RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ ABDERRAHMANE MIRA DE BÉJAÏA

FACULTÉ DES SCIENCES EXACTES
DÉPARTEMENT D'INFORMATIQUE

MEMOIRE

EN VUE DE L'OBTENTION DU DIPLOME DE MASTER

DOMAINE : MATHEMATIQUE ET INFORMATIQUE **FILIÈRE :** INFORMATIQUE

SPÉCIALITÉ : INTELLIGENCE ARTIFICIELLE

Thème

APPRENTISSAGE MACHINE POUR LA PRÉDICTION DE LA CONSOMMATION DU GAZ NATUREL À BEJAIA

Présenté par :

M. BOUZERA Rafik & Mlle. AKSOUH Anais

Soutenu devant le jury composé de :

Président	M. Nabil DJEBARI	M.C.B	U. A/Mira de Béjaïa
Encadrante	Mme. Samiha AIT TALEB	M.C.B	ESTIN de Béjaïa
Co-Encadrant	M. Abderrazak SEBAA	M.C.A	ESTIN de Béjaïa
Examinateur	M. Zoubeyr FARAH	M.C.A	U. A/Mira de Béjaïa
Invité	M. Dalil HADJOUT	Ingénieur	SONALGAZ de Béjaïa

Promotion 2022-2023

*** Remerciements ***

Nous tenons tout d'abord à exprimer notre gratitude et nos remerciements au Dieu tout puissant, pour nous avoir accordé la santé, la sagesse et la persévérance nécessaires pour accomplir ce travail.

Nous souhaitons exprimer notre profonde reconnaissance à nos encadrants, Mme Samiha Ait Taleb et M. Abderrazak Sebaa, pour leur encadrement attentif, leurs précieux conseils et leur soutien constant tout au long de ce projet. Leur expertise et leur disponibilité ont grandement contribué à la réalisation de ce travail.

Nous tenons également à remercier chaleureusement M. Dalil Hadjout, notre promoteur durant le stage, pour sa confiance, son accompagnement et ses orientations judicieuses. Son expertise et son suivi ont été d'une grande valeur pour la réussite de notre projet.

Nous adressons nos sincères remerciements à nos familles, nos amis et tous ceux qui nous ont soutenus tout au long de cette expérience. Leur présence, leurs encouragements et leur soutien moral ont été une source de motivation essentielle.

Enfin, nous souhaitons exprimer notre reconnaissance envers tous les enseignants, les membres du jury et toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail. Leurs conseils, leurs remarques constructives et leur intérêt pour notre projet ont été d'une grande importance.

Dédicaces

À nos trés chers parents, qui ont toujours été
un soutien inconditionnel
tout au long de notre parcours.

À nos amis, qui nous ont accompagnés
dans les moments de doute
et de joie.

À tous ceux qui ont contribué
à notre apprentissage
et à notre réussite.

Nous dédions ce travail.

Anais & Rafik

Table des matières

Table des figures

Liste des tableaux

LISTE DES SIGLES ET ACRONYMES

ML	*Machine learning*
AS	*Apprentissage supervisé*
ANS	*Apprentissage Non supervisé*
RNN	*Reccurent Network Neuronal*
SVM	*Support Vector Machine*
RNA	*Réseaux de neurones Artificiels*
SVR	*la régression vectorielle de support*
DNN	*Réseaux de neurones profond*
MLR	*Reccurent Network Neuronal*
ANN	*Artificial Neuronal Network*
CNN	*Réseaux de neurones a convolution*
AG	*Algorithme génétique*
MLP	*Perceptron multicouche*
LR	*Linear Regression*
RF	*Random Forest*
GNV	*Gaz Naturel Véhicules*
GPU	*Graphics Processing Unit*
TPU	*Tensor Processing Unit*

SES *Simple Exponential Smoothing*

DES *Double Exponential Smoothing*

TES *Triple Exponential Smoothing*

IDE *integrated development environment*

HP *Haute Pression*

MP *Moyenne pression*

SARIMA *Seasonal Autoregressive Integrated Moving Average*

LSTM *Long short-Term Memory*

ADF *Augmnte Dickey-Fuller*

API *Application Programming Interface*

INTRODUCTION GÉNÉRALE

L'énergie gazière est une ressource précieuse et d'une importance cruciale pour l'économie de l'Algérie et du monde. En tant que pays faisant partie des principaux exportateurs de gaz, il est essentiel d'améliorer les méthodes de prévision afin de mieux gérer cette ressource.

Cependant, de nombreux experts nous alertent sur le risque de perdre notre position d'exportateur en raison de la possibilité que la demande locale dépasse la part réservée à l'exportation. Une telle situation aurait des conséquences économiques potentiellement désastreuses.

Face à ces défis, comment pouvons-nous améliorer les méthodes de prévision de la consommation future de l'énergie gazière à Béjaia ? Quelles seraient les conséquences économiques si la demande locale surpassait la part réservée à l'exportation ? Comment pouvons-nous utiliser l'intelligence artificielle pour prédire avec précision la consommation future de gaz naturel ? Quelles mesures proactives pouvons-nous prendre pour optimiser l'utilisation de cette ressource et préserver notre position d'exportateur ? Comment pouvons-nous minimiser les impacts économiques négatifs d'une éventuelle pénurie de gaz à Béjaia ?

Dans le cadre de cette étude, nous proposons une solution moderne basée sur l'intelligence artificielle pour prédire avec précision la consommation future de l'énergie gazière à Béjaïa. Nous commencerons par analyser la consommation de gaz naturel des clients économiques haute pression de la wilaya de Béjaïa en exploitant les données détaillées de ces clients. Cette approche nous permettra ensuite d'estimer la consommation globale du gaz naturel.

Le premier chapitre de notre étude sera consacrée à l'examen des aspects généraux du gaz naturel, y compris son origine, son histoire et ses caractéristiques techniques. Comprendre ces éléments est essentiel pour appréhender l'importance et les applications du gaz naturel dans différents secteurs.

De plus, nous nous concentrerons sur la SONELGAZ, une entreprise majeure dans le domaine de la distribution de gaz naturel. Nous retracerons son historique, étudierons son organisation, sa structure et ses principales activités. Ensuite, nous examinerons la distribution du gaz naturel par la SONELGAZ en fonction des types de clients. Nous présenterons également l'organisme d'accueil, la SONELGAZ Distribution CD Béjaïa, en détaillant son organisation.

Dans le deuxième chapitre, nous explorerons les méthodes traditionnelles et les méthodes d'apprentissage automatique utilisées pour la prédiction de la consommation de gaz naturel. Nous discuterons des séries temporelles, de leurs composantes et de leurs limitations, ainsi que des méthodes de prédiction basées sur l'apprentissage automatique, telles que l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. Nous mettrons également l'accent sur l'évo-

lution de l'apprentissage automatique vers l'apprentissage en profondeur, en examinant les réseaux de neurones artificiels, les réseaux de neurones à convolution, les réseaux de neurones récurrents et les mémoires à long terme à court terme.

Le troisième chapitre sera consacrée à l'examen de l'état de l'art dans le domaine de la prédiction de la consommation du gaz naturel, en présentant des travaux connexes et en fournissant un tableau comparatif des solutions existantes. Nous discuterons des avantages et des limites de ces solutions.

Ensuite, dans le quatrième chapitre nous aborderons la modélisation et l'evaluation des modèles de prédiction de la consommation du gaz naturel.Nous décrirons l'environnement de développement utilisé, y compris le matériel et les logiciels, ainsi que les étapes de modélisation pour un seul client ensuite nous avons analysé quelques clients individuelle avec une consommation midiocre et avons modilisé la consommation pour le reste des clients. Nous avons comparé les résultats obtenus et avons met en place notre approche.

Enfin, dans le cinquième chapitre, constitue une étape importante de notre recherche, où nous détaillons la mise en œuvre de notre approche hybride et fournissons une analyse approfondie des résultats obtenus. Ces informations serviront de base solide pour les discussions et les conclusions finales de notre étude sur la prédiction de la consommation du gaz naturel en utilisant notre approche.

En conclusion, cette étude vise à améliorer les méthodes de prévision de la consommation future de l'énergie gazière à Béjaïa grâce à l'utilisation de l'intelligence artificielle. En anticipant les fluctuations de la consommation de gaz naturel, nous pourrons prendre des mesures proactives pour optimiser son utilisation et préserver notre position d'exportateur. De plus, nous chercherons à minimiser les impacts économiques négatifs d'une éventuelle pénurie du gaz à Béjaïa.

CHAPITRE

1

GÉNÉRALITÉS SUR LE GAZ NATUREL ET PRÉSENTATION DE LA SONELGAZ

1.1 Introduction

Le gaz naturel est une source d'énergie largement utilisée dans le monde entier pour ses nombreux avantages en termes de disponibilité, d'efficacité et de respect de l'environnement. Avant de plonger dans la présentation de la Sonelgaz, il est important de comprendre les bases du gaz naturel. Dans ce chapitre, nous explorerons les généralités sur le gaz naturel notamment ses origines et son histoire, sa description et ses caractéristiques techniques, ainsi que son utilisation dans différents secteurs.

Nous commencerons par examiner les origines et l'histoire du gaz naturel, remontant à des millions d'années avec la formation de matières organiques dans des formations géologiques souterraines. Nous aborderons ensuite la description et les caractéristiques techniques du gaz naturel, y compris sa composition chimique, ses propriétés physiques et ses méthodes d'extraction et de traitement.

Ensuite, nous nous intéresserons à l'utilisation du gaz naturel dans différents secteurs, tels que l'industrie, les transports, la production d'électricité et le chauffage. Nous explorerons les avantages et les applications du gaz naturel dans ces domaines.

Ensuite, nous nous concentrerons sur la présentation de la Sonelgaz, une entreprise algérienne en charge de la production, la distribution et la commercialisation d'électricité et de l'achat, le transport, la distribution et la commercialisation de gaz naturel. Nous aborderons son historique, son organisation, sa structure, ainsi que ses principales activités. Nous examinerons également la distribution de gaz naturel par la Sonelgaz en fonction des types de ses clients, notamment son organisme d'accueil Sonelgaz Distribution CD BEJAIA, en analysant son organisation et sa structure.

Enfin, nous discuterons des projets d'avenir de la Sonelgaz et des perspectives pour cette entreprise dans le contexte de l'évolution du secteur de l'énergie et des défis auxquels elle est confrontée. Ce chapitre fournira une vue d'ensemble complète sur les généralités sur le gaz naturel et la présentation de la Sonelgaz, jetant ainsi les bases pour une compréhension approfondie des sujets abordés dans ce mémoire.

1.2 Généralités sur le Gaz Naturel

1.2.1 Origine et Histoire

Il y a quelques milliers d'années, le gaz naturel a été découvert au Moyen-Orient, où l'apparition de flammes soudaines et brûlantes était assimilée à des sources ardentes. En Perse, en Grèce et en Inde,

des temples ont été construits autour de ces feux pour des pratiques religieuses, mais l'importance de cette découverte n'a pas été immédiatement reconnue. Ce n'est qu'environ en 900 av. J.-C. que la Chine a compris l'importance de ce produit et a foré le premier puits de gaz naturel vers 211 av. J.-C.

En Europe, il a fallu attendre jusqu'en 1659 pour que la Grande-Bretagne découvre et commence à commercialiser le gaz naturel à partir de 1790. En 1821, à Fredonia (États-Unis), les habitants ont découvert le gaz naturel dans une crique en observant des bulles de gaz remontant à la surface. William Hart est considéré comme le "père du gaz naturel", car il a creusé le premier puits de gaz naturel en Amérique du Nord.

Au XIXe siècle (19ème), le gaz naturel a été principalement utilisé comme source d'éclairage. Sa consommation était limitée en raison du manque d'infrastructures de transport qui rendait difficile l'acheminement de grandes quantités de gaz naturel sur de longues distances. Ce n'est qu'en 1890, avec l'invention des joints étanches aux fuites, que des progrès importants ont été réalisés. Cependant, les techniques existantes ne permettaient pas de transporter le gaz naturel sur des distances de plus de 160 kilomètres, et une grande quantité de gaz naturel a été gaspillée pendant des années en étant brûlée sur place. Le transport du gaz naturel sur de longues distances s'est généralisé dans les années 1920 grâce aux progrès technologiques dans la construction de gazoducs. Après la Seconde Guerre mondiale, la consommation de gaz naturel a rapidement augmenté en raison du développement des réseaux de pipelines et des systèmes de stockage [3].

1.2.2 Description et caractéristiques techniques

Le gaz naturel est principalement composé de méthane (CH4), un hydrocarbure léger, ainsi que d'autres hydrocarbures tels que l'éthane, le propane, les butanes et les pentanes. Il peut également contenir d'autres composés chimiques en petites quantités. Sa composition chimique varie en fonction de la source et du lieu d'extraction.

Sur le plan physique, le gaz naturel est incolore, inodore et insipide à l'état pur. Cependant, un additif chimique appelé mercaptan est souvent ajouté pour donner une odeur d'œuf pourri, ce qui permet de détecter facilement les fuites de gaz pour garantir la sécurité lors de son utilisation. Le gaz naturel est plus léger que l'air et se présente sous sa forme gazeuse à des températures supérieures à -161°C [31].

Le gaz naturel est considéré comme un combustible propre par rapport à d'autres combustibles fossiles en raison de ses faibles émissions de polluants lors de sa combustion. Il émet moins de dioxyde de soufre (SO2) car il contient peu de soufre. De plus, les émissions d'oxydes d'azote (NOx)

et de gaz à effet de serre, tels que le dioxyde de carbone (CO2), sont généralement inférieures à celles du pétrole et du charbon, ce qui en fait une option plus respectueuse de l'environnement.

En termes de caractéristiques techniques, le gaz naturel possède une température de combustion élevée, ce qui signifie qu'il brûle facilement et presque totalement, produisant de la chaleur et de l'énergie. Il a également une densité plus basse que celle de l'air, ce qui lui permet de se dissiper rapidement en cas de fuite, minimisant ainsi les risques pour la santé et l'environnement [39].

En ce qui concerne l'extraction et le traitement du gaz naturel, il existe plusieurs méthodes utilisées. L'extraction du gaz naturel se fait principalement par forage de puits dans les gisements de gaz. Il peut être extrait à partir de gisements de gaz associé, c'est-à-dire situés à proximité de gisements de pétrole brut, ou à partir de gisements de gaz non associé, c'est-à-dire des gisements de gaz pur.

Une fois extrait, le gaz naturel est généralement transporté par pipelines vers les installations de traitement. Le traitement du gaz naturel comprend généralement l'élimination des impuretés telles que l'eau, le dioxyde de carbone, le soufre et d'autres contaminants. Ce processus permet de purifier le gaz naturel avant qu'il ne soit distribué aux consommateurs.

1.2.3 Utilisations du gaz naturel

Le gaz naturel est utilisé dans de nombreux secteurs, notamment le marché résidentiel et tertiaire, le secteur industriel, la production d'électricité et les véhicules [46].

Marché résidentiel et tertiaire :

Le gaz naturel est largement utilisé pour le chauffage et la cuisson dans les maisons et les bâtiments commerciaux. Les installations au gaz naturel, comme les chaudières à condensation et les pompes à chaleur, sont de plus en plus performantes, ce qui permet de réaliser des économies d'énergie et de réduire les émissions de gaz à effet de serre. De plus, le gaz naturel offre une source d'énergie continue et fiable pour le chauffage et la cuisson, ce qui en fait une option populaire dans le secteur résidentiel et tertiaire.

Secteur industriel :

Le gaz naturel est largement utilisé dans l'industrie chimique, la pétrochimie et le raffinage. Il est utilisé comme matière première pour la production de divers produits chimiques et plastiques, tels que l'ammoniac, l'urée et le méthanol. Le gaz naturel est également utilisé comme source d'énergie pour les procédés industriels nécessitant une chaleur intense, comme la fusion de métaux, la cuisson du verre et la production d'électricité dans les centrales thermiques. Son utilisation dans l'industrie

permet de réduire les émissions de polluants et de gaz à effet de serre par rapport à d'autres sources d'énergie fossile.

Production d'électricité :

Le gaz naturel est de plus en plus utilisé dans la production d'électricité, car il offre plusieurs avantages. Il émet moins de dioxyde de carbone (CO2) et d'autres polluants atmosphériques que le charbon, ce qui en fait une option plus propre pour la production d'électricité. De plus, les coûts d'investissement et de fonctionnement des centrales à gaz naturel sont souvent inférieurs à ceux des centrales à charbon, ce qui permet un rendement supérieur et une rentabilité accrue. Le gaz naturel est également utilisé dans les centrales électriques pour la production d'électricité d'appoint en cas de pics de demande.

Véhicules :

Le gaz naturel est utilisé comme carburant pour les véhicules sous forme de Gaz Naturel Véhicules (GNV), offrant une alternative plus propre aux carburants traditionnels comme l'essence ou le diesel. Les véhicules au GNV émettent moins de polluants atmosphériques, tels que les oxydes d'azote (NOx) et les particules fines, comparés aux véhicules utilisant des carburants fossiles, ce qui contribue à améliorer la qualité de l'air. De plus, le gaz naturel est généralement moins cher que l'essence ou le diesel, ce qui peut permettre des économies de coûts pour les propriétaires de véhicules et contribuer à la réduction des émissions de gaz à effet de serre.

1.3 Présentation de la Sonelgaz

1.3.1 Historique

Sonelgaz a joué un rôle majeur dans le développement économique et social du pays en contribuant à la concrétisation de la politique énergétique nationale par la réalisation de nombreux programmes d'électrification rurale et de distribution publique de gaz. Grâce à ces effort, le taux de couverture en électricité a atteint 98%
pour 10 983 538 clients, et le taux de pénétration du gaz a atteint 60% pour 6 886 407 clients.

1.3.2 Organisation,structure et principales activités de la Sonalgaz

Le groupe Sonelgaz est composé de 14 sociétés filiales, gérées directement par la holding, ainsi que de 9 sociétés en participations avec des tiers. Parmi ces filiales, on retrouve notamment :

1. La Société de Production de l'Électricité (SPE) : Responsable de la production d'électricité en Algérie à partir de différentes sources d'énergie, y compris les centrales thermiques, hydrauliques et renouvelables.
2. Sharikat Kahraba wa Takat Moutadjadida (SKTM) : Chargée de la construction et de l'exploitation de centrales de production d'électricité en Algérie.
3. La Société de l'Ingénierie de l'Électricité et du Gaz (CEEG) : Spécialisée dans l'ingénierie et la réalisation de projets d'électricité et de gaz, ainsi que dans la formation et le développement des compétences techniques.
4. La Société de Gestion du Réseau de Transport de l'Électricité (GRTE) : Responsable de la gestion et de l'exploitation du réseau de transport d'électricité en Algérie, ainsi que du développement du réseau de transport.
5. La Société de Gestion du Réseau de Transport Gaz (GRTG) : En charge de la gestion et de l'exploitation du réseau de transport de gaz en Algérie, ainsi que du développement du réseau de transport gazier.
6. L'Opérateur Système Électrique (OS) : Responsable de la conduite du système de production et de transport de l'électricité en Algérie, assurant l'équilibre entre la production et la consommation d'électricité.
7. Sonelgaz-Distribution : Anciennement Société Algérienne de Distribution de l'Électricité et du Gaz (SADEG), créée en 2017 par fusion des sociétés SDC, SDA, SDE et SDO, elle est en charge de la distribution de l'électricité et du gaz aux clients en Algérie.
8. La Société des Grands Travaux d'Électricité et de Gaz (Kahragaz) : Spécialisée dans la réalisation de grands travaux d'infrastructures électriques et gazières en Algérie, y compris la construction de nouvelles centrales électriques et de gazoducs.
9. La Société de Réalisation des Infrastructures Énergétiques et Industrielles (Inerkib) : Responsable de la réalisation des infrastructures énergétiques et industrielles nécessaires au développement du secteur de l'électricité et du gaz en Algérie [35].

1.3.3 Distribution du gaz naturel par la Sonalgaz en fonction des types de clients :

La distribution du gaz naturel par la Sonelgaz en Algérie se fait selon différents niveaux de pression en fonction des besoins et des catégories de clients :

- **Basse pression :** C'est le niveau de pression le plus bas utilisé pour la distribution du gaz naturel. Il est généralement utilisé pour les clients résidentiels, les petites entreprises et les petits consommateurs. La pression du gaz à ce niveau est relativement faible et est généralement suffisante pour les besoins domestiques et les petits équipements de consommation.

- **Moyenne pression :** La moyenne pression est utilisée pour les clients industriels, les grandes entreprises et les gros consommateurs. Elle offre une pression de gaz plus élevée que la basse pression pour répondre aux besoins énergétiques plus importants des installations industrielles et commerciales.

- **Haute pression :** La haute pression est utilisée pour les clients industriels avec des besoins énergétiques très importants, tels que les grandes centrales électriques, les installations de production d'énergie et d'autres grands consommateurs industriels. La pression du gaz à ce niveau est la plus élevée et nécessite des équipements spécifiques pour la distribution et l'utilisation du gaz naturel à des niveaux de pression élevés.

1.3.4 Présentation de l'organisme d'accueil (Sonalgaz Distribbution CD BEJAIA)

La Direction de Distribution de Bejaia est rattachée à la société Algériennes de Distribution de l'Électricité et du Gaz de l'Est (SDE), dont le siège se trouve à Constantine. Cette dernière est composée d'une Direction à laquelle sont reliés directement :

- Le secrétariat.
- Les assistants du Directeur de Distribution.
- Le chargé des affaires juridiques.
- Le chargé de la communication.
- Le chargé de la sécurité.

En outre, la SDE est composée de 9 Divisions et 10 Agences Commerciales réparties dans différentes localités, notamment : BEJAIA CITE TOBAL, BEJAIA 4 CHEMAIN, EL KSEUR, AMIZOUR, SIDI AICH, SEDDOUK, AKBOU, TAZMALT, AOKAS, KHERRATA.

De plus, la SDE est également organisée en 5 Districts, à savoir : BEJAIA, AKBOU, SIDI AICH, AMIZOUR, KHERRATA.

1.3.5 Organisation de la Sonalgaz Distribution CD Bejaia

La structure organisationnel des différentes divisions de la Sonelgaz Distribution CD BEJAIA et présenté dans la figure 1.1

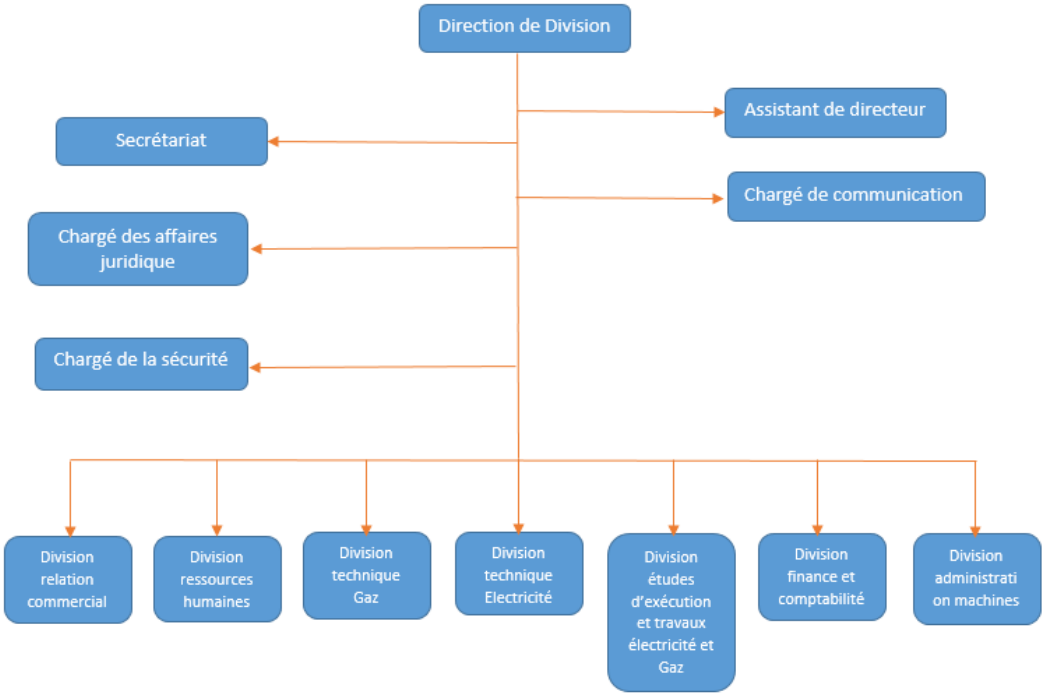


FIGURE 1.1 – Organigramme de la Sonelgaz Bejaia.

1.3.6 Projet d'avenir de la Sonalgaz

Les perspectives d'avenir de la Sonelgaz sont prometteuses, avec un plan de développement ambitieux pour les réseaux de distribution de l'électricité et du gaz. Ce plan intègre les programmes d'électrification et de distribution publique du gaz initiés par l'État, ainsi que les programmes propres de la Sonelgaz, les raccordements de nouvelles clientèles, les équipements de maintenance et d'exploitation, ainsi que les projets de modernisation de la gestion et de l'exploitation.

Sur la période 2021-2030, la Sonelgaz prévoit de développer un réseau d'électricité de 101 960 km de lignes et 38 864 postes pour alimenter 4,4 millions de clients supplémentaires. Pour le réseau de gaz, le plan prévoit de développer 56 792 km de réseau pour alimenter 4,3 millions de clients supplémentaires. Ces projets de développement permettront d'élargir l'accès à l'électricité et au gaz à de nouvelles clientèles, contribuant ainsi au développement économique et social du pays. La Sonelgaz vise également à moderniser sa gestion et son exploitation, en utilisant les technologies de pointe et en mettant en place des systèmes de gestion performants pour optimiser ses opérations et améliorer la qualité de service offerte à ses clients. Cela inclut la mise en place de solutions innovantes pour la ges-

tion des réseaux, la maintenance prédictive, la gestion des données et la digitalisation des processus [47].

1.4 Conclusion

Dans ce chapitre, nous avons dressé un aperçu complet du gaz naturel, en examinant son origine, son histoire, ses caractéristiques techniques et ses multiples utilisations dans divers secteurs. Nous avons également présenté la Sonelgaz, une entreprise spécialisée dans la production, la distribution et la commercialisation du gaz naturel en Algérie. Nous avons examiné son historique, son organisation, sa structure et ses activités, ainsi que son rôle dans la distribution du gaz naturel en fonction de ses différents clients. Nous avons également abordé l'organisation de la Sonelgaz Distribution CD BEJAIA, qui est l'organisme d'accueil de notre étude. Enfin, nous avons évoqué les projets d'avenir de la Sonelgaz. Ce chapitre nous a ainsi permis d'acquérir une compréhension approfondie du gaz naturel et de reconnaître l'importance de la Sonelgaz dans la distribution de cette ressource essentielle en Algérie.

CHAPITRE

2

MÉTHODES DE PRÉDICTION

2.1 Introduction

L'objectif de ce chapitre est de présenter les différentes méthodes de prédiction, en mettant l'accent sur les méthodes traditionnelles et les méthodes de Machine Learning.

La prédiction est une technique qui consiste à utiliser les données passées pour faire des prédictions sur des évènements futures. Elle est utilisée dans de nombreux domaines tels que la finance, la météorologie, la médecine, l'énergie etc.

Dans la première partie, nous allons passer en revue les méthodes traditionnelles de prédiction en expliquant leur fonctionnement, leurs principes et leurs applications. Nous verrons également leurs limitations.

Dans la deuxième partie, nous allons nous concentrer sur les méthodes de Machine Learning, qui ont connu une croissance exponentielle ces dernières années grâce aux progrès de l'informatique et à la disponibilité de données massives. Nous allons explorer les types d'apprentissage automatique, les modèles d'apprentissage automatique les plus couramment utilisés et les étapes de l'apprentissage automatique.

Enfin, nous allons passer en revue les métriques d'évaluation des modèles d'apprentissage automatique pour évaluer leur performance.

2.2 Méthodes traditionnelles de prédiction

2.2.1 Introduction générale aux méthodes traditionnelles

Les méthodes de prédiction traditionnelles, également connues sous le nom de méthodes statistiques, sont des techniques utilisées pour analyser et prédire des phénomènes basées sur des données historiques. Elles ont été développées à partir des principes de la théorie des probabilités et de l'inférence statistique (utilisées pour l'analyse des données historiques, déduire des relations entre les variables et prédire les valeurs futures de ces variables).

Ces méthodes sont largement utilisées avant l'évènement de l'apprentissage automatique et elles sont aussi utilisées dans de nombreux domaines tels que l'énergie, la météorologie, la science des données etc.

Les méthodes traditionnelles de prédiction, notamment la régression et les méthodes de lissage, peuvent être utilisées comme modèles de prédiction dans les séries temporelles. Ces méthodes sont fondées sur l'hypothèse que les données historiques contiennent des motifs ou des tendances qui

peuvent être utilisés pour prédire les valeurs futures. En utilisant des techniques statistiques, il est possible d'identifier ces motifs ou tendances et de les utiliser pour prédire les valeurs futures avec une certaine précision.

2.2.2 Séries temporelles

Les séries temporelles (ou encore série chronologique) sont des données qui sont collectées sur une période de temps Xt régulière. Elles sont utilisées pour analyser les tendances, les modèles saisonniers et les fluctuations au fil du temps. Les séries temporelles sont largement utilisées dans les domaines de l'économie, de la finance, de la météorologie, de la santé publique et de l'ingénierie, entre autres.

Une série temporelle peut être représentée sous format d'un graphe construit de la manière suivante :

- en abscisse le temps.
- en ordonnée la valeur de l'observation à chaque instant [10].

2.2.2.1 Composantes d'une Série chronologique

On considère qu'une série chronologique (Xt) est la résultante des différentes composantes fondamentales :

1. La tendance (ou trend) : représente l'évolution à long terme de la série étudiée. Elle traduit le comportement moyen de la série.
2. La composante saisonnière (ou saisonnalité) : correspond à un phénomène se répétant dans des intervalles de temps réguliers (périodiques). Elle est donc liée au rythme imposé par les saisons (toutes les 12 périodes pour des données mensuelles, toutes les 4 périodes pour des données trimestrielles).
3. La composante résiduelle (ou bruit ou résidu) : correspond à des fluctuations irrégulières, en général de faible intensité mais de nature aléatoire.

2.2.2.2 Types des séries temporelles

les series temporelles sont souvent modelisées par les modèles suivants :

1. **Modèle additif :** Un modèle additif d'une série temporelle est une approche qui considère que la série temporelle observée, notée X_t, peut être décomposée en la somme des trois composantes.

L'équation du modèle additif de la série temporelle est donnée par :

$$X_t = Z_t + S_t + \epsilon_t, \quad \text{pour } t = 1, \ldots, T \tag{2.1}$$

2. **Modèle Multiplicatif :** Le modèle multiplicatif d'une série temporelle est une approche qui considère que la série temporelle observée, notée X_t, peut être décomposée en tois composantes distinctes multipliées entre elles.

   L'équation du modèle multiplicatif de la série temporelle est donnée par :

$$X_t = T_t \times S_t \times \epsilon_t, \quad \text{pour } t = 1, \ldots, T \tag{2.2}$$

3. **Modèle Mixt :** Il s'agit des différentes combinaisons de modèles additifs et de modèles multiplicatifs, par exemple :

$$X_t = T_t \times S_t + \epsilon_t, \quad \text{pour } t = 1, \ldots, T \tag{2.3}$$

Differents modèles statistiques sont employs pour l'analyse et la prédiction des séries temporelles. Parmi ces modèles, on retrouve les approches traditionnelles telles que le méthode de la moyenne mobile, la régression, le lissage et la méthode SARIMA Ces méthodes sont utilisées pour modéliser la série temporelle en fonction de sa tendance, de sa saisonnalité et de ses composantes résiduelles.

2.2.3 Exemples, Principes et Fonctionnement des méthodes de prédiction traditionnelles

1. **La Moyenne Mobile :**

   La méthode de moyenne mobile est une technique statistique couramment utilisée pour analyser et prédire les tendances dans les séries chronologiques qui consiste à calculer la moyenne des valeurs observées sur une période de temps donnée. Cette période de temps est appelée la fenêtre de temps ou la fenêtre de lissage.

   Le principe de base de la méthode de moyenne mobile est d'estimer la tendance en calculant une moyenne mobile centrée sur chaque point dans le temps.

   Il existe plusieurs types de moyennes mobiles, notamment :

   - La moyenne mobile simple : qui calcule la moyenne des observations précédentes sur une

période donnée.

- La moyenne mobile pondérée : qui attribue des pondérations différentes aux observations précédentes en fonction de leur âge ou de leur importance relative.
- La moyenne mobile exponentielle : qui utilise une formule de lissage exponentiel pour estimer la tendance à court terme de la série temporelle.

Le fonctionnement de la méthode de la moyenne mobile est relativement simple. Tout d'abord, une période de temps est choisie pour laquelle la moyenne sera calculée. Ensuite, pour chaque période de temps suivante, la moyenne des observations précédentes sur la période choisie est calculée et utilisée pour prédire la valeur pour la période suivante [48].

2. **La regression :**

La régression est une technique statistique également utilisée comme modèle de prédiction, peut être utilisé pour modéliser à la fois les séries temporelles saisonnières et non saisonnières dans les séries temporelles. utilisé pour modéliser la relation entre une variable dépendante (Y) et une ou plusieurs variables indépendantes (X).

Il existe deux types principaux de régression : la régression linéaire et la régression non linéaire.

Le principe de la régression est de trouver une fonction qui peut prédire la valeur de la variable dépendante en fonction de la valeur ou des valeurs des variables indépendantes. Pour cela, on utilise les données historiques disponibles et on ajuste les coefficients de la fonction pour minimiser l'erreur de prédiction.

- **La régression linéaire :**

  est une méthode statistique utilisée pour modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes.

  L'objectif de la régression linéaire est de trouver la droite de régression qui décrit le mieux la relation entre les variables en estimant les valeurs des coefficients.

  La droite est définie par une équation de la forme :

$$y = mx + b \tag{2.4}$$

  Où y est la variable dépendante, x est la variable indépendante, m est la pente de la droite et b est l'ordonnée à l'origine.

  La régression linéaire tente de modéliser la relation entre deux variables en ajustant une équation linéaire aux données observées.

La forme la plus simple de l'équation de régression avec une variable dépendante et une variable indépendante est définie par la formule :

$$y = b0 + b1 * x \tag{2.5}$$

où y est le score de la variable dépendante estimé, b0 est la constante, b1 est le coefficient de régression et x est le score de la variable indépendante.

Le but de la régression linéaire est de trouver les valeurs des coefficients b0 et b1 qui minimisent la somme des carrés des résidus, c'est-à-dire la différence entre les valeurs observées et prédites de Y.

Cela peut être fait en utilisant la méthode des moindres carrés ordinaires (MCO).

La régression linéaire est largement utilisée dans divers domaines tels que l'économie, la finance et les sciences sociales pour analyser et prédire les tendances et les relations entre les variables.

- **La régression non linéaire :**

La régression non linéaire est utilisée lorsque la relation entre la variable dépendante et les variables indépendantes n'est pas linéaire. Elle peut prendre plusieurs formes fonctionnelles, telles que :

$$Y = b_0 + b_1 X + b_2 X^2 + \varepsilon \tag{2.6}$$

Où Y est la variable dépendante, x est la variable indépendante, b_0, b_1 et b_2 sont les coefficients de la fonction non linéaire, et ϵ est l'erreur résiduelle.

Le but de la régression non linéaire est de trouver les valeurs des coefficients b0, b1 et b2 qui minimisent la somme des carrés des résidus, tout comme la régression linéaire [22].

3. **Le lissage exponentiel :** La méthode de lissage exponentiel est une technique de modélisation de séries temporelles qui est largement utilisée pour prédire les tendances futures. Elle est basée sur l'hypothèse que la valeur actuelle de la série temporelle est influencée par les valeurs passées, mais avec une décroissance exponentielle de l'importance de ces valeurs à mesure qu'on s'éloigne dans le temps.

Le principe de base de la méthode de lissage exponentiel est de calculer une moyenne pondérée

des valeurs passées, en donnant plus de poids aux valeurs les plus récentes de la série temporelle. Les poids sont déterminés par le paramètre alpha, qui est une constante comprise entre 0 et 1, et qui détermine l'importance relative des observations passées et présentes.

Exemple de types de méthodes de lissage :

Il existe plusieurs variantes de la méthode de lissage exponentiel, qui diffèrent par la manière dont les poids sont calculés et par la façon dont la tendance et la saisonnalité sont prises en compte

- Lissage exponentiel simple (SES) :(Simple Exponential Smoothing) utilise un seul paramètre alpha pour calculer la moyenne pondérée des valeurs passées.

  elle est principalement utilisée pour modéliser les séries temporelles sans tendance ou saisonnalité significative.

- Lissage exponentiel double (DES) : (Double Exponential Smoothing) est une extension de la méthode de lissage exponentiel simple qui permet de prendre en compte à la fois la tendance et la saisonnalité de la série temporelle. Elle utilise deux paramètres alpha pour calculer la moyenne pondérée des valeurs passées, l'un pour la tendance et l'autre pour la saisonnalité.

  Cette méthode est plus adaptée aux séries temporelles avec une tendance, mais sans saisonnalité prononcée.

- Lissage exponentiel triple (TES) : (Triple Exponential Smoothing) également connu sous le nom de méthode de Holt-Winters, est une autre extension du lissage exponentiel simple qui prend en compte la tendance, la saisonnalité et les effets cycliques de la série temporelle. elle utilise trois paramètres alpha : un pour le lissage de la moyenne pondérée des valeurs passées, un autre pour le lissage de la tendance, et un troisième pour le lissage de la saisonnalité.

  Cette méthode est utilisée lorsque la série temporelle présente une tendance, une saisonnalité et des effets cycliques significatifs

Le fonctionnement de la méthode de lissage exponentiel consiste donc à ajuster les poids de chaque observation passée en fonction de sa distance à l'instant présent, de manière à donner plus d'importance aux observations les plus récentes et moins d'importance aux observations plus anciennes. La méthode calcule ensuite une moyenne pondérée de ces observations pour prédire la valeur future de la série temporelle [42].

4. De AR au SARIMA :

- **Modèle AR (Autoregressive) :** Le modèle AR est un modèle de séries temporelles qui prédit la valeur future en fonction des valeurs passées de la série. Il utilise une combinaison linéaire des observations précédentes (retards) pour estimer la valeur.
- **Modèle MA (Moving Average) :** Le modèle MA est un modèle de séries temporelles qui prédit la valeur future en fonction des erreurs précédentes de prédiction. Il utilise une combinaison linéaire des erreurs résiduelles précédentes pour estimer la valeur future. Le terme "moyenne mobile" fait référence au fait que le modèle utilise une moyenne mobile des erreurs passées.
- **Modèle ARMA (Autoregressive Moving Average) :** Le modèle ARMA est une combinaison du modèle AR et du modèle MA. Il utilise à la fois les valeurs passées de la série et les erreurs résiduelles passées pour prédire la valeur future. Le modèle ARMA est généralement utilisé lorsque la série présente à la fois des dépendances autoregressives et des erreurs résiduelles.
- **Modèle SARMA (Seasonal Autoregressive Moving Average) :** Le modèle SARMA est une extension du modèle ARMA pour les séries temporelles saisonnières. Il prend en compte les motifs saisonniers de la série en utilisant des retards et des erreurs résiduelles saisonnières. Le terme "saisonnier" fait référence aux cycles récurrents observés dans la série sur une période de temps fixe.
- **Modèle ARIMA (Autoregressive Integrated Moving Average) :** Le modèle ARIMA est un modèle de séries temporelles plus général qui combine les éléments des modèles AR, MA et d'une différenciation intégrée.

  La différenciation intégrée est utilisée pour rendre la série temporelle stationnaire en prenant la différence entre les observations successives.

  Le modèle ARIMA est utilisé pour modéliser les séries temporelles non stationnaires.
- **Modèle SARIMA (Seasonal ARIMA) :** Le modèle SARIMA est une extension du modèle ARIMA pour les séries temporelles saisonnières.

  Il combine les éléments des modèles SARMA et ARIMA pour prendre en compte les dépendances saisonnières et les tendances générales de la série.

  Le modèle SARIMA est généralement utilisé pour modéliser des séries temporelles avec des motifs saisonniers et des tendances non linéaires.

le choix du modèle approprié pour modéliser une série temporelle dépend de ses caractéristiques spécifiques, notamment la présence de motifs saisonniers, de tendances et de stationnarité [7].

2.2.4 Limitations des méthodes de prédictions traditionnelles

Bien que ces méthodes de prédiction traditionnelles soient encore largement utilisées aujourd'hui, elles présentent certaines limitations.

L'une des principales limitations est qu'elles ne sont souvent pas assez flexibles pour capturer des relations complexes entre les variables dans les données. Elles peuvent également avoir du mal à gérer des données de grande dimension et des données non linéaires.

Ces limitations ont conduit à l'émergence de méthodes plus avancées telles que l'apprentissage automatique, qui utilise des algorithmes pour apprendre à partir des données sans être explicitement programmé pour cela. Cependant, les méthodes traditionnelles de prédiction restent pertinentes dans de nombreuses applications et constituent une base importante pour comprendre les concepts de la modélisation statistique.

2.3 Méthodes de Machine Learning pour la prédiction

2.3.1 Introduction générale a l'apprentissage automatique

L'intelligence artificielle (IA) est un processus d'imitation de l'intelligence humaine qui repose sur la création et l'application d'algorithmes exécutés dans un environnement informatique dynamique. Son but est de permettre à des ordinateurs de penser et d'agir comme des êtres humains.

L'IA comprend plusieurs sous-domaines, dont l'Apprentissage Automatique (AA ou Machine Learning en anglais).

2.3.2 Définition de l'apprentissage automatique

Arthur Samuel, l'un des premiers leaders américains dans le domaine des jeux informatiques et de l'intelligence artificielle, a inventé le terme « apprentissage automatique » en 1959 alors qu'il était chez IBM. Il a défini l'apprentissage automatique comme "le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés".

Cependant, il n'existe pas de définition universellement acceptée pour l'apprentissage automatique. Différents auteurs définissent le terme différemment. Nous donnons ci-dessous deux autres définitions :

En 1997, l'informaticien américain « Tom Michael Mitchell » introduit une nouvelle définition

de l'apprentissage automatique. Il a considéré qu'un programme apprend d'une expérience E, par rapport à une classe de tâches T, et avec une mesure de performance P [5].

Avec le temps, la définition de l'apprentissage automatique a commencé à prendre une dimension mathématique et statistique. Selon les auteurs dans (Goodfellow, Bengio, et Courville, 2016), l'apprentissage automatique est essentiellement une forme de statistiques appliquées, mettant davantage l'accent sur l'utilisation d'ordinateurs pour estimer statistiquement les fonctions compliquées et un accent moindre sur la démonstration des intervalles de confiance autour de ces fonctions.

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle qui implique le développement d'algorithmes et de modèles statistiques qui permettent aux ordinateurs d'améliorer leurs performances dans les tâches grâce à l'expérience. Ces algorithmes et modèles sont conçus pour apprendre des données et faire des prédictions ou des décisions sans instructions explicites. Il existe plusieurs types d'apprentissage automatique, notamment l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

2.3.3 Types d'apprentissage automatiques

Il existe trois principaux types d'apprentissage automatique supervisé, non supervisé et apprentissage par renforcement [9].

2.3.3.1 Apprentissage supervisé (AS)

L'apprentissage supervisé se fait en présence d'un superviseur tout comme l'apprentissage effectué par un petit enfant avec l'aide de son professeur. Comme un enfant est formé à reconnaître les fruits, les couleurs, les nombres sous la supervision d'un enseignant, cette méthode est un apprentissage supervisé.

Dans cette méthode, chaque étape de l'enfant est vérifiée par l'enseignant et l'enfant apprend de la sortie qu'il doit produire.

Dans l'AS, l'ordinateur est présenté avec des exemples d'entrainement et leurs sorties souhaitées.

$$D = \{(x(1), y(1)), (x(2), y(2)), \ldots, (x(n), y(n))\} \tag{2.7}$$

L'objectif est d'apprendre un modèle/une fonction qui mappe les entrées aux sorties pour les données nouvellement observées $f\mathbf{x} \rightarrow \mathbf{y}$.

L'AS nécessite moins de données d'apprentissage que les autres méthodes d'apprentissage automatique et facilite l'apprentissage car les résultats du modèle peuvent être comparés aux résultats réels marqués.

Il existe deux sous-catégories au sein de l'apprentissage supervisé :

- **La classification :** la classification consiste à trouver le lien entre une variable d'entrée (X) et une variable de sortie discrète (Y), en suivant une loi multinomiale.
- **Régression :** la régression consiste à prédire une valeur continue pour la variable de sortie.

Les algorithmes les plus célèbres utilisés dans cette approche sont les suivants :

∗ **Machine à vecteurs de support (MVS) :**

La MVS est un algorithme d'apprentissage supervisé qui sépare un ensemble de données en différentes classes à l'aide d'un hyperplan en s'appuyant sur la notion de marge maximale. Les points les plus proches de l'hyperplan dans les différentes classes sont appelés vecteurs de support et ces vecteurs de support sont utilisés pour prédire les classes des nouveaux points. En effet, lorsque qu'un nouveau point est placé sur l'équation de l'hyperplan, il est classé dans une classe en fonction du côté de l'hyperplan où il est tombé sur l'espace vectoriel.

La figuer 2.1 [8] montre le principe général de MVS.

l'hyperplan est la droite noire, les « vecteurs de support » sont les points entourés (les plus proches de l'hyperplan) et la « marge » est la distance entre l'hyperplan et les droites bleue et rouge [23].

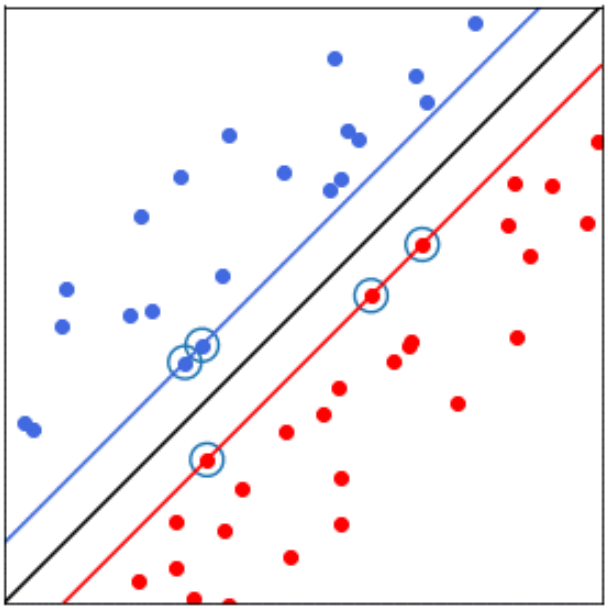


FIGURE 2.1 – Fonctionnement de MVS .

∗ **K plus proches voisins :**

Le K plus proches voisins (KNN) est un algorithme de classification supervisée, utilisé dans le domaine de l'apprentissage automatique. Il s'agit d'une approche non paramétrique, ce qui signifie qu'elle ne fait pas d'hypothèses sur la distribution des données en entrée.

L'idée principale de l'algorithme KNN est de trouver les K échantillons d'entraînement les plus proches de l'échantillon à prédire dans l'espace des caractéristiques. La classe majoritaire parmi les K échantillons les plus proches est alors utilisée pour prédire la classe de l'échantillon à prédire.

La mesure de distance la plus couramment utilisée pour trouver les K échantillons les plus proches est la distance euclidienne. Cependant, d'autres mesures de distance peuvent également être utilisées en fonction du domaine d'application [36].

∗ **Arbres de decesions :**

Un arbre de décision est un modèle de classification qui utilise un arbre pour représenter et classifier les données. Il se compose de nœuds représentant des tests sur des attributs de données, et de branches reliant ces nœuds à des nœuds enfants, qui représentent des résultats de test ou des décisions de classification.

Les arbres de décision sont construits à partir de données d'entraînement et peuvent être utilisés pour prédire la classe ou la valeur d'une nouvelle instance de données en traversant l'arbre à partir de la racine jusqu'à une feuille qui représente une décision de classification.

Les arbres de décision sont souvent utilisés dans les domaines de la classification et de la prédiction en raison de leur simplicité et de leur interprétabilité [16].

2.3.3.2 Apprentissage non supervisé (ANS)

Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé est utilisé pour tirer des conclusions et trouver des modèles à partir des données d'entrée sans se référer aux résultats étiquetés.
Selon [5] , il y a deux types d'apprentissage non supervisé :

- **Regroupement(Clustering) :** Les algorithmes de clustering cherchent à regrouper les données similaires en ensembles distincts appelés clusters. Ces clusters peuvent aider à découvrir des groupes naturels dans les données et à identifier des sous-populations ou des schémas de comportement similaires [5].

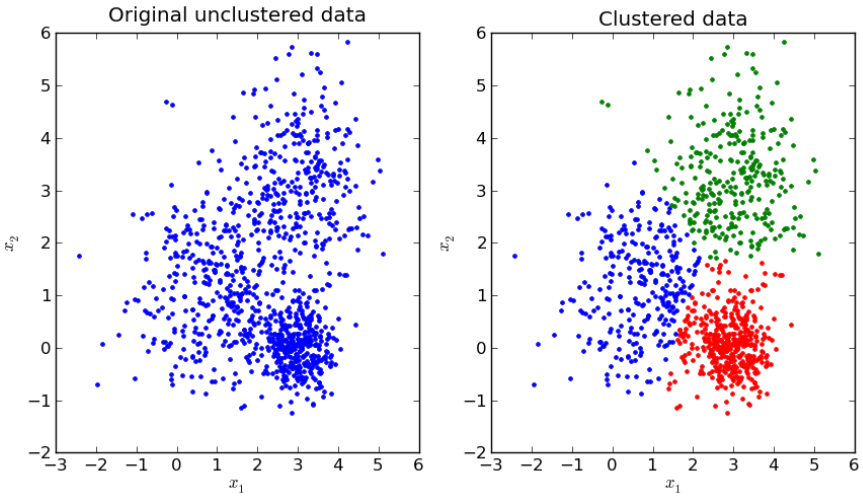


FIGURE 2.2 – prediction 3 Cluster [6].

Le clustering est une technique sans surveillance dans laquelle les points de données sont regroupés. Il est largement utilisé pour la segmentation des clients, la détection des fraudes et la classification des documents. Les techniques de clustering courantes incluent le clustering k-means, le clustering hiérarchique, le clustering à décalage moyen et le clustering basé sur la densité. Bien que chaque technique ait une méthode différente pour trouver des clusters, elles visent toutes à atteindre la même chose.

∗ **K-means :** C'est un type d'algorithme non supervisé qui résout le problème de clustering. Sa procédure suit un moyen simple et facile de classer un ensemble de données donné à travers un certain nombre de clusters (supposons k clusters). Les points de données à l'intérieur d'un cluster sont homogènes et hétérogènes par rapport aux groupes de pairs. En d'autre terme, l'algorithme K-means identifie k nombre de centroïdes, puis alloue chaque point de données au cluster le plus proche, tout en gardant les centroïdes aussi petits que possible
L'algorithme fait une première boucle, en assignant un centroïde aléatoire. Puis, il calcule la distance de chaque point avec le centre de son cluster (une distance mise au carré). Toutes ces distances sont calculées pour chaque point à l'intérieur de chaque cluster. L'objectif étant d'arriver à un modèle où la somme de la distance entre chaque point et son cluster est minimale [38].

- **Réduction de la dimension :** L'objectif est de simplifier les données sans perdre trop d'informations, à titre d'exemple, fusionner plusieurs caractéristiques en un seul caractère.

La réduction de dimensionnalité est la transformation de données d'un espace de grande dimension en un espace de faible dimension de sorte que la représentation de faible dimension conserve

certaines propriétés significatives des données d'origine.

2.3.3.3 Apprentissage par renforcement

Le scénario général de l'apprentissage par renforcement est illustré par la figure 2.3 . Contrairement au scénario de l'apprentissage supervisé, ici, l'apprenant ne reçoit pas passivement un ensemble de données étiquetées. Au lieu de cela, il recueille des informations par le biais d'un cours d'actions en interagissant avec l'environnement. En réponse à une action, l'apprenant ou l'agent reçoit deux types d'informations :

Son état actuel dans l'environnement et une récompense à valeur réelle, spécifique à la tâche et à son objectif correspondant.
L'objectif de l'agent est de maximiser sa récompense et donc de déterminer le meilleur plan d'action, ou la meilleure politique, pour atteindre cet objectif. Cependant, les informations qu'il reçoit de l'environnement ne sont que la récompense immédiate liée à l'action qu'il vient de mener. Un aspect important de l'apprentissage par renforcement consiste à envisager des récompenses ou des pénalités différées. L'agent est confronté à un dilemme entre l'exploration d'états et d'actions inconnus pour obtenir plus d'informations sur l'environnement et les récompenses, et l'exploitation des informations déjà recueillies pour optimiser sa récompense. C'est ce que l'on appelle le compromis entre l'exploration et l'exploitation, qui est lié au l'apprentissage par renforcement [13].

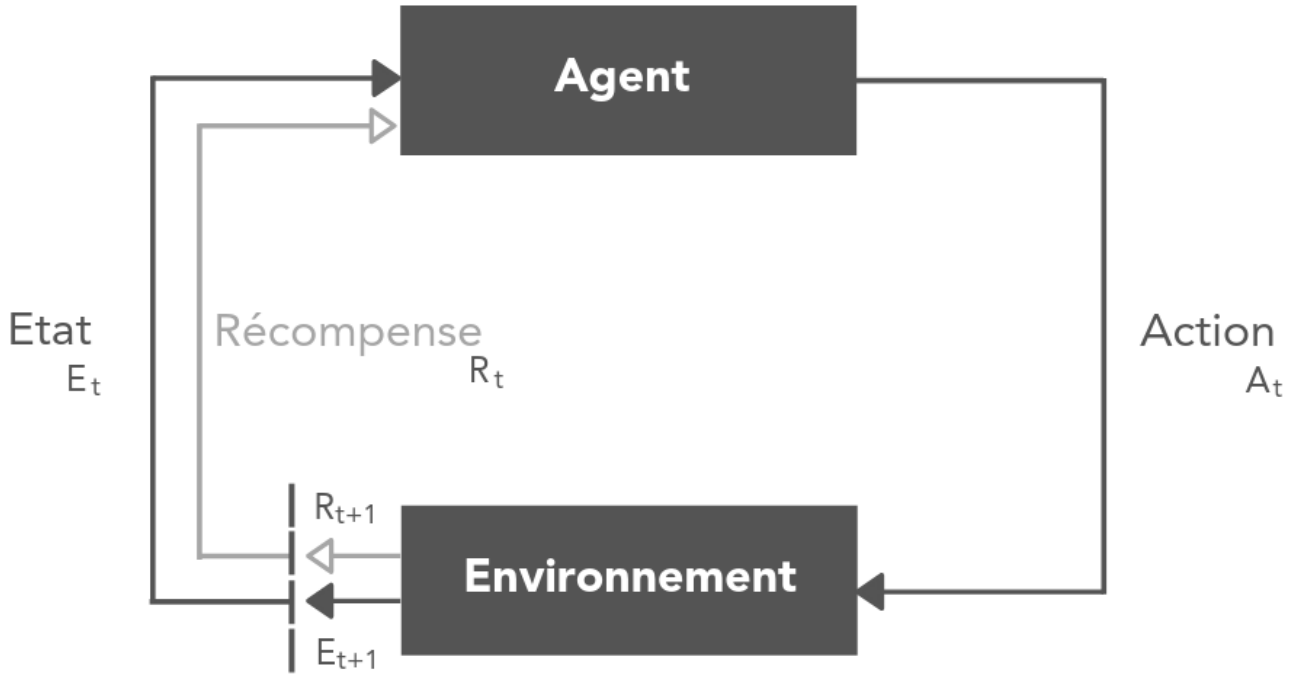


FIGURE 2.3 – Apprentissage par renforcement [52].

Le programme AlphaGo de DeepMind est un bon exemple d'apprentissage par renforcement ,il a

fait l'une des journaux en mars 2016 lorsqu'il a battu le champion du monde Lee Sedol au jeu de Go. Il a appris sa politique gagnante en analysant des millions de parties, puis en jouant de nombreuses parties contre lui-même.

2.3.4 Le choix d'un type d'apprentissage automatique

Avec la présence de différents types de classifieurs pour l'apprentissage automatique, l'opération de choix d'un type est une question typique « Quel algorithme dois-je utiliser ? ». Selon [33], la réponse à cette question varie les facteurs suivants :

- La taille, la qualité et la nature des données.
- Le temps de calcul disponible.
- L'urgence de la tache .
- Le but d'utilisation de ces données.

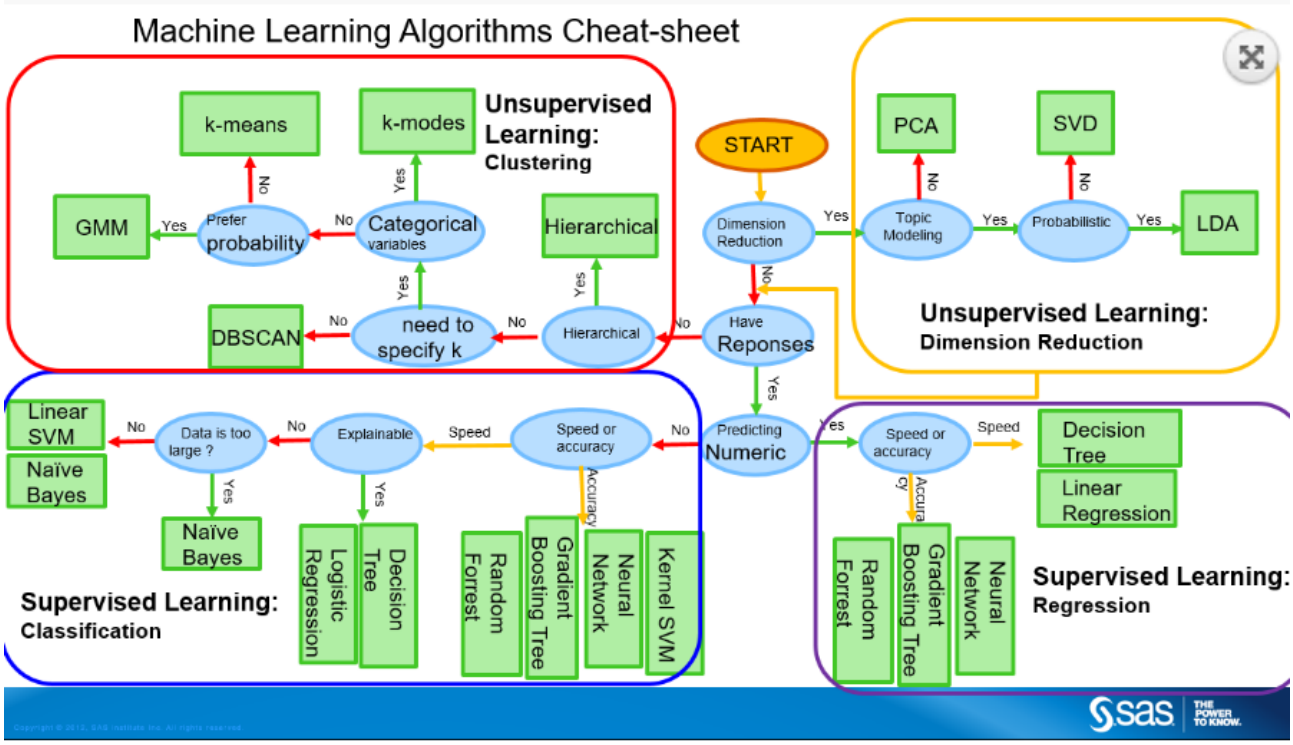


FIGURE 2.4 – Le choix de l'algorithme d'apprentissage selon certains facteurs [51].

2.3.5 De l'apprentissage automatique à l'apprentissage en profondeur

L'apprentissage automatique a prouvé son efficacité dans la résolution d'une variété de problèmes. Cependant, il présente des limites importantes, notamment en termes de temps, de vitesse et d'efficacité. Lorsqu'un algorithme d'apprentissage automatique renvoie une prédiction inexacte, il nécessite

l'intervention d'un ingénieur pour effectuer des ajustements, ce qui peut entraîner une perte de temps et une prévision lente [32].

Les algorithmes simples d'apprentissage automatique fonctionnent bien sur de nombreux problèmes, mais ils ne permettent pas de résoudre certaines clés de l'intelligence artificielle, telles que la reconnaissance de la parole ou la reconnaissance d'objets. Cela a motivé le développement de l'apprentissage en profondeur, qui vise à surmonter les limitations des algorithmes traditionnels. L'apprentissage en profondeur permet de travailler avec des données de grande dimension et de généraliser efficacement sur des tâches complexes de l'intelligence artificielle [27].

Ainsi, l'apprentissage en profondeur offre de nouvelles possibilités pour résoudre des problèmes plus complexes et pour améliorer les performances des systèmes d'intelligence artificielle. Il repose sur des réseaux de neurones artificiels profonds et des architectures complexes qui permettent d'apprendre.

2.3.6 L'apprentissage en profondeur « Deep Learning »

2.3.6.1 Définition et architectures

Le terme « apprentissage profond » a été introduit dans le domaine du l'apprentissage automatique par « Rina Dechter » en 1986, et dans les réseaux de neurones artificiels par « Igor Aizenberg » et ses collègues en 2000, dans le contexte des neurones à seuil booléen, l'apprentissage profond désigne une technique d'apprentissage machine, c'est une sous-branche de l'intelligence artificielle qui vise à construire automatiquement des connaissances à partir de grandes quantités d'information [44] (Voir figure 2.5).

Les caractéristiques essentielles du traitement ne seront plus identifiées par un traitement humain dans un algorithme préalable, mais directement par l'algorithme d'apprentissage profond [4].

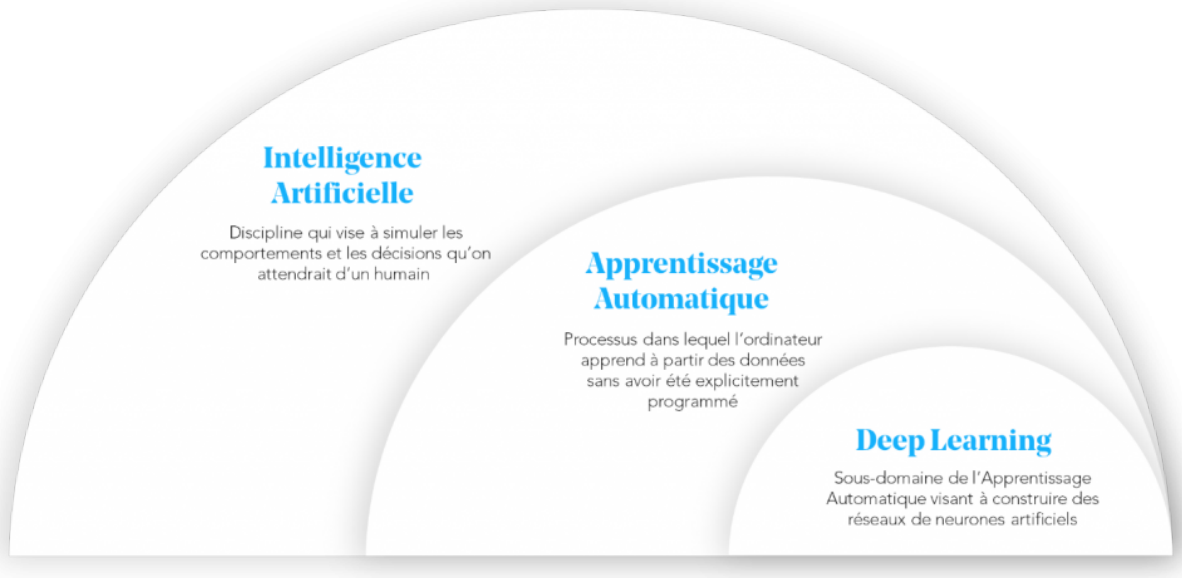


FIGURE 2.5 – les sous-branches de l'intelligence artificielle [14].

L'apprentissage en profondeur permet ainsi de répondre de manière implicite à des questions telles que "Que peut-on déduire de ces données ?" et de découvrir des caractéristiques ou des relations cachées entre les données qui sont souvent difficiles à identifier pour l'homme.

D'après [15], l'apprentissage profond est un réseau neuronal avec un grand nombre de paramètres et de couches. L'exemple de base est le perceptron multicouche MLP « multi-layer perceptron » (voir Figure 2.6).

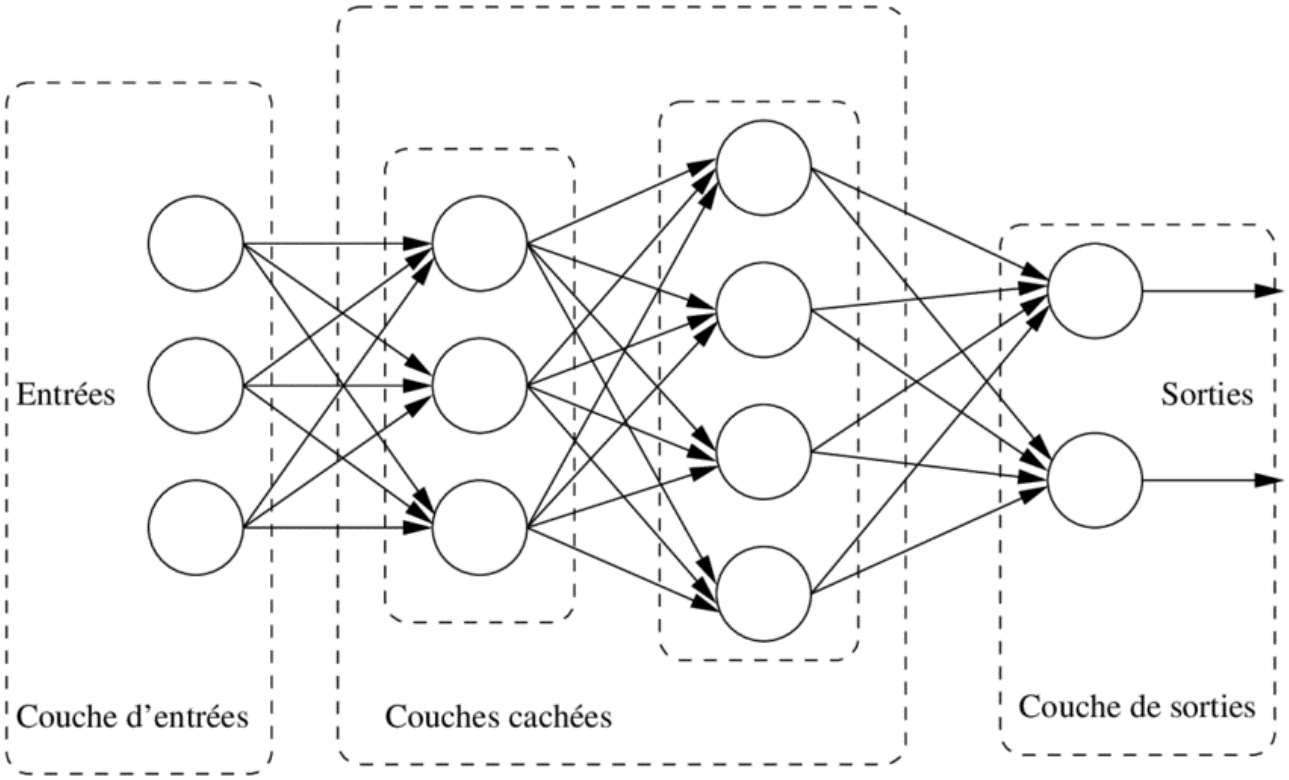


FIGURE 2.6 – Le perceptron multicouche [43].

Dans la suite de ce chapitre, nous allons explorer quelques modèles d'apprentissage en profondeur couramment utilisés dans le domaine de l'intelligence artificielle. Ces modèles comprennent notamment les réseaux de neurones artificiels (ANN), les réseaux de neurones convolutifs (CNN), les réseaux de neurones récurrents (RNN) et les cellules de mémoire à long terme à court terme (LSTM). Chacun de ces modèles présente des caractéristiques spécifiques et est adapté à des types de données et des tâches particulières. Nous allons explorer leurs architectures, leurs fonctionnements et les domaines d'application dans lesquels ils sont excellent.

2.3.7 Modèles d'apprentissage en profondeur

2.3.7.1 Réseaux de Neurones Artificiels (RNA)

Un réseau de neurones artificiels (RNA) est un modèle mathématique inspiré par le fonctionnement du cerveau humain. Il est utilisé pour résoudre des problèmes complexes d'apprentissage automatique en utilisant des algorithmes d'apprentissage.

Le RNA est composé de plusieurs unités de base appelées neurones artificiels, qui sont interconnectés pour former des couches. Chaque neurone reçoit des entrées pondérées, les combine et applique une fonction d'activation pour générer une sortie. Les sorties de certains neurones deviennent les entrées d'autres neurones, formant ainsi des connexions entre les couches. Cette structure en couches

permet au RNA de capturer des relations complexes entre les données d'entrée et de produire des prédictions ou des classifications.

Perceptrons a été inventé en 1958 au « Cornell Aviation Laboratory » par « Frank Rosenblat » financé par le bureau de recherche navale des États-Unis, Le mot vient de verbe latin « Percipio » qui signifie en Anglaise understand ; en Français comprendre, qui montre que Le robot ou l'appareil peut apprendre et comprendre le monde extérieur [30].

Le perceptron est une forme simple de réseau de neurones artificiels, composé d'une seule couche de neurones. Chaque neurone du perceptron prend un ensemble de valeurs d'entrée, les pondère et les passe à travers une fonction d'activation pour produire une sortie. La fonction d'activation peut être une fonction seuil, telle que la fonction d'activation de Heaviside, qui retourne 1 si la somme pondérée des entrées dépasse un seuil et 0 sinon.

Le perceptron est principalement utilisé pour la classification binaire, où il peut apprendre à séparer les données d'entrée en deux catégories distinctes en ajustant les poids des connexions. Il est capable d'apprendre à partir d'exemples étiquetés, où les données d'entrée sont associées à des étiquettes de classe prédéfinies.

Un perceptron multicouche avec plusieurs couches cachées entre la couche d'entrée et la couche de sortie est un réseau de neurones profonds (DNN), le DNN est une fonction mathématique, qui mappe certains ensembles de valeurs d'entrée aux valeurs de sortie. La fonction est formée par la composition de nombreuses fonctions plus simples [27]. Certaines de ses caractéristiques [15] :

- Plus de neurones.
- Des moyens plus complexes de connecter les couches neurones dans les réseaux neuronaux.
- Puissance de calcul.
- Extraction automatique des fonctionnalités.

2.3.7.2 Reseaux de neurones a convolution (CNN)

Le terme "Réseau de neurones à convolution" fait référence à l'utilisation de l'opération mathématique de convolution dans ce type de réseau.

Les réseaux de convolution sont une forme spécialisée de réseaux neuronaux qui utilisent la convolution à la place de la multiplication matricielle générale dans au moins une de leurs couches. Les CNN sont considérés comme l'un des meilleurs algorithmes d'apprentissage pour effectuer cette opé-

ration de convolution, ce qui permet d'extraire des caractéristiques pertinentes à partir de données corrélées localement.

Les résultats des opérations de convolution sont ensuite transmis à des unités de traitement non linéaires, également appelées fonctions d'activation, qui jouent un rôle essentiel dans l'apprentissage des abstractions et l'introduction de non-linéarités dans l'espace des caractéristiques.

Cette non-linéarité permet la génération de modèles d'activation variés pour différentes réponses, ce qui facilite l'apprentissage des différences sémantiques présentes dans les images [17].

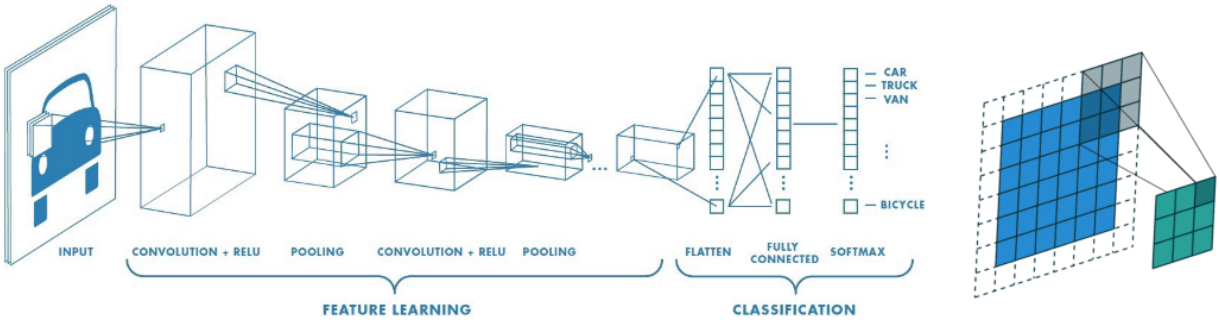


FIGURE 2.7 – L'architecture des réseaux de neurones convolutifs [41].

∗ **L' architecture des réseaux neuronaux convolutifs**

L'architecture des réseaux neuronaux convolutifs (CNNs) compose de trois majors couches :

- **Couche d'entrée :** la couche d'entrée accepte généralement l'entrée tridimensionnelle sous la forme (hauteur*largeur) de l'image et a une profondeur représentant les canaux de couleur (généralement trois pour les canaux de couleur RGB).
- **Couche d'entrainement :** Couche d'entrainement est construite de :
  - **Couches de convolution :** La première couche d'un CNN est utilisée pour extraire les caractéristiques des images d'entrée. Elle effectue une opération mathématique appelée convolution entre l'image d'entrée et un filtre de taille MxM. En glissant ce filtre sur l'image, un produit scalaire est calculé pour chaque partie de l'image correspondant à la taille du filtre.

    La sortie de cette couche est appelée carte des fonctionnalités, qui contient des informations telles que les coins et les bords de l'image. Cette carte de caractéristiques est ensuite transmise à d'autres couches pour apprendre d'autres caractéristiques de l'image.
    La couche de convolution du CNN transmet le résultat à la couche suivante après avoir

appliqué l'opération de convolution sur l'entrée. Les couches convolutives dans un CNN préservent la relation spatiale entre les pixels, ce qui est bénéfique pour l'apprentissage des caractéristiques.

- **Couches d'activation :** Après chaque couche de convolution, une fonction d'activation est appliquée aux résultats obtenus. Cette fonction introduite de la non-linéarité dans le réseau, ce qui permet de modéliser des relations plus complexes entre les caractéristiques extraites.

- **Couches de pooling :** Ces couches réduisent la dimension spatiale des caractéristiques en effectuant une opération de sous-échantillonnage, telle que le max pooling ou le average pooling. Cela permet de réduire la quantité de calcul nécessaire et de conserver les informations essentielles.

- **Couche de classification :** Ces couches sont situées à la fin du CNN et sont responsables de la classification ou de la prédiction finale. Chaque neurone de cette couche est connecté à tous les neurones de la couche précédente, et ils apprennent des combinaisons linéaires des caractéristiques extraites pour effectuer la classification ou la prédiction.

2.3.7.3 Reseaux de Neurones Récurrents(RNN)

Un réseau de neurones récurrents présente une structure similaire à celle d'un réseau de neurones standard, à la différence que des connexions sont établies en boucle dans le réseau. Chaque neurone récurrent reçoit des entrées, génère des sorties, puis se renvoie ces sorties en tant qu'entrées ultérieures.

À chaque pas de temps (ou trame), un neurone récurrent reçoit un vecteur d'entrée x(t), ainsi que sa propre sortie générée à l'étape temporelle précédente, y(t-1). Cela permet aux neurones récurrents de prendre en compte les informations passées lors de la génération des sorties actuelles.

La représentation d'un neurone récurrent sur une dimension temporelle est souvent appelée "déploiement temporel". Cela montre comment les neurones récurrents interagissent au fil du temps, en formant une séquence d'étapes dans le réseau.

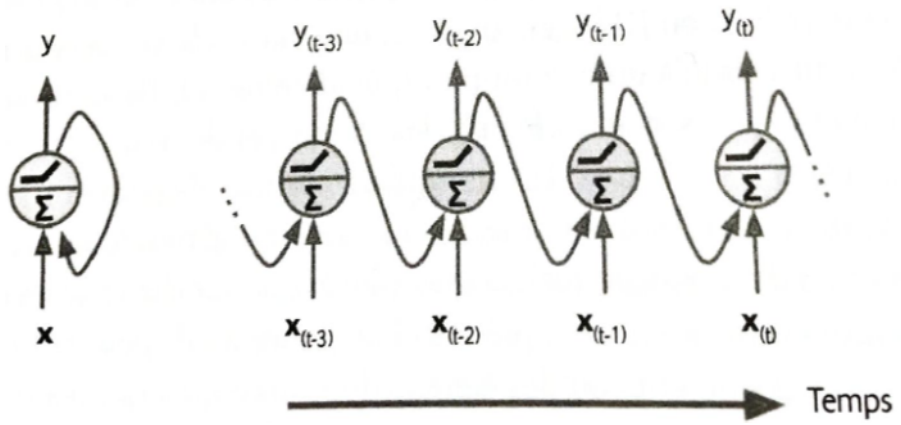


FIGURE 2.8 – Représentation d'un neurone récurrent et son dépilement dans le temps [40].

∗ **Couche de neurones récurrents :**

Les couches de neurones récurrents peuvent être construites en utilisant un vecteur d'entrée $x(t)$ et le vecteur de sortie de l'étape précédente $y(t-1)$ à chaque pas de temps t. Il est important de noter que les entrées et les sorties sont maintenant des vecteurs, contrairement à un neurone individuel où la sortie était un scalaire.

Chaque neurone récurrent possède deux ensembles de poids : un pour les entrées $x(t)$, noté W_x, et un pour les sorties de l'étape précédente $y(t-1)$, noté W_y. En regroupant les vecteurs de poids de tous les neurones dans une couche, nous obtenons les matrices de poids W_x et W_y. La sortie à l'étape t est alors calculée en utilisant l'équation suivante :

$$y(t) = \sigma(W_x^T x(t) + W_y^T y(t-1) + b) \tag{2.8}$$

où σ représente la fonction d'activation et b est le vecteur des termes constants (biais).

La sortie $y(t)$ dépend à la fois de $x(t)$ et de $y(t-1)$, qui dépend à son tour de $x(t-1)$ et de $y(t-2)$, et ainsi de suite. En conséquence, $y(t)$ est une fonction de toutes les entrées depuis le temps $t = 0$, c'est-à-dire $x(0)$, $x(1)$, ..., $x(t)$.

Lors de la première étape temporelle, $t = 0$, les sorties précédentes n'existent pas et sont généralement supposées être toutes égales à zéro.

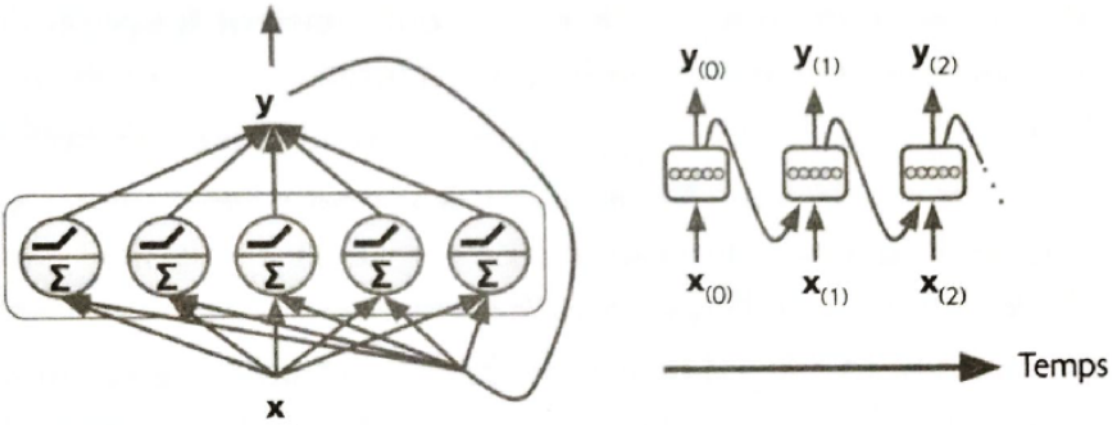


FIGURE 2.9 – Représentation d'une couche de neurones récurrents et son dépilement dans le temps [40].

∗ **Cellules de mémoire :**

Une cellule de mémoire est une partie d'un réseau de neurones qui conserve un état entre plusieurs étapes temporelles. Elle permet à un neurone récurrent de posséder une forme de mémoire en stockant des informations provenant d'étapes temporelles antérieures.

Les cellules de base sont formées soit par un seul neurone récurrent, soit par une couche de neurones récurrents.

L'état d'une cellule à l'étape temporelle t est généralement noté $h(t)$ (h pour "hidden"), et est une fonction de certaines entrées à cette étape temporelle et de son propre état à l'étape temporelle précédente. On peut l'écrire sous la forme :

$$h(t) = f(h(t-1), x(t)) \tag{2.9}$$

La sortie de la cellule à l'étape temporelle t, notée $y(t)$, dépend également de l'état précédent et des entrées courantes.

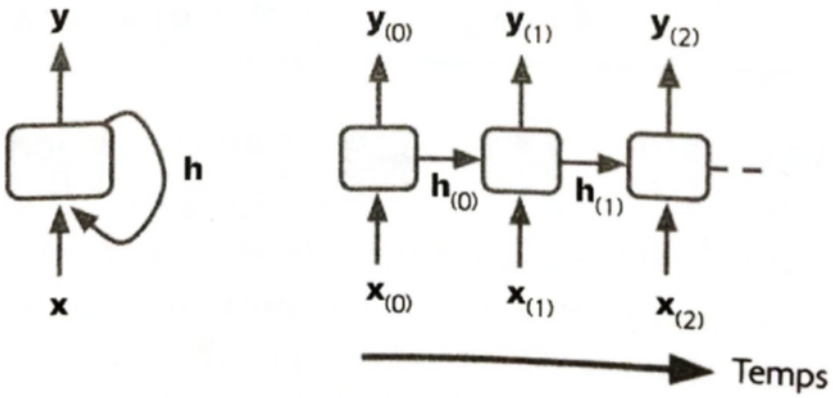


FIGURE 2.10 – Représentation d'une cellule de mémoire et son dépilement dans le temps [40].

* **Types d'entrées et de sorties d'un RNN :**

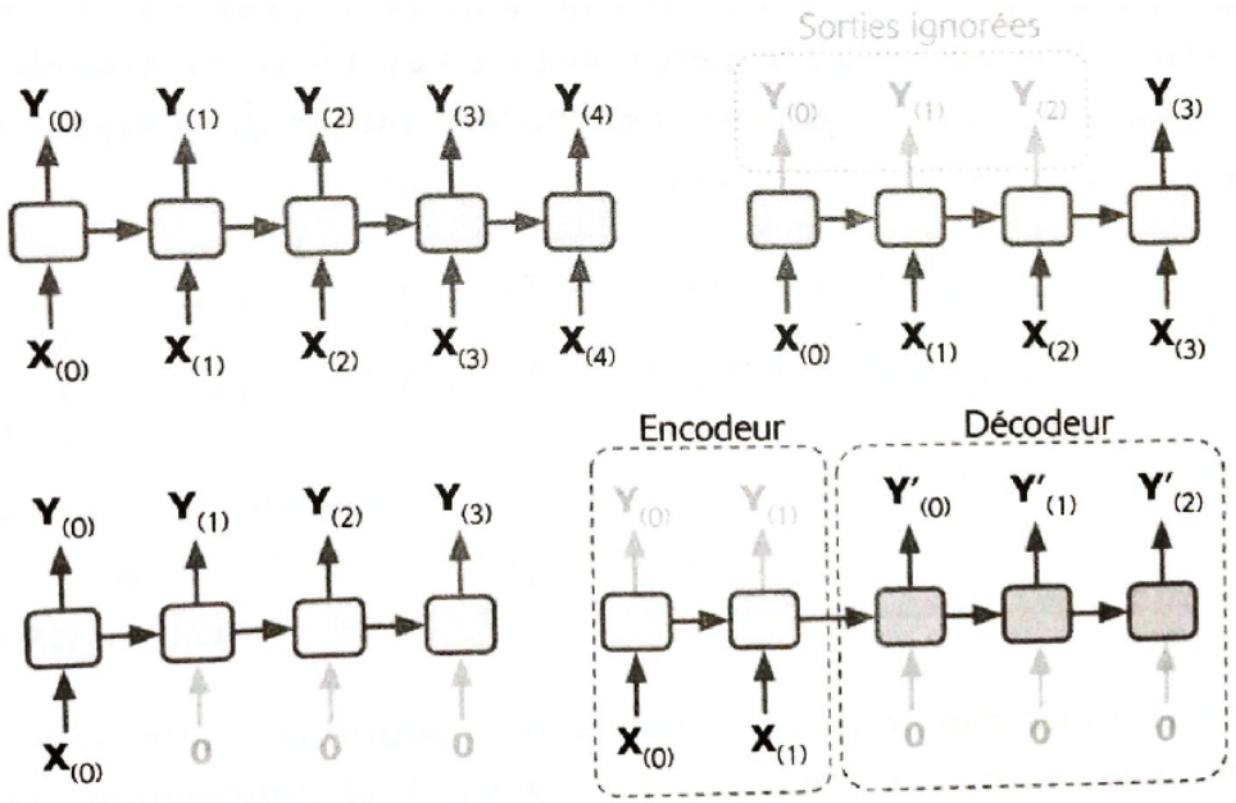


FIGURE 2.11 – différentes types des entrees et des sorties d'un RNN [40].

Série-vers-série (en haut à gauche), série-vers-vecteur (en haut à droite), vecteur-vers-série (en bas à gauche) et série-vers-série différé (en bas à droite).

* **Difficultés d'entrainement sur de nombreuses étapes temporelles :**

Pour entraîner un RNN sur de longues séries, il faut l'exécuter sur de nombreuses étapes temporelles, ce qui résulte en un RNN dépilé extrêmement profond.

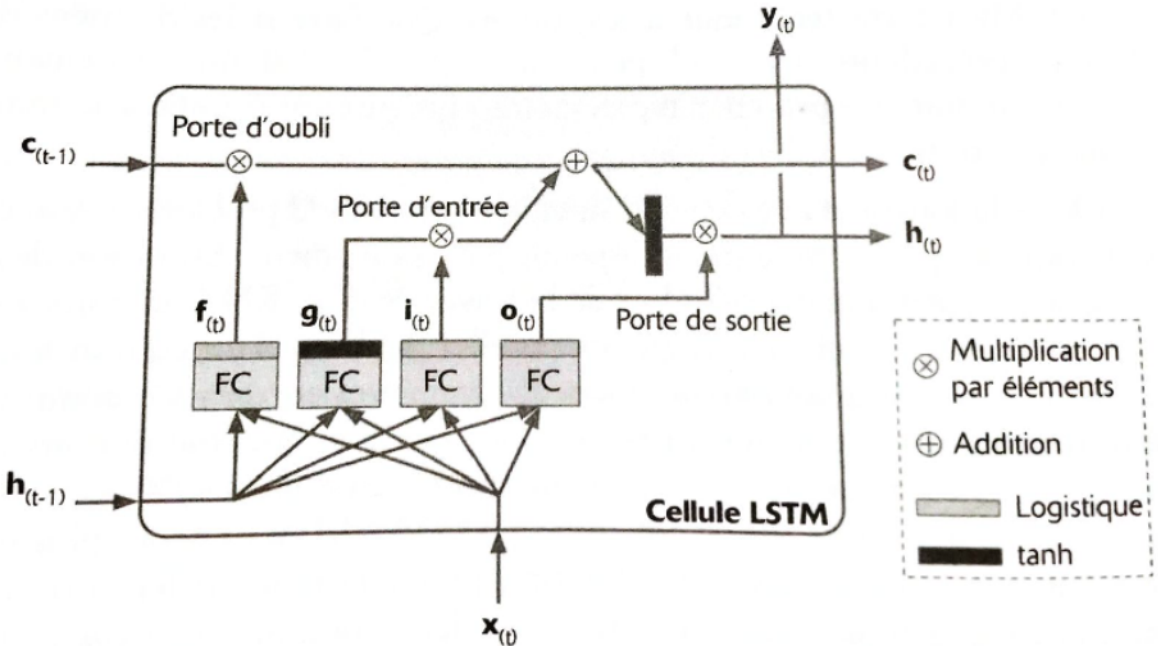
Un RNN profond peut être confronté au problème de disparition/explosion des gradients et l'entrainement peut durer indéfiniment.

Pour résoudre ce problème, plusieurs types de cellules avec une mémoire à long terme ont été imaginés. Leur efficacité a été telle que les cellules de base ne sont quasi plus employées. Les plus populaires de ces cellules sont : longue mémoire à court terme (Long Short-term Memory, LSTM), et unité récurrente à porte (Gated Recurrent Unit, GRU). Ces deux cellules ont énormément contribué au succès des RNNs des dernières années.

2.3.7.4 Longue mémoire à court terme (LSTM)

La cellule de longue mémoire à court terme (Long Short-term Memory, LSTM), proposée en 1997 par Hochreiter et Schumidhuber, a connu des améliorations continues au fil des années par divers chercheurs. Les LSTMs sont particulièrement efficaces pour l'entraînement, convergent plus rapidement et sont capables de capturer les dépendances à long terme dans les données. La structure de la cellule LSTM est similaire à celle d'une cellule classique, à l'exception de son état qui est divisé en deux vecteurs distincts : h(t) et c(t).

Le vecteur h(t) représente l'état à court terme, tandis que le vecteur c(t) représente l'état à long terme.



Représentation d'une cellule LSTM [40].

∗ **Cellule LSTM : idée centrale**

Dans une cellule LSTM, le réseau est capable d'apprendre ce qui doit être stocké dans l'état à long terme, ce qui doit être oublié et ce qui doit être lu. Le flux de l'état à long terme c(t-1) à travers le réseau se déroule de gauche à droite. Tout d'abord, il passe par une porte d'oubli (forget gate) qui élimine certaines informations, puis il ajoute de nouvelles informations via une opération d'addition sélectionnée par une porte d'entrée (input gate). Le résultat c(t) de l'état à long terme est ensuite envoyé directement sans autre transformation.

En outre, après l'opération d'addition, l'état à long terme est copié et passe par une fonction tanh, dont le résultat est filtré par une porte de sortie (output gate). Cela donne l'état à court terme h(t), qui correspond à la sortie de la cellule pour l'étape temporelle y(t). Dans une cellule LSTM,

le vecteur d'entrée courant x(t) et l'état à court terme précédent h(t-1) sont fournis à quatre couches entièrement connectées, chacune ayant un objectif spécifique. La couche principale génère g(t) en analysant les entrées courantes x(t) et l'état précédent à court terme h(t-1). La sortie de cette couche est partiellement stockée dans l'état à long terme.

Les trois autres couches sont des contrôleurs de porte. Elles utilisent la fonction d'activation logistique, de sorte que leurs sorties sont dans la plage de 0 à 1. Ces sorties sont ensuite utilisées dans des opérations de multiplication élément par élément, où une valeur de 0 ferme la porte et une valeur de 1 l'ouvre. Plus précisément : La porte d'oubli (contrôlée par f(t)) décide des parties de l'état à long terme qui doivent être effacées.

La porte d'entrée (contrôlée par i(t)) sélectionne les parties de g(t) qui doivent être ajoutées à l'état à long terme. La porte de sortie (contrôlée par o(t)) choisit les parties de l'état à long terme qui doivent être lues et produites lors de cette étape temporelle (dans h(t) et y(t)).

En résumé, une cellule LSTM est capable d'apprendre à identifier une entrée importante (rôle de la porte d'entrée), la stocker dans l'état à long terme, apprendre à la conserver aussi longtemps que nécessaire (rôle de la porte d'oubli) et apprendre à l'extraire lorsque cela est requis.

2.3.8 Les étapes d'apprentissage automatique

L'apprentissage automatique permet aux machines d'acquérir des connaissances et d'améliorer leurs performances en analysant des données.
Ce processus se déroule à travers différentes étapes clés qui transforment ces données en informations exploitables.
Dans cette section, nous allons présenter les principales étapes de ce processus d'apprentissage.

2.3.8.1 Préparation des données

La préparation des données est une étape cruciale qui vise à rendre les données prêtes à être utilisées par les modèles d'apprentissage. Voici quelques étapes courantes de préparation des données :

* **Collecte des données :** L'étape de collecte des données consiste à rassembler les données nécessaires pour entraîner et tester votre modèle d'apprentissage automatique. Cela peut impliquer la collecte de données à partir de différentes sources telles que des bases de données, des fichiers, des API, des capteurs, des sites web, etc. Il est important de veiller à ce que les données collectées soient représentatives et pertinentes pour votre problème d'apprentissage.

* **Nettoyage des données :** Le nettoyage des données est une étape qui vise à traiter les valeurs manquantes, les valeurs aberrantes et les doublons dans les données collectées. Cela peut impliquer des actions telles que la suppression des lignes ou des colonnes avec des valeurs manquantes, l'imputation des valeurs manquantes en utilisant des méthodes statistiques, la détection et le traitement des valeurs aberrantes, ainsi que la suppression des doublons. Cette étape permet d'assurer la qualité et la cohérence des données utilisées pour l'apprentissage automatique.
* **Transformation des données :** La transformation des données consiste à préparer les variables d'entrée de manière à ce qu'elles soient appropriées pour les algorithmes d'apprentissage automatique. Cela peut inclure des actions telles que la normalisation des variables numériques pour les mettre à la même échelle, la conversion des variables catégorielles en variables indicatrices (one-hot encoding), la discrétisation des variables continues en intervalles, ou la réduction de dimension pour gérer les données de grande dimension. L'objectif est de rendre les données plus adaptées à l'apprentissage et d'améliorer les performances du modèle.
* **Fractionnement des données :** Le fractionnement des données consiste à diviser l'ensemble de données en ensembles distincts pour l'entraînement, la validation et les tests. L'ensemble d'entraînement est utilisé pour ajuster les paramètres du modèle, l'ensemble de validation est utilisé pour évaluer et optimiser les performances du modèle pendant le processus d'apprentissage, et l'ensemble de test est utilisé pour évaluer objectivement les performances finales du modèle. Ce fractionnement permet d'évaluer la généralisation du modèle sur des données qu'il n'a pas vues lors de l'entraînement [45].

la préparation des données pour l'apprentissage automatique nécessite de collecter les données appropriées, de nettoyer les données en traitant les valeurs manquantes et aberrantes, de transformer les variables pour les rendre compatibles avec les algorithmes d'apprentissage, et de fractionner les données en ensembles d'entraînement, de validation et de test pour évaluer les performances du modèle.

2.3.8.2 Choix et implémentation du / des modèles

Une fois les données préparées, la prochaine étape importante de l'apprentissage automatique est le choix du modèle approprié. Avec la disponibilité de nombreuses bibliothèques de Machine Learning, l'implémentation des modèles n'est plus la partie la plus difficile du projet. Des bibliothèques populaires comme Python, R, C++, C ou Julia offrent des outils pour implémenter différents modèles.

Le choix du modèle dépend de l'objectif métier. Si l'interprétation des résultats est essentielle, un modèle simple avec peu de variables, comme un arbre de décision, peut être privilégié. En revanche,

si la précision prédictive est primordiale, des modèles plus complexes basés sur des techniques de Deep Learning peuvent être utilisés pour une analyse fine et précise.

Quel que soit le type de modèle, ils ont tous des hyperparamètres spécifiques. Ce sont des variables d'ajustement qui permettent de contrôler le processus d'apprentissage et d'entraînement du modèle. Ces hyperparamètres doivent être choisis de manière appropriée pour obtenir de bonnes performances du modèle.

Dans la section suivante, nous détaillerons le processus d'apprentissage et d'entraînement du modèle, où ces hyperparamètres joueront un rôle crucial dans l'optimisation des performances du modèle.

2.3.8.3 Entrainement du modèle

L'entraînement d'un modèle d'apprentissage automatique est une étape essentielle du processus. Elle consiste à fournir au modèle les données collectées et nettoyées lors des étapes précédentes. Dans un modèle d'apprentissage supervisé, ces données sont généralement séparées en variables d'entrée (ensemble de caractéristiques) et variables de sortie (ensemble cible).

L'objectif de la phase d'entraînement est d'améliorer progressivement la capacité du modèle à réagir et à résoudre des problèmes complexes en minimisant une fonction d'erreur ou de coût. Il est crucial que les données utilisées soient de haute qualité et représentatives de la situation à analyser afin d'éviter l'introduction de biais dans les résultats et de maintenir une précision optimale.

Pour éviter les biais statistiques, il est courant de diviser les jeux de données en deux ou trois parties distinctes. La première partie est utilisée pour concevoir et entraîner le modèle (ensemble d'entraînement), la deuxième partie est utilisée pour le tester (ensemble de test), et éventuellement une troisième partie est utilisée pour la validation (ensemble de validation). Il est préférable de réaliser cette séparation de manière aléatoire tout en veillant à ce que chaque partie conserve une représentativité similaire des données (par exemple, en utilisant la validation croisée ou cross-validation). Cela est particulièrement critique lorsque l'objectif est de détecter des phénomènes rares, tels que la prédiction d'événements exceptionnels

2.3.8.4 Evaluation et validation du modèle

L'évaluation et la validation du modèle sont des étapes essentielles dans le processus d'apprentissage automatique. Après avoir implémenté et entraîné le modèle sur l'ensemble d'entraînement, il est nécessaire d'évaluer sa performance et de le valider avant de le déployer.

L'évaluation du modèle consiste à mesurer sa capacité à produire des résultats précis et fiables. Cela se fait en utilisant des métriques appropriées en fonction du type de tâche d'apprentissage.

La validation du modèle se fait en utilisant un ensemble de données de test distinct, qui n'a pas été utilisé lors de l'entraînement. Le modèle est appliqué à cet ensemble de test pour évaluer sa capacité à généraliser les connaissances apprises sur de nouvelles données. Cela permet de vérifier si le modèle est capable de produire des résultats cohérents et fiables dans des conditions réelles.

Il est important de sélectionner un ensemble de test représentatif de la situation réelle à analyser ou à reproduire. Cela garantit que les résultats obtenus lors de la validation reflètent les performances réelles du modèle lorsqu'il est déployé en production.

Si les critères de validation ne sont pas atteints, des ajustements peuvent être nécessaires, tels que la modification des hyperparamètres du modèle, l'ajout de nouvelles données d'entraînement ou l'utilisation de techniques d'optimisation. Le processus d'évaluation et de validation est itératif et se poursuit jusqu'à ce que les performances satisfaisantes soient atteintes [12].

2.3.9 Les métriques d'évaluations d'un modèle d'apprentissage automatique

Les métriques de performances sont utilisées dans les modèles d'apprentissage automatique pour évaluer la qualité et la précision des prédictions du modèle.

Les métriques de performances permettent de mesurer à quel point les prédictions du modèle sont proches des valeurs réelles. Elles fournissent des indications quantitatives sur l'exactitude, la précision et la fiabilité du modèle. Ces métriques permettent aux chercheurs, aux ingénieurs et aux praticiens de comprendre les forces et les faiblesses du modèle, d'effectuer des comparaisons entre différents modèles ou différentes configurations de modèles, et de prendre des décisions éclairées pour améliorer les performances du modèle.

Exemples de métriques d'évaluation couramment utilisées pour la prédiction :

- **Erreur absolue moyenne (Mean Absolute Error, MAE)**

  Le MAE mesure l'erreur moyenne absolue entre les valeurs prédites et les valeurs réelles. Il est calculé en prenant la moyenne des valeurs absolues des différences entre les prédictions et les valeurs réelles.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.10}$$

  Où y représente les valeurs réelles,$\hat{y}$ représente les valeurs prédites, et n est le nombre d'échantillons.

- **Erreur quadratique moyenne (Mean Squared Error, MSE)**

  Le MSE mesure l'erreur quadratique moyenne entre les valeurs prédites et les valeurs réelles. Il est calculé en prenant la moyenne des carrés des différences entre les prédictions et les valeurs réelles.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.11}$$

- **Racine carrée de l'erreur quadratique moyenne (Root Mean Squared Error, RMSE)**

  Le RMSE est simplement la racine carrée du MSE. Il fournit une mesure de l'erreur moyenne entre les prédictions et les valeurs réelles, mais avec la même unité que les données d'origine.

$$RMSE = \sqrt{MSE} \tag{2.12}$$

- **Coefficient de détermination (R-squared)** Le R^2 mesure la proportion de la variance totale des valeurs réelles expliquée par le modèle. Il est souvent utilisé pour évaluer la qualité de l'ajustement du modèle aux données.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.13}$$

  Où $\bar{y}$ est la moyenne des valeurs réelles [50].

2.4 Conclusion

Ce chapitre nous a permis de comprendre les différentes méthodes de prédiction utilisées en science des données.

Nous avons tout d'abord abordé les méthodes traditionnelles de prédiction, en expliquant leur fonctionnement, leurs principes et leurs limites.

Ensuite, nous nous sommes concentrés sur les méthodes plus récentes telles que l'apprentissage automatique et en profondeur, qui ont connu une évolution considérable ces dernières années. Nous avons étudié les différents types d'apprentissage automatique, le choix des modèles et les étapes à suivre pour entraîner un modèle.

Enfin, nous avons examiné les métriques couramment utilisées pour évaluer la performance d'un modèle d'apprentissage automatique. Cette compréhension nous permettra de choisir la méthode de

prédiction la plus adaptée en fonction de nos objectifs et des données disponibles.

CHAPITRE

3

ETAT DE L'ART

3.1 Introduction

La prédiction de la consommation de gaz naturel est un domaine important pour les industries du gaz et de l'énergie.

Les méthodes traditionnelles ne peuvent pas toujours prendre en compte les complexités de la consommation de gaz naturel, d'où la nécessité de combiner avec des méthodes plus avancées comme l'apprentissage automatique qui sont largement utilisés dans ce domaine.

Dans ce cadre, ce chapitre vise à explorer l'état de l'art de la prédiction de la consommation de gaz naturel en utilisant les techniques de machine learning et des méthodes statistique .

Pour ce faire, nous avons synthétisé certains travaux connexes existants en identifiant les méthodes utilisées, les facteurs pris en compte et les résultats obtenus. Notre constat révèle la plupart de ces travaux mettent l'accent sur la combinaison des algorithmes de Machine Learning et des méthodes statistiques pour la prédiction.

Nous avons ensuite établi un tableau de comparaison des travaux connexes qui comporte la technique utilisée dans chaque article, la sortie prédite et le point fort de la technique utilisée.

Enfin, nous avons présenté une synthèse de cette comparaison en expliquant les méthodes les plus efficaces pour la prédiction de la consommation du gaz naturel.

Dans cette étude, nous contribuerons à une meilleure compréhension des méthodes de prédiction de la consommation de gaz naturel en utilisant les techniques de machine learning et des méthodes statistiques . Les résultats de cette étude pourraient être utiles aux fournisseurs de gaz naturel pour optimiser leur production et leur capacité en fonction de la demande de leurs clients.

3.2 Travaux connexes

E.F. Sánchez-Úbeda et A. Berzosa [55] ont proposé une nouvelle approche sur la prévision à moyen terme de la consommation finale du gaz naturel industriel basée sur un modèle hybride qui combine des aspects non paramétriques (Supersmoother) et paramétriques (polynômes en morceaux) pour ajuster les données de consommation de gaz naturel. Ce modèle doit être capable de capturer les motifs de demande dans un grand nombre de profils historiques différents et de fournir des prévisions jusqu'à 3 ans à l'avance. L'approche vise également à être simple et interprétable, permettant aux ingénieurs et aux opérateurs du système de comprendre le comportement du modèle. Elle cherche à exploiter les avantages des prévisions basées sur le jugement et les prévisions statistiques tout en

évitant leurs inconvénients .

A. Azadeh et al. [25] ont présenté une approche combinant ANFIS (Adaptive Network-Based Fuzzy Inference System) et SFA (Stochastic Frontier Analysis) pour la prédiction à long terme de la consommation de gaz naturel. Le modèle utilise à la fois des données historiques de consommation de gaz et des variables d'entrée telles que le PIB et la population pour prévoir la demande future de gaz naturel dans quatre pays du Moyen-Orient (Bahreïn, Arabie saoudite, Syrie et Émirats arabes unis) sur la période 1980-2007. Les résultats montrent que le modèle ANFIS est capable de produire des résultats précis en termes d'erreur absolue moyenne (MAPE) pour la prévision de la demande de gaz naturel dans les pays étudiés. Les variables d'entrée du PIB et de la population ont une forte capacité explicative pour estimer la demande réelle de gaz naturel. Cette approche peut être appliquée à d'autres cas pour estimer et prévoir de manière optimale la consommation du gaz.

M. Salehi et al. [18] ont proposé une nouvelle approche intitulé 'predicting national gas consumption in Iran using a hierarchical combinaison of neural network and genetic algorithm' en utilisant une combinaison hiérarchique de reseaux de neurones artificiels (ANN) et d'algorithmes génétique (AG) a pour objectif de prédire la consommation annuelle du gaz naturel en Iran, les résultats obtenus sont efficaces et précises, Ensuit les resultats de l'approche proposé sont comparés a des modeles ANN traditionnels et il est montré que la combinaison d'ANN et d'AG est superieure aux méthodes traditionnelles en termes de précision.

O. Laib et al. [19] ont proposé une étude nommé 'forecasting yearly natural gaz consumption using artificial neural netwok for the algérian market' qui utilise des réseaux de neurones artificiels (ANN) pour prédire la consommation annuelle de gaz naturel en Algérie, en se concentrant sur les trois secteurs de pression (basse, moyenne et haute). Les données de cette recherche ont été obtenues auprès de la société nationale algérienne de distribution du gaz naturel sonalgaz. Au lieu d'utiliser un modèle ANN unique pour l'ensemble du marché, les auteures ont proposé une approche plus détaillée dans laquelle chaque division de distribution au sein de sonalgaz est analysée séparément. Des entrées influentes sont sélectionnées pour chaque division, et un modèle spécifique de Perceptron multicouche (MLP) est développé et entraîné à l'aide de l'algorithme d'apprentissage de Levenberg-Marquardt. Les résultats de tous les modèles MLP sont ensuite additionnés pour donner la consommation totale de chaque secteur, En conclusion, les résultats obtenus avec les modèles MLP sont très satisfaisants et encourageants par rapport à la méthode de régression linéaire. Les MLP peuvent être utilisées de manière fiable et précise pour prédire la consommation de gaz naturel et constituent donc un excellent outil pour aider la compagnie de distribution à prendre les meilleures décisions.

O.F. Beyca et al. [20] ont proposé une étude sur la prédiction du la consommation du gaz naturel. Cette étude examine les différents modèles d'apprentissage automatique dans la province d'Istanbul en Turquie. Trois techniques sont utilisées, la régression linéaire multiple (MLR) c'est un modèle traditionnelle à été utilisé pour comparer et démontrer les performances prédictives des méthodes ANN et SVR, une approche de réseau neuronal artificiel (ANN) sont inspirés en fonctions neurologique du cerveau humain et sont formulés sur le système cognitifs humain classer en différentes versions telles que les techniques d'apprentissage supervisé et non supervisée, la régression vectorielle de support(SVR) une technique d'exploration des données robuste et précise dans les problèmes de reconnaissance de formes et de classification et de régression. Les résultats montrent que le SVR est supérieur aux autres techniques, fournissant des résultats plus fiables et précis en termes d'erreurs de prédiction pour la consommation de gaz naturel, Cependant, il est important de noter que les prévisions de consommation du gaz naturel basées sur l'apprentissage automatique ne sont que des estimations basées sur les données historiques. Les incertitudes peuvent entraîner des erreurs dans les prévisions et il est donc important de les surveiller et de les ajuster régulièrement. Cette étude pourrait être utile pour d'autres pays en développement en raison de la structure de leurs données de consommation.

M. Akpinar et N. Yumuşak [28] ont proposé une étude sur la prédiction de la demande du gaz naturel à moyen term, Les deux auteurs ont évalué différentes techniques univariées de séries chronologiques. Ils ont appliqué les modèles ARIMA/SARIMA, Holt-Winters et la décomposition des séries chronologiques (TSD) à des données historiques de consommation de gaz naturel sur une base quotidienne entre 2011 et 2013 pour prédire l'année 2014. Les résultats ont montré que le modèle ARIMA(1,0,1)1(0,1,1)365 a donné les taux d'erreur les plus bas (24,6% MAPE) et la meilleure conformité (0,802 R2) pour les prévisions quotidiennes, ainsi que les taux d'erreur les plus bas (11,32% MAPE) et la meilleure conformité (R2 - 0,981) pour les prévisions mensuelles. Ces résultats suggèrent que les modèles ARIMA saisonniers sont les plus appropriés parmi les techniques univariées testées. Les chercheurs ont ainsi proposé ces résultats comme référence pour les décideurs afin d'évaluer la cohérence de leurs prédictions et de leurs méthodes, en prenant en compte la complexité de calcul associée à chaque modèle. SARIMA a donné le taux d'erreur le plus bas, suivi de Holt-Winters et de TSD. Cependant les modèles univariés ne prennent en compte qu'une seule variable et ne tiennent pas compte des relations complexes et des influences croisées avec d'autres variables telles que les conditions météorologiques, les prix de l'énergie, l'évolution économique, etc. Cela peut limiter la capacité du modèle à capturer tous les facteurs pertinents qui peuvent affecter la demande du gaz naturel et conduire à des prévisions moins précises. L'inclusion de variables supplé-

mentaires et l'utilisation de modèles multivariés pourraient potentiellement améliorer la précision des prévisions.

V. Sharma et al. [21] ont proposé une nouvelle approche pour la prevision de la demande du gaz naturel a court terme appelé « Data-driven short-term natural gas demand forecasting with machine learning techniques » en utilisant les techniques d'apprentissage automatique. Les auteurs ont combiné quatres algorithmes de ML tels que la regression lineaire, les reseaux de neurones artificiels, le support vector machine et le processus gaussien (pour modeliser la correlation entre les donnee s-eco et meteorologique) afin d'amélioré la précision des prévisions, la methode prend en compte plusieurs facteurs tels que les donnees meteorologiques, les données socio-economique et les donneés historiques pour prédire la demande du gaz a court terme.La methode a été testé sur des donnees réelles d'une entreprise de gaz naturel aux États-Unis et ont démontré que la méthode est plus performante que les méthodes de prévision existantes.Bien que la méthode soit capable de produire des prévisions précises, elle nécessite des données précises et fiables pour fonctionner efficacement .

W. Panek et T. Włodek [26] ont présenté une méthode de prédiction de la consommation du gaz naturel intitulé 'Naturel gas consumption forecasting based on the variability of external meteorological factors using machine learning algorithms' cette méthode est basée sur la modélisation de la relation entre la consommation du gaz et les facteurs externes tels que la météorologie (la température et l'humidité). La méthode compare trois algorithmes d'apprentissage automatique tels que la régression linéaire multiple, la régression de forêt aléatoire et les réseaux de neurones pour prédire la consommation de gaz naturel dans une ville moyenne en Pologne. Les résultats montrent que la régression de forêt aléatoire est la méthode la plus précise pour prédire la consommation du gaz naturel à court terme, tandis que les réseaux de neurones sont plus précis pour les prévisions à long terme. (Cela signifie que chaque méthode de prédiction peut avoir ses propres forces et faiblesses en fonction de la durée de la prévision nécessaire). Cette méthode de prédiction de la consommation de gaz naturel peut être utilisée pour optimiser la gestion de l'énergie dans les villes et pour aider les fournisseurs du gaz naturel à planifier la production et la distribution du gaz .

3.3 Tableau comparatif des solutions lus

Auteurs	Dataset	Objectif	Méthodes	Résultats
W. Panek et T. Włodek (2022)	Données réelles du competitions de prévisions nPower 2018	Prévision de la consommation de gaz naturel basée sur la variabilité de facteurs météorologiques externes	Combinaison de trois algorithms tel que MLR, RF et DNN	En combinant les algorithmes de ML, tels que DNN et RF, cette technique permet de prendre en compte plusieurs facteurs et d'obtenir des prévisions plus précises.
V. Sharma et al. (2021)		Utilisation de techniques d'apprentissage automatique pour la prévision de la demande de gaz naturel à court terme	Combinaison de quatres algorithms de ML tel que : ANN, SVM, LR et le processus gaussien	La méthode est plus performante que les méthodes de prévision existantes Bien que la méthode soit capable de produire des prévisions précises.
M. Akpinar et N. Yumuşak (2020)	Des données triennaux sur une base quotidiennes entre 2011 et 2013	La prévision à moyen terme de la demande du gaz naturel sur une base quotidienne et mensuelle à l'aide de méthodes statistqiues de saisonnalité univariées	Décomposition de séries temporelles,lissage exponentiel de Holt-Winters, ARIMA/SARIMA	Les modéles ARIMA saisonniers sont la technique d'estimation la plus appropriée parmi les techniques univariées on comparants leurs taux de réussite MAPE.
O.F. Beyca et al. (2019)	Les variables utilisées comprennent l'indice saisonnier, la température, le prix du gaz naturel, la population de la province d'Istanbul et la consommation de 12 mois	Utilisation de techniques d'apprentissage automatique pour prévoir la consommation de gaz naturel mensuelle dans la province d'istanbul	Utiliser deux puissantes techniques d'apprentissage automatique ANN et SVR et une technique traditionnelle MLR	Le modèle SVR avec fonction de noyau cubique polynomial surpassait les modèles ANN et MLR pour l'estimation mensuelle de la consommation du gaz naturel en utilisant des variable d'entrée .
O. Laib et al. (2016)	Les données couvrent la période entre 2000-2014 , fournies par la compagnie SONALGAZ	Prévision de la consommation annuelle de gaz naturel pour le marché algérien	MLP entrainé à l'aide de l'algorithme Levenberg-Marquardt	En considérant chaque division de distribution individuellement ont pu aboutir à une approche plus personnalisée et précise.
M. Salehi et al. (2014)	Enregistrement quotidienne de la température et l'historique de 2005-2012	Prévision de la consommation de gaz naturel en Iran	Combinaison hiérarchique de ANN et AG	La méthode est capable de s'adapter a des données non linéaires et complexes
A. Azadeh et al. (2011)	Dataset de la période 1980-2007	la prévision de la consommation de gaz naturel, en utilisant des données historiques.	une approche combinant ANFIS (Adaptive Network-Based Fuzzy Inference System) et SFA (Stochastic Frontier Analysis)	Le modèle ANFIS est capable de produire des résultats précis en termes d'erreur absolue moyenne (MAPE).
E.F. Sánchez-Úbeda et A. Berzosa (2007)	Dataset obtenues auprès d'Enagás, la période historique couvre du 1er janvier 1996 au 30 juin 2005	prévision pour la consommation de gaz naturel dans les secteurs industriels pour l'utilisation finale	Approche de décomposition des séries temporelles	un nouveau modèle de prédiction qui permet de fournir des prévisions à moyen terme avec une résolution très élevée.

TABLE 3.1 – Tableau comparatif des solutions lus

3.4 Synthèse de comparaison

Notre étude comparative montre une évolution dans les méthodes utilisées pour la prévision de la consommation de gaz naturel entre 2007 et 2022. Les articles publiés en 2007 (E.F. Sánchez-Úbeda et A. Berzosa) [55] et 2011 (A. Azadeh et al.) [25] utilisent des approches plus traditionnelles, telles que la décomposition des séries temporelles et l'ANFIS combiné au SFA. Ces méthodes ont permis

d'obtenir des résultats suffisantes.

Cependant, à mesure que nous avançons dans le temps, les articles ultérieurs adoptent des approches plus avancées basées sur l'apprentissage automatique (ML). Par exemple, les articles de 2014 (M. Salehi et al.)[18] et 2016 (O. Laib et al.) [19] combinent des techniques traditionnelles, comme les réseaux de neurones artificiels (ANN) et les algorithmes génétiques (AG), avec des données plus complexes et non linéaires pour améliorer les prévisions.

Ensuite, les articles de 2019 (O.F. Beyca et al.) [20] et 2020 (M. Akpinar et N. Yumuşak) [28] utilisent la technique d'apprentissage automatique plus avancée , telle que le support vector regression (SVR) et la méthode tradionnelle la plus performante telle que le modèle ARIMA saisonnier, respectivement. Ces approches permettent de prendre en compte des variables multiples et d'obtenir des résultats plus précis.

Enfin, les articles plus récents de 2021 (V. Sharma et al.) [21] et 2022 (W. Panek et T. Włodek) [26] soulignent l'utilisation de combinaisons d'algorithmes de ML pour améliorer encore les prévisions. Ces études mettent en évidence l'importance de combiner des méthodes traditionnelles et des techniques d'apprentissage automatique pour obtenir des prévisions plus précises et efficaces.

Ainsi, les facteurs pris en compte dans ces études pour prévoir la consommation de gaz naturel sont les suivants : les données historiques de consommation, les indicateurs saisonniers ou calendaires, la température, le prix du gaz naturel, la population ou la taille de la zone géographique étudiée, les données météorologiques externes (comme la température), les variables socio-économiques (le cas échéant) et les variables spécifiques aux secteurs industriels (le cas échéant). En combinant ces facteurs, les chercheurs peuvent obtenir des modèles de prévision plus précis pour anticiper la consommation du gaz naturel et faciliter la planification énergétique.

Enfin, on peut observer une évolution progressive des méthodes utilisées, passant des approches traditionnelles à l'utilisation croissante de techniques d'apprentissage automatique, avec une combinaison des deux pour améliorer les performances de prévision de la consommation du gaz naturel. Cette évolution reflète la tendance générale dans le domaine de la prévision.

3.5 Conclusion

Au cours de ce chapitre, Nous avons examiné plusieurs études portant sur la prévision de la consommation du gaz naturel. Ces études ont permis de mettre en évidence différentes méthodes et approches utilisées pour prédire la demande du gaz naturel dans différents contextes, ainsi que la

combinaison de méthodes traditionnelles et de ML pour améliorer la précision des prévisions. Ces avancées contribuent à une meilleure compréhension des tendances de consommation de gaz naturel et à une utilisation plus efficace des ressources énergétiques.

Dans le prochain chapitre, nous aborderons la modélisation et l'evaluation des modèles de prédiction de la consommation du gaz naturel.

CHAPITRE

4

MODÉLISATION DES MÉTHODES SARIMA & LSTM

4.1 Introduction

Dans ce chapitre, nous allons analyser les graphes de consommation de chaque client afin de comparer les performances des méthodes SARIMA(*Seasonal Autoregressive Integrated Moving Average*) et LSTM (*Long Short-Term Memory*) pour la prédiction de la consommation du gaz naturel. Nous commencerons par modéliser chaque méthode pour un seul client, en sélectionnant un client ayant une consommation optimale et des clients ayant une consommation médiocre. Ensuite, nous étendrons notre analyse pour englober l'ensemble des clients afin d'avoir une consommation précise et fiable.

Nous débuterons en présentant l'environnement matériel et logiciel utilisé pour le développement et la simulation de nos modèles. Ensuite, nous comparerons les résultats obtenus en utilisant la méthode LSTM et la méthode SARIMA pour la prédiction de la consommation de gaz d'un seul client. Cette comparaison mettra en évidence les avantages spécifiques de chaque méthode.

Dans la poursuite de notre étude, nous étendrons notre analyse en effectuant des prédictions pour plusieurs clients simultanément, en utilisant à la fois la méthode LSTM et la méthode SARIMA. Nous sélectionnerons les clients qui présentent une meilleure consommation du gaz naturel. Nous évaluerons les performances des deux méthodes et exploiterons les avantages de chacune d'entre elles.

4.2 Environnement de développement

Dans cette section, nous fournirons un aperçu complet de l'environnement dans lequel notre modèle a été développé et exécuté. Nous présenterons les différents aspects matériels et logiciels qui ont contribué à la création de notre modèle de prédiction de la consommation de gaz naturel.

4.2.1 Matériel

Dans notre projet, nous avons utilisé nos propres ordinateurs portables pour le développement du modèle. Voici les spécifications de nos machines :

PC	Système d'exploitation	Modèle du système	Processeur	Mémoire (RAM)
PC 1	Windows 11 Professionnel	HP EliteBook 840 G1	Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz (4 CPUs), 2.5 GHz	12 GB
PC 2	Kali Linux	HP EliteBook 840 G1	Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz (4 CPUs), 2.5 GHz	12 GB
PC 3	Windows 7 Professionnel	HP EliteBook 8440P	Intel(R) Core(TM) i5-M520 CPU @ 1.4GHz (4 CPUs), 2.4 GHz	4 GB

TABLE 4.1 – Spécifications des ordinateurs utilisés

4.2.2 Logiciels

4.2.2.1 Langages de programmation et packages

Python : Python est un langage de programmation orienté objet. Créé et publié en 1991 par Guido van Rossum, il offre une syntaxe propre et une structure d'indentation facile à apprendre et à utiliser.

Python est largement utilisé par les programmeurs expérimentés et débutants en raison de sa simplicité, qui permet de se concentrer sur la résolution de problèmes. Il est considéré comme l'un des langages de programmation les plus efficaces en IA, notamment dans les domaines de l'apprentissage automatique (ML) et de l'apprentissage en profondeur (DL) [54] [34].

Numpy : Numpy est une bibliothèque fondamentale pour le calcul scientifique en Python. Elle fournit des structures de données et des fonctions de calcul avancées qui ne sont pas disponibles dans le package de base de Python.

Numpy permet la manipulation de tableaux multidimensionnels, la gestion de tableaux masqués, les opérations rapides sur les tableaux, ainsi que des fonctionnalités mathématiques, logiques, de manipulation de forme, de tri, de sélection, d'E/S, de transformées de Fourier discrètes, d'algèbre linéaire de base, de statistiques et de simulation aléatoire [29].

Pandas : Pandas est un package Python largement utilisé pour la science des données, l'analyse de données et les tâches d'apprentissage automatique. Il offre des structures de données et des opérations de manipulation de tableaux numériques, ainsi que la manipulation de séries chronologiques. Le concept fondamental de ce package est le DataFrame, une structure de données tabulaire bidimen-

sionnelle avec des étiquettes de ligne et de colonne.

Pandas offres une haute performance pour la manipulation de données dans des tableurs ou des bases de données relationnelles (bases de données SQL). En utilisant l'indexation avancée, il est facile d'effectuer de nombreuses opérations telles que le découpage en tranches, le remodelage, les agrégations et les sélections de sous-ensembles [29].

Matplotlib : Matplotlib est la bibliothèque Python la plus populaire pour la création de graphiques et d'autres visualisations de données en 2D. L'analyse des données nécessite des outils de visualisation, et Matplotlib est particulièrement adapté à cet usage.

Matplotlib permet de contrôler tous les aspects de l'affichage graphique en programmant les éléments graphiques. Les visualisations sont créées avec Matplotlib, puis exportées dans des formats graphiques courants (PNG, SVG, etc.) pour une utilisation dans d'autres applications, de la documentation, des pages web, etc. [29].

TensorFlow : TensorFlow est une bibliothèque open-source développée par l'organisation de recherche "Google's Machine Learning Intelligence" dans le but de faciliter l'apprentissage automatique et la recherche en analyse approfondie des réseaux de neurones.

TensorFlow combine l'algèbre computationnelle, l'optimisation et les techniques de compilation pour permettre le calcul efficace d'expressions mathématiques. Il fournit une API frontale en Python pour la création d'applications utilisant le framework, tout en permettant l'exécution de ces applications avec une haute performance.

TensorFlow prend en charge le déploiement de modèles en production à grande échelle, en utilisant les mêmes modèles utilisés pour l'entraînement [53].

Keras : Keras est une interface de programmation d'applications (API) de Deep Learning en Python qui permet de créer rapidement et facilement des modèles.[49].

Scikit-learn : Scikit-learn est une bibliothèque d'apprentissage automatique open-source écrite en Python qui permet d'intégrer facilement des méthodes d'apprentissage automatique dans le code Python.

Scikit-learn offre une variété de méthodes de classification, de régression, d'estimation matricielle, de réduction de dimensionnalité, de prétraitement des données, de génération de problèmes de benchmarking, etc. Elle est disponible sur la plupart des systèmes d'exploitation et est facile à installer.

La bibliothèque est constamment améliorée, étendue et largement utilisée dans de nombreuses applications commerciales [37].

4.2.2.2 Environnement de Développement Intégré IDE

Anaconda : est une distribution gratuite et open-source de la programmation Python et R langages pour le calcul scientifique (science des données, applications d'apprentissage automatique, traitement de données à grande échelle, analyse prédictive, etc.) [1]. Il peut être facilement installé sur n'importe quel système d'exploitation tel que Windows, Linux et MAC OS. Il fournit plus de 1500 Python/R paquets.

La distribution Anaconda fournit l'installation de Python avec divers IDE tels que Jupyter Notebook, Spyder, Anaconda prompt, etc. C'est donc une solution très pratique que vous pouvez facilement télécharger et installer. Il installera automatiquement Python, quelques IDE de base et bibliothèques avec. Dans notre memoire, nous allons utilisé Jupyter, google colab et Python pour construire notre modèle.

Jupyter Notebook : est une application Web IDE gratuite développée en 2014. Elle permet le partage de code en direct, d'équations, de visualisations et de texte explicatif dans une seule extension de document ".ipynb" [2].

Google colab : abréviation de "Google Colaboratory", est une plateforme en ligne gratuite qui permet d'exécuter du code Python, de créer et d'exécuter des notebooks Jupyter, et de collaborer facilement avec d'autres personnes. Il offre un environnement de développement interactif, hébergé sur le cloud, qui donne accès à des ressources puissantes de calcul, y compris des GPU et des TPU, sans nécessiter d'installation ni de configuration complexes.

4.3 Modélisation de la prédiction de la consommation de Gaz Naturel

L'objectif principal de notre étude est de développer un modèle pour la prédiction de la consommation de gaz naturel pour les clients HP (Haute Pression) en utilisant une combinaison des modèles LSTM et SARIMA. Dans cette étude, nous nous intéressons à ces deux modèles de prévision des séries temporelles : SARIMA, abordé dans la sous-section 3 de la section 1 du chapitre 2, et LSTM, discuté dans la sous-section 4 de la sous-section 7 de la section 3 du chapitre 2.

Nous commencerons par appliquer ces deux modèles de prédiction sur les données d'un seul client, en sélectionnant un client avec une consommation optimale et nous comparerons ensuite les résultats obtenus par chaque modèle et nous mettons en place notre approche basée sur la combinaison des performances des deux modeles.

Ensuite nous allons séléctionner les clients qui reprensentes des consommations midiocres faire une disscussion, les supprimer et de passer a une analyse globale.

Ensuite, nous étendrons notre analyse en utilisant ces modèles pour prédire la consommation de gaz naturel pour le reste des clients représentant des consommations optimales. Nous évaluerons les performances des deux modèles et comparerons leurs résultats, mettant en évidence les avantages de notre approche basée sur la combinaison de ces deux modèles. Cette analyse globale nous permettra d'évaluer l'efficacité et la précision de notre nouvelle approche par rapport aux résultats des deux modèles dans la prédiction de la consommation de gaz naturel pour différentes situations et profils de clients.

Dans la suite de notre étude, nous détaillerons ces deux méthodes de prédiction et présenterons les résultats de nos prédictions pour un seul client, puis pour l'ensemble des clients représentant des consommations optimales, ainsi que notre approche basée sur la combinaison des deux modèles.

4.3.1 Modélisation pour un seul client

Dans le cadre de notre étude sur la consommation de gaz, nous avons réalisé une analyse approfondie de l'ensemble de nos clients. Notre objectif était de comprendre les tendances de consommation de gaz et d'identifier les variations entre les différents clients. Pour ce faire, nous avons examiné les données historiques de consommation de gaz de chaque client et avons obtenu des informations précieuses sur leurs habitudes de consommation.

Lors de notre analyse, nous avons observé des schémas intéressants qui ont mis en évidence des différences significatives dans les comportements de consommation de nos clients. Pour visualiser ces variations, nous avons généré des graphiques qui représentent la consommation de gaz au fil du temps pour différents clients.

FIGURE 4.1 – graphe de consommation de chaque client.

la figure 4.1 montre que pour certains clients leurs schémas de consommation de gaz naturel sont optimaux, tandis que d'autres ont montré des niveaux de consommation moins favorables.

Dans la suite de notre étude, nous allons approfondir notre analyse en nous concentrant sur deux clients spécifiques, le client "3360060HP" qui represente une consommation optimale et les clients "3360055HP", "3360056HP". Ces clients ont été sélectionnés car ils illustrent des exemples contrastés de consommation de gaz, où le premier semble afficher une consommation optimale, tandis que les deux autres présentent une consommation médiocre.

4.3.1.1 Modélisation de la méthode SARIMA

Nous avons utilisé le modèle SARIMA pour prédire la consommation de gaz naturel pour le client '3360060HP'. Le modèle SARIMA est particulièrement adapté pour modéliser des données saisonnières, telles que la consommation de gaz naturel qui peut varier de manière saisonnière. Il permet de capturer les tendances, les saisons et les variations aléatoires dans les données.

1. **Collecte des données :** Les données que nous avons recueillies sont spécifiquement liées à la consommation de gaz naturel des clients haute pression de Sonelgaz. Ces clients sont généralement des entreprises ou des industries qui utilisent du gaz naturel à des fins commerciales ou industrielles.

Ces données ont été récupérées auprès de Sonelgaz dans leur format d'origine, sans avoir été modifiées ni traitées. Cela garantit que les données conservent leur intégrité et leur fidélité aux mesures réelles de consommation de gaz naturel.

Par la suite, ces données brutes seront utilisées comme base de travail pour les étapes ultérieures de notre modèle. Cela comprend notamment le nettoyage des données, l'exploration, l'analyse et la modélisation. Ces étapes nous permettront de fournir des prédictions précises sur la consommation de gaz naturel des clients haute pression de Sonelgaz.

2. **Nettoyage des données :** L'étape de nettoyage des données consiste à préparer notre ensemble de données pour l'analyse ultérieure. Dans notre cas, nous avons commencé avec ensembles de fichiers Excel pour chaque année de 2014 à 2022, avec chaque fichier contenant les mois de l'année et les consommations des clients.

   Pour commencer, nous avons regroupé tous ces fichiers dans un seul fichier qui contient toutes les années et les mois de consommation pour chaque client. Cela nous permet d'avoir une vue d'ensemble et facilite la manipulation des données.

   Ensuite, nous avons trié les dates en ordre croissant pour chaque client, afin de disposer d'une séquence chronologique cohérente de leurs mois de consommation. Cela nous permet d'analyser les tendances et les variations au fil du temps de manière plus facile et précise.

   Ensuite, nous avons procédé à l'élimination des colonnes dont nous n'avons pas besoin, en conservant uniquement les colonnes essentielles telles que "Date" pour les dates, "Valeur" pour les consommations et "Client" pour chaque code client. Cela nous permet de simplifier et de focaliser l'analyse sur les informations pertinentes.

   Une fois que notre ensemble de données était prêt, nous l'avons enregistré pour une utilisation ultérieure dans le processus de modélisation et d'analyse.

   Dans le cadre du nettoyage des données, nous avons rempli les cases vides avec des zéros pour garantir la cohérence des données. Nous avons également éliminé les valeurs aberrantes, qui sont des valeurs atypiques ou potentiellement erronées, afin d'éviter qu'elles n'affectent négativement les résultats de notre analyse.

   Pour faciliter ces opérations de nettoyage et d'analyse des données, nous avons utilisé la bibliothèque pandas, qui offre des fonctionnalités puissantes pour la manipulation, le filtrage et la visualisation des données tabulaires.

3. **Découpage des données :** Dans cette étape, nous avons découpé les données en utilisant la méthode dataframe.iloc() de la bibliothèque pandas, en deux parties distinctes :

   nous avons réalisé une découpe des données en un ensemble d'entraînement (90% des données) et un ensemble de test (10% des données). Cette séparation nous permet de former le modèle sur les données d'entraînement et de le tester sur les données de test pour évaluer ses performances de prédiction.

4. **Modélisation :** Le processus de construction du modèle SARIMA peut être divisé en plusieurs étapes. Voici comment nous avons procédé :

   Tout d'abord, nous avons observé les composantes de la série temporelle en utilisant la bibliothèque statsmodels.tsa.seasonal et la fonction seasonal_decompose. Cette étape nous a permis d'avoir une vision globale de la chronique que nous souhaitions étudier et modéliser. En analysant le graphique, nous avons pu identifier les variations saisonnières, les tendances et les résidus de la série temporelle.

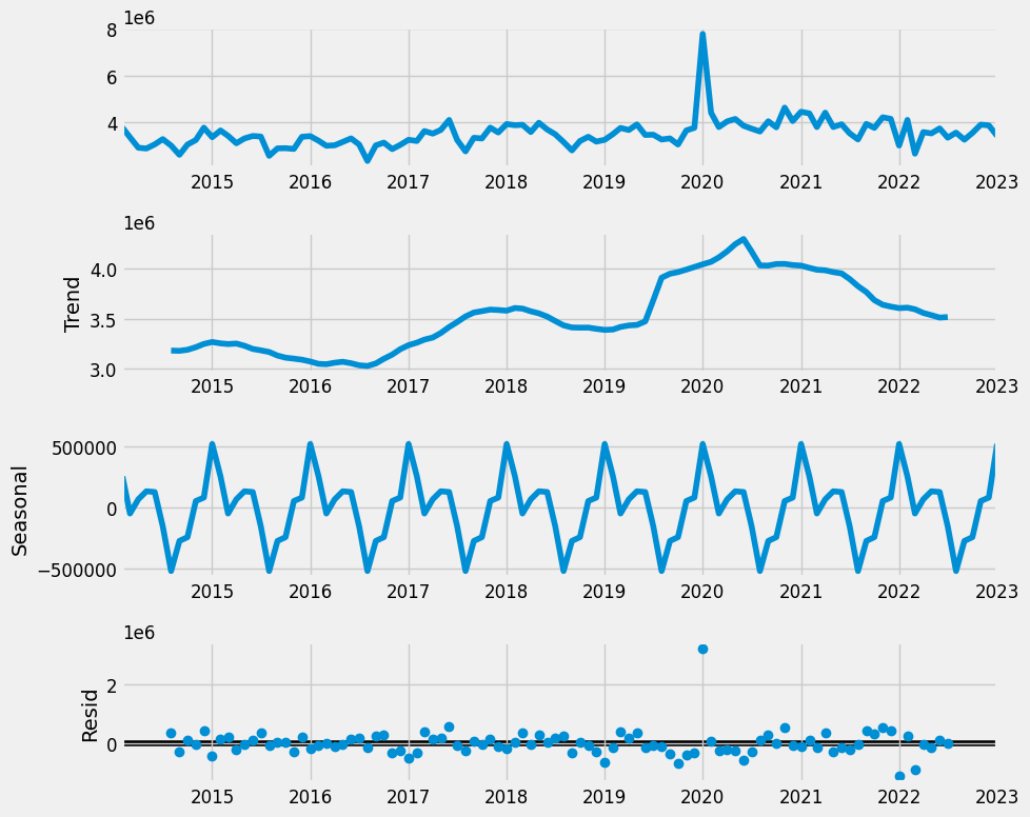


FIGURE 4.2 – Les composantes de la série temporelle du client 3360060HP

Ensuite, nous avons testé la stationnarité de la série temporelle en utilisant le test de Dickey-Fuller augmenté (ADF) à l'aide de la fonction " adfuller " de la bibliothèque " statsmodels.tsa.stattools".

Ce test statistique nous a permis d'évaluer si la série donnée était stationnaire ou non.

- les resultats du test ADF :
  - ADF : -4.509353817537911
  - P-Value : 0.0001891272584280093
  - Nombre de retards (Num Of Lags) : 1
  - Nombre d'observations utilisées pour la régression ADF et le calcul des valeurs critiques : 106
  - Valeurs critiques :
    - 1% : -3.4936021509366793
    - 5% : -2.8892174239808703
    - 10% : -2.58153320754717

Les résultats du test de stationnarité ADF suggèrent fortement que notre série temporelle est stationnaire.

ADF (Augmente Dickey-Fuller) : C'est la statistique du test ADF. Cette statistique est utilisée pour comparer aux valeurs critiques et déterminer si la série est stationnaire. Si la statistique est inférieure aux valeurs critiques, cela suggère que la série est stationnaire.

P-Value : C'est la valeur de p associée au test ADF. La valeur de p représente la probabilité d'obtenir des résultats aussi extrêmes que ceux observés dans les données, sous l'hypothèse nulle que la série est non stationnaire. Une valeur de p inférieure à un certain seuil (généralement 0.05) indique que l'hypothèse nulle peut être rejetée et que la série est stationnaire.

Nombre de retards (Num Of Lags) : C'est le nombre de retards ou de différences inclus dans le modèle ADF.

Nombre d'observations utilisées pour la régression ADF et le calcul des valeurs critiques : C'est le nombre d'observations de la série utilisées pour effectuer le test ADF et calculer les valeurs critiques.

La statistique ADF est fortement négative, la valeur de p est très faible et inférieure à 0,05, et la statistique ADF est inférieure aux valeurs critiques à différents niveaux de signification statistique. Cela indique que notre série temporelle a une tendance et une saisonnalité bien ajustées, et que le modèle SARIMA est approprié pour modéliser ces données. Les résultats du test renforcent l'idée que le modèle SARIMA capturerait efficacement les structures temporelles dans nos données.

La méthode de recherche de grille a été utilisée pour sélectionner les paramètres optimaux pour notre modèle SARIMA. Nous avons généré toutes les combinaisons possibles de paramètres AR, différenciation et MA, ainsi que les paramètres saisonniers, en utilisant la fonction itertools.product().

En ajustant un modèle SARIMA à chaque combinaison de paramètres, nous avons calculé l'AIC pour évaluer la qualité de l'ajustement. L'AIC tient compte de la complexité du modèle et nous permet de comparer les différentes combinaisons de paramètres. Nous avons sélectionné les paramètres qui produisaient le modèle avec le plus faible AIC, ce qui indique le meilleur ajustement aux données.

En utilisant les paramètres sélectionnés, nous avons ajusté un dernier modèle SARIMA qui a été utilisé pour les prédictions et l'analyse ultérieure de nos séries temporelles.

5. **Evaluation :** Dans le processus d'évaluation de la performance de notre modèle, il est crucial de vérifier que les résidus ne présentent pas de corrélation et suivent une distribution normale. Pour ce faire, nous avons utilisé la fonction plot_diagnostics() de la bibliothèque matplotlib, qui nous permet de générer facilement des diagnostics du modèle et d'identifier tout comportement anormal.

   Cette fonction nous a fourni les diagnostics présentés dans la figure 4.3. Ces diagnostics comprennent :

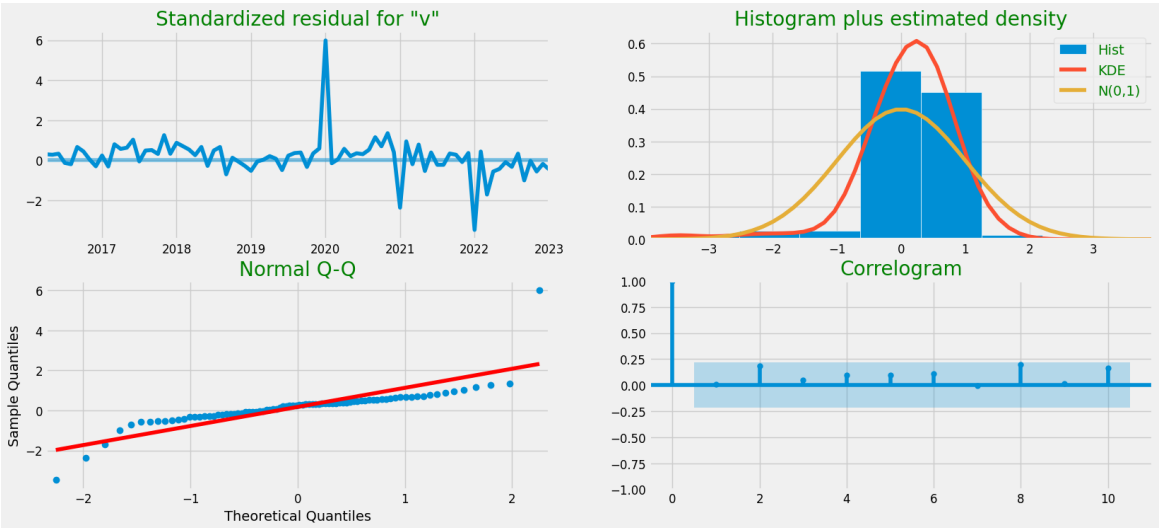


FIGURE 4.3 – Les diagnostics du modèle SARIMA

Comme nous pouvons le constater dans cette figure Matplotlib nous affiche quatre différents graphs, soit :

- **histogamme plus estimated density :** la densité estimée est une estimation de densité de probabilité des résidu.

  Elle estime la probabilité que les résidus prennent une certaine valeur. Si la densité estimée est proche d'une distribution normale, cela indique que le modèle est bon.

  Dans ce cas, la densité estimée semble être proche d'une distribution normale, avec une forme en cloche et une symétrie autour de zéro. Cela suggère que les résidus sont distribués normalement, ce qui est une bonne indication que le modèle est bien ajusté.

- **Graphique Q-Q :** apellé Quantile-Quantile, le graphe Normal Q-Q montre que les résidus semblent suivre une distribution normale. Les points sur le graphe sont alignés sur une ligne droite, ce qui suggère que les résidus sont distribués normalement.

- **Standardized residual :** Le graphe des résidus standardisés montre que les résidus du modèle SARIMA semblent être aléatoires et non corrélés. Les points sur le graphe sont répartis de manière aléatoire autour de zéro, ce qui suggère que les résidus sont aléatoires. De plus, il n'y a pas de tendance claire ou de structure dans les résidus, ce qui indique qu'ils ne sont pas corrélés.

  Cela est confirmé par les autres graphes tels que l'histogramme de densité estimée et le graphe Q-Q, qui montrent également que les résidus sont distribués normalement et aléatoires. Ces résultats suggèrent que le modèle SARIMA est bien ajusté et fournit des prévisions précises.

- **Correlogram :** Le corrélogramme est un outil statistique utilisé pour évaluer la corrélation entre les résidus à différents retards.

  Le corrélogramme montre que les résidus du modèle SARIMA sont aléatoires et non corrélés. Les barres sur le graphe sont toutes situées dans la zone bleue, qui représente la zone de confiance à 95%. Cela suggère que les résidus ne sont pas corrélés à différents retards et qu'ils sont aléatoires.

Ces observations nous amènent à conclure que notre modèle produit un ajustement satisfaisant qui pourrait nous aider à comprendre nos données de séries chronologiqueq et à prévoir les valeurs futures.

6. **Resultats :**

Dans cette section, nous allons discuter les résultats de notre prédiction de la consommation de

gaz pour l'année 2022 en analysant le graphe et en comparant les valeurs prédites aux valeurs réelles fournies par notre client.

Le graphe de la consommation de gaz pour l'année 2022, basé sur nos prédictions, est présenté dans la Figure 4.4.

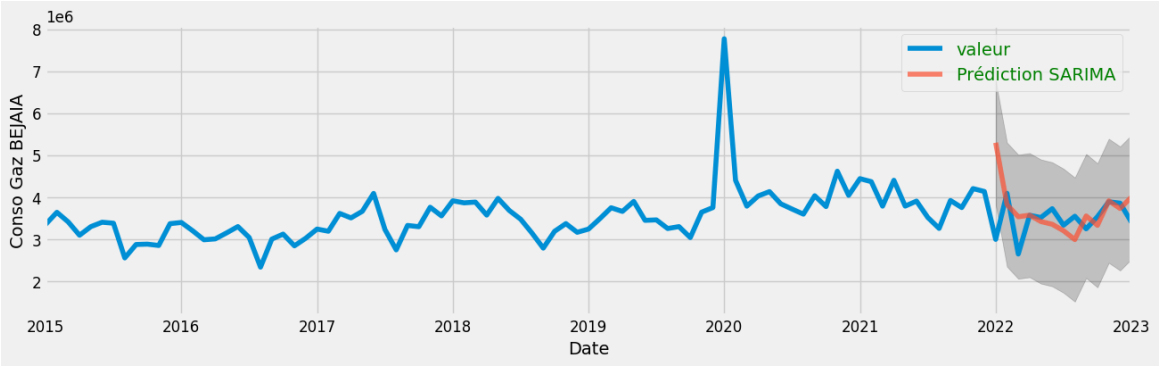


FIGURE 4.4 – Prédiction de la consommation de gaz pour l'année 2022 client '3360060HP'.

Ce graphique représente la prédiction de la consommation de gaz naturel pour l'année 2022. Comme nous l'avons déjà mentionné, nous avons divisé nos données en deux ensembles : un ensemble d'entraînement représenté en bleu et un ensemble de test représenté en orange.

En analysant cette figure, il est possible de constater que les valeurs prédites suivent une tendance presque similaire à celle des valeurs réelles. Cela suggère que la méthode SARIMA utilisée n'a pas été aussi efficace qu'espéré pour prédire la consommation de gaz naturel à Bejaia.

Dans ce qui suit, nous allons discuter les résultats de notre prédiction de la consommation de gaz pour l'année 2022 en analysant le graphe et en comparant les valeurs prédites aux valeurs réelles fournies par notre client.

Le tableau 4.2 présente les valeurs réelles, les valeurs prédites et le pourcentage d'erreur absolu moyen (MAPE) pour la prédiction de la consommation de gaz pour l'année 2022.

Date	Réel	Prédit	MAPE%
2022-01-01	2996640	5299400.427	76.84474703
2022-02-01	4094442	3825210.532	6.575535032
2022-03-01	2648383	3536290.691	33.52640804
2022-04-01	3571148	3573619.732	0.069213921
2022-05-01	3517849	3422157.803	2.720162151
2022-06-01	3728992	3356528.141	9.988325506
2022-07-01	3334771	3203053.307	3.949827226
2022-08-01	3548435	2992235.274	15.67450793
2022-09-01	3248331	3554083.095	9.412590511
2022-10-01	3538105	3332055.362	5.823728737
2022-11-01	3892250	3915790.919	0.604815182
2022-12-01	3859602	3733342.705	3.271303504
	Moyenne MAPE%	14.0384304	

TABLE 4.2 – Résultats de la prédiction de la consommation de gaz pour l'année 2022 avec SARIMA

Les résultats présentés dans le tableau 4.2 sont basés sur les données du client '3360060HP' concernant la consommation de gaz naturel. Nous avons effectué des prédictions pour l'année 2022 et comparé les valeurs prédites aux valeurs réelles.

En examinant les résultats, nous constatons que les valeurs prédites et les valeurs réelles varient pour chaque mois de l'année. Le pourcentage d'erreur absolu moyen (MAPE) est utilisé comme mesure de précision de nos prédictions.

En janvier 2022, la valeur réelle de consommation de gaz pour le client '3360060HP' était de 2 996 640 thermie. Notre prédiction était de 5 299 400,427 thermie, ce qui correspond à une erreur relative de 76,84% (MAPE). Cette erreur élevée peut s'expliquer par divers facteurs tels que des variations saisonnières, des fluctuations du marché du gaz ou des comportements de consommation différents de ceux initialement prévus.

En février 2022, la valeur réelle de consommation de gaz était de 4 094 442 thermie. Notre prédiction était de 3 825 210,532 thermie, avec un MAPE de 6,58%. Cette prédiction montre une précision relativement élevée, indiquant que notre modèle a réussi à capturer correctement la tendance de consommation pour ce mois.

En mars 2022, la valeur réelle était de 2 648 383 thermie, tandis que notre prédiction était de 3

536 290,691 thermie, avec un MAPE de 33,53%. Cette différence importante peut être attribuée à des facteurs imprévus tels que des changements dans les habitudes de consommation ou des événements spécifiques qui ont influencé la consommation de gaz pour ce mois.

En poursuivant l'analyse, nous constatons que les valeurs prédites suivent généralement une tendance similaire à celle des valeurs réelles. Cependant, des variations peuvent survenir d'un mois à l'autre en raison de divers facteurs externes.

En moyenne, le MAPE pour l'ensemble de l'année 2022 pour le client '3360060HP' est de 14,04%. Cela signifie que nos prédictions ont une précision moyenne de 85,96%.

Il convient de souligner que la prédiction de la consommation de gaz naturel est un processus complexe qui dépend de plusieurs facteurs.

En conclusion, les résultats de la prédiction de la consommation de gaz naturel pour notre client montrent une variabilité dans la précision de nos prédictions pour chaque mois de l'année 2022. Malgré des écarts importants dans certaines périodes, notre modèle a réussi à capturer la tendance générale de consommation de gaz pour ce client.

4.3.1.2 Modélisation de la méthode LSTM

Dans le cadre de notre étude, nous avons entrepris de d'utiliser un modèle LSTM (Long Short-Term Memory) pour prédire la consommation de gaz naturel du client '3360060HP'.

Le modèle LSTM est une technique d'apprentissage profond utilisée pour modéliser les séries temporelles et capturer les motifs complexes et les dépendances à long terme. En utilisant cette approche, nous souhaitons fournir des prédictions précises de la consommation de gaz pour ce client.

Dans la suite de notre étude, nous détaillerons toutes les étapes que nous avons suivies pour la modélisation de ce processus.

Remarque :les deux premières étapes Collecte et Nettoyage des données sont les mêmes que celles cité dans la sous section précédente.

3. **Découpage des données :** Nous avons découpé le jeu de données en utilisant la bibliothèque Pandas, en séparant les données en deux parties :

    - Ensemble d'entraînement : Nous avons utilisé les données de l'année 2014 jusqu'au 1er janvier 2022 pour former notre modèle.

- Ensemble de test : Les données de test comprennent la période de 2021 et 2022. Nous avons utilisé l'année 2022 pour faire le test et l'année 2021 qui est indispensable pour prédire 2022.

Pour réaliser cette découpe, nous avons utilisé la méthode dataframe.loc() de Pandas, qui nous permet de sélectionner les données en fonction de critères temporels spécifiques. Ainsi, nous avons pu séparer les données d'entraînement et de test en fonction des années correspondantes.

Cela nous permet de disposer d'un ensemble d'entraînement solide pour former notre modèle LSTM et d'un ensemble de test pour évaluer sa capacité à prédire avec les données futures.

4. **Supervision des données :** Pour la modélisation, le processus utilise la méthode de la fenêtre glissante (sliding window) pour superviser à la fois les données d'entraînement et les données de test. Cette méthode nous permet de prédire le 13ème mois à partir de chaque ensemble de 12 mois consécutifs.

   Pour illustrer cette approche, Nous avons commencé par trier les données par client et par date afin de garantir un ordre croissant. Ensuite, nous avons parcouru chaque client dans l'ordre croissant.

   Pour chaque client, nous avons extrait les données correspondantes et les avons triées par date. Ensuite, nous avons utilisé la fonction ts_to_supervised_overlapping pour créer des paires d'entrées et de sorties supervisées en utilisant une fenêtre de taille 12 mois (W = 12) et un pas de 1 mois (H = 1). Cela signifie que pour chaque fenêtre de 12 mois, nous avons prédit la consommation de gaz pour le 13ème mois.

   Nous avons également créé un DataFrame supplémentaire pour stocker les informations sur les clients correspondant à chaque exemple supervisé. Cela nous permettra de garder une trace de l'association entre les données d'entrée/sortie et les clients respectifs.

   Enfin, nous avons réinitialisé les index des DataFrames pour assurer une continuité dans les données.

   Cette approche de supervision des données avec la méthode de la fenêtre glissante nous permet de capturer les schémas et les tendances temporelles dans les données, en utilisant les informations historiques pour effectuer des prédictions précises.

   Remarque : Bien que nous ayons utilisé la méthode de la fenêtre glissante pour tous les clients

dans cet exemple, le processus est similaire pour modéliser un seul client. La seule différence réside dans les données spécifiques à chaque client utilisées pour la modélisation et la prédiction.

5. **Normalisation des données :** Nous avons effectué une normalisation des données pour garantir une modélisation précise et faciliter l'apprentissage des neurones de notre réseau. Pour cela, nous avons utilisé la fonction "MinMaxScaler" de la bibliothèque sklearn. Cette normalisation a été appliquée séparément aux données d'entraînement et aux données de test.

   Pour les données d'entraînement, nous avons extrait les caractéristiques d'entrée "X" et les valeurs cibles "Y" à partir du DataFrame "tsTrain". Ensuite, nous avons utilisé l'objet "MinMaxScaler" pour normaliser les caractéristiques d'entrée "X" et les valeurs cibles "Y" dans la plage (0, 1).

   De même, pour les données de test, nous avons appliqué la même procédure en utilisant les variables correspondantes. Les caractéristiques d'entrée normalisées pour les données de test ont été stockées dans la variable "xtest".

   Cette normalisation des données assure que toutes les variables sont comparables et contribue à l'efficacité de l'apprentissage de notre modèle.

6. **Modélisation** La figure suivante présente la conception de notre modèle LSTM :

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 12, 512)	1052672
dropout (Dropout)	(None, 12, 512)	0
lstm_1 (LSTM)	(None, 12, 360)	1257120
dropout_1 (Dropout)	(None, 12, 360)	0
lstm_2 (LSTM)	(None, 224)	524160
dropout_2 (Dropout)	(None, 224)	0
dense (Dense)	(None, 1)	225

FIGURE 4.5 – Modélisation du modèle LSTM

On a construit notre modèle LSTM en Utilisant les trois approches suivants :

(a) **Concèption du modèle :**

- L'API Keras Sequential est utilisée pour construire le modèle LSTM.
- Le modèle est composé de trois couches LSTM.
- La première couche LSTM a 512 unités, chacune prenant 12 valeurs en entrée et renvoyant une séquence de sortie.
- Une couche Dropout avec un taux de 0,3 est ajoutée après chaque couche LSTM pour désactiver 20% des unités et éviter le surajustement.
- La deuxième couche LSTM a 360 unités et utilise la fonction d'activation tangente hyperbolique (tanh).
- La troisième couche LSTM a 224 unités et utilise également la fonction d'activation tangente hyperbolique.
- Une couche Dense est ajoutée à la fin du modèle avec un nombre d'unités défini par la variable H=1.

(b) **Configuration du modèle :**

- L'optimiseur Adam est utilisé avec la bibliothèque Keras pour implémenter la descente de gradient et minimiser la fonction de perte.
- Les fonctions de perte utilisées sont l'erreur quadratique moyenne (MSE - mean squared error) inclut à la fois l'exactitude (accuracy).

(c) **Entraînement du modèle :**

- Le modèle est entraîné en utilisant les hyperparamètres suivants :
  - nombre d'échantillion : nous avons utilisé notre modèle pour la prédiction de la consommation pour un seul client.
  - Taille des lots (batch size) : 5 batchs.
  - Nombre d'époques : 510 epochs.
- Une époque correspond à une itération complète sur l'ensemble des données d'apprentissage.

7. **Evaluation :** Nous avons utilisé les métriques d'exactitude (*accuracy*) et d'erreur quadratique moyenne (*MSE*) pour évaluer la performance de notre modèle, ainsi que l'erreur absolue moyenne en pourcentage (*MAPE*) pour évaluer les résultats obtenus après l'entraînement.

MAPE : La métrique d'évaluation MAPE (Mean Absolute Percentage Error) est une mesure couramment utilisée pour évaluer la précision des prédictions en pourcentage pour les modèles de prévision. Elle mesure l'erreur relative moyenne entre les valeurs prédites et les valeurs réelles, exprimée en pourcentage.

L'équation de la MAPE est la suivante :

$$\text{MAPE} = \left(\frac{1}{n}\right) \sum \left| \frac{Y_i - P_i}{Y_i} \right| \times 100$$

où :

— MAPE est le Mean Absolute Percentage Error,

— n est le nombre total d'observations,

— Y_i représente la valeur réelle de l'observation i,

— P_i représente la valeur prédite pour l'observation i.

La MAPE mesure l'erreur relative moyenne entre les valeurs prédites et les valeurs réelles, en pourcentage. Elle donne une indication de la précision moyenne du modèle par rapport aux valeurs réelles.

La MAPE est souvent utilisée pour évaluer les modèles de prévision dans le domaine de la consommation de gaz naturel en raison de ses avantages. Voici quelques raisons pour lesquelles la MAPE pourrait être choisie pour évaluer un modèle de prédiction de la consommation de gaz naturel :

— Sensibilité aux erreurs relatives : La MAPE donne une mesure de l'erreur relative moyenne, ce qui est important pour évaluer la précision des prédictions dans un contexte de consommation de gaz où les variations relatives sont plus significatives que les erreurs absolues.

— Interprétation intuitive : La MAPE est exprimée en pourcentage, ce qui facilite l'interprétation des résultats et permet de comparer les performances des modèles sur différentes échelles de données.

— Évaluation globale : La MAPE calcule une mesure globale de l'erreur relative moyenne pour toutes les observations, fournissant ainsi une vision d'ensemble de la précision du modèle.

Exactitude (*accuracy*) : Une métrique de performance qui se calcule en divisant les résultats de prédiction sur les données réelles.

Erreur quadratique moyenne (*MSE*) : C'est la fonction de perte la plus couramment utilisée pour la régression. La perte est la moyenne des données supervisées des différences au carré entre les valeurs réelles et prédites, définie comme suit :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_{\text{réel}_i} - y_{\text{prédit}_i})^2 \tag{4.1}$$

où :

n : taille de l'échantillon, qui est le nombre de clients.

$y_{\text{réel}_i}$: la valeur réelle des données pour le client i.

$y_{\text{prédit}_i}$: la valeur prédite des données pour le client i.

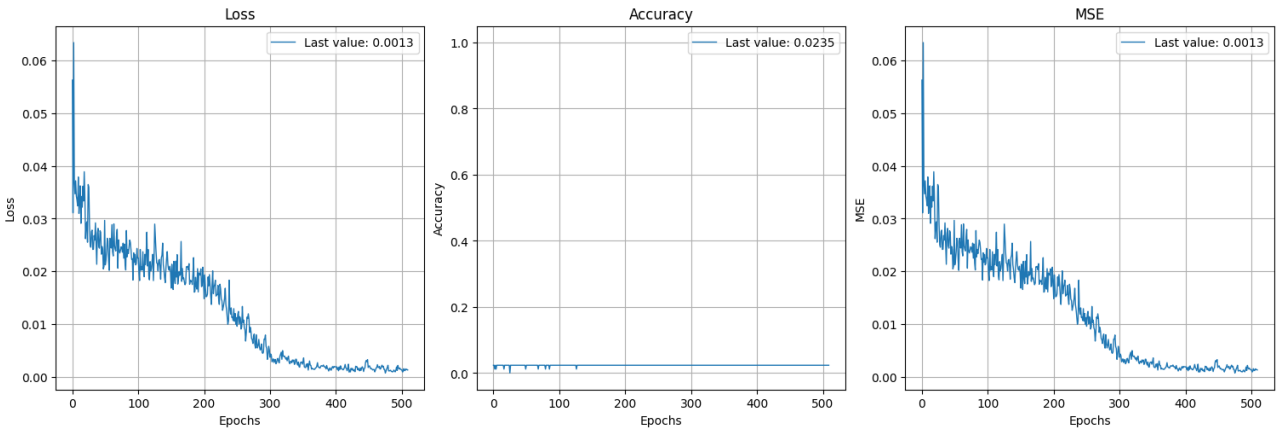


FIGURE 4.6 – graphe d'entrainement de notre modèle

Dans ce graphe on voie que notre modèle a obtenu une MSE de 0.0013 et une accuracy de 0.0235, ce qui indique une très faible erreur quadratique moyenne et une exactitude relativement élevée.

Dans l'ensemble, avec une MSE aussi faible, nous pouvons conclure que notre modèle LSTM a réussi à capturer les motifs et les relations dans les données d'entraînement. Cela signifie que les prédictions de notre modèle sont très proches des valeurs réelles des données, ce qui indique une bonne qualité de prédiction.

Cependant, il est important de noter que l'analyse du graphe d'entraînement ne fournit qu'une vision partielle de la performance du modèle. Pour obtenir une évaluation plus complète, il est nécessaire d'évaluer les résultats sur des données de test ou de validation. Cela nous permettra

de déterminer si notre modèle est capable de généraliser ses prédictions aux nouvelles données, en dehors de l'ensemble d'entraînement.

dans ce qui suit nous alons pêncher sur une analyse des données de test.

8. **Resultat :** Dans cette section nous allons discuter les resultats de notre prédiction de la consommation de gaz naturel pour les mois de l'année 2022 en analysant le graphe et en comparant les valeurs prédites par notre modèle LSTM aux valeurs réelles fournie par notre client .

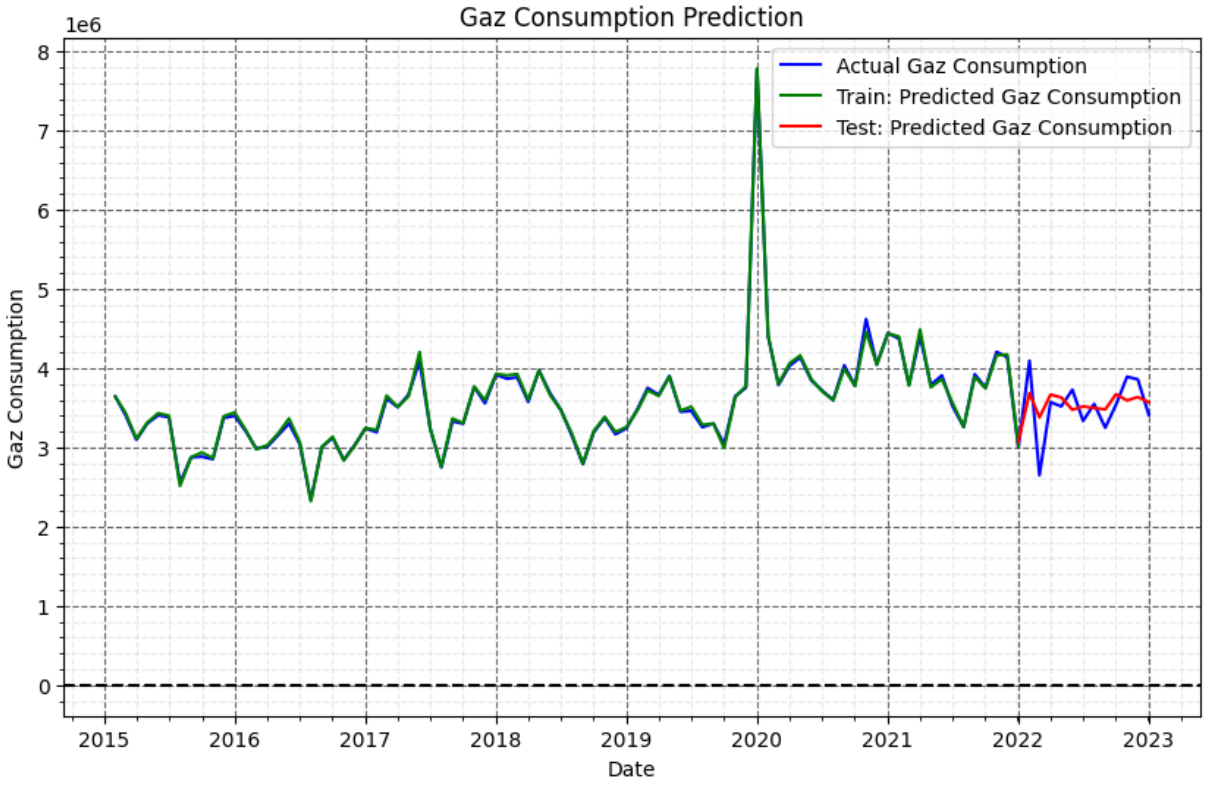


FIGURE 4.7 – prediction de la consommation de gaz par le client '3360060HP' LSTM

La Figure présente les prédictions de la consommation de gaz naturel pour le client '3360060HP' en utilisant le modèle LSTM. Le graphe montre la consommation de gaz naturel réelle en bleu ensuite Les données sont divisées en deux ensembles : un ensemble d'entraînement (représenté en vert) et un ensemble de test (représenté en rouge).

L'analyse de la figure montre que les prédictions du modèle LSTM sont assez proches des valeurs réelles de la consommation de gaz naturel pour les deux ensembles de données (train et test). Cela suggère que le modèle LSTM est capable de capturer les tendances et les modèles de la consommation de gaz naturel pour ce client spécifique, à la fois pour les données d'entraînement et de test.

Dans ce qui suit, nous allons discuter les résultats de notre prédiction de la consommation de gaz pour l'année 2022 en analysant le le tableau et en comparant les valeurs prédites aux valeurs réelles fournies par notre client. Le tableau 4.3 présente les valeurs réelles, les valeurs prédites et le pourcentage d'erreur absolu moyen (MAPE) pour la prédiction de la consommation de gaz pour l'année 2022.

Date	Réel	Prédit	Mois	Année	MAPE%
2022-01-01	2996640	3072259	1	2022	2,523459608
2022-02-01	4094442	3687209,75	2	2022	9,945976766
2022-03-01	2648383	3378731	3	2022	27,57712914
2022-04-01	3571148	3666615	4	2022	2,673286013
2022-05-01	3517849	3630315,5	5	2022	3,197024659
2022-06-01	3728992	3474500	6	2022	6,824686135
2022-07-01	3334771	3517518	7	2022	5,480046456
2022-08-01	3548435	3497830,5	8	2022	1,426107566
2022-09-01	3248331	3479692,5	9	2022	7,122473048
2022-10-01	3538105	3670664,75	10	2022	3,746631318
2022-11-01	3892250	3592767,5	11	2022	7,694328473
2022-12-01	3859602	3636266,5	12	2022	5,786490421
		Moyenne MAPE%	6,9998033		

TABLE 4.3 – la prédiction de la consommation de gaz pour l'annee 2022 avec LSTM un client

Les résultats présentés dans le tableau 4.3 sont basés sur les données du client '3360060HP' concernant la consommation de gaz naturel. Nous avons effectué des prédictions pour l'année 2022 et comparé les valeurs prédites aux valeurs réelles.

En examinant les résultats, nous constatons que les valeurs prédites et les valeurs réelles varient pour chaque mois de l'année. Le pourcentage d'erreur absolu moyen (MAPE) est utilisé comme mesure de précision de nos prédictions.

En janvier 2022, la valeur réelle de consommation de gaz pour le client '3360060HP' était de 2 996 640 thermie. Notre prédiction était de 3 072 259 thermie, ce qui correspond à une erreur relative de 2,52% (MAPE). Cette erreur relativement faible indique une bonne précision de notre modèle pour ce mois.

En février 2022, la valeur réelle de consommation de gaz était de 4 094 442 thermie. Notre

prédiction était de 3 687 209,75 thermie, avec un MAPE de 9,95%. Bien que l'erreur soit un peu plus élevée que pour janvier, la prédiction reste relativement précise.

En mars 2022, la valeur réelle était de 2 648 383 thermie, tandis que notre prédiction était de 3 378 731 thermie, avec un MAPE de 27,58%. Cette différence importante peut être attribuée à des facteurs imprévus qui ont influencé la consommation de gaz pour ce mois.

En poursuivant l'analyse, nous constatons que les valeurs prédites suivent généralement une tendance similaire à celle des valeurs réelles. Cependant, des variations peuvent survenir d'un mois à l'autre en raison de divers facteurs externes.

En moyenne, le MAPE pour l'ensemble de l'année 2022 pour le client '3360060HP' est de 6,99%. Cela signifie que nos prédictions ont une précision moyenne de 93,01%.

En conclusion, notre modèle présente une précision globalement bonne, mais des améliorations sont encore possibles pour une meilleure estimation de la consommation de gaz naturel.

4.3.1.3 Comparaison entre les deux modèles :

La comparaison entre le modèle SARIMA et le modèle LSTM révèle des différences significatives dans leurs performances de prédiction pour la consommation de gaz du client '3360060HP' en 2022.

la figure 4.8 montre une représentation graphique des MAPE%

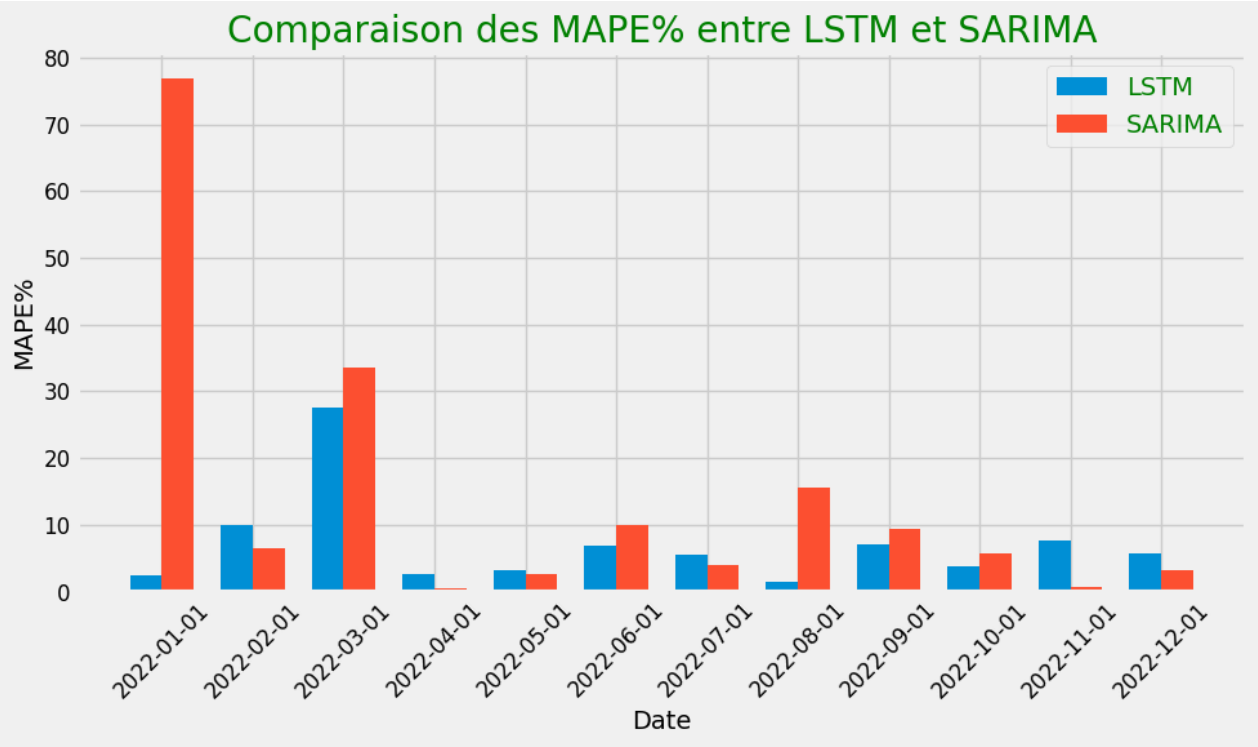


FIGURE 4.8 – comparaison LSTM & SARIMA selons le MAPE%

Le graphe représente la comparaison des valeurs de l'erreur de pourcentage moyen absolu (MAPE) entre les modèles LSTM et SARIMA pour chaque mois de l'année 2022. Les barres bleues représentent les valeurs MAPE% du modèle LSTM, tandis que les barres oranges représentent les valeurs MAPE% du modèle SARIMA.

En observant les hauteurs des barres pour chaque mois, pour les mois de janvier, mars, avril, mai, juillet, août, septembre et novembre, Les barres correspondantes au modèle LSTM seraient relativement hautes, ce qui indique une faible erreur. Cela signifie que le modèle LSTM a réussi à prédire les données de manière précise pour ces mois.

En revanche, pour les mois de février, juin, octobre et décembre,les barres correspondantes au modèle SARIMA seraient également relativement hautes, ce qui indique une faible erreur. Cela suggère que le modèle SARIMA a été plus précis pour la prédiction des données pendant ces mois.

On peut constater alors que les deux modèles, LSTM et SARIMA, sont capables de fournir des prédictions précises, mais leurs performances varient selon les mois. Il est important de prendre en compte les caractéristiques spécifiques des données et les variations saisonnières pour choisir le modèle le plus adapté à chaque mois.

Le tableau ci-dessous représente les valeurs du MAPE% pour chaque mois ainsi que la moyenne MAPE% pour les deux modèles :

Mois	MAPE% LSTM	MAPE% SARIMA
2022-01-01	2.52	76.84
2022-02-01	9.95	6.58
2022-03-01	27.58	33.53
2022-04-01	2.67	0.47
2022-05-01	3.20	2.72
2022-06-01	6.82	9.99
2022-07-01	5.48	3.95
2022-08-01	1.43	15.67
2022-09-01	7.12	9.41
2022-10-01	3.75	5.82
2022-11-01	7.69	0.60
2022-12-01	5.79	3.27
Moyenne MAPE%	6.99	14.85

TABLE 4.4 – Comparaison des MAPE% entre LSTM et SARIMA

Ce tableau présente les valeurs de MAPE% pour LSTM et SARIMA pour chaque mois de l'année 2022. On Observe que les mois de janvier, mars, juin, août, septembre et Octobre présentent une faible erreur MAPE% lorsqu'un modèle LSTM est utilisé, tandis que les mois de février, Avril, Mai, Juillet, Novembre et décembre montrent une faible erreur MAPE% avec le modèle SARIMA. Ces résultats mettent en évidence les avantages et les performances spécifiques de chaque modèle pour la prédiction mensuelle pour la consommation de gaz naturel en 2022.

En tenant compte de ces variations de précision entre les deux modèles, nous avons envisager une approche hybride qui combine les avantages des deux modèles. En utilisant les résultats de notre analyse, nous pourrions développer une approche qui sélectionne automatiquement le modèle le plus performant pour chaque mois spécifique.

Cela pourrait être réalisé en utilisant un mécanisme de pondération où les mois favorables au modèle LSTM obtiennent un poids plus élevé, tandis que les mois favorables à SARIMA obtiennent un poids plus élevé. Cette approche permettrait d'exploiter les forces de chaque modèle dans une approche intégrée, améliorant ainsi la précision des prédictions globales.

Cependant, il convient de noter que la création d'une telle approche hybride nécessiterait une recherche et une expérimentation approfondies pour trouver les meilleurs poids et paramètres qui

maximisent la précision globale du modèle combiné.

4.4 Conclusion

Dans ce chapitre de modélisation , nous avons abordé différents aspects liés à la prédiction de la consommation du gaz naturel. Nous avons d'abord présenté l'environnement de développement utilisé, y compris le matériel, les logiciels et l'environnement de développement intégré (IDE).

Ensuite, nous avons procédé à la modélisation de la consommation de gaz naturel en utilisant deux approches : SARIMA et LSTM. Nous avons évalué chaque modèle individuellement en utilisant des mesures d'évaluation telles que l'erreur quadratique moyenne (MSE) et le pourcentage d'erreur absolue moyenne (MAPE).

Enfin, nous avons comparé les performances des modèles SARIMA et LSTM pour la prédiction de la consommation de gaz naturel pour un seul client. Des analyses supplémentaires sont réalisées sur plusieurs clients individuels avec une consommation midiocre afin de modéliser la consommation pour reste des clients qui représentent une consommation de gaz moyenne. Nous avons constaté que chaque modèle avait ses propres forces et faiblesses, et qu'ils étaient plus performants dans des situations différentes. Parfois, le modèle LSTM était plus précis, tandis que d'autres fois, c'était le modèle SARIMA.

Dans ce qui suit, nous allons détailler notre approche pour la création d'un modèle hybride qui combine les avantages du modèle LSTM et SARIMA.

CHAPITRE

5

APPROCHE PROPOSÉ

5.1 Introduction

Dans ce dernier chapitre, nous présenterons en détail notre approche (*Ensemble_LSTM_SARIMA*) et expliquerons la méthodologie utilisée pour combiner les prédictions des deux modèles. Nous discuterons également des avantages de cette approche combinée par rapport aux méthodes individuelles, en mettant en évidence sa capacité à capturer à la fois les tendances saisonnières et les modèles non linéaires complexes.

Nous évaluerons également les performances de notre approche en utilisant des données réelles de consommation de gaz naturel, en comparant les résultats avec ceux obtenus en utilisant uniquement le modèle LSTM ou le modèle SARIMA.

Grâce à ces comparaisons détaillées, nous démontrerons l'efficacité et la pertinence de notre approche combinée (*Ensemble_LSTM_SARIMA*) pour une prédiction précise de la consommation de gaz naturel. Cette recherche ouvrira de nouvelles perspectives pour l'optimisation de la gestion énergétique dans ce domaine, en offrant une approche plus robuste et complète qui tire parti des avantages des deux méthodes.

5.2 Définition de l'approche

L'agrégation pondérée est une technique de l'ensemble learning qui consiste à combiner les prédictions de plusieurs modèles en attribuant des poids à chacun d'entre eux. Cette méthode vise à améliorer la performance globale en exploitant les forces de chaque modèle et en réduisant l'influence des modèles moins performants. Les poids assignés à chaque modèle déterminent leur contribution relative dans la prédiction finale, ce qui permet d'obtenir une estimation plus précise et fiable. L'agrégation pondérée est souvent utilisée pour combiner des modèles diversifiés et complémentaires, et les poids peuvent être fixés manuellement ou optimisés par des techniques d'optimisation pour obtenir les meilleurs résultats.

5.3 Modélisation de l'approche

Dans notre approche de prédiction de la consommation de gaz, nous utilisons la méthode d'agrégation pondérée pour combiner les prédictions de deux modèles : LSTM (Long Short-Term Memory) et SARIMA (Seasonal Auto-regressive Integrated Moving Average). Voici comment nous utilisons

cette méthode pour obtenir une prédiction finale plus précise :

Entraînement des modèles individuels : Nous entraînons le modèle LSTM et le modèle SARIMA indépendamment sur les données d'apprentissage disponibles. Chaque modèle apprend à capturer les différentes caractéristiques et tendances des données de consommation de gaz.

Attribution des poids : Une fois que les modèles LSTM et SARIMA sont entraînés, nous leur attribuons des poids en fonction de leur performance respective. Les poids peuvent être définis manuellement en fonction de notre expertise ou être optimisés à l'aide de techniques d'optimisation telles que l'apprentissage automatique.

Prédiction pondérée : Pour obtenir la prédiction finale, nous multiplions les prédictions de chaque modèle par leur poids respectif. Cela permet de prendre en compte l'importance relative de chaque modèle dans la prédiction finale.

Prédiction finale = (Poids LSTM * Prédiction LSTM) + (Poids SARIMA * Prédiction SARIMA).

Normalisation des poids : Il est parfois nécessaire de normaliser les poids pour s'assurer qu'ils forment une distribution valide. Par exemple, nous pouvons normaliser les poids de manière à ce qu'ils se somment à 1. Cela garantit que chaque modèle contribue de manière appropriée à la prédiction finale.

Évaluation et ajustement des poids : Nous évaluons les performances de l'ensemble agrégé à l'aide de mesures d'évaluation appropriées, telles que l'erreur moyenne absolue (MAE) ou l'erreur quadratique moyenne (RMSE). Si nécessaire, les poids peuvent être ajustés et optimisés de manière itérative pour améliorer la performance globale de l'ensemble.

L'utilisation de la méthode d'agrégation pondérée nous permet de combiner les forces des modèles LSTM et SARIMA dans la prédiction de la consommation de gaz. En attribuant des poids adaptés à chaque modèle, nous pouvons exploiter les avantages spécifiques de chacun tout en réduisant l'impact des modèles moins performants. Cela conduit à des prédictions plus précises et fiables de la consommation de gaz.

5.4 Test de l'approche

Dans cette section, nous allons explorer les résultats de notre approche de prédiction de la consommation de gaz basée sur une combinaison de modèles LSTM et SARIMA. Nous commencerons par

examiner le graphe résultant des données d'entraînement, qui illustre les prédictions des modèles individuels LSTM et SARIMA, ainsi que de l'approche combinée avec les meilleurs poids optimaux. Ensuite, nous examinerons le graphe résultant des données de test, où nous appliquerons également les meilleurs poids obtenus pour évaluer les performances de notre approche. Enfin, nous discuterons des résultats finaux obtenus en utilisant un tableau qui récapitule les performances de chaque modèle et de l'approche combinée. Cette analyse approfondie nous permettra de mieux comprendre l'efficacité de notre approche et son potentiel à améliorer la précision des prédictions de consommation de gaz.

La figure 5.1 ilustre le graphe d'entrainement du modèle ensemble LSTM-SARIM

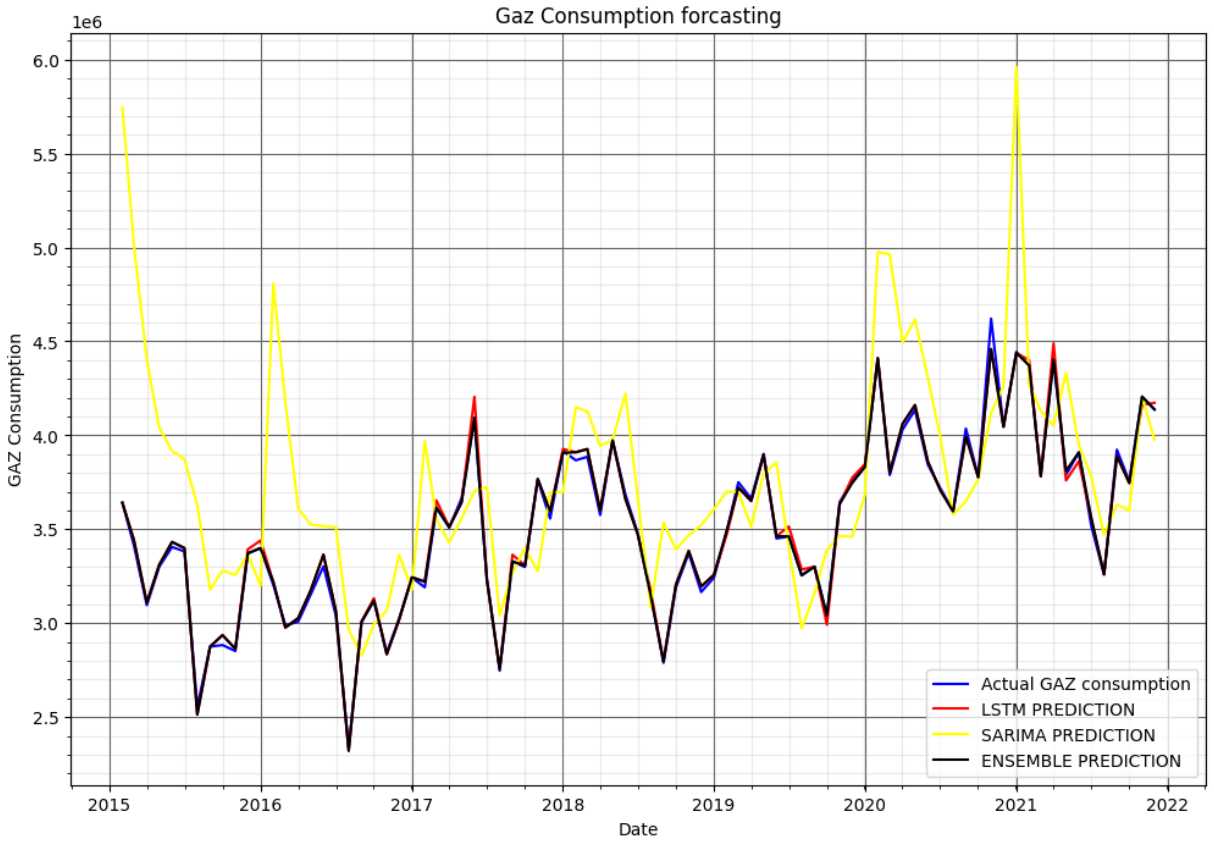


FIGURE 5.1 – entrainement ensemble LSTM_SARIMA

La figure présente l'entrainement du modèle ensemble LSTM_SARIMA pour les donnees d'entraînement de la consommation de gaz naturel pour le client '3360060HP. le graphe montre la consommation réelle de gaz naturel en bleu, l'entrainement du modèle LSTM en rouge, l'entrainement du modèle SARIMA en jaune et enfin l'ensemble des deux modèles est illustré en noire.

On obsèrve que le modèle LSTM s'est bien entraîné lors de la phase d'entraînement donc notre modèle a suivi dans la majorités des temps le modèle LSTM.

La figure 5.2 ilustre les le graphe des resultats de test du modèle ensemble LSTM-SARIM

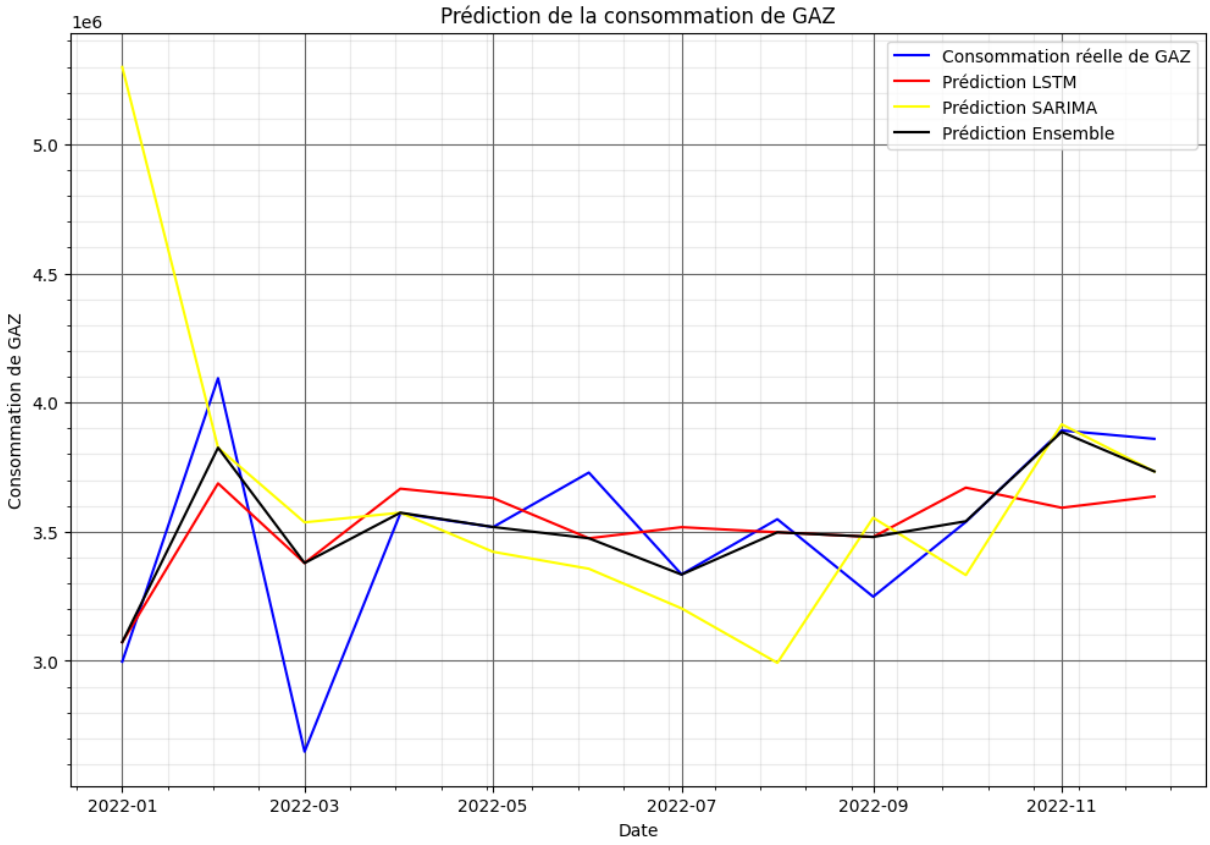


FIGURE 5.2 – partie test d'ensemble LSTM_SARIMA

En observant le graphe, on peut constater que le modèle "ensemble_lstm_sarima" présente une dynamique intéressante. Dans certains mois, on remarque que les prédictions de ce modèle suivent de près les prédictions du modèle LSTM, tandis que dans d'autres mois, elles suivent davantage les prédictions du modèle SARIMA. Cela suggère que notre approche combinée a réussi à combinée les performances des deux modèles individuels. Elle semble être capable de s'adapter aux différentes tendances et structures des données de consommation de gaz au fil du temps. Cette observation renforce l'idée que l'utilisation d'une combinaison de LSTM et SARIMA permet d'obtenir des résultats plus précis et adaptés à différentes périodes.

Dans le tableau 5.1 nous examinons les résultats réels de consommation de gaz ainsi que la comparaison entre les prédictions du modèle "ensemble_lstm_sarima" et les prédictions des modèles individuels (LSTM et SARIMA).

Date	LSTM	SARIMA	real	weights	enspred	score
2022-01-01	3072259	5450636.817	2996640	[1. 0.]	3072259	2.523459608
2022-02-01	3687209.75	3782378.684	4094442	[0. 1.]	3782378	7.621649055
2022-03-01	3378731	3854189.898	2648383	[1. 0.]	3378731	27.57712914
2022-04-01	3666615	3965920.396	3571148	[1. 0.]	3666615	2.673286013
2022-05-01	3630315.5	3819809.197	3517849	[1. 0.]	3630315	3.197010446
2022-06-01	3474500	3791667.534	3728992	[0.2 0.8]	3728234	0.020327209
2022-07-01	3517518	3562065.855	3334771	[1. 0.]	3517518	5.480046456
2022-08-01	3497830.5	3311070.823	3548435	[1. 0.]	3497830	1.426121656
2022-09-01	3479692.5	3954270.531	3248331	[1. 0.]	3479692	7.122457656
2022-10-01	3670664.75	3554474.782	3538105	[0. 1.]	3554474	0.462648791
2022-11-01	3592767.5	4283873.965	3892250	[0.5625 0.4375]	3895126	0.073890423
2022-12-01	3636266.5	3985669.005	3859602	[0.3636 0.6364]	3858613	0.025624404
		Score moyen (%)	4.85			

TABLE 5.1 – reusltat de l'ensemble_LSTM_SARIMA

Dans le tableau 5.1, nous pouvons observer les résultats de prédiction pour les modèles LSTM, SARIMA et l'ensemble LSTM_SARIMA (enspred). Les colonnes "LSTM" et "SARIMA" indiquent les prédictions respectives de chaque modèle, tandis que la colonne "real" représente les valeurs réelles. Les poids utilisés pour l'ensemble LSTM_SARIMA sont donnés dans la colonne "weights". La colonne "enspred" affiche les prédictions obtenues par l'ensemble LSTM_SARIMA.

En examinant les valeurs du tableau, nous constatons que l'ensemble LSTM_SARIMA est capable de capturer les meilleures performances des deux modèles individuels. Dans certains mois, l'ensemble suit la prédiction du modèle LSTM, tandis que dans d'autres mois, il suit la prédiction du modèle SARIMA. Cela suggère que notre modèle ensemble a réussi à exploiter les résultats les plus performants des deux approches.

Enfin, le score moyen (MAPE) est de 4.85%, ce qui indique une bonne précision globale de l'ensemble LSTM_SARIMA par rapport aux valeurs réelles.

Nos résultats ont montré que notre méthode de combinaison LSTM-SARIMA a donné des prédictions plus précises par rapport à chaque modèle pris individuellement, avec une précision de 95.05%. En comparaison, le modèle LSTM avait une précision de 93.01% et le modèle SARIMA avait une précision de 85.05%.

Dans la suit, nous allons analyser les graphiques de consommation de gaz individuels de nos clients mentionées dans la figure 4.1 (sous-section 1 de la section 3 du chapitre 4) afin d'identifier ceux qui ont des performances médiocres, et de leurs soumette un test de prédiction afin d'évaluer leurs resultats.

5.5 Analyse de la consommation de gaz des clients à faible rendement

Dans la figure mentionnée dans la sous-section 1 de la section 3 du chapitre 4, nous pouvons observer que les clients '33600056HP', '3360061HP' et '33600053' présentent une consommation de gaz médiocre. Afin de mieux comprendre cette situation, il est impératif de procéder à des tests spécifiques pour chaque client, dans le but d'analyser les différents facteurs qui pourraient influencer leur consommation. Ces tests nous permettront d'obtenir des résultats détaillés pour chaque client, ce qui sera essentiel pour une évaluation précise de leurs performances individuelles.

1. **Client '3360056HP'**

   Dans la figure 5.3 , nous pouvons observer la consommation de gaz du client '3360056HP' ainsi que sa prédiction.

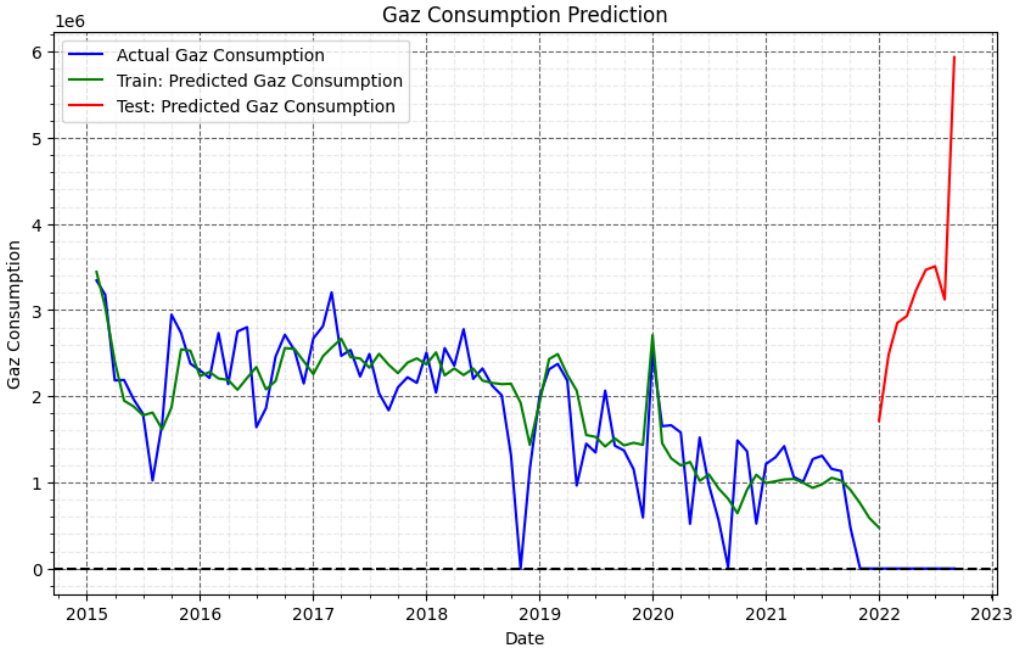


FIGURE 5.3 – consommation de client '3360056HP'

On observe que la consommation de ce client présente des valeurs réelles et prédites pour différentes périodes. Il est remarquable que la consommation réelle de gaz varie au fil du temps,

avec des fluctuations significatives. Les valeurs prédites montrent également des fluctuations, bien qu'elles ne correspondent pas exactement aux valeurs réelles.

Il est possible que les fluctuations de consommation soient dues à plusieurs facteurs. Premièrement, il est plausible que le client ait arrêté son usine pendant les périodes où les valeurs de consommation sont nulles. Deuxièmement, il est également envisageable qu'il ait changé de fournisseur de gaz, ce qui aurait entraîné une interruption de service pendant la transition entre les fournisseurs. Ces circonstances pourraient expliquer pourquoi les valeurs de consommation sont absentes à partir de l'année 2022.

2. **Client '3360061HP'**

   Dans la figure 5.4 , nous pouvons observer la consommation de gaz du client '3360061HP' ainsi que sa prédiction.

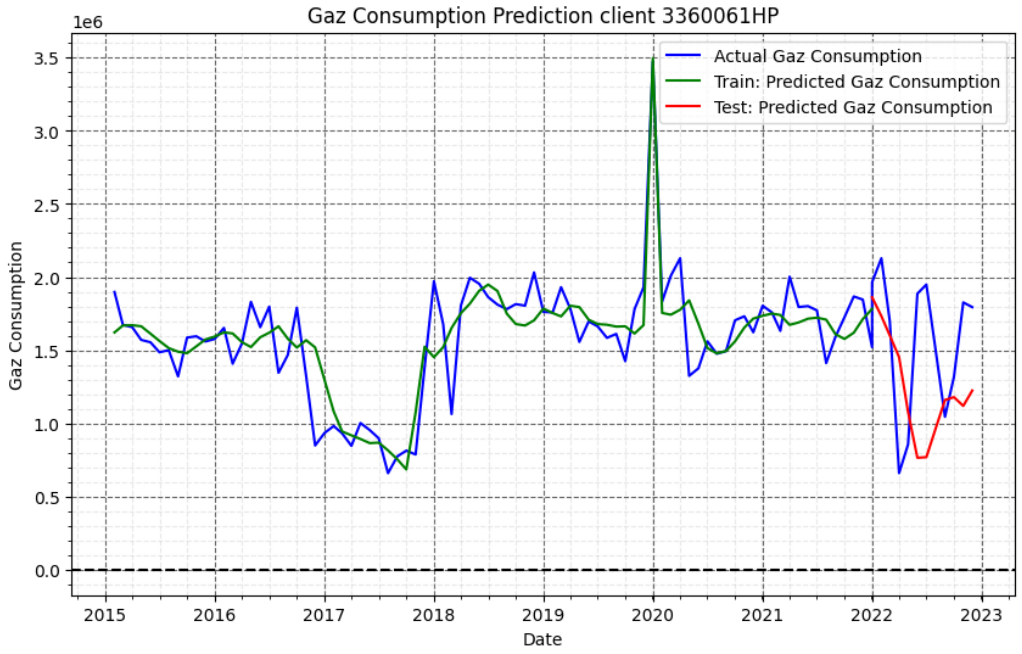


FIGURE 5.4 – consommation de client '3360061HP'

En analysant ce graphe, on peut observer que les valeurs réelles de consommation de gaz du client varient au fil du temps, avec des fluctuations significatives. De plus, les valeurs prédites ne correspondent pas exactement aux valeurs réelles, mais elles présentent également des fluctuations.

La cause possible de ces fluctuations pourrait être liée à certains facteurs, tels que des variations dans les opérations de l'usine du client. Les changements dans les activités de l'usine, tels que l'augmentation de la production ou l'utilisation de nouvelles machines plus économes en énergie, pourraient influencer la consommation de gaz.

3. **Client ’3360053HP’**

   Dans la figure 5.5 , nous pouvons observer la consommation de gaz du client ’3360053HP’ ainsi que sa prédiction.

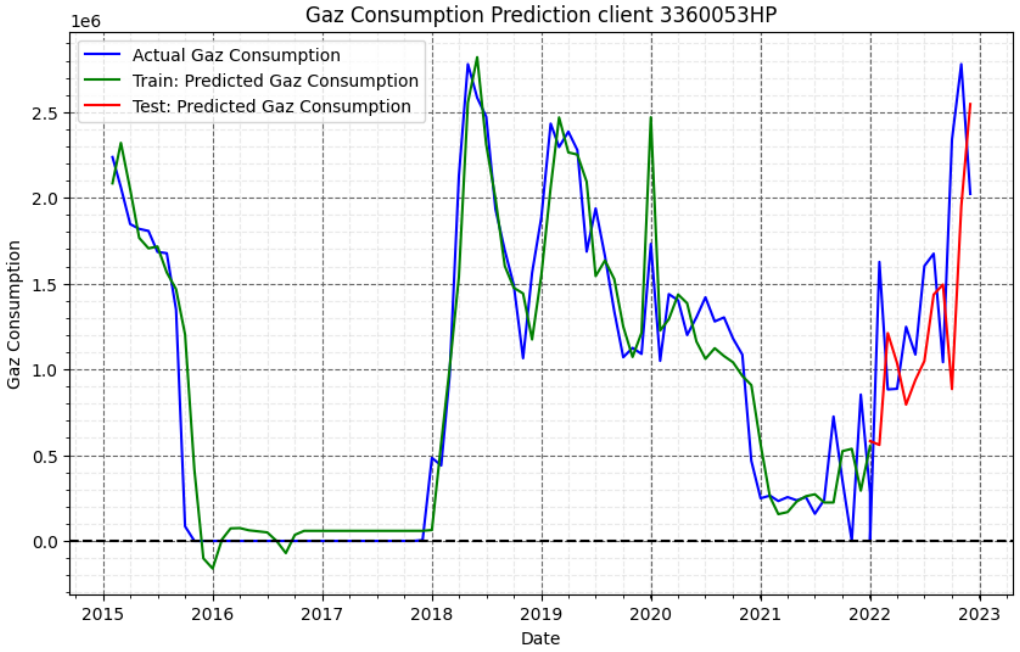


FIGURE 5.5 – consommation de client ’3360053HP’

En observant le graphique, on peut clairement constater des fluctuations significatives entre les valeurs prédites et les valeurs réelles de consommation de gaz pour ce client. Cela indique que les prédictions ne correspondent pas précisément aux données réelles et qu’il existe une certaine variabilité entre les deux.

Cependant, il convient de noter que les valeurs nulles dans les données réelles de ce client, notamment pendant l’année utilisée pour les tests, peuvent avoir un impact important sur les prédictions globales. Lorsque ces valeurs nulles sont utilisées pour estimer les prédictions, cela peut entraîner des valeurs prédites extrêmement élevées ou même des valeurs infinies lors de l’évaluation des prédictions.

Par conséquent, l’analyse de la consommation de gaz des clients à faible rendement met en évidence des fluctuations significatives dans les prédictions des modèles LSTM et SARIMA. Les mauvaises prédictions peuvent être attribuées à plusieurs facteurs, tels que les variations de consommation dues à des arrêts d’usine, des changements de fournisseurs de gaz ou des fluctuations dans les opérations des clients. De plus, la présence de valeurs nulles dans les données réelles peut également affecter les prédictions globales. Il est essentiel de mener des tests spécifiques pour chaque client afin d’analyser en détail les facteurs influençant leur consommation de gaz. Cela permettra d’améliorer la

précision des prévisions et de prendre des mesures correctives appropriées pour améliorer l'efficacité énergétique de ces clients.

5.6 Modélisation pour le reste des clients

Dans cette partie , Nous réaliserons des prédictions simultanées pour le reste des clients en utilisant à la fois la méthode LSTM et la méthode SARIMA.
Notre objectif est de comparer les performances des deux méthodes et de tirer parti de leurs avantages respectifs. De plus, nous nous mettons notre approche de combinaison d'ensemble pour fusionner les prédictions des deux méthodes.

5.6.1 Modélisation de la méthode SARIMA

Dans cette partie, nous allons effectuer une modélisation de la prédiction de la consommation de gaz pour le reste des clients. L'objectif est d'utiliser le modèle SARIMA pour analyser et prévoir les tendances de consommation de gaz des clients restant.

- **Remarque :** toutes les étapes de modélisation et de prévision sont les mêmes que celles déjà mentionnées dans la sous-section précédente (Modélisation SARIMA pour un seul client).

  Cependant, SARIMA est une méthode qui nécessite un modèle distinct pour chaque client. Pour faciliter la comparaison, nous avons effectué les étapes suivantes : nous avons pris toutes les clients restant et avons agrégé leurs données de consommation de gaz par mois. En effectuant la somme des valeurs de consommation pour chaque mois, nous avons créé une série chronologique unique, comme si nous avions un seul client avec des données pour tous les mois. Nous avons ensuite utilisé cette série agrégée pour effectuer les prédictions afin de comparer les performances de cette méthodes en termes de précision et d'efficacité.

Ensuite, Nous avons effectué le test ADF (Augmented Dickey-Fuller) sur les données utilisées pour la modélisation SARIMA.

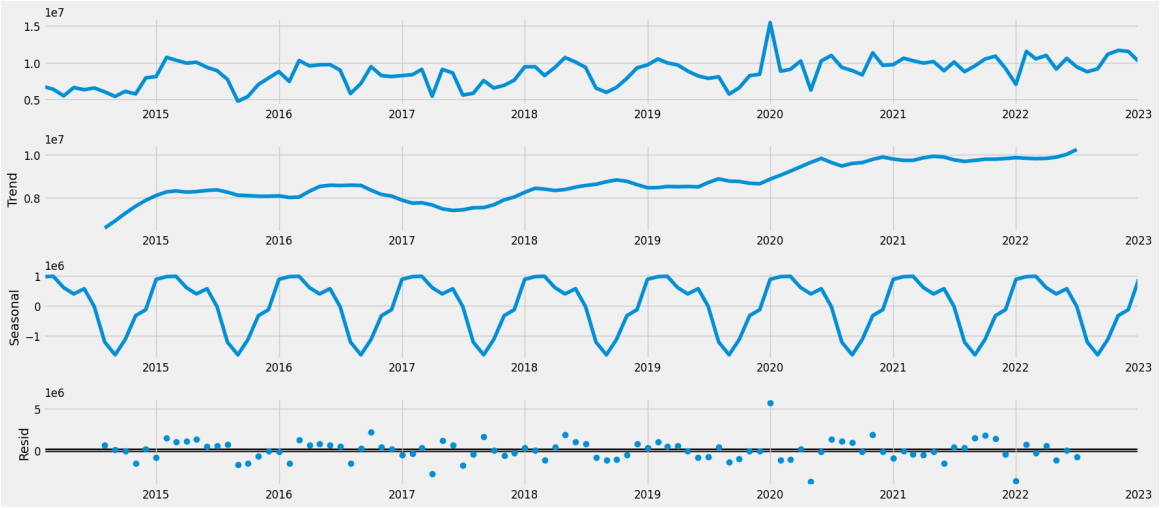


FIGURE 5.6 – composant de notre serie pour le reste des clients

- les resultats du test ADF :

— ADF : -5.642572444906679

— P-Value : 1.0284953204053702e-06

— Nombre de retards (Num Of Lags) : 0

— Nombre d'observations utilisées pour la régression ADF et le calcul des valeurs critiques : 107

— Valeurs critiques :

  — 1% : -3.492995948509562

  — 5% : -2.888954648057252

  — 10% : -2.58139291903223

Les résultats de ce test suggèrent fortement que les données sont stationnaires, ce qui est un prérequis essentiel pour l'application du modèle SARIMA. La statistique ADF est significativement négative et inférieure aux valeurs critiques, tandis que la valeur de p est extrêmement faible. De plus, le nombre de retards utilisés pour le test est nul, ce qui indique que les données ne nécessitent pas de différenciation supplémentaire pour obtenir la stationnarité. En conclusion, ces résultats confirment que les données sont adaptées à l'utilisation du modèle SARIMA dans notre processus de prédiction de la consommation de gaz pour ces clients.

1. **Evaluation**

   tout d'abord nous avons met un diagnostique complet a notre nouvelle série chronologique (agrégation de nos cinq clients) :

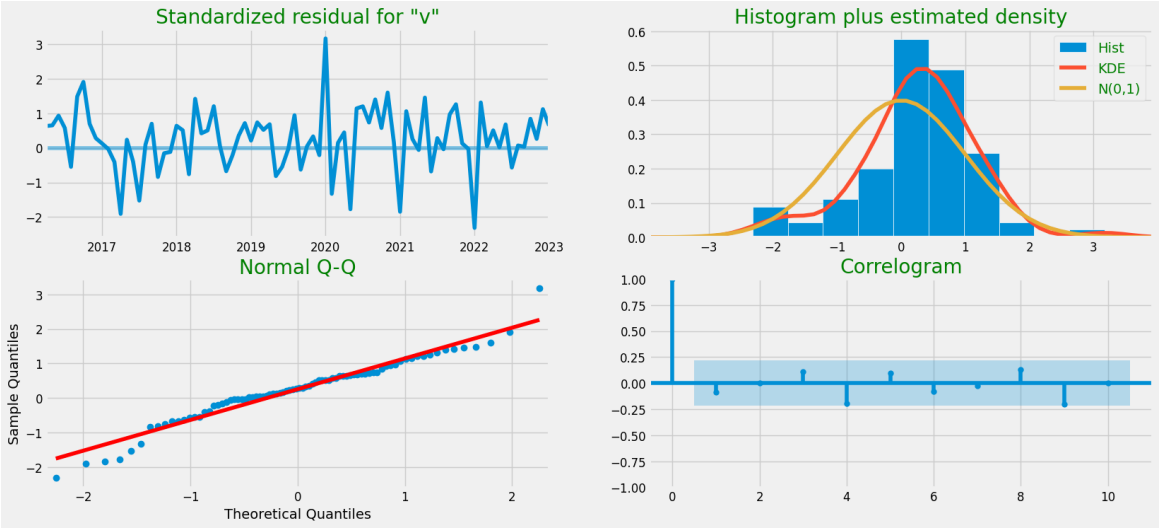


FIGURE 5.7 – Les diagnostics du modèle SARIMA

Comme nous le constatons, la figure nous affiche 4 différents graphes soit :

- **Histogamme plus estimated density :** .

  comme nous pouvons le constater, densité estimée est proche d'une distribution normale, avec une forme en cloche et une symétrie autour de zéro. cela indique que le modèle est bien ajusté.

- **Graphique Q-Q :** le graphe Normal Q-Q montre que les résidus semblent suivre une distribution normale. Les points sur le graphe sont alignés sur une ligne droite, ce qui suggère que les résidus sont distribués normalement.

- **Standardized residual :** Le graphe des résidus standardisés montre que les résidus du modèle SARIMA semblent être aléatoires et non corrélés. Les points sur le graphe sont répartis de manière aléatoire autour de zéro, ce qui suggère que les résidus sont aléatoires. De plus, il n'y a pas de tendance claire ou de structure dans les résidus, ce qui indique qu'ils ne sont pas corrélés.

  Cela est confirmé par les autres graphes tels que l'histogramme de densité estimée et le graphe Q-Q, qui montrent également que les résidus sont distribués normalement et aléatoires. Ces résultats suggèrent que le modèle SARIMA est bien ajusté et fournit des prévisions précises.

- **Correlogramme :**

  Le correlogramme confirme la nature aléatoire et non corrélée des données. Les barres du graphe se situent toutes dans la zone bleue, qui représente la zone de confiance à 95%. Cela

indique qu'il n'y a pas de corrélation significative entre les résidus à différents retards. Les résidus du modèle SARIMA présentent donc un comportement aléatoire et ne montrent pas de structure temporelle résiduelle. Cette observation renforce l'idée que le modèle SARIMA est bien ajusté et produit des prévisions précises pour la consommation de gaz des cinq clients.

2. **Résultats :** Dans cette partie, nous commencerons par présenter le graphique de prédiction de la consommation de gaz pour nos clients. Ensuite, nous procéderons à une analyse de la partie prédite du graphique pour obtenir des informations sur la précision du modéle.

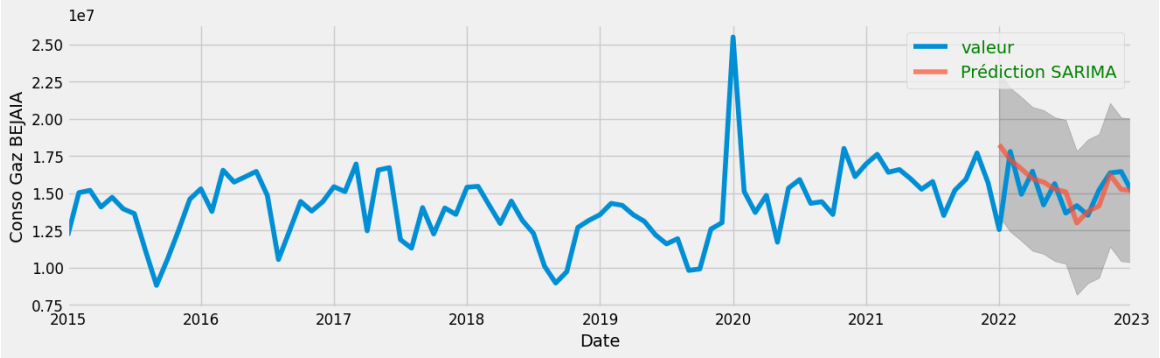


FIGURE 5.8 – Prédiction de la consommation de gaz pour l'année 2022 pour le reste des clients

Ce graphique représente la prédiction de la consommation de gaz naturel pour l'année 2022. Comme nous l'avons déjà fait pour la prédiction pour un seul client, nous avons divisé nos données en deux ensembles : d'entraînement(en bleu) et test(en orange).

En analysant cette figure, on trouve que les valeurs prédites suivent une tendance très proche à celle des valeurs réelles. Cela suggère que la méthode SARIMA utilisée a été efficace pour prédire la consommation de gaz naturel à Bejaia.

Dans ce qui suit, nous allons discuter les résultats de notre prédiction de la consommation de gaz pour l'année 2022 en analysant le tableau des resultats et en comparant les valeurs prédites aux valeurs réelles fournies par nos clients.

Le tableau 5.2 présente les valeurs réelles, les valeurs prédites et le pourcentage d'erreur absolu moyen (MAPE) pour la prédiction de la consommation de gaz agrégée pour le reste des clients en utilisant le modèle SARIMA.

Date	Réel	Prédit	MAPE%
2022-01-01	12556723	18259214,42	45,41385061
2022-02-01	17805534	17243779,44	3,154943651
2022-03-01	14931806	16636891,58	11,4191517
2022-04-01	16476398	15964712,16	3,105568594
2022-05-01	14216402	15753857,21	10,8146577
2022-06-01	15643564	15275487,72	2,352892751
2022-07-01	13670909	15090749,02	10,38584936
2022-08-01	14164602	13002199,27	8,206391746
2022-09-01	13520414	13784852,86	1,955848853
2022-10-01	15216006	14146705,5	7,027471624
2022-11-01	16367034	16240098,13	0,775558157
2022-12-01	16456210	15259008,88	7,275071957
	Moyenne MAPE%	9,32393806	

TABLE 5.2 – la prédiction de la consommation de gaz avec SARIMA pour le reste des clients.

Les résultats présentés dans le tableau 5.2 sont basés sur les données des clients restant agrégés concernant la consommation de gaz naturel. Nous avons effectué des prédictions pour l'année 2022 et comparé les valeurs prédites aux valeurs réelles.

Les valeurs prédites diffèrent des valeurs réelles pour chaque mois de l'année 2022, comme le montrent les colonnes "Réel" et "Prédit". Nous avons utilisé le pourcentage d'erreur absolu moyen (MAPE%) comme mesure de précision des prédictions.

En examinant les valeurs de MAPE%, nous pouvons constater que certaines prédictions ont une précision relativement élevée, tandis que d'autres ont une précision plus faible.

Par exemple, en janvier 2022, la valeur réelle était de 12 556 723, tandis que la prédiction était de 18 259 214,42, ce qui correspond à une MAPE% de 45,41%. Cette erreur élevée suggère que le modèle n'a pas réussi à capturer correctement la tendance de consommation pour ce mois.

D'un autre côté, en novembre 2022, la valeur réelle était de 16 367 034, tandis que la prédiction était de 16 240 098,13, avec une MAPE% de 0,78%. Cette prédiction montre une précision relativement élevée, indiquant que le modèle a réussi à capturer la tendance de consommation de manière plus précise pour ce mois.

En calculant la moyenne de MAPE%, nous obtenons une valeur de 9,32%. Cela suggère que, en moyenne, les prédictions du modèle présentent une erreur relative de 9,32% par rapport aux

valeurs réelles.

Sur la base de ces résultats, il est raisonnable de conclure que le modèle SARIMA utilisé pour prédire la consommation de gaz des clients agrégés donne des prédictions relativement précises, mais elles ne sont pas assez optimales. Une précision de 9,32% suggère que les prédictions du modèle s'écartent significativement des valeurs réelles, indiquant une marge d'erreur considérable. Par conséquent, il est important d'améliorer la précision et la fiabilité du modèle afin d'obtenir des prédictions plus optimales pour la consommation de gaz.

5.6.2 Modélisation du modèle LSTM

Dans cette partie, nous allons aborder la modélisation du modèle LSTM pour la prédiction de la consommation du gaz naturel pour le reste des client.

Remarque : Dans cette partie, toutes les etapes éffectuées dans la sous-section 2 de la sous-section 1 de la section 3 du chapitre 4 'modélisations de la méthode LSTM' sont les mêmes. La différence est que cette fois-ci, nous allons travailler avec reste des clients au lieu d'un seul client.

- Notre modèle a été entraîné avec les hyperparamètres suivants :
  - Nombre d'échantillons : nous avons utilisé notre modèle pour prédire la consommation de gaz naturel pour cinq clients.
  - Taille des lots (batch size) : 128 lots.
  - Nombre d'époques : 510 epochs.

1. **Evaluation :**

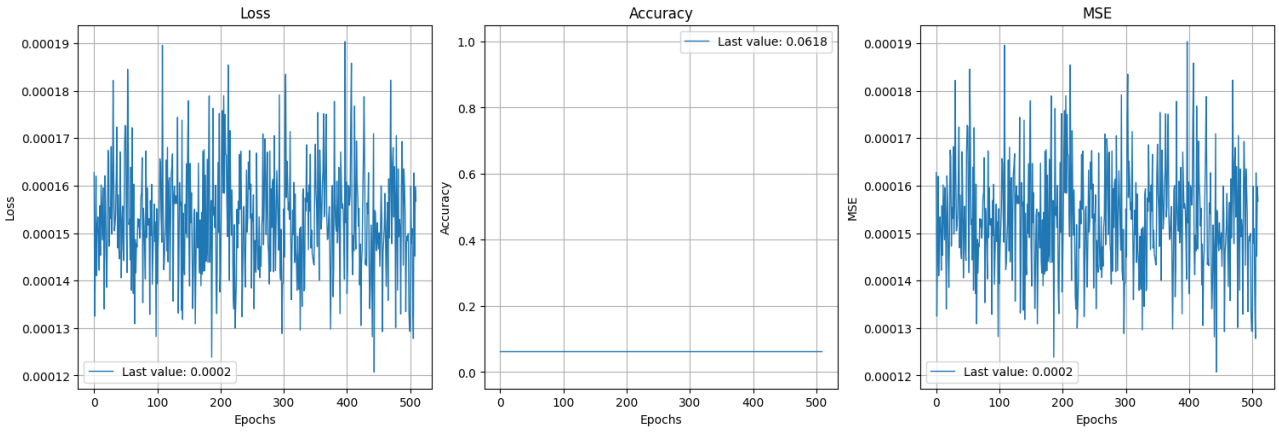


FIGURE 5.9 – entrainement du modèle lstm pour le rest des clients

Dans ce graphe on voie que le modèle lstm a obtenu une MSE de 0.0002 et une accuracy de 0.0618, ce qui indique une très faible erreur quadratique moyenne et une exactitude relativement

élevée. cela signifié que le Modèle a réussi a bien capturer les motifs et les relations dans les données d'entraînements. signifié que les valeurs prédites doivent être proches des valeurs réelles, ce qui indique une bonne qualité de prédition.

Pour une évaluation plus complète nous allons évaluer les résultats des données de test et de validation.

Cela nous permettera de déterminer si notre modèle est capable de générer des valeurs prédites assez proches des valeurs réelles.

2. **Résultat :** Dans cette section, nous allons discuter des résultats de notre prédiction de la consommation de gaz naturel pour les mois de l'année 2022 (les données de test). Nous commençons par l'analyse du graphe de la consommation.

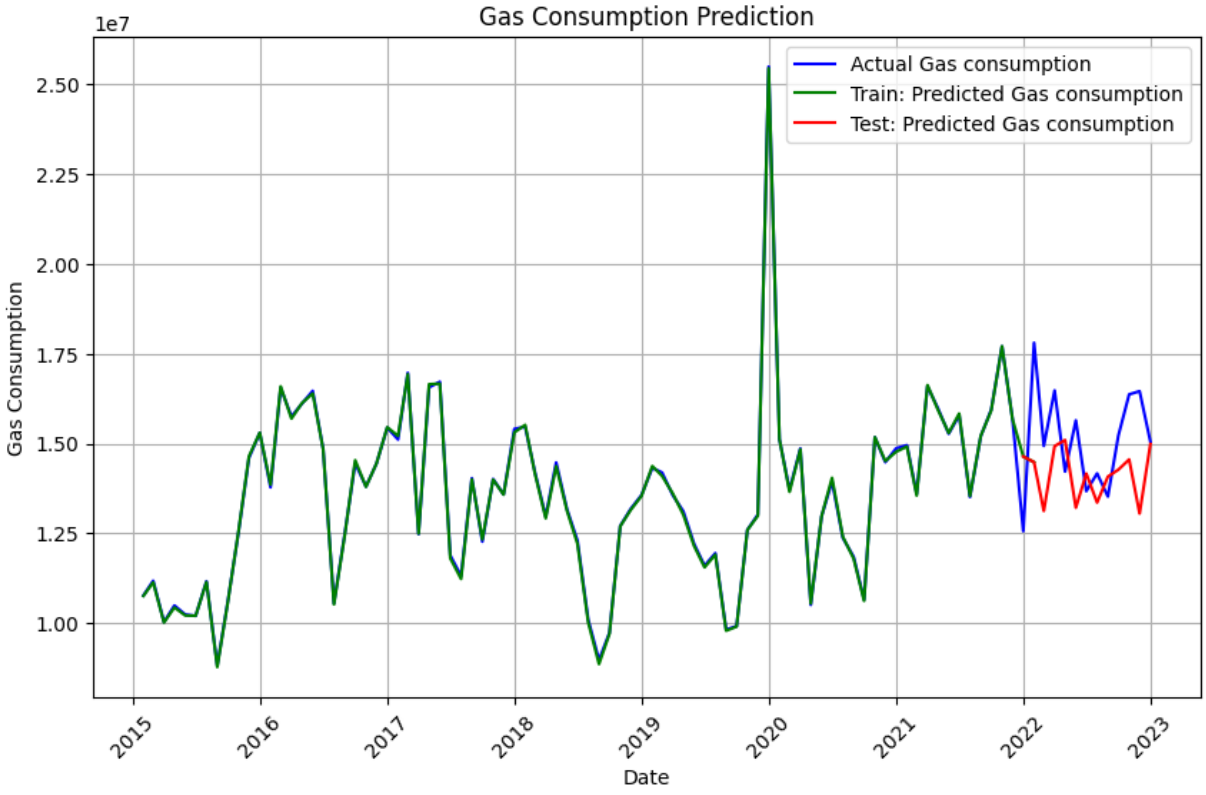


FIGURE 5.10 – Graphe de prediction de la consommation de gaz avec LSTM

Le graphe montre la consommation réelle de gaz naturel en bleu en suite comme nous l'avons dis, nous avons dévisés les données de consommation en deux sous-ensemble : données d'entraînements représenté en vert et les données de test en rouge.

Comme nous pouvons le voir, le modèle LSTM s'est très bien entraîné. Les données d'entraînement correspondent exactement aux données réelles, et les données de test sont très proches des données réelles. Cela suggère que le modèle LSTM est capable de capturer les motifs et les tendances de la consommation de gaz naturel.

dans ce qui suit, nous examinerons de plus près les performances du modèle en utilisant la métrique d'évaluation MAPE. Cela nous permettra d'avoir une mesure quantitative de la précision du modèle.

Le tableau 5.3 présente les valeurs réelles, les valeurs prédites et le pourcentage d'erreur absolu moyen (MAPE) pour la prédiction de la consommation de gaz pour l'année 2022.

Date	Réel	Prédit	MAPE%
2022-01-01	12556723	14636142	16,56020444
2022-02-01	17805534	14478447	18,68569064
2022-03-01	14931806	13119256	12,13885313
2022-04-01	16476398	14925235	9,414454543
2022-05-01	14216402	15094417	6,176070429
2022-06-01	15643564	13209078	15,56222099
2022-07-01	13670909	14161886	3,591399811
2022-08-01	14164602	13353150	5,728731383
2022-09-01	13520414	14076802	4,115169846
2022-10-01	15216006	14268688	6,225799333
2022-11-01	16367034	14554125	11,07658846
2022-12-01	16456210	13047216	20,7155475
	Moyenne MAPE%	10,83256088	

TABLE 5.3 – la prédiction de la consommation de gaz pour l'annee 2022 avec LSTM pour le reste des clients

Le tableau 5.3 présente les résultats de prédiction de la consommation de gaz pour l'année 2022, obtenus à l'aide du modèle LSTM, pour le reste des clients.

En observant les valeurs réelles et prédites, nous constatons une variation entre les deux. Certaines prédictions se rapprochent étroitement des valeurs réelles, tandis que d'autres présentent des écarts plus importants. Cette différence de précision peut être due à plusieurs facteurs, tels que des variations saisonnières, des événements imprévus.

En analysant les résultats de manière approfondie, nous remarquons que la moyenne du MAPE (Mean Absolute Percentage Error) pour l'ensemble des observations est de 10,83%. Cela indique une précision globalement raisonnable du modèle LSTM dans la prédiction de la consommation de gaz pour ces clients. Néanmoins, il est important de souligner que le MAPE varie

d'une observation à l'autre, soulignant ainsi l'importance d'analyser chaque résultat individuellement.

Certains résultats se démarquent par leur précision relativement élevée. Par exemple, en juillet 2022, la valeur réelle est de 13 670 909 et la valeur prédite est de 14 161 886, avec un MAPE de seulement 3,59%. Cette prédiction précise suggère que le modèle LSTM a réussi à capturer de manière fiable la tendance de consommation de gaz pour cette période spécifique.

D'autre part, certaines prédictions présentent des écarts plus importants par rapport aux valeurs réelles. Par exemple, en décembre 2022, la valeur réelle est de 16 456 210 et la valeur prédite est de 13 047 216, avec un MAPE de 20,72%. Ces écarts peuvent être attribués à des circonstances particulières ou à des fluctuations inattendues dans les données.

En conclusion, le modèle LSTM a démontré une précision raisonnable dans la prédiction de la consommation de gaz pour ces clients, avec une moyenne du MAPE de 10,83%. Cependant, il est essentiel de considérer les variations individuelles des prédictions et d'analyser attentivement chaque résultat. Cela permettra d'identifier les facteurs spécifiques qui influencent la précision des prédictions et d'améliorer le modèle en conséquence.

5.6.3 Comparaison entre les deux modèles

Dans cette sous-section nous allons faire une comparaison entre les deux modèles LSTM et SARIMA selone la métrique MAPE% pour le reste des client (agrégés pour le modèle SARIMA).

La figure 5.11 montre un réprésentation graphique des MAPE% des deux modèles.

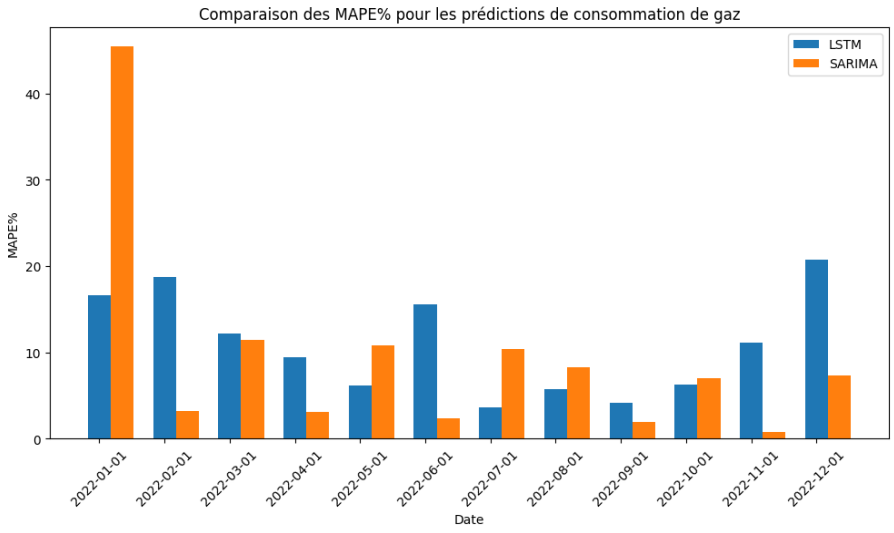


FIGURE 5.11 – comparaison LSTM & SARIMA selons le MAPE% pour le reste des clients

Le graphe représente la comparaison des valeurs de l'erreur de pourcentage moyen absolu (MAPE) entre les modèles LSTM et SARIMA pour chaque mois de l'année 2022 pour le reste des clients. Les barres bleues représentent les valeurs MAPE% du modèle LSTM, tandis que les barres oranges représentent les valeurs MAPE% du modèle SARIMA.

En observant les hauteurs des barres pour chaque mois, on peut constater qu'il y a une variation dans les résultats entre les deux méthodes. Pour certains mois (février, mars, avril, juin, septembre, novembre et décembre), la méthode LSTM présente des valeurs MAPE% plus élevées que la méthode SARIMA, indiquant une moindre précision des prédictions. Cependant, pour les autres mois restant, la méthode LSTM présente des valeurs MAPE% plus faibles que la méthode SARIMA, ce qui suggère une meilleure adéquation avec les valeurs réelles de consommation de gaz pour ces mois-là.

Le tableau 5.4 représente les valeurs du MAPE% pour chaque mois ainsi que la moyenne MAPE% pour les deux modèles :

Mois	MAPE (LSTM)	MAPE (SARIMA)
2022-01-01	16.56%	45.41%
2022-02-01	18.69%	3.15%
2022-03-01	12.14%	11.42%
2022-04-01	9.41%	3.11%
2022-05-01	6.18%	10.81%
2022-06-01	15.56%	2.35%
2022-07-01	3.59%	10.39%
2022-08-01	5.73%	8.21%
2022-09-01	4.12%	1.96%
2022-10-01	6.23%	7.03%
2022-11-01	11.08%	0.78%
2022-12-01	20.72%	7.28%
Moyenne MAPE	10.83%	9.32%

TABLE 5.4 – comparaison MAPE% (LSTM & SARIMA) pour pour nos clients

Ce tableau présente les valeurs de MAPE% pour LSTM et SARIMA pour chaque mois de l'année 2022 pour nos clients .

Après avoir observé les résultats des deux modèles LSTM et SARIMA, il est clair que leurs performances varient en termes de précision pour différents mois. Certains mois montrent une meilleure

précision du modèle LSTM, tandis que d'autres mois montrent une meilleure précision du modèle SARIMA. Cette observation suggère que chaque modèle a ses propres forces et faiblesses en ce qui concerne la prévision de la consommation de gaz pour ces clients.

Pour tirer parti des avantages de chaque modèle et améliorer davantage les performances de prévision, nous allons mettre en oeuvre notre approche de pondération agrégée afin de combinée les performances de ces deux modèles.

Dans la suite de cette étude, nous allons détailler notre approche pour le reste de ces cleints.

5.7 La prédiction pour le reste des clients avec l'ensemble_lstm_sarima

Dans cette section, nous allons explorer les résultats de notre approche de prédiction de la consommation de gaz basée sur la combinaison des deux modèles LSTM et SARIMA.

Remarque

Toutes les étapes de modélisation sont les mêmes que celles deja mentionées dans la sous-section 2 de la section 4 de chapitre 4, sauf que ici nous allons faire la combinaison des deux modèles afin de prédire la consommation de tout le reste des clients.

1. **Resultats**

Dans cette section nous allons discuter des résultats de notre approche pour la prédiction de la consommation de gaz naturel pour les mois de l'année 2022 pour le reste des clients. Nous commencons par l'analyse du graphe resultant de la consommation réel et de prédiction.

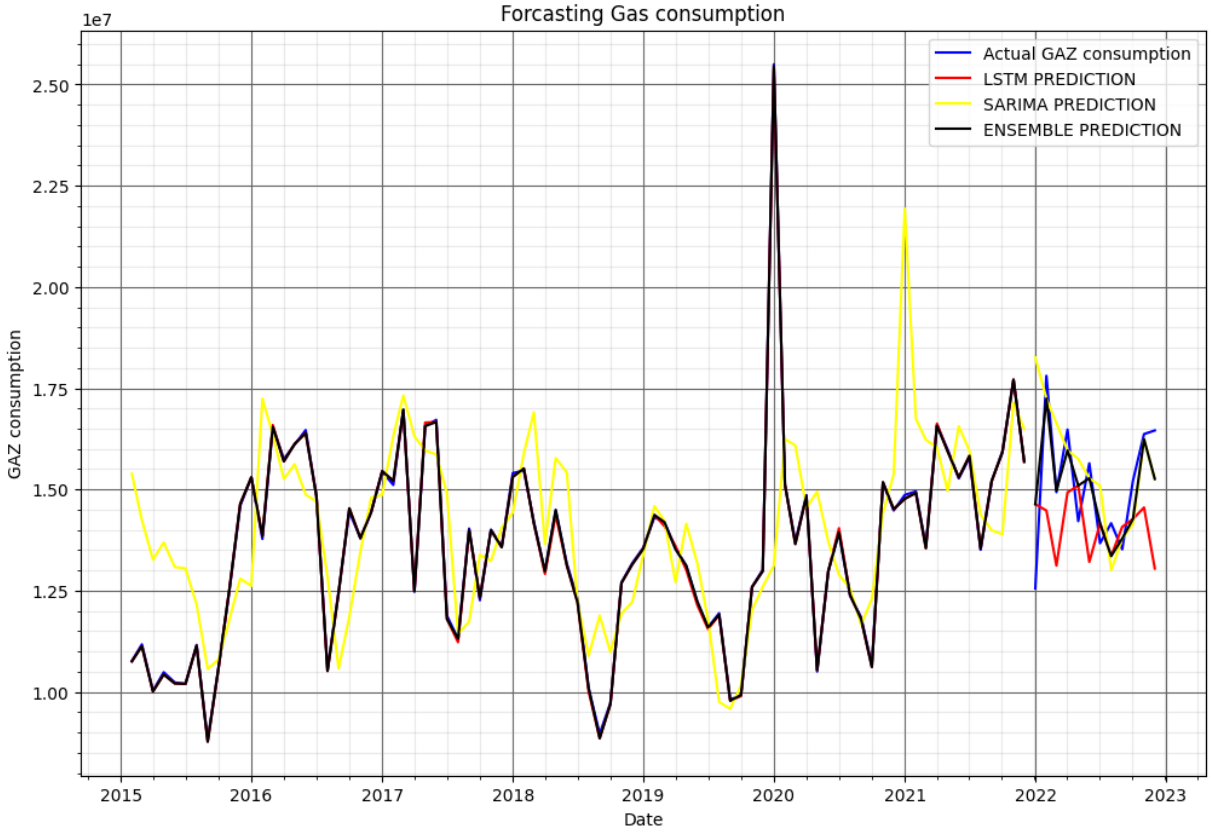


FIGURE 5.12 – Graphe de prediction de la consommation de gaz avec ensemble_lstm_sarima

La figure présente l'entraînement et le test de notre approche sur les données d'entraînement et de test des deux modèles LSTM et SARIMA pour le reste des clients. Les valeurs réelles sont représentées en bleu, les valeurs prédites par le modèle SARIMA sont présentées en jaune, les valeurs prédites par le modèle LSTM en rouge, et les valeurs prédites par notre approche pour l'ensemble des deux modèles sont présentées en noir.

On observe que notre approche a réussi à capturer avec précision et à tirer profit des avantages des deux modèles. Le figure suivante illustre cette observation de manière claire.

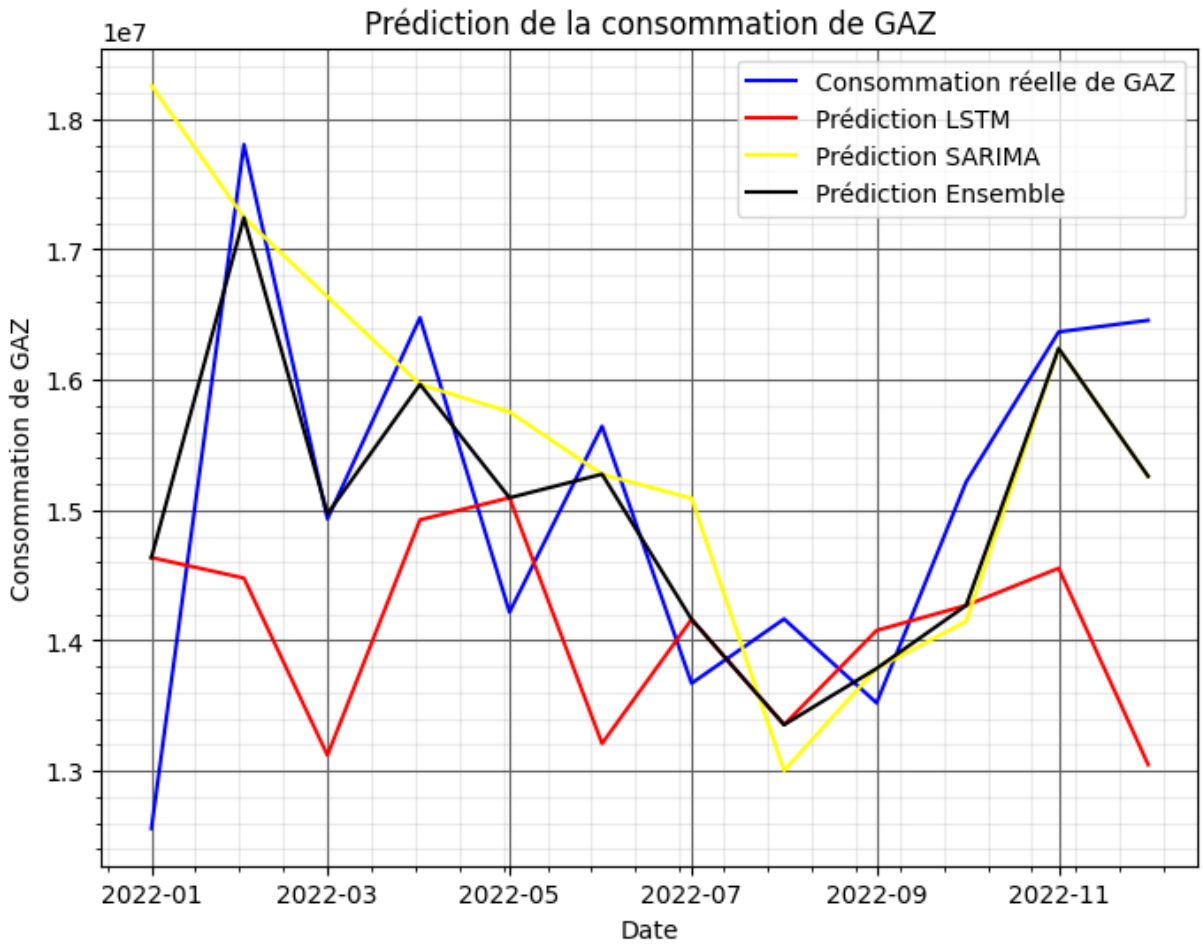


FIGURE 5.13 – Graphe de prediction de la consommation de gaz avec ensemble_lstm_sarima

De plus prés, on observe que les prédictions du modèle Ensemble se situent entre celles du modèle LSTM et du modèle SARIMA pour chaque période donnée. Le modèle LSTM semble mieux performer que le modèle SARIMA dans certaines périodes, tandis que le modèle SARIMA peut être plus précis dans d'autres périodes. Notre modèle semble combiner les avantages des deux modèles pour fournir des prédictions plus précises et cohérentes.

Dans le tableau 5.5 nous examinons les résultats réels de consommation de gaz ainsi que la comparaison entre les prédictions du de notre modèle et les prédictions des modèles individuels (LSTM et SARIMA).

Date	LSTM	SARIMA	real	weights	enspred	score
2022-01-01	14636142	18259214.42	12556723	[1.0 0.0]	14636142	16.56020444
2022-02-01	14478447	17243779.44	17805534	[0.0 1.0]	17243779	3.154946097
2022-03-01	13119256	16636891.58	14931806	[0.47 0.53]	14970643	0.260095798
2022-04-01	14925235	15964712.16	16476398	[0.0 1.0]	15964712	3.105569555
2022-05-01	15094417	15753857.21	14216402	[1.0 0.0]	15094417	6.176070429
2022-06-01	13209078	15275487.72	15643564	[0.0 1.0]	15275487	2.352897332
2022-07-01	14161886	15090749.02	13670909	[1.0 0.0]	14161886	3.591399811
2022-08-01	13353150	13002199.27	14164602	[1.0 0.0]	13353150	5.728731383
2022-09-01	14076802	13784852.86	13520414	[0.0 1.0]	13784852	1.955842476
2022-10-01	14268688	14146705.5	15216006	[1.0 0.0]	14268688	6.225799333
2022-11-01	14554125	16240098.13	16367034	[0.0 1.0]	16240098	0.775558968
2022-12-01	13047216	15259008.88	16456210	[0.0 1.0]	15259008	7.275077311
		Score moyen (%)	4.763516078			

TABLE 5.5 – reusltat de l'ensemble_LSTM_SARIMA

Dans le tableau 5.5, nous pouvons observer les résultats de prédiction pour les modèles LSTM, SARIMA et l'ensemble LSTM_SARIMA (enspred). Les colonnes "LSTM" et "SARIMA" indiquent les prédictions respectives de chaque modèle, tandis que la colonne "real" représente les valeurs réelles. Les poids utilisés pour l'ensemble LSTM_SARIMA sont donnés dans la colonne "weights". La colonne "enspred" affiche les prédictions obtenues par la combinaison LSTM et SARIMA.

En examinant les valeurs du tableau, nous constatons que l'ensemble LSTM_SARIMA est capable de capturer les meilleures performances des deux modèles individuels. Dans certains mois, l'ensemble suit la prédiction du modèle LSTM, tandis que dans d'autres mois, il suit la prédiction du modèle SARIMA. Cela suggère que notre modèle ensemble a réussi à exploiter les résultats les plus performants des deux approches.

Enfin, le score moyen (MAPE) est de 4.76%, ce qui indique une bonne précision globale de l'ensemble LSTM_SARIMA par rapport aux valeurs réelles.

Nos résultats ont montré que notre méthode de combinaison LSTM-SARIMA a donné des prédictions plus précises par rapport à chaque modèle pris individuellement, avec une précision de 95.24%. En comparaison, le modèle LSTM avait une précision de 89.17% et le modèle SARIMA avait une précision de 90.68%.

En conclusion, notre méthode de combinaison LSTM-SARIMA s'avère prometteuse pour les tâches de prédiction, offrant une meilleure précision que les modèles individuels. Cette approche pourrait être appliquée dans d'autres domaines où la combinaison de différents modèles peut conduire à de meilleurs résultats de prédiction.

5.8 Conclusion

En conclusion, afin de tirer parti des avantages des deux approches SARIMA et LSTM, nous avons proposé une approche de combinaison, appelée ensemble LSTM_SARIMA, en utilisant une méthode d'agrégation pondérée de l'ensemble learning. Cette approche nous a permis de créer une prédiction combinée qui exploite les performances des deux modèles.

En testant notre approche de combinaison sur l'ensemble des clients, nous avons constaté que l'ensemble LSTM_SARIMA offrait des prédictions plus précises par rapport à chaque modèle pris individuellement. Le pourcentage d'erreur absolue moyenne (MAPE) de l'ensemble LSTM_SARIMA était de 4.76%, ce qui démontrait une bonne précision globale.

Enfin, notre étude a montré que la combinaison des modèles LSTM et SARIMA à travers l'ensemble LSTM_SARIMA permet d'améliorer la précision des prédictions par rapport à chaque modèle pris individuellement. Cette approche offre une solution prometteuse pour la prédiction de la consommation de gaz naturel, en exploitant les avantages des deux approches. Des recherches supplémentaires peuvent être menées pour affiner les techniques de combinaison et explorer d'autres modèles et méthodes d'ensemble afin d'améliorer encore davantage les performances de prédiction.

CONCLUSION GÉNÉRALE & PERSPECTIVES

En conclusion, cette étude met en évidence l'importance cruciale de l'énergie gazière pour l'économie de l'Algérie et du monde. En tant que pays exportateur de gaz, il est essentiel d'améliorer les méthodes de prévision de la consommation future afin de mieux gérer cette ressource précieuse. Les risques associés à une demande locale dépassant la part réservée à l'exportation sont préoccupants, car ils pourraient avoir des conséquences économiques potentiellement désastreuses.

Face à ces défis, l'utilisation de l'intelligence artificielle se présente comme une solution moderne et prometteuse pour prédire avec précision la consommation future de l'énergie gazière à Béjaïa. Cette approche repose sur l'analyse détaillée de la consommation de gaz naturel des clients économiques haute pression de la wilaya de Béjaïa, ce qui permet d'estimer la consommation globale de gaz naturel de manière plus fiable.

L'étude commence par une exploration des aspects généraux du gaz naturel, en mettant en évidence son origine, son histoire et ses caractéristiques techniques. Une compréhension approfondie de ces éléments est essentielle pour évaluer l'importance et les applications du gaz naturel dans divers secteurs.

La SONELGAZ, en tant qu'entreprise majeure dans la distribution de gaz naturel, joue un rôle central dans cette étude. Son historique, son organisation, sa structure et ses principales activités sont analysés, de même que la répartition de la distribution du gaz naturel en fonction des types de clients. L'organisme d'accueil, la SONELGAZ Distribution CD Béjaïa, est également présenté en détail.

Les méthodes traditionnelles et les méthodes d'apprentissage automatique utilisées pour la prédiction de la consommation de gaz naturel sont examinées dans la troisième partie de l'étude. Les séries temporelles et les méthodes d'apprentissage automatique telles que l'apprentissage supervisé, non supervisé et par renforcement sont discutées, mettant l'accent sur l'évolution vers l'apprentissage en profondeur, qui comprend les réseaux de neurones artificiels, les réseaux de neurones à convolution, les réseaux de neurones récurrents et les longues mémoires à court terme.

L'état de l'art dans le domaine de la prédiction de la consommation de gaz naturel est examiné, en présentant des travaux connexes et en fournissant un tableau comparatif des solutions existantes. Les avantages et les limites de ces solutions sont discutés pour évaluer leur applicabilité dans le contexte de Béjaïa.

Ensuite, l'étude se concentre sur la modélisation et l'évaluation des modèles de prédiction de la consommation du gaz naturel. L'environnement de développement utilisé, y compris le matériel et les logiciels, et les étapes de modélisation pour un seul client sont détaillées. Des analyses supplé-

mentaires sont réalisées sur plusieurs clients individuels avec une consommation modérée afin de modéliser la consommation pour le reste des clients et les résultats obtenus sont comparés,

Enfin, une approche spécifique est proposée pour améliorer la précision de la prédiction de la consommation du gaz naturel. Cette approche repose sur l'utilisation d'un modèle hybride combinant les avantages des modèles SARIMA et LSTM.

En conclusion, cette étude vise à utiliser l'intelligence artificielle pour améliorer les méthodes de prévision de la consommation future de l'énergie gazière à Béjaïa. En anticipant les fluctuations de la consommation de gaz naturel, il est possible de prendre des mesures proactives pour optimiser son utilisation et préserver la position d'exportateur de l'Algérie. De plus, en minimisant les impacts économiques négatifs d'une éventuelle pénurie de gaz à Béjaïa, cette étude contribue à assurer une gestion plus efficace et durable de cette ressource essentielle.

Pendant la réalisation de notre projet de fin d'étude, nous avons réussi à apprendre beaucoup de choses et à mettre en pratique tout ce que nous avons étudié durant notre parcours. Nous avons également appris à manipuler le langage Python ainsi que les différentes étapes de réalisation d'un modèle de machine learning. Ces compétences peuvent être utilisées dans le domaine professionnel en tant que data scientist.

Dans les perspectives futures, nous envisageons de développer notre approche pour l'utilisation avec des big data, c'est-à-dire des ensembles de données volumineux et complexes, afin d'améliorer encore la précision de nos prévisions. De plus, nous souhaitons étendre notre modèle pour la détection de fraudes, en particulier dans le contexte des clients frauduleux dans la consommation de gaz naturel. Cela nous permettrait de contribuer à la lutte contre les pratiques illégales et d'assurer une utilisation plus juste et transparente de cette ressource précieuse.

BIBLIOGRAPHIE

[1] Documentation anaconda. https://docs.anaconda.com/. Consulté le 25 mai 2023.

[2] Documentation jupyter. https://docs.jupyter.org/en/latest/. Consulté le 25 mai 2023.

[3] A3E. histoire gaz. https://www.encyclopedie-energie.org/gaz-naturel-une-histoire-tres-ancienne/, 2020. [Online; accessed 15/05/2023].

[4] acteurs du cloud. *deep learning : définition , concept et usages potentiels*. Nuageo, 2020.

[5] A.Géron. Hands-on machine learning with scikit-learn and tensorflow : Concepts. *Tools, and Techniques to build intelligent systems*, 2017.

[6] A. Allmang. Prédiction 3 cluster. [Online; accessed 23/06/2023].

[7] B. Artley. Prévision de séries chronologiques avec arima, sarima et sarimax. https://towardsdatascience.com/time-series-forecasting/, 2022.

[8] J. Audiffren. Fonctionnement msv. https://dataanalyticspost.com/Lexique/svm/, 2017. [Online; accessed 23/06/2023].

[9] J.J. Avoce. *Apprentissage profond distribué sécurisé : application à la détection de fraudes bancaires sur internet*. PhD thesis, Université du Québec en Outaouais, 2021.

[10] F. Benabbas. *Méthode heuristiques pour la prédiction des séries temporelles*. PhD thesis, Annaba, 2011.

[11] N. Bhandari. Extratreesclassifier. https://medium.com/@namanbhandari/extratreesclassifier-8e7fc0502c7/, urldate=03/06/2023, 2022.

[12] G. CHALONS. *Le cycle de vie d'un projet de Machine Learning en 8 étapes*. kaizenSolutions, 2022.

[13] L. Chen. Mehryar mohri, afshin rostamizadeh, and ameet talwalkar : Foundations of machine learning : The mit press, cambridge, ma, 2018, 504 pp., cdn 96.53 (hardback), isbn 9780262039406, 2019.

[14] B. Ejzenberg. Les sous-branches de l'intelligence artificielle, Septembre 2020. [Online ; accessed 20/05/2023].

[15] J. Patterson et A. Gibson. *Deep learning : A practitioner's approach*. " O'Reilly Media, Inc.", 2017.

[16] J.R. Gaikwad et A.B. Deshmane et H.V.Somavanshi et S.V. Patil et R.A. Badgujar. Credit card fraud detection using decision tree induction algorithm. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 4(6) :2278–3075, 2014.

[17] K. Asifullah et al. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53 :5455–5516, 2020.

[18] M. Salehi et al. Predicting national gas consumption in iran using a hierarchical combination of neural networks and genetic algorithms. *International Gas Research Conference Proceedings*, 2 :1165–1169, 01 2014.

[19] O. Laib et al. Forecasting yearly natural gas consumption using artificial neural network for the algerian market. In *2016 4th International Conference on Control Engineering & Information Technology (CEIT)*, pages 1–5. IEEE, 2016.

[20] O.F. Beyca et al. Using machine learning tools for forecasting natural gas consumption in the province of istanbul. *Energy Economics*, 80 :937–949, 2019.

[21] V. Sharma et al. Data-driven short-term natural gas demand forecasting with machine learning techniques. *Journal of Petroleum Science and Engineering*, 206 :108979, 2021.

[22] P. Grosjean et G. Engels. Sciences des données biologiques. https://wp.sciviews.org/sdd-umons2/?iframe=wp.sciviews.org/sdd-umons2-2022/, 2023. [Online ; accessed le 18 mai 2023].

[23] O.A. Bensiah et M.Berkane. La proposition d'une nouvelle approche basée deep learning pour la prédiction du cancer du sein. 2020.

[24] S. Gelper et R. Fried et C. Croux. Robust forecasting with exponential and holt-winters smoothing. juin 2007.

[25] A. Azadeh et Sm. Asadzadeh et M. Saberi et V. Nadimi et A. Tajvidi et M. Sheikalishahi. A neuro-fuzzy-stochastic frontier analysis approach for long-term natural gas consumption forecasting and behavior analysis : The cases of bahrain, saudi arabia, syria, and uae. *Applied Energy*, 88(11) :3850–3859, 2011.

[26] W. Panek et T. Wlodek. Natural gas consumption forecasting based on the variability of external meteorological factors using machine learning algorithms. *Energies*, 15(1) :348, 2022.

[27] I. Goodfellow et Y. Bengio et A. Courville. *Deep learning*. MIT press, 2016.

[28] M. Akpinar et Y. Nejat. Günlük temelli orta vadeli şehir doğal gaz talebinin tek değişkenli istatistik teknikleri ile tahmine. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 35 :725 – 742, 2020.

[29] N. Fabio. *Python Data Analytics*. Apress Media, California, 2018.

[30] R. Frank. The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6) :386, 1958.

[31] GNL101. Initiation au gaz naturel liquéfié. `https://wwww.energiesaguenay.com/upload/filer_public/d1/84/d1847f94-5096-481e-a713-6ee8397977df/fwebenergie_saguenay_2020uploadcms_page_media27gnl_101_-_initiation_au_gaz_naturel_liquefie.pdf/`, 2016.

[32] B. Grossfeld. Deep learning vs. machine learning : a simple way to learn the difference. *Zendesk Blog*, 23, 2020.

[33] L. Hui. Quels algorithmes ml utiliser ?, 2017.

[34] Imaginovation. 8 sectors that benefit from iot development in 2021. `https://imaginovation.net/blog/8-sectors-benefit-from-iot-development-in-2021`, 2022. (consulté le 03/06/2023).

[35] A. Keramane. *L'Électrification de l'Algérie. De la lumière dans les ksours*. Éditions L'Harmattan, octobre 2020. [Online ; accessed 10-April-2023].

[36] A. Labiad. *Sélection des mots clés basée sur la classification et l'extraction des règles d'association*. PhD thesis, Université du Québec à Trois-Rivières, 2017.

[37] Scikit learn Contributors. Scikit-learn documentation. `https://scikit-learn.org/stable/`, N/A.

[38] Education Ecosystem (LEDU). Understanding k-means clustering in machine learning. 2018.

[39] les membres de CAPP. Extraction de gaz naturel. https://www.capp.ca/natural-gas/drilling-and-fracturing/. [Online ; accessed 10-May-2023].

[40] B. Pesquet. *Réseaux de neurones récurrents*. 2021.

[41] F. Pouyet. *Deep Learning et classification d'images*. PhD thesis, Université Paris Sud, 2023.

[42] A. Roche. *Lissage exponentiel*. 2018. [Online ; accessed 15-April-2023].

[43] M. Sauget. *Parallélisation de problèmes d'apprentissage par des réseaux neuronaux artificiels. Application en radiothérapie externe*. PhD thesis, Université de Franche-Comté, 2007.

[44] J. Schmidhuber. Deep learning. *Scholarpedia*, 10(11) :32832, 2015.

[45] V. Shashkina. *Préparation des données pour l'apprentissage automatique : un guide étape par étape*. itrexgroup, 2023.

[46] société eni. Le gaz naturel : une énergie propre aux multi-usages. https://www.eni.com/fr_FR/produits-services/gaz-naturel/energie-propre/energie-propre.shtml, février 2013. [Online ; accessed 10-May-2023].

[47] Sonalgaz. Plan de dÉveloppement 2021-2030. https://www.sonelgaz.dz/fr/835/plan-de-developpement-2021-2030, 2020.

[48] Auteurs spécialisés Ooreka. Moyenne mobile. https://epargne.ooreka.fr/astuce/voir/605103/moyenne-mobile/, 2023. [Online ; accessed 18-May-2023].

[49] Keras Team. Keras documentation. https://keras.io/, N/A.

[50] Data TechNotes. Regression model accuracy (mae, mse, rmse, r-squared) check in r, 2019.

[51] P. Vipul. Le choix de l'algorithme d'apprentissage selon certains facteurs, mai 2022. [Online ; accessed 22/06/2023].

[52] R. Warlop. Apprentissage par renforcement, avril 2019. [Online ; accessed 01/06/2023].

[53] S. Yegulalp. What is tensorflow ? the machine learning library explained. https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html, urldate=03/06/2023.

[54] G. Zaccone. *Getting Started with TensorFlow*, volume 1. Packt Publishing Birmingham, 2016.

[55] E.F. Sánchez Úbeda et A. Berzosa. Modeling and forecasting industrial end-use natural gas consumption. *Energy Economics*, 29(4) :710–742, 2007.

Résumé

Ce mémoire se concentre sur l'utilisation de l'intelligence artificielle pour améliorer la prévision de la consommation future de gaz naturel à Béjaïa, en Algérie. L'objectif principal était de développer un modèle prédictif précis en utilisant une approche combinant l'apprentissage en profondeur (LSTM) et l'analyse de séries temporelles (SARIMA) "Ensemble_LSTM_SARIMA". Le modèle a été appliqué à un ensemble de données fourni par la SONELGAZ CD Béjaïa, comprenant les données de consommation de gaz naturel des clients haute pression de 2014 à 2023. Les résultats ont montré un taux d'erreur de 4,85% pour la prédiction de la consommation d'un seul client et un taux d'erreur de 4,76% pour l'ensemble des clients. Ces résultats démontrent l'efficacité de notre approche dans la prédiction précise de la consommation future de gaz naturel. Cette approche offre une amélioration significative par rapport aux méthodes individuelles. Ce mémoire contribue ainsi à résoudre la problématique de la gestion efficace de cette ressource précieuse en anticipant les fluctuations de la demande et en permettant une utilisation optimale du gaz naturel à Béjaïa.
Mots clés = Consommation de gaz naturel, Sonalgaz, Prévision, Apprentissage automatique, Méthode statistique, Ensemble_LSTM_SARIMA.

Abstract

This thesis focuses on the use of artificial intelligence to improve the prediction of future natural gas consumption in Béjaïa, Algeria. The main objective was to develop an accurate predictive model using a combined approach of deep learning (LSTM) and time series analysis (SARIMA) called "Ensemble_LSTM_SARIMA" The model was applied to a dataset provided by SONELGAZ CD Béjaïa, which included natural gas consumption data from high-pressure customers from 2014 to 2023. The results showed an error rate of 4.85% for predicting the consumption of a single customer and an error rate of 4.76% for the entire customer set. These results demonstrate the effectiveness of our approach in accurately predicting future natural gas consumption. This approach offers a significant improvement compared to individual methods. This thesis contributes to addressing the challenge of effectively managing this valuable resource by anticipating demand fluctuations and enabling optimal use of natural gas in Béjaïa
Key words = Natural gas consumption, Sonalgaz, Forecasting, Machine Learning, statistical methods, _LSTM_SARIMA ensemble.

.