République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

**Université Abderrahmane Mira – Béjaïa**

Faculté des Sciences Exactes

Département d'Informatique


# Mémoire de Fin de Cycle

En vue de l'obtention du diplôme de Master Professionnel en Informatique

Option : Génie Logiciel

# Thème

---

*Conception et développement d'un système de prédiction et de visualisation des clients à risque de churn*

***Cas : Algérie Télécom***

---

**Réalisé par :**

BOUKAMOUM Sarah

**Encadré par :**

Mme. AIT HACEN Souhila

**Membres du jury :**

Mme. EL BOUHISSI Houda
Mme. BELKHIRI Louiza
M. MOHAMMEDI Mohammed
Mme. KESSIRA Dalila

**Année universitaire : 2024 / 2025**

# Remerciements

*Je tiens tout d'abord à exprimer ma profonde reconnaissance à Dieu Tout-Puissant, qui m'a donné la force, la patience et la volonté de mener à bien ce travail.*

*Je remercie sincèrement madame AIT HACEN Souhila, ma promotrice, pour sa disponibilité, ses conseils précieux, son encadrement rigoureux et ses orientations pertinentes tout au long de ce projet.*

*Je tiens à adresser mes sincères remerciements à l'équipe technique et commerciale d'Algérie Télécom – Direction Opérationnelle de Béjaïa pour leur accueil chaleureux, leur disponibilité et leur collaboration tout au long de mon projet.*

*J'exprime toute ma gratitude aux membres du jury pour le temps qu'ils consacrent à lire avec soin ce mémoire, ainsi que pour l'intérêt qu'ils portent à l'évaluation de ce travail.*

*Mes remerciements vont également à mes amis et camarades de promotion, avec qui j'ai partagé ces années d'apprentissage, de défis et de réussite. Leur soutien moral, leur bienveillance et nos échanges constructifs ont rendu cette expérience universitaire bien plus enrichissante.*

*Je tiens à adresser une pensée toute particulière à ma famille, et plus précisément à mes chers parents, pour leur amour inconditionnel, leurs encouragements constants et les sacrifices qu'ils ont consentis afin de m'offrir les meilleures conditions pour réussir. Leur soutien a été pour moi une source de force et de motivation tout au long de mon parcours.*

*Enfin, je souhaite me remercier moi-même, pour ma persévérance, les efforts fournis malgré les difficultés, et ma capacité à croire en mes objectifs jusqu'au bout. Ce projet est le fruit d'un véritable engagement personnel, dont je suis fière.*

# Table des matières

# Table des figures

# Chapitre 1

# Introduction Générale

Ce premier chapitre présente d'abord le cadre général de l'étude en décrivant le contexte de travail, puis s'attarde sur les aspects théoriques et conceptuels liés à la notion de churn, ses différentes formes, ses causes principales . Ce cadre théorique permettra de mieux comprendre les dynamiques liées à la perte de clients et de situer le présent mémoire dans son contexte d'analyse. Il servira également de point de départ pour définir les objectifs visés.

## 1.1 Contexte et problématique

Dans un contexte économique de plus en plus concurrentiel, la fidélisation des clients représente un enjeu stratégique pour les entreprises, en particulier dans le secteur des télécommunications. La capacité à anticiper le départ des clients, connu sous le terme de *churn*, est devenue essentielle afin de mettre en place des actions préventives ciblées.

Le churn client désigne le phénomène par lequel un utilisateur décide de résilier son contrat ou de ne plus utiliser les services d'un fournisseur. Ce comportement peut être influencé par divers facteurs, tels que la qualité du service, le coût, ou encore l'émergence d'alternatives plus attractives. Pour les fournisseurs d'accès à Internet, en particulier dans les pays en développement, ce phénomène peut avoir un impact significatif sur la rentabilité et la pérennité de l'activité.

Face à cette problématique, l'exploitation des données clients à l'aide d'approches d'intelligence artificielle, et plus spécifiquement les algorithmes d'apprentissage automatique (machine learning), offre de nouvelles perspectives prometteuses. Ces méthodes permettent de modéliser et de prédire le comportement des clients, en identifiant les signaux faibles annonciateurs d'un éventuel départ.

Ce travail s'inscrit dans cette démarche. Il propose une solution basée sur l'application de modèles de machine learning pour la prédiction du churn, en s'appuyant sur l'analyse de données clients historiques. L'objectif principal est de concevoir un système capable de détecter les clients à risque, afin de permettre aux décideurs de mettre en œuvre des stratégies de rétention plus

efficaces et personnalisées. Dans ce qui suit, nous présentons quelques définitions nécessaires à la compréhension de notre travail :

## 1.2 Quelques définitions

### 1.2.1 Définition de churn (perte de clients)

Le terme *churn*, issu de la contraction des mots anglais *change* et *turn*, désigne le phénomène de perte de clientèle. Il est généralement mesuré à travers le **taux de churn**, un indicateur clé pour les organisations. Ce taux représente le pourcentage de clients perdus sur une période donnée, par rapport au nombre total de clients au début de cette période [1].

$$\text{Taux de churn} = \frac{\text{Nombre de clients perdus pendant une période}}{\text{Nombre total de clients au debut de la période}} \times 100 \tag{1.1}$$

### 1.2.2 Typologies de churn

On distingue plusieurs formes essentielles de churn selon la nature du départ du client :

- Churn volontaire : le client décide de son propre chef de mettre fin à la relation (ex. : changement de fournisseur ou abandon du service).
- Churn involontaire : le départ est indépendant de la volonté du client (ex. : décès, impayés, résiliation automatique). [2]

### 1.2.3 Les facteurs clés liés à la perte de clients

Une étude réalisée par **Lanseur Akila** et **Ait Sidhoum Houria**, enseignantes-chercheuses à l'Université de Bejaia, s'est intéressée aux déterminants du churn dans les trois opérateurs télécom en Algérie [3].Les auteures ont identifié plusieurs facteurs influençant le départ des clients. Les principaux déterminants sont :

- La qualité de service : notamment la couverture réseau, le débit Internet et la stabilité des connexions. Une mauvaise qualité perçue pousse fortement les clients à changer d'opérateur.
- Les offres intéressantes chez l'autre opérateur : La disponibilité de promotions attractives, de forfaits mieux adaptés ou de bonus plus généreux chez un opérateur concurrent influence fortement la décision de changement.

## 1.3 Objectifs de l'étude

Ce mémoire s'inscrit dans le cadre de notre projet de fin d'études et a pour objectif la conception et la mise en place d'un système de prédiction des clients à risque de churn chez

Algérie Télécom. L'approche proposée repose sur l'élaboration d'un modèle prédictif capable d'analyser les comportements réels des abonnés à partir de données historiques, dans le but d'identifier ceux qui présentent un risque élevé de résiliation.

Pour ce faire, nous mettrons en œuvre des techniques d'apprentissage automatique, afin d'assurer la précision, la fiabilité et la robustesse du modèle prédictif. L'objectif final de notre système est non seulement de détecter de manière proactive les clients susceptibles de se désabonner, mais aussi de fournir aux responsables marketing un tableau de bord interactif leur permettant de visualiser clairement les segments à risque.

Ce système offrira également aux analystes de données la possibilité d'explorer les différentes variables explicatives du churn et de comprendre les facteurs qui poussent les clients à quitter l'opérateur. À travers cette solution, Algérie Télécom pourrait mettre en œuvre des stratégies de fidélisation plus ciblées et plus efficaces, dans une optique de réduction du taux de désabonnement et d'amélioration de la satisfaction client.

## 1.4 Structure du rapport

Ce rapport est structuré comme suit :

— **Chapitre 1 : Introduction Générale** : Présentation du contexte, de la problématique et des objectifs de l'étude.

— **Chapitre 2 : Présentation de l'organisme d'accueil** — Description de l'entreprise, de son environnement et de ses services.

— **Chapitre 3 : État de l'art** : Présentation des concepts clés du machine learning, des algorithmes utilisés, des techniques de validation et d'évaluation, ainsi qu'une revue des travaux existants sur la prédiction du churn.

— **Chapitre 4 : Conception** : Description de la méthode agile UP adoptée, ainsi que des outils de modélisation UML à travers divers diagrammes (cas d'utilisation, séquence, classes, etc.).

— **Chapitre 5 : Réalisation et résultats** : Présentation des étapes de développement du modèle prédictif et de l'application, illustrée par des captures d'écran et une explication des principales fonctionnalités implémentées.

# Chapitre 2

# Présentation de l'organisme d'accueil

## 2.1 Introduction

Après avoir défini dans le premier chapitre les concepts clés liés à la problématique du churn dans le secteur des télécommunications, ce deuxième chapitre se concentre sur le contexte réel dans lequel s'inscrit notre travail, à savoir celui d'Algérie Télécom, l'organisme d'accueil de notre stage.

Nous y présentons les principales caractéristiques de l'entreprise, ses missions, ses domaines d'activité, ainsi que les défis qu'elle rencontre en matière de fidélisation de la clientèle. Enfin, nous introduisons la solution proposée dans ce mémoire, conçue pour répondre aux besoins spécifiques de l'entreprise face à cette problématique.

## 2.2 Présentation du groupe Algérie Télécom

Algérie Télécom est l'opérateur historique des télécommunications en Algérie, créé en 2003 après la restructuration du secteur des postes et télécommunications. Il est responsable de la gestion, de l'exploitation et du développement des réseaux de télécommunications fixes, d'internet haut débit, ainsi que des services sans fil à travers le pays. Algérie Télécom joue un rôle clé dans la transformation numérique du pays, en fournissant des services à la fois aux particuliers et aux entreprises, tout en s'engageant à renforcer la qualité de ses services et à étendre sa couverture, notamment dans les zones rurales. L'entreprise est également un acteur majeur dans le domaine de la fibre optique et des télécommunications par satellite.

### 2.2.1 Les filiales du groupe Algérie Télécom

- **ALGÉRIE TÉLÉCOM MOBILE** : Opérateur de réseau mobile et fournisseur d'internet haut débit sans fil.
- **ALGÉRIE TÉLÉCOM SATELLITE (ATS)** : Spécialiste des services de télécommunications par satellite.

- **ALGÉRIE TÉLÉCOM EUROPE (ATE)** : Société responsable de la gestion du câble sous-marin « ORVAL/ALVAL », reliant les réseaux télécoms algériens aux réseaux européens.
- **COMINTAL SPA** : Entreprise spécialisée dans la fourniture de solutions et équipements de fibre optique brute (fibre optique noire).
- **SATICOM SPA** : Entreprise qui propose des solutions technologiques modernes permettant aux entreprises de communiquer de manière plus efficace, tant en interne qu'en externe.

### 2.2.2 Missions et Objectifs d'Algérie Télécom

Les missions et objectifs d'Algérie Télécom s'inscrivent dans une stratégie de modernisation et de développement des infrastructures numériques à l'échelle nationale. Elle joue un rôle central dans le déploiement et la démocratisation de l'accès aux technologies de l'information et de la communication (TIC).

#### Missions principales

- Fournir des services de télécommunications fiables et accessibles à l'ensemble du territoire national, aussi bien en milieu urbain que rural.
- Développer et entretenir les infrastructures de télécommunications, notamment les réseaux de fibre optique, l'ADSL, la 4G LTE Fixe et d'autres solutions haut débit.
- Offrir une large gamme de services incluant la téléphonie fixe, l'accès à Internet, les services de données, et les solutions d'hébergement cloud.
- Assurer un service public de qualité, en veillant au respect des normes de sécurité, de confidentialité et de performance.
- Accompagner la transformation numérique du pays en participant activement aux projets de digitalisation des institutions, des entreprises et des citoyens.

#### Objectifs stratégiques

- Moderniser les réseaux de communication à travers l'introduction continue de nouvelles technologies (ex : FTTH, IPv6, data centers. . .).
- Améliorer la qualité de service et la satisfaction client, en mettant en place des solutions plus rapides, stables et adaptées aux besoins des utilisateurs.
- Étendre la couverture réseau nationale, afin de réduire la fracture numérique entre les différentes régions du pays.
- Renforcer la position d'Algérie Télécom comme acteur clé du développement économique, en soutenant les projets nationaux d'e-administration, d'éducation numérique, et de cybersécurité.

- S'orienter vers l'innovation en diversifiant ses services (tels que les solutions cloud, la visioconférence, les services VoIP, etc.).

**Algérie Télécom en chiffres**

Voici quelques statistiques clés concernant Algérie Télécom :

- Plus de 6,5 millions de clients raccordés à Internet, ce qui reflète l'importance de sa présence dans les foyers algériens.
- 1,9 million de clients sont connectés via la fibre optique, témoignant des efforts de modernisation des infrastructures.
- 2,7 millions de clients utilisent les technologies ADSL/VDSL, encore largement déployées dans plusieurs régions.
- 1,9 million de clients utilisent le service Idoom 4G, offrant une connectivité alternative dans les zones moins couvertes par la fibre.
- Un réseau de plus de 500 agences commerciales et points de présence assure une proximité avec les clients.
- 91% des sites d'accueil sont labellisés "Fi Khidmatikom", gage de qualité de service et d'engagement envers la satisfaction client.[4]

### 2.2.3 Services proposés par Algérie Télécom

| Catégorie | Services proposés |
|---|---|
| Internet | - IDOOM ADSL : Internet haut débit via ligne téléphonique (jusqu'à 20 Mbps).<br>- IDOOM Fibre (FTTH) : Internet très haut débit via fibre optique (jusqu'à 300 Mbps).<br>- IDOOM 4G LTE : Internet sans fil via modem 4G.<br>- Wi-Fi Outdoor : Hotspots publics (aéroports, universités, etc.). |
| Téléphonie Fixe | - Téléphonie classique : Appels via ligne fixe.<br>- IDOOM Fixe : Forfaits appels nationaux/internationaux.<br>- Téléphonie sur IP (VoIP) : Pour entreprises. |
| Téléphonie Mobile (Mobilis) | - Appels voix GSM/3G/4G.<br>- SMS et MMS.<br>- Forfaits voix et data prépayés et postpayés.<br>- Services 4G LTE mobile.<br>- Services roaming et appels internationaux. |
| Services aux Entreprises | - Liaisons spécialisées : VPN MPLS, connexions dédiées.<br>- Hébergement web et noms de domaine.<br>- IDC (Internet Data Center) : Cloud, colocation, sauvegarde.<br>- Fibre Pro / SDSL : Connexion dédiée avec garantie de service.<br>- Numéros spéciaux : Numéros verts, courts, etc. |
| Services Numériques | - E-paiement : Paiement des factures en ligne.<br>- Application mobile Algérie Télécom.<br>- Portail client : Suivi abonnement, factures.<br>- Services SMS : Notifications, alertes, solde. |
| Formules et Offres | - IDOOM Pack : ADSL + téléphonie fixe.<br>- Cartes de recharge : ADSL, 4G, VoIP.<br>- Offres promotionnelles : Bonus de volume, réduction, débit temporaire. |

TABLE 2.1 – Services proposés par Algérie Télécom et ses filiales

### 2.2.4 L'organigramme Général de l'entreprise

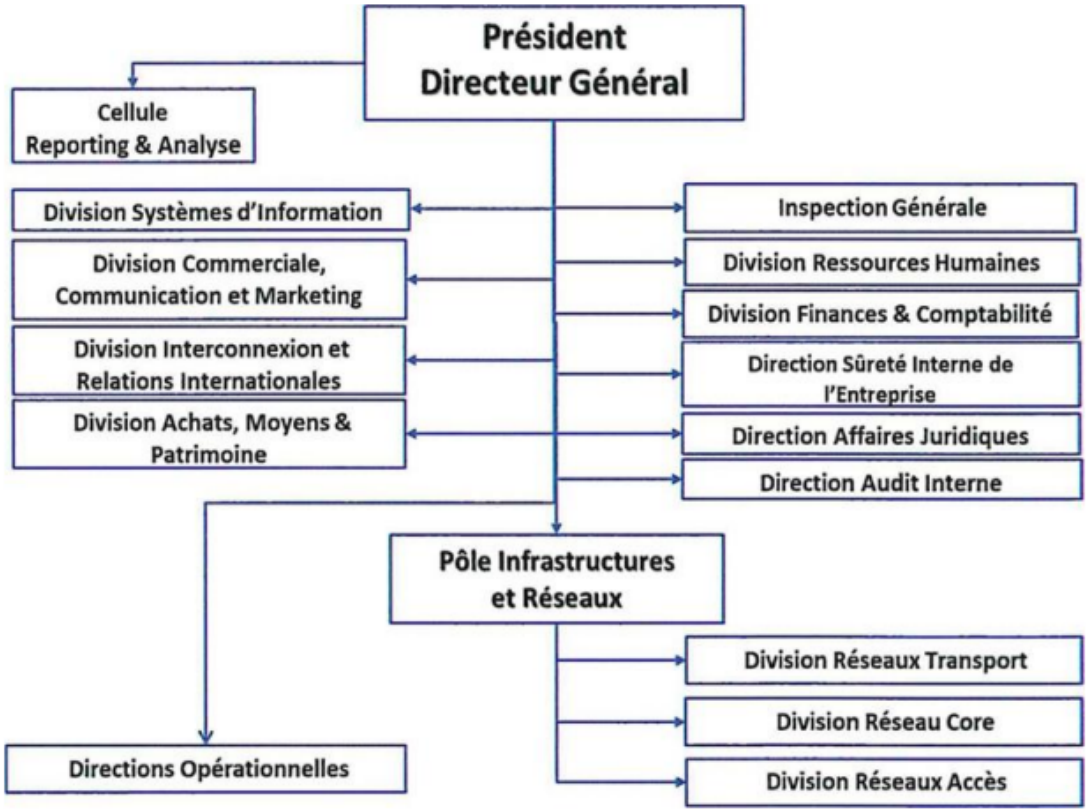



FIGURE 2.1 – Organigramme Général d'Algérie Télécom

## 2.3 Enjeux liés à la fidélisation client

Dans un marché des télécommunications de plus en plus concurrentiel, où les clients sont très sollicités, notamment par les opérateurs privés, la fidélisation devient un enjeu majeur pour les acteurs historiques comme Algérie Télécom

### Définition de la fidélisation

La fidélisation désigne l'ensemble des actions marketing mises en œuvre par une entreprise afin d'inciter les clients à rester fidèles à sa marque, à ses produits ou à ses services. Elle vise à renforcer la satisfaction, à créer un attachement durable, et à encourager la répétition d'achats, même en présence de sollicitations concurrentielles. [5]

### L'importance de la fidélisation des clients

La fidélisation des clients est un enjeu stratégique majeur pour toute entreprise. Elle permet non seulement de renforcer la relation avec les clients existants, mais aussi d'assurer une croissance durable en réduisant les coûts et en optimisant la rentabilité à long terme. Ci-dessous, les différents aspects de son importance :

- Optimisation des coûts : fidéliser un client existant revient nettement moins cher que d'en

conquérir un nouveau. Les coûts de fidélisation peuvent représenter seulement un tiers, voire un dixième, des coûts d'acquisition.

- Amélioration de la rentabilité : Les clients fidèles ont tendance à accroître leurs achats au fil du temps, en particulier dans les secteurs des services et du B2B, ce qui renforce la rentabilité globale de l'entreprise.
- Stabilisation des revenus : La fidélité client contribue à maintenir un chiffre d'affaires plus constant, car ces clients sont généralement moins réceptifs aux sollicitations concurrentes.
- Effet levier par le bouche-à-oreille : Les clients fidèles agissent souvent comme des ambassadeurs de la marque, partageant spontanément leur satisfaction avec leur entourage et générant ainsi une publicité gratuite et crédible.[6]

### 2.3.1 Le phénomène de Churn chez Algérie Télécom

Algérie Télécom fait face au phénomène de churn depuis plusieurs années, un défi majeur qui affecte l'ensemble de ses services , notamment les offres d'accès à Internet . Ce phénomène de perte de clients est un enjeu stratégique pour l'entreprise, impactant non seulement sa part de marché mais également sa rentabilité.

### 2.3.2 Les méthodes traditionnelles utilisées par Algérie Télécom pour comprendre le Churn

Algérie Télécom s'appuie sur plusieurs méthodes pour comprendre les raisons du churn :

- **Analyse des motifs de résiliation :** Cette méthode consiste à examiner les raisons évoquées par les clients lors de la résiliation, en s'appuyant sur des données issues des enquêtes, des appels au service client ou des commentaires sur les réseaux sociaux. Le taux de résiliation est un indicateur clé qui permet d'estimer la proportion de clients perdus.
- **Analyse des comportements :** Cette approche étudie les actions des clients avant leur départ, comme la baisse d'utilisation des services, la fréquence des réclamations ou des factures inhabituelles. Elle aide à repérer les signaux précurseurs d'un départ probable.
- **Analyse des profils :** Algérie Télécom se concentre sur les caractéristiques personnelles des clients (âge, sexe, localisation, segment...) afin d'identifier les groupes les plus susceptibles de résilier. Cette méthode croise à la fois les comportements et les motifs pour mieux cibler les actions de fidélisation.

Bien que ces méthodes traditionnelles soient utiles pour comprendre les raisons du churn, elles présentent certaines limites. Elles demeurent essentiellement réactives et ne permettent pas d'anticiper efficacement les départs futurs. L'analyse des données ainsi collectées se limite à l'observation de tendances passées, sans offrir de capacité prédictive suffisante. Cela rend difficile la mise en place de stratégies de fidélisation proactives et personnalisées.

## 2.4 Solution proposée

Dans le cadre de notre projet de fin d'études, nous proposons de concevoir et de développer une solution innovante répondant efficacement aux exigences de l'entreprise et aux problématiques liées à la perte de clients. Cette solution prendra la forme d'un **système de prédiction et de visualisation des clients à risque de churn**.

Les méthodes classiques de détection du churn atteignent aujourd'hui leurs limites, rendant nécessaire l'adoption d'une démarche plus proactive. Pour une entreprise comme Algérie Télécom, l'utilisation de modèles statistiques et d'algorithmes de machine learning permet d'anticiper les départs en identifiant les signes avant-coureurs dans les données clients. Une telle approche représente un levier stratégique, en offrant à l'entreprise la possibilité d'intervenir avant que la décision de résiliation ne devienne définitive.

Notre approche repose sur les axes suivants :

- Une analyse approfondie des comportements réels des clients, afin d'identifier les tendances et signaux précurseurs du churn.
- La création d'un modèle prédictif fiable et précis, capable de détecter proactivement les clients susceptibles de quitter l'entreprise, en s'appuyant sur des techniques de machine learning.
- Le développement d'un tableau de bord stratégique, permettant de visualiser clairement les résultats des prédictions et les informations clés pour une prise de décision efficace.

En outre, cette solution permettrait à Algérie Télécom de mieux comprendre le comportement de ses clients à travers l'analyse de données issues de l'utilisation des services (consommation, réclamations, etc.). Elle faciliterait l'identification des facteurs d'insatisfaction, tout en améliorant l'efficacité opérationnelle en concentrant les efforts marketing sur les segments les plus vulnérables.

Ce système offrira à l'entreprise, et plus particulièrement au service commercial , la capacité de cibler les clients à risque, d'anticiper leur départ et de mettre en œuvre des stratégies de rétention adaptées. Il constituera également un outil d'aide à la décision pour les spécialistes en data analytics, en facilitant la compréhension des comportements clients et l'identification des facteurs de leur insatisfaction.

## 2.5 Conclusion

Ce chapitre a permis de présenter en détail Algérie Télécom, en mettant en lumière son rôle clé dans le secteur des télécommunications en Algérie ainsi que les défis associés à la fidélisation des clients et à la gestion du churn. Nous avons également introduit la solution proposée pour prédire les départs des clients en utilisant des techniques avancées d'apprentissage automatique.

Le chapitre suivant se concentrera sur les concepts fondamentaux de l'apprentissage automatique, notamment les approches supervisées et non supervisées, ainsi que les techniques d'évaluation des modèles. Cette transition est essentielle pour mieux comprendre les fondements des méthodes utilisées dans notre solution de prédiction.

# Chapitre 3

# État de l'art sur les techniques de l'apprentissage automatique

## 3.1 Introduction

Après avoir présenté la problématique du churn dans le secteur des télécommunications et le contexte spécifique d'Algérie Télécom, ce chapitre est consacré à l'étude des fondements théoriques et des approches existantes en matière de prédiction de la perte de clients.

L'objectif est de fournir un aperçu des principales techniques d'apprentissage automatique utilisées dans ce domaine, en mettant en évidence les concepts clés, les types d'algorithmes, les méthodes d'évaluation des modèles ainsi que les recherches antérieures menées dans le secteur des télécoms. Cette exploration permettra de justifier les choix méthodologiques adoptés dans la suite de ce mémoire et de situer notre solution dans un cadre scientifique rigoureux.

## 3.2 l'intelligence artificielle (IA)

L'intelligence artificielle est un champ de recherche en informatique qui a pour objectif de concevoir des machines ou des programmes capables d'accomplir des tâches qui, jusqu'à récemment, nécessitaient une intelligence humaine. Ces tâches peuvent inclure la résolution de problèmes, la compréhension du langage naturel, la reconnaissance visuelle ou encore la capacité d'apprentissage à partir de l'expérience.[7]

## 3.3 L'Apprentissage Automatique

### 3.3.1 Définition

L'apprentissage automatique, ou machine learning, constitue un domaine fondamental de l'intelligence artificielle. Son objectif principal est de développer des systèmes informatiques

capables d'extraire automatiquement des connaissances ou des modèles à partir de données, afin d'adapter leur comportement ou de prendre des décisions sans qu'ils soient explicitement programmés pour chaque situation.[8]

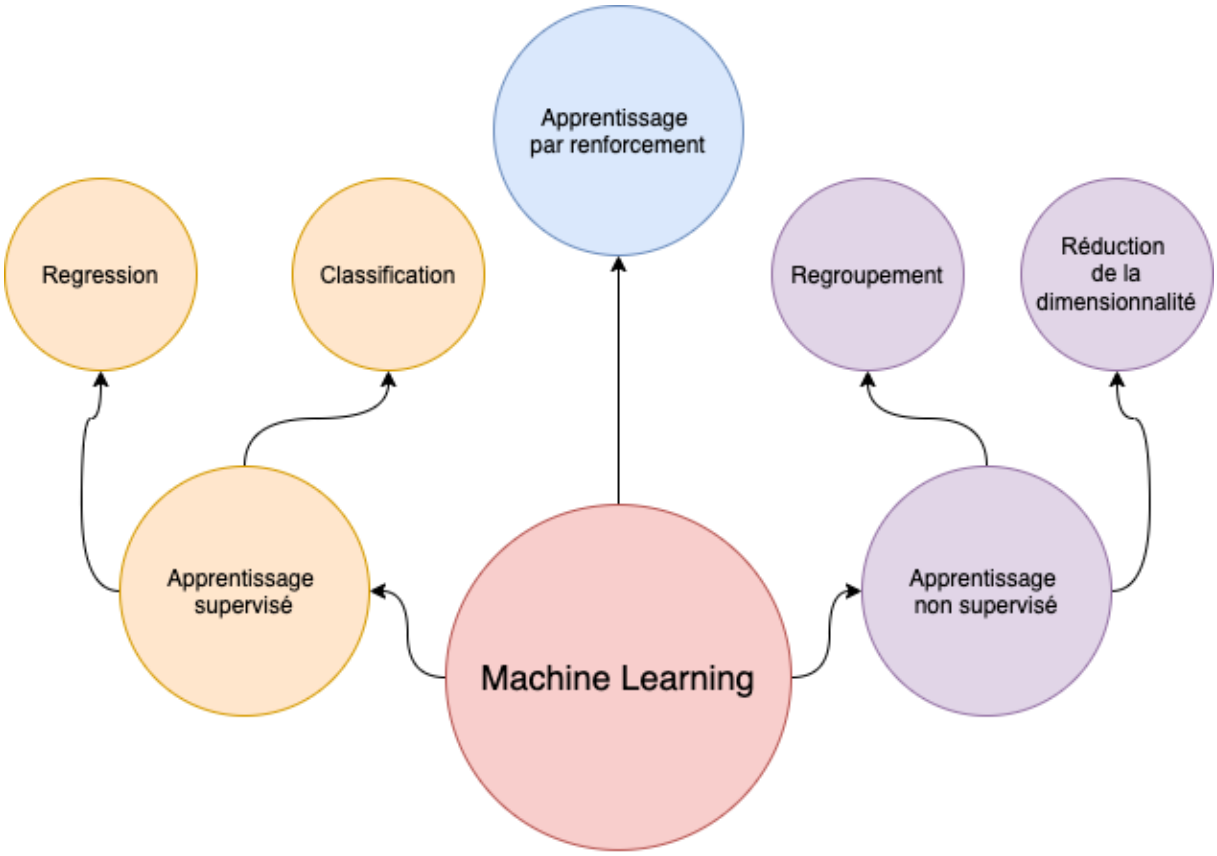



FIGURE 3.1 – Types d'Apprentissage Automatique

La figure 3.1 illustre les principaux types d'apprentissage automatique, que nous détaillerons dans les sections suivantes :

### 3.3.2 Apprentissage supervisé

L'apprentissage supervisé est une branche du machine learning qui consiste à apprendre à partir de données déjà annotées. Autrement dit, on dispose d'un ensemble d'exemples pour lesquels on connaît à la fois les entrées (c'est-à-dire les caractéristiques décrivant chaque situation ou objet) et les sorties attendues (les réponses ou étiquettes associées).L'objectif est de construire un modèle capable de produire des prédictions fiables sur de nouvelles données jamais vues auparavant. On suppose qu'il existe une relation entre les données et les réponses, mais cette relation est inconnue et perturbée par des éléments imprévisibles (comme du bruit ou des erreurs).Dans la plupart des cas, les données sont numériques, mais l'apprentissage supervisé peut également s'appliquer à des données plus variées, comme des catégories, des textes, ou des structures complexes comme des graphes. Le rôle de l'algorithme est donc d'apprendre cette relation sous-jacente pour fournir des prédictions les plus justes possibles. [9]

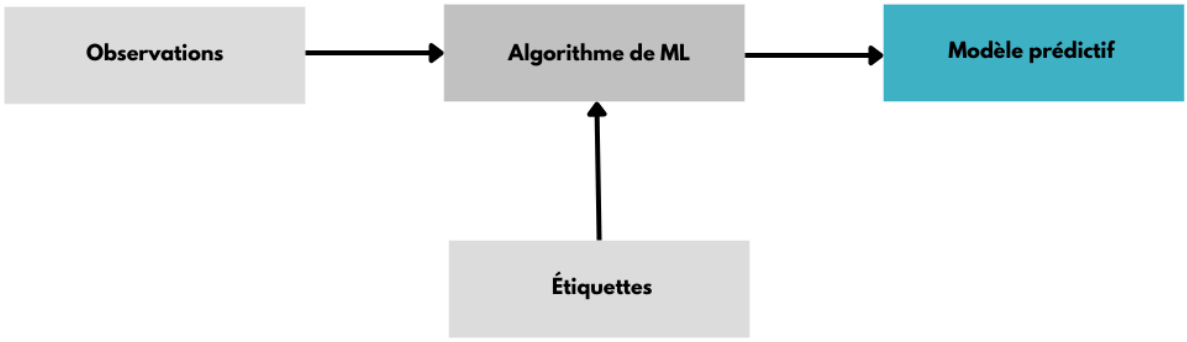



FIGURE 3.2 – Apprentissage supervisé

L'apprentissage supervisé regroupe plusieurs types de problèmes selon la nature des étiquettes à prédire. Dans ce qui suit, nous présentons les principales catégories rencontrées :

1. **Classification** : il s'agit des problèmes pour lesquels les étiquettes appartiennent à un ensemble fini de classes distinctes. Ce type de tâche vise à affecter chaque observation à l'une des catégories prédéfinies.[9]
2. **Régression** : ce sont les problèmes où les étiquettes à prédire sont des valeurs continues. Le but est alors de modéliser une relation permettant d'estimer une variable quantitative à partir des données d'entrée.[9]

### 3.3.3 Apprentissage non supervisé

L'apprentissage non supervisé est une branche du machine learning dans laquelle les données utilisées pour entraîner le modèle ne sont pas accompagnées d'étiquettes ou de réponses connues. L'objectif est d'explorer ces données pour en découvrir la structure cachée, comme regrouper des éléments similaires ou détecter des modèles récurrents.Contrairement à l'apprentissage supervisé, où l'on connaît les résultats attendus, ici le modèle doit apprendre par lui-même à interpréter les données, en identifiant des relations, des regroupements ou des schémas sans aide extérieure[9]


FIGURE 3.3 – Apprentissage non supervisé

Ce type d'apprentissage englobe plusieurs approches, parmi lesquelles on distingue :

1. **Clustering (regroupement) :** Le clustering, ou partitionnement, est une technique d'apprentissage non supervisé qui consiste à regrouper des données similaires en plusieurs ensembles appelés clusters. L'objectif est d'identifier des groupes cohérents au sein des données, sans connaître à l'avance les catégories ou classes auxquelles elles appartiennent.Chaque groupe ainsi formé rassemble des observations qui partagent des caractéristiques communes, ce qui permet de mieux comprendre la structure globale des données. Une fois ces groupes identifiés, il est aussi possible d'interpréter ou d'anticiper certaines propriétés d'une nouvelle donnée en fonction du groupe auquel elle est associée. En pratique, le clustering cherche à diviser un ensemble de données en plusieurs sous-ensembles pertinents, en se basant sur des critères comme la proximité, la densité ou la distribution des points dans l'espace des données.[9]
2. **Réduction de dimension :**

   La réduction de dimension est une méthode d'apprentissage non supervisé qui consiste à transformer des données initialement représentées dans un espace de grande dimension en un espace de dimension plus réduite, tout en préservant au maximum les caractéristiques essentielles des données.Cette transformation vise à simplifier la structure des données, faciliter leur visualisation, accélérer les traitements, ou encore améliorer la performance d'autres algorithmes d'apprentissage. Les nouvelles représentations obtenues dans cet espace réduit doivent conserver les informations les plus pertinentes selon les objectifs de l'analyse, comme la proximité entre les points ou la variance des données.En résumé, la réduction de dimension permet d'alléger la complexité des données tout en conservant leur signification principale, ce qui en fait un outil précieux pour l'exploration de données complexes et volumineuses.[9]
3. **Estimation de densité :** L'estimation de densité est une tâche d'apprentissage non supervisé qui consiste à modéliser la distribution sous-jacente des données. Autrement dit, on cherche à estimer une loi de probabilité à partir d'un ensemble de données, en supposant que celles-ci constituent un échantillon aléatoire représentatif.Cette approche permet de comprendre comment les données sont réparties dans l'espace, d'identifier des zones de forte ou de faible densité, et de détecter d'éventuelles anomalies ou structures particulières. L'estimation de densité joue un rôle fondamental dans plusieurs applications, notamment en détection de valeurs aberrantes, en génération de données ou encore dans l'analyse exploratoire.[9]

### 3.3.4 Apprentissage par renforcement

L'apprentissage par renforcement est une méthode d'apprentissage automatique dans laquelle un agent (système) intelligent interagit avec un environnement en effectuant des actions, et reçoit des récompenses en fonction de la pertinence de ses décisions. Une action appropriée est suivie d'une récompense positive, tandis qu'une mauvaise décision entraîne une récompense négative.

Dans de nombreux cas, ces récompenses ne sont pas immédiates, mais apparaissent après une succession d'actions, comme c'est le cas dans des jeux stratégiques tels que le go ou les échecs. Le but de cet apprentissage est d'amener l'agent à élaborer une stratégie optimale, appelée politique, lui permettant de maximiser les récompenses cumulées sur le long terme.[9]

## 3.4 Présentation des principaux algorithmes de Machine Learning appliqués au prédiction de Churn

Dans cette section, nous présentons les principaux algorithmes de machine learning utilisés pour résoudre des problèmes de prédiction, et plus particulièrement ceux liés au churn. Ce type de prédiction correspond généralement à une tâche de classification supervisée, où l'objectif est de déterminer si un client est susceptible de se désabonner (classe 1) ou non (classe 0), à partir de ses données comportementales ou contractuelles.

### 3.4.1 La régression logistique

La régression logistique est un modèle statistique et un algorithme d'apprentissage automatique utilisé pour prédire un résultat binaire, comme "oui" ou "non". Elle permet d'analyser la relation entre plusieurs variables explicatives et une variable cible, qui ne peut prendre que deux valeurs (0 ou 1).Ce modèle repose sur une fonction appelée sigmoïde (figure 3.4), qui transforme une combinaison des variables en une probabilité comprise entre 0 et 1. Si cette probabilité dépasse un certain seuil (généralement 0,5), le modèle prédit une catégorie, sinon il prédit l'autre . L'objectif de l'apprentissage est d'ajuster les paramètres du modèle pour améliorer la précision des prédictions en minimisant les erreurs. [10].

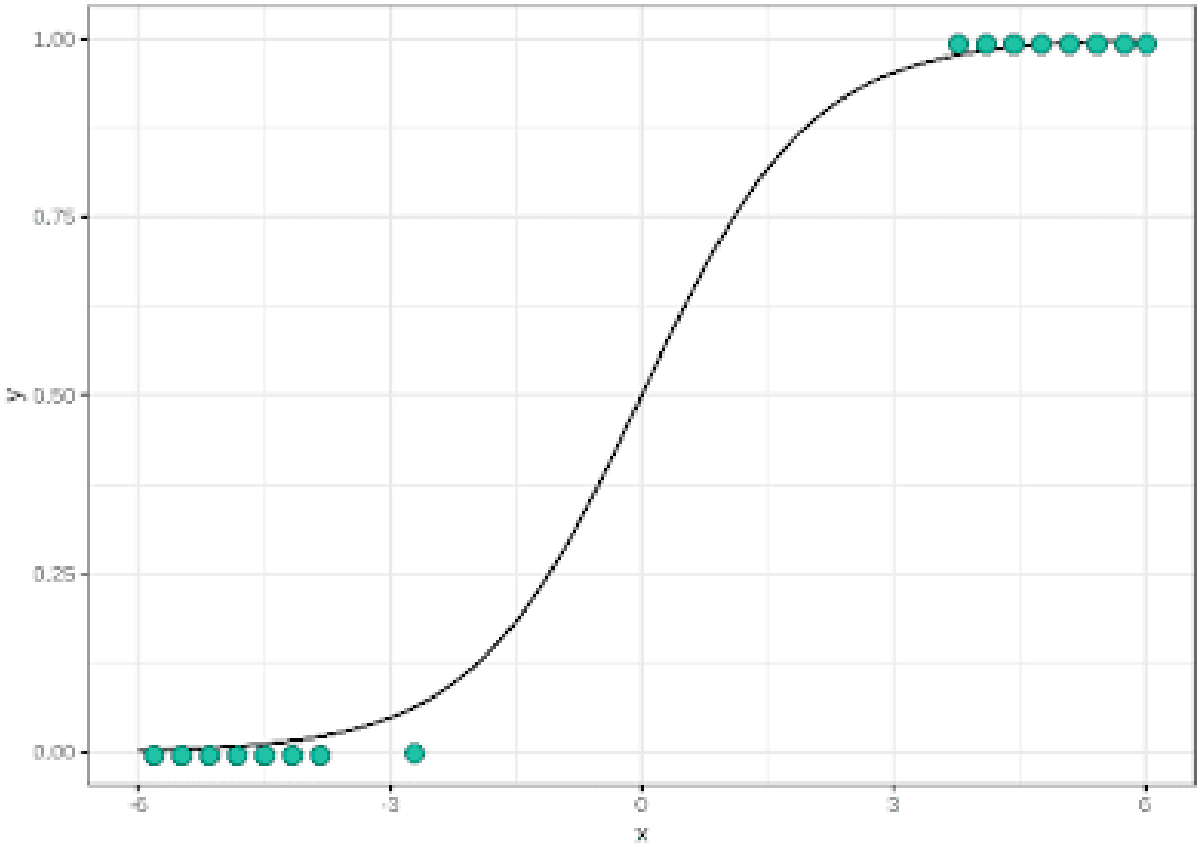



FIGURE 3.4 – La fonction Sigmoïde [10]

### 3.4.2 Arbres de décision (Decision Trees)

Un arbre de décision est un algorithme d'apprentissage supervisé non paramétrique, utilisé aussi bien pour la classification que pour la régression. Il se structure de manière hiérarchique sous forme d'un arbre composé d'un nœud racine, de branches, de nœuds internes et de nœuds feuilles.Comme illustré dans la figure 3.7 , l'arbre commence par un nœud racine (root node), qui ne possède aucune branche entrante. À partir de ce nœud, des branches sortantes conduisent vers des nœuds internes (internal nodes), également appelés nœuds de décision. Ces derniers appliquent des critères de séparation basés sur les caractéristiques des données, divisant progressivement l'ensemble en sous-groupes homogènes. Enfin, ces subdivisions aboutissent aux nœuds feuilles (leaf nodes) ou nœuds terminaux, qui représentent les différentes prédictions possibles du modèle.[11]

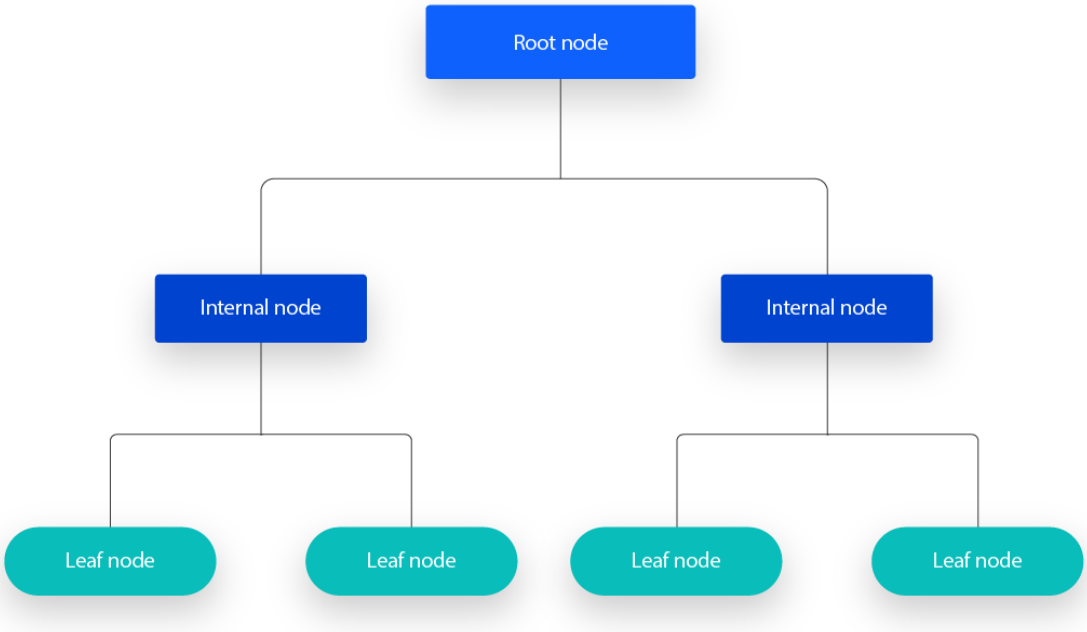



FIGURE 3.5 – Arbres de décision.[11]

### 3.4.3 Les forêts aléatoires (Random Forest)

La Random Forest est une méthode d'apprentissage automatique utilisée pour résoudre des problèmes de classification et de régression. Elle repose sur l'apprentissage ensembliste, qui combine plusieurs arbres de décision pour améliorer la précision des prédictions. Plus le nombre d'arbres est élevé, plus le modèle devient précis en prenant la moyenne des résultats obtenus par chaque arbre.Un des avantages de cet algorithme est qu'il offre des résultats fiables sans nécessiter d'ajustements complexes des hyperparamètres. Il est donc particulièrement adapté pour la prédiction du churn .

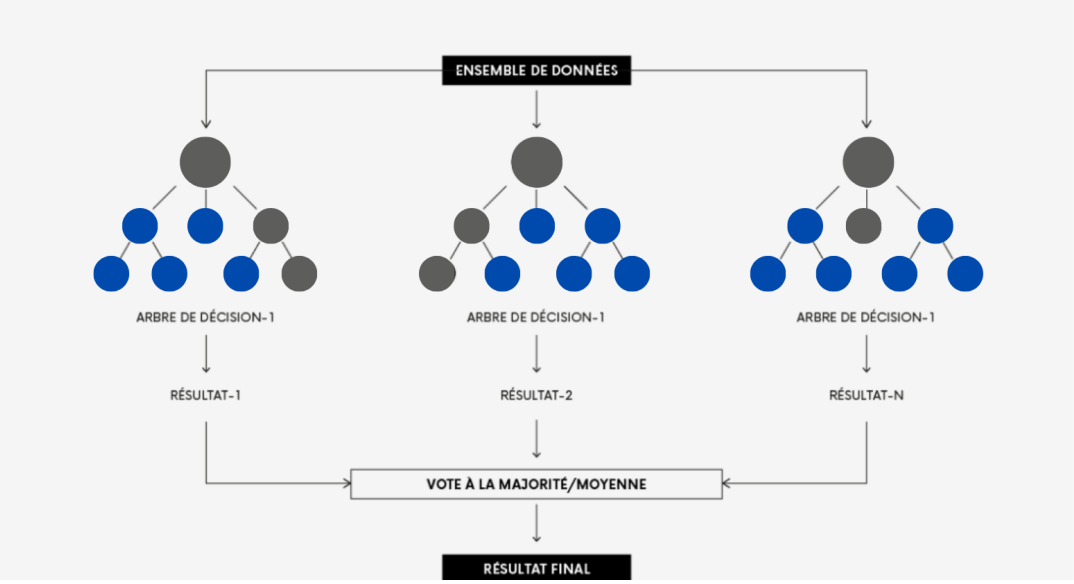



FIGURE 3.6 – Les forêts aléatoires

### 3.4.4 XGBoost

XGBoost (Extreme Gradient Boosting) est un algorithme d'apprentissage automatique basé sur la méthode du gradient boosting, qui vise à améliorer la performance prédictive d'un modèle en combinant plusieurs arbres de décision faibles de manière séquentielle. À chaque itération, XGBoost construit un nouvel arbre pour corriger les erreurs résiduelles des arbres précédents en minimisant une fonction de perte, souvent le logarithme de vraisemblance pour la classification ou l'erreur quadratique pour la régression.[12]

## 3.5 Etapes du processus de développement d'un modèle prédictif

Cette section décrit les différentes étapes nécessaires à la mise en œuvre d'un modèle de prédiction en apprentissage automatique. De la compréhension des données jusqu'à l'évaluation finale des performances, chaque phase joue un rôle essentiel pour garantir la qualité et la fiabilité du modèle obtenu.

### 3.5.1 L'analyse exploratoire des données EDA

L'analyse exploratoire des données (Exploratory Data Analysis – EDA) constitue une étape essentielle dans tout projet de science des données. Elle permet de mieux comprendre la structure du dataset, de détecter d'éventuelles anomalies, valeurs manquantes ou distributions déséquilibrées, et d'identifier les relations entre les variables. Cette phase vise à formuler des premières

hypothèses et à orienter les choix méthodologiques pour la modélisation prédictive. Dans cette section, nous examinons les principales caractéristiques des données utilisées, à travers des statistiques descriptives, des visualisations graphiques, ainsi qu'une étude des corrélations entre variables.

### 3.5.2 Entraînement d'un Modèle

La phase d'entraînement constitue l'étape initiale du processus de développement d'un modèle d'apprentissage automatique. Elle consiste à fournir au modèle un ensemble de données étiquetées, appelées données d'entraînement , afin qu'il puisse apprendre les relations ou motifs existants entre les variables d'entrée (caractéristiques) et les sorties attendues (étiquettes).Durant cette étape, l'algorithme ajuste ses paramètres internes pour minimiser une fonction de coût, en utilisant une méthode d'optimisation comme la descente de gradient. Le but est que le modèle généralise bien à de nouvelles données, sans se contenter de mémoriser l'ensemble d'entraînement.Le choix de l'algorithme, la qualité des données et les techniques de prétraitement jouent un rôle crucial dans la réussite de cette phase.

### 3.5.3 Validation d'un Modèle

La phase de validation intervient après l'entraînement du modèle et a pour objectif d'évaluer sa capacité à généraliser sur des données qu'il n'a jamais vues auparavant **données de validation**, tout en ajustant ses paramètres internes appelés hyperparamètres (par exemple, la profondeur d'un arbre ou le taux d'apprentissage). Les données de validation sont séparées des données d'entraînement afin d'assurer une évaluation impartiale. Cette étape permet de détecter le surapprentissage (overfitting) .Dans ce qui suit, nous présentons les principales techniques de validation utilisées .

#### A. Validation Croisée ($k$-fold Cross-Validation)

La k-fold cross-validation est une méthode utilisée pour évaluer la performance d'un modèle d'apprentissage automatique . Dans cette méthode, l'ensemble de données est divisé en $k$ sous-ensembles (ou folds) de taille similaire. Ensuite, le modèle est entraîné sur $k-1$ de ces sous-ensembles, et évalué sur le sous-ensemble restant, qui sert de jeu de validation. Ce processus est répété $k$ fois, de sorte que chaque sous-ensemble est utilisé $k-1$ fois pour l'entraînement et une fois pour le test. À la fin, la performance du modèle est estimée en calculant la moyenne des résultats obtenus lors de ces $k$ validations.L'objectif est d'utiliser toutes les données à la fois pour l'entraînement et pour la validation, ce qui permet d'obtenir une évaluation plus précise et moins biaisée de la capacité du modèle à généraliser sur de nouvelles données. [13]

**B. Validation croisée imbriquée (Nested Cross-Validation)**

La validation croisée imbriquée est une méthode d'évaluation avancée utilisée pour optimiser les hyperparamètres d'un modèle tout en évitant les biais dans l'estimation de sa performance. Elle consiste en deux niveaux de validation croisée : une validation croisée externe et plusieurs validations croisées internes. Dans chaque fold de la validation croisée externe, l'ensemble de données est divisé en un ensemble d'entraînement et un ensemble de validation. L'ensemble d'entraînement est ensuite utilisé pour effectuer une validation croisée interne afin d'optimiser les hyperparamètres du modèle. Une fois les meilleurs hyperparamètres trouvés à l'intérieur de chaque fold, le modèle est entraîné sur l'ensemble d'entraînement et testé sur l'ensemble de validation externe. Ce processus est répété pour chaque fold externe, permettant ainsi d'obtenir une estimation fiable de la performance du modèle, sans risque de contamination des données entre l'entraînement et l'évaluation. La validation croisée imbriquée est particulièrement utile lorsque l'on souhaite évaluer un modèle et choisir ses hyperparamètres de manière précise, sans introduire de biais dans l'estimation de la performance.[13]

### 3.5.4 Évaluation des performances d'un modèle (Test)

La phase de test constitue l'étape finale du processus de développement d'un modèle d'apprentissage automatique. Elle consiste à examiner le comportement du modèle sur un jeu de données totalement indépendant, appelé **données de test** , qui n'a été utilisé ni pour l'entraînement ni pour la validation. Cette étape permet d'obtenir une indication fiable de la capacité du modèle à produire des prédictions pertinentes sur de nouvelles données, dans un contexte réel d'utilisation. Pour cela, on utilise des mesures quantitatives, appelées métriques d'évaluation, qui comparent les résultats prédits aux valeurs attendues. Dans ce qui suit, nous présentons les principales métriques couramment employées pour l'évaluation des modèles de classification.

1. **Matrice de confusion :** La matrice de confusion est un outil d'évaluation de la performance d'un modèle, représenté sous forme de tableau. Elle aide les analystes de données à mieux comprendre les performances du modèle, en identifiant ses erreurs et ses points faibles. Son analyse permet d'affiner et d'améliorer le modèle.[14]

    - Vrai Positif (TP) : Le modèle prédit correctement la classe positive
    - Vrai Négatif (TN) : Le modèle identifie correctement la classe négative
    - Faux Positif (FP) : Le modèle classe à tort un élément comme positif
    - Faux Négatif (FN) : Le modèle ne détecte pas un élément réellement positif

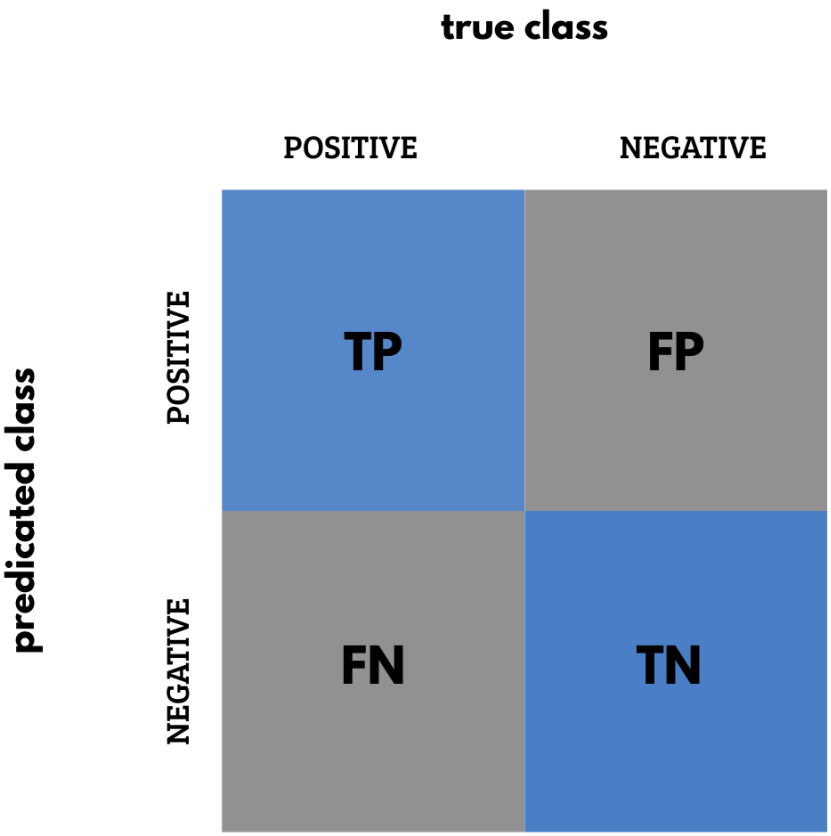



FIGURE 3.7 – Matrice de confusion

2. **Précision** Elle correspond au rapport entre le nombre de prédictions correctes et le nombre total de prédictions réalisées par le modèle , Elle se calcule à l'aide de la formule suivante [15] :

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. **Recall(Rappel)** Le recall mesure la proportion de cas positifs correctement identifiés par le modèle.Autrement dit, il correspond au rapport entre le nombre de vrais positifs prédits et le nombre total d'exemples positifs (vrais positifs + faux négatifs). Mathématiquement, il est défini par la formule suivante :

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. **Le score F1** Le score F1 représente la moyenne harmonique entre la précision et le rappel, offrant un compromis entre ces deux métriques. Il est particulièrement pertinent lorsque l'on recherche une seule mesure pour évaluer les performances du modèle, notamment en cas de déséquilibre des classes [14] . Mathématiquement, il est défini par la formule

suivante :

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

5. **La courbe ROC**

   La courbe ROC illustre la relation entre deux métriques essentielles : la sensibilité (ou taux de vrais positifs, TP) et le taux de faux positifs, FP (complémentaire à la spécificité). Lorsque le seuil de décision du modèle varie, ces valeurs changent, formant une courbe qui met en évidence sa capacité à distinguer les classes.

   L'aire sous la courbe ROC, appelée AUC (Area Under the Curve), offre une mesure unique de la performance globale du modèle. Un classificateur parfait atteint une AUC de 1.0, tandis qu'une prédiction aléatoire correspond à une AUC de 0.5, représentée par une ligne diagonale [16] .

   La figure 3.8 illustre une courbe ROC typique, avec la zone sous la courbe et différents seuils de décision.

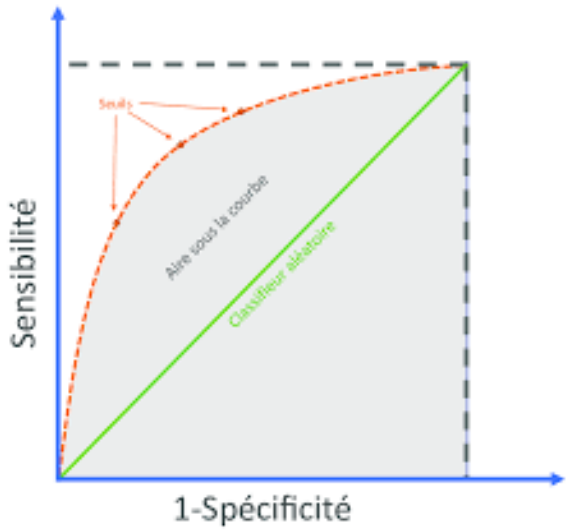



FIGURE 3.8 – Représentation graphique d'une courbe ROC

## 3.6 Sur-ajustement et Sous-ajustement

Lors du développement d'un modèle d'apprentissage automatique, il est essentiel de trouver un bon compromis entre complexité et performance. Un modèle peut en effet être trop complexe et s'adapter excessivement aux données d'entraînement, ou au contraire être trop simple pour en extraire des informations utiles. Dans cette section, nous présentons ces deux cas extrêmes, appelés respectivement sur-ajustement et sous-ajustement, ainsi que leurs impacts sur la qualité des prédictions.

### 3.6.1 Sur-ajustement (Overfitting)

Le sur-ajustement se produit lorsqu'un modèle s'ajuste trop précisément aux données d'entraînement, capturant même le bruit ou les détails non pertinents, ce qui nuit à sa capacité à généraliser sur de nouvelles données.

### 3.6.2 Sous-ajustement (Underfitting)

Le sous-ajustement se produit lorsqu'un modèle est trop simple pour capturer les relations importantes dans les données d'entraînement, ce qui le rend moins performant sur les données d'entraînement et de test.

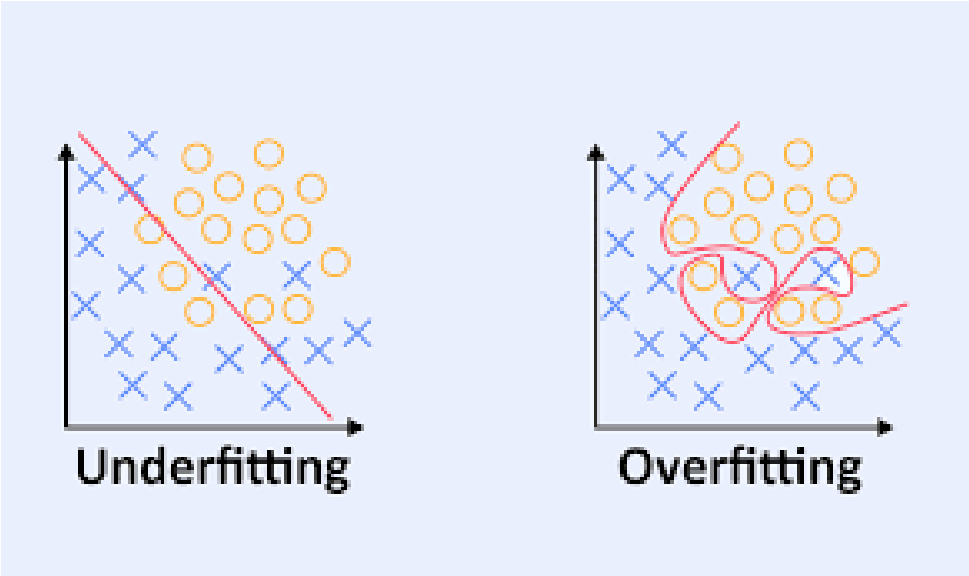



FIGURE 3.9 – Sous-ajustement et Sur-ajustement

## 3.7 L'optimisation des performances d'un modèle

L'optimisation des performances d'un modèle d'apprentissage automatique vise à améliorer sa capacité à effectuer des prédictions précises et fiables. Cela implique l'ajustement minutieux des hyperparamètres du modèle, tels que le taux d'apprentissage, la taille des couches dans les réseaux de neurones, ou d'autres configurations spécifiques, afin de maximiser les performances. Des techniques d'évaluation, comme la validation croisée, sont utilisées pour tester différentes combinaisons d'hyperparamètres et identifier la configuration optimale. Avant de pouvoir ajuster correctement un modèle, il est essentiel de comprendre la différence entre les paramètres qu'il apprend automatiquement et les hyperparamètres que l'on doit définir manuellement. Cette distinction permet d'agir efficacement sur les leviers d'optimisation.

### 3.7.1 Paramètres

Ce sont les variables internes que le modèle apprend directement à partir des données d'entraînement. Pendant l'entraînement, ces paramètres sont ajustés pour minimiser l'erreur entre les prédictions du modèle et les valeurs réelles.

### 3.7.2 Hyperparamètres

Contrairement aux paramètres, les hyperparamètres sont des configurations externes définies avant l'entraînement du modèle. Ils régissent le processus d'apprentissage et la structure du modèle. Des exemples incluent le taux d'apprentissage, le nombre de couches et de neurones dans un réseau neuronal, la profondeur d'un arbre de décision ou le nombre de clusters dans une méthode de regroupement. Le choix approprié des hyperparamètres est crucial, car il peut influencer la capacité du modèle à généraliser et à éviter le surajustement ou le sous-ajustement.

### 3.7.3 Techniques de recherche d'hyperparamètres

La recherche d'hyperparamètres consiste à identifier l'ensemble optimal d'hyperparamètres qui maximise les performances du modèle tout en minimisant la perte et le sur-ajustement. Parmi les méthodes les plus couramment utilisées , nous citons :

#### A. La recherche par grille

Cette technique consiste à définir une grille de valeurs possibles pour chaque hyperparamètre et à tester exhaustivement toutes les combinaisons. Bien que simple, elle peut être coûteuse en temps de calcul, surtout lorsque le nombre d'hyperparamètres et leurs valeurs possibles sont élevés [17] .

#### B. La recherche aléatoire

Au lieu d'explorer toutes les combinaisons, cette méthode sélectionne aléatoirement un sous-ensemble d'hyperparamètres à tester. Cela peut être plus efficace que la recherche exhaustive, en particulier lorsque l'espace des hyperparamètres est vaste [17] .

#### C. L'optimisation bayésienne

Cette approche modélise la fonction de performance du modèle en fonction des hyperparamètres et utilise ces informations pour guider la recherche vers les zones prometteuses de l'espace des hyperparamètres. Elle est souvent plus efficace que les méthodes précédentes, car elle nécessite moins d'évaluations du modèle pour trouver des configurations performantes.[18]

## 3.8 Travaux traitant la prédiction du churn

Afin de mieux comprendre les approches existantes dans le domaine de la prédiction de l'attrition client, notamment dans le secteur des télécommunications, cette section présente une sélection d'études récentes qui illustrent les méthodologies adoptées, les types de données exploitées, ainsi que les algorithmes d'apprentissage automatique utilisés et leurs performances respectives , ainsi que les résultats obtenus en termes de performance prédictive .

1. L'étude de Lackeshwar Bachan et Tarek Gaber (2021) s'intéresse à la prédiction du churn des clients dans l'industrie des fournisseurs d'accès Internet (ISP) des pays en développement, en se concentrant sur Trinidad et Tobago. La recherche explore l'application de trois modèles d'apprentissage automatique, à savoir l'arbre de décision, la régression logistique et la machine à vecteurs de support (SVM), pour prédire si un client quittera son fournisseur de services. Le jeu de données utilisé comprend 16 attributs et 31 728 entités, dont 27 166 non-churners et 4 562 churners. L'étude suit plusieurs étapes, notamment la préparation et le prétraitement des données, avec l'utilisation de techniques comme l'encodage one-hot pour les variables catégorielles et la corrélation de Pearson pour la sélection des attributs. Les résultats montrent que l'arbre de décision obtient la meilleure performance avec une précision de 89,5 %, un AUC de 81,9 % et un score F1 de 86,9 %, surpassant ainsi la régression logistique et la SVM. Les auteurs concluent que l'arbre de décision est le modèle le mieux adapté pour prédire le churn des clients dans ce contexte particulier, et suggèrent que des études comparatives dans d'autres pays pourraient apporter des éclairages intéressants sur les comportements des consommateurs [19].

2. Dans leur étude intitulée « Customer Churning Analysis Using Machine Learning Algorithms » (2023), Prabadevi, Shalini et Kavitha ont mené une comparaison entre quatre algorithmes d'apprentissage automatique visant à prédire le churn dans le secteur des télécommunications. Les auteurs ont utilisé un jeu de données issu de Kaggle, contenant 7043 instances, qu'ils ont divisé en 70 % pour l'entraînement et 30 % pour les tests. L'évaluation des performances a été effectuée à l'aide de plusieurs métriques, dont la courbe ROC et l'AUC. Les résultats ont montré que le modèle Gradient Booster obtenait les meilleures performances avec une AUC de 0,84, tandis que le KNN s'est révélé le moins performant, avec une AUC de 0,781. Les auteurs soulignent néanmoins que l'optimisation des hyperparamètres à l'aide de Grid Search CV s'est avérée coûteuse en temps, et que Randomized Search CV, bien que plus rapide, ne garantit pas forcément une configuration optimale. En conclusion, ils recommandent que les recherches futures portent davantage sur le prétraitement des données et l'optimisation des hyperparamètres afin d'améliorer la performance des modèles [20]

# Conclusion

Ce chapitre a permis de poser les bases théoriques nécessaires à la compréhension des méthodes utilisées pour la prédiction du churn. Nous avons présenté les différents types d'apprentissage automatique, les algorithmes les plus couramment employés dans ce contexte, ainsi que les critères d'évaluation permettant de juger leur efficacité. Une revue des travaux antérieurs a également permis de mettre en lumière les approches les plus pertinentes adoptées dans le secteur des télécommunications.

Ces éléments fournissent un socle solide pour la suite de ce mémoire. Le prochain chapitre sera consacré à la phase de conception de notre système de prédiction, dans laquelle nous traduirons ces concepts en une architecture fonctionnelle .

# Chapitre 4

# Conception

## 4.1 Introduction

Ce chapitre est dédié à la phase de conception du système, une étape essentielle pour structurer la solution avant sa mise en œuvre. Après avoir abordé les notions théoriques sur la prédiction du churn, nous adoptons ici la méthode Unified Process (UP). Le chapitre expose les principes du processus UP, ses phases, ainsi que son application dans notre projet. Il se conclut par la modélisation UML du système à travers plusieurs diagrammes, afin de représenter les interactions, les fonctionnalités et la structure globale.

## 4.2 Méthodologie de conception : Unified Process (UP)

Le Unified Process (UP) est une méthodologie de développement logiciel utilisée pour organiser et structurer les différentes étapes d'un projet, depuis l'analyse des besoins jusqu'à la livraison du produit final. Elle fournit un cadre général pour la planification, la modélisation, le développement et le suivi des projets logiciels . Dans ce qui suit, nous allons présenter les principes clés ainsi que les différentes phases de cette méthode.

### 4.2.1 Principes clés

UP est fondée sur une approche itérative, structurée autour de l'architecture du système et guidée par les cas d'utilisation, avec pour objectif principal la réduction des risques.Il s'agit d'un cadre méthodologique flexible, capable de s'adapter à une grande variété de systèmes logiciels, de domaines d'application, de contextes organisationnels, de niveaux de compétence et de tailles d'entreprise.[21]

#### Le Unified Process est itératif

L'approche itérative du Unified Process repose sur la répétition contrôlée d'une séquence d'activités ou de traitements. Chaque itération correspond à une version partielle mais fonctionnelle du système, permettant d'intégrer progressivement de nouvelles fonctionnalités tout en affinant celles déjà existantes. Cette répétition se poursuit jusqu'à ce qu'un objectif précis soit atteint ou qu'une condition de validation soit remplie, favorisant ainsi une meilleure maîtrise du projet et une détection précoce des erreurs.[21]

#### Le Unified Process est centré sur l'architecture

Selon Philippe Kruchten (IEEE, 1995), le processus unifié met l'accent sur la définition d'une architecture logicielle robuste dès les premières phases du projet. Cette architecture constitue le socle du développement, assurant à la fois la stabilité technique et la cohérence du système tout au long du cycle de vie. Elle est élaborée à partir de plusieurs vues complémentaires, permettant de structurer le système de manière efficace et durable.[21]

#### Le Unified Process est piloté par les cas d'utilisation

Le processus unifié place les besoins des utilisateurs au cœur du développement. À travers les cas d'utilisation d'UML, les exigences fonctionnelles du système sont identifiées, modélisées et structurées du point de vue de l'utilisateur. Ces cas d'utilisation servent de base à l'analyse, à la conception, aux tests, et même à la documentation, assurant ainsi que le système développé réponde effectivement aux attentes du client.[21]

### 4.2.2 Cycle de vie de Processus Unifié

Le Unified Process organise le développement logiciel selon deux axes complémentaires (voir figure 4.1) :

- **L'axe vertical** : il illustre la dimension statique du processus. Il regroupe les activités selon leur nature (par exemple : modélisation, conception, implémentation, tests, etc.) et met en évidence les composants, les processus, les artefacts produits ainsi que les rôles des différents intervenants.
- **L'axe horizontal** : il représente la dimension dynamique. Il retrace l'évolution du projet dans le temps, à travers les différentes phases du cycle de vie (inception, élaboration, construction, transition), les itérations successives et les jalons clés à atteindre.

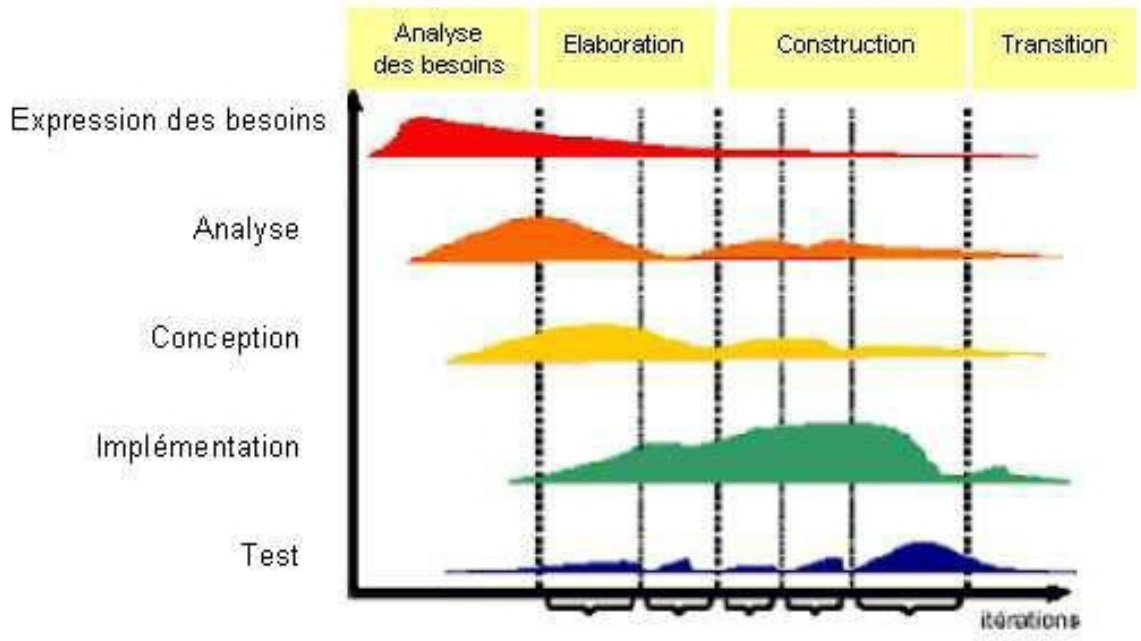



FIGURE 4.1 – Représentation des deux axes du Unified Process [21]

### 4.2.3 Les phases du Unified Process

Le Unified Process (UP) est structuré en quatre grandes phases qui se succèdent tout au long du cycle de vie du projet. Ces phases représentent (présentées dans la figure 4.1, elles constituent l'axe horizontal du processus). Elles permettent une progression itérative et incrémentale du développement, avec des objectifs bien définis pour chacune.

- **Inception (Analyse des besoins)**

Cette phase initiale vise à définir la portée du projet et à valider sa faisabilité. Elle se concentre sur l'identification des besoins essentiels, la définition d'une architecture générale, l'évaluation des risques majeurs ainsi que l'estimation des coûts et des délais. Elle répond notamment aux questions suivantes : Que va faire le système ? Quels services offrira-t-il ? Quelle sera son architecture cible ? Quels moyens faut-il mobiliser ?

- **Élaboration**

La phase d'élaboration approfondit les éléments recueillis durant l'inception pour aboutir à une spécification détaillée de la solution à mettre en œuvre. Elle permet de concevoir une architecture de référence stable, de préciser les cas d'utilisation, d'évaluer les risques techniques et d'établir un plan réaliste pour la suite du projet. Elle sert de fondation technique et organisationnelle.

- **Construction**

Il s'agit de la phase où le système est effectivement développé. L'architecture de référence est transformée en une version opérationnelle du logiciel. On implémente les cas d'utilisation, on intègre les composants, on effectue les tests nécessaires et on construit le produit final à livrer.

C'est la phase de réalisation concrète du système.

- **Transition**

Le produit est livré aux utilisateurs finaux. Cette phase inclut la validation en conditions réelles, la formation des utilisateurs, la correction des anomalies, la mise en production ainsi que la mise en place du support. Elle vise à assurer une adoption efficace du système par les utilisateurs cibles.

## 4.3 Application de la méthode UP dans le cadre de notre projet

Afin d'organiser efficacement les différentes étapes de notre projet, nous avons choisi d'adopter la méthode Unified Process (UP). Cette méthode itérative et incrémentale a été retenue pour sa simplicité de mise en œuvre, sa souplesse d'adaptation au contexte académique, et surtout parce qu'elle ne nécessite pas un grand effectif d'équipe, ce qui correspond parfaitement à notre cas. UP permet une structuration progressive du projet à travers quatre phases essentielles. Dans ce qui suit, nous présentons de manière détaillée l'application concrète de chacune de ces phases à notre projet de prédiction du churn .

| Phase | Étapes réalisées | Livrables produits | Durée estimée |
|---|---|---|---|
| **Inception** | — Compréhension du problème du churn chez Algérie Télécom<br>— Analyse des besoins des utilisateurs (admin, spécialiste)<br>— Définition des objectifs du projet (prédiction, visualisation, segmentation)<br>— Délimitation du périmètre (données, fonctionnalités principales) | — Cahier des charges fonctionnel simplifié<br>— Identification des acteurs et rôles<br>— Objectifs du projet validés | 2 semaines |
| **Élaboration** | — Collecte du dataset (Excel)<br>— Étude des variables (`OFFER_NAME`, `Mois_De_Contrat`, ...)<br>— Modélisation UML (cas d'utilisation, séquence, classes)<br>— Choix technologiques : Python/Flask, React/Redux, etc. | — Dataset nettoyé<br>— Diagrammes UML (cas d'utilisation, séquence, classes)<br>— Stack technique validée<br>— Planification du développement | 3 semaines |
| **Construction** | — Prétraitement : gestion des valeurs manquantes, encodage, etc.<br>— Équilibrage des classes avec SMOTE<br>— Entraînement et évaluation des modèles (Random Forest, XGBoost, ...)<br>— Développement de l'API Flask (upload, prédiction, statistiques)<br>— Développement du frontend React avec Redux | — Modèle ML sauvegardé (.pkl)<br>— API Flask fonctionnelle<br>— Interface utilisateur React opérationnelle<br>— Visualisations intégrées | 4 semaines |
| **Transition** | — Tests fonctionnels (chargement CSV, réponses API, affichage)<br>— Gestion des cas d'erreur<br>— Préparation de la démonstration (jeu de données test, scénarios)<br>— Rédaction du rapport et des supports de soutenance | — Application testée<br>— Rapport final<br>— Présentation PowerPoint | 2 semaines |

TABLE 4.1 – Application des phases de la méthode UP dans le projet

## 4.4 Modélisation UML

Afin de mieux comprendre et formaliser la structure et le fonctionnement de notre application, nous avons utilisé le langage UML pour représenter les différents aspects du système, tels que les cas d'utilisation, les interactions entre les composants (Les diagrammes de séquence) et l'organisation des classes.

### 4.4.1 Présentation du langage UML

Le langage UML (Unified Modeling Language) est un langage de modélisation visuelle standardisé, utilisé pour représenter la structure et le comportement des systèmes logiciels, notamment orientés objet. Il se compose de plusieurs types de diagrammes permettant de décrire les éléments du système, leurs interactions et leurs fonctionnalités [22] .

### 4.4.2 Diagramme de cas d'utilisation

Un diagramme de cas d'utilisation est une représentation graphique d'une ou plusieurs fonctionnalités spécifiques d'un système. Il illustre la manière dont ces fonctionnalités sont liées entre elles et identifie les acteurs internes ou externes qui interagissent avec le système pour les activer ou en bénéficier.[22]

Dans ce qui suit, nous illustrons les principaux acteurs du système ainsi que les cas d'utilisation associés, afin de mettre en évidence les différentes interactions entre les utilisateurs et les fonctionnalités offertes par l'application.

**Les acteurs :**

| Acteur | Cas d'utilisation |
|---|---|
| **Spécialiste commercial** | — Consulter la liste des clients à risque de churn.<br>— Télécharger la liste des clients à risque.<br>— Accéder aux statistiques du tableau de bord :<br>  — Taux de churn prédit<br>  — Niveau de satisfaction client<br>  — Consommation mensuelle en Go<br>  — Répartition du churn par type d'abonnement (avec filtres)<br>— Consulter la segmentation des clients :<br>  — Persuadables<br>  — Sure Thing<br>  — Lost Cause<br>  — Do Not Disturb<br>— Visualiser la réaction des clients aux campagnes marketing.<br>— Télécharger les résultats d'analyse marketing.<br>— Télécharger la liste des clients à cibler par les campagnes marketing. |
| **Administrateur** | — Tous les droits du spécialiste commercial.<br>— Mettre à jour le dataset (ajouter, supprimer ou modifier les données).<br>— Gérer les utilisateurs :<br>  — Afficher la liste des utilisateurs.<br>  — Ajouter un utilisateur.<br>  — Modifier les informations d'un utilisateur.<br>  — Supprimer un utilisateur. |

TABLE 4.2 – Acteurs et cas d'utilisation associés

### 4.4.3 Diagramme de cas d'utilisation général

Le diagramme 4.2) représente les cas d'utilisation de notre tableau de bord en spécifiant les différentes fonctionnalités qu'il englobe suivant les rôles des utilisateurs

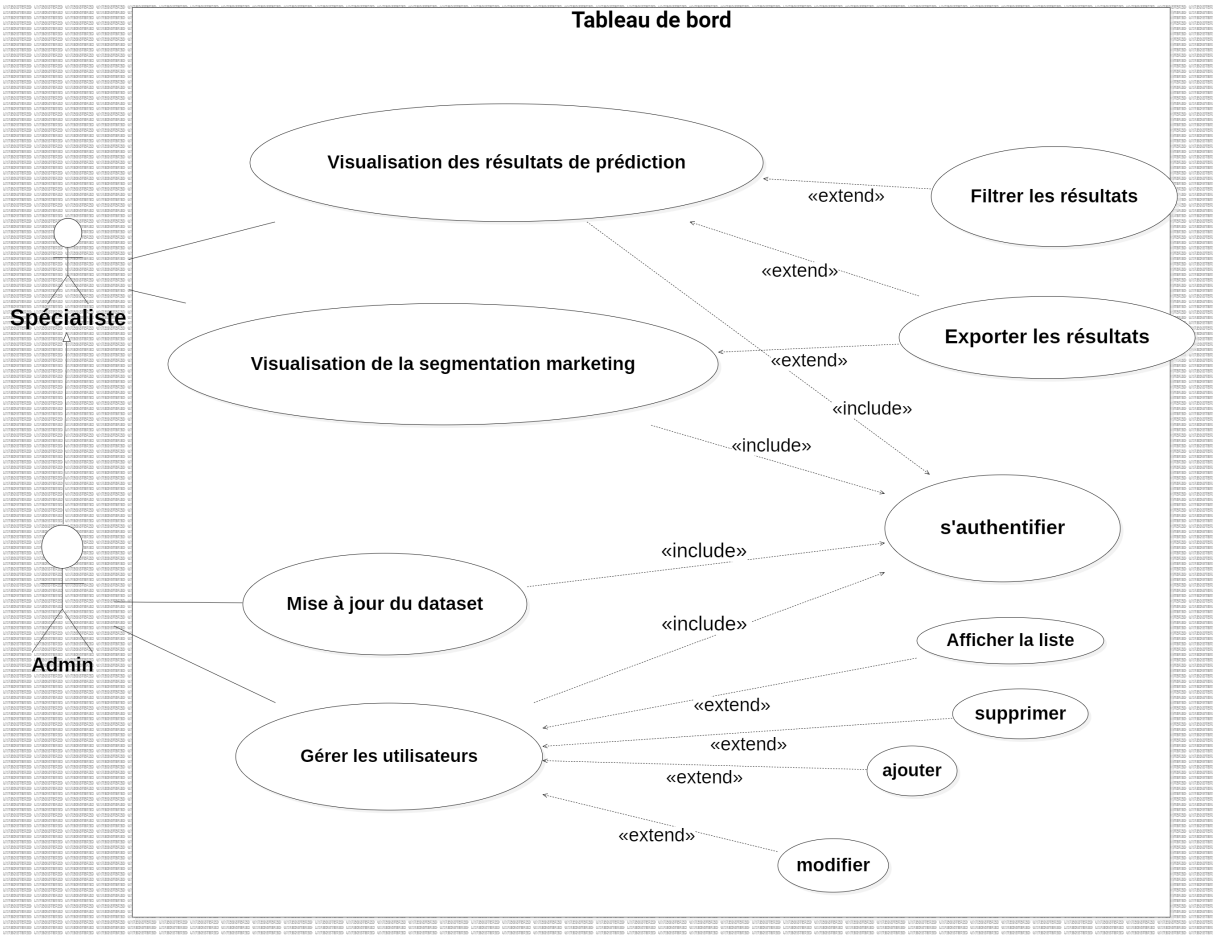



FIGURE 4.2 – Diagramme de cas d'utilisation général

### 4.4.4 Diagrammes de séquences

Le diagramme de séquence est un type de diagramme UML qui représente l'interaction entre différents objets au cours d'un scénario spécifique, en mettant en évidence l'ordre chronologique des messages échangés. Il permet de visualiser comment les objets collaborent entre eux pour accomplir une fonction ou un cas d'utilisation donné.[22]

#### Diagramme de séquence pour le cas d'utilisation "Authentification"

Le diagramme suivant décrit l'enchainement d'actions réalisées lors de l'authentification d'un utilisateur :

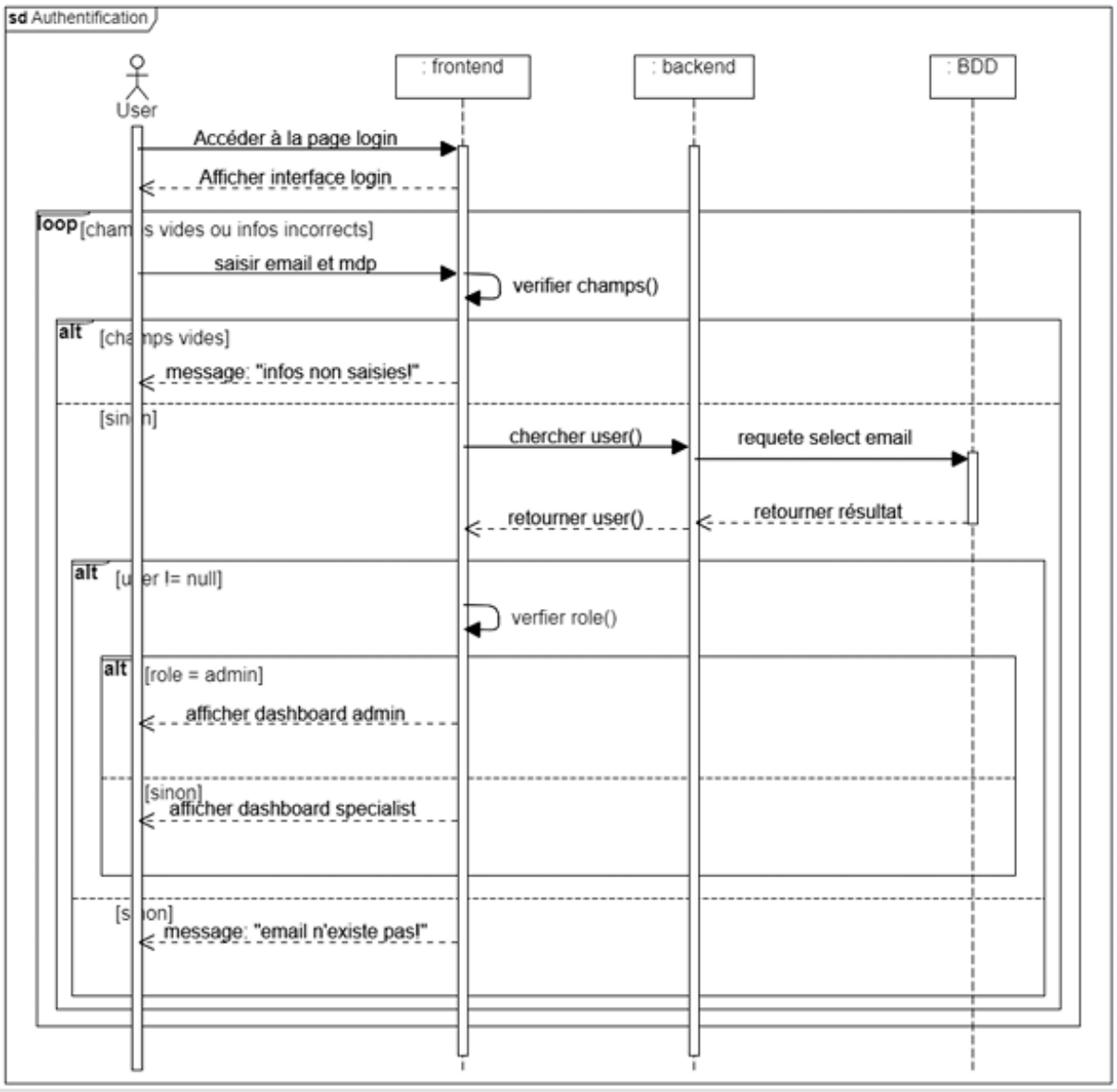



FIGURE 4.3 – Diagramme de séquence pour le cas d'utilisation "Authentification"

### 4.4.5 Diagramme de séquence pour le cas d'utilisation "Mise à jour d'un dataset"

Le diagramme suivant décrit l'enchaînement des actions effectuées lors de la mise à jour d'un ensemble de données par l'administrateur, en vue de son injection dans le modèle prédictif.

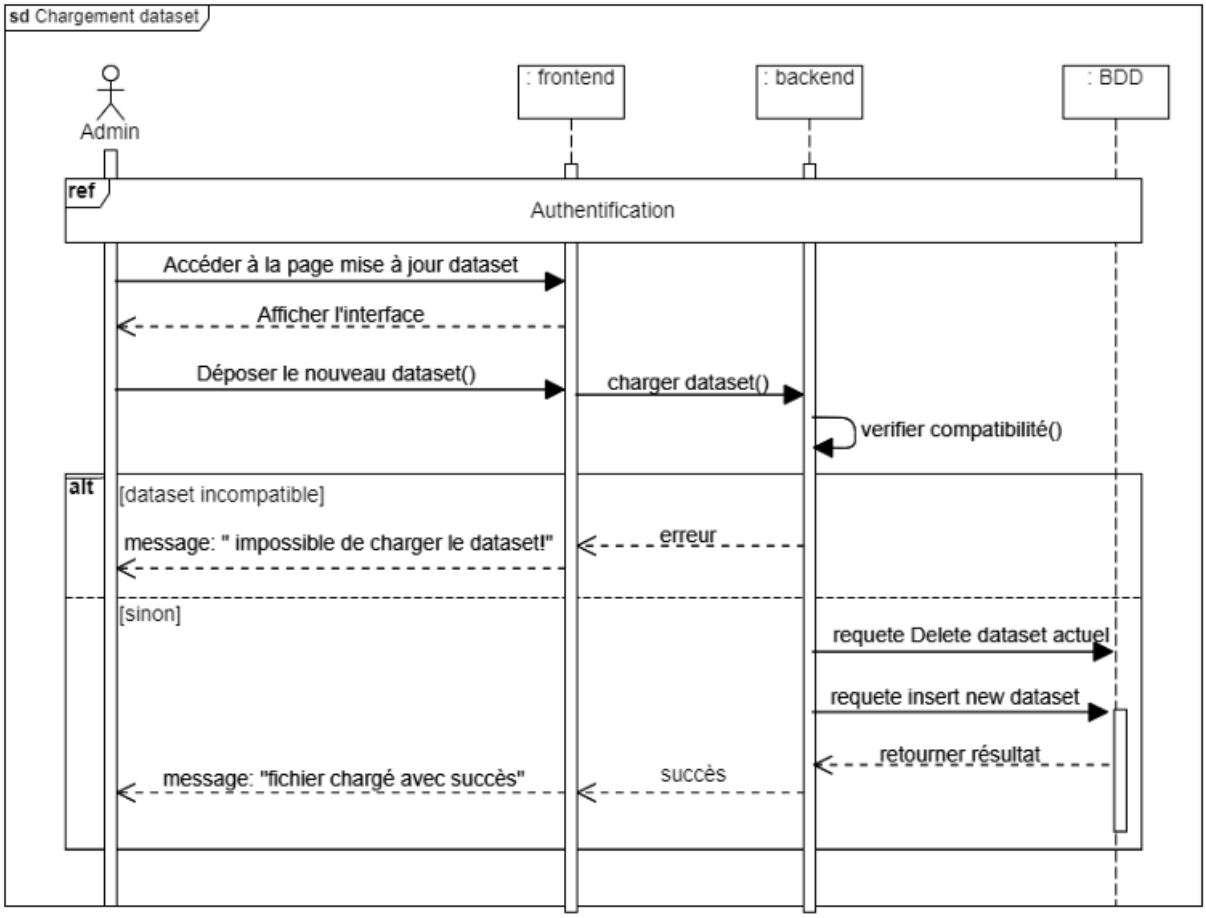



FIGURE 4.4 – Diagramme de séquence pour le cas d'utilisation "Mise à jour d'un dataset"

### 4.4.6 Diagramme de séquence "Ajout d'un utilisateur"

Le diagramme suivant décrit l'enchainement d'actions réalisées lors de l'ajout d'un nouveau utilisateur par l'administrateur :

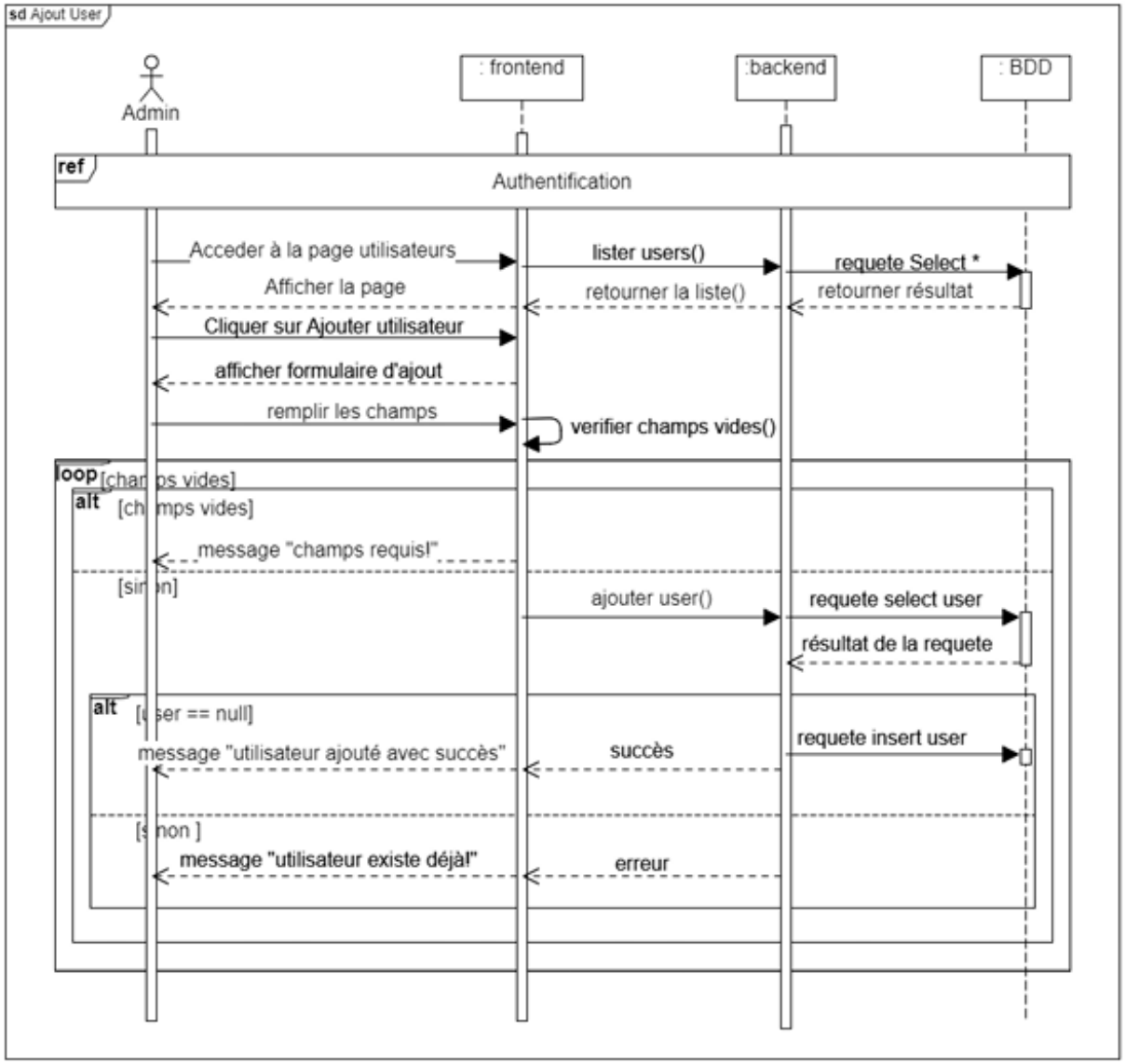



FIGURE 4.5 – Diagramme de séquence "Ajout d'un utilisateur"

### 4.4.7 Diagramme de classes

Le diagramme de classes est un type de diagramme structurel utilisé pour représenter la structure statique d'un système. Il montre les classes, leurs attributs, opérations, ainsi que les relations entre les objets. [22].

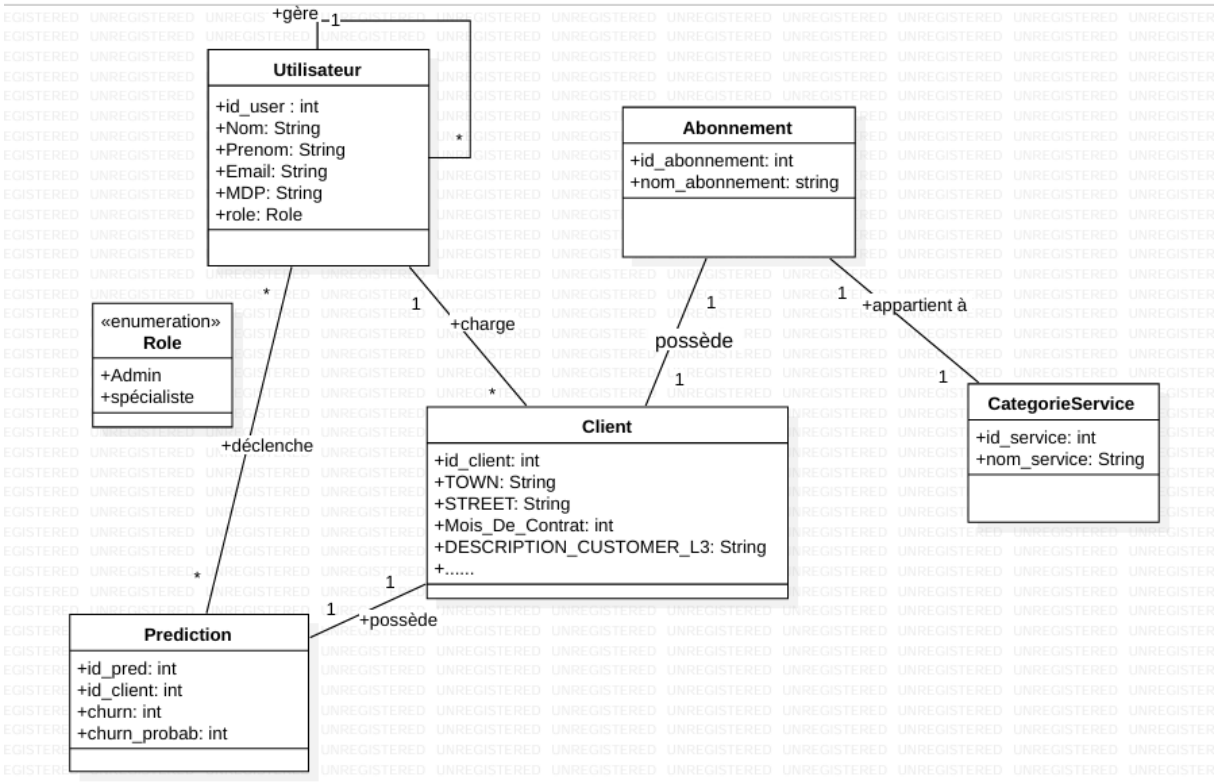



FIGURE 4.6 – Diagramme de classes

Ce diagramme de classes représente notre application web dans sa globalité en évoquant les classes essentielles ainsi que les relations qui les relient entre elles. Pour ce faire, nous allons représenter les différentes règles de gestion qui correspondent à chaque classe :

— Un utilisateur invoque la prédiction des clients. Il peut avoir l'un des deux rôles suivants : administrateur ou spécialiste. Le rôle est défini par une énumération (Role) et permet de restreindre ou étendre les fonctionnalités accessibles. Un utilisateur peut gérer plusieurs clients.

— Un client représente l'abonné d'Algérie Télécom. Il est identifié par un identifiant unique et dispose d'un ensemble important d'attributs décrivant ses caractéristiques. Parmi ceux-ci, on peut citer notamment la ville (TOWN), la rue (STREET), la durée de son contrat exprimée en mois (Mois_De_Contrat), ainsi qu'un champ descriptif (DESCRIPTION_CUSTOMER_L3) représentant son segment ou son type. Ces attributs ne constituent qu'un extrait représentatif des données disponibles sur chaque client. Un client est géré par un utilisateur, possède un seul abonnement Chaque client est ainsi associé à une seule prédiction à un moment donné.

— La prédiction contient les résultats générés par le modèle de machine learning. Elle regroupe l'identifiant de la prédiction, une référence au client concerné, une variable binaire indiquant s'il y a churn (churn), ainsi qu'une probabilité de churn (churn_probab) exprimée en pourcentage. Chaque prédiction est donc associée à un seul client.

— Un abonnement correspond à l'offre commerciale souscrite par un client. Il est caractérisé par un identifiant et un nom d'abonnement. Chaque client possède un seul abonnement, et chaque abonnement appartient à une seule catégorie de service.

— La catégorie de service regroupe les différentes offres commerciales sous des types plus larges (par exemple : ADSL, VDSL, Fibre, 4G LTE, etc.). Elle est définie par un identifiant de service et un nom. Une catégorie de service peut regrouper plusieurs abonnements.

Ce diagramme permet ainsi de visualiser l'architecture logique de l'application, en mettant en évidence les dépendances entre utilisateurs, clients, prédictions, abonnements et catégories de service. Il facilite la compréhension du fonctionnement global du système, notamment pour la gestion des prédictions de churn et la catégorisation des offres commerciales.

## 4.5 Conclusion

Ce chapitre a présenté la phase de conception de notre projet en s'appuyant sur la méthode Unified Process, choisie pour sa souplesse et sa structuration itérative. Nous avons détaillé les phases du processus, leur application à notre projet, ainsi que les diagrammes UML utilisés pour modéliser les interactions et la structure du système. Cette base conceptuelle prépare le terrain pour la phase suivante : la réalisation concrète de l'application et l'expérimentation des modèles prédictifs.

# Chapitre 5

# Réalisation et résultats

## 5.1 Introduction

CCe chapitre présente l'ensemble du processus de conception du modèle prédictif de churn ainsi que son intégration au sein d'une application web dédiée. Il se divise en deux grandes sections complémentaires : la première décrit les différentes étapes techniques de construction du modèle. La seconde section est consacrée à l'implémentation des principales fonctionnalités de l'interface utilisateur, illustrée par des captures d'écran représentatives de l'application développée.

## 5.2 Réalisation du modèle prédictif

Dans cette première partie, nous présentons le processus global de réalisation du modèle prédictif destiné à anticiper le churn des clients. Ce processus englobe plusieurs étapes successives, allant de lanalyse exploratoire des données à la sélection du modèle final. La figure 5.1 illustre de manière synthétique les principales étapes suivies dans ce cadre.

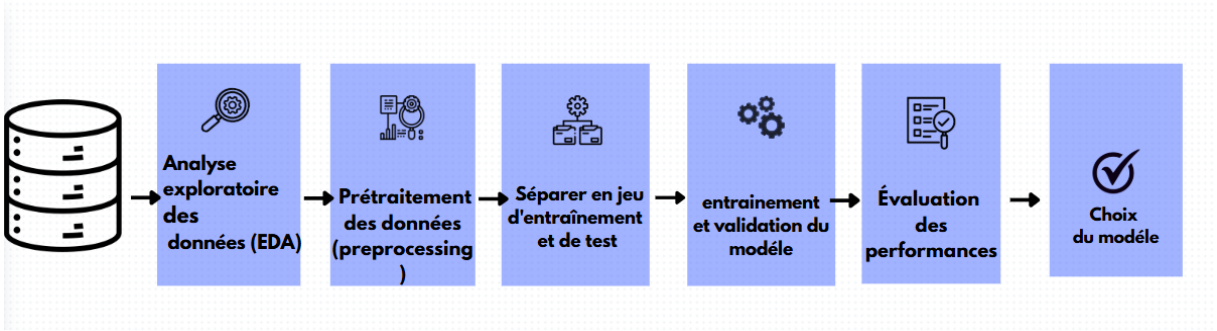



FIGURE 5.1 – Schéma représentant les étapes de création du modèle prédictif

### 5.2.1 Analyse exploratoire des données (EDA)

Cette section présente un aperçu détaillé du dataset utilisé, incluant sa structure, la distribution des clients selon différentes caractéristiques, ainsi que les premières observations statistiques utiles à l'analyse du churn.

1. **Description initiale du jeu de données :** Le dataset utilisé dans cette étude a été collecté lors de notre stage effectué au sein de la Direction Opérationnelle Territoriale (DOT) d'Algérie Télécom à Béjaïa, entre les services commercial et technique. Il contient exactement 330 019 lignes et 39 colonnes, chacune représentant un enregistrement lié au comportement des clients de l'entreprise. Ces colonnes correspondent à divers attributs relatifs aux abonnements, à l'utilisation des services, ainsi qu'à des informations contractuelles permettant d'analyser les facteurs influençant le churn.La figure 5.2 présente un extrait représentatif de ce dataset, illustrant la structure et le type d'informations disponibles.

| | CREATION_DATE | OFFER_NAME | STREET | TOWN | DESCRIPTION_CUSTOMER_L3 | TELECOM_TYPE | Contract_Month | Monthly_Consumed_Volume_GB | Daily_Connected |
|---|---|---|---|---|---|---|---|---|---|
| 186013 | 2020-12-17 15:17:38.999999866 | 10M'_ADSL_Prepaid_indiviual | BEJAIA, . Lycee technicum | BEJAIA | Residential | xDSL | 2.0 | 57.35 | |
| 250917 | 2021-09-25 13:50:51.000000115 | Pack Data_Voix 4G LTE (RÃ©sidentiel) | AMIZOUR, Vlage Ahmam | AMIZOUR | Residential | LTE | 32.0 | 136.25 | |
| 86932 | 2021-10-25 13:44:41.000000375 | PACK VOIP Migration TDM | BEJAIA, 56 LOGTS AADL BRANDY | BEJAIA | Officially agreed DGSN | VOIP | 34.0 | 62.06 | |
| 33814 | 2017-04-25 00:00:00.000000000 | Pack Internet et Voix 4G LTE (RÃ©sidentiel) | KHERRATA, KHERRATA..VILLAGE MEROUAHA. | KHERRATA | Residential | LTE | 3.0 | 66.44 | |
| 175346 | 2018-07-25 00:00:00.000000000 | Pack Internet 4G LTE (RÃ©sidentiel) | AMIZOUR, AMIZOUR..LOTISSEMENT MERDJ OUAMENE. | AMIZOUR | Residential | LTE | 72.0 | 28.69 | |
| 227373 | 2012-12-25 16:10:15.999999196 | 15M_ADSL_Prepaid_indiviual | AMIZOUR, ,E Cite 50 Logts LSP | AMIZOUR | Residential | xDSL | 26.0 | 41.25 | |
| 250780 | 2023-05-10 17:09:43.000000076 | PACK_Idoom_FTTH_Migration_MSAN_15M | AKBOU, CITE DES PINS | AKBOU | Residential | FTTx | 22.0 | 99.88 | |
| 19895 | 2016-02-15 00:00:00.000000000 | Pack Internet 4G LTE (RÃ©sidentiel) | AKOURMA, AMALOU..VILLAGE BENI DJEMHOUR. | AMALOU | Residential | LTE | 88.0 | 39.52 | |
| 162073 | 2025-01-16 15:52:16.000000078 | MIXTEAbonnement IDOOM FIXE 15Mbs et plus | BEJAIA LIBERTE, Cite Naceria | BEJAIA | Residential | PSTN | 5.0 | 64.82 | |
| 222873 | 2023-08-27 15:21:53.000000150 | Idoom VOIP PROMO 15M et plus | BEJAIA, n15 3200 LOGTS AADL IGHZER OUZARIF | BEJAIA | Residential | VOIP | 3.0 | 72.29 | |
| 87202 | 2023-05-15 15:34:58.999999699 | FTTc-b_15M | EL KSEUR, Ilot 12 | EL KSEUR | Residential | xDSL | 14.0 | 91.36 | |
| 306874 | 2016-04-06 00:00:00.000000000 | Pack Internet 4G LTE (RÃ©sidentiel) | GUENDOUZE, AIT R'ZINE..VILLAGE TIGHILT. | AIT R'ZINE | Residential | LTE | 96.0 | 95.86 | |

FIGURE 5.2 – Extrait représentatif du dataset utilisé

Le dataset comporte un total de 39 colonnes. Le tableau ci-dessous présente la signification de quelques-unes de ces colonnes, afin de mieux comprendre les informations qu'elles contiennent.

TABLE 5.1 – Description des colonnes du dataset

| **Nom de la colonne** | **Description** |
|---|---|
| CREATION_DATE | Date de création du contrat. |
| OFFER_NAME | Nom de l'offre commerciale souscrite par le client. |

| Nom de la colonne | Description |
|---|---|
| STREET | Rue de résidence du client. |
| TOWN | Commune de résidence du client. |
| DESCRIPTION_CUSTOMER_D3 | Description du client selon une segmentation interne. |
| TELECOM_TYPE | Type de service télécom souscrit (ADSL, FTTH, etc.). |
| Contract_Month | Durée du contrat en mois. |
| Monthly_Consumed_Volume_GB | Volume moyen de données consommé par le client chaque mois (en Go). |
| Daily_Connected_Time | Temps moyen de connexion par jour. |
| Number_of_Recharges | Nombre total de recharges effectuées par le client. |
| Months_Without_Recharge | Nombre de mois sans recharge. |
| Monthly_Outages | Nombre moyen de pannes signalées par mois. |
| Average_Repair_Time | Temps moyen nécessaire pour réparer une panne. |
| Support_Calls | Nombre d'appels effectués vers le service client. |
| Online_Payment | Indique si le client a déjà utilisé le paiement en ligne. |
| MyIdoom_Usage | Utilisation de la plateforme MyIdoom. |
| Gender | Sexe du client. |
| Region_Score | Score attribué à la région de résidence du client. |
| Technology_Type | Type de technologie utilisée (Fibre, ADSL, VDSL, etc.). |
| Technology | Technologie d'accès du client (FTTH, LTE, FTTc. . .). |
| Churn | Variable cible indiquant si le client a résilié. |

2. **Répartition selon le churn :** L'étiquette Churn indique si un client a quitté l'entreprise ou non. Comme l'illustre la figure 5.3, la grande majorité des clients semble encore active (72,64 %), tandis qu'une fraction plus réduite a effectivement résilié son contrat (27,36 %). Cette distribution est significative, car elle révèle un déséquilibre de classes qu'il convient de prendre en compte lors de l'entraînement des modèles de classification. En effet, la prédominance de la classe des clients non churners peut fausser les performances du modèle prédictif si l'équilibre entre les classes n'est pas rigoureusement maintenu.

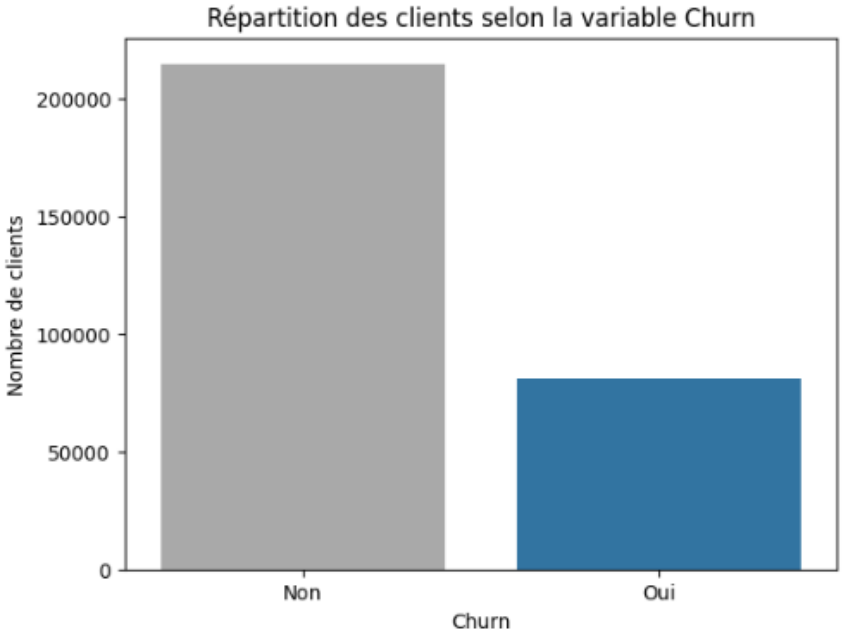



FIGURE 5.3 – Répartition des clients selon le churn

3. **Distribution des clients par région :** La Figure 5.4 illustre la répartition des clients d'Algérie Télécom à travers les différentes communes de la wilaya de Béjaïa. Cette représentation met en évidence les zones ayant une forte densité d'abonnés, avec en tête la commune de Béjaïa, suivie d'Akbou, El Kseur, Amizour, et d'autres localités telles que Tazmalt, Oued Ghir ou encore Ifri Ouzellaguen. Afin de garantir une meilleure lisibilité du graphique, seules les communes ayant un nombre d'abonnés supérieur à 3 500 ont été retenues dans cette visualisation.

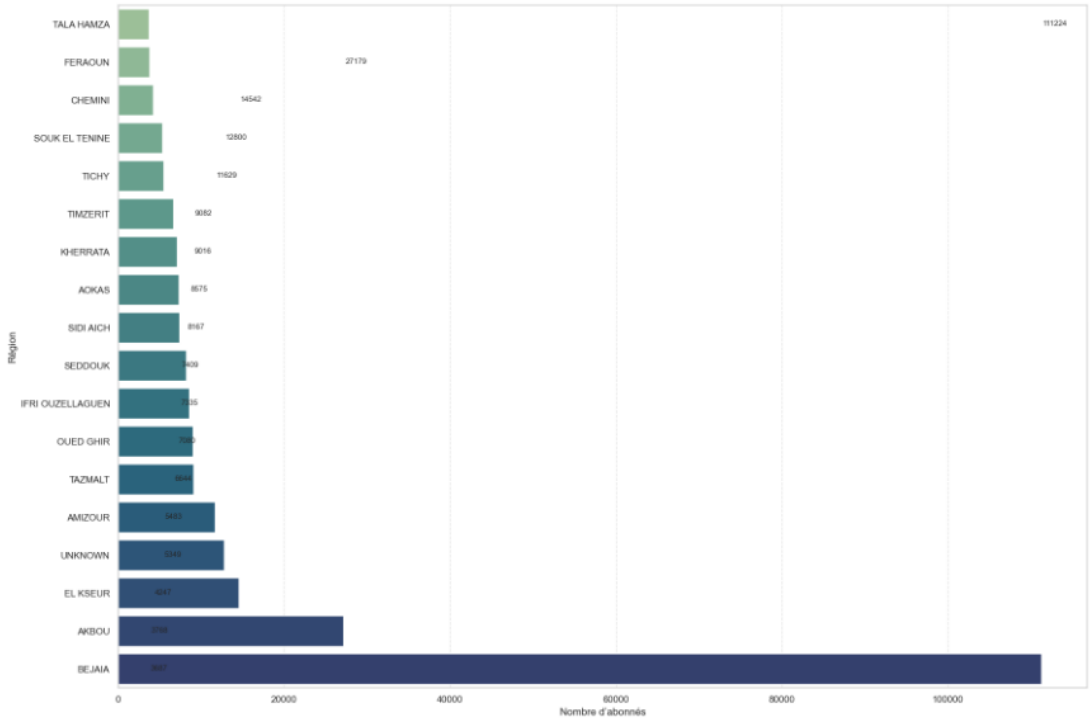



FIGURE 5.4 – Répartition des clients d'Algérie Télécom par région (communes de la wilaya de Béjaïa)

4. **corrélation :** L'analyse de corrélation constitue une étape essentielle de l'analyse exploratoire des données. Elle permet d'évaluer la relation statistique entre les variables numériques et de mesurer leur degré de dépendance. Dans le cadre de cette étude, une matrice de corrélation a été générée afin d'identifier les variables numériques ayant le plus d'influence sur la variable cible churn. La figure 5.5 illustre ces relations :

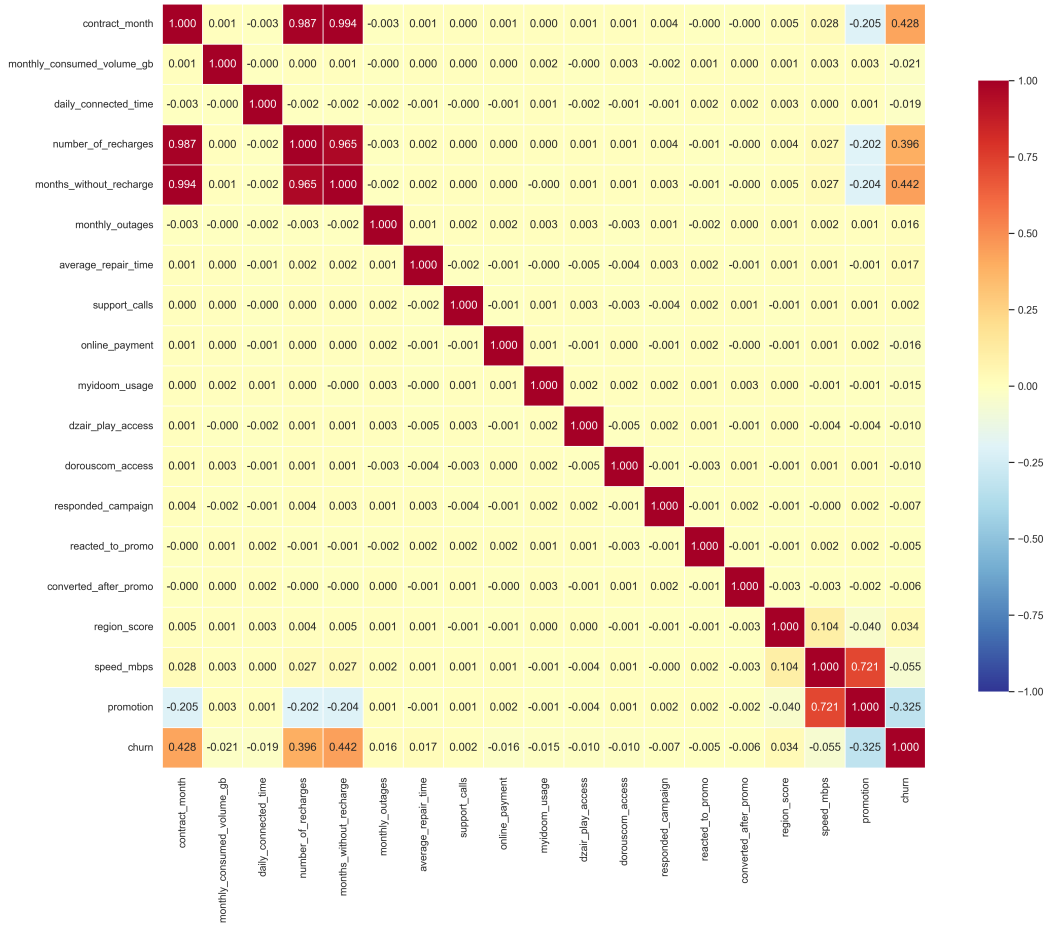



FIGURE 5.5 – matrice de corrélation

Les valeurs de corrélation varient entre –1 et +1. Une valeur proche de +1 indique une forte relation linéaire positive, tandis qu'une valeur proche de –1 traduit une forte relation linéaire négative. Une valeur proche de 0 signifie une absence ou une très faible relation linéaire entre les deux variables.

### 5.2.2 Prétraitement des données

Avant d'entraîner un modèle de machine learning, il est essentiel de préparer correctement les données afin d'assurer leur qualité et leur cohérence. Cette étape de prétraitement comprend plusieurs opérations visant à nettoyer, structurer et enrichir le jeu de données, afin d'en extraire les informations les plus pertinentes pour la prédiction du churn.

1. **Suppression des doublons :** Un doublon dans un dataset signifie qu'une même ligne apparaît plus d'une fois, exactement à l'identique dans toutes les colonnes. Dans notre cas, nous avons détecté 1289 doublons, que nous avons ensuite supprimés afin d'éviter toute redondance susceptible de fausser les résultats ou de biaiser la prédiction.
2. **Gestion des valeurs manquantes :** Les valeurs manquantes (ou missing values) sont des cellules vides ou absentes dans un jeu de données. Elles peuvent être représentées

par NaN, (Not a Number), null, `None`, etc. Nous avons détecté la présence de valeurs manquantes dans 4 colonnes spécifiques, ce qui est illustré dans la Figure 5.6. Ces valeurs ont été supprimées afin de garantir la qualité des données et d'éviter toute perturbation lors de la phase d'entraînement des modèles de prédiction.

```
contract_month                9330
monthly_consumed_volume_gb    7775
number_of_recharges           4509
churn                        15239
dtype: int64
```

FIGURE 5.6 – Valeurs manquantes

3. **Ingénierie des caractéristiques (Feature Engineering) :**

   Le Feature Engineering constitue une étape essentielle du prétraitement des données. Elle consiste à créer de nouvelles variables (ou caractéristiques) à partir de celles déjà existantes, dans le but de mieux représenter l'information pertinente pour les modèles de machine learning. Dans notre cas, la colonne `OFFER_NAME` contenait des chaînes de caractères complexes regroupant plusieurs informations imbriquées. Un aperçu de cette colonne est présenté dans la Figure 5.7, illustrant la nécessité de simplifier et structurer ces informations.

   À l'aide de techniques de traitement textuel (comme la recherche de mots-clés ou l'utilisation d'expressions régulières), nous avons extrait plusieurs colonnes dérivées, notamment :

   — `TECHNOLOGIE` : indique la technologie utilisée (ex. : ADSL, VDSL, FTTH, 4G LTE, PSTN, etc.) ;
   — `DEBIT` : représente le débit de l'offre (ex. : 10M, 15M, 100M, 1.2 Gbps, etc.) .
   — `TYPE_ABONNEMENT` : précise la nature de l'abonnement (ex. : prépayé, postpayé, résidentiel, professionnel, etc.) .
   — `MIGRATION` : indique si l'offre est liée à une opération de migration technique (ex. : Migration MSAN ou TDM).

   Ces nouvelles variables structurées permettent une meilleure exploitation des données dans les modèles prédictifs, en rendant explicites des informations auparavant implicites.

| | offer_name |
|---|---|
| 245853 | MIXTEAbonnement IDOOM FIXE |
| 119265 | PACK Idoom Fibre VOIP 10M |
| 65332 | Pack Data_Voix 4G LTE (RÃ©sidentiel) |
| 179673 | 10M'_ADSL_Prepaid_indiviual |
| 141334 | Pack Internet 4G LTE (RÃ©sidentiel) |
| 16253 | 15M_ADSL_Prepaid_indiviual |
| 305044 | 10M'_ADSL_Prepaid_indiviual |
| 111314 | Pack Data_Voix 4G LTE (RÃ©sidentiel) |
| 26964 | MIXTEAbonnement IDOOM FIXE |
| 102555 | IDOOM Fibre VOIP 3Mois_ONT PROMO |

FIGURE 5.7 – Extrait de la colonne `OFFER_NAME`

4. **Encodage des Données :**

   Certaines colonnes de notre jeu de données, telles que `TOWN` et `DESCRIPTION_CUSTOMER_L3`, contenaient des valeurs textuelles (catégorielles) qui ne peuvent pas être directement exploitées par les algorithmes de machine learning. Pour les rendre interprétables, nous avons appliqué un encodage numérique en remplaçant chaque modalité par une valeur entière à l'aide de la méthode *LabelEncoder* [23] fournie par la bibliothèque `scikit-learn` 5.4. La Figure 5.8 présente un extrait des données avant transformation, tandis que la Figure 5.9 illustre le résultat après encodage, où chaque modalité textuelle a été convertie en valeur numérique.

| | town | description_customer_l3 |
|---|---|---|
| 0 | BENI MELIKECHE | Residential |
| 1 | BEJAIA | Residential |
| 2 | TICHY | Residential |
| 3 | BEJAIA | Residential |
| 4 | SIDI AICH | Residential |

FIGURE 5.8 – Extrait des données avant encodage

| town | description_customer_l3 |
|---|---|
| 39 | 77 |
| 33 | 77 |
| 173 | 77 |
| 33 | 77 |
| 151 | 77 |

FIGURE 5.9 – Extrait des données après encodage

5. **Équilibrage des Données**

   Notre dataset présentait un déséquilibre modéré entre les classes de la variable cible churn 2.Étant donné la taille importante du dataset initial (environ 300 000 lignes), nous avons opté pour une méthode de *sous-échantillonnage aléatoire* (*Random UnderSampling*) [24] , consistant à réduire aléatoirement le nombre d'échantillons de la classe majoritaire afin d'obtenir un équilibre avec la classe minoritaire. Après cette opération, la taille du jeu de données a diminué à environ 200 000 lignes. La Figure 5.10 illustre la nouvelle répartition de la variable `Churn`, où l'on observe un équilibre relatif entre les deux classes, permettant de garantir un apprentissage plus équitable pour les modèles de classification.

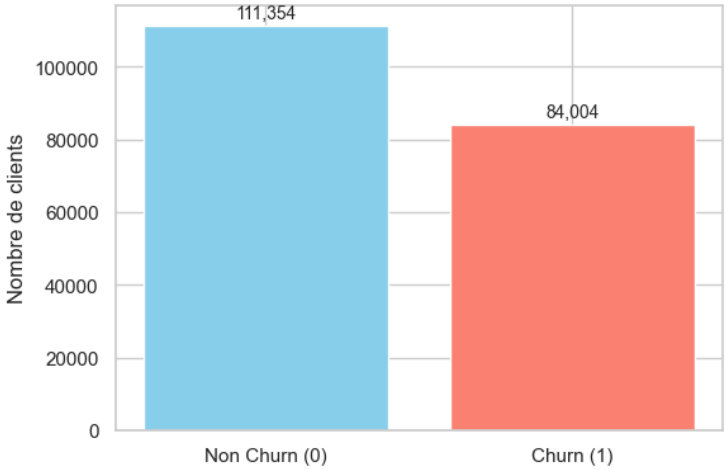



FIGURE 5.10 – Extrait de la colonne `OFFRE_NAME`

### 5.2.3 Construction de l'ensemble d'entraînement et de test

Dans un premier temps, nous avons divisé l'échantillon obtenu à l'issue de la phase de prétraitement en deux sous-ensembles : un ensemble d'entraînement représentant 70 % des données, et un ensemble de test représentant les 30 % restants. Chaque ensemble a ensuite été séparé en deux parties : les variables indépendantes (appelées `X`), et la variable cible `y` (dans notre cas, la variable `churn`). Cette séparation permet de préparer les données d'entrée et les étiquettes nécessaires à l'apprentissage supervisé.À l'issue de cette étape, nous disposons de quatre ensembles : `X_train`, `y_train`, `X_test` et `y_test`, qui seront utilisés pour entraîner et évaluer les différents algorithmes de classification.

### 5.2.4 Entraînement des modèles et choix des algorithmes

Après avoir préparé, divisé le jeu de données, nous avons entamé la phase d'entraînement des modèles de classification dans le but de prédire la probabilité de Churn des clients. Pour ce faire, nous avons sélectionné trois algorithmes de classification binaire, reconnus pour leur robustesse et leur efficacité dans des contextes similaires :

— Régression logistique

— Forêt aléatoire

— XGBoost

### 5.2.5 Ajustement des Hyperparamètres

Afin d'optimiser les performances de chaque modèle, nous avons recours à la technique de recherche en grille (GridSearchCV)[25], qui permet d'explorer systématiquement différentes combinaisons d'hyperparamètres propres à chaque algorithme. Cette méthode vise à identifier les configurations les plus performantes.

### 5.2.6 Évaluation des performances

Les modèles entraînés ont ensuite été évalués sur l'ensemble de test à l'aide de plusieurs métriques d'évaluation. Ces indicateurs permettent de mesurer la qualité des prédictions, notamment la capacité à identifier correctement les clients susceptibles de résilier leur abonnement.

— **Compratif des performances des modéles suivant les métriques "Precision , Recall , F1-score , Specificity , Accuracy " :** Le tableau 5.2 présente un résumé des performances des modèles évalués selon plusieurs indicateurs clés. Cette comparaison permet d'identifier le modèle le plus performant en fonction des critères suivants :

TABLE 5.2 – Comparaison réaliste des performances des modèles selon plusieurs métriques

| **Modèle** | **Précision (Precision)** | **Rappel (Recall)** | **F1-score** | **Spécificité** | **Exactitude (Accuracy)** |
|---|---|---|---|---|---|
| Logistic Regression | 0.7435 | 0.8780 | 0.8050 | 0.8650 | 0.8800 |
| Random Forest | 0.8750 | 0.9400 | 0.9060 | 0.9300 | 0.9250 |
| XGBoost | **0.9020** | **0.9600** | **0.9300** | **0.9500** | **0.9400** |

En observant les valeurs du tableau, on constate que le modèle **XGBoost** surpasse clairement les deux autres en obtenant les meilleures performances sur toutes les métriques, en particulier en termes de *recall* (0,9600) et de *F1-score* (0,9300), ce qui indique sa grande capacité à détecter efficacement les clients churners.

— **La courbe ROC (AUC) :** Le graphe 5.11 représente les résultats obtenues par la courbe ROC suivant les valeurs AUC :

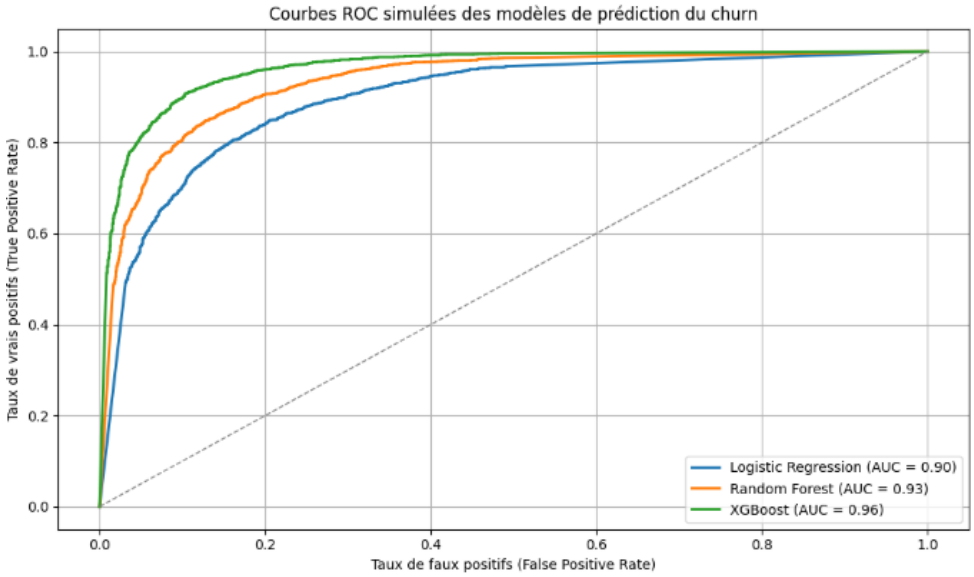



FIGURE 5.11 – Courbe ROC

La figure 5.11 montre que le modèle **XGBoost** présente la meilleure courbe ROC, indiquant une excellente capacité de classification. Toutefois, la différence de performance avec le modèle **Random Forest** reste minime, ce qui suggère que les deux modèles offrent des résultats comparables en termes de discrimination entre les clients churners et non churners.

— **Matrices de confusion :** La figure 5.12 présente les matrices de confusion pour les trois modèles. Chaque matrice affiche la répartition des prédictions en quatre catégories : les vrais négatifs (en vert), les faux positifs (en rouge), les faux négatifs (en jaune) et les vrais positifs (en bleu).

On observe que le modèle XGBoost fournit les meilleurs résultats en termes de précision globale, avec un nombre élevé de vrais positifs (19200) et peu de faux négatifs (800), ce qui est crucial dans un contexte de churn où l'identification correcte des clients susceptibles de partir est primordiale. De plus, XGBoost minimise les fausses alertes (seulement 870 faux positifs), ce qui en fait un choix optimal parmi les trois modèles évalués.

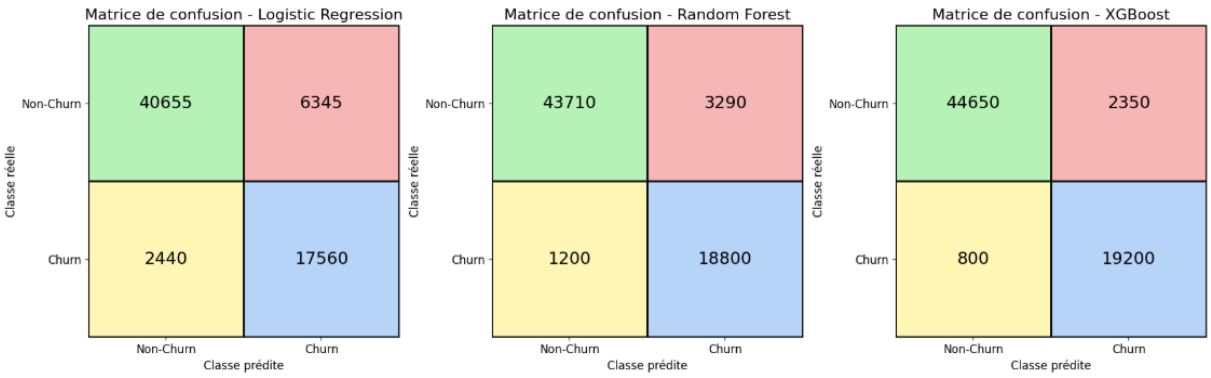



FIGURE 5.12 – Les matrices de confusion pour les trois modèles

### 5.2.7 Choix Final du Modèle

Le choix final du modèle s'est basé sur la comparaison de leurs performances respectives. XGBoost s'est démarqué par une meilleure accuracy, un rappel élevé et un bon score F1, ce qui en fait la solution retenue pour la prédiction du churn.Le modèle final, entraîné à l'aide de l'algorithme XGBoost, a été sauvegardé dans le fichier MON-MODEL-XGBOOST.pkl. Ce fichier est ensuite utilisé dans le backend Flask de notre application web afin d'effectuer des prédictions de churn à partir des données utilisateurs fournies via l'interface.

## 5.3 Réalisation du tableau de bord

Cette section décrit la mise en œuvre du tableau de bord interactif permettant de visualiser les résultats issus du modèle de prédiction, ainsi que d'explorer différentes dimensions liées au phénomène de churn. L'objectif est de fournir un outil accessible, dynamique et utile à la prise de décision pour l'équipe commerciale .

### 5.3.1 Architecture du tableau de bord

La figure ci-dessous illustre l'architecture de notre tableau de bord. Les outils technologiques utilisés pour sa réalisation sont détaillés en annexe.

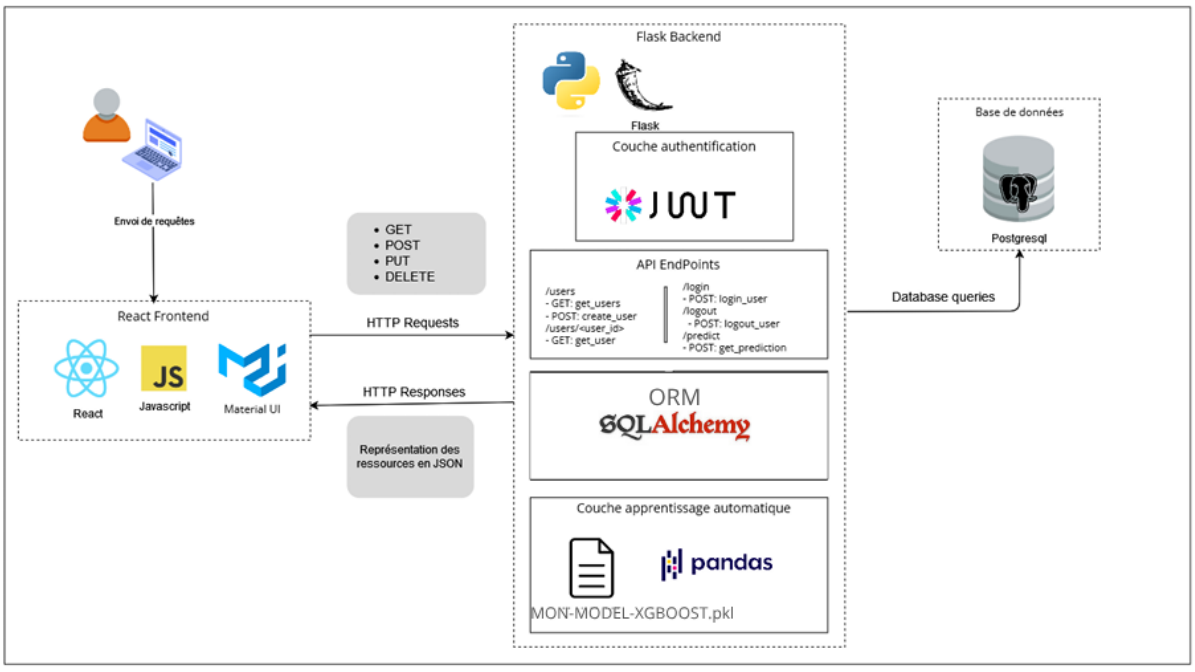



FIGURE 5.13 – Architecture du tableau de bord

### 5.3.2 Page d'authentification :

Cette page constitue la porte d'entrée de l'application. Elle permet aux utilisateurs autorisés d'accéder aux différentes fonctionnalités du système via un formulaire simple et intuitif. L'interface présente deux champs de saisie : l'un pour l'adresse e-mail et l'autre pour le mot de passe. Une fois les informations saisies, l'utilisateur peut se connecter en toute sécurité. Cette étape garantit la protection des données et restreint l'accès aux personnes habilitées.

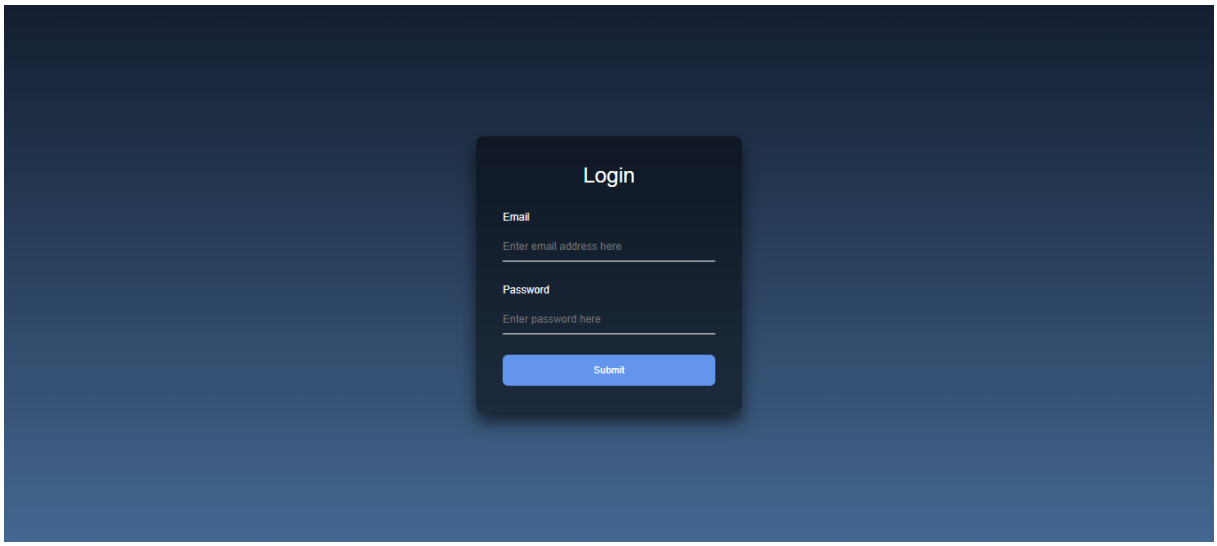



FIGURE 5.14 – Interface de la page d'authentification .

### 5.3.3 Page d'accueil (Dashboard) :

La page d'accueil de l'application constitue le centre de visualisation des informations clés issues du dataset client. Elle est conçue pour offrir une vue d'ensemble claire et interactive du phénomène de churn.

Elle se compose de plusieurs éléments :

- État global du dataset 5.15 affiché sous forme de cartes synthétiques :
  - Nombre total de clients dans le dataset
  - Taux global de churn prédit
  - Nombre de clients présentant un risque élevé de churn
  - Indice moyen de satisfaction
  - Volume moyen de données consommées par mois
  - Durée moyenne de connexion mensuelle
  - Nombre moyen d'appels au service après-vente (SAV)
- Un graphique illustrant le taux de churn en fonction du type de service internet 5.15 souscrit (adsl, fibre, 4g lte, etc.) est proposé. Il permet de visualiser, pour chaque offre internet, le nombre de clients churneurs et non churneurs. De plus, un filtrage par catégorie

de service est disponible, facilitant ainsi l'analyse ciblée selon les types d'abonnement (offres adsl, mixtes, fibre, 4g lte, etc.).

- Une section présente une carte montrant la répartition géographique du churn selon les régions, accompagnée d'un graphique illustrant les scores de satisfaction des clients. Ces visualisations permettent d'identifier les zones les plus touchées par le churn et de mieux comprendre le ressenti des abonnés. Elles sont illustrées dans la figure 5.16.
- Une section illustre, à l'aide d'un graphique linéaire, l'évolution mensuelle de la consommation moyenne d'internet en Go. Elle met en évidence les tendances globales, les pics d'utilisation ainsi que les périodes de faible activité. Cette visualisation est présentée dans la figure 5.17.
- Une section présente les facteurs les plus importants influençant le churn, identifiés grâce aux modèles d'apprentissage automatique. Elle permet de comprendre les variables les plus déterminantes dans la décision de résiliation des clients. La visualisation correspondante est affichée dans la figure 5.17.

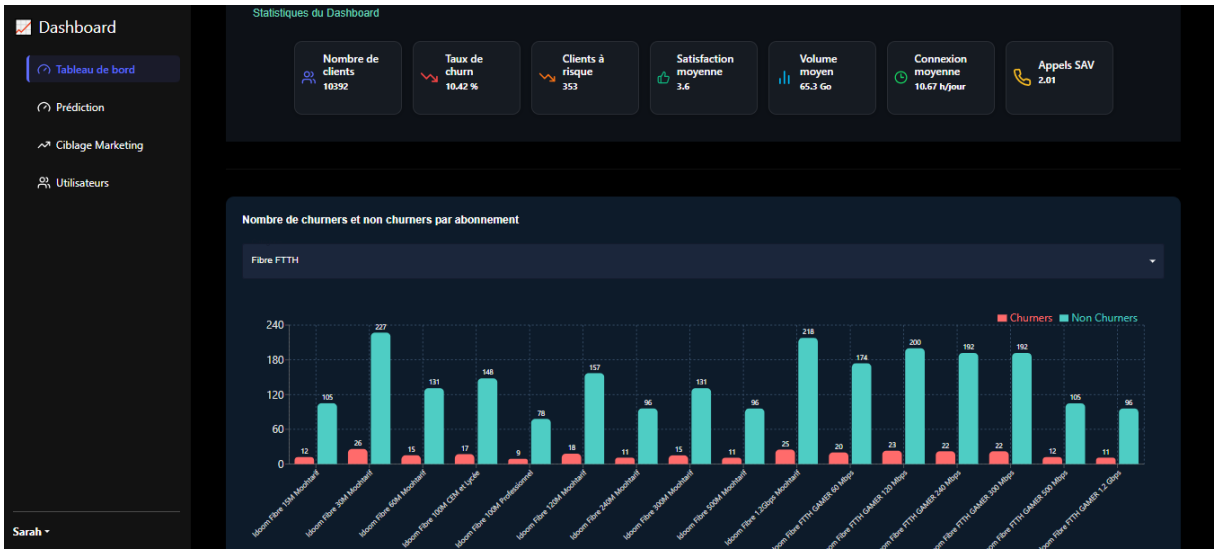



FIGURE 5.15 – Page d'accueil — Sections État global du dataset et Churn par offre Internet

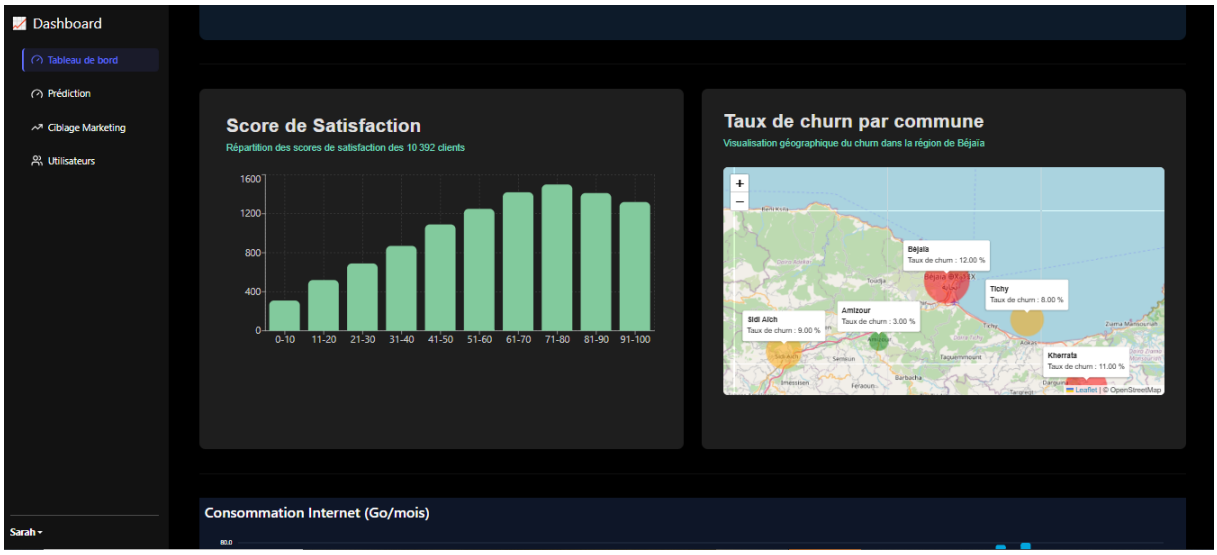



FIGURE 5.16 – Page d'accueil — Carte de répartition du churn et scores de satisfaction des clients

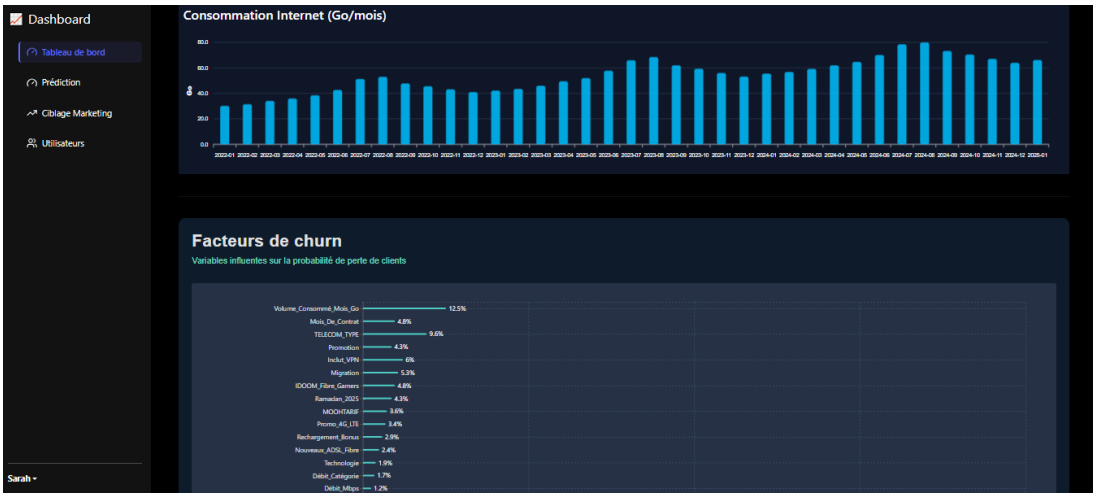



FIGURE 5.17 – Page d'accueil — Facteurs de churn et répartition des scores de satisfaction

### 5.3.4 Page prédiction

Cette page est dédiée à la mise à jour du dataset et à la visualisation des résultats de prédiction. Elle permet d'importer de nouvelles données, de supprimer l'ancien dataset si nécessaire, et d'identifier les clients à risque de churn à travers un tableau interactif. Elle constitue donc une interface essentielle pour exploiter le modèle prédictif sur des données réelles.

• Mise à jour du dataset : une section permet à l'utilisateur de charger un nouveau dataset au format CSV, tout en offrant la possibilité de supprimer le dataset existant. Cette action déclenche automatiquement la mise à jour des prédictions . La figure **??** présente un aperçu du contenu de cette section. .

• Affichage des clients à risque : une seconde section présente la liste des clients à risque de churn sous forme de tableau. Ce tableau contient des informations détaillées pour chaque client

et peut être exporté au format CSV pour un usage ultérieur .La figure **??** présente un aperçu du contenu de cette section. .

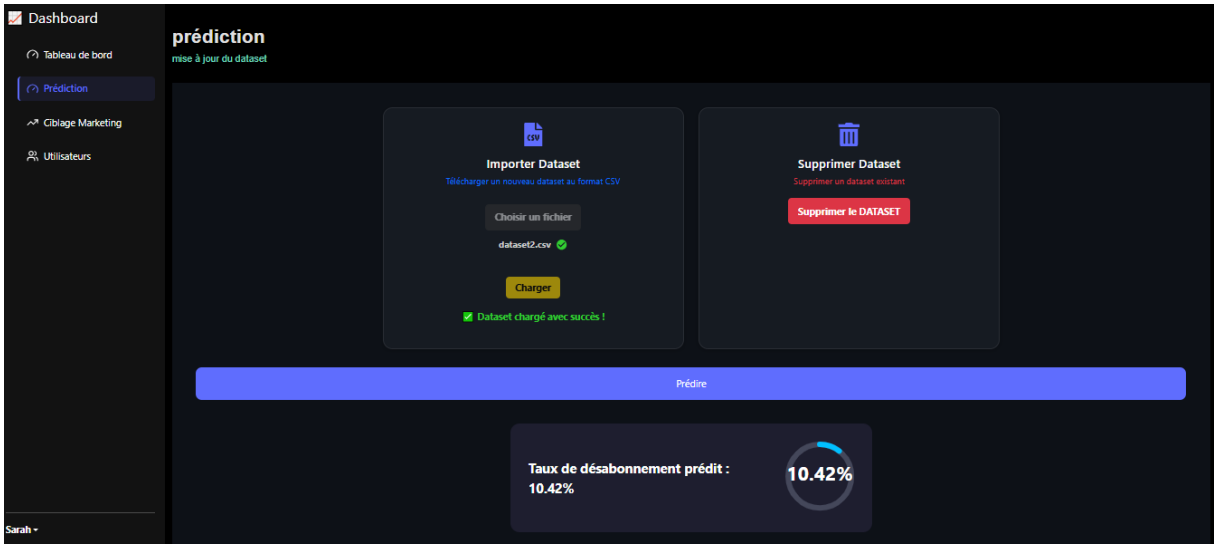



FIGURE 5.18 – Page Prédiction — Mise à jour du dataset

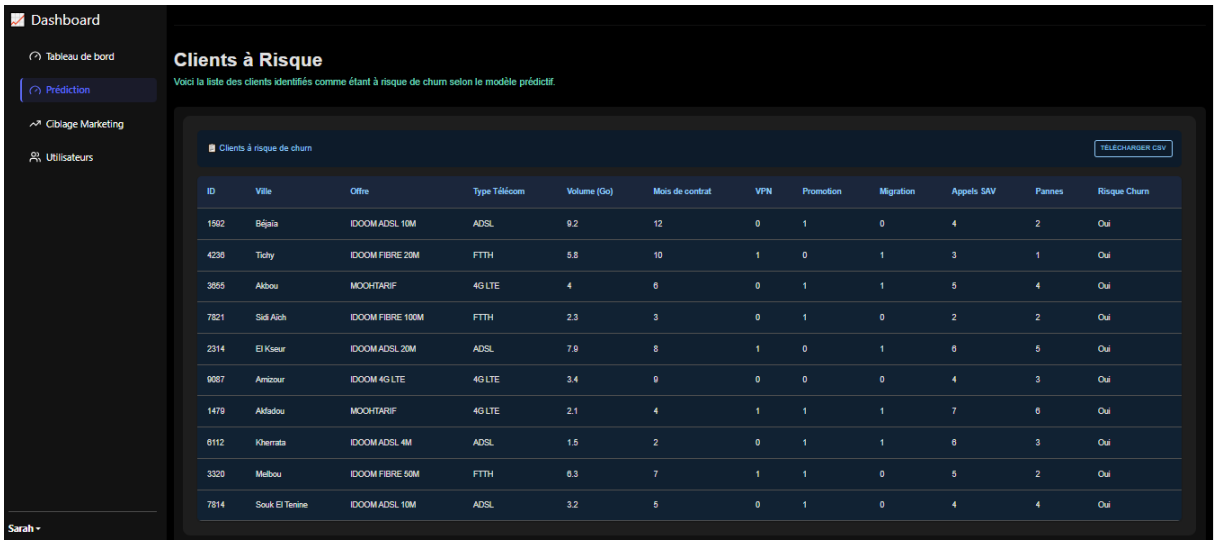



FIGURE 5.19 – Page prédiction — Tableau des clients à risque de churn

### 5.3.5 Page segmentation marketing

Cette page est consacrée à l'analyse des comportements clients, dans le but d'identifier les segments les plus susceptibles de réagir favorablement aux campagnes marketing. Elle vise à optimiser les actions promotionnelles en évitant les dépenses inutiles et en ciblant les clients les plus influençables.

• Segmentation : une première section présente les quatre segments de clients (Persuadables, Sure Thing, Lost Cause, Do Not Disturb) à l'aide d'un graphique. Cette segmentation est basée

sur un score Uplift calculé pour chaque abonné, et permet de visualiser la répartition des clients dans chaque catégorie.

• Explication des segments : pour chaque segment, une brève explication est fournie afin de guider l'interprétation marketing des comportements clients attendus.

• Réaction aux campagnes marketing : un graphique vertical permet de comparer, pour chaque campagne, le nombre de clients ciblés et ceux ayant effectivement accepté l'offre. Cette visualisation facilite l'évaluation de l'impact de chaque action promotionnelle menée par Algérie Télécom.

L'interface de cette page est illustrée dans les figures 5.20 et 5.21.


FIGURE 5.20 – Page Segmentation Marketing — Visualisation des segments


FIGURE 5.21 – Page Segmentation Marketing — Réaction aux campagnes marketing

### 5.3.6 Page utilisateurs

Une section est dédiée à la gestion des utilisateurs de l'application. Elle permet de visualiser la liste des utilisateurs enregistrés dans le système, ainsi que d'effectuer les opérations classiques de gestion : ajout, modification et suppression. Ces actions sont accessibles via des boutons interactifs accompagnés de modales (fenêtres contextuelles) pour la saisie ou l'édition des informations.

Cette interface facilite l'administration des comptes, notamment dans le cadre d'un accès multi-utilisateur ou d'un usage collaboratif.

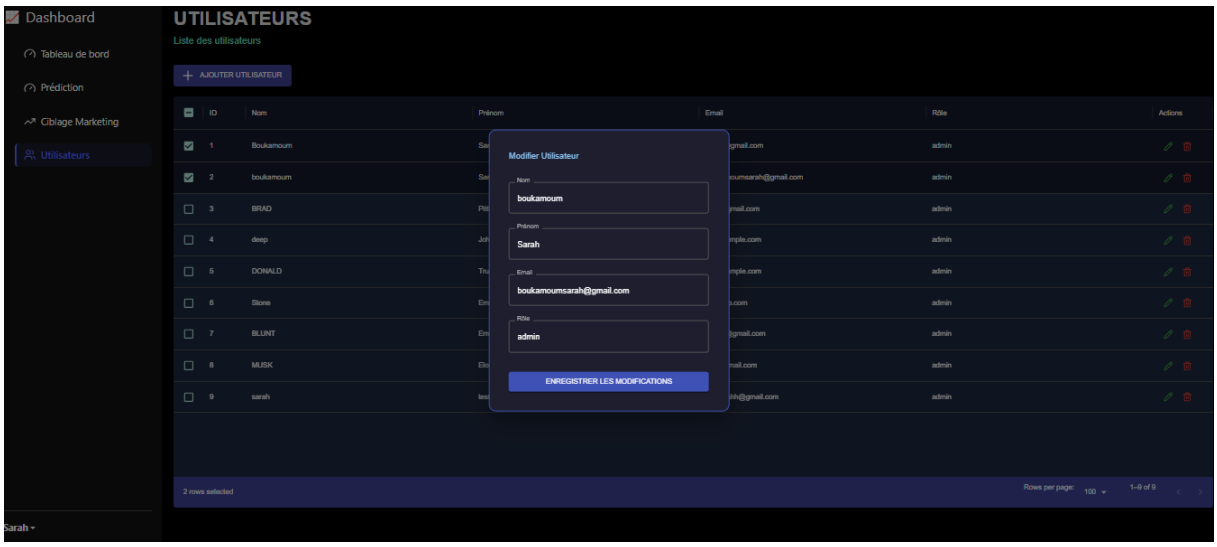



FIGURE 5.22 – Page Utilisateurs — Gestion des comptes (ajout, modification, suppression)

## 5.4 Conclusion

Pour conclure, ce chapitre a permis de détailler l'ensemble du processus menant à la réalisation de notre solution proposée ( expliquée dans la section 2.4 ) , depuis l'analyse exploratoire des données jusqu'à l'implémentation d'une application web fonctionnelle. Grâce aux différentes étapes de traitement, de modélisation et de visualisation, nous avons pu construire une solution capable de détecter les clients à risque de départ et de fournir des indicateurs décisionnels clairs aux équipes concernées. Cette intégration technique et fonctionnelle constitue une base solide pour une exploitation opérationnelle des résultats, tout en ouvrant la voie à d'éventuelles améliorations futures.

# Conclusion Générale

Dans Ce projet de fin d'étude nous avons conçu et mis en œuvre une solution de prédiction du churn permettant d'identifier les clients présentant un risque élevé de résiliation, en exploitant des méthodes d'apprentissage automatique.

L'étude s'est focalisée sur le cas d'Algérie Télécom, où nous avons analysé les principaux facteurs influençant le départ des clients, notamment la qualité du service, la politique commerciale ainsi que la gestion de la relation client. Après un travail rigoureux de préparation, de nettoyage et d'équilibrage des données extraites du système d'information, 3 algorithmes de machine learning ont été appliqués. Leur performance a été évaluée selon des indicateurs standards, nous permettant de sélectionner le modèle le plus performant.

Ce modèle a ensuite été intégré dans une application web complète, basée sur une architecture Flask pour le backend et React pour le frontend. Cette plateforme permet aux utilisateurs d'explorer les résultats de manière claire et interactive, et constitue un outil décisionnel pertinent pour les responsables marketing.

Ce travail a ainsi démontré qu'il est possible de renforcer la stratégie de fidélisation grâce à une exploitation intelligente des données et des techniques de modélisation prédictive. En anticipant les comportements de résiliation, Algérie Télécom pourra déployer des actions ciblées et proactives afin de limiter le churn.

Enfin, ce projet ouvre la voie à plusieurs perspectives d'amélioration, telles que l'intégration de modèles dynamiques, l'analyse en temps réel, ainsi que le développement de systèmes de recommandation personnalisés permettant de proposer aux clients des offres ciblées en fonction de leur profil et de leur comportement.

# Annexe A : Technologies Utilisées

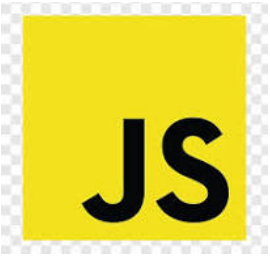
## A.1 Les langages de programmation

### - Python

C'est un langage de programmation interprété de haut niveau, facile à apprendre et doté d'une syntaxe simple. Il est largement utilisé dans les universités pour les projets de recherche, le prototypage et le calcul, ainsi que dans le domaine de la Data Science, l'apprentissage automatique, le développement web et l'automatisation. Python est connu pour sa simplicité, sa lisibilité et sa polyvalence. Sa popularité est largement due à sa vaste bibliothèque standard et au grand nombre de bibliothèques et de cadres tiers qui existent. [26]

### - JavaScript

JavaScript (abrégé par JS) est un langage de programmation dynamique complet, qui permet de développer des pages web réactives et interactives. À l'inverse des autres langages serveurs, JS exécute ses tâches au niveau du navigateur lui-même, du côté de l'utilisateur et non du serveur web. Il rejoint ainsi le rang d'ECMA Script, qui désigne les langages de scripts orientés vers le client. [27]




## A.2 Les bibliothèques Python

### - Pandas

Pandas est une bibliothèque Python largement utilisée pour la manipulation, le nettoyage et l'analyse des données. Elle repose sur NumPy et offre des structures de données efficaces comme les `DataFrame` et `Series`, permettant de gérer facilement des données

tabulaires. Pandas facilite le traitement de fichiers CSV, Excel, ou encore JSON, ce qui en fait un outil indispensable pour les data scientists. [28]

### - NumPy

NumPy est une bibliothèque de calcul numérique offrant un support pour des tableaux multidimensionnels ainsi qu'une panoplie de fonctions mathématiques. Grâce à ses performances optimisées, elle est au cœur de nombreuses bibliothèques scientifiques Python, notamment Pandas, SciPy ou Scikit-learn.[29]

### - Matplotlib


Matplotlib permet de générer des visualisations statiques, interactives ou animées. Elle est souvent utilisée pour tracer des graphiques 2D (lignes, barres, nuages de points, etc.), mais peut également produire des visualisations 3D. Son intégration avec NumPy en fait une bibliothèque puissante pour l'analyse visuelle de données scientifiques. [30]

### - Scikit-learn


Scikit-learn est l'une des bibliothèques les plus complètes pour l'apprentissage automatique. Elle propose des outils pour la classification, la régression, le clustering, la réduction de dimension, la sélection de caractéristiques, ainsi que l'évaluation de modèles. Son interface simple en fait un choix populaire pour les projets de machine learning en Python.[31]

### - Seaborn

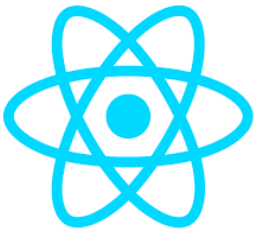

Seaborn est construite sur Matplotlib et fournit une interface haut niveau pour créer des visualisations statistiques attractives. Elle permet notamment de visualiser des relations complexes entre plusieurs

variables avec peu de code, tout en exploitant les structures de données Pandas.

## A.3 Frameworks et bibliothèques Web

### - ReactJS

ReactJS est une bibliothèque JavaScript développée par Facebook, conçue pour la création d'interfaces utilisateur dynamiques et performantes. Elle repose sur une architecture basée sur les composants réutilisables, et utilise un DOM virtuel pour minimiser les mises à jour réelles du navigateur, ce qui améliore les performances. Grâce à sa modularité et à sa large adoption, React est devenue une solution incontournable pour le développement d'applications web monopages (SPA). [32]



### - Material UI

Material UI est une bibliothèque de composants React qui applique les principes du Material Design de Google. Elle propose une large gamme de composants visuels modernes, personnalisables et accessibles, tout en assurant la cohérence visuelle de l'interface. Elle permet un développement rapide et professionnel d'interfaces utilisateur réactives, compatibles SEO et responsive. [33]

### - Flask

Flask est un microframework web open source pour Python, apprécié pour sa légèreté et sa flexibilité. Il fournit les fonctionnalités essentielles pour créer une application web, tout en laissant à l'utilisateur la liberté d'ajouter des extensions selon ses besoins (authentification, base de données, formulaire, etc.). Flask est particulièrement adapté pour le développement d'API RESTful, de prototypes rapides, et d'applications backend légères.

## A.4 Logiciels Utilisés

### - Jupyter Notebook

Jupyter Notebook est un environnement interactif basé sur le web, permettant de créer des documents contenant du code exécutable, du texte (Markdown), des formules mathématiques et des visualisations. Les fichiers ont l'extension `.ipynb` et sont très utilisés en data science et en recherche.


### - Visual Studio Code (VS Code)

VS Code est un éditeur de code source gratuit développé par Microsoft. Il supporte plusieurs langages, propose des fonctionnalités avancées (débogage, terminal intégré, extensions) et une excellente intégration avec Git. Très populaire, il est hautement personnalisable.

### - Postman

Postman est un outil collaboratif de test d'API, très utilisé pour envoyer des requêtes HTTP, analyser les réponses, et automatiser les scénarios de test. Il est essentiel dans le développement backend et dans la validation des endpoints RESTful. [34]


### - StarUML

StarUML est un logiciel de modélisation UML qui prend en charge de nombreux diagrammes standards (classes, cas d'utilisation, séquences, etc.). Il permet aussi la génération de code et favorise la conception structurée des systèmes logiciels.

## A.5 Autres Outils

### - SQLAlchemy

SQLAlchemy est un ORM (Object-Relational Mapper) pour Python, qui établit une correspondance entre les classes Python et les tables d'une base de données relationnelle. Cette couche d'abstraction permet d'interagir avec les bases de données via des objets, au lieu d'écrire directement des requêtes SQL. SQLAlchemy offre une grande flexibilité, tout en assurant des interactions cohérentes, performantes et idiomatiques avec les bases de données. [35]

### - JSON Web Token (JWT)

Le JSON Web Token (JWT) est un standard (RFC 7519) permettant l'échange sécurisé d'informations sous forme de jetons signés. Il est composé de trois parties : un en-tête, une charge utile et une signature, qui garantit l'authenticité des données transmises. JWT est largement utilisé dans les applications web modernes pour l'authentification sans état (stateless), notamment dans les applications monopages (SPA) et mobiles. [36]

### - Redux

Redux est une bibliothèque JavaScript utilisée pour gérer l'état global des applications, en particulier dans les projets React. Elle repose sur le principe d'un store centralisé qui stocke l'état de l'application, rendant les flux de données prévisibles. Redux facilite la gestion de données complexes, la synchronisation entre composants et le débogage.

### - Bootstrap

Bootstrap est un framework CSS open-source conçu pour développer des interfaces web réactives et modernes. Il propose un ensemble de composants, grilles, et utilitaires prêts à l'emploi, facilitant la création de pages responsives. Même si Material UI est le principal framework utilisé ici, Bootstrap est parfois intégré pour sa rapidité de mise en œuvre et sa compatibilité avec divers navigateurs.

# Bibliographie


[1] HARVESTR. *Churn : qu'est-ce que c'est et comment le réduire ?* Consulté le 6 juillet 2025. 2023. URL : https://blog.harvestr.io/fr/churn.

[2] Jonathan BUREZ et Dirk Van den POEL. "Separating financial from commercial customer churn : A modelling step towards resolving the conflict between the sales and credit department". In : *Expert Systems with Applications* (2008), p. 497-514.

[3] Akila LANSEUR et Houria Ait SIDHOUM. "Les déterminants du Churn client dans le secteur des télécommunications : étude des trois opérateurs de la téléphonie mobile en Algérie". In : (2021). URL : https://asjp.cerist.dz/en/downArticlepdf/180/9/3/168009.

[4] https ://www.algerietelecom.dz/fr/page/mot-du-president-directeur-general-p141.

[5] Jean-Marc LEHU. *Stratégie de fidélisation*. Paris : Éditions d'Organisation, 2004. URL : https://www.academia.edu/6907217/Strat%C3%A9gie_de_fid%C3%A9lisation.

[6] Jacques LENDREVIE, Julien LÉVY et Denis LINDON. *Mercator : Théorie et pratique du marketing*. 8e édition. Dunod, 2006.

[7] Stuart RUSSELL et Peter NORVIG. *Artificial Intelligence : A Modern Approach*. 4e éd. Hoboken, NJ : Pearson, 2021.

[8] COURSERA STAFF. *What Is Machine Learning ? Definition, Types, and Examples*. Consulté le 24 juin 2025. Mai 2025.

[9] Cécile CAZENCOTT. *Introduction au Machine Learning*. https://cazencott.info/dotclear/public/lectures/IntroML_Azencott.pdf. 2020.

[10] DATASCIENTEST. *Régression Logistique : Qu'est-ce que c'est ?* 2024. URL : https://datascientest.com/regression-logistique-quest-ce-que-cest.

[11] IBM. *Arbres de décision : Définition et explication*. Consulté le 26 février 2025. 2024. URL : https://www.ibm.com/fr-fr/think/topics/decision-trees.

[12] Will KOEHRSEN. *XGBoost Explained — Everything You Need to Know*. Towards Data Science. 2021. URL : https://towardsdatascience.com/xgboost-explained-everything-you-need-to-know-5ed36f2fc437.

[13] *Cross-validation : A Brief Survey*. Accessed : 2025-04-06. URL : https://dberrar.github.io/papers/Berrar_EBCB_2nd_edition_Cross-validation_preprint.pdf.

[14] INNOVATIANA. *Démystifier la matrice de confusion en IA*. URL : https://www.innovatiana.com/post/understand-confusion-matrix-in-ai.

[15] GOOGLE FOR DEVELOPERS. *Classification : justesse, rappel, précision et métriques associées*. URL : https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=fr.

[16] DATASCIENTEST. *Comprendre la courbe ROC AUC*. URL : https://datascientest.com/courbe-roc-auc-tout-savoir.

[17] Ilyes TALBI. *3 méthodes pour optimiser les hyperparamètres de vos modèles de machine learning*. La Revue IA. Consulté le 24 juin 2025. 2021. URL : https://larevueia.fr/3-methodes-pour-optimiser-les-hyperparametres-de-vos-modeles-de-machine-learning/.

[18] Pınar CIHAN. *Bayesian Hyperparameter Optimization of Machine Learning Models*. URL : https://www.researchgate.net/publication/388271121_Bayesian_Hyperparameter_Optimization_of_Machine_Learning_Models_for_Predicting_Biomass_Gasification_Gases/figures.

[19] Lackeshwar BACHAN et Tarek GABER. "Predicting Customer Churn in the Internet Service Provider Industry of Developing Nations : A Single, Explanatory Case Study of Trinidad and Tobago". In : *Advances in Intelligent Systems and Computing*. Springer, 2021. DOI : 10.1007/978-3-030-69717-4_77. URL : https://www.researchgate.net/publication/349812555_Predicting_Customer_Churn_in_the_Internet_Service_Provider_Industry_of_Developing_Nations_A_Single_Explanatory_Case_Study_of_Trinidad_and_Tobago.

[20] C. PRABADEVI, R. SHALINI et V. KAVITHA. "Customer Churning Analysis Using Machine Learning Algorithms". In : *Heliyon* 9.3 (2023), e14088. DOI : 10.1016/j.heliyon.2023.e14088. URL : https://www.sciencedirect.com/science/article/pii/S2666603023000143.

[21] Philippe KRUCHTEN. *The Rational Unified Process : An Introduction*. https://www.ibm.com/docs/en/engineering-lifecycle-management-suite/lifecycle-management/7.0.2?topic=overview-rational-unified-process. Consulté en juin 2025. 2003.

[22] LUCIDCHART. *Le langage UML (Unified Modeling Language)*. Consulté en juin 2025. 2024. URL : https://www.lucidchart.com/pages/fr/langage-uml#:~:text=Le%20langage%20UML%20(Unified%20Modeling,et%20riche%20s%C3%A9mantiquement%20et%20syntaxiquement..

[23] SCIKIT-LEARN DEVELOPERS. *sklearn.preprocessing.LabelEncoder — scikit-learn 1.4.2 documentation*. Consulté en juin 2025. 2024. URL : https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html.

[24] IMBALANCED-LEARN DEVELOPERS. *imblearn.under_sampling.RandomUnderSampler¯imbalanced − learn*0.11.0*documentation*. Consulté en juin 2025. 2024. URL : https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html.

[25] Scikit learn DEVELOPERS. *sklearn.model_selection.GridSearchCV¯scikit − learnDocumentation*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. Consulté le 25 juin 2025. 2024. URL : https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[26] PYTHON SOFTWARE FOUNDATION. *Tutoriel Python — Documentation Python 3.13*. Consulté en juin 2025. 2024. URL : https://docs.python.org/fr/3.13/tutorial/index.html.

[27] WIKIPÉDIA. *JavaScript*. Consulté en juin 2025. 2025. URL : https://fr.wikipedia.org/wiki/JavaScript.

[28] THE PANDAS DEVELOPMENT TEAM. *Pandas - Python Data Analysis Library*. Consulté en juin 2025. 2025. URL : https://pandas.pydata.org/.

[29] DATASCIENTEST. *NumPy*. Consulté en juin 2025. 2025. URL : https://datascientest.com/numpy.

[30] DATASCIENTEST. *Matplotlib : tout savoir sur la bibliothèque de tracé en Python*. Consulté en juin 2025. 2025. URL : https://datascientest.com/matplotlib-tout-savoir.

[31] REACT TEAM. *React – Documentation Française*. Consulté en juin 2025. 2025. URL : https://fr.react.dev/.

[32] MATERIAL-UI TEAM. *Material-UI (MUI) – Getting Started*. Consulté en juin 2025. 2025. URL : https://mui.com/material-ui/getting-started/.

[33] DATASCIENTEST. *Avantages et fonctionnement de Flask*. Consulté en juin 2025. 2025. URL : https://datascientest.com/avantages-et-fonctionnement-de-flask.

[34] DATASCIENTEST. *Postman : tout savoir sur l'outil de tests d'API*. Consulté en juin 2025. 2025. URL : https://datascientest.com/postman-tout-savoir.

[35] DATASCIENTEST. *SQLAlchemy : tout savoir sur l'ORM Python*. Consulté en juin 2025. 2025. URL : https://datascientest.com/sqlalchemy-tout-savoir.

[36] IONOS DIGITALGUIDE. *JSONWebToken (JWT) – Sécurisation et fonctionnement*. Consulté en juin 2025. 2025. URL : https://www.ionos.fr/digitalguide/sites-internet/developpement-web/json-web-token-jwt/.

# Résumé

La fidélisation des clients représente aujourd'hui un défi stratégique pour l'ensemble des entreprises, en particulier dans le secteur des télécommunications où la concurrence est forte et les attentes des clients en constante évolution. **Algérie Télécom**, acteur majeur du marché national, n'échappe pas à ce phénomène de résiliation croissante. Dans ce contexte, ce projet de fin d'études vise à concevoir un système intelligent de prédiction du churn à l'aide des techniques d'apprentissage automatique. En s'appuyant sur un dataset représentatif des offres et services proposés par l'opérateur, plusieurs étapes ont été réalisées : préparation des données, encodage, normalisation, équilibrage, puis entraînement et évaluation de plusieurs modèles. Le modèle retenu a été intégré dans une application web interactive, développée avec *React* pour l'interface utilisateur et *Flask* pour le backend. Cette solution permet aux responsables de visualiser les clients à risque et de prendre des décisions éclairées pour améliorer les stratégies de rétention. Le projet constitue ainsi un outil concret et évolutif, adapté aux besoins d'Algérie Télécom et plus largement à ceux des opérateurs confrontés au churn.

**Mots-clés :** prédiction du churn, fidélisation client, apprentissage automatique, télécommunications, Algérie Télécom, React, Flask.

# Abstract

Customer retention has become a strategic challenge for businesses across all sectors, especially in the telecommunications industry, where competition is intense and customer expectations continue to evolve. **Algérie Télécom**, a major player in the Algerian market, is no exception to this growing churn phenomenon. In this context, this final year project aims to develop an intelligent system capable of predicting customer churn using machine learning techniques. Based on a dataset reflecting the operator's services and commercial offers, several steps were carried out : data preprocessing, encoding, normalization, balancing, model training, and evaluation. The most effective model was then integrated into an interactive web application, developed with *React* for the frontend and *Flask* for the backend. This tool enables business teams to visualize at-risk customers and make informed decisions to improve retention strategies. The resulting solution is practical, scalable, and aligned with the current needs of Algérie Télécom and other operators facing similar churn-related challenges.

**Keywords :** churn prediction, customer retention, machine learning, telecommunications, Algérie Télécom, React, Flask.