République Algérienne Démocratique et Populaire
Université Abderrahmane MIRA de Béjaia
Faculté des Sciences Exactes

Département de Recherche Opérationnelle

Mémoire Présenté pour l'obtention du diplôme de Master en Mathématiques Appliquées

Spécialité : Sciences de données et aide à la décision

Analyse des séries temporelles avec les algorithmes de machine learning

Présenté par :Maroua Redouane

Sous la direction de : Mme S.Amroun

Défendu le 30 /06/2025, devant le jury composé de :

M^{r} N.Zougab	Professeur Président du jury	UAMB - Béjaia
M^{elle} K.Bouchebbah	M.C. Classe B Examinatrice	UAMB - Béjaia
M^{me} L.Djerroud	M.C Classe B Examinatrice	UAMB -Béjaïa

Année Universitaire 2024-2025

Remerciments

بسم الله الرحمن الرحيم
"يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنْكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ درجات"

Je remercie Dieu de m'avoir guidée, soutenue, et d'avoir placé en moi la force, la patience et le courage nécessaires pour mener à bien ce parcours, jusqu'à la réalisation de ce mémoire.

Je tiens particulièrement à me remercier pour ma persévérance, ma patience et ma détermination. Malgré les épreuves, les doutes et la fatigue, j'ai continué à avancer, pas à pas, avec courage et confiance. Aujourd'hui, je suis fière du chemin parcouru et du travail accompli.

À l'issue de ce parcours, je ressens profondément le besoin d'exprimer ma gratitude à celles et ceux qui ont été ma lumière dans les moments sombres, ma force dans les instants de doute, et mon refuge dans les tempêtes.

Je tiens à exprimer toute ma reconnaissance à mon encadrante, Mme **Sonia Amroun**, pour son encadrement bienveillant, sa disponibilité, ses conseils précieux et sa confiance tout au long de ce travail. Son accompagnement a largement contribué à la qualité de ce mémoire.

Je remercie du fond du cœur **ma famille**, en particulier mes parents, pour leur amour inconditionnel, leurs prières, leur patience et leur soutien indéfectible durant toutes ces années.

Je suis également profondément reconnaissante envers mes tantes et mes cousines **Cyrine** et **Nada**, dont les encouragements constants et la bienveillance m'ont été d'un grand réconfort.

Un merci tout spécial à mon meilleur **R.I**, pour sa présence, ses mots de réconfort, sa motivation et sa confiance en moi, même dans les moments les plus difficiles

Enfin, je remercie sincèrement tous les **enseignants** qui m'ont transmis leur savoir tout au long de mon parcours universitaire. Grâce à leur dévouement, j'ai pu acquérir les connaissances nécessaires à l'aboutissement de ce mémoire.

À toutes et à tous, merci infiniment.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Le monde industriel actuel est marqué par une volatilité croissante des marchés et une concurrence internationale renforcée. Pour les grandes entreprises comme Cevital, leader algérien du secteur agroalimentaire et industriel, anticiper l'évolution de la demande est devenu un enjeu stratégique important. Dans ce contexte, l'analyse des séries temporelles constitue un outil classique et essentiel pour prévoir la demande et planifier efficacement la production, la logistique et la gestion des stocks. Parallèlement, les avancées récentes en machine learning offrent de nouvelles perspectives pour améliorer les prévisions, en exploitant les capacités d'apprentissage des données historiques pour modéliser des dynamiques complexes.

Dans le domaine de la prévision, l'analyse des séries temporelles constitue depuis longtemps des méthodes classique et largement utilisée. Elle permet d'anticiper l'évolution de variables dans le temps, telles que la demande, les ventes ou la production, en s'appuyant sur les données historiques. Ces approches, comme les modèles autorégressif intégré à moyennes mobiles (ARIMA) ou les méthodes de lissage exponentiel, ont prouvé leur efficacité dans de nombreux contextes industriels.

Cependant, avec l'emergonce des technologies de l'information, la croissance exponentielle des volumes de données disponibles et la puissance accrue des systèmes de calcul, de nouvelles approches basées sur l'intelligence artificielle ont émergé. En particulier, le machine learning s'est imposé comme une alternative prometteuse pour affiner les prévisions. Grâce à leur capacité à apprendre des schémas complexes à partir de grandes quantités de données, ces algorithmes permettent de modéliser des dynamiques temporelles plus subtiles que celles captées par les méthodes traditionnelles. Appliqué aux séries temporelles, le machine learning ouvre ainsi de nouvelles perspectives pour améliorer la précision des prévisions, notamment dans le domaine industriel où les enjeux de planification et d'optimisation sont cruciaux.

La mise en place d'un modèle de prévision basé sur les séries temporelles et les techniques de machine learning suit généralement un processus structuré en plusieurs étapes clés. La première étape consiste en la collecte et la préparation des données : il s'agit de rassembler les données historiques pertinentes, de les nettoyer, de traiter les valeurs manquantes, et de structurer les séries temporelles de manière exploitable. Ensuite, une analyse exploratoire est réalisée pour comprendre les tendances, les saisons et les anomalies éventuelles dans les données. La troisième étape concerne le choix et la conception du modèle : selon la nature du problème, on peut opter pour des approches statistiques classiques comme le modèle autorégressif intégré à moyennes mobiles saisonnier (SARIMA), ou pour des algorithmes d'apprentissage automatique tels que méthode de boosting de gradient extrême (XGBoost) ou mémoire à long court terme (LSTM). Une fois les modèles entraînés, une phase d'évaluation permet de comparer

leurs performances à l'aide de plusieurs indicateurs . Et la dernière étape consiste à valider le modèle retenu et à le mettre en application dans un environnement opérationnel, afin de générer des prévisions exploitables pour la prise de décision. Et enfin la visualisation des données et des résultats. Elle permet non seulement d'interpréter plus facilement les prédictions, mais aussi de communiquer efficacement les conclusions auprès des décideurs. En tant que data scientist, cette capacité à traduire les résultats complexes en représentations claires et accessibles est un levier clé pour orienter les actions stratégiques de l'entreprise.

L'objectif principal de ce mémoire est de développer un système de prévision mensuelle des quantités (ou ventes) pour l'année 2025, en s'appuyant à la fois sur des méthodes classiques d'analyse de séries temporelles et sur des techniques avancées de machine learning ; et sur une approche hybride. Ce travail vise à comparer ces approches afin d'identifier les modèles les plus performants, dans le but d'améliorer la précision des prévisions et de soutenir la prise de décision stratégique.

Ce mémoire est structuré en quatre chapitres complémentaires :
Le premier chapitre présente l'entreprise de Cevital dont lequel nous examinerons son histoire, sa structure organisationelle, ses principales filiales et ses accoplissement .
Le deuxième chapitre est consacré aux séries temporelles, avec une introduction aux modèles classiques tels que ARIMA et SARIMA, ainsi qu'aux étapes nécessaires à leur mise en œuvre.
Le troisième chapitre traite des fondements du machine learning, en particulier des algorithmes adaptés à la prévision comme XGBoost et LSTM.
Enfin, le quatrième chapitre constitue la partie pratique du mémoire, où sont décrites les étapes d'implémentation, d'expérimentation, d'évaluation et de comparaison des modèles appliqués à la prévision mensuelle des quantités pour l'année 2025, y compris le développement du modèle hybride et la création du tableau de bord avec Power BI.

Ce mémoire se termine par une conclusion générale qui résume les principaux résultats obtenus concernant l'application des modèles des séries temporelles et du machine learning dans la prévision.

1

Présentation de l'entreprise Cevital

Introduction

Dans un contexte économique mondial en constante évolution, marqué par la libéralisation des marchés et la diversification des activités industrielles, certaines entreprises émergent comme des piliers du développement national.
En Algérie, le Groupe **Cevital** incarne cette dynamique. Fondé en 1998 par l'entrepreneur Issad Rebrab, Cevital s'est rapidement imposé comme le premier groupe privé du pays, jouant un rôle central dans la transformation économique de l'Algérie.

Initialement spécialisé dans l'agro-industrie, notamment la production de sucre et d'huiles végétales, Cevital a su diversifier ses activités pour devenir un conglomérat présent dans plusieurs secteurs clés tels que l'électroménager, la logistique, la distribution, la construction et les services. Cette diversification stratégique a permis au groupe de renforcer sa résilience face aux fluctuations économiques et de s'étendre au-delà des frontières nationales, avec des implantations en Europe, en Afrique et au Moyen-Orient. Ce chapitre a pour objectif de fournir une vue d'ensemble du Groupe Cevital, en explorant son historique, sa structure organisationnelle, ses domaines d'activité et son positionnement économique. Cette analyse permettra de mieux comprendre les facteurs qui ont contribué à son succès et les défis auxquels il est confronté dans un environnement économique en mutation.[**1**]

1.1 Identité de l'entreprise

1.1.1 Dénomination sociale, forme juridique, date de création et fondateur :

Le Groupe Cevital est une société par actions (SPA) fondée en mai 1998 par Issad Rebrab, un entrepreneur algérien visionnaire. Issad Rebrab, né en 1944, a débuté sa carrière en créant un cabinet d'expert-comptable en 1968, avant de se lancer dans l'entrepreneuriat en 1971 en

fondant des entreprises dans la métallurgie et la sidérurgie. Sous sa direction, Cevital est devenu le premier groupe privé algérien, opérant dans divers secteurs industriels et de services . [**1**]

1.1.2 Siège social et implantations principales :

Le siège social du Groupe Cevital est situé à **l'Îlot D, N°6 ZHUN Garidi II, Kouba, 16005 Alger, Algérie**, dans la capitale du pays. C'est à partir de cette adresse que sont coordonnées les principales décisions stratégiques et les activités administratives du groupe. [**1**]

FIGURE 1.1 – Le siège social du Groupe Cevital

Le Groupe Cevital dispose également d'**implantations internationales** qui témoignent de sa stratégie de diversification et d'expansion à l'échelle mondiale. Parmi les plus importantes, on peut citer :

— **OXXO Evolution** à *Cluny, France* et **Alas Iberia** à *Langréo, Asturies, Espagne*,**Aferpi / Lucchini** à *Piombino, Italie*.

Ces implantations reflètent la volonté de Cevital de s'imposer comme un acteur industriel global, en renforçant sa présence sur les marchés européens et méditerranéens.[**1**]
La carte 1.2 illustre cette répartition géographique de ses principales filiales et investissements :

FIGURE 1.2 – Carte des implantations internationales du Groupe Cevital

1.2 Historique et évolution de Cevital

1.2.1 Les débuts (1998–2005)

Le Groupe Cevital a été fondé en 1998 . La première activité du groupe a été l'agro-industrie, avec l'établissement d'une raffinerie d'huile et de sucre à Béjaïa, marquant ainsi le début de son expansion industrielle .[**1**]

1.2.2 Diversification industrielle (2006–2012)

Au cours de cette période, Cevital a connu une croissance significative, diversifiant ses activités au-delà de l'agroalimentaire. En 2007, la création de Mediterranean Float Glass (MFG) a marqué l'entrée du groupe dans l'industrie du verre . La même année, Cevital a lancé Numilog, une filiale spécialisée dans la logistique et la gestion de la chaîne d'approvisionnement .

1.2.3 Expansion internationale (2013–2015)

Cevital a poursuivi son expansion internationale en acquérant en 2013 l'entreprise française Oxxo, spécialisée dans la menuiserie PVC . En 2014, le groupe a repris les activités françaises du groupe Fagor-Brandt, renforçant ainsi sa présence dans le secteur de l'électroménager . La même année, Cevital a acquis les aciéries Lucchini à Piombino, en Italie .

1.2.4 Consolidation et gouvernance (2016–2022)

En juin 2022, Issad Rebrab a officiellement quitté la direction de Cevital, cédant la présidence à son fils Malik Rebrab . Cette transition marque une nouvelle ère pour le groupe, qui continue de consolider sa position en tant que leader industriel en Algérie et en Afrique.

1.3 Structure organisationnelle de Cevital

Le Groupe Cevital adopte une structure organisationnelle **multidivisionnelle**, lui permettant de gérer efficacement ses diverses activités industrielles, commerciales et de services à travers ses nombreuses filiales nationales et internationales.[**1**][**hafsi2013cevital**]

La gouvernance du groupe est assurée par un **Conseil d'administration**, présidé par **Malik Rebrab**, avec la participation d'autres membres de la famille Rebrab, notamment Issad, Omar, Salim, Lynda et Yassine Rebrab.

a) Les principaux pôles d'activités Cevital est structuré autour de plusieurs pôles stratégiques Pôle l'agro-industrie et industrie et le services.

b) Fonctionnement transversal Des fonctions transversales, telles que les ressources humaines, la finance, la communication, la conformité et la stratégie, sont centralisées au niveau du siège social pour assurer la cohérence de la politique du groupe.

c) Gouvernance et pilotage Le pilotage stratégique est assuré par :

- Le **Président Directeur Général (PDG)** du Groupe
- Un **comité de direction** composé des directeurs des principaux pôles
- Des **conseils d'administration** dans certaines filiales importantes

Cette structure offre à Cevital la flexibilité nécessaire pour s'adapter aux évolutions du marché tout en conservant un pilotage centralisé et cohérent de ses activités.
Afin de mieux comprendre la répartition des rôles au sein de cevital , l'organigramme de la macro-structure d Agro-Industrie, détaillant la répartition des principales directions et pôles au sein de l'unité de production agroalimentaire de Cevital . [**2**]

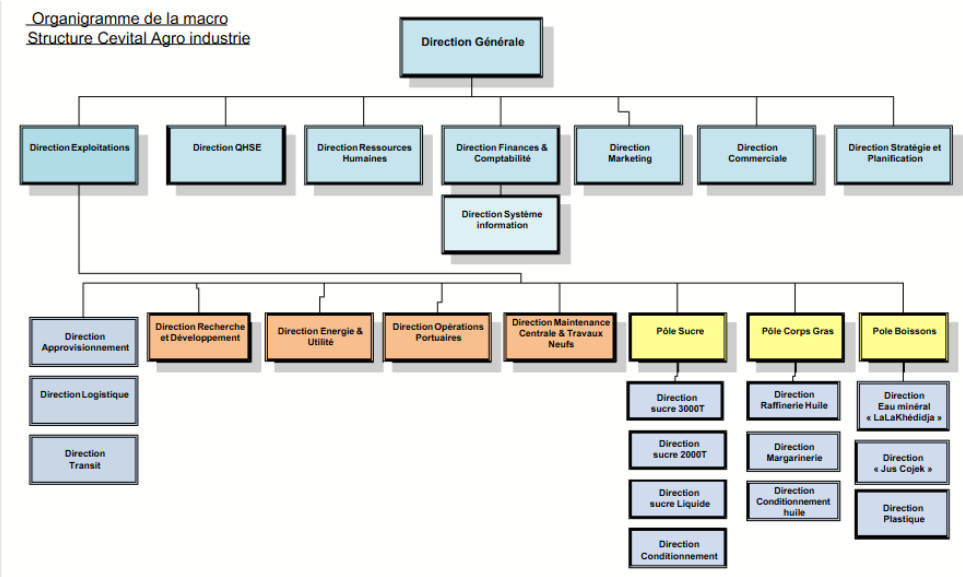

FIGURE 1.3 – Organigramme de la structure organisationnelle du Groupe Cevital

1.3.1 Pôle agro-industriel : cœur de l'activité de Cevital :

1.3.1.1 Cevital Agro-Industri :

Le pôle **Cevital Agro-Industrie** Créée en 1998, situé dans la zone portuaire de Béjaïa , est le plus grand complexe agroalimentaire privé d'Algérie a 200 ML du quai. Il comprend un ensemble d'unités modernes et automatisées, structurées autour de la transformation et du conditionnement de produits de grande consommation.

Voici les principales unités de production :

- **2 raffineries de sucre** , avec une capacité de production 3000 et 3500 tonnes (A Béjaia)
- **2 Unités de conditionnement de sucre**.
- **Unités de sucre liquide , et une unitésde production de sucre roux .**
- **1 raffinerie d'huile végétale**(A Béjaia), assurant la transformation des huiles brutes importées .

FIGURE 1.4 – Vue aérienne du port de Béjaïa, principal site industriel de Cevital Agro-Industrie

— **Unités de conditionnement d'huile**, destinées à l'embouteillage et à la distribution nationale et internationale .
— **1 margarinerie**(A Béjaia), pour la fabrication de différentes formes de margarine industrielle et domestique .
— **Unités de fabrication et de conditionnement de boissons fruitées et de production de conserves et de confitures**,(A El Kseur).
— **Unités de production de souces**(A El Kseur).
— **1 unité de conditionnement d'eau minérale et gazeifiée** .Lala Khedidja

FIGURE 1.5 – Bouteilles d'huile végétale produites par Cevital

FIGURE 1.6 – Bouteilles d'eau minérale et de jus produites par Cevital

Grâce à cette infrastructure, Cevital répond aux besoins nationaux tout en développant fortement ses exportations, notamment dans les domaines du sucre raffiné, de l'huile, de la margarine et des boissons. Le complexe est relié directement au port de Béjaïa, facilitant l'importation des matières premières et l'exportation des produits finis à grande échelle.

1.3.2 Pôle Industrie

— **Brandt Algérie** : Spécialisée dans la fabrication d'appareils électroménagers.

— **Mediterranean Float Glass (MFG)** : Créée en 2007, MFG est le plus grand producteur de verre plat en Afrique, avec une capacité de production de 600 tonnes/jour. Elle exporte vers plusieurs pays, notamment en Europe.

— **Oxxo Évolution** : Acquise en 2013, cette filiale française est spécialisée dans la fabrication de menuiseries en PVC et aluminium.

1.3.3 Pôle Services et Distribution

Ce pôle englobe des activités liées à la logistique, la grande distribution et les services :

— **Numilog** : Créée en 2007, cette filiale offre des services logistiques complets, incluant le transport terrestre, le stockage, la gestion des stocks, l'emballage et le co-packing. Elle dispose de plusieurs plateformes en Algérie et à l'international.

— **Numidis** : Spécialisée dans la grande distribution, cette filiale gère les enseignes "Uno" et "Unocity", avec plusieurs supermarchés et hypermarchés à travers l'Algérie.

— **Sierra Cevital** : Joint-venture créée en 2011 cette filiale est spécialisée dans le développement et la gestion de centres commerciaux.

1.4 Chiffres clés et performances

Le Groupe Cevital, en tant que premier groupe privé algérien, affiche des performances économiques notables et une forte présence sur le marché national et international.

a) Effectifs et répartition géographique

Cevital emploie environ **18 000 salariés** répartis sur l'ensemble du territoire national, avec une concentration importante dans les unités de production situées à Béjaïa, Sétif, Blida et Alger. À l'international, le groupe compte également des centaines d'employés en France, en Italie et en Espagne.

b) Chiffre d'affaires et parts de marché

Le chiffre d'affaires annuel du groupe s'élève à environ **3,5 milliards de dollars** (données variables selon les années). Cevital détient des parts de marché dominantes dans plusieurs secteurs , comme le Premier producteur de sucre en Afrique. Leader de la distribution de produits agroalimentaires en Algérie et acteur majeur de l'électroménager via Brandt Algérie.

c) Investissements et projets majeurs

Cevital a investi massivement dans la modernisation de ses installations, la logistique, et l'expansion à l'international. Parmi les projets majeurs la plateforme logistique de Béjaïa. Le complexe sidérurgique Aferpi en Italie et le développement de pôles industriels intégrés à Sétif et Blida.

1.5 Responsabilité sociétale et développement durable

Conscient de son rôle au sein de la société, Cevital s'engage activement dans une démarche de responsabilité sociétale, à travers plusieurs initiatives :

1.5.1 Engagements environnementaux

Cevital investit dans des technologies propres et dans le recyclage de ses déchets industriels, notamment dans ses unités de production de sucre et d'huile. Le groupe œuvre à réduire ses émissions et sa consommation énergétique par des processus plus durables.[**3**]

1.5.2 Initiatives sociales et communautaires

Le groupe participe à des actions de mécénat, de soutien aux écoles, de développement local, et emploie une politique de recrutement favorisant les jeunes diplômés algériens. Il soutient également des événements culturels, sportifs et éducatifs.

1.5.3 Retour d'expérience : stage au sein de Cevital

Dans le cadre de mon stage de fin d'études chez Cevital, j'ai initialement intégré le service de gestion des stocks, où j'ai été impliqué dans le suivi dans les deux départements des produits finis et des matières premières . Cette première phase m'a permis de comprendre les mécanismes de gestion des flux physiques et les outils utilisés pour assurer une disponibilité optimale des produits.

Durant ces deux phases, j'ai utilisé le logiciel de gestion intégré **Sage**, qui s'est révélé être un outil essentiel pour le suivi en temps réel des mouvements de stock, l'optimisation des niveaux de stock et la génération de rapports détaillés facilitant la prise de décision.

Par la suite, afin de collecter et d'analyser des données pertinentes pour mon mémoire, j'ai été amené à collaborer avec le département commercial. Cette transition m'a offert une vision plus globale de la chaîne logistique, en mettant en évidence l'interconnexion entre la gestion des stocks et les opérations commerciales.

Cette expérience m'a permis de développer une compréhension approfondie des processus opérationnels de Cevital, en particulier de la manière dont la digitalisation, via des outils comme Sage, contribue à l'efficacité et à la réactivité de l'entreprise.

Conclusion

Le Groupe Cevital se distingue comme un pilier de l'économie algérienne, combinant diversification sectorielle, expansion internationale et engagement sociétal. Avec plus de 26 filiales opérant dans des domaines variés tels que l'agroalimentaire, l'industrie, la distribution et les services, Cevital illustre une stratégie de croissance ambitieuse et cohérente.

Son modèle de gouvernance, marqué par une gestion familiale et une vision à long terme, lui a permis de s'imposer comme le premier groupe privé algérien. Les investissements stratégiques, tant sur le plan national qu'international, témoignent de sa volonté de renforcer sa position sur le marché et de contribuer activement au développement économique du pays. Par ailleurs, Cevital accorde une importance particulière à la responsabilité sociétale, en initiant des actions en faveur de l'environnement et en soutenant des projets sociaux et communautaires. Cette approche intégrée, alliant performance économique et engagement citoyen, positionne Cevital comme un acteur clé dans la construction d'un avenir durable et équitable pour l'Algérie.

2
Séries temporelles

Introduction

Dans le domaine de la prédiction temporelle, les séries temporelles constituent un domaine clé de l'analyse statistique, se concentrant sur l'étude de données indexées par le temps. Une série temporelle est une suite d'observations collectées au cours du temps, Dans le cadre de l'analyse des séries temporelles, la prévision vise à anticiper les valeurs futures d'une série chronologique en exploitant ses données antérieures.
Ce chapitre explorera les composantes des séries temporelles, leur importance et quelques modèles statistiques classiques pour la prédiction temporelle.

2.1 Série temporelle

Une **série temporelle** (ou chronologique) est une suite de valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps, enregistrée à intervalles réguliers ou irréguliers. On supposera qu'il s'agit d'une réalisation d'un processus X, c'est-à-dire d'une suite $\{X_i\}$ de variables aléatoires, souvent exprimée mathématiquement pour analyser son comportement passé et prévoir son comportement futur.[**5**]

2.2 Objectifs de l'analyse des séries temporelles

Les objectifs les plus importants des séries chronologiques sont les suivants :

1. **Description de la série chronologique** : En représentant leur valeur dans un graphique et en trouvant quelques métriques statistiques pour l'estimation des composants des séries chronologiques. Leurs propriétés sont reconnues comme : croissantes, décroissantes, ou stables, etc.

2. **Interprétation de la série chronologique** : En interprétant une variable par le temps ou en comparant son comportement avec celui de la même variable dans le passé.
3. **Prévision** : La prévision est l'un des objectifs majeurs de l'analyse des séries temporelles, et repose sur la dépendance des données par rapport aux séries historiques.
4. **Contrôler le phénomène aléatoire si possible** : L'analyse des séries temporelles permet de mieux comprendre et modéliser les phénomènes aléatoires qui influencent les données, tels que les chocs externes ou les variations imprévisibles. Cela permet de réduire l'incertitude et d'optimiser les prises de décisions.
5. **Prise de décisions** : Les séries temporelles permettent de prendre des décisions stratégiques basées sur des modèles de prévision précis, ce qui est essentiel pour la gestion des stocks, l'allocation des ressources, la planification financière, etc. Elles aident à minimiser les risques et à maximiser l'efficacité des processus décisionnels.

Exemples des séries temporelles

Les séries temporelles représentent des ensembles de données recueillies à intervalles au cours du temps. Elles sont utiles pour analyser la progression d'un phénomène et réaliser des prévisions. Voici quelques exemples de séries temporelles, organisés par secteur :
Économie et Finance : Prix des actions : Fluctuation quotidienne du prix d'une action sur le marché boursier.
Agroalimentaire et Logistique : Demande d'un produit : Volume vendu quotidiennement ou mensuellement.
Stock de matières premières : Quantité d'huile brute disponible dans les réserves
Météorologie et Climatologie : Températures quotidiennes : Variations des températures moyennes journalières.
Santé : Cas de grippe ou de COVID-19 : Total des cas rapportés chaque jour/semaine.
Industrie et Production : Production mensuelle d'une usine : Quantité de tonnes produites au cours du mois.
Transport et Mobilité : Nombre de passagers dans un aéroport : Quantité de voyageurs comptabilisés chaque jour.

2.3 Composantes d'une série temporelle

Tendance

Une tendance notée $T(t)$ existe quand il y a une augmentation, une diminution ou une stabilisation à long terme dans la variable observée. En d'autres termes, la tendance traduit le comportement moyen de la série temporelle. Il n'y a pas de techniques bien spécifiques pour identifier la tendance dans une série temporelle. Cependant, si la tendance est strictement croissante ou décroissante, il est facile de l'identifier visuellement avec une représentation temporelle. [**9**][**10**]

Saisonnalité

La saisonnalité notée$S(t)$, est un phénomène dans une série temporelle qui se répète à des intervalles réguliers, généralement inférieurs à un an, comme des cycles hebdomadaires, mensuels ou trimestriels. Elle est souvent causée par des facteurs récurrents tels que la météo, les vacances ou les habitudes de consommation, et se caractérise par des variations périodiques et prévisibles des données. [**9**]

Bruit (ou Résidu)

Le Bruit notée ϵ_t, dans l'analyse des séries temporelles, les bruits représentent la variation irrégulière, souvent assimilée au bruit statistique, similaire aux erreurs dans les modèles statistiques. Ce bruit aléatoire peut être corrélé dans le temps ou non. Lorsqu'il n'y a pas de corrélation, on parle de bruit blanc.il revêt une importance fondamentale dans la théorie et la pratique de l'analyse des séries temporelles car il permet d'évaluer la qualité des modèles. Les bruits d'un modèle ajusté forment une série temporelle propre, où des bruits non corrélés et distribués normalement indiquent un bon ajustement du modèle aux données.[**9**]
Bruit blanc : Un bruit blanc est un processus $\{\varepsilon_t\}_{t=1}^{T}$ qui vérifie :

$$E[\varepsilon_t] = 0 \quad , \quad E[\varepsilon_t^2] = \sigma^2, \quad E[\varepsilon_t \varepsilon_s] = 0, \quad \text{pour } s \neq t.$$

Cela signifie que la série a une moyenne constante (nulle), une variance constante et une covariance nulle entre deux observations, quelle que soit leur distance .

Cycle

Une variation cyclique notée $C(t)$ se produit lorsque la variable observée présente des mouvements d'augmentation et de diminution à une fréquence irrégulière. De façon générale, la durée moyenne d'un cycle est beaucoup plus longue que celle de la composante saisonnière. De plus, les mouvements durant un cycle ont tendance à avoir plus de variabilité que pendant la composante saisonnière.[**10**]

Exemple La figure 2.1 présente la décomposition de la série temporelle en ses différentes composantes :

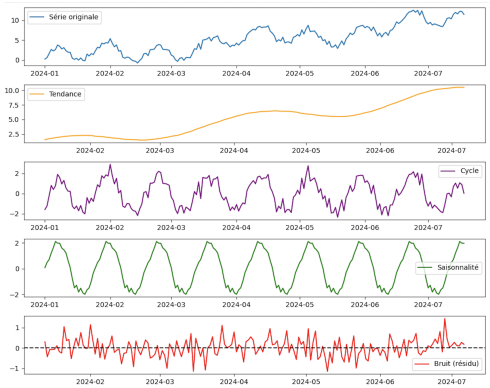

FIGURE 2.1 – Décomposition de la série temporelle

2.4 Techniques de décomposition-recomposition

Les techniques de décomposition-recomposition sont des schémas utilisées pour analyser les séries temporelles .Voici les trois techniques de décomposition-recomposition :

2.4.1 Modèle additif

Une méthode de décomposition fondamentale est la suivante :

$$X_t = T_t + S_t + \epsilon_t, \quad t = 1, \ldots, T.$$

— X_t est la valeur observée à l'instant t.

— $T(t)$ est une composante de tendance déterministe qui reflète le comportement à long terme de la variable observée, comme une croissance ou une décroissance linéaire ou quadratique. Cette composante peut également varier selon les périodes, par exemple en adoptant une forme affine par morceaux.

— $S(t)$ est une séquence périodique correspondant à une composante saisonnière, par exemple avec une période de 12 pour les séries de trafic de passagers .

— ϵ_t est le bruit il représente une composante irrégulière et aléatoire, généralement de faible amplitude par rapport à la composante saisonnière, mais significative en pratique car elle est souvent autocorrélée.

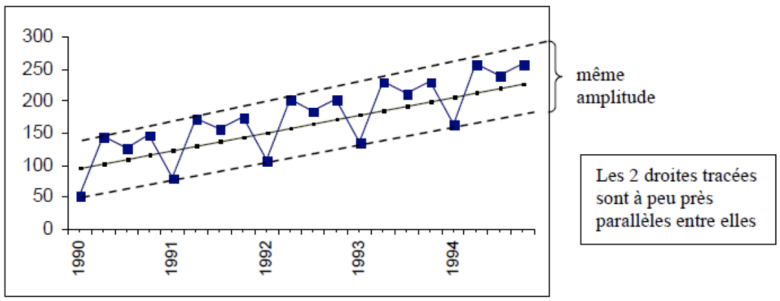

FIGURE 2.2 – Représentation graphique d'un modèle additif

Graphiquement, l'amplitude des variations est constante autour de la tendance.

2.4.2 Modèle multiplicatif

il est le plus utilisé en Èconomie , qui est sous la forme suivante :

$$X_t = T_t \times S_t \times \epsilon_t, \quad , t = 1, \ldots, T.$$

Cette formulation indique que les variations saisonnières et les résidus dépendent de la tendance. Si la tendance augmente, les variations saisonnières augmentent également, ce qui n'est pas le cas dans un modèle additif. Graphiquement, l'amplitude des variations (saisonnières) varie.

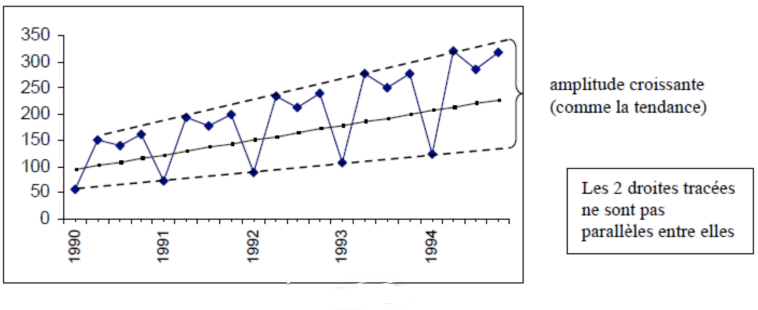

FIGURE 2.3 – Représentation graphique d'un modèle multiplicatif

2.4.3 Modèle multiplicatif complet

Il s'agit des différentes combinaisons de modéle additif et multiplicatif, par exemple :

$$X_t = t_t(1 + s_t)(1 + \epsilon_t), t = 1, \ldots, T$$

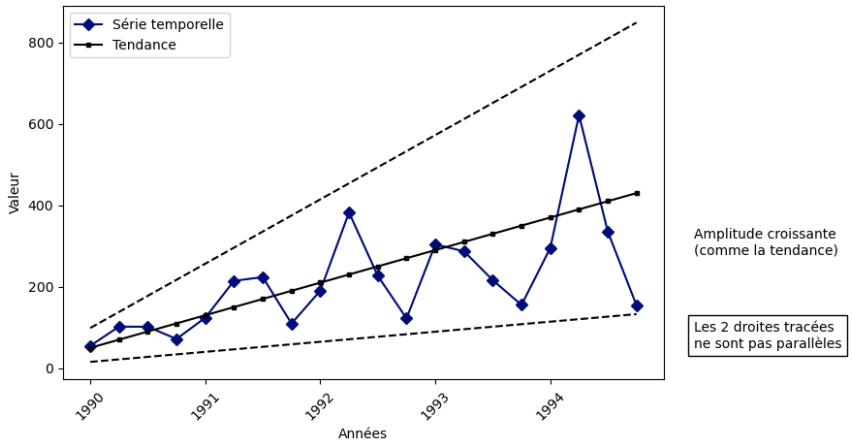

FIGURE 2.4 – Représentation graphique d'un modèle multiplicatif complet

Remarque :

Afin de faire la distinction entre les types de modèle, on peut se baser sur une méthode graphique. Sur le graphique de la série, on trace une droite passant par les minima de la courbe et une autre droite passant par les maxima pour chaque saison. Cette méthode s'appelle la méthode de Bande. Si ces droites sont paralléles, c-à-d l'amplitude de la composante saisonnière et de l'erreur reste constante autour de la tendance au cours du temps, on est en présence d'un modèle additif.
Par contre, si les droites ne sont pas paralléles, c-à-d l'amplitude de la composante saisonniére varie de façon exponentielle dans le temps, on est en présence d'un modèle multiplicatif.

2.5 Stationnarité des séries temporelles

La stationnarité dans les séries temporelles est une propriété importante qui indique que les caractéristiques statistiques de la série ne changent pas avec le temps. L'importance de la stationnarité réside dans le fait que de nombreux modèles de prévision, supposent que la série est stationnaire pour fonctionner correctement. Une série non stationnaire peut conduire à des résultats erronés et à des prévisions peu fiables.

2.5.1 Types de stationnarité

2.5.1.1 Stationnarité strict

Une série temporelle est strictement stationnaire si sa distribution statistique (moyenne, variance et corrélation) reste constante dans le temps. X_t est dite **strictement** (ou fortement) stationnaire si, pour tout entier positif k et pour tout $(t_1, \ldots, t_k) \in \mathbb{T}^k$ et tout décalage $h \in \mathbb{Z}$, les distributions jointes de $(X_{t_1}, \ldots, X_{t_k})$ et $(X_{t_1+h}, \ldots, X_{t_k+h})$ sont identiques.

2.5.1.2 Stationnarité du second ordre

Une série temporelle X_t est dite **stationnaire du second ordre** (ou **stationnaire faible**) si elle vérifie les trois conditions suivantes :

1. **Moyenne constante** : La moyenne de la série ne dépend pas du temps.

$$\mathbb{E}[X_t] = \mu, \quad \forall t.$$

2. **Variance constante et finie** : La dispersion des valeurs autour de la moyenne est constante.

$$\mathrm{Var}(X_t) = \sigma^2, \quad \forall t.$$

3. **Autocorrélation ne dépendant que du décalage temporel** : La covariance entre deux valeurs de la série ne dépend que du décalage h et non des instants t et $t + h$.

$$\gamma(h) = \mathrm{Cov}(X_t, X_{t+h}).$$

La stationnarité joue un rôle essentiel dans l'analyse des séries temporelles et est particuliè-rement pertinente en ce qui concerne les modèles des séries temporelles. On suppose que les données sont stationnaires pour pouvoir appliquer ces modèles de façon appropriée. Cela peut

être évalué par des tests statistique, ou vérifié graphiquement dans la série chronologique pour détecter toute tendance.Si la série ne présente pas de stationnarité, il existe des techniques de transformation pour rendre la série stationnaire.

2.5.2 Tests de stationnarité

Pour vérifier si une série est stationnaire, on utilise des tests statistiques :

2.5.2.1 Test de Dickey-Fuller Augmenté (ADF)

Ce test vérifie si la série possède une racine unitaire, ce qui indique une tendance non stationnaire.
Hypothèses :

— H_0 : La série a une racine unitaire (non stationnaire).

— H_1 : La série est stationnaire.

Décision : Si la p-valeur < 0.05, on rejette H_0 et on considère la série comme stationnaire.

2.5.2.2 Test KPSS (Kwiatkowski-Phillips-Schmidt-Shin)

Contrairement au test ADF, ce test vérifie si la série est stationnaire autour d'une tendance déterministe.
Hypothèses :

— H_0 : La série est stationnaire.

— H_1 : La série est non stationnaire.

Décision : Si la p-valeur < 0.05, on rejette H_0 et on considère la série comme non stationnaire.

2.5.2.3 Test de Phillips-Perron (PP)

Ce test est similaire au test ADF mais il est plus robuste aux hétéroscédasticités (variances non constantes).
Hypothèses :

— H_0 : La série a une racine unitaire (non stationnaire).

— H_1 : La série est stationnaire.

Décision : Comme pour le test ADF, une p-valeur < 0.05 indique que la série est stationnaire.

2.5.3 Comment rendre une série stationnaire ?

Si une série temporelle est **non stationnaire**, on peut appliquer plusieurs transformations pour la rendre stationnaire :

2.5.3.1 Différenciation (Differencing)

— La **différenciation** consiste à soustraire une valeur passée de la valeur actuelle :

$$X'_t = X_t - X_{t-1}.$$

— Cela permet d'éliminer les tendances linéaires.

— Pour des tendances plus complexes, on peut appliquer une **différenciation d'ordre supérieur** :

$$X''_t = X'_t - X'_{t-1}.$$

2.5.3.2 Transformation logarithmique

— Utilisée lorsque la **variance** de la série change au fil du temps.

— On applique une **transformation logarithmique** pour stabiliser la variance :

$$X'_t = \log(X_t).$$

— Fonctionne bien pour les séries avec une **croissance exponentielle**.

2.5.3.3 Transformation par racine carrée ou puissance

— Alternative à la transformation logarithmique si les valeurs contiennent des zéros ou des nombres négatifs :

$$X'_t = \sqrt{X_t},$$

$$X'_t = X_t^{\lambda}, \quad 0 < \lambda < 1.$$

2.5.3.4 Suppression de la tendance (Trend Removal) :

— Si une série suit une tendance, on peut la modéliser et la supprimer avec une **régression linéaire** :

$$X_t = at + b + \varepsilon_t.$$

— Après ajustement, on analyse les **résidus**, qui doivent être stationnaires.

2.5.3.5 Suppression de la saisonnalité :

— Si la série présente une **saisonnalité**, on peut appliquer une **différenciation saisonnière** :

$$X'_t = X_t - X_{t-s}.$$

— où s est la période saisonnière .

2.6 Analyse et caractérisation des séries temporelles

L'analyse et la caractérisation des séries temporelles sont essentielles pour comprendre l'évolution des données dans le temps et faire des prédictions éclairées. Et les principales méthodes utilisées sont : [**10**]

2.6.1 Fonction d'auto-correlation ACF

La fonction d'autocorrélation (ACF), souvent utilisée pour comparer plusieurs séries entre elles, mesure la corrélation entre une série et elle- même à différents intervalles de temps. Dans le cas d'un processus discret $X = X_1, ..., X_n$.
L'ACF est particulièrement utile pour choisir l'ordre **q** pour l'un des modèle des séries temporelle, en observant le dernier décalage significatif avant que les valeurs ne deviennent non significatives. la fonction d'autocorrélation $\rho(t, t+h)$ est définie comme la covariance corrigée de la variance :

$$\rho(h) = \frac{\mathrm{Cov}(X_t, X_{t+h})}{\sqrt{\mathrm{Var}(X_t) \cdot \mathrm{Var}(X_{t+h})}}.$$

L'estimation de la fonction d'autocorrélation se fait par la fonction d'autocorrélation empirique :

$$\hat{\rho}(t, t+h) = \frac{\hat{C}(t, t+h)}{\sqrt{\hat{C}(t,t) \cdot \hat{C}(t+h, t+h)}},$$

où :

— $\hat{C}(t, t+h)$ est l'estimation empirique de la covariance :

$$\hat{C}(t, t+h) = \frac{1}{n} \sum_{i=1}^{n-h} (X_i - \bar{X})(X_{i+h} - \bar{X}).$$

— $\hat{C}(t,t)$ et $\hat{C}(t+h, t+h)$ sont les estimations de la variance :

$$\hat{C}(t,t) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

avec $\bar{X}$ la moyenne empirique :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Si $\rho(t, t+h)$ ne dépend pas du temps (séries stationnaires d'ordre deux), alors on estime $\hat{\rho}(h) = \hat{\rho}(t, t+h)$ pour tout t. Pour une réalisation donnée,

les corrélations $\hat{\rho}(t, t+h)$ peuvent être visualisées en traçant des nuages de points de (x_i, x_{i+h}) pour $i = 1, \ldots, n-h$.

La représentation graphique de la fonction d'autocorrélation est appelée corrélogramme comme l'illustre la figure 2.5 :

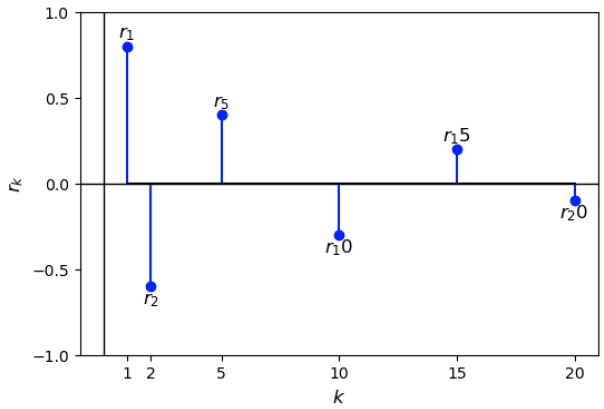

FIGURE 2.5 – Exemple de corrélogramme.

2.6.2 Fonction d'autocorrélation partielle (PACF)

La fonction d'autocorrélation partielle (PACF) évalue la corrélation entre $y(t)$ et son retard $y(t-k)$, en tenant compte de l'influence des valeurs $y(t-k-i)$, pour tout $i > k$. Le corrélogramme partiel, représentant le PACF pour différents ordres k, permet de visualiser cette relation. La PACF est particulièrement utile pour identifier l'ordre **p** pour l'un des modèle des séries temporelle, en repérant le dernier retard significatif avant que les valeurs ne deviennent non significatives, après avoir retiré l'influence des termes intermédiaires. Elle est obtenue en résolvant la relation de Yule-Walker :

$$\phi_{kk} = \rho(k) - \sum_{i=1}^{k-1} \phi_{ki}\rho(k-i);$$

où :

— ϕ_{kk} est l'autocorrélation partielle à l'ordre k.

— $\rho(k)$ est l'autocorrélation simple à l'ordre k.

On utilise souvent la méthode de **Durbin-Levinson** pour estimer ces valeurs.

2.7 Modèles de prévision en séries temporelles

2.7.1 Modèles Autorégressifs (AR) :

En statistique et en économétrie, un modèle autorégressif (AR) est un type de modèle utilisé pour analyser et prédire des séries temporelles. Il repose sur l'idée que les valeurs passées d'une variable peuvent être utilisées pour expliquer ou prédire ses valeurs futures.

La définition formelle d'un modèle AR d'ordre p (noté AR(p)) est la suivante :

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t,$$

où :

— X_t est la valeur de la série à un instant t,

— c est une constante,

— ϕ_i sont les coefficients du modèle qui mesurent l'influence des valeurs passées,

— p est l'ordre du modèle, c'est-à-dire le nombre de valeurs passées prises en compte,

— ε_t est un terme d'erreur ou bruit blanc, supposé indépendant et identiquement distribué.

Le modèle AR est dit « autorégressif » car il se réfère à lui-même, en utilisant ses propres valeurs passées comme variables explicatives. Ces modèles sont largement utilisés dans divers domaines comme la finance, l'économie et la météorologie pour modéliser des phénomènes dynamiques.

2.7.2 Modèle moyenne mobile(MA) :

Dans un modèle MA, la variable de sortie est une combinaison linéaire des erreurs passées, ce qui signifie que chaque valeur de la série est influencée par les erreurs aléatoires $\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$, qui se sont produites à des moments précédents. Ce modèle est souvent noté MA(q), où q représente le nombre de termes d'erreur passés utilisés dans le modèle.

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

— X_t est la valeur de la série temporelle à l'instant t,

— μ est la moyenne de la série,

— ε_t est le terme d'erreur à l'instant t,

— θ_i sont les coefficients de moyenne mobile pour les erreurs passées.

— q est l'odre du modèle.

2.7.3 Modèle autoregressive moving average (ARMA) :

Il combine deux composantes principales la partie autorégressive et la partie moyenne mobile :
AR : La partie autorégressive, qui modélise la relation entre une observation et ses valeurs passées.
MA : La partie moyenne mobile, qui modélise la relation entre une observation et les erreurs passées du modèle.
Un modèle ARMA est noté ARMA(p, q), où :
p est l'ordre de la partie autorégressive (AR), et **q** est l'ordre de la partie moyenne mobile (MA).
La formule générale d'un modèle ARMA(p, q) est :

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

où :

— X_t est la valeur de la série à l'instant t,
— $\phi_1, \phi_2, \ldots, \phi_p$ sont les coefficients de la partie AR,
— $\theta_1, \theta_2, \ldots, \theta_q$ sont les coefficients de la partie MA,
— ε_t est un bruit blanc (terme d'erreur aléatoire indépendant et identiquement distribué).

Remarque : Le modèle ARMA suppose que la série temporelle est stationnaire, c'est-à-dire que ses propriétés statistiques ne changent pas au fil du temps. Si la série n'est pas stationnaire, il peut être nécessaire de la différencier pour obtenir un processus stationnaire, ce qui conduit au modèle ARIMA.

2.7.4 Modèle autoregressive integrated moving average (ARIMA) :

Le modèle ARIMA est une combinaison de ce processus de différenciation et du processus ARMA classique, le **I** du modèle ARIMA signifie **'integrated'** pour intégration. En différenciant les séries temporelle pour la stationnariser, c'est-à-dire éliminer les tendances linéaires ou quadratiques .
Un modèle ARIMA est noté ARIMA(p, d, q), où :

— p est l'ordre de la partie (AR),
— d est l'ordre de **différenciation** nécessaire pour stationnariser la série,
— q est l'ordre de la partie (MA).

La formule générale d'un modèle **ARIMA**(p, d, q) peut être exprimée comme suit :

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)(1 - L)^d X_t = \mu + (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t$$

où :

— X_t est la valeur de la série temporelle à l'instant t,
— ϕ_i sont les coefficients **autorégressifs** (AR),
— θ_i sont les coefficients de **moyenne mobile** (MA),
— L est l'**opérateur de retard**, défini par : $LX_t = X_{t-1}$,
— ε_t est le **terme d'erreur** à l'instant t,
— μ est la **moyenne** de la série.

2.7.5 Modèle Autorégressif intégré à moyennes mobiles saisonnier (SARIMA)

Le modèle **SARIMA** (Seasonal ARIMA) est une extension du modèle ARIMA qui intègre des composantes saisonnières pour modéliser les séries temporelles présentant des motifs répétitifs à intervalles réguliers.
Il est défini par sept paramètres : $(p, d, q) \times (P, D, Q)_s$. Les trois premiers paramètres (p, d, q) correspondent aux composantes non saisonnières du modèle ARIMA, tandis que les trois suivants (P, D, Q) décrivent les composantes saisonnières. Le paramètre s représente la période de saisonnalité, c'est-à-dire le nombre d'observations après lesquelles le motif saisonnier se répète.

SARIMA$(p, d, q) \times (P, D, Q)_s$

Les composantes non saisonnières (p, d, q) sont définies comme suit :

— p : ordre du processus autorégressif (AR).

— d : degré de différenciation nécessaire pour rendre la série stationnaire.

— q : ordre du processus à moyenne mobile (MA).

Les composantes saisonnières $(P, D, Q)_s$ sont définies de manière similaire, mais elles s'appliquent aux observations espacées de s périodes :

— P : ordre du processus autorégressif saisonnier.

— D : degré de différenciation saisonnière.

— Q : ordre du processus à moyenne mobile saisonnier.

Le modèle SARIMA s'écrit de manière générale sous la forme :

$$\Phi_P(B^s)\,\phi_p(B)\,(1-B)^d\,(1-B^s)^D\,X_t = \Theta_Q(B^s)\,\theta_q(B)\,\epsilon_t;$$

où :

— B est l'opérateur de retard ($By_t = y_{t-1}$),

— $\phi(B)$ et $\theta(B)$ sont les polynômes AR et MA non saisonniers,

— $\Phi(B^s)$ et $\Theta(B^s)$ sont les polynômes AR et MA saisonniers,

— s représente la période de la saisonnalité,

— ε_t est un bruit blanc centré et homoscédastique.

2.8 Méthodologie Box-Jenkins

Box et Jenkins (1976)[**19**] ont promu une méthodologie consistant à modéliser les séries temporelles univariées au moyen des processus ARMA. Ces processus sont efficaces et constituent une bonne approximation de processus plus généraux pourvu que l'on se restreigne au cadre linéaire. Lorsqu'une composante saisonnière est présente dans les données, cette approche peut être étendue aux **modèles SARIMA (Seasonal ARIMA)**, où les paramètres saisonniers permettent de mieux capturer les fluctuations périodiques. Ces modèles sont particulièrement efficaces pour la **prévision**, à condition que la série temporelle soit correctement identifiée et stationarisée. La méthodologie Box-Jenkins repose sur un processus **itératif** comprenant quatre étapes fondamentales 2.6.

2.8.1 Identification :

La stationnarité de la série X_t est d'abord testée soit graphiquement, soit théoriquement à l'aide de tests de racines unitaires (ou de non-stationnarité). Les plus connus sont :

— le test de Dickey-Fuller simple,

— le test de Dickey-Fuller augmenté (ADF),

— le test de Phillips-Perron (PP).

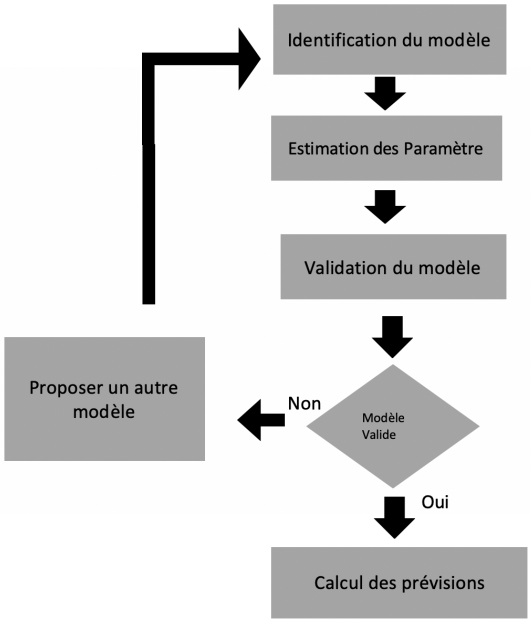

FIGURE 2.6 – Organigramme de la méthode *Box Jenkins*

Si la série n'est pas stationnaire, il convient de la transformer par différenciation :

$$\Delta X_t, \quad \Delta^2 X_t, \quad \Delta^d X_t, \quad \text{etc.}$$

afin d'obtenir une série stationnaire.

L'étape d'identification d'un processus ARMA (choix entre AR, MA et ARMA, ainsi que la sélection des ordres p et q) selon la méthodologie de Box et Jenkins repose sur la comparaison des caractéristiques théoriques des modèles ARMA avec leurs équivalents empiriques.

1. Pour un modèle stationnaire satisfaisant une représentation **AR(p)**, les autocorrélations partielles deviennent identiquement nulles au-delà de l'ordre p.
2. Pour un modèle satisfaisant une représentation **MA(q)**, les autocorrélations s'annulent à partir de l'ordre q.
3. Pour déterminer les ordres p et q d'un modèle **ARMA(p,q)**, on peut se baser sur les graphes de l'autocorrélation (ACF) et de l'autocorrélation partielle (PACF) :
   — L'ordre de la composante **MA** correspond au plus grand ordre d'autocorrélation significative.
   — L'ordre de la composante **AR** correspond au plus grand ordre d'autocorrélation partielle significative.

Dans le cas de données saisonnières, le modèle de référence devient le **SARIMA(p,d,q)(P,D,Q)$_s$**, qui prend en compte à la fois les composantes non saisonnières et saisonnières :

— p, d, q : ordres non saisonniers du modèle ARIMA,

— P, D, Q : ordres saisonniers des composantes AR, I et MA,

— s : périodicité saisonnière (par exemple, $s = 12$ pour des données mensuelles).

L'identification du modèle SARIMA s'appuie également sur les graphes ACF et PACF, mais cette fois en tenant compte des lags multiples de s (par exemple : 12, 24, 36, etc.). Ces lags permettent d'identifier la présence d'une structure saisonnière et d'estimer les ordres P et Q.

2.8.2 Estimation des paramétres :

Une fois le modèle identifié, X_t admettant une représentation ARMA(p, q), l'estimation des paramètres ϕ_j pour $j = 1, \ldots, p$, θ_k pour $k = 1, \ldots, q$ et σ^2 du (des) modèle(s) sélectionné(s) peut se réaliser par les méthodes suivantes :

1. Méthode du maximum de vraisemblance (MLE).
2. Méthode des moindres carrés non linéaires.
3. Dans le cas d'un AR(p), on utilise les équations de Yule-Walker.
   Les paramètres à estimer dans les modèles ARMA et ARIMA sont :
   — p : L'ordre du composant autorégressif (AR).
   — d : L'ordre de différenciation nécessaire pour la stationnarité.
   — q : L'ordre du composant de moyenne mobile (MA).
   — $\phi_1, \phi_2, \ldots, \phi_p$: Les coefficients AR.
   — $\theta_1, \theta_2, \ldots, \theta_q$: Les coefficients MA.

Dans le cas des modèles **SARIMA(p,d,q)(P,D,Q)$_s$**, des paramètres saisonniers supplémentaires sont également estimés :

— P : l'ordre de la composante saisonnière autorégressive (SAR),

— D : le degré de différenciation saisonnière,

— Q : l'ordre de la composante saisonnière de moyenne mobile (SMA),

— s : la périodicité saisonnière (ex. $s = 12$ pour des données mensuelles),

— $\Phi_1, \Phi_2, \ldots, \Phi_P$: les coefficients saisonniers AR,

— $\Theta_1, \Theta_2, \ldots, \Theta_Q$: les coefficients saisonniers MA.

L'estimation de ces paramètres se fait également en utilisant les méthodes susmentionnées, notamment le maximum de vraisemblance, en prenant en compte les composantes saisonnières et non saisonnières.

2.8.3 Validation :

Cette étape vise à vérifier si le modèle estimé est adéquat (ARIMA ou SARIMA). Les résidus du modèle sont analysés pour s'assurer qu'ils se comportent comme du bruit blanc, généralement à l'aide du plusieurs test sur les paramètres et sur les rèsidus.

2.8.3.1 Test de signficativitè des coefficients :

Une fois les coefficients d'un modèle ARIMA(p,d,q) ou SARIMA(p,d,q)(P,D,Q,s) estimés, On vèrifie tout d'abord :

1. que les racines des polynômes AR (et SAR) ainsi que MA (et SMA)ne sont pas egales à 1 ;
2. la significativité des coefficients ϕ_j, θ_k, Φ_ℓ, Θ_m au test de Student (ou Wald).

2.8.3.2 Test sur le bruit blanc :

Pour que les modèles obtenus soient valides, il convient de vérifier que les résidus estimés, notés $\hat{\varepsilon}_t$, suivent un bruit blanc, non autocorrélé et de même variance σ^2, et qu'ils suivent une loi normale. Si ces hypothèses ne sont pas rejetées, on peut alors mener des tests sur les paramètres.

Test sur la moyenne des résidus :

Pour vérifier que les résidus estimés sont centrés, comme il se doit, il s'agit de confronter les hypothèses :

$$\begin{cases} H_0 : E(\varepsilon_t) = 0; \\ H_1 : E(\varepsilon_t) \neq 0. \end{cases}$$

Le test est bassé sur la statistique ε_t définie par :

$$\hat{\varepsilon} = \frac{\bar{\varepsilon}}{S_\varepsilon / \sqrt{T}} \quad \text{où} \quad \bar{\varepsilon} = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t, \quad S_\varepsilon^2 = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t^2.$$

La statistique $\hat{\varepsilon}$ suit approximativement une loi normale centrée réduite $N(0,1)$.
Ainsi, on rejette H_0 si $|\hat{\varepsilon}| > z_{\alpha/2}$, où $z_{\alpha/2}$ est le quantile d'ordre $\alpha/2$ de la loi normale 4 $N(0,1)$. [**14**]

Test de Box-Pierce et Ljung-Box :

Le test de **Box-Pierce** permet d'identifier les processus de bruit blanc. Ce test s'écrit :

$$H_0 : \rho(1) = \rho(2) = \cdots = \rho(h) = 0.$$

contre l'alternative :

$$H_1 : \exists\, i \text{ tel que } \rho(i) \neq 0.$$

Pour effectuer ce test, on utilise la statistique de **Box et Pierce** (1970)[**14**], donnée par :

$$Q = T \sum_{k=1}^{h} \hat{\rho}_k^2.$$

où

— h est le nombre de retards,

— T est le nombre d'observations,

— $\hat{\rho}(k)$ est l'autocorrélation empirique d'ordre k.

Sous l'hypothèse H_0 (les résidus suivent un bruit blanc), la statistique Q suit une loi du Khi-deux à h degrés de liberté, notée χ^2_h.

Ainsi, on rejette H_0 si :

$$Q > \chi^2_h(1-\alpha).$$

où $\chi^2_h(1-\alpha)$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2_h.
Ljung et Box (1978) ont amélioré le test de Box-Pierce en considérant la statistique suivante :[**14**]

$$Q_{LB} = T(T+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{T-k}.$$

Test de normalité :

De nombreux modèles de séries temporelles supposent que les résidus sont indépendants et distribués selon la loi normale. Un des tests permettant de vérifier cette hypothèse est le **test de Jarque-Bera**.
Les hypothèses à tester sont :

$H_0 : \varepsilon_t \sim \mathcal{N}(0,1)$;

$H_1 : \varepsilon_t \nsim \mathcal{N}(0,1)$.

La statistique du test est définie par :

$$JB = \frac{T}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

où :

— S est le coefficient d'asymétrie (*skewness*),

— K est le coefficient d'aplatissement (*kurtosis*), définis par :

$$S = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t-\bar{X}}{S_X}\right)^3, \quad K = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t-\bar{X}}{S_X}\right)^4.$$

avec $\bar{X}$ et S_X étant respectivement la moyenne et l'écart-type empirique des données.
Sous H_0, la statistique JB suit asymptotiquement une loi du Khi-deux à deux degrés de liberté, soit χ^2_2. On rejette l'hypothèse de normalité si :

$JB > \chi^2_2(1-\alpha)$;

où $\chi^2_2(1-\alpha)$ est le quantile d'ordre $1-\alpha$ de la loi χ^2_2.

Visualisation graphique :

— Graphique des résidus dans le temps : Permet de voir s'il existe des motifs, tendances ou variations de la variance.

— Histogramme des résidus : Permet de vérifier la normalité des résidus (utile pour certains tests statistiques).

— Graphique ACF/PACF des résidus : Permet de détecter la présence d'autocorrélation restante.

2.8.3.3 Choix du modèle :

Une fois le modèles choisis, l'erreur de prèvision dèpend de la variance , notre but est donc de choisir le meilleur modèle qui minimise cette erreur par líintermèdiaire de l'un des critères suivants : Parmi les modèles ARIMA et SARIMA candidats, on retient celui minimisant :

Critères standard :

— Erreur quadratique moyenne (MSE : Mean Square Error)

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2$$

— Racine de l'erreur quadratique moyenne (RMSE : Root Mean Square Error)

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2}$$

— Erreur absolue moyenne (MAE : Mean Absolue Error)

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\varepsilon_t|$$

Critères d'information :

— Critère d'information d'Akaike (AIC) :Mesure la qualité d'un modèle en tenant compte de l'ajustement et du nombre de paramètres.

$$AIC = \ln(\hat{\sigma}^2) + \frac{2k}{T},$$

— Critère d'information bayésien de Schwarz (BIC) :Comme l'AIC, mais il pénalise plus fortement les modèles complexes (avec beaucoup de paramètres).

$$BIC = \ln(\hat{\sigma}^2) + \frac{k\ln T}{T},$$

— Critère de Hannan-Quinn : Critère intermédiaire entre AIC et BIC, moins strict que BIC mais plus que AIC.

$$HQ = \ln(\hat{\sigma}^2) + \frac{k\,c\,\ln\ln T}{T},$$

où $k = p + q + P + Q$ pour SARIMA

Le modèle retenu est celui qui minimise ces critères tout en présentant des résidus conformes aux tests précédents.

2.8.4 Prévision :

Étant donnée une série stationnaire X_t, observée entre 1 et T, on cherche à faire de la prévision à horizon h : $X_{T+1}, \ldots, X_{T+h}$. Tous les processus AR, MA et ARMA seront supposés sous forme canonique, et n'avoir aucune racine unité. De plus, les polynômes ϕ et θ ont leurs racines de module strictement supérieur à 1.

Prévisions à l'aide d'un modèle AR(p)

Soit un modèle AR(p) :

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t, \quad \text{avec } \varepsilon_t \sim bruitblanc(0, \sigma^2).$$

La prévision optimale à la date $T + h$ est :

$$\hat{X}_T(h) = X_{T+h} = E[X_{T+h}|\mathcal{F}(X)];$$

où $\mathcal{F}(X)$ est l'information disponible $\{X_1, \ldots, X_T\}$; On trouve ensuite :

$$\begin{aligned}
\hat{X}_{T+1} &= c + \phi_1 X_T + \phi_2 X_{T-1} + \cdots + \phi_p X_{T+1-p};\\
\hat{X}_{T+2} &= c + \phi_1 \hat{X}_{T+1} + \phi_2 X_T + \cdots + \phi_p X_{T+2-p};\\
&\vdots\\
\hat{X}_{T+h} &= c + \phi_1 \hat{X}_{T+h-1} + \phi_2 \hat{X}_{T+h-2} + \cdots + \phi_p \hat{X}_{T+h-p}.
\end{aligned}$$

Prévisions à l'aide d'un modèle MA(q)

Considérons un modèle MA(q) :

$$X_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

La prévision optimale à la date $T + h$ est donnée par :

$$\hat{X}_T(h) = X_{T+h} = E[X_{T+h}|\varepsilon_T, \varepsilon_{T-1}, \ldots];$$

On trouve alors :

$$\begin{aligned}
\hat{X}_{T+1} &= c + \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1} + \cdots + \theta_q \varepsilon_{T+1-q};\\
\hat{X}_{T+2} &= c + \theta_2 \varepsilon_T + \cdots + \theta_q \varepsilon_{T+2-q};\\
&\vdots
\end{aligned}$$

Si $h > q$, alors :

$$\hat{X}_{T+h} = c.$$

Prévisions à l'aide d'un modèle ARMA(p,q)

Considérons un modèle ARMA(p,q) :

$$\phi(L)X_t = c + \theta(L)\varepsilon_t, \quad \text{avec } \varepsilon_t \sim Bruitblanc(0, \sigma^2);$$

où :

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p, \quad \theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q.$$

La prévision optimale est donnée par :

$$\hat{X}_{T+h} = c + \sum_{j=1}^{p} \phi_j \hat{X}_{T+h-j} + \sum_{k=1}^{q} \theta_k \hat{\varepsilon}_{T+h-k}.$$

Prévisions à l'aide d'un modèle ARIMA(p,d,q)

— Différenciation d fois :

$$Y_t = (1 - L)^d X_t.$$

— Modélisation ARMA(p,q) sur sur la série différenciée Y_t :

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) Y_t = \left(1 + \sum_{j=1}^{q} \theta_j L^j\right) \varepsilon_t.$$

— Prévision h pas :

$$\hat{Y}_{T+h|T}, \quad \text{via ARMA}$$

— Reconstruction de X_t par sommation :

$$\hat{X}_{T+1|T} = X_T + \hat{Y}_{T+1|T};$$

$$\hat{X}_{T+2|T} = \hat{X}_{T+1|T} + \hat{Y}_{T+2|T}$$

$$\vdots$$

Prévisions à l'aide d'un modèle SARIMA :

Une fois le modèle SARIMA identifié, estimé et validé conformément à la méthodologie de Box et Jenkins, il peut être utilisé pour produire des prévisions à court ou moyen terme. Cette étape constitue l'aboutissement du processus de modélisation, et repose sur l'exploitation des relations temporelles détectées dans la série.
SARIMA$(p, d, q)(P, D, Q)_s$, intègre à la fois les composantes non saisonnières et saisonnières de la série temporelle. Sa forme générale s'écrit :

$$\Phi(B^s)\,\phi(B)\,(1 - B)^d\,(1 - B^s)^D\,y_t = \Theta(B^s)\,\theta(B)\,\varepsilon_t;$$

Formulation de la prévision

L'objectif de la prévision est d'estimer les valeurs futures $\hat{y}_{T+h}$ de la série à l'horizon h, en utilisant l'information disponible jusqu'au temps T. Mathématiquement, cela revient à calculer l'espérance conditionnelle suivante :

$$\hat{y}_{T+h|T} = \mathbb{E}(y_{T+h} \mid y_T, y_{T-1}, \dots).$$

Par développement du modèle, la prévision à l'horizon h peut être exprimée sous la forme :

$$\hat{y}_{T+h|T} = \sum_{i=1}^{p} \varphi_i \, \hat{y}_{T+h-i|T} + \sum_{j=1}^{P} \Phi_j \, \hat{y}_{T+h-js|T} + c - \sum_{k=1}^{q} \theta_k \, \hat{\varepsilon}_{T+h-k|T} - \sum_{\ell=1}^{Q} \Theta_\ell \, \hat{\varepsilon}_{T+h-\ell s|T}$$

avec $\hat{\varepsilon}_t = 0$ pour tout $t > T$, car les erreurs futures ne sont pas observables et sont supposées de moyenne nulle.

Évaluation de la précision des prévisions

L'incertitude associée aux prévisions peut être mesurée par la variance de l'erreur de prévision :[**13**]

$$\mathrm{Var}(y_{T+h} - \hat{y}_{T+h|T}) = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2$$

où ψ_j sont les coefficients de la représentation MA(∞) du modèle.

Sous l'hypothèse de normalité des erreurs ($\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$), un **intervalle de confiance au niveau** $(1 - \alpha)$ pour la prévision est donné par :

$$\hat{y}_{T+h|T} \pm z_{\alpha/2} \cdot \sqrt{\mathrm{Var}(y_{T+h} - \hat{y}_{T+h|T})};$$

où $z_{\alpha/2}$ est le quantile de la loi normale standard (par exemple $z_{0.025} = 1{,}96$ pour un intervalle à 95 %).

La prévision par le modèle SARIMA permet ainsi de tirer parti des composantes structurelles de la série temporelle (tendance, saisonnalité, dépendances temporelles) pour anticiper l'évolution future du phénomène étudié. L'utilisation rigoureuse de la méthodologie Box-Jenkins garantit la robustesse et la fiabilité des prévisions, à condition que le modèle ait été correctement spécifié et validé.

Conclusion

En conclusion, ce chapitre a mis en place tous les concepts clés pour modéliser et prévoir une série chronologique selon la méthodologie de Box et Jenkins, de la décomposition initiale

jusqu'à la validation finale. Nous avons ainsi vu comment isoler les composantes de tendance, de saisonnalité et de bruit, appliquer les différenciations nécessaires pour atteindre la stationnarité, identifier les ordres appropriés à l'aide des fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF), puis estimer et valider différents modèles ARIMA et SARIMA. La phase de validation a permis de contrôler la significativité des coefficients, de vérifier l'absence d'autocorrélation résiduelle et de comparer divers candidats à l'aide des critères d'information (AIC, BIC, etc.).

Dans le chapitre suivant, nous verrons comment les techniques de machine learning viennent compléter, et parfois surpassent, ces approches traditionnelles de séries temporelles, en offrant une plus grande flexibilité pour capturer des relations complexes et améliorer la qualité des prévisions.

3
Machine learning

Introduction

Machine Learning, ou l'apprentissage automatique, est une branche de l'intelligence artificielle qui se concentre sur le développement d'algorithmes capables d'apprendre et de s'améliorer à partir des données. il repose sur des approches mathématiques et statistiques pour identifier des motifs au sein de grands ensembles de données, permettant ainsi aux systèmes de prendre des décisions ou de faire des prédictions sans être explicitement programmés pour chaque tâche. Les applications du Machine learning sont vastes, englobant des domaines tels que la reconnaissance vocale, les systèmes de recommandation, la détection de fraudes et les véhicules autonomes, etc. Cette technologie joue un rôle central dans la transformation numérique actuelle, offrant des outils puissants pour analyser et interpréter des volumes massifs de données, et ouvrant la voie à des innovations dans divers secteurs.

3.1 Machine learning

Est un processus de construction d'un modèle général à partir de données (observations) particulières du monde réel. Ainsi, le but est double :

— Prédire un comportement face à une nouvelle donnée.

— Approximer une fonction ou une densité de probabilité.

L'apprentissage est défini comme étant la capacité à améliorer les performances au fur et à mesure de l'exercice d'une activité. il fait référence au développement, à l'analyse et à l'implémentation de méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qui sont difficile ou impossible d'accomplir par des moyens algorithmiques plus classiques.
L'objectif : est d'extraire et d'exploiter automatiquement l'information présente dans un jeu de données.

C'est une technique de science des données qui permet aux ordinateurs d'utiliser des données existantes afin de prévoir les tendances, les résultats et les comportements futurs.
Arthur Samuel a défini Machine learning comme étant :
"Un champ d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés ".[**11**]

3.2 Objectifs de machine learning

— **Prédiction** : Créer des modèles capables d'établir des prédictions précises sur de nouvelles données. Cela inclut l'estimation de valeurs manquantes ou l'anticipation de tendances futures.

— **Classification** : Classer les données dans des catégories prédéfinies. Cela est utilisé pour identifier des e-mails spam, détecter des fraudes financières ou diagnostiquer des maladies à partir de symptômes.

— **Reconnaissance de Formes** : Identifier des motifs et structures complexes dans les données, comme la reconnaissance d'images, de la parole, ou la détection d'objets.

— **Recommandation** : Analyser les préférences et comportements des utilisateurs pour leur suggérer des produits ou services pertinents. C'est couramment utilisé en e-commerce et par les services de streaming.

— **Optimisation** : Améliorer des processus et des décisions en analysant les données historiques pour optimiser la gestion des stocks, l'affectation des ressources, etc.

— **Apprentissage Autonome** : Permettre aux ordinateurs d'apprendre sans programmation explicite, en s'adaptant aux données et en améliorant leurs performances au fil du temps.

— **Clustering et Association** : Regrouper des données similaires ou identifier des relations entre elles, souvent utilisé pour comprendre les comportements des clients.

3.3 Étapes de Machine learning

Pour développer et gérer un modèle de machine learning adapté à une mise en production, il est généralement nécessaire de suivre les étapes suivantes :

1. **Collecte des données :** La collecte des données est la première étape essentielle dans le développement d'un modèle de Machine learning. La quantité et la qualité des données collectées déterminent la qualité du modèle final. Les données peuvent être collectées à partir de diverses sources telles que des fichiers, des bases de données ou des capteurs. Cependant, les données collectées nécessitent souvent une préparation, car elles peuvent contenir des champs manquants, des valeurs aberrantes ou être non structurées. Ainsi, la préparation des données est essentielle pour rendre les données utilisables par le modèle.[**15**]
2. **Prétraitement des données :** Le prétraitement des données est essentiel pour nettoyer les données brutes et les rendre utilisables par les modèles de machine learning. Voici quelques techniques de prétraitement de base :

- Conversion des données :Les données catégorielles et ordinales sont converties en données numériques, car les modèles de machine learning ne peuvent traiter que des fonctionnalités numériques.
- Ignorer les valeurs manquantes : Les lignes ou colonnes contenant des données manquantes peuvent être supprimées, bien que cette méthode ne doit pas être utilisée de manière excessive.
- Remplissage des valeurs manquantes : Les données manquantes peuvent être remplacées par la valeur moyenne, médiane ou la plus fréquente de la variable concernée.
- Machine learning : Les données manquantes peuvent être prédites en utilisant des techniques de machine learning, permettant de prédire les valeurs manquantes en se basant sur les données existantes.
- Détection des valeurs aberrantes : Les valeurs aberrantes,qui s'écartent considérablement des autres observations, peuvent être identifiées et supprimées ou remplacées. Par exemple, un poids humain de $800kg$ en raison d'une faute de frappe doit être corrigé.

Ces techniques de prétraitement des données contribuent à améliorer la qualité des données et à garantir que le modèle de machine learning peut obtenir des résultats précis et fiables .[**15**]

3. **Recherche du modèle de machine learning :** La recherche du modèle de machine learning consiste à sélectionner un algorithme ou un modèle qui est capable de prédire une sortie pour une entrée donnée. Il existe une variété d'algorithmes disponibles, chacun étant adapté à un type spécifique d'apprentissage .
[**15**]
4. **Formation et test du modèle :** Pour former un modèle, les données sont divisées en trois sections : l'ensemble **d'apprentissage** (Training Set), l'ensemble de **validation** (Validation Set) et l'ensemble de **test** (Testing Set). Le classificateur est entraîné avec l'ensemble d'apprentissage, les paramètres sont ajustés avec l'ensemble de validation, puis les performances sont évaluées avec l'ensemble de test. L'ensemble de test ne doit pas être utilisé pendant l'entraînement du classificateur, étant réservé uniquement pour le test final.

5. **L'évaluation :** L'évaluation est une étape essentielle du processus de développement du modèle. Elle permet de déterminer quel modèle représente le mieux les données et à quel point ce modèle sélectionné fonctionnera efficacement à l'avenir .[**15**]

3.4 Types de machine learning

Machine learning est un domaine vaste qui se divise en plusieurs types selon la manière dont un modèle apprend à partir des données. Il s'intéresse à différentes classes de problèmes, chacune correspondant à une approche spécifique d'apprentissage.
On distingue principalement trois grandes catégories de machine learning :

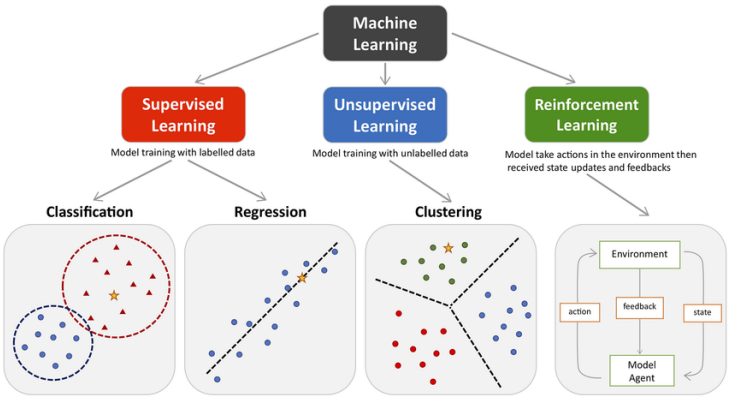

FIGURE 3.1 – Types de machine learning

3.4.1 Apprentissage supervisé :

Consiste à enseigner à un ordinateur à reconnaître des schémas dans les données en lui fournissant des exemples étiquetés. Ces exemples sont des paires de données, où chaque entrée est associée à une sortie attendue.
L'apprentissage supervisé s'intéresse aux problèmes pouvant être formalisés de la façon suivante :

Étant données n observations $\{\mathbf{x}_i\}_{i=1,\dots,n}$ décrites dans un espace X, et leurs étiquettes $\{y_i\}_{i=1,\dots,n}$ décrites dans un espace Y, on suppose que les étiquettes peuvent être obtenues à partir des observations grâce à une fonction $\varphi : X \to Y$ fixe et inconnue :

$$y_i = \varphi(\mathbf{x}_i) + \varepsilon_i,$$

où ε_i est un bruit aléatoire.
L'objectif est alors d'utiliser les données pour déterminer une fonction
$f : X \to Y$ telle que, pour tout couple $(\mathbf{x}, \varphi(\mathbf{x})) \in X \times Y$, on ait :

$$f(\mathbf{x}) \approx \varphi(\mathbf{x}).$$

L'espace sur lequel sont définies les données est le plus souvent $X = \mathbb{R}^p$. Nous verrons cependant aussi comment traiter d'autres types de représentations, comme des variables binaires, discrètes, catégoriques, voire des chaînes de caractères ou des graphes.

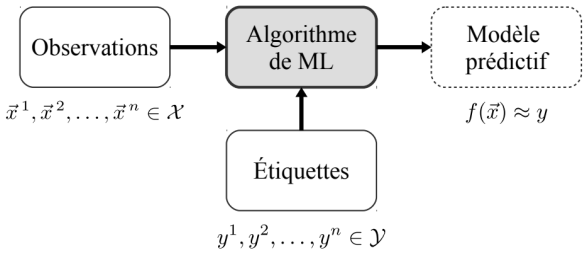

FIGURE 3.2 – Apprentissage supervisé.

L'apprentissage supervisé est de deux types Classification et Régression

3.4.1.1 Classification :

La classification est une méthode d'apprentissage supervisé qui vise à attribuer une étiquette ou une classe à une donnée d'entrée en fonction de ses caractéristiques. Elle s'appuie sur des données étiquetées pour former un modèle capable de prédire la classe d'une nouvelle donnée inédite. les principaux algorithme de classifications sont :[**16**]

- Arbre de décision.
- Machine à support de vecteur
- k-Plus Proches Voisins (k-NN).
- Classificateurs Bayésiens Naïfs .

En classification, il existe plusieurs types de méthodes, chacune avec ses propres caractéristiques et approches. Voici quelques-uns des types de classification les plus courants :

Classification binaire :

Un problème d'apprentissage supervisé où l'espace des étiquettes est binaire, c'est-à-dire $Y = \{0, 1\}$, est appelé un problème de classification binaire. Voici quelques exemples de tels problèmes :

- Identifier si un email est un spam ou non.
- Identifier si une transaction financière est frauduleuse ou non .

Le modèle de classification binaire attribue à chaque exemple une étiquette 0 ou 1 selon ses caractéristiques, afin de prendre une décision entre deux options possibles.

Classification multi-classe :

Un problème d'apprentissage supervisé où l'espace des étiquettes est fini et discret, c'est-à-dire $Y = \{1, 2, \ldots, C\}$, est appelé un problème de classification multi-classe. C représente le nombre de classes disponibles. Voici quelques exemples de tels problèmes :

- Identifier en quelle langue un texte est écrit.
- Identifier l'expression d'un visage parmi une liste prédéfinie de possibilités (colère, tristesse, joie, etc.).

Le modèle apprend à choisir la classe la plus appropriée selon les caractéristiques de chaque exemple.

3.4.1.2 Régression :

Un problème d'apprentissage supervisé dans lequel l'espace des étiquettes est appartiennent à l'ensemble des valeurs réelles $Y = \mathbb{R}$ est appelé un problème de régression. il est utilisée lorsque la sortie à prédire peut prendre des valeurs continues, c'est-à-dire qu'il s'agit d'une variable réelle.[**16**]
Voici quelques exemples de problèmes de régression :

- Prédire le nombre d'utilisateurs d'un service en ligne à un moment donné.
- Prédire le prix d'une action en bourse.

Dans ces exemples, le modèle de régression cherche à prédire une valeur continue en fonction des caractéristiques des données d'entrée, permettant ainsi de réaliser des prédictions sur des phénomènes qui varient de manière continue.

Types de régression :

La régression , bien qu'elle repose sur un modèle fondamental, peut être utilisée de différentes manières selon le contexte et les objectifs spécifiques d'une analyse. Voici quelques types courants de régression linéaire :

— Régression linéaire simple et multiple.
— Régression polynomiale.
— Régression linéaire pondérée.
— Régression linéaire régularisée.
— Régression linéaire robuste.

3.4.2 Apprentissage non supervisé :

L'apprentissage non-supervisé consiste à analyser des données sans étiquettes ni réponses préétablies. Contrairement à l'apprentissage supervisé, où le modèle est entraîné sur des exemples étiquetés pour prédire des résultats, l'apprentissage non-supervisé cherche à découvrir des structures ou des modèles intrinsèques dans les données elles-mêmes.Il s'intéresse aux problèmes pouvant être formalisés de la façon suivante :
Étant données n observations $\{\mathbf{x}_i\}_{i=1,\dots,n}$ décrites dans un espace X, il s'agit d'apprendre une fonction sur X qui vérifie certaines propriétés.

FIGURE 3.3 – Apprentissage non supervisé

Voici quelques-uns des principaux algorithmes d'apprentissage non-supervisé :

3.4.2.1 Clustering :

Dans le clustering, un algorithme analyse les données sans étiquettes pour détecter des groupes similaires. Par exemple, dans le contexte des visiteurs d'un blog, l'algorithme peut identifier des groupes tels que les amateurs de bandes dessinées ou les passionnés de science-fiction, en se basant sur des caractéristiques communes telles que les habitudes de consultation. Avec un algorithme hiérarchique, ces groupes peuvent être subdivisés en sous-groupes plus spécifiques, ce qui facilite le ciblage de messages adaptés à chaque segment .[**21**] Concernant les algorithmes spécifiques de clustering :

— K-Means
— Analyse des clusters hiérarchiques (HCA)
— Maximisation des attentes

Dans le graphique suivant , nous montrons comment un ensemble de points peut être classé pour former trois sous-ensembles :

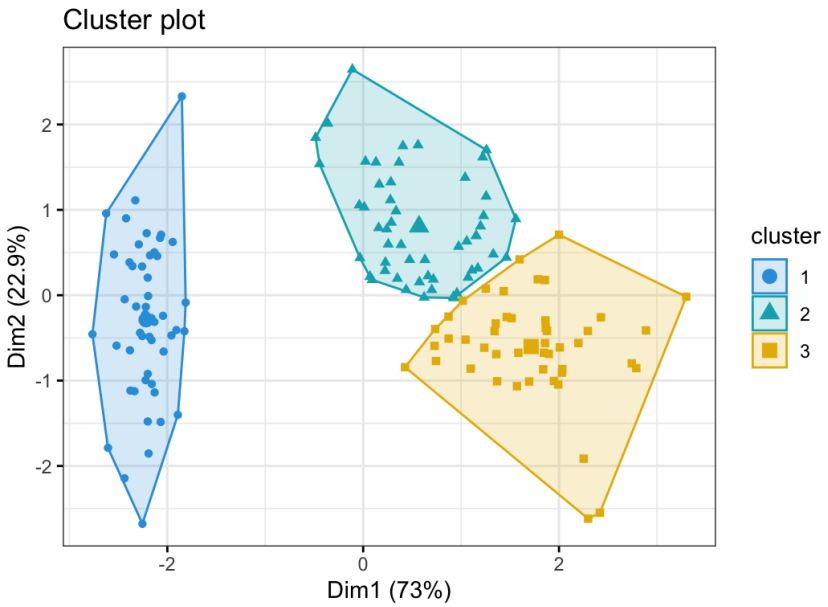

FIGURE 3.4 – Clustering

3.4.2.2 Réduction de la dimensionnalité :

La réduction de la dimensionnalité vise à simplifier les données tout en préservant autant d'informations que possible. Une méthode pour y parvenir consiste à regrouper plusieurs caractéristiques corrélées en une seule. Par exemple, le kilométrage d'une voiture peut être fortement corrélé avec son âge, de sorte que l'algorithme de réduction de la dimensionnalité les fusionnera en une seule caractéristique représentant l'usure de la voiture. Cette technique est connue sous le nom d'extraction de caractéristiques.[**21**]

3.4.3 Apprentissage par renforcement :

Dans le cadre de l'apprentissage par renforcement, le système d'apprentissage peut interagir avec son environnement et accomplir des actions. En retour de ces actions, il obtient une récompense, qui peut être positive si l'action était un bon choix, ou négative dans le cas contraire. La récompense peut parfois venir après une longue suite d'actions , c'est le cas par exemple pour un système apprenant à jouer au go ou aux échecs. Ainsi,l'apprentissage consiste dans ce cas à définir une politique, c'est-à-dire une stratégie permettant d'obtenir systématiquement la meilleure récompense possible.
Les applications principales de l'apprentissage par renforcement se trouvent dans les jeux (échecs, go, etc) et la robotique.[**20**]

3.5 Critères de performance :

Il existe de nombreuses façons d'évaluer la performance prédictive d'un modèle de machine learning. Cette section présente les principaux critères utilisés.

3.5.1 Matrice de confusion :

La matrice de confusion (ou matrice d'erreur) est l'un des nombreux indicateurs qui permettent d'évaluer la performance des modèles d'apprentissage automatique il permet de visualiser les résultats d'un algorithme de classification. Plus précisément, il s'agit d'une table qui décompose le nombre d'instances de vérité terrain d'une classe donnée par rapport au nombre d'instances de classes prédites. On peut s'en servir pour calculer un certain nombre d'autres indicateurs de performance , tels que la précision et le rappel, peuvent être utilisées avec n'importe quel algorithme de classification.

Voici ce à quoi pourrait ressembler un modèle de matrice de confusion standard pour un classificateur binaire :

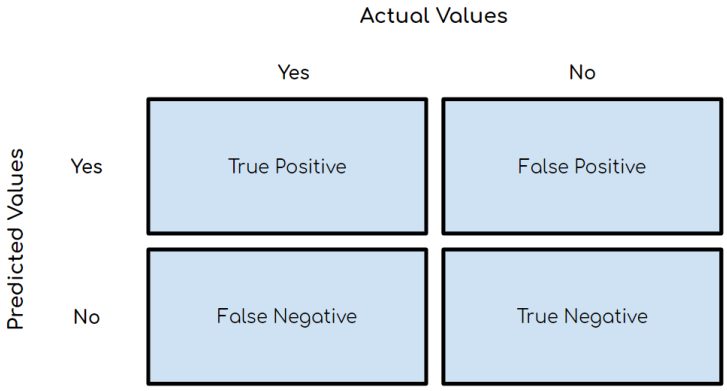

FIGURE 3.5 – Matrice de Confusion

3.5.2 Accuracy :

L'exactitude Accuracy est une mesure fondamentale et à la fois simple en machine learning. Elle représente le pourcentage de prédictions correctes par rapport au nombre total de prédictions (multiplié par 100). Un score élevé d'exactitude indique que le modèle effectue des prédictions précises, tandis qu'un score faible suggère une efficacité moindre du modèle. [**18**]

$$\text{Accuracy} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}}.$$

3.5.3 Précision :

La précision est définie comme la proportion de prédictions positives correctes parmi toutes les prédictions positives effectuées par le modèle. Mathématiquement, elle est calculée comme suit :[**18**]

$$\text{Précision} = \frac{\text{Vrais Positifs (VP)}}{\text{Vrais Positifs (VP)} + \text{Faux Positifs (FP)}}.$$

Une précision élevée signifie que lorsque le modèle prédit une classe positive, il est très souvent correct, et c'est la précision est faible signifie que le modèle fait beaucoup d'erreurs en classant des éléments négatifs comme positifs .

3.5.4 Le rappel (Recall) :

Le rappel (ou recall en anglais) est une mesure de performance utilisée dans les modèles de classification. Il permet d'évaluer la capacité d'un modèle à identifier correctement toutes les occurrences positives d'un jeu de données.
Un rappel élevé signifie que le modèle détecte bien les cas positifs et minimise les faux négatifs,et un rappel faible signifie que de nombreux exemples positifs sont manqués (prédits comme négatifs).[**18**]

$$\text{Rappel} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Négatifs}}.$$

3.5.5 Le score F1 :

Le score F1 est une évaluation unique qui combine précision et rappel ,il est représenté comme suit :[**18**]

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}.$$

— Si $F1 = 1$: Précision et rappel sont parfaits .
— Si $F1$ est proche de 0 : Mauvaise performance du modèle.
— Si $F1$ est élevé ($\sim 0.8 - 0.9$) :Bon équilibre entre précision et rappel.

3.5.6 MAE (Mean Absolute Error) :

L'erreur absolue moyenne (MAE) est une mesure largement utilisée en statistiques et l'analyse des données qui quantifie l'ampleur moyenne des erreurs dans un ensemble de prédictions, sans tenir compte de leur direction. Elle est calculée comme la moyenne des différences absolues entre les valeurs prédites et les valeurs réelles.
est définie comme suite :

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

Un **MAE** plus faible indique que les prédictions du modèle sont plus proches des valeurs réelles, ce qui suggère une meilleure précision.

3.5.7 MSE (Mean Squared Error) :

L'erreur quadratique moyenne (MSE) est une mesure largement utilisée pour évaluer les performances des modèles de régression, appréciée pour sa capacité à accentuer des erreurs plus importantes. Il mesure la différence quadratique moyenne entre les valeurs prédites et les valeurs réelles, en mettant l'accent sur les erreurs plus importantes que sur les erreurs plus petites.

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

— MSE élevé : Indique que le modèle a de grandes erreurs de prédiction, suggérant une performance médiocre.

— MSE faible : Indique que le modèle a de petites erreurs de prédiction, suggérant une bonne performance.

MAPE (Mean Absolute Percentage Error) : est une métrique de régression exprimée en pourcentage, calculée comme la moyenne des erreurs absolues relatives :

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|,$$

3.5.8 RMSE (Root Mean Squared Error) :

est une métrique statistique couramment utilisée pour évaluer la précision des modèles prédictifs, notamment en régression. Elle mesure la racine carrée de la moyenne des carrés des écarts entre les valeurs prédites par le modèle et les valeurs observées.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

— RMSE élevé : Indique que le modèle a de grandes erreurs de prédiction, suggérant une performance médiocre.

— RMSE faible : Indique que le modèle a de petites erreurs de prédiction, suggérant une bonne performance.

3.5.9 R^2 (coefficient de détermination) :

Une mesure statistique qui indique la proportion de la variance dans la variable dépendante qui est prévisible à partir des variables indépendantes.

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

Ces métriques aident à évaluer et à comparer la précision et la performance des modèles de régression, permettant ainsi de choisir le modèle le plus approprié pour les prédictions.

3.6 Surapprentissage et sous-apprentisage :

Surapprentissage et sous-apprentissage(overfitting and underfitting) sont des concepts essentiels en machine learning. Ils sont les causes principales des mauvaises performances des modèles prédictifs générés par les algorithmes de machine learning.

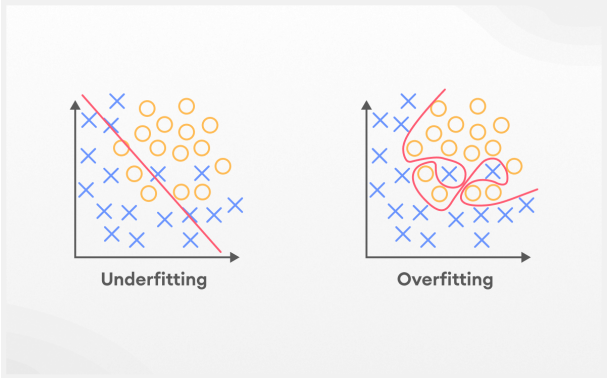

FIGURE 3.6 – Exemple sur le Surapprentissage et Sous-apprentisage

3.6.1 Surapprentissage :

Cela se produit lorsqu'un modèle mémorise les données d'entraînement au lieu d'apprendre à généraliser à de nouvelles données. En conséquence, il fonctionne bien sur les données d'entraînement, mais échoue sur les données de validation, car il ne peut pas faire de bonnes prédictions sur des données nouvelles et différentes. Pour y remédier, il est nécessaire d'améliorer la capacité de généralisation du modèle.[**17**]

Techniques de Prévention :

Il existe plusieurs méthodes pour éviter le surapprentissage d'un modèle prédictif. En voici quelques-unes :

1. **Augmenter la quantité de données d'entraînement :** Fournir plus de données au modèle durant la phase d'entraînement permet aux algorithmes de mieux généraliser, réduisant ainsi le risque qu'ils s'adaptent trop aux échantillons spécifiques.
2. **Limiter la complexité du modèle :** Un modèle trop complexe risque de mémoriser les données d'entraînement plutôt que d'apprendre des tendances générales. Garder le modèle aussi simple que possible aide à éviter cela.
3. **Utiliser la validation croisée (cross-validation) :** Cette technique consiste à diviser les données d'entraînement en plusieurs sous- groupes. Un des sous-groupes est utilisé pour le test, tandis que les autres sont utilisés pour l'apprentissage. Ce processus est répété jusqu'à ce que chaque sous-groupe ait été utilisé pour le test. La validation croisée améliore la performance des algorithmes de machine learning en permettant une évaluation plus rigoureuse du modèl

3.6.2 Sous-apprentisage :

Le sous-apprentisage peut être traduit par le sous-ajustement ou underfitting. Dans ce cas, on dit aussi que le modèle souffre d'un grand Bias (ou biais).
Cela survient lorsque l'on nomme uniquement des variables pertinentes en lien avec le problème ou lorsque l'on force le modèle à arrêter d'apprendre prématurément. Dès lors, le modèle créé est trop simpliste et risque de passer à côté de la tendance générale.
En effet,Un modèle sous-ajusté s'adapte mal aux données d'entraînement. Il présentera des prédictions biaisées faute d'entraînement suffisant et laissera apparaître de nombreuses erreurs en phase d'apprentissage.[**17**]

En d'autres termes, c'est un modèle d'apprentissage trop généraliste qui ne parvient pas à fournir des analyses prédictives précises.

Techniques de Prévention :

Pour éviter le sous-apprentisage, plusieurs techniques peuvent être employées :

— Utiliser des modèles plus complexes .
— Ajouter des caractéristiques ou des variables explicatives pertinentes.
— Augmenter la quantité de données d'entraînement disponibles.
— Appliquer des techniques de prétraitement des données pour améliorer leur qualité.
— Changer l'algorithme d'apprentissage ou ajuster les hyperparamètres pour affiner le modèle

3.7 Machine learning pour la prédiction temporelle

La prévision des séries temporelles est essentielle dans de nombreux domaines. Bien que dominées par les modèles statistiques comme ARIMA, les données réelles exigent souvent des approches non linéaires, favorisant l'usage croissant de machine learning. Cependant, cette intégration reste encore peu explorée et offre un fort potentiel de recherche.

3.7.1 Prédire avec XG Boost :

XGBoost (eXtreme Gradient Boosting) est un algorithme supervisé open source de machine learning réputé pour sa rapidité et sa polyvalence. Il s'agit d'un modèle très puissant, largement utilisé pour résoudre des tâches de classification et de régression. Il combine plusieurs arbres de décision selon un schéma de gradient boosting régularisé, ce qui permet de limiter le surapprentissage et d'améliorer la capacité de généralisation du modèle. Grâce à ses optimisations (parallélisation, gestion efficace des ressources, etc.), XGBoost offre d'excellentes performances sur les données structurées ou tabulaires.

Dans le cadre des séries temporelles, XGBoost peut également fournir des prévisions très précises, à condition d'être alimenté avec des caractéristiques temporelles appropriées. Par exemple, la création de variables retardées (lags), de moyennes mobiles ou de composantes

fréquentielles permet de capturer les motifs temporels présents dans les données. Cette ingénierie de variables permet au modèle de détecter des tendances historiques ainsi que des relations non linéaires cachées. Plusieurs expérimentations montrent d'ailleurs que l'ajout de ces caractéristiques réduit significativement les erreurs de prévision.

XGBoost est donc un modèle rapide et efficace, capable de traiter des ensembles de données volumineux et bruités. Cependant, il présente une limite notable dans le contexte des séries temporelles : il ne modélise pas de manière native la dépendance entre les observations successives. En conséquence, ses performances peuvent se dégrader sur des horizons de prévision à long terme, ce qui le rend généralement moins performant que des modèles spécifiquement conçus pour les données séquentielle.[**8**]

Hyperparamètres principaux :

Les hyperparamètres de XGBoost contrôlent le nombre d'arbres, leur complexité et le processus d'apprentissage. Parmi les plus importants, on compte :

- **`n_estimators` :**Nombre d'arbres à construire. Un compromis est nécessaire entre sous-apprentissage et surapprentissage.
- **`max_depth`** : Profondeur maximale des arbres. Une grande valeur permet au modèle de mieux apprendre des relations complexes, mais augmente le risque de surapprentissage.
- **`eta`** (ou `learning_rate`) : Taux d'apprentissage. Il contrôle à quel point chaque nouvel arbre corrige les erreurs des précédents. Une petite valeur rend l'apprentissage plus lent mais plus précis.
- **`subsample`** : Fraction des données utilisée pour entraîner chaque arbre. Réduit le surapprentissage si inférieur à 1.
- **`colsample_bytree`** : Fraction des variables (colonnes) utilisées pour construire chaque arbre.
- **`lambda`** : Paramètre de régularisation L2 (ridge) sur les poids.
- **`alpha`** : Paramètre de régularisation L1 (lasso) sur les poids. Ces deux paramètres permettent de lutter contre le surapprentissage.

3.7.2 Prédire avec les réseaux de neurones :

Les réseaux de neurones sont des modèles complexes utilisés en apprentissage supervisé pour effectuer à la fois de la régression et de la classification.
Les réseaux de neurones sont des outils reconnus pour leur puissance dans la prédiction temporelle. Ils constituent un sous-ensemble crucial de machine learning et sont au cœur des algorithmes de deep learning. Inspirés du fonctionnement du cerveau humain, où les neurones biologiques se transmettent des signaux les uns aux autres, les réseaux de neurones artificiels sont conçus pour imiter ce processus. Un réseau de neurones artificiels se compose de plusieurs couches de neurones, comprenant une couche d'entrée, une ou plusieurs couches cachées, et une couche de sortie. Cette structure permet au réseau d'apprendre des modèles complexes à partir des données en ajustant les poids des connexions entre les neurones.[**16**]
À titre de comparaison, le cerveau humain est caractérisé par un vaste réseau de neurones divers, estimé à environ 100 milliards. Ces cellules neuronales se distinguent par leur capacité à

échanger des impulsions électriques et à communiquer à travers un réseau de fibres nerveuses interconnectées. Chaque neurone est connecté à de nombreux autres neurones et possède la capacité unique de stocker, classer et traiter les informations de manière complexe.

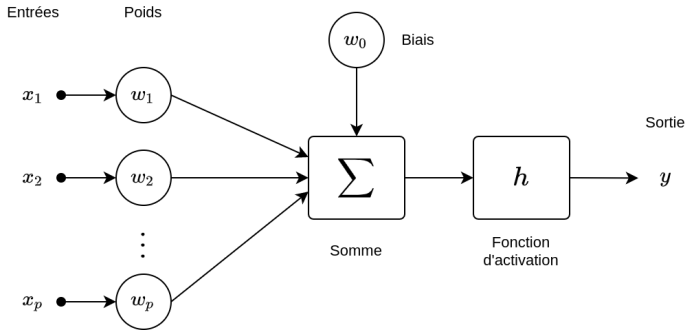

FIGURE 3.7 – réseaux de neurones

Conclusion

Ce chapitre explore le machine learning, ses types (supervisé, non supervisé, par renforcement) et ses étapes clés, de la collecte des données à l'évaluation des modèles. Il aborde les défis du surapprentissage et du sous-apprentissage ainsi que les métriques de performance. Enfin, il met en avant plusieurs techniques de prévision des séries temporelles, comme XGBoost et les réseaux de neurones. Ces modèles seront étudiés plus en détail dans le chapitre suivant, à travers leur application à notre problématique.

4

Implémentation et évaluation des modèles de prévision appliqués à un cas réel

Introduction

Dans ce chapitre, nous appliquons différentes méthodes de machine learning et d'analyse des séries temporelles afin de prédire la demande future en quantités de produits. L'objectif principal est de prédire les quantités mensuelles de queleque produits et de fournir à l'entreprise Cevital des prévisions fiables pour l'année 2025, dans le but d'optimiser la gestion des stocks et des approvisionnements. Après une phase de préparation des données, plusieurs modèles seront développés et évalués, notamment SARIMA, Réseaux de neurones, XGBoost. Les résultats permettront d'identifier les approches les plus adaptées aux besoins de prévision .
Ce travail pratique illustre ainsi l'apport essentiel de la data science dans la prise de décision logistique et stratégique dans Cevital.

4.1 Outils et environnements de travail :

4.1.1 Langage de programmation python :

Le langage Python a été retenu pour cette étude en raison de sa simplicité syntaxique, de sa polyvalence et de la richesse de son écosystème. Il est largement utilisé dans le domaine de la data science, offrant une multitude de bibliothèques adaptées à l'analyse de données, à la modélisation statistique et à l'apprentissage automatique. Sa syntaxe claire facilite la rédaction et la compréhension du code, tandis que sa communauté active assure une mise à jour continue des outils et une vaste documentation.

4.1.2 Outil de développement : Jupyter Notebook via Anaconda

L'environnement de développement utilisé est Jupyter Notebook, lancé via Anaconda. Cette combinaison permet une exécution interactive du code, une visualisation immédiate des résultats et une gestion efficace des bibliothèques, facilitant ainsi l'analyse exploratoire et la reproductibilité du travail.

4.1.3 Bibliothèques utilisées :

Dans ce travail, plusieurs bibliothèques mobilisées . Voici un aperçu :

```
# Manipulation de données
import pandas as pd  # Pour lire, manipuler et analyser des données tabulaires (DataFrame).
import numpy as np  # Pour les calculs numériques (vecteurs, matrices, fonctions mathématiques).

# Visualisation
import matplotlib.pyplot as plt  # Pour créer des graphiques (lignes, barres, histogrammes, etc.).
import seaborn as sns  # Pour des visualisations statistiques plus stylisées (heatmaps, boxplots, etc.).

# Évaluation et validation
from sklearn.model_selection import train_test_split  # Pour diviser les données en ensembles d'entraînement/test.
from sklearn.metrics import mean_squared_error, mean_absolute_error# Pour évaluer la performance des régression.
from sklearn.model_selection import cross_val_score  # Pour effectuer la validation croisée.
from sklearn.model_selection import TimeSeriesSplit  # Pour la validation croisée sur séries temporelles.

# Modèles de machine learning
from sklearn.ensemble import RandomForestRegressor # Pour appliquer le modèle de prévision Random Forest.
from sklearn.linear_model import LinearRegression # Pour appliquer la régression linéaire comme modèle de base.
from xgboost import XGBRegressor  # Pour appliquer le modèle XGBoost, efficace pour les données tabulaires.

# Traitement des séries temporelles
from statsmodels.tsa.stattools import adfuller  # Pour tester la stationnarité des séries avec le test ADF.
from statsmodels.tsa.seasonal import seasonal_decompose # Pour décomposer les séries temporelles .
from statsmodels.tsa.statespace.sarimax import SARIMAX  # Pour modéliser les séries temporelles avec SARIMA.

# Prétraitement
from sklearn.preprocessing import MinMaxScaler# Pour normaliser les données avant de les utiliser dans les RN.

# Modèle de réseau de neurones (LSTM)
from tensorflow.keras.models import Sequential  # Pour créer un modèle LSTM séquentiel.
from tensorflow.keras.layers import Dense, LSTM  # Pour construire des couches de neurones et LSTM.
from tensorflow.keras.optimizers import Adam  # Pour optimiser l'entraînement du réseau de neurones (LSTM).
```

FIGURE 4.1 – Importation des bibliothèques nécéssaires

4.2 Collecte des données :

Dans cette étude, deux bases de données au format excel ont été récupérées auprès du département commercial de l'entreprise Cevital. Ces données constituent la base sur laquelle repose l'ensemble des analyses et prévisions réalisées. :

Première base de données :

Elle contient le chiffre d'affaires mensuel de 56 produits sur une période de 12 mois. Cette base permet d'étudier l'évolution des ventes au cours de l'année **2024** et de détecter d'éventuelles tendances ou saisonnalités.

Deuxième base de données :

Elle regroupe des informations détaillées sur tout les 56 produits, telles que :

	Produit	CA Janvier	CA Février	CA Mars	CA Mai	CA Juin	CA Juillt	CA Aout	CA Sept	CA Oct	CA Nov	CA Déc
0	SUCRE SKOR 1KG	352670000	439005000	1065340000	308806000	300010000	300990500	319544000	300363200	605452800	381000000	380040000
1	SUCRE SKOR 2Kg	16137000	24205500	109974000	99046600	68411000	50831550	51638400	15491520	61966080	59096000	59044000
2	SUCRE SKOR 5KG	29937600	44906400	99875200	43887680	89812800	84303440	95800320	28740096	114960384	39500800	39251200
3	SUCRE SKOR 10Kg	3524400	5286600	7048800	6343920	10573200	11101860	11278080	3383424	13533696	28195200	22292800

FIGURE 4.2 – Extrait de quelque ligne de la Première base de données

— Le prix unitaire (Prix HT),
— Le nombre d'unités par pack (UN/PACK),
— Le nombre de packs par palette (PACK/PLTS),
— Le nombre d'unités par palette (UN/PLT),

	Caption	UN/PACK	PACK/PLTS	UN / PLT	Prix HT
0	SUCRE SKOR 1KG	10	120	1200	83.0
1	SUCRE SKOR 2Kg	4	150	600	170.0
2	SUCRE SKOR 5KG	5	224	1120	430.0
3	SUCRE SKOR 10Kg	1	100	120	90.0
4	SKOR EN SACHET VERSEUR 1KG	6	85	510	108.0

FIGURE 4.3 – Extrait de quelque ligne de la deuxième base de données

Remarque : Afin de préserver la confidentialité des données commerciales de l'entreprise, les valeurs présentes dans les bases des données utilisée ont été modifiées. Les chiffres d'affaires ont été normalisés ou exprimés à une autre échelle , ce qui signifie qu'ils ne correspondent pas directement aux volumes réels mais permettent de mener une analyse correcte de la tendance et de la saisonnalité.

4.3 Exploration et préparation des données

4.3.1 Reconstruction des données et fusion des sources

Reconstruction du mois manquant

Lors de l'importation de la base de données du chiffre d'affaires 2024, il a été constaté que les données du mois d'avril étaient absentes. Afin d'assurer la continuité temporelle, le chiffre d'affaires d'avril 2024 a été reconstitué en prenant la moyenne des mois de mars et mai. Cette estimation permet de conserver la structure saisonnière et de limiter l'impact de valeurs manquantes sur la modélisation.

Génération des données pour l'année 2023

Les données initiales ne couvrant qu'une seule année (2024), dans le cas spécifique de CE-VITAL, l'augmentation annuelle des objectifs commerciaux se situe généralement entre 10% à 15%. En se basant sur des indications transmises par l'entreprise, les chiffres d'affaires de 2023 ont été approximés à partir de ceux de 2024.

	Produit	CA Janvier_2023	CA Février_2023	CA Mars_2023	CA Avril_2023	CA Mai_2023	CA Juin_2023	CA Juillt_2023	CA Aout_2023	CA Sept_2023	...	CA Avril_2024
0	SUCRE SKOR 1KG	3.206091e+08	3.990955e+08	9.175818e+08	5.537027e+08	1.898236e+08	2.727364e+08	3.636277e+08	4.554082e+08	2.812393e+08	...	609073000.0
1	SUCRE SKOR 2Kg	1.467000e+07	2.200500e+07	1.817945e+08	1.359185e+08	9.004236e+07	6.219182e+07	4.621050e+07	4.694400e+07	1.408320e+07	...	149510300.0

2 rows × 27 columns

FIGURE 4.4 – Extrait de la base consolidée.

Fusion des deux bases

Les deux bases de départ4.2et 4.3 ont été combinées afin d'obtenir une base consolidée, qui reliait les chiffres d'affaires et les informations détaillées sur chaque produit. Par la suite, les **quantités vendues** (par unités)ont ètè calculées comme suit :

$$\text{Quantité vendue} = \frac{\text{Chiffre d'affaires}}{\text{Prix unitaire HT}}$$

4.3.2 Nettoyage des données

Valeurs manquantes

Les données manquantes ont été traitées comme suit : le mois d'avril manquant a été imputé par la moyenne, et les autres valeurs absentes ont été supprimées si elles étaient marginales.

Types de données

Les colonnes de date ont été converties au format `datetime`, permettant des opérations temporelles comme le tri ou l'extraction de mois.

Doublons

Un contrôle a été effectué pour détecter des doublons. Aucun doublon significatif n'a été trouvé.

4.3.3 Normalisation et standardisation

— **Normalisation Min-Max** : mise à l'échelle entre 0 et 1.

— **Standardisation Z-score** : centrage-réduction (moyenne = 0, écart-type = 1).

4.3.4 Valeurs aberrantes

Une analyse des valeurs aberrantes a été réalisée sur les chiffres d'affaires totaux annuels pour les années 2023 et 2024. Plusieurs produits présentent des valeurs de Chiffre d'affaires (CA) nettement supérieures à la majorité, identifiées comme outliers selon la méthode de l'écart interquartile (IQR). Toutefois, ces valeurs sont interprétées comme des résultats normaux pour des produits à fort volume ou à forte valeur, et non comme des anomalies de données. Ces observations ont donc été conservées pour l'analyse.

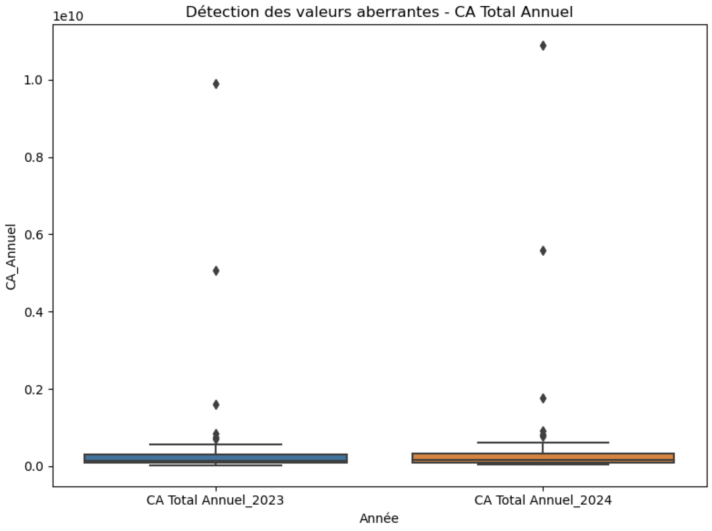

FIGURE 4.5 – boxplot du chiffre d'affaires total annuel(2023/2024)

4.3.5 Création de variables temporelles et Lags

Extraction : mois, année, trimestre

Variables cycliques :

$$\text{sin\_mois} = \sin\left(\frac{2\pi \cdot \text{mois}}{12}\right) \tag{4.1}$$

$$\text{cos\_mois} = \cos\left(\frac{2\pi \cdot \text{mois}}{12}\right) \tag{4.2}$$

Variables de décalage (lags) : pour chaque observation mensuelle, des variables ont été créées représentant les quantités vendues aux périodes précédentes :

- $t-1$: quantité du mois précédent.
- $t-2$: quantité de deux mois avant.

Enrichissement temporel :

- Jour Férié : Indicateur binaire pour modéliser l'impact des jours non ouvrés sur la demande.
- Ramadan : Indicateur binaire pour capturer l'effet saisonnier particulier des habitudes de consommation durant le Ramadan.
- Saison : Variable catégorielle (hiver, printemps, été, automne) pour modéliser les effets saisonniers.

Séparation des données : Pour évaluer la performance prédictive des modèles, les données ont été divisées en deux ensembles :

- **Ensemble d'entraînement (70 %)** : utilisé pour entraîner le modèle.
- **Ensemble de test (30 %)** : utilisé pour évaluer les performances du modèle sur des données non vues.

Le découpage s'est effectué à l'aide du code suivant :

	Produit	Date	Quantité	Jour_Ferie	Ramadan	Mois	Saison
0	SUCRE SKOR 1KG	2023-01-01	3.862760e+06	1	0	1	1
1	SUCRE SKOR 1KG	2023-02-01	4.808379e+06	0	0	2	1
2	SUCRE SKOR 1KG	2023-03-01	1.204096e+07	0	0	3	2
3	SUCRE SKOR 1KG	2023-04-01	7.437820e+06	0	1	4	2
4	SUCRE SKOR 1KG	2023-05-01	2.834677e+06	1	0	5	2
...	...	...	...	...	...	...	...
1339	E. F NECTAR DE MANGUE 1L PET	2024-08-01	3.340835e+04	0	0	8	3
1340	E. F NECTAR DE MANGUE 1L PET	2024-09-01	1.002250e+04	0	0	9	4
1341	E. F NECTAR DE MANGUE 1L PET	2024-10-01	4.009001e+04	0	0	10	4
1342	E. F NECTAR DE MANGUE 1L PET	2024-11-01	8.352086e+04	1	0	11	4
1343	E. F NECTAR DE MANGUE 1L PET	2024-12-01	1.194796e+04	0	0	12	1

1344 rows × 7 columns

FIGURE 4.6 – Le résultat final de la base donnée modifié.

```
#1. Division des données en train/test (70%/30%)
train_size = int(len(serie_mensuelle) * 0.7)
train, test = serie_mensuelle[:train_size], serie_mensuelle[train_size:]
print(f"Taille train: {len(train)} mois ({train.index.min()} à {train.index.max()})")
print(f"Taille test: {len(test)} mois ({test.index.min()} à {test.index.max()})")
```

```
Taille train: 16 mois (2023-01-31 00:00:00 à 2024-04-30 00:00:00)
Taille test: 8 mois (2024-05-31 00:00:00 à 2024-12-31 00:00:00)
```

FIGURE 4.7 – Séparation train/test

4.3.6 Agrégation et série temporelle

Les quantités ont été agrégées par mois, en sommant la quantité totale de tous les produits. On obtient ainsi une série temporelle mensuelle de janvier 2023 à décembre 2024 (24 points).

4.3.7 Analyse exploratoire des données

4.3.7.1 Évaluation mensuelle des performances de vente :

Dans cette section, on a effectué une évaluation mensuelle des performances de vente, en mettant en évidence le chiffre d'affaire total et la quantité vendue en unités pour chaque mois de de janvier 2023 jusqu'à décembre 2024. La figure4.8 représente une comparaison 2023-2024 pour le chiffre d'affaire et la quantité :

L'analyse comparative entre 2023 et 2024 révèle une nette amélioration des performances commerciales, marquée par une hausse régulière du chiffre d'affaires et des quantités vendues. Cette croissance est particulièrement prononcée durant les mois d'activité intense (mars, juillet, novembre). La progression semble principalement liée à l'augmentation des volumes, suggérant une stratégie axée sur la quantité ou une demande accrue sur le marché.

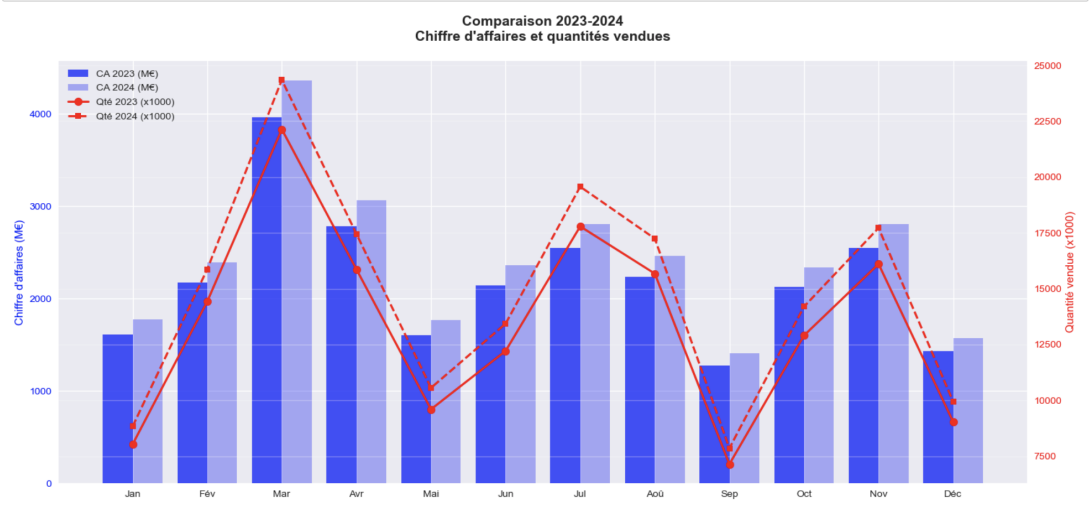

FIGURE 4.8 – CA mensuel et la quantité (2023 et 2024

4.3.7.2 produits les plus vendus :

L'analyse des volumes mensuels des produits les plus vendus avec leurs quantité en unités au cours de l'année 2024. met en évidence des dynamiques commerciales marquées par une forte saisonnalité. Le Sucre SKOR 1kg se distingue par des ventes particulièrement élevées, avec un pic significatif enregistré au mois d'mars, vraisemblablement en lien avec des périodes de forte consommation, telles que le Ramadan. Les formats d'huile ELIO II (1L et 5L) affichent également des hausses de ventes notables en début d'année, notamment en février et mars, l'eau minérale 1,5L connaît un pic de consommation en juillet, ce qui s'explique logiquement par les besoins accrus durant la saison estivale. Les variations soulignent l'importance d'une gestion des stocks adaptée aux cycles de consommation afin d'optimiser la disponibilité des produits et répondre efficacement à la demande

4.3.7.3 Analyse statistique descriptive des produits phares :

Afin de compléter l'analyse graphique des produits les plus vendus, la figure 4.10 présente les statistiques descriptives des ventes mensuelles pour les cinq produits les plus demandés. On y observe des différences notables en termes de moyenne, de dispersion (écart-type) et de valeurs extrêmes :

— Sucre SKOR 1KG présente une moyenne très élevée et une forte dispersion, ce qui confirme son statut de produit phare avec des pics de ventes importants.

— Huile ELIO II 1L montre également une moyenne élevée mais avec une variation plus modérée, ce qui suggère une consommation régulière.

— L'eau minérale 1.5L et Sucre SKOR 2KG ont des moyennes plus faibles et des écarts-types plus réduits, traduisant une demande plus stable ou saisonnière.

Ces statistiques mettent en lumière les dynamiques commerciales de chaque produit, utiles pour adapter la gestion des stocks selon leur variabilité et volume moyen de vente.

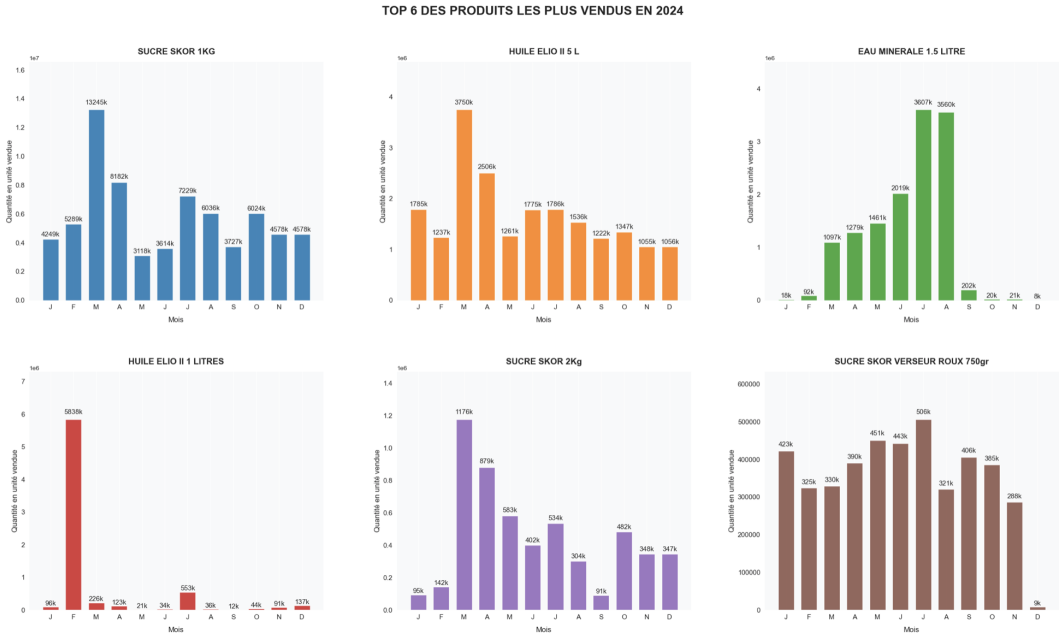

FIGURE 4.9 – Tendances historique des ventes pour les 5 top produit

Produit	count	mean	std	min	25%	50%	75%	max
EAU MINERALE 1.5 LITRE	24.0	1.523344e+06	2.016544e+06	7.130140e+03	1.965946e+04	7.476698e+05	1.963772e+06	6.821300e+06
HUILE ELIO II 1 LITRES	24.0	5.735532e+05	1.547828e+06	1.071030e+04	3.395331e+04	8.943674e+04	1.539235e+05	5.837748e+06
HUILE ELIO II 5 L	24.0	1.569156e+06	6.747128e+05	9.595180e+05	1.141302e+06	1.339042e+06	1.660459e+06	3.570818e+06
SUCRE SKOR 1KG	24.0	5.262171e+06	2.378064e+06	2.287032e+06	3.828887e+06	4.584578e+06	6.170823e+06	1.216072e+07
SUCRE SKOR 2Kg	24.0	4.095898e+05	3.018451e+05	8.284235e+04	2.394662e+05	3.316695e+05	4.940303e+05	1.176318e+06

FIGURE 4.10 – Analyse statistique descriptive des 5 top produits.

4.4 Prévision avec les séries temporelles

4.4.1 Méthodologie de Box et jenkins pour la Prévision :

La méthode de Box et Jenkins repose sur une approche systématique en quatre étapes pour modéliser et prévoir les séries temporelles, principalement utilisée pour les modèles de type ARIMA et ses variantes saisonnières (SARIMA).

4.4.1.1 Identification :

Cette étape consiste à analyser la stationnarité de la série (via la décomposition, le test de Dickey-Fuller, etc.) et à observer l'ACF et la PACF pour déterminer l'ordre des composantes .

Décomposition de la séries temporelle :

la figure 4.11 illustre la décomposition de la séries temporelle en ses compasantes principale : tendance, saisonnalité et résidu .

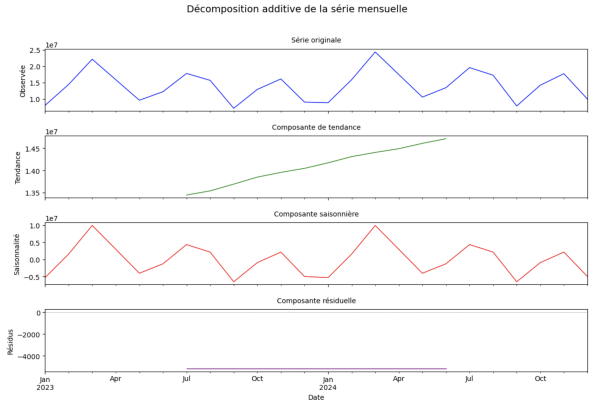

FIGURE 4.11 – Décomposition de la séries temporelle

Analyse de la stationnarité de la série temporelle à l'aide du test de Dickey-Fuller augmenté (ADF) :

```
ANALYSE DE STATIONNARITE ET SAISONNALITE
ADF Statistic: -2.488574139307292
p-value: 0.11827916890606827
 La série n'est pas stationnaire. Une différenciation est nécessaire.
```

FIGURE 4.12 – Analyse de la stationnarité

Remarque : La série temporelle n'est pas stationnaire, comme l'indique la p-value du test ADF (0.118 >0.05). Il est donc nécessaire d'appliquer une différenciation afin de stabiliser la moyenne et rendre la série stationnaire, condition essentielle pour la modélisation avec des méthodes comme ARIMA et SARIMA.

Différenciation :

```
Différenciation de la série
p-value après différenciation : 0.0457
La série est stationnaire.
```

FIGURE 4.13 – Différenciation de la sérire temporelle.

Analyse des graphiques ACF et PACF

L'identification des ordres du modèle SARIMA repose sur l'analyse conjointe des composantes non saisonnières et saisonnières à l'aide des graphiques d'autocorrélation (ACF) et autocorrélation partiell(PACF).

Pour les paramètre non saisonniers :

l'observation des graphiques ACF et PACF a permis d'identifier les valeurs suivantes :

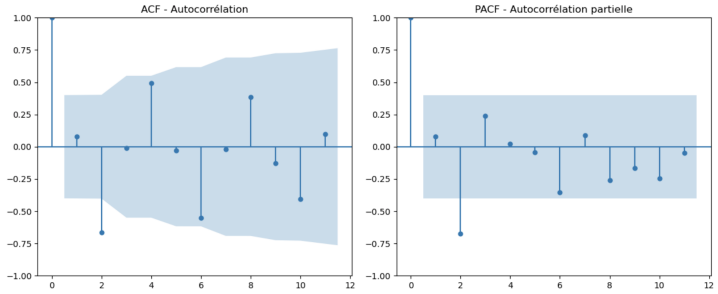

FIGURE 4.14 – Graphiques ACF et PACF de la série temporelle(Pour les paramètre non saisonniers)

— $p = 1$ (ordre autorégressif),
— $d = 1$ (car une différenciation a été appliquée),
— $q = 1$ ou 2 (ordre de la moyenne mobile),

que ce soit dans le cadre d'un modèle ARIMA ou SARIMA.

Pour les paramètre saisonniers :

Pour les paramètres saisonniers, une analyse distincte a été réalisée en appliquant une différenciation saisonnière de la série avec une période de 12 mois (diff(12)), afin d'éliminer les effets saisonniers et de mieux identifier les composantes saisonnières du modèle.

l'observation des graphiques ACF et PACF elle a permis de déterminer les paramètres suivants :

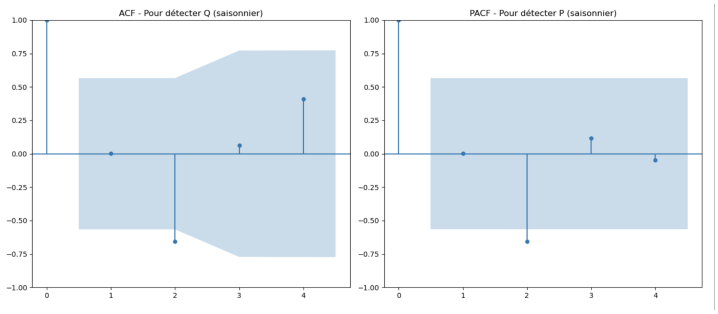

FIGURE 4.15 – Graphiques ACF et PACF de la série temporelle(Pour les paramètre saisonniers)

— P=0 : ordre saisonnier de l'AR,
— D=1 : ordre saisonnier de différenciation,
— Q=0 : ordre saisonnier de la MA.

4.4.1.2 Estimation :

Une fois les ordres du modèle SARIMA(1,1,1)(0,1,0,12) déterminés à partir de l'analyse des ACF, PACF et des tests de stationnarité, l'estimation des **coefficients** a été réalisée à l'aide de la **méthode maximum de vraisemblance** . Cette méthode permet d'ajuster automatiquement les

paramètres du modèle (coefficients AR, MA, variance de l'erreur, etc.) de manière à maximiser l'adéquation avec les données observées. La figure 4.16, généré par la fonction `summary()` de `statsmodels`, présente les coefficients estimés ainsi que leurs statistiques associées :

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                          Quantité   No. Observations:                   16
Model:             SARIMAX(1, 1, 1)x(0, 1, [], 12)   Log Likelihood                -43.614
Date:                             Tue, 06 May 2025   AIC                             93.229
Time:                                     21:46:37   BIC                             90.525
Sample:                                 01-31-2023   HQIC                            87.793
                                      - 04-30-2024
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3262      0.870      0.375      0.708      -1.379       2.031
ma.L1         -0.4003      0.719     -0.557      0.578      -1.809       1.008
sigma2      2.976e+11   2.13e-12   1.39e+23      0.000    2.98e+11    2.98e+11
===================================================================================
Ljung-Box (L1) (Q):                   2.08   Jarque-Bera (JB):                 0.29
Prob(Q):                              0.15   Prob(JB):                         0.87
Heteroskedasticity (H):                nan   Skew:                             0.10
Prob(H) (two-sided):                   nan   Kurtosis:                         1.50
===================================================================================
```

FIGURE 4.16 – Résultats du modèle SARIMA(1,1,1)(0,1,0)[12]

Les résultats montrent que :

— Le coefficient AR(1) (ar.L1) est estimé à **0.3262**,

— Le coefficient MA(1) (ma.L1) à **–0.4003**,

— La variance résiduelle (σ^2) est significative.

Bien que certains coefficients ne soient pas significatifs au seuil de 5 % (p-value > 0.05), le modèle global présente un bon ajustement. En complément, le **log de la vraisemblance** est de **–43.614**, et les critères AIC (93.229), BIC (90.525) et HQIC (87.793) confirment que le compromis entre qualité de l'ajustement et complexité du modèle est raisonnable.

4.4.1.3 Vérification :

Après l'ajustement du modèle SARIMA, une analyse des résidus a été menée pour évaluer la qualité du modèle. Un bon modèle doit produire des résidus aléatoires , sans autocorrélation significative, de moyenne nulle et de variance constante. Les résidus du modèle SARIMA se comportent effectivement comme un bruit blanc, ce qui confirme la validité du modèle.

Analyse des résidus :

L'analyse visuelle des résidus (tracé dans le temps) ne montre aucune tendance marquée, ni dérive, ce qui indique que le modèle a bien capturé la structure de la série temporelle. Le graphique de densité (histogramme + courbe KDE "Kernel Density Estimate") et le Q-Q plot suggèrent une distribution normale des résidus, avec un alignement satisfaisant sur la droite théorique.

Tests statistiques :

Le modèle SARIMA(1,1,1)(0,1,0,12) a été validé, le comportement des résidus comme un bruit blanc , ce qui conforte l'idée que le modèle est adéquat pour produire des prévisions fiables. Ainsi, nous avons réussi notre validation méthodologique.

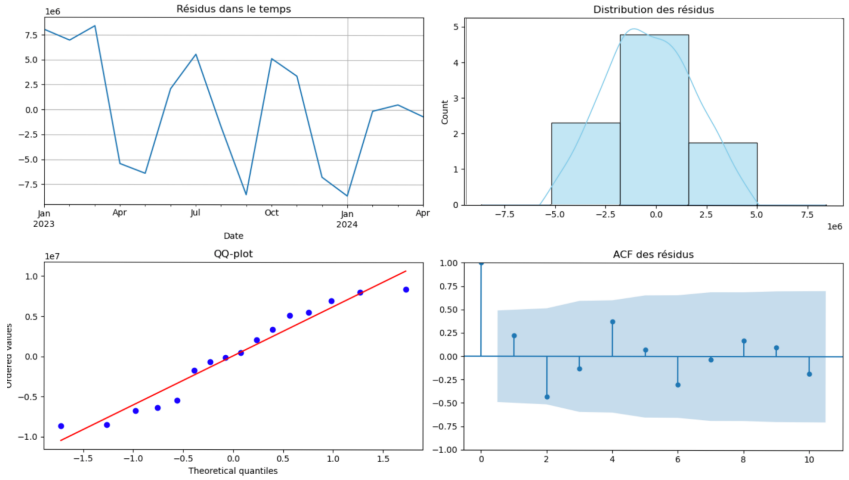

FIGURE 4.17 – Analyse diagnostique des résidus du modèle SARIMA

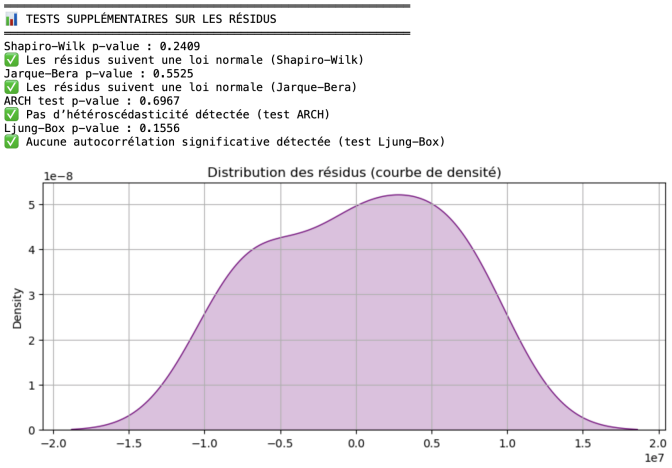

FIGURE 4.18 – Analyse des résidus : normalité et stabilité de la variance

4.4.1.4 Prévision :

Une fois le modèle validé, il peut être utilisé pour générer des prévisions fiables sur les périodes futures.

Évaluation des performances (sur données de test - année 2024) :

Le modèle a été testé sur les données de 2024, non utilisées lors de l'entraînement.
Les indicateurs obtenus sont :

Le graphique montre la comparaison entre les observations réelles de test (2024) et les prévisions du modèle SARIMA. Les deux courbes sont très proches, ce que confirment les métriques calculées dans la figure 4.20 :

Nous pouvons donc considérer que notre modèle parvient à bien décrire la dynamique de notre série temporelle. Ces résultats montrent une excellente capacité prédictive du modèle, avec des erreurs faibles et une variance expliquée élevée.

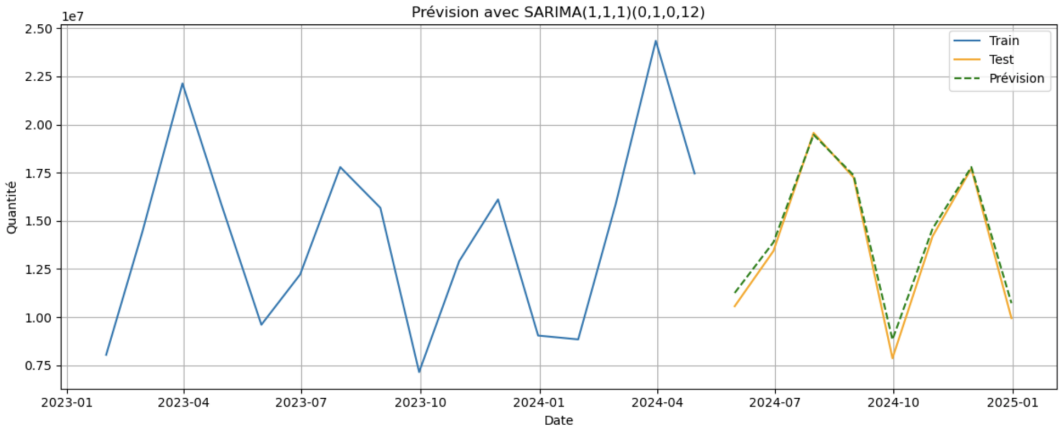

FIGURE 4.19 – Prévision du modèle SARIMA sur la période de test (année 2024)

```
📊 Évaluation sur le jeu de test (2024) :
RMSE : 49 174,79
MAPE : 3,30  %
R²   : 0,9801
```

FIGURE 4.20 – Performances du modèle SARIMA

Génération de prévisions pour 2025 :

Le modèle final a été réentraîné sur la totalité des données disponibles (2023–2024), puis utilisé pour prévoir les 12 mois de l'année 2025.

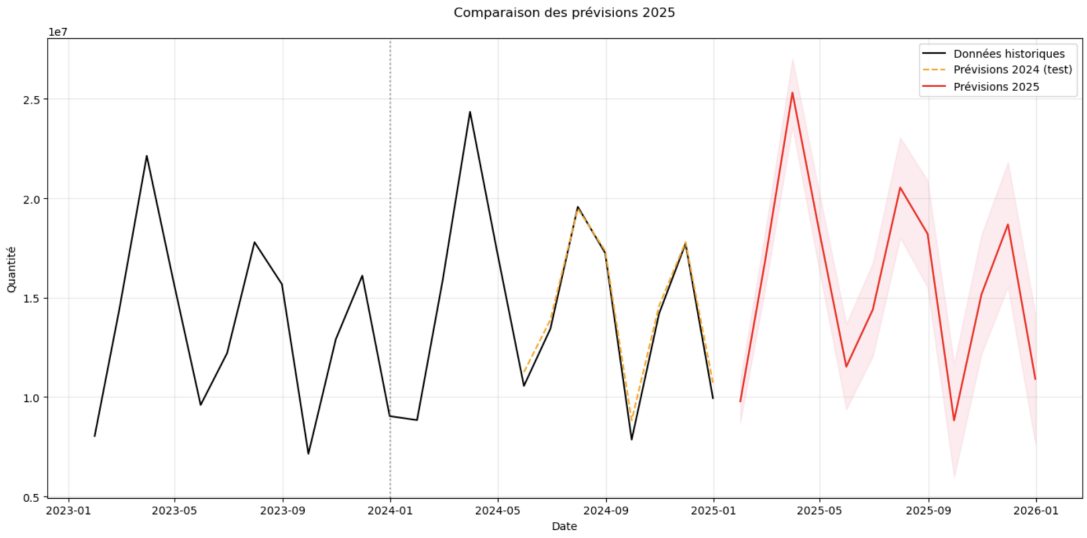

FIGURE 4.21 – Prévision du modèle SARIMA sur l'année 2025

Les résultats suggèrent une tendance haussière modérée, avec des pics de demande attendus autour de mars et juillet, probablement liés à des effets saisonniers (Ramadan, été).

L'analyse mensuelle révèle une demande particulièrement élevée prévue pour le mois de juillet 2025, tandis que janvier et novembre affichent des volumes relativement stables.

la figure 4.22 représente. les valeurs des quantités prédites mois par mois pour 2025 :

```
Prévisions mensuelles des quantités pour l'année 2025 (modèle SARIMA) :
      Date  Quantité Prédite (SARIMA)
2025-01-31               9.789814e+06
2025-02-28               1.683013e+07
2025-03-31               2.531284e+07
2025-04-30               1.842108e+07
2025-05-31               1.152857e+07
2025-06-30               1.440213e+07
2025-07-31               2.053426e+07
2025-08-31               1.821079e+07
2025-09-30               8.826306e+06
2025-10-31               1.516241e+07
2025-11-30               1.868396e+07
2025-12-31               1.091230e+07
```

FIGURE 4.22 – Prévisions mensuelles des quantités pour l'année 2025 (modèle SARIMA)

Interprétation :

Le modèle SARIMA s'est avéré performant pour modéliser la série temporelle, avec de faibles erreurs de prévision et un R^2 élevé (0.9801). Il a capturé les effets saisonniers et permis de générer des prévisions fiables pour l'année 2025. Ces résultats fournissent une base solide pour comparer SARIMA à d'autres approches comme XGBoost ou LSTM dans la suite de l'analyse.

4.5 Prévision avec le machine learning

4.5.1 Modélisation avec XGBoost :

Le modèle XGBoost a été entraîné sur l'ensemble des données mensuelles après suppression des valeurs manquantes induites par les variables de retard (*lags*). Les variables explicatives comprenaient notamment des encodages temporels, des indicateurs calendaires et des variables issues de l'historique de la série. L'entraînement a été effectué avec les hyperparamètres suivants :

— `n_estimators = 200` et `learning_rate = 0.05`

— `max_depth = 4` et `subsample = 0.8` et `colsample_bytree = 0.8`

Ces paramètres ont été choisis de manière empirique dans le but de garantir un bon compromis entre le biais (sous-apprentissage) et la variance (sur-apprentissage), tout en tenant compte de la structure de la série temporelle et des caractéristiques disponibles. Le modèle a ensuite été ajusté à l'aide de la fonction `fit()` du module `XGBRegressor`, en utilisant comme cible la variable `Quantité`.

4.5.2 Évaluation du modèle :

Le modèle a été évalué sur l'ensemble d'apprentissage et sur les données de l'année 2024. Les résultats sont les suivants :

```python
# Entraînement XGBoost
X = quantite_totale_mensuelle[features].dropna()
y = quantite_totale_mensuelle.loc[X.index, 'Quantité']

xgb_model = XGBRegressor(
    n_estimators=200,
    learning_rate=0.05,
    max_depth=4,
    subsample=0.8,
    colsample_bytree=0.8,
    random_state=42
)

xgb_model.fit(X, y)
```

```
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=0.8, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             feature_weights=None, gamma=None, grow_policy=None,
             importance_type=None, interaction_constraints=None,
             learning_rate=0.05, max_bin=None, max_cat_threshold=None,
             max_cat_to_onehot=None, max_delta_step=None, max_depth=4,
             max_leaves=None, min_child_weight=None, missing=nan,
             monotone_constraints=None, multi_strategy=None, n_estimators=200,
```

FIGURE 4.23 – Entraînement du modèle XGBoost

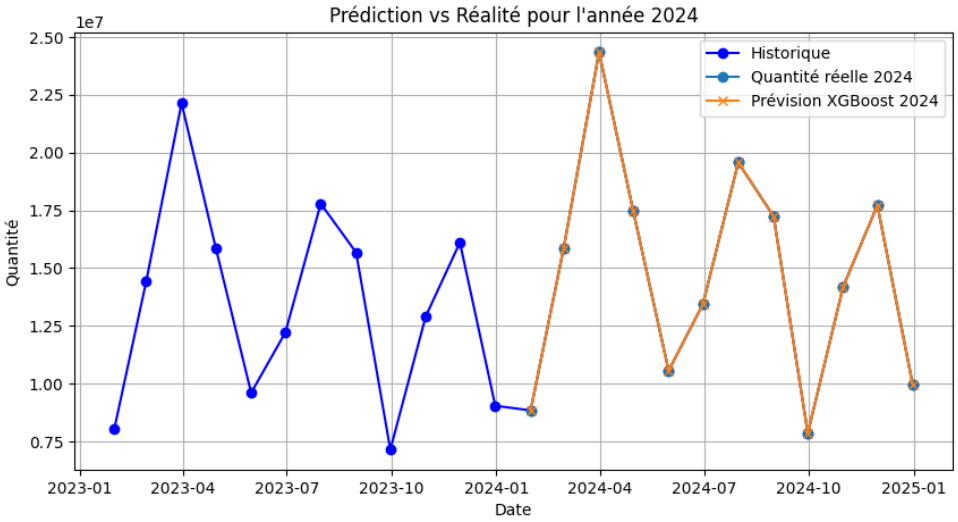

FIGURE 4.24 – Comparaison des quantités réelles et des prédictions XGBoost sur l'année 2024

Les performances du modèle **XGBoost** sont résumées dans la figure 4.25, montrant une excellente précision prédictive.

```
📊 Évaluation sur le jeu de test (2024) :
RMSE : 1281,02
MAPE : 1,22 %
R²   : 0,9998
```

FIGURE 4.25 – Performances du modèle XGBoost

4.5.3 Prévision pour l'année 2025 :

Pour effectuer les prévisions sur l'année 2025, les mêmes transformations ont été appliquées aux dates futures. Les valeurs des lags ont été dérivées à partir des 12 derniers mois disponibles. Les variables exogènes ont été remplies avec des hypothèses raisonnables (par exemple, 0 pour les jours fériés et le Ramadan par défaut). Le modèle a ensuite été utilisé pour prédire la demande mensuelle de janvier à décembre 2025.

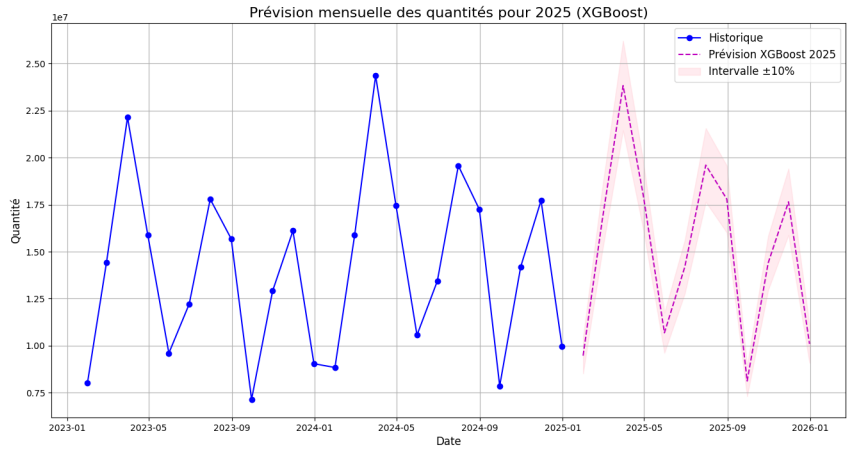

FIGURE 4.26 – Prévision pour l'année 2025 avec un intrevalle de confiance.

Ce graphique 4.26 : confirme la capacité du modèle à capturer efficacement les dynamiques saisonnières et les effets exogènes dans les séries temporelles.

La figure 4.27, représente les valeurs de la prévisions mensuelles de la quantité totale pour l'année 2025 :

Prévisions mensuelles de la quantité totale pour l'année 2025 avec XGBoost :

	Date	Quantité Prédite
0	2025-01-31	9147441.0
1	2025-02-28	16519568.0
2	2025-03-31	23929646.0
3	2025-04-30	17924514.0
4	2025-05-31	10979689.0
5	2025-06-30	14415532.0
6	2025-07-31	19436492.0
7	2025-08-31	18008130.0
8	2025-09-30	8180173.5
9	2025-10-31	14164582.0
10	2025-11-30	17726472.0
11	2025-12-31	10469907.0

FIGURE 4.27 – Prévisions mensuelles des quantités pour l'année 2025 (modèle XGBoost)

Un cas spécifique :

Afin d'aller plus loin dans l'analyse, une segmentation par produit a été appliquée. Deux produits stratégiques ont été identifiés comme étant les plus vendus ,Les résultats obtenus pour ces deux produits seront présentés dans les sections suivantes, accompagnés d'une évaluation des performances et d'une interprétation des variations mensuelles.

— **Sucre SKOR 1 kg**

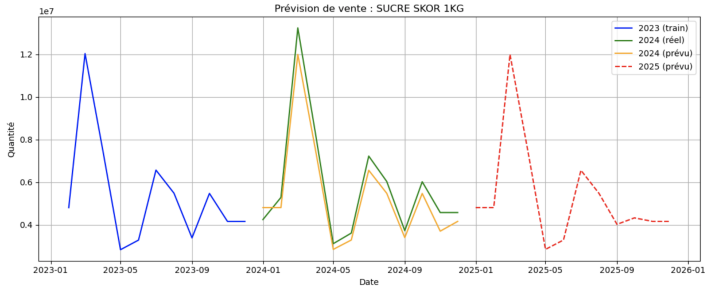

FIGURE 4.28 – Prévision mensuelle pour « Sucre SKOR 1 kg » avec XGBoost.

— **Huile 5L**.

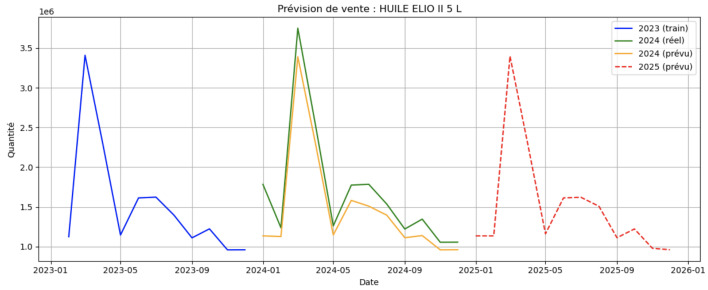

FIGURE 4.29 – Prévision mensuelle pour « Huile ELIO II 5L » avec XGBoost.

TABLE 4.1 – Performances du modèle XGBoost pour la prévision de l'huile et du sucre

Produit	R^2	**RMSE**	**MAPE**
Huile	0.94	1 303,02	2,8 %
Sucre	0.94	1 303,02	2,8 %

4.5.4 Modélisation avec les réseaux de neurones (LSTM) :

4.5.4.1 Mise en place du modèle

Le modèle LSTM (Long Short-Term Memory) a été utilisé pour modéliser la série chronologique mensuelle de ventes. Cette architecture de réseau de neurones récurrent est capable de capturer des dépendances temporelles longues, ce qui la rend particulièrement adaptée à la prévision de séries saisonnières.

Les principales étapes de préparation des données sont les suivantes :

— Agrégation des ventes mensuelles ;

— Normalisation des données avec un `MinMaxScaler` ;

— Génération de séquences avec une fenêtre glissante de 12 mois (`look_back = 12`) ;

— Construction d'un modèle LSTM simple.

TABLE 4.2 – Paramètres du modèle LSTM

Paramètre	**Valeur**
Nombre de neurones (LSTM)	100
Fonction d'activation	ReLU
Optimiseur	Adam
Taux d'apprentissage	0,005
Fonction de perte	MSE
Nombre d'époques	200
Taille de batch	1

Les hyperparamètres du modèle sont récapitulés dans le tableau 4.2.

L'entraînement a été effectué sur l'ensemble des données disponibles, l'objectif étant de générer des prévisions futures, plutôt que d'évaluer une généralisation sur échantillon indépendant.

4.5.4.2 Évaluation du modèle

Le modèle a été évalué sur un jeu de test issu des 20 % les plus récents de la série temporelle. Les prédictions ont été dénormalisées pour comparaison avec les vraies valeurs.
Les métriques d'évaluation sur l'année 2024 sont les suivantes , la figure 4.30 :
Ces résultats montrent une très bonne capacité du modèle à suivre la dynamique de la série. La

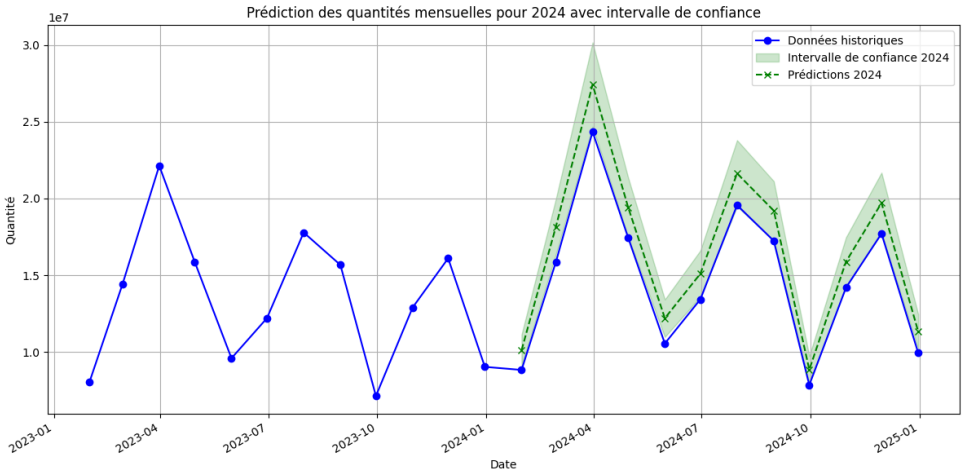

FIGURE 4.30 – Prévision LSTM pour l'année 2024 avec intervalle de confiance ±10 %

prévision de l'année 2024 a été réalisée de manière récursive à partir des 12 derniers mois de 2023.

4.5.4.3 Prévision pour l'année 2025

Une seconde phase de prévision a été conduite pour l'année 2025. Les valeurs prédites de 2024 ont servi de base pour générer les mois suivants, selon une approche récursive. Un intervalle de confiance théorique de ±10 % a été ajouté aux prédictions pour illustrer l'incertitude.

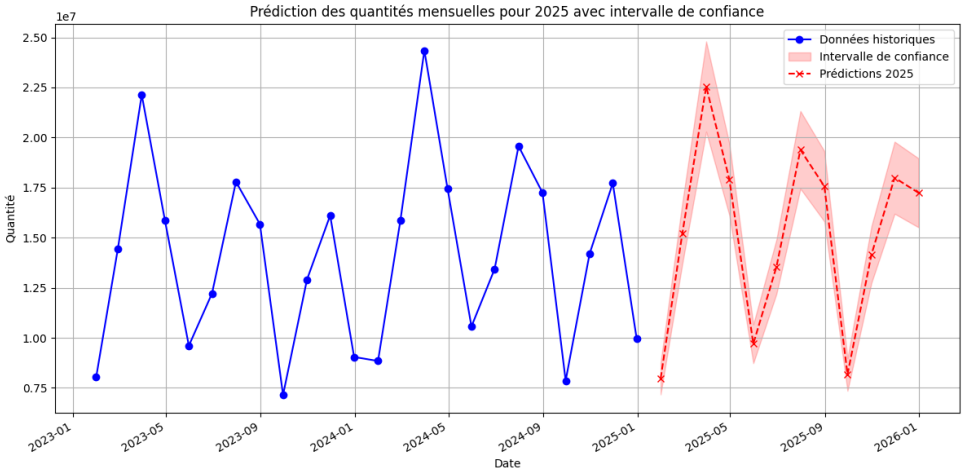

FIGURE 4.31 – Prévision des ventes mensuelles en 2025 selon le modèle LSTM

Les prévisions pour 2025 restent cohérentes et stables, illustrées par un comportement saisonnier réaliste et un intervalle de confiance raisonnable.

Globalement, le modèle est bien adapté à la série et constitue un outil fiable pour la prévision de la demande. le tableau4.32. suivante les valeur prédir pour les mois de 2025 :

```
Prévisions mensuelles de la quantité totale pour l'année 2025 :
         Date  Quantité_prédite
0  2025-01-31         9221540.0
1  2025-02-28        18052744.0
2  2025-03-31        29149960.0
3  2025-04-30        19885194.0
4  2025-05-31        11298479.0
5  2025-06-30        14715262.0
6  2025-07-31        22097820.0
7  2025-08-31        19515556.0
8  2025-09-30         7319841.5
9  2025-10-31        15104556.0
10 2025-11-30        19988266.0
11 2025-12-31         9963530.0
```

FIGURE 4.32 – Prévisions mensuelles des quantités pour l'année 2025 (modèle LSTM)

```
📊 Évaluation sur le jeu de test (2024) :
RMSE : 1 5728,07
MAPE : 0,01 %
R²   : 1,00
```

FIGURE 4.33 – Performances du modèle LSTM

"*Bien que LSTM ait donné un R^2 de 1.00 et un MAPE très bas, ces résultats pourraient indiquer un possible surapprentissage , car l'erreurs absolue (RMSE) restent élevées par rapport à XGBoost. Ce comportement est typique des modèles LSTM mal régularisés ou surentraînés.*"

4.5.4.4 Interprétation :

Les approches de Machine Learning, notamment XGBoost et LSTM, ont démontré une forte capacité à modéliser les tendances saisonnières et les effets calendaires complexes de la demande mensuelle. XGBoost s'est avéré particulièrement performant pour intégrer des variables exogènes et historiques (lags). Donc le modèle XGBoost est le meilleur dans les approches de Machine Learning.

4.6 Comparaison des modèles de prévision :

Afin d'identifier le modèle le plus adapté pour la prévision de la demande des produits Cevital, nous avons comparé les performances des différentes approches. Les critères d'évaluation utilisés sont :

— **RMSE** : Root Mean Squared Error.

— **MAPE** : Mean Absolute Percentage Error.

— **R^2** : Coefficient de détermination.

Modèle	RMSE	MAPE	R^2
SARIMA	~49174.79	~3,30 %	~0,9801
XGBoost	~1281.02	~1,22 %	~0,9998

TABLE 4.3 – Comparaison des performances des modèles de prévision sur l'année 2024

Interprétation

— Le modèle **XGBoost** est légèrement meilleur aux autres sur tous les critères de performance.

— **SARIMA** reste très fiable, surtout pour des données strictement saisonnières et sans variables exogènes.

Au vu des résultats obtenus, **XGBoost** émerge comme le meilleur compromis entre performance, robustesse et flexibilité. Il constitue ainsi le modèle recommandé pour la prévision opérationnelle de la demande chez Cevital, tout en laissant la possibilité d'exploiter SARIMA pour des analyses plus simples ou LSTM pour des scénarios plus complexes.

4.7 Modèle hybride SARIMA–XGBoost

Les modèles hybrides en séries temporelles combinent plusieurs approches de modélisation afin de tirer parti des forces de chacune. Ils visent à améliorer la précision des prévisions en capturant à la fois les composantes linéaires (telles que les tendances et les saisonnalités) et non linéaires (comme les interactions complexes ou les anomalies) des données temporelles. [**12**]
Dans le but d'améliorer la précision des prévisions mensuelles des quantités demandées, un **modèle hybride SARIMA–XGBoost** a été mis en œuvre, le modèle **SARIMA**, qui permet de capturer les composantes linéaires et saisonnières, et l'algorithme **XGBoost**, qui modélise les relations non linéaires et les résidus non expliqués par le SARIMA. Cette combinaison permet de modéliser efficacement les séries temporelles présentant à la fois des structures saisonnières régulières et des comportements non linéaires. Deux approches sont généralement possibles :

— **Modélisation séquentielle (résiduelle)** : SARIMA est appliqué en premier pour modéliser la partie linéaire, et XGBoost est ensuite entraîné sur les résidus (les erreurs de prévision de SARIMA).
— **Modélisation parallèle ou en feature stacking** : les prédictions de SARIMA sont utilisées comme une variable d'entrée (feature) parmi d'autres dans le modèle XGBoost.

4.7.1 Application du modèle hybride SARIMA–XGBoost à la prédiction des quantités mensuelles :

Dans le cadre de ce mémoire, l'objectif est de prévoir la demande mensuelle de certains produits pour l'année 2025 à partir des données historiques . Deux modèles ont été explorés individuellement :

— **Étape 1**, On commence par ajuster un modèle sur la série temporelle y_t afin de capturer les composantes linéaires et saisonnières. Le modèle SARIMA fournit une prévision $\hat{y}_t^{\text{SARIMA}}$.

$$\hat{y}_t^{\text{SARIMA}} = \text{SARIMA}(y_t)$$

— **Étape 2 :** Calcul des résidus
Les résidus e_t représentent la différence entre les valeurs observées et les prévisions du modèle SARIMA :

$$e_t = y_t - \hat{y}_t^{\text{SARIMA}}$$

— **Étape 3 :** Modélisation des résidus avec XGBoost On utilise l'algorithme XGBoost pour modéliser les résidus e_t en fonction de variables explicatives pertinentes (par exemple, des retards de la série, des indicateurs saisonniers, des variables exogènes, etc.). Le modèle XGBoost fournit une estimation des résidus :

$$\hat{e}_t^{\text{XGBoost}} = \text{XGBoost}(X_t)$$

où X_t représente les variables explicatives utilisées pour la modélisation.
— **Étape 4 :** Prévision finale du modèle hybride est obtenue en combinant la prévision du modèle SARIMA et l'estimation des résidus par XGBoost :

$$\hat{y}_t^{\text{Hybride}} = \hat{y}_t^{\text{SARIMA}} + \hat{e}_t^{\text{XGBoost}}$$

4.7.2 Résultats de la prévisions pour l'année 2025 :

Le modèle hybride a été utilisé pour générer des prévisions mensuelles de la quantité pour l'année 2025. La Figure 4.34 suivante compare graphiquement les valeurs réelles de 2023 et 2024 avec les prévisions hybrides pour 2025 :

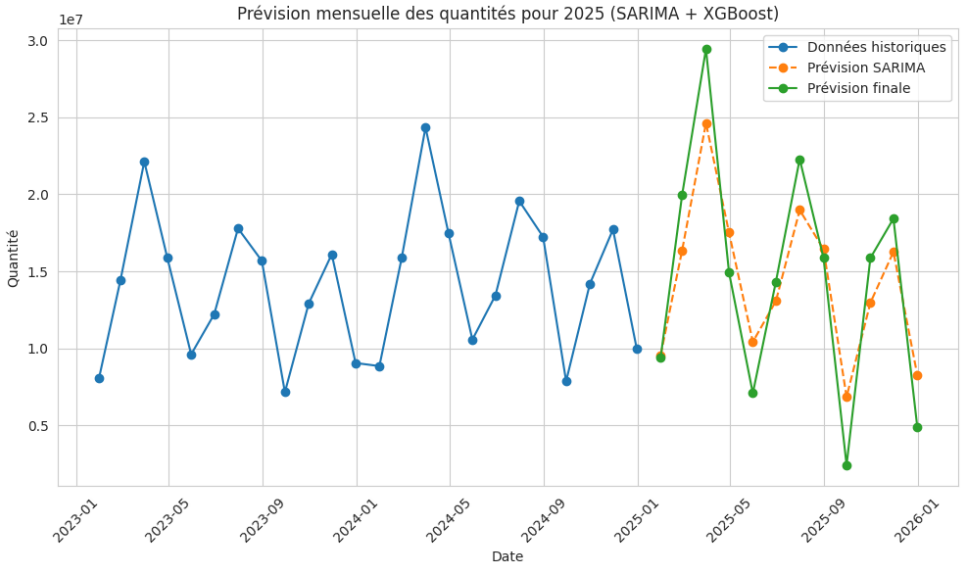

FIGURE 4.34 – La quantités mensuelle réelle (2023–2024) et prévisionnée pour 2025 à l'aide du modèle hybride SARIMA-XGBoost.

La courbe en bleue représente la série réelle pour les années 2023 et 2024, tandis que la courbe verte en pointillés correspond aux prévisions générées par le modèle hybride pour l'année 2025.
Les valeurs prédites par mois sont présentées dans la Figure 4.35 :

	Quantité_Prévue
2025-01-31	8.262249e+06
2025-02-28	1.530257e+07
2025-03-31	2.378528e+07
2025-04-30	1.689351e+07
2025-05-31	1.000101e+07
2025-06-30	1.287456e+07
2025-07-31	1.900670e+07
2025-08-31	1.668323e+07
2025-09-30	7.298742e+06
2025-10-31	1.363485e+07
2025-11-30	1.715639e+07
2025-12-31	9.384736e+06

FIGURE 4.35 – Quantités mensuelles prévues pour 2025 selon le modèle SARIMA-XGBoost

4.7.3 Évaluation des performances du modèle hybride :

Les performances du modèle SARIMA-XGBoost ont été évaluées à l'aide de plusieurs métriques standards :

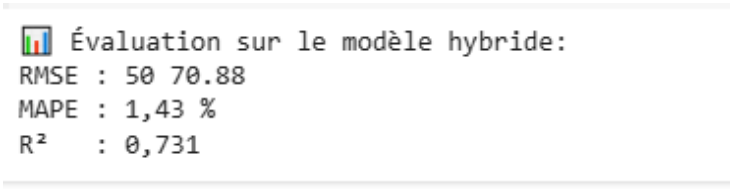

FIGURE 4.36 – Mesures de performance du modèle hybride XGBoost + SARIMA

4.7.4 Interprétation :

Le modèle hybride SARIMA + XGBoost a été mis en œuvre dans l'objectif de combiner les avantages des deux approches : la capacité du modèle SARIMA à modéliser les composantes linéaires et saisonnières de la série, et la puissance du modèle XGBoost à capturer des relations non linéaires plus complexes. Concrètement, les prédictions issues de SARIMA ont été utilisées comme base, et les erreurs résiduelles ont été modélisées par XGBoost afin de corriger les limites du modèle statistique.
Le modèle hybride présente des performances correctes , mais n'atteint pas la précision du XGBoost . Cela s'explique par le caractère peu structuré des résidus de SARIMA, qui contiennent peu d'informations supplémentaires exploitables. En conséquence, la composante XGBoost du modèle hybride ne bénéficie pas d'un gain significatif. Ce résultat illustre que l'ajout d'un modèle statistique à un algorithme de machine learning performant n'apporte pas nécessairement d'amélioration. Il conviendra toutefois de tester cette approche hybride sur d'autres séries présentant des résidus plus informatifs.

4.8 Tableau de bord de visualisation des prévisions avec Power BI

Dans le cadre de cette étude, nous avons développé un tableau de bord interactif à l'aide de **Power BI**(Power BI est un logiciel développé par Microsoft qui offre la possibilitéde collecter, analyser et visualiser des données à travers des tableaux de bord interactifs et des rapports dynamiques. On l'utilise dans le domaine de business intelligence et contribue à la prise de décision basée sur des données.), afin de visualiser les résultats des prévisions générées par le **modèle XGBoost**, qui s'est révélé être le plus performant selon les métriques d'évaluation (R^2, RMSE, MAE, etc.).
Ce dernier permet de comparer les quantités réelles enregistrées durant les années 2023 et 2024, et les quantités prévues pour l'année 2025.

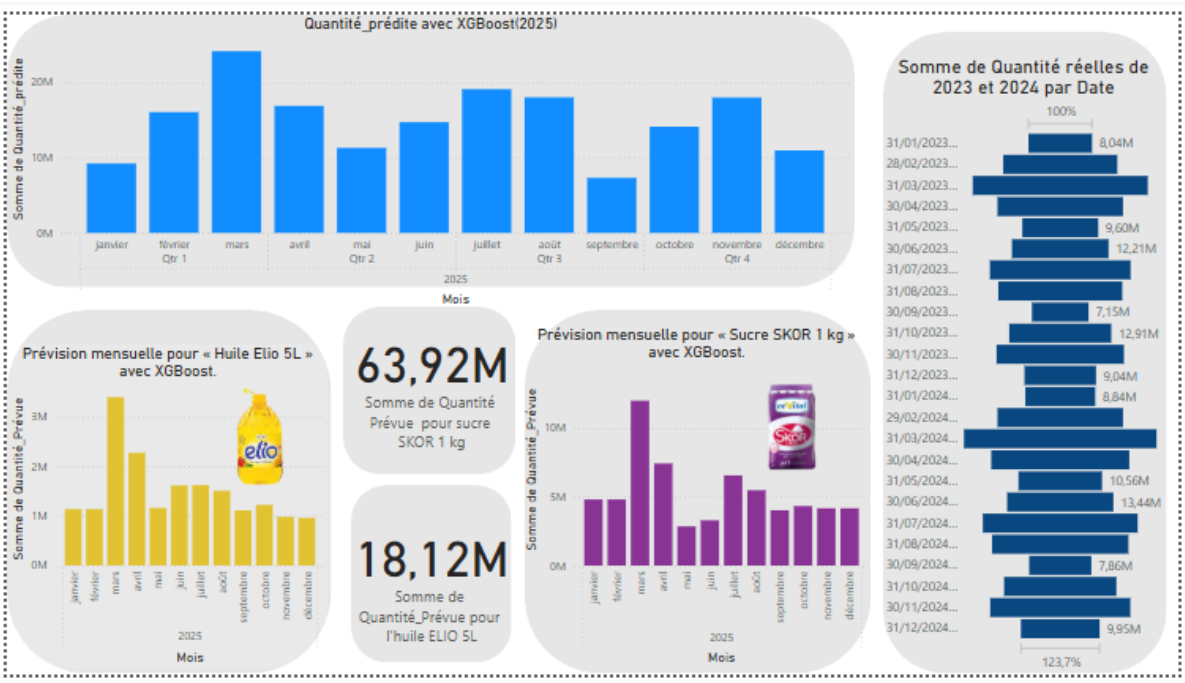

FIGURE 4.37 – Tableau de bord des prévisions de quantités pour 2025 avec XGBoost

Ce tableau de bord comprend plusieurs éléments clés :

- **Histogramme principal (haut à gauche)** : Visualisation des quantités mensuelles prévues en 2025, tous produits confondus. On observe notamment des pics en avril, juillet et décembre, indiquant une forte demande potentielle pendant ces périodes.
- **Graphique des quantités réelles (à droite)** : Comparaison des ventes mensuelles réelles sur les années 2023 et 2024. Ce graphique permet de constater l'évolution saisonnière et les tendances du marché sur deux ans.
- **Prévisions spécifiques par produit** :
  - **Huile Elio 5L** (en bas à gauche) : Le modèle prédit un total de **18,12 millions d'unités** vendues sur l'année 2025, avec une demande accrue au premier trimestre.

— **Sucre SKOR 1 kg** (en bas à droite) : Une forte demande est attendue en avril, avec une estimation totale de **63,92 millions d'unités** pour 2025.

Cette visualisation permet d'aider à la prise de décision stratégique, notamment en matière de gestion des stocks, d'achats et de distribution. Elle illustre l'apport concret du machine learning à la prévision des séries temporelles dans un contexte commercial.

Conclusion

Ce chapitre a permis de démontrer l'efficacité de différentes approches de prévision appliquées à la demande mensuelle de produits chez Cevital. Plus précisément, en utilisant des modèles classiques des séries temporelles tels que SARIMA, ainsi que des méthodes modernes d'apprentissage automatique telles que XGBoost et LSTM, nous avons pu évaluer leurs performances respectives face aux spécificités des données disponibles. Les résultats ont clairement mis en évidence la supériorité du modèle XGBoost en termes de précision et de robustesse, notamment grâce à sa capacité à intégrer des variables exogènes et à capturer la complexité des dynamiques de vente. Toutefois, les modèles SARIMA et LSTM conservent leur intérêt selon les contextes d'application . Cette analyse comparative offre ainsi une base solide pour orienter les futures décisions en matière de prévision et de gestion des stocks dans un cadre industriel. Concernant les modèles hybrides, bien que **SARIMA–XGBoost** ait été testé pour combiner structure saisonnière et non-linéarités, il n'a pas dépasser les performancesde modèle XGBoost. Cela s'explique par le fait que les résidus de SARIMA étaient peu structuré, laissant peu d'information supplémentaire à exploiter par le second modèle. Cette observation illustre que, même si l'approche hybride est théoriquement pertinente, elle ne s'avère efficace que lorsque les résidus contiennent véritablement des motifs exploitable. Enfin, afin de valoriser les résultats obtenus et d'en faciliter l'interprétation, un **tableau de bord interactif** a été élaboré. Ce dernier permet de visualiser les prévisions mensuelles pour l'année **2025**. Ce tableau constitue un outil précieux d'aide à la décision.

Conclusion générale

Au terme de ce mémoire consacré à l'analyse des séries temporelles à l'aide des algorithmes de machine learning, il apparaît clairement que la combinaison des méthodes classiques et des approches modernes d'intelligence artificielle offre un cadre robuste et performant pour la prévision des données séquentielles dans un contexte professionnel.

Dans un premier temps, la présentation de l'entreprise a permis de contextualiser les enjeux métiers liés à la prévision, soulignant l'importance stratégique d'anticiper les évolutions futures pour optimiser la prise de décision.

Le deuxième chapitre a été consacré aux séries temporelles, avec une attention particulière portée sur leurs caractéristiques fondamentales, les étapes de leur modélisation, ainsi que sur la méthode de Box-Jenkins et les modèles SARIMA. Cette partie théorique a posé les bases nécessaires à la mise en œuvre des méthodes statistiques de prévision.

Dans le troisième chapitre, nous avons exploré les principales techniques de machine learning appliquées à la prévision, en insistant sur les algorithmes XGBoost et LSTM. Leurs atouts en matière de modélisation de relations non linéaires et de capture des dépendances temporelles ont été mis en évidence, notamment dans le cadre de séries de données limitées.

Le quatrième chapitre a présenté la démarche pratique de notre étude. À partir d'une base de données mensuelle couvrant une période de 24 mois (2023–2024), nous avons développé et comparé plusieurs modèles de prévision afin d'estimer la demande mensuelle pour l'année 2025. Les modèles SARIMA, XGBoost et LSTM ont été mis en œuvre, puis évalués à l'aide de plusieurs métriques statistiques telles que le coefficient de détermination (R^2), l'erreur quadratique moyenne (MSE), la racine de l'erreur quadratique moyenne (RMSE) et le pourcentage d'erreur moyen (MPE). Les résultats ont révélé que le modèle **XGBoost** offrait la meilleure performance en termes de précision et de stabilité, suivi du modèle **SARIMA**, qui s'est montré efficace pour les séries présentant une composante saisonnière bien marquée. Le modèle **LSTM**, quant à lui, a présenté des performances moins satisfaisantes dans ce cas d'étude, probablement en raison du volume limité des données et de la complexité de son entraînement.

Finalement, dans cette partie pratique, nous avons également développé une classe spécifique mettant en œuvre un **modèle hybride** combinant SARIMA et XGBoost. L'objectif était de tirer parti des points forts de chaque méthode : SARIMA pour la capture des tendances linéaires et saisonnières, et XGBoost pour modéliser les relations non linéaires et intégrer des variables exogènes. Cependant, les résultats obtenus avec ce modèle hybride n'ont pas été à

la hauteur des attentes. Contrairement à ce que suggèrent certaines études, la performance du modèle hybride s'est révélée inférieure à celle des modèles individuels, notamment SARIMA et XGBoost pris séparément. Ce constat met en évidence que l'efficacité des modèles hybrides dépend fortement du contexte d'application, de la nature des données et du soin apporté à leur conception, afin de valoriser les résultats obtenus et d'en faciliter l'interprétation, un **tableau de bord interactif** a été élaboré. Ce tableau constitue un outil précieux d'aide à la décision pour la planification et la gestion des stocks dans un contexte industriel.

En conclusion, l'intégration des algorithmes de machine learning dans l'analyse des séries temporelles ouvre de nouvelles perspectives pour la prévision en entreprise. Ces méthodes permettent non seulement d'améliorer la précision des prédictions, mais aussi de s'adapter rapidement à des environnements changeants et à des données en constante évolution. Toutefois, il demeure essentiel de continuer à évaluer et à ajuster les modèles au fil du temps, et de garder un regard critique sur les approches innovantes, afin de garantir leur pertinence et leur efficacité opérationnelle dans un contexte où la donnée devient un levier clé de compétitivité. L'expérimentation du modèle hybride enrichit ainsi la réflexion méthodologique de ce mémoire et ouvre des pistes pour de futurs travaux visant à mieux exploiter les complémentarités entre modèles statistiques et algorithmes de machine learning.

Bibliographie

[1] GROUPE CEVITAL (2025). *Site officiel du Groupe Cevital*. [En ligne]. Disponible sur : https://www.cevital.com/en/ (consulté en 2025).

[2] CEVITAL (2025). *Organigramme Cevital*. [En ligne]. Disponible sur : https://fr.scribd.com/document/730676325/organigrame-cevital (consulté en 2025).

[3] BOULHRAOUAT, Lidia (2023). *Valorisation des déchets solides issus du complexe CEVITAL*. Mémoire de Master, Université des Frères Mentouri Constantine 1, Constantine, Algérie.

[4] HAFSI, Taïeb et KHELIF, Mouloud (2013). *CEVITAL (C) : Les Enjeux de Gouvernance et de Préparation de la Relève*. Cas pédagogique, HEC Montréal, juillet 2013. Document PDF.

[5] MONBET, V. (2017). *Modélisation des séries temporelles - Notes de cours*. Master 1, Statistique et Économétrie, Université Rennes 1..

[6] DESBOIS, Dominique (2005). *Une introduction à la méthodologie de Box et Jenkins : l'utilisation de modèles ARIMA avec SPSS*. Revue MODULAD, n° 33.

[7] Chloé-Agathe Azencott. *Introduction au Machine Learning*. Éditions Dunod, 2021.

[8] Sandha,*XGBoost for Time Series Extrapolation – An Approach in Python*, Medium, 10 juillet 2020.

[9] Andrew T. Jebb and Louis Tay. Introduction to time series analysis for organizational research : Methods for longitudinal analyses. *Organizational Research Methods*, 20(1) :61–94, 2017.

[10] Ameni Mezni. *Approches d'apprentissage automatique pour la prédiction de la qualité de performance dans les réseaux optiques opérationnels*. Mémoire de maîtrise, École de technologie supérieure, Université du Québec, Montréal, 2020.

[11] Samuel, A. L. (1959). *Some Studies in Machine Learning Using the Game of Checkers*. IBM Journal of Research and Development, 3(3), 210–229.

[12] Zhao, J., Zhao, Q., Wang, L., & Wang, X. (2022). *Hybrid time series model based on SARIMA, support vector regression and firefly algorithm for forecasting building energy consumption*. Scientific Reports.

[13] Pennsylvania State University, *: Seasonal Models*, STAT510 – Applied Time Series Analysis. Online STAT510 course notes, Department of Statistics, Penn State University.

[14] Box, G.E.P. & Pierce, D.A. (1970). *Distribution of Residual Autocorrelations in Autoregressive Moving Average Time Series Models*. Journal of the American Statistical Association, **65**(332), 1509–1526.

[15] Aiboud, L., & Laskri, S. (2020). *Appréciation de la qualité des leads dans le marketing numérique à l'aide de l'apprentissage profond*. Mémoire de Master, Novembre 2020.

[16] De Matteis, L., Steeven, M., Janny, S., Nathan, S., & Shu-Quartier, W. (2022). *Introduction à l'apprentissage automatique*. Culture Sciences de l'Ingénieur, **108**, mai 2022.

[17] Maurice, B. (2018). *Comprendre l'overfitting et l'underfitting*. [En ligne]. Disponible sur : https://deeplylearning.fr/cours-theoriques-deep-learning/comprendre-overfitting-et-underfitting/ (consulté en 2025).

[18] De Matteis, L., Steeven, M., Janny, S., Nathan, S., & Shu-Quartier, W. (2022). *Introduction à l'apprentissage automatique*. Culture Sciences de l'Ingénieur, **108**, mai 2022.

[19] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis : Forecasting and Control*. Holden-Day, San Francisco.

[20] Chloé-Agathe, Azencott (2022). *Introduction au Machine Learning* (Vol. 263, Éd. 2100834762).

[21] Bellahmer, H. (2020). *Implémentation et évaluation d'un modèle d'apprentissage automatique pour l'estimation de la valeur marchande de propriétés immobilières*.

Résumé

Ce mémoire porte sur la prévision des quantités mensuelles de produits au sein de l'entreprise Cevital, en combinant les méthodes classiques d'analyse de séries temporelles (comme SARIMA) et les algorithmes de machine learning (tels que XGBoost et LSTM). Après une présentation de l'entreprise, l'étude explore les fondements théoriques des séries temporelles et du machine learning. Une application pratique est réalisée à partir de données réelles, impliquant la préparation, l'analyse, la modélisation et la comparaison des performances des différents modèles. Le travail se conclut par le développement d'un modèle hybride SARIMA-XGBoost, jugé performant pour la prévision. L'objectif final est d'améliorer la prise de décision stratégique grâce à des prévisions plus précises. Ce mémoire s'inscrit dans le cadre d'un Master en Sciences de données et aide à la décision.

Mots-clés :

Séries temporelles,Prévision,Machine Learning,Modèle hybride,Tableau de bord interactif,Cevital,Prédiction de la demande,prise de décision.

This thesis focuses on forecasting monthly product quantities at the Cevital company by combining classical time series analysis methods (such as SARIMA) with machine learning algorithms (like XGBoost and LSTM). After presenting the company, the study explores the theoretical foundations of time series and machine learning. A practical application is conducted using real data, including data preparation, analysis, modeling, and performance comparison of different models. The work concludes with the development of a hybrid SARIMA-XGBoost model, which proves to be effective for forecasting. The main goal is to enhance strategic decision-making through more accurate predictions. This thesis is part of a Master's program in Data Science and Decision Support.

Keywords :

Time series, Forecasting, Machine Learning, Hybrid model, Interactive dashboard, Cevital, Demand prediction, Decision-making.