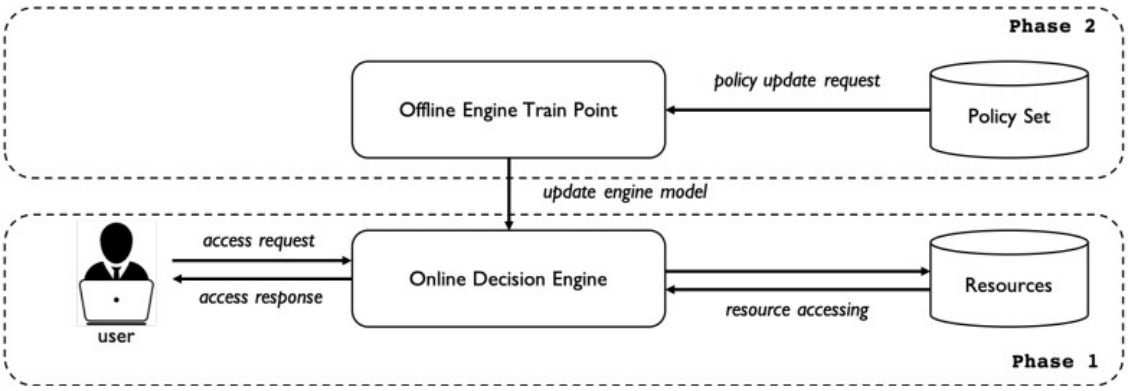


FIGURE 3.3 – Méthode de décision d'autorisation d'EPDE-ML..

3. Amélioration des politiques : Employe le machine learning afin d'améliorer les politiques d'accès déjà en place.

RBAC (Role-Based Access Control)

Extraction de rôles : Identifie des rôles basés sur l'analyse des permissions d'accès des utilisateurs. Affectation de rôles/permissions : Utilise le ML pour assigner des rôles et des permissions.

ReBAC (Relationship-Based Access Control)

Extraction de politiques ReBAC : Génère des politiques basées sur les relations entre les entités.

Policy Verification and Testing (Vérification et test des politiques)

Utilise le ML pour vérifier et tester l'efficacité et la conformité des politiques d'accès existantes.

Administration et surveillance des politiques

Administration des politiques : Applique le ML pour aider à gérer les politiques d'accès. Surveillance des politiques : Utilise le ML pour surveiller et détecter des anomalies ou des violations dans l'application des politiques.

ML for Access Control ML pour le contrôle d'accès

Utilise le ML pour prendre des décisions d'accès basées sur des modèles prédictifs et des données contextuelles[3] .

Illustration

En résumé, cette taxonomie montre comment l'apprentissage automatique est intégré dans différents aspects du contrôle d'accès, depuis l'extraction et l'optimisation des politiques jusqu'à la vérification, l'administration, la surveillance et la prise de décisions d'accès. Chaque branche représente un domaine où le ML apporte une valeur ajoutée en améliorant l'efficacité et la précision des systèmes de contrôle d'accès.

Modèle de sous-arbres généré par le classificateur RFC

Les arbres de décision (DT) et les forêts aléatoires (RFC) sont deux algorithmes de classification majeurs en machine learning utilisés pour vérifier les politiques de contrôle d'accès. Contrairement à d'autres algorithmes de classification mieux adaptés à l'analyse de régression pour les données numériques, les algorithmes DT et RFC utilisent des structures arborescentes binaires efficaces pour traiter l'analyse de données binaires non régressives. La figure 1 illustre un modèle d'arbre binaire généré par l'algorithme RFC.

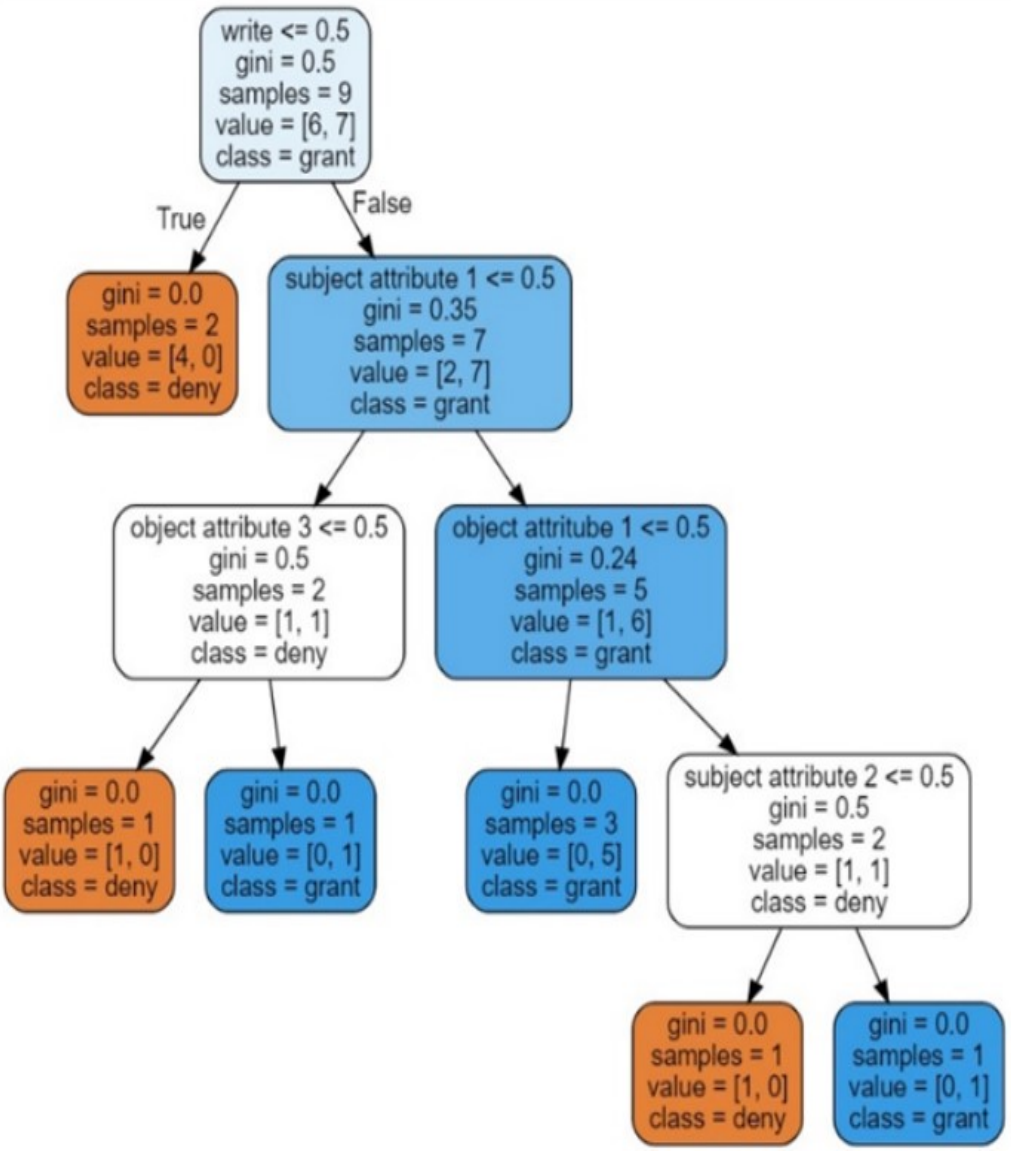


FIGURE 3.4 – 1a (ci-dessus) 1b (ci-dessous) – Modèle de sous-arbres généré par le classificateur RFC

Interprétation de l'Arbre

— Racine (write ≤ 0.5) :
  — Si True : Accès refusé (class = deny) avec 4 échantillons.
  — Si False : Continue l'évaluation avec `subject attribute 1` $\leq$ `0.5`.
— Sous-nœuds :
  — Par exemple, pour `subject attribute 1` $\leq$ `0.5` :
    — Si True : Continue l'évaluation avec `object attribute 3` $\leq$ `0.5`.
    — Si False : Continue l'évaluation avec `object attribute 1` $\leq$ `0.5`.
— Feuilles :
  — Les branches finissent par des feuilles qui classifient soit en `grant` soit en `deny`. Par exemple, `object attribute 1` $\leq$ `0.5` se divise en trois feuilles, déterminant `grant` ou `deny` selon les attributs.

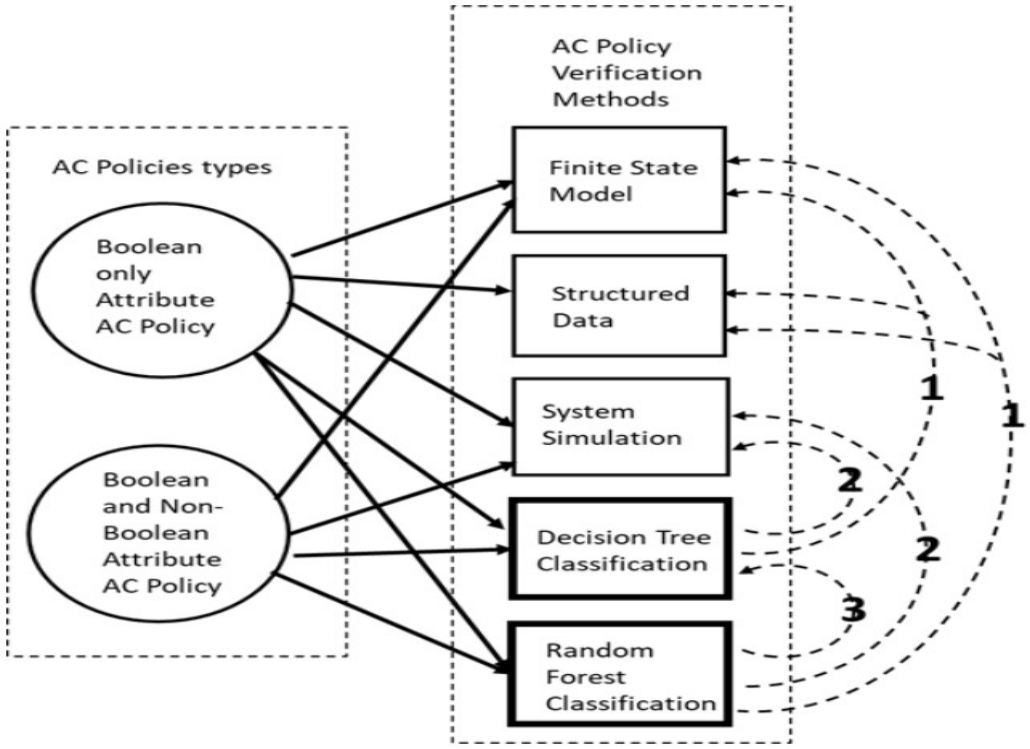
Cet arbre est utilisé pour décider l'accès en évaluant séquentiellement les conditions basées sur les attributs des sujets (utilisateurs) et des objets (ressources). En suivant les branches selon les valeurs des attributs, on détermine si l'accès est accordé (`grant`) ou refusé (`deny`).

Comparaison RFC/ DT

En résumé, les avantages de l'algorithme RFC par rapport à l'algorithme DT et aux méthodes de vérification traditionnelles pour valider les politiques de contrôle d'accès sont illustrés à la figure 2. Les chiffres associés aux lignes pointillées identifient les défis suivants auxquels répondent les méthodes de classification ML représentées par les carrés :

1. Nécessité d'un cas de test/exigences ou d'un oracle pour détecter toutes les erreurs potentielles de stratégie.
2. Difficulté de mise en œuvre et de mise à jour du système simulé.
3. Impossibilité de séparer les règles de stratégie du modèle avec un risque de sur-ajustement aux valeurs d'attributs non binaires.

Les lignes continues indiquent les méthodes de vérification traditionnelles ou les algorithmes de classification ML pertinents pour les stratégies de contrôle d'accès reliées par un cercle aux types d'attributs spécifiés.



1. Lignes Pointillées : Correspondent à des défis spécifiques.
2. Carrés : Représentent les solutions de classification ML.
3. Lignes Continues : Reliées aux méthodes traditionnelles et algorithmes applicables.

3.4.2 Applications de ML par Type de Politique de CA

L'intégration du machine learning (ML) dans les politiques de contrôle d'accès (CA) révolutionne la gestion des accès dans les systèmes d'information modernes. Cette enquête se penche sur l'application de l'apprentissage automatique dans les principaux modèles d'AC en identifiant les avantages obtenus, les défis rencontrés et les solutions proposées.

Type de Politique de CA	Applications de ML	Avantages	Défis	Solutions
ACL	• Utilisation de la classification pour prédire le décisions d'accès • Détection des anomalies en identifiant des schémas d'accès inhabituels	• Amélioration de l'efficacité • Renforcement de la sécurité	• Données bruyantes ou incomplètes	• - Prétraitement des données pour améliorer la qualité
DAC	• Utilisation de la classification pour détermine les permissions d'accès • Surveillance en temps réel pour détecter les comportements anormaux	• Flexibilité accrue • Réduction des erreurs humaines	• Complexité des modèles ML	• Adoption de modèles explicables et réduction de la dimensionnalité
RBAC	• Utilisation du clusterin pour regrouper les utilisateurs en rôles. • Gestion dynamiqu des politiques d'accès	• Adaptation dynamique aux changements • Réactivité face aux nouvelles menaces	• Scalabilité pour gérer de grands volumes de données	• Algorithmes parallélisés et traitements distribués
ABAC	• Utilisation du deep learning pour analyser les attributs prendre des décisions d'acc • Régression pour prédire les niveaux de risque	• Optimisation des décisions en temps réel • Surveillance continue des comportements d'accès	• Intégration avec les systèmes existants	• Développement d'API standardisées et plateformes modulaires

3.4.3 Examen bref des approches

Les recherches actuelles montrent les bénéfices de l'application de modèles de machine learning dans le contrôle d'accès où les décisions sont prises par un modèle ML formé plutôt que par des politiques écrites. Typiquement, ces modèles utilisent les métadonnées et attributs des utilisateurs et des ressources pour déterminer si l'accès doit être accordé ou refusé. Ces métadonnées et attributs sont les caractéristiques apprises par le modèle pour prendre des décisions futures. Nous examinons brièvement ces approches ci-dessous et les résumons dans le tableau ci-dessous[3] .

Reference	Application	Problem Considered	Access Control Model	ML Approach	Dataset Type
Chang et al. 2006 [28]	Not specified	A novel access control model with time constraint	Time-constraint Access Control	SVM	Syn
Outchakoucht et al. 2017 [103]	IoT	Blockchain based access control policy	Blockchain based Access Control	Reinforcement Learning	No Evaluation
Cappelletti et al. 2019 [24]	Not specified	Inferring ABAC policies from access logs	ABAC	DT, RF, SVM, MLP	RW (SL: 1.4, 1.8, 1.13)
Khilar et al. 2019 [82]	Cloud Computing	Policy for cloud resources based on the access history and behaviour	Trust-Based Access Control	RF, DT, SVM, Neural Network, etc.	Not Specified
Srivastava et al. 2020 [118]	Defense, airport, and healthcare	A novel access control framework to decide accesses based on the genuineness of the requester	Risk Adaptive Access Control (RAdAC)	Neural Network, RF	Not Specified
Liu et al. 2021 [88]	Big Data & IoT	Improves the policy decision point (PDP) of the ABAC model	ABAC	RF	RW (SL: 1.8)
Karimi et al. 2021 [78]	IoT	Adaptive ABAC policy learning	ABAC	Reinforcement Learning	Syn & RW (SL: 1.8)
Nobi et al. 2022 [101]	Not specified	A deep neural network based access control model	DLBAC	ResNet	Syn & RW (SL: 1.4, 1.8)

FIGURE 3.5 – Résumer l'apprentissage automatique pour la décision de contrôle d'accès

3.5 Conclusion

On conclut ce chapitre par un résumé des avantages :

— Automatisation et réduction de la charge administrative

— Détection proactive et sécurité renforcée

— Adaptation et flexibilité en temps réel

Chapitre 4

Méthodologie & résultats

4.1 Introduction

Dans les grandes entreprises comme Google et Amazon, la gestion efficace des demandes d'accès aux ressources par les employés est essentielle pour assurer la sécurité et le bon fonctionnement des opérations. Avec des environnements organisationnels complexes, il devient impératif d'automatiser le processus de gestion des demandes d'accès. Ce chapitre explore le développement d'un système automatisé de contrôle d'accès des employés à l'aide d'algorithmes d'apprentissage machine.

Dans ce chapitre, nous explorons le développement d'un système automatisé de contrôle d'accès des employés en utilisant des techniques d'apprentissage machine. L'objectif est de prédire si une demande d'accès à une ressource d'un employé doit être approuvée ou refusée.

4.2 Implémentation

4.2.1 Environnement et Outils

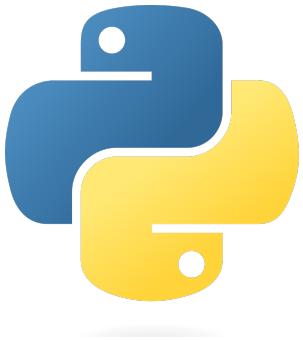
Matérielles On a utilisé le GPU NVDIA Tesla T4 qu'est sur la plateforme Colab de Google

processeur : NVDIA Tesla T4 a 2560 Cores avec 16 GB de RAM

RAM : 16GB

Logiciels

— **Langage de programmation Python :**



Python, en tant que langage de programmation interprété, orienté objet et de haut niveau, se prête parfaitement au développement d'un système de contrôle d'accès efficace. Sa syntaxe claire et facile à apprendre favorise la lisibilité du code, réduisant ainsi les coûts de maintenance associés aux systèmes complexes de gestion des accès. Python est largement utilisé dans le développement rapide d'applications (RAD), offrant un cycle d'édition-test-débogage rapide grâce à l'absence d'étape de compilation.

— **Environnement Google Colab :**

c'est révélé être un outil précieux pour le développement efficace du projet. Son infrastructure basée sur le cloud nous a permis d'accéder à des ressources informatiques puissantes, y compris des GPU, ce qui a considérablement accéléré nos expériences d'apprentissage automatique pour l'analyse des sentiments et le traitements de texte exactement l'entraînement des modèles.

Le modèle gratuit de Colab a éliminé le besoin d'investissements initiaux dans l'infrastructure matérielle, nous permettant de nous concentrer sur les activités de recherche et développement. De plus, la facilité de partage des notebooks a facilité la collaboration au sein de notre équipe, permettant une itération efficace du code et un transfert de connaissances .[29]

4.2.2 Les bibliothèques utilisées

— **NumPy (Numerical Python)**

est une bibliothèque Python fondamentale pour le calcul scientifique. Elle offre un ensemble complet de structures de données et de fonctions pour des opérations numériques efficaces, ce qui en fait un outil essentiel pour les scientifiques, les ingénieurs, les analystes de données et les praticiens de l'apprentissage automatique .[30]

En résumé, les utilisations principales de NumPy tournent autour de sa capacité à gérer et manipuler des tableaux multidimensionnels de manière performante, ce qui en fait un outil incontournable pour le calcul scientifique dans de nombreux domaines.

— **Pandas**

Pandas est une bibliothèque Python open-source conçue pour l'analyse et la manipulation de données. Elle offre des structures de données performantes et des outils intuitifs pour simplifier les tâches complexes liées au traitement de données volumineuses. Pandas est devenue un outil incontournable pour les data scientists, les analystes de données et les développeurs qui travaillent avec des données en Python.

En intégrant Pandas à notre workflow Python, nous gagnons un outil puissant et polyvalent pour l'analyse et la manipulation de données. Sa facilité d'utilisation, ses fonctionnalités riches et sa communauté active en font un choix incontournable pour notre équipe de data scientists, analystes de données et développeurs qui travaillent avec des données en Python.

— **matplotlib**

Matplotlib est une bibliothèque Python largement utilisée pour la création de visualisations de données de haute qualité. Elle offre un ensemble complet d'outils pour générer divers types de graphiques, notamment des courbes, des histogrammes, des diagrammes à barres, des camemberts et des cartes thermiques.

Matplotlib est un outil essentiel pour les data scientists, les analystes de données,

les ingénieurs et les chercheurs qui souhaitent communiquer efficacement des informations à partir de données complexes .[31]

— **Scikit-learn (SkLearn)**

La bibliothèque Scikit-learn, également connue sous le nom de sklearn, est l'une des ressources les plus précieuses pour les praticiens de l'apprentissage automatique. Développée en Python, cette bibliothèque open source offre une panoplie d'outils et d'algorithmes pour la modélisation des données, la classification, la régression, le clustering et bien d'autres tâches liées à l'apprentissage supervisé et non supervisé.

4.3 Implémentation

4.3.1 Aperçu sur le dataset

Les données se composent de données historiques réelles collectées en 2010 et 2011. Les employés sont manuellement autorisés ou refusés l'accès aux ressources au fil du temps. Vous devez créer un algorithme capable d'apprendre à partir de ces données historiques pour prédire l'approbation ou le refus pour un ensemble d'employés non vus.

4.3.2 Pseudo code

Le pseudo code ci-dessous montre les étapes de la réalisation de notre travail. [1] **Initialisation du Processus** *Étape 1 : Chargement des données Étape 2 : Exploration des données Étape 3 : Prétraitement des données Étape 4 : Construction des modèles Étape 5 : Évaluation des modèles*

Étape 1 : Chargement des Données `Charger le jeu de données contenant les informations sur les employés et leur accès aux ressources.`

Étape 2 : Exploration des Données `Visualiser les premières lignes du jeu de données pour comprendre sa structure. Afficher des statistiques descriptives pour chaque variable pour obtenir un aperçu. Représenter graphiquement la distribution des autorisations d'accès accordées et refusées.`

Étape 3 : Prétraitement des Données `Supprimer les variables redondantes`

TABLE 4.1 – Description des données des employés

Action	ACTION est 1 si la ressource a été approuvée, 0 si la ressource n'a pas été approuvée
Ressource	Un ID pour chaque ressource
ID du manager	L'ID EMPLOYÉ du manager de l'enregistrement ID EMPLOYÉ actuel ; un employé peut n'avoir qu'un seul manager à la fois
Catégorie de rôle 1	ID de catégorie de regroupement des rôles de l'entreprise 1 (par exemple, Ingénierie US)
Catégorie de rôle 2	ID de catégorie de regroupement des rôles de l'entreprise 2 (par exemple, Commerce de détail US)
Nom du département	Description du département du rôle de l'entreprise (par exemple, Commerce de détail)
Intitulé du poste	Description de l'intitulé du poste dans l'entreprise (par exemple, Directeur principal de l'ingénierie commerciale)
Description de la famille de rôles	Description étendue de la famille de rôles de l'entreprise (par exemple, Directeur du commerce de détail, Ingénierie logicielle)
Famille de rôles	Description de la famille de rôles de l'entreprise (par exemple, Directeur du commerce de détail)
Code de rôle	Code de rôle de l'entreprise ; ce code est unique pour chaque rôle (par exemple, Directeur)

`ou fortement corrélées pour simplifier le modèle. Diviser les données en ensembles de formation et de test pour évaluer la performance du modèle.`

Étape 4 : Construction des Modèles State `Utiliser différents modèles de machine learning`

Étape 5 : Évaluation des Modèles `Comparer les performances des différents modèles en utilisant des mesures telles que le score F1 et la précision.`

4.3.3 Pré-traitement :

Dans le processus de prétraitement des données pour la modélisation prédictive, la sélection des caractéristiques et la division des données jouent des rôles cruciaux. Premièrement, lorsqu'une corrélation élevée est identifiée entre deux variables, comme c'est le cas entre `ROLE_CODE` et `ROLE_TITLE`, il est souvent judicieux de supprimer l'une de ces variables pour éviter la redondance d'informations similaires. Cette étape de réduction de dimensionnalité contribue à simplifier le modèle sans compromettre sa capacité prédictive.

Une fois que le jeu de données a été préparé en retirant la variable `ROLE_CODE`, nous procédons à la séparation des données en variables dépendantes (`y`) et indépendantes (`x`). Cela permet de définir clairement quelles caractéristiques seront utilisées pour prédire la variable cible `ACTION`, tandis que `y` représente les étiquettes de classe correspondantes pour chaque observation.

4.3.4 Visualisation des données :

Figure 4.1 montre la répartition des employés selon l'accès accordé.

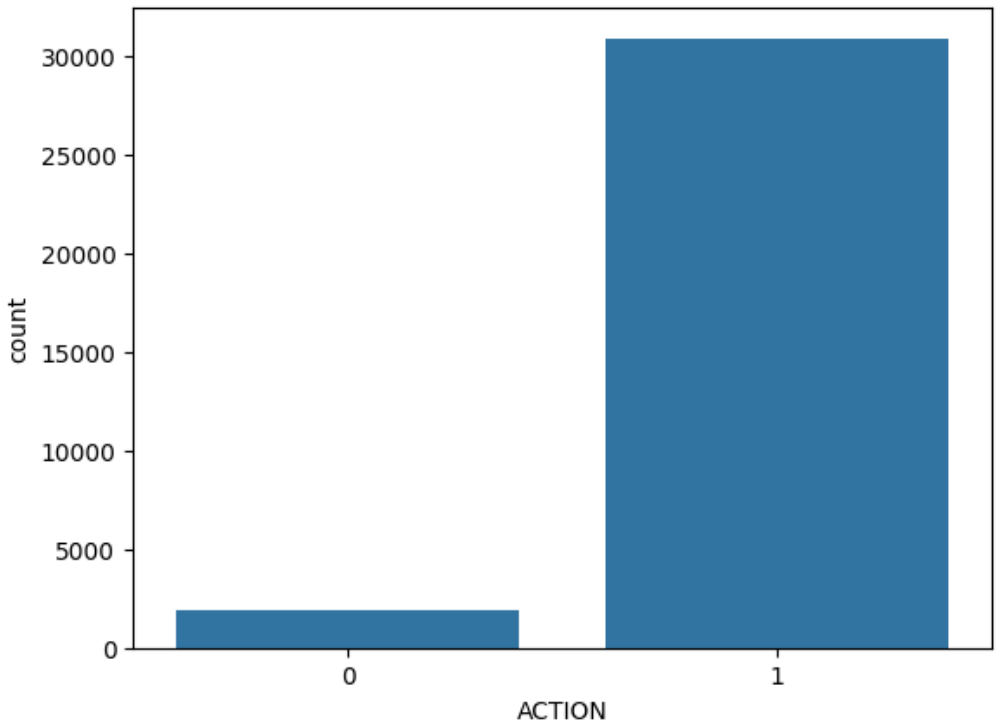


FIGURE 4.1 – Répartition des Employés selon l'Accès Accordé

Figure 4.2 montre les corrélations observées.

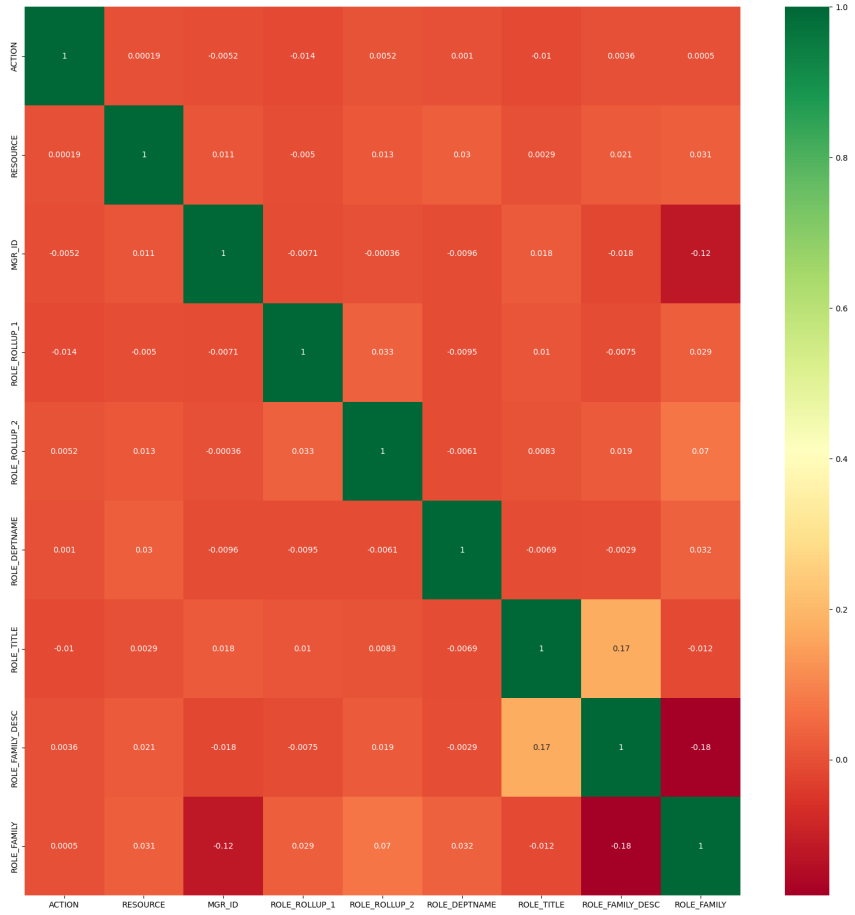


FIGURE 4.2 – Carte Thermique des Corrélations

Figure 4.2 montre une analyse des corrélations observées.

4.3.5 Entraînement :

Régression Logistique : La régression logistique est un algorithme de classification en apprentissage automatique utilisé pour prédire la probabilité d'une variable dépendante catégorielle. Le modèle de régression logistique prédit $P(Y = 1)$ en fonction de X. Il utilise la fonction sigmoïde pour prédire des valeurs de probabilité.

Forêt Aléatoire (Ensemble Bagging) : La forêt aléatoire est un algorithme de classification constitué de nombreux arbres de décision. Il utilise le bagging et la randomisation des caractéristiques lors de la construction de chaque arbre

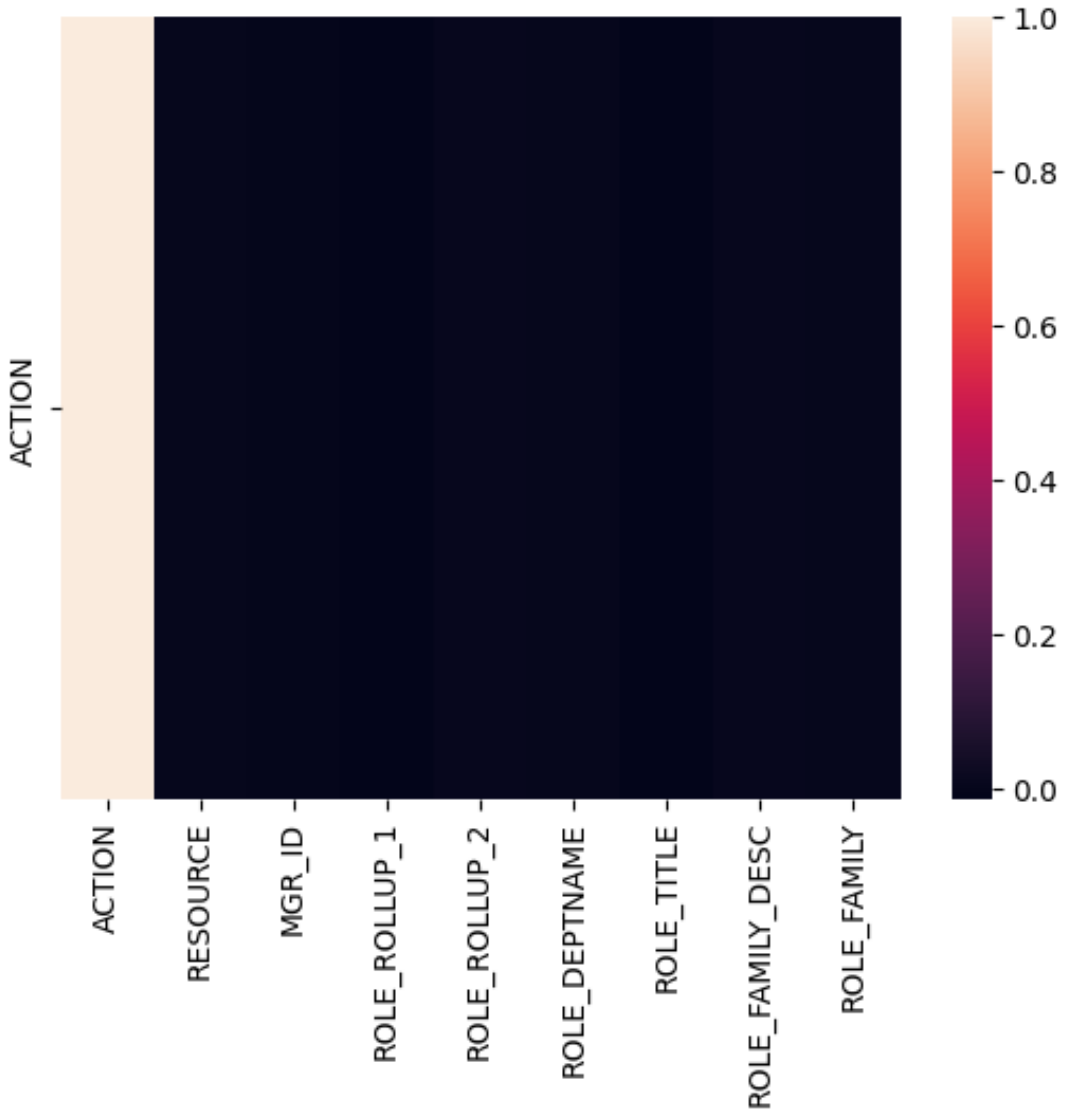


FIGURE 4.3 – Analyse des Corrélation

individuel pour essayer de créer une forêt d'arbres non corrélés dont la prédiction par comité est plus précise que celle de n'importe quel arbre individuel.

K-Nearest Neighbour (KNN) : L'algorithme des k plus proches voisins (KNN) est un type d'algorithme d'apprentissage supervisé qui peut être utilisé pour les problèmes de classification ainsi que de régression. Le KNN est un algorithme non paramétrique (c'est-à-dire qu'il ne fait aucune hypothèse sous-jacente sur la distribution des données). Pour les problèmes de régression, l'algorithme KNN utilise les moyennes des voisins les plus proches plutôt que le vote des voisins les plus proches.

Classificateur de Bayes Naïf : Les classificateurs de Bayes naïfs sont un groupe d'algorithmes d'apprentissage automatique qui utilisent tous le théorème de Bayes pour classer les points de données. Ils sont dits "naïfs" car ils supposent que les caractéristiques d'un point de données sont complètement indépendantes les unes des autres. Les classificateurs de Bayes naïfs utilisent les probabilités de certains événements étant vrais - étant donné que d'autres événements sont vrais - pour faire des prédictions sur de nouveaux points de données. C'est ce qui rend cette formule si unique par rapport aux autres algorithmes de classification en apprentissage automatique.

4.4 Contribution

Au cours de notre projet, on ne s'est pas limité à l'application directe des méthodes du machine learning mentionnées au dessus, mais on a choisi d'améliorer ces techniques en utilisant des approches scientifiques utilisées pas les chercheurs en IA afin d'augmenter les résultats et ça sans chercher à augmenter notre dataset intial.

4.4.1 Grid Search

Grid search constitue une méthode rigoureuse pour le réglage des hyperparamètres dans le cadre de l'apprentissage supervisé, visant à améliorer les performances de généralisation d'un modèle. En utilisant cette technique, toutes les combinaisons possibles des paramètres d'intérêt sont explorées systématiquement afin d'identifier celles offrant les meilleures performances. Autrement dit, il s'agit d'un processus d'optimisation des hyperparamètres, où différentes valeurs pour chaque hyperparamètre sont évaluées de manière exhaustive pour déterminer la configuration optimale du modèle d'apprentissage automatique.[32]
Table 4.2 montre les meilleures valeurs des paramètres utilisés pour l'algorithme du KNN, grâce à la technique de grid search.

TABLE 4.2 – Meilleurs paramètres trouvés par GridSearchCV pour KNN

Paramètre	**Valeur**
Métrique	Manhattan
Nombre de voisins (*n_neighbors*)	20
Poids (*weights*)	Distance

4.4.2 K-folds

La validation croisée k-fold est une procédure de rééchantillonnage employée pour évaluer les modèles d'apprentissage automatique sur un échantillon de données restreint. Ce processus se déroule comme suit :[33]

1. **Division en plis (folds) :** L'ensemble de données est subdivisé en k groupes (plis). Chaque pli est alternativement utilisé comme ensemble de test, tandis que les autres plis constituent l'ensemble d'entraînement.[33]
2. **Itérations :** Le modèle est entraîné k fois, chaque itération utilisant un pli différent comme ensemble de test. Les performances du modèle sont évaluées sur l'ensemble de test correspondant.[33]
3. **Moyenne des performances :** Les performances du modèle sur les k ensembles de test sont moyennées pour obtenir une estimation globale de la performance du modèle.[33]

La validation croisée k-fold est particulièrement utile car elle permet d'évaluer le modèle sur plusieurs ensembles de test, réduisant ainsi le risque d'obtenir des résultats biaisés par hasard. Cette méthode est couramment utilisée pour estimer les performances d'un modèle sur des données non utilisées lors de l'entraînement. Par exemple, avec $k = 5$, on parle de validation croisée à 5 plis.[33]

4.5 Résultats

Table 4.3 donne les résultats obtenus après l'entraînement sur différents modèles.

Modèle	**Score de Test**	**Score d'Entraînement**
2*Régression Logistique	F1 score : 0.9708 Accuracy score : 0.9433	F1 score : 0.9699 Accuracy score : 0.9416
2*Random Forest Classifier	F1 score : 0.9720 Accuracy score : 0.9466	F1 score : 0.9999 Accuracy score : 0.9999
2*Gaussian Naive Bayes	F1 score : 0.9557 Accuracy score : 0.9154	F1 score : 0.9567 Accuracy score : 0.9172
2*KNN	F1 score : 0.9708 Accuracy score : 0.9433	F1 score : 0.9699 Accuracy score : 0.9416
2*KNN Amélioré	F1 score : 0.9710 Accuracy score : 0.9442	F1 score : 1.0 Accuracy score : 1.0

TABLE 4.3 – Scores de performance des modèles de classification

4.6 Conclusion

En conclusion, notre KNN amélioré semble offrir les meilleures performances globales avec des scores élevés tant pour le test que pour l'entraînement, suggérant qu'il pourrait être le choix optimal pour construire un système automatique de contrôle d'accès aux ressources des employés. Cependant, les autres méthodes montrent également des résultats prometteurs et peuvent constituer des alternatives viables en fonction des besoins spécifiques en termes de rapidité d'exécution et d'interprétabilité du modèle.

Conclusion générale

Conclusion

Cette thèse explore l'application de techniques avancées de Machine Learning dans le domaine du contrôle d'accès aux ressources des employés. L'objectif principal était de construire un système automatique capable de prendre des décisions d'approbation ou de rejet des demandes d'accès, en optimisant à la fois la précision et l'efficacité opérationnelle.

Les modèles étudiés, y compris la régression logistique, le Random Forest Classifier, le Naïve Bayes Classifier et le K-Nearest Neighbours, ont été évalués en termes de performances telles que la précision et le F1-score sur des ensembles de données de test et d'entraînement. Chacun de ces modèles présente des caractéristiques uniques en matière de rapidité, d'interprétabilité et de capacité à gérer des données complexes.

Le Random Forest Classifier a émergé comme le modèle le plus performant, démontrant des scores élevés pour la précision et le F1-score sur les ensembles de test et d'entraînement. Cela suggère que ce modèle pourrait être optimal pour une implémentation effective du système de contrôle d'accès dans un environnement réel.

Cependant, la régression logistique et le K-Nearest Neighbours ont également montré des performances prometteuses et pourraient être considérés comme des alternatives viables en fonction des exigences spécifiques du déploiement.

En intégrant ces modèles de Machine Learning dans le cadre du contrôle d'accès, cette thèse a non seulement contribué à renforcer la sécurité des informations sen-

sibles au sein de l'organisation, mais a également ouvert la voie à de nouvelles possibilités d'innovation et d'amélioration continue dans ce domaine crucial de la sécurité informatique.

Pour l'avenir, des perspectives incluent l'amélioration continue des modèles existants, l'exploration de nouvelles architectures de réseaux neuronaux adaptées aux données complexes, et l'intégration de techniques d'apprentissage fédéré pour une gestion distribuée et sécurisée du contrôle d'accès.

En conclusion, l'application de techniques avancées de Machine Learning pour le contrôle d'accès représente une avancée significative dans la sécurisation des informations sensibles, offrant des solutions innovantes pour répondre aux défis croissants de la cybersécurité dans le monde moderne.

Perspectives

Pour aller de l'avant, plusieurs perspectives méritent d'être explorées :

— **Amélioration continue des modèles** : Adapter les modèles existants pour intégrer de nouvelles données et améliorer la précision des prédictions.

— **Sécurité et éthique** : Renforcer les aspects de sécurité et d'éthique dans le développement et l'utilisation de systèmes basés sur l'intelligence artificielle pour le contrôle d'accès.

— **Exploration de nouvelles architectures** : Investiguer des architectures de réseaux de neurones plus avancées pour traiter efficacement les données complexes et hétérogènes.

— **Intégration de l'apprentissage fédéré** : Explorer l'utilisation de techniques d'apprentissage fédéré pour permettre un contrôle d'accès distribué et sécurisé.

Bibliographie

[1] Worachet UTTHA. *Étude des politiques de sécurité pour les applications distribuées : le problème des dépendances transitives.* PhD thesis, Université d'Aix-Marseille, 2016.

[2] Faiza AINENNAS and Nassima ZIDI. Contrôle d'accès aux services sensibles au contexte. Master's thesis, Université Abderahman Mira de Béjaia, 2015.

[3] Anonymous. Title of the article. *arXiv*, 2207.01.739V1, 2022.

[4] Meriem BOUGHELIT and Soumia MOUSSOUS. Développement d'un système d'accès intelligent en utilisant les méthodes d'apprentissage automatique. Master's thesis, Université Saad Dahlab Blida, 2020.

[5] Sourour JEMILI. Analyse de risque dans les systèmes de contrôle d'accès. Master's thesis, Université du Québec en Outaouais, 2013.

[6] Nor Eddine KHELIFA. Intégration du modèle de contrôle d'accès rbac (role based access control) dans les diagrammes uml (cas d'utilisation et séquence). Master's thesis, Université d'Oran, Year.

[7] Sofiene Boulares. Validation des politiques de sécurité par rapport aux modèles de contrôle d'accès. Master's thesis, Université du Québec en Outaouais, 2010.

[8] Philip WL Fong. Relationship-based access control : protection model and policy language. In *ACM CODASPY*, 2011.

[9] Unknown. L'intelligence artificielle : Machine learning, deep learning. Blog, 04 2020. Consulté le 17/04/2020.

[10] Unknown. Apprentissage supervisé vs non supervisé. Blog, 05 2020. Consulté le 10/05/2020.

[11] Djamila HAMMOUD. *Apprentissage Automatique dans un Agent*. PhD thesis, Université Mentouri Constantine.

[12] Islem Zara. L'intelligence artificielle principe, outils et objectifs. Master's thesis, Université Badji Mokhtar Annaba, 2019.

[13] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining : Concepts and Techniques*. Elsevier, USA, 3rd edition, 2012.

[14] Ali LABIAD. Sélection des mots clés basée sur la classification et l'extraction des règles d'association. Master's thesis, Université du Québec À Trois-Rivières, 2017.

[15] Unknown. Algorithme k-means.

[16] ZAHRA YAHIAOUI. Etude et implémentation de l'algorithme c moyenne floue et ses variantes. Master's thesis, Université de M'sila, 2013.

[17] Unknown. Les svm (support vector machine). Consulté le 20/02/2019.

[18] Ahmed GHALI. Amélioration de la reconnaissance par le visage. Master's thesis, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, 2015.

[19] Samir Amir. Une mesure de similarité entre phrases basée sur des noyaux sémantiques, janvier 2016.

[20] Cisco. Types of authentication, 2021.

[21] Elearning Online Academy. Types of supervised machine learning, 2021.

[22] Bruno Agard and Andrew Kusiak. Exploration des bases de données industrielles à l'aide du data mining – perspectives. In *9ème Colloque National AIP PRIMECA*, 04 2005.

[23] Daniel T. Larose and Chantal D. Larose. *Data Mining and Predictive Analytics*. John Wiley & Sons, Canada, 2nd edition, 2015.

[24] Minimol Anil Job. Data mining techniques applying on educational dataset to evaluate learner performance using cluster analysis. *EJERS, European Journal of Engineering Research and Science*, 3(11), 11 2018.

[25] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining : Concepts and Techniques, 3rd Edition Solution Manual*, 2011.

[26] O. Mehdi and K. Salim. Classification d'objets avec le deep learning. Université de Bouira, 2018.

[27] Tom Mitchell. *Machine Learning*. McGraw Hill, 1997.

[28] Y. Bengio, A. Courville, and P. Vincent. Representation learning : A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38 :1798–1828, 2013.

[29] Google Colaboratory. Google colaboratory. https ://colab.research.google.com/, 2021.

[30] NumPy Developers. Numpy. https ://numpy.org/, 2021.

[31] Matplotlib Developers. Matplotlib : Visualization with python. https ://matplotlib.org/, 2021.

[32] Author's Name. Title of the paper on grid search. *Journal Name*, Volume :Pages, Year.

[33] Author's Name. Title of the paper on k-folds. *Journal Name*, Volume :Pages, Year.

[34] A. M. Giancarlo Zaccone and Md. Rezaul Karim. *Deep Learning with TensorFlow : Explore neural networks with Python*. 2017.

[35] Françoise Fessant. Apprentissage non supervisé. TECH/SUSI, 09 2006.

[36] Unknown. Exemple de classification binaire linéaire et non linéaire dans r^2 − *problème*.

[37] Soumia CHIKOUCHE. Système de détection d'intrusion basé sur la classification comportementale des processus. Master's thesis, Université de M'sila, 2012.

ملخص

الاستخدام التعلم الآلي هو موضوع ذو أهمية كبيرة يشمل جميع المجالات، ويعتمد على تنفيذ سياسة التحكم في الوصول. أثبت التعلم الآلي فعاليته الكبيرة في حل المشكلات الشاقة في العديد من المجالات. في السنوات الأخيرة، تم استخدام التعلم الآلي لفحص والتحقق من سياسات الدخول، مما يمكن أن يحسن من جودة سياسات التحكم في الوصول ويخفف من الأعباء المتعلقة بالمهام الشاقة، مثل إدارة الوصول إلى المعلومات، من خلال توتير عمليات التحكم في الوصول. يجري أيضًا البحث حاليًا في كيفية استخدام التعلم الآلي لمراقبة سياسة التحكم في الوصول بشكل تلقائي وإنذار مديري النظام في حال اكتشاف أنشطة مشبوهة. في عملنا، نقترح طريقة أو نموذجًا للتحكم في الوصول يستند إلى التعلم الآلي (تفاصيل من المروج لنا)، حيث يمكن لهذا النموذج أن يقرر ما إذا كان يجب منح طلب الوصول أم رفضه

Abstract

The use of machine learning is a topic of major importance that affects all domains, including the environment and users. It is based on the implementation of an access control policy. Machine learning has proved to be extremely effective in solving tedious problems in many areas. Recent years have seen the use of machine learning for access policy verification, which could improve the quality of the implemented access control policy and reduce the burden of the underlying laborious tasks of access control, such as information access management, by automating access control processes. Research is also underway on how machine learning can automatically monitor the deployed access control policy and alert system administrators if it detects suspicious activities. In our work, we propose a method or model of access control based on machine learning (details from our promoter), where this model will be able to decide whether an access request should be granted or refused.

Keywords : Access control, Access control policy, Machine learning, Deep learning

Résumé

L'utilisation du Machine Learning est un sujet d'une importance majeure qui touche tous les domaines, y compris l'environnement et les utilisateurs. Elle repose sur la mise en place d'une politique de contrôle d'accès. L'apprentissage automatique (Machine Learning) s'est révélé extrêmement efficace pour résoudre des problèmes fastidieux dans de nombreux domaines. Ces dernière années ont vu l'utilisation du Machine Learning pour la vérification des politique d'accès, ce qui pourrait améliorer la qualité de la politique de contrôle d'accès mise en œuvre et alléger les charges des tâches laborieuses sous-jacentes du contrôle d'accès, telles que la gestion de l'accès à l'information, en automatisant les processus de contrôle d'accès. Des recherches sont également menées sur la manière et les méthodes de Machine Learning qui peuvent surveiller automatiquement la politique de contrôle d'accès déployée et avertir les administrateurs du système s'il détecte des activités suspectes. Dans notre travail nous proposons une méthode ou un modèle de contrôle d'accès basée sur la machine Learning est consacré à proposer une méthode ou un modèle de contrôle d'accès, que nous testerons avec différentes approches de machine Learning, en intégrant des améliorations potentielles pour atteindre une meilleure performance. ou ce model pourra décider si une demande d'accès doit être accordée ou refusée.

Mots clés : contrôle d'accès, politique de contrôle d'accès, Machine Learning, deep learning