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

Département de Recherche Opérationnelle

Mémoire de Fin d'Études

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

Spécialité : Sciences de Données et Aide à la Décision

Utilisation du Machine Learning pour la Prévision des Ventes : Application Pratique chez Cevital

---

Présenté par : Omaira Chenouf

Sous la direction de : Mme L. Djerroud

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

Mr	N. Zougab	Professeur	Président du jury	UAMB - Béjaia
Mme	K. Bouchebah	M.C. Classe B	Examinatrice	UAMB - Béjaia
Mme	S. Amroun	M.C. Classe B	Examinatrice	UAMB - Béjaia

Année Universitaire 2024–2025

Remerciements

Je tiens tout d'abord à remercier Dieu pour m'avoir donné la force et le courage d'achever ce travail.

Je souhaite exprimer ma profonde gratitude à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire.

Je remercie chaleureusement **Madame Lamia Djerroud** pour son accompagnement précieux, sa disponibilité ainsi que ses conseils éclairés et constructifs tout au long de ce travail.

J'adresse également mes sincères remerciements à **Monsieur Yassine Samah** et **Monsieur Atman Chikh**, de l'entreprise **Cevital**, pour leur accueil chaleureux, leur soutien constant et leur accompagnement tout au long de mon stage. Leur disponibilité et leurs conseils ont été d'une aide précieuse dans la réalisation de ce travail.

Enfin, je remercie tous mes **enseignants** et mes **camarades de promotion**, pour leur partage de connaissances, leur collaboration et leur présence, qui ont contribué à enrichir mon apprentissage et mon développement personnel.

Omaira

DÉDICACES

Du fond du cœur, je dédie ce modeste travail à tous qui sont chers :

À mes chers parents,
Merci pour votre amour inconditionnel, vos sacrifices silencieux, votre patience et votre soutien constant.
Vous êtes les piliers de ma réussite. Ce mémoire est aussi le vôtre.

À mes chers frères « Adem, Yahia, Yazid »
Merci d'avoir toujours été là, de m'avoir encouragée avec bienveillance et protégée avec affection. Votre présence m'a donné la force d'aller de l'avant.

À ma chère sœur « Maha »
Ma sœur, ma meilleure amie, qui n'a pas cessé de m'écouter dans les moments difficiles, me conseiller, encourager et me soutenir tout au long de mes études.

Ma petite nièce « Léa »
Ta joie de vivre, ton innocence et ton sourire ont été une source de réconfort et de bonheur durant cette aventure académique.

À mes amies fidèles : Wawa, Lala
Merci pour votre présence, vos mots d'encouragement, vos rires partagés et votre amitié sincère.
Votre soutien a rendu ce parcours plus doux et plus fort.

Table des matières

Liste des abréviations **8**

Introduction Générale **9**

1 Présentation de l'entreprise Cevital Agro-Industriel **11**
1.1 Introduction . . . 11
1.2 Historique . . . 11
1.3 Situation Géographique . . . 12
1.4 Chiffres Clés du Groupe Cevital . . . 13
1.5 La Stratégie de Développement de Cevital . . . 14
1.6 Activités . . . 14
1.6.1 Activités à Béjaïa . . . 14
1.6.2 Activités à El Kseur . . . 15
1.6.3 Activités à Tizi Ouzou . . . 15
1.7 Produits Agroalimentaires de Cevital . . . 15
1.8 Liste des Filiales en Activité du Groupe Cevital . . . 16
1.9 La structure organisationnelle de Cevital . . . 17
1.10 Conclusion . . . 18

2 Généralités et application de l'apprentissage automatique dans la prévision des ventes **20**
2.1 Introduction . . . 20
Partie 1 : Généralités sur l'apprentissage automatique . . . 21
2.2 Définition . . . 21
2.3 Historique . . . 21
2.4 Types d'apprentissage automatique . . . 22
2.4.1 Apprentissage supervisé . . . 22
2.4.1.1 La Régression . . . 23
2.4.1.2 La Classification . . . 27
2.4.2 Apprentissage non-supervisé . . . 34
1.Algorithmes de Clustering . . . 34
2.Algorithmes de Réduction de Dimensionnalité . . . 35
2.4.3 Apprentissage par renforcement . . . 36
2.5 Domaines d'application du Machine Learning . . . 37
Partie 2 : Prévision des Ventes avec le Machine Learning . . . 37
2.6 La prévision des ventes, qu'est -ce que c'est ? . . . 37
2.7 Objectifs de la prévision des ventes . . . 37
2.8 Enjeux et impacts pour l'entreprise . . . 38

2.9 Apports du Machine Learning 38
2.9.1 Pourquoi utiliser le Machine Learning ? 38
2.10 Avantages du Machine Learning dans la prévision des ventes 38
2.11 Défis liés à la prévision des ventes avec le Machine Learning 39
2.11.1 Données incomplètes ou bruitées 39
2.11.2 Variabilité du marché et événements externes 39
2.11.3 Choix du bon modèle et sur-apprentissage 39
2.12 Étapes de mise en œuvre d'un modèle ML de prévision 39
2.12.1 Collecte des données de ventes 39
2.12.2 Prétraitement et nettoyage 40
2.12.3 Analyse exploratoire 40
2.12.4 Sélection et entraînement du modèle ML 40
2.12.5 Évaluation des performances 40
2.12.6 Déploiement et intégration dans le processus décisionnel 40
2.13 Problématique 41
2.14 Conclusion 41

3 Implémentation d'un Modèle de Prévision des Ventes par Machine Learning chez Cevital 42
3.1 Introduction 42
Partie 1 : Prévision des ventes pour l'année 2025 42
3.2 Outils Utilisés pour le Développement 43
3.2.1 Langage de programmation : Python 43
3.2.2 Environnement de développement : Jupyter Notebook 43
3.2.3 Bibliothèques utilisées 43
3.3 Collecte des données 44
3.4 Prétraitement des données 44
3.4.1 Suppression des valeurs nulles et négatives 45
3.4.2 Détection et traitement des valeurs aberrantes par la méthode IQR 45
3.5 Analyse Exploratoire des Données (AED) 47
3.5.1 Statistiques Descriptives 47
3.5.2 Répartition des ventes totales par produit (2023–2024) 48
3.5.3 Les produits les plus vendus 48
3.6 Modélisation 49
3.6.1 Préparation et séparation des données 49
3.6.2 Choix du modèle 50
3.6.3 Optimisation des hyperparamètres 50
3.6.4 Évaluation des performances globales 50
3.6.5 Prévision pour l'année 2025 50
3.6.5.1 Prédictions pour l'année 2025 51
3.6.5.2 Répartition des Ventes Prévues par Produit 2025 . . . 52
3.6.6 Prévision pour les Top 5 sauces les plus vendus 52
Partie 2 : Analyse Visuelle des Résultats de Vente via un Tableau de Bord BI 55
3.7 Outils utilisés 55
3.8 Description des éléments du tableau de bord 55
3.8.1 Graphique de comparaison des ventes 55

3.8.2 Carte de score : quantité totale vendue 56
3.8.3 Premier produit "MAYONNAISE FULL FAT (VERRE) 450g" . 56
3.8.4 histogramme des meilleures ventes 57
3.8.5 Courbe mensuelle : évolution des ventes en 2025 58
3.9 Présentation du tableau de bord complet 58
3.10 Conclusion 59

Conclusion Générale **60**

Bibliographie **60**

Table des figures

1.1 Historique du Groupe Cevital 12
1.2 Localisation de Cevital à Béjaïa 13
1.3 Déploiement des gammes agroalimentaires de Cevital (1998–2023) . . . 16
1.4 Structure organisationnelle de Cevital 17

2.1 Relation entre l'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond 21
2.2 Les types d'apprentissage automatique 22
2.3 Types d'apprentissage supervisé 23
2.4 Exemple de régression linéaire 24
2.5 Exemple sur la régression logistique. 25
2.6 Exemple de régression polynomiale. 26
2.7 Arbre de décision pour classifier les animaux 28
2.8 Caractéristiques principales de XGBoost 28
2.9 Illustration de la Forêt Aléatoire. 30
2.10 Séparation linéaire par un SVM avec hyperplan et marges. 31
2.11 Principe de classification avec l'algorithme des K plus proches voisins (KNN). 31
2.12 Matrice de confusion pour un problème de classification binaire 32
2.13 Algorithme d'apprentissage non supervisé 34
2.14 Illustration du clustering en apprentissage non supervisé. 35
2.15 Illustration de l'ACP (Analyse en Composantes Principales) 36
2.16 Schéma du processus d'apprentissage par renforcement 36

3.1 Bibliothèques utilisées pour la mise en œuvre du modèle de prévision . 44
3.2 Aperçu des premières lignes du jeu de données 44
3.3 Nombre de lignes restantes après la suppression des valeurs négatives et nulles 45
3.4 Détection des valeurs aberrantes par la méthode IQR pour chaque variable numérique 46
3.5 Comparaison des boîtes à moustaches avant et après suppression des outliers (méthode IQR) 46
3.6 Résumé statistique des variables 47
3.7 Répartition des ventes totales par produit (2023–2024) 48
3.8 Les 10 produits les plus vendus (2023–2024) 49
3.9 Visualisation des meilleurs paramètres 50
3.10 Métriques globales sur toutes les ventes. 50
3.11 Comparaison entre les ventes réelles et prédites sur l'ensemble de test (20 %). 51

3.12 Prévisions mensuelles des ventes pour l'année 2025. 51
3.13 Répartition des ventes prévues par produit pour l'année 2025 52
3.14 Comparaison des valeurs réelles et prédites sur l'ensemble de test pour les cinq produits les plus vendus. 53
3.15 Prévisions pour l'année 2025 des cinq produits les plus vendus. 54
3.16 Répartition des volumes de ventes 55
3.17 Somme totale des ventes prévues pour 2025 56
3.18 Produit le plus vendu en 2025 56
3.19 Top produits les plus vendus en 2025 57
3.20 Évolution mensuelle des ventes en 2025 58
3.21 Vue d'ensemble du tableau de bord de suivi des ventes (2023–2024) . . 59

Liste des abréviations

Abréviation	**Définition**
IA	Intelligence Artificielle
LASSO	Least Absolute Shrinkage and Selection Operator
Q-Learning	Quality Learning
DQN	Deep Q-Network
PPO	Proximal Policy Optimization
DDPG	Deep Deterministic Policy Gradient
Power BI	Power Business Intelligence
MAYO FF VER 450g	MAYONNAISE FULL FAT (VERRE) 450g
MAYO ELIO PET 220g	SAUCE MAYONNAISE ELIO (PET) 220g
MAYO FF VER 220g x12	MAYONNAISE FULL FAT (VERRE) 220g, colis de 12 bocaux
MAYO FF PET 395g	MAYONNAISE FULL FAT (PET) 395g
MAYO FF PET 200g	MAYONNAISE FULL FAT (PET) 200g

Introduction Générale

Dans un environnement économique de plus en plus compétitif, les entreprises industrielles doivent faire face à de multiples défis : anticiper la demande, optimiser les ressources, réduire les coûts, tout en garantissant la satisfaction des clients. Ces enjeux sont d'autant plus cruciaux dans les secteurs à forte variabilité de la demande, comme l'agroalimentaire. Dans ce contexte, la prévision des ventes devient un élément stratégique incontournable, permettant une meilleure planification de la production, une gestion plus efficiente des stocks, et une optimisation globale de la chaîne logistique.

L'entreprise Cevital, acteur majeur de l'industrie agroalimentaire en Algérie, est confrontée à ces problématiques au quotidien. À travers ses différentes unités de production et ses dizaines de références produits, la prévision précise de la demande constitue un levier important pour améliorer la performance opérationnelle et répondre efficacement aux attentes du marché.

L'évolution technologique a favorisé l'émergence de nouvelles approches, notamment celles issues de l'intelligence artificielle (IA). Cette dernière regroupe un ensemble de techniques permettant aux machines d'imiter certaines capacités humaines, telles que l'apprentissage, le raisonnement ou la prise de décision. Parmi ces techniques, le Machine Learning (ML), ou apprentissage automatique, occupe une place prépondérante. Il s'agit d'un sous-domaine de l'IA qui permet aux systèmes d'apprendre à partir de données historiques afin d'effectuer des prédictions ou de prendre des décisions sans être explicitement programmés pour chaque situation.

Ce mémoire se compose de trois chapitres :

Le premier est dédié à la présentation de l'entreprise Cevital, en abordant son historique, ses implantations géographiques, ses principales activités, ses filiales ainsi que sa structure organisationnelle.

Le deuxième chapitre introduit les notions fondamentales de l'apprentissage automatique, en mettant en évidence ses différentes catégories (supervisé, non supervisé, par renforcement) et ses applications concrètes dans le domaine de la prévision des ventes. Nous y abordons également les étapes clés de mise en œuvre d'un modèle prédictif, les défis associés à l'utilisation des données réelles, et les bénéfices apportés par ces nouvelles approches.

Enfin, le dernier chapitre portera sur la mise en œuvre d'un modèle de machine learning pour prévoir les ventes de sauces de l'année 2025 chez Cevital. Il décrira les étapes du projet, depuis la collecte et le prétraitement des données historiques (2023 et 2024), jusqu'à l'entraînement du modèle avec Python. Les résultats seront ensuite intégrés dans un tableau de bord interactif développé avec Power BI, afin de visualiser les prévisions de manière claire.

Ce mémoire se termine par une conclusion générale qui résume les principaux résultats obtenus dans l'application du machine learning à la prévision des ventes.

Chapitre 1

Présentation de l'entreprise Cevital Agro-Industriel

1.1 Introduction

Cevital est une entreprise privée algérienne fondée en 1998 par Issad Rebrab. En quelques décennies, elle est devenue un acteur majeur de l'économie algérienne et internationale grâce à une stratégie de diversification et d'innovation constante.

Ce chapitre a pour objectif de présenter l'entreprise Cevital Agro-Industrie. Il s'agit d'un acteur majeur de l'agroalimentaire en Algérie, dont l'envergure industrielle, l'implantation géographique et la stratégie de développement méritent une attention particulière. La présentation commence par un aperçu historique de la société, suivi de sa localisation géographique, de ses chiffres clés, et de ses principales activités réparties sur plusieurs sites (notamment à Béjaïa, El Kseur et Tizi Ouzou). Nous abordons ensuite la gamme variée de produits agroalimentaires proposée par Cevital, ses différentes filiales en activité, ainsi que sa structure organisationnelle.

Cette présentation vise à donner une vision globale de l'entreprise et à montrer comment elle a su allier performance économique, innovation et responsabilité sociétale [1].

1.2 Historique

Première entreprise privée algérienne à avoir investi dans des secteurs d'activités diversifiés, elle a traversé d'importantes étapes historiques pour atteindre sa taille et sa notoriété actuelle.

Industrie agroalimentaire et grande distribution, électronique et électro-ménager, sidérurgie, industrie du verre plat, construction industrielle, automobile, services, médias… Le Groupe Cevital s'est construit, au fil des investissements, autour de l'idée forte de constituer un ensemble économique.

Porté par 18 000 employés répartis sur 3 continents, il représente le fleuron de l'économie algérienne, et œuvre continuellement dans la création d'emplois et de richesse.
La figure suivante retrace les principales étapes de l'histoire de Cevital, illustrant l'évolution du groupe à travers les différentes étapes clés de son développement.

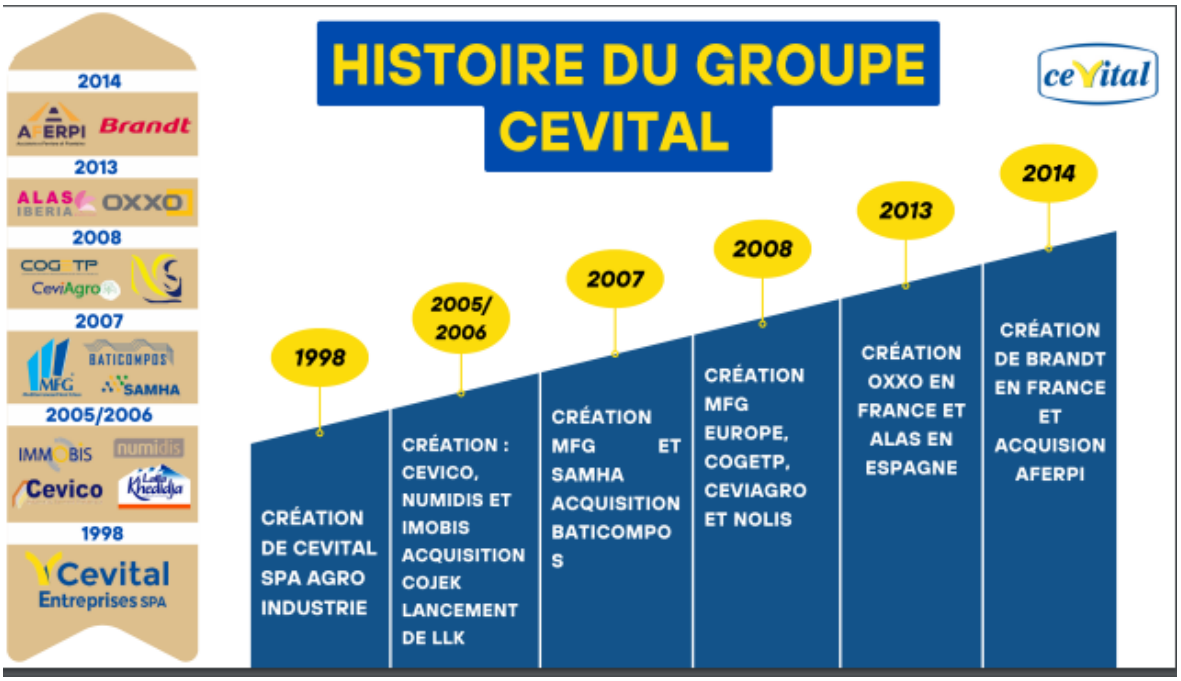


FIGURE 1.1 – Historique du Groupe Cevital

1.3 Situation Géographique

Cevital est l'une des plus grandes entreprises d'Algérie et le leader du secteur agroalimentaire. Son complexe de production est situé sur le nouveau quai du port de Béjaïa, à 3 km au sud-ouest de la ville, à proximité des routes nationales RN 26 et RN 9. Cette localisation stratégique lui confère un avantage économique majeur en raison de sa proximité avec les infrastructures de transport.

En effet, le complexe est proche à la fois du port et de l'aéroport de Béjaïa, facilitant ainsi l'importation des matières premières et l'exportation des produits finis. Il s'étend sur une superficie de 45 000 m², faisant de lui le plus grand complexe privé en Algérie. Sa capacité de stockage s'élève à 182 000 tonnes par an (silos portuaires), tandis que son terminal de déchargement portuaire permet une réception de matières premières à un débit de 200 000 tonnes par heure.

De plus, Cevital possède un vaste réseau de distribution comptant plus de 52 000 points de vente à travers tout le territoire national (voir la Figure 1.2).

FIGURE 1.2 – Localisation de Cevital à Béjaïa

1.4 Chiffres Clés du Groupe Cevital

Le groupe Cevital, en tant que premier groupe privé algérien, occupe une place prépondérante dans l'économie du pays. Ces chiffres reflètent son poids économique et son rayonnement tant au niveau national qu'international.

- **Leader privé en Algérie** : Premier groupe privé algérien, acteur clé du développement économique.
- **Présence internationale** : 26 filiales réparties sur 3 continents.
- **Poids en termes d'emploi** : Plus de 18 000 employés, premier groupe employeur privé.
- **Performances financières** : 4 milliards de dollars de chiffre d'affaires.
- **Leader agro-industriel** : Premier groupe agro-industriel en Afrique.
- **Impact économique** : Premier exportateur hors hydrocarbures et premier contributeur privé au budget de l'État.
- **Diversité des métiers** : Plus de 10 métiers différents.
- **Croissance continue** : Croissance annuelle moyenne de 30%.

1.5 La Stratégie de Développement de Cevital

Cevital s'est construit autour de l'ambition et de la vision de son fondateur, Issad Rebrab, de bâtir un groupe industriel d'envergure mondiale, très compétitif, tourné vers l'exportation et l'international.

Le groupe dispose aujourd'hui d'unités de production de taille mondiale, équipées des technologies les plus avancées. Sa stratégie repose sur une forte compétitivité en matière de prix, de qualité, de volumes, de logistique, de robotisation et de co-localisation.

Une place importante est accordée à la recherche et développement, à l'innovation ainsi qu'au talent de ses collaborateurs. Ces facteurs constituent le socle d'une industrie dynamique, exportatrice, créatrice d'emplois et attractive pour la jeunesse algérienne.

Selon Issad Rebrab, le succès du groupe repose sur les **sept piliers suivants** :

- Le **réinvestissement systématique** des gains dans des secteurs porteurs à forte valeur ajoutée ;
- La **recherche et l'adoption de technologies de pointe** ;
- Une attention particulière au **choix, à la formation et au transfert de compétences** des collaborateurs ;
- L'encouragement de l'**esprit d'entreprise** ;
- Le **sens de l'innovation** ;
- La **recherche permanente de l'excellence** ;
- La **fierté et la passion de contribuer à l'économie nationale**.

1.6 Activités

Cevital exerce différentes activités industrielles dans plusieurs localités d'Algérie.

1.6.1 Activités à Béjaïa

Dans la commune de Béjaïa, Cevital développe plusieurs activités dans le secteur de l'industrie agroalimentaire, notamment :

- La production de margarine.
- Le raffinage du sucre.
- Le raffinage des huiles alimentaires.

1.6.2 Activités à El Kseur

À El Kseur, Cevital a réhabilité l'unité de production de jus de fruits COJEK. Cette unité, initialement mise en exploitation en 1978 sous l'égide de SOGEDIA, a été reprise par ENAJUC en 1982 avant d'être cédée à Cevital en novembre 2006. Elle fonctionne sous le statut de société par actions avec un capital de 1 007 000 000 DA.

— Capacité de production actuelle : 14 400 tonnes par an.
— Objectif de développement : 150 000 tonnes par an en 2010.

1.6.3 Activités à Tizi Ouzou

Dans la wilaya de Tizi Ouzou, plus précisément dans la commune d'Agouni Gueghrane, située au cœur du massif montagneux du Djurdjura à plus de 2300 mètres d'altitude, Cevital possède une unité de production et de conditionnement d'eaux minérales sous la marque **Lala Khedidja**. Cette unité a été inaugurée en juin 2007.

1.7 Produits Agroalimentaires de Cevital

Le pôle agroalimentaire de Cevital constitue l'un des segments les plus importants de ses activités industrielles. Grâce à des installations modernes et à une stratégie axée sur la qualité et l'innovation, Cevital propose une large gamme de produits destinés aussi bien au marché national qu'à l'export.

Les principales catégories de produits agroalimentaires sont :

1. Huiles végétales
2. Margarinerie et graisses végétales
3. Sucre blanc
4. Sucre liquide
5. Silos portuaires
6. Boissons
7. Sauces

La figure ci-dessous présente les différents lancements de produits agroalimentaires de Cevital, répartis sur la période allant de 1998 à 2023.

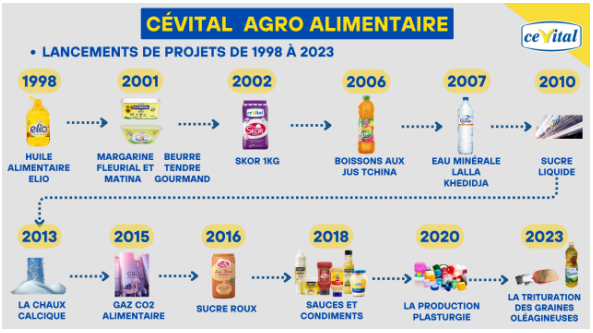


FIGURE 1.3 – Déploiement des gammes agroalimentaires de Cevital (1998–2023)

1.8 Liste des Filiales en Activité du Groupe Cevital

Afin de mieux comprendre l'envergure et la diversification du Groupe Cevital, il est utile de passer en revue ses principales filiales. Ces dernières couvrent une large gamme de secteurs allant de l'agroalimentaire à l'électroménager, en passant par la logistique, la métallurgie, le bâtiment, l'énergie, la finance et la communication. Le tableau ci-dessous présente les filiales actuellement en activité, ainsi que leur domaine d'intervention.

Filiale	**Secteur d'activité**
CEVITAL AGRO	Groupe agroalimentaire
SAMHA	Électroménager
BRANDT	Électroménager
WEG Algérie	Électroménager
MFG	Industrie du verre
SOLARIS	Panneaux solaires
OXXO	Fenêtres PVC
ALSEV	Bâtiment
CEVITAL MINERAL	Minerai
ALLIANCE GLASS	Glace et miroir
NUMILOG	Logistique
CT LOG	Logistique
METAL SIDER	Métallurgie
METAL STRUCTURE	Métallurgie
BATICOMPOS	Bâtiment
COGETP	Équipements BTP
AAC	Automobile
IMMOBIS	Gestion immobilière
PROMOTION IMMO	Gestion immobilière
KEEP CONTACT	Services
ANTEI	Trading d'actifs financiers
Futur Media	Communication

TABLE 1.1 – Liste des principales filiales en activité du Groupe Cevital

1.9 La structure organisationnelle de Cevital

Cevital agro-alimentaire est organisée en différentes composantes pour assurer une gestion efficace et une production optimale. dont la structure organisationnelle est représentée dans la Figure 1.4

FIGURE 1.4 – Structure organisationnelle de Cevital

1.Direction Générale

Elle assure la coordination globale du groupe, fixe les grandes orientations stratégiques, veille à la performance économique et à la cohérence des actions de toutes les directions.

2.Direction Achats et Approvisionnements

Elle gère l'acquisition de matières premières, biens et services nécessaires à l'activité de l'entreprise. Elle optimise les coûts, la qualité et les délais des achats.

3.Direction Supply Chain

Elle est chargée de la planification, de la gestion des stocks, de la logistique et du transport, afin d'assurer la disponibilité des produits tout au long de la chaîne de valeur.

4.Direction Marketing

Elle conçoit et met en œuvre les stratégies de communication, de marque, de positionnement et d'études de marché pour promouvoir les produits du groupe.

5.Direction Qualité

Elle garantit que les produits et processus respectent les normes de qualité exigées (internes et externes), en mettant en place des contrôles et démarches d'amélioration continue.

6.Direction Commerciale B2C (Business to Consumer)

Elle gère les relations commerciales avec les clients particuliers, notamment via les canaux de distribution directe (points de vente, supermarchés, e-commerce…).

7.Direction Commerciale B2B (Business to Business)

Elle développe les partenariats et ventes avec d'autres entreprises, revendeurs ou distributeurs en gros. Elle gère les grands comptes clients.

8.Direction des Ressources Humaines

Elle gère la politique de recrutement, la formation, le développement des compétences, la gestion des carrières et le climat social au sein de l'entreprise.

9.Direction Finances et Comptabilité

Elle supervise la gestion budgétaire, la trésorerie, la comptabilité générale et analytique, les audits internes, ainsi que la conformité financière.

10.Direction Systèmes d'Informations (DSI)

Elle développe et maintient les infrastructures informatiques, les systèmes de gestion (ERP, CRM…), la cybersécurité et la digitalisation des processus internes.

1.10 Conclusion

En conclusion, Cevital se positionne comme un acteur majeur dans le secteur industriel en Algérie, avec une stratégie axée sur l'innovation, la diversification de ses activités et la recherche de l'excellence dans la gestion de ses ressources. Son expertise s'étend à plusieurs secteurs clés, dont l'agroalimentaire, la distribution, et les technologies, ce qui lui permet de jouer un rôle déterminant dans l'économie nationale. L'engagement de l'entreprise envers

la modernisation de ses processus industriels et la mise en place de solutions technologiques de pointe reflète son ambition de devenir un leader régional et international.

Chapitre 2

Généralités et application de l'apprentissage automatique dans la prévision des ventes

2.1 Introduction

L'apprentissage automatique représente une avancée majeure dans le domaine de la science des données, transformant profondément la manière dont les entreprises anticipent et planifient leurs ventes. En analysant de vastes volumes de données historiques et en détectant des schémas complexes, il permet d'établir des prévisions plus précises et fiables de la demande future.

Grâce à ces capacités prédictives, les entreprises peuvent optimiser la gestion de leurs stocks, adapter leurs stratégies commerciales, améliorer la satisfaction client et réagir plus rapidement aux fluctuations du marché. l'apprentissage automatique offre également des informations en temps réel, facilitant une prise de décision stratégique plus réactive et éclairée.

Dans un environnement économique de plus en plus volatil, la prévision des ventes demeure une fonction essentielle . L'intégration des techniques de l'apprentissage automatique permet de surmonter certains défis liés à l'incertitude, à la variabilité de la demande et à la complexité des données, en s'appuyant sur des méthodes robustes de traitement et d'analyse.

Ce chapitre s'intéressera d'abord aux principes de base de l'apprentissage automatique, puis explorera en détail son application concrète à la prévision des ventes, en soulignant ses apports, ses limites et les différentes étapes de sa mise en œuvre.

Partie 1 : Généralités sur l'apprentissage automatique

2.2 Définition

L'apprentissage automatique "en anglais :machine learning" ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'« apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes. On parle d'apprentissage statistique car l'apprentissage consiste à créer un modèle dont l'erreur statistique moyenne est la plus faible possible[2].
La figure(2.1) illustre la relation hiérarchique entre l'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond.

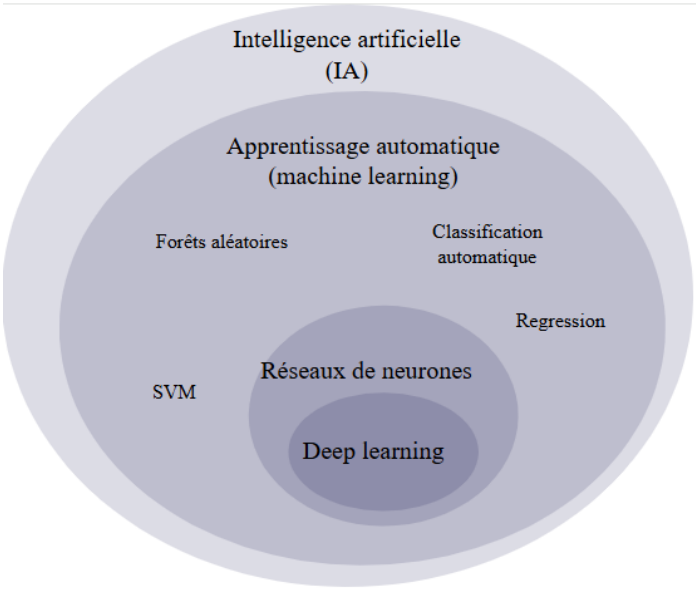


FIGURE 2.1 – Relation entre l'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond

2.3 Historique

Le Machine Learning, intimement lié à l'essor de l'intelligence artificielle, trouve ses racines dans les années 1950. En 1950, Alan Turing propose son célèbre test de Turing, ouvrant la voie à l'idée de machines intelligentes. Le terme Machine Learning apparaît pour la première fois en 1959 avec Arthur Samuel, qui développe un programme capable d'apprendre à jouer aux dames[3] [4].

En 1957, Frank Rosenblatt introduit le Perceptron, premier modèle de réseau de neu-

rones. Cependant, les limites de ces premiers modèles entraînent un déclin temporaire de l'intérêt pour l'IA, connu sous le nom d'hiver de l'IA [5].

Les années 1990 marquent un renouveau avec l'essor d'internet, offrant un accès massif aux données, indispensable au développement du Machine Learning. Depuis les années 2000, des algorithmes comme SVM, Random Forest et XGBoost se généralisent et sont adoptés dans divers domaines.

Parallèlement, le développement des réseaux de neurones artificiels permet des avancées majeures dans le traitement de données complexes, renforçant l'importance du Machine Learning dans des applications telles que la prévision des ventes, l'analyse des tendances et l'optimisation des ressources.

2.4 Types d'apprentissage automatique

L'apprentissage automatique repose sur plusieurs approches qui varient selon la manière dont les algorithmes exploitent les données. La classification des types d'apprentissage se base essentiellement sur le type d'informations disponibles dans les données et la façon dont l'algorithme apprend à résoudre une tâche. On distingue généralement trois catégories principales d'apprentissage automatique, chacune répondant à des objectifs et des contextes spécifiques.
La figure 2.2 illustre les trois types d'apprentissage automatique.

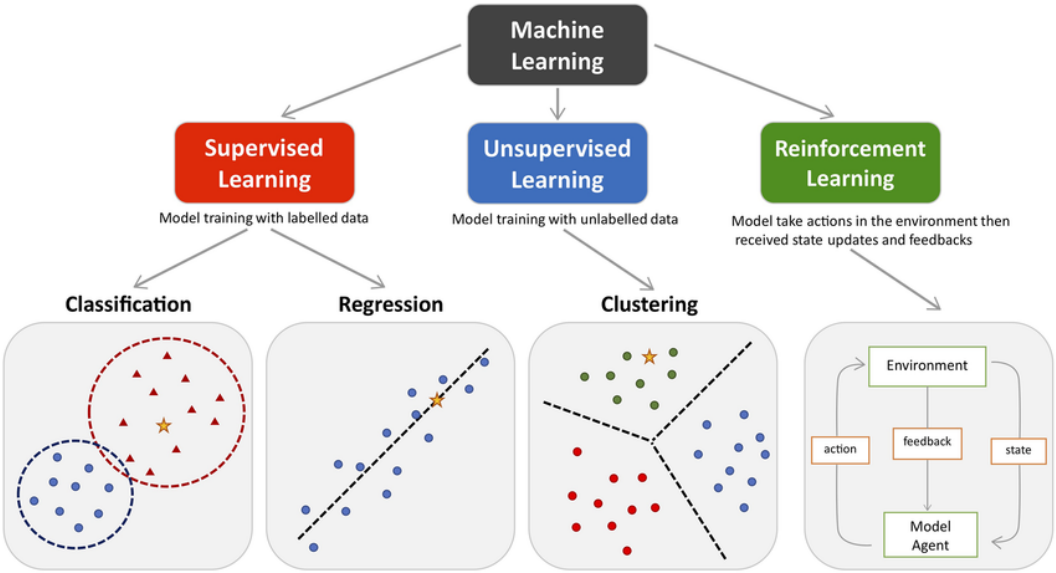


FIGURE 2.2 – Les types d'apprentissage automatique

2.4.1 Apprentissage supervisé

L'apprentissage supervisé est une branche du Machine Learning où un modèle est entraîné à partir de données étiquetées (ou labels en anglais). L'objectif est que l'algorithme

apprenne à faire des prédictions sur de nouvelles données en généralisant les relations observées dans l'ensemble d'entraînement.
Il existe deux types principaux d'apprentissage supervisé, voir la figure 2.3

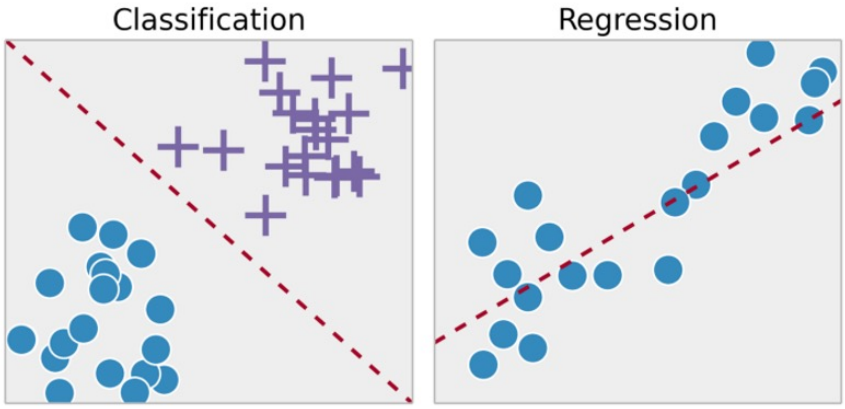


FIGURE 2.3 – Types d'apprentissage supervisé

2.4.1.1 La Régression

La régression consiste à prédire une valeur numérique continue en fonction des données d'entrée.
Il existe de nombreux types d'algorithmes de régression, chacun ayant ses propres forces et faiblesses. Parmi les plus courants, on trouve

a-Régression linéaire

L'algorithme de régression linéaire (RL) est un algorithme *paramétrique* qui fait des prédictions en ajustant un nombre fini de paramètres à partir des données d'entraînement.

Il suppose que la variable cible continue est une combinaison linéaire des variables prédictives [6], formulée par l'équation suivante :

$$y = \beta_0 + X\beta + \varepsilon \tag{2.1}$$

où :

- y est la variable cible,
- β_0 est le biais (ou intercept),
- X représente les variables prédictives,
- β est le vecteur des coefficients de régression,
- ε est l'erreur résiduelle.

Dans la figure 2.4 , on peut observer un exemple de régression linéaire

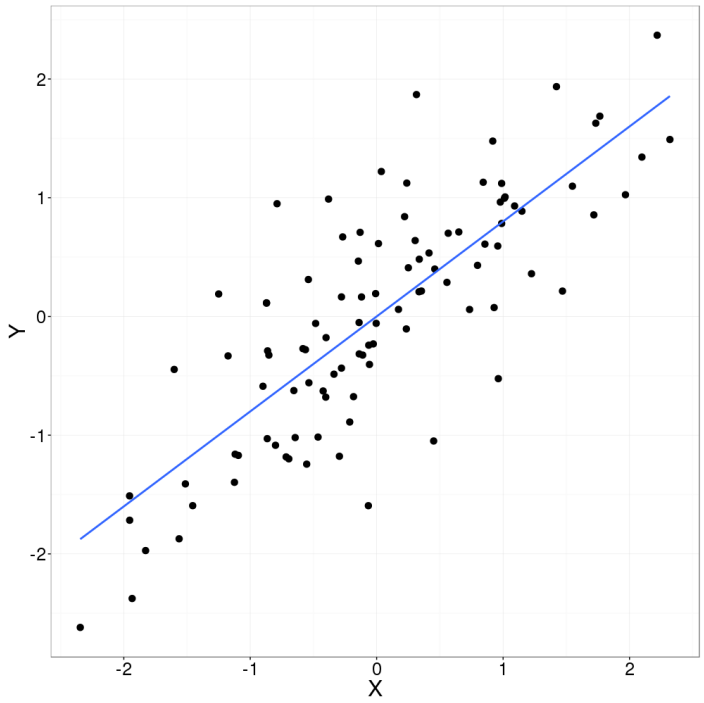


FIGURE 2.4 – Exemple de régression linéaire

b- Régression Logistique

La régression logistique (RG) est un algorithme paramétrique qui utilise une combinaison linéaire des variables prédictives pour estimer la probabilité qu'une variable cible binaire appartienne à l'une des deux classes (positive ou négative). Cette combinaison linéaire est transformée en une probabilité grâce à la fonction logistique, qui comprime la sortie en une valeur entre 0 (classe négative) et 1 (classe positive). Cette probabilité permet ensuite d'effectuer une classification binaire en utilisant un seuil, souvent fixé à 0.5, pour décider de la classe assignée [7].

Mathématiquement, cela s'exprime par la fonction logistique suivante :

$$p(y) = \frac{1}{1 + e^{-y}}, \quad ou\ y = \beta_0 + X\beta \tag{2.2}$$

avec :

- β_0 : le biais (ou intercept),
- X : les variables prédictives,
- β : les coefficients de régression.

Contrairement à la régression linéaire, les paramètres β_0 et β sont déterminés en maximisant la fonction de vraisemblance suivante :

$$L = \prod_{k:y_k=1} p_k \prod_{k:y_k=0} (1 - p_k) \tag{2.3}$$

Cette fonction représente la probabilité d'appartenance de chaque instance de données à la classe correspondante.

L'algorithme peut également être régularisé en utilisant des méthodes telles que *LASSO*, *Ridge* ou *Elastic-Net*, afin de pénaliser les coefficients β et ainsi éviter le sur-apprentissage, La figure 2.5 illustre un exemple de régression logistique.

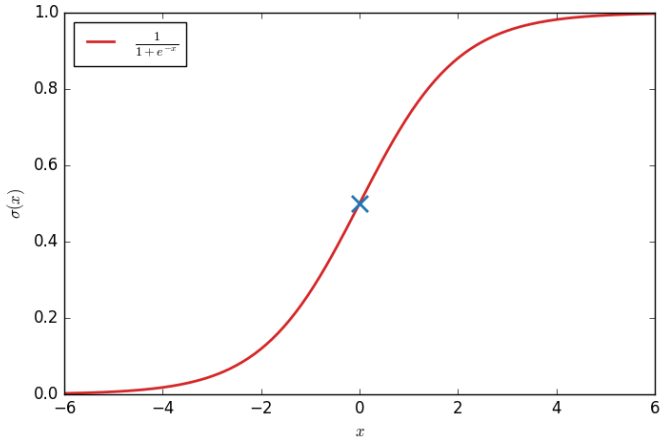


FIGURE 2.5 – Exemple sur la régression logistique.

c- Régression polynomiale

C'est une extension de la régression linéaire qui permet de modéliser une relation non linéaire entre la variable cible y et les variables prédictives x. [8]

L'équation générale prend la forme suivante :

$$y = a + b_1 x_1 + b_2 x_2^2 + b_3 x_3^3 + \cdots + b_k x_k^n + \varepsilon$$

Dans cette équation, x^n représente les termes polynomiaux, b_k sont les coefficients, et ε représente l'erreur.
Dans la figure 2.6, on peut observer un exemple de régression linéaire

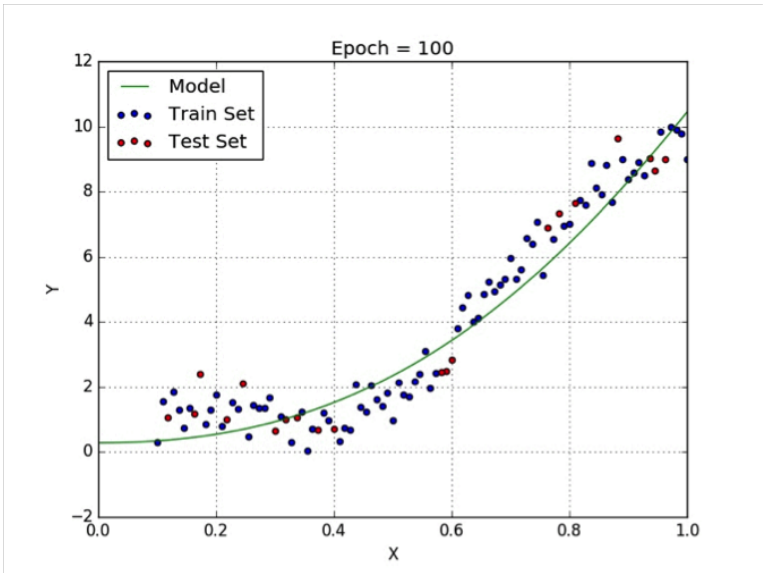


FIGURE 2.6 – Exemple de régression polynomiale.

d- Régression Ridge et Lasso

Pour éviter le surapprentissage, nous utilisons la régression Ridge et Lasso en présence d'un grand nombre de caractéristiques. Ce sont des techniques de régularisation utilisées dans le domaine de la régression. Elles fonctionnent en pénalisant l'amplitude des coefficients des caractéristiques tout en minimisant l'erreur entre les prédictions et les observations réelles. Les coefficients ont tendance à augmenter pour s'ajuster à un modèle complexe, ce qui peut conduire à un surapprentissage. Ainsi, lorsqu'ils sont pénalisés, cela limite leur croissance pour éviter de telles situations [9].

Régression Ridge / Régularisation L2

La régression Ridge ajoute un terme de pénalité (λw_i^2) à la fonction de coût, ce qui aide à éviter le surapprentissage. Par conséquent, notre fonction de coût est désormais exprimée comme suit :

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}^{(i)} - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{p} w_j^2 \tag{2.4}$$

Lorsque $\lambda = 0$, nous revenons à un surapprentissage, et lorsque $\lambda \to \infty$, trop de poids est ajouté, ce qui mène à un sous-apprentissage. Par conséquent, λ doit être choisi avec soin pour éviter ces deux scénarios.

Régression Lasso / Régularisation L1

Dans la régression Lasso, ou régularisation L1, c'est une valeur absolue (λw_i) qui est ajoutée au lieu d'un coefficient au carré. Cela représente un opérateur de réduction sélective moins sélectif.

La fonction de coût devient alors :

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})^2 + \lambda\sum_{j=1}^{p}|w_j| \tag{2.5}$$

2.4.1.2 La Classification

La classification est une tâche où l'algorithme apprend à associer une entrée à une catégorie parmi plusieurs possibles.

Il existe de nombreux algorithmes de classification :

a- Arbre de décision

Un arbre de décision est une technique utilisée pour approximer une fonction cible à valeurs discrètes en représentant la fonction apprise sous forme d'un arbre. Il classe les instances en les triant du nœud racine vers les nœuds feuilles en fonction de leurs valeurs de caractéristiques. Chaque nœud représente une décision ou une condition de test sur un attribut, et chaque branche représente une valeur possible pour cette caractéristique.

Le processus de classification commence au nœud racine et se poursuit vers le bas de l'arbre en fonction des résultats des tests effectués sur les caractéristiques. La décision finale est déterminée au niveau du nœud feuille, qui représente la catégorie de classification. Pour construire l'arbre de décision, des mesures statistiques telles que le gain d'information, l'indice de Gini, le Chi-carré et l'entropie sont calculées à chaque nœud afin d'évaluer son importance [10].

Exemple : On classifie les animaux en fonction de certaines caractéristiques (plumes, poils, écailles).

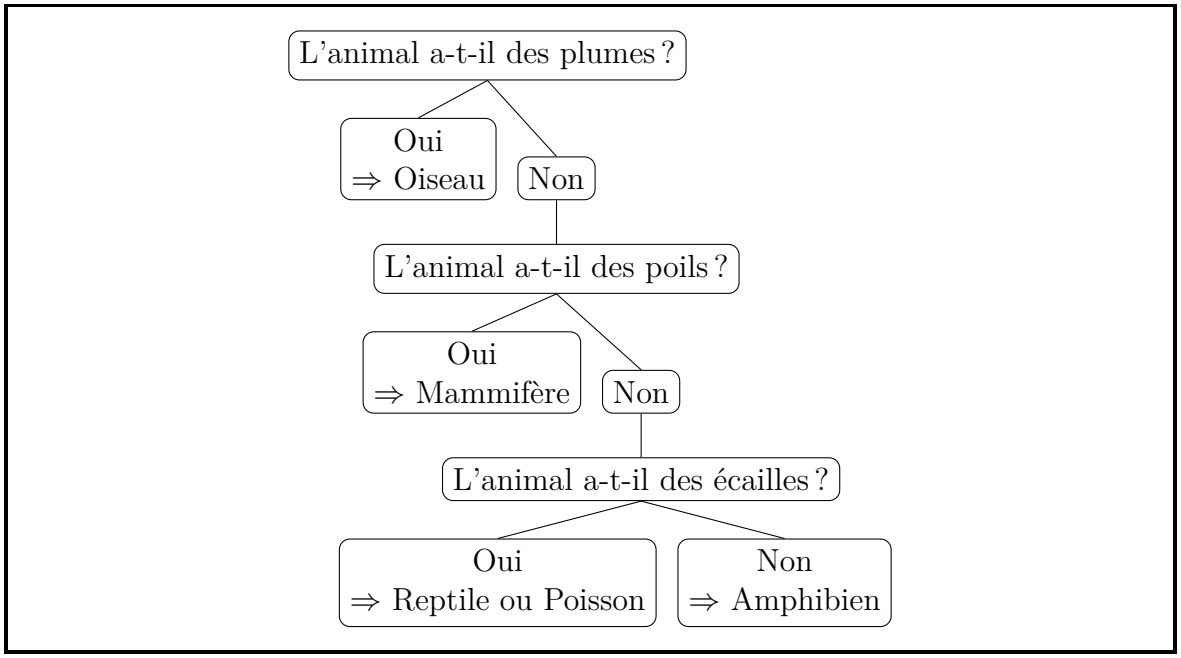


FIGURE 2.7 – Arbre de décision pour classifier les animaux

XGBoost (eXtreme Gradient Boosting)

est une bibliothèque open source d'apprentissage automatique qui implémente l'algorithme de gradient boosting basé sur des arbres de décision. Elle est conçue pour être rapide, efficace, et hautement performante, notamment sur de grands ensembles de données. XGBoost est largement utilisée pour les tâches de classification et de régression, grâce à sa robustesse et ses performances élevées [11].

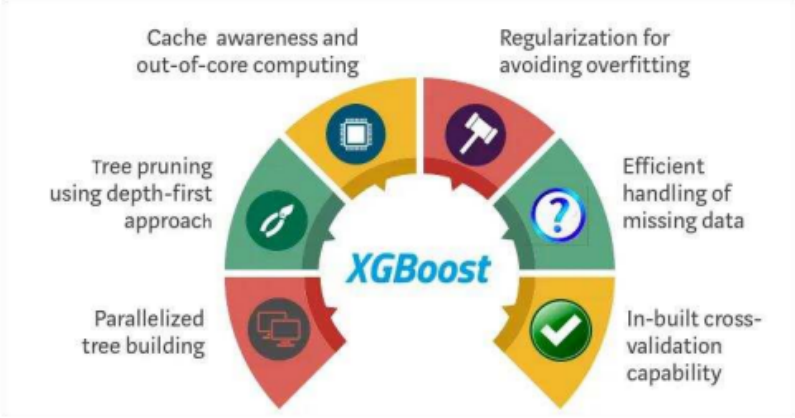


FIGURE 2.8 – Caractéristiques principales de XGBoost

Description de la figure 2.8 : La figure ci-dessus illustre les principales fonctionnalités qui font de *XGBoost* (Extreme Gradient Boosting) l'un des algorithmes les plus performants en apprentissage automatique. Parmi ses atouts majeurs, on retrouve :

- **Une gestion efficace de la mémoire**, grâce à une sensibilisation au cache et un traitement *out-of-core*, permettant de manipuler de très grands jeux de données.
- **Une construction parallèle des arbres**, accélérant considérablement le temps d'entraînement.
- **Un élagage optimal** (pruning) basé sur une approche en profondeur pour supprimer les branches peu pertinentes.
- **Une régularisation intégrée** (L1 et L2) permettant de limiter le surapprentissage.
- **Une gestion automatique des valeurs manquantes**, en déterminant dynamiquement le meilleur chemin à suivre.
- **Une validation croisée intégrée**, facilitant l'évaluation des performances pendant l'apprentissage.

Ces caractéristiques font de XGBoost un algorithme à la fois rapide, robuste et adapté à des contextes industriels exigeants.

b- Forêt Aléatoire (Random Forest)

La forêt aléatoire est une méthode d'apprentissage par ensemble couramment utilisée pour les tâches de classification . Elle fonctionne en créant plusieurs arbres de décision, chacun étant entraîné sur un sous-ensemble aléatoire des données à l'aide d'une technique appelée bagging. La prédiction finale est obtenue en combinant les résultats de tous les arbres de décision de la forêt aléatoire.

L'algorithme de la forêt aléatoire se compose de deux étapes :

Construction de la forêt aléatoire : Un grand ensemble d'arbres de classification est généré en entraînant chaque arbre sur des échantillons bootstrap des données.
Prédiction : Les résultats de l'ensemble des arbres sont agrégés, généralement en prenant la moyenne des probabilités estimées pour chaque classe.
Cette approche améliore la précision et la généralisation du modèle en réduisant le surajustement (overfitting) et en capturant une diversité de motifs dans les données [5].

En résumé, Forêt Aléatoire est un algorithme puissant qui exploite la force de plusieurs arbres de décision pour fournir des prédictions robustes,La figure suivante montre un exemple de forêt aléatoire.

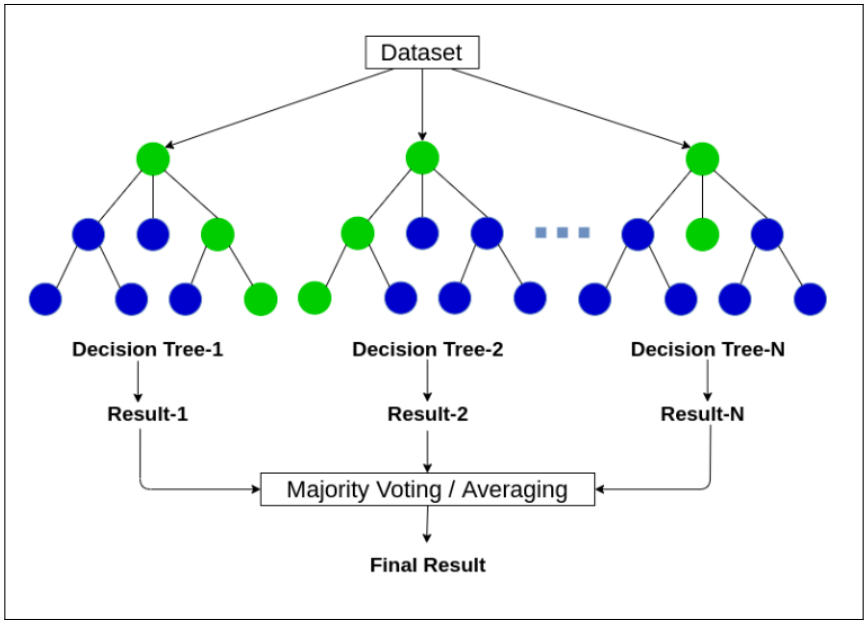


FIGURE 2.9 – Illustration de la Forêt Aléatoire.

c- Machines à vecteurs de support (SVM)

Les Machines à Vecteurs de Support (SVMs) sont des modèles d'apprentissage automatique puissants et polyvalents, largement utilisés pour diverses tâches telles que la classification et la détection d'anomalies. Ils excellent dans le traitement de jeux de données complexes, en particulier lorsque la taille des données est petite à moyenne.

Les SVMs sont des algorithmes d'apprentissage supervisé basés sur le risque, qui construisent des frontières de décision non linéaires, appelées hyperplans, afin de séparer les classes avec une marge maximale. Le problème d'optimisation des SVMs vise à minimiser les pénalités tout en maximisant la largeur de la marge.

Grâce à l'utilisation de l'astuce du noyau (kernel trick), les SVMs peuvent gérer efficacement les problèmes de classification linéaire et non linéaire en projetant les données dans un espace de caractéristiques de dimension supérieure, permettant ainsi de trouver des classes linéairement séparables[12].

Les SVMs ont démontré leur efficacité dans de nombreux domaines d'application, notamment la catégorisation de textes, la détection d'événements, l'analyse de l'expression des gènes et la modélisation des choix de modes de transport, mettant en évidence leur grande utilité en reconnaissance de formes et en analyse de données, voir la figure 2.10.

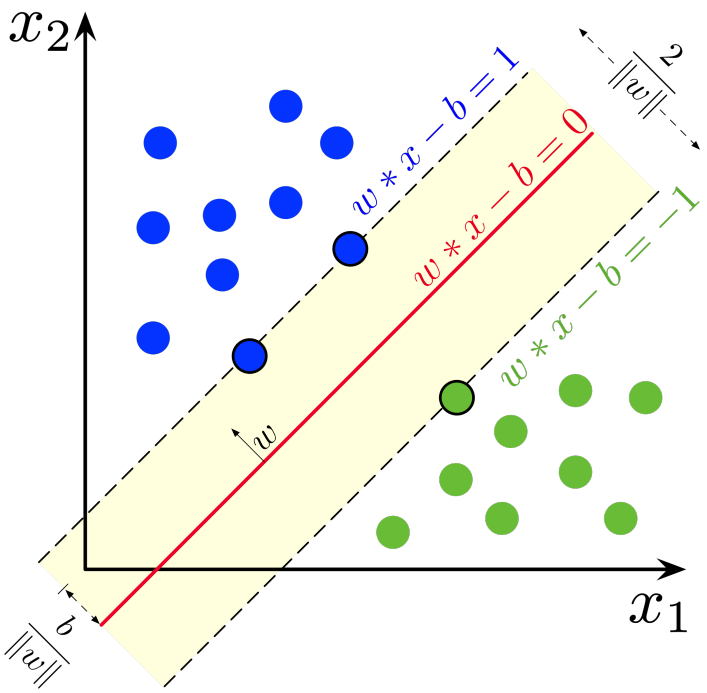


FIGURE 2.10 – Séparation linéaire par un SVM avec hyperplan et marges.

- L'hyperplan de décision (ligne rouge : $w \cdot x - b = 0$)
- Les deux marges (lignes parallèles : $w \cdot x - b = \pm 1$)
- Les vecteurs supports (les points touchant les marges)

d- k-Plus Proches Voisins (KNN)

L'algorithme des k plus proches voisins (KNN) est un classificateur d'apprentissage non paramétrique et supervisé qui s'appuie sur la notion de proximité pour réaliser des classifications ou des prédictions sur le regroupement d'un point de données[13], la figure suivante montre le principe de l'algorithme KNN.

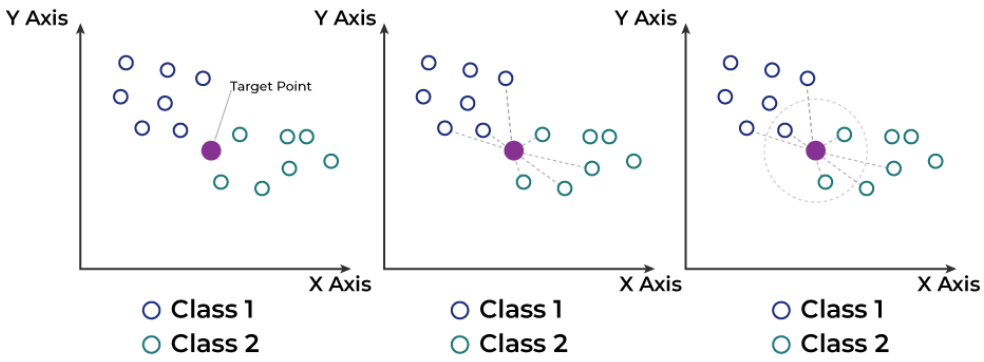


FIGURE 2.11 – Principe de classification avec l'algorithme des K plus proches voisins (KNN).

Mesure d'évaluation pour les modèles de classification

Une fois qu'un modèle a été déterminé et implémenté, il est essentiel d'évaluer sa qualité. Pour cela, différentes mesures d'évaluation peuvent être utilisées, et le choix de la mesure influence directement l'interprétation des performances du modéle [14].

A.La matrice de confusion

L'un des outils les plus couramment utilisés pour évaluer la performance d'un modèle de classification est la **matrice de confusion**. Il s'agit d'un tableau qui résume le nombre de prédictions correctes et incorrectes faites par le modèle.

Dans cette matrice, chaque ligne représente une classe réelle, tandis que chaque colonne représente une classe prédite (estimée) par le modèle.

Elle comprend les éléments suivants :

- **Vrais positifs (True Positive, TP)** : la classe réelle est positive et la prédiction est également positive.
- **Vrais négatifs (True Negative, TN)** : la classe réelle est négative et la prédiction est également négative.
- **Faux positifs (False Positive, FP)** : la classe réelle est négative, mais le modèle a prédit une classe positive (erreur de Type I).
- **Faux négatifs (False Negative, FN)** : la classe réelle est positive, mais la prédiction est négative (erreur de Type II).

Dans le cas d'une **classification binaire**, la matrice de confusion prend la forme d'une matrice 2×2 ,voir la figure 2.12.

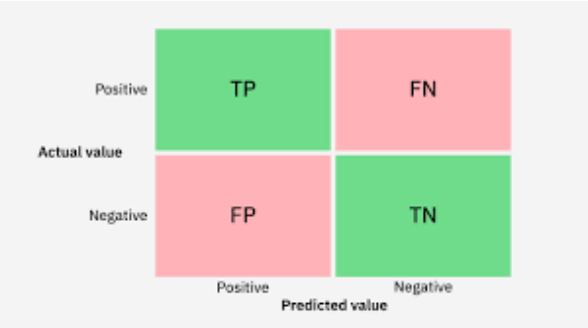


FIGURE 2.12 – Matrice de confusion pour un problème de classification binaire

B.Mesures de classification

Une fois que la matrice de confusion a été établie, elle peut être utilisée pour des mesures plus approfondies afin d'obtenir une meilleure évaluation de la qualité du modèle. Parmi

les mesures de classification, on trouve : l'**accuracy**, la **précision**, le **rappel**, la **spécificité** et le **score F1**.

Accuracy

L'accuracy correspond au nombre de prédictions correctes faites par le modèle. Elle représente le ratio entre le nombre de prédictions correctes et le nombre total de prédictions. Cette mesure est utilisée lorsque les *True Positive (TP)* et *True Negative (TN)* sont les plus importants. La formule est donnée par :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Précision (Precision)

La précision mesure la proportion de prédictions positives qui sont correctes. Elle est utile lorsque le nombre de *False Positives (FP)* est important. Sa formule est :

$$Précision = \frac{TP}{TP + FP}$$

Rappel (Recall)

Le rappel, ou sensibilité, détermine la proportion de vraies valeurs positives correctement identifiées. Il est utilisé lorsque le nombre de *False Negatives (FN)* est élevé. La formule est :

$$Rappel = \frac{TP}{TP + FN}$$

Spécificité (Specificity)

La spécificité mesure la proportion de vraies valeurs négatives correctement identifiées. Elle se calcule comme suit :

$$Spécificité = \frac{TN}{TN + FP}$$

Score F1

Le score F1 est la moyenne harmonique entre la précision et le rappel. Il est utilisé lorsque l'on souhaite trouver un équilibre entre ces deux mesures, notamment lorsque les *FP* et *FN* sont importants. Il est défini par :

$$F1score = 2 \times \frac{Précision \times Rappel}{Précision + Rappel}$$

2.4.2 Apprentissage non-supervisé

L'apprentissage non supervisé est un type d'apprentissage automatique où les données d'entrée ne possèdent pas de labels ou de catégories prédéfinis. L'objectif de l'apprentissage non supervisé est d'explorer et de découvrir des motifs, des structures ou des relations dans les données sans guidage ou supervision explicite [15], La figure suivante schématise le principe d'un algorithme d'apprentissage non supervisé.

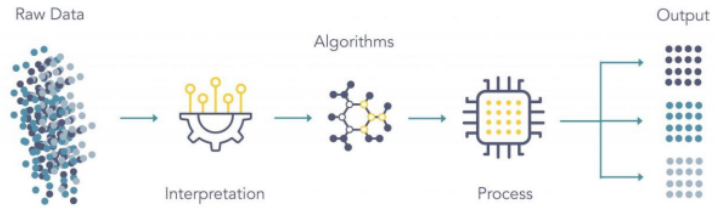


FIGURE 2.13 – Algorithme d'apprentissage non supervisé

Quelques algorithmes d'apprentissage non supervisé :

1.Algorithmes de Clustering

Le clustering, également appelé analyse de regroupement (Cluster Analysis), est une méthode utilisée pour regrouper des objets ou des points de données en clusters distincts en fonction de leurs similitudes. Il s'agit d'une technique d'apprentissage non supervisé, car elle ne nécessite pas de labels ou de classes prédéfinis. L'objectif du clustering est de partitionner l'ensemble de données en sous-groupes ou clusters, où les points de données au sein de chaque cluster partagent des caractéristiques communes ou présentent une similarité selon une mesure de distance définie.

Il existe différents types de méthodes de clustering, notamment le clustering hiérarchique et le clustering partitionnel. Un exemple d'algorithme de clustering partitionnel est K-means, qui utilise une approche basée sur les centroïdes pour affecter les points de données aux clusters.

L'analyse de regroupement (Cluster Analysis - CA) vise à diviser les données en groupes présentant des caractéristiques similaires. L'objectif principal de cette technique est de maximiser l'hétérogénéité entre les clusters tout en maximisant la similarité au sein de chaque cluster. Elle permet ainsi de révéler la structure sous-jacente des données sans nécessiter de connaissances préalables ou de prédictions.

De manière générale, les techniques de clustering sont des outils précieux en apprentissage non supervisé, car elles permettent d'identifier des regroupements naturels ou des motifs dans les données, sans information ou supervision préalable. Elles fournissent ainsi des informations essentielles sur la structure et les relations au sein d'un ensemble de données [15] [16],voir la figure 2.14.

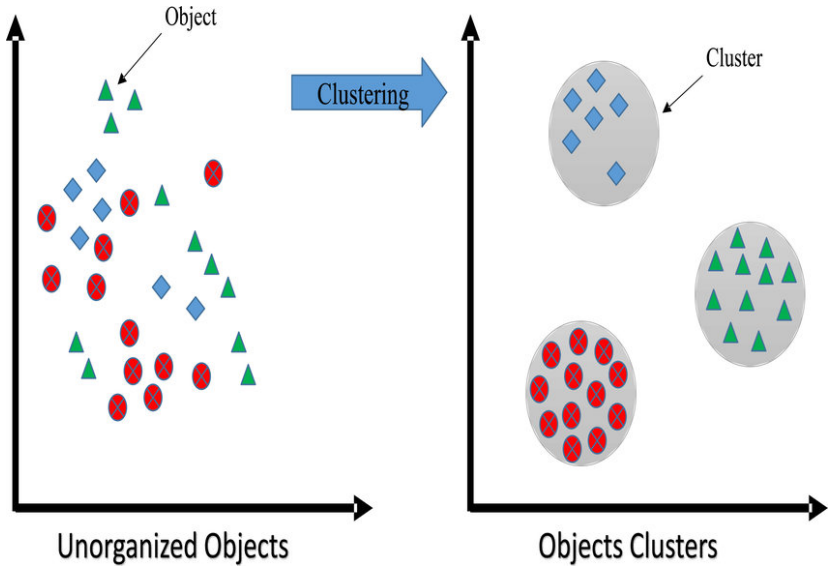


FIGURE 2.14 – Illustration du clustering en apprentissage non supervisé.

2.Algorithmes de Réduction de Dimensionnalité

La réduction de dimensionnalité est le processus qui consiste à minimiser la perte d'information tout en réduisant le nombre de caractéristiques (ou dimensions) d'un ensemble de données. Cette approche permet de diminuer la complexité d'un modèle, d'améliorer les performances d'un système d'apprentissage et de faciliter la visualisation des données. Plusieurs techniques existent pour effectuer cette réduction[17].

2.1 Analyse en Composantes Principales (ACP - PCA)

Les étapes principales de l'ACP sont les suivantes [17][18] :

- Standardisation : Normaliser les variables afin qu'elles contribuent équitablement, en les centrant à une moyenne de 0 et une variance de 1.
- Calcul de la Matrice de Covariance : Construire la matrice de covariance des données.
- Décomposition en Valeurs Propres : Calculer les vecteurs propres de cette matrice.

— Sélection des Composantes Principales : Choisir les composantes principales associées aux plus grandes valeurs propres, car elles capturent le plus de variance.
— Projection : Projeter l'ensemble de données initial sur ces composantes principales afin d'obtenir une meilleure représentation des données .

La figure suivante illustre le principe de l'Analyse en Composantes Principales (ACP).

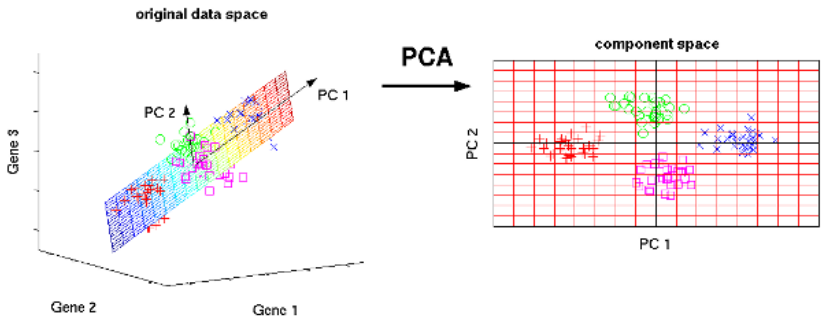


FIGURE 2.15 – Illustration de l'ACP (Analyse en Composantes Principales)

2.4.3 Apprentissage par renforcement

L'apprentissage par renforcement est une méthode d'apprentissage automatique où un agent apprend à prendre des décisions en interagissant avec un environnement.

À chaque action, l'agent reçoit des récompenses ou des pénalités, l'amenant à affiner ses stratégies pour maximiser ses gains,Ce principe est illustré dans la figure ci-dessous.[19]

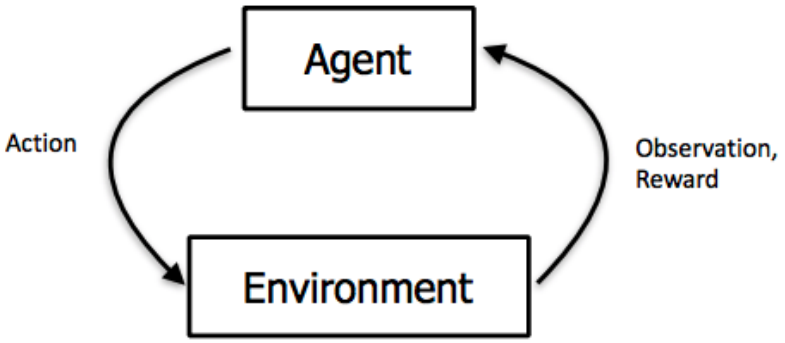


FIGURE 2.16 – Schéma du processus d'apprentissage par renforcement

Les algorithmes les plus utilisés en apprentissage par renforcement sont Q-Learning, DQN, PPO et DDPG. Le Q-Learning est simple et adapté aux petits environnements. Le DQN utilise des réseaux de neurones pour gérer des espaces d'états plus grands. Le PPO est aujourd'hui le plus populaire pour sa stabilité, et le DDPG est efficace pour les actions continues.

2.5 Domaines d'application du Machine Learning

Le *Machine Learning* connaît aujourd'hui un large éventail d'applications dans de nombreux secteurs. Grâce à sa capacité à traiter de grandes quantités de données et à découvrir des modèles complexes, il permet de résoudre des problèmes concrets et d'automatiser des tâches autrefois réservées à l'intelligence humaine.

Voici quelques domaines d'application majeurs :

- **Santé** : détection de maladies, aide au diagnostic, prédiction de l'évolution de maladies, analyse d'images médicales.
- **Finance** : détection de fraudes, scoring de crédit, analyse prédictive des marchés financiers.
- **Marketing et e-commerce** : recommandation de produits, segmentation de clients, prédiction du comportement d'achat.
- **Industrie** : maintenance prédictive, automatisation de la production, contrôle qualité.
- **Transport et logistique** : prévision de la demande, optimisation des itinéraires, gestion des stocks.
- **Reconnaissance vocale et traitement du langage naturel** : assistants vocaux, traduction automatique, analyse de sentiments.

Le Machine Learning continue d'évoluer et d'ouvrir de nouvelles perspectives dans des domaines toujours plus variés.

Partie 2 : Prévision des Ventes avec le Machine Learning

2.6 La prévision des ventes, qu'est -ce que c'est ?

la prévision des ventes est une méthode qui consiste à estimer les ventes à venir en fonction des données passées et des études comparatives correspondant à un secteur d'activité particulier. La prévision des ventes est un processus qui permet aux entreprises de générer les données nécessaires à la préparation des plans et des stratégies futures.

2.7 Objectifs de la prévision des ventes

L'objectif principal de cet outil est l'aide à la décision stratégique qui vise à améliorer la croissance de l'entreprise en anticipant l'état de la demande future, de fixer avec plus de précision des objectifs réalistes et surtout atteignables. En effet, elle permet de prévoir le chiffre d'affaires qui sera réalisé par l'entreprise sur une période donnée dans le but d'orienter les

ressources à d'autres fins utiles (quantité de marchandise, gestion des stocks, recrutement, distribution…). Un autre objectif de la prévision des ventes est son impact positif sur l'amélioration de la gestion des stocks de l'entreprise. Cet outil permet notamment de planifier ses activités sur du court, moyen ou long terme et bien sûr d'affiner sa stratégie globale afin de rester compétitif sur le marché.

2.8 Enjeux et impacts pour l'entreprise

La prévision des ventes représente un enjeu majeur pour les entreprises. Elle influence plusieurs aspects de la gestion et de la stratégie.

- Optimisation des stocks.
- Réduction des coûts.
- Meilleure planification des ressources.
- Amélioration du service client.
- Aide à la prise de décisions stratégiques.

2.9 Apports du Machine Learning

Le Machine Learning apporte une nouvelle approche dans la prévision des ventes en s'appuyant sur l'apprentissage automatique à partir des données historiques. Il permet de mieux capturer les tendances complexes et de s'adapter aux évolutions du marché.

2.9.1 Pourquoi utiliser le Machine Learning ?

Le Machine Learning permet d'automatiser les prévisions, de traiter de grandes quantités de données, et de détecter des relations non linéaires que les méthodes classiques ne peuvent pas facilement modéliser.

2.10 Avantages du Machine Learning dans la prévision des ventes

L'utilisation du Machine Learning dans la prévision des ventes présente plusieurs avantages concrets pour les entreprises :

- Amélioration de la précision des prévisions.
- Prise en compte de nombreux facteurs internes et externes.

— Détection automatique des tendances et des saisonnalités.
— Capacité à s'adapter rapidement aux changements du marché.
— Réduction du temps d'analyse et d'intervention manuelle.

2.11 Défis liés à la prévision des ventes avec le Machine Learning

Malgré ses avantages, l'utilisation du Machine Learning dans la prévision des ventes présente plusieurs défis qu'il convient de prendre en compte.

2.11.1 Données incomplètes ou bruitées

Les données de ventes peuvent contenir des valeurs manquantes, des erreurs ou du bruit, ce qui impacte la qualité des modèles.

2.11.2 Variabilité du marché et événements externes

Les changements brusques du marché ou les événements imprévus peuvent réduire la fiabilité des modèles prédictifs.

2.11.3 Choix du bon modèle et sur-apprentissage

Il est souvent difficile de choisir le modèle adapté et d'éviter le sur-apprentissage, notamment avec des données limitées ou peu représentatives.

2.12 Étapes de mise en œuvre d'un modèle ML de prévision

La mise en place d'un modèle de Machine Learning pour la prévision des ventes suit un processus structuré composé de plusieurs étapes.

2.12.1 Collecte des données de ventes

Les données utilisées dans cette étude proviennent de l'historique des ventes mensuelles de sauces produites par Cevital, couvrant les années 2023 et 2024. Ces données ont été

extraites du système interne de gestion des ventes de l'entreprise. Chaque enregistrement contient le nom du produit, le mois de vente, ainsi que la quantité vendue (en tonnes).

2.12.2 Prétraitement et nettoyage

Le prétraitement a inclus la suppression des valeurs nulles et négatives, la détection des valeurs aberrantes par la méthode IQR, ainsi qu'un nettoyage général des incohérences. Ces opérations visent à améliorer la qualité des données et la performance du modèle de prévision.

2.12.3 Analyse exploratoire

L'analyse exploratoire des données permet d'avoir une première compréhension des comportements et des relations présentes dans le jeu de données. Elle s'appuie sur des outils graphiques et statistiques pour identifier les tendances, les variables significatives et les corrélations. Cette étape inclut également des tests comme la stationnarité ou l'autocorrélation, essentiels pour orienter le choix des modèles prédictifs.

2.12.4 Sélection et entraînement du modèle ML

Une fois les données préparées, il est important de sélectionner l'algorithme de machine learning le mieux adapté aux spécificités du problème. Pour notre projet, nous avons choisi d'utiliser XGBoost, un algorithme de boosting reconnu pour sa capacité à fournir des prédictions précises et robustes en fonction des données disponibles.

2.12.5 Évaluation des performances

L'évaluation des performances du modèle se fait à l'aide de différents indicateurs d'erreur, tels que l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE) ou le coefficient de détermination (R^2), en fonction du type de problème (régression ou classification). Ces métriques permettent de mesurer la précision des prédictions du modèle et d'évaluer sa capacité à généraliser sur de nouvelles données.

2.12.6 Déploiement et intégration dans le processus décisionnel

Une fois le modèle entraîné et validé, il est essentiel de le déployer dans l'environnement de production. Cela permet d'utiliser les prévisions générées dans leurs outils quotidiens. L'intégration du modèle dans les systèmes d'information de l'entreprise facilite l'automatisation des prévisions et leur prise en compte dans les décisions stratégiques, tactiques et opérationnelles.

2.13 Problématique

Dans un contexte économique marqué par une forte incertitude et une concurrence accrue, les entreprises doivent anticiper la demande avec précision pour optimiser leurs opérations. Les méthodes classiques de prévision, souvent basées sur des modèles statistiques linéaires, montrent leurs limites face à la complexité croissante des données et à la volatilité du marché.

Comment les techniques de Machine Learning peuvent-elles permettre à une entreprise agroalimentaire comme Cevital d'améliorer la précision et la fiabilité des prévisions de ventes, et ainsi renforcer la planification, la gestion des stocks et la prise de décision stratégique ?

2.14 Conclusion

Dans ce chapitre, nous avons d'abord présenté les fondements du Machine Learning, ses catégories, ainsi que ses principaux algorithmes, afin de mieux comprendre son fonctionnement et ses domaines d'application. Ensuite, nous avons abordé l'application du Machine Learning à la prévision des ventes, en mettant en évidence ses apports, ses avantages, ainsi que les étapes nécessaires à la mise en œuvre d'un modèle performant.

Chapitre 3

Implémentation d'un Modèle de Prévision des Ventes par Machine Learning chez Cevital

3.1 Introduction

Ce chapitre décrit la mise en œuvre d'un modèle de prévision des ventes pour l'année 2025 chez Cevital, en s'appuyant sur les données historiques de ventes de sauces des années 2023 et 2024. Le modèle, développé avec l'algorithme XGBoost, vise à anticiper la demande future afin d'optimiser la production, la gestion des stocks et la prise de décision.

Par la suite, les résultats obtenus ont été intégrés dans un tableau de bord interactif conçu avec Power BI. Cet outil permet une visualisation claire des tendances de vente mensuelles, facilite l'analyse des performances commerciales et constitue un véritable support pour le pilotage stratégique de l'entreprise.

Partie 1 : Prévision des ventes pour l'année 2025

Dans cette section, nous décrivons le processus de développement d'un modèle de prévision des ventes pour l'année 2025, au sein de l'entreprise Cevital.

3.2 Outils Utilisés pour le Développement

3.2.1 Langage de programmation : Python

Python est un langage de programmation interprété, polyvalent et orienté objet. Il est reconnu pour sa syntaxe claire et sa facilité de prise en main, ce qui en fait un outil idéal pour le développement rapide d'applications.

Dans le cadre de ce projet, Python a été utilisé pour manipuler les données, entraîner le modèle de prévision des ventes, et visualiser les résultats. Grâce à ses nombreuses bibliothèques spécialisées, il constitue un choix efficace pour les tâches de *Machine Learning* et d'analyse de données.

3.2.2 Environnement de développement : Jupyter Notebook

Jupyter Notebook est une application web open-source qui permet de créer des documents contenant à la fois du code exécutable, des visualisations et du texte. Il offre un environnement interactif permettant d'écrire, d'exécuter et de documenter du code de manière dynamique, facilitant ainsi les expérimentations et la présentation des résultats.

3.2.3 Bibliothèques utilisées

Le développement du modèle s'est appuyé sur plusieurs bibliothèques Python spécialisées dans la manipulation de données, la modélisation et la visualisation :

- **Pandas** : utilisée pour la manipulation et le nettoyage des données tabulaires (DataFrame).
- **NumPy** : indispensable pour le calcul scientifique. Elle permet de manipuler efficacement des tableaux numériques multidimensionnels et propose un grand nombre de fonctions mathématiques optimisées.
- **Matplotlib** et **Seaborn** : utilisées pour la visualisation des données. Ces bibliothèques permettent de créer des graphiques clairs et esthétiques pour explorer les tendances, les corrélations et la distribution des ventes.
- **Scikit-learn** : utilisée pour le prétraitement des données, le découpage en ensembles d'entraînement et de test, ainsi que pour le calcul des métriques d'évaluation des performances du modèle.
  La figure suivante montre les bibliothèques utilisées.

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

#  Scikit-Learn : Outils pour l'apprentissage automatique, le prétraitement des données et l'évaluation des modèles
from sklearn.model_selection import train_test_split, GridSearchCV  # Séparation des données et recherche des meilleurs hyperparamètres
from sklearn.metrics import mean_absolute_error, mean_squared_error  # Évaluation des performances des modèles
from sklearn.preprocessing import MinMaxScaler  # Mise à l'échelle des données pour un meilleur entraînement des modèles

# XGBoost : Algorithme d'optimisation des arbres de décision, utilisé pour des prévisions puissantes et rapides
import xgboost as xgb
```

FIGURE 3.1 – Bibliothèques utilisées pour la mise en œuvre du modèle de prévision

3.3 Collecte des données

Le jeu de données utilisé pour la prévision des ventes concerne les différentes variétés de sauces commercialisées par l'entreprise **Cevital**. Il s'agit d'un historique mensuel des ventes de 40 produits répartis sur les années 2023 et 2024. Les données sont structurées sous forme matricielle dans un fichier *Excel*, avec les produits en lignes et les mois en colonnes.

Ci-dessous, un aperçu des premières lignes du jeu de données utilisé.

	Produit	Jan 2023	Feb 2023	Mar 2023	Apr 2023	May 2023	Jun 2023	Jul 2023	Aug 2023	Sep 2023	...	Apr 2024	May 2024	Jun 2024	Jul 2024	Au
0	HARISSA (VERRE) 150g.	8.452800	10.143425	11.319110	11.445546	10.058220	8.026365	8.038573	7.255267	6.562483	...	11.009247	12.293208	11.841820	15.301507	15.
1	KETCHUP (PET) 220g.	7.463104	7.267203	9.356990	9.406408	7.577180	4.827304	5.329028	9.175278	9.621579	...	10.113875	10.954691	10.800698	10.855075	10.
2	KETCHUP (PET) 435g.	8.427168	8.179719	10.254780	9.278383	9.735843	6.606797	5.140746	9.264550	10.045452	...	11.422021	12.273431	12.075008	11.798369	11.
3	MAYONNAISE 1L « MAYONA »	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	13.812480	13.
4	MAYONNAISE AIL ET FINES HERBES (PET) 395g.	9.586176	10.188602	12.717888	11.238730	11.269382	10.131376	13.541053	15.327340	13.156116	...	13.950800	14.269941	11.720669	13.407327	14.
5	MAYONNAISE FULL FAT (PET) 200g.	19.418880	20.077488	24.007296	18.815014	17.438899	13.004544	18.201830	19.746144	20.194304	...	24.454272	18.735667	12.771405	15.365338	18.
6	MAYONNAISE FULL FAT (PET) 395g.	16.517952	16.461442	24.480849	22.739183	22.433472	14.636393	17.924835	20.269871	20.160092	...	32.708694	21.467536	15.820589	18.501939	20.
7	MAYONNAISE FULL FAT (VERRE) 220g. / 12 BOCAUX	21.473936	23.382740	28.145156	27.836638	22.978560	16.264503	16.655523	20.819188	20.389680	...	30.194050	20.181934	16.472592	16.793290	17.

FIGURE 3.2 – Aperçu des premières lignes du jeu de données

3.4 Prétraitement des données

Le prétraitement des données constitue une étape essentielle pour assurer la qualité et la fiabilité des résultats obtenus par le modèle de prévision. Dans cette étude, plusieurs opérations ont été menées pour nettoyer et préparer le jeu de données avant l'entraînement.

3.4.1 Suppression des valeurs nulles et négatives

Dans le cadre du nettoyage des données, les lignes contenant des valeurs nulles (égales à zéro) ou négatives ont été supprimées. Ces valeurs sont généralement considérées comme non exploitables ou correspondant à des retours de consignation.la figure ci-dessous illustre la réduction du nombre total de lignes après cette opération de filtrage.

FIGURE 3.3 – Nombre de lignes restantes après la suppression des valeurs négatives et nulles

3.4.2 Détection et traitement des valeurs aberrantes par la méthode IQR

Une fois les données nettoyées des valeurs nulles et négatives, la détection des valeurs aberrantes a été réalisée à l'aide de la méthode IQR (*Interquartile Range*). Cette méthode repose sur la dispersion des données entre le premier quartile (Q1) et le troisième quartile (Q3), soit :

$$IQR = Q3 - Q1$$

Les observations considérées comme extrêmes sont celles qui se situent en dehors de l'intervalle :

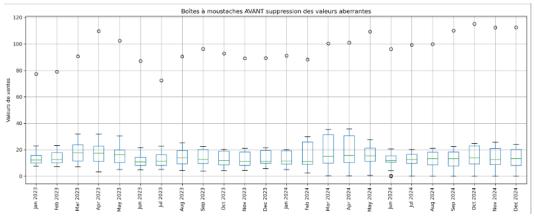
$$[Q1 - 1.5 \times IQR,\ Q3 + 1.5 \times IQR]$$

Ces valeurs, souvent dues à des anomalies ou à des comportements inhabituels, peuvent fausser l'entraînement du modèle. Selon le cas, elles peuvent être supprimées, voir la figure suivante.

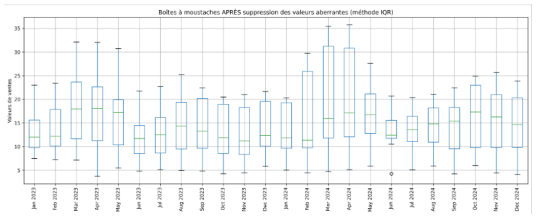
```
Valeurs supprimées :
     Colonne  Valeur_supprimée
0   Jan 2023         77.364960
1   Feb 2023         78.983955
2   Mar 2023         90.516119
3   Apr 2023        109.609282
4   May 2023        102.407266
5   Jun 2023         87.020450
6   Jul 2023         72.383538
7   Aug 2023         90.427039
8   Sep 2023         96.276794
9   Oct 2023         92.811229
10  Nov 2023         89.088398
11  Dec 2023         89.308648
12  Jan 2024         91.160777
13  Feb 2024         88.216009
14  Mar 2024        100.312757
15  Apr 2024        100.943790
16  May 2024        109.275293
17  Jun 2024         96.087629
18  Jun 2024          0.359821
19  Jun 2024          0.287950
20  Jun 2024          0.054219
21  Jul 2024         99.240256
22  Aug 2024         99.831500
23  Sep 2024        110.026312
24  Oct 2024        115.157059
25  Nov 2024        112.404061
26  Dec 2024        112.425138

Nombre de lignes avant nettoyage : 18
Nombre de lignes après nettoyage  : 14
```

FIGURE 3.4 – Détection des valeurs aberrantes par la méthode IQR pour chaque variable numérique

La figure 3.5 compare la distribution des ventes mensuelles avant et après suppression des valeurs aberrantes selon la méthode IQR. On observe une réduction significative des valeurs extrêmes, rendant les données plus représentatives et mieux adaptées à l'analyse statistique et à la modélisation prédictive.



(a) Avant suppression des valeurs aberrantes



(b) Après suppression des valeurs aberrantes

FIGURE 3.5 – Comparaison des boîtes à moustaches avant et après suppression des outliers (méthode IQR)

Cette étape est cruciale pour améliorer la robustesse du modèle et garantir que les prévisions soient basées sur des données fiables.

3.5 Analyse Exploratoire des Données (AED)

3.5.1 Statistiques Descriptives

Nous avons réalisé une analyse descriptive des données en utilisant la fonction describe() de Pandas, qui résume les statistiques descriptives mensuelles sur la période allant de janvier 2023 à décembre 2024. Chaque mois contient 14 observations. Nous analysons ici la moyenne, l'écart-type, les quartiles, ainsi que les valeurs extrêmes,voir les resultats dans la figure suivante.

	count	mean	std	min	25%	50%	75%	max
Jan 2023	14.0	13.264708	4.967512	7.463104	9.799200	12.020640	15.614034	22.988880
Feb 2023	14.0	13.965626	5.511036	7.267203	10.154719	12.191092	17.858406	23.382740
Mar 2023	14.0	18.273540	7.562410	7.147012	11.668805	17.989005	23.656075	32.135453
Apr 2023	14.0	17.551771	7.902774	3.715920	11.290434	18.107573	22.650380	32.044240
May 2023	14.0	16.216768	6.903810	5.453021	10.361010	17.293363	19.859438	30.700980
Jun 2023	14.0	11.704035	4.614423	4.827304	8.552618	11.717937	14.443490	21.806793
Jul 2023	14.0	12.652354	5.214416	5.140746	8.639757	12.517729	16.196643	22.685050
Aug 2023	14.0	14.519020	5.880003	4.963443	9.502451	14.333017	19.408268	25.218562
Sep 2023	14.0	13.791606	6.106403	4.849402	9.630426	13.276091	20.185751	22.375142
Oct 2023	14.0	12.897604	5.778524	4.274756	8.559881	11.918660	18.957192	20.488075
Nov 2023	14.0	12.500662	5.490185	4.468101	8.332947	11.199916	18.303667	21.021370
Dec 2023	14.0	13.903306	5.322372	5.801318	10.109679	12.382873	19.535601	21.641252
Jan 2024	14.0	13.705146	5.214972	5.009523	9.661918	11.896575	19.298066	20.255582
Feb 2024	14.0	16.727768	9.019727	4.454574	9.788496	11.361883	25.880101	29.738918
Mar 2024	14.0	20.410921	10.730160	4.741080	11.800292	15.998681	31.250814	35.454904
Apr 2024	14.0	20.591205	10.390424	5.113730	12.054216	17.157628	30.791345	35.750475
May 2024	14.0	17.112162	5.929002	5.871993	12.787391	16.765559	21.146135	27.625018
Jun 2024	14.0	13.391867	4.108673	4.295725	11.740313	12.423207	15.575569	20.712737
Jul 2024	14.0	13.733227	4.141916	5.097726	11.090898	13.560409	16.436302	20.369353
Aug 2024	14.0	14.216313	4.881419	5.892551	10.966250	14.821510	18.150168	21.072840
Sep 2024	14.0	13.954399	5.920604	4.270182	9.525501	15.385790	18.282523	22.429270
Oct 2024	14.0	16.165344	6.760518	6.014557	9.822584	17.323445	22.984707	24.892857
Nov 2024	14.0	15.484846	6.752482	4.430471	9.816010	16.282022	20.982810	25.702734
Dec 2024	14.0	14.791519	6.136074	4.094808	9.874094	14.636168	20.292701	23.906447

FIGURE 3.6 – Résumé statistique des variables

Selon les résultats :

- **Tendance centrale :** La moyenne mensuelle des quantités varie entre environ 11,70 (juin 2023) et 20,59 (avril 2024), ce qui reflète une tendance globale à la hausse en 2024. La médiane suit une évolution similaire.
- **Dispersion :** L'écart-type est plus élevé au cours des mois de février à avril 2024 (autour de 10), ce qui indique une plus grande variabilité dans les données observées durant cette période. À l'inverse, les mois comme juin et juillet 2024 présentent une dispersion plus faible (écart-type 4).

- **Valeurs extrêmes :** Les valeurs maximales dépassent les 35 tonnes en mars et avril 2024, ce qui suggère des pics ponctuels importants. En revanche, certaines valeurs minimales restent proches de 4, notamment en avril, septembre et décembre 2024.
- **Distribution :** Dans plusieurs mois, la moyenne est significativement supérieure à la médiane (ex. mars 2024 : moyenne = 20,41, médiane = 15,99), indiquant une distribution asymétrique avec des valeurs extrêmes à droite. Cela montre que certaines observations exceptionnellement élevées influencent la moyenne.
- **Comparaison annuelle :** En comparant les mêmes mois sur deux années consécutives, on note une progression significative. Par exemple :
  - *Mars 2023* : moyenne = 18,27 vs *Mars 2024* : moyenne = 20,41
  - *Avril 2023* : moyenne = 17,55 vs *Avril 2024* : moyenne = 20,59
  - *Mai 2023* : moyenne = 16,21 vs *Mai 2024* : moyenne = 17,11

  Cette évolution suggère une amélioration des performances ou une augmentation de la demande.

3.5.2 Répartition des ventes totales par produit (2023–2024)

Le diagramme en camembert ci-dessous présente la répartition des ventes totales par produit pour les années 2023 et 2024. On observe que certains produits dominent largement les ventes, ce qui reflète une forte demande pour ces références. À l'inverse, d'autres produits représentent une part plus faible, suggérant une consommation plus modérée ou une moindre disponibilité. Cette visualisation permet ainsi d'identifier les produits stratégiques et d'orienter les décisions en matière de gestion des stocks et de planification.

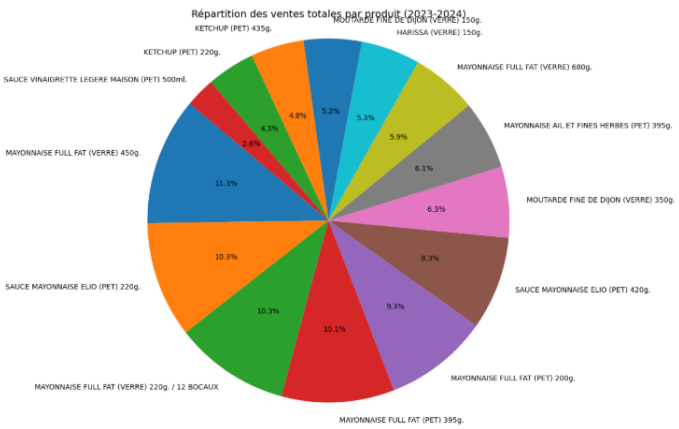


FIGURE 3.7 – Répartition des ventes totales par produit (2023–2024)

3.5.3 Les produits les plus vendus

La figure 3.8 illustre la répartition des ventes, en tonnes, des dix sauces les plus vendues au cours de la période 2023–2024.

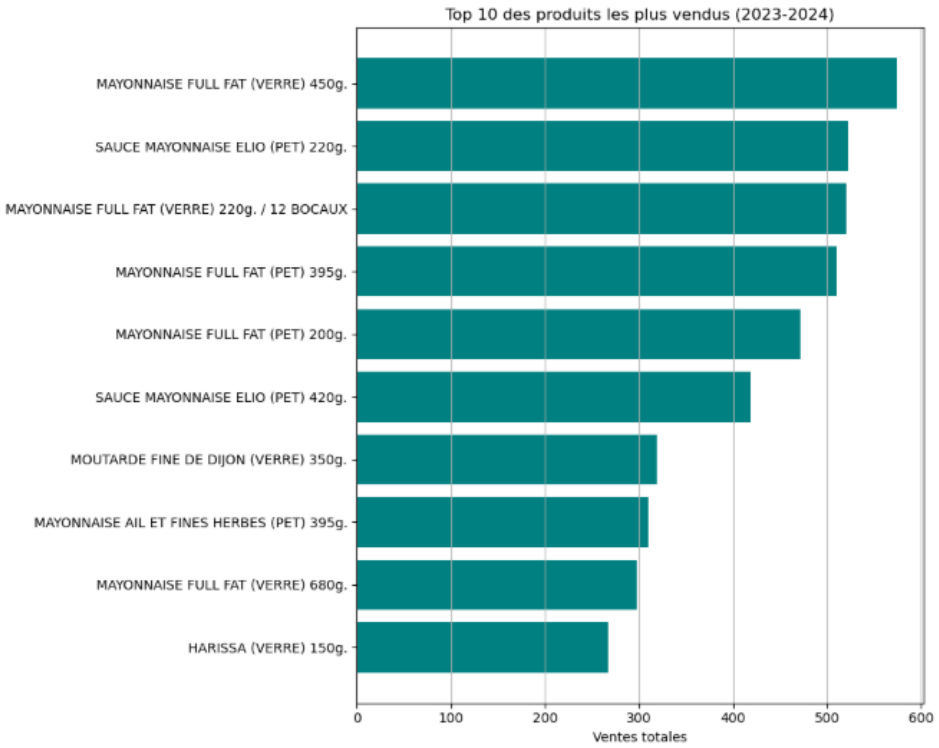


FIGURE 3.8 – Les 10 produits les plus vendus (2023–2024)

3.6 Modélisation

L'objectif principal de cette modélisation est de prédire les quantités mensuelles vendues (exprimées en tonnes) pour chaque sauce produite par Cevital, à partir de l'historique des ventes sur les années 2023 et 2024. Ces prévisions permettront d'optimiser la gestion des stocks, la planification de la production et les approvisionnements.

3.6.1 Préparation et séparation des données

Les données ont été prétraitées afin de corriger les éventuelles valeurs manquantes et assurer une bonne cohérence temporelle. L'ensemble des données a ensuite été scindé en deux sous-ensembles dans le but de garantir une évaluation rigoureuse :

- 80% des données ont été utilisées pour l'apprentissage du modèle (ensemble d'entraînement) ;
- 20% restantes ont servi à l'évaluation des performances (ensemble de test).

3.6.2 Choix du modèle

Étant donné que notre problème relève de la régression, nous avons choisi d'utiliser *XGBoost*, un modèle particulièrement performant. Grâce à sa capacité à gérer efficacement le surajustement, cet algorithme s'avère être une solution robuste pour notre tâche de prédiction.

3.6.3 Optimisation des hyperparamètres

Pour optimiser les performances du modèle *XGBoost*, nous avons utilisé la méthode *GridSearchCV* avec validation croisée. Cette méthode nous a permis de tester plusieurs combinaisons d'hyperparamètres afin d'obtenir les meilleurs paramètres pour notre modèle. Les meilleurs paramètres sélectionnés sont les suivants :

```
Meilleurs paramètres : {'colsample_bytree': 0.6, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 300, 'subsample': 0.6}
```

FIGURE 3.9 – Visualisation des meilleurs paramètres .

3.6.4 Évaluation des performances globales

Pour évaluer la performance globale du modèle, nous avons calculé plusieurs métriques de référence sur l'ensemble des ventes . Ces métriques permettent de mesurer la précision du modèle de manière objective. Les résultats suivants(figure3.10) montrent l'excellente capacité de notre modèle à prédire les quantités de ventes :

```
R² global   : 0.9999
RMSE global : 0.0002
MSE global  : 0.0064
```

FIGURE 3.10 – Métriques globales sur toutes les ventes.

3.6.5 Prévision pour l'année 2025

Avant de réaliser la prévision pour l'année 2025, il est essentiel de confirmer la performance de notre modèle. Nous avons utilisé une division des données en 80 % pour l'apprentissage et 20 % pour le test, ce qui nous permet de valider la capacité du modèle à généraliser sur de nouvelles données. Cette approche garantit que le modèle n'est pas surajusté et peut fournir des prévisions fiables sur des données non vues, comme celles de l'année 2025.

La figure 3.11 présente une comparaison entre les ventes réelles et les ventes prédites sur les 20 % de données de test, sous forme d'histogramme. Cette visualisation permet de juger visuellement de la qualité des prédictions.

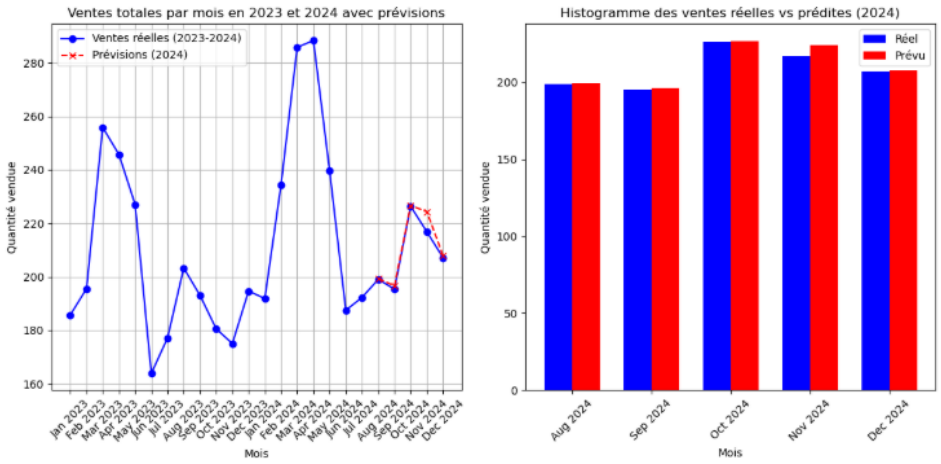


FIGURE 3.11 – Comparaison entre les ventes réelles et prédites sur l'ensemble de test (20 %).

Les résultats obtenus sur l'ensemble de test montrent que le modèle parvient à prédire les ventes avec une grande précision, ce qui renforce la confiance dans les prévisions à venir.

3.6.5.1 Prédictions pour l'année 2025

À partir de ces résultats validés, nous avons généré des prévisions pour l'année 2025. Ces prévisions sont basées sur les données historiques des années 2023 et 2024, et permettent d'anticiper les quantités de chaque sauce vendue sur l'année à venir.

Les résultats sont illustrés dans la figure suivante :

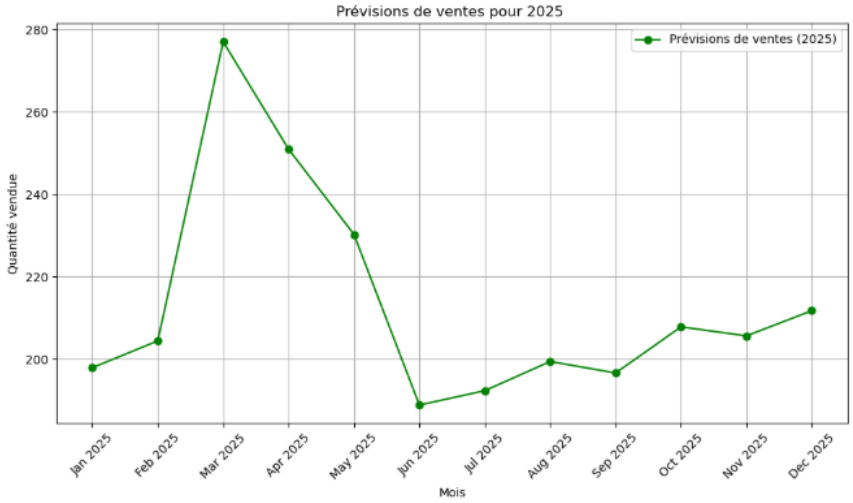


FIGURE 3.12 – Prévisions mensuelles des ventes pour l'année 2025.

Un pic de la demande est observé au mois de mars 2025, ce qui correspond à la pé-

riode du Ramadan. Cette hausse est cohérente avec les habitudes de consommation pendant ce mois, qui entraîne traditionnellement une augmentation des ventes de sauces.

3.6.5.2 Répartition des Ventes Prévues par Produit 2025

Le graphique en camembert (figure 3.13) présente la part estimée de chaque produit dans le total des ventes prévues en 2025. Ce visuel permet d'identifier les produits les plus contributeurs au chiffre d'affaires projeté.

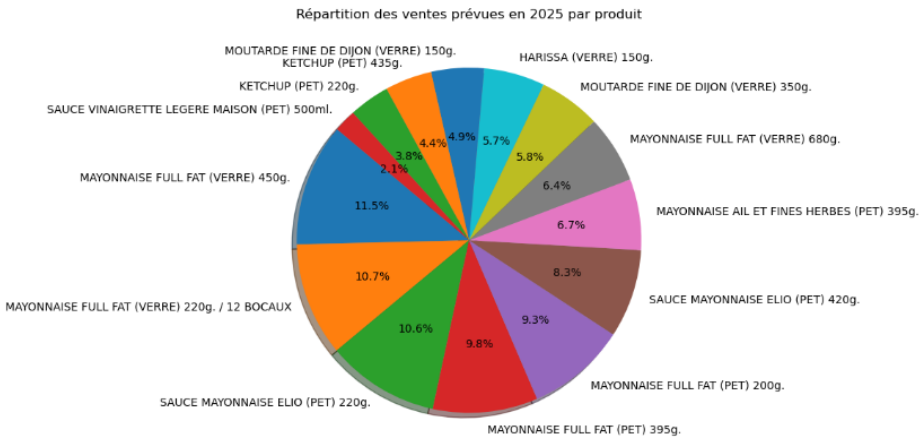


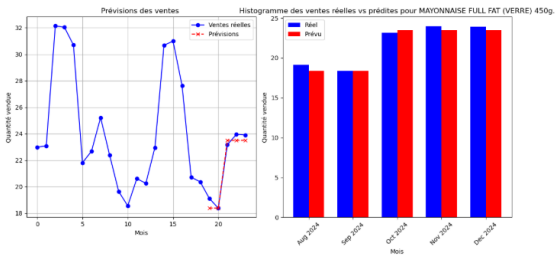
FIGURE 3.13 – Répartition des ventes prévues par produit pour l'année 2025

3.6.6 Prévision pour les Top 5 sauces les plus vendus

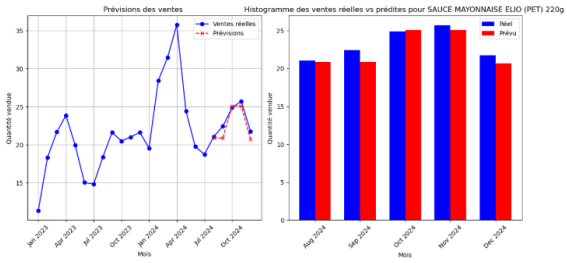
Les prévisions globales sont utiles pour avoir une vue d'ensemble, mais il est également pertinent d'analyser les produits les plus performants individuellement. Nous avons donc sélectionné les cinq produits les plus vendus en 2023 et 2024, afin de détailler leur comportement de vente prévu pour l'année 2025.

Pour valider la performance de notre modèle XGBoost. Les données ont été divisées en deux sous-ensembles : 80 % pour l'apprentissage et 20 % pour le test. Cela permet de mesurer la capacité du modèle à généraliser.

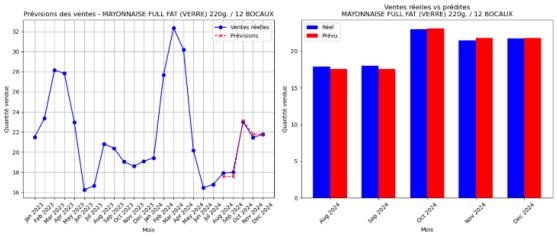
La figure suivante illustre la comparaison entre les valeurs réelles et les valeurs prédites sur l'ensemble de test pour ces cinq produits.



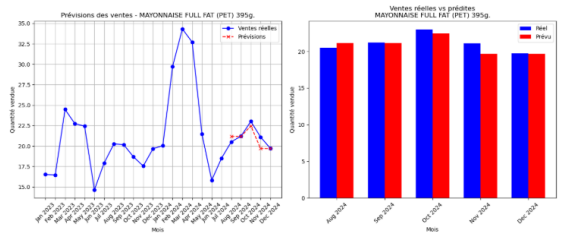
MAYO FF VER 450g



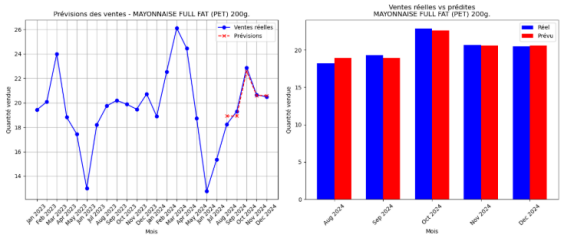
MAYO ELIO PET 220g



MAYO FF VER 220g x12



MAYO FF PET 395g

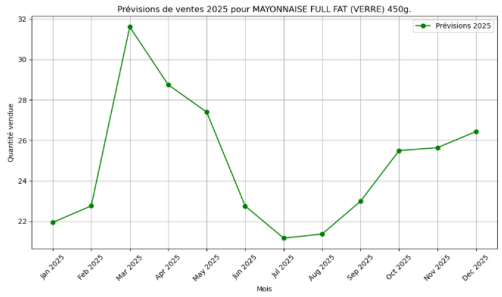


MAYO FF PET 200g

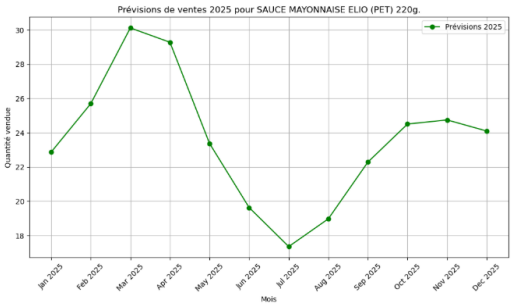
FIGURE 3.14 – Comparaison des valeurs réelles et prédites sur l'ensemble de test pour les cinq produits les plus vendus.

Suite à la validation des performances du modèle, nous avons procédé à la génération des prévisions pour l'année 2025, en nous focalisant sur les cinq produits les plus vendus. Ces prédictions mensuelles constituent un outil stratégique pour anticiper la demande, améliorer la planification de la production, ajuster les niveaux de stock, et optimiser les approvisionnements tout au long de l'année.

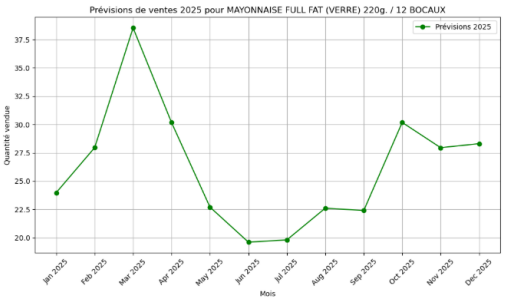
La figure 3.15 illustre les quantités prédites pour chaque mois de l'année 2025 pour ces cinq produits.



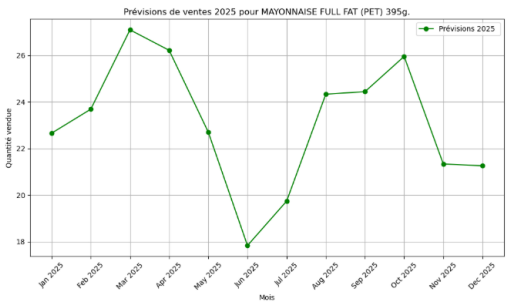
Prévision 2025 -MAYO FF VER 450g



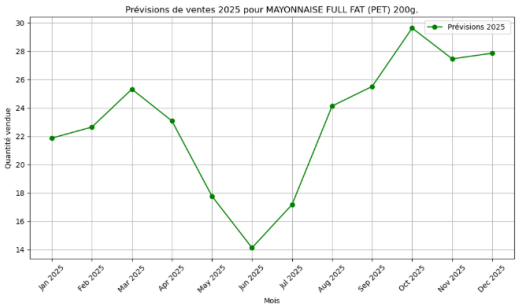
Prévision 2025 -MAYO ELIO PET 220g



Prévision 2025 -MAYO FF VER 220g x12



Prévision 2025 -MAYO FF PET 395g



Prévision 2025 -MAYO FF PET 200g

FIGURE 3.15 – Prévisions pour l'année 2025 des cinq produits les plus vendus.

Les courbes présentées dans la figure ci-dessus illustrent les prévisions mensuelles des ventes pour l'année 2025, produit par produit. On observe des tendances spécifiques à chaque produit, avec des variations saisonnières marquées. Par exemple, un pic de la demande est clairement visible pour certains produits au mois de mars, ce qui peut être attribué à la période du Ramadan, où la consommation de sauces augmente traditionnellement.

Ces résultats confirment la capacité du modèle à capter les comportements de vente saisonniers et à fournir des estimations cohérentes et utiles pour la prise de décision opérationnelle.

Partie 2 : Analyse Visuelle des Résultats de Vente via un Tableau de Bord BI

Dans cette seconde partie, nous présentons la mise en place d'un tableau de bord interactif destiné à l'analyse mensuelle des ventes de sauces pour l'année 2025.

3.7 Outils utilisés

L'outil Power BI a été utilisé pour sa capacité à se connecter facilement à des sources de données variées (CSV, Excel, bases de données), à proposer une large palette de visualisations interactives et à permettre la création de mesures avancées en DAX. Son intégration au sein de l'écosystème Microsoft (Excel, Power Query, Power Automate) facilite également le partage, la publication et l'actualisation automatisée des rapports, garantissant ainsi une veille commerciale en temps réel.

3.8 Description des éléments du tableau de bord

Dans cette section, nous détaillons chacun des principaux visuels présents dans le dashboard.

3.8.1 Graphique de comparaison des ventes

La figure suivante illustre une comparaison des ventes totales réalisées en 2023 et 2024. Elle permet d'observer l'évolution du volume des ventes avant les prévisions pour l'année 2025.

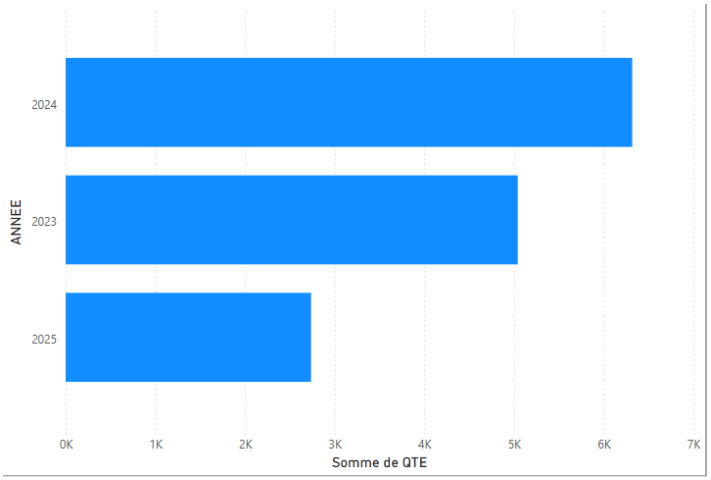


FIGURE 3.16 – Répartition des volumes de ventes

On observe une progression notable entre 2023 et 2024, avec une augmentation du volume total de ventes. En revanche, les prévisions pour 2025 affichent un volume inférieur, ce qui peut refléter soit une estimation prudente, soit une variation attendue de la demande. Cette visualisation permet ainsi d'avoir un aperçu clair de l'évolution des ventes d'une année à l'autre, et d'orienter les décisions stratégiques en conséquence.

3.8.2 Carte de score : quantité totale vendue

FIGURE 3.17 – Somme totale des ventes prévues pour 2025

La figure 3.17 présente la somme totale des ventes prévues pour l'année 2025, estimée à **14 097,53** tonnes. Cet indicateur synthétique permet d'avoir une vision globale du volume annuel attendu, facilitant la planification stratégique.

3.8.3 Premier produit "MAYONNAISE FULL FAT (VERRE) 450g"

FIGURE 3.18 – Produit le plus vendu en 2025

La figure 3.18 met en avant le produit ayant réalisé les meilleures performances de vente durant l'année 2025 : **MAYONNAISE FULL FAT (VERRE) 450g**. Cette information est précieuse pour orienter les efforts commerciaux et logistiques vers les références les plus rentables.

3.8.4 histogramme des meilleures ventes

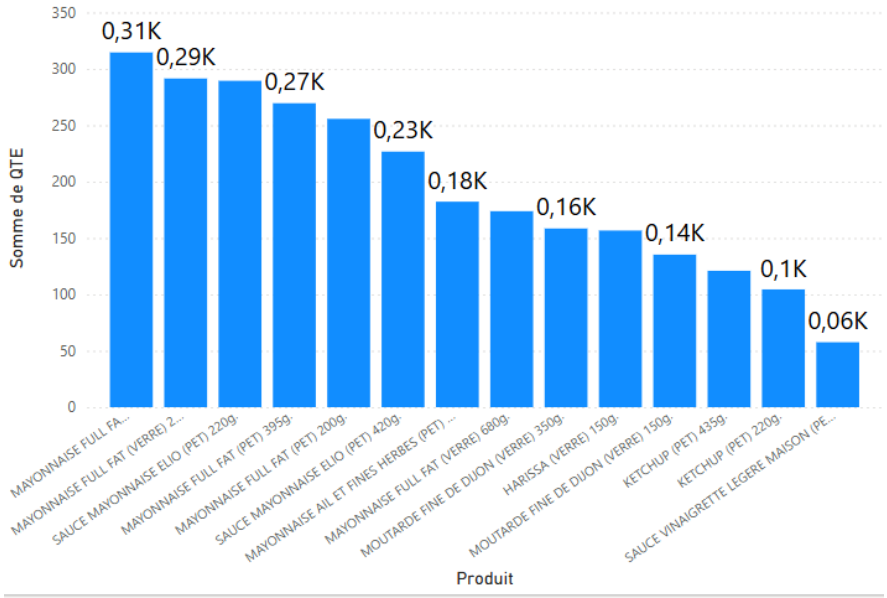


FIGURE 3.19 – Top produits les plus vendus en 2025

Ce graphique met en évidence les produits ayant enregistré les volumes de vente les plus élevés en 2025. En tête du classement, on retrouve la MAYONNAISE FULL FAT (VERRE) 450g, avec un total avoisinant 0.31K tonnes vendues, suivie de près par plusieurs autres références de mayonnaise en différents formats. Cette concentration des meilleures performances autour d'un même type de produit (la mayonnaise) révèle une tendance forte de consommation, qui peut être exploitée pour renforcer la stratégie de production, de distribution et de promotion. De plus, ces résultats permettent de prioriser les produits à fort rendement et d'optimiser les ressources sur les références les plus rentables.

3.8.5 Courbe mensuelle : évolution des ventes en 2025

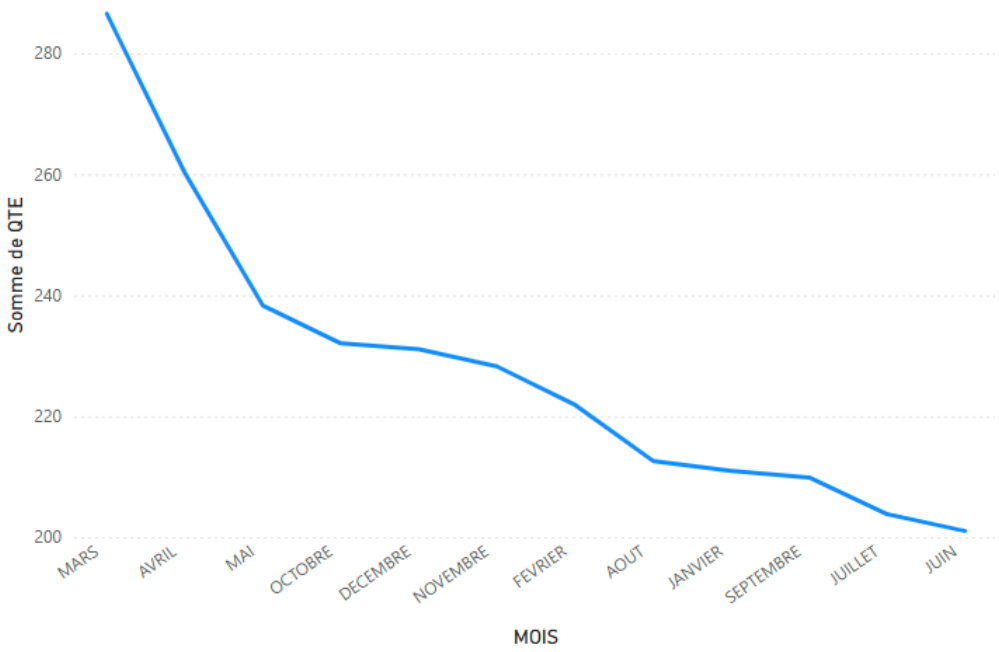


FIGURE 3.20 – Évolution mensuelle des ventes en 2025

La figure 3.20 présente l'évolution des ventes mensuelles au cours de l'année 2025. L'axe horizontal représente les mois de l'année, tandis que l'axe vertical indique la somme des quantités vendues (QTE).

On constate une tendance décroissante des ventes depuis le mois de mars. Cette évolution peut résulter d'une baisse saisonnière de la demande ou d'un déséquilibre entre l'offre et les besoins réels du marché.

Il est important de noter que les mois affichés ne suivent pas l'ordre chronologique traditionnel, mais sont classés par ordre décroissant des quantités vendues. Cela permet de visualiser en priorité les mois ayant généré les volumes les plus élevés, mais nécessite une vigilance lors de l'interprétation temporelle.

Ce type de représentation offre un aperçu précieux pour ajuster les politiques de vente et anticiper les fluctuations de la demande.

3.9 Présentation du tableau de bord complet

Afin de faciliter la lecture et l'analyse des prévisions de ventes pour l'année 2025, un tableau de bord interactif a été conçu à l'aide de Power BI. Ce tableau de bord regroupe plusieurs indicateurs visuels permettant de synthétiser les informations clés issues des résultats de modélisation. Il permet une vision globale des performances attendues, tout en

offrant un niveau de détail suffisant pour orienter les décisions stratégiques.

La figure suivante présente une vue d'ensemble du tableau de bord :

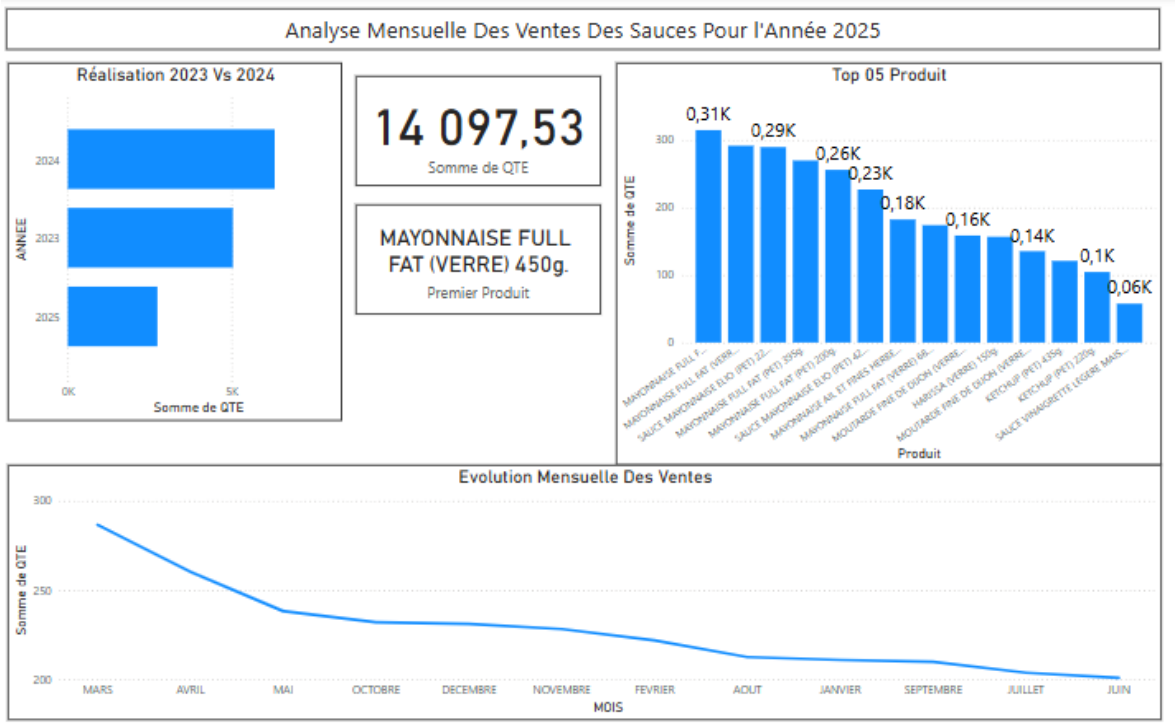


FIGURE 3.21 – Vue d'ensemble du tableau de bord de suivi des ventes (2023–2024)

3.10 Conclusion

Au cours de ce chapitre, nous avons d'abord développé un modèle XGBoost sous Python pour réaliser la prévision des ventes de sauces pour l'année 2025 à partir des historiques de 2023 et 2024, garantissant des estimations à la fois robustes et précises. Nous avons ensuite conçu un tableau de bord interactif sous Power BI dédié à l'analyse des données réelles de 2023 et 2024 ; ce dashboard synthétise à la fois la répartition annuelle, les performances mensuelles et le classement des produits. Cet outil combine exploration visuelle et indicateurs clés, facilitant l'analyse des tendances passées et la prise de décision stratégique pour la planification future.

Conclusion Générale

Ce mémoire a porté sur la mise en œuvre d'un modèle de machine learning pour la prévision des ventes au sein de l'entreprise Cevital,Après une présentation détaillée de l'entreprise et de son environnement industriel, nous avons introduit les fondements théoriques du machine learning, en mettant en lumière les principes qui régissent les modèles prédictifs. Nous avons ensuite appliqué ces connaissances à un cas pratique, en développant un modèle de prévision basé sur les données de ventes de sauces de Cevital, couvrant les années 2023 et 2024.

Dans l'implémentation du modèle, les données exploitées ont constitué la base du travail. Il s'agit ici des historiques de ventes de sauces de l'entreprise Cevital, couvrant les années 2023 et 2024. Après un prétraitement minutieux visant à assurer la qualité et la cohérence des informations, ces données ont été utilisées pour entraîner un modèle de machine learning en vue de prédire la demande pour l'année 2025. Les résultats ont ensuite été intégrés dans un tableau de bord interactif conçu avec Power BI, offrant une visualisation claire des tendances et un support décisionnel précieux pour l'entreprise.

En conclusion, ce mémoire a mis en évidence l'apport concret des techniques de machine learning dans l'amélioration de la précision des prévisions de ventes. Une telle approche permet à Cevital de mieux anticiper la demande, d'optimiser la gestion des stocks, de planifier les opérations de manière plus efficace, et ainsi de renforcer leur performance globale.

Bibliographie

[1] Cevital. (2025). Site officiel de Cevital. https ://www.cevital.com

[2] Géron, A. (2017). Machine learning avec Scikit-Learn. O'Reilly Media.

[3] Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, s2–42(1), 230–265. https ://doi.org/10.1112/plms/s2-42.1.230

[4] Arthur L.S.History Computer. https ://history.computer.org/pioneers/samuel.html

[5] Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms : An overview. Journal of Physics : Conference Series, 1142(1). https ://doi.org/10.1088/1742-6596/1142/1/012012

[6] McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (2e éd.). Chapman and Hall/CRC.

[7] Berkson, J. (1944). Application of the logistic function to bio-essay. Journal of the American Statistical Association, 39, 357–365.

[8] Fan, J., Heckman, NE, et Wand, MP (1995). Régression polynomiale locale à noyau pour modèles linéaires généralisés et fonctions de quasi-vraisemblance. Journal of the American Statistical Association , 90 (429), 141–150. https ://doi.org/10.1080/01621459.1995.10476496

[9] Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. Contemporary Mathematics, 443(7), 59-72.

[10] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 3–18). IEEE.

[11] Chen, T., & Guestrin, C. (2016). XGBoost : A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https ://doi.org/10.1145/2939672.2939785

[12] Sucre, A. (2022, octobre 10). Support vector machines : A brief introduction. The Beginner's Guide. https ://medium.com/the-beginnersguide/support-vector-machines-a-brief-introduction-784bbf97cdce

[13] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., et al. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1–37.

[14] Allmang, A. (2025, avril 3). Apprentissage supervisé et classification. Linedata. https ://fr.linedata.com/apprentissage-supervise-et-classification

[15] Dhage, S. N., & Raina, C. K. (2016). A review on machine learning techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 4(3), Article n°3. https ://doi.org/10.17762/ijritcc.v4i3.1902

[16] Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (2e éd.). O'Reilly Media, Inc.

[17] GeeksforGeeks. (2017, juin 1). Introduction to dimensionality reduction. https ://www.geeksforgeeks.org/dimensionality-reduction/

[18] Murphy, K. P. (2012). Machine learning : A probabilistic perspective. MIT Press.

[19] Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

Résumé

Ce mémoire traite de l'application des techniques de machine learning à la prévision des ventes dans le secteur agroalimentaire, à travers une étude de cas réalisée au sein de l'entreprise Cevital. Â partir de données historiques de sauces, des années 2023 et 2024, un modèle prédictif basé sur l'algorithme XGBoost a été développé afin d'anticiper les ventes pour l'année 2025. Le travail comprend les étapes de collecte, de prétraitement, de modélisation et de visualisation à travers un tableau de bord interactif. Les résultats obtenus montrent que l'intelligence artificielle, et plus particulièrement l'apprentissage automatique, constitue un outil performant et stratégique pour la prise de décision en entreprise.

Abstract

This thesis explores the application of machine learning techniques for sales forecasting in the agri-food sector, using a case study at the company Cevital. Based on historical data from 2023 and 2024, a predictive model using the XGBoost algorithm was developed to forecast sales for the year 2025. The study covers data collection, preprocessing, modeling, and visualization through an interactive dashboard. The results demonstrate that artificial intelligence, and more specifically machine learning, is a powerful and strategic tool for business decision-making.