République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Béjaïa

Faculté des Sciences Exactes

Département d'Informatique

Mémoire de Fin d'Etude

En vue de l'obtention du diplôme de Master recherche en Informatique

Option : Intelligence Artificielle

Thème

Détection de fraude électrique en utilisant le machine Learning

Réalisé par :

- M[lle] KHENNICHE Chimsine
- M[lle] OUHADDAD Bahia

Devant le jury composé de

Président :	*D[r] MIR Foudil*	M.A.A - Université de Béjaïa.
Examinateur :	*D[r] BOUCHEBAH Fateh*	M.C.B - Université de Béjaïa.
Encadrant :	*D[r] SEBAA Abderazzak*	M.C.A - Université de Béjaïa.

Promotion 2022 - 2023

Remerciements

Tout d'abord, nous remercions Allah, tout puissant, de nous avoir donné la force et la foi pour achever ce travail

Nous tenons à exprimer notre profonde gratitude à notre encadrant Monsieur Sebaa Abderazzak, nous le remercions de nous avoir encadré, orienté, aidé, et pour sa disponibilité

Nous adressons nos sincères remerciements à Madame Azzouguer Dalila de nous avoir aidé et donné le principe pour entamer notre mémoire

Nous tenons à remercier toutes les personnes qui ont contribué par leurs paroles, leurs écrits durant nos recherches.

Dédicaces

À mon très cher père

Quoi que je fasse ou Quoi que je dise

Je ne saurai point te remercier comme il se doit

Ton affection me couvre, ta bienveillance me guide et ta présence à mes cotés a toujours été ma source de Force pour affronter

les différents obstacles

À ma très chère Mère

et qui a été mon pilier et ma source d'inspiration tout au long de ma vie

qui m'a toujours encouragé dans mes études , que Dieu la garde pour nous

et sans elle ma réussite n'aurra pas eu lieu

À Mes soeurs Fahima et Salima qui ont toujours été la pour moi

À ma famille, mes oncles, cousins, cousines

Au petit 'AXEL' de mon cher cousin samir

Et Au petit 'IRATH' de mon cher oncle Amine qui rejoignent notre famille

À ma binôme BAHIA,

je tiens à souligner à quel point notre partenariat a été exceptionnel.

Ta détermination et ton travail acharné ont été une source d'inspiration constante.

Ensemble, nous avons relevé les défis et réalisé ce travail avec succès

.

Et je suis honoré d'avoir eu l'occasion de travailler à tes côtés.

Chimsine.

Dédicaces

Je tiens à dédier ce travail

À Mes très chers parents, qui m'ont comblé avec leur amour, sacrifices et précieux conseils. Qui m'ont soutenu moralement et financièrement jusqu' à ce jour, qui m'ont encouragé tout au long de mon parcours d'études. Que ce travail soit pour eux un modeste témoignage de ma gratitude.

À ma grande sœur DALILA, qui a toujours été un modèle inspirant de persévérance et de réussite. Votre exemple m'a guidé à travers les hauts et les bas de ce voyage académique, et je vous remercie d'avoir toujours été là pour me soutenir, tout comme votre mari NASSIM et vos adorables enfants, ANIA et AYLAN, à qui je souhaite un avenir radieux et couronné de succès.

À mes deux frères AMIROUCHE et HAMIMI, pour leur présence réconfortante. Votre soutien a été une source de réconfort et de motivation.

À ma binôme CHIMSINE, je tiens à souligner à quel point notre partenariat a été exceptionnel. Ta détermination et ton travail acharné ont été une source d'inspiration constante. Ensemble, nous avons relevé les défis et réalisé ce travail avec succès, et je suis honoré d'avoir eu l'occasion de travailler à tes côtés.

À tous mes amis, qui ont apporté leur contribution de près ou de loin à ce mémoire. Votre encouragement, vos conseils et votre amitié m'ont porté à travers les moments les plus exigeants de cette aventure académique. Votre présence a rendu ce voyage mémorable et significatif.

Chacun de vous a joué un rôle essentiel dans cette réalisation, et je vous en suis profondément reconnaissante.

Bahia.

Table des matières

Table des figures

Liste des abréviations

IA	**I**ntelligence **A**rtificielle
SNE	**S**ociété **N**ationale d'**E**lectricité
CRE	**C**ommission de **R**égulation de l'**E**lectricité
CoBC	**Co**-training **B**y l'**C**ommittee
TL	**T**echnical **L**osses
NTL	**N**on **T**echnical **L**osses
ML	**M**achine **L**earning
AS	**A**pprentissage **S**upervisé
ANS	**A**pprentissage **N**on **S**upervisé
SVM	**S**upport **V**ector **M**achine
RF	**R**andom **F**orest
KNN	**K**- **N**earest **N**eighbors
CNN	**C**onvolutional **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
DNN	**D**eep **N**eural **N**etwork
LSTM	**L**ong **S**hort **T**erm **M**emory
LOF	**L**ocal **O**utlier **F**actor

Introduction Générale

L'électricité est au cœur de tout développement économique et technologique. Les avancées dans la gestion des réseaux électriques et la sécurité ont marqué un progrès significatif à l'échelle mondiale. Cependant, malgré ces avancées, les services publiques d'électricité sont aux prises avec un problème persistant : la fraude énergétique. Le vol d'électricité et les difficultés à recouvrer les factures représentent un défi majeur. Les tactiques frauduleuses évoluent constamment, obligeant ainsi les systèmes anti-fraude à s'adapter et à trouver un moyen de lutter contre ces fraudes dans les brefs délais et au moindre coût.

C'est dans ce contexte que l'intelligence artificielle se distingue. L'IA représente un domaine de recherche et d'innovation visant à créer des systèmes capables de simuler l'intelligence humaine. L'apprentissage automatique, une branche de l'IA, se concentre spécifiquement sur le développement de modèles qui apprennent à partir de données pour effectuer des tâches spécifiques, sans être explicitement programmés.

En utilisant le Machine Learning, nous avons exploré une approche pour lutter contre la fraude électrique. Cette méthode repose sur la capacité des modèles à analyser de vastes ensembles de données et à identifier des schémas souvent indiscernables pour l'œil humain. Ces modèles peuvent ainsi repérer des anomalies dans la consommation d'électricité, signe révélateur de fraudes potentielles.

Cette convergence entre l'énergie et l'innovation technologique constitue le cœur de notre démarche. Elle incarne une réponse essentielle aux enjeux cruciaux que représentent la sécurité énergétique et la viabilité économique.

La structure de ce mémoire s'articule autour de quatre chapitres distincts :

Le premier chapitre offre une vue d'ensemble de l'électricité, en définissant ses différents modes de production et en présentant les acteurs clés du marché de l'électricité. Nous abordons également la signification cruciale de la détection de la fraude électrique.

Le deuxième chapitre se concentre sur les séries temporelles, en expliquant en détail leurs composantes et les modèles qui les caractérisent. Nous introduisons également les concepts fondamentaux de l'apprentissage automatique et de l'apprentissage profond, ainsi que leurs algorithmes associés.

Le troisième chapitre consiste en une revue de l'état de l'art, synthétisant les travaux les plus pertinents dans le domaine. Une analyse comparative des différentes approches est également présentée.

Le quatrième et dernier chapitre expose les modèles utilisés, l'environnement de travail, les outils déployés, et présente les résultats obtenus dans le cadre de la détection de la fraude électrique à l'aide de l'apprentissage automatique.

Cette structure permettra une exploration complète et approfondie de notre proposition novatrice pour la détection de la fraude électrique. Nous aspirons ainsi à contribuer activement à la préservation et à l'optimisation de nos ressources énergétiques pour les générations futures.

Chapitre 1

Généralités sur l'énergie électrique

1.1 Introduction

L'électricité est une forme d'énergie présente dans notre vie quotidienne qui joue un rôle essentiel, Elle est devenue indispensable dans de nombreux domaines comme le chauffage, l'éclairage, la communication, les transports, l'industrie. Il existe principalement deux types de moyens de production de l'électricité : le premier est la production d'électricité centralisé qui se devise en deux catégories la première est basée sur l'énergie primaire utilisée, la deuxième est basée sur sa méthode de conversion. Le deuxième moyen est la production d'électricité décentralisée qui se classe aussi selon l'énergie primaire utilisée La première catégorie comprend les technologies basées sur des sources d'énergie renouvelables la deuxième catégorie utilise des technologies à base d'énergies fossiles aussi contrôlables que conventionnelles. Il existe aussi plusieurs principaux acteurs du secteur sur le marché d'électricité comme les producteurs, les consommateurs, Le Gestionnaire du réseau de transport, Le Gestionnaire de réseau de distribution, les fournisseurs, L'agence de régulation et Les traders et courtiers. Le système de comptage de l'électricité comprend plusieurs types de compteur exemple Compteurs électromécaniques qui est basé sur le modèle de conception Thompson, Il

existe aussi les Compteurs électroniques qui sont plus avancés, ils utilisent des circuits électroniques pour la facturation. Malgré tous ces progrès de sécurité des réseaux de distribution, les sociétés nationales d'électricité se plaignent sur l'énergie non facturée qui est provenu des pertes techniques fait référence á la perte d'énergie électrique, elle se produit lorsque l'électricité est convertie en énergie thermique lors de son passage dans divers composants du système ou des pertes non techniques qui est principalement associé au vol d'électricité ou á la fraude. .

1.2 Définition de l'électricité

L'électricité est un phénomène énergétique associé à la mobilité ou au repos de particules chargées positivement ou négativement. L'électricité est un Phénomène directement lié à la structure de la matière, dû aux différentes charges électriques. Les atomes sont formés d'un noyau positif autour duquel tournent un ou plusieurs électrons négatifs, ils sont électriquement neutres, c'est-à-dire qu'ils contiennent autant de charges positives que de charges négatives [30].

1.3 Différents types de moyens de production

la production d'électricité est produite à partir de diverses sources, telles que :

1.3.1 Production d'électricité centralisée

Elle classe les installations de production centralisée en deux catégories, la première catégorie selon l'énergie primaire utilisée et la seconde selon son mode de conversion. Seules les technologies les plus développées et éprouvées sont prises en compte, telles que les centrales nucléaires, les centrales hydroélectriques, les centrales thermiques à

vapeur et les turbines à gaz [30].

1.3.2 Production d'électricité décentralisée

elle classe les installations de production décentralisées selon l'énergie primaire utilisée. La première catégorie regroupe les technologies basées sur les énergies renouvelables telles que le solaire (photovoltaíque, thermique), l'éolien, l'hydraulique, l'hydrolienne, la marémotrice, la géothermie, la biomasse. D'autre part, la deuxième catégorie utilise des technologies basées sur les combustibles fossiles, qui sont contrôlables comme traditionnels (turbines á gaz, moteurs thermiques á combustion, à combustion et á explosion), les piles à combustible, la cogénération. Ils permettent une production d'énergie plus stable et prévisible [30].

1.4 Acteurs du marché d'électricité

Le marché de lélectricité implique plusieurs acteurs qui jouent des rôles différents. Voici Les principaux acteurs du secteur de lélectricité :

1.4.1 Producteurs

sont des productions d'électricité centralisée ou décentralisée, qui produisent et distribuent de l'électricité [30].

1.4.2 Consommateurs

sont des individus qui utilisent l'électricité pour leurs besoins quotidiens, que ce soit dans les résidences, les entreprises, les industries ou les secteurs publics [30].

1.4.3 Gestionnaire du réseau de transport

Chargé d'investir et de renforcer le réseau de transport en prévision de l'évolution de la consommation d'électricité. Il vise également à minimiser la congestion et les pertes afin de ne pas affecter l'erreur de prévision du lendemain [30].

1.4.4 Gestionnaire du réseau de distribution

responsable de l'intégrité du réseau de distribution d'électricité, de la gestion et de l'exploitation du réseau, ainsi que du respect des normes de qualité de l'énergie fournie aux clients finaux. Il participe également á la maîtrise des consommations d'énergie et peut sanctionner les clients en cas de non-respect des obligations contractuelles [30].

1.4.5 Fournisseurs

sont l'intermédiaire financier entre le producteur et le client final [30].

1.4.6 Agence de régulation

géré par la Commission de régulation de l'électricité (CRE), il a pour mission de définir les règles de gestion du réseau de transport et de favoriser un accès équitable et non discriminatoire à tous les acteurs du marché [30].

1.4.7 Traders et courtiers

les premiers spéculent sur les marchés de l'énergie, proposant des contrats d'achat et de vente avec des options garantissant les prix, L'autre effectue des transactions financières [30].

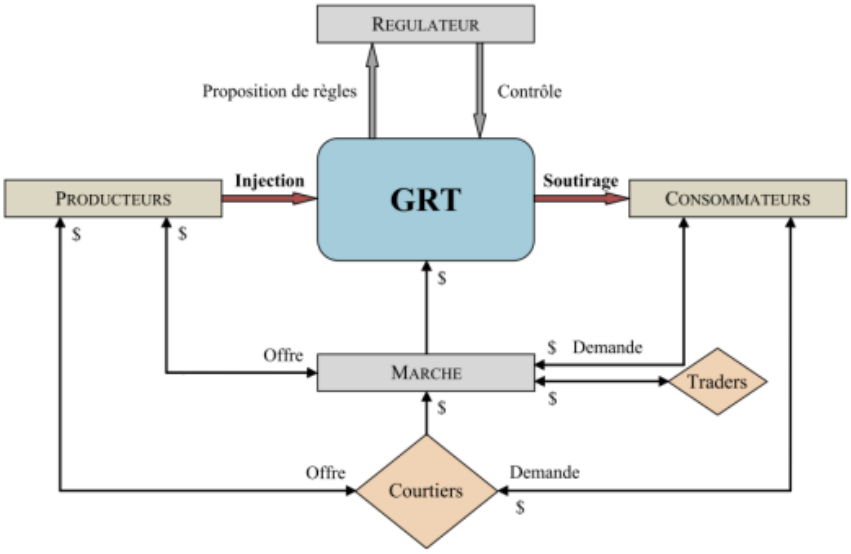

FIGURE 1.1 – Acteurs de marché d'éléctricité
[30].

1.5 Système de mesure de l'énergie électrique

C'est un élément essentiel du système de mesure de l'électricité. Il collecte, traite et communique les données de consommation, permettant une facturation précise aux clients et fournissant des informations pour gérer et optimiser la consommation d'électricité. Il comprend plusieurs types de compteurs, voici quelques-uns des types de systèmes de mesure couramment utilisés en Algérie :

1.5.1 Compteurs électromécaniques

Ce sont des appareils de mesure de l'électricité selon le modèle de construction Thompson, Ils utilisent des composants électriques et mécaniques pour mesurer la consommation d'énergie, ils peuvent être monophasés ou multiphasés Ces compteurs ont été utilisés pendant longtemps pour leur fiabilité, maintenant ils ont été remplacés par des compteurs plus avancés tels que les compteurs électroniques et les compteurs intelligents [30].

FIGURE 1.2 – Un compteur électromécanique
[32]

1.5.2 Compteurs électroniques

Ces compteurs électriques utilisent des circuits électroniques pour mesurer la consommation d'électricité. Ils offrent généralement des fonctionnalités plus avancées que les compteurs électromécaniques telles que la communication de données et l'affichage numérique de la consommation [30].

FIGURE 1.3 – un compteur électronique
[32].

1.6 Catégories de vol d'électricité

Les pertes d'énergie dans un réseau de distribution d'électricité se divisent en deux catégories qui sont :

1.6.1 Pertes techniques

Les pertes techniques(TL)sont des pertes d'énergie électrique qui se produisent lorsque l'électricité est convertie en chaleur lors de son passage dans divers composants du système électrique, tels que les lignes de transmission, les tableaux de distribution et les transformateurs. Ces pertes sont causées par des facteurs tels que l'usure des lignes, leur longueur, le nombre de charges connectées, la connexion entre les lignes et la qua-

lité des transformateurs. Bien qu'il ne soit pas possible de les éliminer complètement, il est possible de les réduire en utilisant des technologies avancées et en effectuant un entretien régulier de l'infrastructure du réseau électrique. Les pertes techniques sont généralement estimées et prises en compte dans la conception et la construction des installations électriques [30].

1.6.2 Pertes non techniques

Les pertes techniques (NTL)représentent l'électricité consommée mais non facturée. NTL est principalement associé au vol d'électricité ou à la fraude, mais peut également être causé par des appareils de mesure défectueux et des erreurs de lecture et de facturation. NTL se produit lorsqu'un consommateur déclare de manière erronée sa consommation d'énergie à l'entreprise de services publics. Une pratique courante de NTL consiste à altérer le compteur et à modifier les lectures pour masquer la consommation d'énergie et réduire les factures. La figure montre les différents endroits oú les NTL peuvent se produire dans le cadre d'un réseau intelligent [30].

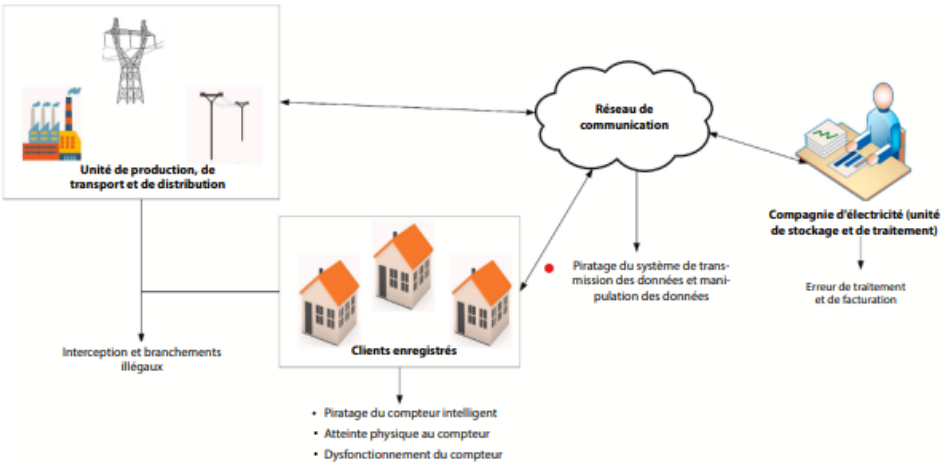

FIGURE 1.4 – Objectifs probables pour NTL dans le domaine des réseaux intelligents [30].

Les attaques se manifestent dans les systèmes de distribution, les appareils de me-

sure et les réseaux de communication. L'une des principales raisons des attaques est de manipuler les données afin de réduire la consommation d'énergie et donc les factures énergétiques. Il existe deux types d'attaques NT L qui sont :

1.6.2.1 Attaques cybernétiques

Les systèmes intelligents sont de plus en plus vulnérables aux cyber-attaques qui peuvent intervenir à différentes étapes telles que l'enregistrement, la transmission ou le stockage des données. Les cyberattaques ciblant les réseaux intelligents sont variées et ont des objectifs spécifiques tels que l'écoute clandestine, le déni de service, le camouflage, l'injection de logiciels malveillants et l'injection de fausses données. Certaines de ces attaques peuvent entraîner des interruptions de service, la destruction de l'infrastructure, ainsi que le vol d'électricité et d'informations. La saisie de fausses informations est une méthode courante de vol d'énergie. Les pirates tentent de manipuler les données des compteurs intelligents en modifiant la consommation d'énergie. Pour réussir ces attaques, les pirates doivent identifier les vulnérabilités du réseau, obtenir des privilèges d'accès, fournir des informations de configuration du système et injecter de fausses données. Il est essentiel de mettre en place des mesures de sécurité solides pour protéger les réseaux intelligents contre les cyberattaques et maintenir les systèmes à jour pour éviter les vulnérabilités [30].

1.6.2.2 Attaques physiques

Elles nécessitent une manipulation physique des compteurs afin de modifier les relevés des compteurs intelligents.

Les méthodes d'attaque physique sont : contourner le compteur en tressant des tuyaux ou en utilisant des câbles, des connexions non autorisées, placer un aimant puis-

sant, renverser le compteur et rendre le compteur inutilisable.

Voici quelques méthodes de manipulation qui affectent les relevés des compteurs intelligents :

- **Les interférences magnétiques :** impliquent l'utilisation d'aimants à proximité du compteur pour perturber son fonctionnement. Cela ralentira la mesure et faussera la consommation d'énergie en utilisant la quantité enregistrée. L'effet dépend de la taille et de la force de l'aimant, voire de l'annulation complète de la lecture du compteur par des aimants plus puissants.
- **Inversion de courant :** elle se fait en inversant le branchement du compteur électrique pour annuler la puissance active et réduire l'énergie totale. Dans le cas d'une installation monophasée, le compteur affichera des lectures négatives ou nulles pour les longues connexions inverses. Dans la conception triphasée, la phase négative annule les mesures positives des autres phases et signale une consommation d'énergie réduite.
- **Pontage du compteur :** le pontage implique la connexion de conducteurs sous tension ou de métal entre les lignes et agit comme un diviseur de courant. En conséquence, moins de courant circule dans le shunt et le compteur lit moins qu'il n'en prend réellement (contournement partiel) ou empêche le compteur de prendre une lecture complète (contournement complet) [30].

1.7 Conclusion

Dans ce chapitre, nous avons fourni des informations générales sur les systèmes électriques et un aperç général du vol d'électricité. Dans un premier temps, nous avons abordé les différents types de moyens connus dans la littérature pour la production

d'énergie électrique. Ensuite, nous avons décrit les principaux acteurs du marché de l'électricité. Et nous avons également présenté les différents types de compteurs électriques utilisés en Algérie. Enfin, nous nous sommes concentrés sur la question du vol d'électricité.

Dans le chapitre suivant, nous établirons les définitions du Machine learning et du DeepLearning et quelques algorithmes qui nous aideront à détecter cette fraude.

Chapitre 2

Généralités sur les séries temporelles et machine Learning

2.1 Introduction

La plupart des entreprises utilisent encore des systèmes basés sur des règles comme principal outil de détection de la fraude, ces règles permettent de découvrir très facilement les tendances connues, mais elles sont peu efficaces face aux schémas de fraude inconnus ou aux techniques de plus en plus sophistiquées des fraudeurs. C'est là que l'analytique et le Machine Learning (encore appelé l'apprentissage automatisé), devient nécessaires pour la prévention et la détection de la fraude.

L'apprentissage automatique est un champ d'étude de l'intelligence artificielle qui peut être la meilleure solution pour la détection d'anomalies Dans ce chapitre, nous présentons dans un premier temps tout ce qui concerne les séries temporelles, ses déférentes composantes et comment se modélisent, ensuite le Machine Learning et ses Paradigmes qui se divisent en deux types (supervisé et non supervisé), ainsi les Graphiques d'une série chronologique et par la suite nous donnons une vue générale sur le Deep Learning et ses déférents paradigmes.

2.2 Séries temporelles

La prévision de séries temporelles est devenue un domaine de recherche très intensif, qui est même en augmentation ces dernières années. Les Réseaux de neurones profonds se sont révélés puissants et atteignent une grande précision dans de nombreux domaines d'application. Pour ces raisons ils sont l'une des méthodes d'apprentissage automatique les plus largement utilisées pour résoudre les problèmes de Big Data aujourd'hui.

Très souvent, pour des raisons économiques, nous prévoyons la consommation d'électricité de manière à faire correspondre au mieux la production à ce qui se passait avant le début de la prévision. Et parfois en utilisant des informations supplémentaires.

Vous ne pouvez pas prédire parfaitement, il y aura toujours une erreur, et les bonnes méthodes ne fournissent pas de prédiction, mais un intervalle de prédiction. Il arrive souvent qu'une petite amélioration de la qualité des prévisions ait un impact important sur les coûts [20].

2.2.1 Définition

Une série chronologique est définie comme une séquence de valeurs, triées chronologiquement et suivies dans le temps. Bien que le temps soit une quantité mesurée en continu, les valeurs de la série chronologique sont échantillonnés à intervalles constants.

Cette définition s'applique à de nombreuses applications, mais pas à toutes les séries chronologiques ne peuvent pas être modélisées de cette manière, pour certaines des raisons suivantes :

1. Les données manquantes dans une série chronologique sont un problème très courant en raison de la fiabilité de la collecte des données. Pour faire face à ces valeurs, il existe de nombreux stratégies, mais celles basées sur l'imputation d'informations

manquantes et sur le fait de sauter tout l'enregistrement sont les plus utilisés.

2. Les valeurs aberrantes sont aussi un problème qui revient souvent en séries chronologiques. Méthodes basée sur des statistiques robustes doivent être sélectionnées pour les supprimer valeurs ou simplement les incorporer dans le modèle.

3. Si les données sont collectées à des heures irrégulièrement, cela peut être le cas appelées séries temporelles irrégulièrement espacées ou, si elles sont des flux de données suffisamment importants [20].

2.3 Composantes d'une série temporelle

Une série chronologique est constituée de trois composantes suivantes :

- **Tendance :** C'est le mouvement général que la série chronologique présente au cours de la période d'observation, sans tenir compte de la saisonnalité et des irrégularités.

  Dans certains textes, cette composante est également connue sous le nom de variation à long terme. Bien qu'il existe différents types de tendances dans les séries chronologiques, les plus populaires sont les tendances linéaires, exponentielles ou paraboliques [20].

- **Saisonnalité :** Ce composant identifie les variations qui se produisent à des intervalles réguliers spécifiques et peut fournir des informations utiles lorsque des périodes de temps présentent des schémas similaires. Il intègre des effets raisonnablement stables en fonction du temps, de l'amplitude et de la direction.

  La saisonnalité peut être causée par plusieurs facteurs tels que le climat ou les cycles économiques,voire les festivités [20].

- **Résidus :** Une fois la tendance et les oscillations cycliques calculées et supprimées, certaines valeurs résiduelles subsistent. Ces valeurs peuvent être, parfois, suffisamment élevées pour masquer la tendance et la saisonnalité.

  Dans ce cas, le terme valeur aberrante est utilisé pour désigner ces résidus, et des statistiques robustes sont généralement appliquées pour y faire face20. Ces fluctuations peuvent être d'origines diverses, ce qui rend la prédiction presque impossible. Cependant, si par hasard, cette origine peut être détectée ou modélisée, elles peuvent être considérées comme des précurseurs des changements de tendance [20].

2.4 Graphiques d'une série chronologique

Considérons la série trimestrielle du chiffre d'affaires en milliers de francs des ventes d'un magasin de 2015 à 2019.

- **Graphique de séries chronologiques** : nous montrons les points (t,Yt), que nous relions par des lignes. L'évolution de la quantité considérée sur toute la période surveillée est indiqué [6].

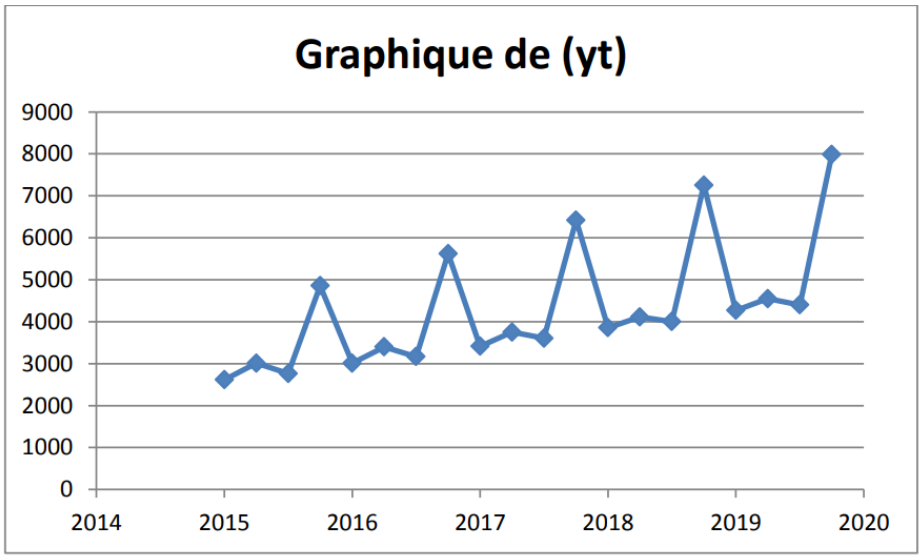

FIGURE 2.1 – Graphe de la série chronologique du chiffre d'affaires en milliers de francs des ventes d'un magasin de 2015 à 2019

[6].

- **Courbe superposée** : les points (j,Yi,j) connectés sont représentés par des segments de ligne pour chaque année. Cela représente la variation annuelle de hauteur au fil des mois (pour chaque année). On peut ainsi comparer le même mois d'années différentes, mais on ne voit pas l'évolution globale.

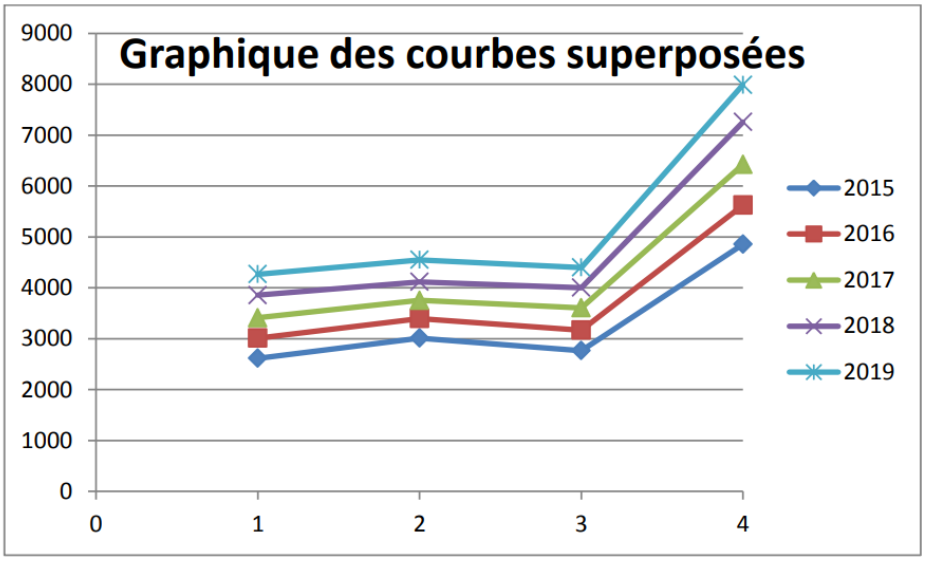

FIGURE 2.2 – Graphe des courbes seperposées de la série trimestrielle du chiffre d'affaires en milliers de francs des ventes d'un magasin de 2015 à 2019

[6].

2.5 Modélisation des séries temporelles

Les deux modéles de la modélisation sont :

2.5.1 Modèle additif

Dans le modèle additif, 3 composantes : tendance, saisonnalité et valeur résiduelle sont indpendantes les unes des autres. La série Xt s'écrit comme la somme de ces trois composantes :

$Xt = Ct + St + \varepsilon t$ [6]. (2.1)

2.5.2 Modèle multiplicatif

lorsque les fluctuations saisonniéres dépendent de la tendance. Xt s'écrit ainsi :

$Xt = Ct \times St + \varepsilon t$ [6]. (2.2)

2.6 Choix du modèle

Avant toute modélisation et étude approfondie du modèle, on cherche d'abord à déterminer si on est en présence d'une série dans laquelle pour une observation donnée X. Si la variation saisonnière de S s'ajoute simplement à la tendance de Z, alors elle est un modèle additif. Si la variation saisonnière de S est proportionnelle à la tendance de Z, alors il s'agit d'un modèle multiplicatif. Pour faire cette distinction, vous pouvez vous

appuyer sur une méthode graphique ou utiliser une méthode analytique [6].

2.6.1 Méthode de la bande

Nous utiliserons un graphique en série et une ligne passant par les bas et une ligne passant par les hauts. Si ces 2 droites sont approximativement parallèles : le modèle est additif. Si ces 2 droites ne sont pas parallèles : le modèle est multiplicatif [6].

2.6.2 Méthode du profil

Nous allons utiliser un graphique de courbes superposées Si les différentes courbes sont quasiment parallèles : le modèle est additif. Sinon (les pics et les creux sont mis en évidence) : le modèle est multiplicatif [6].

2.6.3 Méthode du tableau de Buys et Ballot

Pour chaque année, nous calculons la moyenne et l'écart type. Les points sur l'axe horizontal sont la moyenne et l'ordonnée l'écart type de la même année. Une ligne des moindres carrés de ces points est tracée. Si l'écart type est indépendant de la moyenne, le modèle est additif. Et la pente (a) de la droite des moindres carrés est très proche de 0. Si l'écart type est fonction de la moyenne, le modèle est multiplicatif. Et la pente (a) de la droite des moindres carrés n'est pas nulle [6].

2.7 Domaines de l'Intelligence artificielle

L'apprentissage automatique (ML), est un domaine en plein essor de l'intelligence artificielle (IA) qui utilise des techniques statistiques pour donner à une machine la

capacité d'apprendre sans la programmer explicitement, c'est-à-dire de faire des prédictions à partir de données [16].

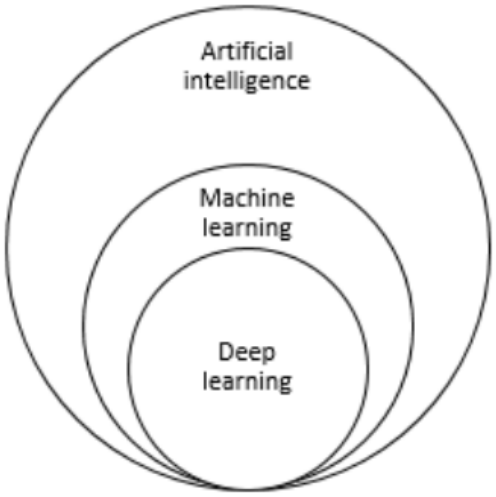

FIGURE 2.3 – l'intelligence artificielle et le machine learning

[14]

2.8 Machine Learning

L'apprentissage automatique est un domaine qui s'est maintenant imposé dans notre société. Il a été utilisé pendant des décennies dans la reconnaissance automatique de caractéres ou les filtres anti-spam, et maintenant il est utilisé pour se protéger contre la fraude, identifier les visages dans notre viseur d'appareil photo ou traduire automatiquement des textes d'une langue à l'autre.

L'apprentissage automatique peut être utilisé pour résoudre des problèmes :

1. Que nous ne savons pas résoudre.
2. Qu'on sait résoudre, mais qu'on ne peut pas formaliser en termes algorithmiques, comment on les résout (c'est le cas par exemple de la reconnaissance d'images ou de la comprchension du langage naturel)

3. Que l'on sait résoudre, mais avec des procédures trop consommatrices de ressources informatiques (c'est par exemple le cas de la prédiction d'interactions entre grosses molécules, pour lesquelles les simulations sont très lourdes).

Ainsi, le machine learning est utilisé lorsque les données sont abondantes (relativement) mais que les connaissances sont peu disponibles ou peu développées [10].

2.8.1 Catégories de Machine Learning

Le machine Learning apporte des réponses aux :

1. Enjeux sécuritaires
2. Enjeux économiques
3. Enjeux environnementaux
4. Enjeux organisationnels
5. Enjeux technologiques [37].

2.9 Paradigmes d'apprentissage en machine learning

Les algorithmes sont les moteurs du machine Learning. En général, deux principaux types d'algorithmes de machine Learning sont utilisés aujourd'hui : l'apprentissage supervisé et l'apprentissage non supervisé. La différence entre les deux se définit par la méthode employée pour traiter les données afin de faire des prédictions.

- **Apprentissage supervisé** : L'apprentissage supervisé (AS) est un processus en apprentissage automatique qui consiste à apprendre à une fonction c faire des prédictions à partir d'une liste d'exemples étiquetés, c'est-à-dire accompagnés de la valeur á prédire (voir Figure 2.3). Les balises servent de "professeur" et supervisent l'apprentissage de l'algorithme [10].

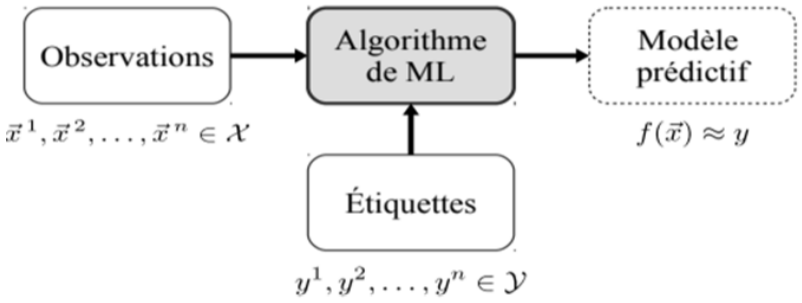

FIGURE 2.4 – Apprentissage supervisé [10]

.

En apprentissage supervisé, on distingue deux types de tâches :

1. **Classification :** dans un problème de classification, on essaie de classer un objet en différentes classes, c'est-à-dire qu'on essaie de prédire la valeur d'une variable discrète (qui ne prend qu'un nombre fini de valeurs).
2. **Régression :** les tâches de régression se concentrent sur la recherche d'une prédiction de la valeur d'une variable continue, c'est-à-dire une variable pouvant prendre une infinité de valeurs [10].

- **Apprentissage non supervisé** : Dans l'apprentissage non supervisé (ANS), les données ne sont pas étiquetées. Il s'agit alors de modéliser les observations pour mieux les comprendre. Aucun exemple requis. L'algorithme doit découvrir par lui-même la structure en fonction des données et ainsi parvenir à trouver par lui-même les bons poids.

  1. **Clustering :** un sous-problème d'apprentissage non supervisé est le regroupement en classes homogènes constituées d'une représentation en nuage de points d'un espace arbitraire dans un ensemble de groupes appelé cluster. Il s'agit d'un traitement sur un ensemble d'objets qui n'ont pas été marqués par le gestionnaire. L'algorithme doit découvrir par lui-même la structure en fonction des données [10].

- **Apprentissage semi supervisé** : L'apprentissage semi-supervisé est un mélange d'apprentisage supervisé et non-supervisé. Ils classent un jeu de données non étiquette à l'aide d'un jeu de données étiquetées [10].
- **Apprentissage par renforcement** : L'apprentissage par renforcement est un processus d'apprentissage automatique dans lequel un logiciel apprend â effectuer une tâche en fonction de ses échecs et de ses succès[10].

2.10 Algorithmes d'apprentissage automatique

L'apprentissage automatique comprends plusieurs algorithmes, les plus utilisées sont :

2.10.1 Support Vector Machine

Support-Victor Machine (SVM) est un algorithme d'apprentissage supervisé classé dans les techniques de classification. Il s'agit d'une technique de classification binaire qui utilise un ensemble de données d'apprentissage pour prédire l'hyperplan optimal dans un espace à n dimensions.

Les SVM sont une généralisation des classificateurs linéaires. Les SVM ont été développés dans les années 1990 sur la base des considérations théoriques de Vladimir Vapnik sur le développement de la théorie de l'apprentissage statistique : Théorie de Vapnik Chervonenkis. Les SVM ont été rapidement adoptés pour leur capacité à travailler avec des données volumineuses, leur faible nombre d'hyperparamètres, leurs garanties théoriques et leurs bons résultats en pratique.

Les SVM ont été appliqués dans de nombreux domaines (bio-informatique, recherche d'information, vision par ordinateur, finance...). Selon les données, les performances des machines â vecteurs de support sont du même ordre, voire meilleures,

que les performances du réseau de neurones ou du modèle de mélange gaussien. Un hyperplan qui est le centroïde des points x satisfaisant +b=0. L'orientation de l'hyperplan correspond à la règle de décision consistant à observer de quel côté de l'hyperplan se trouve l'exemple x. On voit que le vecteur w définit la pente de l'hyperplan (w est perpendiculaire à l'hyperplan). Pendant ce temps, le terme b permet de transformer l'hyperplan parallèlement à lui-même [9].

2.10.2 K-Nearest Neighbor(KNN)

L'algorithme k-proche voisin, également connu sous le nom de KNN ou k-NN, est un apprentissage discriminant supervisé non paramétrique qui utilise la proximité pour effectuer des classifications ou des prédictions sur le regroupement de points de données individuels. Il est généralement utilisé comme algorithme de classification, en supposant que des points similaires peuvent être trouvés les uns à cété des autres. Le but de l'algorithme du plus proche voisin est d'identifier les voisins les plus proches d'un point de requête donné afin que nous puissions attribuer une étiquette de classe à ce point [28].

2.10.3 Random Forest

Radom Forest est un algorithme d'apprentissage automatique couramment utilisé qui combine la sortie de plusieurs arbres de décision pour produire un seul résultat. Sa facilité d'utilisation et sa flexibilité ont alimenté son adoption car il résout à la fois les problèmes de classification et de régression.

Les arbres de décision sont des modèles hiérarchiques qui se comportent comme une série sq́uentielle de tests conditionnels dans lesquels chaque test dépend de ses prédécesseurs. Ils sont couramment utilisés en dehors du monde du machine learning,

par exemple pour décrire les étapes d'un diagnostic ou d'un choix de traitement pour un médecin, ou des cheminements possibles dans un "livre dont vous êtes le héros".

Un arbre de décision est un modèle de prédiction qui peut être représenté sous la forme d'un arbre. Chaque nœud de l'arbre teste une condition sur une variable, et chacun de ses enfants correspond à une réponse possible à cette condition. Les feuilles de l'arbre correspondent à l'étiquette. Pour prédire l'étiquette d'une observation, nous "suivons" les réponses aux tests depuis la racine de l'arbre et renvoyons l'étiquette de la feuille que nous atteignons [13].

2.10.4 Régression linéaire

Le principe de la régression linéaire est de modéliser la variable dépendante quantitative Y par une combinaison linéaire de p variables explicatives quantitatives, X1, X2, Xp.

Il comporte trois étapes clés pour créer le meilleur modèle de régression linéaire :

1. Définir une fonction de coût : c'est une fonction mathématique qui mesure l'erreur que l'on commet dans l'approximation des données. On parle aussi de l'erreur causée par le modèle.
2. Minimiser cette fonction de coût : nous devons trouver les bons paramètres de notre modèle pour minimiser l'erreur de modélisation.
3. Choisir une méthode de résolution de problème (gradient) [10].

2.10.5 K-means Clustering

K-Means est un algorithme de clustering non supervisé. Divise les données d'image en K clusters. Contrairement à d'autres méthodes dites hiérarchiques qui créent une structure « en arbre de cluster »pour décrire des groupements, k-Means ne crée qu'un

seul niveau de clusters. L'algorithme renvoie une distribution des données dans laquelle les objets de chaque cluster sont aussi proches que possible les uns des autres et aussi éloignés que possible des objets des autres clusters. Chaque cluster de la partition est défini par ses objets et son centre de gravité. K-Means est un algorithme itératif qui minimise la somme des distances entre chaque objet et son centre de gravité de cluster.

La position initiale des centroïdes détermine le résultat final, de sorte que les centroïdes doivent être initialement placés aussi loin que possible pour optimiser l'algorithme. K-Means modifie les objets du cluster jusqu'à ce que la somme ne puisse plus diminuer. Le résultat est un ensemble de clusters compacts et clairement séparés, à condition que la valeur K correcte pour le nombre de clusters ait été choisie [25].

Les principales étapes de l'algorithme des k-moyennes sont :

1. Sélection aléatoire de la position initiale de K clusters.
2. Réaffectez les objets au cluster selon le critère de minimisation de distance (généralement selon la mesure de distance euclidienne).
3. Une fois tous les objets placés, recalculez les K barycentres.
4. Répétez les étapes 2 et 3 jusqu'à ce que vous n'apportiez plus de modifications d'affectation [25].

2.11 L'intelligence Artificielle et le Machine Learning

L'intelligence artificielle implique l'idée d'une machine capable d'imiter l'intelligence humaine, ce que l'apprentissage automatique ne peut pas faire. L'objectif de l'apprentissage automatique est d'apprendre à une machine à effectuer une tâche spécifique et à fournir des résultats précis en identifiant des modèles.

Intelligence Artificielle	Machine Learning
- L'IA permet à une machine de simuler l'intelligence humaine pour résoudre des problèmes.	- Le ML permet à une machine d'apprendre de manière autonome à partir des données passées.
- L'objectif est de développer un système intelligent capable d'effectuer des tâches complexes.	- L'objectif est de créer des machines capables d'exploiter les données pour améliorer la précision du résultat.
- Nous créons des systèmes capables de réaliser des tâches complexes comme un humain.	- Nous entraînons des machines avec des données pour exécuter des tâches spécifiques et obtenir des résultats précis.
- L'IA couvre un large éventail d'applications.	- Le champ d'application des applications de Machine Learning est limité.
- L'IA utilise des technologies dans un système de manière à imiter la prise de décision humaine.	- Le ML génère des modèles prédictifs à l'aide d'algorithmes d'autoapprentissage.
- L'IA est compatible avec tous les types de données tels que structurées, semi-structurées et non structurées.	- Le ML ne peut utiliser que des données structurées et semi-structurées.
- Les systèmes d'IA s'appuient sur une logique et des arbres de décision pour apprendre, raisonner et corriger.	- Les systèmes de ML s'appuient sur des modèles statistiques pour apprendre et peuvent corriger automatiquement les nouvelles données.

TABLEAU 2.1 – La diférence entre machine learning et intelligence artificielle [1]

2.12 Deep Learning

Le deep learning permet à l'ordinateur de construire des concepts complexes à partir de concepts plus simples.

L'apprentissage en profondeur est une classe de techniques d'apprentissage automatique qui modélisent les données avec un haut niveau d'abstraction sur des architectures multicouches.

L'apprentissage en profondeur fournit également des outils utiles pour traiter d'énormes quantités de données et faire des prédictions utiles dans les domaines scientifiques.

Cela a été utilisé avec succès pour prédire comment les molécules interagissent [38].

2.13 Catégorisation de l'apprentissage profond

Selon la manière dont les architectures et les techniques doivent être utilisées, le Deep learning peut généralement être divisé en trois grandes catégories :

2.13.1 Réseaux profonds pour l'apprentissage supervisé

Ils sont destinés à fournir directement un pouvoir discriminant pour la classification des modèles, souvent en caractérisant la distribution a posteriori des classes conditionnelles aux données visibles exemple :

- **Réseaux neuronaux convolutifs :** sont un type spécialisé de réseau neuronal conçu pour traiter des données ayant une topologie en grille connue. Cela inclut les données en séries temporelles, qui peuvent être considérées comme une grille unidimensionnelle avec des échantillons à intervalles de temps réguliers, ainsi que les données d'images, qui peuvent être vues comme une grille bidimensionnelle

de pixels. Les réseaux de convolution sont simplement des réseaux neuronaux qui utilisent la convolution à la place de la multiplication matricielle générale dans au moins l'une de leurs couches[38].

2.13.2 Réseaux profonds pour l'apprentissage non supervisé

Ils sont destinés à capturer la forte corrélation des données observées pour l'analyse ou la synthèse du modèle lorsqu'aucune information de sortie d'étiquette n'est disponible exemple :

- **Réseaux neuronaux récurrents** Les réseaux de neurones récurrents (Recurrent Neural Networks, RNN) forment une classe de réseaux qui permettent de prédire le futur. On trouve les LSTM qui sont un type de réseau neuronal récurrent qui peut apprendre et mémoriser des dépendances à long terme. Se souvenir d'informations passées pendant de longues périodes est le comportement par défaut [16].
- **Autoencoders** Les auto-encodeurs sont des réseaux de neurones artificiels capables d'apprendre des représentations efficaces des données d'entrée, appelées encodages, sans aucune supervision. Ces encodages sont généralement de dimensions inférieures aux données d'entrée.Les auto-encodeurs fonctionnent en apprenant simplement à copier leurs entrées vers leurs sorties. Ce qui semble être une tâche triviale. Mais en réalité, ce sera difficile si nous fixons une restriction sur le réseau. Par exemple, on peut limiter la taille de la représentation interne ou ajouter du bruit aux entrées et former le réseau pour récupérer les entrées d'origine[15].

2.14 Conclusion

Dans ce chapitre, nous avons donné une vision générale sur les sèries temporelles, les déférents Paradigmes de Machine Learning qui est le noyau de ce projet ainsi que ses types (supervisées ou non supervisées) et quelques algorithmes (KNN, KMeans, Random Forest) et enfin nous avons vu le principe de l'apprentissage profond (DeepLearning) et sa catégorisation.

Dans le chapitre suivant, nous allons établir un état de l'art des principales approches relatives à la détection de fraude électrique et une étude comparative pour les principaux travaux déjà réalisés.

Chapitre 3

Etat de l'art sur la détetcion de fraude

3.1 Introduction

Grâce à l'intelligence artificielle, les fraudeurs de l'électricité sont devenus de plus en plus facile à les détecter, ce qui facilite le travail des collaborateurs des entreprise. Cela encourage les chercheurs à développer davantage des systèmes intelligents qui contribuent à une meilleure détection des fraudes.

La détection des fraudes est un sujet largement étudié par les chercheurs, comme en témoignent de nombreuses études et publications sur le sujet.

Dans ce chapitre, nous allons présenter les principaux travaux liés à la détection des fraudes et les méthodes utilisées pour les réaliser. Dans ces études, ils sont appuyés sur la classification comme principal facteur de comparaison entre les algorithmes d'apprentissage utilisés.

Afin de classer un consommateur comme fraudeur, il est nécessaire d'utiliser des techniques d'apprentissage automatique pour identifier les facteurs de risque pouvant conduire à la fraude électrique.

Il existe plusieurs méthodes de détection de fraude, dont certaines que nous verrons dans le résumé des travaux connexes dans ce chapitre.

3.2 Différentes méthodes de détection de fraude

La consommation frauduleuse d'électricités est une pratique existante depuis de nombreuses années. Afin de limiter leurs consommations et ainsi leurs factures, certains clients modifient ou perturbent leurs installations de comptage, cela entraîne des pertes importantes pour les services publics. C'est pourquoi les savants ont pensè à l'utilisation de Machine Learning pour une meilleure détection de fraude.

L'objectif de Machine Learning est de rendre la machine capable de traiter une quantité volumineuse et inimaginable de données d'une manière rapide et d'effectuer des tâches extrêmement complexes et d'obtenir des résultats en temps réel, ce qui est difficile à obtenir avec des algorithmes classiques.

3.2.1 Méthodes supervisées

Bernat Coma-Puig et al [18] : Cet article présente une approche supervisée pour détecter la fraude dans Gas Natural Fenosa qui est une entreprise de services publics espagnole fournissant de l'électricité et du gaz. Les auteurs ont mené des campagnes expérimentales dans trois sites de taille moyenne pour tester l'efficacité de leurs méthodes de détection de fraude. Les résultats ont montré une nette amélioration par rapport à la méthode de référence.

Les campagnes ont utilisé des algorithmes d'apprentissage automatique, notamment l'algorithme de Gradient Boosting, pour détecter les modèles de fraude impliquant plusieurs indicateurs. Ces modèles ont été capables de détecter des schémas de fraude que même les employés spécialisés avaient du mal à repérer.

L'utilisation de compteurs communicants s'est avérée bénéfique, avec une performance supérieure d'environ 15% par rapport aux compteurs traditionnels. Cependant, le système développé a surpassé les deux méthodologies (gaz et électricité) en utilisant

les commentaires des campagnes pour apprendre en continu et s'adapter aux nouvelles techniques de fraude.

Les auteurs ont également élargi leurs campagnes au niveau des pays, en utilisant le même algorithme de Gradient Boosting. Les résultats obtenus ont été significativement meilleurs, avec un facteur de performance de 11,3 par rapport au modèle de référence.

L'article souligne également les perspectives d'amélioration du système, telles que l'exploration de nouveaux classificateurs, la localisation des modèles en tenant compte des caractéristiques géographiques et des tarifs, et l'utilisation des compteurs intelligents pour une meilleure détection des fraudes.

Les auteurs reconnaissent également certains défis à relever, tels que le compromis entre l'exploitation et l'exploration, la diversification des campagnes pour éviter la surexploitation de certaines niches, et la prise en compte de l'évolution des comportements frauduleux au fil du temps.

En résumé, cet article présente une approche prometteuse pour détecter la fraude dans les services publics, en utilisant des algorithmes d'apprentissage automatique et en tirant parti des commentaires des campagnes pour améliorer en continu les résultats. Des perspectives d'amélioration et des défis futurs sont également abordés pour rendre le système encore plus efficace .

3.2.2 Méthodes semi-supervisé :

MR Barrosun et al [34] : ont proposé un cadre de travail afin d'identifié des pertes non techniques dans les systèmes d'alimentation électrique (NTL) en utilisant différents classificateurs d'apprentissage automatique. Dans cette étude, 23 classificateurs différents ont été évalués en termes de performances, d'éxécution et de fiabilité. Les classificateurs sont dérivés des 10 algorithmes les plus utilisés en bibliographie plus deux

algorithmes d'ensemble qui n'ont pas encore été utilisés pour l'identification (NTL). De plus, différentes approches ont également été évaluées, l'un d'eux consiste à effectuer un regroupement préliminaire des données avant le processus de classification et un autre est un critère de vote, combinant les résultats de nombreux classificateurs. L'étude a été réalisée à partir des données de 261 489 consommateurs d'un service public d'électricité brésilien, les données contiennent des informations historiques sur les systèmes commerciaux, techniques et opérationnels. À partir des résultats obtenus, il a été possible de conclure que les classificateurs basés sur des méthodes d'ensemble sont les plus appropriés pour l'identification non technique des pertes. Le Gradient Boosted Three a présenté un score F1 de 0,45 et la Rotation Forest a présenté une précision de 66,50 lors d'inspections réelles sur le terrain .

Joaquim L. Viegas et al [19] : Les pertes non techniques d'électricité les plus courantes sont les fraudes causées par la falsification des compteurs, qui entraînent des pertes importantes pour les compagnies d'énergie, qui ne peuvent effectuer qu'un nombre limité de contrôles, d'où un manque de données représentant les cas de fraud. Cet article propose l'utilisation d'un cadre d'apprentissage semi-supervisé (co-training by committee) pour développer un détecteur de fraude à l'électricité à partir de données ne contenant que peu d'informations sur la présence de fraude pour la majorité des échantillons. Les performances de détection sont améliorées par rapport à l'utilisation de modèles supervisés uniquement entraînés avec des données étiquetées. Dans cet article les auteurs ont utilisé le cadre d'apprentissage semi-supervisé Random Forest (RF) qui sont des modèles d'ensemble basés sur la méthode de bagging. Ces modèles combinent des arbres de décision générés à partir d'échantillons de données aléatoires pour maximiser leurs capacités de généralisation. Les arbres de décision utilisés dans les ensembles sont des arbres de classification et de régression. Les RF, utilisés comme

modèles de comité dans le cadre Co-Training by Committee (CoBC). Ces échantillons de données des consommateurs d'électricité, y compris ceux qui ne sont pas étiquetés, sont transformés à l'aide de l'ingénierie des caractéristiques pour obtenir des variables pertinentes pour l'entraînement du modèle. Le cadre RF Co-Training by Committee (CoBC) utilise ces données pour générer le modèle de comité final et les étiquettes pour les données d'entraînement non étiquetées. Le modèle de comité final peut ensuite être utilisé pour détecter la présence de fraude dans les échantillons de données des consommateurs en cours d'analyse (test). Les auteurs ont évalué un jeu de données qui contient 4232 foyers irlandais qui est composé de données de consommation d'électricité enregistrées toutes les 30 minutes pendant un an et demi. Les résultats ont montré que l'approche proposée peut atteindre de bonnes performances de classification avec un taux de vrais positifs TPR = 84%, un taux de faux positifs de FPR=11% et une aire sous la courbe AUC=0,89 avec un équilibre de classe positive de 5% et 90% d'échantillons non étiquetés dans les données d'entraînement. Avec l'apprentissage supervisé en utilisant le même modèle de base, une performance de TPR=84, FPR=16% et une aire sous la courbe AUC= 0,88 est obtenue. Les auteurs ont conclu que les performances de l'approche proposée correspondent à celles de l'apprentissage supervisé pour des pourcentages plus élevés d'échantillons étiquetés, ce qui suggère fortement que des améliorations peuvent être obtenues grâce à l'optimisation du cadre utilisé ou à l'application d'approches plus sophistiquées .

Abdel-nassir mahmat nassour [30] : Dans ce travail, l'auteur propose une méthode qui se base sur des données collectées en temps réel, décompose le réseau de distribution basse tension de la Société Nationale d'électricité (SNE) en un nombre fini de sous-réseaux de distribution, puis installe trois capteurs avec des postes très spécifiques dans ce réseau : Capteur intelligent interne (NS) : ce capteur est installé sur chaque

consommateur, collecte et transmet des données telles que l'identifiant du client, son adresse et sa consommation. Capteur externe intelligent (NH) : Ce capteur est installé sur un poteau situé au niveau du câble inférieur chez le consommateur et reçoit les données transmises par différents capteurs (NS) des clients raccordés à une même dérivation du poteau. Capteur collecteur intelligent (CH) : Il est installé sur chaque branche de la distribution. Il collecte des données aux entrées NS et NH et transmet ces données à la station de base. Ce modèle est basé sur trois valeurs principales correspondant aux nœud (CH, NS, NH). Après avoir installé ces capteurs, la fonction "sélectionner" regroupera les n oeuds après avoir calculé les distances. Ce modèle permet de déterminer la consommation de n'importe quel n oeud du réseau. Il compare ensuite la somme totale des consommations d'énergie pour toutes les sorties UT et vérifie si ce résultat est plus ou moins égal à la consommation totale d'énergie à l'entrée UT et cette comparaison fournit un écart fort qui inclut un cas de fraude existant Le gestionnaire du réseau de distribution enquête .

3.2.3 Méthodes non-supervisé

George M. Messinis et Nikos D. Hatziargyriou [27] : Dans cet article, les auteurs ont utilisé des techniques de détection de fraude non supervisée. Ils ont également utilisé la bibliothèque de détection d'évasion de Twitter pour la simuler et l'appliquer sur une base de données des consommateurs. Ils ont commencé par l'application de l'algorithme LOF sur 3639 consommateurs résidentiels issus de l'ensemble de données CER. Cependant, cette approche détecte des valeurs aberrantes qui ne sont pas nécessairement dues à une fraude. LOF a été combiné avec deux règles. La première règle consiste à calculer la différence entre la consommation moyenne avant et après la rupture. Cette différence est ensuite normalisée en la divisant par la consommation

moyenne annuelle. La deuxième règle vérifie l'écart type de la consommation après un événement de fraude. Si l'écart type est inférieur à celui d'avant, la valeur est considérée comme une fraude. Cette approche permet de calculer la densité pour tous les consommateurs. Un pourcentage fixe de consommation commettant une fraude est fixé à 5%. Le début de la fraude est choisi de manière aléatoire entre le jour 40 et le jour 290, tandis que l'intensité de la fraude suit une distribution normale avec une moyenne de 0,4 et un écart-type de 0,08. Cela donne un taux de détection de fraude BDR = 68,7%, un taux de faux positifs FPR = 1,64%, et un taux de détection DR = 68,6%.

Les auteurs ont également testé l'approche de la distribution gaussienne multivariée (MGD) pour modéliser les données. Ils ont calculé la fonction de densité de probabilité pour chaque échantillon de consommateurs, puis classé les échantillons par ordre croissant de probabilité. En combinant cette approche avec les deux règles précédentes pour exclure les valeurs aberrantes qui ne sont pas des fraudes de la liste, les résultats montrent un score F1 = 88,12%, un taux de détection de 92,77% et un taux de faux positifs de 0,93%. Ainsi, l'approche MGD se présente comme un candidat prometteur.

Les auteurs ont également proposé le clustering en utilisant les algorithmes k-Means et fuzzy c-means (FCM) avec 2 clusters, ce qui donne un score F1 de 81,3% et un taux de faux positifs de 1,67%, et un taux de détection de 90,55%. Ils ont utilisé DBSCAN pour détecter les fraudes en produisant deux clusters, dont l'un représente les valeurs aberrantes. En déterminant la densité des clusters avec les paramètres (MinPts et ε), et en appliquant la deuxième règle, les performances de l'approche sont améliorées.

Les résultats montrent un score F1 de 75,4%, un taux de détection de 71,7% et un taux de faux positifs de 0,9%. Ainsi, il est démontré que DBSCAN peut être efficace pour détecter les fraudes dans les données.

Après la comparaison de tous ces algorithmes, les auteurs ont décidé de combiner

l'algorithme de la carte auto-organisatrice (SOM), qui est un réseau de neurones souvent utilisé pour la réduction de la dimensionnalité, avec la méthode k-Means. Ils ont choisi la taille de la grille de manière heuristique (400 nœuds au total) en fonction de la taille de l'ensemble de données. Tous les échantillons appartenant au cluster le plus petit produit par k-Means sont marqués comme frauduleux. Ce dernier est exécuté 100 fois et le meilleur regroupement est ensuite choisi. Le score F1 pour cet ensemble de données particulier est de 89,66%, et le taux de faux positifs est de 0,8%. Ainsi, la combinaison SOM-k-Means montre des perspectives prometteuses pour la détection de fraudes .

Oleksandr et al [32] : Dans cet article les auteurs présente trois approches de détection d'anomalies : l'algorithme SVM à classe unique appliqué aux données audio, la décomposition LOESS des séries temporelles et l'algorithme d'apprentissage non supervisé basé sur les auto-encodeurs.

Dans la première approche, les auteurs utilisent une transformée de Fourier discrète pour réduire la dimensionnalité des données audio brutes et calculer le spectre de chaque segment. Ensuite, un modèle One-Class SVM est entraîné en utilisant les amplitudes correspondant aux fréquences des anomalies. Les tests sont effectués sur une base de données audio contenant des coups de feu. Les résultats montrent une précision relativement faible, avec une exactitude de 55% à 65% sur l'ensemble d'entraînement, mais une amélioration légère par rapport à un lancer de pièce aléatoire.

La deuxième approche repose sur la décomposition LOESS des séries temporelles. Cependant, cette méthode s'est avérée inadaptée aux données audio en raison de l'absence de tendance claire et de la présence de multiples fréquences. Les tests sur des données artificielles ont montré des résultats variables et nécessité un ajustement manuel des paramètres. Cette approche peut être utilisée dans des cas spécifiques où la structure du signal, la fréquence dominante et les propriétés des anomalies restent stables au fil du

temps, mais elle présente l'inconvénient de devoir choisir manuellement les paramètres appropriés.

Enfin, l'algorithme d'apprentissage non supervisé basé sur les auto-encodeurs est testé sur deux ensembles de données. Pour les signaux artificiels, des fenêtres de largeur 600 avec un pas de 50 sont utilisées, tandis que pour l'ensemble de données DCASE contenant des détonations, des fenêtres de déplacement de 16000 avec un pas égal sont utilisées. Le réseau de neurones comprend 8 couches LSTM avec 5 unités LSTM chacune, et des techniques d'ajustement moyen et de réponse du signal sont appliquées. La détection des anomalies se fait en utilisant le score de quantile à 99%. Les résultats montrent une précision de 87% et une correspondance précise de l'emplacement des anomalies dans 91,7% des cas.

En conclusion, l'approche des auto-encodeurs s'avère être une méthode générale et puissante pour la détection d'anomalies dans les séries temporelles, offrant de bons résultats et une grande adaptabilité aux différents types de données .

Sebastian Schmidl [36] : Dans cet article les auteurs ont testé 76 algorithmes et ils ont évalués sur 976 ensembles de données de séries temporelles. Ils ont recueilli 158 publications, chacune décrivant une approche unique pour la détection d'anomalies dans les séries temporelles, ils ont constaté que les algorithmes de détection d'anomalies dans les séries temporelles peuvent également être regroupés en trois types d'apprentissage, on a les non supervisés séparent les points anormaux de la partie normale de la série temporelle sans connaissance préalable, Les algorithmes supervisés modélisent le comportement normal et anormal dans la série temporelle et nécessitent une étape d'entraînement avant de pouvoir être utilisés sur une nouvelle série temporelle. Les algorithmes semi-supervisés tentent d'apprendre uniquement le comportement normal d'une série temporelle d'entraînement. Avant tout d'appliquer ces algorithmes ils ont

d'abords découpée les données en sous-séquences de longueur fixe. Les auteurs ont commencé à tester ces ensembles de données par les méthodes de prévision qui utilisent un modèle (continuellement) appris pour prévoir un certain nombre de pas de temps. Après Ils ont collecté un total de 1 354 ensembles de données provenant de 24 collections différentes. Ils ont développé leur propre générateur d'anomalies appelé "Good Time Series Anomaly Generator" (GutenTAG) et ont généré 194 nouvelles séries temporelles synthétiques avec des anomalies bien étiquetées. Puis ils ont évalué les algorithmes en réalisant un processus systématique de réglage des hyperparamètres afin d'obtenir le meilleur pour l'ensemble de l'évaluation. La plupart des algorithmes (87 %) ont traité avec succès plus de 70 % des ensembles de données, et de nombreux algorithmes (35 %) ont même traité plus de 99 % des ensembles de données. La grande majorité des mesures de qualité rapportées sont donc fiables. Par exemple, DWT-MLEAD a un score AUC-ROC moyen de 0,83 avec une fiabilité de 100 % et aura très probablement de bonnes performances également sur d'autres ensembles de données. En revanche, RobustPCA a un score AUC-ROC moyen de seulement 0,54 avec une fiabilité de 100 % et aura donc très probablement de mauvaises performances également sur d'autres ensembles de données. Parmi ces algorithmes ils ont trouvé k-Means, qui utilise un regroupement simple des sous-séquences de séries temporelles, est une approche multivariée très efficace qui se comporte aussi bien que d'autres représentants de la famille de distance. La détection des anomalies sur les séries temporelles univariées est en moyenne plus facile que sur les séries temporelles multivariées, avec un score AUC-ROC moyen supérieur de 0,06 pour les séries temporelles univariées. Ces résultats soulignent l'importance de comprendre les caractéristiques des ensembles de données lors du choix d'un algorithme d'anomalie approprié .

3.3 Analyse et comparaison :

Dans le tableau ci-dessous nous effectueons une étude comparative des approches proposées ci-dessus selon les 5 facteurs suivants :

- **Dataset :** indique les sources de données utilisées pour l'implémentation de l'approche pour la détection d'anomalie.
- **Approches :** les algorithmes utilisés pour détecter les anomalies.
- **Résultats :** les résultats de l'approche.
- **Avantages :** avantages de l'approche abordée.
- **Inconvénients :** inconvénients de l'approche abordée.

Titre	Auteurs	Dataset	Approches	Résultats	Avantages	Inconvénients
Unsupervised Anomaly Detection in Time Series Using LSTM-Based Autoencoders (2019)	-Oleksandr I.Provotar -Yaroslav M.Linder -Maksym M.Veres	-une base de données d'événements de test de sons rares audio -des données artificielles -jeux de données artificiels -DCASE	-SVM -STL -Auto-encodeur basé sur RNN et LSTM	-Nous testons SVM sur une base de données d'événements de test de sons rares audio.et on obtient une faible précision de 55% á 65%. -nous testons STL sur des données artificielles qui se composent de 58 fichiers avec des données de séries chronologiques de nature différente Sur certains fichiers, l'algorithme a donné des résultats décents, tandis que sur d'autres échoue complétement. - La méthode de détection d'anomalies non supervisée basée sur des autoencodeurs testée sur des jeux de données artificiels. Nous pouvons voir que le score d'anomalie de l'auto-encodeur sur l'anomalie est bien supérieur aux scores d'anomalie des autres parties du signal. -La méthode de détection d'anomalies non supervisée basée sur des autoencodeurs testée sur l'ensemble de données DCASE, La précision est de 87 %, pour les anomalies correctement détectées, le lieu exact est correct dans 91,7 % des cas.	-SVM est légérement meilleur que le tirage au sort aléatoire. -STL est simple par rapport à un algorithme d'apprentissage automatique basé sur des auto-encodeurs. -l'approche de l'auto-encodeur est suffisamment générale et puissante pour être utilisée dans tous les types de séries chronologiques.	-STL nécessite le choix manuel des paramètres. -pour SVM le Temps de calcul est grand quand K augmente[33]
Electricity fraud detection using committee semi-supervised learning(2018)	-Joaquim L.Viegas -Nuno M.Cepeda -Susana M.Vieira	- un ensemble de données basé sur des lectures de consommation d'électricité réelle de 4232 ménages irlandais enregistrées à des intervalles de 30 minutes, pendant un an et demi.	-CoBC -Random Forest (RF)	avec l'approche non supervisé : -TPR = 84% - FPR=11% - AUC=0,89% avec l'approche supervisé : - TPR=84% -FPR=16% - AUC= 0,88%	-Il permet d'obtenir une prédiction fiables grâce à son système d'arbres décisionnels. - Une gestion efficace de grands ensembles de données	-Entrainement plus lent[31] - Stratégie d'élagage délicate[29]
Fraud Detection in Energy Consumption : A Supervised Approach (2016)	- Bernat Coma-Puig - Josep Carmona - Ricard Gavalda - Santiago Alcoverro - Victor Martin	- Données des Entreprises de Services Publics	-K-nearest neighbors -Support Vector Machines -Random Forests -Gradient Boosting -AdaBoost	-Précision jusqu'à 15x par rapport à la méthodologie de référence	- Utilisation de l'apprentissage automatique pour détecter la fraude	- nécessite d'un ajustement minutieux des hyperparamètres [18]
Unsupervised Classification for Non-Technical Loss Detection	- George M. Messinis - Nikos D. Hatziargyriou	- réseau de distribution hellénique (HEDNO)	-LOF -FCM -DBSCAN -SOM-k-Means	avec LOF : -BDR = 68,7% -FPR = 1,64% - DR = 68,6% avec FCM : -DR= 92,77% - FPR= 0,93% - F1 = 88,12% avec FCM : -DR = 90,55% -FPR = 1,67% - F1 = 81,3% avec DBSCAN : -DR = 71,7% -FPR = 0,93%. -F1 =75,4% avec SOM-k-Means : -FPR = 0,8% - F1= 89,66%	-L'algorithme est très simple -DBSCAN ne nécessite pas qu'on lui précise le nombre de clusters à trouver -Il est capable de gérer les données aberrantes en les éliminant du processus de partitionnement	- LOF détecte les valeurs aberrantes qui ne sont pas nécessairement dues à la fraude.[26] -DBSCAN n'est pas capable de gérer des clusters de densités différentes.[26]
Abdel-nassir mahamat nassour	Algorithmes de structuration et de collecte de données électriques, application à la détection du vol d'électricité sur le réseau de la SNE	réseau électrique de la ville de N'Djamena	-algorithme basé sur le clustering	-Trouver une grande variation entre l'énergie consommée à l'entrée de l'énergie consommée à la sortie, ce que signifier le vol d'électricité	-il s'applique à n'importe quel type de réseau.	- manque de communication du puits vers les capteurs[30]
Anomaly Detection in Time Series : A Comprehensive Evaluation	-Sebastian Schmidl -Phillip Wenig -Thorsten Papenbrock	-976 ensembles de données de séries temporelles	-k-Means -KNN -LOF -LSTM-AD - ...	-La détection des anomalies sur les séries temporelles univariées est en moyenne plus facile que sur les séries temporelles multivariées.	- KNN est facile à comprendre et à interpréter[4] -Il est utile pour les données non linéaires et est considéré comme un algorithme polyvalent puisqu'il est utile pour la classification et la régression.[4] - le K-means peut identifier des groupes de données inconnus à partir d'ensembles de données complexes [5]	- les grandes tailles de jeu de données entrainera un temps beaucoup plus large[4] - nombre voisins K n'est pas quelque chose d'évidente[?] - Manque de cohérence[5] - Ensemble non optimal de clusters[5]
Évaluation des classificateurs pour l'identification des pertes non techniques dans les systèmes d'alimentation électrique	-Raphaël MR Barrosun -Edson G. da Costab -Jalberth F. Araujob	un ensemble de données de 261 489 consommateurs d'un service public d'électricité brésilien.	-SVM -GBT -XGBT -RF ...	-La détection des anomalies sur les séries temporelles univariées est en moyenne plus facile que sur les séries temporelles multivariées.	- KNN est facile à comprendre et à interpréter[4] -Il est utile pour les données non linéaires et est considéré comme un algorithme polyvalent puisqu'il est utile pour la classification et la régression.[4] - le K-means peut identifier des groupes de données inconnus à partir d'ensembles de données complexes [5]	- les grandes tailles de jeu de données entrainera un temps beaucoup plus large[4] - nombre voisins K n'est pas quelque chose d'évidente[?] - Manque de cohérence[5] - Ensemble non optimal de clusters[5]

TABLEAU 3.1 – Étude comparative des travaux connexes 1

3.4 Conclusion

Dans ce chapitre, nous avons établi l'état actuel de l'art de la détection de la fraude électrique, qui représente une étude comparative de tous les travaux connexes, que nous avons abrégés, nous avons présenté dans un tableau détaillé décrivant chaque accès à des documents de synthèse, suivi de chaque tâche un court paragraphe qui le résume. Dans le chapitre suivant, nous présentons notre approche et ses différentes étapes.

Chapitre 4

Approche proposée

4.1 Introduction

La fraude électrique est un défi majeur dans le secteur de l'énergie, entraînant des pertes financières importantes. Les méthodes traditionnelles de détection de fraudes reposent souvent sur des règles prédéfinies et des processus manuels qui ont des limites en termes d'efficacité et de précision. Grâce à l'intelligence artificielle, de nouvelles approches émergent pour améliorer la détection de fraudes en électronique.

Dans ce chapitre, nous présentons notre approche de la détection de fraudes électrique à l'aide de l'apprentissage automatique. Notre objectif est de proposer l'une des techniques d'apprentissage automatique pour analyser les données électriques en adoptant une approche pour détecter les clients fraudeurs.

4.2 Plateformes et outils de développement

Dans cette section nous présentons l'environnement de développement,le langage de programmation, et les Bibliothèques que nous avons utilisé :

4.2.1 Environnement de développement

- **Anaconda** : Anaconda est une distribution gratuite et open source des langages de programmation Python et R appliqués au développement d'applications dédiées au machine learning, qui vise à simplifier la gestion et le déploiement des packages[11].
- **Jupyter notebook** : Jupyter Notebook est un environnement de développement interactif basé sur le code et les données. Son interface flexible permet aux utilisateurs de configurer et d'organiser des flux de travail en science des données, en informatique scientifique, en journalisme informatique et en apprentissage automatique. La conception modulaire invite les extensions à étendre et enrichir la fonctionnalité[2].

4.2.2 Langage de programmation

- **Python** :Python est un langage de programmation puissant et facile à apprendre. Il a des structures de données de haut niveau efficaces et une approche simple mais efficace de la programmation orientée objet. La syntaxe et le typage dynamique de Python, ainsi que sa nature interprétée, en font un langage idéal pour les scripts et le développement rapide d'applications dans de nombreux domaines sur la plupart des plates-formes [35].

4.2.3 Bibliothèques python

- **Pandas :** utilisée pour la manipulation des données, en particulier pour lire les données a partir d'un fichier csv ou Excel et effectuer des opérations sur les tableaux de données.

- **Numpy :** une bibliothèque utilisée pour les calculs numériques et les opérations sur les tableaux multidimensionnels [3].
- **Matplotlib :** une bibliothèque de visualisation utilisée pour tracer des graphiques et des figures.
- **Sklearn :** Scikit-learn est une bibliothèque Python qui fournit une interface standard pour la mise en œuvre d'algorithmes d'apprentissage automatique. Il comprend d'autres fonctions d'assistance qui font partie intégrante du pipeline d'apprentissage automatique, telles que les étapes de prétraitement des données, les techniques de rééchantillonnage des données,paramètres d'évaluation et une interface de recherche pour régler/optimiser les performances de l'algorithme[**?**].

4.3 Contribution

Notre projet consiste à détecter la fraude électrique, c'est-à-dire de permettre a l'entreprise SONELGAZ de savoir les clients fraudeurs. Pour atteindre cet objectif il faut suivre plusieurs étapes qui doivent être effectuées pour obtenir de meilleurs résultats.

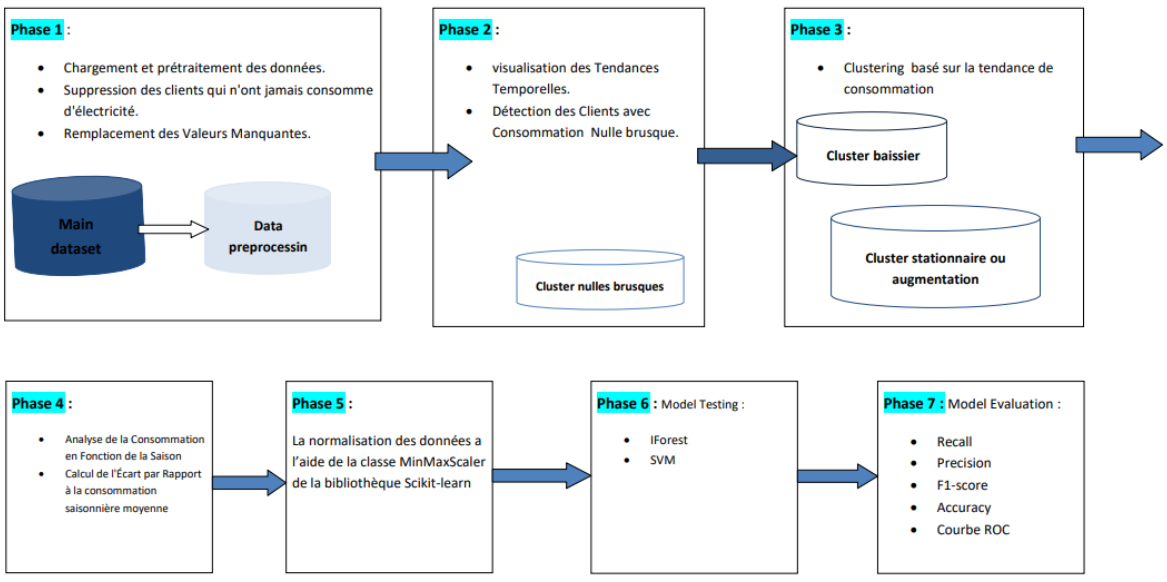

FIGURE 4.1 – Schéma global des différentes étapes à suivre lors de l'implémentation

4.3.1 Collecte des données

La collecte des données constitue une étape fondamentale dans notre projet de détection de fraude. Elle établit les bases pour le traitement et l'évaluation des informations sélectionnées. Pour notre projet de détection de fraude électrique, nous avons choisi une approche qui repose sur l'utilisation de deux jeux de données distincts, chacun ayant un rôle spécifique au sein de notre analyse.

Ces deux jeux de données authentiques ont été extraits en collaboration avec l'entreprise SONELGAZ de BEJAIA, renforçant ainsi leur pertinence et leur authenticité. Ils nous fournissent un aperçu précis des schémas de consommation électrique et servent de fondation à notre analyse approfondie.

4.3.1.1 Ensemble de Données 1

Notre premier ensemble de données est le fondement de notre exploration et de notre développement. Ce jeu de données n'est pas étiqueté, ce qui signifie qu'il ne contient pas d'informations préexistantes sur les consommations en tant que frauduleuses ou non frauduleuses. Ce choix nous permet d'explorer des méthodes non supervisées pour la détection de fraude, en sondant les subtilités des habitudes de consommation électrique. Sa taille est de 6.5 MB et il contient un total de 1699 clients frauduleux et non frauduleux de consommations électriques. Ces enregistrements couvrent la période de janvier/2006 à décembre/2019.

L'ensemble de données 1 se compose de trois colonnes principales :

- **Client :** cette colonne représente l'identifiant client associé à chaque consommation. Chaque client est attribué un numéro unique pour permettre l'identification et le suivi des consommations spécifique a un client.
- **Concat :** cette colonne contient des valeurs de dates au format yyyy-mm-dd, re-

présentant la date de chaque consommation.

- **Valeur :** cette colonne représente la consommation d'électricité de chaque client en kilowattheure (KWH).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 285432 entries, 0 to 285431
Data columns (total 3 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   client  285432 non-null  int64
 1   concat  285432 non-null  int64
 2   valeur  193438 non-null  float64
dtypes: float64(1), int64(2)
memory usage: 6.5 MB
None
```

	client	concat	valeur
count	285432.000000	285432.000000	1.934380e+05
mean	849.000000	67012.500000	3.241955e+04
std	490.459828	34520.586002	1.615338e+05
min	0.000000	12006.000000	1.000000e+00
25%	424.000000	39509.250000	2.299000e+03
50%	849.000000	67012.500000	6.494000e+03
75%	1274.000000	94515.750000	2.132475e+04
max	1698.000000	122019.000000	7.404162e+06

TABLEAU 4.1 – Les informations du data-frame non étiquetées et le résumé statistique des colonnes numériques

4.3.1.2 Ensemble de Données 2

Le deuxième ensemble de données, extrait du premier, joue un rôle crucial dans l'évaluation de nos modèles. Il couvre une période de deux années spécifiques, de janvier/2017 à décembre/2018, et est étiqueté. Chaque consommation dans cet ensemble est classée comme frauduleuse ('1') ou non frauduleuse ('0'). Cet étiquetage nous permet d'évaluer la performance réelle de nos modèles en utilisant des métriques de classification standard. Sa taille est de 1.2 MB, avec un total de 1699 clients.

L'ensemble de données 2 est composé de quatre colonnes, incluant une colonne supplémentaire par rapport à l'ensemble de données 1 :

- **Fraud :** Cette colonne contient des variables binaires, indiquant une consommation frauduleuse ('1') ou non frauduleuse ('0'). Cette variable cible est celle que nous cherchons à prédire à l'aide de nos modèles d'apprentissage automatique.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40805 entries, 0 to 40804
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   client  40805 non-null  int64
 1   concat  40776 non-null  datetime64[ns]
 2   valeur  34686 non-null  float64
 3   fraud   40805 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(2)
memory usage: 1.2 MB
None
```

	client	valeur	fraud
count	40805.000000	3.468600e+04	40805.000000
mean	849.099326	3.513378e+04	0.019115
std	490.466669	1.833813e+05	0.136932
min	0.000000	1.000000e+00	0.000000
25%	424.000000	2.376000e+03	0.000000
50%	849.000000	6.547500e+03	0.000000
75%	1274.000000	2.175925e+04	0.000000
max	1698.000000	4.602835e+06	1.000000

TABLEAU 4.2 – Les informations du data-frame étiquetées et le résumé statistique des colonnes numériques

Avec ces deux ensembles de données complémentaires, notre projet s'articule autour d'une approche globale de détection de fraude électrique. Nous explorons des méthodes non supervisées pour la détection de fraude tout en évaluant la performance de nos modèles sur des données étiquetées, renforçant ainsi l'efficacité et la pertinence de notre projet. L'objectif le plus important de la collecte de données est de s'assurer que les données sont complètes et fiables en informations qui permet de mieux évaluer des résultats et de mieux anticiper les probabilités et les tendances à venir [7].

4.3.2 Prétraitement

Après la collecte des données, l'étape suivante réalisée est le prétraitement, ce dernier est très important pour préparer les données avant l'entrainement de notre modèle d'apprentissage automatique. Ce prétraitement consiste en plusieurs étapes pour nettoyer, normaliser et préparer les données de manière approprie. Les étapes du prétraitement que nous avons effectué sont les suivantes :

4.3.2.1 Exploration de données

l'exploration visuelle des donnée aide à obtenir et connaître des informations et caractéristiques approfondies et clairs sur les ensemble de données et les variables. On remarque que :

- **Ensemble de données 1 :**
  - Le nombre de clients (1699 clients).
  - Les valeurs manquantes (91994 valeurs manquantes sur 285432 dans la colonne valeur).
  - La taille de dataset (6.5 MB).
  - Les colonnes existantes(client, concat et valeur).
- **Ensemble de données 2 :**
  - Le nombre de clients (1699 clients) dont 164 sont fraudeurs durant l'année 2017 et 2018.
  - Les valeurs manquantes (29 valeurs manquantes sur 40805 dans la colonne concat et 6119 sur 40805 valeurs manquantes dans la colonne valeur).
  - La taille de dataset (1.2 MB).
  - Les colonnes existantes(client, concat, valeur et fraud)
- **Conversion de la clonne concat en type datetime :** dans cette étape nous avons extrait les informations temporelles de la colonne concat afin d'obtenir les années, les mois et les jours correspondants de manière plus appropriée pour l'analyse ultérieur.

4.3.2.2 Nettoyage des données

Le nettoyage des données est une étape essentielle pour garantir la qualité de notre ensemble de données. Dans notre cas, nous avons effectué plusieurs actions pour nettoyer les données. Tout d'abord, nous avons identifié les clients qui n'ont jamais consommé d'électricité au cours de toutes les années et les avons supprimés pour l'ensemble de données 1, ensuite nous avons éliminé ces clients de l'ensemble de données 2 , car ces enregistrements ne seraient pas pertinents pour notre analyse. De plus, nous avons supprimé les lignes qui contenaient des valeurs manquantes dans la colonne concat pour l'ensemble de données 2, car ces lignes représentaient une répétition de la ligne précédente, mais sans la valeur dans la colonne concat.

Pour traiter les valeurs manquantes, nous avons regroupé les données par client et par année, puis nous avons évalué le nombre de valeurs manquantes pour chaque groupe, si le nombre de valeurs manquantes est inférieur ou égale à trois on les remplace par la moyenne de consommation pour l'année correspondante sinon nous avons pris la décision de les remplacer par des zéro.

Cette étape permis d'obtenir un ensemble de données complet et prêt à être utilisée.

client	0	client	0
concat	0	concat	0
valeur	91994	valeur	0
dtype: int64		annee	0
		mois	0
		dtype: int64	

(a) Avant et après le nettoyage de l'ensemble de données 1

client	0	client	0
concat	29	concat	0
valeur	6119	valeur	0
fraud	0	fraud	0
dtype: int64		annee	0
		mois	0
		jour	0
		dtype: int64	

(b) Avant et après le nettoyage de l'ensemble de données 2

TABLEAU 4.3 – Comparaison avant et après le nettoyage de données

4.3.3 Visualisation des Tendances temporelles et division des clusters

Dans le cadre de notre étude sur la consommation d'électricité, une analyse approfondie des tendances temporelles nous a permis de mieux comprendre les variations de la consommation au fil du temps. Cette exploration s'est révélée cruciale pour identifier les changements significatifs dans les habitudes de consommation des clients.

4.3.3.1 Analyse Individuelle des Clients

Nous avons commencé par regrouper les données par client et tracé les graphiques de consommation d'électricité au fil des mois pour chaque client individuel. Cette approche nous a permis d'observer les fluctuations spécifiques à chaque client, révélant des schémas uniques de consommation.

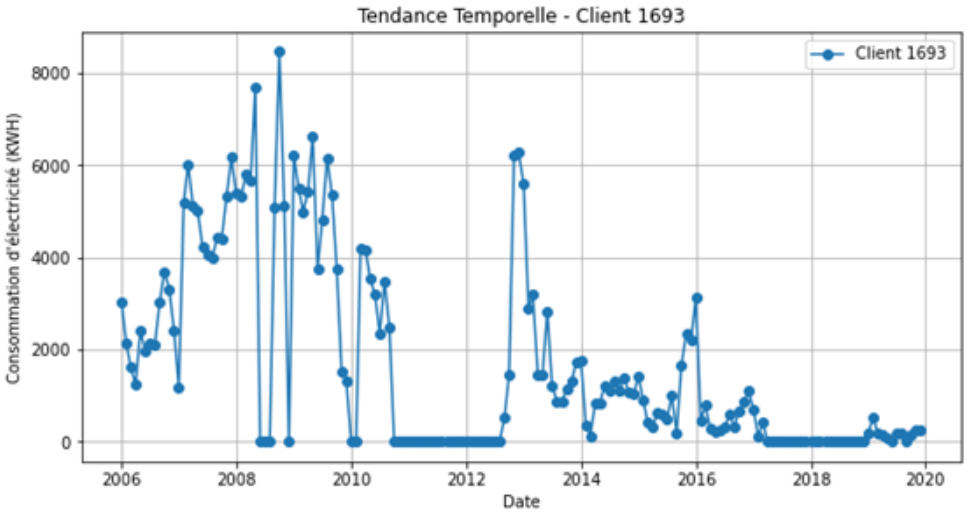

FIGURE 4.2 – La consommation d'électricité du client 1693

4.3.3.2 Tendances Temporelles Mensuelles

Par la suite, nous avons agrégé les données par mois, ce qui nous a donné une vue d'ensemble des tendances mensuelles de la consommation d'électricité. Cette visualisation globale nous a permis de repérer les mois de consommation élevée ou faible à travers la période étudiée.

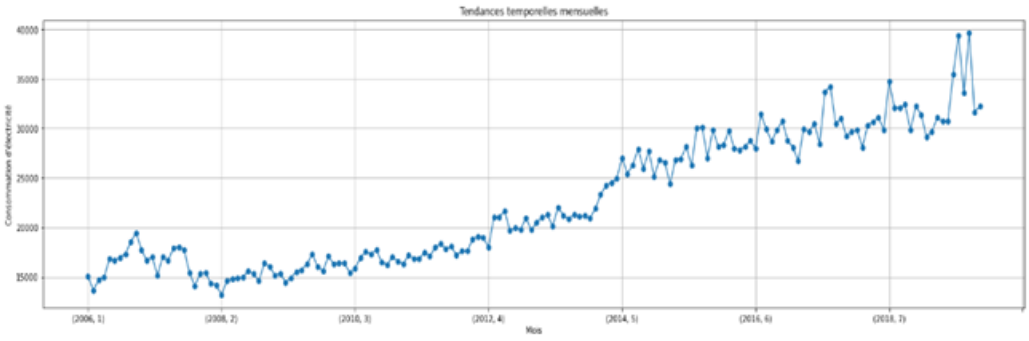

FIGURE 4.3 – Tendances temporelles mensuelles

4.3.4 clustering et Analyse des Tendances de Consommation

Après avoir effectué une analyse approfondie des tendances temporelles de la consommation d'électricité, nous avons constaté que trois grandes tendances caracté-

risent les majorité des clients donc Nous avons également exploré les critères qui ont sous-tendu cette division, en mettant en lumière les caractéristiques clés qui ont influencé la séparation des clusters.

4.3.4.1 Identification des Clients à Consommation Nulle Brusque

Avant d'entamer le processus de division en clusters, nous avons préalablement identifié les clients dont les schémas de consommation étaient atypiques. À cet effet, nous avons utilisé une méthodologie rigoureuse pour détecter les clients présentant des mois consécutifs de consommation nulle au cours d'une année. Ces clients, dénommés "clients à consommation nulle brusque", ont été écartés de notre analyse afin d'assurer la crédibilité de nos résultats.

client	Tendance de consommation
993	Clients nul brusque

TABLEAU 4.4 – Clients à consommation nulle brusque

4.3.4.2 Regroupement en Clusters selon les Tendances de Consommation

Pour catégoriser les clients restants en clusters distincts, nous avons recouru à l'algorithme K-Means. L'objectif était de regrouper les clients partageant des caractéristiques de tendances de consommation similaires. Cette étape nous a permis d'obtenir un aperçu plus précis des différentes dynamiques de consommation au sein de notre ensemble de données.

4.3.4.3 Facteurs Déterminants la Division en Clusters

La division des clusters a été basée sur l'évaluation des tendances de consommation au fil du temps pour chaque client. Plus spécifiquement, nous avons calculé les pentes de régression linéaire ajustées à chaque série temporelle de consommation mensuelle.

La pente résultante a été utilisée pour établir si la tendance était à la hausse, à la baisse ou stationnaire.

4.3.4.4 Étiquetage des Clusters

Suite à l'évaluation des tendances de consommation, nous avons attribué des étiquettes distinctes aux clusters. Les clients présentant une tendance de consommation à la baisse ont été rassemblés dans le cluster "Tendance baissière", tandis que ceux présentant une tendance stationnaire ou en augmentation ont été regroupés dans le cluster "Tendance stationnaire ou augmentation".

client	Tendance de consommation
269	Tendance baissière
402	Tendance stationnaire ou augmentation

TABLEAU 4.5 – Étiquetage des clusters

En poursuivant notre analyse, nous plongerons plus profondément dans les caractéristiques distinctives de chaque cluster, en évaluant les consommations moyennes, les écarts saisonniers et les implications générales des tendances observées. Cette étape nous permettra de tirer des conclusions plus approfondies quant aux facteurs influençant les schémas de consommation d'électricité de nos clients.

4.3.5 Analyse des Variations Saisonnières

Après avoir segmenté nos clients en clusters distincts en fonction de leurs tendances de consommation, nous avons cherché à explorer davantage les variations saisonnières de la consommation d'électricité. Comprendre comment la consommation fluctue au fil des saisons peut fournir des aperçus précieux pour ajuster les stratégies tarifaires,

anticiper la demande et mieux servir nos clients. Pour ce faire, nous avons réalisé les étapes suivantes :

- **Attribution de la Saison en Fonction du Mois :** Nous avons commencé par ajouter une nouvelle colonne à notre ensemble de données, appelée "saison", qui attribue une saison à chaque mois en fonction du mois lui-même. Par exemple, les mois de janvier, février et mars ont été attribués à la saison 1, tandis que les mois d'avril, mai et juin ont été attribués à la saison 2, et ainsi de suite. Cette étape nous a permis de classer les consommations en fonction de la saisonnalité.

	client	concat	valeur	annee	mois	saison
0	0	2006-01-01	0.0	2006	1	1
1	0	2006-02-01	0.0	2006	2	1
2	0	2006-03-01	0.0	2006	3	1
3	0	2006-04-01	0.0	2006	4	2
4	0	2006-05-01	0.0	2006	5	2
...	...	...	...	...	...	...
285427	999	2019-08-01	343.0	2019	8	3
285428	999	2019-09-01	200.0	2019	9	3
285429	999	2019-10-01	153.0	2019	10	4
285430	999	2019-11-01	138.0	2019	11	4
285431	999	2019-12-01	169.0	2019	12	4

279552 rows × 6 columns

TABLEAU 4.6 – Capture après l'attribution de la saison

- **Calcul de la Consommation Moyenne par Saison :** Nous avons ensuite calculé la consommation moyenne pour chaque client et pour chaque saison. Cela a été accompli en regroupant les données par client et par saison, puis en calculant la moyenne des valeurs de consommation. Cette étape a généré un tableau qui illustre la consommation moyenne de chaque client pour chaque saison de l'année.

client	consommation_moyenne_saison_1	consommation_moyenne_saison_2	consommation_moyenne_saison_3	consommation_moyenne_saison_4
0	7841.939394	6209.619048	8665.690476	7852.738095
1	23776.642857	22127.976190	19753.190476	21950.571429
2	12469.809524	14832.238095	12934.952381	16410.000000
3	52727.214286	39008.428571	27683.675325	54176.452381
4	3669.547619	2070.428571	2369.567100	3361.619048
...	...	...	...	...
1694	1433.738095	2586.804762	4205.452381	3334.952381
1695	4699.309524	5259.261905	7129.642857	4774.523810
1696	14598.642857	17799.119048	17293.761905	22818.738095
1697	9629.320346	7296.142857	8772.857143	12099.476190
1698	2128.904762	1858.928571	1846.619048	2003.571429

1664 rows × 4 columns

TABLEAU 4.7 – Capture de la base de données après avoir calculer les consommations moyennes saisonnières

- **Calcul de l'écart saisonnier :** Pour évaluer plus en détail les variations saisonnières, nous avons calculé l'écart entre la consommation réelle et la consommation moyenne de la saison correspondante pour chaque consommation individuelle. Cette différence, appelée "écart saisonnier", nous permet de quantifier la déviation d'une consommation par rapport à la moyenne saisonnière attendue. Un écart positif indique une consommation supérieure à la moyenne saisonnière, tandis qu'un écart négatif indique une consommation inférieure.

saison	consommation_moyenne_saison_1	consommation_moyenne_saison_2	consommation_moyenne_saison_3	consommation_moyenne_saison_4	ecart_saison
1	7841.939394	6209.619048	8665.690476	7852.738095	-7841.939394
1	7841.939394	6209.619048	8665.690476	7852.738095	-7841.939394
1	7841.939394	6209.619048	8665.690476	7852.738095	-7841.939394
2	7841.939394	6209.619048	8665.690476	7852.738095	-6209.619048
2	7841.939394	6209.619048	8665.690476	7852.738095	-6209.619048
...	...	...	...	...	...
3	36.523810	55.928571	82.619048	42.833333	260.380952
3	36.523810	55.928571	82.619048	42.833333	117.380952
4	36.523810	55.928571	82.619048	42.833333	110.166667
4	36.523810	55.928571	82.619048	42.833333	95.166667
4	36.523810	55.928571	82.619048	42.833333	126.166667

TABLEAU 4.8 – Capture de la base de données après avoir calculer l'écart saison

En résumé, cette analyse des variations saisonnières nous a permis de mieux comprendre comment la consommation d'électricité varie en fonction des saisons pour chaque client et chaque cluster. Les écarts saisonniers obtenus fournissent des informations précieuses sur les habitudes de consommation saisonnières des clients.

4.3.6 Normalisation des données

La normalisation des données consiste en un lissage, une normalisation et une agrégation des données [17] qui vise à mettre les données entre 0 et 1. Dans notre cas, la normalisation est effectuée à l'aide de la classe MinMaxScaler de la bibliothèque Scikit-learn (sklearn.preprocessing) sur les colonnes 'valeur', 'consommation_moyenne_saison_1', 'consommation_moyenne_saison_2', 'consommation_moyenne_saison_3' ,'consommation_moyenne_saison_4' , 'ecart_saison'.

4.4 Sélection de la méthode

Après la division de données, l'étape suivante est la sélection de la méthode pour la détection de fraude. Cette étape est celle qui permet de mettre en œuvre divers algorithmes d'exploration de données. Deux modèles ont été testé pour la détection des clients frauduleux sont les suivants :

4.4.1 Isolation forest

Est un algorithme d'apprentissage automatique non supervisé, Il est souvent utilisé dans les domaines tels que la détection de fraude, la détection d'intrusion ou la détection d'anomalies dans les données financières. La forêt d'isolation est un ensemble d'arbres d'isolation qui sont employés en commun pour séparer les données. Construire une forêt d'isolation (iForest) revient à construire un nombre t d'arbre binaires. Il suffit donc de relancer les mêmes étapes pour construire plusieurs arbres binaires qui vont former la forêt d'isolation[12] .

L'algorithme iForest a des hyper-paramètres qui contrôlent le comportement de l'algorithme lors de la création des sous-ensembles et de l'évaluation de l'anormalité des instances.

Voici les hyper-paramètres que nous avons ajustés lors de l'utilisation de cet algorithme :

- **n_estimators :** il s'agit du nombre d'arbres à utiliser dans la forêt d'isolation. Dans notre étude on l'a spécifié en utilisant la méthode DBSCAN qui est un algorithme de Clustering qui peut être utilisé pour estimer le nombre de clusters présents dans les données. En utilisant DBSCAN, nous avons identifié le nombre de clusters pertinents dans nos données et nous avons utilisé ce nombre comme valeur pour n_estimators.

- **max_samples :** c'est le nombre de caractéristiques à utiliser lors de la division des données à chaque étape de l'algorithme.
- **contamination :** c'est le pourcentage d'instance suspectées d'être des anomalies dans nos données.
- **random_state :** est une graine aléatoire utilisée pour l'initialisation des nombres aléatoire dans l'algorithme. Fixer random_forest a une valeur spécifique garanti la productibilité des résultats.

Dans le cadre de notre projet visant à détecter les fraudes dans la consommation d'électricité, nous avons employé l'algorithme de l'Isolation Forest comme outil pour identifier les fraudeurs potentiels. Pour ce faire, nous avons entraîné cet algorithme sur notre base de données, en utilisant des colonnes spécifiques telles que l'écart saisonnier, la valeur réelle de la consommation et les consommations moyennes par saison (saisons 1 à 4).

L'approche que nous avons mise en place pour déterminer quels clients pourraient être considérés comme fraudeurs repose sur deux conditions spécifiques, appliquées sur les prédictions de l'algorithme Isolation Forest :

Première Condition : Si l'écart saisonnier est négatif et que la consommation réelle est inférieure de 20% à la consommation moyenne de la saison correspondante, alors le client est classé comme un candidat fraudeur. Cette condition vise à identifier les cas où un client consomme moins que la normale pendant une saison donnée, ce qui pourrait être un indicateur d'une activité frauduleuse.

Deuxième Condition : Si l'écart saisonnier est positif et que la consommation réelle est supérieure de 20% à la consommation moyenne de la saison correspondante, alors le client est également considéré comme un candidat fraudeur. Cette condition tente de repérer les situations où un client consomme davantage que prévu, ce qui pourrait

indiquer une tentative de manipulation frauduleuse.

En utilisant ces deux conditions, nous avons pu améliorer l'efficacité de l'algorithme Isolation Forest pour la détection des fraudeurs

4.4.2 One-Class Support Vector Machine (SVM)

Un autre algorithme crucial que nous avons utilisé pour renforcer notre projet de détection de fraudes dans la consommation d'électricité est la Machine à Vecteurs de Support à Classe Unique, communément appelée One-Class SVM. Cet algorithme, comme l'Isolation Forest, est également un modèle d'apprentissage automatique non supervisé, mais avec une approche différente pour détecter les anomalies.

SVM choisit les points / vecteurs extrêmes qui aident à créer l'hyperplan. Ces cas extrêmes sont appelés vecteurs de support, et donc l'algorithme est appelé machine de vecteur de support [23].

Dans notre mise en œuvre, nous avons suivi les étapes suivantes :

Préparation des données : nous avons converti les colonnes nécessaires en types numériques pour que les données soient compatibles avec l'algorithme.

Seuil de tolérance : Nous avons défini un seuil de tolérance de 20% pour les écarts saisonniers. Ce seuil sera utilisé pour déterminer si une consommation est anormale en fonction de la consommation moyenne saisonnière correspondante.

Création du modèle : Nous avons instancié un modèle One-Class SVM en spécifiant les hyper-paramètres suivants :

- **nu :** c'est le paramètre qui définit la proportion d'observations à considérer comme des anomalies.
- **kernel :** nous avons choisi le noyau RBF (Radial Basis Function) pour le modèle SVM, car il peut capturer des relations non linéaires entre les données.

- **gamma :** ce paramètre contrôle l'influence des points de données dans la fonction de décision du modèle.

Entraînement de la méthode : Le méthode a été entraîné sur les données des clients baissiers en utilisant les mêmes colonnes que dans le cas de l'Isolation Forest.

Prédiction des anomalies : Nous avons utilisé le modèle entraîné pour prédire les anomalies pour tous les clients dans notre base de données.

Identification des fraudeurs : Les clients considérés comme fraudeurs par la One-Class SVM sont ceux pour lesquels les conditions définies sur l'écart saisonnier et la consommation réelle sont satisfaites. Si l'écart saisonnier est négatif et que la consommation réelle est inférieure au seuil défini, ou si l'écart saisonnier est positif et que la consommation réelle dépasse le seuil défini, alors le client est marqué comme un candidat fraudeur.

Grâce à cette approche, nous avons pu améliorer davantage notre capacité à détecter les comportements frauduleux parmi les clients de notre base de données de consommation d'électricité. L'utilisation combinée de l'algorithme One-Class SVM et des conditions spécifiques a contribué à augmenter notre précision dans la détection des fraudeurs potentiels, tout en offrant une approche complémentaire à celle de l'Isolation Forest.

4.5 Évaluation des méthodes

Après la détection de fraude, une évaluation est nécessaire pour bien déterminer les métriques de performances de l'approche choisie, les résultats de modèle ont été évaluer en analysant quelques critères a savoir la précision, rappel et le f1-score.

L'évaluation des performances est une étape très nécessaire pour tester la qualité de modèle, afin d'assurer la fiabilité des résultats prédicatifs de modèle.

Le tableau ci-dessous représente la précision, le rappel et le F1-score pour les classifieurs iForest et OneClassSVM :

<table>
<tr><th>méthodes</th><th>Clusters</th><th>Hyperparamètres</th><th>Résultats</th></tr>
<tr><td rowspan="3">OneClass SVM</td><td rowspan="3">Cluster 1 : Les clients qui ont des nuls brusques</td><td>nu=0.2
gamma=0.1
seuil = 0.2</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.90</td><td>0.67</td><td>0.77</td><td>894</td></tr>
<tr><td>fraudeur</td><td>0.09</td><td>0.31</td><td>0.15</td><td>99</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.63</td><td>993</td></tr>
<tr><td>macro avg</td><td>0.50</td><td>0.49</td><td>0.46</td><td>993</td></tr>
<tr><td>weighted avg</td><td>0.82</td><td>0.63</td><td>0.70</td><td>993</td></tr>
</table>
</td></tr>
<tr><td>nu=0.1
gamma=0.1
seuil = 0.2</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.90</td><td>0.67</td><td>0.77</td><td>894</td></tr>
<tr><td>fraudeur</td><td>0.09</td><td>0.31</td><td>0.15</td><td>99</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.63</td><td>993</td></tr>
<tr><td>macro avg</td><td>0.50</td><td>0.49</td><td>0.46</td><td>993</td></tr>
<tr><td>weighted avg</td><td>0.82</td><td>0.63</td><td>0.70</td><td>993</td></tr>
</table>
</td></tr>
<tr><td>nu=0.2
gamma=0.1
seuil sup = 1.1
seuil inf = 0.9</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.89</td><td>0.62</td><td>0.73</td><td>894</td></tr>
<tr><td>fraudeur</td><td>0.09</td><td>0.32</td><td>0.14</td><td>99</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.59</td><td>993</td></tr>
<tr><td>macro avg</td><td>0.49</td><td>0.47</td><td>0.43</td><td>993</td></tr>
<tr><td>weighted avg</td><td>0.81</td><td>0.59</td><td>0.67</td><td>993</td></tr>
</table>
</td></tr>
<tr><td colspan="4">Suite sur la page suivante</td></tr>
</table>

Tableau 4.9 – Suite de la page précédente

<table>
<tr><th>méthodes</th><th>Cluster</th><th>Hyperparamètre</th><th>Résultats</th></tr>
<tr><td></td><td>Cluster 2 : Les clients avec une certaine station-narité ou une légère augmen-tation</td><td>nu=0.2
gamma=0.1
seuil = 0.2</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.89</td><td>0.97</td><td>0.93</td><td>359</td></tr>
<tr><td>fraudeur</td><td>0.08</td><td>0.02</td><td>0.04</td><td>43</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.87</td><td>402</td></tr>
<tr><td>macro avg</td><td>0.48</td><td>0.49</td><td>0.48</td><td>402</td></tr>
<tr><td>weighted avg</td><td>0.80</td><td>0.87</td><td>0.83</td><td>402</td></tr>
</table>
</td></tr>
<tr><td></td><td></td><td>nu=0.1
gamma=0.1
seuil = 0.2</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.89</td><td>0.97</td><td>0.93</td><td>359</td></tr>
<tr><td>fraudeur</td><td>0.08</td><td>0.02</td><td>0.04</td><td>43</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.87</td><td>402</td></tr>
<tr><td>macro avg</td><td>0.48</td><td>0.49</td><td>0.48</td><td>402</td></tr>
<tr><td>weighted avg</td><td>0.80</td><td>0.87</td><td>0.83</td><td>402</td></tr>
</table>
</td></tr>
</table>

Suite sur la page suivante

Tableau 4.9 – Suite de la page précédente

<table>
<tr><th>méthodes</th><th>Cluster</th><th>Hyperparamètre</th><th>Résultats</th></tr>
<tr><td rowspan="3"></td><td></td><td>nu=0.2
gamma=0.1
seuil sup = 1.1
seuil inf=0.9</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.89</td><td>0.96</td><td>0.92</td><td>359</td></tr>
<tr><td>fraudeur</td><td>0.12</td><td>0.05</td><td>0.07</td><td>43</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.86</td><td>402</td></tr>
<tr><td>macro avg</td><td>0.51</td><td>0.50</td><td>0.50</td><td>402</td></tr>
<tr><td>weighted avg</td><td>0.81</td><td>0.86</td><td>0.83</td><td>402</td></tr>
</table>
</td></tr>
<tr><td>Cluster 3 : Les clients qui ont une tendance baissière</td><td>nu=0.2
gamma=0.1
seuil = 0.2</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.93</td><td>1.00</td><td>0.97</td><td>250</td></tr>
<tr><td>fraudeur</td><td>1.00</td><td>0.05</td><td>0.10</td><td>19</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.93</td><td>269</td></tr>
<tr><td>macro avg</td><td>0.97</td><td>0.53</td><td>0.53</td><td>269</td></tr>
<tr><td>weighted avg</td><td>0.94</td><td>0.93</td><td>0.90</td><td>269</td></tr>
</table>
</td></tr>
<tr><td></td><td>nu=0.1,
gamma=0.1,
seuil = 0.2</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.93</td><td>1.00</td><td>0.97</td><td>250</td></tr>
<tr><td>fraudeur</td><td>1.00</td><td>0.05</td><td>0.10</td><td>19</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.93</td><td>269</td></tr>
<tr><td>macro avg</td><td>0.97</td><td>0.53</td><td>0.53</td><td>269</td></tr>
<tr><td>weighted avg</td><td>0.94</td><td>0.93</td><td>0.90</td><td>269</td></tr>
</table>
</td></tr>
<tr><td colspan="4">Suite sur la page suivante</td></tr>
</table>

Tableau 4.9 – Suite de la page précédente

<table>
<tr><th>méthodes</th><th>Cluster</th><th>Hyperparamètre</th><th>Résultats</th></tr>
<tr><td></td><td></td><td>nu=0.2,
gamma=0.1,
seuil sup = 1.1,
seuil inf=0.9</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.93</td><td>1.00</td><td>0.97</td><td>250</td></tr>
<tr><td>fraudeur</td><td>1.00</td><td>0.05</td><td>0.10</td><td>19</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.93</td><td>269</td></tr>
<tr><td>macro avg</td><td>0.97</td><td>0.53</td><td>0.53</td><td>269</td></tr>
<tr><td>weighted avg</td><td>0.94</td><td>0.93</td><td>0.90</td><td>269</td></tr>
</table>
</td></tr>
<tr><td rowspan="2">IForest</td><td rowspan="2">Cluster 1 : Les clients qui ont des nuls brusques</td><td>contamination=0.2
n_estimators=3
random_state=42
seuil = 0.01</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.90</td><td>0.58</td><td>0.70</td><td>894</td></tr>
<tr><td>fraudeur</td><td>0.10</td><td>0.40</td><td>0.15</td><td>99</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.56</td><td>993</td></tr>
<tr><td>macro avg</td><td>0.50</td><td>0.49</td><td>0.43</td><td>993</td></tr>
<tr><td>weighted avg</td><td>0.82</td><td>0.56</td><td>0.65</td><td>993</td></tr>
</table>
</td></tr>
<tr><td>contamination=0.1
n_estimators=3
random_state=42
seuil = 0.01</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.90</td><td>0.86</td><td>0.88</td><td>894</td></tr>
<tr><td>fraudeur</td><td>0.09</td><td>0.12</td><td>0.10</td><td>99</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.78</td><td>993</td></tr>
<tr><td>macro avg</td><td>0.49</td><td>0.49</td><td>0.49</td><td>993</td></tr>
<tr><td>weighted avg</td><td>0.82</td><td>0.78</td><td>0.80</td><td>993</td></tr>
</table>
</td></tr>
</table>

Suite sur la page suivante

Tableau 4.9 – Suite de la page précédente

<table>
<tr><th>méthodes</th><th>Cluster</th><th>Hyperparamètre</th><th>Résultats</th></tr>
<tr><td></td><td></td><td>contamination=0.2
n_estimators=3
random_state=42
seuil_sup = 1.1
seuil_inf=0.9</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.90</td><td>0.83</td><td>0.86</td><td>894</td></tr>
<tr><td>fraudeur</td><td>0.09</td><td>0.15</td><td>0.11</td><td>99</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.76</td><td>993</td></tr>
<tr><td>macro avg</td><td>0.49</td><td>0.49</td><td>0.49</td><td>993</td></tr>
<tr><td>weighted avg</td><td>0.82</td><td>0.76</td><td>0.79</td><td>993</td></tr>
</table>
</td></tr>
<tr><td></td><td>Cluster 2 : Les clients avec une certaine station-narité ou une légère augmen-tation</td><td>contamination=0.2
n_estimators=3
random_state=42
seuil = 0.01</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.89</td><td>0.72</td><td>0.80</td><td>359</td></tr>
<tr><td>fraudeur</td><td>0.09</td><td>0.23</td><td>0.13</td><td>43</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.67</td><td>402</td></tr>
<tr><td>macro avg</td><td>0.49</td><td>0.48</td><td>0.46</td><td>402</td></tr>
<tr><td>weighted avg</td><td>0.80</td><td>0.67</td><td>0.73</td><td>402</td></tr>
</table>
</td></tr>
</table>

Suite sur la page suivante

Tableau 4.9 – Suite de la page précédente

<table>
<tr><th>méthodes</th><th>Cluster</th><th>Hyperparamètre</th><th>Résultats</th></tr>
<tr><td rowspan="3"></td><td rowspan="2"></td><td>contamination=0.1
n_estimators=3
random_state=42
seuil = 0.01</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.89</td><td>0.82</td><td>0.85</td><td>359</td></tr>
<tr><td>fraudeur</td><td>0.09</td><td>0.14</td><td>0.11</td><td>43</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.75</td><td>402</td></tr>
<tr><td>macro avg</td><td>0.49</td><td>0.48</td><td>0.48</td><td>402</td></tr>
<tr><td>weighted avg</td><td>0.80</td><td>0.75</td><td>0.77</td><td>402</td></tr>
</table>
</td></tr>
<tr><td>contamination=0.2
n_estimators=3
random_state=42
seuil_sup = 1.1
seuil_inf=0.9</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.90</td><td>0.99</td><td>0.94</td><td>359</td></tr>
<tr><td>fraudeur</td><td>0.40</td><td>0.05</td><td>0.08</td><td>43</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.89</td><td>402</td></tr>
<tr><td>macro avg</td><td>0.65</td><td>0.52</td><td>0.51</td><td>402</td></tr>
<tr><td>weighted avg</td><td>0.84</td><td>0.89</td><td>0.85</td><td>402</td></tr>
</table>
</td></tr>
<tr><td>Cluster 3 : Les clients qui ont une tendance baissière</td><td>contamination=0.2
n_estimators=3
random_state=42
seuil = 0.01</td><td>
<table>
<tr><th></th><th>precesion</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>non fraudeur</td><td>0.93</td><td>0.71</td><td>0.81</td><td>250</td></tr>
<tr><td>fraudeur</td><td>0.08</td><td>0.32</td><td>0.12</td><td>19</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.68</td><td>269</td></tr>
<tr><td>macro avg</td><td>0.50</td><td>0.51</td><td>0.47</td><td>269</td></tr>
<tr><td>weighted avg</td><td>0.87</td><td>0.68</td><td>0.76</td><td>269</td></tr>
</table>
</td></tr>
<tr><td colspan="4">Suite sur la page suivante</td></tr>
</table>

Tableau 4.9 – Suite de la page précédente

méthodes	Cluster	Hyperparamètre	Résultats
		contamination=0.1 n_estimators=3 random_state=42 seuil = 0.01	precesion / recall / f1-score / support non fraudeur 0.92 0.86 0.89 250 fraudeur 0.03 0.05 0.04 19 accuracy 0.01 269 macro avg 0.48 0.46 0.46 269 weighted avg 0.85 0.81 0.83 269
		contamination=0.2 n_estimators=3 random_state=42 seuil_sup = 1.1 seuil_inf=0.9	precesion / recall / f1-score / support non fraudeur 0.93 1.00 0.97 250 fraudeur 1.00 0.05 0.10 19 accuracy 0.93 269 macro avg 0.97 0.53 0.53 269 weighted avg 0.94 0.93 0.90 269

TABLEAU 4.9 – Comparaison des Performances des Méthodes avec Différents Hyperparamètres

Après avoir comparé les résultats obtenus en utilisant de différents hypers paramètres et algorithmes pour la détection des anomalies dans chaque cluster, nous avons formulé un choix final pour optimiser la précision de notre modèle.

Dans le Cluster 1, regroupant les clients présentant des nuls brusques, l'algorithme Isolation Forest s'est avéré performant avec une contamination de 0.2 et 3 estimateurs. L'utilisation d'un seuil de 0.01 pour déterminer les anomalies a permis d'obtenir une précision de 10%, un rappel de 40% et un score F1 de 15%, avec une précision globale (accuracy) de 56% avec une courbe ROC (AUC=0.48).

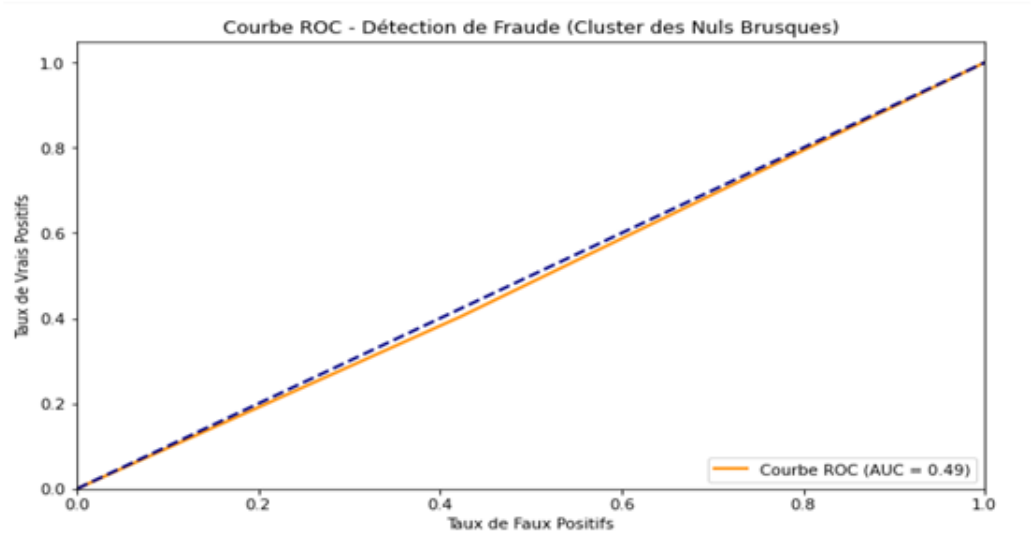

FIGURE 4.4 – Courbe ROC du cluster des clients nuls brusques

Pour le Cluster 2, englobant les clients avec une certaine stationnarité ou une légère augmentation, l'algorithme Isolation Forest a été préféré avec une contamination de 0.2 et 3 estimateurs. Les seuils de 1.1 pour les valeurs supérieures et de 0.9 pour les valeurs inférieures par rapport à la consommation moyenne saisonnière ont conduit à une précision de 40% , un rappel de 5% et un score F1 de 8%, avec une précision globale (accuracy) de 89%s avec une courbe ROC (AUC=0.52).

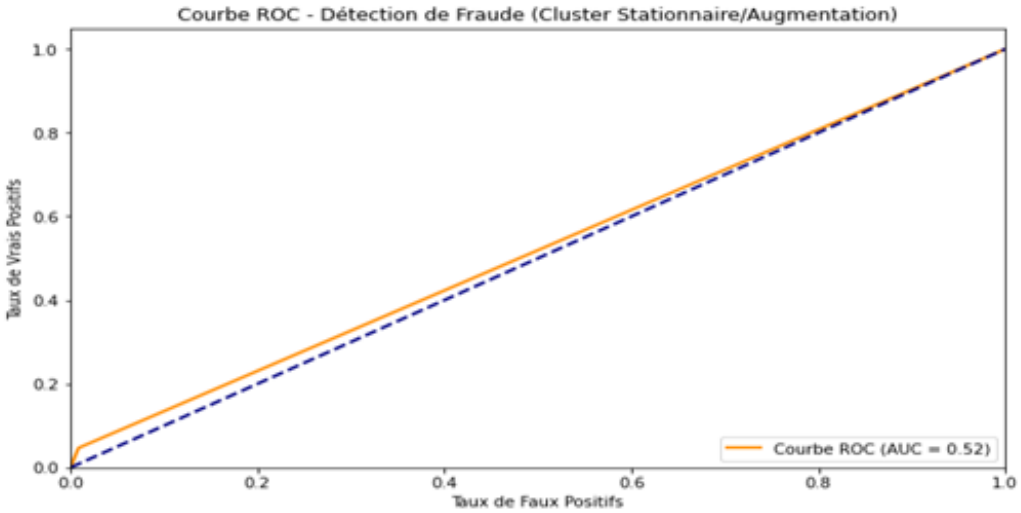

FIGURE 4.5 – Courbe ROC du cluster stationnaire ou augementation

En ce qui concerne le Cluster 3, composé des clients présentant une tendance baissière, l'algorithme Isolation Forest a été retenu avec une contamination de 0.2 et 3 estimateurs. L'application d'un seuil de 0.01 pour la détection des anomalies a généré une précision de 8% , un rappel de 32% et un score F1 de 12%, avec une précision globale (accuracy) de 68% avec une courbe ROC(AUC=0.51).

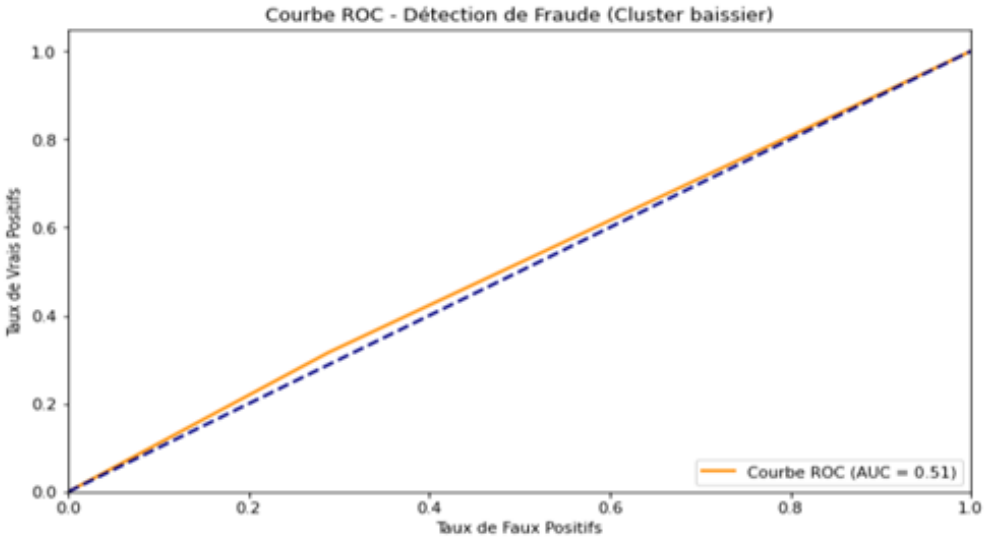

FIGURE 4.6 – Courbe ROC du cluster baissier

En prenant en compte ces résultats et les spécificités de chaque cluster, l'algorithme Isolation Forest nous donne de meilleurs résultats d'implémentation et d'évaluation pour l'ensemble des clusters .

4.6 Conclusion

En conclusion, ce chapitre a détaillé le processus complet de développement de notre système de détection de fraudes dans la consommation d'électricité. Nous avons examiné en profondeur les différentes étapes, de la préparation des données à l'application d'algorithmes de machine learning pour identifier les comportements anormaux. Après avoir comparé les performances de l'Isolation Forest et du OneClassSVM pour chaque

cluster, nous avons constaté que l'Isolation Forest offre globalement des résultats plus prometteurs en termes de détection de fraudes.

Conclusion Générale

Ce travail a été réalisé dans le cadre de notre projet de fin de cycle Master en informatique option intelligence artificielle. Il représente une exploration approfondie dans le domaine de la détection de la fraude électrique en utilisant des techniques d'apprentissage automatique.

La détection de la fraude électrique est un enjeu fondamental pour les services publiques, visant à identifier et à contrôler les comportements frauduleux. Notre travail est de résoudre ce défi en appliquant les méthodes du Machine Learning pour détecter la consommation d'électricité anormale des clients. Deux classifieurs majeurs, le modèle One-Class SVM et l'Isolation Forest, ont été évalués et améliorer dans le but de développer une solution robuste.

Nos résultats démontrent que l'approche basée sur l'Isolation Forest se distingue en termes de performances dans la détection des fraudes électriques. Ce modèle s'est révélé capable d'identifier avec une grande précision les clients suspects, réduisant ainsi les risques de pertes pour les entreprises de services publics et contribuant à garantir la légalité dans le secteur de l'électricité.

Cependant, il est essentiel de noter que notre recherche présente certaines limitations. La précision de nos modèles dépend en grande partie de la qualité des données et des caractéristiques extraites. De plus, la détection de la fraude électrique est un domaine en

constante évolution, nécessitant des améliorations continues pour contrôler les tactiques de plus en plus sophistiquées des fraudeurs.

Même si nous avons identifié l'Isolation Forest comme le modèle le plus performant dans notre étude, il est important de noter que ses résultats actuels ne sont pas entièrement satisfaisants. Ces résultats soulignent clairement la nécessité d'investir davantage dans la recherche et le développement futurs pour améliorer ce modèle.

En conclusion, ce mémoire met en évidence l'importance cruciale de l'apprentissage automatique dans la lutte contre la fraude électrique. Notre recherche démontre que des modèles tels que l'Isolation Forest peuvent devenir des outils efficaces pour les entreprises de services publics et les fournisseurs d'électricité. Nous espérons que ce travail contribuera à l'amélioration continue de la détection de la fraude électrique, protégeant ainsi les ressources énergétiques pour les générations à venir.

Bibliographie

[1] Quelle est la différence entre l'ia et le machine learning ? https://aws.amazon.com/fr/compare/the-difference-between-artificial-intelligence-and-machine-learning/, Consulté 28 avril 2023.

[2] A next-generation notebook interface. https://jupyter.org, miseenligne2022, Consulté le 11 aout 2023.

[3] Datascientest. numpy : la bibliothèque python la plus utilisée en data science. https://datascien-test.com/numpy, Consulté le 14 aout 2023.

[4] Tuto python scikit-learn : Knn (k-nearest neighbors). https://www.cours-gratuit.com/, Consulté le 15 mai 2023.

[5] K-means : Definition avantages / inconvénients. https://brightcape.co/k-means/, Consulté le 19 mai 2023.

[6] Généralités sur les séries chronologiques. https://www.i3s.unice.fr, Consulté le 21 Mars 2023.

[7] Techtarget. collecte de données. https://www.lemagit.fr/definition/Collecte-de-donnees::text=La%20collecte%20des%20donn%C3%A9es%20permet,et%20les%20tendances20%C3%A0%20venir, Consulté le 25 juillet 2023.

[8] Quels sont les algorithmes de deep learning ? https://mobiskill.fr, Consulté le 3 avril 2023.

[9] LABIAD ALI. Sélection des mots clés basée sur la classification et l'extraction des rÈgles d'association. 2017.

[10] Chloé-Agathe Azencott. Introduction au machine learning. https://www.dunod.com/sciences-techniques/introduction-au-machine-learning-0/. Consulté en avril 2023.

[11] Milena S Golshan Xiaowei Xu Bernadette M Randles, Irene V Pasquetto. Using the jupyter notebook as a tool for open science : An empirical study. in 2017 acm/ieee joint conference on digital libraries (jcdl). 2017.

[12] Ekaba Bisong. Introduction to scikit-learn. in building machine learning and deep learning models on google cloud platform. 2022.

[13] L Breiman. Random forests. machine learning. 2001.

[14] Anthony Gachagan Chigozie Enyinna Nwankpa, Winifred Ijomah. Activation functions : Comparison of trends in practice and research for deep learning. 2018.

[15] cours de Mr BOUCHEBBAH Fatah. Autoencodeurs. 2022-2023.

[16] cours de Mr BOUCHEBBAH Fatah. Réseaux de neurones récurrents. 2022-2023.

[17] MIT Critical Data. Secondary analysis of electronic health records. 2016.

[18] Bernat Coma Puig et al. Détection de fraude dans la consommation d'énergie : Une approche supervisée, 2016.

[19] Joaquim L. Viegas et al. Electricity fraud detection using committee semi-supervised learning. 2013.

[20] José F. Torres et al. évaluation des classificateurs pour l'identification des pertes non techniques dans les systèmes d'alimentation électrique. *international de l'alimentation électrique et des systèmes énergétiques*, 2000.

[21] Raphael MR Barros et al. Évaluation des classificateurs pour l'identification des pertes non techniques dans les systèmes d'alimentation électrique. *international de l'alimentation électrique et des systèmes énergétiques*, 2021.

[22] TOLEDO Fabio. Guide international du comptage intelligent. *Lavoisier*, 2012.

[23] Sébastien Gavois. des nano-neurones pour « repenser l'architecture interne de l'électro- nique».les réseaux de neurones artificiels. `https://www.nextinpact.com/article/27283/105231-ia-nano-neurones-pour-repenser-larchitecture-interne-lelectroniqu` Consulté 18 juillet 2023.

[24] M. F. A. Hady and F. Schwenker. "co-training by committee : A new semi-supervised learning framework. *Proceedings - IEEE International Conference on Data Mining Workshops*, 2008.

[25] Miin-Shen Yang Kristina P. Sinaga. Unsupervised k-means clustering algorithm. 20 April 2020.

[26] Jörg Sander Xiaowei Xu Martin Ester, Hans-Peter Kriegel. A density-based algorithm for discovering clusters in large spatial databases with noise. 2002.

[27] George M. Messinis. Classification non supervisée pour perte non technique détection. 2021.

[28] Zhi-Hua Zhou Min-Ling Zhang. A k-nearest neighbor based algorithm for multi-label classification. 20 April 2005.

[29] Pr. Fabien Moutarde. Arbres de décision et forêts aléatoires. 2017.

[30] ABDEL-NASSIR MAHAMAT NASSOUR. Algorithmes de structuration et de collecte de données électriques, application à la détection du vol d'électricité sur le réseau de la sne. 2021.

[31] Szilard Pafka. 'benchmarking random forest implementations'. https ://www.r-bloggers.com/2015/05/benchmarking-random-forest-implementations/, Mise en ligne Mai 19 2015, consulté le 30 Mai 2022.

[32] Oleksandr I. Provotar. Unsupervised anomaly detection in time series using lstm-based autoencoders. 2019.

[33] Ricco Rakotomalala. Gradient boosting.inconvénients (16p),université lumière lyon 2. 2016.

[34] Jalberth F. Araujob Raphaël MR Barrosun, Edson G. da Costab. Évaluation des classificateurs pour l'identification des pertes non techniques dans les systèmes d'alimentation électrique. 2021.

[35] Guido Van Rossum and Fred L Drake. An introduction to python. network theory ltd. bristol. 2003.

[36] Thorsten Papenbrock Sebastian Schmidl, Sebastian Schmidl. Anomaly detection in time series : A comprehensive evaluation. 2022.

[37] Pr.patruce wira. Etat de l'art du machine learning. 13 février 2018.

[38] Aaron Courvil Yoshua Bengio, Ian Goodfellow. Deep learning. October 03, 2015.

Résumé

Dans notre mémoire, nous examinons l'importance de la détection de la fraude électrique et explorons les progrès réalisés dans le domaine de la sécurité des réseaux électriques. Malgré les avancées dans la distribution de l'électricité et les réseaux informatiques, la fraude électrique demeure un défi persistant pour les entreprises de services publics. Nous présentons l'intelligence artificielle, en particulier l'apprentissage automatique, comme une solution pour la détection en temps réel de la consommation d'énergie non facturée.

Dans notre travail, nous proposons une approche basée sur l'apprentissage automatique pour la détection de la fraude électrique. Nous évaluons les modèles One-Class SVM et Isolation Forest, en apportant des mises à jour pour renforcer leur robustesse. Nos résultats montrent que le modèle Isolation Forest se distingue par ses performances dans la détection des fraudes électriques, réduisant ainsi le risque de pertes pour les fournisseurs de services publics.

Cependant, nous reconnaissons certaines limites dans notre étude, notamment la qualité des données et la nécessité d'améliorations continues pour contrer les tactiques sophistiquées des fraudeurs. Nous soulignons l'importance des investissements futurs dans la recherche pour affiner les modèles existants.

En conclusion, notre mémoire met en lumière le rôle essentiel de l'apprentissage automatique dans la lutte contre la fraude électrique. Nous aspirons à contribuer à l'amélioration continue de la détection de la fraude électrique, préservant ainsi les ressources énergétiques pour les générations futures.

Abstract

In our thesis, we examine the significance of electricity fraud detection and explore advancements in the field of electrical network security. Despite progress in electricity distribution and information technology networks, electricity fraud remains a persistent challenge for utility companies. We present artificial intelligence, particularly machine learning, as a solution for real-time detection of unbilled energy.

In our work, we present a machine learning-based approach for electricity fraud detection. We evaluate the One-Class SVM and Isolation Forest models, with updates to enhance their robustness. Our results indicate that the Isolation Forest model excels in detecting electrical fraud, thereby reducing the risk of losses for utility providers.

Nevertheless, we acknowledge certain limitations in our research, including data quality and the ongoing need for improvements to combat the sophisticated tactics employed by fraudsters. We emphasize the importance of investing in future research to refine existing models.

In conclusion, our thesis underscores the pivotal role of machine learning in combating electricity fraud. We aspire to contribute to the continual enhancement of electricity fraud detection, safeguarding energy resources for future generations.