

DIPLÔME NATIONAL D'INGENIEUR

SPÉCIALITÉ : INFORMATIQUE | ERP/BI

**Développement d'un système
d'affectation de produits aux
clients au profit d'une banque**

Nature du stage : Projet de Fin d'Etudes

Réalisé par : Fatma Ezzahra CHAMMEM

Encadrée par :

Encadrant en entreprise : M. Mohamed Salah Baccouche

Encadrante pédagogique : Mme Amani Lamine



EY

Année universitaire : 2021-2022



J'autorise l'étudiante Fatma Ezzahra Chammem à faire le dépôt de son rapport de stage en vue d'une soutenance.

Encadrant professionnel : M. Mohamed Salah Baccouche

A handwritten signature in blue ink, enclosed in an oval, belonging to Mohamed Salah Baccouche.



J'autorise l'étudiante Fatma Ezzahra Chammem à faire le dépôt de son rapport de stage en vue d'une soutenance.

Encadrante académique : Mme Amani Lamine

Remerciements

Je ne saurais commencer la rédaction de ce rapport sans remercier toutes les personnes, qui par leurs conseils, leurs suggestions ou par leur simple présence m'ont permis de rendre mon travail aussi instructif qu'efficace.

Tout d'abord, je tiens à adresser ma reconnaissance à mon encadrante pédagogique Mme Amani Lamine pour ses conseils et sa disponibilité tout au long du stage.

Mes vifs remerciements s'adressent à mon encadrant Mohamed Salah Baccouche, qui malgré son emploi du temps chargé et ses préoccupations a réussi à rendre cette expérience plus enrichissante tant sur le plan professionnel que personnel à travers sa disponibilité, ses conseils et son sens de l'écoute.

Je tiens également à adresser ma profonde gratitude envers les membres de l'équipe qui a réussi à donner vie à ce projet, notamment Mayssa Jebari, Baha Arfaoui et Firas El Mehdi, pour m'avoir accueilli si chaleureusement au sein de l'équipe, et pour le partage de connaissances qu'ils m'ont fourni.

Je tiens aussi à remercier toute l'équipe EY Tunisie pour l'environnement de travail convivial ainsi que toutes les connaissances que j'ai acquises durant ces 6 derniers mois.

Finalement, j'adresse ma gratitude et mes profonds respects à toute l'équipe pédagogique de l'école supérieure privée d'ingénierie et de technologie (ESPRIT) pour la qualité de la formation dont j'ai pu bénéficier au cours de mon cursus, ainsi qu'aux membres du jury pour avoir accepté d'évaluer mon travail.

Dédicaces

C'est avec une profonde gratitude que je dédie ce travail à...

Mon père, Yadh

Aucune dédicace ne saurait exprimer l'amour, le respect et l'attachement que j'ai pour toi. Rien au monde ne pourra compenser les sacrifices que tu as faits n'épargnant ni santé ni efforts pour mon éducation, mes études et mon bien-être. J'espère que tu es fier de moi et J'espère ne jamais te décevoir, ni trahir ta confiance et tes sacrifices.

Ma mère, Dorra

Une maman unique et extraordinaire, une source inépuisable de tendresse et de bonne humeur, merci de m'avoir toujours soutenue et encouragée. Merci d'avoir toujours cru en moi, ta présence à mes côtés a toujours été une source de force.

Mon frère Zayn et ma sœur Maha

Pour leur amour, tendresse et encouragements, je ne saurai traduire sur du papier l'affection que j'ai pour vous. Merci d'avoir toujours été à mes côtés.

Au reste de ma famille et à mes amis qui se reconnaîtront

Qui ont toujours été à mes côtés aux bons comme aux mauvais moments, merci pour le soutien et le réconfort que vous m'avez offert à chaque fois que j'en ai eu besoin.

Puisse Dieu tout puissant, vous préserver et vous accorder santé, longue vie et bonheur.

Résumé

Le présent document est la synthèse du travail que j'ai réalisé dans le cadre de mon projet de fin d'études, intitulé « Développement d'un système d'affectation de produits aux clients au profit d'une banque ». Ce système intervient dans l'anticipation du comportement de la clientèle d'une banque, en utilisant des techniques de Machine Learning telles le modèle K-means, utilisé pour la répartition des clients similaires en groupes, afin de proposer une recommandation des produits ainsi qu'une nouvelle segmentation dans le but d'améliorer le taux d'équipement des clients au sein de la banque.

Mots clés : Segmentation clients, Marketing prédictif, Visualisation de données, Machine Learning, Kmeans

Abstract

This document is the synthesis of the work that I carried out as part of my end of studies project, entitled "Development of a system for assigning products to customers for the benefit of a bank". This system intervenes in the anticipation of the behavior of a bank's customers, using Machine Learning techniques such as the K-means model, used for the distribution of similar customers into groups, in order to propose a recommendation of products as well as a new segmentation in order to improve the rate of customer equipment in the bank.

Keywords: Customer segmentation, Predictive marketing, Data visualization, Machine Learning, Kmeans

Table des matières

Introduction Générale	1
Chapitre 1 : Etude préliminaire.....	2
Introduction	2
1. Présentation de l'organisme d'accueil	2
1.1. EY Global	2
1.2. EY Tunisie.....	5
2. Cadre du projet	6
2.1. Marketing prédictif.....	6
2.2. Machine Learning ^[4]	7
2.3. Data visualisation	7
2.4. Développement Web	7
3. Contexte, problématique et solution proposée.....	7
3.1. Contexte et problématique.....	7
3.2. Solution retenue.....	8
4. Méthodologie de travail.....	9
4.1. Les méthodes agiles ^[5]	9
4.2. Méthodologie Scrum ^[6]	10
Conclusion.....	12
Chapitre 2 : Identification des besoins et de l'environnement de travail.....	13
Introduction	13
1. Connaissances métiers : Segmentation de la clientèle.....	13
2. Connaissances techniques.....	14
2.1. Visualisation des données.....	14
2.2. Machine Learning.....	14
2.2.1. Pré-traitement et nettoyage des données ^[7]	14
2.2.2. Les types d'apprentissage ^[8]	16
2.2.3. Les algorithmes d'apprentissage non supervisé	17
2.3. Développement web	19
3. Analyse fonctionnelle du système	19
3.1. Identification des acteurs	19
3.2. Besoins fonctionnels.....	19
3.3. Besoins non fonctionnels.....	20
4. Backlog de produit.....	21
5. Environnement de travail.....	23
6. Diagramme de Gantt.....	25
Conclusion.....	25
Chapitre 3 : Visualisation des données	26
Introduction	26
1. Extraction des données	26
a. Base de données des clients	26
b. Base de données des dépôts bancaires	27

c. Base de données des équipements clients.....	27
d. Base de données des financements	28
e. Base de données des opérations bancaires.....	28
2. Indicateurs clé de performance	29
2.1. Identification des KPIs	29
2.2. Mise en place des KPIs.....	30
3. Modélisation	32
4. Tableaux de bord	33
4.1. Vue globale.....	33
4.2. Portefeuille	34
4.3. Clients « professionnels ».....	36
4.4. Clients « particuliers ».....	37
Conclusion.....	38
Chapitre 4 : Mise en œuvre du modèle	39
Introduction	39
1. Méthodologie de travail.....	39
2. Pré-traitement et nettoyage des données.....	40
2.1. Transformation de la base de données des clients	40
2.2. Transformation de la base de données des dépôts	42
2.3. Transformation de la base de données des opérations.....	43
3. Consolidation des bases de données	44
4. Choix du modèle.....	44
4.1. Le modèle K-means.....	46
5. Application de K-means	47
6. Génération des recommandations	48
7. Résultat	51
Conclusion.....	51
Chapitre 5 : Développement de la plateforme	52
Introduction	52
1. Conception de la plateforme	52
1.1. Le langage UML.....	52
1.2. Les diagrammes utilisés	52
2. Architecture de la solution.....	54
3. Conception de la maquette.....	55
4. Développement de la plateforme	57
Conclusion.....	60
Conclusion	61
Bibliographie	62

Table des figures

Figure 1- Répartition globale de EY	3
Figure 2- Services de EY	4
Figure 3- Identité visuelle de EY	4
Figure 4- EY Tunisie en chiffres	5
Figure 5- Organigramme EY Tunisie.....	6
Figure 6- Services de l'équipe Intelligent Automation & Analytics	6
Figure 7- Méthodologie Scrum	10
Figure 8- Diagramme de Gantt.....	25
Figure 9- Base de données Clients	26
Figure 10- Base de données Dépôts	27
Figure 11- Base de données Equipement	27
Figure 12- Base de données des Financements	28
Figure 13- Base de données des Opérations.....	28
Figure 14- Calcul des indicateurs de complétude	30
Figure 15- Calcul des produits éligibles consommés	30
Figure 16- Calcul du taux d'équipement client.....	31
Figure 17- Calcul du taux d'équipement en produits éligibles	31
Figure 18- Calcul du taux d'équipement en produits non éligibles	31
Figure 19- Calcul de la moyenne d'un mouvement bancaire	31
Figure 20- Calcul de la somme des revenus.....	31
Figure 21- Calcul de la moyenne des revenus.....	31
Figure 22- Calcul de la moyenne des visites par semaine par client.....	31
Figure 23- Affectation des statuts aux clients	32
Figure 24- Modélisation des données.....	32
Figure 25- Dashboard 1_ Données collectées	33
Figure 26- Dashboard 2_ Vue globale	34
Figure 27- Dashboard 3_ Portefeuille	35
Figure 28- Dashboard 4_ Encours.....	35
Figure 29- Dashboard 5_ Professionnels 1-1	36
Figure 30- Dashboard 5_ Professionnels 1-2	36
Figure 31- Dashboard 6_ Professionnels 2-1	37
Figure 32- Dashboard 7_ Haut de gamme - 1	37
Figure 33- Dashboard 7_ Haut de gamme - 2	38
Figure 34- Dashboard 8_ Classe moyenne.....	38
Figure 35- Dashboard 9_ Grand public	38
Figure 36- Méthodologie CRISP-DM.....	39
Figure 37- Elimination des colonnes inefficaces	40
Figure 38- Traitement de la base de données des clients	41
Figure 39- Traitement de la base de données des dépôts	42
Figure 40- Traitement de la base de données des opérations	43
Figure 41- Consolidation des bases de données.....	44
Figure 42- Elbow method pour la détermination du nombre de clusters	47
Figure 43- Application de K-means	47
Figure 44- Récupération des produits des clients.....	48
Figure 45- Calcul des similarités.....	49
Figure 46- Recommandation d'une nouvelle segmentation	50
Figure 47- Recommandation de produits de la segmentation cible	51
Figure 48- Résultat du modèle	51
Figure 49- Diagramme de cas d'utilisation global.....	53

Figure 50- Diagramme de séquences	54
Figure 51- Architecture de l'application	55
Figure 52- Maquette des pages Accueil et Authentification	56
Figure 53- Maquette des pages d'Upload	56
Figure 54- Maquette de la page de résultat	56
Figure 55- Page d'accueil	57
Figure 56- Page d'authentification.....	57
Figure 57- Page de chargement de la base de données des clients.....	58
Figure 58- Etape de chargement de la base de données des clients	58
Figure 59- Page de chargement des bases de données des dépôts et des opérations.....	58
Figure 60- Page de résultat du modèle - 1	59
Figure 61- Page de résultat du modèle - 2	60

Table des tableaux

Tableau 1- Terminologies de Scrum	11
Tableau 2- Rôles et responsabilités dans Scrum	12
Tableau 3- Backlog de produit	22
Tableau 4- Technologies utilisées pour la création des tableaux de bord	23
Tableau 5- Technologies utilisées pour le développement du modèle.....	24
Tableau 6- Technologies utilisées pour le développement de la plateforme.....	24
Tableau 7- Indicateurs clé de performance	30
Tableau 8- Comparaison des modèles.....	45

Introduction Générale

En vue de l'énorme explosion des données à laquelle nous avons assisté durant cette dernière décennie, Internet est devenu l'outil d'information et de communication en évolution, avec des perspectives de croissance exceptionnelles. C'est devenu un grand moyen de communication, de commerce et même d'analyse dans plusieurs domaines.

C'est de cette explosion qu'apparût le contexte d'Intelligence Artificielle et plus précisément le Machine Learning, méthode d'analyse incontournable des grandes masses de données qui ne cessent d'émerger. Aucun organisme ne se voit exprimer son indifférence quant à ces changements majeurs, et toute entreprise souhaitant s'aventurer dans un marché concurrentiel en pleine évolution se voit obligée d'avoir recourt à ces techniques afin de fixer des objectifs stratégiques, d'améliorer sa rentabilité, et de garder une longueur d'avance sur ses concurrents.

Avec tous ces développements, il est beaucoup plus commode de travailler avec des documents numérisés, parfaitement analysés, tout en respectant tous les détails, dans l'objectif d'améliorer l'exploitation des données, des moyens de communication et de commerce. De là, naît le principe de marketing prédictif, qui est l'ensemble des dispositifs permettant d'anticiper les comportements des clients par des prévisions basées sur des données et des probabilités de réussite.

Au sein de EY, il est question de mettre en place un système d'affectation de produits aux clients au profit d'une banque, en analysant le comportement client et en recommandant le bon produit au bon client grâce au Machine Learning. Mon projet est de développer ce système et ce rapport le documentera.

Le présent rapport qui s'étale sur cinq chapitres, décrit en détail la progression du projet. Le premier chapitre est consacré à la présentation de l'organisme d'accueil, de la problématique, ainsi que de la solution retenue et la méthodologie adoptée. Le second présente les connaissances métiers et l'analyse des besoins et les connaissances techniques nécessaires pour la réalisation du projet. Finalement, les trois derniers chapitres sont consacrés à la réalisation pratique du projet, respectivement pour les phases de visualisation de données, de développement du modèle, ainsi que de développement de la plateforme réalisée pour la présentation de notre système.

Chapitre 1 : Etude préliminaire

Introduction

L'étude préliminaire est une étape essentielle dans tout cycle de développement de logiciels. L'objectif de cette phase est de connaître l'environnement du produit.

Ce chapitre aura pour but de poser la problématique, de présenter la solution proposée, choisir la méthodologie la plus appropriée pour le projet et identifier les besoins fonctionnels et non fonctionnels. Au cours de ces phases, nous essayons d'abord de comprendre et de décrire précisément les besoins du client. « Quelles caractéristiques ? », « Comment ? », « Quoi ? », « De quelle façon ? », « Pourquoi ? » sont les questions auxquelles cette section doit répondre.

1. Présentation de l'organisme d'accueil

EY, aussi connu sous son ancien nom *Ernst & Young et associés*, est un cabinet d'audit financier et de conseil, dont le siège se situe à Londres, en Angleterre.

1.1. EY Global

L'entreprise est reconnue comme un des plus importants réseaux à l'échelle mondiale. Avec *Deloitte*, *KPMG* et *PricewaterhouseCoopers*, elle fait partie des quatre grands cabinets comptables connus sous le nom des « *Big Four* »

1.1.1. Historique^[1]

Les racines d'*EY* remontent au début du 20^{ème} siècle, plus précisément en 1903, lorsque les deux frères Alwin C. et Theodore Ernst se joignent afin de créer leur cabinet *Ernst & Ernst* en Ohio. En 1979, le cabinet s'unit à la firme britannique *Whinney Smith & Whinney*, afin de donner naissance à un partenariat anglo-américain, nommé *Ernst & Whinney* en 1979, classée alors quatrième plus grande société comptable au monde.

Quelque part ailleurs, trois ans après la naissance de *Ernst & Ernst*, le cabinet *Arthur Young & Co.* voit le jour grâce au comptable écossais Arthur Young en 1906, et s'unit par la suite avec la firme britannique *Young à Broads Paterson & Co.*

C'est alors qu'en 1989 qu'*Ernst & Whinney* fusionne avec la cinquième plus grande entreprise au monde à l'époque, *Arthur Young & Co.*, pour donner naissance à *Ernst &*

Young.

1.1.2. Activités^[2]

EY regroupe plus de 310 000 salariés, et fournit des prestations d'audit financier, de consulting (IT, RH, organisation, finance, stratégie...). Elle fonctionne comme un réseau de cabinets membres structurés en entités juridiques distinctes en partenariat, réparties à travers plus de 700 bureaux dans 152 pays dans le monde et regroupés dans quatre zones géographiques à savoir :

- EMEA : L'Europe, le Moyen-Orient, L'Inde et L'Afrique
- L'Asie-Pacifique
- Les Amériques

La figure ci-dessous décrit la répartition mondiale de EY.

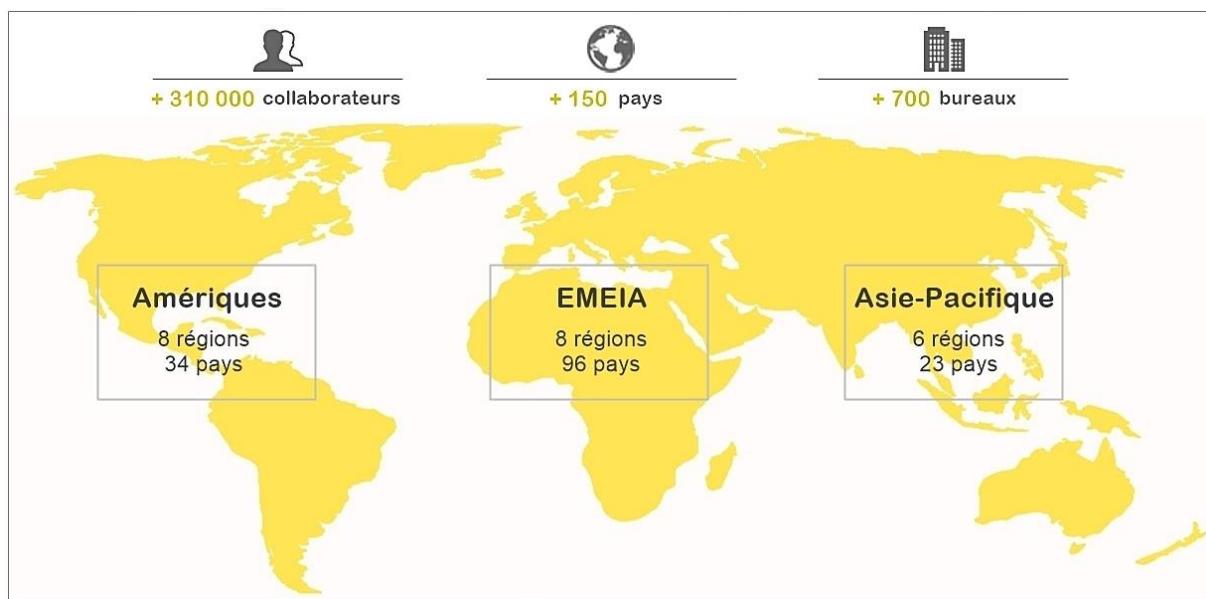


Figure 1- Répartition globale de EY

EY opère dans les domaines de :

- Audit : certification, maîtrise des risques, amélioration de la performance financière, accompagnement et externalisation (expertise-comptable),
- Consulting : marketing & innovation, performance financière, performance opérationnelle, systèmes d'information,
- Droit et fiscalité : fiscalité des entreprises, droit des affaires, droit social, mobilité internationale,
- Transactions : évaluations, fusions & acquisitions.

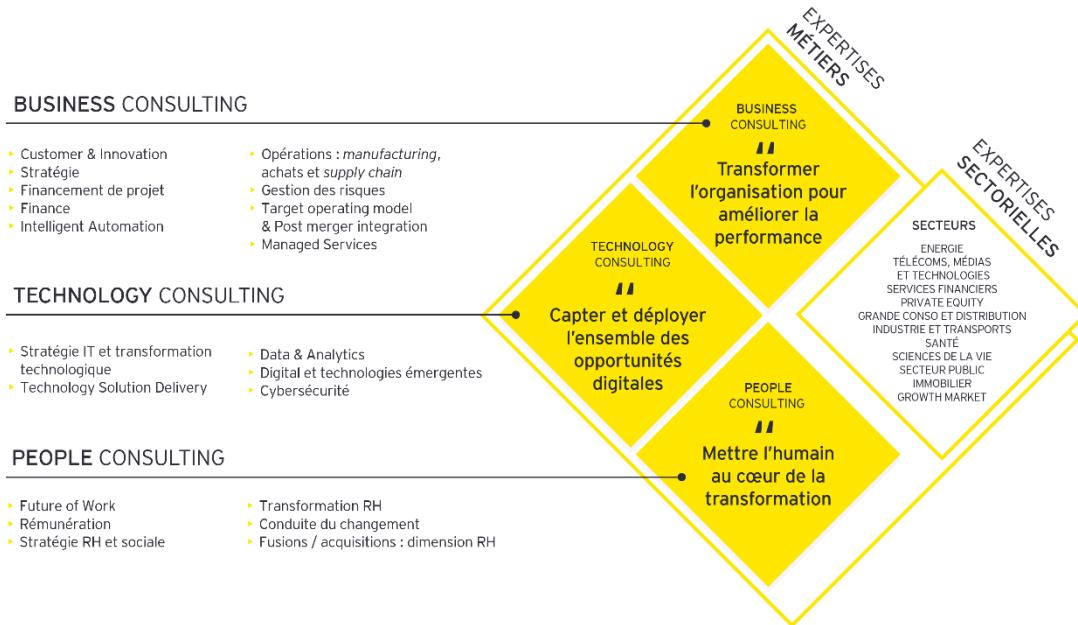


Figure 2- Services de EY

Comme bon nombre des plus grands cabinets d'experts-comptables ces dernières années, *EY* s'est étendu sur des marchés comptables adjacents, notamment la stratégie, les opérations, les ressources humaines, la technologie et les services. Et, comme tout membre des *Big Four*, l'entreprise est également active dans le secteur du conseil fiscal. Le cabinet est l'auditeur de nombreuses entreprises de premier plan, notamment AOL Time Warner, Wal-Mart, Amazon.com, 3M, Oracle, McDonalds, Google, Intel, Hewlett-Packard, Coca-Cola, et Verizon.

Depuis 2019, *EY* devint la septième plus grande institution privée aux États-Unis, et figure sur la liste Fortune des meilleures entreprises où travailler depuis 21 ans, plus longtemps que tout autre cabinet comptable.

1.1.3. Identité visuelle^[3]

Il s'appelait *Ernst & Young* jusqu'à ce que cette campagne de rebranding change officiellement son nom en *EY* 2013, bien que ce sigle, qui était auparavant utilisé de manière non officielle, ait été adopté.

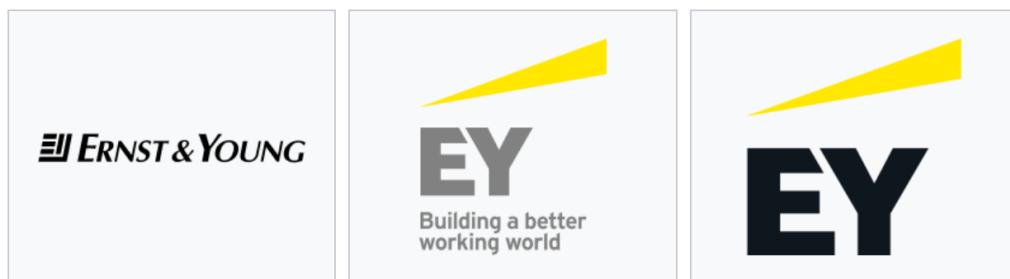


Figure 3- Identité visuelle de EY

1.2. EY Tunisie

Avec une présence locale depuis 1987, AMC EY Tunisie est l'un des principaux cabinets de conseil en Tunisie. Ses relations avec le secteur public (ministères, banques, etc.) fait de lui le conseiller dévoué, en plus de son accompagnement aux investisseurs étrangers dans leurs implantations en Tunisie et son assistance aux entreprises tunisiennes dans leurs projets d'internationalisation, notamment au Maghreb et en Afrique.

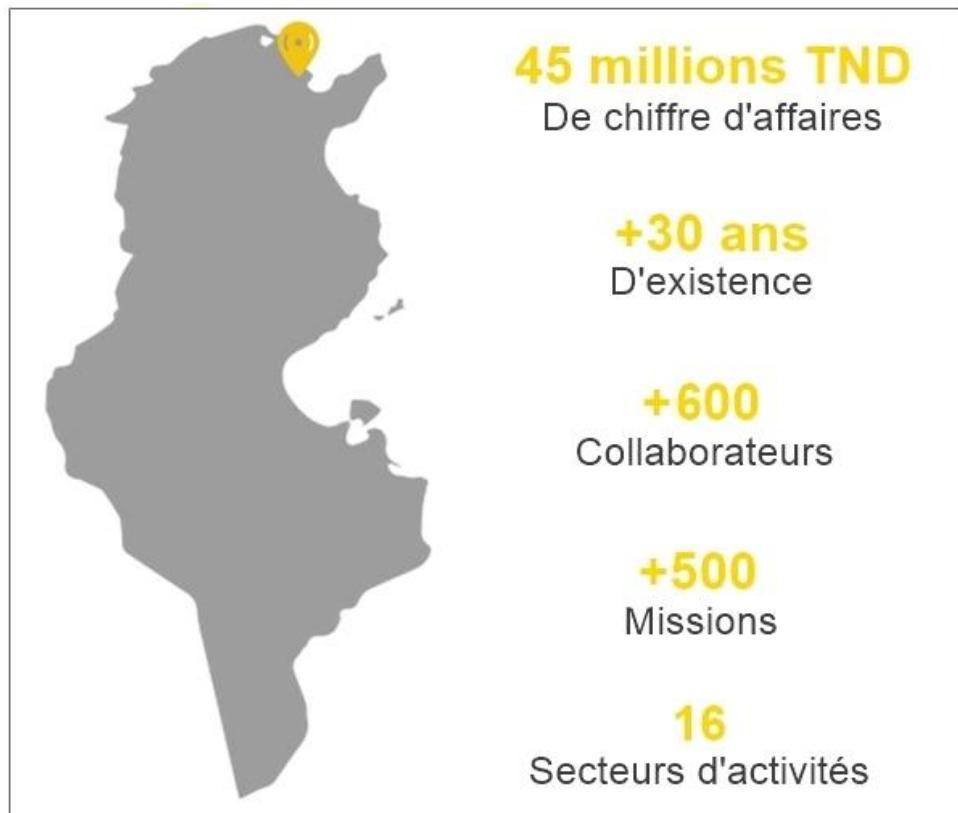


Figure 4- EY Tunisie en chiffres

Avec plus de 30 ans d'expérience sur le marché local, EY Tunisie rassemble plus de 500 collaborateurs travaillant dans le domaine de l'audit, du conseil (financiers, stratégiques et technologique), de la fiscalité, du droit et des transactions. EY possède donc les connaissances et l'expertise nécessaires pour guider ses clients dans leurs :

- Conception et mise en œuvre de la stratégie
- Amélioration de la performance
- Evaluation des risques
- Transformation IT
- Information Technology
- Services financiers
- Conduite du changement
- Réformes Publiques

La figure suivante illustre l'organigramme de EY Tunisie.

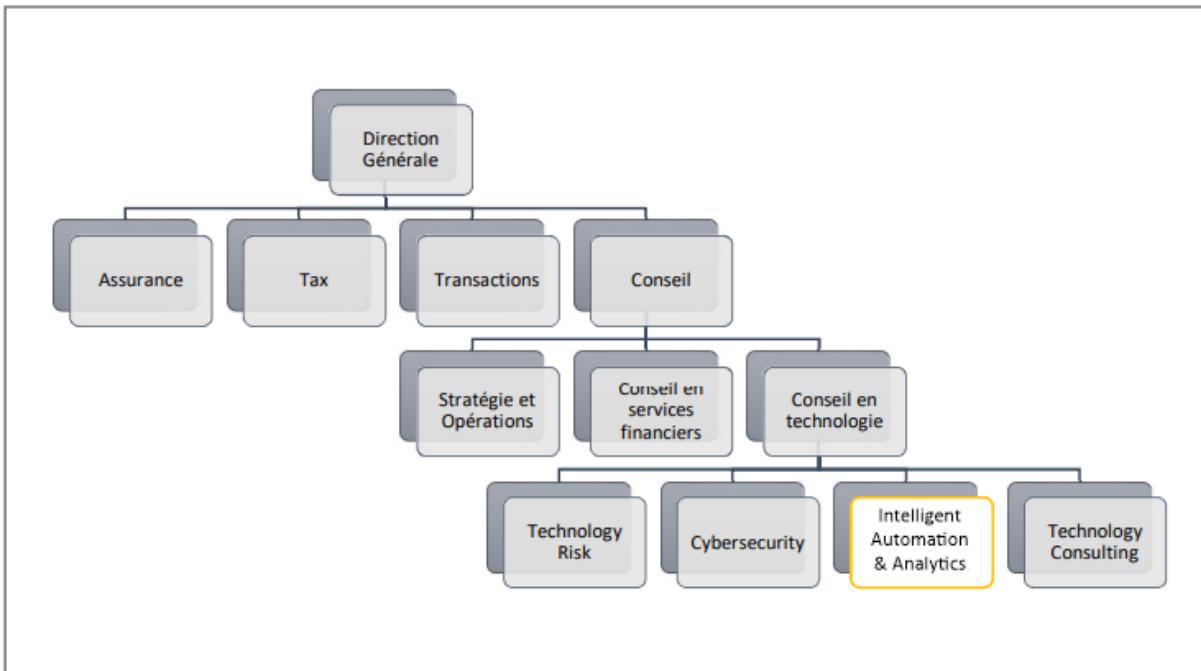


Figure 5- Organigramme EY Tunisie

Le département « *Intelligent Automation & Analytics* » intervient dans l'accompagnement des clients à digitaliser les processus au sein de leur organisation pour développer, protéger et optimiser leur entreprise en exploitant les dernières technologies. L'équipe opère dans les domaines suivants :



Figure 6- Services de l'équipe Intelligent Automation & Analytics

2. Cadre du projet

Dans cette partie, il est question de présenter les mots clés intervenant dans la définition du projet afin de le placer dans son contexte et de mieux comprendre les parties à venir.

2.1. Marketing prédictif

Le marketing prédictif est la méthode d'anticipation des comportements de consommations d'une clientèle potentielle ou déjà existante, basée sur l'analyse des données, permettant aux organismes de prendre en mesure des stratégies et actions marketing adaptées. Cette technique a donc pour objectif de collecter les données relatives aux consommateurs et prédire leurs intentions en termes d'achat afin de leur proposer une offre adéquate.

2.2. Machine Learning^[4]

Etant encore un nouveau terme pour de nombreuses personnes, le Machine Learning est une science moderne permettant de découvrir des répétitions (appelés *patterns*) dans un flux de données et d'en tirer des prédictions basées sur des statistiques. Il permet donc d'accélérer le processus d'analyse et le rendre plus précis tout en s'améliorant de manière autonome au fil du temps.

Réaliser ces prédictions permet aux entreprises de prendre de meilleures décisions stratégiques, en respectant les besoins des consommateurs et les futures tendances.

2.3. Data visualisation

« Une image vaut mille mots. » – Confucius.

Cette phrase est aujourd'hui plus réelle que jamais étant donné les montagnes de données collectées durant les dernières années. Avec les heures à passer à essayer de les comprendre et les déchiffrer, elles sont plus que jamais complexes à analyser, expliquer et partager sous leur forme brute. En transformant les informations brutes en objets visuels clairs et compréhensibles : points, barres, courbes, cartographies, la data visualisation offre un gain de temps conséquent dans la recherche et l'analyse.

2.4. Développement Web

Le développement Web va de la création de pages en texte brut à des applications Web complexes, en fonction des exigences du client. Ce processus comprend l'analyse des besoins, la conception, le développement du contenu ainsi que l'élaboration de scripts côté client ou côté serveur et la configuration de la sécurité du réseau. Il s'agit de l'étape incontournable pour le déploiement des tableaux de bord relatifs à la data visualisation, ainsi que l'algorithme de Machine Learning mis en place.

3. Contexte, problématique et solution proposée

Aujourd'hui, on parle beaucoup d'analyse décisionnelle, en raison de la naissance de nouvelles formes d'organisations qui s'appuient sur la numérisation et l'analyse de l'information. Dans le cadre de ma formation pédagogique, j'ai étudié les initiations au Machine Learning, et eu l'occasion de mettre en pratique mes connaissances théoriques en développant un système d'affectation de produits aux clients au profit d'une banque dans le cadre de mon projet de fin d'études.

3.1. Contexte et problématique

À mesure de l'augmentation de la quantité de données de façon exponentielle, celles-ci deviennent le nouvel or noir et jouent un rôle de plus en plus important dans l'efficacité des activités de vente et de marketing. Anticiper les besoins et les comportements des internautes

devient essentiel afin d'accélérer la conversion des consommateurs et garantir la croissance de l'entreprise.

Dans notre cas, nous nous intéressons plus particulièrement aux banques et à leur manière de segmenter leur clientèle et leur recommander les produits adéquats. C'est à ce moment qu'apparaît la complexité : les commerciaux n'ont ni le temps ni la capacité mentale de filtrer des millions d'articles et de rapports chaque jour, créant une énorme surcharge d'informations.

3.2. Solution retenue

Aujourd'hui, la connaissance et l'analyse en temps réel du passé et du présent est devenue un facteur clé de succès pour les entreprises évoluant dans un environnement compétitif. L'évolution du domaine des données, les progrès technologiques, et la forte demande en services commerciaux automatisés obligent les entreprises à s'adapter à la transformation digitale pour survivre et profiter des actualités.

Dans cet environnement, *EY Tunisie*, figurant parmi les leaders dans le domaine du conseil à l'échelle nationale et internationale, propose des solutions de prédiction dans ce domaine, afin d'aider dans l'analyse d'informations et mettre en place des recommandations personnalisées de produits et de services. Parmi ces solutions, et afin de répondre au besoin des banques qui cherchent à atteindre une certaine évolution digitale afin de fidéliser leur clientèle, nous intervenons dans la mise en place d'un système d'affectation de produits à la clientèle d'une banque.

L'une des causes qui nous poussent à développer cette solution est la volonté de changer ou d'améliorer l'existant. Les acteurs responsables d'analyser les comportements font souvent face à des contraintes étant donné l'importance et la délicatesse de la tâche. Par conséquent, la méthode classique prend beaucoup de temps et beaucoup plus de risques d'erreurs.

L'objectif de cette solution consiste à assimiler tout un process qui souscrit à l'entreprise de suivre de près le comportement de ses clients et ainsi recueillir, analyser et interpréter leurs données afin d'accentuer leur taux d'équipement.

Le système prend comme entrée les données des clients, de leurs comptes et de leurs opérations, les traitent afin de les regrouper par similarités, ensuite recommande des produits qui leur sont adéquats, une nouvelle segmentation, ainsi que le taux d'équipement que chaque client devrait atteindre.

Nous proposons également un tableau de bord résumant les données de la clientèle, la segmentation, les chiffres d'affaires et les indicateurs clé de performance les plus pertinents, et pour finir une plateforme web dans le rôle du regroupement et d'affichage des solutions mentionnées plus haut.

4. Méthodologie de travail

Afin de satisfaire les besoins du client et de l'administrateur, garantir une bonne qualité et éviter tout retard en termes de délais, il est nécessaire d'adopter une méthodologie de développement qui définit les règles de conduite, les rôles des différents acteurs, la chaîne d'actions, etc.

4.1. Les méthodes agiles^[5]

Ces dernières années, le mot « agilité » semble avoir envahi tous les projets : les méthodes agiles ont éclipsé toute autre forme de gestion de projet. Ce qui distingue ces méthodes de celles traditionnelles, est le fait qu'elles partent d'un principe itératif, répondant aux besoins évolutifs des clients.

Appelées également méthodes productives, leur principe est simple : planifier des objectifs à court terme, se lancer sur la route sans tarder, sans spécifier l'ensemble du produit et rentrer dans ses détails, afin d'assurer une flexibilité au niveau de la réalisation. Une fois cet objectif atteint, l'adaptation de l'itinéraire du deuxième se fait en fonction de la situation du moment.

Les méthodes agiles utilisent un cycle de développement qui tourne autour du client. Ce dernier est donc impliqué dans la réalisation du début à la fin du projet pour permettre à l'équipe d'obtenir un feedback régulier afin d'appliquer au fur et à mesure les changements nécessaires.

Les méthodes agiles se basent donc sur quatre valeurs fondamentales :

- **Les individus et leurs interactions** plutôt que les processus et les outils : l'équipe est bien plus importante que le matériel et les procédures. Ces méthodes favorisent le travail en équipe en privilégiant le face à face. Même si le télétravail a pris de l'ampleur ces dernières années, cette étape ne cesse de faire ses preuves.
- **Un logiciel opérationnel** plutôt qu'une documentation complète : il est vital que l'application fonctionne. Il est préférable de commenter le code et de transférer les compétences au sein de l'équipe mais cela reste secondaire.
- **La collaboration avec le client** plutôt que la contractualisation : le client doit impérativement être impliqué dans le développement pour fournir un feedback continu sur l'adaptation du système à ses attentes, afin de garantir un taux d'échec minimal, la satisfaction du client étant l'un des éléments clés du développement.
- **L'adaptation au changement** plutôt que le suivi d'un plan : la planification initiale doit être flexible aux demandes du client tout au long du projet, le but étant la revue rapide, si nécessaire, de la stratégie sans perturber le déroulement du projet.

Les méthodes agiles sont donc des méthodes adaptées pour les projets complexes qui durent longtemps et qui nécessitent que l'équipe de développement reste constamment en interaction directe avec le client.

Dans notre cas, les revues du système devront être effectuées à la fin de chaque phase, allant de l'observation des indicateurs clé de performance au sein des tableaux de bord, jusqu'à la vérification de proposition d'une nouvelle segmentation des clients ainsi que leur éligibilité quant aux produits recommandés. De ce fait, nous serons constamment en relation avec le client, toujours prêts à s'adapter aux changements que ce dernier proposera, et en priorisant le côté opérationnel de l'application. Nous opterons donc pour la méthode agile Scrum pour le développement de notre projet.

4.2. Méthodologie Scrum [6]

De nos jours, Scrum apparaît sur toutes les bouches, c'est la méthodologie de travail utilisée au sein de presque toutes les entreprises, faisant du client le personnage principal, à l'instar des autres méthodes agiles. « Scrum » est un terme anglais, signifiant « mêlée », inspiré du rugby, un sport qui requiert une équipe solidaire et avançant dans une même direction. Cette « mêlée » se traduit alors par un « sprint », au terme duquel, une « revue de sprint » est organisée afin de faire le point sur l'état de l'avancement du projet, la vérification des fonctionnalités finalisées ainsi que d'éventuels changements et adaptations, avant de passer à l'identification des objectifs du sprint suivant.

Cette méthode encourage les membres de l'équipe à s'auto-organiser et se montrer soudés lors de la résolution de problèmes, mais également à interagir au fur et à mesure avec le client dans le but de proposer les améliorations nécessaires, rendant ainsi le processus nettement plus productif. La figure suivante illustre le cycle à suivre dans le cadre d'un projet adoptant la méthodologie Scrum.

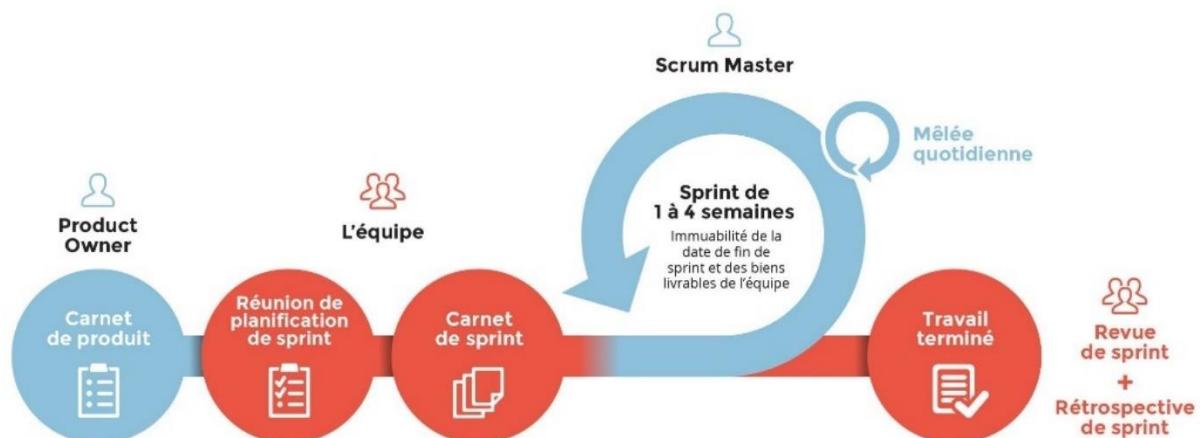


Figure 7- Méthodologie Scrum

4.2.1. Terminologies

Afin d'assurer une bonne application de cette méthode, il faudra initialement procéder à l'explication des termes et mots-clés relatifs à Scrum, présentés au sein du tableau 1 suivant :

▪ Sprint	Ce terme désignant une phase de développement d'une durée d'une à quatre semaines, il vise à concentrer l'équipe sur un seul objectif, celui de clôturer une brique du projet.
▪ Product Backlog	Se résumant comme la « to-do list » de l'équipe, il représente la liste des fonctionnalités, exigences et améliorations. Géré par le Product Owner, le Backlog Produit est constamment remis en question et ses priorités sont redéfinies.
▪ User Story	En mettant les utilisateurs finaux au centre de la discussion, les stories décrivent leur point de vue tout en utilisant un langage non technique afin de simplifier le travail.
▪ Daily Scrum	Appelée aussi « stand-up quotidien », elle représente une mini-réunion quotidienne de 15 minutes, dans le but de réunir les membres de l'équipe pour que chacun réponde aux questions « qu'est-ce que j'ai fait hier ? », « qu'est-ce que je prévois de faire aujourd'hui » et « ai-je rencontré des obstacles ? ».
▪ Revue de Sprint	A la fin de chaque sprint, l'équipe se rassemble afin d'assister à une première présentation de la partie finalisée du produit dans le but de la collecte d'avis et de l'apport de modifications si nécessaire.
▪ Backlog de sprint	Il représente un ensemble de user stories prises en charge par l'équipe de développement pendant le sprint en question.

Tableau 1- Terminologies de Scrum

4.2.2. Rôles et responsabilités

Comme tout autre méthodologie, Scrum identifie plusieurs rôles au sein de l'équipe apportant chacun leur expertise, et présentés au sein du tableau 2 suivant :

▪ Product Owner	Également appelé chef de projet digital, celui-ci est le garant de l'atout essentiel du projet : la satisfaction client. Ses principales fonctions sont la définition de la vision du produit mais aussi agir en tant qu'intermédiaire entre la partie métier et la partie technique du projet, tout en priorisant la qualité du développement des fonctionnalités prédéfinies.
▪ Scrum Master	Endossant le rôle de coach de l'équipe, il n'est pas à confondre avec le chef de projet. Celui-ci joue l'animateur pour les <i>daily meetings</i> (réunions quotidiennes de 15 minutes) en facilitant la communication en interne afin de faire avancer efficacement le projet et essayer de tirer le meilleur de chaque membre de l'équipe.

▪ Développeurs	N'ayant pas de hiérarchie en interne, ce groupe de personnes est chargé de la mise en pratique des objectifs tout en respectant les spécifications techniques prédéfinies afin de préparer une version livrable du projet.
----------------	--

Tableau 2- Rôles et responsabilités dans Scrum

Conclusion

L'étape de l'étude préalable est nécessaire pour chaque projet, cela permet de mettre dans le cadre l'architecture du système et les besoins requis. Au sein de ce premier chapitre, nous avons présenté l'organisme d'accueil. Ensuite, nous avons détaillé le cadre et le contexte du projet afin d'assurer une bonne compréhension de ce dernier, en énumérant les différents domaines d'activités, la problématique, ainsi que la solution retenue. Pour finir, nous avons présenté la méthodologie de travail et le planning d'exécution. Le chapitre suivant sera consacré aux connaissances métiers et à l'analyse des besoins techniques.

Chapitre 2 : Identification des besoins et de l'environnement de travail

Introduction

Avant d'entamer la première phase de notre projet, nous devons avoir une idée très claire des fonctionnalités demandées. De ce fait, il est très important de consacrer une partie à la compréhension du projet d'un point de vue commercial, mais aussi d'un point de vue technique. Au sein de ce chapitre, nous visons à bien comprendre et à décrire précisément les besoins du client. « Quelles fonctionnalités ? », « Comment ? », « Par quel moyen ? », « Pour quel usage ? » sont les questions auxquelles cette partie doit répondre.

1. Connaissances métiers : Segmentation de la clientèle

Le préalable incontournable à la conception fonctionnelle d'un logiciel est l'identification des besoins métiers auxquels ce dernier répond. Il s'agit de très bien comprendre l'activité et l'environnement de l'organisation à laquelle il est destiné.

La segmentation de la clientèle est une étape essentielle pour garantir l'efficacité de l'activité commerciale de la banque. Dans la plupart des cas, la répartition se fait premièrement en différenciant les types de clients : Professionnel, Particulier ou Entreprise. Un « particulier » est une personne physique indépendante, tandis qu'un « professionnel » est soit une personne physique entrepreneur individuelle soit une personne morale (une entreprise réalisant un chiffre d'affaires en-dessous d'un seuil déterminé). Dans notre cas, nous nous intéressons à ces deux-là.

Au sein de la banque, la segmentation se fait selon un calcul bien précis des caractéristiques du client : son âge, sa profession, son flux créiteur, l'historique de ses opérations bancaires, etc. Les clients sont donc regroupés par catégories : Grand Public, Commerciaux, Agriculteurs, Prestige, Associations, etc. De ce fait, nous nous retrouvons avec des groupes parfaitement homogènes, au sein duquel les comportements des clients se ressemblent. Une segmentation peut être qualifiée de fiable lorsqu'elle tient compte de la spécificité de chaque segment, incluant les produits et services répondant aux besoins de la clientèle de chacun. En développant un système qui garantit ces critères, nous pouvons garantir la pertinence du processus commercial au sein de la banque.

2. Connaissances techniques

Cette partie sera consacrée à présentations des différents aspects techniques de notre projet.

2.1. Visualisation des données

La visualisation de données est en plein essor. Elle évolue pour s'adapter au développement des mégadonnées, de sorte qu'il ne disparaît jamais. Si la visualisation des données était importante il y a quelques années, elle est très importante aujourd'hui. À l'ère du big data, les milliards de données que les entreprises peuvent collecter chaque jour sont affichées sur des lignes distinctes avant ce processus de conversion et ne sont pas immédiatement disponibles.

Avec une représentation simple et facile à comprendre des données, la visualisation des données vous permet de visualiser les tendances, les phénomènes, les connexions et de les utiliser de manière stratégique. En un sens, il raconte une histoire qui devrait être intégrée au plan d'action. En classant visuellement, en segmentant et en scénarisant les données, les organisations peuvent découvrir en un coup d'œil des informations auparavant inaccessibles. Par conséquent, la visualisation des données permet à toute entreprise de gérer ses activités plus efficacement en adoptant une stratégie agile axée sur les données.

Par conséquent, la visualisation de données est une représentation visuelle des données afin qu'elles puissent être reconnues et comprises car il est difficile d'interpréter et d'utiliser les données brutes. Ce processus est réalisé à l'aide d'un outil d'analyse dédié et prend la forme d'une infographie regroupée dans un tableau, un graphique, une carte visuelle ou un tableau de bord.

2.2. Machine Learning

L'univers de l'intelligence artificielle est un univers de nomenclature technique et d'anglicisme. Parfois, 2 termes pourtant très voisins peuvent signifier des termes totalement opposés. Pour cela, il est nécessaire d'étudier de près le problème à traiter.

2.2.1. Pré-traitement et nettoyage des données^[7]

L'augmentation de la collecte de données et son traitement systématique ont permis de développer des techniques d'apprentissage automatique qui nécessitent de grandes quantités de données pour s'exécuter et s'entraîner. On peut simplement penser qu'une grande quantité de données est suffisante pour que l'algorithme réussisse, mais dans la plupart des cas, les données ne sont pas adaptées et nécessitent un prétraitement avant de pouvoir être utilisées. Il s'agit d'une étape de prétraitement.

Le prétraitement des données est une technique d'exploration de données utilisée pour transformer des données brutes en formats utiles et efficaces. Les données peuvent contenir

de nombreuses parties non pertinentes et manquantes. Un nettoyage des données est effectué pour gérer cette partie. Cela inclut le traitement des données manquantes, des données bruyantes, etc. En effet, les erreurs de collecte liées à des erreurs humaines ou techniques peuvent corrompre le jeu de données et affecter la formation. Parmi ces erreurs, on peut citer des informations incomplètes, des valeurs manquantes ou incorrectes, ou encore des bruits parasites liés à l'acquisition des données. Par conséquent, il est souvent essentiel de passer par l'étape de prétraitement des données à partir de données brutes pour obtenir des données exploitables qui fournissent un modèle plus efficace.

Le nettoyage des données améliore l'intégrité et la pertinence des données en réduisant les incohérences, en prévenant les erreurs et en permettant une prise de décision plus éclairée et plus précise. Il s'agit d'un processus qui vise à identifier et corriger les données corrompues, inexactes ou non pertinentes. Cette étape de base du traitement des données améliore la cohérence, la fiabilité et la valeur des données. Les causes les plus courantes d'inexactitude des données sont les valeurs manquantes, les entrées qui n'apparaissent pas aux bons endroits et les fautes de frappe. De ce fait, le nettoyage des données renforce l'intégrité et la pertinence de nos données en réduisant les incohérences, en évitant les erreurs et en permettant de prendre des décisions mieux avisées et plus précises.

Ensuite, l'étape de transformation des données résume les modifications apportées à la structure réelle des données. Ces transformations concernent la définition mathématique de l'algorithme et la façon dont les données sont traitées pour optimiser les performances. Parmi ces techniques, nous pouvons citer le lissage des données si elles sont bruitées, la discréétisation des variables continues (par division en intervalles), mais aussi la normalisation et la standardisation des données qui mettent à l'échelle les données numériques à une échelle plus petite afin de permettre une centralisation de la moyenne et une réduction de la variance.

Pour donner suite à l'étape de nettoyage des données, nous passons à celle de la consolidation des bases de données. Cette étape consiste à regrouper des sources multiples en une seule base de données. Elle est effectuée dans un cadre de gestion des données pour la création de bases exploitables.

Finalement, nous arrivons à l'étape de réduction de dimensionnalité. En apprentissage automatique, la réduction de dimensionnalité consiste à passer d'un espace d'apprentissage de dimension supérieure à un espace de calcul plus petit. En d'autres termes, réduire le nombre de variables qui nous permettent d'entraîner notre modèle pour obtenir un modèle d'intelligence artificielle plus robuste et des temps de traitement plus rapides.

Lorsque les données sont affichées dans un tableau, la réduction de dimension est obtenue en réduisant le nombre de colonnes. Les modèles 3D tels que les cubes et les sphères peuvent être réduits à un seul plan, carré ou cercle. S'il y a trop de variables dans le modèle d'apprentissage automatique, il y a un risque de surapprentissage. Dans ce cas, le modèle se limite à reconnaître des exemples entraînés et ne peut pas identifier de nouveaux

exemples.

2.2.2. Les types d'apprentissage^[8]

En matière de Machine Learning, on distingue deux grandes familles d'algorithmes : les algorithmes d'apprentissage supervisé, et les algorithmes d'apprentissage non supervisé, selon la labellisation ou non des données.

a. L'apprentissage supervisé

Reconnu comme étant le type d'apprentissage le plus utilisé ces dernières années, l'apprentissage supervisé consiste en l'entraînement d'un modèle en utilisant des données étiquetées. En d'autres termes, le modèle disposera d'exemples à suivre en s'entraînant, afin de lui permettre de prédire par la suite le label de nouvelles données non étiquetées. Au fur et à mesure de l'enrichissement du modèle, le résultat gagne en précision et pertinence, réduisant la marge d'erreur.

Il existe deux types d'algorithmes dans le cas de l'apprentissage supervisé : les algorithmes de classification et les algorithmes de régression. Les algorithmes de classification servent à prédire la famille de catégorie à laquelle appartiennent des données de test à la suite de l'étude de caractères spécifiques (exemple : classification de spams dans une boîte mail). Les classificateurs linéaires, les machines à vecteurs de support, les arbres de décision et les forêts d'arbres décisionnels sont tous des types courants d'algorithmes de classification. D'autre part, les algorithmes de régression sont utilisés pour la même cause que les algorithmes de classification, à la différence que ceux-ci sont utilisés pour prédire des valeurs continues (des nombres) et non des variables discrètes (des catégories). Les algorithmes de régression sont par exemple la régression linéaire, la régression logistique et la régression polynomiale.

Les algorithmes d'apprentissage supervisé ont alors beau être plus utilisés en raison de leur simplicité, ils font face à quelques limites qui les rendent de moins en moins populaires. Parmi ces limites, on peut citer les difficultés d'étiquetage des données lorsqu'elles sont en grande quantité, les problèmes de sur-apprentissage dans le cas où le modèle rencontre des données anormales, etc.

Pour résumer, les algorithmes d'apprentissage supervisé ne sont pas parfaits mais reste l'un des meilleurs moyens de résoudre des problématiques complexes dans divers domaines de la finance à la santé.

b. L'apprentissage non supervisé

Dans la plupart des cas, les modèles d'apprentissage non supervisé permettent de répondre à des problématiques complexes. À la différence de l'apprentissage supervisé, l'apprentissage non supervisé est celui où l'algorithme doit opérer à partir d'exemples non étiquetés. Dans ce cadre, la machine ne dispose pas d'exemples de résultat, les données

collectées sont traitées comme des variables aléatoires et le modèle découvre alors par lui-même la structure et les tendances disponibles au sein de la base de données, devenant ainsi complètement indépendante et ne nécessitant donc presque plus d'intervention humaine. Les différents faits énumérés font qu'il est impossible pour la machine de calculer ses scores de réussite. De ce fait, cette technique est utilisée dans le but d'effectuer des regroupements de données en fonction de leur ressemblance, ou dans certains cas en fonction de leur différence.

Ce type d'apprentissage est utilisé dans le partitionnement des données (data clustering), dans le développement des moteurs de recommandations et des systèmes de suggestions, en se basant sur les données d'un groupe de personnes, leur comportement et leur ressemblance comparés aux autres individus.

L'utilisation de l'apprentissage non supervisé peut être réunie en problèmes de « clustering » et « d'association ». Dans le cas du clustering, l'utilisateur attend de la machine qu'elle rassemble sous forme de clusters (groupes) des données de la manière la plus efficace selon la problématique. D'autre part, le système d'association permet de trier et de regrouper des données partageant certaines caractéristiques (reconnaissance visuelle ou vocale par exemple).

→ Dans notre cas, étant donné que nous ne disposons pas des « réponses correctes », et que nous cherchons à recommander des produits à partir d'une analyse comportementale, nous nous orientons vers les algorithmes d'apprentissage non supervisé, qui s'avèrent être la solution idéale à notre problématique. En fournissant au modèle l'ensemble des données des clients, leurs informations personnelles, leur historique d'achat ainsi que leurs revenus, nous donnons à la machine la liberté de classer ces données en fonction de différentes caractéristiques afin de proposer elle-même la liste des produits que chaque individu serait susceptible de consommer et le segment qui lui conviendrait le plus.

2.2.3. Les algorithmes d'apprentissage non supervisé

Parmi les méthodes d'apprentissage non supervisé auxquelles on s'intéresse, on peut citer la clustering basé sur le partitionnement, le clustering basé sur la hiérarchie et le clustering basé sur la densité. La sélection d'un algorithme de clustering approprié aux données est souvent difficile en raison du nombre de choix disponibles. Certains facteurs importants incluent les caractéristiques des clusters et de l'ensemble de données, le nombre de valeurs aberrantes et le nombre d'objets de données.

a. Clustering basé sur la hiérarchie

Cette catégorie consiste à former pas à pas des connexions entre des individus, et utilisent une matrice de distances entre individus pour retrouver le regroupement le plus proche d'un autre. A l'instar du clustering partitionné, le nombre de clusters (k) est souvent prédéterminé par l'utilisateur. Et contrairement à de nombreuses techniques de clustering

partitionné, le clustering hiérarchique est un processus déterministe : les affectations de cluster ne changeront pas lors de l'exécution d'un algorithme deux fois sur les mêmes données d'entrée. Cette méthode de clustering se charge de partitionner la base de données hiérarchiquement par une approche ascendante ou descendante :

- Le clustering agglomératif (approche ascendante) : fusionnement des deux points les plus similaires jusqu'à ce que tous les points aient été fusionnés en un seul cluster.
- Le clustering par division (approche descendante) : Initialisation de tous les points comme un seul cluster puis division des moins similaires jusqu'à ce qu'il ne reste que des points de données uniques.

Ces méthodes se terminent par la fourniture d'un dendrogramme (arborescence de points) servant à l'interprétation du nombre de clusters. Néanmoins, celles-ci sont coûteuses en calcul et sensibles au bruit et aux valeurs aberrantes.

b. Clustering basé sur la densité

En fonction de la densité des points de données dans une région, cette catégorie de clustering détermine les affectations de cluster, séparés par des régions à faible densité. Contrairement aux autres catégories, cette approche ne nécessite pas que l'utilisateur spécifie le nombre de clusters. Au lieu de cela, on opère suivant la densité, où celle-ci est relativement élevée : les zones où les individus sont plus proches les uns des autres. En plus de former des classes d'individus, l'algorithme repère par la même occasion les valeurs hors du commun (bruits).

Il prend 2 paramètres en entrée : la distance ϵ (la distance maximale qui peut définir 2 individus comme voisins) et n (le nombre d'individus minimal nécessaire pour former un groupe). Parmi les avantages de cette méthode, nous pouvons déclarer que cet algorithme dispose d'une excellente identification de clusters de formes non sphériques et est résistants aux valeurs aberrantes. D'autre part, on peut constater des difficultés à identifier des clusters de densités variables et dans des espaces de grande dimension.

c. Clustering basé sur le partitionnement

Cette technique de classification non supervisée est une collection d'algorithmes d'apprentissage visant à regrouper des données non étiquetées ayant des propriétés similaires. Elle consiste à diviser les objets de données en groupes non chevauchés : aucun objet ne peut être membre de plus d'un cluster, et chaque cluster doit avoir au moins un objet, où il est demandé à l'utilisateur de spécifier le nombre k de clusters. Par conséquent, pour mettre en œuvre une telle approche, nous avons besoin d'un moyen d'observer et de mesurer les différences entre ces données, et à partir de cette analyse, nous pouvons proposer de regrouper les données en plusieurs classes distinctes et cohérentes. Ce type de clustering est utilisé notamment lorsqu'il est coûteux d'étiqueter les données. Il est évolutif en ce qui concerne la complexité de l'algorithme. Néanmoins, il n'est pas adapté aux clusters de formes complexes et de tailles différentes.

2.3. Développement web

A l'heure où le digital est devenu omniprésent, le secteur est en pleine explosion et les entreprises courtisent massivement ces nouveaux profils. Le développement web est apparu avec l'avènement d'internet. Dans un secteur digital en pleine expansion, c'est une compétence très recherchée par les entreprises actuellement. La mission du développeur est d'être capable de créer et de développer un site web ou une application mobile. C'est lui qui, à partir d'une maquette, réalise toute la partie technique des différentes pages d'un site internet. Il développe la face visible d'un site, que l'on appelle Front End. Mais également l'interface utilisateur, que l'on nomme le Back Office. La programmation web comme le développement mobile passe par la maîtrise de langages informatiques spécifiques comme le HTML, le CSS, le PHP ou le Javascript. Les métiers du développement web sont en perpétuelle évolution. Ils demandent une maîtrise parfaite d'internet et de réelles compétences en informatique. Patience, curiosité, logique et organisation sont les qualités requises pour s'engager dans ce secteur.

3. Analyse fonctionnelle du système

Pour réussir à concevoir et à mettre en œuvre ce système de recommandation, il est essentiel de recueillir les informations nécessaires au développement afin de parvenir à un système opérationnel et évolutif en même temps, qui offre les informations nécessaires en temps réel et surtout qui peuvent répondre aux exigences de tous les décideurs.

3.1. Identification des acteurs

Un acteur représente une entité externe qui interagit directement ou indirectement avec le système étudié. Notre projet consiste en le développement d'une plateforme de recommandation de produits, qui donne l'accès uniquement à l'administrateur. Il va donc interagir avec un seul acteur.

3.2. Besoins fonctionnels

Tout au long de la phase de compréhension, nous en sommes venus à la conclusion que le projet est fondé sur un objectif principal : l'amélioration du taux d'équipement par la recommandation des produits adéquats.

a. Première phase du projet

Les tableaux de bord destiné à la visualisation des données fournies devront répondre aux besoins suivants :

- Un résumé des données collectées et leur taux de complétude,
- Indicateurs de répartition des clients par type,
- Une vue globale sur les taux d'équipement actuels et l'éligibilité des clients par rapport aux produits consommés,

- Un bref résumé de la catégorie de clientèle (âge, profession, revenus, etc.),
- Une vue globale de ces mêmes indicateurs par segment de clientèle.
- Les revenus, le chiffre d'affaires confié et la moyenne de visite de chaque segment de clientèle.

Tout cela en ayant la possibilité de trier les vues selon plusieurs filtres bien définis : segments, types, catégories, et année ; et en garantissant une utilisation simple et facile à manipuler avec un minimum de clics possible.

b. Deuxième phase du projet

Le système de recommandation des produits doit répondre aux besoins suivants :

- Le calcul du taux d'équipement de chaque client,
- La recommandation de produits aux clients tout en garantissant l'éligibilité de ces derniers,
- Le calcul du taux d'équipement estimé de chaque client si ce dernier consomme les produits recommandés,
- La recommandation d'une nouvelle segmentation tout en garantissant l'éligibilité des clients,
- La recommandation de produits aux clients selon la nouvelle segmentation,
- Le calcul du taux d'équipement estimé de chaque client si ce dernier change de segment et consomme les produits recommandés.

c. Troisième phase du projet

La plateforme web devra répondre aux besoins suivants :

- L'authentification de l'utilisateur à l'aide de son adresse électronique et de son mot de passe,
- L'importation des bases de données nécessaires au lancement du moteur,
- La génération et l'affichage des résultats attendus par le système de recommandation.

3.3. Besoins non fonctionnels

Les besoins non fonctionnels sont les besoins qui sont indirectement liés au fonctionnement du système et qui sont aussi très importants.

- Performance : Avant tout, le système doit être efficace par ses différentes fonctionnalités, c'est-à-dire qu'il répond de manière optimale à toutes les exigences des utilisateurs.
- Autonomie : Le système doit fonctionner dans son intégralité sans avoir recours à d'autres acteurs logiciels externes.
- Convivialité : Le système doit être facile à utiliser, notamment les interfaces

utilisateur doivent être conviviales, c'est-à-dire simples, ergonomiques et adaptées à l'utilisateur.

- Facilité : La navigation doit être claire et facile à gérer par l'administrateur.

4. Backlog de produit

Le backlog de produit est une liste élaborée et mise à jour par le Product Owner, contenant des fonctionnalités considérées prioritaires dans l'atteinte des objectifs du projet. Elle contient les éléments suivants :

- ID User Story : c'est l'identifiant unique de chaque user story.
- Feature : c'est un ensemble de user stories rassemblant l'objectif d'un même module.
- User story : c'est une fonctionnalité demandée par le client.
- Priorité : c'est la priorité souhaitée par le client. Dans ce cas, nous avons utilisé la méthode « MOSCOW ». La lettre M (Must) désigne une fonctionnalité qui doit être faite, la lettre S (Should) en désigne une qui devrait être faite, la lettre C (Could) en désigne une qui pourrait être faite, et la lettre W (Won't) en désigne une qui ne sera pas faite pour le moment, mais sera peut-être faite dans le cadre de collaborations futures.

Le tableau suivant présente les user stories et les différentes fonctionnalités qui seront mises en œuvre.

Sprint	ID	Feature	User story	Priorité
Sprint 1	ODG.1	Observation des données globales	En tant qu'administrateur, je souhaite observer la répartition des clients de la banque par type.	M
	ODG.2		En tant qu'administrateur, je souhaite mesurer la performance du service commercial de la banque en observant les différents taux d'équipement actuels des clients.	M
	ODG.3		En tant qu'administrateur, je souhaite observer les données et la catégorie de clientèle de la banque.	M
	ODG.4		En tant qu'administrateur, je souhaite observer un résumé des données collectées auprès des bases de données de la banque, ainsi que leur taux de complétude.	C
	ODS.1	Observation des données par segment.	En tant qu'administrateur, je souhaite observer des tableaux de bord des informations des clients par segment.	M

	ODS.2		En tant qu'administrateur, je souhaite observer les chiffres d'affaires annuels confiés par chaque segment de clientèle.	M
Sprint 2 + 3	RP.1	Recommandation des produits	En tant qu'administrateur, je souhaite que le système puisse recommander des produits aux clients tout en garantissant l'éligibilité de ces derniers.	M
	RP.2		En tant qu'administrateur, je souhaite que le système puisse calculer le taux d'équipement estimé de chaque client si ce dernier consomme les produits recommandés.	M
	RP.3		En tant qu'administrateur, je souhaite que le système puisse calculer le taux d'équipement actuel de chaque client.	M
	PNS.1	Proposition d'une nouvelle segmentation	En tant qu'administrateur, je souhaite que le système puisse recommander une nouvelle segmentation des clients tout en garantissant l'éligibilité de ces derniers.	M
	PNS.2		En tant qu'administrateur, je souhaite que le système puisse recommander des produits aux clients selon la nouvelle segmentation.	S
	PNS.3		En tant qu'administrateur, je souhaite que le système puisse calculer le taux d'équipement estimé de chaque client si ce dernier change de segment et consomme les produits recommandés.	S
Sprint 4 + 5	AU.1	Authentification	En tant qu'administrateur, je souhaite m'authentifier à la plateforme.	S
	LM.1	Lancement du modèle	En tant qu'administrateur, je souhaite importer les différentes bases de données pour le lancement du moteur.	S
	ORM.1	Observation du résultat du modèle	En tant qu'administrateur, je souhaite observer le résultat du moteur dans un tableau.	S
	ORM.2		En tant qu'administrateur, je souhaite filtrer le tableau selon mes préférences.	S

Tableau 3- Backlog de produit

5. Environnement de travail

Dans un monde où la digitalisation a bien posé ses marques et commence à régner, le public demande une utilisation claire et simple des outils et logiciels qui envahissent son quotidien. Pour cela, il est crucial de choisir au préalable les technologies à l'aide desquelles nous allons subvenir aux besoins de notre client. Dans cette partie, nous menons une étude technique où nous décrivons les ressources logicielles utilisées dans le développement de notre projet. Les tableaux présentent les différentes technologies que nous avons choisi pour les nombreux avantages qu'ils offrent.

Power BI Desktop 	Power BI est un ensemble de services logiciels, d'applications et de connecteurs qui travaillent ensemble pour transformer les sources de données non liées en informations cohérentes, visuellement immersives et interactives. Les données peuvent être une feuille de calcul Excel ou un ensemble d'entreposés de données hybrides basés sur le cloud et sur site. Power BI permet de se connecter facilement aux sources de données, de visualiser et de découvrir ce qui est important, et de le partager avec nos collaborateurs.
Power Query 	L'éditeur Power Query représente l'interface utilisateur Power Query, où il est possible d'ajouter ou modifier des requêtes, gérer des requêtes en regroupant ou en ajoutant des descriptions aux étapes de requête, ou visualiser vos requêtes et leur structure avec différentes vues.

Tableau 4- Technologies utilisées pour la création des tableaux de bord

Anaconda 	Anaconda est une plateforme de distribution gratuite et open source du langage de programmation Python. Anaconda possède ses propres outils de gestion, d'installation et de mise à jour des paquets invoqués avec la commande « conda ». Anaconda offre un moyen facile de créer différents environnements Python et de passer d'un environnement à l'autre.
Jupyter 	Jupyter est une application web utilisée pour programmer dans plus de 40 langages de programmation, dont Python, R ou Scala. Jupyter est une évolution du projet IPython, qui permet de créer des blocs-notes : programmes contenant à la fois du texte en démarque et du code en Python, R, etc.

Python  python	<p>Python est le langage de programmation le plus pertinent pour la science des données. Il est considéré comme le langage de programmation le plus en vogue pour le traitement de texte et le traitement numérique. Comparé à R, Python est beaucoup plus rapide et mature dans le traitement de données qu'elles soient de type textuel ou numérique très volumineuses grâce à des packages complets tels que Scikit-learn et NLTK.</p>
---	---

Tableau 5- Technologies utilisées pour le développement du modèle

Figma 	<p>Une des étapes essentielles dans le développement d'un projet consiste à réfléchir à la forme que prendra ce projet. En effet, les maquettes permettent d'établir un cahier des charges visuel. Pour cela, nous avons utilisé Figma, un outil de design d'interface en ligne, collaboratif et qui se présente en temps réel. En plus de ses capacités de prototypage, Figma permet la génération du code (SVG, CSS, iOS et Android) pour le transfert.</p>
Django 	<p>Django est un framework de développement web en Python consacré au développement rapide de sites internet, sécurisés, et maintenables. Il est gratuit, open source, a une communauté active, une bonne documentation, et plusieurs options pour du support gratuit ou non.</p>
HTML et CSS 	<p>Les sigles « HTML » sont l'abréviation de « HyperText Markup Language » ou « langage de balisage hypertexte » en français. C'est un langage de balisage, c'est-à-dire un langage permettant de définir les différents contenus d'une page. Et tandis que le HTML sert à définir les différents éléments d'une page, à leur donner du sens. Le CSS, lui, va servir à mettre en forme les différents contenus définis par le HTML en leur appliquant des styles.</p>

Tableau 6- Technologies utilisées pour le développement de la plateforme

6. Diagramme de Gantt

Couramment utilisé en gestion de projet, ce diagramme est un outil très utile et efficace pour l'identification et la représentation visuelles des différentes tâches à effectuer durant le projet.

Les unités de temps sont énumérées dans la ligne d'en-tête afin d'assurer une planification prévisionnelle simple et claire, tandis que chaque tâche est représentée par une barre horizontale dont la position et la longueur représentent la date de début, la durée et la date de fin selon l'unité de temps mentionnée précédemment. La figure suivante sert de représentation visuelle de notre planning approximatif :

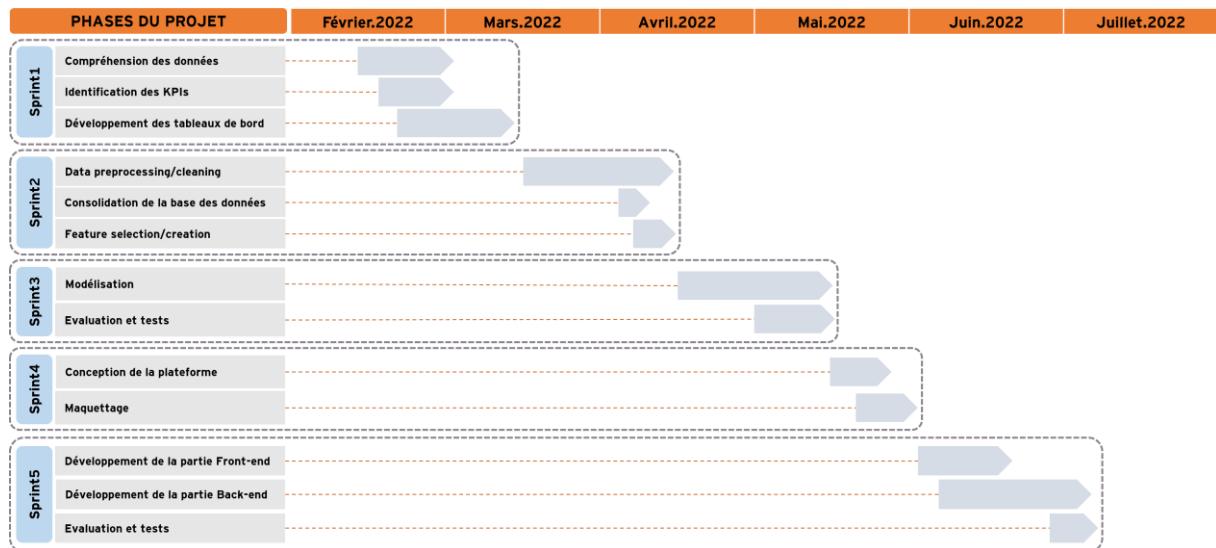


Figure 8- Diagramme de Gantt

Conclusion

L'objectif de ce chapitre est d'établir une vision globale du produit afin d'identifier les besoins et les fonctionnalités. C'est là que se terminent les rôles des directions fonctionnelles et techniques. Le prochain chapitre traitera la première partie du projet.

Chapitre 3 : Visualisation des données

Introduction

L'analyse descriptive permet d'exploiter les énormes quantités de données collectées auprès des clients, des applications en temps réel, etc. En diffusant des informations tangibles, elles nous permettent de garder une longueur d'avance sur nos projets.

Au sein de ce chapitre, nous allons entamer la présentation de la réalisation pratique, en commençant par la phase de visualisation des données, représentant le premier sprint. Pour cela, nous allons commencer par identifier les indicateurs clés de performance qu'exigent le client, afin de passer à la mise en place des tableaux de bord nécessaires.

1. Extraction des données

La première étape avant l'intégration des données est leur extraction. Lors de l'étape d'extraction des données, les données sont exportées des emplacements sources vers une zone de transit.

a. Base de données des clients

La figure ci-dessous montre la base de données des clients sous format Excel.

CODECLIENT	NUMEROCOMPTE	DEVISE	AGENCE	GENRE	SEGMENT	DATE_ENTREE_EN_RELATION	DATE_DE_NAISSANCE	NOTATION_INTERNE_DU_CLIENT	CSP	IMPAYES_ANTE_RIEURS
				M		16/09/2010	30/07/1968	invalide	820 0.00	
				M		10/04/2006	15/03/1957	agréments Directeu	410 0.00	
				M		11/02/2013	15/03/1957	agréments Directeu	410 0.00	
				M		15/07/2015	15/03/1957	agréments Directeu	410 -16,008.00	
				M		11/07/2005	16/04/1960	Primes En Recours	440 0.00	
				M		18/06/2020	14/07/1957	Redevance =	510 0.00	
				M		02/07/2010	01/01/1960	invalide =	510 0.00	
				F		11/12/1992	29/05/1963	valide	530 0.00	
				F		07/06/2012	31/12/1955	Redevance	950 0.00	
				F		19/09/2012	15/06/1960	invalide -	351 0.00	
				M		03/04/2007	12/10/1962	Redevance =	440 0.00	
				M		28/08/2006	26/06/1962	invalide -	317 0.00	
				M		25/06/2007	26/06/1962	invalide -	317 0.00	
				M		23/08/2011	26/06/1962	invalide -	317 0.00	
				F		29/05/2002	03/10/1970	valide	316 -15,000.00	
				F		03/01/2000	23/03/1984	Faillite	270 0.00	
				M		15/05/2015	12/08/1989	valide	899 0.00	
				M		04/02/2014	26/10/1958	Redevance =	351 -195,000.00	
				M		07/09/2016	26/10/1958	Redevance =	351 0.00	
				M		11/01/2011	17/03/1970	Redevance =	499 0.00	
				M		03/02/2012	17/03/1970	Redevance =	499 0.00	
				M		11/03/2013	17/03/1970	Redevance =	499 0.00	
				F		09/08/2012	02/07/1962	valide	510 0.00	
				F		18/12/2017	02/07/1962	Redevance	510 0.00	
				M		06/01/1997	01/01/1941	Primes En Recours	530 -35,145.00	
				M		05/10/2006	17/12/1967	Redevance =	351 0.00	
				M		20/02/2014	17/12/1967	Redevance =	351 0.00	
				F		19/07/2011	31/12/1963	valide =	510 0.00	
				M		20/10/2014	20/06/1970	valide	331 0.00	

Figure 9- Base de données Clients

Cette base de données contient 315 mille lignes alimentées par les informations des différents clients, notamment : le code client – le numéro de compte – le sexe du client – la date de naissance – le statut matrimonial – la catégorie socio-professionnelle – la catégorie – l'employeur – l'activité – la date d'entrée en relation du client – le segment du client – la devise utilisée – l'agence – la notation interne – le type de compte – le numéro de compte associé – les mouvements créditeurs – les mouvements débiteurs – les impayés antérieurs – l'autorisation de découvert – commissions perçus – commissions encourus.

b. Base de données des dépôts bancaires

La figure suivante montre la base de données des dépôts bancaires sous format Excel

CODECLIENT	Column1	NUMEROCOMPTE	1	DEVISE	_2	AGENCE	_3	TYPE_DE_DEPOT	_4	ENCOURS	_5	DEPOTS	_6	TAUX_COMMISI0NS
						CSE				200374				
						CSE				0				
						CSE				0				
						CSE				0				
						CSE				1003906				
						CSE				0				
						CSE				2530504				
						CSE				10581953				
						CSE				0				
						CSE				0				
						CSE				0				
						CSE				904143				
						PEL				2719373				
						CSE				0				
						CSE				0				
						CSE				6702319				
						CSE				6708226				

Figure 10- Base de données Dépôts

La base de données des dépôts bancaires contient 315 mille lignes alimentées par les informations des différents dépôts bancaires, notamment : le code client – le numéro de compte – la devise – l'agence – le type du dépôt – les encours – les dépôts – les taux commissions.

c. Base de données des équipements clients

La figure suivante montre la base de données des équipements sous format Excel.

Figure 11- Base de données Equipement

La base de données des équipements bancaires contient 437 mille lignes alimentées par les informations des différentes consommations des clients en termes de produits, notamment : le code client – le code siège – la racine – les différents produits de la banque.

d. Base de données des financements

La figure suivante montre la base de données des financements sous format Excel.

CODECLIENT	NUMEROCOMPTE	DEVISE	AGENCE	TYPE_DE_FINANCEMENT	MONTANT_DU_FINANCEMENT	ENCOURS	DATE_DE_DEBLOCAGE
00000000000000000000	00000000000000000000	CONGO			1500000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			3000000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			2500000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			10000000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			3500000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			6500000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			8150000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			1500000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			2000000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			2700000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			700000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			7500000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			5500000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			3500000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			250000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			1000000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			1200000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			5300000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			4000000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			3000000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			3300000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			5000000	0	02/01/2017
00000000000000000000	00000000000000000000	CONGO			2500000	0	02/01/2017

Figure 12- Base de données des Financements

La base de données des financements bancaires contient 87 mille lignes alimentées par les informations des financements que la banque a accordé à ses clients, notamment : le code client – le numéro de compte – la devise – l'agence – le type de financement – le montant du financement – les encours – la date de déblocage.

e. Base de données des opérations bancaires

La figure suivante montre la base de données des opérations bancaires sous format Excel.

Figure 13- Base de données des Opérations

La base de données des opérations bancaires contient 17,8 millions de lignes alimentées par les informations des opérations bancaires des clients, notamment : le code client – le numéro de compte – le libellé du mouvement bancaire – le montant du mouvement – la date du mouvement – le segment du client – l'année de l'opération.

2. Indicateurs clé de performance

Afin d'assurer un bon compte rendu des informations, il est essentiel d'identifier les données qui seront utilisés par la suite pour étudier les écarts, les résultats, les évolutions et les différentes performances des activités commerciales de la banque. Dans ce cas, nous aurons recours aux indicateurs clé de performance, plus communément appelés KPIs.

2.1. Identification des KPIs

Nous pouvons identifier 2 types de KPIs dans notre cas :

- Les indicateurs de performances *business* indiquent le chiffre d'affaires confiés et la fréquence de l'interaction des clients avec la banque.
- Les indicateurs de performance commerciaux sont utilisés afin de déterminer la performance commerciale de la banque en termes de vente de produits.

Le tableau 3 suivant présente les indicateurs clé de performance utilisés afin de mettre en place les tableaux de bord :

▪ Clients retail	Nombre total des clients de la banque.
▪ Clients actifs	Nombre des clients ayant effectué au moins une opération pendant les 90 derniers jours.
▪ Clients dormants	Nombre des clients ayant effectué moins d'une opération pendant les 90 derniers jours.
▪ KPIs de complétude	Taux de complétude des données existantes.
▪ Taux d'équipement client	Pourcentage du nombre total des produits consommés par rapport au nombre des produits dédiés.
▪ Taux d'équipement produits éligibles	Pourcentage du nombre des produits éligibles consommés par rapport au nombre des produits dédiés.
▪ Taux d'équipement produits non éligibles	Pourcentage du nombre des produits non éligibles consommés par rapport au nombre des produits dédiés.
▪ Les produits les plus consommés	Les 3 produits les plus vendus et leur nombre de vente.
▪ Les produits les moins consommés	Les 3 produits les moins vendus et leur nombre de vente.
▪ Chiffre d'affaires	Somme d'argent confiée à la banque par client.

annuel confié	
▪ Moyenne d'opérations par client par semaine	Nombre d'opérations qu'effectue un client par semaine.

Tableau 7- Indicateurs clé de performance

2.2. Mise en place des KPIs

Afin de répondre aux besoins analytiques particuliers de l'organisation, les figures qui suivent montrent la mise en place des différents indicateurs de performance cités au sein du Tableau 3.

a. Calcul des indicateurs de complétude

Les différents indicateurs d'incomplétude, faisant référence aux données non renseignées en ce qui concerne l'activité socio-professionnelle, le statut matrimonial et la date de naissance, ont été calculés comme suit :

CSPBlank = CALCULATE(COUNTBLANK('ClientGroupé'[CSP]))

Figure 14- Calcul des indicateurs de complétude

b. Calcul des taux d'équipement

Afin de calculer les différents taux d'équipement, nous avons d'abord créé 4 mesures, toutes calculées de la même manière que l'illustre la figure suivante. Notons que le terme « produit éligible » fait référence à un produit destiné au segment dont le client fait partie.

- PEL_OK : Nombre de produits éligibles et consommés par le client.
 - PEL_NONOK : Nombre de produits éligibles mais non consommés par le client.
 - PNONEL_OK : Nombre de produits non éligibles mais consommés par le client.
 - PNONEL_OK : Nombre de produits non éligibles et non consommés par le client.

```
1 PEL_OK =
2 var erreur = 0 return
3 var _ = IF(groupByequipement[_____]=1&&groupByequipement[ProduitR  el._____]=1,erreur+1,erreur) return
4 var _ = IF(groupByequipement[_____]=1&&groupByequipement[ProduitR  el._____]=1,erreur+1,erreur) return
5 var SMS = IF(groupByequipement[SMS]=1&&groupByequipement[ProduitR  el.SMS]=1,erreur+1,erreur) return
6 var AFFAIRE = IF(groupByequipement[AFFAIRE]=1&&groupByequipement[ProduitR  el.AFFAIRE]=1,erreur+1,erreur)
.
.
.
43 var resultat = _____+_____+SMS+AFFAIRE+_____+AUTRES_CARTES+GOLD+PLATINUM+ASSURANCE+PACK_+_____+PACK_PRESTIGE+PACK_PRO+PACK_
44 +_____+ECOMMERCE+PACK_+ASSURANCE_AUTOMOBILE+_____+CAPITAL_ETUDE+PACK_+_____+CARTES_SALAIRES+TPE+CSE+CONSO+DAT
45 +HABIT+EQUIP+LOYER+AUTAV+PE+PEL+PER+PACK_+_____+plus+PACK_F+_____+plus+PACK_PRO_plus+PACK_F+_____+plus return resultat
```

Figure 15- Calcul des produits éligibles consommés

- Produits consommés : PEL OK + PNQNEL OK

Les différents taux d'équipement (taux d'équipement client/ taux d'équipement en produits éligibles/ taux d'équipement en produits non éligibles) ont été respectivement été calculés comme suit :

```
Cons/Dédiès = DIVIDE(groupByequipement[Produit consommé],groupByequipement[Produit dédiès])
```

Figure 16- Calcul du taux d'équipement client

```
just/Dédiès = DIVIDE(groupByequipement[PEL_OK],groupByequipement[Produit dédiès])
```

Figure 17- Calcul du taux d'équipement en produits éligibles

```
Erroné/dédiès = DIVIDE(groupByequipement[PNoNEL_OK],groupByequipement[Produit dédiès])
```

Figure 18- Calcul du taux d'équipement en produits non éligibles

c. Calcul des revenus

Afin de procéder au calcul des indicateurs de performances *business*, nous avons procédé comme suit :

- Calcul de la somme moyenne effectuée par mouvement bancaire :

```
1 MoyMvt =
2
3 var OPpos = CALCULATE(SUM(OPERATIONS_ano[MONTANT_MVT_CPTABLE]), FILTER ('OPERATIONS_ano', ('OPERATIONS_ano'[MONTANT_MVT_CPTABLE] > 0)))
4 var OPneg = CALCULATE(SUM(OPERATIONS_ano[MONTANT_MVT_CPTABLE]), FILTER ('OPERATIONS_ano', ('OPERATIONS_ano'[MONTANT_MVT_CPTABLE] < 0)))
5
6 var TotMvt = OPpos + ABS(OPneg)
7
8 var result = DIVIDE(TotMvt, 17843007) return result
```

Figure 19- Calcul de la moyenne d'un mouvement bancaire

- Calcul de la somme des revenus (les sommes déposées à la banque) :

```
SumRevenus = CALCULATE(SUM('New OP'[MONTANT_MVT_CPTABLE]), FILTER ('New OP', ('New OP'[MONTANT_MVT_CPTABLE] > 0)))
```

Figure 20- Calcul de la somme des revenus

- Calcul de la moyenne des revenus (sommes déposées à la banque) :

```
1 MoyRevenus =
2 var _year = SELECTEDVALUE('New OP'[Year]) return
3 Var _2020 = CALCULATE(([SumRevenus]/DISTINCTCOUNT('New OP'[CODECLIENT]),'New OP'[Year]=2020) return
4 Var _2021 = CALCULATE(([SumRevenus]/DISTINCTCOUNT('New OP'[CODECLIENT]),'New OP'[Year]=2021) return
5
6 If(_year=2020, CALCULATE([SumRevenus]/DISTINCTCOUNT('New OP'[CODECLIENT]), 'New OP'[Year]=2020),
7 IF(_year=2021, CALCULATE([SumRevenus]/DISTINCTCOUNT('New OP'[CODECLIENT]), 'New OP'[Year]=2021), ((_2020+_2021)/2)))
```

Figure 21- Calcul de la moyenne des revenus

- Calcul de la moyenne des visites par semaine par client :

```
1 MoyVisitSemCli =
2
3 var _year = SELECTEDVALUE('New OP'[Year]) return
4 Var _2020 = CALCULATE(DIVIDE([MoyVisiteSemaine],DISTINCTCOUNT('New OP'[CODECLIENT]),'New OP'[Year]=2020) return
5 Var _2021 = CALCULATE(DIVIDE([MoyVisiteSemaine],DISTINCTCOUNT('New OP'[CODECLIENT]),'New OP'[Year]=2021) return
6
7 If(_year=2020, CALCULATE(DIVIDE([MoyVisiteSemaine], DISTINCTCOUNT('New OP'[CODECLIENT])), 'New OP'[Year]=2020),
8 IF(_year=2021, CALCULATE(DIVIDE([MoyVisiteSemaine], DISTINCTCOUNT('New OP'[CODECLIENT])), 'New OP'[Year]=2021), ((_2020+_2021)/2)))
```

Figure 22- Calcul de la moyenne des visites par semaine par client

d. Affectation des statuts aux clients

Comme dernière étape, il nous est demandé de répartir la totalité des clients en clients actifs et d'autres dormants.

```
1 statut client =
2
3 var NB_T = Calculate(MAX('Opération2020'[Index]),
4 FILTER('Opération2020','Opération2020'[CODECLIENT]=EARLIER('Opération2020'[CODECLIENT])))
5 ) return
6
7 var NB_TR = Calculate(
8 DISTINCTCOUNT('Opération2020'[Trimestre]),
9 FILTER('Opération2020','Opération2020'[CODECLIENT]=EARLIER('Opération2020'[CODECLIENT]))
10 ) return
11
12 If (
13 NB_T>3 && NB_TR=4,"Actif",
14 "Non Actif"
15 )
```

Figure 23- Affectation des statuts aux clients

3. Modélisation

Pour programmer un système, il ne suffit pas de commencer immédiatement dans la phase de pratique : il faut d'abord organiser nos idées, les décrire, les relier pour faciliter le travail. C'est ce qu'on appelle la modélisation. La modélisation d'un système avant sa réalisation permet de mieux comprendre son fonctionnement. L'objectif est de maîtriser sa complexité en la transformant en graphisme, et ainsi d'établir une vision globale du produit. Le DataWarehouse sur lequel sera basée cette phase du projet est représenté comme suit :

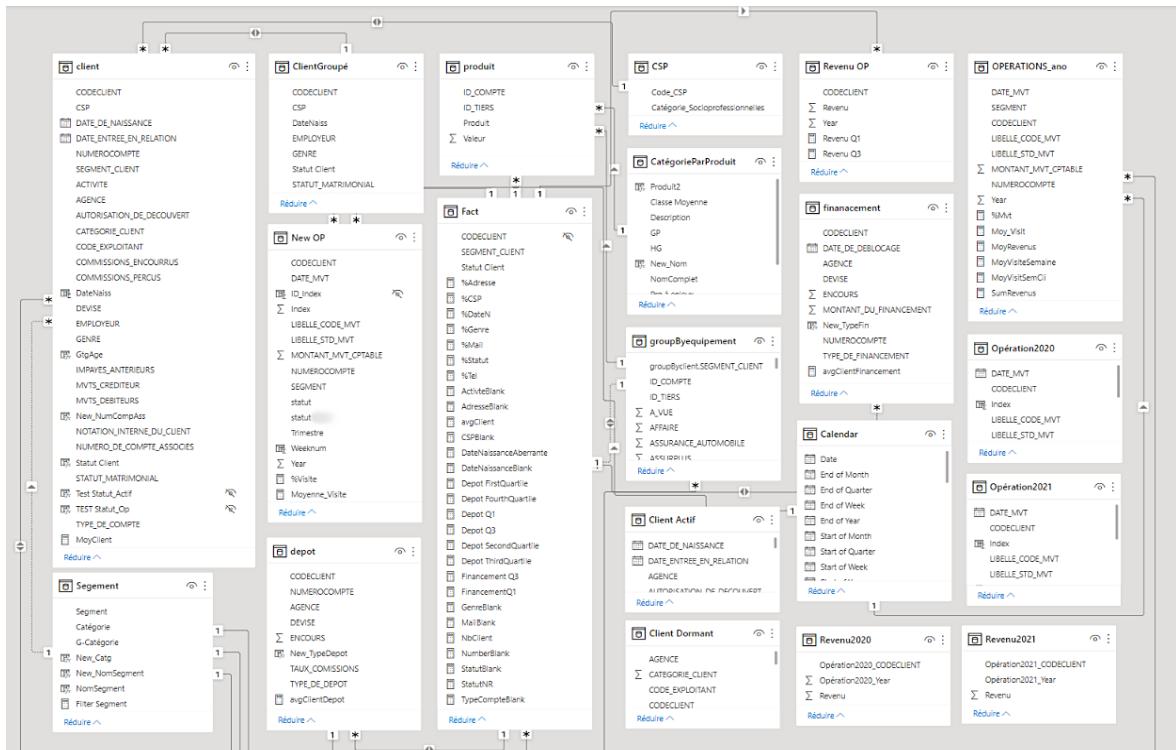


Figure 24- Modélisation des données

4. Tableaux de bord

Au fur et à mesure de l'évolution des activités de la banque, le volume de données à traiter augmente tellement qu'il devient impossible pour les commerciaux d'assimiler ces informations et les analyser en un simple coup d'œil. En tant que première étape du projet, il est nécessaire d'extraire les informations nécessaires et chercher à comprendre et à synthétiser les résultats des indicateurs de performance dans le but d'établir une meilleure stratégie et faciliter la prise de décision. Nous allons procéder à la présentation des 14 tableaux de bord élaborés dans le cadre de cette partie du projet.

4.1. Vue globale

Pour commencer, nous avons tout d'abord mis en place un résumé du projet et des données collectées. La figure suivante le représente.

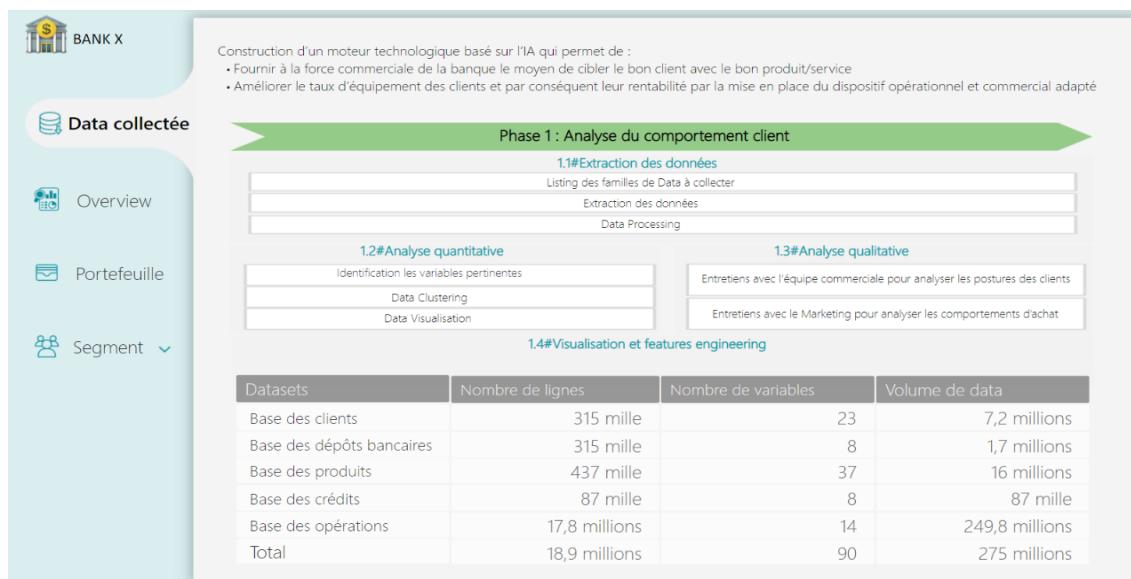


Figure 25- Dashboard 1_Données collectées

Comme communiqué, nous disposons de 5 bases de données : une base des clients, une base des dépôts bancaires, une base des produits de la banque, une base des crédits fournis aux clients, et une base des opérations bancaires effectuées durant les années 2020 et 2021. Nous avons donc traité au total 275 millions de données, collectées parmi 18,9 millions de lignes et 90 variables.

La figure ci-dessous présente une vue globale sur les informations collectées. Notons qu'il est possible de filtrer ce tableau de bord par catégorie du client (particuliers et professionnels), son type (actif ou dormant), ainsi que par les segments de la banque.



Figure 26- Dashboard 2_ Vue globale

Au sein de ce tableau de bord, nous avons présenté le résumé des données collectées et les indicateurs les plus importants concernant le total des clients. Pour commencer, la banque dispose de 176 293 clients dont 92 774 actifs et 83 519 dormants.

Les données des catégories socio-professionnelles des clients, leur statut matrimonial, leur sexe, ainsi que leur date de naissance ont été parfaitement renseignées avec des taux allant de 95% à 99%. Au contraire de ces dernières, seulement 3.6% des adresses mails ont été renseignées, mais cela n'affectera pas le travail à effectuer étant donné leur insignifiance au sein de notre projet.

En ce qui concerne les indicateurs de performances *business*, nous remarquons un taux d'équipement client global de 22%, un taux d'équipement de produits éligibles de 18% et un taux d'équipement de produits non éligibles de 4%. Sous le visuel « Taux de clients équipés », on peut remarquer que 74% des clients sont équipés de 0 à 8 produits parmi 33, 19% sont équipés de 8 à 17 produits, et seulement 6% sont équipés de 17 à 25 produits. Ces taux sont considérés faibles et démontrent l'existence de faiblesses au sein du processus de vente de produits.

4.2. Portefeuille

Les deux figures suivantes présentent le résumé de la répartition des clients en fonction de différentes caractéristiques, mais également une vue globale sur les informations des portefeuille clients.

Notons qu'il est possible de filtrer ces deux tableaux de bord par type de clients (actif ou dormant), ainsi que par les segments de la banque.

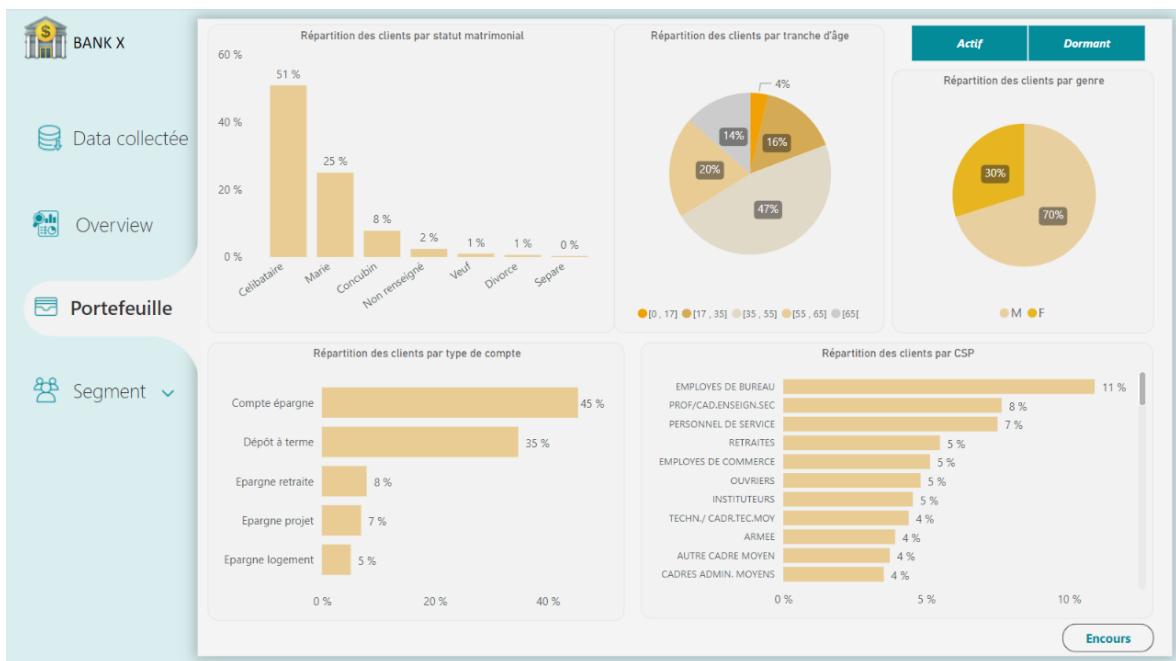


Figure 27- Dashboard 3_Portefeuille

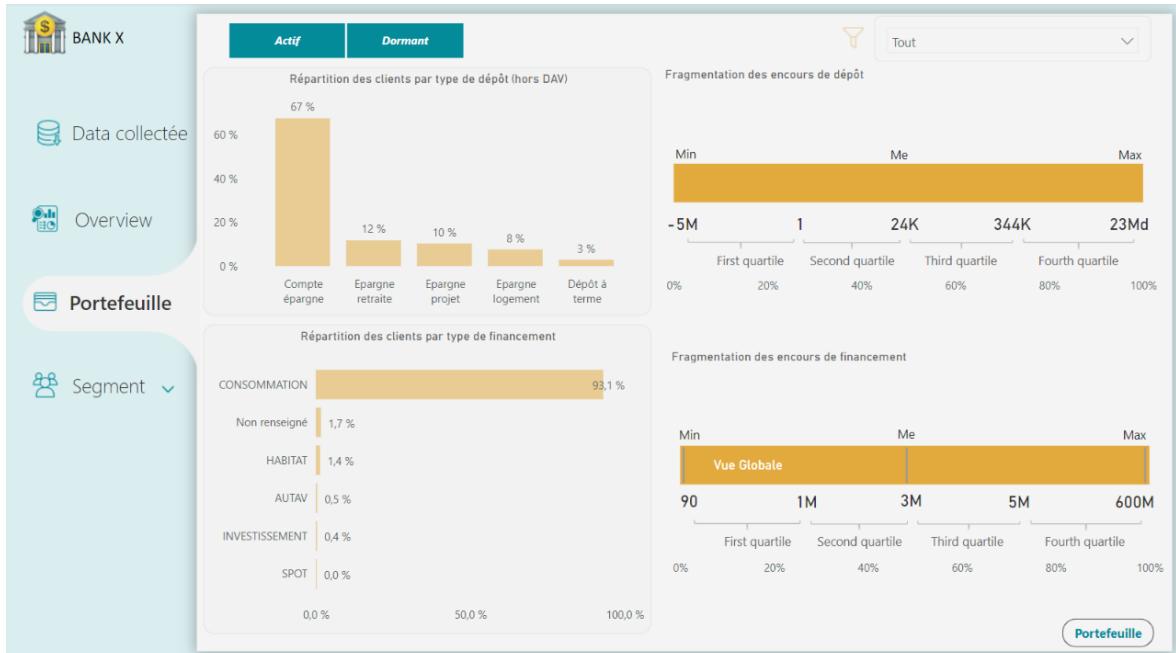


Figure 28- Dashboard 4_Encours

D'après ces tableaux de bord, nous pouvons remarquer que près de 50% des clients de la banque sont âgés de 35 à 55 ans, que la plupart sont célibataires, et que le sexe masculin est plus présent que le féminin. Mais l'information la plus pertinente à relever est que les types de compte les plus fréquents sont le « compte spécial épargne » avec un taux de 45%, suivi du compte « dépôt à terme » avec un taux de 35%. Quant aux financements, les plus fréquents s'avèrent être les crédits à la consommation avec un taux de 93,1%.

4.3. Clients « professionnels »

Comme mentionné précédemment, les clients sont catégorisés en deux grandes catégories : les clients professionnels appartenant aux segments « Professionnels 1 » et « Professionnels 2 » et les clients particuliers appartenant aux segment « Grand public », « Classe moyenne », et « Haut de gamme ».

Les 4 figures suivantes illustrent les tableaux de bord de la catégorie des clients professionnels. Notons qu'il est possible de filtrer ces tableaux de bord par année (2020 ou 2021) ainsi que par les sous-segments.

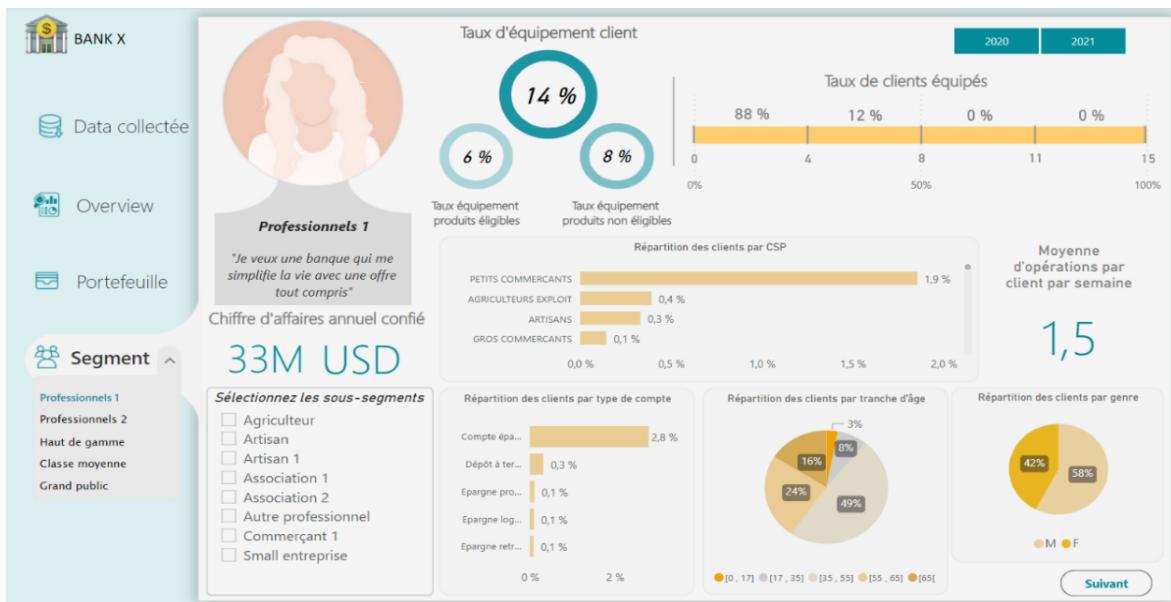


Figure 29- Dashboard 5_ Professionnels 1-1

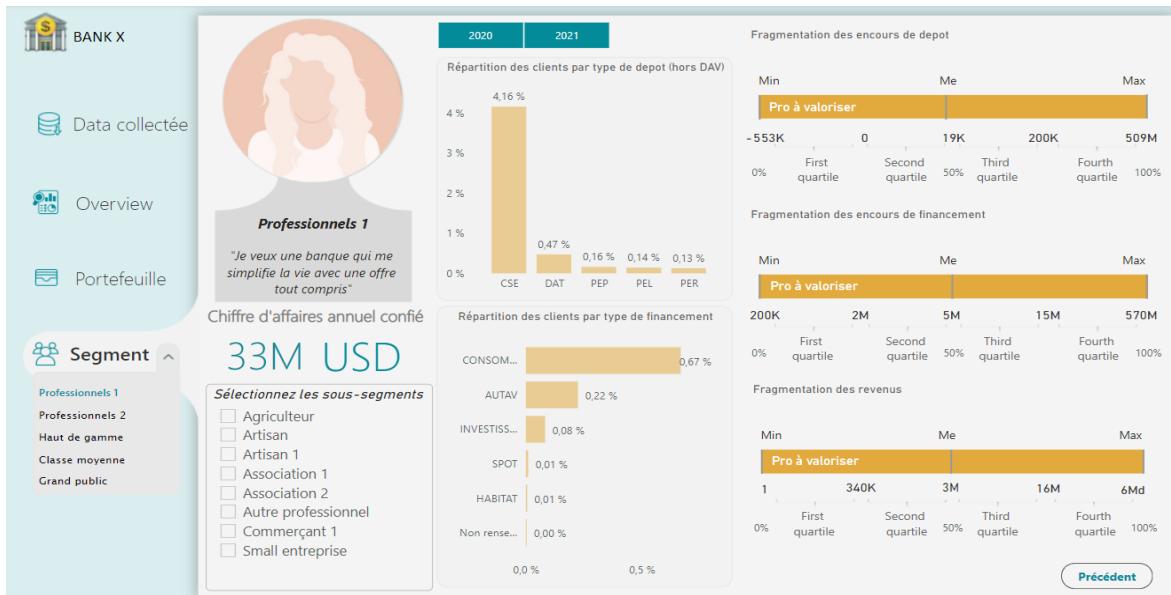


Figure 30- Dashboard 5_ Professionnels 1-2



Figure 31- Dashboard 6_ Professionnels 2-1

Les figures 24 – 26 représentent des vues globales sur les informations confiées concernant les clients appartenant respectivement aux segments « Professionnels 1 » et « Professionnels 2 ». On peut relever une différence remarquable entre le chiffre d'affaires annuel confié par les clients du premier segment (33M USD) pour une moyenne de 1,5 opération par semaine par client et celui confié par les clients du deuxième (136M USD) pour une moyenne de 4,4 opérations par semaine par client. Ces chiffres sont notamment expliqués par les catégories socio-professionnelles des clients ainsi que par les revenus : des opérations pouvant atteindre 48Md pour les « Professionnels 2 » contre 6Md pour les « Professionnels 1 ».

4.4. Clients « particuliers »

Les 6 figures qui vont suivre illustrent les tableaux de bord de la catégorie des clients particuliers. Notons qu'il est possible de filtrer ces tableaux de bord par année (2020 ou 2021) ainsi que par les sous-segments.

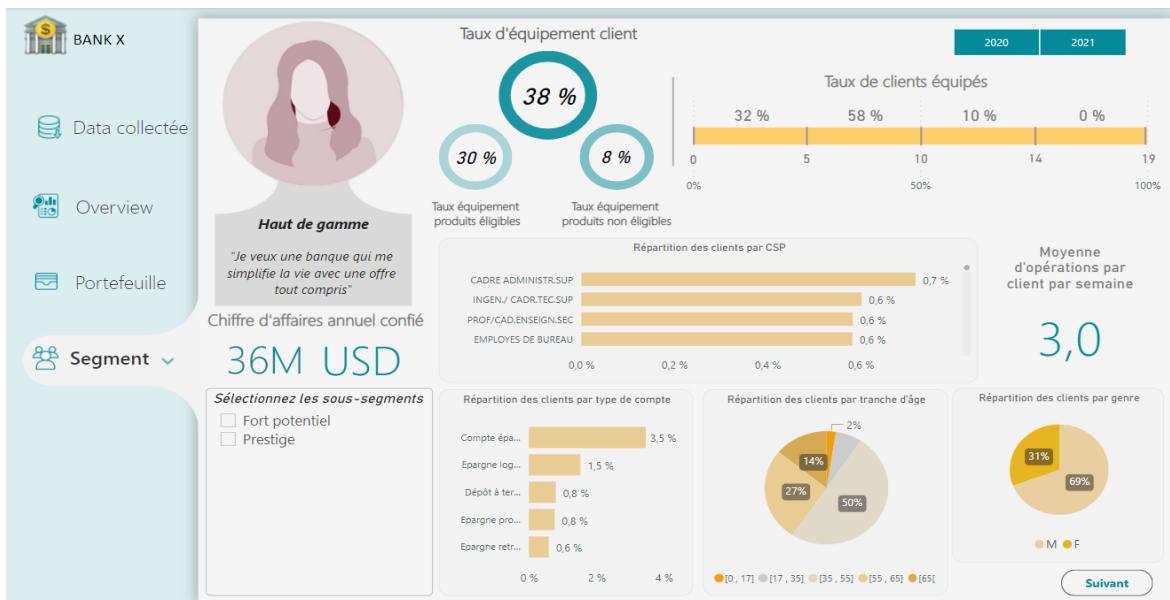


Figure 32- Dashboard 7_ Haut de gamme - 1



Figure 33- Dashboard 7_ Haut de gamme - 2



Figure 34- Dashboard 8_ Classe moyenne



Figure 35- Dashboard 9_ Grand public

Les figures 27 – 30 ci-dessus représentent des vues globales sur les informations confiées concernant les clients appartenant respectivement aux segments « Haut de gamme », « Classe moyenne » et « Grand public ». A l'instar de l'analyse précédente, nous pouvons également remarquer les différences de revenus entre les segments. Ce qui nous intéresse le plus au sein de ces visuels, ce sont les taux d'équipement visiblement très faibles, atteignant les 13% pour le segment « Grand public ».

Conclusion

L'analyse descriptive est essentielle pour chaque projet, permettant aux décideurs d'avoir une compréhension plus approfondie des données et donc de prendre des décisions sur les actions futures dans les entreprises. Ce chapitre a été consacré à la présentation les différents graphiques des tableaux de bord, et nous a permis de constater la faiblesse des taux d'équipement des clients en produits. Le chapitre suivant sera alors consacré à la mise en œuvre du modèle qui nous permettre de proposer des recommandations dans le but d'améliorer ces chiffres.

Chapitre 4 : Mise en œuvre du modèle

Introduction

Au sein du chapitre précédent, nous avons pu observer les différents chiffres résumant les relations de la banque avec ses clients, notamment les taux d'équipement en produits remarquablement faibles. Ce chapitre sera donc consacré à la partie principale de notre projet : le développement du modèle de recommandation de produits aux clients afin d'améliorer le taux d'équipement de ces derniers.

1. Méthodologie de travail

En science des données, il n'existe toujours pas de méthodologie faisant l'unanimité quant à la gestion d'un projet d'exploration de données. Nous sélectionnons alors notre méthodologie selon 3 critères principaux : l'accès aux données, l'interactivité et la concentration sur les tâches (*task-focused*). Dans ce contexte, nous choisissons de travailler selon la méthodologie CRISP-DM, communément répandue ce domaine.

CRISP-DM, qui signifie *Cross-Industry Standard Process for Data Mining*, est un modèle de processus d'exploration de données décrivant une approche couramment utilisée pour résoudre les problèmes d'analyse, d'exploration et de science des données. Elle divise le processus de data mining en six étapes, en expliquant les relations entre ces dernières. L'avantage de cette méthode est qu'elle est facilement adaptable dans le cas où un projet ne traite pas toutes les étapes, concentrant le travail sur les phases demandées. La figure suivante illustre les différentes phases de CRISP-DM.

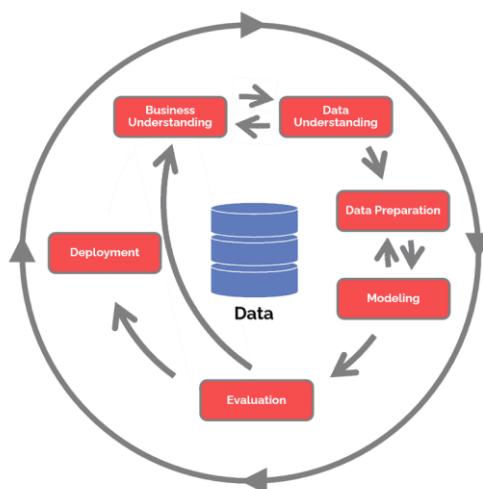


Figure 36- Méthodologie CRISP-DM

Les phases de la méthodologie se déroulent comme suit :

- 1- La compréhension du métier consiste en l'analyse du problème et l'identification des besoins que le modèle cherche à résoudre ou à améliorer.
- 2- La compréhension des données vise à la détermination des données à analyser, à leur qualité et leur signification d'un point de vue métier.
- 3- La préparation des données consiste au pré-traitement et au nettoyage des données, et surtout le recodage et la normalisation pour les rendre compatibles avec l'algorithme utilisé.
- 4- La modélisation représente la phase principale qui comprend le choix de l'algorithme et son enchaînement.
- 5- L'évaluation conduit à la vérification du modèle et des résultats obtenus afin de s'assurer qu'ils répondent aux objectifs prédéfinis, en testant notamment la robustesse et la précision du modèle obtenu.
- 6- Le déploiement s'agit de la mise en production du résultat au profit de l'utilisateur afin de mettre le résultat obtenu dans une forme simple, claire, et facile à comprendre.

2. Pré-traitement et nettoyage des données

L'augmentation de la collecte de données et son traitement systématique ont permis de développer des techniques d'apprentissage automatique qui nécessitent de grandes quantités de données pour s'exécuter et s'entraîner. On peut simplement penser qu'une grande quantité de données est suffisante pour que l'algorithme réussisse, mais dans la plupart des cas, les données ne sont pas adaptées et nécessitent un prétraitement avant de pouvoir être utilisées. Il s'agit d'une étape de prétraitement.

Les figures de l'élément – 1. Extraction des données du chapitre 3 – Visualisation des données – montrent des lignes dupliquées, des dates écrites de façons différentes, des collisions de données au sein des colonnes, des contenus erronés, etc.

2.1. Transformation de la base de données des clients

Nous avons d'abord commencé par le traitement de la base de données des clients, détaillé à la suite des figures suivantes.

```
def remove_espaces(self,df):
    df.columns=df.columns.str.strip()
    return df

def remove_columns(self,df):
    filter_col1 = [col for col in df.columns.tolist() if col.startswith('_')]
    df = df.drop(filter_col1, axis = 1)
    return df
```

Figure 37- Elimination des colonnes inefficaces

```

def preprocessing_client(self,base_client):
    base_client=self.remove_espace(base_client)
    base_client=base_client.drop(['CODECLIENT','DEVISER','GENRE','SEGMENT_CLIENT','NOTATION_INTERNE_DU_CLIENT',
                                'DATE_ENTREE_EN_RELATION','DATE_DE_NAISSANCE','STATUT_MATRIMONIAL','CSP'])
    corpo=[ "CO1","CO2","AET","II","PA","MNC1","MNC2","LC"]
    base_client.drop(base_client[base_client['SEGMENT_CLIENT'].isin(corpo)].index, inplace = True)
    base_client.drop(base_client[base_client['DEVISE'].isin(['EUR','USD'])].index, inplace = True)
    base_client.drop('DEVISE',axis=1,inplace=True)

    #Récupérer le minimum des dates d'entrées en relation par client pour avoir l'ancienneté du client
    relation=base_client[['CODECLIENT','DATE_ENTREE_EN_RELATION']]
    relation=relation.drop(relation[relation.apply(lambda x: x['DATE_ENTREE_EN_RELATION'].year > datetime.date.today().year and
                                                x['DATE_ENTREE_EN_RELATION'].month > datetime.date.today().month, axis = 1)].index)
    relation=relation.groupby("CODECLIENT").min().reset_index()

    data=base_client[['CODECLIENT','NOTATION_INTERNE_DU_CLIENT','GENRE','SEGMENT_CLIENT','STATUT_MATRIMONIAL','CSP']]

    #Merger la relation avec la data
    data=pd.merge(data,relation,on='CODECLIENT', how='inner')

    #Merger la date de naissance avec la data
    naissance=base_client[['CODECLIENT','DATE_DE_NAISSANCE']]
    data=pd.merge(data,naissance,on='CODECLIENT', how='inner')

    #Supprimer les clients dont la date de naissance est supérieure à la date d'entrée en relation
    data['DATE_ENTREE_EN_RELATION']=pd.to_datetime(data['DATE_ENTREE_EN_RELATION'])
    data['DATE_DE_NAISSANCE']=pd.to_datetime(data['DATE_DE_NAISSANCE'])
    df_client1 = data[pd.DatetimeIndex(data['DATE_DE_NAISSANCE']).year < datetime.date.today().year]
    df_client2=df.Client1.drop(df.Client1[df.Client1.apply(lambda x: x['DATE_DE_NAISSANCE'] > x['DATE_ENTREE_EN_RELATION'],
                                                       axis = 1)].index)

    #Supprimer les clients qui ont plus qu'une notation
    df_Client_different=df.Client2.drop_duplicates(['CODECLIENT', 'NOTATION_INTERNE_DU_CLIENT'])
    indexNames=df.Client_different.groupby('CODECLIENT').filter(lambda x: len(x)>= 2).index
    df_client_final=df.Client_different.drop(indexNames)

    #Transformation de la colonne date d'entrée en relation
    year=datetime.date.today().year
    month=datetime.date.today().month
    day=datetime.date.today().day

    df_client_final['DATE_ENTREE_EN_RELATION'] = pd.to_datetime(datetime.datetime(year,month,day))
    df_client_final['Ancienneté Client'] = (df_client_final['DATE_ENTREE_EN_RELATION'] / np.timedelta64(1,'M')).astype(int)
    df_client_final=df.client_final.drop(['DATE_ENTREE_EN_RELATION'],axis=1)
    df_client_final=df.client_final.dropna()

    #Création de la colonne âge
    df_client_final['Age']= datetime.date.today().year - pd.DatetimeIndex(df_client_final['DATE_DE_NAISSANCE']).year
    df_client_final=df_client_final.drop(['DATE_DE_NAISSANCE'], axis=1)

    #Regroupement des notations
    notation={'Acceptable - ':'Acceptable','Mediocre + ':'Mediocre','Mediocre - ':'Mediocre','Mediocre':'Mediocre',
              'Moyen - ':'Moyen','Moyen':'Moyen','Faible + ':'Faible','Faible':'Faible','Faible - ':'Faible',
              'Tres Preoccupant + ':'Tres Preoccupant','Tres Preoccupant - ':'Tres Preoccupant','Preoccupant + ':'Preoccupant',
              'Preoccupant - ':'Preoccupant','Preoccupant - ':'Preoccupant','Preoccupant + ':'Preoccupant',
              'Preoccupant - ':'Preoccupant','Creances En Recouvr.':'Creances En Recouvr.', 'Engagements Douteux - ':'Engagements Douteux'}
    df_client_final['NOTATION_INTERNE_DU_CLIENT']=df_client_final[['NOTATION_INTERNE_DU_CLIENT']].map(notation)

    return(df_client_final)

```

Figure 38- Traitement de la base de données des clients

Comme le montre la figure précédente, nous avons procédé comme suit :

- Pour commencer, nous avons supprimé les espaces qui compromettent les noms des colonnes et éliminé des colonnes commençant par le caractère « _ » étant donné leur inutilité.
- En second lieu, étant donné que la base de données comporte des clients de la catégorie « corporate » (personnes morales), nous avons procédé à l'élimination des lignes de ces derniers.
- Afin d'éviter de compromettre les résultats du modèle, et après échange avec le client, nous avons éliminé les lignes des clients dont les devises utilisées sont l'euro et le dollar.
- Comme quatrième étape, toujours en accord avec le client, nous avons supprimé les lignes des clients dont la date d'entrée en relation est plus récente que la date actuelle, et celle dont la date de naissance est plus récente que la date d'entrée en relation.
- Etant donné que la notation interne du client rentre dans la liste des variables pertinentes pour le développement du modèle, nous avons supprimé les lignes des clients ayant plusieurs notations afin de ne pas confusionner le modèle.
- Ensuite, afin de garantir des données quantitatives pour le modèle, nous avons

transformé la colonne « date d'entrée en relation » en une colonne « ancienneté du client », et la colonne « date de naissance » en colonne « âge du client ».

- Finalement, après consultation du client, nous avons regroupé les notations identiques en une seule (exemple : « Acceptable + » et « Acceptable - » deviennent « Acceptable »).

→ Résultat : Table contenant les colonnes « code client », « notation interne du client », « genre », « segment client », « statut matrimonial », « CSP », « ancienneté client » et « âge ».

2.2. Transformation de la base de données des dépôts

La figure suivante illustre les transformations effectuées sur la base de données des dépôts, que nous détaillerons par la suite.

```
def preprocessing_base_depot(self,base_depot):
    #Traitement
    base_depot=base_depot.remove_espaces(base_depot)
    base_depot =base_depot[['CODECLIENT','TYPE_DE_DEPOT','ENCOURS']]

    #Imputer les encours négatifs par 0
    base_depot['ENCOURS'] = np.where(base_depot['ENCOURS'] < 0, 0, base_depot['ENCOURS'])

    #Regroupement des encours par codeclient et type de dépôt
    base_depot=base_depot[base_depot['TYPE_DE_DEPOT']!='Type de dépôt']
    base_depot=base_depot.groupby(['CODECLIENT','TYPE_DE_DEPOT']).sum().reset_index()

    #Pivoter la base
    base_depot=base_depot.pivot(index='CODECLIENT',columns='TYPE_DE_DEPOT',values='ENCOURS')
    base_depot=base_depot.reset_index()
    base_depot=base_depot.fillna(0)

    #Renommer les colonnes
    base_depot.rename(columns = {'CSE':'Encours_CSE', 'DAT':'Encours_DAT',
                                'PEL':'Encours_PEL','PEP':'Encours_PEP','PER':'Encours_PER'}, inplace = True)

    base_depot['Avoirs Contrôlés']=base_depot['Encours_CSE']+base_depot['Encours_DAT']+base_depot['Encours_PEL']
                                +base_depot['Encours_PEP']+base_depot['Encours_PER']
    base_depot.drop(['Encours_CSE','Encours_DAT','Encours_PEL','Encours_PEP','Encours_PER'],axis=1,inplace=True)
    base_depot=base_depot[['CODECLIENT','Avoirs Contrôlés']]

    return base_depot
```

Figure 39- Traitement de la base de données des dépôts

Pour la manipulation de cette base de données, nous avons procédé comme suit :

- Etant donné que la base de données des dépôts fait face aux mêmes problèmes de nomenclature des colonnes, nous avons commencé par la suppression des espaces à la fin des noms des colonnes et éliminé les colonnes commençant par le caractère « _ ».
- En second lieu, nous avons supprimé les colonnes inutiles et gardé seulement les colonne « code client », « type de dépôt » et « encours ».
- Etant donné qu'il n'est pas possible pour un client d'avoir des encours négatifs, et après consultation de la banque, nous avons remplacé ces derniers par une valeur nulle (zéro).
- Pour finir, nous avons créé la colonne « Avoir contrôlés » comportant la somme des encours de chaque client, quel que soit le type de dépôt.

→ Résultat : Table contenant les colonnes « code client » et « avoir contrôlés ».

2.3. Transformation de la base de données des opérations

La figure suivante illustre les transformations effectuées sur la base de données des opérations, que nous détaillerons par la suite.

```
def preprocessing_operation(self,base_operation):
    base_operation=self.remove_columns(base_operation)
    corpo=["C01","C02","AEI","II","PA","MNC1","MNC2","LC"]
    base_operation.drop(base_operation[base_operation['SEGMENT'].isin(corpo)].index, inplace = True)

    #Changement du type de la colonne Date_MVT en datetime et l'ajout de la colonne Année et Mois
    base_operation.DATE_MVT=pd.to_datetime(base_operation.DATE_MVT)

    #Suppression des codes devise EUR et USD puisqu'ils sont négligeable dans notre échantillon
    base_operation.drop(base_operation[base_operation['CODE_DEVISE'].isin(["EUR","USD"])].index, inplace = True)
    #On introduit la lité des opérations systèmes pour les enlever après
    systemes=['ANNULOPE','FRAIS','AGIOS PERCUS','INTERETS SERVIS']
    base_operation.drop(base_operation[base_operation['LIBELLE_STD_MVT'].isin(systemes)].index, inplace = True)

    #On s'intéresse uniquement aux 3 derniers mois
    d = datetime.datetime.strptime("2021-12-31", "%Y-%m-%d")
    d2 =( d - dateutil.relativedelta.relativedelta(months=3)) + datetime.timedelta(days=1)
    base_operation=base_operation.loc[base_operation['DATE_MVT']>d2]

    #Ajout des deux colonnes Mvt_Crediteur et MVT_Débiteur
    base_operation[['MVT_Crediteur']] = np.where(base_operation['MONTANT_MVT_CPTABLE'] > 0, base_operation['MONTANT_MVT_CPTABLE'], 0)
    base_operation[['MVT_Débiteur']] = np.where(base_operation['MONTANT_MVT_CPTABLE'] < 0, -base_operation['MONTANT_MVT_CPTABLE'], 0)

    base_operation=base_operation[['CODECLIENT','SEGMENT','MVT_Crediteur']]
    base_operation=base_operation.groupby('CODECLIENT').sum().reset_index()
    base_operation['Flux_Crediteur']=base_operation['MVT_Crediteur']/3

    base_operation.drop('MVT_Crediteur',axis=1,inplace=True)

    return(base_operation)
```

Figure 40- Traitement de la base de données des opérations

Pour le traitement de cette base de données, nous avons procédé comme suit :

- Etant donné que la base de données des opérations fait face aux mêmes problèmes de nomenclature des colonnes, nous avons commencé par la suppression des espaces à la fin des noms des colonnes.
- En second lieu, étant donné que la base de données comporte des opérations des clients de la catégorie « *corporate* » (personnes morales), nous avons procédé à l'élimination des lignes de ces dernières.
- Afin de donner suite à la suppression des lignes des clients dont les devises utilisées sont l'euro et le dollar, nous l'avons également fait pour les lignes des opérations faites en ces mêmes devises.
- Ensuite, étant donné que la base de données comporte des opérations système (frais, agios, intérêts, etc.), et que ces dernières s'avèrent inefficaces pour le modèle, nous avons procédé à leur élimination.
- Pour faire suite à la demande de la banque – pour ne garder que les clients actifs – nous n'avons gardé que les lignes des opérations des 3 derniers mois.
- La colonne « MVT COMPTABLE » fait référence à deux types de mouvement bancaires : les mouvements créditeurs (positifs) et les mouvements débiteurs (négatifs). Nous avons donc séparé ces deux colonnes pour ne garder que celle des mouvements créditeurs.
- Finalement, nous avons créé la colonne « flux créditeurs », qui représente la moyenne des mouvements créditeurs par mois.

→ Résultat : Table contenant les colonnes « code client » et « flux créditeurs ».

3. Consolidation des bases de données

Pour donner suite aux différents traitements des trois bases de données fournies, ces dernières doivent être regroupées en une même base afin de faciliter le processus de développement : c'est ce qu'on appelle la consolidation des données, un processus consistant à combiner les données provenant de différentes sources, à les nettoyer et à les vérifier en supprimant les erreurs et à les stocker dans un emplacement unique. En regroupant toutes les données en un même endroit, une vue degré 360 est générée afin de faciliter l'observation, examiner les tendances, accélérer l'exécution des processus et simplifier l'accès aux informations.

La figure suivante montre l'étape de consolidation des bases, que nous détaillerons.

```
def consolidation_base(self,base_client,base_depot,base_operation):
    data_consolide=pd.merge(base_client,base_depot,on='CODECLIENT')
    data_consolide=pd.merge(data_consolide,base_operation,on="CODECLIENT")
    #Create a set storing outliers
    outliers = set()

    #Calculate z_scores for age, work_experience, and family_size
    scores = pd.DataFrame(columns = ['CODECLIENT','Age'])
    scores['CODECLIENT'] = data_consolide['CODECLIENT']
    for var in ['Age']:
        scores[var] = np.abs(stats.zscore(data_consolide[var]))
        scores[var] = np.abs(stats.zscore(data_consolide[var]))
        scores[var] = np.abs(stats.zscore(data_consolide[var]))
    #Find and remove outliers
    for i, row in scores.iterrows():
        if np.max(row[['Age']]) > 3:
            outliers.add(row['CODECLIENT'])
    data_consolide = data_consolide[data_consolide['CODECLIENT'].isin(outliers)==False]

    particulier=['GRP', 'JEU', 'TRAD', 'DEV', 'FPO', 'PRE']
    professionnel=['ASS1','APRO','ART','ART1','AGR','COM1','SE','ASS2','ART2','COM2','PLS','APL']
    pro=data_consolide[data_consolide['SEGMENT_CLIENT'].isin(professionnel)]
    part=data_consolide[data_consolide['SEGMENT_CLIENT'].isin(particulier)]
    part['Age']=np.where(part['Age']<30,"< 30 ans",np.where(part['Age']<45 , " 30-45 ans",> 45 ans"))
    return(part,pro)
```

Figure 41- Consolidation des bases de données

Comme le montre la figure précédente, nous avons procédé comme suit :

- Tout d'abord, nous avons commencé par la consolidation des bases de données par code client.
- En second lieu, nous nous sommes focalisés sur la détection des valeurs aberrantes avec la méthode « z_scores » ainsi que leur élimination.
- Ensuite, étant donné que la recommandation diffère de la catégorie « particuliers » à la catégorie « professionnels », nous avons séparé les clients de ces deux dernières en une base de données chacune.

Pour finir, nous avons réparti les clients en tranche d'âge afin de faciliter le travail.

4. Choix du modèle

Comme expliqué au sein de la partie – 2.1. Machine Learning – du chapitre 2 : « Identification des besoins et de l'environnement », il nous faut d'abord choisir le bon modèle pour la réalisation ; et pour donner suite aux brèves descriptions mentionnées, nous mettons en place un tableau comparatif des trois différents types de clustering, afin de choisir lequel nous allons adopter.

Méthode	Clustering basé sur le partitionnement	Clustering basé sur la hiérarchie	Clustering basé sur la densité
Principe	S'appuie sur une mesure de distance pour créer k clusters en minimisant la distance entre les voisins d'un même cluster.	Partitionner le dataset hiérarchiquement en agrégant à chaque étape les deux clusters les plus proches.	Diviser les points en k clusters, homogènes et compacts, selon la densité, en s'appuyant sur 2 paramètres : la distance ϵ (la distance minimale entre deux voisins) et MinPts le nombre minimum de voisins pour former un cluster.
Avantages	- Grande simplicité/Rapidité - Non affecté par les densités variables des points de données. - Plus efficace pour les grands datasets. Peu gourmandes en calcul.	- Plus simple de définir le nombre de clusters à partir du dendrogramme.	- Détecte et isole de lui-même les valeurs aberrantes. - Pas besoin de prédefinir le nombre de clusters
Inconvénients	- Besoin de définir le nombre de clusters.	- Non adapté à un grand volume de données : les temps de calcul explosent.	- Incapable de travailler avec une base multi-dimensionnelle (plusieurs attributs). - Sensible au choix de ϵ et de MinPts → Risque d'erreur.

Tableau 8- Comparaison des modèles

Nous disposons d'une quantité de données très volumineuses, notre projet requiert une méthode simple, rapide et capable de prendre en charge une base multi-dimensionnelle. Nous allons donc opter pour le modèle K-means pour le développement de notre système.

4.1. Le modèle K-means

Le clustering K-means est un type d'apprentissage non supervisé utilisé lorsque les données ne sont pas étiquetées. Son but est de classifier les données en k groupes, appelés clusters, et fonctionne de manière itérative pour attribuer chaque point de données à l'un des k clusters en fonction des caractéristiques fournies. Il est utilisé pour trouver des groupes qui n'ont pas été explicitement étiquetés dans les données. Une fois l'algorithme exécuté et les clusters définis, toute nouvelle donnée peut être facilement affectée au bon groupe.

Il s'agit d'un algorithme polyvalent qui peut être utilisé pour :

- La segmentation comportementale (historique d'achat, activités, des intérêts, etc.)
- La catégorisation d'inventaire (regroupement des stocks par activité de vente, métrique de fabrication, etc.)
- Le tri des mesures des capteurs (détection des types d'activité des capteurs de mouvement, etc.)
- La détection de bots ou d'anomalies (regrouper d'activités pour nettoyer la détection des valeurs aberrantes, etc.)

Fonctionnement de l'algorithme :

Les paramètres de K-means sont le nombre de clusters K et l'ensemble de données. L'algorithme commence par des estimations initiales pour les centroïdes K sélectionnés au hasard dans l'ensemble de données. L'algorithme itère ensuite entre deux étapes :

- 1- Étape d'affectation des données : Chaque centroïde définit l'un des clusters, où, chaque point est affecté à son centre de gravité le plus proche, en fonction de la distance euclidienne au carré.
- 2- Étape de mise à jour du centroïde : Les centroïdes sont recalculés en fonction de la moyenne de tous les points de données attribués au cluster de ce centroïde.

L'algorithme itère entre les étapes 1 et 2 jusqu'à ce qu'un critère d'arrêt soit satisfait : aucun point de données ne change de cluster, la somme des distances est minimisée ou un certain nombre maximum d'itérations est atteint.

Choix du nombre de clusters k :

En général, il n'y a pas de méthode pour déterminer la valeur exacte de K, mais une estimation précise peut être obtenue en exécutant l'algorithme de clustering pour une plage de valeurs k et comparer les résultats. La mesure la plus utilisée pour cela est la distance moyenne entre les points de données et leur centroïde de cluster. Mais, étant donné que l'augmentation du nombre de clusters réduira toujours la distance entre points de données, l'augmentation de k réduira toujours cette métrique. Ainsi, cette métrique ne peut pas être utilisée comme seul caractéristique.

Au lieu de cela, selon la courbe de la distance moyenne au centre de gravité en fonction de k, le elbow point (point du coude), où le taux de diminution se déplace brusquement, peut être utilisé pour déterminer approximativement k.

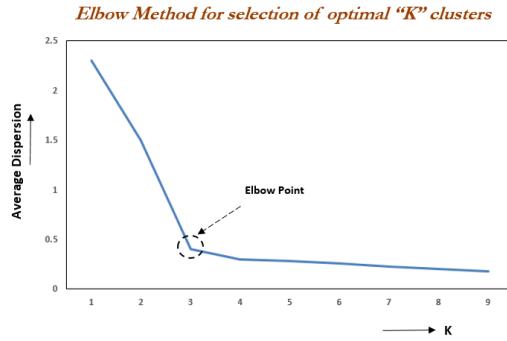


Figure 42- Elbow method pour la détermination du nombre de clusters

5. Application de K-means

Nous arrivons maintenant à la partie de l'application de l'algorithme K-means, comme le montre la figure suivante :

```
def kmeans_part(self,particulier):
    df=particulier.copy()
    df.viz = df.copy(deep=True)
    df.drop(['SEGMENT_CLIENT'],axis=1,inplace=True)
    #encode categorical variables
    le = LabelEncoder()
    cat_var = ['GENRE', 'STATUT_MATRIMONIAL']
    for var in cat_var:
        df[var] = le.fit_transform(df[var])
    df = pd.concat([df, pd.get_dummies(df['NOTATION_INTERNE_DU_CLIENT'])], axis=1)
    df = df.drop(['NOTATION_INTERNE_DU_CLIENT'], axis=1)
    df = pd.concat([df, pd.get_dummies(df['CSP'])], axis=1)
    df = df.drop(['CSP'], axis=1)
    df = pd.concat([df, pd.get_dummies(df['Age'])], axis=1)
    df = df.drop(['Age'], axis=1)
    #normalize data with MinMaxScaler
    mms = MinMaxScaler()
    X = df.drop(['CODECLIENT'], axis=1)
    X_mms = mms.fit_transform(X)
    #reduce dimensionality with PCA
    pca = PCA(.92)
    X_pca = pca.fit_transform(X_mms)
    #create a k-means model and assign each customer to a cluster
    #Dans ce cas nous devons choisir Le nombre de cluster
    kmeans = KMeans(n_clusters=4, random_state=42)
    prediction = kmeans.fit_predict(X_pca)
    #add the cluster and PCA components to the dataframe
    df_viz['cluster'] = prediction
    return(df_viz)
```

Figure 43- Application de K-means

Comme le montre la figure précédente, nous avons procédé comme suit :

- Etant donné que le système que nous allons développer va également recommander une nouvelle segmentation des clients, nous avons éliminé la colonne des segments.
- Deuxièmement, nous avons procédé à l'encodage des variables catégorielles comme le genre, le statut matrimonial, la notation, la catégorie socio-professionnelle et l'âge ; puis à la normalisation de ces données.
- Nous avons ensuite procédé à la réduction de la dimensionnalité à l'aide de la méthode d'analyse en composantes principales (ACP) afin de ne garder que les variables pertinentes au fonctionnement du modèle.
- Pour finir, nous avons initialisé le nombre de cluster à 4 comme l'a indiqué le résultat de la « elbow method », et ajouté une colonne « cluster » à la base pour la

spécification de l'appartenance de chaque client.

6. Génération des recommandations

Cette dernière phase comprend plusieurs étapes avant la génération du résultat final, comme la récupération des produits des clients, le calcul des similarités entre les clients, la génération des produits recommandés ainsi que leur éligibilité, la recommandation d'une nouvelle segmentation et enfin le calcul du taux d'équipement estimé.

a. Récupération des produits

Comme le montre la figure suivante, nous avons tout d'abord commencé par la récupération des listes de produits par client, par segment et par pack. Ensuite nous avons supprimé les colonnes vides et inutiles pour notre modèle, pour ensuite générer un *dataframe* contenant tous les clients ainsi que leurs produits.

```
#Fonction qui retourne la liste des produits par client
def get_product_per_customer(self,ID_client,df_equipement):
    data=df_equipement[df_equipement["ID_TIERS"]==ID_client]
    product=[]
    for col in data.columns.tolist():
        if data[str(col)].values[0]==1:
            product.append(col)
    return(product)

#Fonction qui retourne la liste des produits par segment
def get_product_per_segment(self,segment,produit_segment):
    data=produit_segment[produit_segment["Segment"]==segment]
    product=[]
    for col in data.columns.tolist():
        if data[str(col)].values[0]==1:
            product.append(col)
    return(product)

#Fonction qui retourne une liste des produits par pack
def get_product_per_pack(self,pack,prod_pack):
    data=prod_pack[prod_pack["Pack"]==pack]
    product=[]
    for col in data.columns.tolist():
        if data[str(col)].values[0]==1:
            product.append(col)
    return(product)

def traitement_base_equip_pack_pdtsegment(self,equipement,produit_segment,produit_pack):
    produit_pack=self.remove_espace(produit_pack)
    produit_pack=self.remove_columns(produit_pack)
    produit_pack=produit_pack.rename(columns={"Unnamed: 0":"Pack"})
    produit_pack.Pack=produit_pack.Pack.apply(lambda x : x.strip())
    equipement=self.remove_espace(equipement)
    produit_segment=self.remove_espace(produit_segment)
    equipement=self.remove_columns(equipement)
    equipement.drop(['ID_CLIENT','Column1'],axis=1,inplace=True)
    equipement=equipement.fillna(0)

    return(equipement,produit_segment,produit_pack)

def produit_client(self,data_client,df_equipement):
    dic_prod={}
    for client in data_client.CODECLIENT.unique().tolist():
        produit_test=self.get_product_per_customer(client,df_equipement)
        dic_prod[client]=produit_test
    clients_prod=[k for k,v in dic_prod.items()]
    produit_prod=[v for k,v in dic_prod.items()]
    d={"CODECLIENT":clients_prod,"Produits":produit_prod}
    df=pd.DataFrame(d)
    return df
```

Figure 44- Récupération des produits des clients

Nous avons ensuite procédé à la génération du taux d'équipement actuel des clients, en calculant le taux des produits éligibles consommés par le client par rapport au nombre de l'ensemble des produits éligibles pour ce segment.

b. Calcul des similarités

Afin de procéder à la recommandation de produits basée sur les 100 clients les plus similaires, nous passons par une étape de calcul des similarités des clients, qui retourne pour chaque client les produits consommés, les produits recommandés totaux, ainsi que les produits recommandés éligibles, comme le montre la figure suivante.

```

def top_cosine_similarity_one_nv_äge(self,data_finale_33,data_cluster,df_equipement,produit_segment,prod_clients,prod_pack,
                                     codeclient,top_n):
    data=data_finale_33.copy()
    cluster=data[data['CODECLIENT']==codeclient]
    segment_client_cible=cluster['SEGMENT_CLIENT'].values[0]
    cluster=cluster.drop(['SEGMENT_CLIENT'],axis=1)
    cluster_number=cluster['Cluster'].values[0]
    data_cluster=data_cluster[data_cluster['Cluster']==cluster_number]
    data_cluster=data_cluster.reset_index()
    data_cluster.drop('index',axis=1,inplace=True)
    cluster_sim=data_cluster.drop(['CODECLIENT','Cluster','SEGMENT_CLIENT'],axis=1)

    le = LabelEncoder()
    cat_var = ['GENRE', 'STATUT_MATRIMONIAL']
    for var in cat_var:
        cluster_sim[var] = le.fit_transform(cluster_sim[var])

    cluster_sim = pd.concat([cluster_sim, pd.get_dummies(cluster_sim[['NOTATION_INTERNE_DU_CLIENT']])], axis=1)
    cluster_sim = cluster_sim.drop(['NOTATION_INTERNE_DU_CLIENT'], axis=1)
    cluster_sim = pd.concat([cluster_sim, pd.get_dummies(cluster_sim['CSP'])], axis=1)
    cluster_sim = cluster_sim.drop(['CSP'], axis=1)
    cluster_sim = pd.concat([cluster_sim, pd.get_dummies(cluster_sim['Age'])], axis=1)
    cluster_sim = cluster_sim.drop(['Age'], axis=1)
    cluster_sim=cluster_sim.reset_index()
    cluster_sim.drop('index',axis=1,inplace=True)

    index = data_cluster[data_cluster['CODECLIENT']==codeclient].index.tolist()[0]
    client_row = cluster_sim.iloc[index, :]
    magnitude = np.sqrt(np.einsum('ij, ij -> i', cluster_sim, cluster_sim))
    similarity = np.dot(client_row, cluster_sim.T) / (magnitude[index] * magnitude)
    sort_indexes = np.argsort(-similarity)
    top_index = sort_indexes[1:top_n]
    ar = np.array([top_index,-np.sort(-similarity)[1:top_n]])
    d = pd.DataFrame({'CODECLIENT':ar[0],'similarity':ar[1]})
    d = d[['CODECLIENT','similarity']]
    d.CODECLIENT=d.CODECLIENT.astype('int')
    for i in range(len(d)):
        d['CODECLIENT'][i]=data_cluster.iloc[d['CODECLIENT'][i],0]

    products=[]
    clients=d.CODECLIENT.unique().tolist()
    data_prod=prod_clients[prod_clients['CODECLIENT'].isin(clients)][['Produits']].tolist()
    for item in data_prod:
        for prod in item:
            if prod not in products:
                products.append(prod)
    products=[item.strip() for item in products]
    dictionnaire_segment=dict(zip(data_finale_33.CODECLIENT,data_finale_33.SEGMENT_CLIENT))
    segment_clients=data_finale_33[data_finale_33['CODECLIENT'].isin(clients)][['SEGMENT_CLIENT']].tolist()
    d.insert(2,'Segment',segment_clients)
    dictio_seg=dict(d.Segment.value_counts())
    test_seg_number=[for v,k in dictio_seg.items()]
    test_seg=[v for v,k in dictio_seg.items()]
    segment_data_finale_33[data_finale_33['CODECLIENT']==codeclient]['SEGMENT_CLIENT'].values[0]
    panier_one = list(set(products))
    product_client_cible=self.get_product_per_customer(codeclient,df_equipement)
    packs=[item for item in product_client_cible if 'pack' in item.lower()]
    panier_final=[value for value in panier_one if value not in product_client_cible]
    produit_par_pack=[]
    if len(packs) != 0:
        for pack_test in packs:
            for item in self.get_product_per_pack(pack_test,prod_pack):
                produit_par_pack.append(item)
    pack_recom=[item for item in panier_final if "pack" in item.lower()]
    if len(pack_recom) !=0:
        for pack_test in pack_recom:
            products_pack_recom= self.get_product_per_pack(pack_test,prod_pack)
            for item in products_pack_recom:
                if item in product_client_cible:
                    produit_par_pack.append(item)
    produit_par_pack=list(set(produit_par_pack))
    panier_final=[item for item in panier_final if item not in produit_par_pack ]
    segment_data_finale_33[data_finale_33['CODECLIENT']==codeclient]['SEGMENT_CLIENT'].values[0]
    seg=self.get_product_per_segment(segment,produit_segment)
    panier_final_elig=[value for value in panier_final if value in seg]

    return(produit_client_cible,panier_final,panier_final_elig)

```

Figure 45- Calcul des similarités

- La fonction commence par récupérer la ligne du client et son segment avant de supprimer ce dernier. Elle récupère ensuite le numéro du cluster afin d'extraire les clients de ce dernier.
- Ensuite, elle procède à la suppression du code client, du numéro de cluster, le segment et les variables PCA1 et PCA2 (outputs de l'algorithme K-means), étant donné que celles-ci n'interviendront pas dans le calcul, et à l'encodage des variables

notation, catégorie socio-professionnelle et âge.

- Nous allons ensuite récupérer la ligne du client (âge, CSP, notation, situation matrimonial, ancienneté client, avoirs contrôlés et flux créditeurs) et créer la matrice de similarité en multipliant la matrice du client par la matrice des autres clients.
- Comme prochaine étape, nous procéderons au tri des 100 premiers clients du plus similaire au moins similaire dans un tableau les index des clients et la valeur de leur similarité.
- Nous récupérons ensuite les codes clients ainsi que les produits qu'ils ont consommé avant de procéder au tri pour ne garder que les produits éligibles de leur segment.
- Arrivé à cette étape, et si un des produits recommandés fait partie d'un pack, nous sommes confrontés à deux cas :
 - Si le client possède déjà ce pack, alors nous supprimons le produit de la liste de recommandation,
 - Si le client ne possède pas le pack, alors nous lui recommandons, séparément, le produit ainsi que le pack.

Pour faire suite à la recommandation des produits, nous avons procédé à la génération du taux d'équipement estimé, en calculant la somme des produits éligibles consommés par le client avec les produits éligibles recommandés que le client ne possède pas, par rapport au nombre de l'ensemble des produits éligibles pour ce segment.

c. Recommandation des produits éligibles de la nouvelle segmentation

Les figures suivantes illustrent respectivement le processus de recommandation de la nouvelle segmentation suivant des critères prédéfinis (somme des avoirs contrôlés et/ ou du flux créditeur), et le processus de recommandation des produits de ce même segment cible.

```
def new_segment_critère_part(self,data_client):
    new_segment=[]
    for client in data_client.CODECLIENT.tolist():
        item=''
        age=data_client[data_client['CODECLIENT']==client]['Age'].values[0]
        flux=data_client[data_client['CODECLIENT']==client]['Flux_Créditeur'].values[0]
        ac=data_client[data_client['CODECLIENT']==client]['Avoirs Contrôle'].values[0]
        if age=='< 30 ans':
            if ac > _____ or flux > _____:
                item='PRE'
            else:
                item='JEU'
        elif age=='30-45 ans':
            if ac > _____ or flux > _____:
                item='PRE'
            elif (ac > _____ and ac<=_____ ) or (flux>_____ and flux <=_____):
                item='FPO'
            elif (ac > _____ and ac<=_____ ) or (flux>_____ and flux <=_____):
                item='DEV'
            else:
                item='GRP'
        else:
            if ac > _____ or flux > _____:
                item='PRE'
            elif (ac > _____ and ac<=_____ ) or (flux>_____ and flux <=_____):
                item='FPO'
            elif (ac > _____ and ac<=_____ ) or (flux>_____ and flux <=_____):
                item='TRAD'
            else:
                item='GRP'
        new_segment.append(item)
    data_client['Segment Cible']=new_segment
    return(data_client)
```

Figure 46- Recommandation d'une nouvelle segmentation

```

def produit_cible(self,data_client,produit_segment):
    prod_cible=[]
    for client in data_client.CODECLIENT.tolist():
        products=[]
        segment_cible=data_client[data_client['CODECLIENT']==client]['Segment Cible'].values[0]
        segment_cible=segment_cible.strip()
        prod=data_client[data_client['CODECLIENT']==client]['Produits Recommandés( tot )'].values[0]
        product_cible=self.get_product_per_segment(segment_cible,produit_segment)

        for elem in prod :
            if elem in product_cible :
                products.append(elem)
        prod_cible.append(products)
    data_client["Produits Recommandés Cibles"]=prod_cible

    return(data_client)

```

Figure 47- Recommandation de produits de la segmentation cible

Pour faire suite à la recommandation des produits du segment cible, nous avons procédé à la génération du taux d'équipement cible, en calculant la somme des produits éligibles consommés par le client avec les produits éligibles recommandés que le client ne possède pas, par rapport au nombre de l'ensemble des produits éligibles pour ce segment.

7. Résultat

Finalement, la figure suivante représente le résultat généré par le modèle.

CODECLIENT	NOTATION_IN_TERNE_DU_CLE_IENT	GENRE	SEGMENT_CLIENT	STATUT_MATRIMONIAL	CSP	MONTHS_AS_CLIE_NT	Age_Disc	Flux_CrédiTe_ur	Avoir contrôles	Taux d'équipement tactual (%)	Produits Disponibles	Produits Recommandés (tot)	Produits Recommandés (éligibles)	Taux d'équipement estimé (%)	Segment Cible	Produits recommandé s Cibles	Taux d'équipement Cible (%)
Acceptable	M	Prestige	Marie	820	338 > 45	53	2597191,33	6432	25	25	['SMS', 'CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'ASSU', 'VUE']	['CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'ASSU', 'VUE']	62,5	Prestige	['CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'ASSU', 'VUE']	62,5	
Mediocre	M	Senior	Marie	510	396 > 45	64	942478	749401	31,25	25	['SMS', 'CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'ASSU', 'VUE']	['CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'ASSU', 'VUE']	62,5	Senior	['CONSO', 'DAT', 'GOLD', 'PACK', 'ASSU', 'VUE']	62,5	
Mediocre	F	Senior	Cellibataire	950	383 > 45	61	3637951,33	274891	31,25	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	62,5	Prestige	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	62,5	
Acceptable	F	Haut de	Marie	351	409 > 45	61	995366	12347636	43,75	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	62,5	Haut de	['GOLD', 'PACK', 'CONSO', 'DAT', 'VUE']	62,5	
Mediocre	M	Prestige	Cellibataire	440	263 > 45	59	8795076	50817378	37,5	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	62,5	Prestige	['GOLD', 'PACK', 'CONSO', 'DAT', 'VUE']	62,5	
Mediocre	M	Senior	Marie	351	298 > 45	63	1163175	1354889	37,5	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	62,5	Senior	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	62,5	
Mediocre	F	Senior	Veuf	510	242 > 45	59	259142	8385218	37,5	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	50	Senior	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	50	
Mediocre	M	Senior	Marie	351	295 > 45	54	1486469,33	322051	37,5	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	62,5	Haut de	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	68,75	
Mediocre	F	Senior	Marie	440	306 > 45	59	2018130	5126338	25	25	['SMS', 'CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	['CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	56,25	Haut de	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	62,5	
Acceptable	M	Senior	Cellibataire	430	240 30-45	43	820038,667	1282706	25	25	['SMS', 'CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	['CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	56,25	Haut de	['CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	56,25	
Acceptable	F	Prestige	Marie	358	248 > 45	52	13033858	5415169	37,5	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	68,75	Prestige	['PLATINUM', 'GOLD', 'PACK', 'CONSO', 'VUE']	68,75	
Moyen	M	Prestige	Marie	950	227 > 45	61	1918109,67	1069161	56,25	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	68,75	Haut de	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	75	
Mediocre	F	Haut de	Marie	530	238 > 45	47	2558366,67	0	43,75	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	75	Prestige	['CSE', 'CONSO', 'DAT', 'GOLD', 'VUE']	75	
Mediocre	M	Moyenne	Marie	510	170 > 45	53	0	17501	13,33	25	['A_VUE', 'CSE', 'PE']	['A_VUE', 'CSE', 'PE']	46,67	Moyenne	['ASSU', 'CONSO', 'VUE']	46,67	
Acceptable	M	Grand	Cellibataire	355	118 30-45	42	8155017	600453	43,75	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	62,5	Prestige	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	62,5	
Faible	M	Moyenne	Marie	355	149 > 45	47	0	54656	13,33	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	33,33	Moyenne	['A_VUE', 'SMS', 'VUE']	33,33	
Mediocre	M	Senior	Cellibataire	354	263 > 45	53	566666,667	1810114	18,75	25	['AUTRES_CARTES', 'CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	['AUTRES_CARTES', 'CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	68,75	Senior	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	68,75	
Mediocre	F	Senior	Marie	440	306 > 45	59	2018130	5126338	25	25	['SMS', 'CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	['CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	56,25	Haut de	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	62,5	
Acceptable	M	Senior	Cellibataire	430	240 30-45	43	820038,667	1282706	25	25	['SMS', 'CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	['CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	56,25	Senior	['CSE', 'CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	56,25	
Acceptable	F	Prestige	Marie	358	248 > 45	52	13033858	5415169	37,5	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	68,75	Prestige	['PLATINUM', 'GOLD', 'PACK', 'CONSO', 'VUE']	68,75	
Moyen	M	Prestige	Marie	950	227 > 45	61	1918109,67	1069161	56,25	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	68,75	Haut de	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	75	
Mediocre	F	Haut de	Marie	530	238 > 45	47	2558366,67	0	43,75	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	75	Prestige	['CSE', 'CONSO', 'DAT', 'GOLD', 'VUE']	75	
Mediocre	M	Moyenne	Marie	510	170 > 45	53	0	17501	13,33	25	['A_VUE', 'CSE', 'PE']	['A_VUE', 'CSE', 'PE']	46,67	Moyenne	['ASSURPLUS', 'CONSO', 'VUE']	46,67	
Acceptable	M	Haut de	Cellibataire	355	118 30-45	42	8155017	600453	43,75	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	62,5	Prestige	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	62,5	
Faible	M	Moyenne	Marie	355	149 > 45	47	0	54656	13,33	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	33,33	Moyenne	['A_VUE', 'SMS', 'VUE']	33,33	
Mediocre	M	Senior	Cellibataire	354	263 > 45	53	566666,667	1810114	18,75	25	['SMS', 'CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	['CSE', 'PACK', 'CONSO', 'DAT', 'GOLD', 'VUE']	68,75	Senior	['CONSO', 'DAT', 'GOLD', 'PACK', 'VUE']	68,75	

Figure 48- Résultat du modèle

Conclusion

Au sein de ce chapitre, nous avons présenté les différentes étapes du développement du système de recommandation de produits aux clients, en passant par les étapes du choix du modèle et de la méthode adéquate, jusqu'à l'étape de développement et la génération du résultat. Le chapitre suivant sera alors consacré au développement de la plateforme dédiée à l'exposition des résultats générés par le modèle, afin de garantir une utilisation claire et simple aux administrateurs.

Chapitre 5 : Développement de la plateforme

Introduction

Après avoir élaboré la conception, nous abordons dans ce chapitre le dernier volet de ce rapport, qui a pour objectif d'exposer la phase de réalisation, considérée comme étant la concrétisation finale de toute la méthode de conception.

1. Conception de la plateforme

Il est nécessaire d'éclaircir au mieux les besoins fonctionnels, présenter les acteurs et identifier les cas d'utilisation ainsi que les classes afin de comprendre au mieux le concept du projet. Au sein de cette partie, nous présentons la première étape de la phase de développement qu'est la conception, y compris le langage UML et les diagrammes utilisés.

1.1. Le langage UML

Pour programmer un système, il ne suffit pas de se lancer tout de suite dans la phase de pratique : il faut d'abord organiser ses idées, les décrire, les lier entre elles afin de se faciliter le travail. Modéliser un système avant sa réalisation permet de mieux comprendre son fonctionnement, le but étant de maîtriser sa complexité en le transformant en graphiques, et donc établir une vision globale du produit.

Pour ce faire, nous avons utilisé le langage UML, acronyme anglais pour « *Unified Modeling Language* » qu'on traduit par « Langage de modélisation unifié ». Ce dernier est un langage visuel constitué d'un ensemble de schémas explicatifs, les diagrammes, qui présentent chacun une étape du projet : son fonctionnement et les différentes actions susceptibles d'être effectuées par les acteurs de la plateforme.

Un acteur représente une entité externe qui interagit directement ou indirectement avec le système étudié. Notre projet consiste en le développement d'une plateforme qui donne l'accès seulement à l'administrateur. Il va donc interagir avec un seul acteur.

1.2. Les diagrammes utilisés

Tout comme la construction d'une maison, la réalisation d'un site web nécessite un plan, qui dans notre cas, est représenté par plusieurs diagrammes. À ce jour, il existe plusieurs diagrammes en UML, mais nous n'allons en présenter que les essentiels pour

clarifier l'usage de notre plateforme.

Pour la conception, nous avons utilisé les diagrammes suivants :

- Diagramme des cas d'utilisation : représentation des possibilités d'interaction entre le système et les acteurs.
- Diagramme de séquence : représentation de façon séquentielle du déroulement des interactions entre les éléments du système et/ou de ses acteurs.

a. Diagramme de cas d'utilisation

La figure ci-dessous représente le diagramme de cas d'utilisation global.

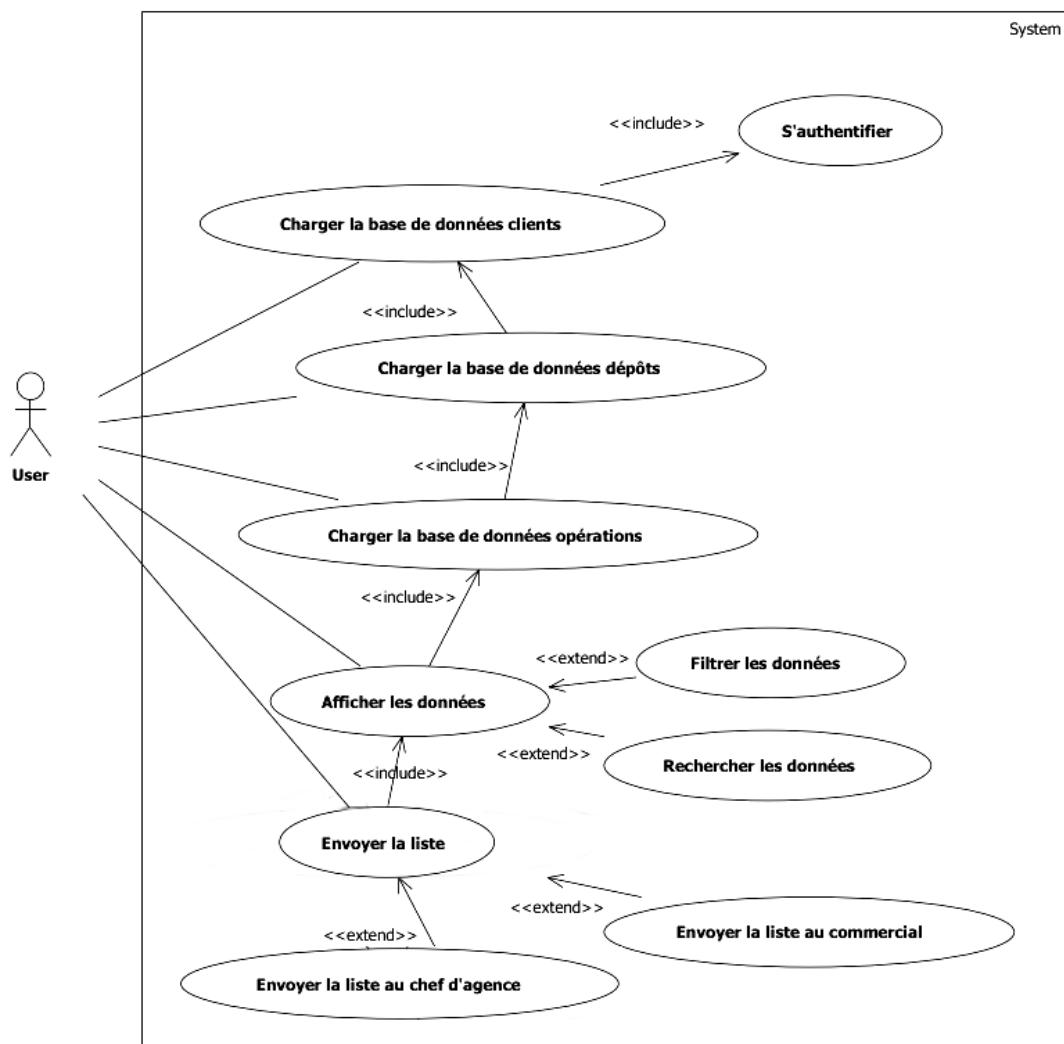


Figure 49- Diagramme de cas d'utilisation global

Ce diagramme regroupe tous les cas d'utilisation relativs à notre système : comme première étape, l'utilisateur s'authentifie en fournissant ses informations de connexion ; ensuite, il devra importer les trois bases de données demandées (la base de données des clients, la base de données des dépôts bancaires et la base de données des opérations

bancaires). Le résultat du modèle est ensuite généré dans un tableau, où l'utilisateur peut filtrer les résultats, avant d'envoyer au chef d'agence la liste des clients de cette dernière, ou bien d'envoyer à chaque commercial la liste des clients desquels il est chargé.

b. Diagramme de séquence

La figure ci-dessous représente le diagramme de séquence qui comporte un acteur : l'administrateur, ainsi que trois objets, l'interface d'authentification, l'interface de chargement des bases de données et l'interface du résultat. L'opération réalisée est la connexion de l'utilisateur et le chargement des bases de données.

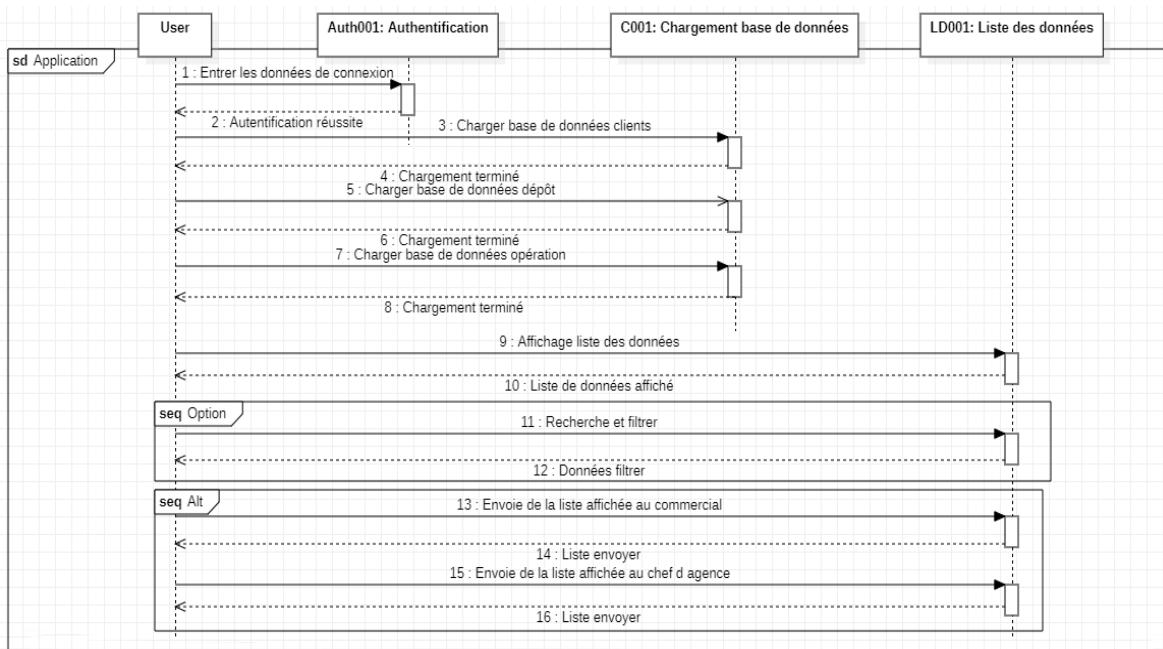


Figure 50- Diagramme de séquences

2. Architecture de la solution

Cette section vise à présenter et à justifier l'architecture physique sur laquelle nous nous sommes basés pour notre solution, afin de présenter quelques éléments de base pour comprendre comment l'application fonctionne avant de commencer le développement.

La figure 51 illustre l'architecture technique de la plateforme développée en suivant le processus suivant :

- Une fois l'URL saisie par l'utilisateur dans la barre d'adresse du navigateur Web, le serveur envoie le fichier au navigateur sur demande. Le navigateur exécute ensuite ces fichiers pour afficher la page demandée.
- Le modèle aide à gérer la base de données. C'est une couche d'accès aux données qui a pour fonctionnalité de fournir l'interface pour les données stockées dans la base de données en interagissant avec cette dernière via un ORM (Object Relational Mapping, un programme informatique qui se place entre la couche de stockage des données et la

couche applicative). Tous les modèles sont réunis dans un fichier python « *models.py* ».

- La vue est le pont entre les données du modèle et les modèles eux-mêmes. Il est géré pour exécuter la logique métier et interagir avec un modèle pour fournir des données et restituer un modèle. Elle reçoit une requête HTTP et renvoie une réponse HTTP convenable. Les vues se trouvent dans le fichier « *views.py* ».
- Le template est une couche de présentation qui gère entièrement la partie interface utilisateur. Il gère toutes les parties statiques de la page Web ainsi que le HTML, que les utilisateurs qui demandent la page Web percevront.

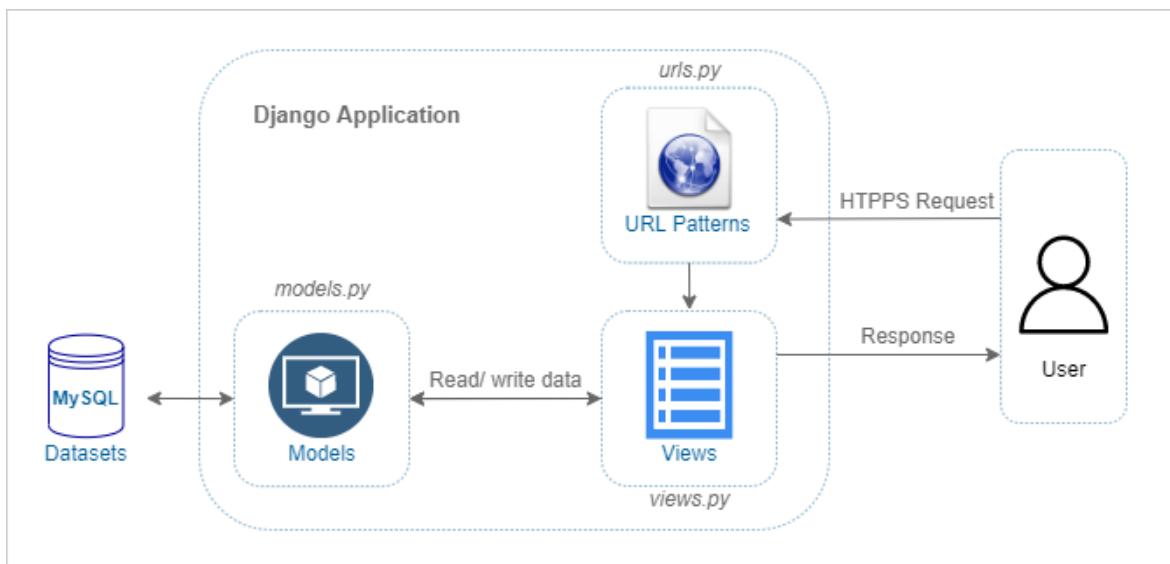


Figure 51- Architecture de l'application

3. Conception de la maquette

Cette phase vise à purifier les processus et doit donc aboutir à une maquette, une étape indispensable pour se faire une idée plus précise de l'utilisation du système. Le maquettage est une méthode de conception d'interface qui permet d'avoir des interfaces conformes aux attentes et aux besoins du client, en disposant avec précision les éléments graphiques sur différentes pages, travailler la dimension de ceux-ci, définir l'emplacement des icônes, etc. Cela permet également, de valider ou de corriger des choix fonctionnels ou techniques avant d'entrer dans la phase de réalisation.

Nous présentons ci-dessous quelques-unes des interfaces dont nous avons réalisé la maquette avant de commencer notre travail.

Les figures suivantes représentent respectivement les maquettes de la page d'accueil de la plateforme, ainsi que de la page d'authentification.

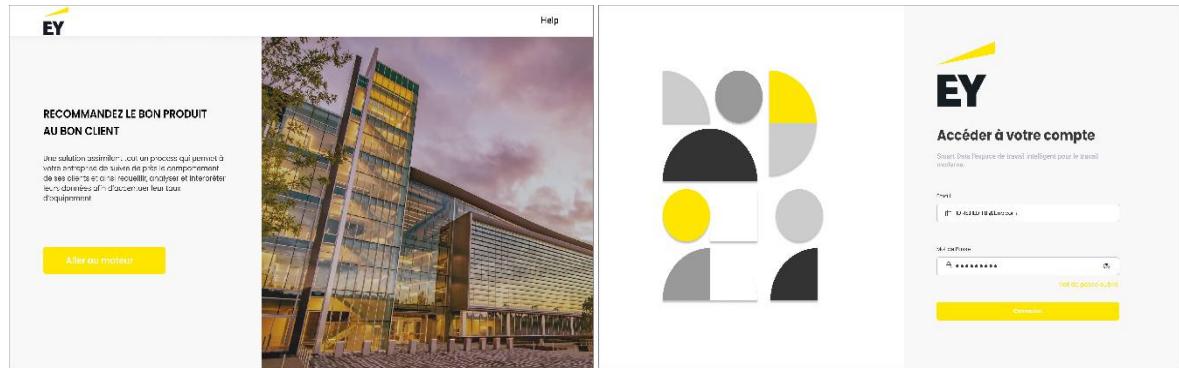


Figure 52- Maquette des pages Accueil et Authentification

Les figures suivantes représentent les maquettes des pages de chargement des bases de données, notamment la base de données des clients, la base de données des opérations bancaires, ainsi que la base de données des dépôts bancaires, avant de conduire au résultat généré par le système.

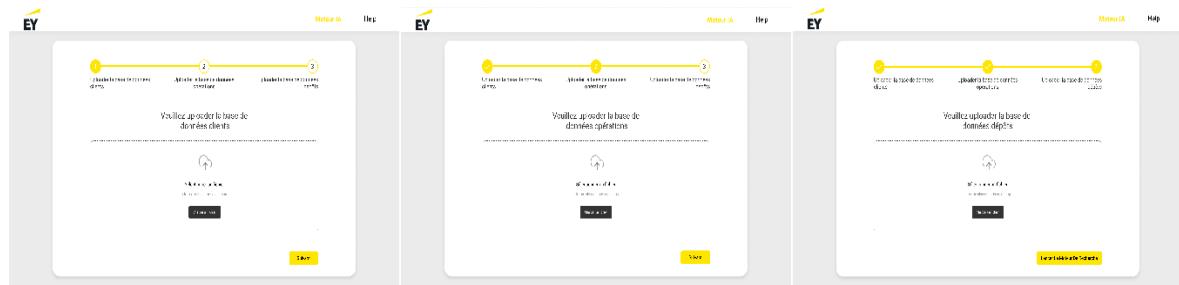


Figure 53- Maquette des pages d'Upload

Finalement, les figures suivantes présentent le tableau résultant du moteur contenant les différentes informations clients fournies, ainsi que les recommandations de segmentation et de produit, et le calcul des taux d'équipement.

Dashboard Moteur IA Help									
Modifier la liste de clients									
Objet	Statut segment client	Statut segment recommandé	Agence	Entité	Commercial associé	Produits achetés clients par secteur	Produits recommandés clients par secteur	Taux d'équipement client	Taux d'équipement client recommandé
Personne 1	HBO	GRI	Agence 1	Agence 1	Commercial 1	GOLD	BANK NE	100%	100%
Personne 2	CM	STN	Agence 2	Agence 2	Commercial 2	PLATINUM	BANK MOBILE	100%	100%
Personne 3	HBO	HNG	Agence 3	Agence 3	Commercial 3	H&I AI	SMS	100%	100%
Personne 4	IIG	PRE	Agence 4	Agence 4	Commercial 4	SMS	ASSURANCE	100%	100%
Personne 5	PV	ART	Agence 5	Agence 5	Commercial 5	ASSURANCE	HABITAT	100%	100%
Personne 6	CM	DEV	Agence 6	Agence 6	Commercial 6	BANK MOBILE	ASSURANCE	100%	100%
Personne 7	JPI	APRO	Agence 7	Agence 7	Commercial 7	RANK NFT	BCD	100%	100%

Dashboard Moteur IA Help									
Modifier la liste de clients									
Objet	Statut segment client	Statut segment recommandé	Agence	Entité	Commercial associé	Produits achetés clients par secteur	Produits recommandés clients par secteur	Taux d'équipement client	Taux d'équipement client recommandé
Personne 1	HBO	GRI	Agence 1	Agence 1	Commercial 1	GOLD	BANK NE	100%	100%
Personne 2	CM	STN	Agence 2	Agence 2	Commercial 2	PLATINUM	BANK MOBILE	100%	100%
Personne 3	HBO	HNG	Agence 3	Agence 3	Commercial 3	H&I AI	SMS	100%	100%
Personne 4	IIG	PRE	Agence 4	Agence 4	Commercial 4	SMS	ASSURANCE	100%	100%
Personne 5	PV	ART	Agence 5	Agence 5	Commercial 5	ASSURANCE	HABITAT	100%	100%
Personne 6	CM	DEV	Agence 6	Agence 6	Commercial 6	BANK MOBILE	ASSURANCE	100%	100%
Personne 7	JPI	APRO	Agence 7	Agence 7	Commercial 7	RANK NFT	BCD	100%	100%

Figure 54- Maquette de la page de résultat

4. Développement de la plateforme

Après avoir élaboré la conception, nous abordons dans cette partie le dernier volet de ce rapport, qui a pour objectif d'exposer la phase de réalisation de la plateforme, considérée comme étant la concrétisation finale de toute la méthode de conception.

La figure suivante représente la page d'accueil de la plateforme contenant un bouton pour la redirection à une page d'authentification.

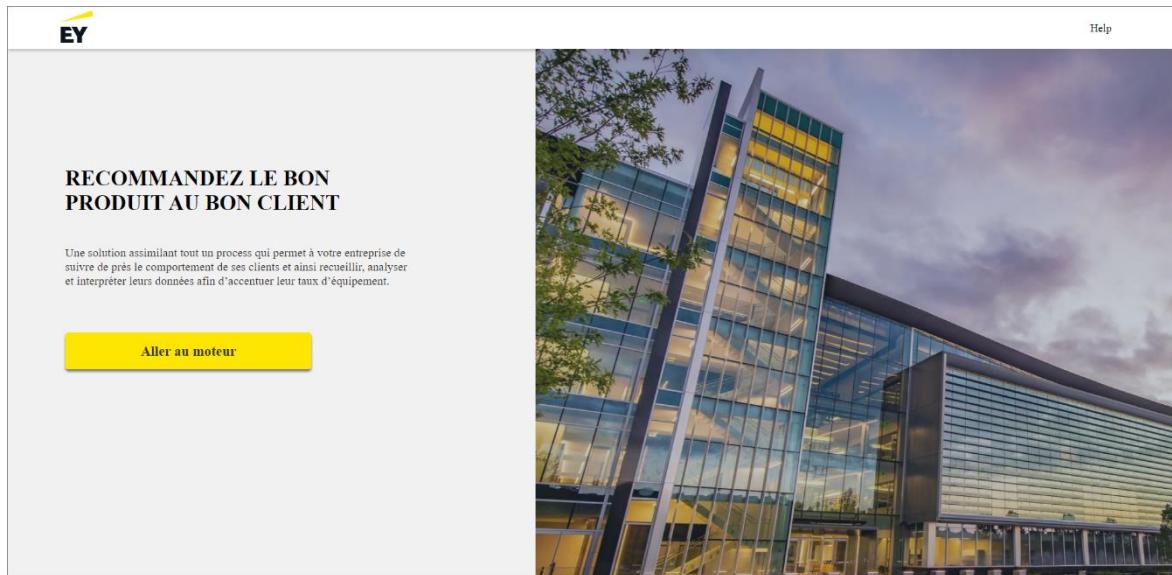


Figure 55- Page d'accueil

La page suivante est la page d'authentification, où l'utilisateur devra entrer ses informations de connexion, notamment son adresse électronique et son mot de passe, avant de se connecter.



Figure 56- Page d'authentification

L'utilisateur sera ensuite redirigé vers les pages de chargement des bases de données avant de mettre en marche le modèle. Il va d'abord importer la base de données des clients, ensuite celle des dépôts, et enfin celle des opérations.

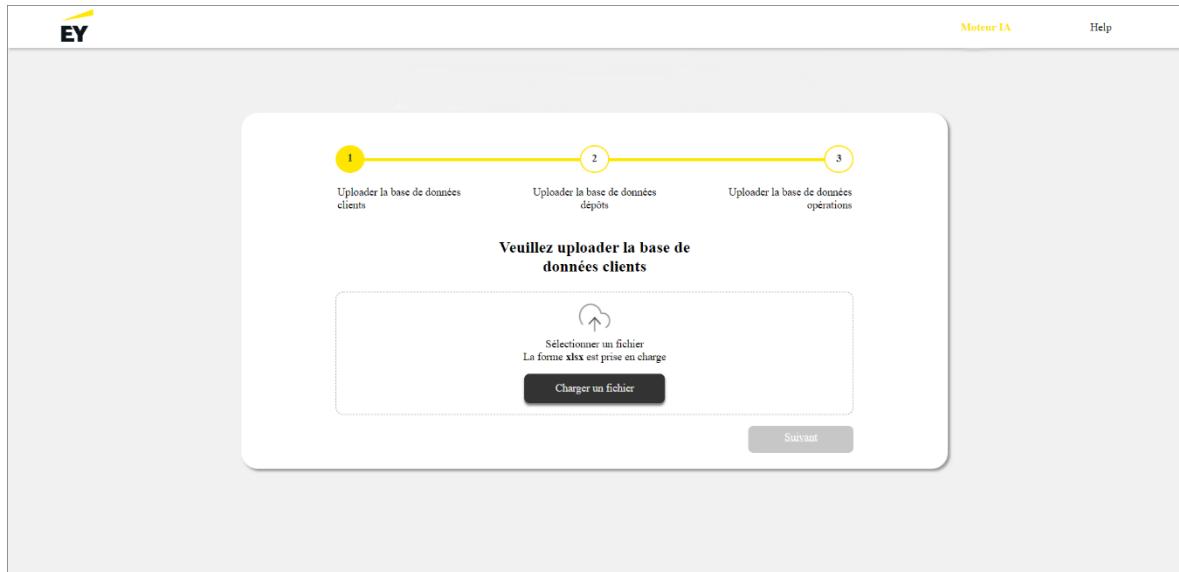


Figure 57- Page de chargement de la base de données des clients

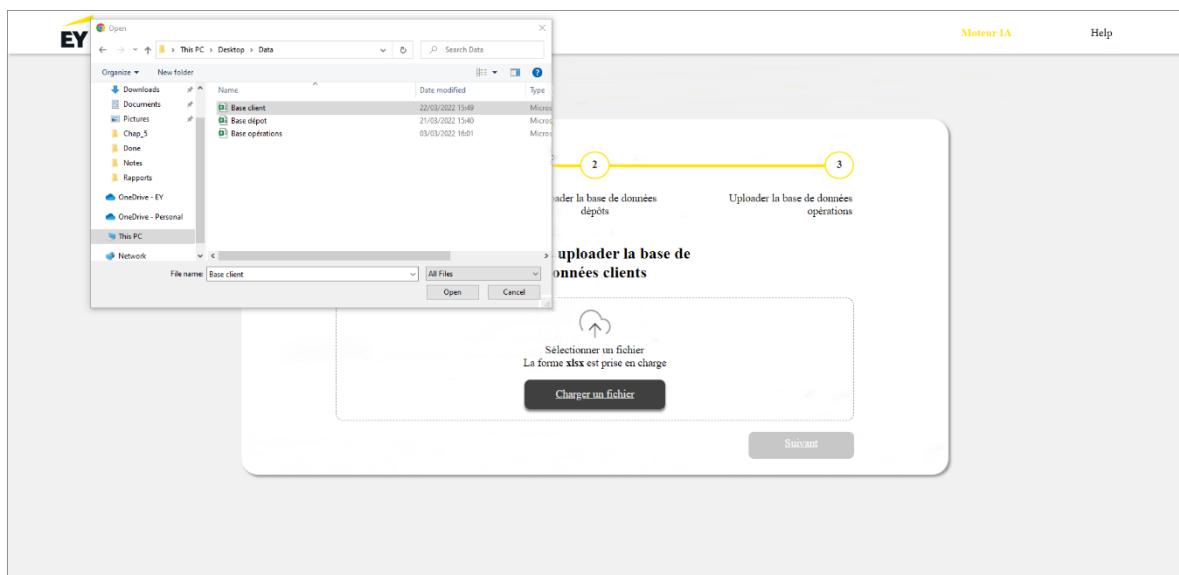


Figure 58- Etape de chargement de la base de données des clients

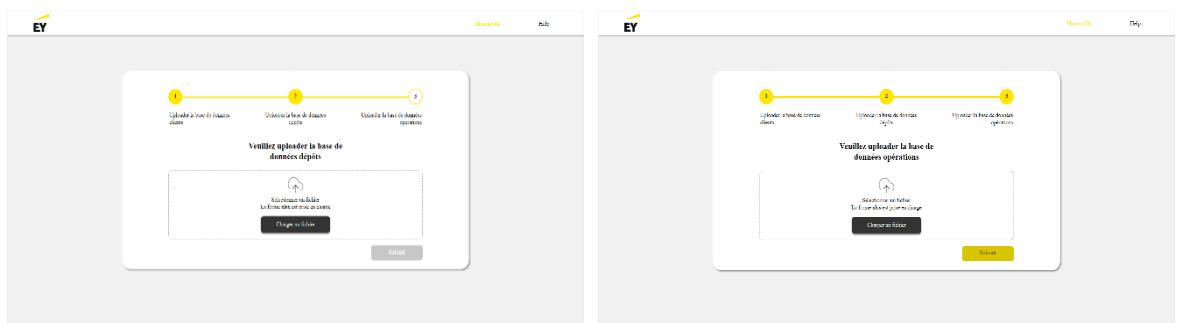


Figure 59- Page de chargement des bases de données des dépôts et des opérations

La dernière interface de ce cas d'utilisation est celle présentant le résultat généré par le modèle une fois les différentes bases de données importées. Le tableau comporte 10 colonnes et autant de lignes que le nombre de clients dont nous disposons.

Les colonnes sont les suivantes :

- Client : contient le code du client.
- Segment actuel : contient le segment auquel le client fait partie.
- Segment recommandé : contient le segment recommandé par le modèle selon des critères bien définis.
- Produits actuels : liste les produits consommés par le client.
- Produits recommandés : liste les produits recommandés au client.
- Taux d'équipement actuel : contient le taux d'équipement du client avant la recommandation des produits.
- Taux d'équipement cible : contient le taux d'équipement estimé si le client consomme tous les produits recommandés.
- Produits vendus : contient un menu déroulant destiné au commercial, listant tous les produits recommandés et offrant la possibilité de cocher les produits consommés.
- Client contacté : contient une case à cocher si le client a été contacté ou non.
- Commentaire : pour donner la possibilité au commercial de laisser un commentaire concernant le processus de recommandation, l'avis du client, et d'éventuels avis ou suggestions à prendre en compte.

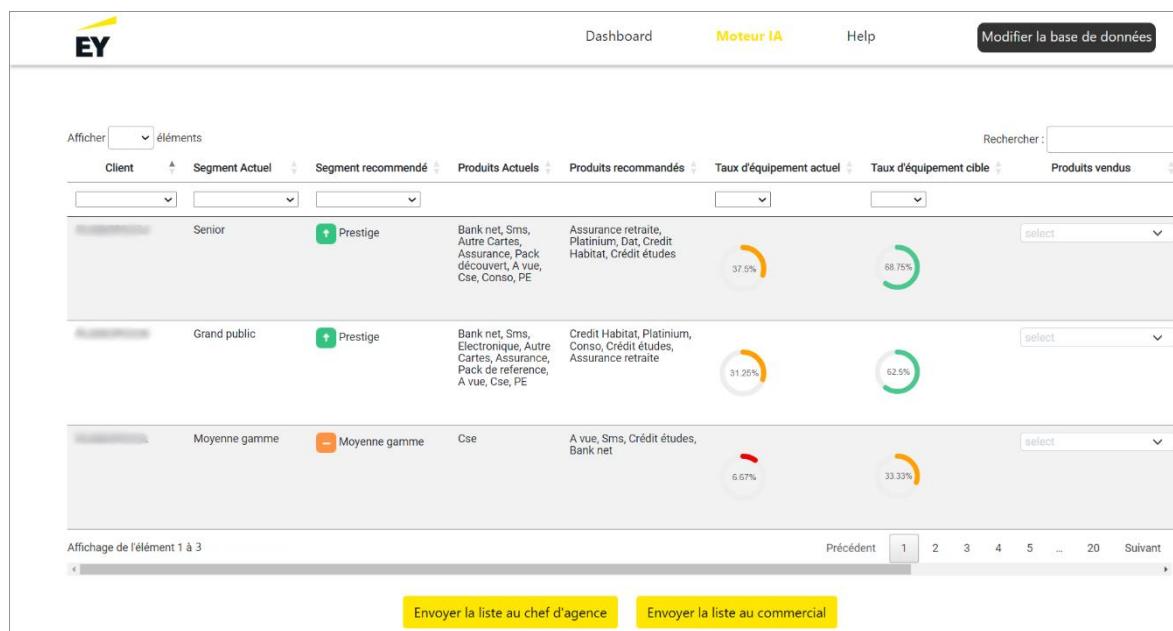


Figure 60- Page de résultat du modèle - 1

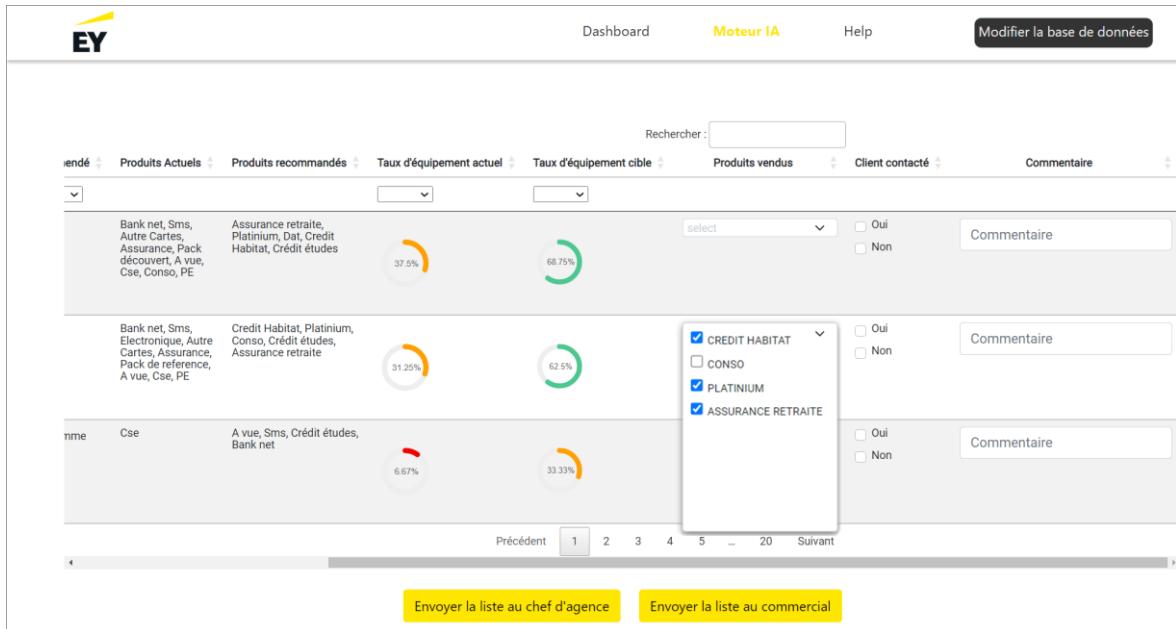


Figure 61- Page de résultat du modèle - 2

Conclusion

Tout au long de ce chapitre, nous avons présenté les principaux résultats de cette dernière phase de réalisation allant de la conception jusqu'aux différentes interfaces de la plateforme en illustrant ces dernières par quelques figures. Nous passons dans ce qui suit à une conclusion générale.

Conclusion

Le présent rapport a été rédigé dans le cadre de mon projet de fin d'études à la suite d'un stage de 6 mois effectué au cabinet Ernst & Young pour l'obtention du diplôme d'ingénieur en information de l'école supérieure privée d'ingénierie et de technologie.

J'ai eu l'occasion de participer à un projet de grande envergure et de travailler avec une équipe tout aussi professionnelle que coopérative et dévouée, et ce document présente les principaux enseignements et les principales conclusions issues de ce projet, qui m'a permis d'acquérir de nouvelles compétences et de mettre en application les connaissances théoriques acquises durant mon cursus.

Ce stage m'a aussi permis d'acquérir de nouvelles compétences organisationnelles ainsi que techniques grâce aux formations que j'ai suivies, aux réunions auxquelles j'ai eu l'occasion de participer et à l'environnement de travail au sein duquel j'ai évolué. J'ai notamment apprécié la diversité des tâches effectuées, loin d'être routinières, allant de la visualisation des données jusqu'au développement de la plateforme, en passant par le développement du modèle.

L'élaboration n'a pourtant pas été des plus simples. Nous avons dû confronter des obstacles qui nous ont rendu la tâche un peu difficile tel que les changements à effectuer sur le modèle, le développement de la plateforme, l'ajout de fonctionnalités, l'adaptation aux changements à chaque étape du projet et notamment la pression des deadlines. Cette expérience m'a donc permis de faire mes premiers pas dans le monde professionnel tout en étant confronté aux difficultés réelles du monde du travail, et, malgré les difficultés rencontrées, nous avons atteint les objectifs spécifiés au début. Et comme tout autre projet, le nôtre n'atteint pas encore la perfection, c'est pour cela que nous prévoyons d'avancer en ajoutant d'autres fonctionnalités tels que la gestion de profils et la mise en place de tableaux de bord récapitulatifs des données des clients suite à l'intervention du système dans la recommandation.

Je tiens finalement à remercier et à féliciter les membres de l'équipe avec lesquels j'ai passé ces six derniers mois pour leur professionnalisme, leur total implication dans leur mission d'encadrement, ainsi que le soutien dont ils n'ont cessé de faire preuve à mon égard. Je garde de cette expérience un agréable souvenir étant donné qu'elle constitue une expérience professionnelle valorisante et encourageante pour mon avenir m'ayant permis d'enrichir mes connaissances, mon savoir-faire et surtout mon savoir-être.

Bibliographie

[1] : <https://web.archive.org/web/20090715172635/http://www.ey.com/GL/en/About-us/Our-history>

[2] : https://www.ey.com/en_gl/about-us#our-values

[3] : [https://fr.wikipedia.org/wiki/EY_\(entreprise\)](https://fr.wikipedia.org/wiki/EY_(entreprise))

[4] : <https://ia-data-analytics.fr/machine-learning/>

[5] : <https://asana.com/fr/resources/agile-methodology>

[6] : <https://www.journaldunet.fr/web-tech/guide-de-l-entreprise-digitale/1443834-scrum-guide-de-la-methode-agile-star/>

[7] : <https://datascientest.com/preprocessing>

[8] : <https://mobiskill.fr/blog/conseils-emploi-tech/apprentissage-supervise-vs-apprentissage-non-supervise/>



ESPRIT SCHOOL OF ENGINEERING

www.esprit.tn - E-mail : contact@esprit.tn

Siège Social : 18 rue de l'Usine - Charguia II - 2035 - Tél. : +216 71 941 541 - Fax. : +216 71 941 889

Annexe : 1-2 rue André Ampère - 2083 - Pôle Technologique - El Ghazala - Tél +216 70 250 000 - Fax +216 70 685454