



## **Projet de Fin d'études**

**Titre du sujet :** Automatisation du modèle de profiling des clients basé sur les données collectées.

**Organisme d'accueil :**



Troisième année de la Licence en Informatique de gestion

**Parcours :** Business Intelligence

**Réalisé par :** Imen SOUIDI et Mariem BEN NACEF

**Encadrés par :**

Encadrante académique : Madame Jihen TOUNSI

Encadrant professionnel : Monsieur Aymen MECHERGUI

**Année universitaire :** 2022-2023

## Dédicaces

---

**À mes chères sœurs** qui n'ont jamais cessé de me motiver et de me pousser à faire de mon mieux. Je suis reconnaissante pour tout l'amour, le soutien indéfectible et l'encouragement que vous m'avez offert. Il n'y a pas de mots qui puissent exprimer toute ma joie, ma gratitude et ma reconnaissance pour tout ce que vous avez fait pour moi.

**À mes ami(e)s** qui m'ont beaucoup soutenu durant cette période, pour l'amitié qui nous unit et pour tous les merveilleux moments et les bons souvenirs que nous avons partagés ensemble. Je vous souhaite le succès dans vos vies.

**À tous ceux que j'aime et qui m'aiment, je dédie ce modeste travail.**

**Imen Souidi**

**À mon père, Hedi,** ma raison d'être, ma source inépuisable d'inspiration.

**À ma mère, Thouraya,** ma muse éternelle, la femme qui a fait de moi ce que je suis aujourd'hui.

**À mon frère Bechir,** l'épaule sur laquelle je me repose.

**À mes oncles Ahmed, Slim, Khaled et ma tante Amel,** vous êtes et vous resterez toujours des figures paternelles dans ma vie.

**À ma cousine Sarra et ma meilleure amie Farah,** qui, malgré la distance, ont toujours été présentes pour me soutenir.

Je remercie particulièrement **Belhsan** pour sa présence et son soutien constant qui m'a permis de rester motivée et concentrée sur mes objectifs.

Je tiens finalement à adresser mes remerciements à ma deuxième famille, **mes amies Sarra, Souleima, Sinda, Azza** qui n'ont jamais cessé de me soutenir et de me pousser vers l'avant.

**Mariem BEN NACEF**

## Remerciements

---

Au terme de ce travail, nous tenons à remercier **le cadre enseignant** de l'institut supérieur des hautes études commerciales de Carthage ainsi que tous nos enseignants qui ont largement contribué à l'atteinte de nos objectifs.

Nous remercions notre encadrante académique **Madame Jihen TOUNSI** pour son accompagnement, ses encouragements, sa patience, pour le temps qu'elle nous a consacré pour ses précieux conseils.

Un grand merci à **Monsieur Mohamed Aymen Mechergui** pour son encadrement, l'effort qu'il a fourni pour nous accueillir au sein de Queney.

Nous adressons particulièrement nos remerciements à notre encadrant technique **Mohamed Berrima** qui nous a accompagné tout au long de notre stage et qui nous a assuré un suivi continu.

Nous saisissons cette occasion pour remercier **les membres du jury** tout en espérant qu'ils trouvent dans ce rapport les qualités de la clarté et motivation qu'ils attendent.

Enfin, nous souhaitons exprimer notre gratitude envers **le club Radio Libertad** qui a marqué notre parcours étudiant et nous a apporté plusieurs leçons et expériences.

# Table des matières

---

<b>DEDICACES .....</b>	<b>2</b>
<b>REMERCIEMENTS .....</b>	<b>3</b>
<b>TABLE DES FIGURES.....</b>	<b>7</b>
<b>TABLE DES TABLEAUX.....</b>	<b>9</b>
<b>ACRONYMES.....</b>	<b>10</b>
<b>INTRODUCTION GENERALE .....</b>	<b>11</b>
<b>I. PRESENTATION DU CADRE DE PROJET .....</b>	<b>14</b>
<i>I. 1. Introduction.....</i>	<i>14</i>
<i>I. 2. Cadre général.....</i>	<i>14</i>
<i>I. 3. Présentation de l'organisme d'accueil.....</i>	<i>14</i>
I. 3.1. Description de l'organisme .....	14
I. 3.2. Les services offerts par Naxxum .....	2
I. 3.3. Les technologies adoptées par Naxxum .....	2
I. 3.4. Organigramme de la société.....	2
I. 3.5. Description de Queney .....	3
<i>I. 4. Analyse de l'existant .....</i>	<i>3</i>
<i>I. 5. Problématique .....</i>	<i>3</i>
<i>I. 6. Solution proposée .....</i>	<i>4</i>
<i>I. 7. Méthodologie de travail.....</i>	<i>5</i>
I. 7.1. Exploration des méthodologies .....	5
I. 7.1.1. KDD .....	5
I. 7.1.2. SEMMA.....	6
I. 7.1.3. CRISP-DM.....	7
I. 7.2. Comparaison entre les méthodes .....	9
I. 7.3. Méthode adoptée : .....	10
<i>I. 8. Planification du projet .....</i>	<i>10</i>
<i>I. 9. Conclusion.....</i>	<i>11</i>
<b>II. ÉTUDE PRELIMINAIRE.....</b>	<b>14</b>
<i>II. 1. Introduction.....</i>	<i>14</i>
<i>II. 2. Spécification des besoins.....</i>	<i>14</i>
II. 2.1. Identification des acteurs .....	14
II. 2.2. Spécification des besoins fonctionnels.....	15
II. 2.3. Spécification des besoins non fonctionnels .....	15
<i>II. 3. Architecture.....</i>	<i>15</i>
II. 3.1. Architecture physique trois-tiers .....	15
II. 3.2. Architecture logique .....	16
<i>II. 4. Environnement de travail.....</i>	<i>17</i>
II. 4.1. Environnement matériel.....	17

II. 4.2.	Environnement logiciel.....	18
II. 4.3.	Technologies adoptées.....	19
II. 4.3.1.	Technologies Front-end .....	19
II. 4.3.2.	Technologies Back-end.....	19
II. 5.	<i>Intégration du Traitement automatique des langues et de l'apprentissage automatique</i> .....	20
II. 5.1.	Traitement automatique des langues.....	20
II. 5.1.1.	Avantages du TAL.....	20
II. 5.1.2.	Application pratique du TAL.....	21
II. 5.2.	Apprentissage automatique :.....	21
5.2.1.	Les techniques d'apprentissage automatique.....	21
II. 5.2.1.1.	Régression : .....	21
II. 5.2.1.2.	Regroupement (Clustering) :.....	21
II. 5.2.1.3.	Classification : .....	22
II. 5.2.2.	Technique choisie.....	22
II. 5.2.3.	Les algorithmes de regroupement .....	23
II. 5.2.3.1.	K-means .....	23
II. 5.2.3.2.	Meanshift.....	24
II. 5.2.3.3.	DBSCAN.....	25
II. 5.2.4.	Comparaison entre les algorithmes et l'algorithme choisi .....	27
II. 5.3.	Rôle de l'apprentissage automatique dans le traitement automatique des langues .....	28
II. 6.	<i>Bibliothèques, modules et classes utilisés</i> .....	28
II. 7.	<i>Conclusion</i> .....	30
III.	PROPOSITION D'UNE SOLUTION D'APPRENTISSAGE AUTOMATIQUE .....	32
III. 1.	<i>Introduction</i> .....	32
III. 2.	<i>Collecte des données</i> .....	32
III. 2.1.	Source de stockage de données .....	32
III. 2.2.	Extraction de données.....	33
III. 3.	<i>Compréhension des données</i> .....	34
III. 3.1.	Analyse exploratoire des données .....	34
III. 3.1.1.	La campagne « Pour mieux vous servir ».....	35
III. 3.1.2.	La campagne « Votre profil ».....	36
III. 3.1.3.	La campagne « Queney ».....	36
III. 3.2.	Évaluation de la qualité des données .....	36
III. 4.	<i>Préparation des données</i> .....	38
III. 4.1.	Création d'un dictionnaire de synonymes .....	38
III. 4.2.	Structuration des données .....	41
III. 4.2.1.	Opération de jointure .....	41
III. 4.2.2.	La normalisation de la trame de données .....	41
III. 4.2.3.	Organisation et qualification des questions selon des critères spécifiques dans la trame de données.....	42
III. 4.2.3.1.	Recherche de questions répondant à un même critère.....	42
III. 4.2.3.2.	Croisement des réponses .....	45
III. 4.2.4.	Création d'une nouvelle trame de données avec les réponses correctes pour chaque critère .....	46
III. 4.2.5.	Création d'une nouvelle trame de données avec les réponses correctes pour les autres questions .....	47
III. 5.	<i>Modélisation</i> .....	48

III. 5.1.	Encodage et traitement des données .....	48
III. 5.2.	Paramètres d'entrée de l'algorithme de Mean--Shift .....	49
III. 5.3.	Application du Mean-Shift.....	50
III. 5.4.	Evaluation du modèle .....	52
III. 5.4.1.	Métriques d'évaluation .....	52
III. 5.4.2.	Interprétation des clusters .....	52
III. 6.	<i>Conclusion</i> .....	53
IV.	DEPLOIEMENT .....	55
IV. 1.	<i>Introduction</i> .....	55
IV. 2.	<i>Modélisation conceptuelle</i> .....	55
IV. 2.1.	Diagramme de cas d'utilisation globale .....	56
IV. 2.1.1.	Description textuelle du cas d'utilisation global .....	56
IV. 2.1.1.1.	Description textuelle du diagramme de cas d'utilisation « Consulter le tableau de bord ».....	56
IV. 2.1.1.2.	Description textuelle du diagramme de cas d'utilisation « Télécharger une trame de données » 57	
IV. 2.1.1.3.	Description textuelle du diagramme de cas d'utilisation « Gérer les administrateurs » .....	57
IV. 2.1.1.4.	Description textuelle du diagramme de cas d'utilisation « S'inscrire » .....	59
IV. 2.1.1.5.	Description textuelle du diagramme de cas d'utilisation « S'authentifier » .....	60
IV. 2.2.	Diagrammes de séquences du cas d'utilisation.....	61
IV. 2.2.1.	Diagramme de séquences du cas d'utilisation « S'inscrire » .....	61
IV. 2.2.2.	Diagramme de séquences du cas d'utilisation « S'authentifier » .....	62
IV. 2.2.3.	Diagramme de séquences du cas d'utilisation « Activer un compte ».....	63
IV. 2.2.4.	Diagramme de séquences du cas d'utilisation « Désactiver un compte » .....	63
IV. 2.3.	Diagramme de classes .....	64
IV. 2.3.1.	Dictionnaire de données.....	64
IV. 2.3.2.	Diagramme de classe.....	65
IV. 3.	<i>Réalisation :</i> .....	65
IV. 3.1.	TopBar .....	66
IV. 3.1.1.	Rechercher.....	66
IV. 3.1.2.	Mode jour .....	66
IV. 3.1.3.	Profil .....	67
IV. 3.2.	Interface d'inscription.....	1
IV. 3.3.	Interface d'authentification .....	1
IV. 3.4.	Interface de Dashboard .....	2
IV. 3.5.	Les interfaces des trames de données .....	3
IV. 3.5.1.	Interface de la table « Informations générales » .....	3
IV. 3.5.2.	Interface de la table des autres questions .....	4
IV. 3.5.3.	Interface du tableau des clusters.....	5
IV. 3.5.4.	Interface de la liste des utilisateurs .....	5
IV. 3.6.	Interface de Chart Line .....	6
IV. 4.	<i>Conclusion</i> .....	7
<b>CONCLUSION GENERALE ET PERSPECTIVES .....</b>		<b>8</b>
<b>BIBLIOGRAPHIE.....</b>		<b>10</b>

## Table des figures

Figure I-1. Logo de la société Naxxum.....	2
Figure I-2. Les services offerts par Naxxum [1].....	2
Figure I-3. Les technologies utilisées par Naxxum [1] .....	2
Figure I-4. Organigramme de la société Naxxum Mea .....	2
Figure I-5. Logo de Queney [2] .....	3
Figure I-6. Logo de JAYEG [3].....	3
Figure I-7. Les conditions pour effectuer la qualification des réponses.....	2
Figure I-8. Regroupement en fonction du critère « Genre ».....	3
Figure I-9. Architecture de la méthodologie de KDD [5] .....	6
Figure I-10. Architecture de la méthodologie de SEMMA [7].....	7
Figure I-11. Architecture de la méthodologie de CRISP-DM [9].....	8
Figure I-12. Diagramme de Gantt.....	11
Figure II-1. Architecture trois-tiers [12] .....	16
Figure II-2. Architecture MVC de l'application Web [14] .....	17
Figure II-3. K-means avec différents nombres de clusters .....	24
Figure II-4. Mean-Shift pour un nombre de clusters égal à 3.....	25
Figure II-5. DBSCAN avec 3 valeurs différentes d'épsilon [38] .....	26
La figure III-1 représente un document ayant l'identifiant de la campagne « Votre profil » .....	33
Figure III-2. Reporting d'une campagne dans le Backoffice de l'application "JAYEG" .....	34
Figure III-3 Type de données et réponses possibles pour chaque question de la campagne "Pour mieux vous servir" .....	35
Figure III-4 Type de données et réponses possibles pour chaque question de la campagne "Votre profil" .....	36
Figure III-5 Type de données et réponses possibles pour chaque question de la campagne "Queney" .....	36
Figure III-6. Exemple de réponses incohérentes pour les questions relatives au genre .....	37
Figure III-7. Nombre de réponses vides .....	38
Figure III-8: Le dictionnaire .....	40
Figure III-9. Fonction pour effectuer la jointure entre les trois campagnes .....	41
Figure III-10. Un exemple illustrant avant et après la normalisation.....	42
Figure III-11. Fonction compatible_columns .....	43
Figure III-12. Fonction compatible_columns_lignes .....	43
La figure III-13 représente le résultat de la fonction compatible_columns_lignes : .....	44
Figure III-14 . Résultat de la fonction compatible_columns_lignes .....	44

Figure III-15. Fonction compatible_columns_colonnes .....	44
La figure III-16 représente le résultat de la fonction compatible_columns_colonnes : .....	45
Figure III-17. Résultat de la fonction compatible_columns_lignes .....	45
Figure III-18. Trame de données structurée pour les critères "Genre" et "Age". .....	46
Figure III-19. Trame de données des critères .....	47
Figure III-20. Trame de données des autres questions.....	48
Figure III-21. Traitement des colonnes catégorielles.....	49
Figure III-22. Trame de données encodée.....	49
Figure III-23. Application du Mean-Shift .....	50
Figure III-24. Nombre de clusters obtenus.....	50
Figure III-25. Les caractéristiques d'un cluster.....	51
Figure III-26. Trame de données des caractéristiques des clusters.....	51
Figure III-27: Valeur de l'indice de silhouette moyen .....	52
Figure IV-1: Diagramme du cas d'utilisation .....	56
Figure IV-2. Diagramme de séquences du cas d'utilisation « S'inscrire » .....	61
Figure IV-3. Diagramme de séquence du cas d'utilisation « S'authentifier ».....	62
Figure IV-4. Diagramme de séquences du cas d'utilisation « Activer un compte ».....	63
Figure IV-5. Diagramme de séquences du cas d'utilisation « Désactiver un compte » .....	64
Figure IV-6. Diagramme de classe .....	65
Figure IV-7: Capture du TopBar .....	66
Figure IV-8: Alerte de recherche 'Aucun résultat trouvé'.....	66
Figure IV-9: Mode nuit .....	67
Figure IV-10: Mode jour .....	67
Figure IV-11: Profil .....	67
Figure IV-12. Interface d'inscription.....	1
Figure IV-13. Interface d'authentification .....	2
Figure IV-14. Interface de Dashboard .....	3
Figure IV-15. Interface de la table "Informations générales" .....	4
Figure IV-16. Interface de la table des autres questions posées .....	4
Figure IV-18. Interface des Clusters .....	5
Figure IV-19. Interface de la liste des utilisateurs .....	6
Figure IV-20. Interface de Line Chart.....	6



## Table des tableaux

Tableau I-1. Tableau comparatif des méthodologies .....	9
Tableau II-1. Environnement matériel .....	17
Tableau II-2: Environnements logiciels .....	18
Tableau II-3. Technologies Front-end.....	19
Tableau II-4. Technologies Back-end.....	19
Tableau II-5 Comparaison des techniques d'apprentissage automatique .....	22
Tableau II-6. Les étapes du K-means [31] .....	23
Tableau II-7. Étapes pour appliquer l'algorithme de Mean-Shift [34] .....	25
Tableau II-8. Les étapes pour appliquer l'algorithme de DBSCAN [37] .....	26
Tableau II-9. Comparaison des algorithmes de regroupement .....	27
Tableau II-10: Bibliothèques utilisées .....	29
Tableau II-11 : Modules utilisés .....	29
Tableau II-12. Classes utilisées .....	30
Tableau III-1. Les étapes de l'extraction des données .....	34
Tableau III-2. Composition des fichiers CSV .....	35
Tableau III-3. Description de l'utilité de chaque caractéristique dans le profiling .....	39
Tableau IV-1. Description textuelle du cas d'utilisation « Consulter le tableau de bord » .....	57
Tableau IV-2. Description textuelle du cas d'utilisation « Télécharger une trame de données ».....	57
Tableau IV-3. Description textuelle du cas d'utilisation « Gérer les administrateurs ».....	58
Tableau IV-4. Description textuelle du cas d'utilisation « S'inscrire » .....	60
Tableau IV-5. Description textuelle du cas d'utilisation « S'authentifier » .....	61
Tableau IV-6. Dictionnaire de données.....	65

## Acronymes

---

SEMMA: Sample, Explore, Modify, Model, Assess

CRISP-DM: Cross-Industry Standard Process

KDD: Knowledge, Discovery, Data Mining

ML: Machine Learning

TAL : traitement des langues automatique

MinPts : nombre minimum de points

NLP: Natural language processing

DBSCAN: Density-based Spatial Clustering of Applications with Noise

Ast : Arbres Syntaxiques Abstraits

Nltk : Natural Language Toolkit

MVC : Model-View-Controller

# Introduction générale

---

Le profilage des clients consiste à recueillir, analyser et exploiter des données pertinentes sur les clients afin de dresser leur portrait de manière détaillée. Cette pratique permet d'obtenir des informations précieuses sur leurs comportements, leurs préférences, leurs habitudes d'achat, et bien d'autres éléments clés. Grâce à ces connaissances, les entreprises peuvent ajuster leur offre, personnaliser leurs communications, et offrir des expériences client plus pertinentes et satisfaisantes.

La gestion efficace de la relation client est essentielle pour assurer la réussite d'une entreprise. Dans un contexte concurrentiel où les clients sont de plus en plus exigeants et informés, il est primordial de développer des approches stratégiques visant à mieux comprendre et anticiper leurs besoins. C'est dans cette optique que le profilage client et l'apprentissage automatique ont pris une place prépondérante.

Avec la quantité massive de données disponibles de nos jours, il devient complexe de traiter ces informations de manière manuelle et efficace. C'est là que l'apprentissage automatique entre en jeu. L'apprentissage automatique est un domaine de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir des données et de prendre des décisions ou d'effectuer des prédictions de manière autonome. En utilisant des algorithmes sophistiqués, l'apprentissage automatique permet d'analyser et d'exploiter les données de profilage client de manière rapide et précise, facilitant ainsi la prise de décisions éclairées et la mise en place de stratégies efficaces.

Ce projet ambitieux vise à développer un système de profilage automatisé en utilisant des techniques de Machine Learning. L'objectif principal est de créer un modèle prédictif capable d'analyser des données client complexes et de générer des profils précis. Nous cherchons à exploiter efficacement les données disponibles pour comprendre les utilisateurs de l'application JAYEG. Grâce à ce modèle prédictif, QUENEY pourra prendre des décisions éclairées et personnalisées, adaptées aux besoins spécifiques de chaque client.

Ce rapport présentera en détail la méthodologie et les étapes nécessaires à la mise en place de ce système de profilage automatisé. Il s'articule autour de quatre chapitres :

Le premier chapitre est un chapitre introductif dans lequel nous présenterons en détail les différentes étapes de ce projet. Nous commencerons par une introduction du cadre général, en mettant en lumière l'importance du profilage client et les avantages de son automatisation. Une analyse de l'existant est réalisée pour mettre en évidence la problématique à résoudre, suivie d'une proposition de solution. La méthodologie de travail adoptée et la planification du projet sont également abordées.

Le deuxième chapitre décrit les objectifs spécifiques du projet, les outils et les technologies utilisés mises en œuvre. L'environnement de travail, tant matériel que logiciel, est également décrit, suivi d'une explication de l'intégration du Traitement Automatique des Langues (TAL) et de l'apprentissage automatique dans le projet.

Le troisième chapitre propose une solution basée sur le ML. Nous présenterons en détail la solution proposée, en décrivant les modèles de Machine Learning utilisés et les techniques de prétraitement des données.

Le quatrième chapitre, quant à lui, se concentre sur le déploiement de la solution proposée. Il présente une modélisation conceptuelle à l'aide de différents diagrammes pour représenter la structure du système. La réalisation pratique de l'application est également abordée, avec des descriptions des interfaces utilisateur et des fonctionnalités clés.

Enfin, nous évaluerons les résultats obtenus, en analysant les performances du modèle, sa capacité à générer des profils précis et sa valeur ajoutée pour les entreprises. Nous discuterons également des limites et des perspectives d'amélioration du système, ouvrant ainsi la voie à de futures recherches et développements.

# Chapitre I

---

## Présentation du cadre de projet

---

## **I. Présentation du cadre de projet**

---

### **I. 1. Introduction**

Le premier chapitre de ce rapport est consacré à l'introduction du cadre général du projet et à la description des exigences. Nous entamerons cette section en présentant l'organisation d'accueil, qui joue un rôle central dans la réalisation de ce projet. Ensuite, nous aborderons la problématique qui motive notre démarche et l'étude approfondie de l'existant. Enfin, nous présenterons la solution que nous proposons pour répondre à cette problématique.

Ce chapitre constitue une étape essentielle pour comprendre le contexte dans lequel s'inscrit notre projet et les objectifs que nous visons. Il nous permettra également de poser les bases nécessaires à la compréhension des chapitres suivants, où nous détaillerons les étapes de mise en œuvre de notre solution.

### **I. 2. Cadre général**

Nous allons développer une solution innovante qui permet l'automatisation du modèle de profiling des clients basé sur les données collectées. Ce modèle a pour objectif de réaliser des croisements des réponses d'un même utilisateur suite à sa participation aux campagnes lancées par l'application JAYEG en vue de vérifier la validité de ses réponses et ajuster son profil selon les informations vérifiées. Cette solution vise à améliorer l'efficacité et la productivité en automatisant les processus clés et en offrant des outils d'analyse avancés.

### **I. 3. Présentation de l'organisme d'accueil**

Dans cette partie nous allons présenter l'organisme d'accueil tout en mentionnant les services et les technologies adoptées par la société.

#### **I. 3.1. Description de l'organisme**

Fondée en 2014, Naxxum [1] est une société mondiale de conseil en informatique et de développement de logiciels personnalisés avec plus de 400 projets livrés. Naxxum offre à ses clients et partenaires une suite de produits numériques permettant de transformer une idée de l'état embryonnaire à la concrétisation et au développement d'un business réel. Présente dans huit pays, dont le Qatar, le Canada, la France, les Philippines, Madagascar, la Tunisie, le Maroc

et l'Algérie, Naxxum bénéficie d'une portée internationale pour répondre aux besoins et aux exigences de ses clients à travers le monde.



Figure I-1. Logo de la société Naxxum

### I. 3.2. Les services offerts par Naxxum

Naxxum offre une gamme variée de service et se distingue par sa capacité à s'adapter en fonction des besoins changeants et des avancées technologiques. Grâce à sa flexibilité, la société est en mesure d'offrir des solutions performantes, tout en maintenant un haut niveau de qualité et de satisfaction client.

La figure I-2 extraite du site officiel de Naxxum représente les différents services offerts par Naxxum :

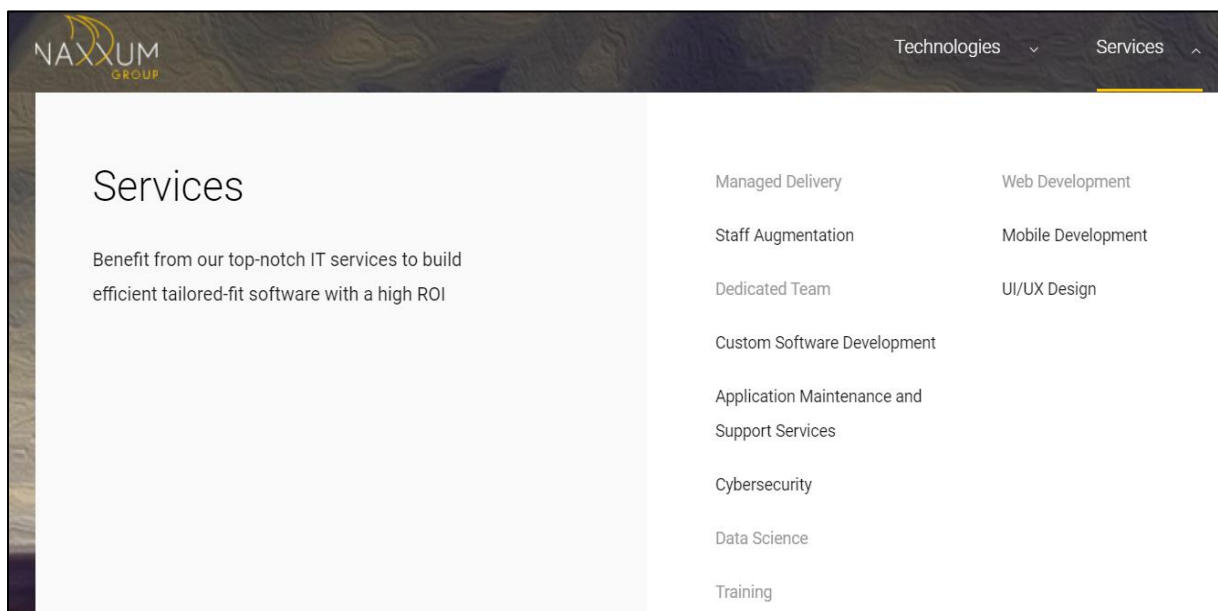
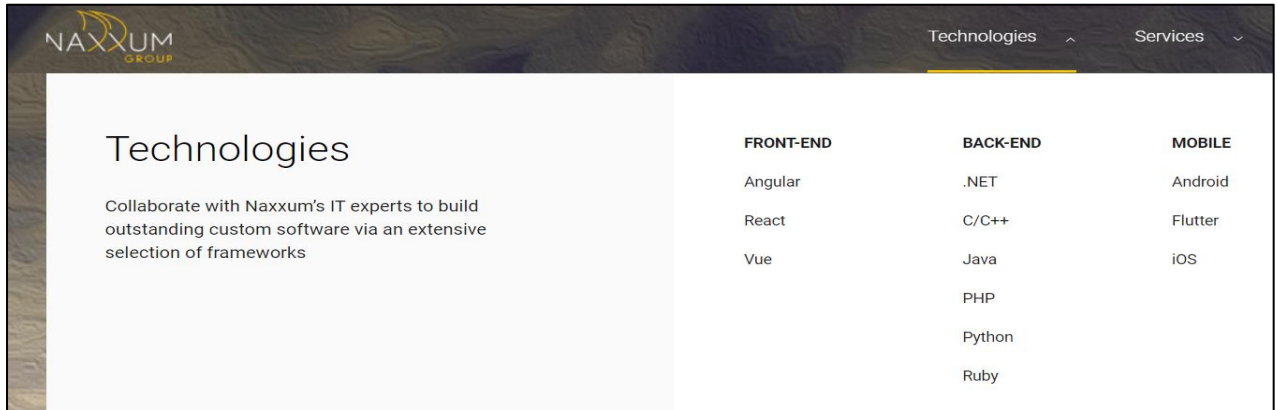


Figure I-2. Les services offerts par Naxxum [1]

### I. 3.3. Les technologies adoptées par Naxxum

Naxxum reste constamment à l'affût des dernières avancées technologiques, elle utilise des technologies de pointe pour le développement de ses solutions, que ce soit pour le front-end web, le back-end web ou les applications mobiles.

La figure I-3 représente les différentes technologies utilisées par Naxxum :



Technologies	FRONT-END	BACK-END	MOBILE
Collaborate with Naxxum's IT experts to build outstanding custom software via an extensive selection of frameworks	Angular React Vue	.NET C/C++ Java PHP Python Ruby	Android Flutter iOS

Figure I-3. Les technologies utilisées par Naxxum [1]

### I. 3.4. Organigramme de la société

Notre stage est réalisé au sein de Naxxum Mea en Tunisie. Nous présentons la hiérarchie de la société par l'organigramme suivant :

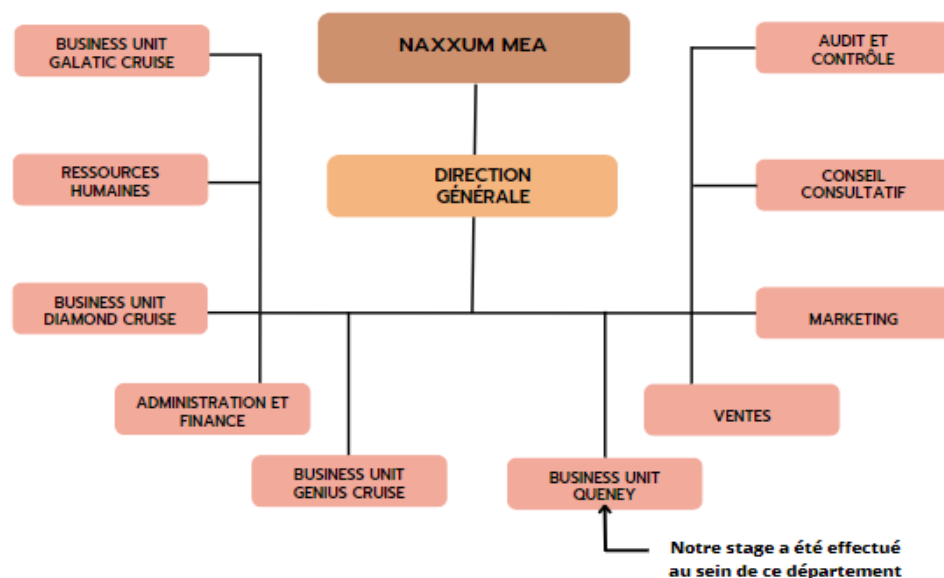


Figure I-4. Organigramme de la société Naxxum Mea



### **I. 3.5. Description de Queney**

Queney (Question for money) est une Business Unit au sein de Naxxum Mea, qui explore de nouvelles méthodes de collecte d'informations sur les utilisateurs. Queney a choisi l'utilisation de la gamification pour encourager les utilisateurs à partager davantage d'informations sur eux-mêmes en lançant des challenges amusants et interactifs qui offrent des récompenses en échange d'informations sur leurs préférences et leurs comportements. Grâce à cette approche, nous sommes en mesure de collecter des données plus précises des utilisateurs, sans qu'ils aient l'impression d'être surveillés ou de subir une enquête. Ces données nous permettent de mieux comprendre les besoins et les attentes des utilisateurs.



*Figure I-5. Logo de Queney [2]*

Dans le contexte de notre projet, l'application étudiée sera « JAYEG » :

JAYEG est une application qui permet aux abonnés mobile Orange Tunisie de gagner instantanément de l'internet mobile en regardant des vidéos et en participant à des campagnes contenant des questionnaires qui permettent par la suite de collecter des informations sur les utilisateurs pour mieux les connaître et comprendre leurs besoins.



*Figure I-6. Logo de JAYEG [3]*

### **I. 4. Analyse de l'existant**

Nous entamons maintenant l'analyse de l'existant afin d'identifier les points faibles à corriger. Le processus du profiling existant suit ses étapes :

1. L'extraction des réponses des utilisateurs, qui peuvent atteindre un nombre important, pouvant dépasser les 10 000 utilisateurs, et nous les exportons dans un fichier Excel au format xlsx.
2. L'établissement des relations entre les questions et croisement des réponses possibles pour créer un tableau représentant toutes les conditions envisageables afin de vérifier la cohérence. Cette étape implique une classification manuelle des questions et leur regroupement, ce qui permet d'organiser les questions de manière structurée et de faciliter leur analyse ultérieure.

La Figure I-7 tirée du fichier préparé montrant les croisements des questions possibles suivis de la qualification pour chaque cas :

Quelle est ton âge	Année de naissance	Qualif
Q=62709874307c7001f0838ed Moins de 18 ans	Q=6267664772e4c0002a4b003 Moins de 2004	COHERENCE
Entre 18 et 24 ans	1998 - 2004	COHERENCE
Entre 25 et 34 ans	1988 - 1997	COHERENCE
Je suis ?	Je suis unit	
Q=626fadb9a43c670020803235 ; C=62709874307c7001f0838ed	Q=62669903072e4c0002a4b003 ; C=6267664772e4c0002a4b003	
Les réponses doivent être identiques		COHERENCE
Je suis	Année de naissance	
Q=626fadb9a43c670020803235 ; C=62709874307c7001f0838ed	Q=6267664772e4c0002a4b003 ; C=6267664772e4c0002a4b003	
collège	Moins de 2004	COHERENCE
lycée	Moins de 2004	COHERENCE
Université	1988 - 2004	COHERENCE
Parmi les catégories suivantes, laquelle décrit le mieux votre statut professionnel actuel ?	Je suis	
Q=62669903072e4c0002a4b003 ; C=6267664772e4c0002a4b003	Q=626fadb9a43c670020803235 ; C=62709874307c7001f0838ed	
Etudiant	Au collège	COHERENCE
Etudiant	Au lycée	COHERENCE
Etudiant	Université	COHERENCE
Quelle est votre plus haut niveau d'éducation ?	Je suis	
Q=6272965d817db3001f654538 ; C=6279a424a43c670020805557	Q=626fadb9a43c670020803235 ; C=62709874307c7001f0838ed	
Ecole primaire	université	INCOHERENCE
Ecole secondaire	université	INCOHERENCE
Université	Collège /lycée	INCOHERENCE
3 <sup>ème</sup> cycle	Collège /lycée	INCOHERENCE
Doctorat	Collège /lycée	INCOHERENCE
Avez-vous un permis de conduire	Etes-vous motorisé	
Q=626f7678174b3001f653a011 ; C=62709874307c7001f0838ed	Q=6272a9f-c0000-0000047ef7a ; C=6279a424a43c670020805557	
Oui	Oui	COHERENCE
Non	Oui	INCOHERENCE

Figure I-7. Les conditions pour effectuer la qualification des réponses

3. La comparaison manuelle des réponses correspondantes : Cela implique un examen attentif des réponses fournies par les utilisateurs, à la recherche de similitudes, de différences ou de contradictions en fonction des conditions établies. Cette étape permet de détecter d'éventuelles incohérences ou divergences dans les réponses des utilisateurs.
4. La prise de décisions est basée sur l'analyse et la comparaison des réponses, quant à leur cohérence. Les réponses sont considérées comme cohérentes si elles concordent et nous les jugeons incohérentes si elles présentent des divergences significatives en fonction

des conditions définies. Cette évaluation se fait pour chaque critère spécifique dans une feuille unique.

La figure I-8 montre le résultat de l'attribution de la qualification pour chaque utilisateur pour le critère « Genre » :

User_id	nsrOne.campa	nsrOne.questi	rOne.quest	nsrOne.an	nsrTwo.campa	nsrTwo.questi	rTwo.quest	nsrTwo.a	Qualification GENDER
62663031cb2dac002	627008c74307c7001	626fadd9a43c67002	(je suis?)	Un garçon	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	COHERENCE
620eb098441324002	627008c74307c7001	626fadd9a43c67002	(je suis?)	Un garçon	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	COHERENCE
623b144102b958001	627008c74307c7001	626fadd9a43c67002	(je suis?)	Un garçon	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	COHERENCE
6221fbac203e20001	627008c74307c7001	626fadd9a43c67002	(je suis?)	Un garçon	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	COHERENCE
620e2fac712bb4002	627008c74307c7001	626fadd9a43c67002	(je suis?)	Un garçon	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	COHERENCE
6265b9bb58da3b001	627008c74307c7001	626fadd9a43c67002	(je suis?)	Une fille	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Femme	COHERENCE
62728110360d8c001	627008c74307c7001	626fadd9a43c67002	(je suis?)	Un garçon	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	COHERENCE
626932ad0d7189001	627008c74307c7001	626fadd9a43c67002	(je suis?)	Une fille	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Femme	COHERENCE
627402f51be8ce001	627008c74307c7001	626fadd9a43c67002	(je suis?)	Une fille	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Femme	COHERENCE
6212eaa6619553001	627008c74307c7001	626fadd9a43c67002	(je suis?)	Un garçon	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	COHERENCE
628b7ffc74e9ee002	627008c74307c7001	626fadd9a43c67002	(je suis?)	Une fille	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	INCOHERENCE
623d0aafaf93ec002	627008c74307c7001	626fadd9a43c67002	(je suis?)	Un garçon	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	COHERENCE
628ba834be5df9002	627008c74307c7001	626fadd9a43c67002	(je suis?)	Un garçon	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	COHERENCE
6268195f261941001f688c9f					6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Homme	NON APPLICABLE
623210596d818c001	627008c74307c7001	626fadd9a43c67002	(je suis?)	Une fille	6267b64772e4c00c	6266903072e4c0002a4	(je suis un(e))	Femme	COHERENCE

Figure I-8. Regroupement en fonction du critère « Genre »

## I. 5. Problématique

Le processus de profilage des clients est actuellement réalisé de manière manuelle, ce qui présente plusieurs limitations et peut avoir un impact négatif sur l'efficacité des campagnes. Ces limitations comprennent :

- **Temps et ressources** : Le processus manuel de profilage des clients est fastidieux et demande beaucoup de temps et de ressources. Les équipes marketing doivent collecter, analyser et interpréter manuellement un grand volume de données pour identifier les caractéristiques et les comportements des clients. Cela peut entraîner des retards dans la mise en place des campagnes marketing et une utilisation inefficace des ressources humaines.

- **Erreurs et subjectivité :** Lorsque le profilage est effectué manuellement, il y a un risque accru d'erreurs et de subjectivité. Les interprétations individuelles peuvent varier, ce qui peut entraîner une segmentation inexacte des clients et une compréhension incomplète de leurs besoins et préférences. Ces erreurs peuvent conduire à des actions marketing mal ciblées et à un gaspillage de ressources.
- **Manque de scalabilité :** Le processus manuel de profilage des clients n'est pas facilement scalable pour des volumes de données importants. Avec la croissance des données disponibles, il devient de plus en plus difficile de gérer efficacement le profilage manuel pour un grand nombre de clients. Cela limite la capacité des entreprises à analyser et à exploiter pleinement les données pour des campagnes marketing personnalisées et ciblées.

Ces limitations ont des conséquences directes sur l'efficacité des campagnes lancées. En raison du processus manuel, Queney peut manquer de comprendre réellement les besoins et les préférences des utilisateurs, ce qui conduit à des actions marketing moins pertinentes. Cela peut entraîner une baisse du taux de conversion, une augmentation des coûts marketing et une insatisfaction des utilisateurs. De plus, l'incapacité à gérer efficacement les volumes de données peut limiter la capacité des entreprises à s'adapter rapidement aux évolutions du marché et à saisir de nouvelles opportunités.

## **I. 6. Solution proposée**

Le projet consiste à mettre en place un système d'aide à la décision et plus précisément le ciblage dédié à la déduction de la cohérence des informations des utilisateurs à travers leurs réponses à des campagnes lancées par l'application JAYEG afin de pouvoir analyser leurs comportements pour anticiper leurs besoins et intérêts.

La solution se base sur :

- La mise en place d'un système d'analyse des réponses fournies par les utilisateurs :
  - Traitement des données.
  - Définition les critères de profiling.
  - Implémentation d'un modèle d'apprentissage automatique.

- Implémentation d'un dictionnaire afin d'améliorer les résultats du modèle.
- Déduction de la cohérence d'une réponse à travers la comparaison avec des questions similaires.
- Intégration des données en récupérant les différents fichiers Excel et de les regrouper dans une seule base unique.
- Restitution des données en créant des tableaux de bords afin de visualiser et analyser les réponses des utilisateurs pour mettre à jour et affiner leurs profils, en les classant de manière plus précise en fonction de leurs caractéristiques.

## **I. 7. Méthodologie de travail**

Dans notre projet, nous avons suivi un processus de conception méthodique pour assurer la productivité et l'optimisation de la phase de réalisation, tout en estimant correctement le temps de développement. Dans cette partie, nous expliquons notre approche justifiée afin d'atteindre les objectifs fixés.

### **I. 7.1. Exploration des méthodologies**

La définition des méthodes SEMMA, KDD et CRISP-DM sera abordées dans la partie suivante

#### **I. 7.1.1. KDD**

KDD [4], abréviation de "Knowledge Discovery in Databases" (découverte de connaissances dans les bases de données), est un processus largement utilisé qui répond aux besoins des entreprises. Cette technique englobe la préparation, la sélection et le nettoyage des données, ainsi que l'intégration de connaissances préexistantes sur de grandes quantités de données, et l'interprétation de solutions précises à partir des résultats observés.

Le processus KDD fait appel à des méthodes de Data Mining et comporte cinq étapes :

- Sélection :  
Cette étape a pour objectif de choisir les données à analyser.
- Prétraitement :

Il s'agit d'une étape visant à obtenir des données cohérentes en nettoyant et en prétraitant les données ciblées.

- Transformation :

Cette étape vise à transformer les données à l'aide de méthodes afin de réduire et de transformer les dimensions.

- Data Mining :

Cette étape a pour but de trouver des modèles qui correspondent aux objectifs du Data Mining.

- Interprétation / Évaluation :

Il s'agit de l'étape où les modèles construits sont interprétés et évalués.

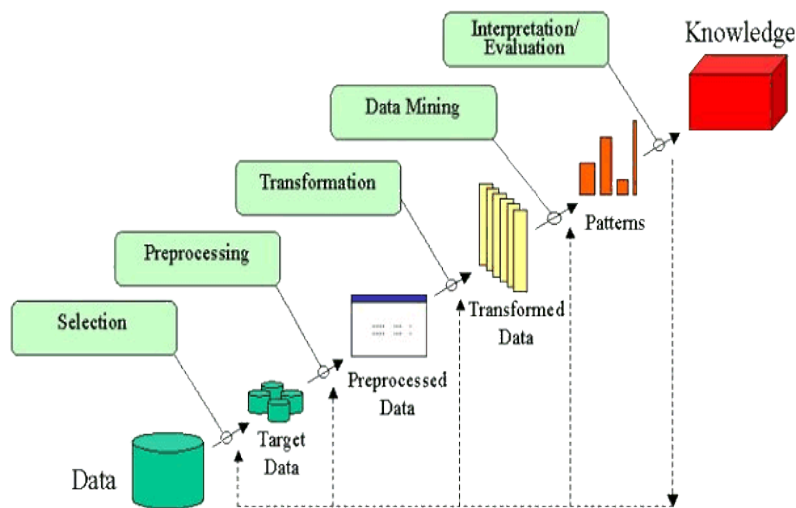


Figure I-9. Architecture de la méthodologie de KDD [5]

### I. 7.1.2. SEMMA

SEMMA [6] est une méthode développée par l'institut SAS qui facilite et rend plus compréhensible le processus d'exploration, de visualisation, de sélection, de transformation et de modélisation des données pour un data scientist. La signification de SEMMA, "Échantillon, Explorer, Modifier, Modèle, Évaluer", se réfère aux cinq étapes du processus d'un projet de Data Mining :

- Échantillon : L'échantillonnage consiste à extraire une partie significative et moins volumineuse d'un ensemble de données pour faciliter leur manipulation rapide.

- Explorer : Cette étape vise à rechercher les tendances et les anomalies inattendues lors de l'exploration des données, dans le but d'acquérir une compréhension approfondie.
- Modifier : La modification des données consiste à créer, sélectionner et transformer des variables afin de garantir le processus de sélection du modèle.
- Modèle : Cette étape consiste à construire un modèle adapté pour résoudre les problématiques du Data Mining.
- Évaluer : L'évaluation de l'utilité et de la fiabilité des résultats du processus de Data Mining, ainsi que l'estimation des performances des données.

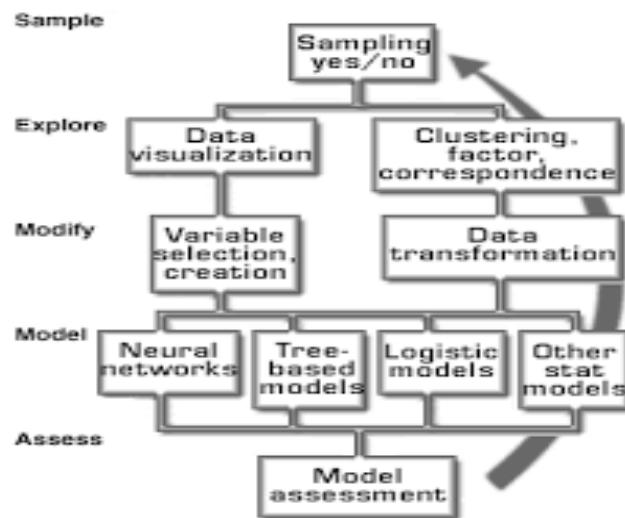


Figure I-10. Architecture de la méthodologie de SEMMA [7]

### I. 7.1.3. CRISP-DM

CRISP-DM [8] est une méthode qui permet de gérer les projets de Data Mining. Plusieurs entreprises utilisent cette méthode car elle a prouvé son efficacité dans le domaine et est devenue le processus le plus couramment utilisé pour les projets de Data Mining.

La méthode comprend un cycle de six étapes :

- Compréhension du domaine : Cette première étape consiste à bien identifier le périmètre, la nécessité et les ressources du projet. Elle comprend l'évaluation des risques, des avantages, des coûts et la planification du projet.

- Compréhension des données : Cette étape consiste à explorer et à obtenir une vue d'ensemble de la qualité des données en identifiant les besoins et en effectuant une collecte de données.
- Préparation des données : Il s'agit d'une étape qui vise à sélectionner, nettoyer, formater et intégrer les données. La préparation des données nécessite des transformations et des enrichissements afin de permettre une analyse plus approfondie. Cette partie est souvent la plus longue des projets de Data Mining.
- Modélisation : Dans cette étape, une technique de modélisation est choisie et ses paramètres sont calibrés pour obtenir les meilleures valeurs possibles.
- Évaluation : Cette étape implique une évaluation approfondie du modèle obtenu et la révision des étapes de construction du modèle pour s'assurer qu'il atteint avec succès les objectifs commerciaux.
- Déploiement : Dans cette étape, nous devons déployer notre projet sur un système où tous nos besoins seront exécutés. Nous effectuons également un examen final du projet suivi d'une planification pour les phases suivantes.

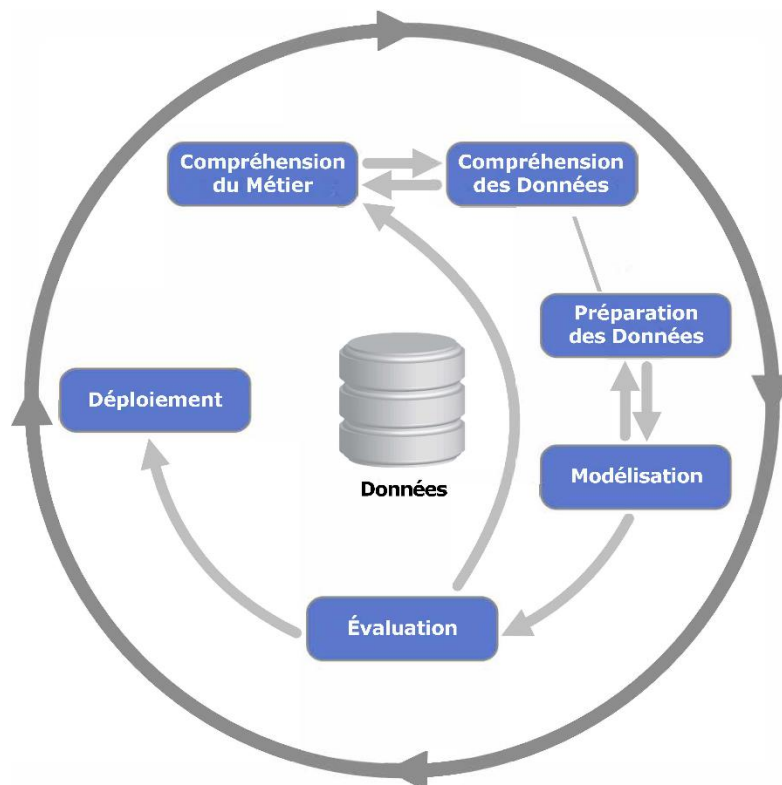


Figure I-11. Architecture de la méthodologie de CRISP-DM [9]



### I. 7.2. Comparaison entre les méthodes

Le tableau I-1 représente une comparaison faite entre les trois méthodologies [10] en fonction de leur origine, de leur domaine d'application, de leur flexibilité, de leur popularité, ainsi que de leurs avantages et limites.

<b>Méthodologie</b>	<b>KDD</b>	<b>CRISP-DM</b>	<b>SEMMA</b>
<b>Origine</b>	Domaine de la découverte de connaissances dans les bases de données	Développée par le consortium CRISP-DM	Développée par SAS Institute
<b>Domaine d'application</b>	Analyse de grandes quantités de données pour la découverte de connaissances	Projet de Data Mining dans divers domaines	Analyse de données pour la résolution de problèmes complexes
<b>Flexibilité</b>	Offre une approche générale et flexible pour la découverte de connaissances	Offre une méthodologie structurée et itérative pour les projets de Data Mining	Offre une approche spécifique et structurée pour la résolution de problèmes complexes
<b>Popularité</b>	bien connue et largement utilisée	Couramment utilisée dans l'industrie du Data Mining	Utilisée dans les produits logiciels de SAS Institute
<b>Avantages</b>	Prise en compte de l'intégration de connaissances préexistantes, résultats précis	Approche itérative, flexibilité, adaptation aux projets de Data Mining	Approche spécifique pour les problèmes complexes, mise en œuvre pratique
<b>Les limites</b>	Peut nécessiter une préparation intensive des données, complexité de l'interprétation des résultats	Peut-être trop générique pour certains projets spécifiques, dépendant du contexte	Nécessite l'utilisation d'outils spécifiques de SAS Institute

Tableau I-1. Tableau comparatif des méthodologies

### **I. 7.3. Méthode adoptée :**

Nous avons choisi de suivre le processus CRISP-DM pour plusieurs raisons. Parmi celles-ci, ce modèle se distingue des deux autres modèles principalement par sa phase de compréhension du métier, qui joue un rôle essentiel dans la réussite d'un projet de Data Mining. Il met également l'accent sur le déploiement du modèle pour mettre la solution en production, nous permettra de créer une stratégie à long terme qui améliore la stratégie préalablement développée. De plus, CRISP-DM propose un processus itératif avec la possibilité d'effectuer des allers-retours entre les différentes étapes ce qui nous permettra de procéder à des adaptations continues pour une amélioration continue de notre stratégie et de nos résultats.

### **I. 8. Planification du projet**

La durée de notre stage s'est déroulée sur une période de 4 mois, débutant du début février jusqu'à la fin du mois de juin. La planification du projet se fera en suivant une méthodologie rigoureuse et structurée. Nous diviserons le travail en différentes phases afin de garantir une progression cohérente et efficace.

1. Au cours de la phase initiale, nous nous concentrerons sur la compréhension approfondie du problème métier, en analysant les besoins et en définissant clairement les objectifs à atteindre.
2. Ensuite, nous procéderons à la phase de collecte et d'exploration des données, où nous examinerons attentivement les sources de données disponibles, leur qualité et leur pertinence pour notre projet. Cette étape sera suivie par la phase de préparation des données, au cours de laquelle nous effectuerons des opérations de nettoyage, de transformation et de normalisation pour assurer la qualité et la cohérence des données utilisées dans notre projet.
3. Par la suite, nous entamerons la phase de modélisation, où nous sélectionnerons les techniques et les algorithmes d'apprentissage automatique les plus adaptés à notre problème. Nous entraînerons et ajusterons ces modèles en utilisant les données préalablement préparées. Une fois les modèles développés, nous évaluerons leurs performances à l'aide de métriques appropriées pour mesurer l'efficacité de notre approche.

4. Enfin, nous procéderons au déploiement de la solution, intégrant les résultats de notre projet dans une application fonctionnelle prête à être utilisée par les utilisateurs finaux. Tout au long du projet, nous veillerons à itérer entre les différentes étapes, en effectuant des ajustements et des améliorations au besoin, afin d'assurer une progression continue et une adaptation aux besoins du projet.

La rédaction du rapport s'est étalée sur tout le long de la durée du stage

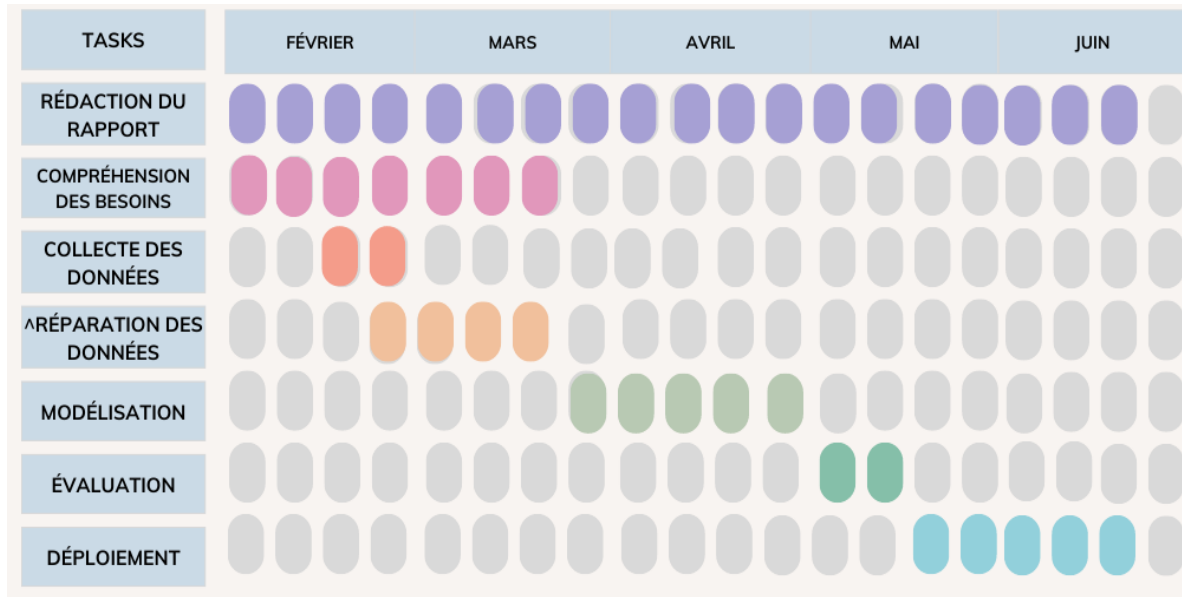


Figure I-12. Diagramme de Gantt

## I. 9. Conclusion

Dans ce premier chapitre, nous avons introduit le cadre général du projet et décrit les exigences qui le guident. Nous avons commencé par présenter l'organisation d'accueil. Ensuite, nous avons exposé la problématique à laquelle nous nous confrontons, en soulignant l'importance de trouver une solution adaptée. Nous avons également effectué une étude de l'existant, ce qui nous a permis de mieux comprendre le contexte et les enjeux liés à notre projet.

Pour répondre à cette problématique, nous avons développé une solution que nous présenterons en détail dans les chapitres suivants. Cette solution vise à apporter des réponses concrètes et efficaces, en utilisant des méthodologies et des outils appropriés.

En somme, ce premier chapitre constitue une base solide pour la suite de notre rapport. Il nous permet de situer notre projet dans son contexte global, de comprendre les enjeux et les besoins, et de définir les fondements de notre approche. Dans les chapitres à venir, nous développerons les différentes étapes de mise en œuvre de notre solution, en mettant en évidence les choix et les résultats obtenus.

# Chapitre II

---

## Étude préliminaire

---

## **II . Étude préliminaire**

---

### **II. 1. Introduction**

Ce chapitre constitue une étape essentielle de notre projet, car il se concentre sur l'étude préliminaire et la spécification des besoins. Nous débuterons en posant les bases nécessaires à la mise en œuvre de notre solution, en identifiant clairement les exigences et les objectifs à atteindre. Nous aborderons également deux aspects fondamentaux de notre étude, à savoir l'architecture logique et physique, ainsi que l'environnement de travail.

Au cours de cette étude préliminaire, nous mettrons en évidence les différentes techniques et algorithmes que nous avons sélectionnés pour l'apprentissage automatique et le traitement automatique de langues. Nous expliquerons en détail les technologies que nous avons choisies pour le développement du back-end et du front-end de notre application. Ces choix technologiques ont été faits en fonction de leur pertinence et de leur adéquation aux besoins spécifiques de notre projet.

### **II. 2. Spécification des besoins**

Nous allons nous concentrer sur la spécification des besoins liés au développement de la solution.

#### **II. 2.1. Identification des acteurs**

Pour chaque rôle bien précis, nous désignons un acteur.

Les acteurs se désignent comme suit :

Admin : Il dispose des droits de visualiser les tableaux de bords, les trames de données nettoyées avant et après l'application des techniques d'apprentissage automatique.

Super Admin : Il s'agit du gestionnaire de l'application Web, il peut gérer les comptes des utilisateurs, il est responsable du maintien du bon fonctionnement de l'application.

## II. 2.2. Spécification des besoins fonctionnels

Dans cette partie, nous aborderons les besoins fonctionnels de la solution. Les principaux objectifs sont les suivants :

- Collecte des données stockées dans des fichiers Excel.
- Alimentation de la solution par un modèle d'apprentissage.
- Réalisation d'une solution permettant d'enrichir un dictionnaire de critères de profiling pour renforcer le modèle.
- Visualisation du résultat
- Création des tableaux de bords analytique pour aider l'entreprise à obtenir une idée générale sur les utilisateurs de l'application et mieux les comprendre.

## II. 2.3. Spécification des besoins non fonctionnels

La solution proposée devrait avoir les qualités suivantes :

- **Fiabilité** : le système doit fournir aux administrateurs des résultats et des analyses fiables et valides.
- **Précision** : les calculs des prédictions doivent être aussi précis que possible.
- **Performance** : le système doit fournir des données en termes de rapidité et de vélocité pour améliorer ses performances.
- **Utilisation** : l'utilisation doit être simple.

## II. 3. Architecture

Nous allons aborder l'architecture de la solution proposée.

### II. 3.1. Architecture physique trois-tiers

Une architecture physique en trois tiers [11] sépare les différentes couches fonctionnelles en trois niveaux distincts. Chaque niveau est responsable de tâches spécifiques et interagit avec les

autres niveaux de manière bien définie. Les trois niveaux typiques d'une architecture physique en trois tiers sont les suivants :

- Couche de présentation : Ce niveau est responsable de la présentation de l'interface utilisateur à l'utilisateur final. Il gère l'interaction avec l'utilisateur et fournit une interface conviviale pour entrer et afficher les données.
- Couche logique : Ce niveau contient la logique métier de l'application. Il traite les demandes de l'utilisateur, effectue des calculs, applique des règles métier et accède aux données nécessaires. Cette couche est responsable de la manipulation et de la transformation des données en fonction des règles métier spécifiques de l'application.
- Couche de données : Ce niveau est responsable de la gestion des données utilisées par l'application. Il stocke et récupère les données à partir d'une base de données.

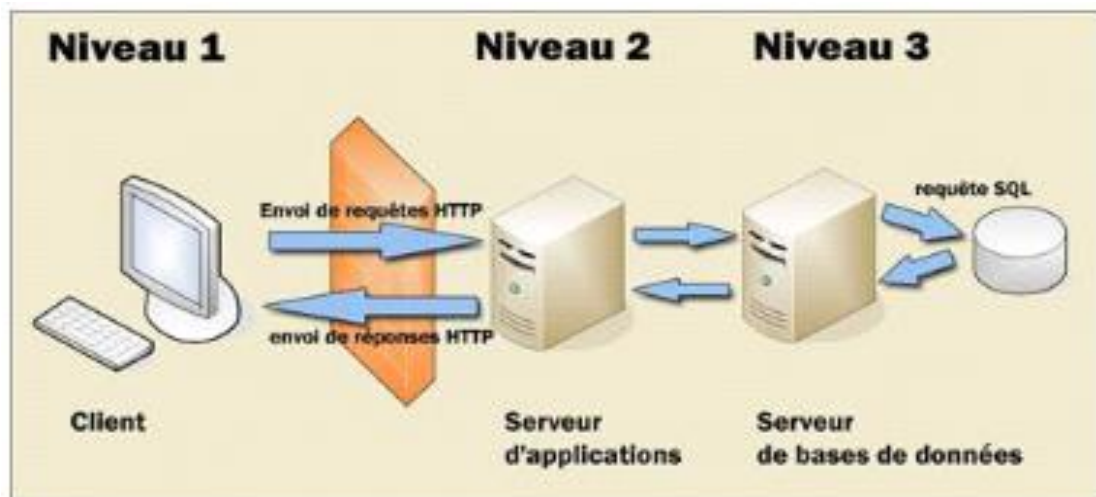


Figure II-1. Architecture trois-tiers [12]

### II. 3.2. Architecture logique

Pour clarifier davantage l'interaction entre les calques pendant le traitement effectué dans l'application, nous avons opté pour l'architecture MVC [13] qui illustre les interactions entre les modules.

MVC est un modèle architectural composé de trois parties : Modèle, Vue, Contrôleur.

- Modèle : Gère la logique des données.
- Vue : Affiche les informations du modèle à l'utilisateur.
- Contrôleur : Synchronisation du modèle et de la vue.



Il met l'accent sur une séparation entre la logique métier du logiciel et les détails de la présentation.

La figure II-2 illustre cette architecture :

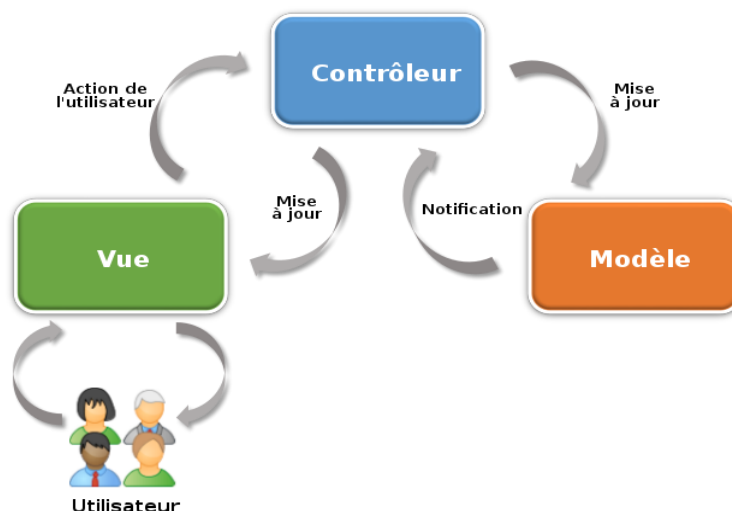


Figure II-2. Architecture MVC de l'application Web [14]

## II. 4. Environnement de travail

Dans cette section, nous exposons l'environnement matériel et logiciel exploité.

### II. 4.1. Environnement matériel

Tout au long de ce projet nous avons utilisé deux ordinateurs portables pour assurer tous les traitements nécessaires à la réalisation de ce projet. Ces deux ordinateurs ont les caractéristiques suivantes :

	Ordinateur portable 1	Ordinateur portable 2
<b>Processeur</b>	Intel(R) Core (TM) i7-10510U CPU @ 1.80GHz 2.30 GHz	11th Gen Intel(R) Core (TM) i5-11300H @ 3.10GHz 3.11 GHz
<b>Système d'exploitation</b>	Système d'exploitation 64 bits, processeur x64	Système d'exploitation 64 bits, processeur x64
<b>Mémoire RAM</b>	8,00 Go	8,00 Go
<b>Disque Dur</b>	WDC WD10SPZX-08Z10	
<b>Carte graphique</b>	NVIDIA GeForce MX330	NVIDIA GEFORCE GTX

Tableau II-1. Environnement matériel

## II. 4.2. Environnement logiciel

D'après les spécifications de notre projet nous avons eu besoin des logiciels suivants :






Logiciel	Description	Objectif
<b>Visual Studio Code</b> <b>[15]</b> 	Un éditeur de code cross Platform, open source et gratuit supportant une dizaine de langages.	Nous avons utilisé ce logiciel pour développer la partie front-end et back-end de l'application.
<b>Draw.io [16]</b> 	C'est un logiciel gratuit disponible en ligne. Il permet de concevoir toutes sortes de dessins vectoriels et de les enregistrer en format XML puis les exporter.	Nous avons utilisé ce logiciel pour tracer les diagrammes de séquence.
<b>Visual Paradigm [17]</b> 	C'est un outil de conception de diagrammes en ligne.	Nous avons utilisé ce logiciel pour tracer le diagramme de cas d'utilisation.
<b>Postman [18]</b> 	C'est une application permettant de tester les API.	Nous avons utilisé ce logiciel pour tester non API.
<b>Jupyter Notebook</b> <b>[19]</b> 	Jupyter Notebook est un environnement de développement interactif largement utilisé pour l'analyse de données, la visualisation, l'apprentissage automatique et d'autres tâches liées à la programmation.	Nous avons utilisé Jupyter Notebook pour analyser , mieux comprendre nos données ainsi que visualiser le résultat pour notre travail.

Tableau II-2: Environnements logiciels

## II. 4.3. Technologies adoptées

### II. 4.3.1. Technologies Front-end

Le tableau II-3 présente l'ensemble des technologies utilisées pour la réalisation du front-end.



Technologie	Description
<b>React [20]</b> 	React est une bibliothèque JavaScript conçu par Meta pour la création d'interface utilisateurs à partir de composants.
<b>JavaScript [21]</b> 	JavaScript est un langage de script léger, orienté objet utilisé pour le développement d'application web interactives. Ce langage permet aux concepteurs d'ajouter du comportement et de l'interactivité aux sites Web en manipulant le contenu et la structure des pages en réponse aux actions des utilisateurs.

Tableau II-3. Technologies Front-end

### II. 4.3.2. Technologies Back-end

Le tableau II-4 ci-dessous présente l'ensemble des technologies utilisées pour la réalisation du back-end :




Technologie	Description
<b>MongoDB [22]</b> 	MongoDB est un programme de base de données multiplateforme orienté document disponible en source. Classé comme programme de base de données NoSQL, MongoDB utilise des documents de type JSON avec des schémas facultatifs. MongoDB est développé par MongoDB Inc. et sous licence publique côté serveur.
<b>Flask [23]</b> 	Flask est un micro framework open-source de développement web en Python. Il est classé comme microframework car il est très léger.
<b>Python [24]</b> 	Python est un langage de programmation interprété, polyvalent, dynamique et facile à apprendre. Python est un excellent choix en tant que langage de backend pour le développement d'applications web. Sa bibliothèque standard riche, sa syntaxe claire et sa communauté active en font un langage puissant et apprécié des développeurs.

Tableau II-4. Technologies Back-end

## **II. 5. Intégration du Traitement automatique des langues et de l'apprentissage automatique**

### **II. 5.1. Traitement automatique des langues**

Le Traitement Automatique des Langues [25] (TAL), également connu sous le nom de Traitement du Langage Naturel (NLP) en anglais, est un sous-domaine de la linguistique, des technologies de l'information et de l'intelligence artificielle qui se concentre sur les interactions entre les ordinateurs et le langage humain. Il englobe la programmation informatique visant le traitement et l'analyse des données exprimées dans un langage naturel. Le NLP vise à permettre aux machines de comprendre, d'interpréter et de générer du langage humain de manière efficace, en utilisant des techniques telles que la reconnaissance automatique de la parole, la traduction automatique, l'analyse sémantique, la génération de texte, la compréhension du langage naturel et bien d'autres. Ces avancées dans le domaine du NLP ouvrent de nombreuses possibilités dans des domaines tels que la recherche d'informations, les chatbots, l'analyse de sentiments, la classification de textes, le résumé automatique, et bien d'autres applications liées au langage humain et à l'interaction homme-machine.

#### **II. 5.1.1. Avantages du TAL**

Ci-après quelques avantages [26] offerts par le TAL :

- Traitement efficace d'une grande quantité de données textuelles en analysant rapidement d'énormes volumes de textes, permettant ainsi d'extraire des informations pertinentes et d'en tirer des connaissances précieuses.
- Automatisation des tâches linguistiques telles que la traduction, la génération ou la classification de texte et l'analyse sémantique.
- Faciliter les interactions homme-machine en comprenant et en traitant le langage naturel.

### **II. 5.1.2. Application pratique du TAL**

- Analyse des réponses des utilisateurs afin de comprendre leurs opinions, leurs sentiments et leurs besoins. Cela peut aider à améliorer les produits ou les services proposés en tenant compte des retours des utilisateurs.
- Classification automatique des données textuelles dans des catégories prédéfinies, ce qui peut être utile pour organiser et structurer les informations.
- Extraction d'informations importantes à partir de grandes quantités de données textuelles, facilitant ainsi l'analyse et l'interprétation des données.

### **II. 5.2. Apprentissage automatique :**

L'apprentissage automatique [27], également connu sous le nom de Machine Learning, est un domaine de l'intelligence artificielle qui se concentre sur le développement de techniques permettant aux ordinateurs d'apprendre à partir des données et d'accomplir des tâches spécifiques sans être explicitement programmés. Il repose sur l'idée de permettre aux machines d'acquérir des connaissances à partir de l'expérience et de s'améliorer au fil du temps au lieu de suivre des instructions précises. Les algorithmes peuvent analyser des ensembles de données volumineux pour extraire des informations précieuses, même dans des situations où les relations entre les variables sont complexes et non linéaires.

#### **5.2.1. Les techniques d'apprentissage automatique**

Nous allons explorer les techniques d'apprentissage automatique [28] afin de choisir la technique convenable à notre projet

##### **II. 5.2.1.1. Régression :**

La régression est une technique d'apprentissage automatique utilisée pour modéliser la relation entre une variable de sortie continue et des variables d'entrée. L'objectif est de prédire des valeurs numériques en se basant sur les relations et les tendances présentes dans les données d'entraînement.

##### **II. 5.2.1.2. Regroupement (Clustering) :**

Le regroupement est une méthode d'apprentissage non supervisé utilisée pour identifier des structures ou des groupes similaires dans un ensemble de données. Les algorithmes de

regroupement cherchent à regrouper les données en fonction de leurs similarités, sans avoir d'étiquettes de classe préexistantes.

### II. 5.2.1.3. Classification :

La classification est une technique d'apprentissage automatique utilisée pour prédire une variable de sortie discrète ou catégorique en fonction des caractéristiques d'entrée. L'objectif est de trouver des modèles et des règles permettant de classer correctement de nouvelles instances dans des classes prédéfinies.

### II. 5.2.2. Technique choisie

Comparons ces trois techniques en fonction de certains critères pour déterminer laquelle est la mieux adaptée à notre sujet.

Algorithme	Critères	
	Nature des données	Objectif de l'analyse
<b>Régression</b>	Les variables de sortie sont continues, c'est-à-dire des valeurs numériques.	L'objectif de la régression est de prédire des valeurs continues et de modéliser la relation entre les variables d'entrée et de sortie.
<b>Regroupement</b>	Utilisé lorsque nous souhaitons identifier des structures similaires ou des groupes au sein de nos données, sans avoir de variables de sortie prédéfinies.	L'objectif du regroupement est de trouver des structures similaires ou des groupes dans les données sans avoir d'étiquettes de classe préexistantes.
<b>Classification</b>	Utilisée lorsque nous avons des variables de sortie discrètes ou catégoriques.	L'objectif de la classification est de prédire la classe ou la catégorie d'une variable de sortie en se basant sur les caractéristiques d'entrée.

*Tableau II-5 Comparaison des techniques d'apprentissage automatique*

Pour notre sujet, nous cherchons à regrouper des données similaires, notre objectif principal est de regrouper les profils similaires à partir de nos données, le regroupement s'avère la technique la plus adaptée.

### **II. 5.2.3. Les algorithmes de regroupement**

Suite à notre choix du regroupement, nous allons maintenant explorer les algorithmes [29] de cette technique et leurs étapes d'application.

#### **II. 5.2.3.1. K-means**

K-means [30] partitionne les données en  $k$  clusters où  $k$  est un nombre prédéfini par l'utilisateur. Il converge vers une solution finale où les points de données sont regroupés de manière à minimiser la distance moyenne entre les points et les centres de cluster.

Le tableau II-6 présente les étapes du K-means:

<b>Étapes</b>	<b>Description</b>
<b>Initialisation</b>	Sélectionner aléatoirement $k$ centres de clusters à partir des données d'entrée.
<b>Attribution des points</b>	Chaque point de données est attribué au centre de cluster le plus proche en calculant la distance entre eux.
<b>Mise à jour des centres de cluster</b>	Recalculer les positions des $k$ centres de cluster.
<b>Boucle itérative</b>	Répéter les deux étapes précédentes jusqu'à ce que les centres ne bougent plus.
<b>Sortie</b>	Les centres de cluster déterminent les groupes dans lesquels les points de données sont regroupés.

Tableau II-6. Les étapes du K-means [31]

La figure II-3 représente un exemple d'application de K-means avec différentes valeurs de clusters :

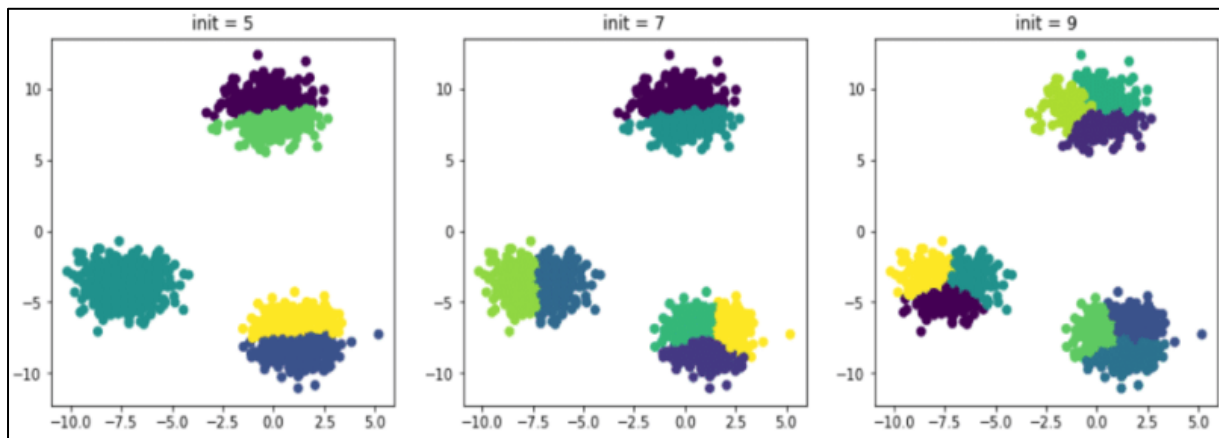


Figure II-3. K-means avec différents nombres de clusters

### II. 5.2.3.2. Meanshift

Mean-Shift [33] est un algorithme itératif qui a pour objectif de faire converger un point vers le maximum local le plus proche. Cet algorithme est basé sur une approche de déplacement de densité. Il fonctionne en déplaçant itérativement les centres de cluster vers les régions de densité maximale des données.

Le tableau II-7 présente les étapes du Mean-Shift :

Étapes	Description
<b>Initialisation</b>	Chaque point de données est utilisé comme centre initial.
<b>Calcul de la densité</b>	Pour chaque centre, une fenêtre de noyau (kernel window) est définie autour de ce centre, sa taille est déterminée par un paramètre de bande passante (bandwidth) qui contrôle l'influence des points à l'intérieur qui sont considérés comme des voisins.
<b>Déplacement des centres</b>	Déplacement en fonction de la moyenne pondérée des positions des points voisins. Cette étape est répétée jusqu'à ce que les centres ne se déplacent plus.
<b>Attribution des points</b>	Chaque point est attribué au centre le plus proche en fonction du calcul de la distance.
<b>Réduction du nombre des clusters</b>	Étape supplémentaire : les centres très proches peuvent être fusionnés en un seul cluster pour éviter la fragmentation.



Tableau II-7. Étapes pour appliquer l'algorithme de Mean-Shift [34]

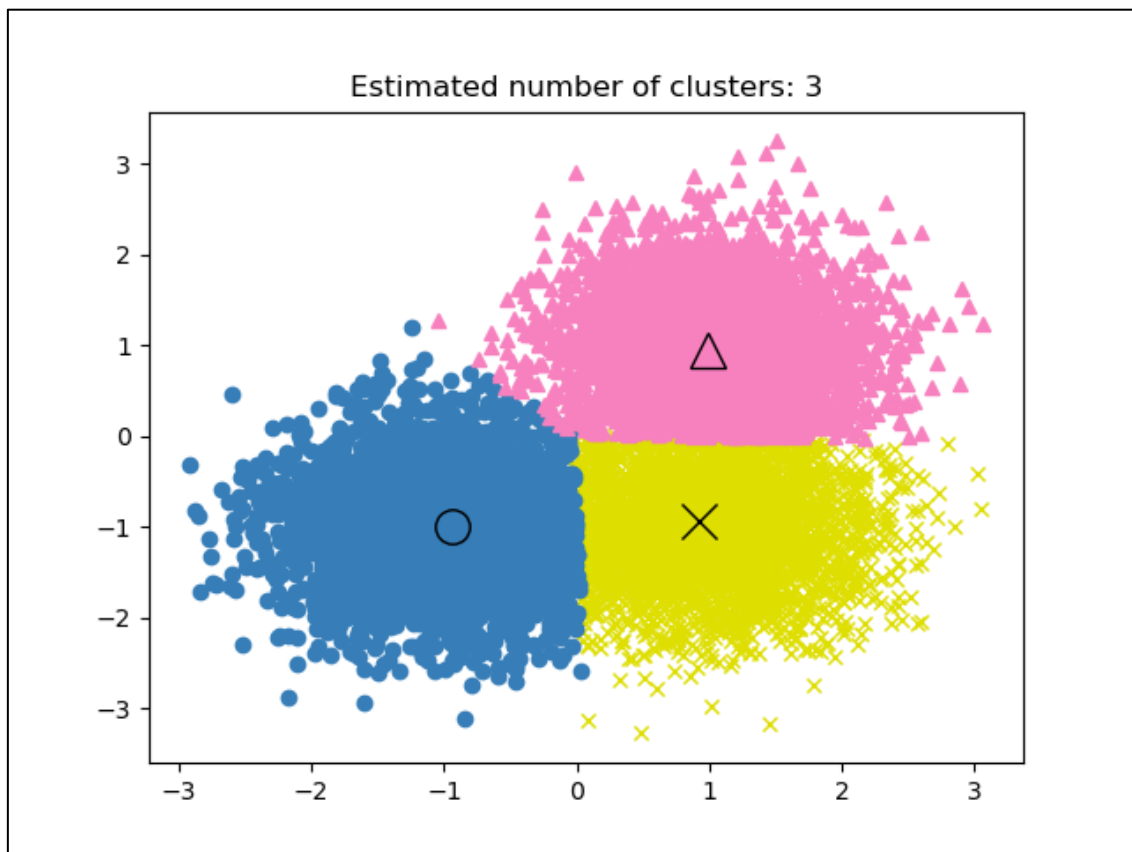


Figure II-4. Mean-Shift pour un nombre de clusters égal à 3

### II. 5.2.3.3. DBSCAN

DBSCAN [36] identifie les régions denses de points de données et les regroupe en clusters en utilisant l'estimation de la densité locale. Il peut également détecter les points de données qui ne font partie d'aucun cluster. Son fonctionnement repose sur deux paramètres principaux :

1. Epsilon ( $\epsilon$ ) : une mesure de la distance maximale qui sépare deux points. Lorsque tous les points ont une distance inférieure ou égale à  $\epsilon$  d'un autre point, ils sont considérés comme appartenant à son voisinage direct.
2. Le nombre minimum de points (MinPts) : C'est le nombre minimal de voisins pour qu'un point soit considéré comme un point central et pas un point aberrant.

Le tableau II-8 présente les étapes du DBSCAN :

	Étapes	Description
<b>Répétitions de ces étapes pour tous les points non visités</b>	<b>Sélection d'un point de départ</b>	Un point de départ non visité est choisi aléatoirement parmi les données d'entrée.
	<b>Recherche des voisins</b>	Examiner l'épsilon du point de départ pour déterminer les points voisins de ce dernier.
	<b>Vérification de la densité</b>	<ul style="list-style-type: none"> <li>- Nombre de voisins <math>\geq</math> MinPts : Le point de départ est considéré comme un point central et un nouveau cluster est créé.</li> <li>- Sinon : Le point de départ est marqué comme un point de bruit</li> </ul>
	<b>Expansion du cluster</b>	Ajout des points voisins du point central au cluster.
<b>Exploration des autres points</b>		L'arrêt du processus de répétitions des étapes précédentes lorsque tous les points soient attribués à un cluster ou marqués comme un bruit.
<b>Sortie</b>		Les points attribués à un même cluster sont regroupés ensembles.

Tableau II-8. Les étapes pour appliquer l'algorithme de DBSCAN [37]

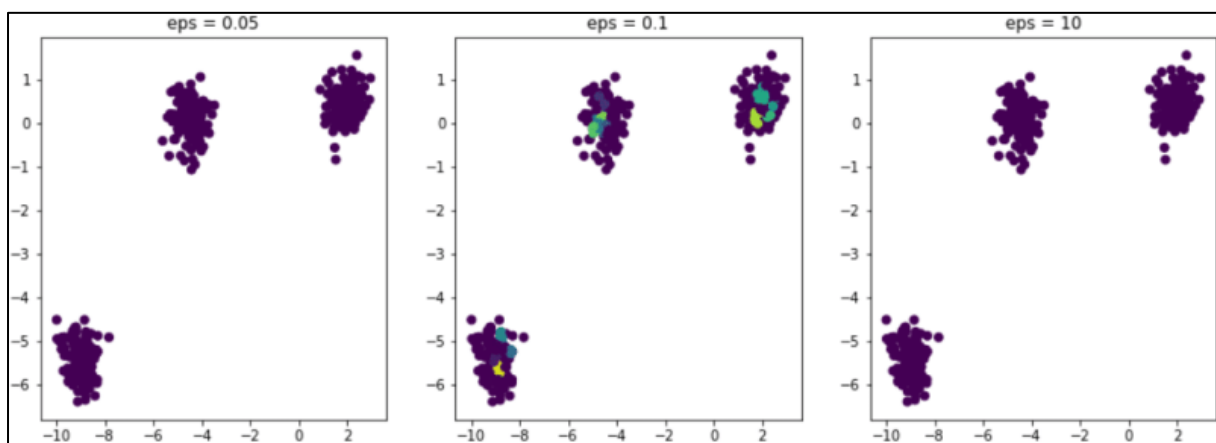


Figure II-5. DBSCAN avec 3 valeurs différentes d'épsilon [38]

## II. 5.2.4. Comparaison entre les algorithmes et l’algorithme choisi

Le tableau II-9 présente une comparaison faite entre K-means, DBSCAN et Mean-Shift :

Critères	K-means	DBSCAN	Mean-Shift
Paramètres	Nombre de clusters (k)	Epsilon MinPts	bandwidth
Choix de clusters	L'utilisateur doit spécifier le nombre de clusters.	Détermine automatiquement le nombre de clusters.	
Inconvénients	Sensible aux valeurs initiales des centres.	Sensible aux paramètres epsilon et minPts	Le choix de la bande passante est crucial
Résultat	Les centres finaux déterminent les groupes dans lesquels les points sont regroupés	Les clusters sont déterminés en fonction de la densité et de la connectivité des points	Les centres finaux déterminent les groupes dans lesquels les points sont regroupés

*Tableau II-9. Comparaison des algorithmes de regroupement*

Après cette comparaison, il ressort que Mean-Shift présente des caractéristiques uniques. Cet algorithme est choisi parce qu'il est capable d'identifier le nombre des clusters de forme arbitraire en fonction de la densité des données. De plus, Mean-Shift est capable de gérer des données présentant des variations de densité et n'est pas limité à des données spécifiques ou à des formes prédéfinies de clusters. Cela permet une flexibilité et une précision accrues dans la détection des structures de données complexes.

### **II. 5.3. Rôle de l'apprentissage automatique dans le traitement automatique des langues**

L'apprentissage automatique joue un rôle [39] essentiel dans la réalisation des tâches de TAL. Il permet de développer des modèles et des algorithmes capables de traiter et d'analyser efficacement les données en langage naturel. Il fournit les outils nécessaires pour comprendre et interpréter les structures linguistiques, extraire des informations, effectuer des traductions automatiques, effectuer des classifications de textes, et bien plus encore.

En combinant le TAL et l'apprentissage automatique, il est possible d'obtenir des systèmes plus performants et adaptatifs dans le traitement du langage humain.

### **II. 6. Bibliothèques, modules et classes utilisés**

Nous avons utilisé plusieurs bibliothèques, modules et classes qui nous ont permis d'accomplir différentes tâches pour effectuer le traitement automatique des langues et de l'apprentissage automatique. Ces outils puissants offrent des fonctionnalités spécifiques pour la manipulation, le traitement et l'analyse des données linguistiques, ainsi que pour la préparation des données et la création de modèles d'apprentissage automatique.

Le tableau II-10 présente les bibliothèques [40] que nous avons importées et leurs utilités :

<b>Bibliothèque</b>	<b>Description et utilité</b>
pandas	Manipulation et analyse des données.
numpy	Manipuler et analyser les données numériques.
Collections	Fournit des types de données spécialisés et efficaces pour la manipulation d'objets itérables, tels que les listes, les chaînes de caractères et les dictionnaires.
String	Utilisée pour effectuer des opérations sur les chaînes.
Nltk	Utilisée pour le traitement du langage naturel.
Ast	Utilisée pour convertir des chaînes de caractères en objets Python.

Scikit-learn	Fournit un large éventail d'outils et de fonctionnalités pour faciliter le développement, l'évaluation et le déploiement de modèles d'apprentissage automatique.
--------------	--

*Tableau II-10: Bibliothèques utilisées*

Le tableau II-11 présente les modules que nous avons importés et leurs descriptions :

Module	Description
Nltk.corpus	Fournit des ressources linguistiques pour NLTK
Nltk.stem	Fournit des classes pour la normalisation des mots
sklearn.cluster	Contient des classes et des fonctions pour le clustering des données.
sklearn.preprocessing	Contient des classes et des fonctions pour la préparation des données.
Sklearn.metrics	Fournit des métriques et des fonctions pour évaluer les performances des modèles d'apprentissage automatique.

*Tableau II-11 : Modules utilisés*

Le tableau II-12 présente les classes que nous avons utilisé et leurs descriptions :

Classe	Description
nltk.corpus.wordnet	Elle contient des synsets (ensembles de synonymes) et des relations sémantiques entre les mots. Elle est utilisée pour l'analyse sémantique et la recherche de synonymes.
nltk.corpus.stopwords	Contient des mots couramment utilisés et considérés comme non informatifs ou sans importance dans l'analyse de texte. Ces mots, tels que "le", "de", "et", sont souvent supprimés lors du prétraitement du texte.
nltk.stem.WordNetLemmatizer	Outil de lemmatisation, utilisé pour réduire les mots à leur forme de base ou lemmes.

sklearn.preprocessing.LabelEncoder	LabelEncoder est une classe du module sklearn.preprocessing. Il est utilisé pour encoder les variables catégorielles en nombres entiers. Cela permet de représenter les catégories de manière numérique dans les modèles d'apprentissage automatique.
sklearn.cluster.MeanShift	Utilisé pour la segmentation des données et l'identification de clusters.
collections.Counter	Counter est utilisé pour le comptage des éléments dans une liste ou une séquence. Il permet d'analyser la distribution et la fréquence des éléments.

*Tableau II-12. Classes utilisées*

## II. 7. Conclusion

Dans ce chapitre, nous avons réalisé une étude préliminaire approfondie afin de spécifier d'une part les besoins et de poser d'autre part les bases de notre projet. Nous avons ainsi identifié les exigences clés et les objectifs à atteindre, ce qui nous a permis de définir une vision claire de la solution cible. Nous avons également abordé deux aspects essentiels de notre étude, à savoir l'architecture logique et physique, ainsi que l'environnement de travail.

En ce qui concerne l'apprentissage automatique et le traitement automatique de langues, nous avons choisi avec soin les techniques et les algorithmes appropriés pour répondre aux exigences de notre projet. De plus, nous avons sélectionné les technologies adéquates pour le développement du back-end et du front-end de notre application, en tenant compte de leur pertinence et de leur adéquation aux besoins spécifiques.

Grâce à cette étude préliminaire approfondie, nous disposons désormais d'une base solide sur laquelle nous pourrons construire et mettre en œuvre notre solution de manière efficace.

# Chapitre III

---

## Proposition d'une solution d'apprentissage automatique

---

## **III . Proposition d'une solution d'apprentissage automatique**

---

### **III. 1. Introduction**

Dans ce chapitre, nous avons entrepris une série d'étapes clés pour analyser les données de profilage. Tout d'abord, nous avons extrait les données pertinentes de MongoDB, une base de données flexible et évolutive. Ensuite, nous avons effectué une compréhension approfondie des données et les avons traitées pour les rendre plus structurées et exploitables. Nous avons également identifié les critères de profiling pertinents et créé un dictionnaire pour les catégoriser. Enfin, nous avons appliqué l'algorithme de clustering MeanShift pour regrouper les réponses des utilisateurs similaires et détecter des tendances significatives.

### **III. 2. Collecte des données**

Dans cette section, nous explorerons le processus de collecte des données, en commençant par examiner la source de stockage des données ensuite nous aborderons l'étape de l'extraction des réponses.

#### **III. 2.1. Source de stockage de données**

Les réponses de tous les utilisateurs qui ont participé à une campagne spécifique sont enregistrées dans une collection MongoDB nommée "campaignanswers". Cette collection est constituée de plusieurs documents, chaque document représente la participation d'un utilisateur à une campagne spécifique, identifiée par leurs identifiants respectifs. Chaque document contient également l'identifiant de chaque question de la campagne, accompagné de la valeur correspondante qui représente la réponse de l'utilisateur.



La figure III-1 représente un document ayant l'identifiant de la campagne « Votre profil »

```
_id: ObjectId('634c8d18d2e2ca002b3f22c7')
score: 0
▼ answers: Array
  ▼ 0: Object
    _id: ObjectId('634c8d2ad34855002af9640c')
    ► questionAnswers: Array
      questionId: "634c6b9ebeccf0002a7d0f9a"
      updatedAt: 2022-10-16T23:00:58.708+00:00
      createdAt: 2022-10-16T23:00:58.708+00:00
    ► 1: Object
    ► 2: Object
    ► 3: Object
    ► 4: Object
  campaignId: ObjectId('634c6b9fbeccf0002a7d0fa7')
  status: "done"
  progress: 100
  ► deviceConfiguration: Object
    userId: "634c499bc1caa5002b90bada"
    createdAt: 2022-10-16T23:00:40.173+00:00
    updatedAt: 2022-10-16T23:00:58.708+00:00
    __v: 0
  ► geoLocation: Object
```

Figure III-1. Un document de la collection "campaignanswers »

### III. 2.2. Extraction de données

Le processus d'extraction des données des réponses des utilisateurs de trois campagnes « Queney », « Votre profil » et « Pour mieux vous servir » a été réalisé selon les étapes suivantes :

Étape	Description
<b>Sélection des campagnes</b>	Dans le backoffice de l'application JAYEG, nous avons choisi trois campagnes spécifiques pour lesquelles nous souhaitons extraire les réponses des utilisateurs.
<b>Activation de l'extraction</b>	Une fois la campagne sélectionnée, nous avons cliqué sur le bouton "Export to CSV" pour déclencher le processus d'extraction des données.
<b>Extraction des données à partir de MongoDB</b>	Le processus d'extraction des données associées à la campagne sélectionnée est déclenché automatiquement, extrayant ainsi les réponses des utilisateurs.
<b>Génération du fichier CSV</b>	Une fois l'extraction des données terminée, nous avons reçu un courrier électronique contenant un lien permettant de télécharger le fichier CSV contenant les données extraites. Ce fichier CSV est

	structuré de manière à fournir une représentation tabulaire des réponses des utilisateurs pour chaque campagne.
--	---

Tableau III-1. Les étapes de l'extraction des données

La figure III-2 illustre l'interface du Backoffice de l'application « JAYEG » :

Reporting Pour mieux vous servir!!!

Campaigns List / Pour mieux vous servir!!! / Reporting

Export to CSV

Global

Score LeaderBoard

Questions

List

Q Search in the list des utilisateurs

id	je suis?	Quel est ...	Dans quel...	je suis...	j'habite?...	Pour me d...	Que faite...	Quel spor...	Parmi ces...	As-tu acc...	Avez- vou...	Vous aime...	Votre typ...	vous préf...
63e37513e...	Un homme...	moins de ...	Jendouba...	au collég...	Dans un l...	le vélo...	autre	Football...	CLUB AFRI...	false	true	Non	action...	Nike, Lac...
63e764bb1...	Un homme...	entre 18 ...	Tozeur...	au lycée...	Dans un f...	le vélo...	Internet...	Football...	ES SAHEL...	true	true	Oui	comédie...	Chanel, S...
63e771458...	Un homme...	moins de ...	Nabeul...	au collég...	Chez mes ...	autre   l...	Internet...	Football...	CLUB AFRI...	true	false	Oui	action...	Nike, Lac...
63e3f7af7...	Un homme...	entre 18 ...	sidi bouz...	à l'unive...	Chez mes ...	autre	Activités...	Football ...	CLUB AFRI...	false	false	Non	tragédie...	Nike, Lac...
63e7d6de8...	Un homme...	moins de ...	Jendouba...	au lycée...	Chez mes ...	autre	Activités...	Football...	CLUB AFRI...	false	false	Oui	action...	Nike, Lac...
63e8c8ad8...	Un homme...	entre 18 ...	Tunis	à l'unive...	autre	le taxi  ...	Internet...	Football...	ES TUNIS...	true	false	Oui	action...	Nike, Lac...
63e3e54d7...	Une femme...	entre 18 ...	Kef	à l'unive...	Chez mes ...	le taxi...	Activités...	Football...	CLUB AFRI...	true	false	Oui	action...	Chanel, S...

Activier Windows

Items per page: 10

1 - 10 of 3500

Accédez aux paramètres pour activer Windows.

Figure III-2. Reporting d'une campagne dans le Backoffice de l'application "JAYEG"

### III. 3. Compréhension des données

Dans cette section, nous aborderons la phase de compréhension des données.

#### III. 3.1. Analyse exploratoire des données

Nous avons à notre disposition 3 fichiers CSV correspondant à trois campagnes distinctes. Chaque fichier est structuré en colonnes comprenant les questions posées, une colonne "User id" contenant l'identifiant de chaque utilisateur, une colonne "S.No" contenant un identifiant généré par MongoDB, ainsi que les colonnes "Created at" et "Updated at" pour indiquer les

dates de création et de mise à jour des données. Les lignes de chaque fichier représentent les réponses associées à chaque question posée. Voici les détails spécifiques de chaque fichier :

Fichiers	Queney	Votre profil	Pour mieux vous servir
Lignes	38 221	12 728	38 955
Colonnes	20	10	18

Tableau III-2. Composition des fichiers CSV

Pour chaque campagne, nous avons examiné le type de données correspondant à chaque colonne ainsi que les réponses possibles pour chaque question.

### III. 3.1.1. La campagne « Pour mieux vous servir »

Type de données		Réponses possibles
je suis?	object	[Un homme , Une femme , -]
Quel est ton âge?	object	[moins de 18 ans , entre 18 et 24 ans , entre ...]
Dans quelle region vous habitez?	object	[sidi bouzid , Tunis , Siliana , Nabeul , Sous...]
je suis	object	[au lycée , à l'université , au collège , entr...]
j'habite?	object	[Chez mes parents , Dans un foyer , Dans un ap...]
Pour me déplacer je prends	object	[le bus , autre , la voiture , le taxi , le tr...]
Que faites vous pendant votre temps libre?	object	[Activités sportives , autre , TV/jeux vidéos ...]
Quel sport pratiquez vous?	object	[Football , autre , Football -Basketball -Nata...]
Parmi ces clubs quel est votre club préféré ?	object	[CLUB AFRICAIN , ES TUNIS , ES SAHEL , US MONA...]
As-tu accès à internet facilement?	object	[false , true ]
Avez- vous un permis de conduire?	object	[false , true ]
Vous aimez les jeux videos?	object	[Non , Oui , -]
Votre type de film préféré	object	[action , comédie , science fiction , horreur ...]
vous préférez quelles marques	object	[Chanel, Saint laurent, Dior et Louis vuitton ...]

Figure III-3 Type de données et réponses possibles pour chaque question de la campagne "Pour mieux vous servir"

### III. 3.1.2. La campagne « Votre profil »

Type de données		Réponses possibles
Je suis un (e)	object	[Homme , Femme ]
le revenue de votre famille	object	[entre 1200 et 1800 dtn par mois , entre 3000 ...
Vous disposez (plusieurs réponses possible)	object	[une machine a laver automatique , aucun , un ...
vous êtes	object	[propriétaire , locataire , -]
Etes vous motorisé?	object	[true , false ]
combien de voitures disposez vous dans votre famille	object	[aucune , 1, 2, plus que 2 , -]

Figure III-4 Type de données et réponses possibles pour chaque question de la campagne "Votre profil"

### III. 3.1.3. La campagne « Queney »

Type de données		Réponses possibles
je suis?	object	[Un garçon , Une fille , -]
Quel est ton âge?	object	[entre 18 et 24 ans , moins de 18 ans , entre ...
je suis	object	[au lycée , à l'université , au collège , sala...
j'habite?	object	[Chez mes parents , Dans un internat , Dans un...
Pour me déplacer je prends	object	[le bus , le vélo , le metro , la voiture , la...
Que fais tu pendant ton temps libre?	object	[Activités sportives -TV/jeux vidéos , Interne...
Quel sport pratiquez vous?	object	[Football , Aucun , Basketball , Football -Nat...
Parmi ces clubs quel est votre club préféré ?	object	[ES TUNIS , ES SAHEL , CLUB AFRICAINE , CS SFAX...
As-tu accès à internet facilement?	object	[false , true ]
Vous avez un compte?	object	[Facebook , Facebook -Instagram -Tik Tok -Snap...
En général quel type d'information as tu besoin?	object	[Etudes/Formations , Santé , Vie pratique (tra...
Plus globalement, as-tu des projets personnels ?	object	[projet d'etudes , Voyage , pas de projet en p...
T'intéresses tu à la politique ?	object	[false , true ]
Vois-tu un intérêt à voter ?	object	[false , true ]
D'après toi, qu'est-ce qui manque le plus aux jeunes de ton âge?	object	[Des évènements culturels , Des structures spo...
Avez- vous un permis de conduire?	object	[false , true ]

Figure III-5 Type de données et réponses possibles pour chaque question de la campagne "Queney"

### III. 3.2. Évaluation de la qualité des données

Après avoir exploré les données, nous avons remarqué que certaines colonnes telles que "S.No", "Created at" et "Updated at" sont présentes mais ne sont pas nécessaires pour notre analyse.

Nous avons aussi identifié un schéma récurrent : certaines questions sont présentes dans les trois campagnes, souvent formulées de manière similaire, mais les réponses fournies par les utilisateurs ne sont pas cohérentes. Cette incohérence peut être attribuée à plusieurs causes potentielles :

- Différences d'interprétation des questions : Les utilisateurs peuvent comprendre les questions de manière différente.
- Influence des circonstances : Les réponses peuvent être influencées par des facteurs externes tels que l'humeur, les événements récents ou les préoccupations personnelles.
- Réponses impromptues pour obtenir des points : Il est possible que certains utilisateurs répondent de manière précipitée et sans réflexion dans le seul but de terminer leurs réponses et d'accumuler des points.

La figure III-6 présente un exemple de réponses incohérentes d'un utilisateur pour trois questions relatives au genre :

User id	je suis?	je suis?	Je suis un (e)
627e8554e0f7fe002b52e5af	homme	fille	femme

*Figure III-6. Exemple de réponses incohérentes pour les questions relatives au genre*

De plus, lors de l'exploration des données, nous avons également constaté la présence de valeurs manquantes pour certaines questions. Cela indique que certains utilisateurs ont choisi de ne pas répondre à ces questions ou ont omis de le faire, ce qui peut également introduire des incohérences dans les données.

La figure III-7 présente le nombre de réponses vide d'un utilisateur pour les trois campagnes :

<code>queney.stack().str.count('-').sum()</code>
149233
<code>profil.stack().str.count('-').sum()</code>
1116
<code>servir.stack().str.count('-').sum()</code>
32513

*Figure III-7. Nombre de réponses vides*

### III. 4. Préparation des données

Dans cette section, nous aborderons la phase de préparation des données.

#### III. 4.1. Création d'un dictionnaire de synonymes

Dans le but d'améliorer la performance de notre modèle d'apprentissage automatique, nous avons mis en place un dictionnaire de synonymes. Ce dictionnaire joue un rôle crucial dans la comparaison des réponses fournies par les utilisateurs et dans la vérification de leur cohérence. Il est spécifiquement conçu en fonction des critères de profilage que nous avons sélectionnés. Chaque critère de profilage est associé à des sous-éléments qui regroupent les synonymes pertinents.

Le tableau III-4 présente les critères de profiling définis et leurs descriptions

Critère	Description
<b>Âge</b>	Cette caractéristique peut être utilisée pour segmenter les individus en groupes générationnels et fournir des indications sur leurs préférences, leurs comportements et leurs besoins spécifiques en fonction de leur stade de vie.
<b>Genre</b>	Cette caractéristique peut être utilisée pour comprendre les différences dans les préférences, les comportements et les besoins en fonction du genre.

<b>Permis de conduire</b>	Cette caractéristique peut être pertinente pour des analyses liées à la mobilité, aux préférences de transport et aux comportements liés à la conduite.
<b>Niveau d'études</b>	Cette caractéristique peut fournir des informations sur les compétences, les connaissances et les intérêts particuliers d'une personne.
<b>Situation familiale</b>	Cette caractéristique peut être utile pour comprendre les responsabilités familiales, les priorités et les besoins spécifiques d'une personne.
<b>Revenue</b>	Cette caractéristique peut fournir des informations sur le pouvoir d'achat, le niveau de vie et les préférences de consommation d'une personne.
<b>Zone géographique</b>	Cette caractéristique peut être utilisée pour comprendre les différences culturelles, les influences régionales et les habitudes de consommation spécifiques à une zone géographique donnée.
<b>Interactions en lignes</b>	Cette caractéristique peut être utilisée pour comprendre les préférences en ligne, les intérêts spécifiques et les comportements de navigation.
<b>Langues parlées</b>	Cette caractéristique peut être pertinente pour la personnalisation des communications et des offres, ainsi que pour la compréhension des influences culturelles.
<b>Occupation</b>	Cette caractéristique peut fournir des informations sur les intérêts professionnels, les compétences et les influences socio-économiques.
<b>Intérêts</b>	Cette caractéristique peut être utilisée pour cibler des produits, des services ou des expériences en fonction des intérêts personnels et des préférences.

*Tableau III-3. Description de l'utilité de chaque caractéristique dans le profiling*

La figure III-8 représente le dictionnaire créé :

```
{
  "Genre": {
    "similarite": ["homme", "femme", "masculin", "féminin", "garçon", "fille"],
    "synonymes": {
      "masculin": ["garçon", "homme"],
      "féminin": ["fille", "femme"]
    }
  },
  "Club": {
    "similarite": [],
    "synonymes": {}
  },
  "Age": {
    "similarite": ["an"],
    "synonymes": {}
  },
  "Situation familiale": {
    "similarite": [],
    "synonymes": {}
  },
  "Interactions en ligne": {
    "similarite": ["facebook", "instagram", "tik tok", "snapshat"],
    "synonymes": {}
  },
  "Revenu": {
    "similarite": ["dtn", "mois"],
    "synonymes": {}
  },
  "Niveau d'études": {
    "similarite": [],
    "synonymes": {}
  },
  "Permis de conduire": {
    "similarite": ["permis", "conduire"],
    "synonymes": {}
  },
  "Motorisé": {
    "similarite": ["motorisé?"],
    "synonymes": {}
  },
  "Langues": {
    "similarite": ["français", "anglais", "arabe"],
    "synonymes": {
      "fr": ["français", "french"],
      "eng": ["anglais", "english"],
      "ar": ["arabic", "arabe"]}
  },
  "Film": {
    "similarite": ["science", "fiction", "horreur", "action", "comédie", "tragédie", "romantique", "policier", "film"],
    "synonymes": {
      "fiction": ["science", "fiction"],
      "horreur": ["horreur"],
      "action": ["action"],
      "comédie": ["comédie"],
      "tragédie": ["tragédie"],
      "romantique": ["romantique"],
      "policier": ["policier"]
    }
  },
  "Zone géographique": {
    "similarite": ["nabeul", "jendouba", "tunis", "sousse", "siliana", "sfax", "kairouan", "gafsa", "mahdia", "kef", "bizerte", "gabes", "medenine", "zaghouan", "sidi bouzid", "tataouine", "tozeur", "ariana", "ben arous", "monastir", "kasserine", "manouba", "beja"],
    "synonymes": {
      "na": ["nabeul"],
      "je": ["jendouba"],
      "tu": ["tunis"],
      "sou": ["sousse"],
      "sf": ["sfax"],
      "ka": ["kairouan"],
      "gaf": ["gafsa"],
      "ma": ["mahdia"],
      "ke": ["kef"],
      "biz": ["bizerte"],
      "gab": ["gabes"],
      "med": ["medenine"],
      "zag": ["zaghouan"],
      "sidi": ["sidi bouzid"],
      "ta": ["tataouine"],
      "to": ["tozeur"],
      "ar": ["ariana"],
      "ba": ["ben arous"],
      "mon": ["monastir"],
      "kass": ["kasserine"],
      "man": ["manouba"],
      "be": ["beja"]
    }
  },
  "Occupation": {
    "similarite": ["entrepreneur", "recherche d'un emploi", "lycée", "l'université", "salariée", "collège", "métier", "occupation", "job", "place", "travail", "emploi", "profession", "tâche", "fonction"],
    "synonymes": {
      "entr": ["entrepreneur"],
      "rech": ["recherche d'un emploi"],
      "ly": ["lycée"],
      "univ": ["l'université"],
      "sal": ["salariée"],
      "co": ["collège"]
    }
  },
  "Sport": {
    "similarite": ["handball", "football", "volleyball", "basketball", "gymnastique", "natation", "tennis", "golf", "danse", "activités sportives"],
    "synonymes": {}
  }
}
```

Figure III-8: Le dictionnaire



### III. 4.2. Structuration des données

Dans cette section, nous allons explorer les différentes étapes que nous avons effectué pour contribuer à la structuration des données.

#### III. 4.2.1. Opération de jointure

Afin de structurer nos données, nous commençons par effectuer une opération de jointure des trois fichiers CSV, en utilisant l'identifiant de l'utilisateur (User id) comme clé commune pour rassembler les données des utilisateurs qui ont participé aux trois campagnes sélectionnées.

La figure III-9 représente la fonction que nous avons utilisé afin de rassembler les trois campagnes dans une même trame de données.

```
def jointure():
    servir=lecture_fichier('servir.csv')
    profil=lecture_fichier('profil.csv')
    queney=lecture_fichier('queney.csv')
    df = pd.merge(servir, queney, on='User id', suffixes=('', ' '))
    df = pd.merge(df, profil, on='User id', suffixes=('', ' '))
    df.insert(0, 'User id', df.pop('User id'))
    return df
```

Figure III-9. Fonction pour effectuer la jointure entre les trois campagnes

#### III. 4.2.2. La normalisation de la trame de données

La normalisation [41] fait référence à un ensemble de techniques utilisées pour standardiser les données textuelles. Dans ce contexte, nous avons normalisé la trame de données résultante de la jointure en convertissant les noms de colonnes en minuscules et en supprimant la ponctuation. De plus, nous avons défini les stopwords et initialisé le lemmatizer pour la langue française. Ensuite, nous avons prétraité les données en lemmatisant les mots et en supprimant les stopwords, à la fois pour chaque valeur dans la trame de données et pour les noms de colonnes. Cela nous permet d'avoir une représentation cohérente des données pour une analyse ultérieure.

La figure III-11 représente avant et après l'application de la normalisation :

Avant normalisation	Après normalisation
Un homme	homme
Une femme	femme

Figure III-10. Un exemple illustrant avant et après la normalisation

### III. 4.2.3. Organisation et qualification des questions selon des critères spécifiques dans la trame de données

L'objectif des étapes suivantes est de regrouper les questions répondant à un même critère afin de procéder à l'étape de comparaison des réponses et avoir pour chaque critère une réponse vérifiée.

#### III. 4.2.3.1. Recherche de questions répondant à un même critère

Dans un premier temps, nous avons structuré notre jeu de données en regroupant les questions en fonction des critères définis dans notre dictionnaire. Pour atteindre cet objectif, nous avons créé trois fonctions qui nous a permis d'organiser notre trame de données en identifiant les colonnes compatibles avec les critères définis. Voici une explication de ces trois fonctions en détaillant le résultat obtenu :

- La fonction **compatible\_columns** parcourt les colonnes d'une ligne d'un DataFrame et compare les réponses avec le dictionnaire de synonymes. Elle recherche les mots similaires et les synonymes dans chaque valeur de colonne. Si un mot similaire ou un synonyme est trouvé, la colonne correspondante est considérée comme compatible avec un critère correspondant dans le dictionnaire.

```

def compatible_columns(row, synonym_dict):
    compatible_columns = {} # Dictionnaire pour stocker les colonnes compatibles avec les critères
    for col in row.index:
        #variable value pour stocker la réponse de la colonne
        value = row[col]
        if isinstance(value, str): # Vérifier si la réponse est une chaîne de caractères
            words = value.split()
            for critere, details in synonym_dict.items():
                similarite = details.get('similarite', []) # Liste des mots similaires
                synonymes = details.get('synonymes', {}) # Dictionnaire des synonymes
                # Vérifier si un mot similaire est présent dans les mots de la valeur
                if any(synonyme in words for synonyme in similarite):
                    if critere not in compatible_columns:
                        compatible_columns[critere] = [col]
                    else:
                        compatible_columns[critere].append(col)
            else:
                for synonyme, synonyme_values in synonymes.items():
                    # Vérifier si un synonyme est présent dans les mots de la réponses
                    if any(syn in words for syn in synonyme_values):
                        if critere not in compatible_columns:
                            compatible_columns[critere] = [col]
                        else:
                            compatible_columns[critere].append(col)

    if compatible_columns: # Vérifier si des colonnes compatibles ont été trouvées
        return True, compatible_columns

    return False, None

```

Figure III-11. Fonction compatible\_columns

- La fonction compatible\_columns\_lignes parcourt les lignes d'un DataFrame et applique la fonction compatible\_columns à chaque ligne. Elle utilise un dictionnaire de synonymes pour déterminer les colonnes compatibles pour chaque ligne.

```

def compatible_columns_lignes(df):
    for i in range(df.shape[0]):
        # Récupérer une ligne du DataFrame à l'indice i
        result = compatible_columns(df.iloc[i], lecture_dictionnaire('dic.txt'))

        # Vérifier si des colonnes compatibles ont été trouvées pour la ligne actuelle
        if result is not None and result[0] == True:
            # Retourner True et le dictionnaire des colonnes compatibles
            return result[0], result[1]
    # Si aucune ligne ne contient de colonnes compatibles, retourner False et None
    return False, None

```

Figure III-12. Fonction compatible\_columns\_lignes

La figure III-13 représente le résultat de la fonction `compatible_columns_lignes` :

```
{'Genre': ['je suis?', 'je suis? ', 'Je suis un (e)'],
 'Age': ['Quel est ton âge?', 'Quel est ton âge? '],
 'Zone géographique': ['Dans quelle region vous habitez?'],
 'Occupation': ['je suis', 'je suis '],
 'Sport': ['Quel sport pratiquez vous?', 'Quel sport pratiquez vous? '],
 'Film': ['Votre type de film préféré'],
 'Interactions en ligne': ['Vous avez un compte?'],
 'Revenu': ['le revenue de votre famille']}
```

Figure III-14. Résultat de la fonction `compatible_columns_lignes`

- La fonction **`compatible_columns_colonnes`** parcourt les colonnes d'un DataFrame et recherche des critères de compatibilité dans les noms de colonnes. Elle utilise le dictionnaire pour déterminer si un mot similaire ou un synonyme est présent dans le nom de la colonne. Si un critère de compatibilité est satisfait, la colonne est ajoutée au dictionnaire des colonnes compatibles correspondantes.

```
def compatible_columns_colonnes(df, synonym_dict):
    compatible_columns = {} # Dictionnaire pour stocker les colonnes compatibles avec les critères
    for col in df.columns:
        value = str(col)
        words = value.split()
        for critere, details in synonym_dict.items():
            similarite = details["similarite"] # Liste des mots similaires
            synonymes = details["synonymes"] # Dictionnaire des synonymes
            # Vérifier si un mot similaire est présent dans les mots de la réponse
            if any(synonyme in words for synonyme in similarite):
                if critere not in compatible_columns:
                    compatible_columns[critere] = [col] # Ajouter la colonne au dictionnaire des colonnes compatibles
                else:
                    # Ajouter la colonne à la liste des colonnes compatibles pour le critère
                    compatible_columns[critere].append(col)
            for syn, syn_values in synonymes.items():
                # Vérifier si un synonyme est présent dans les mots de la valeur
                if any(syn_value in words for syn_value in syn_values):
                    if critere not in compatible_columns:
                        # Ajouter la colonne au dictionnaire des colonnes compatibles
                        compatible_columns[critere] = [col]
                    else:
                        # Ajouter la colonne à la liste des colonnes compatibles pour le critère
                        compatible_columns[critere].append(col)

    if compatible_columns: # Vérifier si des colonnes compatibles ont été trouvées
        return compatible_columns

    return None
```

Figure III-15. Fonction `compatible_columns_colonnes`

La figure III-16 représente le résultat de la fonction compatible\_columns\_colonnes :

```
{'Permis de conduire': ['Avez- vous un permis de conduire?',  
  'Avez- vous un permis de conduire? '],  
  'Film': ['Votre type de film préféré'],  
  'Motorisé': ['Etes vous motorisé?']}
```

*Figure III-17. Résultat de la fonction compatible\_columns\_lignes*

### **III. 4.2.3.2. Croisement des réponses**

Dans cette étape, nous avons créé une trame de données contenant le résultat du regroupement des questions suivies de deux colonnes supplémentaires :

1. "critere\_Qualification" : Dans cette colonne, nous avons évalué la cohérence des réponses en utilisant une fonction de comparaison, attribuant les valeurs "Cohérent" ou "Incohérent" selon le cas.
2. "nom\_critere". Cette colonne a été remplie en déterminant la réponse appropriée pour chaque critère, grâce à une fonction dédiée. Ainsi, la trame de données a été structurée de manière à faciliter l'analyse ultérieure.

Suite à cette étape, cette trame de données est composée de 37 colonnes.

La figure III-18 illustre le résultat des fonctions qui ont permis de déterminer la cohérence des réponses et le choix d'une valeur correcte.

User id	je suis?	je suis?	Je suis un (e)	Genre Qualification	Genre	Quel est ton âge?	Quel est ton âge?	Age Qualification	Age
6234f7b96d818c001e35edc9	homme	garçon	homme	Cohérent	homme	entre 25 34 an	entre 25 34 an	Cohérent	entre 25 34 an
62057143cbad6e0020e11f6e	femme	filie	femme	Cohérent	femme	entre 18 24 an	entre 18 24 an	Cohérent	entre 18 24 an
6218aa29a2acb3001fb2cab5	homme	garçon	homme	Cohérent	homme	moins 18 an	moins 18 an	Cohérent	moins 18 an
620d56c47db2fa002a6ab6ae	homme	garçon	homme	Cohérent	homme	entre 18 24 an	entre 18 24 an	Cohérent	entre 18 24 an
627e8554e0f7fe002b52e5af	homme	filie	femme	Incohérent	None	entre 25 34 an	entre 18 24 an	Incohérent	None

Figure III-18. Trame de données structurée pour les critères "Genre" et "Age".

### III. 4.2.4. Création d'une nouvelle trame de données avec les réponses correctes pour chaque critère

Dans cette étape, nous avons créé une nouvelle trame de données en extrayant les colonnes correspondant à chaque critère et en incluant uniquement les réponses correctes de la trame de données précédent. Cette nouvelle trame de données est structurée de manière à regrouper les réponses correctes associées à chaque critère spécifique. Cela nous permet d'avoir une vision claire et organisée des réponses cohérentes pour chaque critère, ce qui facilite les analyses ultérieures et la prise de décision basée sur des données fiables.

Nous avons comme résultat une trame de données composée de 3120 lignes et 12 colonnes.

User id	Club	Genre	Age	Zone géographique	Occupation	Sport	Film	Interactions en ligne	Revenu	Permis de conduire	Motorisé
6234f7b96d818c001e35edc9	ES, SAHEL	homme	entre 25 34 an	nabeul	entrepreneur	football	science fiction	facebook instagram tik tok	plus der 5000 dtn mois	true	true
62057143cbad6e0020e11f6e	ES, TUNIS	femme	entre 18 24 an	jendouba	None	None	horreur	facebook instagram tik tok	entre 650 1200 dtn mois	false	false
6218aa29a2acb3001fb2cab5	ES, TUNIS	homme	moins 18 an	tunis	lycée	None	horreur	facebook tik tok	entre 3000 5000 dtn mois	None	None
620d56c47db2fa002a6ab6ae	ES, TUNIS	homme	entre 18 24 an	tunis	None	aucun	action	facebook instagram autre	entre 1800 3000 dtn mois	None	false
627e8554e0f7fe002b52e5af	None	None	None	nabeul	l'université	handball	horreur	tik tok instagram facebook snapchat	entre 3000 5000 dtn mois	true	true
...	...	...	...	...	...	...	...	...	...	...	...
6231bf3a6d818c001e346bf2	CLUB, AFRICAÏN	homme	entre 25 34 an	tunis	None	None	action		entre 400 650 dtn mois	true	false
62c767dc3317a600295fd1fd	None	homme	entre 18 24 an	tunis	l'université	None	action	facebook instagram snapchat	entre 1800 3000 dtn mois	None	false
6206015acbad6e0020e15750	CS, SFAXIEN	homme	entre 25 34 an	sfax	salariée	None	action	facebook instagram tik tok autre	entre 1200 1800 dtn mois	true	true
6208fda0f75cbe001f0a800a	None	homme	entre 18 24	sfax	None	volleyball	horreur	facebook instagram tik tok	entre 1000 1800 dtn mois	None	false

Figure III-19. Trame de données des critères

### III. 4.2.5. Création d'une nouvelle trame de données avec les réponses correctes pour les autres questions

Dans cette étape, nous avons créé une nouvelle trame de données en extrayant les colonnes qui ne correspondent à aucun critère et en incluant uniquement les réponses correctes de la trame de données résultante de la comparaison faite précédemment.

Nous avons comme résultat une trame de données composée de 3120 lignes et 17 colonnes.

User id	Que faites vous pendant votre temps libre?	Vous aimez les jeux vidéos?	vous préférez quelles marques	Que fais tu pendant ton temps libre?	En général quel type d'information as tu besoin?	Plus globalement, as-tu des projets personnels ?	T'intéresses tu à la politique ?	Vois-tu un intérêt à voter ?	D'après toi, qu'est-ce qui manque le plus aux jeunes de ton âge?	Vous disposez (plusieurs réponses possible)
6234f7b96d818c001e35edc9	Sorties entre amis	Oui	Nike, Lacoste, Boss, Prada	Internet - TV/jeux vidéos - Sorties entre amis -...	Sport et loisirs - Projets à l'étranger (voyages...	projet professionnel -Voyage	true	true	Des espaces pour jeunes - Des lieux où s'infor...	une machine a laver automatique
62057143cbad6e0020e11f6e	Activités sportives - TV/jeux vidéos - Internet	Oui	Chanel, Saint laurent, Dior et Louis vuitton	Activités sportives	Etudes/Formations -Offres d'emploi	projet d'etudes - Voyage	false	false	Des événements culturels	un congélateur hors celui du réfrigérateur
6218aa29a2acb3001fb2cab5	Activités associatives	Non	Nike, Lacoste, Boss, Prada	Internet	Etudes/Formations	projet professionnel	false	true	Des événements culturels	une machine a laver semi automatique
620d56c47db2fa002a6ab6ae	Internet - TV/jeux vidéos	Oui	Nike, Lacoste, Boss, Prada	Activités sportives - Internet - TV/jeux vidéos ...	Etudes/Formations	Voyage - projet professionnel	false	false	Des transports - La sécurité - Des instances p...	une machine a laver automatique
627e8554e0f7fe002b52e5af	Activités associatives	Oui	Nike, Lacoste, Boss, Prada	Activités associatives	Vie pratique (transports, horaires d'ouverture...	Voyage - projet d'etudes - projet professionnel	true	true	Des transports	une machine a laver automatique

Figure III-20. Trame de données des autres questions

### III. 5. Modélisation

Dans cette partie nous nous intéressons à l'application de l'algorithme de regroupement Mean-Shift.

#### III. 5.1. Encodage et traitement des données

Avant d'appliquer l'algorithme de MeanShift, nous avons effectué un encodage et un traitement des données afin de les rendre compatibles avec l'algorithme d'apprentissage automatique et prêtes à être utilisées. Durant cette étape nous avons identifié les colonnes catégorielles et nous avons utilisé la technique d'encodage Label Encoding.



La figure III-21 présente les instructions nécessaires pour encoder les colonnes catégorielles :

```
# Prétraitement des colonnes catégorielles
colonnes_catégorielles = tableau_copie.select_dtypes(include=["object"]).columns
for colonne in colonnes_catégorielles:
    # Encoder les valeurs catégorielles en numériques
    label_encoder = LabelEncoder()
    tableau_copie[colonne] = label_encoder.fit_transform(tableau_copie[colonne])
```

Figure III-21. Traitement des colonnes catégorielles

La figure III-22 illustre le résultat de l'encodage :

User id	Club	Genre	Age	Zone géographique	Occupation	Sport	Film	Interactions en ligne	Revenu	Permis de conduire	Motorisé
6234f7b96d818c001e35edc9	10	1	1	9	2	9	6	23	8	1	1
62057143cbad6e0020e11f6e	12	0	0	4	7	41	3	23	6	0	0
6218aa29a2acb3001fb2cab5	12	1	2	16	4	41	3	33	4	2	2
620d56c47db2fa002a6ab6ae	12	1	0	16	7	0	1	17	3	2	0
627e8554e0f7fe002b52e5af	18	2	3	9	3	32	3	77	4	1	1
...	...	...	...	...	...	...	...	...	...	...	...
6231bf3a6d818c001e346bf2	5	1	1	16	7	41	1	0	5	1	0
62c767dc3317a600295fd1fd	18	1	0	16	3	41	1	20	3	2	0
6206015acbad6e0020e15750	3	1	1	10	6	41	1	24	2	1	1
6208fda0f75cbe001f0a800a	18	1	0	10	7	39	3	25	1	2	0
6208e04169ce2d0020d04ee6	12	1	0	10	4	9	3	11	8	2	1

Figure III-22. Trame de données encodée

### III. 5.2. Paramètres d'entrée de l'algorithme de Mean--Shift

Les paramètres d'entrée de l'algorithme de Mean-Shift sont :

- La bande passante (bandwidth) qui représente un paramètre clé, contrôlant la distance maximale autorisée entre un point et ses voisins pour être regroupé dans le même cluster.

Dans notre cas, nous avons choisi une valeur de 20 pour avoir une taille de clusters relativement grande.

- L'option bin-seeding (amorçage par regroupement) est une technique utilisée pour initialiser les centres de cluster. Dans notre cas, nous avons bin\_seeding a été activée pour améliorer l'efficacité de l'algorithme en accélérant l'initialisation des centres de cluster, contribuant ainsi à une exécution plus rapide de l'algorithme.

### III. 5.3. Application du Mean-Shift

Après avoir défini les paramètres d'entrée, nous avons entraîné le modèle de Mean-Shift sur les données préalablement préparées. Le modèle a été ajusté aux données pour identifier les clusters.

La figure III-23 présente l'estimation de bandwidth et l'application du Mean-Shift :

```
#Estimation de bandwidth
est_bandwidth = 20
#Application de l'algorithme
ms = MeanShift(bandwidth=est_bandwidth, bin_seeding=True).fit(tableau_copie)
```

*Figure III-23. Application du Mean-Shift*

La figure III-24 présente le nombre de clusters obtenu suite à l'application de l'algorithme :

```
nombre_clusters = len(np.unique(labels))
print("Nombre de clusters obtenus :", nombre_clusters)

Nombre de clusters obtenus : 15
```

*Figure III-24. Nombre de clusters obtenus*

Nous avons ensuite extrait des informations sur chaque cluster, telles que les valeurs les plus courantes pour chaque colonne. Le nombre d'utilisateurs dans chaque cluster a également été enregistré.

La figure III-25 présente un exemple illustrant les caractéristiques du cluster 0 :

Cluster Label: 0	
Nombre d'utilisateurs: 1197	
Caractéristiques communes:	
Club	ES, TUNIS
Genre	homme
Age	entre 18 24 an
Zone géographique	tunis
Occupation	lycée
Sport	football
Film	action
Interactions en ligne	facebook
Revenu	plus der 5000 dtn mois
Permis de conduire	false
Motorisé	false
Number of Users	1197.0

Figure III-25. Les caractéristiques d'un cluster

Nous avons également créé une trame de données pour mieux visualiser le résultat contenant pour chaque numéro de cluster les caractéristiques communes ainsi que le nombre d'utilisateurs.

La figure III-26 présente 5 lignes de cette trame de données :

cluster	Club	Genre	Age	Zone géographique	Occupation	Sport	Film	Interactions en ligne	Revenu	Permis de conduire	Motorisé	Number of Users
0	ES, SAHEL	homme	entre 25 34 an	nabeul	entrepreneur	football	science fiction	facebook instagram tik tok	plus der 5000 dtn mois	true	true	1197
1	ES, TUNIS	femme	entre 18 24 an	jendouba	salariée	volleyball	horreur	facebook instagram tik tok	entre 650 1200 dtn mois	false	false	698
2	ES, SAHEL	femme	moins 18 an	siliana	entrepreneur	tennis	tragédie	autre	entre 3000 5000 dtn mois	true	false	294
3	ES, TUNIS	homme	moins 18 an	tunis	lycée	handball	horreur	facebook tik tok	entre 3000 5000 dtn mois	false	true	209
4	ES, TUNIS	homme	entre 18 24 an	bizerte	l'université	football	action	instagram	entre 1200 1800 dtn mois	false	false	191

Figure III-26. Trame de données des caractéristiques des clusters

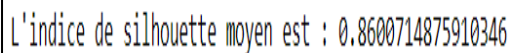
### III. 5.4. Evaluation du modèle

Dans cette section, nous aborderons l'évaluation de notre modèle.

#### III. 5.4.1. Métriques d'évaluation

Nous allons évaluer les mesures de performance en utilisant la méthode d'indice de silhouette qui mesure la cohésion et la séparation des clusters. Cette mesure combine la distance moyenne des points à l'intérieur d'un cluster et la distance moyenne des points entre les clusters adjacents. Suite au calcul, nous aurons une valeur dans l'intervalle de 1 et -1.

En utilisant la classe `silhouette_score` pour effectuer ce calcul nous avons eu ce résultat illustré dans la figure III-27 :



```
L'indice de silhouette moyen est : 0.8600714875910346
```

*Figure III-27: Valeur de l'indice de silhouette moyen*

➔ Cette valeur indique un niveau élevé de cohérence et de séparation entre les clusters.

#### III. 5.4.2. Interprétation des clusters

Nous avons identifié un total de 15 clusters dans notre analyse. Le nombre d'utilisateurs dans chaque cluster varie de 1 à 1197, ce qui indique une répartition inégale des utilisateurs entre les clusters. Certains clusters peuvent contenir un nombre significatif d'utilisateurs, tandis que d'autres ne contiennent qu'un seul utilisateur. Cela suggère que notre application JAYEG présente une diversité dans les profils d'utilisateurs, avec certains clusters regroupant un grand nombre d'individus partageant des caractéristiques similaires, tandis que d'autres sont plus spécifiques et représentent des profils d'utilisateurs uniques.

### **III. 6. Conclusion**

Ce chapitre a joué un rôle crucial dans notre processus d'analyse de données de profilage. Nous avons réussi à extraire et à traiter les données brutes, puis à les regrouper en utilisant l'algorithme de clustering MeanShift. Les résultats obtenus ont fourni des informations précieuses pour résoudre le problème initial. Dans les chapitres suivants, nous explorerons d'autres méthodes d'apprentissage automatique et d'analyse afin d'approfondir notre compréhension des données de profilage et d'obtenir des résultats plus précis et significatifs.

# Chapitre IV

---

## Déploiement

---

## **IV . Déploiement**

---

### **IV. 1. Introduction**

Dans ce chapitre consacré au déploiement de notre projet, nous aborderons les différentes étapes qui nous ont permis de concrétiser notre application. Nous présenterons la modélisation conceptuelle sous forme de diagrammes, mettant en évidence la structure et les fonctionnalités de l'application. De plus, nous détaillerons la mise en œuvre pratique de l'application, en mettant l'accent sur les choix technologiques et les étapes de développement.

### **IV. 2. Modélisation conceptuelle**

Nous commencerons par présenter les diagrammes de modélisation conceptuelle, tels que le diagramme de cas d'utilisation, le diagramme de classes, et les diagrammes de séquence. Ces diagrammes nous permettront de visualiser l'interaction entre les différents acteurs et les fonctionnalités clés de l'application.

## IV. 2.1. Diagramme de cas d'utilisation globale

La figure IV-1 représente le diagramme de cas d'utilisation global :

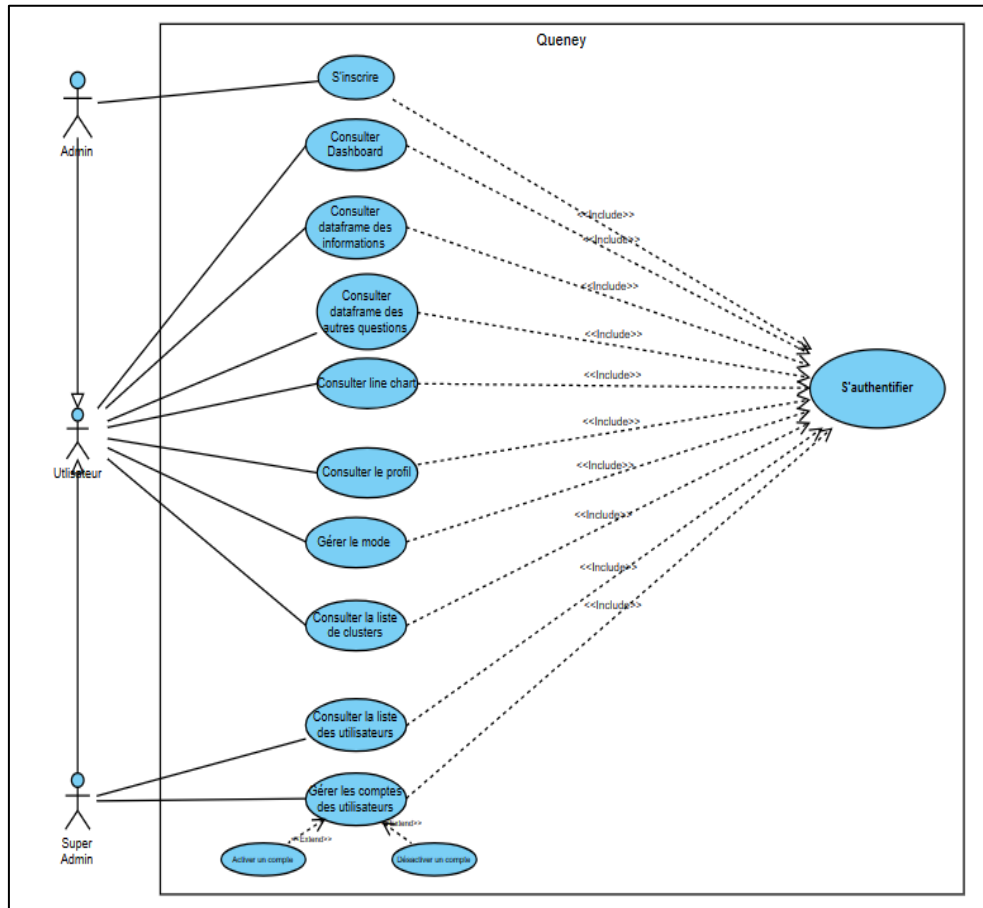


Figure IV-1: Diagramme du cas d'utilisation

### IV. 2.1.1. Description textuelle du cas d'utilisation global

Dans cette partie nous allons décrire textuellement chaque cas d'utilisation.

#### IV. 2.1.1.1. Description textuelle du diagramme de cas d'utilisation « Consulter le tableau de bord »

Le tableau IV-1 suivant représente la description textuelle du cas d'utilisation « Consulter le tableau de bord »



Titre	Consulter le tableau de bord
Acteur principal	Administrateur
Résumé	À travers ce cas, l'utilisateur peut consulter le tableau de bord.
Précondition	L'administrateur doit s'authentifier.
Scénario principal	L'administrateur accède à la page d'accueil où se trouve le tableau de bord.
Post condition	Le tableau de bord sera affiché.

Tableau IV-1. Description textuelle du cas d'utilisation « Consulter le tableau de bord »

#### IV. 2.1.1.2. Description textuelle du diagramme de cas d'utilisation « Télécharger une trame de données »

Le tableau IV-2 représente la description textuelle du cas d'utilisation « Télécharger une trame de données »

Titre	Télécharger une trame de données
Acteur principal	Administrateur
Résumé	À travers ce cas, l'administrateur peut télécharger une trame de données sous format CSV.
Précondition	L'administrateur doit s'authentifier.
Scénario principal	<ol style="list-style-type: none"> <li>1. L'administrateur accède à la page qui affiche la table qu'il souhaite télécharger.</li> <li>2. L'utilisateur clique sur le bouton « Télécharger ».</li> </ol>
Post condition	La trame de donnée choisie sera téléchargée.

Tableau IV-2. Description textuelle du cas d'utilisation « Télécharger une trame de données »

#### IV. 2.1.1.3. Description textuelle du diagramme de cas d'utilisation « Gérer les administrateurs »

Le tableau IV-3 représente la description textuelle du cas d'utilisation « Gérer les administrateurs »

Titre		Gérer les utilisateurs
Acteur principal		Super Admin
Résumé		À travers ce cas, le super admin peut afficher la liste des utilisateurs, activer, désactiver ou supprimer un compte désactivé.
Précondition		Le super admin doit s'authentifier.
Affichage de la liste des administrateurs	Scénario principal	<ol style="list-style-type: none"> <li>1. Accéder à la page « Administration »</li> <li>2. Le système affiche la liste des administrateurs de la base de données.</li> </ol>
	Post condition	La liste sera automatiquement affichée.
Activer un compte	Précondition	Affichage de la liste des utilisateurs avec des comptes désactivés.
	Scénario principal	<ol style="list-style-type: none"> <li>1. Cliquer sur le bouton « Activer » dans la ligne de l'administrateur à activer</li> <li>2. Le système modifie le statut de l'administrateur choisi dans la base de données</li> </ol>
	Post-condition	L'administrateur possède le statut « Activé » et peut accéder à l'application.
Désactiver un compte	Précondition	Affichage de la liste des utilisateurs avec des comptes activés.
	Scénario principal	<ol style="list-style-type: none"> <li>1. Cliquer sur le bouton « Désactiver » dans la ligne de l'administrateur à activer</li> <li>2. Le système modifie le statut de l'administrateur choisi dans la base de données.</li> </ol>
	Post-condition	L'administrateur possède le statut « Désactivé » et ne peut plus accéder à l'application.

Tableau IV-3. Description textuelle du cas d'utilisation « Gérer les administrateurs »

#### IV. 2.1.1.4. Description textuelle du diagramme de cas d'utilisation « S'inscrire »

Le tableau IV-4 représente la description textuelle du cas d'utilisation « S'inscrire »

Titre	S'inscrire
<b>Acteur principal</b>	Administrateur
<b>Résumé</b>	A travers ce cas, l'administrateur peut créer un compte.
<b>Précondition</b>	L'administrateur fait partie de Queney.
<b>Scénario principal</b>	<ol style="list-style-type: none"> <li>1. L'admin choisit l'action « Sign up ».</li> <li>2. Le système affiche un formulaire.</li> <li>3. L'admin remplit le formulaire.</li> <li>4. Le système vérifie les données saisies.</li> <li>5. Le système enregistre les données dans la base.</li> </ol>
<b>Post condition</b>	Le compte est ajouté à la base de données avec le statut « Désactivé »
<b>Scénario alternatif</b>	<p>A1 : Les champs sont vides</p> <p>L'enchaînement A1 démarre après le point 4</p> <p>5. Le système affiche un message d'erreur pour informer l'admin qu'il y a des champs obligatoires vides.</p> <p>Le scénario nominal reprend au point 2.</p> <p>A2 : L'adresse e-mail ne représente pas une adresse des employés de Naxxum.</p> <p>5. Le système affiche un message d'erreur pour informer l'admin que son e-mail ne représente pas une adresse e-mail professionnelle de Naxxum.</p> <p>Le scénario nominal reprend au point 2.</p> <p>A3 : L'adresse e-mail est déjà utilisée par un autre compte</p> <p>L'enchaînement A3 démarre après le point 4</p> <p>5. Le système affiche un message d'erreur pour informer l'admin que l'adresse e-mail est utilisée par un autre compte</p> <p>Le scénario nominal reprend au point 2.</p> <p>A4 : Le mot de passe est faible.</p>

	<p>L'enchaînement A4 démarre après le point 4</p> <p>5. Le système affiche un message d'erreur pour informer l'admin que le mot de passe saisi est faible.</p> <p>Le scénario nominal reprend au point 2.</p>
--	---

Tableau IV-4. Description textuelle du cas d'utilisation « S'inscrire »

#### IV. 2.1.1.5. Description textuelle du diagramme de cas d'utilisation « S'authentifier »

Le tableau IV-5 représente la description textuelle du cas d'utilisation « S'authentifier »

Titre	S'authentifier
<b>Acteur principal</b>	Administrateur
<b>Résumé</b>	A travers ce cas, l'administrateur peut accéder à l'application.
<b>Précondition</b>	L'administrateur possède un compte
<b>Scénario principal</b>	<p>L'admin choisit l'action « Sign in ».</p> <ol style="list-style-type: none"> <li>1. Le système affiche un formulaire d'authentification.</li> <li>2. L'admin remplit le formulaire.</li> <li>3. Le système vérifie les données saisies.</li> <li>4. Le système dirige l'admin vers la page d'accueil.</li> </ol>
<b>Post condition</b>	L'admin est dirigé vers la page d'accueil.
<b>Scénario alternatif</b>	<p>A1 : Les données saisies n'existent pas dans la base de données : mot de passe ou adresse e-mail peut être saisi incorrectement.</p> <p>L'enchaînement A1 démarre après le point 3</p> <p>4. Le système affiche un message d'erreur pour informer l'admin que ce compte n'existe pas ou les données fournies sont incorrectes.</p> <p>Le scénario nominal reprend au point 1.</p> <p>E1 : Le compte est « Désactivé »</p> <p>L'enchaînement A2 démarre après le point 3</p> <p>4. Le système affiche un message d'erreur pour informer l'admin que son compte est « Désactivé »</p>
<b>Scénario d'exception</b>	E1 : Le compte est « Désactivé »

	<p>L'enchaînement A2 démarre après le point 3</p> <p>4. Le système affiche un message d'erreur pour informer l'admin que son compte est « Désactivé »</p>
--	---

Tableau IV-5. Description textuelle du cas d'utilisation « S'authentifier »

## IV. 2.2. Diagrammes de séquences du cas d'utilisation

Nous allons maintenant passer à la présentation des diagrammes de séquence, qui nous permettront de représenter de manière visuelle les interactions et les séquences d'événements entre les différents éléments de notre système.

### IV. 2.2.1. Diagramme de séquences du cas d'utilisation « S'inscrire »

La figure IV-2 présente Diagramme de séquences du cas d'utilisation « S'inscrire » :

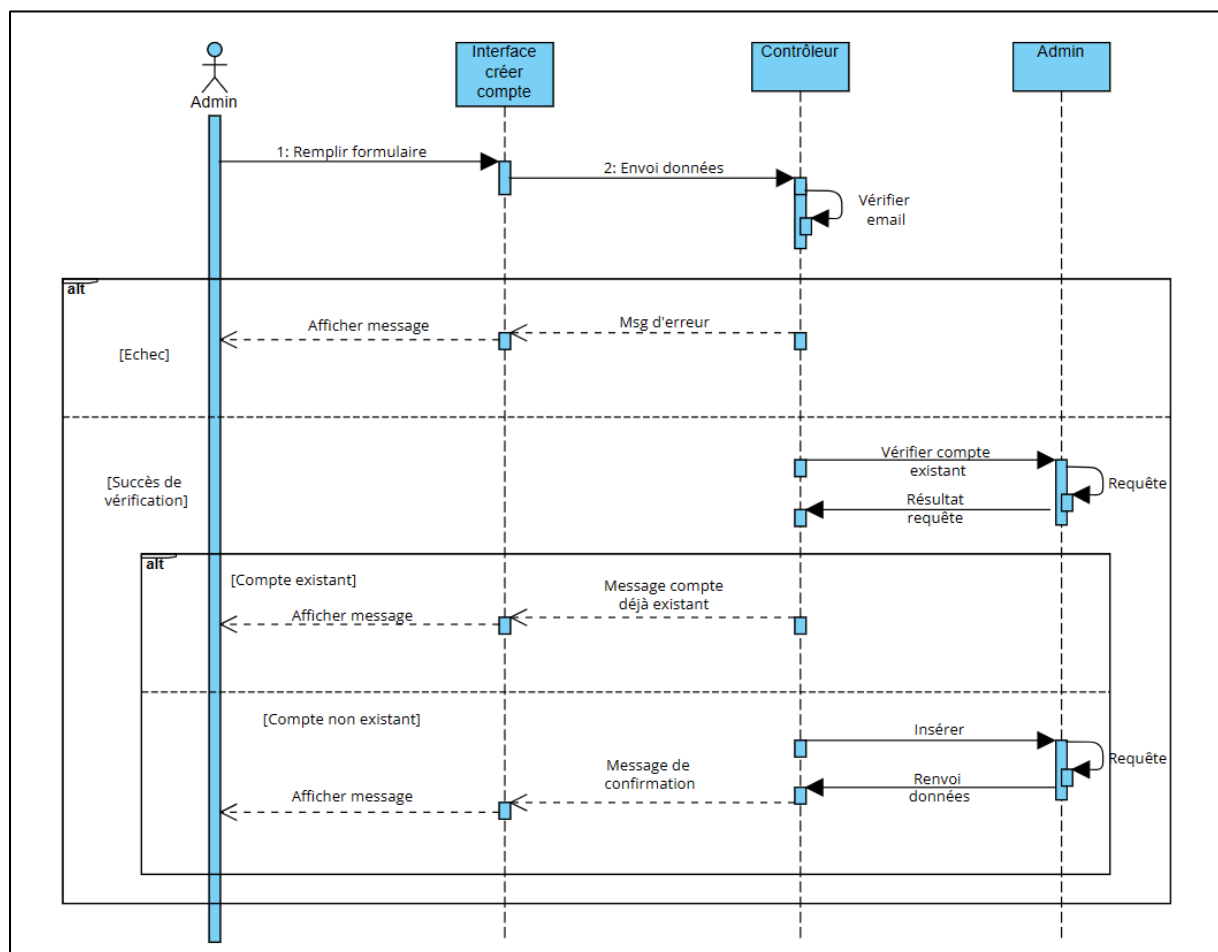


Figure IV-2. Diagramme de séquences du cas d'utilisation « S'inscrire »

#### IV. 2.2.2. Diagramme de séquences du cas d'utilisation « S'authentifier »

»

La figure IV-3 présente Diagramme de séquences du cas d'utilisation « S'authentifier » :

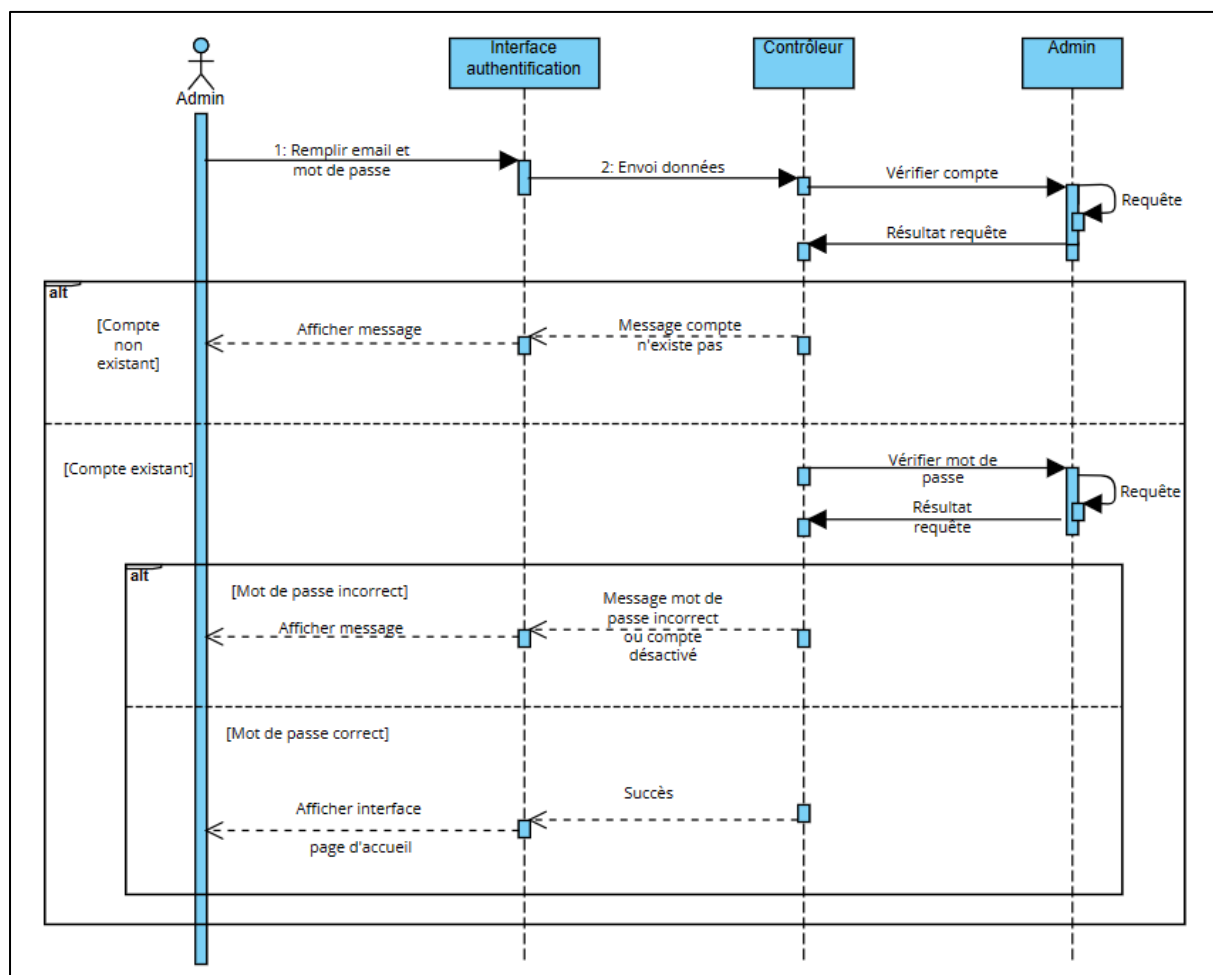


Figure IV-3. Diagramme de séquence du cas d'utilisation « S'authentifier »

#### IV. 2.2.3. Diagramme de séquences du cas d'utilisation « Activer un compte »

La figure IV-4 présente Diagramme de séquences du cas d'utilisation « Activer un compte »

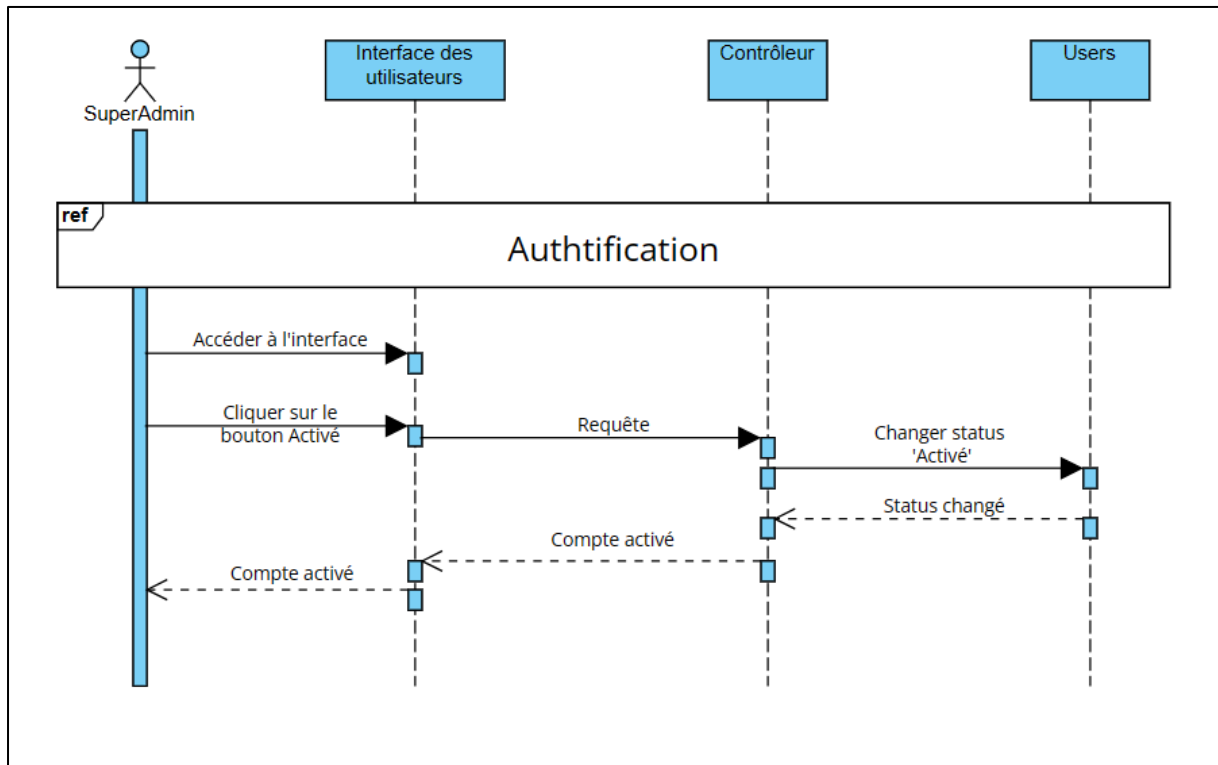


Figure IV-4. Diagramme de séquences du cas d'utilisation « Activer un compte »

#### IV. 2.2.4. Diagramme de séquences du cas d'utilisation « Désactiver un compte »

La figure IV-5 présente Diagramme de séquences du cas d'utilisation « Désactiver un compte »

»

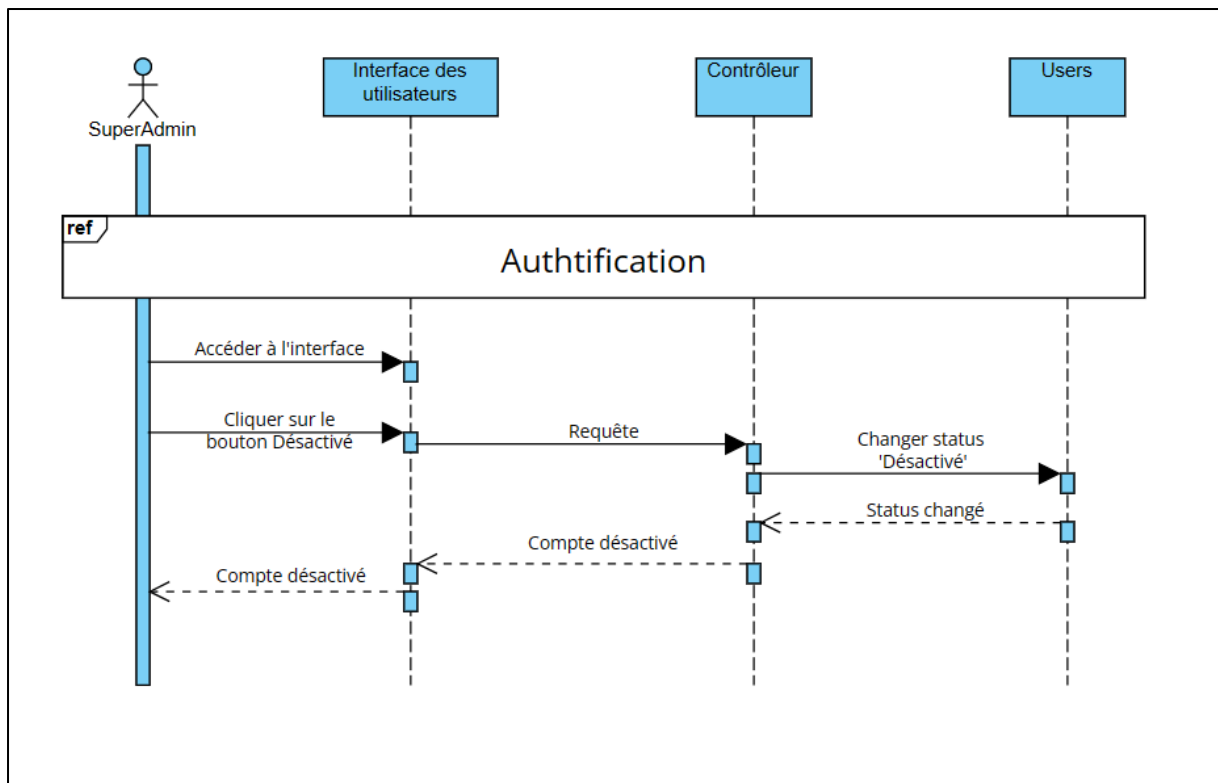


Figure IV-5. Diagramme de séquences du cas d'utilisation « Désactiver un compte »

### IV. 2.3. Diagramme de classes

Nous allons maintenant présenter le diagramme de classe qui offre une représentation structurée des entités principales de notre système, ainsi que de leurs attributs et relations.

#### IV. 2.3.1. Dictionnaire de données

Numéro	Attribut	Libellé	Type
1	Id_utilisateur	Identifiant d'un utilisateur de l'application.	String
2	Email	L'email d'un utilisateur.	String
3	Last_name	Le nom d'un utilisateur.	String
4	First_name	Le prénom d'un utilisateur.	String
5	Phone_number	Le numéro de téléphone d'un utilisateur.	String
6	Status	Le statut du compte d'un utilisateur (activé/désactivé).	String
7	Password	Le mot de passe d'un utilisateur	String



8	Rôle	La désignation du rôle des utilisateurs (Admin/ Super Admin).	String
---	------	---	--------

Tableau IV-6. Dictionnaire de données

#### IV. 2.3.2. Diagramme de classe

Notre diagramme de classe représenté par la figure IV-6 est composé de trois entités principales : Utilisateur, Super Admin et Admin.

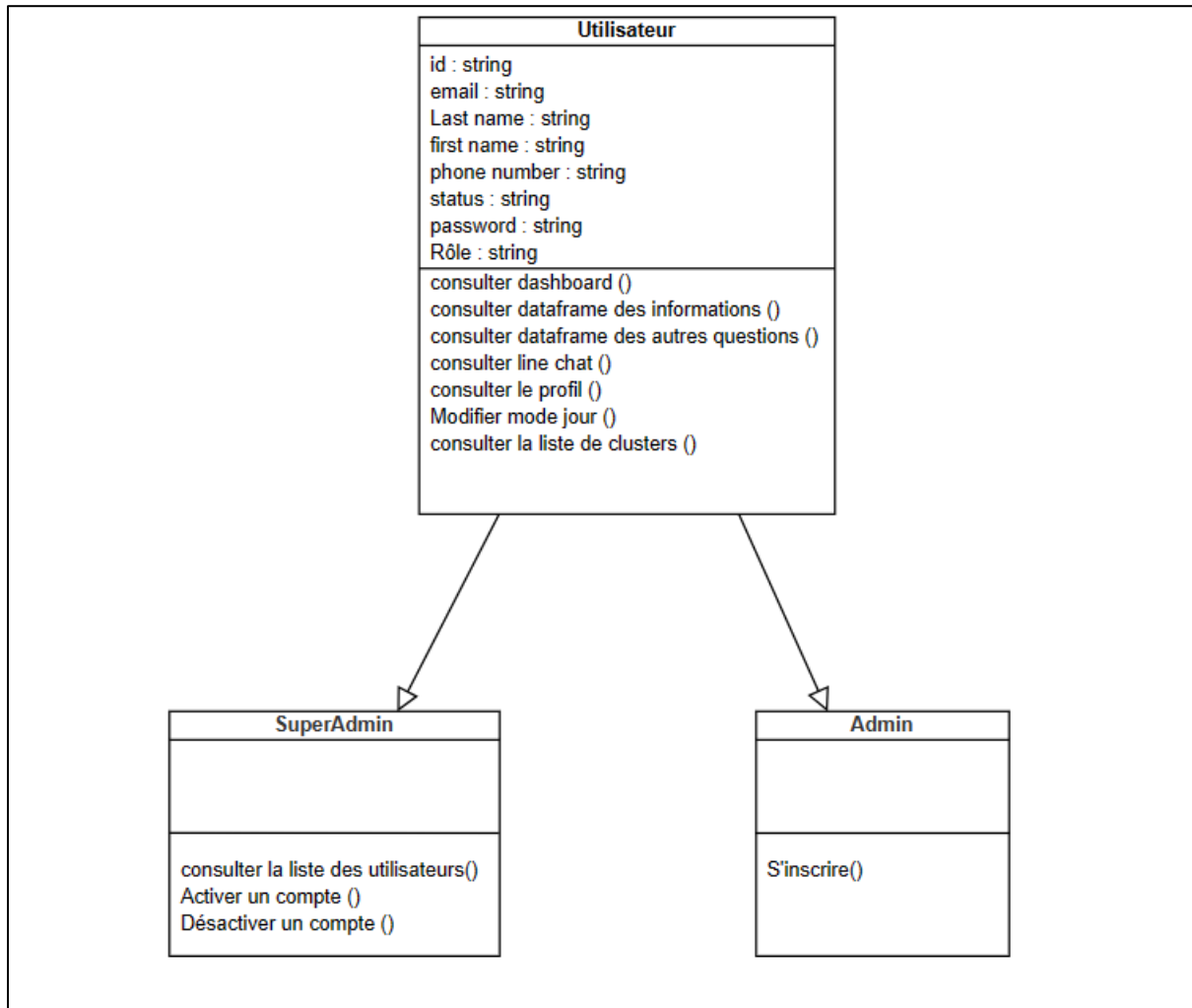


Figure IV-6. Diagramme de classe

#### IV. 3. Réalisation :

Dans cette section nous allons exposer les diverses interfaces de notre application.

### IV. 3.1. TopBar



Figure IV-7: Capture du TopBar

Cette figure IV-6 représente TopBar qui est une barre de navigation située en haut de l'interface, qui offre aux utilisateurs un accès rapide aux éléments suivant :

#### IV. 3.1.1. Rechercher

Un champ interactif permettant aux utilisateurs d'entrer des mots-clés qui sont les éléments de SideBar pour accéder directement à l'interface recherchée.

Dans le cas où le terme recherché n'est pas trouvé, une alerte s'affiche pour en informer l'utilisateur.

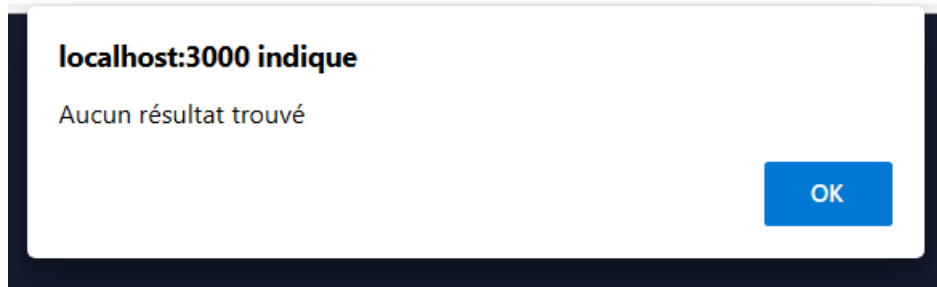


Figure IV-8: Alerte de recherche 'Aucun résultat trouvé'

#### IV. 3.1.2. Mode jour

Un bouton qui permet aux utilisateurs de basculer entre le mode jour et le mode nuit de l'application. En cliquant sur ce bouton, l'interface peut passer d'un thème clair à un thème sombre, offrant ainsi une expérience visuelle adaptée aux préférences de l'utilisateur.



Figure IV-9: Mode nuit

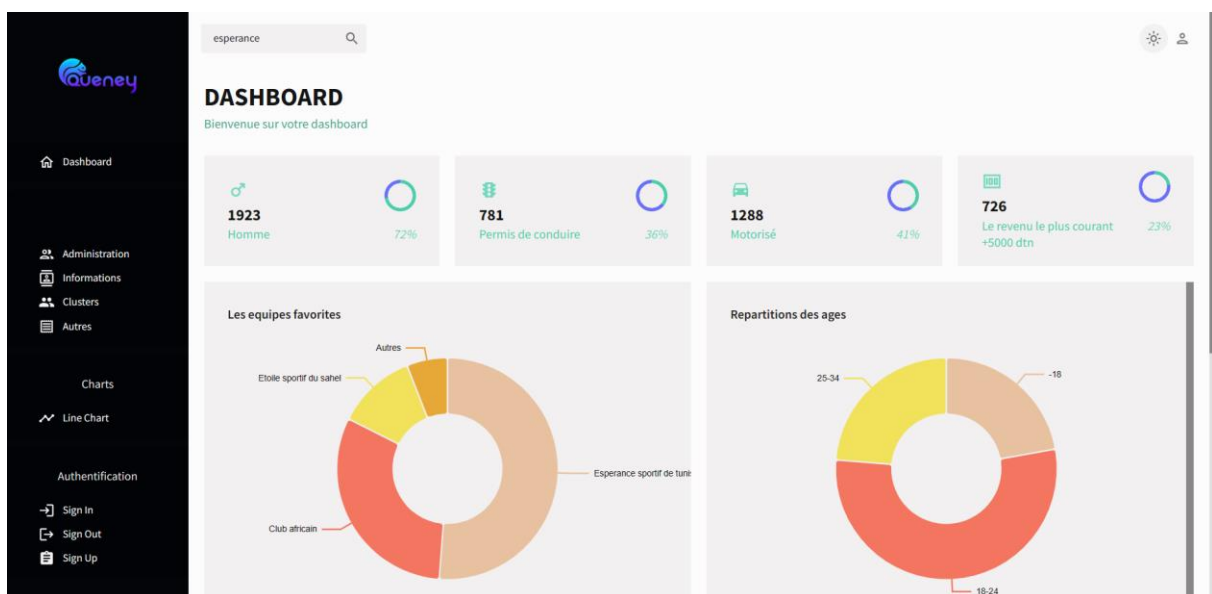


Figure IV-10: Mode jour

### IV. 3.1.3. Profil

L'icône de la figure IV-11 représente un utilisateur, qui permet d'afficher le rôle de l'utilisateur :

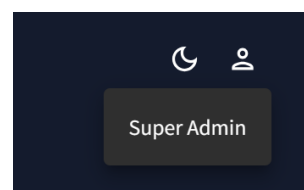
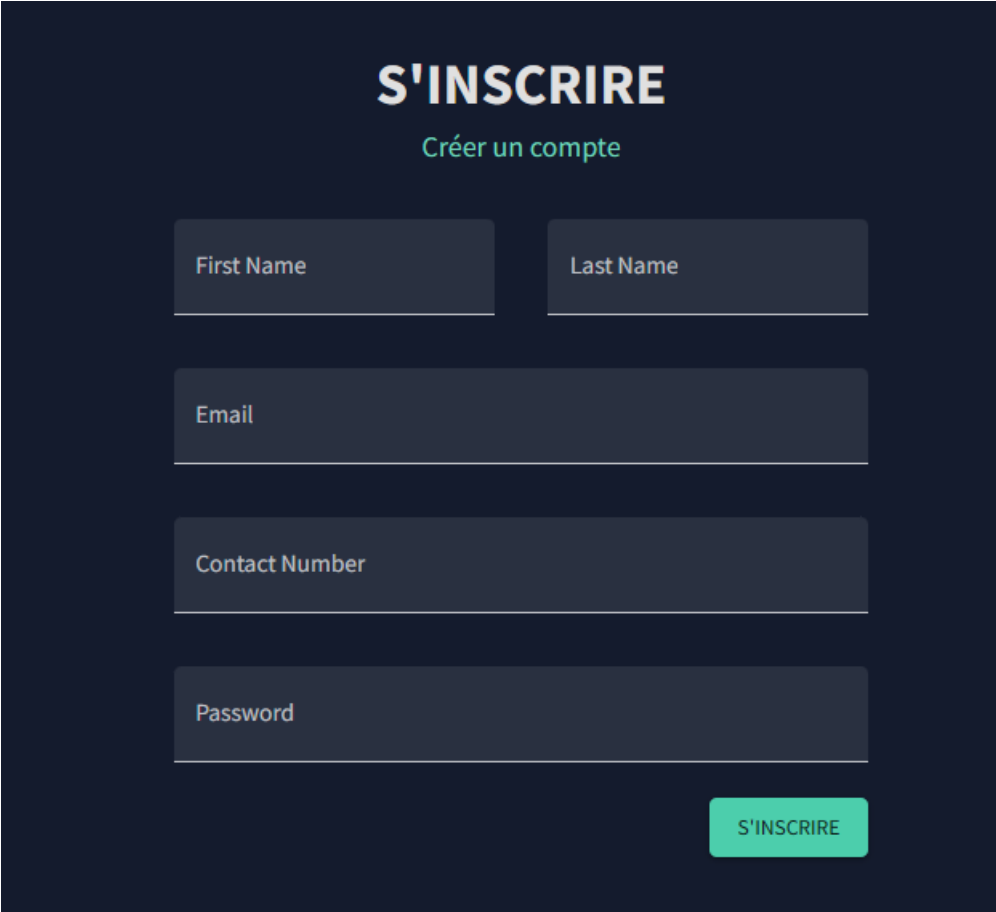


Figure IV-11: Profil

### IV. 3.2. Interface d'inscription

Cette figure IV-12 représente l'interface qui permet à l'administrateur de créer un nouveau compte :



The image shows a registration form titled "S'INSCRIRE" (Sign Up) with the subtitle "Créer un compte" (Create an account). The form is set against a dark blue background. It contains five input fields: "First Name", "Last Name", "Email", "Contact Number", and "Password". A green "S'INSCRIRE" button is located at the bottom right of the form.

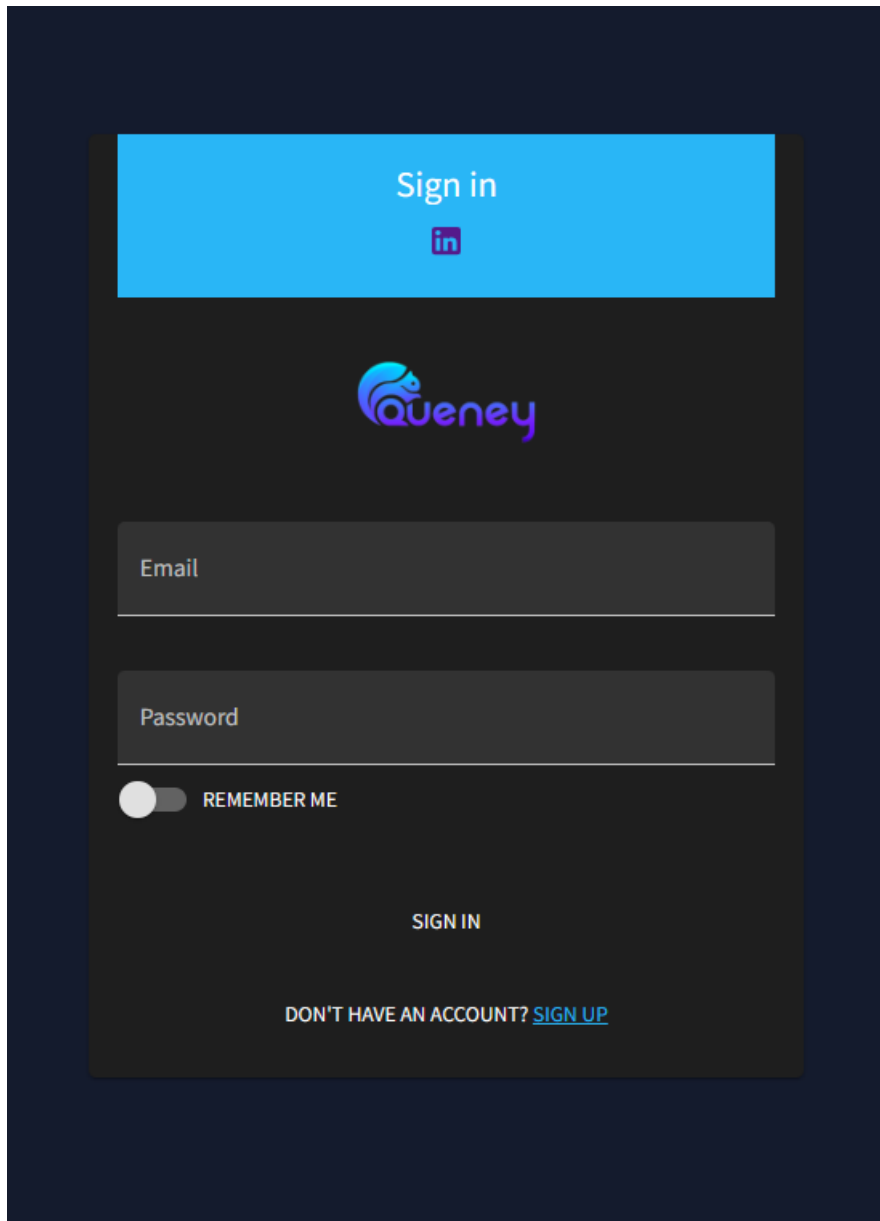
Field	Label
First Name	First Name
Last Name	Last Name
Email	Email
Contact Number	Contact Number
Password	Password

S'INSCRIRE

*Figure IV-12. Interface d'inscription*

### IV. 3.3. Interface d'authentification

L'utilisateur utilise cette interface pour s'authentifier en saisissant son adresse e-mail et son mot de passe, afin d'accéder à l'application :



*Figure IV-13. Interface d'authentification*

#### **IV. 3.4. Interface de Dashboard**

L'interface IV-14 représente la page d'accueil de notre application, où l'on trouve un tableau de bord contenant plusieurs composants :

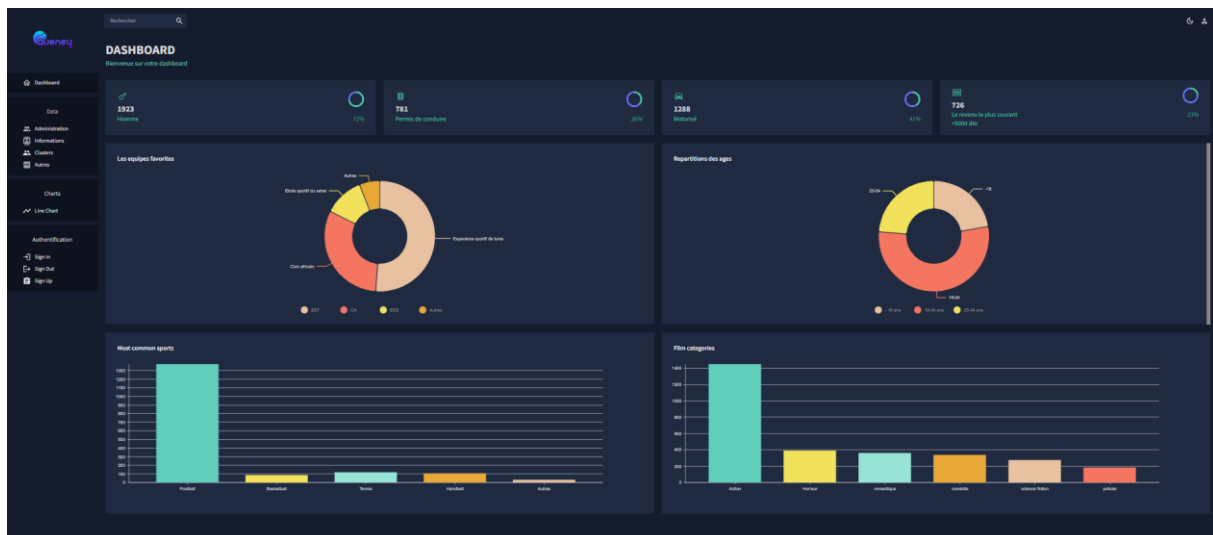


Figure IV-14. Interface de Dashboard

### IV. 3.5. Les interfaces des trames de données

Les interfaces qui seront présentées dans cette section sont regroupées dans un même box dans la barre-latérale de notre application nommée « Data » où nous trouverons les différentes trames de données

#### IV. 3.5.1. Interface de la table « Informations générales »

Cette interface présentée par la figure IV-15 permet de visualiser la trame de données contenant les réponses vérifiées de chaque utilisateur pour chaque critère de profiling et d'afficher un bouton de téléchargement pour obtenir les données au format CSV.

User id	Club	Genre	Age	Zone géograph...	Occupation	Sport	Film	Interactions en ...	Revenu	Permis de cond...	Motorisé
62347fb96d8...	ES, SAHEL	homme	entre 25 34 an	nabeul	entrepreneur	football	science fiction	facebook inst...	plus der 5000...	true	true
62057143cba...	ES, TUNIS	femme	entre 18 24 an	jendouba			horreur	facebook inst...	entre 650 120...	false	false
6218aa29a2a...	ES, TUNIS	homme	moins 18 an	tunis	lycée		horreur	facebook tik t...	entre 3000 50...		
620d56c47db...	ES, TUNIS	homme	entre 18 24 an	tunis		aucun	action	facebook inst...	entre 1800 30...		false
627e8554e0f...				nabeul	université	handball	horreur	tik tok instag...	entre 3000 50...	true	true
622504dba92...	US, MONASTI...	homme	entre 25 34 an	sousse	salarée		comédie	facebook inst...	entre 650 120...	true	true
623cb801c64...	CLUB_AFRICA...	homme	entre 18 24 an	nabeul	lycée	football	action	facebook inst...	inferieur 400 ...		false
621b963ba2a...		homme		tunis			comédie	tik tok	entre 3000 50...		false

Figure IV-15. Interface de la table "Informations générales"

#### IV. 3.5.2. Interface de la table des autres questions

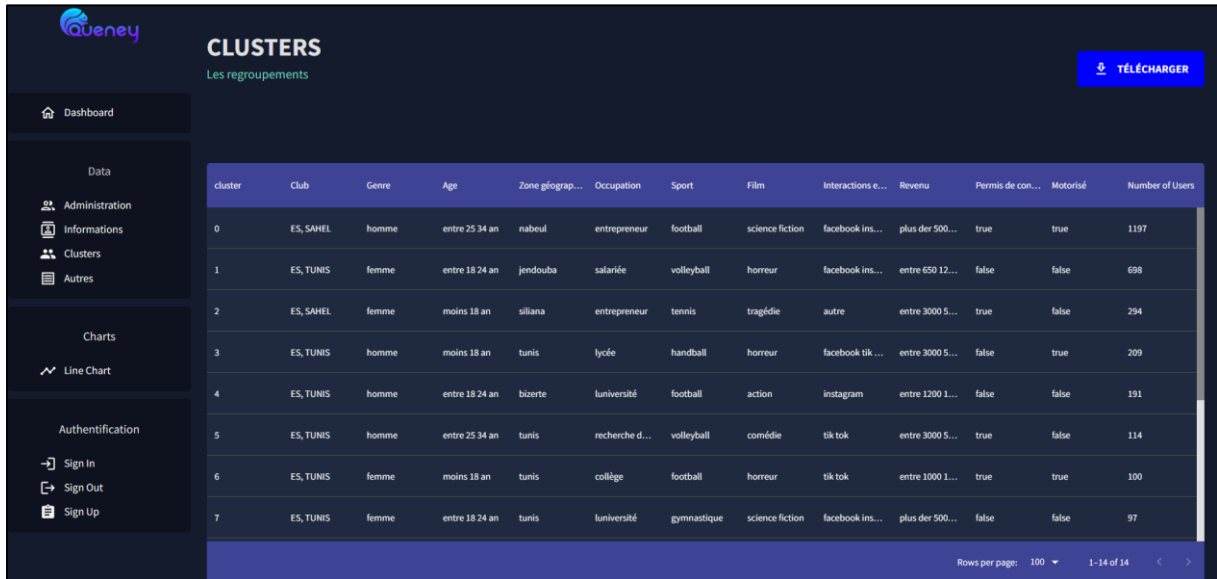
Cette interface permet de visualiser la trame de données contenant les réponses vérifiées de chaque utilisateur à d'autres questions posées et d'afficher un bouton de téléchargement pour obtenir les données au format CSV.

User id	Que faites ...	Vous aime...	vous préf...	Que fais t...	En général...	Plus globa...	T'intéress...	Vois-tu un ...	D'après toi...	Vous disp...	vous êtes	combien d...	?l'habite?	Pour me d...	As-tu accè...
62347fb...	sortie e...	oui	nike lac...	internet ...	sport lo...	projet p...	true	true	espaces ...	machin...	propriéta...	2	apparte...	voiture	true
6205714...	activités...	oui	chanel s...	activités...	etudesf...	projet d...	false	false	évènem...	congèlat...	locataire	1	chez, pa...		true
6218aa2...	activités...	non	nike lac...	internet	etudesf...	projet p...	false	true	évènem...	machin...	locataire	1		metro	
620d56c...	internet ...	oui	nike lac...	activités...	métiers ...	voyage ...	false	false	transpor...	machin...	locataire	1			true
627e855...	activités...	oui	nike lac...	activités...	vie prati...	voyage ...	true	true	transport	machin...	locataire	1		metro	
622504d...	internet	non	nike lac...	internet	métiers	projet p...	false	false	autre	machin...	locataire	1			true
623cb80...	internet	oui	nike lac...	internet ...	etudesf...	projet d...	false	false	espaces ...	climatis...	locataire	1	chez, pa...	taxi	
621b963...	internet	non	nike lac...	internet ...	offres d...	voyage	true	false	évènem...	congèlat...	locataire	1			false

Figure IV-16. Interface de la table des autres questions posées

#### IV. 3.5.3. Interface du tableau des clusters

La figure IV-17 présente le résultat de la trame de données générée suite à l'application de l'algorithme de Mean-Shift. Chaque cluster est accompagné de ses caractéristiques spécifiques et du nombre d'utilisateurs correspondants. De plus, un bouton de téléchargement est disponible pour obtenir les données au format CSV.



cluster	Club	Genre	Age	Zone géograp...	Occupation	Sport	Film	Interactions e...	Revenu	Permis de con...	Motorisé	Number of Users
0	ES, SAHEL	homme	entre 25 34 an	nabeul	entrepreneur	football	science fiction	facebook ins...	plus der 500...	true	true	1197
1	ES, TUNIS	femme	entre 18 24 an	jendouba	salarée	volleyball	horreur	facebook ins...	entre 650 12...	false	false	698
2	ES, SAHEL	femme	moins 18 an	siliana	entrepreneur	tennis	tragédie	autre	entre 3000 5...	true	false	294
3	ES, TUNIS	homme	moins 18 an	tunis	lycée	handball	horreur	facebook tik ...	entre 3000 5...	false	true	209
4	ES, TUNIS	homme	entre 18 24 an	bizerte	université	football	action	instagram	entre 1200 1...	false	false	191
5	ES, TUNIS	homme	entre 25 34 an	tunis	recherche d...	volleyball	comédie	tik tok	entre 3000 5...	true	false	114
6	ES, TUNIS	femme	moins 18 an	tunis	collège	football	horreur	tik tok	entre 1000 1...	true	true	100
7	ES, TUNIS	femme	entre 18 24 an	tunis	université	gymnastique	science fiction	facebook ins...	plus der 500...	false	false	97

Figure IV-17. Interface des Clusters

#### IV. 3.5.4. Interface de la liste des utilisateurs

La figure IV-18 illustre l'interface qui affiche la liste des administrateurs inscrits dans l'application. Cette liste permet au Super Admin d'accéder aux informations des administrateurs, de les activer ou de les désactiver, de les supprimer, ainsi que de télécharger la liste au format CSV.



email	first_name	last_name	Status
mberima@naxum.fr	Muhamad	Berlima	ACTIVE
mbenacof@naxum.fr	Marlem	Ben Nacof	ACTIVE
lsould@naxum.fr	Imen	Souldi	DESACTIVE

Figure IV-18. Interface de la liste des utilisateurs

### IV. 3.6. Interface de Chart Line

Cette interface IV-19 est présentée dans la section "Data" de la barre latérale de notre application, où l'on trouve un graphique en ligne des salaires en fonction de l'occupation des utilisateurs.

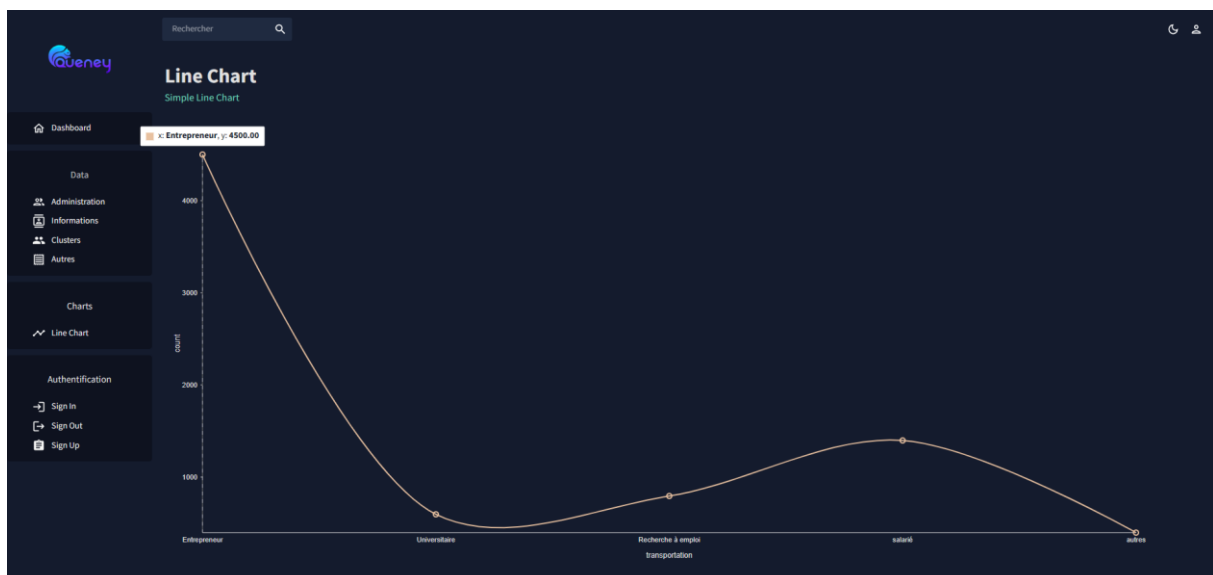


Figure IV-19. Interface de Line Chart

#### **IV. 4. Conclusion**

Ce chapitre consacré au déploiement de notre application a été crucial pour concrétiser notre projet. Nous avons pu présenter les diagrammes de modélisation conceptuelle qui ont permis de définir la structure et les fonctionnalités clés de l'application. Ces diagrammes ont été des outils précieux pour visualiser l'interaction entre les acteurs et les différents cas d'utilisation.

De plus, nous avons détaillé la réalisation de l'application en mettant en avant les choix technologiques et les étapes de développement. Nous avons pu concrétiser les fonctionnalités principales de l'application et mettre en place une interface utilisateur conviviale et intuitive.

## Conclusion générale et perspectives

---

Ce rapport présente les résultats du travail accompli au sein de la startup Queney dans le cadre de notre projet de fin d'études de la licence fondamentale en Business Intelligence à l'IHEC Carthage. Pendant le stage de quatre mois, nous avons contribué à la conception et au développement d'une application web dédiée au profilage des clients. Ce rapport de stage résume les différentes étapes et activités entreprises tout au long de cette expérience enrichissante.

L'objectif principal de ce stage était de mettre en place un système de profilage automatique des clients de Queney. Cette solution permet d'identifier les caractéristiques des profils des clients et offre aux administrateurs de la startup la possibilité de visualiser le tableau de bord, les composants graphiques et les tables directement sur le portail développé. Le projet a débuté par des réunions pour définir les objectifs et les besoins fonctionnels et non fonctionnels de notre système. Nous avons ensuite divisé le projet en étapes en choisissant la méthodologie CRISP-DM, la plus adaptée pour sa réalisation.

À la fin de ce stage, nous avons réussi à automatiser le modèle de profilage des clients en utilisant les données collectées par les campagnes. L'étape de prétraitement des données a joué un rôle important dans les résultats obtenus, et l'implémentation d'un dictionnaire a amélioré les performances du modèle. Nous avons également obtenu des informations cohérentes et vérifiées. Ainsi, nous avons créé un tableau de bord présentant les différents axes d'analyse pour répondre aux besoins.

Ce projet nous a permis de mettre en évidence nos connaissances acquises lors de notre cursus et nous a offert l'occasion de développer de nouvelles compétences. Dans l'ensemble, ce stage a été très enrichissant, nous permettant d'améliorer nos compétences techniques et humaines. Cette expérience exceptionnelle nous a également familiarisé avec le milieu professionnel.

Notre processus n'est pas simple, et nous devons faire face à certaines difficultés telles que la compréhension du sujet, la compréhension des données et l'étude des technologies. Malgré ces obstacles, nous avons réussi à mettre en œuvre l'application.

Cependant, notre travail ne s'arrête pas là, et plusieurs perspectives d'amélioration peuvent être envisagées. Nous reconnaissons que certaines données complexes peuvent encore poser des défis en termes de précision du profilage. Ainsi, nous croyons fermement qu'en élargissant

progressivement le dictionnaire utilisé, nous pourrions améliorer de manière significative la précision de notre application. De plus, l'exploration d'autres méthodes et techniques avancées d'apprentissage automatique pourrait être bénéfique pour une meilleure compréhension et interprétation des données clients.

En poursuivant nos efforts pour affiner et enrichir notre approche, nous nous engageons à fournir des résultats plus précis et pertinents, répondant ainsi aux besoins de notre startup et de ses utilisateurs. Une perspective supplémentaire consisterait à intégrer des filtres permettant de spécifier les critères selon lesquels le clustering doit opérer, offrant ainsi une personnalisation plus fine et des résultats encore plus adaptés aux attentes des utilisateurs. En résumé, notre objectif continuera d'être l'amélioration constante de notre application de profilage des clients, en explorant de nouvelles méthodes, en élargissant nos ressources et en prenant en compte les commentaires des utilisateurs pour une expérience toujours plus satisfaisante.

Finalement, nous espérons avoir réalisé un travail sérieux et convenable tout en ayant laissant une bonne impression.

---

# Bibliographie

---

[1] Naxxum

<https://naxxum.com/>

[2] Queney

[https://www.google.com/search?q=queney&tbm=isch&ved=2ahUKEwi\\_zcv6pt7\\_AhVPpCcCHVP5CDUQ2-cCegQIABAA&oq=queney&gs\\_lcp=CgNpbWcQAzIECCMQJzIECCMQJzIGCAAQBRAeMgYIABAFEB4yBggAEAUQHjIHCAAQGBCABDIJCAAQGBCABBAMgcIABAYEIAEMgkIABAYEIAEEAo6BwgAEBMQgAQ6CAgAEAUQHhAToggIABCABBCxAzoFCAAQgARQ1QRYpwlg4wpoAHAAeACAACyBiAHQB5IBAzAuN5gBAKABAaoBC2d3cy13aXotaW1nwAEB&sclient=img&ei=NiOYZL\\_aMM\\_InsEP0\\_KjqAM&bih=688&biw=1536&rlz=1C1GCEU\\_frTN1045TN1045#imgrc=-3lq4mNnkHBjjM](https://www.google.com/search?q=queney&tbm=isch&ved=2ahUKEwi_zcv6pt7_AhVPpCcCHVP5CDUQ2-cCegQIABAA&oq=queney&gs_lcp=CgNpbWcQAzIECCMQJzIECCMQJzIGCAAQBRAeMgYIABAFEB4yBggAEAUQHjIHCAAQGBCABDIJCAAQGBCABBAMgcIABAYEIAEMgkIABAYEIAEEAo6BwgAEBMQgAQ6CAgAEAUQHhAToggIABCABBCxAzoFCAAQgARQ1QRYpwlg4wpoAHAAeACAACyBiAHQB5IBAzAuN5gBAKABAaoBC2d3cy13aXotaW1nwAEB&sclient=img&ei=NiOYZL_aMM_InsEP0_KjqAM&bih=688&biw=1536&rlz=1C1GCEU_frTN1045TN1045#imgrc=-3lq4mNnkHBjjM)

[3] JAYEG

[https://www.google.com/search?rlz=1C1GCEU\\_frTN1045TN1045&sxsrf=APwXEdfd50dR9nDVqFftlIOz2PkHOBdj-w:1687692084952&q=jayeg&tbm=isch&sa=X&ved=2ahUKEwihtdr5pt7\\_AhWNTqQEHQ1gAliQ0pQJegQICxAB&biw=1536&bih=688&dpr=1.25#imgrc=3OYjNNuaHq075M](https://www.google.com/search?rlz=1C1GCEU_frTN1045TN1045&sxsrf=APwXEdfd50dR9nDVqFftlIOz2PkHOBdj-w:1687692084952&q=jayeg&tbm=isch&sa=X&ved=2ahUKEwihtdr5pt7_AhWNTqQEHQ1gAliQ0pQJegQICxAB&biw=1536&bih=688&dpr=1.25#imgrc=3OYjNNuaHq075M)

[4] KDD

[https://en.wikipedia.org/wiki/Data\\_mining](https://en.wikipedia.org/wiki/Data_mining)

[5] Architecture de la méthodologie de KDD

[https://www.google.com/search?rlz=1C1GCEU\\_frTN1045TN1045&hl=fr&sxsrf=APwXEdfnjyAZPDP\\_D22DotZuBxX5SUYM0w:1687692187666&q=kdd&tbm=isch&sa=X&ved=2ahUKEWjz3Neqp97\\_AhWZaqQEHzR2AP0Q0pQJegQIDRAB&biw=767&bih=679&dpr=1.25#imgrc=Oeyjiu\\_FfEkI6M](https://www.google.com/search?rlz=1C1GCEU_frTN1045TN1045&hl=fr&sxsrf=APwXEdfnjyAZPDP_D22DotZuBxX5SUYM0w:1687692187666&q=kdd&tbm=isch&sa=X&ved=2ahUKEWjz3Neqp97_AhWZaqQEHzR2AP0Q0pQJegQIDRAB&biw=767&bih=679&dpr=1.25#imgrc=Oeyjiu_FfEkI6M)

[6] SEMMA

<https://www.datascience-pm.com/semma/>

[7] Architecture de la méthodologie de SEMMA

[https://www.google.com/search?q=Architecture+de+la+m%C3%A9thodologie+de+SEMMA&tbm=isch&ved=2ahUKEwicn9hMv\\_AhWc7LsIHXIFDIQ2-cCegQIABAA&oq=Architecture+de+la+m%C3%A9thodologie+de+SEMMA&gs\\_lcp=CgNpbWcQAzIECCMQJ1D5G1j5G2CAI2gAcAB4AIABjgGIAZcCkgEDMC4ymAEAoAEBqgELZ3dzLXdpei1pbWfAAQE&sclient=img&ei=cwmOZNyDDZzZ7\\_UP8oq4kAU&bih=746&biw=1536&rlz=1C1GCEU\\_frTN1045TN1045#imgsrc=zDRxPUBBBED-iM](https://www.google.com/search?q=Architecture+de+la+m%C3%A9thodologie+de+SEMMA&tbm=isch&ved=2ahUKEwicn9hMv_AhWc7LsIHXIFDIQ2-cCegQIABAA&oq=Architecture+de+la+m%C3%A9thodologie+de+SEMMA&gs_lcp=CgNpbWcQAzIECCMQJ1D5G1j5G2CAI2gAcAB4AIABjgGIAZcCkgEDMC4ymAEAoAEBqgELZ3dzLXdpei1pbWfAAQE&sclient=img&ei=cwmOZNyDDZzZ7_UP8oq4kAU&bih=746&biw=1536&rlz=1C1GCEU_frTN1045TN1045#imgsrc=zDRxPUBBBED-iM)

[8] CRISP-DM

<https://www.datascience-pm.com/crisp-dm-2/>

[9] Architecture de la méthodologie de CRISP-DM

[https://www.google.com/search?q=crispdm+methodology&tbm=isch&ved=2ahUKEwjw7L6cqN7\\_AhXUtEwKHZykBkQQ2-cCegQIABAA&oq=crispdm+methodology&gs\\_lcp=CgNpbWcQAzIHCAAQExCABDoECCMQJzoGCAAQBxAeOggIABAIEAcQHjoFCAAQgARQIQVYvRtg4h1oAnAAeACAAaEBiAHFCpIBBDAuMTCYAQCgAQGqAQtn3Mtd2l6LWltZ8ABAQ&sclient=img&ei=iiSYZPD\\_E9TpsgKcyZqgBA&bih=662&biw=750&rlz=1C1GCEU\\_frTN1045TN1045&hl=fr#imgsrc=aUZik5jA29VJwM](https://www.google.com/search?q=crispdm+methodology&tbm=isch&ved=2ahUKEwjw7L6cqN7_AhXUtEwKHZykBkQQ2-cCegQIABAA&oq=crispdm+methodology&gs_lcp=CgNpbWcQAzIHCAAQExCABDoECCMQJzoGCAAQBxAeOggIABAIEAcQHjoFCAAQgARQIQVYvRtg4h1oAnAAeACAAaEBiAHFCpIBBDAuMTCYAQCgAQGqAQtn3Mtd2l6LWltZ8ABAQ&sclient=img&ei=iiSYZPD_E9TpsgKcyZqgBA&bih=662&biw=750&rlz=1C1GCEU_frTN1045TN1045&hl=fr#imgsrc=aUZik5jA29VJwM)

[10] La comparaison entre KDD, CRISP-DM et SEMMA

<https://core.ac.uk/download/pdf/47135941.pdf>

[11] Architecture trois tiers

<https://www.techno-science.net/definition/5266.html>

[12] Représentation de l'architecture trois tiers

[https://www.memoireonline.com/02/17/9661/m\\_Conception-et-realisation-dun-robot-virtuel-marketiste11.html](https://www.memoireonline.com/02/17/9661/m_Conception-et-realisation-dun-robot-virtuel-marketiste11.html)

[13] MVC

<https://www.irif.fr/~carton/Enseignement/InterfacesGraphiques/Cours/Swing/mvc.html>

[14] Figure de MVC

<https://fr.wikipedia.org/wiki/Mod%C3%A8le-vue-contr%C3%B4leur>

[15] Visual Studio Code

<https://framalibre.org/content/visual-studio-code>

[16] Draw.io

<https://www.tice-education.fr/tous-les-articles-et-ressources/articles-internet/819-draw-io-un-outil-pour-dessiner-des-diagrammes-en-ligne>

[17] Visual paradigm

<https://digitiz.fr/outil/visual-paradigm-online/>

[18] Postman

<https://www.postman.com/product/what-is-postman/>

[19] Jupyter Notebook

<https://realpython.com/jupyter-notebook-introduction/>

[20] React

<https://fr.wikipedia.org/wiki/React>

[21] JavaScript

<https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203585-javascript/>

[22] MongoDB

<https://www.mongodb.com/fr-fr/what-is-mongodb>

[23] Flask

<https://pythonbasics.org/what-is-flask-python/>

[24] Python

<https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/>

[25] Le traitement automatique des langues

<https://www.demotal.fr/le-traitement-automatique-des-langues/#:~:text=Avec%20l'av%C3%A8nement%20d'internet,%C3%A0%20une%20analyse%20s%C3%A9mantique%20fine.>

[26] Avantages du TAL

<https://www.itconvergence.com/blog/natural-language-processing-key-benefits-and-use-cases/>

[27] Apprentissage automatique

<https://www.ibm.com/fr-fr/topics/machine-learning>

[28] Les techniques d'apprentissage automatique

[https://www.trendmicro.com/fr\\_fr/what-is/machine-learning.html](https://www.trendmicro.com/fr_fr/what-is/machine-learning.html)

[29] Algorithmes de regroupement

<https://www.javatpoint.com/clustering-in-machine-learning#:~:text=Mean%2Dshift%20algorithm%3A%20Mean%2D,points%20within%20a%20given%20region.>

[30] K-means

<https://mrmint.fr/algorithme-k-means>

[31] Les étapes de K-means

<http://www.metz.supelec.fr/metz/personnel/vialle/course/BigData-2A-CS/slides-pdf/13-MachineLearning-Clustering-2spp.pdf>

[32] K-means avec différents nombres de clusters

<https://datascientest.com/algorithme-des-k-means>

[33] Mean-Shift

[https://en.wikipedia.org/wiki/Mean\\_shift](https://en.wikipedia.org/wiki/Mean_shift)

[34] Les étapes de Mean-Shift

[https://elearn.univ-tlemcen.dz/pluginfile.php/108519/mod\\_resource/content/1/MID-RdF-06.pdf](https://elearn.univ-tlemcen.dz/pluginfile.php/108519/mod_resource/content/1/MID-RdF-06.pdf)

[35] Mean-Shift pour un nombre de clusters égal à 3

[A demo of the mean-shift clustering algorithm — scikit-learn 1.2.2 documentation](#)



[36] DBSCAN

<https://datascientest.com/machine-learning-clustering-dbscan#:~:text=Le%20DBSCAN%20est%20un%20algorithme%20simple%20qui%20d%C3%A9finit%20des%20clusters,%20voisinage%20de%20l'observation.>

[37] Les étapes de DBSCAN

<https://fr.wikipedia.org/wiki/DBSCAN>

[38] DBSCAN avec 3 valeurs différentes d'épsilon

<https://datascientest.com/machine-learning-clustering-dbscan>

[39] Le rôle du ML dans le NLP

<https://ts2.space/fr/le-role-de-lapprentissage-automatique-dans-le-traitement-et-la-comprehension-du-langage-naturel/>

[40] Les bibliothèques Python

<https://geekflare.com/fr/popular-python-libraries-modules/>

[41] La normalisation

<https://www.actuia.com/contribution/victorbigand/tutoriel-tal-pour-les-debutants-classification-de-texte/>