ECOLE NATIONALE SUPERIEURE
D'INFORMATIQUE ET D'ANALYSE DES SYSTEMES

FILIÈRE BUSINESS INTELLIGENCE AND ANALYTICS

Rapport de stage :

Amélioration de l'ERP avec un Système de Prévision des Ventes

---

Réalisé par:
BELLAHBIB Mhamed El Ghali

Encadrants :
Mme. LEKHDASSI Meryem
M. FQIHI Ayyoub
Membres du jury :
PR. DAADAOUI Latifa
PR. AMRANI Ibrahim

Année Universitaire : 2024-2025

Remerciements

Avant de développer davantage sur cette expérience professionnelle, je tiens tout d'abord à exprimer ma sincère gratitude envers ceux qui m'ont énormément appris tout au long de ce stage, et même envers ceux qui ont contribué à en faire un moment extrêmement bénéfique.

Ainsi, je souhaite tout d'abord exprimer ma profonde gratitude à mes encadrants, Mme Meryem Lekhdassi et M. Ayyoub Fqihi, pour leur précieux soutien, leur dévouement et leur disponibilité tout au long de ce projet. Leur expertise et leurs conseils ont été inestimables pour mener à bien ce travail dans les meilleures conditions. Je tiens également à remercier chaleureusement l'ensemble du personnel de l'entreprise Cogitas Solution pour leur générosité et leur partage d'informations, qui ont grandement contribué à la réussite de ce projet.

Je souhaiterais également adresser mes remerciements à mon chef de filière, M. Yasser EL MADANI EL ALAMI, ainsi qu'à tous mes enseignants, pour la formation précieuse qu'ils m'ont dispensée.

Résumé

Le présent rapport présente le travail effectué lors de mon stage de première année au sein de Cogitas Solution. Le projet avait pour objectif d'améliorer l'ERP en y intégrant une capacité de prévision des ventes pour aider les entreprises à optimiser la gestion de leurs stocks et la planification de leurs stratégies commerciales. Dans ce cadre, j'ai développé un modèle de prévision des ventes et conçu un tableau de bord interactif permettant de visualiser les résultats..

Dans le cadre de ce projet, j'ai comparé plusieurs modèles de machine learning en utilisant Python afin de sélectionner celui offrant les meilleures performances pour la prévision des ventes. J'ai également réalisé une segmentation des clients et une classification des remises, enrichissant ainsi l'analyse des comportements d'achat. La visualisation des données a joué un rôle crucial dans ce projet en rendant les informations plus accessibles et compréhensibles, mettant en lumière les tendances, les modèles et les relations entre les différentes variables de vente et de comportement client.

Pour cela, des tableaux de bord et des visualisations interactives ont été développés afin de présenter les statistiques et les informations clés issues de nos analyses et prévisions.Le base de données a été manipulé a l' aide de SQL Server et les résultats obtenus ont été visualisés à travers un tableau de bord interactif conçu avec Streamlit.

Mots clés : Prévision des ventes, Machine learning,Python,SQL Server, Streamlit, Tableau de bord, ERP.

Abstract

This report outlines the work conducted during my first-year internship at Cogitas Solutions. The project aimed to enhance the ERP by integrating a sales forecasting capability to help companies optimize inventory management and strategic planning. As part of this initiative, I developed a sales forecasting model and designed an interactive dashboard to visualize the results.

For this project, I compared various machine learning models using Python to select the one offering the best performance for sales forecasting. Additionally, I carried out customer clustering and discount classification, thereby enriching the analysis of purchasing behavior. Data visualization played a crucial role in this project by making information more accessible and understandable, highlighting trends, patterns, and relationships between different sales and customer behavior variables.

To achieve this, dashboards and interactive visualizations were developed to clearly and intuitively present the key statistics and insights derived from our analyses and forecasts. The database was managed using SQL Server, and the results were visualized through an interactive dashboard built with Streamlit.

KEYWORDS :Sales Forecasting, Machine Learning, Python, SQL Server, Streamlit, Dashboard, ERP.

Table des matières

Table des figures

Introduction générale

Dans un contexte commercial de plus en plus complexe et dynamique, la gestion efficace des stocks et la planification stratégique sont devenues des impératifs pour les entreprises. L'utilisation de prévisions de ventes dans un système ERP (Enterprise Resource Planning) qui se definit comme un système de gestion intégré qui centralise et coordonne les processus opérationnels clés d'une organisation, [3] est un atout majeur pour optimiser ces processus. Les prévisions de ventes permettent aux entreprises d'anticiper les besoins futurs, de mieux gérer leurs ressources et de prendre des décisions éclairées basées sur des données précises. Les tableaux de bord interactifs jouent également un rôle crucial en fournissant une visualisation claire et en temps réel des tendances et des indicateurs clés, facilitant ainsi la prise de décision stratégique.

Dans ce cadre, mon stage au sein de Cogitas Solution a été l'occasion de développer un modèle de prévision des ventes pour l'ERP de l'entreprise. Ce projet visait à offrir une solution innovante pour prévoir les ventes futures, en utilisant des techniques avancées de machine learning. Le modèle de prévision des ventes a été conçu pour être applicable aux données historiques extraites de l'ERP, permettant ainsi une analyse en temps réel des tendances de vente.

Pour réaliser ce projet, j'ai utilisé Python pour comparer différents modèles de machine learning et sélectionner celui offrant les meilleures performances pour la prévision des ventes. J'ai également importé et préparé les données à partir de SQL Server, avant de les visualiser à l'aide d'un tableau de bord interactif créé avec Streamlit. Le processus a inclus des étapes de préparation, d'exploration et de création de visualisations pour présenter les résultats de manière intuitive et accessible.

Le rapport est structuré en trois chapitres principaux : le premier chapitre introduit le contexte général du projet, présente l'entreprise d'accueil et les objectifs du stage. Dans le deuxième chapitre ,nous avons réalisé à la fois l'analyse de l'existant et la conception du projet, en mettant en évidence les besoins. Enfin, le troisième chapitre décrit les outils et technologies utilisés, ainsi que les résultats obtenus à travers le tableau de bord interactif.

1 Chapitre 1 :Présentation générale du projet

Le chapitre suivant introduit le contexte général dans lequel le projet s'est déroulé, cela dit une présentation de l'organisme d'accueil ainsi que le projet en question et enfin la problématique du projet, suivi des objectifs et de la méthodologie de travail.

1.1 Présentation de l'organisme d'accueil

Présentation de l'entreprise

Cogitas Solutions est une société marocaine fondée en 2010, spécialisée dans l'édition et l'intégration de solutions IT pour les PME. L'entreprise se consacre à démocratiser l'accès aux outils de gestion avancés, permettant aux PME de bénéficier de solutions dignes des grandes entreprises tout en maîtrisant leurs budgets. Son objectif est d'offrir aux PME les meilleures pratiques pour optimiser leur performance et gérer efficacement leur patrimoine.

Avec une vision ambitieuse d'expansion internationale, Cogitas Solutions se distingue par son engagement à intégrer les dernières avancées technologiques et les meilleures pratiques organisationnelles. L'entreprise propose des solutions de gestion innovantes, spécifiquement adaptées au contexte des entreprises marocaines, afin de libérer leur plein potentiel et de soutenir leur croissance.

FIGURE 2 – Cadre Stratégique

Projets

Cogitas Solutions a pu, en si peu de temps, décrocher plusieurs partenariats avec différentes communautés sur différents secteurs d'activités. Parmi les déploiements réalisés par l'enttreprise, on peut citer :

— RayOne ERP

FIGURE 3 – Logo : RayOne

RayOne est un progiciel de gestion intégrée dédié aux entreprises de taille moyenne qui cherchent une solution complète et optimale pour gérer leurs activités opérationnelles et piloter leur performance en toute sérénité.

— RH-P

FIGURE 4 – Logo :RH-P

RH-P est une solution spécialement conçue pour permettre aux entreprises de valoriser leur capital humain, d'exceller dans la gestion administrative, et de se conformer pleinement aux exigences légales et réglementaires en vigueur. Elle se distingue par sa capacité à harmoniser les processus RH avec le cadre réglementaire marocain, tout en assurant simplicité d'utilisation et efficacité opérationnelle.

— SIMPL-LIYAS

FIGURE 5 – Logo : Simple-Liyas

Simpl-Liyas est une solution de télédéclarations sociales et fiscales de référence destinée aux entreprises de toute taille et configuration, experts comptables, comptables agréés etc. afin de leur permettre de se conformer aux exigences déclaratives légales en vigueur au Maroc.

Quelques clients de Cogitas Solutions

FIGURE 6 – Quelques clients

Après avoir présenté l'organisme d'accueil, je passe maintenant en revue notre projet.

1.2 Présentation du projet

Je décris dans ce qui suit mon projet en exposant sa problématique et ses objectifs.

1.2.1 Problématique

Dans un contexte où les entreprises cherchent constamment à améliorer leur compétitivité, l'efficacité de la gestion des stocks et la précision des prévisions de vente sont devenues des enjeux cruciaux pour assurer une planification stratégique optimale. L'intégration de solutions de prévision des ventes au sein des systèmes ERP (Enterprise Resource Planning) peut fournir aux entreprises un avantage concurrentiel significatif en leur permettant d'ajuster rapidement leurs opérations aux fluctuations du marché.

Cependant, l'intégration de ces capacités prédictives dans un ERP existant pose plusieurs défis : Quels modèles de machine learning sont les mieux adaptés à ce type de prédiction dans le contexte spécifique des PME marocaines ? Et comment présenter ces résultats de manière claire et accessible à travers un tableau de bord interactif, afin de faciliter la prise de décision pour les utilisateurs finaux ? Mon projet chez Cogitas Solutions s'inscrit dans cette problématique, en explorant et développant une solution de prévision des ventes intégrée, adaptée aux besoins spécifiques des entreprises marocaines.

1.2.2 Objectifs du projet

Le projet de prévision des ventes intégré au système ERP de Cogitas Solutions nommé RAYONE poursuit plusieurs objectifs essentiels pour répondre aux besoins des entreprises marocaines en matière de gestion des stocks et de planification stratégique.

1. **Développer un modèle de prévision des ventes précis et adapté** : Concevoir et implémenter un modèle de machine learning capable de prédire avec précision les ventes futures en fonction des données historiques et des tendances du marché.
2. **Création de tableaux de bord interactifs** : Développer des visualisations claires et intuitives à travers des tableaux de bord interactifs, facilitant la prise de décision en offrant une vue d'ensemble des prévisions de vente et des recommandations associées.
3. Aider les entreprises à optimiser leur planification des stocks et leurs stratégies de vente en se basant sur des prévisions précises et fiables

Après avoir présenter mon projet, j'aborderais la méthodologie de travail suivie.

1.3 Méthodologie de travail

1.3.1 Méthodologie de développement

La méthodologie du projet a débuté par une exploration approfondie de l'ERP RAY-ONE à l'aide d'une version de démonstration. Cette phase initiale a permis de se familiariser avec le fonctionnement général du système ainsi que la structure de la base de données associée. Comprendre ces aspects était crucial pour préparer les étapes suivantes du projet. En acquérant une vue d'ensemble des fonctionnalités et des données disponibles, cette étape a fourni les bases nécessaires pour une manipulation efficace des données.

Suite à cette première étape, le projet a évolué vers l'utilisation d'une base de données démo pour préparer et explorer les données en profondeur. Cette phase a inclus le nettoyage des données,

l'analyse des caractéristiques clés et l'identification des patterns pertinents. Ce travail de préparation a été essentiel pour assurer la qualité des données et faciliter les analyses ultérieures.

Avec la version finale de la base de données à disposition, plusieurs modèles de machine learning ont été évalués afin de prédire les ventes de manière précise. Cette phase a impliqué la comparaison des performances des différents modèles pour sélectionner celui offrant les meilleures prévisions. Les critères de sélection ont inclus la précision des prévisions et leur adéquation avec les besoins spécifiques du projet.

Pour finaliser le projet, Streamlit a été utilisé pour développer un tableau de bord interactif, permettant de visualiser à la fois les prévisions de ventes et des visualisations de données de manière claire et accessible. Cette intégration a facilité la présentation des résultats et leur interprétation, fournissant ainsi aux utilisateurs une interface conviviale pour la prise de décisions basée sur les prévisions générées.

1.3.2 Planification du déroulement

Après avoir pris connaissance du sujet du projet et afin de structurer le déroulement du travail, il était crucial de bien planifier le projet en le divisant en étapes chronologiques. J'ai opté pour l'utilisation d'un diagramme de Gantt, qui offre une visualisation claire de toutes les tâches planifiées pour le projet.

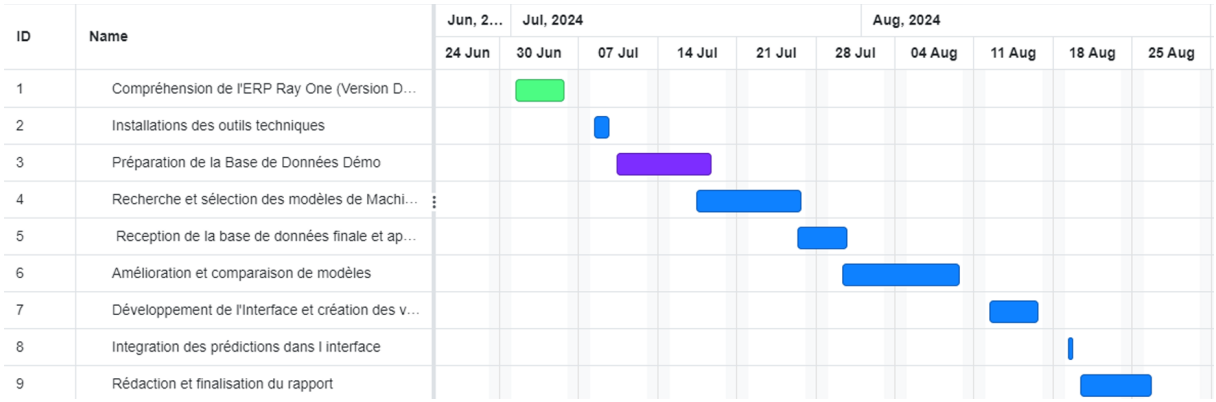


FIGURE 7 – Diagramme de GANTT

Conclusion

Dans ce chapitre, nous avons présenté l'organisme d'accueil, Cogitas Solutions, en mettant en lumière ses projets phares et sa clientèle. Nous avons ensuite exposé la problématique et les objectifs de notre projet, avant de détailler la méthodologie suivie pour sa réalisation. Le chapitre suivant sera consacré à l'analyse et à la conception du projet .

2 Chapitre 2 : Analyse et conception

Ce chapitre a pour objectif d'introduire à la fois l'analyse de l'existant et la conception du projet, en mettant en évidence la base de données utilisée. Il présente également les manipulations effectuées sur la base de données ainsi que les choix concernant les visuels et les mesures utilisées.

2.1 Analyse de l'existant

La base de données contient des informations sur tous les aspects de l'entreprise, incluant les achats, les ventes, le stock, la logistique, et la production, réparties en plusieurs tables, chacune dédiée à des données spécifiques. Ces tables fournissent des informations essentielles sur les ventes et sur la manière dont l'entreprise est gérée et organisée. Toutefois, pour obtenir une vue d'ensemble et comprendre comment les ventes sont structurées, je me suis concentré sur l'analyse de cinq tables principales :

- **Table des achats (Achat_DA_Ent) :** Cette table représente les achats effectués par l'entreprise. Elle regroupe des informations détaillées sur les articles achetés, leur provenance, la date d'achat, le fournisseur, ainsi que de nombreux autres indicateurs et détails liés aux achats.
- **Produit_Art :** Cette table offre une fiche détaillée de chaque article, incluant son prix minimum et maximum, ainsi que la famille de produits à laquelle il appartient.
- **tables de stocks (Stock_Mvt_His) :**Cette table enregistre en détail les mouvements de stock, c'est-à-dire les articles qui entrent et sortent, les dates de ces mouvements, et leurs emplacements respectifs.
- **Vente_Ent_Pie :**Cette table, qui constitue la base de mon projet, contient toutes les informations relatives aux ventes, notamment les données sur les clients, leurs emplacements, et les remises accordées.
- **Vente_Lig_Pie :**Cette table complète la précédente en fournissant les détails des actions liées aux ventes, en enregistrant les lignes de chaque transaction, par opposition à la première table qui représente les en-têtes des bons de livraison ou des factures.

Pour assurer la cohérence et l'intégrité des données, les différentes tables sont soigneusement liées entre elles par des clés. Pour des raisons de clarté, je me concentrerais uniquement sur l'analyse et l'exploitation des tables déjà expliquées précédemment. Cette structuration permet une analyse croisée des informations et facilite les requêtes complexes

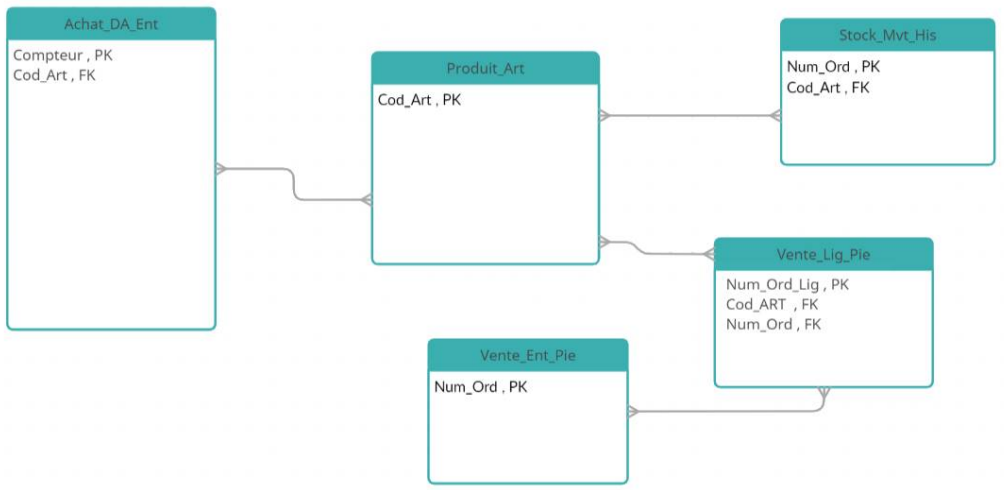


FIGURE 8 – Le modele conceptuel

2.2 Exigences fonctionnelles

1. Gestion des données de vente
2. Prévision des ventes
3. Visualisation des données
4. Clustering des clients
5. Interface utilisateur

2.3 Exigences non fonctionnelles

1. Performance : Assurer un temps de réponse rapide pour les requêtes et les prévisions
2. Sécurité : Garantir la protection des données sensibles des clients
3. Fiabilité : Assurer un fonctionnement sans interruption et des calculs précis

2.4 Besoins techniques

1. Base de données
2. Environnement de développement
3. Bibliothèques Machine Learning
4. Visualisation

2.5 Conception

2.5.1 Acteurs

le client est le principal acteur, utilisant l'interface pour accéder à des informations clés telles que les ventes, les commandes, et les prévisions de ventes. Cette interface permet au client d'ajuster ses stratégies de vente en fonction des prévisions et de prendre des décisions éclairées pour optimiser la gestion de son entreprise. Le système est conçu pour offrir une expérience utilisateur intuitive et efficace, facilitant l'accès et l'exploitation des données essentielles pour le succès commercial du client.

2.5.2 Diagramme de cas d'utilisation

Le diagramme des cas d'utilisation est un diagramme UML qui sert à définir les entités qui agissent avec le système, appelés acteurs, ainsi que les actions contenues dans le système, appelées cas d'utilisation ou use cases.

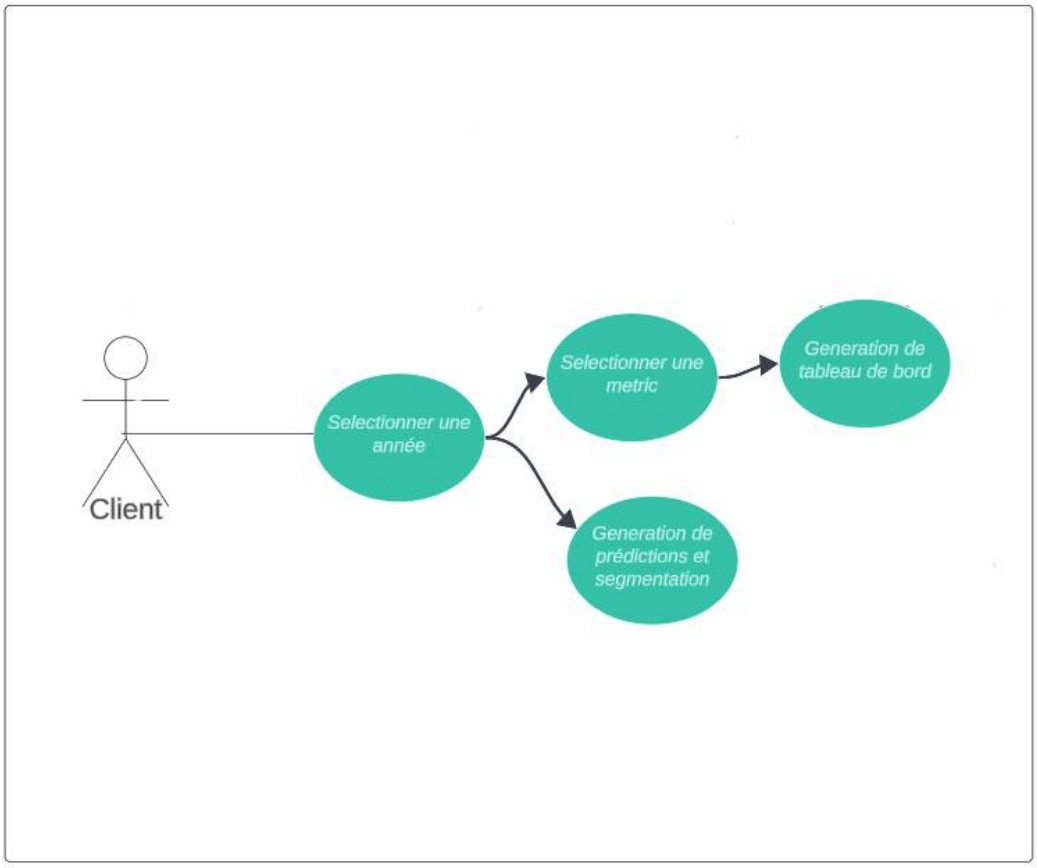


FIGURE 9 – Diagramme de cas d 'utilisation

2.5.3 Etapes de conception

— **Importation des données**
Récupération des données pertinentes à partir des différentes tables de la base de données pour constituer un ensemble de données complet et représentatif.

— **Préparation des données**
Nettoyage et transformation des données importées pour les rendre exploitables, y compris la gestion des valeurs manquantes, la normalisation des variables, et la création de nouvelles caractéristiques si nécessaire.[7]

— **Modélisation des prévisions de ventes**
Application et évaluation de plusieurs modèles de machine learning pour prédire les ventes futures. Cela inclut la sélection et l'entraînement de modèles appropriés, l'évaluation de leur performance, et l'ajustement des hyperparamètres.

— **Prédiction des classifications de remises**
Développement et évaluation de modèles de classification pour prédire les tranches de remise que les clients pourraient obtenir, en fonction de leurs caractéristiques et comportements d'achat.

— **Clustering des clients**
Utilisation de techniques de clustering pour segmenter les clients en groupes homogènes basés sur leur comportement d'achat, leur fréquence d'achat, et d'autres critères pertinents.

— **Génération de tableaux de bord**
Création de visualisations et de tableaux de bord pour présenter les résultats des analyses et des prévisions de manière claire et interactive.

— **Affichage via Streamlit**
Développement d'une interface utilisateur avec Streamlit pour permettre aux clients de visualiser les prévisions de ventes, les clusters de clients, et la classification de remises de manière interactive et intuitive.

Conclusion

Dans ce chapitre, j'ai détaillé les différentes tables de la base de données en soulignant leur importance pour notre projet. J'ai également abordé les exigences fonctionnelles, les besoins techniques, ainsi que les acteurs clés du système, illustrés par un diagramme d'utilisation. Ensuite, j'ai expliqué les étapes de la conception de l'interface utilisateur et de la modélisation des prédictions, en mettant l'accent sur la sélection des modèles de machine learning pour obtenir des résultats précis et pertinents pour l'entreprise.

3 Chapitre 3 : Réalisation et restitution

Ce chapitre détaille les étapes de la réalisation de notre projet, depuis l'importation des données jusqu'à la création de l'interface utilisateur et la prédiction des ventes . Nous décrivons les outils utilisés ainsi que les étapes clés de notre processus de développement.

3.1 Outils de travail

Python

Python est un langage de programmation interprété, de haut niveau et polyvalent, créé par Guido van Rossum et publié pour la première fois en 1991. Connu pour sa syntaxe simple et claire, Python favorise une programmation efficace et lisible, ce qui le rend idéal pour les débutants tout en étant suffisamment puissant pour les développeurs expérimentés. Il dispose d'une vaste collection de bibliothèques standard et externes qui facilitent le développement dans divers domaines tels que le web, les sciences des données, l'intelligence artificielle, et l'analyse financière. Python est également apprécié pour sa grande communauté d'utilisateurs et de développeurs, son interopérabilité avec d'autres langages de programmation, et sa portabilité à travers différentes plateformes. Ces caractéristiques font de Python un outil précieux pour le développement rapide et la maintenance de projets complexes.

FIGURE 10 – Logo :Python

VS CODE

Visual Studio Code est un éditeur de code open-source et cross-platform développé par Microsoft. Conçu pour les développeurs web, il prend en charge de nombreux langages et offre de nombreuses fonctionnalités comme la coloration syntaxique, l'auto-complétion et le débogage. Grâce à son vaste écosystème d'extensions, VSCode est hautement personnalisable et peut être adapté aux besoins spécifiques de chaque développeur.

FIGURE 11 – Logo :VS CODE

Jupyter

Jupyter est une application open-source qui permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte narratif. Utilisé principalement dans les domaines de la science des données, de l'enseignement et de la recherche, Jupyter prend en charge plus de 40 langages de programmation, dont Python, R et Julia. L'interface interactive des notebooks Jupyter permet d'exécuter des blocs de code individuellement et de voir immédiatement les résultats, ce qui facilite l'exploration des données, le prototypage rapide et la collaboration. Les notebooks Jupyter sont particulièrement appréciés pour leur capacité à combiner du code exécutable avec des explications détaillées, des graphiques et des visualisations, offrant ainsi un environnement complet pour le développement et la présentation de projets analytiques.

FIGURE 12 – Logo :Jupyter

SQL Server Management Studio

SQL Server Management Studio (SSMS) est un outil complet développé par Microsoft pour gérer et administrer les bases de données SQL Server. Il fournit une interface graphique qui permet aux utilisateurs de créer, modifier, et exécuter des requêtes SQL, ainsi que de gérer les objets de la base de données comme les tables, vues, et procédures stockées. SSMS est conçu pour répondre aux besoins des administrateurs de bases de données, des développeurs, et des analystes de données, en offrant des fonctionnalités avancées pour la configuration de la sécurité, l'analyse des performances, et l'automatisation des tâches. Il inclut également des outils de diagnostic pour surveiller l'état du serveur et optimiser les requêtes, ce qui en fait un outil essentiel pour toute personne travaillant avec SQL Server.

FIGURE 13 – Logo : SQL Server Management Studio

3.2 Réalisation

3.2.1 Importation et Préparation des Données

Dans cette étape initiale, les données ont été importées depuis SQL Server. La préparation des données a consisté à supprimer les valeurs dupliquées et les valeurs nulles, ainsi qu'à retirer les colonnes non nécessaires pour le projet. Les données ont été travaillées sur deux tables principales :

1. **Table de jointure :** La table Vente_Ent_Pie a été liée à la table Vente_Lig_Pie sur la colonne Num_Ord, avec un filtrage des lignes de type "pièce" pour ne conserver que les bons de livraison.

```
query = '''
SELECT z.*, t.*
FROM [data_fin].[dbo].[Vente_Ent_Pie] AS z
JOIN [data_fin].[dbo].[Vente_Lig_Pie] AS t
    ON z.Num_Ord = t.Num_Ord
WHERE z.Typ_Pie = 'BL' AND t.Typ_Lig = 'L'
'''
df = pd.read_sql(query, conn)
```

FIGURE 14 – Code d'importation des données

2. **Table de vente :** La table Vente_Ent_Pie a été utilisée seule pour les cas où les détails de chaque commande n'étaient pas requis.

```
conn = pyodbc.connect(connection_str)
query2 = '''
    select  z.*
    from[data_fin].[dbo].[Vente_Ent_Pie] as z

    where  z.Typ_Pie='BL'
    '''

df2= pd.read_sql(query2, conn)
```

FIGURE 15 – Code d'importation des données2

3.2.2 Prédiction des ventes

La prédiction des ventes est un problème de séries temporelles, ce qui nécessite d'organiser les données de manière spécifique pour obtenir les meilleures prévisions possibles. Ainsi, j'ai structuré mon DataFrame en deux colonnes : l'une contenant les dates et l'autre les ventes. J'ai regroupé les données de manière mensuelle. Cependant, certains mois manquaient de données, c'est-à-dire qu'il n'y avait pas de ventes enregistrées pour ces périodes. Pour ces mois, j'ai ajouté des lignes avec des ventes à 0, car il est crucial de définir une fréquence régulière pour analyser correctement les données de séries temporelles.

—

	Date	Mnt_Lig_Net
0	2011-02-01	120.00
1	2011-03-01	1500.00
2	2011-04-01	33600.00
3	2011-06-01	20094.00
4	2011-07-01	6230.00
5	2011-08-01	45928.00
6	2011-09-01	101586.95
7	2011-10-01	80810.50
8	2011-11-01	122444.00
9	2011-12-01	75596.50

	Date	Mnt_Lig_Net
0	2011-02-01	120.00
1	2011-03-01	1500.00
2	2011-04-01	33600.00
3	2011-05-01	0.00
4	2011-06-01	20094.00
5	2011-07-01	6230.00
6	2011-08-01	45928.00
7	2011-09-01	101586.95
8	2011-10-01	80810.50
9	2011-11-01	122444.00

FIGURE 16 – Dataframe des ventes mensuelles

Analyse des Composants

Afin d éviter des problemes par la suite j ai rempalce les valeurs de 0 par 100 .

Decomposition

Décomposer des données de séries temporelles signifie les diviser en trois composantes : la tendance, qui représente la direction générale des données (qu'elles soient à la hausse ou à la baisse), la saisonnalité, qui correspond aux motifs cycliques, et les résidus, qui sont les variations aléatoires. Ces composantes peuvent être traitées différemment selon la question posée et le contexte, en les additionnant dans un modèle additif ou en les multipliant dans un modèle multiplicatif. Dans un modèle multiplicatif, la tendance générale augmente à un rythme croissant au fil du temps, et les fluctuations saisonnières (les pics et les creux) deviennent de plus en plus importantes. En revanche, dans un modèle additif, la tendance augmenterait de manière constante, tout comme les variations saisonnières resteraient stables au fil du temps.[2]

— **Decomposition Seasonal_Decompose**

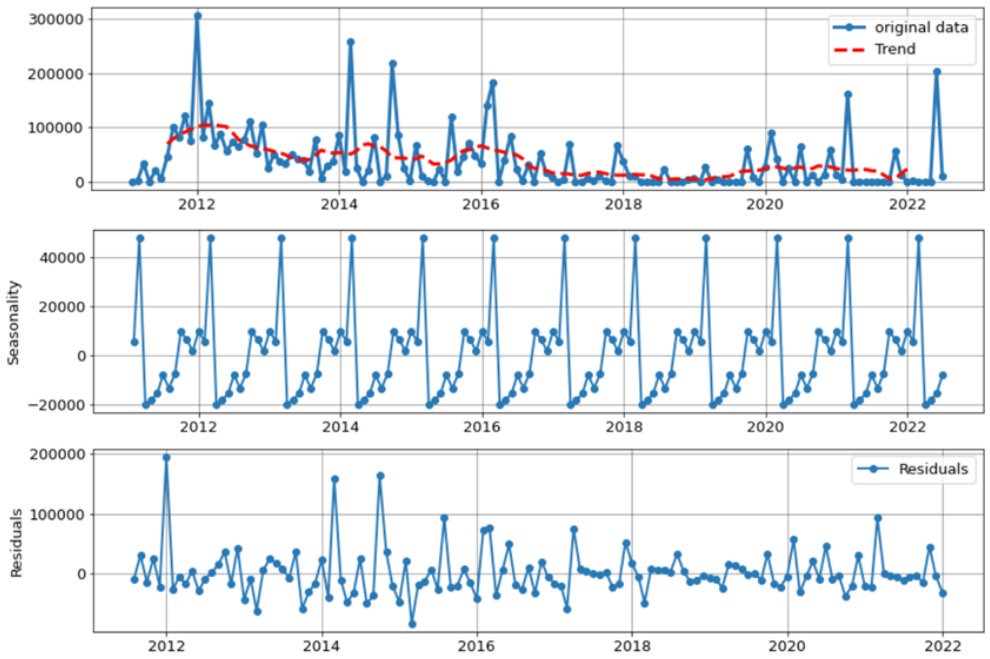


FIGURE 17 – SESONAL DECOMPOSE

— **Decomposition STL**

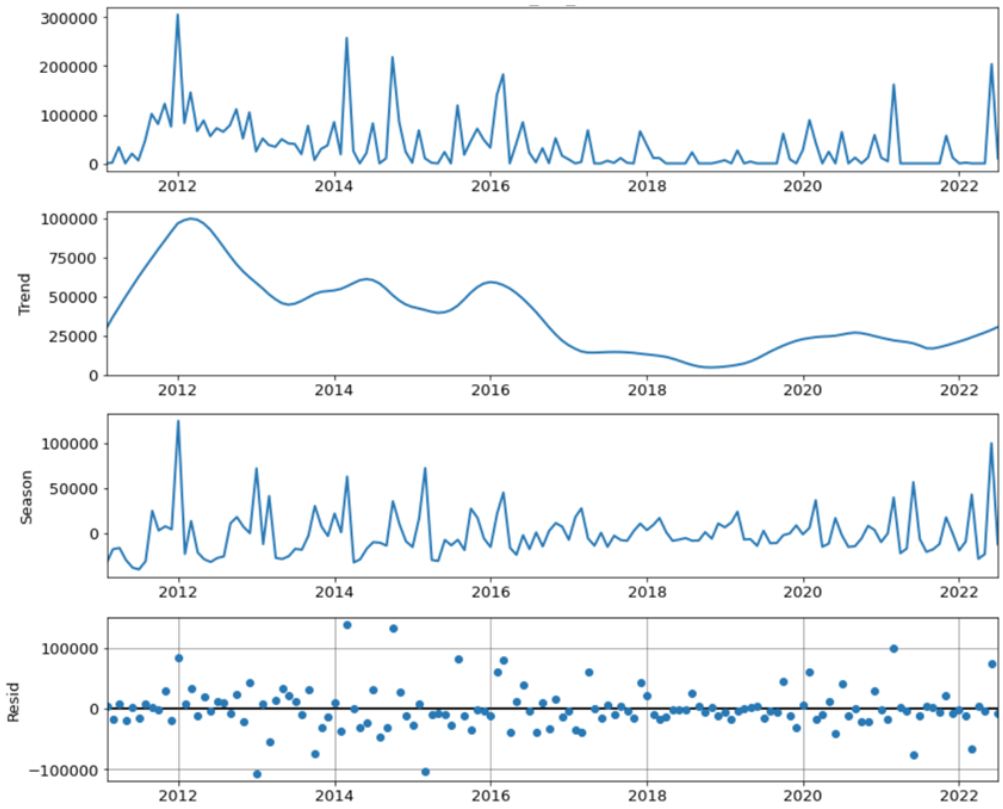


FIGURE 18 – STL

Dans la décomposition STL, le graphique de tendance reflète toutes les données sans aucune donnée manquante (contrairement à seasonal_decompose, qui montre une année de données manquantes). La saisonnalité dans STL reflète les tendances, tandis que dans seasonal_decompose, elle est supposée se répéter d'année en année.

— **Interpretation**
Comme nous pouvons constater notre data ne suit pas un trend et non plus une seasonalité

Stationnarisation des Données
lors de l'analyse des séries temporelles, il est crucial de vérifier la stationnarité. En termes simples, une série temporelle est considérée comme stationnaire si ses caractéristiques statistiques, telles que la moyenne, la variance ,restent constantes dans le temps. Pour tester la stationnarité, l'une des méthodes les plus courantes est le test KPSS. L'hypothèse nulle de ce test est que la série temporelle est stationnaire. Si la valeur p obtenue est inférieure au niveau de signification choisi (généralement 0,05), nous rejetons l'hypothèse nulle et concluons que les données ne sont pas stationnaires.

Voici le p-value de notre data

```
*******Results of kpss Test*******
Test Statistic           0.99
p-value                  0.01
Lags Used                4.00
Critical Value (10%)     0.35
Critical Value (5%)      0.46
Critical Value (2.5%)    0.57
Critical Value (1%)      0.74
dtype: float64
****************
```

FIGURE 19 – TEST KPSS1

Puisque le p-value est inferieur ici a 0.05 alors notre data n est pas stationnaire et par conséquent il faut la rendre stationnaire Pour le faire nous avons fait deux tansfortmation le logarithme et puis en differenciant les données ces deux opérations géneralement rendent la moyenne et la variance constantes

Voici le test après les modifications

```
*******Results of kpss Test*******
Test Statistic            0.17
p-value                   0.10
Lags Used                32.00
Critical Value (10%)      0.35
Critical Value (5%)       0.46
Critical Value (2.5%)     0.57
Critical Value (1%)       0.74
dtype: float64
************
```

FIGURE 20 – TEST KPSS2

Donc on remarque le la valuer de p est maintenant superieur a 0.05 ce qui siginifie que notre data est devenu maintenant stationnaire.

Modèles utilisés

Après avoir traité et modifié nos données, il est temps de sélectionner le modèle de machine learning le plus approprié pour obtenir les meilleurs résultats dans notre projet. Nous avons travaillé avec trois modèles : ARIMA, FBProphet, et NeuralProphet.

— **ARIMA**

L'ARIMA (AutoRegressive Integrated Moving Average) est un modèle statistique utilisé pour prévoir les séries temporelles en combinant autorégression, différenciation pour stationnariser la série, et moyenne mobile. Il est efficace pour les séries avec des tendances ou des cycles saisonniers.[1] dans le deux figures suivantes je vais illustrer le score que j ai eu plus precisement le rmse et le mae de ce modele et puis dans un tableau une comparaison entre les valeurs predites et actuelles

```
RMSE: 57040.113637390794
MAE: 23654.34692990719
```

FIGURE 21 – RMSE et MAE d'ARIMA

date	predictions	actual
2021-11-01	187.21	56790.00
2021-12-01	190.34	12010.00
2022-01-01	194.12	100.00
2022-02-01	197.71	1601.00
2022-03-01	201.33	100.00
2022-04-01	205.14	100.00
2022-05-01	208.94	100.00
2022-06-01	212.83	203775.00
2022-07-01	216.81	10360.00

FIGURE 22 – Tableau comparatif entre les valeurs actuelles et predites

— **FBPROPHET**
Prophet, développé par Facebook, est un modèle de prévision de séries temporelles simple à utiliser, conçu pour gérer des données avec des tendances, des saisons, et des événements spéciaux. Il est particulièrement adapté aux données irrégulières et aux prévisions business[4]. Afin de pouvoir entrainé ce modele il fallait tout d abord renommer les colonnes en ds et y car c est une phase obligatoire . J' ai amelioré le modele en ajoutant des variables externes a l aide de (add.regressor) pour signaler le modele des zeros présents.

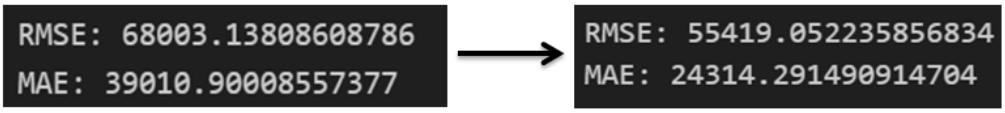


FIGURE 23 – RMSE et MAE de FBPROPHET

ds	yhat	y_test
2021-11-30	0.000000	56790.0
2021-12-31	28428.165125	12010.0
2022-01-31	22245.533009	0.0
2022-02-28	1925.761226	1601.0
2022-03-31	0.000000	0.0
2022-04-30	0.000000	0.0
2022-05-31	0.000000	0.0
2022-06-30	0.000000	203775.0
2022-07-31	0.000000	10360.0

FIGURE 24 – Tableau comparatif entre les valeurs acutelles et predites

— **NEURALPROPHET**
NeuralProphet est une extension de Prophet qui intègre des réseaux neuronaux pour mieux capturer les patterns complexes des séries temporelles. Il combine la facilité d'utilisation de Prophet avec la puissance du deep learning pour des prévisions plus précises.
j ai pu ameliore les scores en precisant le type de growth en logistic et le seasonality_mode en multiplicative

```
RMSE: 50247.24546081796
MAE: 34142.34275716146
```

FIGURE 25 – RMSE et MAE de NEURALPROPHET

— **Comparaison et choix du modeles**
Comme nous pouvons le constater, les modèles FBProphet et ARIMA présentent des scores très proches. Cependant, en comparant les valeurs prédites aux valeurs réelles, il apparaît que les prévisions d'ARIMA sont nettement inférieures, ce qui semble peu rationnel. En revanche, FBProphet fournit des valeurs plus proches de la réalité, malgré des scores élevés pour les deux modèles. Cela est probablement dû à la nature des données, qui, selon notre analyse, ne présente aucune saisonnalité ou tendance identifiable. Ainsi, **nous avons choisi d'utiliser FBProphet** pour la prédiction des ventes.
Il est à noter que les prévisions s'interrompent en juillet 2023, ce qui est dû au fait que les données disponibles s'arrêtent en juillet 2022. Voici le schéma de prédiction pour les 12 mois à venir

FIGURE 26 – Graphe de prédiction des ventes

3.2.3 Classification des remises

La prédiction des classes de remises constitue un problème de classification multi-classe, car les remises ont été classées en quatre catégories : remise de 0 (0), remise inférieure à 2000 (1), remise entre 2000 et 4000 (2), et remise supérieure à 4000 (3). Pour ce faire, nous avons prétraité les données en conservant uniquement les colonnes pertinentes pour la prédiction, réduisant ainsi le nombre de colonnes de 121 à 20. Pour avoir une meilleur précision nous avons appliqué deux codages :

— **Label encoding :**convertit chaque catégorie d'une variable en un entier unique. Cette méthode est particulièrement adaptée aux données ordinales, où les catégories ont un ordre naturel.

— **Hot-encoding :**transforme chaque catégorie en une colonne binaire distincte, où une colonne est créée pour chaque catégorie et contient des valeurs 0 ou 1 pour indiquer la présence ou l'absence de cette catégorie. Cette méthode est idéale pour les données nominales.[5]

Dans notre projet le codage de labels a été utilisé pour les colonnes client et id_article, tandis que le codage one-hot a été appliqué à la colonne ville pour faciliter une meilleure interprétation par les modèles. Ensuite, nous avons divisé notre DataFrame en caractéristiques (features) et en cible (target), cette dernière étant la classe de remise.

— **Choix du modèle**
Après la transformation des données et le choix des Features, nous avons procédé à la sélection des modèles de classification. Nous avons expérimenté plusieurs modèles, notamment RandomForestClassifier, XGBoostClassifier, LogisticRegression, et DecisionTreeClassifier. Voici les scores que nous avons obtenus pour chaque modèle :

```
Random Forest Accuracy: 0.9776785714285714
XGBoost Accuracy: 0.9732142857142857
Logistic Regression Accuracy: 0.9508928571428571
DecisionTree Accuracy: 0.9464285714285714
```

FIGURE 27 – Mesures d'evaluation des modèles de classification

Comme nous pouvons le constater, le **RandomForestClassifier** affiche la précision la plus élevée parmi les modèles testés. Ce modèle de classification pourrait offrir des informations précieuses, notamment si nous l'appliquons à des tâches telles que la classification des quantités de retours. Bien que notre base de données actuelle ne contienne pas de valeurs spécifiques pour cette colonnes, il reste intéressant d'explorer son utilisation. Par exemple, si nous détectons qu'un client reçoit systématiquement des remises qui ne correspondent pas à celles prévues, cela pourrait nous aider à identifier des anomalies potentielles.

3.2.4 Clustering de clients

La segmentation des clients, ou clustering, est une technique d'apprentissage non supervisé qui vise à diviser les données en groupes distincts appelés clusters. L'objectif est de regrouper les données de manière à ce que les objets au sein de chaque groupe soient plus similaires entre eux qu'avec ceux des autres groupes.[6] Pour ce faire, nous avons travaillé avec la table Vente_Ent_Pie, en filtrant sur les bons de livraison. Avant de procéder à la segmentation, une analyse RFM est nécessaire, et elle sera expliquée dans la partie suivante.

— **L'anlyse RFM** :
L'analyse RFM (Recency, Frequency, Monetary) est une approche basée sur le comportement des clients, permettant de les regrouper en segments selon leurs transactions d'achat passées[8]. Cette méthode évalue trois critères : Recency (R), Frequency (F) et Monetary Value (M).

- **Recency (R) :** Mesure la récente activité d'achat du client. Il s'agit du nombre de jours écoulés depuis le dernier achat ; un faible nombre de jours indique une faible récence et donc une récente activité.
- **Frequency (F) :** Évalue la fréquence des achats du client. Il correspond au nombre total d'achats effectués ; une fréquence élevée indique que le client achète régulièrement.
- **Monetary Value (M) :** Indique le montant total dépensé par le client. Une valeur monétaire élevée signifie que le client a dépensé beaucoup d'argent.

Alors nous avons appliqué cette analyse a notre dataframe et voici le résultat obtenu :

Nom_Clt_Fac	recency	frequency	monetary
ADM MOROCCO	1646	2	12720.0
AFRIC LIGHT	1025	6	319453.2
AGQ Maroc s.a.r.l.	2929	10	219909.0
AGUALANDIA Maroc sarl.	3857	1	10308.0
AKA GOLF	532	4	22800.0
...	...	...	...

FIGURE 28 – Dataframe apès RFM

— **Determination des clusters** :
Pour déterminer les segments, nous avons choisi l'algorithme K-means, qui divise les données en k clusters, chaque cluster étant représenté par la moyenne des points qui lui sont attribués. Les points sont assignés au cluster dont le centre est le plus proche[8]. Afin de déterminer le nombre optimal de clusters, nous avons opté pour la méthode du coude (elbow), qui nous aide à sélectionner le k le plus approprié.
La méthode du coude est utilisée après scaling de la data pour choisir le nombre optimal de clusters en évaluant la variation de la variance intra-cluster à mesure que le nombre de clusters change
Afin d'identifier le k on analyse le graphique pour identifier le "coude" ou le point où la courbe commence à s'aplatir. Ce point indique le nombre optimal de clusters, car il représente un équilibre entre la réduction de la variance intra-cluster et la complexité du modèle .

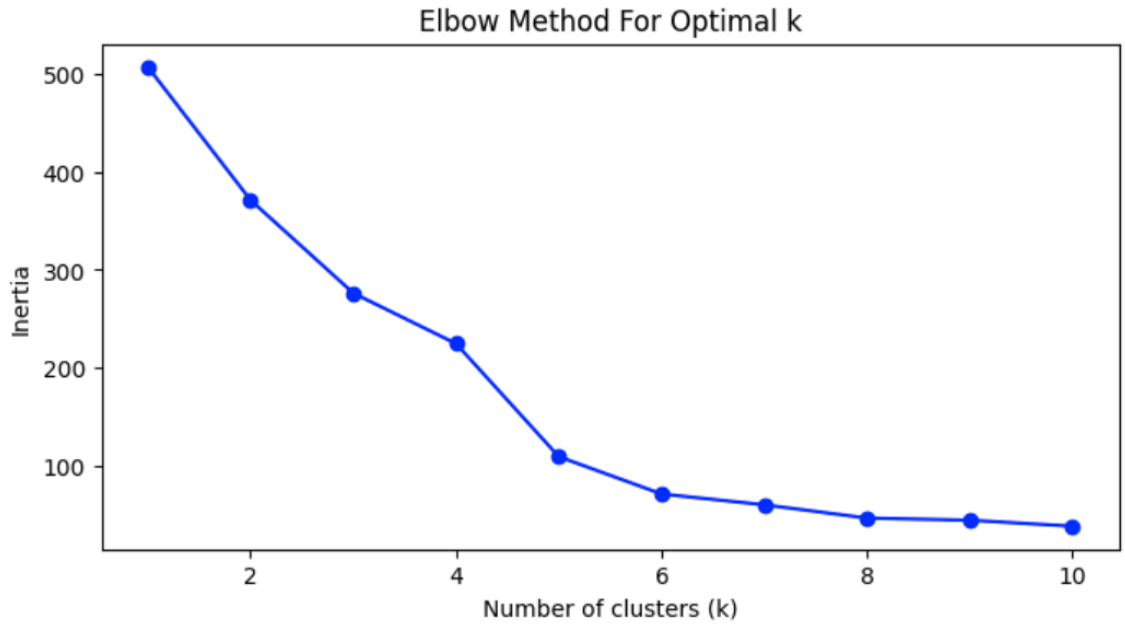


FIGURE 29 – Methode de l'Elbow

Comme nous pouvons le constater, le point de coude est à 5, ce qui signifie que nous formerons 5 groupes de clients.

Voici ci-dessous les moyennes pour chacun de ces groupes.

Cluster	recency	frequency	monetary
0	624.307692	3.576923	6.640754e+04
1	3571.575758	2.090909	1.995830e+04
2	2766.000000	9.000000	1.171467e+06
3	2209.571429	1.900000	2.364137e+04
4	2739.500000	21.000000	2.101462e+05

FIGURE 30 – Les moyennes de chaque cluster

En se basant sur ces moyennes, nous avons attribué à chaque groupe un nom qui reflète leur fidélité ou leur risque de perte. La figure ci-dessous présente ces noms ainsi que le nombre de clients dans chaque catégorie.

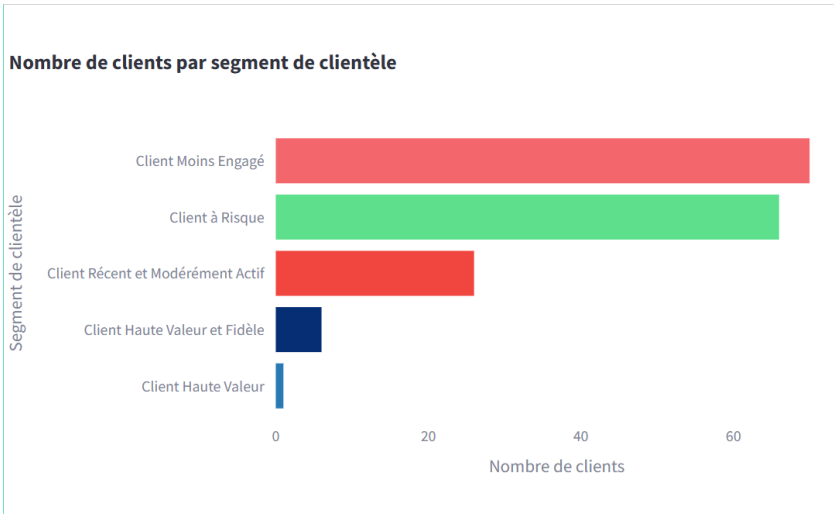


FIGURE 31 – Nombre de clients par segment de clientèle

3.2.5 Interface

Parallèlement aux modèles de prédiction, j'ai créé plusieurs visualisations pour constituer un tableau de bord intégrant des graphiques, des KPI et des prévisions. En utilisant la bibliothèque Python Streamlit, j'ai pu afficher ces visualisations et prédictions de manière efficace.

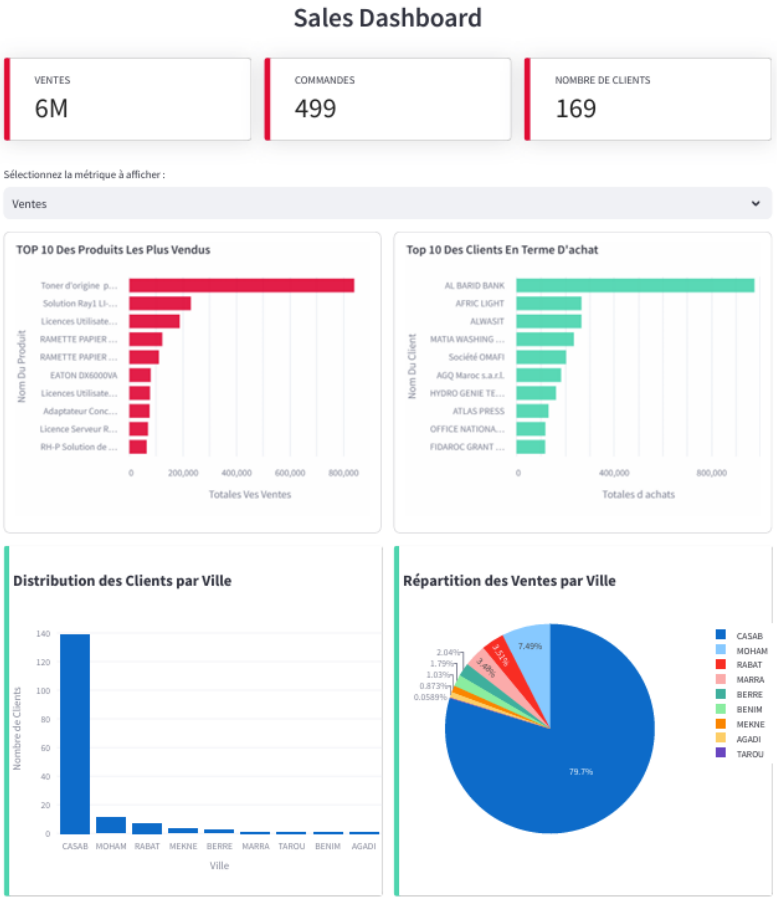


FIGURE 32 – Interface

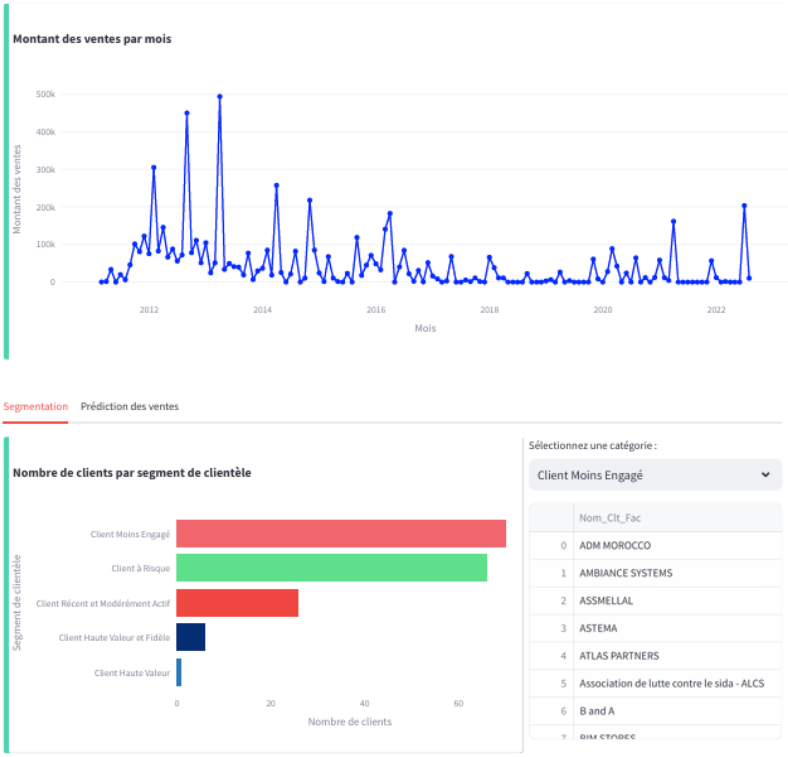


FIGURE 33 – Interface

1. **Visualisations :** Avant d'afficher les visualisations, le client est invité à sélectionner soit une année spécifique, soit l'ensemble de l'historique des ventes à l'aide d'un bouton de sélection. Une fois la sélection effectuée, le tableau de bord se charge et présente des métriques telles que le montant des ventes, le nombre de clients et le nombre de commandes, tout en affichant les pourcentages d'évolution par rapport à l'année précédente.

FIGURE 34 – Visualisations

Ensuite, le client choisit entre les ventes ou les quantités commandées. Deux graphiques s'affichent alors, montrant les dix principaux produits et clients en fonction du critère sélectionné,en dépendant bien sur de l'année choisie

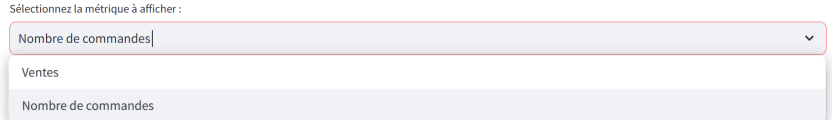
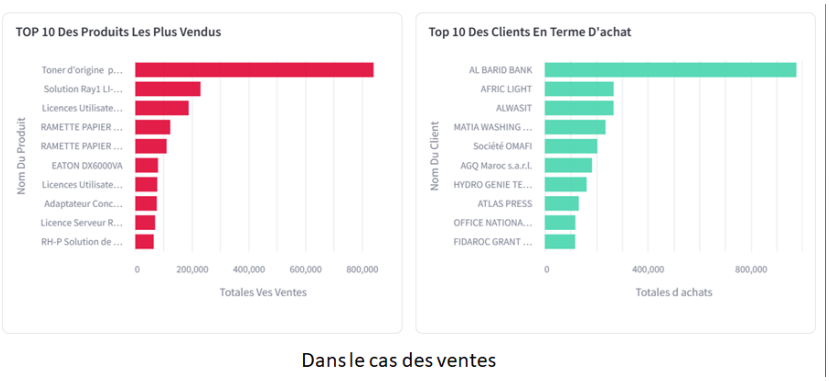


FIGURE 35 – Select box



Ensuite, un graphique à barres et un graphique circulaire illustrent la répartition des clients et des ventes par ville.

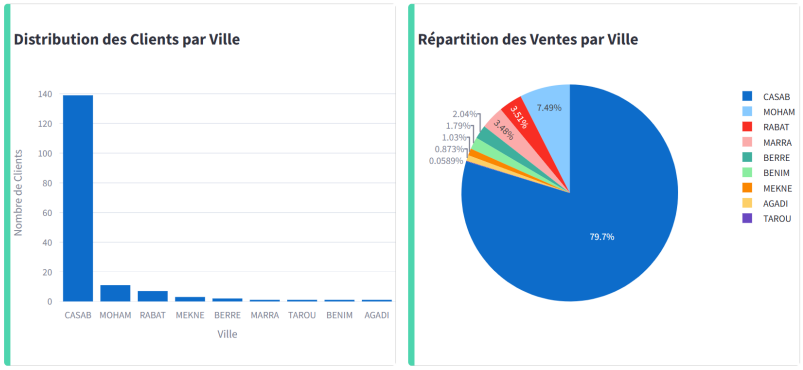
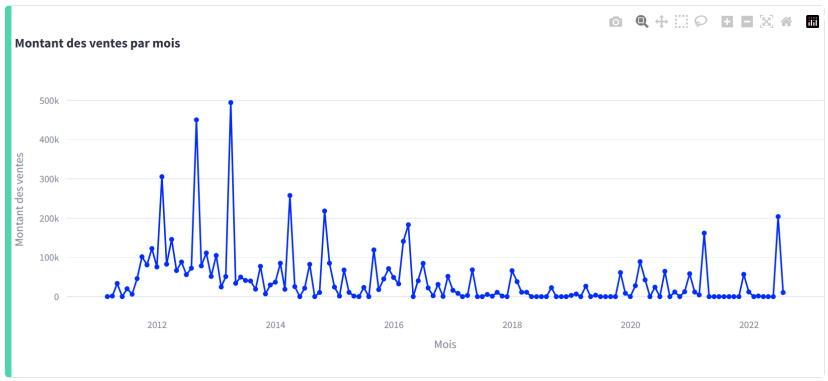


FIGURE 36 – Visualisations

et enfin un graphique representant le montant des ventes chaque mois



2. **Prédictions :**
   Cette interface m'a permis d'afficher les prévisions que j'ai déjà réalisées, notamment les prédictions des ventes pour les 12 prochains mois ainsi que la segmentation des clients.
   — **Prédiction des ventes pour les 12 prochains mois**

FIGURE 37 – Prédictions des ventes

— **Nombre de clients par groupes de clients**

Pour cette présentation, j'ai ajouté un selectbox permettant au client de connaître les noms exactes des clients regroupés dans chaque cluster.

FIGURE 38 – Segmentation de clients

Conclusion

Dans ce chapitre, nous avons etayé la réalisation de notre projet. en commençant par les technologies utilisées dans la mise en œuvre de cette application. Ensuite nous avons présenté comment nous avons préparé et analysé les données, sélectionné et appliqué des modèles de prévision adaptés. Enfin, une interface interactive a été créée pour présenter les analyses et les prévisions, offrant ainsi une vue enrichie des ventes et des comportements des clients, bien que cette fonctionnalité ne soit pas intégrée directement dans l'ERP.

Conclusion générale

Ce rapport a présenté en détail les différentes étapes de notre projet de prévision des ventes au sein de Cogitas Solutions, dans le cadre d'un contexte commercial complexe nécessitant une gestion efficace et une planification stratégique. En utilisant des techniques de machine learning, nous avons préparé et analysé les données afin de sélectionner le modèle le plus performant pour les prévisions. L'interface interactive crée permet de visualiser les analyses et les prévisions de manière claire, bien que cette fonctionnalité ne soit pas intégrée directement dans l'ERP de l'entreprise. Ce projet illustre l'importance des outils modernes dans l'optimisation des processus décisionnels et la gestion des ressources, en offrant des prévisions de ventes précises et en facilitant la prise de décisions stratégiques. Ce projet a été une opportunité enrichissante pour développer mes compétences en matière de machine learning et de visualisation de données .

Pour aller plus loin, nous envisageons l'optimisation des algorithmes existants afin d'améliorer encore la performance et la précision des prévisions de ventes et de la classification des remises. De plus, l'amélioration de l'interface utilisateur est une priorité, dans le but d'offrir une expérience plus fluide et accessible, facilitant l'interprétation des données et la navigation dans les tableaux de bord interactifs. L'exploration d'algorithmes de machine learning plus avancés serait également un axe d'amélioration pour perfectionner la segmentation des clients et obtenir des insights plus précis sur leurs comportements d'achat. Enfin, l'intégration de nouvelles sources de données permettrait d'enrichir les analyses et d'affiner les stratégies de prise de décision.

Références

[1] Autoregressive integrated moving average (ARIMA) : définition. https://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligence-artificielle/1501315-autoregressive-integrated-moving-average-arima-definition/.

[2] Décomposition d'une série temporelle. https://blog.statoscop.fr/timeseries-4.html.

[3] ERP : définition, rôle, fonctionnement et avantages. https://axelor.com/fr/erp-definition/.

[4] Facebook Prophet : Tout ce qu'il faut savoir. https://datascientest.com/facebook-prophet-tout-savoir.

[5] One Hot Encoding vs. Label Encoding in Machine Learning. https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/.

[6] Qu'est-ce que le clustering ? https://www.50a.fr/0/clustering.

[7] Tout savoir sur le traitement des données. https://www.talend.com/fr/resources/what-is-data-processing/.

[8] Understanding RFM Segmentation and K-Means Clustering for Effective Marketing Strategy. https://medium.com/@radhityan/understanding-rfm-analysis-for-effective-marketing-c21bc1e3b6bb.