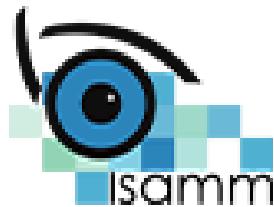


Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de la Mannouba

**Institut Supérieur des Arts Multimédias**



Rapport De projet de fin d'études  
Présenté en vue de l'obtention du diplôme de

**Licence en Big Data et Analyse de données**

Sujet :

Prédiction du Churn des clients Tunisie Télécom

Élaboré par :

Fatma Ezzahra Ben Abdallah

Organisme d'accueil :

Tunisie Télécom



Encadrant Universitaire : M. Bechikh Chedi

Encadrant Professionnel : M. Jemai Hatem

Année Universitaire 2023 / 2024

# Dédicace

## **À mes très chers parents Mohamed et Leila**

Je vous dois ce que je suis aujourd’hui grâce à votre amour, à votre patience vos innombrables sacrifices. Que ce modeste travail soit pour vous une petite compensation et reconnaissance envers ce que vous avez fait d’incroyable pour moi. Que dieu, le tout puissant, vous préserve et vous procure santé et longue vie afin que je puisse, à mon tour, vous combler.

## **À mon frère Ayoub**

Celui que j'aime à la folie, celui qui n'a jamais hésité à me soutenir et à m'encourager. Que Dieu te réserve une longue vie pleine de bonheur, de santé et de succès dans ta vie professionnelle et familiale.

## **À mes meilleures amies**

En témoignage de l'amitié sincère qui nous a li et des bons moments que nous avons passés ensemble. Je vous dédie ce travail en vous souhaitant un avenir radieux et plein de bonnes promesses.

# Remerciements

Je tiens, avant de présenter mon travail, à exprimer ma grande reconnaissance envers les personnes qui m'ont, de près ou de loin, apporté leur soutien. Qu'ils trouvent ici collectivement et individuellement l'expression de toute ma gratitude. J'adresse mes vifs remerciements dans un premier temps, à toute l'équipe pédagogique de l'ISAMM, et spécialement mon encadrant Monsieur Bechikh Chedi qui a bien voulu assurer un encadrement continu et rigoureux à ce travail tout en mettant à ma disposition tous les moyens possibles et nécessaires. De plus, ses compétences et ses conseils étaient extrêmement judicieux et constructifs.

J'exprime mes sincères gratitude à toute l'équipe de cadre de stage, pour l'expérience enrichissante et pleine d'intérêt qu'ils m'ont fait vivre durant ces quatre mois au sein de l'entreprise Tunisie Télécom, et spécialement mon encadrant Monsieur Hatem Jemai pour m'avoir intégré rapidement au sein de l'entreprise et accordé toute sa confiance, son encouragement, ses remarques et pour le temps qu'il m'a consacré tout au long de cette période à toutes mes interrogations malgré ses grandes occupations.

J'adresse également mes remerciements aux membres du Jury pour avoir accepté de juger ce travail.

# Résumé

L'objectif principal de ce rapport de stage est de développer un modèle de prédiction de churn (taux de désabonnement) qui identifiera les clients les plus proches d'annuler leurs abonnements. Nous utiliserons des techniques d'apprentissage automatique pour prédire l'attrition des clients et intégrerons le modèle dans une application Web.

Cette application aidera les agences de télécommunications tunisiennes, dans notre cas la Tunisie Télécom, à faire des prévisions facilement et contiendra un lien vers un rapport BI qui permet de visualiser les données historiques des clients. En fait, pour mener à bien ce projet, nous avons suivi la méthodologie CRISP-DM.

Mots clés : client churn, apprentissage automatique, prédiction churn, CRISP-DM, Python, BI

# Abstract

The main objective of this internship report is to develop a churn (churn rate) prediction model that will identify the customers closest to cancel their subscriptions. We will use machine learning techniques to predict customer churn and embed the model in a web application.

This application will help Tunisian telecommunications agencies, in our case Tunisia Telecom, to make forecasts easily and will contain a link to a BI report which allows to visualize the historical data of the customers. In fact, to carry out this project, we followed the CRISP-DM methodology.

Keywords: churn costumers, machine learning, churn prediction, CRISP-DM, Python, BI

# Table des matières

Introduction générale	1
I Étude du projet	3
1.1 Introduction .....	4
1.2 Présentation de l'organisme d'accueil .....	4
1.2.1 Présentation de Tunisie Télécom .....	4
1.2.2 Historique .....	5
1.2.3 Organisation .....	5
1.2.4 Présentation du complexe HACHED .....	6
1.3 Présentation du projet .....	6
1.3.1 Cadre général du projet .....	6
1.3.2 Problématique .....	6
1.3.3 Solution proposée .....	7
1.4 Méthodologie de gestion de projet adoptée .....	7
1.4.1 Méthodologie SEMMA .....	7
1.4.2 Méthodologie TDSP .....	8
1.4.3 Méthodologie CRISP-DM .....	9
1.4.4 Comparaison des méthodologies .....	10
1.5 Architecture de projet .....	11
1.6 Environnement de travail .....	12
1.6.1 Environnement logiciel .....	12
1.6.1.1 Logiciels utilisés .....	12
1.6.1.2 Bibliothèque utilisées .....	13

1.6.1.3 Langages utilisés .....	15
1.6.2 Environnement matériel .....	15
1.7 Conclusion .....	16
2 Compréhension du problème métier et des données	17
2.1 Introduction .....	18
2.2 Notions théoriques liées au projet .....	18
2.2.1 Churn .....	18
2.2.2 Apprentissage automatique .....	19
2.3 Les défis de la modélisation en Machine Learning .....	19
2.3.1 Overfitting .....	19
2.3.2 Underfitting .....	20
2.4 Choix de la base de données .....	20
2.5 Exploration des données .....	23
2.6 Analyse des données .....	27
2.6.1 Analyse univarié .....	27
2.6.2 Analyse bivariée .....	35
2.7 Conclusion .....	40
3 Prétraitement des données	41
3.1 Introduction .....	42
3.2 Nettoyage des données .....	42
3.2.1 Suppression des colonnes inutiles .....	42
3.2.2 Les valeurs manquantes .....	44
3.2.2.1 Traitement des valeurs manquantes .....	45

3.2.2.2 Vérification des valeurs manquantes .....	48
3.2.3 Les valeurs dupliquées .....	49
3.3 Encodage des données .....	49
3.4 Normalisation et standardisation .....	52
3.5 Sélection des fonctionnalités .....	53
3.5.1 Matrice de corrélation .....	53
3.5.2 Importance des variables .....	54
3.6 Conclusion .....	55
4 La modélisation	56
4.1 Introduction .....	57
4.2 Découpage de base de données .....	57
4.2.1 Extraction des variables prédictive et cible : .....	57
4.2.2 Données d'entraînement et de test .....	57
4.3 Modélisation .....	58
4.3.1 Random Forest .....	58
4.3.2 Arbre de décision .....	60
4.3.3 Différence entre arbre de décision et Random Forest .....	61
4.3.4 Logistic Regression .....	62
4.3.5 K-nearest neighbors: KNN .....	63
4.3.6 eXtreme Gradient Boosting: XGboost .....	64
4.3.7 Naïve Bayes .....	66
4.4 La validation croisée .....	67
4.5 Les étapes de construction du modèle .....	68

4.6 Les mesures de performance .....	68
4.6.1 Matrice de confusion .....	68
4.6.1.1 Precision .....	69
4.6.1.2 Accuracy .....	70
4.6.1.3 Recall .....	70
4.6.1.4 F1-Score .....	70
4.6.2 La courbe ROC-AUC .....	71
4.6.3 La courbe Learning Curve .....	72
4.7 Réglage des hyperparamètres .....	73
4.8 Conclusion .....	74
<b>5 Évaluation et optimisation de la performance des modèles</b>	<b>75</b>
5.1 Introduction .....	76
5.2 Évaluation de chaque modèle .....	76
5.2.1 Random Forest. ....	76
5.2.2 Decision Tree. ....	78
5.2.3 Logistic Regression. ....	79
5.2.4 XGBoost. ....	80
5.2.5 K-nearest neighbors: KNN. ....	82
5.2.6 Naïve Bayes .....	83
5.3 Comparaison et évaluation des algorithmes utilisés .....	85
5.4 L'Ensemble Learning .....	86
5.5 Ajustement de modèle .....	88
5.6 Conclusion .....	88

6 Déploiement	89
6.1 Introduction .....	90
6.2 Déploiement .....	90
6.3 Extraction du modèle .....	90
6.4 Développement de l'interface Web .....	91
6.4.1 Les interfaces de l'application .....	94
6.5 Construction de tableau de bord .....	97
6.6 Diagramme de gantt .....	99
6.7 Conclusion .....	100
Conclusion générale	101

# Liste des figures

1.1	Logo de Tunisie Télécom [1] .....	4
1.2	Organisation fonctionnelle de Tunisie Télécom [2] .....	5
1.3	Le cycle de vie de SEMMA [4] .....	8
1.4	Le cycle de vie de TDSP [5] .....	9
1.5	Le cycle de vie de CRISP-DM [7] .....	10
1.6	Architecture du projet .....	11
1.7	Les bibliothèques utilisées [18] .....	14
2.1	Différence entre apprentissage supervisé et non-supervisé [26] .....	19
2.2	Les problèmes de la modélisation [33] .....	20
2.3	La base de données .....	22
2.4	Importation du dataset .....	22
2.5	Les dimensions du dataset .....	23
2.6	Les 5 premiers lignes du dataset .....	23
2.7	Résumé du dataset .....	24
2.8	Les statistiques sur les colonnes numériques .....	25
2.9	Les Statistiques du colonnes catégoriques .....	26
2.10	Le nombre des valeurs uniques par colonne .....	26
2.11	Les variables catégorielles .....	27
2.12	Répartition des clients en fonction de leur statut de churn .....	27
2.13	Visualisation de chaque catégorie de genre .....	28
2.14	Distribution des clients en fonction de leur activation des messages vocaux .....	29
2.15	Répartition des clients selon leur situation civile .....	29
2.16	Répartition des clients selon leur offre tarifaire suivie .....	30
2.17	Répartition des clients selon leur âge .....	31
2.18	Répartition des clients selon leur période d'abonnement .....	31
2.19	Répartition des clients selon leur nombre d'appel pour chaque type d'appel .....	32
2.20	Répartition des clients selon leurs durées des appels pour chaque type d'appel .....	33

2.21	Répartition des clients selon le coût des appels pour chaque type d'appel .....	34
2.22	Répartition du churn selon nombre de réclamations .....	34
2.23	Répartition du churn selon l'âge.....	35
2.24	Répartition du churn selon l'état civile .....	36
2.25	Répartition du churn selon le nombre de réclamations .....	36
2.26	Répartition du churn selon l'activation des messages vocaux .....	37
2.27	Répartition du churn selon le nombre des messages vocaux .....	37
2.28	Répartition du churn selon la période d'abonnement en jours.....	38
2.29	Boîte à moustaches du coût d'appel des différents types d'appels .....	39
2.30	Comparaison de la durée d'appel jour/nuit/soirée entre clients churn et non churn. ....	40
3.1	Les valeurs unique de chaque colonne .....	43
3.2	Code et résultat de suppression de colonne inutiles "id_client" et "num_tel" .....	43
3.3	Nombre des valeurs manquantes par colonne .....	44
3.4	Traitemennt des valeurs manquante de colonne « nb_jours_abonne » .....	45
3.5	Traitemennt des valeurs manquante de colonne « nb_reclamation » .....	46
3.6	Traitemennt des valeurs manquante des colonnes du nombre d'appels .....	46
3.7	Traitemennt des valeurs manquante de colonne « nb_appel_inter » .....	47
3.8	Traitemennt des valeurs manquante de colonne « active_msg_vocaux ».....	47
3.9	Traitemennt des valeurs manquante de colonne « churn » .....	47
3.10	Vérification des valeurs manquantes .....	48
3.11	Nombre des valeurs dupliquées .....	49
3.12	Vérification sur les types des données .....	50
3.13	Application du Label Encoder sur les colonnes «active_msg_vocaux» et «churn» .....	50
3.14	Encodage de la colonne 'offre_type' avec la moyenne cible de 'churn'.....	51
3.15	Dataset avant l'encodage .....	51
3.16	Résultat de l'encodage .....	52
3.17	Code de standardisation .....	53
3.18	La matrice de corrélation de la Dataset .....	54
3.19	Corrélation entre « churn » et les autres variables .....	55

4.1	Extraction de Feature et Target .....	57
4.2	Division de base de données .....	58
4.3	Le fonctionnement de Random Forest [47] .....	59
4.4	Le fonctionnement d'Arbre de décision .....	60
4.5	Modèle de régression logistique [41] .....	62
4.6	Algorithme K-nearest neighbors [43].....	63
4.7	Le processus de XGBoost [45] .....	65
4.8	Le processus de Naïve Bayes classifier [50] .....	67
4.9	La matrice de confusion .....	69
4.10	Lecture de courbe ROC .....	71
4.11	Exemple de courbe Learning Curve .....	72
5.1	Matrice de confusion de Random Forest .....	76
5.2	Precision-Recall Curve de Random Forest .....	77
5.3	Learning curve de Random Forest .....	77
5.4	Matrice de confusion de Decsion Tree .....	78
5.5	Learning curve de Decsion Tree .....	79
5.6	Matrice de confusion de Logistic Regression .....	79
5.7	Learning curve de Logistic Regression .....	80
5.8	Matrice de confusion de XGBoost .....	81
5.9	Learning curve de XGBoost .....	81
5.10	Matrice de confusion de KNN .....	82
5.11	Learning curve de KNN .....	83
5.12	Matrice de confusion de Naïve Bayes .....	84
5.13	Learning curve de Naïve Bayes .....	84
5.14	Comparaison des performances des différents modèles .....	85
5.15	Résultat de la méthode Voting Classifier .....	86
5.16	Résultat de la méthode Stacking Classifier .....	86
5.17	Diagramme ROC/AUC de tous les modèles .....	87
5.18	Comparaison entre le score du partie Train et Test .....	88

6.1	Architecture de déploiement du modèle . . . . .	90
6.2	Méthode d'enregistrement le modèle . . . . .	91
6.3	La structure de la partie développement . . . . .	91
6.4	Chargement du modèle flask . . . . .	92
6.5	Code source de fonction "predict" . . . . .	93
6.6	Script de l'exécution de l'application . . . . .	94
6.7	L'interface de prédiction . . . . .	95
6.8	Résultat de prédiction positive . . . . .	96
6.9	Résultat de prédiction négative . . . . .	96
6.10	Interface de tableau de bord 1 . . . . .	97
6.11	Interface de tableau de bord 2 . . . . .	98
6.12	Interface de tableau de bord 3 . . . . .	99
6.13	Diagramme de Gantt . . . . .	100

# Liste des Tableaux

1.1	La comparaison des méthodologies .....	11
1.2	Les logiciels utilisés .....	12
1.3	Les bibliothèques utilisées .....	13
1.4	Les langages de programmation utilisées .....	15
1.5	Environnement matériel .....	16
2.1	Les variables de la base de données.....	20
4.1	Différence entre arbre de décision et Random Forest .....	61
4.2	Les formules des distances.....	64
4.3	Les hyperparamètres de chaque modèle .....	73

## Liste des abréviations

- AUC = Area Under the Curve
- BI = Business Intelligence
- CRISP-DM = Cross-Industry Standard Process for Data Mining
- ISAMM= Institut Supérieur des Arts Multimédia de la Manouba
- GBM = Gradient Boosting Machine
- IA = Intelligence Artificielle
- KNN = K-Nearest Neighbors
- ML = Machine Learning
- ROC = Receiver Operating Characteristics
- SAS = Statistical Analysis System
- SEMMA = Sample Explore Modify Model Assess
- TDSP = Team Data Science Process

# Introduction générale

De nos jours, le secteur des télécommunications connaît une concurrence intense et une évolution technologique rapide. Cette situation a un impact considérable sur le taux de résiliation des clients, connu sur le nom de churn, ce qui en fait une préoccupation majeure pour ce secteur. Le churn se produit lorsque les clients décident de résilier leur abonnement et de passer à un autre fournisseur de services téléphoniques.

Pour les entreprises de télécommunications, il est devenu essentiel de mettre en place une gestion efficace des relations clients afin d'accroître leurs revenus. En fait, la perte de clients ou d'abonnés reste un défi majeur pour l'industrie des télécommunications, car les clients n'hésitent pas à se désabonner ou changer d'opérateur s'ils ne sont pas satisfaits.

De nombreuses études ont démontré l'efficacité du Machine Learning dans la prévision de cette situation. L'intelligence artificielle joue un rôle crucial dans la prédiction du churn des clients. En exploitant les techniques d'apprentissage automatique, il est possible de traiter les données et de développer des modèles de prédiction du taux de désabonnement. Ces modèles aident les opérateurs de télécommunications à identifier les clients les plus susceptibles de résilier leur abonnement, ce qui leur permet de prendre des mesures préventives pour les retenir.

Actuellement, l'opérateur Tunisie Télécom ne dispose d'aucun système de prédiction du churn. C'est pour cette raison que les dirigeants de Tunisie télécom ont décidé de trouver une solution informatisée pour le problème de l'arrition des clients. Alors, dans le but de résoudre ce problème, notre projet de fin d'étude intitulé « Prédiction du churn des clients Tunisie Télécom » a été initié. Ce projet s'inscrit dans le cadre d'une Licence en Big Data et Analyse de données à l'ISAMM. Il consiste à proposer un modèle de prédiction de churn pour identifier les clients les plus susceptibles à arrêter leurs abonnements de ligne téléphonique avec Tunisie Télécom en se basant sur des techniques de Machine Learning et de Data Science. Par la suite, nous évaluerons les performances de ce modèle en analysant les résultats de la prédiction, en suivant la méthodologie de travail CRISP-DM.

Le présent rapport comporte cinq chapitres :

- Le premier chapitre de notre rapport, intitulé "Étude de projet et compréhension du métier ", est consacré à la présentation du cadre de notre projet, l'organigramme d'accueil, la problématique que nous traitons, la solution que nous proposons, ainsi que les méthodologies utilisées en Data Science et les

2 outils adoptés pour la réalisation du projet. Ensuite, il se focalise sur les techniques de base de la Data Science.

- Le deuxième chapitre, intitulé "Compréhension des données", se concentre sur l'exploration et de l'analyse des données de notre data base.
- Le troisième chapitre, intitulé "Prétraitement des données", présente en détail toutes les étapes nécessaires pour prétraiter et nettoyer les données avant de les modéliser.
- Le quatrième chapitre, intitulé "La modélisation", nous explorons en détail les différents modèles et les mesures de performance appropriées à appliquer dans notre projet.
- Le cinquième chapitre, intitulé "Évaluation, optimisation de la performance des modèles et déploiement du modèle choisi", se concentre sur les résultats obtenus dans chaque modèle, ainsi que la comparaison en utilisant les mesures nécessaires.
- Le dernier chapitre, intitulé "Déploiement", se focalise sur la création d'une interface web et la construction d'un tableau de bord afin de rendre notre projet facilement accessible aux utilisateurs.

Finalement, ce rapport est clôturé par une conclusion générale et quelques perspectives.

# Chapitre 1 : Étude du projet

## Plan

1.	Introduction .....	4
2.	Présentation de l'organisme d'accueil.....	4
3.	Présentation du projet.....	6
4.	Méthodologie de gestion de projet adoptée.....	7
5.	Architecture de projet.....	11
6.	Environnement de travail.....	12
7.	Conclusion .....	16

## 1.1 Introduction

Dans ce chapitre, nous exposons l'objectif de notre projet à travers le contexte général des télécommunications en Tunisie, en mettant l'accent sur Tunisie Télécom. Nous débuterons par présenter l'organisme d'accueil, puis la problématique spécifique et la solution proposée. Ensuite, nous décrirons brièvement l'architecture du projet, la méthodologie adaptée et l'environnement de travail. Enfin, nous aborderons les concepts essentiels du projet.

## 1.2 Présentation de l'organisme d'accueil

### 1.2.1 Présentation du Tunisie Télécom

Tunisie Télécom est une entreprise de télécommunications tunisienne qui fournit des services de téléphonie fixe et mobile, d'internet et de transmission de données. Actuellement, c'est l'un des plus grands opérateurs des télécommunications de la région. Tunisie Télécom a été fondée en 1995 et est ouverte au Grand public qu'aux entreprises et opérateurs tiers. [1]



Figure 1.1 : Logo de Tunisie Télécom [1]

### 1.2.2 Historique

L'Agence nationale des télécommunications a été créée par la promulgation de la loi n° 36 du 17 avril 1995. Le bureau a ensuite changé son statut juridique pour devenir une société anonyme en vertu du décret n° 30 du 5 avril 2004 dénommé "Tunisie Télécom". En juillet 2006, le capital de Tunisie Télécom a cédé 35% de son capital au profit de consortium émirati « TeCom-DIG ». L'entreprise vise à accroître la rentabilité de Tunisie Télécom et à en faire l'un des principaux opérateurs internationaux.

### 1.2.3 Organisation :

Tunisie Télécom est composée de 24 directions régionales, 80 Actuels et points de vente, Plus de 13 000 emplacements privés. Elle emploie plus de 8 000 agents. Cet opérateur historique dispose aussi six centres de support clients de téléphonie fixe et Mobile et données. La figure 1.2 présente l'organisation fonctionnelle de Tunisie Télécom.

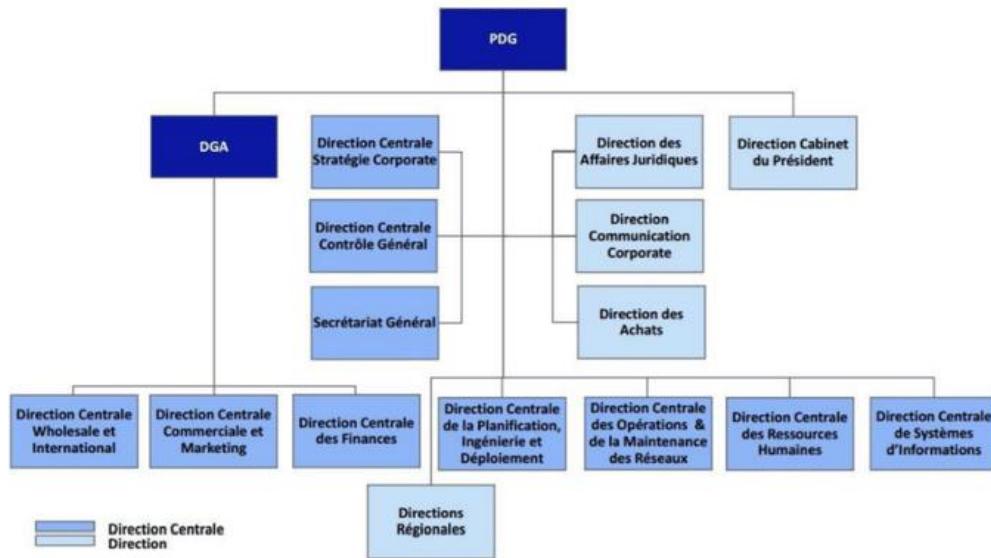


Figure 1.2 : Organisation fonctionnelle de Tunisie Télécom [2]

### 1.2.4 Présentation du complexe HACHED

Le complexe HACHED est l'un des complexes les plus importants de Tunisie Télécom. Il est caractérisé par une grande variété de services :

- Centre de transmission numérique CTN.
- Centre de Commutation d'Abonnées CCA.
- Centres d'exploitation et de maintenance des réseaux radio national OMC (operating maintenance center) et régional.
- Salle de MSC (Mobile service Switching Center).
- Centre de donnée.

## **1.3 Présentation du projet**

### **1.3.1 Cadre général du projet**

Ce projet se présente dans le cadre d'un projet fin d'étude à l'Institut Supérieur des Arts Multimédias pour le but d'obtenir le diplôme d'une Licence en Big Data et Analyse de données. Le stage est effectué au sein de la société Tunisie Télécom.

### **1.3.2 Étude de l'existant :**

La phase d'étude de l'existant revêt une grande importance dans le projet. Elle permet d'obtenir une vision claire et précise des concepts de l'environnement de travail afin d'identifier les besoins et de déterminer les problématiques à résoudre. L'objectif primordial de Tunisie Telecom est de satisfaire ses clients, ce qui nécessite une approche basée sur la compréhension de leurs besoins et de leurs comportements d'achat. Dans le domaine des télécommunications, les clients recherchent généralement les produits ou services qui leur procurent le maximum de satisfaction.

Pour assurer la satisfaction de ses clients, Tunisie Telecom utilise différentes méthodes motivées par diverses raisons, notamment la détection des causes d'insatisfaction et des dysfonctionnements potentiels, ainsi que l'évaluation des opinions des clients. Actuellement, ces méthodes sont mises en œuvre de manière manuelle, où une équipe dédiée analyse les retours des clients et identifie les points d'amélioration nécessaires.

Actuellement, Tunisie Telecom met en place une démarche qualité visant à mesurer la satisfaction de ses clients et à fournir un diagnostic précis de leur satisfaction et insatisfaction. Ces évaluations conduisent à des actions concrètes pour améliorer le niveau de satisfaction.

Ainsi, la mesure de la satisfaction client devient un outil essentiel pour guider l'entreprise dans sa démarche de satisfaction client, en permettant de déterminer dans quelle mesure les clients sont satisfaits des services et d'évaluer le niveau de réponse de l'entreprise aux attentes des clients.

### **1.3.3 Problématique**

Tunisie Télécom, en tant que premier opérateur de télécommunication en Tunisie, propose différentes offres et une variété de services pour répondre aux besoins de sa clientèle. Malgré les services proposés, certains clients ne sont pas satisfaits et ont même décidé de résilier leur abonnement et de passer à un opérateur concurrent, ce qui représente un problème critique car le taux de désabonnement entraîne une baisse des revenus et une diminution de la part de marché.

Les raisons de résiliation incluent les coûts élevés des appels, des problèmes de couverture réseau et un service client insatisfaisant. Ces facteurs peuvent influencer leur décision de choisir un autre opérateur.

**Comment pouvons-nous donc trouver les clients qui peuvent résilier leurs contrats et quittent l'opérateur Tunisie Télécom ?**

### 1.3.4 Solution

Tunisie Télécom s'engage à développer des moyens permettant de prédire le taux de désabonnement de ses clients. L'entreprise cherche à identifier les différents segments de sa clientèle, car le coût d'acquisition d'un nouveau client est généralement plus élevé que celui de fidélisation d'un client existant.

La solution envisagée consiste à créer un modèle de prédiction du taux de désabonnement en utilisant divers algorithmes d'apprentissage supervisé. L'objectif est de sélectionner le meilleur modèle en fonction de son taux de précision, en analysant les données historiques des clients et en visualisant les résultats de la prédiction ainsi que les performances du modèle. Les modèles seront évalués en termes de précision dans la prédiction des clients résiliant leur abonnement et de ceux qui décident de rester.

Un data scientist recherche toujours la satisfaction de l'utilisateur. Dans ce contexte, notre meilleur modèle de prédiction sera intégré dans une page web simple et facile à utiliser.

## 1.4 Méthodologie de gestion de Projet adoptée

Avant de commencer la réalisation de notre projet, il est crucial d'examiner et d'évaluer la méthode de gestion de projet à adopter afin de garantir une coordination efficace entre les parties prenantes et les différentes tâches. Alors, nous allons étudier trois méthodologies qui sont : SEMMA, TDSP et CRISP-DM. Notre objectif est d'analyser chaque méthode afin de choisir celle qui conviendra le mieux à notre projet.

### 1.4.1 Méthodologie SEMMA

*SEMMA* est une méthode utilisée en Data Mining pour résoudre des problèmes d'analyse de données en cinq étapes : échantillonnage, exploration, modification, modélisation et évaluation. Cette méthodologie a été développée par SAS pour aider les analystes à utiliser les données pour résoudre des problèmes commerciaux et analytiques. [3]

1. *Échantillonner les données (Sample)* : Cette étape consiste à choisir les données pertinentes qui seront utilisées pour l'analyse et d'en extraire un échantillon représentatif.

2. *Explorer et visualiser les données (Explore)* : Exploration et visualisation des données pour comprendre

leur structure et permet de détecter les anomalies et identifier les relations entre les variables.

3. *Modifier et nettoyer les données (Modify)* : Nettoyage et préparation des données pour l'analyse.
4. *Modéliser les données (Model)* : Construction des modèles de Machine Learning pour prédire ou expliquer les phénomènes étudiés.
5. *Analyser les résultats (Assess)* : L'évaluation des résultats obtenus à l'aide des modèles et les interpréter en les comparant à des critères de performance prédéfinis.

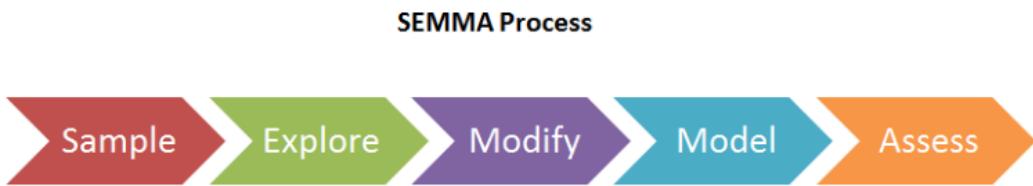


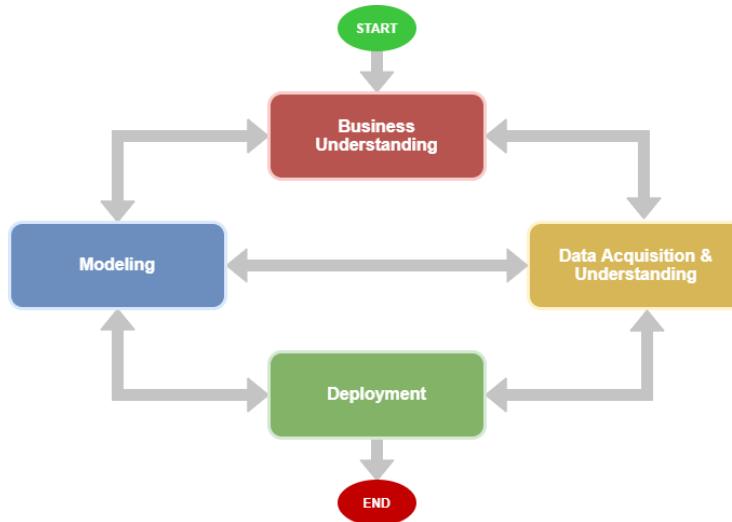
Figure 1.3 : Le cycle de vie de SEMMA [4]

#### 1.4.2 Méthodologie TDSP

Le TDSP est une méthodologie de gestion de projet agile et itérative développée spécifiquement pour la science des données. Il a été créé par Microsoft, pour faciliter la collaboration entre les membres d'une équipe. [5]

La méthode TDSP se divise en cinq étapes :

1. *Business Understanding* : L'objectif de cette première phase est de bien comprendre les enjeux métiers, d'identifier les objectifs du projet ainsi que les critères de réussite et les parties prenantes impliquées.
2. *Data Acquisition and Understanding* : Collecte, nettoyage et préparation des données nécessaires pour répondre aux objectifs métier.
3. *Modeling* : Au cours de cette phase, des modèles de données sont développés et testés pour répondre aux objectifs métiers du projet et sélectionner le modèle le plus performants.
4. *Deployment* : Une fois que les modèles ont été créés, ils sont prêts à être déployés dans un environnement opérationnel pour l'utilisation.
5. *Acceptance* : Cette étape a pour but de mesurer les résultats du déploiement des modèles pour vérifier la réussite du projet et la réalisation des objectifs métier.



**Figure 1.4 :** Le cycle de vie de TDSP [5]

### 1.4.3 Méthodologie CRISP-DM

CRISP-DM a été publié en 1999 pour normaliser les processus d'exploration de données dans tous les secteurs. C'est devenu la méthode la plus populaire pour les projets d'exploration de données, d'analyse et de science des données. [6]

Le cycle de vie de CRISP-DM est divisé en 6 étapes :

1. *Business Understanding* : La première étape est d'identifier le problème que l'organisation est en train d'essayer de résoudre en se basant sur données et établir une architecture bien définie pour la mise en œuvre du projet.
2. *Data UnderStanding* : Cette étape comprend la collecte des informations initiales, la compréhension et la description du type de données à analyser et établir des liens entre les données et leur signification d'un point de vue métier.
3. *Data Preparation* : Cette phase a pour but de préparer les données à analyser. Il consiste notamment à nettoyer les données, à les transformer afin qu'ils soient compatibles avec les algorithmes qui seront utilisés.
4. *Modeling* : La modélisation est basée sur le choix de paramétrage ainsi que le modèle qui sera utilisé après l'essai de plusieurs modèles.

Cette étape comprend 4 fonctions : sélection de la technique de modélisation, conception du prototype, construction du modèle, évaluation du modèle.

5. *Evaluation* : Avant de passer au déploiement définitif du modèle, il faut contrôler et vérifier les modèles

ou les connaissances obtenues pour s'assurer qu'ils atteignent les objectifs énoncés dès le début du processus, permet aussi de prendre la décision de déploiement du modèle ou de l'améliorer.

6. *Deployment* : C'est la phase finale du processus. Elle s'agit de déployer les analyses pour une utilisation effective. Son objectif est de mettre en forme les connaissances obtenues par la modélisation et les intégrer dans le processus de prise de décision.

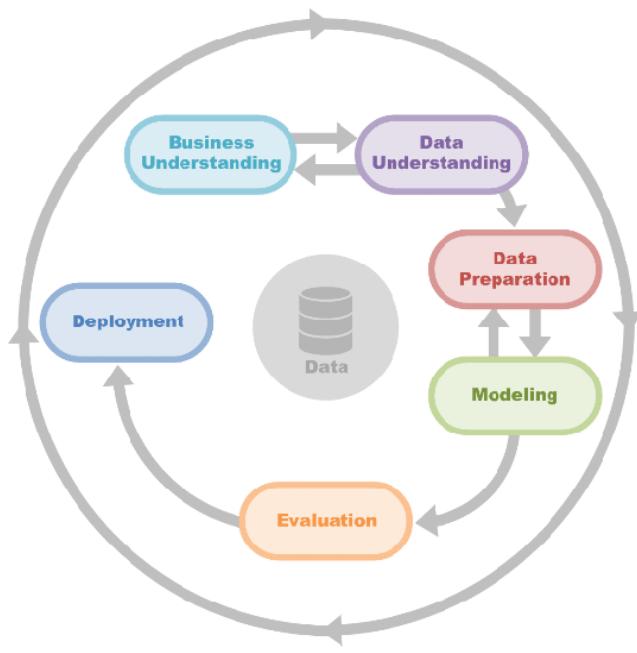


Figure 1.5 : Le cycle de vie de CRISP-DM [7]

#### 1.4.4 Comparaison des méthodologies

Les trois méthodologies de gestion de projet de science des données SEMMA, TDSP et CRISP-DM ont le même objectif, mais il y a quelques différences entre eux.

Le tableau 1.1 présente la comparaison entre ces trois méthodologies.

Tableau 1.1 : La comparaison des méthodologies

Critère	SEMMA	TDSP	CRISP-DM
Domaine d'application	Data Mining et analyse prédictive	Projet de science des données en équipe	Data Mining et exploration de données
Phases	5 phases	5 phases	6 phases
Approche itérative	Non	Oui	Oui
Flexibilité	Faible	Moyenne	Élevée
Priorités	Accent sur l'exploitation et la modification	Accent sur l'acquisition et la préparation	Accent sur la compréhension et la préparation

Il est crucial de choisir la méthodologie de gestion de projet adéquate pour garantir le succès d'un projet de science des données. Après une comparaison des trois méthodologies disponibles, nous avons opté pour **CRISP-DM** en raison de sa souplesse et de son approche itérative, qui permet une adaptation aux changements tout au long du projet.

## 1.5 Architecture de Projet

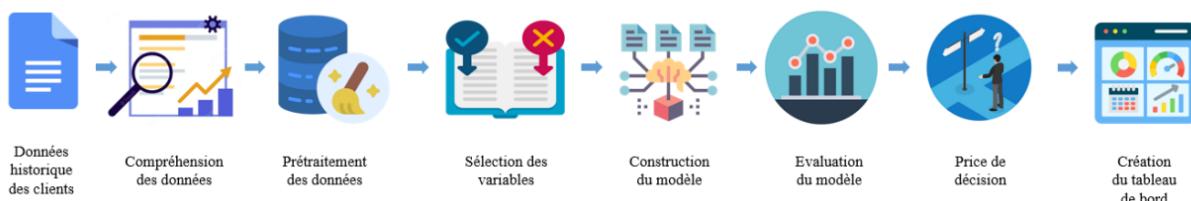


Figure 1.6 : Architecture de projet

Notre système de prédiction de churn repose sur une architecture composée de huit étapes essentielles, assurant ainsi des prévisions précises et efficaces :

- 1- Collecte des données des clients.
- 2- Compréhension des données clients.
- 3- Prétraitement des données clients.
- 4- Sélection des fonctionnalités pertinentes pour le développement des modèles de prédiction.
- 5- Construction des modèles.
- 6- Évaluation et calcul de la précision des modèles.
- 7- Prise de décision pour la rétention des clients.
- 8- Création du tableau de bord.

## 1.6 Environnement de travail

Dans cette section, nous déterminons les outils logiciels et matériels qui ont été utilisés pour la réalisation et la mise en place de notre projet.

### 1.6.1 Environnement logiciel

Nous allons présenter les différents logiciels, langage de programmation et bibliothèques utilisés.

#### 1.6.1.1 Logiciels utilisés

Le tableau 1.2 représente les logiciels utilisés.

**Tableau 1.2 : Les logiciels utilisés**

Environnement	Description
 <b>ANACONDA</b>	ANACONDA : Il s'agit d'un environnement de distribution de logiciels et d'une plateforme de gestion de packages pour les langages de programmation Python et R, spécifiquement adaptés au développement d'applications dans le domaine de la science des données et de l'apprentissage automatique. [8]
	JUPYTER NOTEBOOK : Un notebook de calcul open source permet de créer et de partager des documents qui contiennent du code interactif, des visualisations de données et des textes explicatifs. [9]

	POWER BI DESKTOP : Plateforme de Business Intelligence qui permet aux utilisateurs de collecter, d'analyser et de visualiser des données provenant de multiples sources. Il permet de créer des tableaux de bord ainsi que des visualisations de données. [10]
	VISUAL STUDIO CODE : Un éditeur de code utilisé pour écrire, modifier et déboguer du code dans différents langages de programmation. Il offre une gamme de fonctionnalités pour faciliter le processus de développement logiciel. [11]
	pgAdmin : pgAdmin est une plateforme d'administration open-source pour les bases de données PostgreSQL. Elle fournit une interface graphique conviviale et intuitive qui permet aux administrateurs et aux développeurs de gérer facilement et efficacement leurs bases de données PostgreSQL. [12]

### 1.6.1.2 Bibliothèques utilisés

Le tableau 1.3 représente les bibliothèques utilisées.

Tableau 1.3 : Les bibliothèques utilisés

Bibliothèque	Description
PANDAS	Un package open source pour Python qui se concentre sur la fourniture de structures de données et d'outils d'analyse de données performants pour l'analyse, le nettoyage et la préparation des données dans le domaine de la science des données. [13]
NUMPY	Une bibliothèque open source permet de manipuler des tableaux et des matrices de données multidimensionnelles. Elle est réputée pour son efficacité, sa rapidité et sa facilité d'utilisation. [14]

SEABORN	Bibliothèque de visualisation de données basée sur Matplotlib, qui offre une interface de haut niveau pour la création de graphiques statistiques à la fois attrayants et informatifs. [15]
MATPLOTLIB	Bibliothèque de visualisation de données qui permet de tracer différents types de graphiques, tels que des histogrammes, des graphiques à barres, des diagrammes de dispersion, etc. [16]
SKLEARN	Bibliothèque Python fournit une variété d'outils et d'algorithmes pour créer, former et évaluer des modèles d'apprentissage automatique. [17]
FLASK	Framework de développement web en Python qui permet de créer des applications web de manière rapide et simple. Il fournit les fonctionnalités de base nécessaires pour gérer les requêtes et les réponses HTTP, gérer les routes URL, etc. [18]
PICKLE	Ce module permet la sérialisation et la désérialisation d'objets Python, convertissant ainsi ces objets en un flux de bytes pour le stockage ou le transfert de données, facilitant ainsi la sauvegarde et le chargement de données complexes dans les applications. [19]
IPYWIDGETS	C'est une bibliothèque Python pour créer des widgets interactifs dans les notebooks Jupyter et les applications web. Ils rendent l'exploration et l'analyse de données plus dynamiques. [20]

La figure 1.7 présente quelques bibliothèques utilisées :



Figure 1.7 : Les bibliothèques utilisées [21]

### 1.6.1.3 Langages utilisés

Le tableau 1.4 représente les langages utilisés.

Tableau 1.4 : Les langages utilisés

Langage	Description
	PYTHON : Un langage de programmation multiplateforme, interprété, orienté objet et de haut niveau. Il est utilisé dans de nombreux domaines différents tels que la science des données, l'intelligence artificielle, la visualisation de données, etc. [22]
	HTML : (HyperText Markup Language) est un langage de balisage standard utilisé pour créer et structurer le contenu d'une page web. Il définit la structure et le contenu des éléments d'une page web en utilisant des balises prédéfinies.[23]
	CSS : (Cascading Style Sheets) est une feuille de style css est désignée pour la réalisation de la charte graphique de plateforme. Il permet de contrôler les couleurs, les polices, les marges, les positions, les arrière-plans. [24]
	JAVASCRIPT : Langage de script orienté objet principalement utilisé dans les pages HTML. Il permet d'ajouter des animations, des formulaires dynamiques, des manipulations du DOM (Document Object Model), des requêtes AJAX, des interactions avec des API. [25]
	BOOTSTRAP : Framework CSS gratuit pour un développement web plus rapide et plus facile permet de faire la présentation graphique, comprend des modèles HTML, CSS et JavaScript. [26]

### 1.6.2 Environnement matériel

Pour la réalisation de projet, nous avons utilisé une machine avec les caractéristiques suivantes :

**Tableau 1.5 : Environnement matériel**

<b>Marque</b>	Lenovo Legion
<b>Processeur</b>	Intel® Core™ i7-9750H CPU @ 2.60GHz 2.59GHz
<b>Mémoire RAM</b>	16 Go
<b>Disque Dur</b>	256 GB
<b>Système d'exploitation</b>	Microsoft Windows 10 Professional, 64 bits

## 1.7 Conclusion

Dans ce premier chapitre, nous avons présenté l'organisme de travail. Par la suite nous avons définis la problématique de notre sujet. Nous avons spécifié la solution la plus adéquate pour résoudre cette problématique. Puis nous avons définis la méthodologie de travail à adopter tous au long de la réalisation de projet de fin d'étude.

# Chapitre 2 : Compréhension du problème métier et des données

## Plan

1.	Introduction .....	18
2.	Notions théoriques liées au projet .....	18
3.	Les défis de la modélisation en Machine Learning .....	19
4.	Choix de la base de données .....	20
5.	Exploration des données .....	23
6.	Analyse des données .....	27
7.	Conclusion .....	40

## 2.1 Introduction

Dans ce chapitre, nous allons présenter les concepts fondamentaux nécessaire à la compréhension de notre projet. Nous présentons ensuite une analyse détaillée les données collectés nécessaire à la réalisation du projet. Nous mettrons en œuvre des techniques de visualisation pour explorer et interpréter les données, dans le but d'enrichir leur compréhension.

## 2.2 Notions théoriques liées au projet

Dans cette section, nous allons explorer les différents concepts liés à notre projet. Les principaux concepts qui nous intéressent sont le churn, l'apprentissage automatique et les défis de la modélisation. Nous allons commencer par la définition du churn.

### 2.2.1 Churn

Selon B. Bathelot Churn est un terme anglais qui désigne la perte de clientèle ou d'abonnées. Le taux de churn, ou taux d'attrition, est un indicateur important pour mesurer la performance des services d'une entreprise. Il représente le pourcentage de clients qui cessent d'utiliser les produits ou services de l'entreprise sur une période donnée. Cet indicateur permet de déterminer si le service proposé par l'entreprise est en phase avec les attentes des clients ou non. [27]

$$\text{Taux de churn} = \frac{\text{Nombre de clients perdus}}{\text{Nombre de clients total}}$$

**Equation 1 :** Formule de taux de churn

Les clients churn de télécommunication peuvent être divisés en deux catégories principales : involontaire et volontaire.

- **Client volontaire** : Décrit comme la fin du service par l'abonnée.
- **Client involontaire** : Les abonnées que la société de télécommunications décide de supprimer pour diverses raisons, notamment la fraude et le non-paiement de leurs factures.

## 2.2.2 Apprentissage automatique

L'apprentissage automatique, ou Machine Learning en anglais, est un domaine de l'intelligence artificielle axé sur le développement de modèles qui permettent à une machine de comprendre les problèmes métiers et les données, en apprenant à partir de ces données pour s'améliorer de manière automatique.[28]

Les techniques d'apprentissage automatique peuvent être regroupées en deux catégories :

- Supervisé
- Non-supervisé

La figure présentée ci-dessous représente la différence entre les 2 catégories

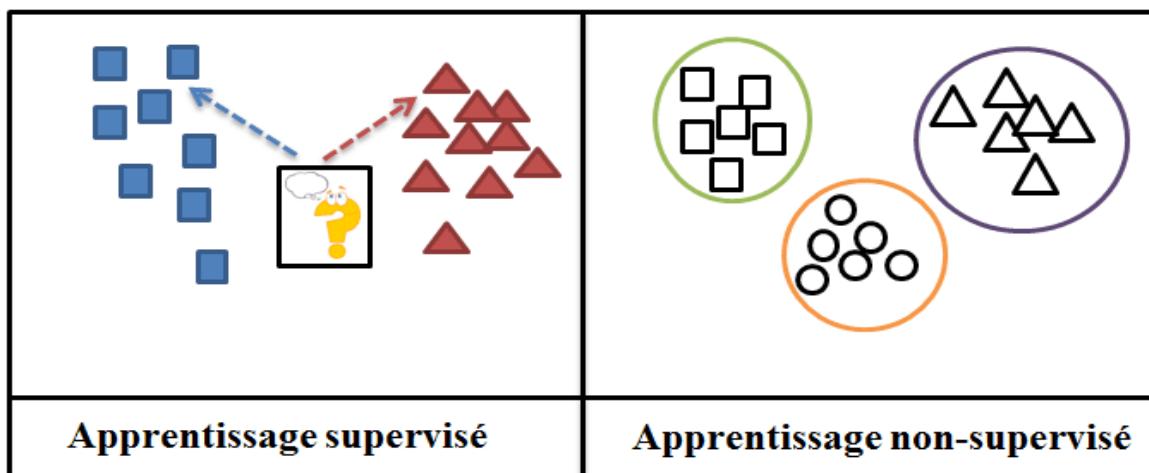


Figure 2.1 : Différence entre apprentissage supervisé et non-supervisé [29]

## 2.3 Les défis de la modélisation en Machine Learning

Lors de l'entraînement d'un modèle de Machine Learning, il y a deux problèmes majeurs qui peuvent se présenter : Surapprentissage (overfitting) et sous-apprentissage (underfitting).

### 2.3.1 Overfitting

L'overfitting se produit lorsqu'un modèle s'adapte trop étroitement aux données d'entraînement, mémorisant le bruit et les erreurs et ne peut pas faire de prédictions précises sur de nouveaux exemples. Le modèle peut afficher de bons résultats sur les données d'entraînement, mais de mauvaises performances sur les données de test ou de nouvelles données. [30]

### 2.3.2 Underfitting

Lorsqu'un modèle n'a pas une capacité suffisante pour apprendre à partir des données d'entraînement, cela peut mener à un underfitting. En conséquence, il peut ne pas être en mesure de généraliser efficacement à de nouvelles données, ce qui entraîne de mauvais résultats sur les données de test. [30]

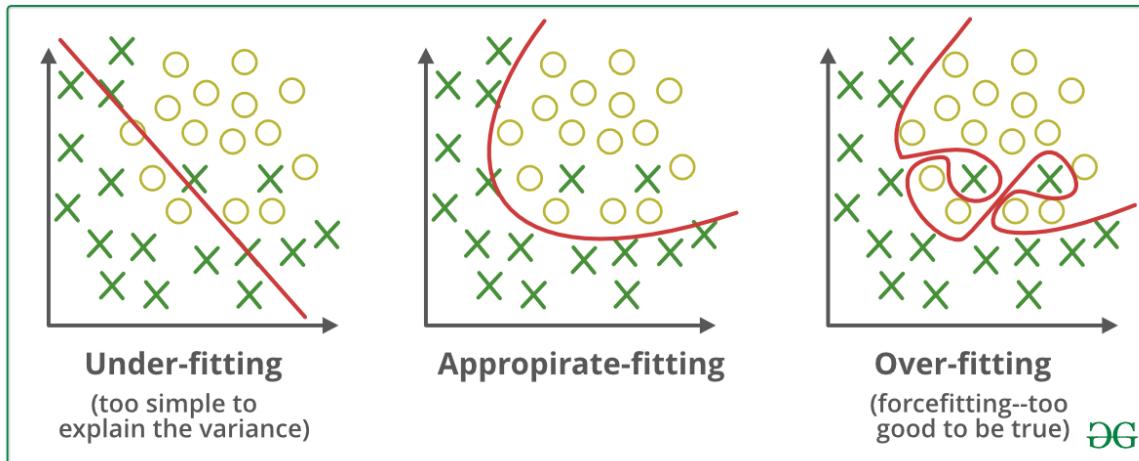


Figure 2.2 : Les problèmes de la modélisation [31]

### 2.4 Choix de la base de données

Notre dataset qui a été collecté par les services de Tunisie Télécom et que nous utilisons dans cette recherche comprend des renseignements sur l'historique des opérations d'appels nationaux et internationaux, les messages vocaux et les réclamations des clients de Tunisie Télécom, ainsi que leurs données démographiques et leur numéro téléphonique.

Le tableau 2.1 donne une vue d'ensemble des données. Il identifie les variables, leurs types et une simple description de leur signification.

Tableau 2.1 : Les variables de la base de données.

Variable	Description	Type
id_client	Identifiant du client	Chaine de caractère
genre	La catégorisation sociale des individus en fonction de leur sexe	Numérique
age	L'âge de chaque client	Numérique
marié	L'état civil de chaque client	Numérique
num_tel	Le numéro téléphonique pour chaque client	Numérique
nb_jours_abonne	Nombre de jours que le client a été abonné à Tunisie Télécom	Numérique
durée_appel_jour (minutes)	La durée totale des appels effectués pendant la journée (en minutes)	Numérique
nb_appel_jour	Le nombre total d'appels effectués pendant la journée	Numérique
cout_appel_jour	Le coût total des appels effectués pendant la journée.	Numérique
durée_appel_soirée (minutes)	La durée totale des appels effectués pendant la soirée (en minutes)	Numérique
nb_appel_soirée	Nombre de jours que le client a été abonné à Tunisie Télécom	Numérique
cout_appel_soirée	Le coût total des appels effectués pendant la soirée	Numérique
durée_appel_nuit (minutes)	La durée totale des appels effectués pendant la nuit (en minutes)	Numérique
nb_appel_nuit	Le nombre total d'appels effectués pendant la nuit	Numérique
cout_appel_nuit	Le coût total des appels effectués pendant la nuit	Numérique
durée_appel_inter (minutes)	La durée totale des appels internationaux (en minutes)	Numérique
nb_appel_inter	Le nombre total d'appels internationaux	Numérique
cout_appel_inter	Le coût total des appels internationaux	Numérique
active_msg_vocaux	Indique si le client a activé la messagerie vocale ('Yes', 'No')	Chaine de caractère
nb_msg_vocaux	Le nombre de messages vocaux reçus par le client	Numérique
nb_reclamation	Le nombre des réclamations déposées par le client	Numérique
churn	False si le client est fidèle, True si le client churn	Chaine de caractère
offer_type	Le type d'offre tarifaire choisis par le client	Chaine de caractère

Notre corpus de données contient 23 variables dont 5 variables "id\_client", "marie" "offer\_type",

"active\_msg\_vocaux" et "churn" qui sont des variables de type "Chaîne de caractères".

id_client	genre	age	marie	num_tel	nb_jours_abo	duree_appel_nb	appel_jou	cout_appel_j	duree_appel_nb	appel_so	cout_appel_s	duree_appel_nb	appel_nu	cout_appel_n	duree_appel_nb	appel_int	cout_appel_i	active_msg_v	nb_msg_v	nb_reclamati	churn	offer_type
382-4657	femme	37 yes	98505453	128	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10	5	2.7 yes	25	1	False	Hayy+		
371-7191	homme	46 no	97321658	107	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7 yes	26	1	False	PRE-1=11		
358-1921	homme	50 no	98653270	137	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32	12.2	5	3.29 no	0	0	False	PRE-900 bonus		
375-9999	homme	78 yes	96303256	84	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78 no	0	2	False	PRE-AHLA		
330-6626	femme	75 yes	96412387	75	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73 no	0	3	False	PRE-Binetra		
391-8027	femme	23 no	98142367	118	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.7 no	0	0	False	PRE-Classic		
355-9993	femme	67 yes	98501260	121	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	2.03 yes	24	3	False	PRE-Club Optimum Plus		
329-9001	homme	52 yes	99606321	147	157	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92 no	0	0	False	PRE-Corporate Optimum Family		
355-4719	femme	68 no	99421753	117	184.5	97	51.37	351.6	80	29.89	215.8	90	9.71	8.7	4	2.35 no	0	1	False	PRE-CSS 1000% New		
330-8173	femme	43 yes	99203170	141	258.6	84	43.96	222	111	18.87	326.4	97	14.69	11.2	5	3.02 yes	37	0	False	PRE-CSS Mobile 1000%		
329-6603	homme	47 no	95606231	65	129.1	137	21.95	228.5	85	19.42	208.8	111	9.4	12.7	6	3.43 no	0	4	True	PRE-CSS 35ml/min		
344-9403	femme	25 yes	98741230	74	187.7	127	31.91	163.4	148	15.89	196	94	8.82	9.1	5	2.46 no	0	0	False	PRE-Day Pass		
363-1107	femme	58 yes	98632140	168	128.8	96	21.9	104.9	71	8.92	141.1	128	6.35	11.2	2	3.02 no	0	1	False	PRE-Double Reinsta		
394-8006	femme	32 no	96321586	95	156.6	88	26.62	247.6	75	21.05	192.3	115	8.65	12.3	5	3.32 no	0	3	False	PRE-E.M. 1000%		
366-9238	femme	39 no	99201369	62	120.7	70	20.52	307.2	76	26.11	203	99	9.14	13.1	6	3.54 no	0	4	False	PRE-Javahara 35ml		
351-7269	femme	58 yes	98512647	161	332.9	67	56.59	317.8	97	27.01	160.6	128	7.23	5.4	9	1.46 no	0	4	True	PRE-Ellisa 300%		
350-8884	femme	52 yes	93601476	85	196.4	139	33.39	280.9	90	23.88	89.3	75	4.02	13.8	4	3.73 yes	27	1	False	PRE-Employe TT		
386-2923	femme	72 no	93602500	93	190.7	114	32.42	218.2	111	18.55	129.6	121	5.83	8.1	3	2.19 no	0	3	False	PRE-ESS 1000% New		
356-2992	homme	79 no	98503260	76	189.7	66	32.25	212.8	65	18.09	165.7	108	7.46	10	5	2.7 yes	33	1	False	PRE-ESS Mobile 35ml/min		
373-2782	homme	67 no	90032100	73	224.4	90	38.15	159.5	88	13.56	192.8	74	8.68	13	2	3.51 no	0	1	False	PRE-EST 1000% New		
396-5800	femme	79 yes	98416870	147	155.1	117	26.37	239.7	93	20.37	208.8	133	9.4	10.6	4	2.86 no	0	0	False	PRE-Javahara 35ml		
393-7984	femme	26 yes	98560321	77	62.4	89	10.61	169.9	121	14.44	209.6	64	9.43	5.7	6	1.54 no	0	5	True	PRE-New Ellissa		
358-1956	femme	30 yes	99652014	130	183	112	31.11	72.9	99	6.2	181.8	78	8.18	9.5	19	2.57 no	0	0	False	PRE-offre 40		
350-2565	femme	22 no	98505463	111	110.4	103	18.77	137.3	102	11.67	189.6	105	8.53	7.7	6	2.08 no	0	2	False	PRE-Offre WLS		
343-4696	femme	34 yes	98505459	132	81.1	86	13.79	245.2	72	20.84	237	115	10.67	10.3	2	2.78 no	0	0	False	PRE-Oulidha 1000%		
331-3693	femme	37 yes	96302514	174	124.3	76	21.13	277.1	112	23.55	250.7	115	11.28	15.5	5	4.19 no	0	3	False	PRE-Oulidha 2000%		
357-3817	femme	37 yes	98505453	57	213	115	36.21	191.1	112	15.24	182.7	115	8.22	9.5	3	2.57 yes	39	0	False	PRE-Pack cle 4G		
418-6412	homme	42 yes	98505453	54	134.3	73	22.83	155.5	100	13.22	102.1	68	4.59	14.7	4	3.97 no	0	3	False	PRE-Pass Etudiant		
359-2630	homme	64 no	98505453	20	190	109	32.3	258.2	84	21.95	181.5	102	8.17	6.3	6	1.7 no	0	0	False	PRE-Taraji Mobile 1500%		
410-7789	femme	47 yes	98505453	49	119.3	117	20.28	215.1	109	18.28	178.7	90	8.04	11.1	1	3 no	0	1	False	PRE-Tawwa		
416-8428	homme	23 yes	98505453	142	84.8	95	14.42	136.7	68	11.62	250.5	148	11.27	14.2	6	5.83 no	0	2	False	PRE-TM 35ml/min		
370-3359	femme	48 yes	98505453	75	226.1	105	38.44	201.5	107	17.13	246.2	98	11.08	10.3	5	2.78 no	0	1	False	PRE-Touriste SIM		
383-1121	homme	28 yes	98505453	172	212	121	36.04	31.2	115	2.65	293.8	78	13.2	12.6	10	3.4 no	0	3	False	PRE-Trankil ELISSA		
360-1596	femme	28 no	98505453	12	249.6	118	42.43	252.4	119	21.45	280.2	90	12.61	11.8	3	3.19 no	0	1	True	PRE-Trankil TT		
395-2854	homme	33 yes	98505453	57	176.8	94	30.06	195	75	16.58	213.5	116	9.61	8.3	4	2.24 yes	25	0	False	PRE-TT 1000%		
362-1407	femme	31 no	98505453	72	220	80	37.4	217.3	102	18.47	152.8	71	6.88	14.7	6	3.97 yes	37	3	False	PRE-TT 1500%		
341-9764	femme	37 yes	98505453	36	146.3	128	24.87	162.5	80	13.81	129.3	109	5.82	14.5	6	3.92 yes	30	0	False	PRE-TT 2000%		

Figure 2.3 : La base de données

Afin de réaliser notre étude, nous disposons d'un échantillon de 5000 clients qui ont été répartis de la manière suivante :

- 4289 : Clients fidèles (non-churn)
- 707 : Clients ayant résilié leur abonnement (churn)

Nous allons utiliser la fonction "read\_csv" de bibliothèque Pandas pour importer les données d'un fichier CSV et les stocker dans un DataFrame.

```
# importation de donnees
df = pd.read_csv("database_TT_Churn.csv")
df
```

Figure 2.4 : Importation du dataset

## 2.5 Exploration des données

L'exploration de données est une étape cruciale dans l'analyse de données qui permet de mieux comprendre les données en les examinant et en les interpréter dans leur contexte. Cette étape comprend généralement des tâches telles que l'analyse de statistiques descriptives et la visualisation de données.

- `data.shape` : Permet de retourner le nombre de lignes et de colonnes du DataFrame

```
df.shape
✓ 0.0s
(5000, 23)
```

Figure 2.5 : Les dimensions du dataset

- `data.head()` : Permet d'afficher les premières lignes d'un DataFrame. Par défaut, renvoie les 5 premières lignes.

```
df.head()
✓ 0.0s
   id_client  genre  age  marie  num_tel  nb_jours_abonne  duree_appel_jour  nb_appel_jour  cout_appel_jour  duree_appel_soiree  ...  nb_appel_nuit  cout_
0    382-4657  femme   37     yes  98505453              128.0        265.1       110.0        45.07      197.4  ...
1    371-7191  homme   46      no  97321658              107.0        161.6       123.0        27.47      195.5  ...
2    358-1921  homme   50      no  98653270              137.0        243.4       114.0        41.38      121.2  ...
3    375-9999  homme   78     yes  96303256              84.0         299.4       71.0        50.90      61.9  ...
4    330-6626  femme   75     yes  96412387              75.0         166.7       113.0        28.34      148.3  ...
5 rows × 23 columns
```

Figure 2.6 : Les 5 premiers lignes du dataset

- **data.info()** : Permet d'afficher un résumé du DataFrame (les noms et types de colonne, le nombre de valeurs non-nulles et l'utilisation de la mémoire)

```
# information sur l'ensemble de donnees
df.info()
✓ 0.0s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   id_client         5000 non-null    object  
 1   genre             5000 non-null    object  
 2   age               5000 non-null    int64  
 3   marie            5000 non-null    object  
 4   num_tel           5000 non-null    int64  
 5   nb_jours_abonne  4990 non-null    float64
 6   duree_appel_jour 5000 non-null    float64
 7   nb_appel_jour    4993 non-null    float64
 8   cout_appel_jour  5000 non-null    float64
 9   duree_appel_soiree 5000 non-null    float64
 10  nb_appel_soiree  4994 non-null    float64
 11  cout_appel_soiree 5000 non-null    float64
 12  duree_appel_nuit 5000 non-null    float64
 13  nb_appel_nuit    4989 non-null    float64
 14  cout_appel_nuit  5000 non-null    float64
 15  duree_appel_inter 5000 non-null    float64
 16  nb_appel_inter   4988 non-null    float64
 17  cout_appel_inter 5000 non-null    float64
 18  active_msg_vocaux 4992 non-null    object  
 19  nb_msg_vocaux    5000 non-null    int64  
 ...
 21  churn             4996 non-null    object  
 22  offer_type        5000 non-null    object  
dtypes: float64(14), int64(3), object(6)
```

Figure 2.7 : Résumé du dataset

- `data.describe()` : Permet d'afficher des statistiques descriptives d'un DataFrame (le nombre de valeurs, la moyenne, l'écart type, les valeurs minimum et maximum, et les quartiles pour chaque colonne)

df.describe().T								
	count	mean	std	min	25%	50%	75%	max
age	5000.0	4.650680e+01	16.743513	19.0	3.200000e+01	46.00	60.00	80.00
num_tel	5000.0	9.844445e+07	371630.462063	90032100.0	9.850545e+07	98505453.00	98505453.00	99652014.00
nb_jours_abonne	4990.0	1.002707e+02	39.704130	1.0	7.300000e+01	100.00	127.00	243.00
duree_appel_jour	5000.0	1.802889e+02	53.894699	0.0	1.437000e+02	180.10	216.20	351.50
nb_appel_jour	4993.0	1.000190e+02	19.828023	0.0	8.700000e+01	100.00	113.00	165.00
cout_appel_jour	5000.0	3.064967e+01	9.162069	0.0	2.443000e+01	30.62	36.75	59.76
duree_appel_soiree	5000.0	2.006366e+02	50.551309	0.0	1.663750e+02	201.00	234.10	363.70
nb_appel_soiree	4994.0	1.002147e+02	19.820534	0.0	8.700000e+01	100.00	114.00	170.00
cout_appel_soiree	5000.0	1.705432e+01	4.296843	0.0	1.414000e+01	17.09	19.90	30.91
duree_appel_nuit	5000.0	2.003916e+02	50.527789	0.0	1.669000e+02	200.40	234.70	395.00
nb_appel_nuit	4989.0	9.991241e+01	19.959576	0.0	8.700000e+01	100.00	113.00	175.00
cout_appel_nuit	5000.0	9.017732e+00	2.273763	0.0	7.510000e+00	9.02	10.56	17.77
duree_appel_inter	5000.0	1.026178e+01	2.761396	0.0	8.500000e+00	10.30	12.00	20.00
nb_appel_inter	4988.0	4.432638e+00	2.456616	0.0	3.000000e+00	4.00	6.00	20.00
cout_appel_inter	5000.0	2.771196e+00	0.745514	0.0	2.300000e+00	2.78	3.24	5.40
nb_msg_vocaux	5000.0	7.755200e+00	13.546393	0.0	0.000000e+00	0.00	17.00	52.00
nb_reclamation	4988.0	1.570569e+00	1.306012	0.0	1.000000e+00	1.00	2.00	9.00

**Figure 2.8 :** Les statistiques sur les colonnes numériques

Nous allons afficher des statistiques pour les colonnes qui contiennent des données de type "Object".

df.describe(include=object).T				
	count	unique	top	freq
id_client	5000	5000	382-4657	1
genre	5000	2	femme	2593
marie	5000	2	no	2552
active_msg_vocaux	4992	2	no	3671
churn	4996	2	False	4289
offer_type	5000	83	PRE - 900 bonus	766

Figure 2.9 : Les statistiques des colonnes catégoriques

- `data.nunique()` : renvoie le nombre de valeurs uniques pour chaque colonne.

# nombre de valeurs uniques par colonne	
df.nunique()	
✓	0.0s
id_client	5000
genre	2
age	62
marie	2
num_tel	33
nb_jours_abonne	218
duree_appel_jour	1961
nb_appel_jour	123
cout_appel_jour	1961
duree_appel_soiree	1879
nb_appel_soiree	126
cout_appel_soiree	1659
duree_appel_nuit	1853
nb_appel_nuit	131
cout_appel_nuit	1028
duree_appel_inter	170
nb_appel_inter	21
cout_appel_inter	170
active_msg_vocaux	2
nb_msg_vocaux	48
nb_reclamation	10
churn	2
offer_type	83
age_interval	5
subscription_interval	5

Figure 2.10 : Le nombre des valeurs uniques par colonne

## 2.6 Analyse des données

### 2.6.1 Analyse univariée

L'analyse univariée est une méthode d'analyse statistique qui permet d'explorer une seule variable à la fois. Elle permet de résumer et de visualiser les caractéristiques de la variable étudiée. Cette analyse nous permet d'améliorer notre compréhension des données.

La répartition des clients en fonction de leur statut de Churn est représentée dans la figure

★ Variables catégorielles :

print(df.describe(include=object))						
✓ 0.0s						
	id_client	genre	marie	active_msg_vocaux	churn	offer_type
count	5000	5000	5000	4992	4996	1000
unique	5000	2	2	2	2	40
top	382-4657	femme	no	no	False	PRE - 900 bonus
freq	1	2593	2552	3671	4289	120

Figure 2.11 : Les variables catégorielles

- Churn : La répartition des clients en fonction de leur statut de churn est représentée dans la figure suivante

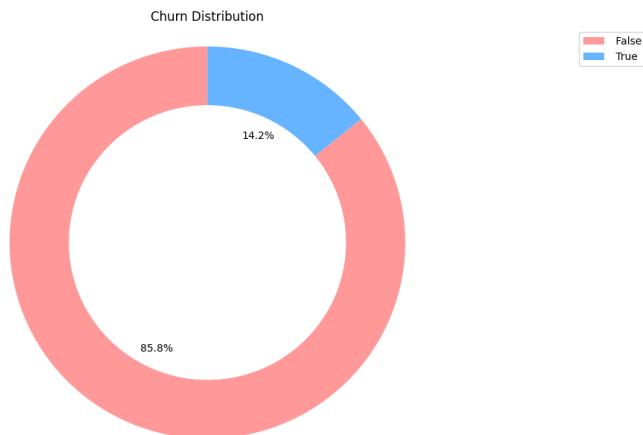
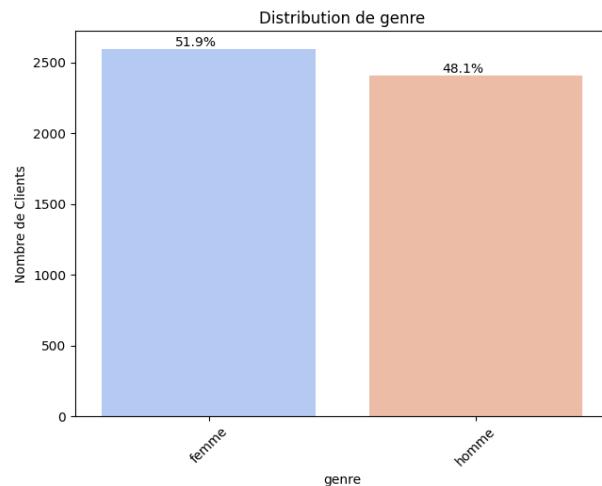


Figure 2.12 : Répartition des clients en fonction de leur statut de churn

Après l'analyse sur le taux de désabonnements de la clientèle, nous remarquons que la majorité des clients restent actifs, il représente une proportion de 85.8%, alors que 14.2% des clients ont résilié leurs services, ce qui peut être considéré comme relativement élevé. Cela signifie qu'il pourrait avoir un impact négatif important sur les modèles finaux.

- Genre :

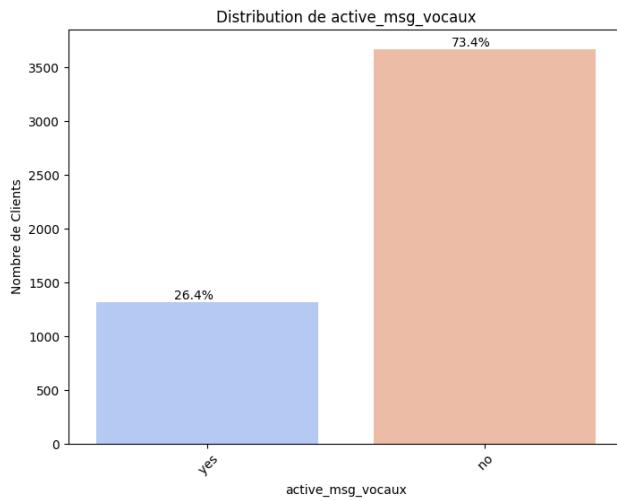


**Figure 2.13 :** Visualisation de chaque catégorie de genre

D'après le graphique, nous pouvons observer que notre jeu de données à une répartition entre les femmes et les hommes presque équilibrée, avec une proportion de 51,9% de femmes et de 48,1% d'hommes.

- Active Messages Vocaux :

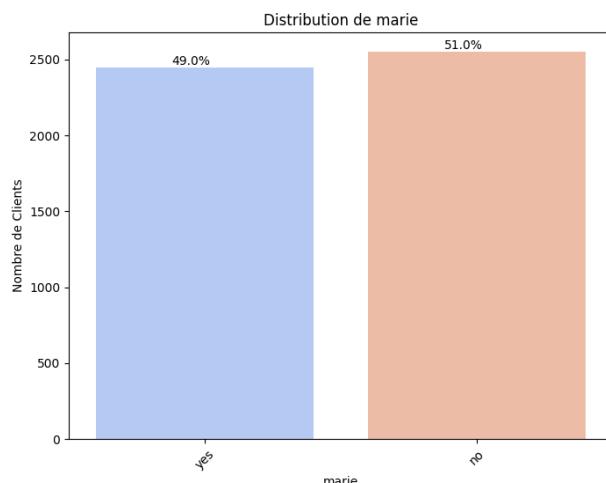
La figure 2.14 présente la distribution des clients en fonction de leur activation des messages vocaux.



**Figure 2.14 :** Distribution des clients en fonction de leur activation des messages vocaux

D'après le graphique, nous pouvons observer que la majorité des clients ne sont pas intéressés par l'activation de la fonctionnalité de messages vocaux, avec un pourcentage de 73.50% qui n'a pas activé cette fonctionnalité. Seulement 26.50% des clients ont activé les messages vocaux. Cette analyse indique que la fonctionnalité de messages vocaux ne semble pas être très utilisée par la majorité des clients.

- Marié :



**Figure 2.15 :** Répartition des clients selon leur situation civile

D'après le graphique, nous pouvons observer que les clients sont répartis, selon leur état civil, d'une manière presque équivalente avec une proportion de 49% sont mariés et de 51% sont non mariés.

- Offre Tarifaire :

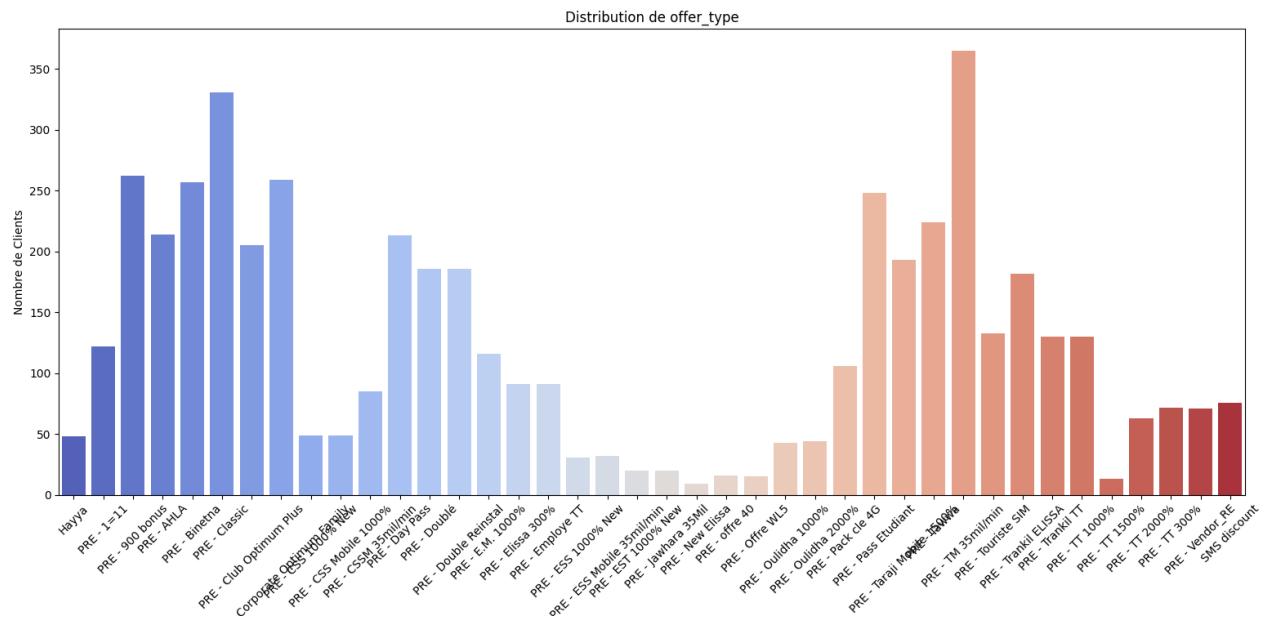


Figure 2.16 : Répartition des clients selon leur offre tarifaire suivie

Le graphique révèle une répartition inégale des offres tarifaires suivies par les clients de Tunisie Télécom. En tête se trouvent les offres PRE - Binetna et PRE - TAWWA, attirant un nombre significatif d'abonnés. Les autres offres sont moins fréquemment souscrites. Cette observation suggère que ces deux offres sont particulièrement attractives pour la clientèle de Tunisie Télécom, peut-être en raison de leurs avantages ou de leur pertinence pour les besoins des utilisateurs.

☆ Variables numériques :

- Age :

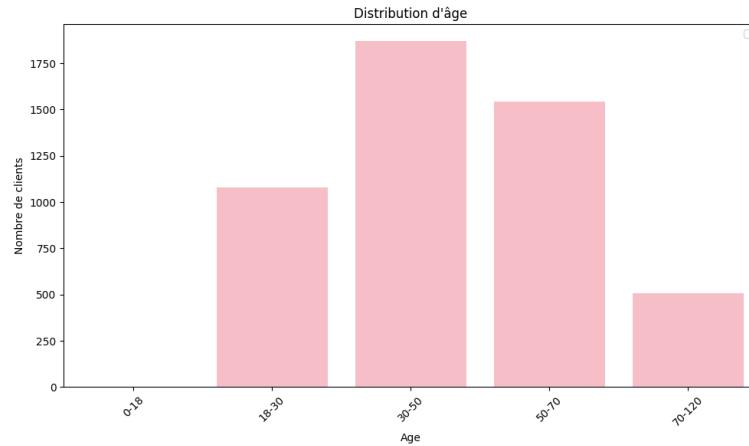


Figure 2.17 : Répartition des clients selon leur âge

Le graphique montre que la majorité des clients de Tunisie Télécom ont entre 30 et 50 ans, suivis par ceux âgés de 50 à 70 ans. Les tranches d'âge de 18 à 30 ans et de 70 à 120 ans sont moins représentées. Cette diversité suggère que Tunisie Télécom cible un large éventail d'âges, ce qui pourrait être bénéfique pour l'entreprise.

- Nombre de jours d'abonnement :

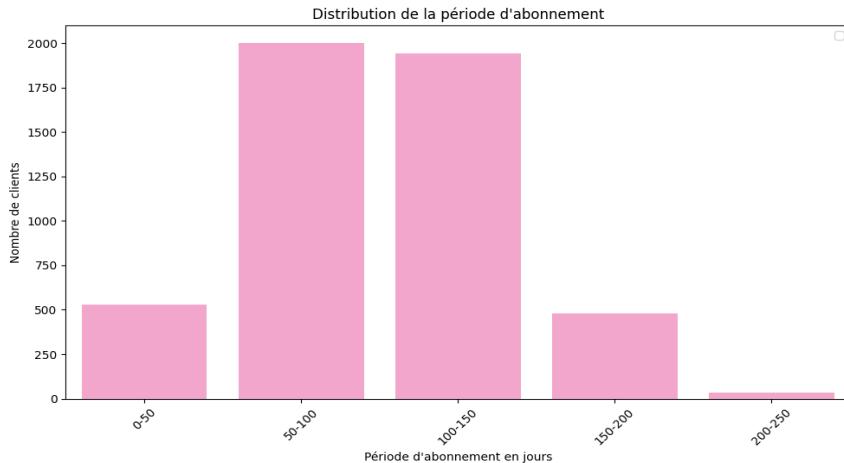
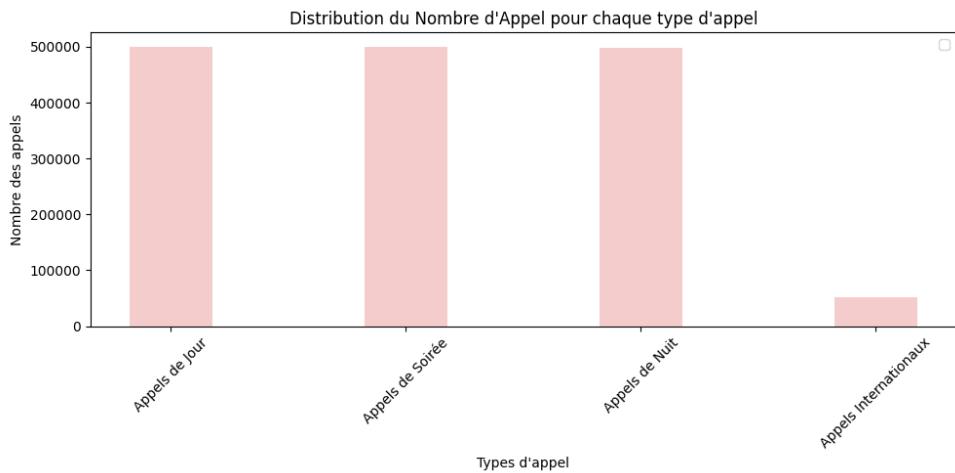


Figure 2.18 : Répartition des clients selon leur période d'abonnement

D'après le graphique, nous pouvons observer que la majorité des clients sont abonnés entre 50 et 100 jours, indiquant l'importance de cette période pour la fidélisation. Les abonnements de 100 à 150 jours sont également significatifs, montrant une bonne rétention. Cependant, un petit groupe de clients préfère des abonnements plus courts, peut-être des essais ou des changements fréquents d'opérateur. Enfin, quelques clients sont fidèles depuis plus de 200 jours, soulignant la satisfaction à long terme avec les services de Tunisie Télécom.

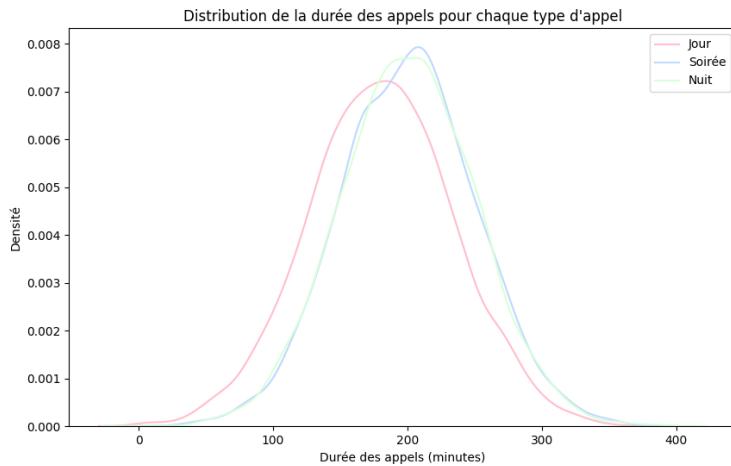
- Nombre des Appels pour chaque type d'appel :



**Figure 2.19 :** Répartition des clients selon leur nombre d'appel pour chaque type d'appel

Le graphique met en lumière une distribution presque égale des appels de jour, de nuit et de soirée, tandis que les appels internationaux sont nettement moins nombreux. Cela suggère que la majorité des clients préfèrent les appels nationaux, avec une préférence marquée pour ces trois plages horaires spécifiques.

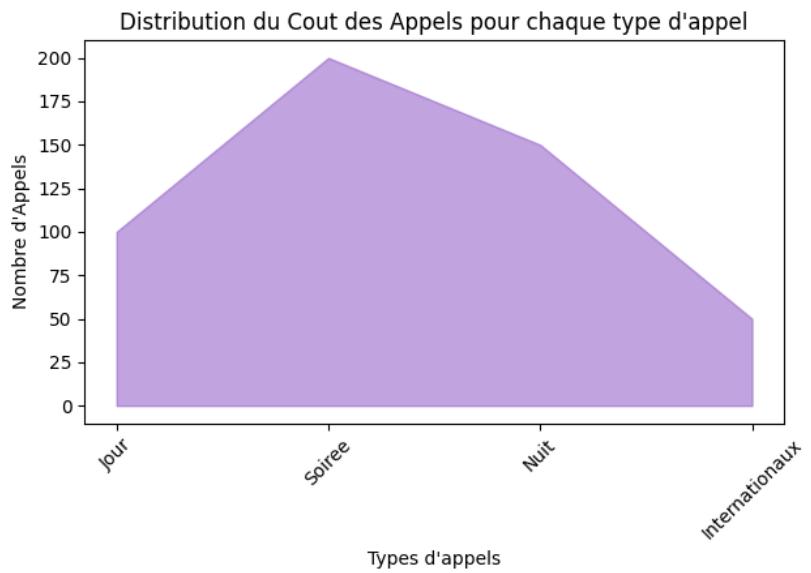
- Durée des Appels pour chaque type d'appel :



**Figure 2.20 :** Répartition des clients selon leurs durées des appels pour chaque type d'appel

Le graphique révèle une répartition presque équitable des durées d'appel pour les plages horaires de jour, de nuit et de soirée, mais il montre également que la durée des appels en soirée est légèrement plus élevée que pour les autres périodes. Cette légère différence suggère que les clients accordent une importance légèrement plus grande aux communications en soirée par rapport aux autres moments de la journée. Cela met en évidence la diversité des besoins et des préférences des clients, tout en soulignant leur engagement continu à rester connectés avec leurs proches et leurs affaires, même tard dans la journée.

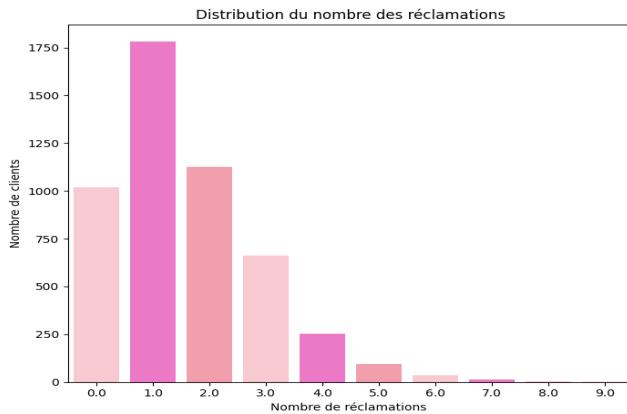
- Cout des Appels pour chaque type d'appel :



**Figure 2.21 :** Répartition des clients selon le coût des appels pour chaque type d'appel

Les coûts des appels en soirée sont considérablement plus élevés que ceux des appels effectués pendant les périodes de jour et de nuit. Cette disparité significative dans les coûts suggère que les utilisateurs peuvent opter pour des conversations plus longues et potentiellement plus coûteuses pendant les heures du soir. Ces coûts plus élevés pourraient être dus à la durée moyenne d'appel plus longue observée pendant cette période, comme le suggère le graphique précédent.

- Nombre de réclamations :



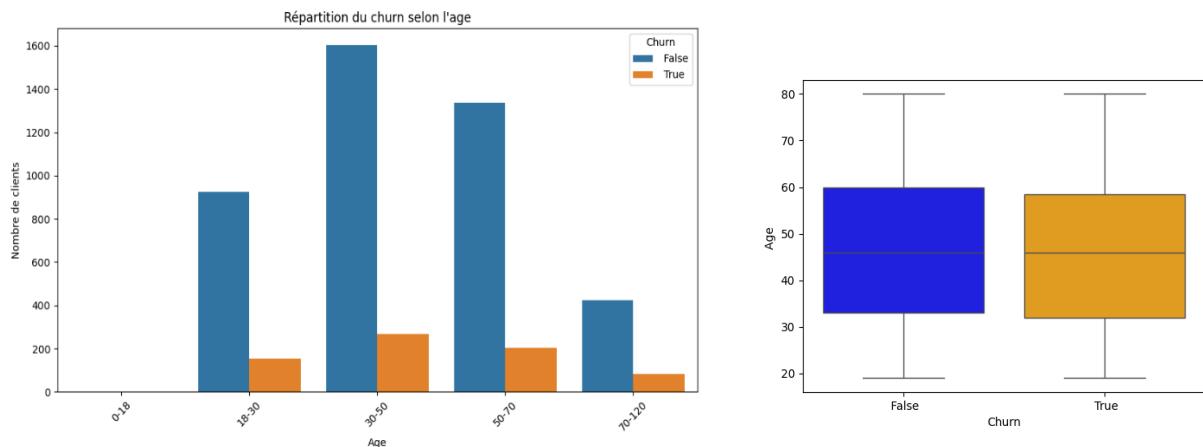
**Figure 2.22 :** Répartition des clients selon le nombre des réclamations

En analysant la distribution des clients par nombre d'appels internationaux, nous pouvons conclure que la plupart des clients ont effectué entre 2 et 5 appels internationaux, ce qui suggère que cette fonctionnalité n'est pas très utilisée.

## 2.6.2 Analyse bivariée

L'analyse bivariée est une méthode statistique qui permet d'examiner la relation entre deux variables en les comparant ou en les associant, ce qui peut aider à comprendre comment elles sont liées et comment elles interagissent.

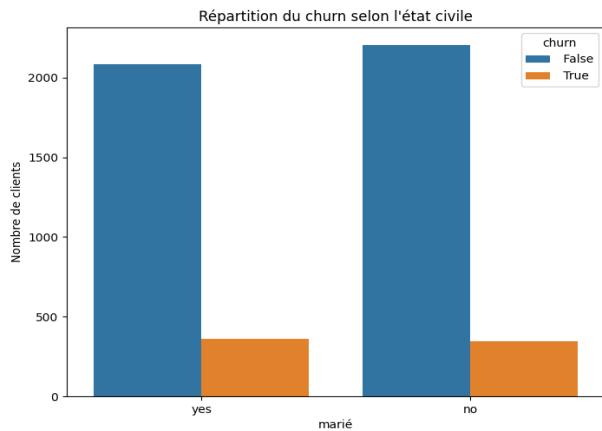
- La figure illustre la distribution du churn selon l'âge.



**Figure 2.23 :** Répartition du churn selon l'âge

Cette analyse révèle une corrélation entre l'âge des clients et leur taux de churn. Les clients âgés de 18 à 30 ans présentent le taux de churn le plus élevé, d'environ 14%. Ce taux diminue progressivement avec l'âge des clients. Les jeunes résilient leur abonnement plus fréquemment que les clients plus âgés. Une analyse approfondie est recommandée pour comprendre les raisons de ces différences d'attrition entre les groupes d'âge.

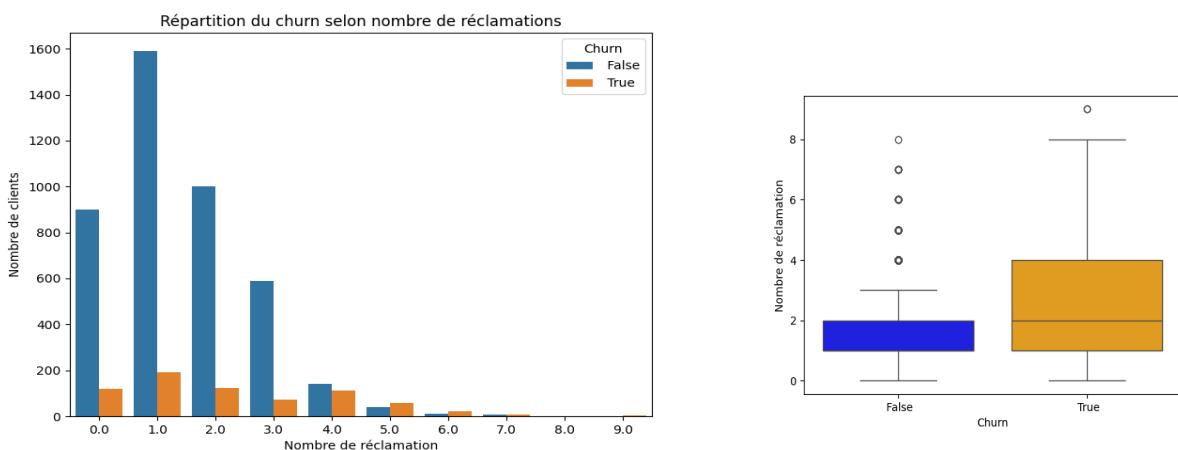
- La figure illustre la distribution du churn selon l'état civile.



**Figure 2.24 :** Répartition du churn selon l'état civile

L'analyse de Tunisie Telecom révèle une corrélation entre l'état civil des clients et leur taux de churn. Les non-mariés ont un taux de churn plus élevé (environ 11%) par rapport aux mariés (environ 7%). Cette différence suggère que l'état civil influence la résiliation d'abonnement. Une exploration approfondie des raisons de cette disparité est nécessaire pour élaborer des stratégies de rétention ciblées.

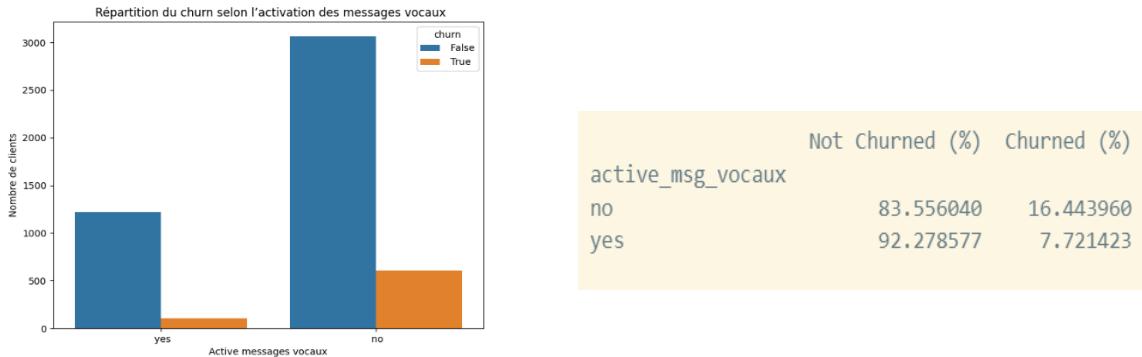
- La figure illustre la distribution du churn selon le nombre de réclamations.



**Figure 2.25 :** Répartition du churn selon le nombre de réclamations

La figure indique clairement que le taux de désabonnement augmente fortement à partir de 5 réclamations.

- La figure illustre la distribution du churn selon l'activation des messages vocaux.



**Figure 2.26 :** Répartition du churn selon l'activation des messages vocaux

D'après la figure ci-dessus, on peut conclure que les clients qui n'utilisent pas la messagerie vocale ont un risque élevé de résiliation de leur abonnement.

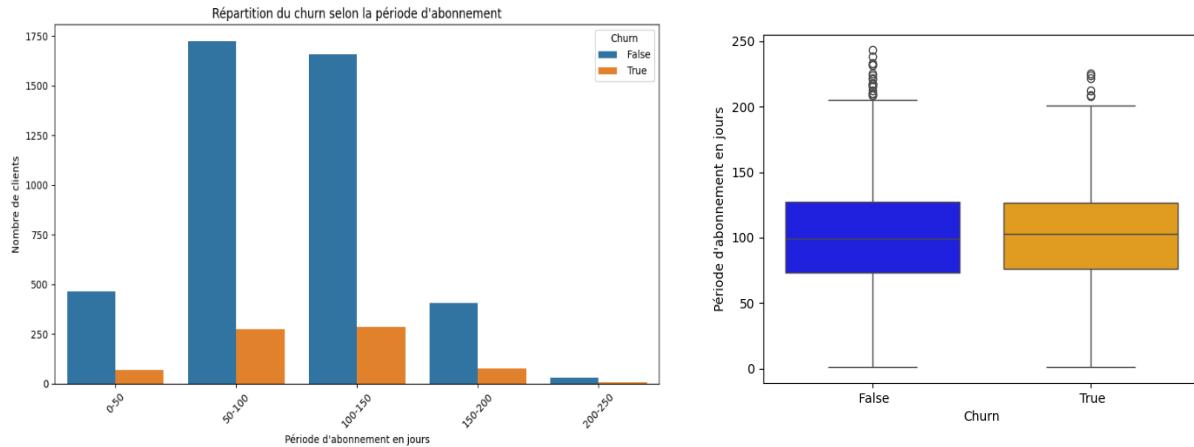
- La figure illustre la distribution du churn selon le nombre des messages vocaux.



**Figure 2.27 :** Répartition du churn selon le nombre des messages vocaux

En analysant les données des messages vocaux, nous pouvons conclure que les clients qui reçoivent moins de 20 messages vocaux présentent un risque plus élevé de résilier leur abonnement.

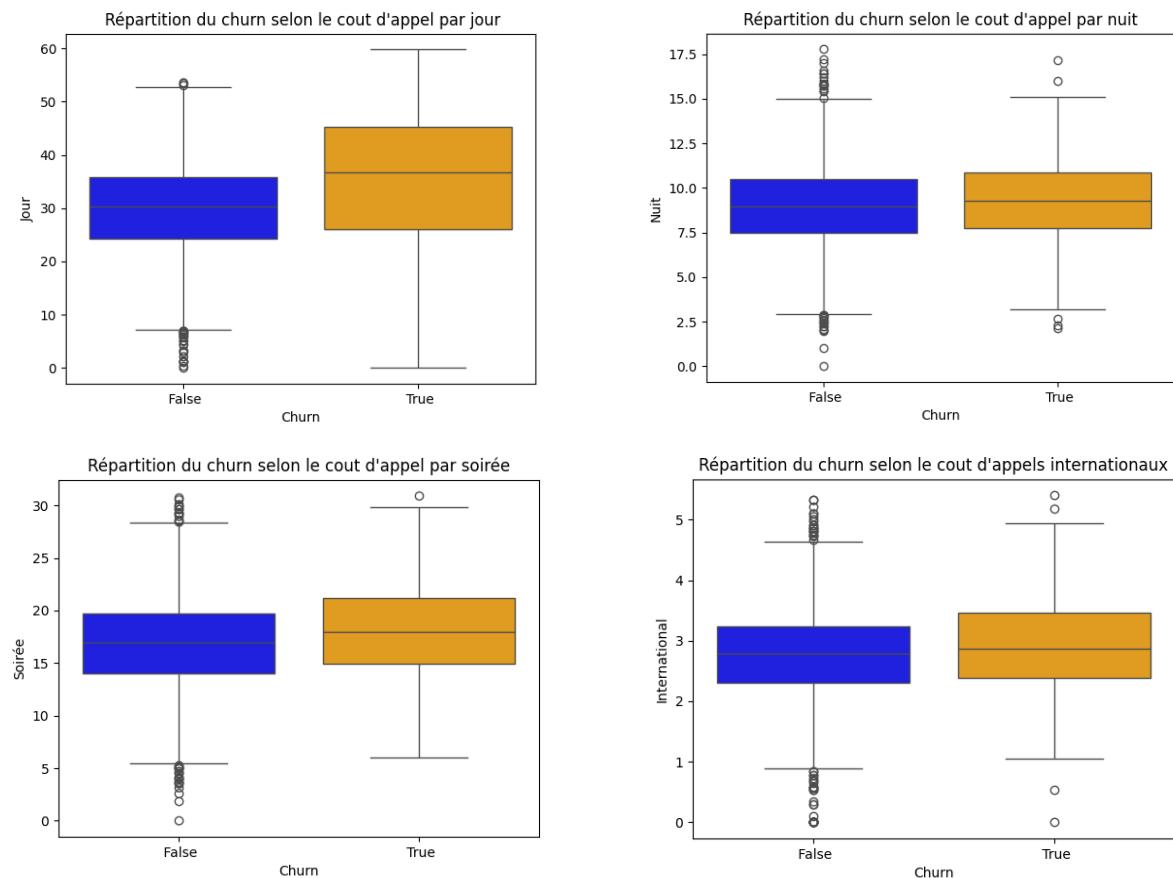
- La figure illustre la distribution du churn selon la période d'abonnement en jours.



**Figure 2.28 :** Répartition du churn selon la période d'abonnement en jours

Cette analyse montre que la majorité des clients ayant résilié leur abonnement l'ont fait au cours des 3 premiers mois. La période de 50 à 150 jours après l'abonnement se révèle être une période critique pour la rétention des clients, car le taux de churn y est particulièrement élevé.

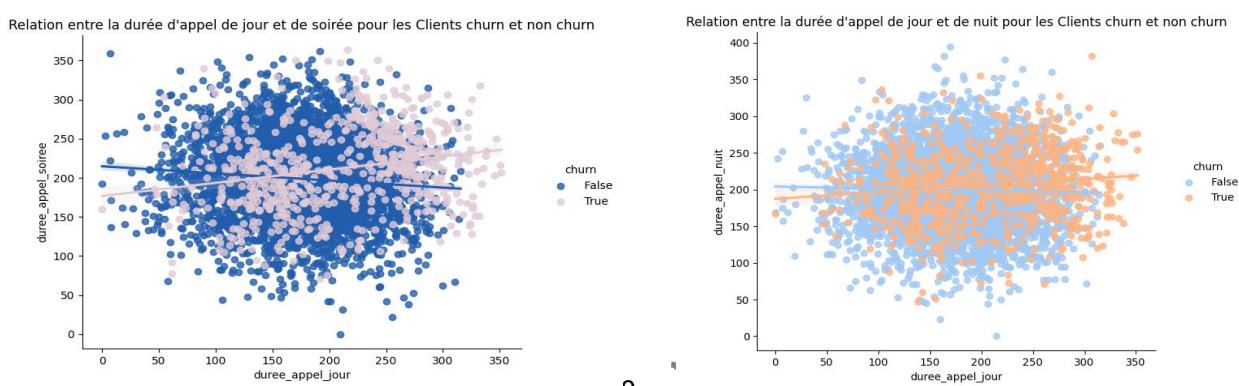
- La figure illustre la distribution du churn selon le coût d'appel des différents types d'appels

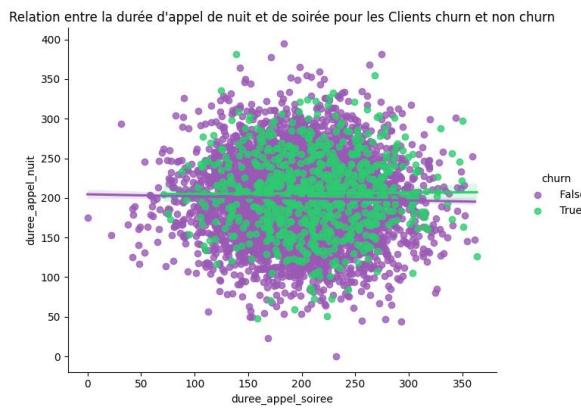


**Figure 2.29 :** Boîtes à moustaches du coût d'appel des différents types d'appels

Les coûts totaux qui correspondent aux clients chameurs sont plus élevées par rapport aux autres qui restent. Cela pourrait montrer que les clients qui quittent l'entreprise ne sont pas satisfaits du montant qu'ils paient pour leur plan.

- La figure illustre la distribution du churn selon la durée d'appel des différents types d'appels





**Figure 2.30 :** Comparaison de la durée d'appel jour/nuit/soirée entre clients churn et non churn

Cette figure montre un nuage de points presque centré, cela suggère qu'il n'y a pas de relation claire ou significative entre les deux variables « durée\_appel\_jour(minutes) », «durée\_appel\_nuit(minutes)» et «durée\_appel\_soirée(minutes) »

## 2.7 Conclusion

Au cours de cette phase, nous avons examiné attentivement et en détaillons notre ensemble de données ainsi nous avons identifié différentes anomalies qui pourraient affecter notre analyse. Cette étape est cruciale car elle nous permet de mieux comprendre notre ensemble de données et de déterminer les mesures à prendre pour nettoyer et préparer les données pour la prochaine phase de la méthodologie CRISP-DM, à savoir la préparation des données qui l'on trouve dans le chapitre suivant.

# Chapitre 3 : Prétraitement des données

## Plan

1.	Introduction . . . . .	42
2.	Nettoyage des données . . . . .	42
3.	Encodage des données . . . . .	49
4.	Normalisation et standardisation. . . . .	52
5.	Sélection des fonctionnalités. . . . .	53
6.	Conclusion . . . . .	55

### **3.1 Introduction**

Ce chapitre met vraiment l'accent sur l'importance de bien préparer les données dans un projet de science des données. On parle ici de tout le processus de nettoyage, de transformation et d'organisation des données afin de les rendre prêtes à être utilisées dans des algorithmes d'apprentissage automatique. L'idée, c'est d'avoir des données qui soient vraiment exploitables pour que les modèles d'intelligence artificielle puissent fonctionner de manière efficace et pertinente.

### **3.2 Nettoyage des données**

L'étape de nettoyage des données, également appelée "data cleaning" en anglais, est une phase essentielle de la préparation des données. Cette étape vise à identifier et à corriger les erreurs, les incohérences et les données manquantes présentes dans le jeu de données. Le nettoyage des données garantit la qualité et la fiabilité des informations utilisées dans l'analyse et la modélisation. En fait, dans cette partie, nous allons réaliser les traitements suivants :

- Suppression des colonnes inutiles
- Traitement des valeurs manquantes
- Suppression des valeurs dupliquées

#### **3.2.1 Suppression des colonnes inutiles**

```

for x in df.columns :
    print(x, len(df[x].unique()))
✓ 0.0s

id_client 5000
genre 2
age 62
marie 2
num_tel 33
nb_jours_abonne 219
duree_appel_jour 1961
nb_appel_jour 124
cout_appel_jour 1961
duree_appel_soiree 1879
nb_appel_soiree 127
cout_appel_soiree 1659
duree_appel_nuit 1853
nb_appel_nuit 132
cout_appel_nuit 1028
duree_appel_inter 170
nb_appel_inter 22
cout_appel_inter 170
active_msg_vocaux 2
nb_msg_vocaux 48
nb_reclamation 11
churn 3
offer_type 40
age_interval 4
subscription_interval 6

```

**Figure 3.1 :** Les valeurs uniques de chaque colonne

La colonne 'id\_client' comporte 5000 valeurs uniques, et la colonne 'num\_tel' en comporte 33. Étant donné que ces deux colonnes sont des chaînes de caractères et qu'elles ne sont pas pertinentes pour notre analyse, il est conseillé de les supprimer.

```

df = df.drop(df.columns[[0, 4]], axis=1)
df.head()

✓ 0.0s
Python

   genre  age  marie  nb_jours_abonne  duree_appel_jour  nb_appel_jour  cout_appel_jour  duree_appel_soiree  nb_appel_soiree  cout_appel_soiree  ...  nb_appel_nuit  cout_ap
0  femme  37    yes           128.0        265.1       110.0      45.07        197.4        99.0      16.78     ...        91.0
1  homme  46     no            107.0        161.6       123.0      27.47        195.5        103.0      16.62     ...       103.0
2  homme  50     no            137.0        243.4       114.0      41.38        121.2        110.0      10.30     ...       104.0
3  homme  78    yes            84.0         299.4       71.0      50.90        61.9         88.0      5.26     ...        89.0
4  femme  75    yes            75.0         166.7      113.0      28.34        148.3        122.0      12.61     ...       121.0

5 rows × 21 columns

```

**Figure 3.2 :** Code et résultat de suppression de colonne inutiles "id\_client" et "num\_tel"

### 3.2.2 Les valeurs manquantes

Une donnée manquante fait référence à une information ou observation qui est absente ou égale à zéro dans un ensemble de données. Ces situations peuvent poser des difficultés lors de l'analyse statistique. La présence de données manquantes dans un modèle d'apprentissage automatique peut s'avérer inefficace et potentiellement risquée pour diverses raisons :

- Réduit la précision du modèle ML.
- Modification de la distribution des données.
- Introduit un biais dans l'estimation du modèle ML.

Avant de commencer à utiliser un algorithme d'apprentissage automatique (ML), il est crucial de repérer les données manquantes dans le jeu de données. Le graphique suivant affiche le nombre de valeurs manquantes pour chaque colonne en utilisant la fonction "isnull()".

#Vérifier les valeurs nuls df.isnull().sum()	
✓	0.0s
genre	0
age	0
marie	0
nb_jours_abonne	10
duree_appel_jour	0
nb_appel_jour	7
cout_appel_jour	0
duree_appel_soiree	0
nb_appel_soiree	6
cout_appel_soiree	0
duree_appel_nuit	0
nb_appel_nuit	11
cout_appel_nuit	0
duree_appel_inter	0
nb_appel_inter	12
cout_appel_inter	0
active_msg_vocaux	8
nb_msg_vocaux	0
nb_reclamation	12
churn	4
offer_type	0
dtype: int64	

Figure 3.3 : Nombre des valeurs manquantes par colonne

### 3.2.2.1 Traitement des valeurs manquantes

D'après la figure, nous constatons qu'il manque des données dans 8 colonnes : "nb\_jours\_abonne", "nb\_appel\_jour", "nb\_appel\_soirée", "nb\_appel\_nuit", "nb\_appel\_inter", 'active\_msg\_vocaux', 'nb\_reclamation' et 'churn'.

Ainsi, nous allons rectifier ces valeurs pour chaque colonne :

- **Traitement de la colonne « nb\_jours\_abonne » :**

Pour remplir les valeurs manquantes dans cette colonne, nous allons les remplacer par la valeur la plus élevée existante, déterminée à l'aide de la fonction "max ()".

```
df['nb_jours_abonne'].value_counts()
✓ 0.0s
nb_jours_abonne
90.0    65
87.0    59
105.0   57
93.0    57
112.0   56
...
215.0    1
238.0    1
216.0    1
208.0    1
233.0    1
Name: count, Length: 218, dtype: int64

df['nb_jours_abonne'] = df['nb_jours_abonne'].fillna(df["nb_jours_abonne"].max())
✓ 0.0s
```

**Figure 3.4 :** Traitement des valeurs manquantes de la colonne "nb\_jours\_abonne"

- **Traitement de colonne « nb\_reclamation »**

Nous remarquons que la grande majorité des clients n'ont envoyé qu'une seule réclamation.

Par conséquent, nous allons utiliser la fonction « fillna () » pour remplir les valeurs manquantes de la colonne "nb\_reclamation" avec la valeur la plus élevée observée dans cette colonne, qui est égale à 1.

```

df['nb_reclamation'].value_counts()
✓ 0.0s

nb_reclamation
1.0    1781
2.0    1126
0.0    1020
3.0     663
4.0     252
5.0      95
6.0      34
7.0      13
9.0      2
8.0      2
Name: count, dtype: int64

df['nb_reclamation'] = df['nb_reclamation'].fillna(1.0)
✓ 0.0s

```

**Figure 3.5 :** Traitement des valeurs manquante de la colonne « nb\_reclamation »

- **Traitement des colonnes du nombre des appels jours, soirée et nuit**

Pour remplir les valeurs manquantes dans ces colonnes, nous allons remplacer les valeurs manquantes par la valeur la plus élevée présente dans chaque colonne, obtenue à l'aide de la fonction "max ()".

```

df['nb_appel_jour'] = df['nb_appel_jour'].fillna(df["nb_appel_jour"].max())
df['nb_appel_soiree'] = df['nb_appel_soiree'].fillna(df["nb_appel_soiree"].max())
df['nb_appel_nuit'] = df['nb_appel_nuit'].fillna(df["nb_appel_nuit"].max())
✓ 0.0s

```

**Figure 3.6 :** Traitement des valeurs manquante des colonnes du nombre d'appels

- **Traitement de colonne « nb\_appel\_inter »**

Nous allons remplir les valeurs manquantes de la colonne "nb\_appel\_inter" avec la moyenne des valeurs à l'aide de la fonction « mean()».

```
mean_value = df['nb_appel_inter'].mean()  
mean_value = round(mean_value)  
mean_value  
✓ 0.0s
```

4

```
df["nb_appel_inter"] = df['nb_appel_inter'].fillna(mean_value)  
✓ 0.0s
```

Figure 3.7 : Traitement des valeurs manquante de colonne « nb\_appel\_inter »

- **Traitement de la colonne « active\_msg\_vocaux » :**

Nous allons remplir les valeurs manquantes dans la colonne "active\_msg\_vocaux" en utilisant la méthode "ffill" (forward fill). Cette fonction consiste à remplacer les valeurs manquantes par la dernière valeur observée avant la valeur manquante.

```
df['active_msg_vocaux'].fillna(method ="ffill", inplace = True)
```

✓ 0.0s

Figure 3.8 : Traitement des valeurs manquante de colonne « active\_msg\_vocaux »

- **Traitement de la colonne « churn » :**

La méthode "bfill" est utilisée pour traiter les valeurs manquantes dans la colonne "churn" qui consiste à remplir ces valeurs en utilisant la première valeur disponible après la valeur manquante.

```
df['churn'].value_counts()  
✓ 0.0s
```

```
churn  
False    4289  
True     707  
Name: count, dtype: int64
```

```
df['churn'].fillna(method ="bfill", inplace = True)  
✓ 0.0s
```

Figure 3.9 : Traitement des valeurs manquante de colonne « churn »

### 3.2.2.2 Vérification des valeurs manquantes

Une fois que nous avons traité les valeurs manquantes dans les colonnes en utilisant les méthodes appropriées, nous allons maintenant vérifier si des valeurs manquantes subsistent dans l'ensemble de données.

df.isnull().sum()	
✓	0.0s
genre	0
age	0
marie	0
nb_jours_abonne	0
duree_appel_jour	0
nb_appel_jour	0
cout_appel_jour	0
duree_appel_soiree	0
nb_appel_soiree	0
cout_appel_soiree	0
duree_appel_nuit	0
nb_appel_nuit	0
cout_appel_nuit	0
duree_appel_inter	0
nb_appel_inter	0
cout_appel_inter	0
active_msg_vocaux	0
nb_msg_vocaux	0
nb_reclamation	0
churn	0
offer_type	0
dtype:	int64

Figure 3.10 : Vérification des valeurs manquantes

En examinant la figure ci-dessous, nous pouvons constater qu'il n'y a plus de valeurs manquantes dans la base de données après le traitement effectué. Cela indique que les données sont prêtes à être utilisées pour la modélisation.

### 3.2.3 Les valeurs dupliquées

La présence de valeurs dupliquées indique que des observations ont les mêmes valeurs pour toutes les variables enregistrées. Cela peut entraîner des biais dans les résultats d'analyse et compromettre la qualité des conclusions qui en découlent. La solution de suppression des valeurs dupliquées inclut :

- Utilisation de la fonction « drop\_duplicates() » : cette fonction est utile car elle permet de ne conserver qu'une seule copie de chaque valeur tout en supprimant les doublons.

La figure 3.10 présente le code et le résultat de la détection de lignes dupliquées.

```
#Vérifier s'il y a des doublons
print(df.duplicated().sum())
✓ 0.0s
0
```

Figure 3.11 : Nombre des valeurs dupliquées

La figure ci-dessus montre que la fonction « duplicated().sum() » renvoie une valeur égale à 0, ce qui signifie qu'il n'y a pas de doublons ou de données en double dans la base de données. En conséquence, chaque ligne est unique et il n'y a pas de répétition d'observations.

## 3.3 Encodage des données

LabelEncoder est une technique de prétraitement des données utilisé Machine Learning. Elle permet de convertir des données catégorielles en données numérique afin de faciliter leur traitement. Elle attribue un identifiant numérique unique à chaque catégorie. [32]

La première étape consiste à identifier les colonnes de type « chaîne de caractères » à encoder, puis l'algorithme LabelEncoder encode chaque colonne sélectionnée en attribuant à chaque catégorie un entier unique.

df.dtypes	
✓	0.0s
genre	object
age	int64
marie	object
nb_jours_abonne	float64
duree_appel_jour	float64
nb_appel_jour	float64
cout_appel_jour	float64
duree_appel_soiree	float64
nb_appel_soiree	float64
cout_appel_soiree	float64
duree_appel_nuit	float64
nb_appel_nuit	float64
cout_appel_nuit	float64
duree_appel_inter	float64
nb_appel_inter	float64
cout_appel_inter	float64
active_msg_vocaux	object
nb_msg_vocaux	int64
nb_reclamation	float64
churn	object
offer_type	object
dtype:	object

Figure 3.12 : Vérification sur les types des données

D'après la figure suivante, Nous constatons que 5 colonnes sont de type « chaîne de caractères » et nécessitent d'être encodées à l'aide de la méthode Label Encoder. Ces 5 colonnes sont : genre, marié, active\_msg\_vocaux , churn et offer\_type. Alors, on va remplacer :

- Les valeurs "femme", "yes" et "True" par 1
- Les valeurs "homme", "no" et "False" par 0

```
label_encoder = LabelEncoder()
df['genre'] = label_encoder.fit_transform(df['genre'])
df['marie'] = label_encoder.fit_transform(df['marie'])
df['active_msg_vocaux'] = label_encoder.fit_transform(df['active_msg_vocaux'])
df['churn'] = label_encoder.fit_transform(df['churn'])
```

✓ 0.0s

Figure 3.13 : Application du Label Encoder sur les colonnes « genre », « marie », « active\_msg\_vocaux » et « churn »

Nous avons utilisé une méthode appelée « encodage cible » pour transformer la colonne 'offer\_type' en une représentation numérique unique pour chaque type d'offre.

Cette approche consiste à calculer le taux moyen de résiliation associé à chaque type d'offre. Ensuite, nous avons remplacé les libellés des offres par ces chiffres, qui reflètent le taux moyen de résiliation pour chaque type d'offre. Cette technique nous permet d'intégrer la relation entre les différents types d'offres et le taux de résiliation dans notre modèle d'apprentissage automatique, ce qui améliore notre capacité à prédire les comportements des clients.

```
# Calculer la moyenne de 'encoded_churn' pour chaque type d'offre
target_mean_df = df.groupby('offer_type')['churn'].mean().reset_index()
offer_type_mapping = dict(zip(target_mean_df['offer_type'], target_mean_df['churn']))

df['offer_type'] = df['offer_type'].map(offer_type_mapping)
```

✓ 0.0s

Figure 3.14 : Encodage de la colonne 'offre\_type' avec la moyenne cible de 'churn'

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	offer_type
1	id_client	genre	age	marie	num_tel	nb_jours_abo_duree_appel	nb_appel	soi_cout_appel	soi_cout_appel	duree_appel	nb_appel	soi_cout_appel	duree_appel	nb_appel	soi_cout_appel	n_duree_appel	nb_appel	inti_cout_appel	il_active_msg_v_nb	msg_voc_nb	reclamati	churn		
2	382-4657	femme	37 yes		98505453	128	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10	3	2.7	yes		25	1	False	Hayya
3	371-7191	homme	46 no		97321658	107	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7	yes		26	1	False	PRE -E1-11
4	358-1921	homme	50 no		98653270	137	243.4	114	41.38	121.2	110	10.3	182.6	104	7.32	12.2	5	3.29	no		0	0	False	PRE -900 bonus
5	375-9999	homme	79 yes		96303255	75	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.8	7	1.78	no		0	2	False	PRE -AHLA
6	330-6662	femme	75 yes		98412387	75	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	no		0	3	False	PRE -Binetra
7	391-8027	femme	23 no		98142387	118	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.7	no		0	0	False	PRE -Classic
8	355-9993	femme	67 yes		98501269	121	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	2.03	yes		24	3	False	PRE -Club Optimum Plus
9	329-9001	homme	52 yes		99606321	147	157	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92	no		0	0	False	PRE -Corporate Optimum Fam
10	335-4719	femme	68 no		99421753	117	184.5	97	31.37	351.6	80	29.89	215.8	90	9.71	8.7	4	2.35	no		0	1	False	PRE -CSS 1000% New
11	330-8173	femme	43 yes		99203170	141	258.6	84	43.96	222	111	18.87	326.4	97	14.69	11.2	5	3.02	yes		37	0	False	PRE -CSS Mobile 1000%
12	329-6603	homme	47 no		95608231	65	120.1	137	21.95	228.5	83	19.42	208.8	111	9.4	12.7	6	3.43	no		0	4	True	PRE -CSSM 35min/min
13	344-9403	femme	25 yes		98741230	74	187.7	127	31.91	163.4	148	13.89	198	94	8.82	9.1	5	2.46	no		0	0	False	PRE -Day Pass
14	363-1107	femme	58 yes		986532140	168	128.8	96	21.9	104.9	71	8.92	141.1	128	6.35	11.2	2	3.02	no		0	1	False	PRE -Doubié
15	394-8006	homme	32 no		96321506	95	156.6	88	26.62	247.6	75	21.05	192.3	115	8.65	12.3	5	3.32	no		0	3	False	PRE -Double Reinstal
16	366-9238	femme	39 no		99201369	62	120.7	70	20.52	307.2	76	26.11	203	98	9.14	13.1	6	3.54	no		0	4	False	PRE -E.M. 1000%
17	351-7269	femme	58 yes		98512647	161	332.9	67	56.59	317.8	97	27.01	160.6	128	7.23	5.4	9	1.46	no		0	4	True	PRE -Elissa 200%
18	350-8884	femme	52 yes		98601476	85	198.4	139	33.39	280.9	90	23.88	89.3	75	4.02	13.8	4	3.73	yes		27	1	False	PRE -Employee TT
19	398-2923	femme	72 no		98602500	98	190.7	114	32.42	218.2	111	18.55	129.6	121	5.83	8.1	3	2.39	no		0	2	False	PRE -ESS 1000% New
20	356-2992	homme	79 no		98502360	76	188.7	66	32.25	212.8	65	18.09	165.7	108	7.46	10	5	2.7	yes		33	1	False	PRE -ESS Mobile 35min/min
21	073-0782	homme	67 no		90022100	73	224.4	90	38.15	159.5	88	13.56	192.8	74	8.86	13	2	3.51	no		0	1	False	PRE -EST 1000% New
22	396-5800	femme	70 yes		98418670	147	155.1	117	26.37	238.7	93	20.37	208.8	133	9.4	10.6	4	2.86	no		0	0	False	PRE -Jawhara 35Min
23	393-7984	femme	26 yes		98502211	77	62.4	89	10.61	186.9	121	14.44	209.5	64	9.43	5.7	6	1.54	no		0	5	True	PRE -New Elissa
24	358-1958	femme	30 yes		98652014	130	183	112	31.11	72.9	99	6.2	181.8	78	8.18	9.5	19	2.57	no		0	0	False	PRE -Offre 40
25	350-2665	femme	22 no		98505453	111	110.4	103	16.77	137.3	102	11.67	189.6	105	8.53	7.7	6	2.08	no		0	2	False	PRE -Offre W45
26	343-4966	femme	34 yes		98505459	102	81.1	86	13.79	245.2	72	20.84	207	115	10.67	10.3	2	2.78	no		0	0	False	PRE -Qualita 1000%
27	337-3968	femme	37 yes		98302514	174	124.3	76	21.18	277.1	112	23.55	250.7	115	11.28	15.5	5	4.19	no		0	3	False	PRE -Qualita 200%
28	357-3817	femme	37 yes		98805453	57	213	115	36.21	191.1	112	16.24	182.7	115	8.22	9.5	3	2.57	yes		39	0	False	PRE -Prix cle 40
29	418-6412	homme	42 yes		98505453	54	134.3	73	22.83	155.5	100	13.22	102.1	68	4.59	14.7	4	3.97	no		0	3	False	PRE -Prix Etudiant
30	383-2630	homme	64 no		98505453	20	190	109	32.3	256.2	84	21.95	181.5	102	8.17	6.3	6	1.7	no		0	0	False	PRE -Tarija Mobile 1500%
31	410-7769	femme	47 yes		98505453	49	119.3	117	20.28	215.1	109	18.28	178.7	90	8.04	11.1	1	3	no		0	1	False	PRE -Tawa
32	416-8428	homme	23 yes		98505453	142	84.8	95	14.42	136.7	63	11.62	250.5	148	11.27	14.2	6	3.83	no		0	2	False	PRE -TM 35Min/min
33	370-3359	femme	48 yes		98505453	75	226.1	105	38.44	201.5	107	17.13	246.2	98	11.08	10.3	5	2.78	no		0	1	False	PRE -Touriste SIM
34	383-1121	homme	28 yes		98505453	172	212	121	36.04	31.2	115	2.65	293.3	78	13.2	12.6	10	3.4	no		0	3	False	PRE -Trankit ELUSA
35	360-1596	femme	28 no		98505453	12	249.6	118	42.43	252.4	119	21.45	280.2	90	12.61	11.8	3	3.39	no		0	1	True	PRE -Trankit TT
36	395-2854	homme	33 yes		98505453	57	176.8	94	30.06	195	75	16.58	213.5	116	9.61	8.3	4	2.24	yes		25	0	False	PRE -TT 100%
37	362-1407	femme	31 no		98505453	72	220	80	37.4	217.3	102	18.47	152.8	71	6.88	14.7	6	3.97	yes		37	3	False	PRE -TT 1500%

Figure 3.15 : Dataset avant l'encodage

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	genre	age	marie	nb_jours_abo	duree_appel	nb_appel	jou_cout	appel_duree	appel_nb	cout_appel	soli_duree	appel_nb	cout_appel_nui	cout_appel_l	duree_appel_nb	appel_int	cout_appel_i	active_msg_nb	msg_nb	reclamati	churn
2	0	37	1	128	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10	3	2.7	1	25	1	0	0.1666666666666666
3	1	46	0	107	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7	1	26	1	0	0.139344262239
4	1	50	0	137	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32	12.2	5	3.29	0	0	0	0	0.122137404581
5	1	78	1	84	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	0	0	2	0	0.14485981308
6	0	75	1	75	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	0	0	0	3	0.116731517501
7	0	23	0	118	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.7	0	0	0	0	0.13595166163
8	0	67	1	121	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	2.03	1	24	3	0	0.14634146341
9	1	52	1	147	157	79	26.69	103.1	94	8.76	211.6	96	9.53	7.1	6	1.92	0	0	0	0	0.1602316602
10	0	68	0	117	184.5	97	31.37	351.6	80	29.89	215.6	90	9.71	8.7	4	2.35	0	0	1	0	0.061224489798
11	0	43	1	141	258.6	84	43.96	222	111	18.87	326.4	97	14.69	11.2	5	3.02	1	37	0	0	0.16326530612
12	1	47	0	65	129.1	137	21.95	228.5	83	19.42	208.8	111	9.4	12.7	6	3.43	0	0	4	0	0.141176470581
13	0	25	1	74	187.7	127	31.91	163.4	148	13.89	196	94	8.82	9.1	5	2.46	0	0	0	0	0.17370892018
14	0	58	1	168	128.8	96	21.9	104.9	71	8.92	141.1	128	6.35	11.2	2	3.02	0	0	1	0	0.134408602151
15	0	32	0	95	156.6	88	26.62	247.6	75	21.05	192.3	115	8.65	12.3	5	3.32	0	0	3	0	0.13440860215
16	0	39	0	62	120.7	70	20.52	307.2	76	26.11	203	99	9.14	13.1	6	3.54	0	0	4	0	0.094827586
17	0	58	1	161	332.9	67	56.59	317.8	97	27.01	180.6	128	7.23	5.4	9	1.46	0	0	4	0	0.14285714285
18	0	52	1	85	196.4	139	33.39	280.9	90	23.88	89.3	75	4.02	13.8	4	3.73	1	27	1	0	0.098901099
19	0	72	0	93	190.7	114	32.42	218.2	111	18.55	129.6	121	5.83	8.1	3	2.19	0	0	3	0	0.22580645161
20	1	79	0	76	189.7	66	32.25	212.8	65	18.09	165.7	108	7.46	10	5	2.7	1	33	1	0	0.125
21	1	67	0	73	224.4	90	38.15	159.5	88	13.56	192.8	74	8.68	13	2	3.51	0	0	1	0	0.1
22	0	79	1	147	155.1	117	26.37	239.7	93	20.37	208.8	133	9.4	10.6	4	2.86	0	0	0	0	0.05
23	0	26	1	77	62.4	89	10.61	169.9	121	14.44	209.6	64	9.43	5.7	6	1.54	0	0	5	0	0.1333333333333333
24	0	30	1	130	183	112	31.11	72.9	99	6.2	181.8	78	8.18	9.5	19	2.57	0	0	0	0	0.25
25	0	22	0	111	110.4	103	18.77	137.3	102	11.67	189.6	105	8.53	7.7	6	2.08	0	0	2	0	0.0666666666666666
26	0	34	1	132	81.1	86	13.79	245.2	72	20.84	237	115	10.67	10.3	2	2.78	0	0	0	0	0.11627906976
27	0	37	1	174	124.3	76	21.13	277.1	112	23.55	250.7	115	11.28	15.5	5	4.19	0	0	3	0	0.204545454545
28	0	37	1	57	213	115	36.21	191.1	112	16.24	182.7	115	8.22	9.5	3	2.57	1	39	0	0	0.13207547169
29	1	42	1	54	134.3	73	22.83	155.5	100	13.22	102.1	68	4.59	14.7	4	3.97	0	0	3	0	0.18548387096
30	1	64	0	20	190	109	32.3	258.2	84	21.95	181.5	102	8.17	6.3	6	1.7	0	0	0	0	0.1138963730
31	0	47	1	49	119.3	117	20.28	215.1	109	18.28	178.7	90	8.04	11.1	1	3	0	0	1	0	0.17410714285
32	1	23	1	142	84.8	95	14.42	136.7	63	11.62	250.5	148	11.27	14.2	6	3.83	0	0	2	0	0.15068493150
33	0	48	1	75	226.1	105	38.44	201.5	107	17.13	246.2	98	11.08	10.3	5	2.78	0	0	1	0	0.14285714285
34	1	78	1	172	719	191	36.74	212	115	9.65	922	78	13.2	12.6	10	2.4	0	0	2	0	0.116731517501

Figure 3.16 : Résultat de l'encodage

### 3.4 Normalisation et standardisation

La normalisation est une technique de prétraitement des données utilisée en apprentissage automatique qui modifie les valeurs des colonnes numériques pour utiliser une échelle commune centrée autour de 0. Cette technique facilite le traitement des données par l'algorithme. [33]

Pour normaliser et standardiser les données, nous avons utilisé la méthode « StandardScaler () » de la bibliothèque « sklearn.preprocessing ».

La formule de StandardScaler est la suivante :

$$z = \frac{x - \mu}{\sigma}$$

Equation 2 : Formule de normalisation StandardScaler

- $z$  : valeur transformée.
- $x$  : valeur originale.
- $\mu$  : la moyenne des valeurs de la colonne.
- $\sigma$  : l'écart-type des valeurs de la colonne.

```
#Standardisation

sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)

✓ 0.0s
```

Figure 3.17 : Code de standardisation

## 3.5 Sélection des fonctionnalités

### 3.5.1 Matrice de corrélation

La matrice de corrélation est un outil statistique qui mesure la relation entre deux variables dans un ensemble de donnée. Elle fournit une mesure numérique de la force de la relation. [34]

Les valeurs de la matrice de corrélation peuvent varier de -1 à 1 :

- -1 indique une corrélation négative parfaite.
- 0 indique l'absence de corrélation.
- 1 indique une corrélation positive parfaite.

Nous avons appliqué la fonction `.corr()` sur l'ensemble des colonnes de la base de données pour effectuer une analyse de corrélation.

Cela la matrice de corrélation est présentée dans la figure 3.18.



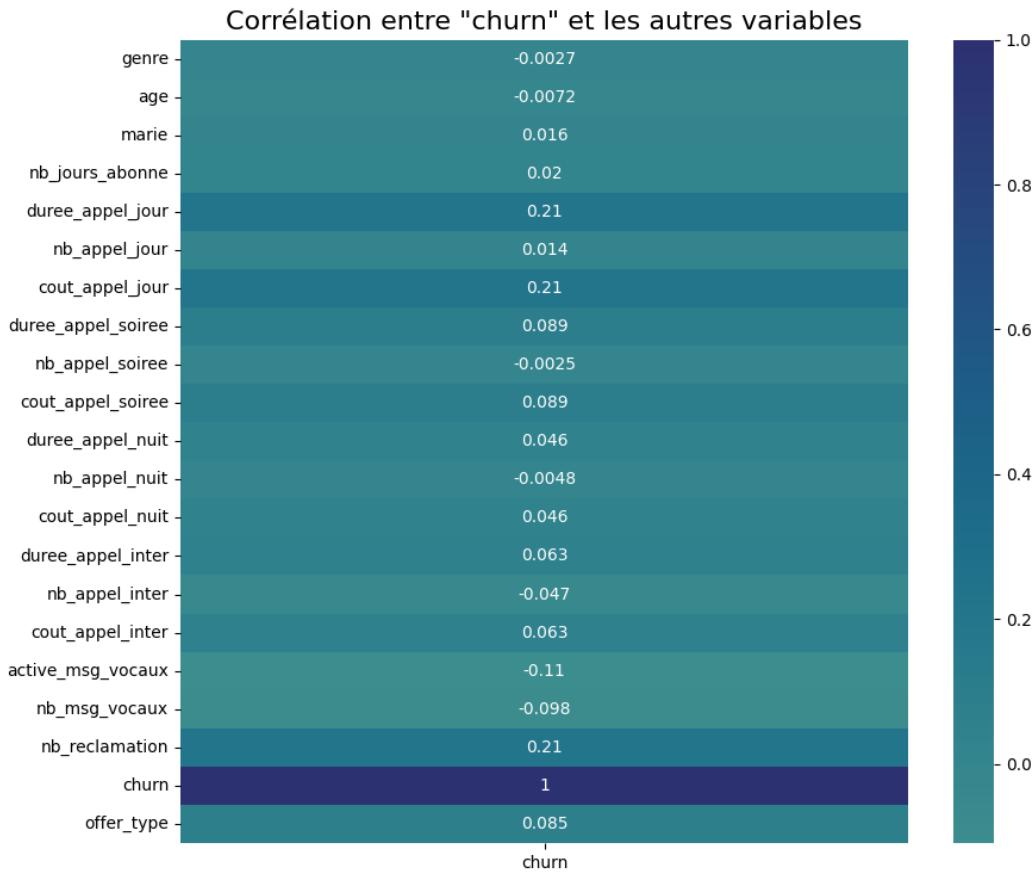
Figure 3.18 : La matrice de corrélation de la Dataset

Chaque carré de la carte thermique illustre l'association entre deux variables. Les couleurs dans la carte thermique indiquent les relations entre ces variables et sont graduées sur une échelle de 1 à 0.

À partir de la matrice de corrélation, nous pouvons constater qu'il existe certaines variables ayant une corrélation de 1.

### 3.5.2 Importance de variables

Dans cette partie, nous allons analyser la relation entre le churn et les autres variables de notre base de données. Cette analyse nous permettra d'identifier les facteurs qui ont le plus d'influence sur le churn. La figure 3.19 représente l'importance de chaque variable.



**Figure 3.19 :** Corrélation entre « churn » et les autres variables

D'après la figure de corrélation entre le "churn" et les autres variables, nous pouvons conclure qu'aucune des colonnes ne présente une corrélation forte avec le "churn".

### 3.6 Conclusion

En résumé, ce chapitre a souligné l'importance cruciale de la préparation des données en apprentissage automatique. Nous avons appris à nettoyer, traiter et transformer nos données pour obtenir des résultats précis. La qualité des données influe directement sur la qualité des résultats, faisant de la préparation des données une étape incontournable.

Nous sommes maintenant prêts à passer à la phase de modélisation pour créer différents modèles et atteindre nos objectifs.

# Chapitre 4 : La modélisation

## Plan

7. Introduction .....	57
8. Découpage de base de données .....	57
9. Modélisation .....	58
10. La validation croisée .....	67
11. Les étapes de construction du modèle .....	68
12. Les mesures de performance .....	68
13. Réglage des hyperparamètres .....	73
14. Conclusion .....	74

## 4.1 Introduction

Une fois que nous avons terminé les étapes de nettoyage et de préparation de nos données de manière efficace, nous sommes prêts à passer à la phase de modélisation. Notre objectif principal dans ce chapitre est de définir et de développer des modèles de machine Learning capables de prédire les taux de désabonnement.

## 4.2 Découpage de base de données

### 4.2.1 Extraction des variables prédictive et cible :

- Variables prédictives : « genre », « âge », « marié », « nb\_jours\_abonne », « nb\_appels\_jour », « cout\_appel\_jour », « nb\_appel\_soirée », « cout\_appel\_soirée », « nb\_appel\_nuit », « cout\_appel\_nuit », « nb\_appel\_inter », « cout\_appel\_inter », « active\_msg\_vocaux », « nb\_msg\_vocaux » , « nb\_reclamation » et « offer\_type ».
- Variable cible : « churn ».

```
#Extraction des variables prédictives et cible
x=data.drop(["churn"], axis =1)
y=data["churn"]
x
```

Figure 4.1 : Extraction de Feature et Target

### 4.2.2 Données d'entraînement et de test

Nous allons utiliser la fonction « train\_test\_split » du package « scikit-learn » pour diviser notre base de données en données d'entraînement et de test. Cette étape est importante pour évaluer les performances des modèles que nous allons créer. Nous avons décidé de segmenter notre base de données en deux parties :

- **Train Set** : 80% des données pour l'apprentissage.
- **Test Set** : 20% des données pour le test dans le but d'évaluer l'efficacité de modèle.

```

np.random.seed(0)
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2, random_state=15, stratify=y)
print("Train set shape:", x_train.shape, y_train.shape)
print("Test set shape:", x_test.shape, y_test.shape)

Train set shape: (4000, 20) (4000,)
Test set shape: (1000, 20) (1000,)

```

**Figure 4.2 :** Division de base de données

La stratification avec l'argument « `stratify=y` » est employée pour faire face au déséquilibre de la variable cible « `churn` ». En effet, la classe “`False`” compte 4289 échantillons, tandis que la classe “`True`” ne compte que 707 échantillons. En utilisant cette méthode, nous assurons une répartition équitable de deux classes dans les ensembles d'apprentissage et de test. Cela permet une modélisation plus équilibrée et précise, en prenant en compte les différentes proportions de chaque classe pour éviter un biais potentiel dans la performance du modèle.

## 4.3 Modélisation

Dans notre cas, nous allons utiliser des modèles de classification, où l'objectif est de prédire à quelle classe appartient une donnée discrète, dans notre cas « `churn` ». Il existe plusieurs algorithmes de classification tels que : Logistic Regression, K-Nearest Neighbors (KNN), eXtreme Gradient Boosting (XGboost), Random Forest, Arbre de décision, Naïve Bayes.

Avant de mettre en œuvre les modèles de classification, nous allons examiner en détail les principes de chaque modèle et analyser leurs avantages et inconvénients.

Dans cette partie, nous allons présenter les différents modèles d'apprentissage supervisé de classification que nous allons employer pour notre analyse de données.

### 4.3.1 Random Forest

Random Forest, connu sous le nom `foret aléatoire`, est un algorithme d'apprentissage automatique qui utilise une technique d'ensemble appelé (`Bootstrap Aggregating`) pour résoudre des problèmes de classification et de régression.

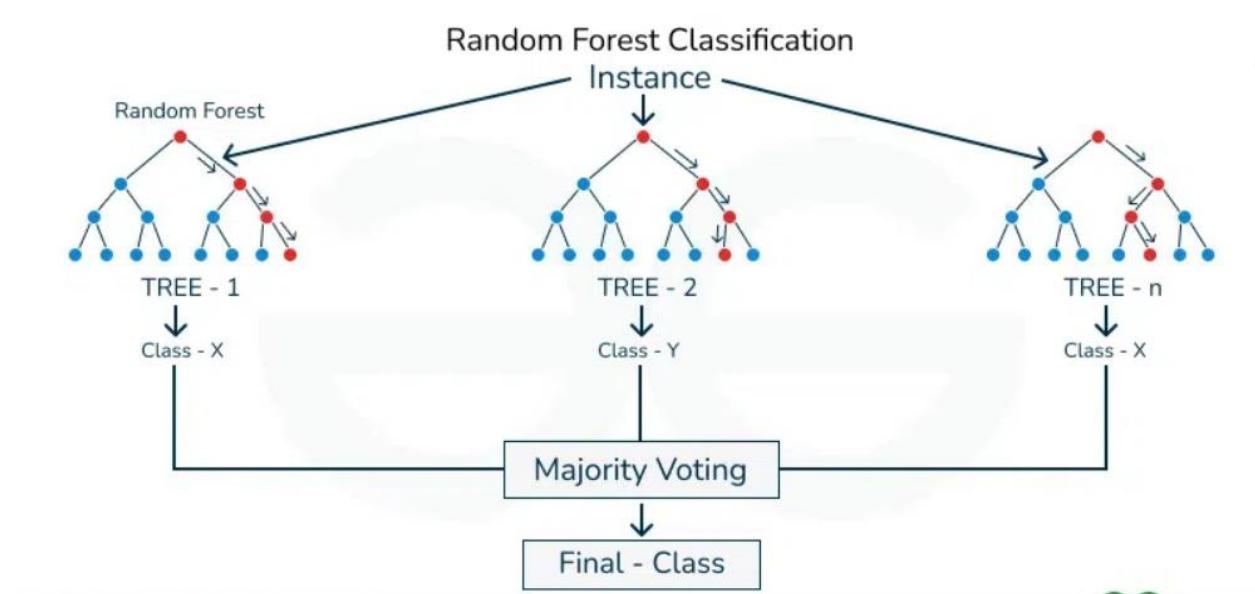
Cette technique consiste à construire plusieurs arbres de décision indépendantes à partir d'échantillons aléatoire de données d'entraînement. Lors de la prédiction, chaque arbre donne une

prédiction et la prédiction finale est obtenue en agrégant les prédictions de tous les arbres, soit par majoritaire pour les problèmes de classification, soit par moyenne pour les problèmes de régressions. [35]

Random Forest utilise cette technique pour améliorer les performances prédictives qui est une méthode permet de créer plusieurs ensembles d'entraînement en effectuant des échantillonnages aléatoires avec remplacement à partir de l'ensemble de données initial. Ensuite, des modèles sont construits en utilisant ces ensembles d'entraînement.

Nous allons décrire le fonctionnement de l'algorithme Random Forest en détaillant les étapes suivantes :

1. Sélectionner aléatoirement des échantillons à partir d'un ensemble de données pour chaque arbre.
2. Construire un arbre de décision pour chaque échantillon et obtenir un résultat de prédiction de chaque arbre.
3. Effectuer un vote majoritaire pour chaque résultat obtenu.
4. Choisir la prédiction ayant obtenu le plus de votes comme prédiction finale.



**Figure 4.3 :** Le fonctionnement de Random Forest [36]

- Les avantages de Random Forest :

- Offre une estimation de l'importance des variables pour identifier les caractéristiques les plus pertinentes pour la prédiction.
- Réduit le risque de surajustement grâce à la méthode d'agrégation.
- Fournit des prédictions plus précises que l'algorithme de l'arbre de décision.
- Capable de gérer efficacement les datasets de grandes tailles.

- Les inconvénients de Random Forest :

- Le temps d'entraînement est plus lent.
- Chaque arbre de décision doit produire une sortie pour les données d'entrée à chaque prédiction, ce qui peut ralentir le processus.
- Plus complexe que les arbres de décision où les décisions peuvent être prises en suivant le chemin de l'arbre.

### 4.3.2 Arbre de décision

L'algorithme des arbres de décision est un type d'apprentissage supervisé qui aide à la prise de décision en construisant un modèle sous forme d'un arbre composé d'un nœud principal, de branches qui partent de ce nœud, de nœuds internes qui représentent les résultats de ces décisions.

Ce modèle est construit en divisant progressivement les données en sous-groupes plus petits en fonction de caractéristiques spécifiques, jusqu'à ce que chaque sous-groupe soit suffisamment homogène dans son score. [37]

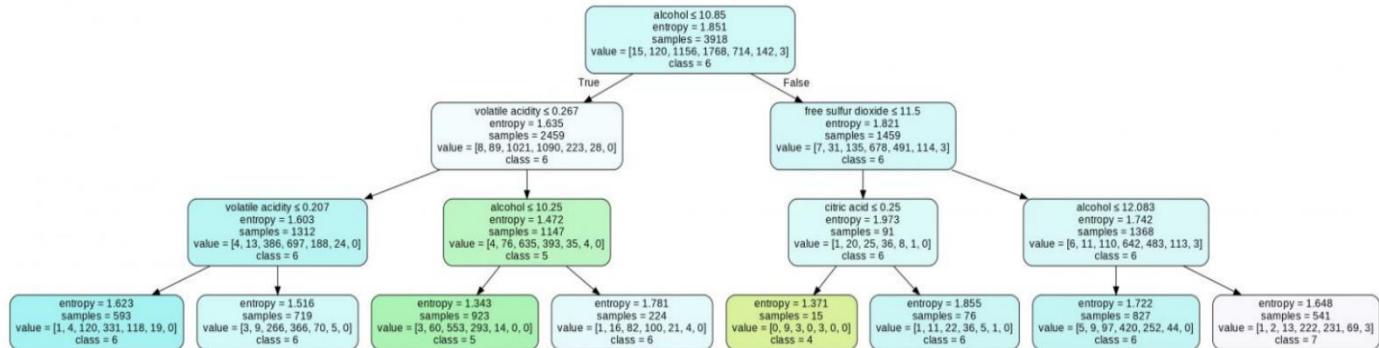


Figure 4.4 : Le fonctionnement d'Arbre de décision

- Les avantages d'Arbre de décision

- Simple à interpréter.
- Non paramétrique.
- Flexibilité pour ajouter de nouvelles décisions.
- Ne nécessite pas de prétraitement compliqué des données.

- Les inconvénients d'Arbre de décision

- Faible performance.
- Risque de sur-apprentissage.
- Difficulté de modéliser des relations complexes.
- Coût d'entraînement élevé.

### 4.3.3 Différence entre arbre de décision et Random Forest

La Random Forest est composée d'une collection d'arbres de décision, mais il existe de nombreuses différences entre eux. Le tableau 4.1 représente les différences entre Random Forest et l'arbre de décision.

**Tableau 4.1 :** Différence entre arbre de décision et Random Forest

Decision Tree	Random Forest
Un seul arbre de décision est construit	Plusieurs arbres sont construits puis combinés
Plus facile à interpréter	Plus difficile à interpréter en raison du nombre d'arbre
Chaque arbre prend une décision	Décision final obtenue par le vote majoritaire ou la moyenne de tous les arbres
Plus rapide à construire	Plus lent à construire plusieurs arbres

#### 4.3.4 Régression Logistique

La régression logistique est un algorithme d'apprentissage supervisé utilisé pour la classification binaire ou multi classe. Son objectif est de prédire la probabilité d'une variable cible en fonction de variables d'entrée, en établissant une relation linéaire entre les deux. Elle se base sur la fonction logistique, également appelée fonction sigmoïde, qui transforme une valeur continue en une probabilité comprise entre 0 et 1. [38]

$$\text{Sigmoid}(x) = \frac{1}{1+e^{-x}}$$

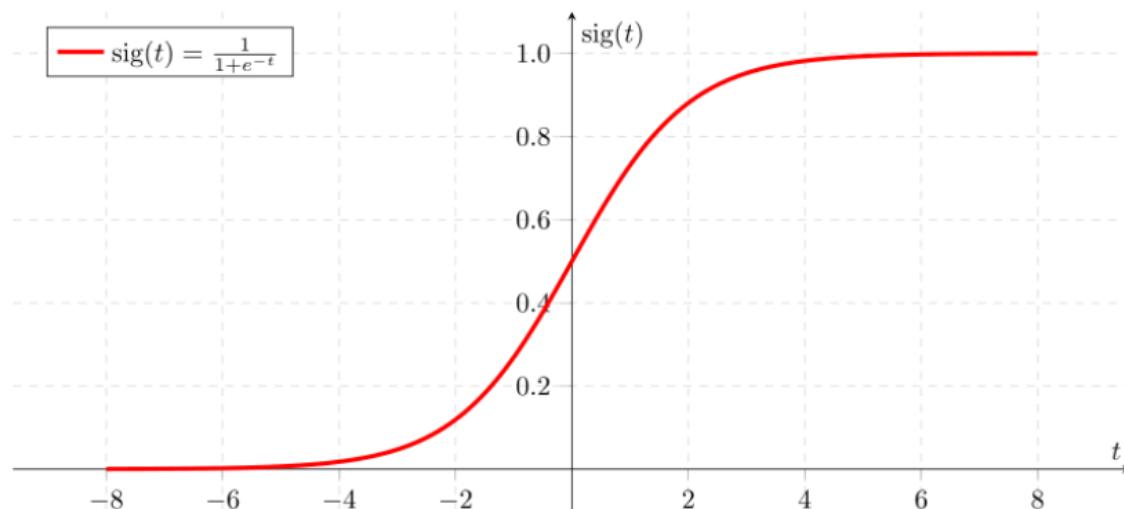


Figure 4.5 : Modèle de régression logistique [39]

- Les avantages de la régression logistique
  - Simple à utiliser et à interpréter
  - Moins sensible aux valeurs manquantes
  - Possibilité de régularisation pour éviter le surapprentissage
- Les inconvénients de la régression logistique
  - Difficulté à modéliser des variables catégorielles avec de nombreuses catégories
  - Difficulté à gérer les interactions complexes
  - Incapable de modéliser des relations non linéaires.

### 4.3.5 K-nearest neighbors : KNN

KNN (K-Nearest Neighbors) ou K plus proches voisins est un algorithme d'apprentissage automatique supervisé simple qui peut être utilisé pour la classification et la régression et utilise la distance entre les échantillons pour faire la prédiction d'un nouvel échantillon. Il est basé sur l'idée que les points qui sont proches les uns des autres dans l'espace des caractéristiques ont plus de chances d'appartenir à la même classe ou d'avoir une valeur de sortie similaire. [40]

Pour prédire la classe ou la valeur de sortie d'un nouvel échantillon, KNN recherche les k échantillons les plus proches dans l'ensemble de données et utilise leur classe ou leur valeur de sortie pour estimer celle du nouvel échantillon.

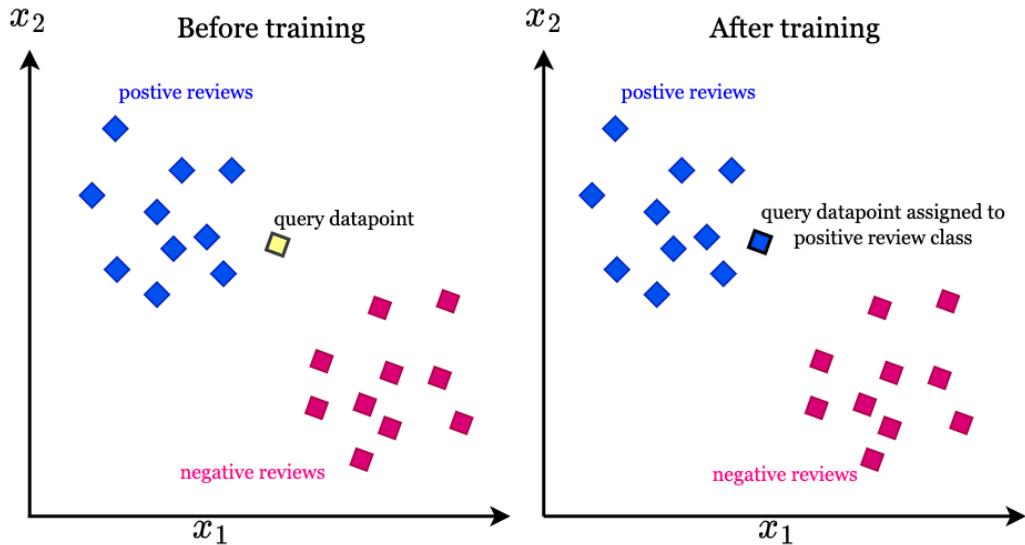


Figure 4.6 : Algorithme K-Nearest Neighbors [41]

Nous allons expliquer le fonctionnement de l'algorithme K-NN en se basant sur les étapes suivantes :

Étape 1 : Choisir le nombre K de voisins.

Étape 2 : Calculer la distance entre le nouvel échantillon et les échantillons existants.

**Tableau 4.2 : Les formules des distances**

Distance euclidienne	Distance Manhattan
$D_e(x, y) = \sqrt{\sum_{j=1}^n (x_j - y_j)^2}$	$D_m(x, y) = \sum_{i=1}^k  x_i - y_i $

Étape 3 : Trier les distances et sélectionner les K voisins les plus proches en fonction de la distance calculée.

Étape 4 : Compter le nombre de points appartenant à chaque classe parmi ces K voisins.

Étape 5 : Attribuez le nouvel échantillon à la classe la plus représentée parmi ces K voisins.

Étape 6 : Notre modèle est prêt pour faire la prédiction.

- **Les avantages de KNN**

- Facile à implémenter et à utiliser.
- Applicable à la classification et à la régression
- Efficace pour les datasets de petite taille.

- **Les inconvénients de KNN**

- Sensible à la présence de valeurs aberrantes.
- Difficile de choisir la bonne méthode de calcul de la distance et le nombre de voisins "k".
- Temps de prédiction lent.
- Stocke toutes les données d'entraînement.

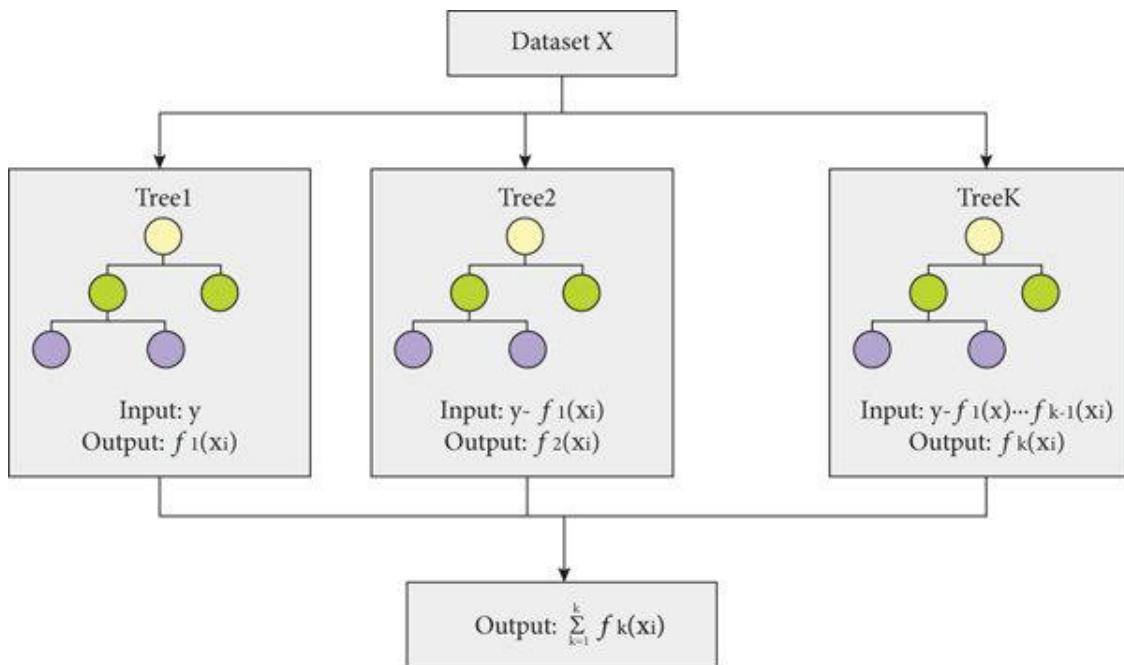
### 4.3.6 eXtreme Gradient Boosting: XGboost

eXtreme Gradient Boosting (XGBoost) est un algorithme qui fait partie de la branche supervisée de l'apprentissage automatique qui utilise des arbres de décision et des méthodes de gradient boosting pour créer des modèles prédictifs précis et utilisée pour résoudre des problèmes de classification et de régression.

Il permet de construire de manière itérative une séquence d'arbre de décision, où chaque arbre tente de corriger les erreurs commises par les arbres précédents. Les méthodes de Gradient Boosting, telles que XGBoost et Gradient Boosting Machines, sont des techniques qui combinent les prédictions de plusieurs modèles plus simples pour former un modèle global plus performant. [42]

Nous allons expliquer le fonctionnement de l'algorithme XGboost en se basant sur les étapes suivantes :

1. Création d'un premier modèle d'arbre de décision.
2. Construction d'un autre arbre de décision pour prédire les erreurs résiduelles du premier modèle.
3. Fusion des deux modèles pour créer un modèle plus précis.
4. Création d'une séquence de modèles d'arbres de décision, où chaque modèle apprend à corriger les erreurs du modèle précédent.
5. Utilisation de techniques de régularisation et d'optimisation pour améliorer les performances du modèle en évitant le surapprentissage.
6. Utilisation du modèle XGBoost pour prédire les sorties pour de nouvelles données.



**Figure 4.7:** Le processus de XGBoost [43]

- **Les avantages de XGBoost**

- Haute performance avec une vitesse de traitement rapide et une précision élevée.
- Gère les données manquantes et les valeurs aberrantes.
- Traite les problèmes de classification avec les datasets déséquilibrés.

- **Les inconvénients de XGBoost**

- Nécessite de données prétraitées.

- Difficile de configurer et d'ajuster les hyperparamètres.
- Sur-apprentissage possible des données d'entraînement.

#### 4.3.7 Naïve Bayes

Naïve Bayes est un algorithme simple de type classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance de hypothèses. Il utilise les probabilités a priori des classes et les probabilités conditionnelles des caractéristiques étant donnée chaque classe pour estimer la probabilité postérieure d'apprentissage à une classe donnée. Il est basé sur le théorème de Bayes. [44]

- Théorème de Bayes : En apprentissage automatique, nous sommes souvent intéressés par la sélection de la meilleure hypothèse ( $h$ ) à partir des données ( $d$ ).

Le théorème de Bayes est formulé comme suit :

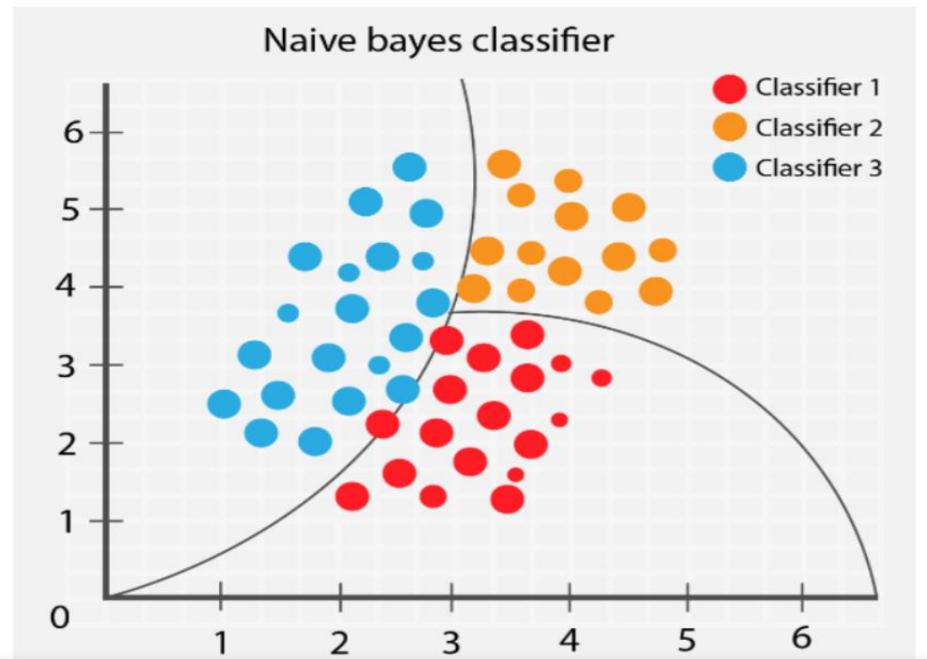
$$P(h | d) = (P(d | h) * P(h)) / P(d)$$

**Equation 3** : Formule de théorème de Bayes

- $P(h | d)$  : la probabilité de l'hypothèse  $h$  étant donnée les données  $d$ .
- $P(d | h)$  : la probabilité pour les données  $d$  étant donné que l'hypothèse  $h$  était vraie.
- $P(h)$  : la probabilité que l'hypothèse  $h$  soit vraie.
- $P(d)$  : la probabilité des données.

Il existe également les Bayes naïfs multinomiaux et les Bayes naïfs de Bernoulli.

Nous choisissons le naïf gaussien parce que c'est le plus simple et le plus populaire.



**Figure 4.8:** Le processus de Naïve Bayes classifier [45]

#### 4.4 La validation croisée

Est une technique de Machine Learning qui permet d'évaluer les performances d'un modèle d'apprentissage automatique et pour estimer sa capacité à généraliser à de nouveaux exemples.

L'un des avantages de la validation croisée, en particulier dans le contexte de la lutte contre le surajustement (overfitting), est qu'elle fournit une estimation plus fiable de la performance du modèle sur des données.

La procédure de validation croisée implique de diviser l'ensemble de données initial en deux parties : un ensemble d'apprentissage et un ensemble de validation. Le modèle est entraîné sur l'ensemble d'apprentissage et ses performances sont évaluées sur l'ensemble de validation. [46]

Après avoir identifié un problème de surapprentissage dans notre algorithme, nous avons pris la décision d'adopter la technique de validation croisée afin d'améliorer la performance de modèle en réduisant le risque de surajustement.

## 4.5 Les étapes de construction du modèle

La construction d'un modèle passe par différentes étapes, telles que :

- **Création du modèle** : Nous créons un modèle.
- **Entraînement du modèle** : Nous entraînons le modèle en utilisant la méthode "fit()" qui prend x\_train et train comme paramètres.
- **Prédiction** : Une fois le modèle entraîné, nous utilisons la méthode "predict()" pour effectuer des prédictions sur de nouvelles données.
- **Évaluation du modèle** : Nous évaluons les performances du modèle en utilisant des métriques telles que la matrice de confusion, l'accuracy, la précision, le F1-score, etc.

Grâce à ces étapes, nous pouvons construire notre modèle et l'évaluer afin d'obtenir des prédictions précises et fiables.

## 4.6 Les mesures de performance

Dans cette section, nous allons mettre en évidence les métriques d'évaluation qui permettent d'effectuer une analyse pertinente de la performance des modèles. Pour chaque type de problème d'apprentissage, il existe diverses mesures qui peuvent être utilisées.

Dans notre étude, nous avons étudié les problèmes de classification, pour lesquels nous avons utilisé un ensemble de métriques telles que l'exactitude (Accuracy), la précision (Precision), le rappel (Recall) et le score F1 (F1-score).

### 4.6.1 Matrice de confusion

La matrice de confusion n'est pas une métrique, mais elle représente l'un des concepts clés dans l'évaluation des performances d'un modèle de classification. Il se présente sous la forme d'un tableau qui permet de mesurer la qualité de la prédiction en comparant les données d'entrée avec les données prédites par le modèle. [47]

Cette matrice se compose de quatre éléments :

- **Vrais positifs (VP)** : La valeur réelle positive et la valeur prédite positive.
- **Vrais négatifs (VN)** : La valeur réelle négative et la valeur prédite négative.
- **Faux positifs (FP)** : La valeur réelle négative et la valeur prédite positive.
- **Faux négatifs (FN)** : La valeur réelle positive et la valeur prédite négative.

La figure 4.9 présente la matrice de confusion.

		Données prédites par l'algorithme	
		Non churn ↗	Churn ✘
Données réelles	Non churn ↗	Vrai négatif	Faux positif
	Churn ✘	Faux négatif	Vrai positif

**Figure 4.9:** La matrice de confusion

En utilisant cette matrice de confusion, il est possible de calculer plusieurs critères de performances pour évaluer un modèle de classification. Ces critères sont présents dans un rapport de classification qui permet de mesurer la qualité des prédictions du modèle.

#### 4.6.1.1 Precision

La précision est une mesure de performance utilisée en classification pour évaluer la qualité des prédictions positives faites par un modèle qui sont effectivement positives. Elle est définie comme le rapport entre le nombre de vrais positifs et la somme des vrais positifs et des faux positifs. [48]

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

**Equation 4 :** Formule de precision

#### 4.6.1.2 Accuracy

L'accuracy est une métrique de performance permet de mesurer la performance d'un modèle en termes de proportion de prédictions correctes par rapport à l'ensemble des prédictions réalisées.

Elle est calculée en utilisant la formule suivante : Nombre de bonnes prédictions / Nombre total de prédictions. [49]

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

**Equation 5 :** Formule d'accuracy

NB : Cette mesure de performance est particulièrement efficace lorsqu'il y a un équilibre dans la base de données.

#### 4.6.1.3 Recall

Le Recall (Rappel) est une métrique qui mesure le nombre de prédictions positives correctes par rapport au nombre total de données positives. Il permet de répondre à la question suivante : sur tous les exemples positifs, combien ont été correctement identifiés par le modèle. [50]

La formule suivante est utilisée pour calculer le rappel :

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

**Equation 6 :** Formule de Recall

#### 4.6.1.4 F1-Score

Le score F1 est une métrique de performance qui combine la précision et le rappel en une moyenne harmonique, ce qui permet d'obtenir une vue d'ensemble de la qualité des prédictions d'un modèle. [50]

Nous calculons le F1 score avec la formule suivante :

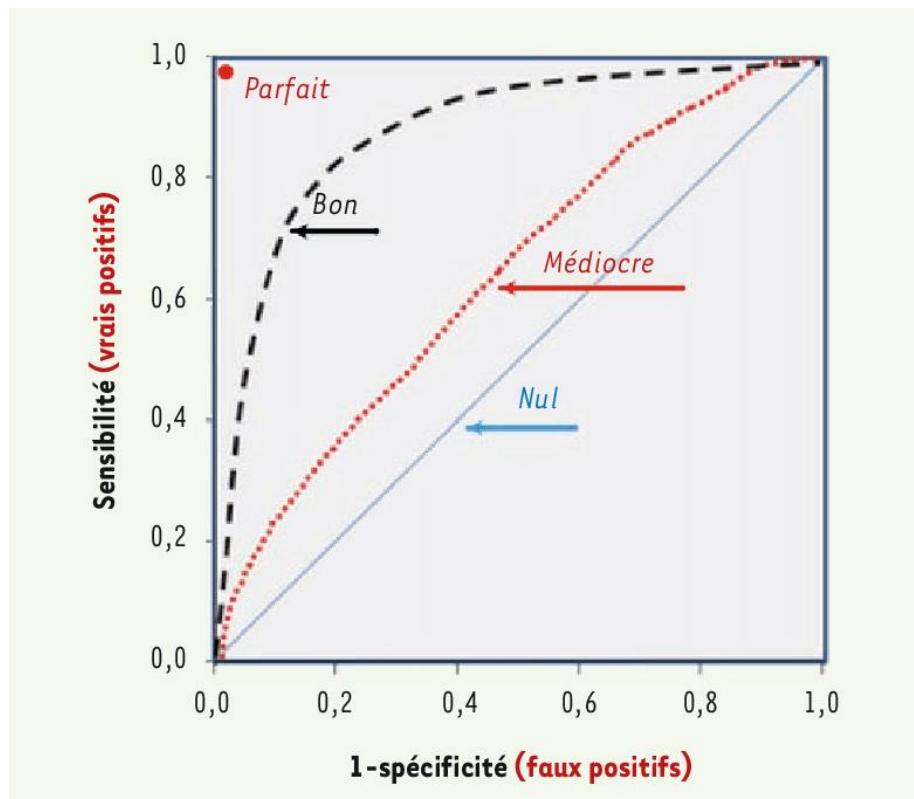
$$F1 - Score = \frac{2TP}{2TP + FP + FN}$$

**Equation 7 :** Formule de F1-Score

#### 4.6.2 La courbe ROC-AUC

La courbe ROC est l'abréviation de Receiver Operating Characteristics, et AUC est l'Area Under the Curve est une courbe de performance qui permet d'évaluer les performances d'un modèle de classification en traçant la sensibilité (taux de vrais positifs) en fonction de la spécificité (taux de faux positifs) pour différents seuils de classification. [51]

La figure 4.10 illustre la courbe ROC-AUC.



**Figure 4.10:** Lecture de courbe ROC

Une courbe ROC idéale se situe dans le coin supérieur gauche du graphique, ce qui signifie qu'elle présente un taux élevé de vrais positifs et un taux faible de faux positifs pour tous les seuils de classification.

#### 4.6.3 La courbe Learning Curve

La courbe d'apprentissage (Learning Curve en anglais) est un graphique qui montre comment la performance d'un modèle évolue en fonction de la quantité de données d'entraînement utilisées. Elle compare l'erreur d'entraînement et l'erreur de test du modèle en fonction de la quantité de données d'entraînement utilisées pour entraîner le modèle. Elle permet aussi de visualiser la relation entre l'erreur d'entraînement et l'erreur de test. [52]

A partir de cette courbe, les développeurs peuvent déterminer si le modèle a besoin de plus de données d'entraînement, d'un ajustement des hyperparamètres ou d'une régularisation pour éviter le surapprentissage.

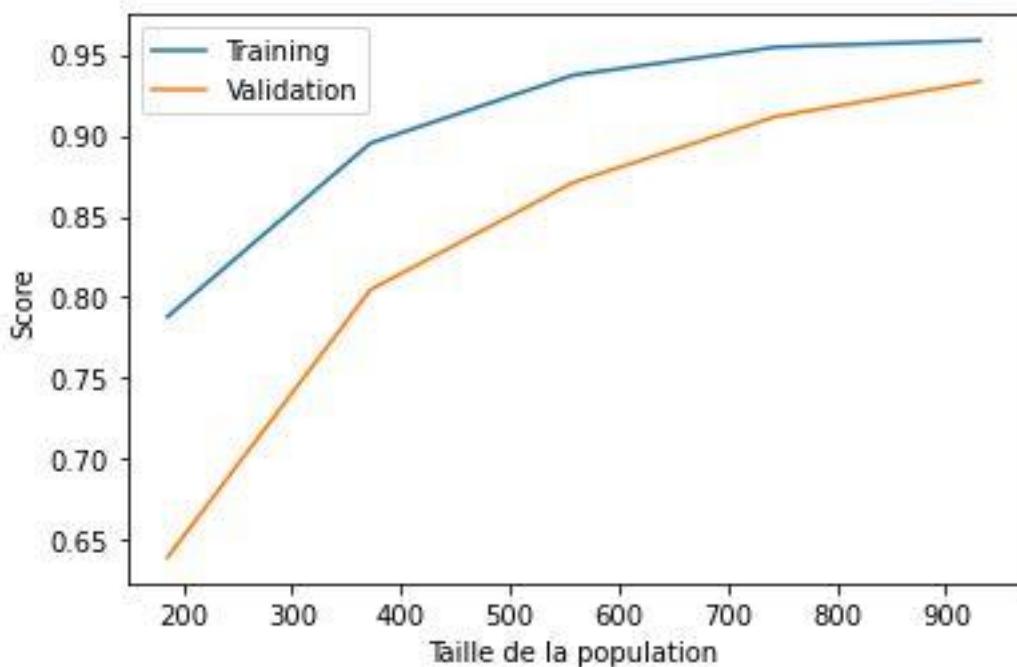


Figure 4.11: Exemple de courbe Learning Curve

## 4.7 Réglage des hyperparamètres

Pour obtenir les meilleurs résultats avec la plupart des modèles de machine Learning, il est important de trouver les meilleurs hyperparamètres pour le modèle. « GridSearch » et « RandomizedSearchCV » sont deux techniques populaires pour trouver les meilleurs hyperparamètres.

- **GridSearch** Une technique de recherche de paramètres pour trouver les meilleurs hyperparamètres pour un modèle de Machine Learning. Elle consiste à tester toutes les combinaisons d'hyperparamètres possibles, à l'aide d'une grille prédéfinie. Le but est de trouver les meilleurs hyperparamètres pour le modèle, qui maximisent ses performances. Lorsque la grille d'hyperparamètres est grande, la recherche de toutes les combinaisons possibles peut être très coûteuse en temps de calcul. [53]
- **RandomizedSearchCV** Une méthode de recherche de paramètres qui évalue un nombre limité de combinaisons d'hyperparamètres en utilisant un échantillonnage aléatoire pour trouver les meilleurs paramètres pour un modèle. Cette technique est plus rapide que GridSearch car elle teste un nombre limité de combinaisons. Cette section présente les informations détaillées sur les hyperparamètres que nous avons choisis d'utiliser dans chaque modèle. [53]

Le tableau 4.3 détaille chaque algorithme utilisé et les hyperparamètres correspondants qui ont été ajustés pour optimiser les performances des modèles.

**Tableau 4.3 :** Les hyperparamètres de chaque modèle

Modèle	Hyperparamètres	Explication d'hyperparamètres
KNN	n_neighbors : 7 metric : Manhattan	Nombre de voisins les plus proches.

XGBOOST	subsample: 0.8 min_child_weight: 3 max_depth: 7 gamma: 1.2 colsample_bytree: 0.5	La fraction d'échantillons d'apprentissage à utiliser lors de la construction de chaque arbre de décision. Le poids minimum pour les nœuds terminaux. La profondeur maximale de l'arbre de décision. La réduction minimale de la fonction de perte. La fraction de colonnes (caractéristiques).
Random Forest	max_depth: 11 n_estimators: 250	La profondeur maximale de l'arbre de décision. Le nombre d'arbres
Decision Tree	criterion: entropy  max_depth: 7	La fonction de mesure de qualité de la division d'un noeud en deux sous-groupes lors de la construction de l'arbre.  La profondeur maximale d'un arbre.
Naive Bayes	var_smoothing 5.76724670052	La fraction de colonnes (caractéristiques).

## 4.8 Conclusion

Ce chapitre traitera des techniques de Machine Learning. Nous explorerons en profondeur les divers modèles de classification. Nous aborderons ensuite la phase d'évaluation, où chaque modèle sera analysé pour identifier celui qui garantit la plus grande précision.

# Chapitre 5 : Évaluation et optimisation de la performance des modèles

## Plan

1.	Introduction .....	76
2.	Évaluation de chaque modèle.....	76
3.	Comparaison et évaluation des algorithmes utilisés .....	85
4.	L'Ensemble Learning .....	86
5.	Ajustement de modèle. ....	88
6.	Conclusion .....	88

## 5.1 Introduction

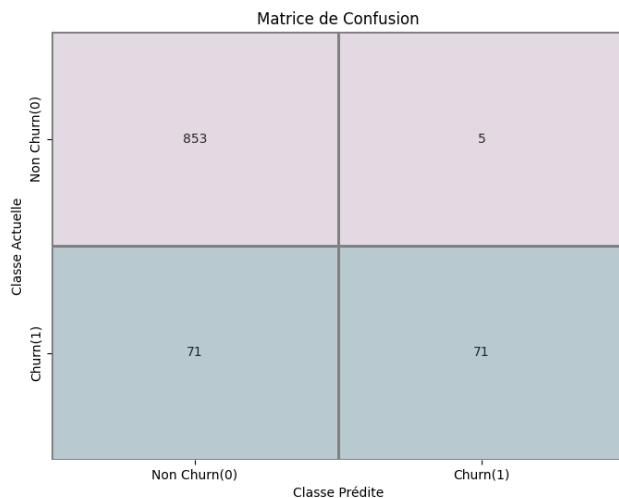
Après avoir développé les modèles d'apprentissage automatique, nous comparons leur performance individuelle pour identifier le modèle le plus performant. Ce chapitre présente les évaluations de chacun des modèles, basées sur des indicateurs de performances spécifiques. En se basant sur ces résultats, nous avons apporté des améliorations à plusieurs modèles pour optimiser leurs résultats.

## 5.2 Évaluation de chaque modèle

Dans cette partie, nous allons présenter les résultats des matrices de confusion de chaque modèle utilisé afin d'évaluer leur performance.

### 5.2.1 Random Forest

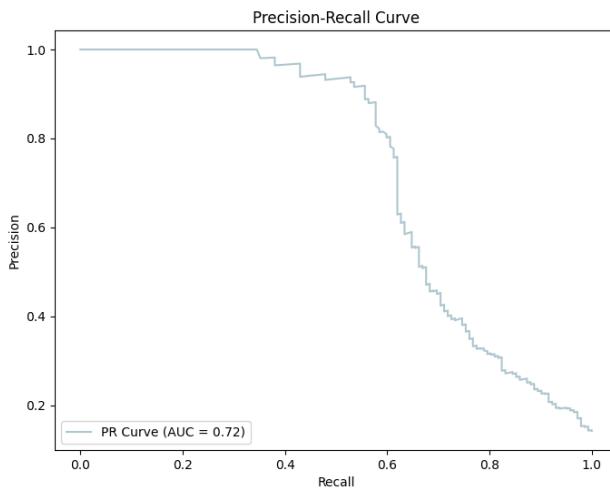
La matrice de confusion du modèle « Random Forest » est illustrée dans la figure suivante :



**Figure 5.1:** Matrice de confusion de Random Forest

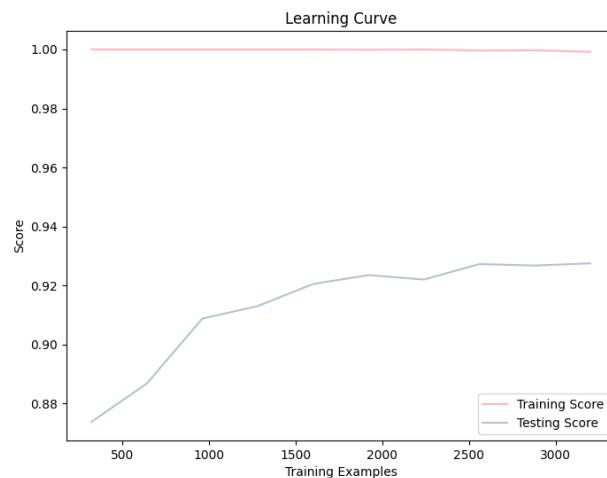
L'algorithme a prédit :

- 853 vrais négatifs et 71 vrais positifs.
- 71 faux négatifs et 5 faux positifs.



**Figure 5.2:** Precision-Recall Curve de Random Forest

La courbe Précision-Rappel avec une Aire Sous la Courbe (AUC) de 0,72 suggère que le modèle de classification peut distinguer avec modération à bien les exemples positifs des négatifs. Bien que cette performance puisse être acceptable pour de nombreux cas, il est crucial de considérer les besoins précis de l'application et l'impact des faux positifs et des faux négatifs. Pour une évaluation plus approfondie, il est conseillé d'analyser en détail les valeurs de précision et de rappel, ainsi que leurs compromis, afin de comprendre comment le modèle se comporte à différents seuils de classification.



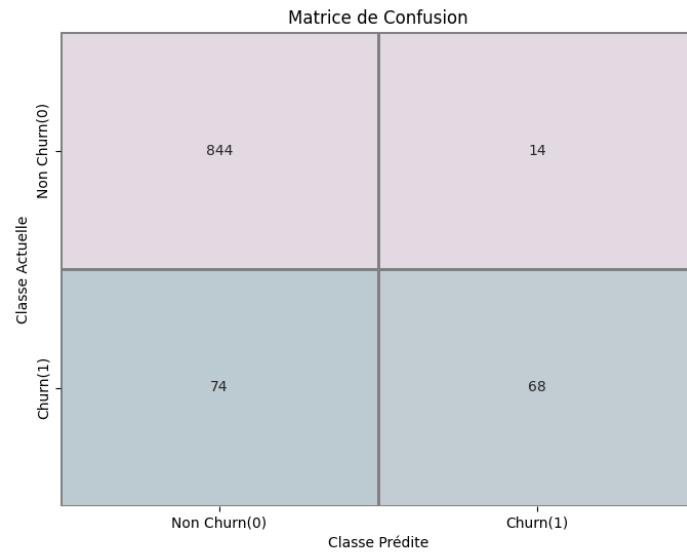
**Figure 5.3:** Learning Curve de Random Forest

Le modèle Random Forest s'adapte bien aux données, car sa précision sur les données de test s'améliore lorsque le nombre de données d'entraînement augmente. Cependant, il atteint un seuil où

l'ajout de données n'améliore plus significativement sa précision. La précision sur les données de test se stabilise alors, ce qui signifie que le modèle ne peut plus extraire d'informations utiles des données d'entraînement supplémentaires.

### 5.2.2 Decision Tree

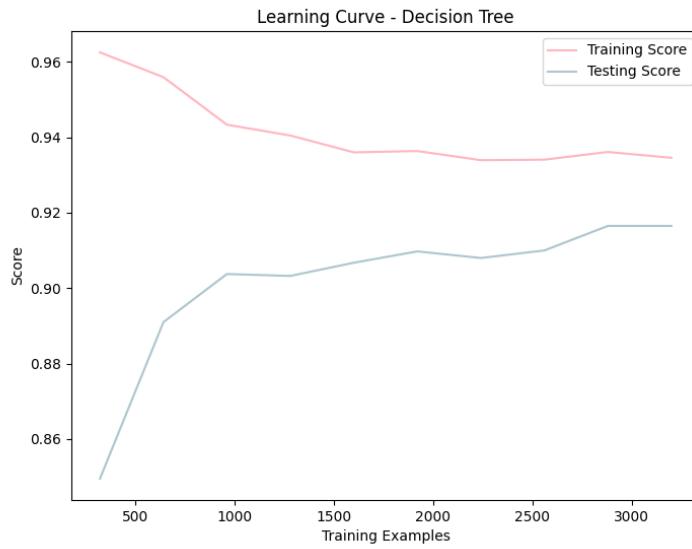
La matrice de confusion du modèle « Decision Tree » est illustrée dans la figure suivante :



**Figure 5.4:** Matrice de confusion de Decision Tree

L'algorithme a prédit :

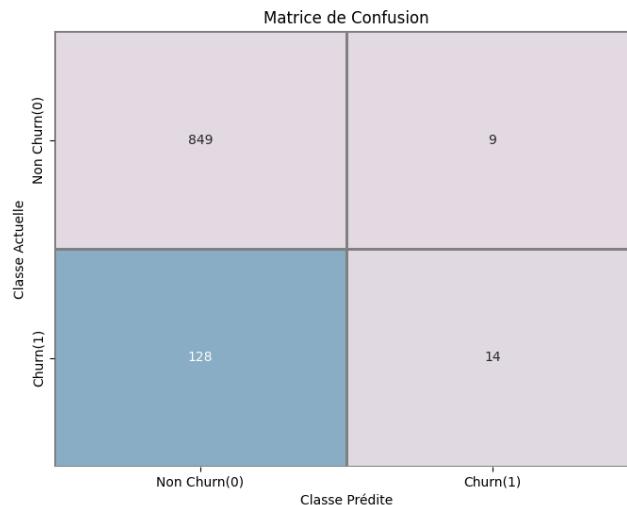
- 844 vrais négatifs et 68 vrais positifs.
- 74 faux négatifs et 14 faux positifs.



**Figure 5.5:** Learning Curve de Decision Tree

L'arbre de décision peut bien généraliser les données, car sa précision sur les données de test s'améliore avec la taille des données d'entraînement. Mais à un moment donné, ajouter plus de données ne conduit plus à une amélioration significative de la précision du test. Le score de test se stabilise et devient presque parallèle au score d'entraînement. Cette stabilisation du score de test indique que le modèle ne peut plus apprendre de nouvelles informations à partir des données d'entraînement supplémentaires.

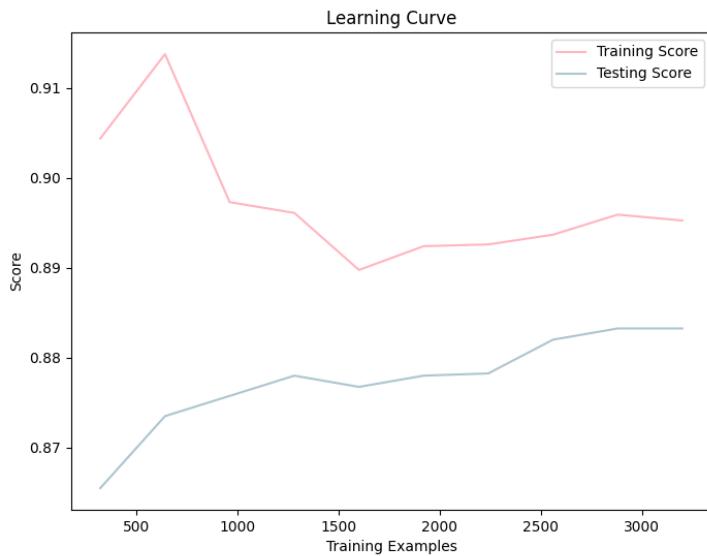
### 5.2.3 Logistic Regression



**Figure 5.6:** Matrice de confusion de Logistic Regression

L'algorithme a prédit :

- 849 vrais négatifs et 14 vrais positifs.
- 128 faux négatifs et 9 faux positifs.

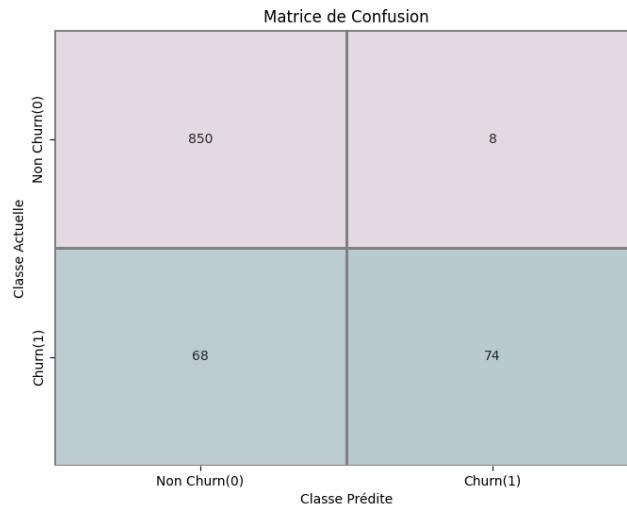


**Figure 5.7:** Learning Curve de Logistic Regression

D'après la courbe suivante, nous pouvons voir qu'au début du processus d'apprentissage, celui-ci est relativement simple et ne parvient pas à capturer les relations complexes présentes dans les données. À mesure que nous ajoutons davantage de données d'entraînement, le modèle commence à s'adapter avec plus de précision aux données. Cependant, il arrive un moment où le modèle commence à trop s'ajuster aux données d'entraînement spécifiques. C'est ce qu'on appelle le surajustement, et la courbe représentant les performances sur l'ensemble d'entraînement reste généralement en dessous de la courbe représentant les performances sur l'ensemble de test.

#### 5.2.4 XGBoost

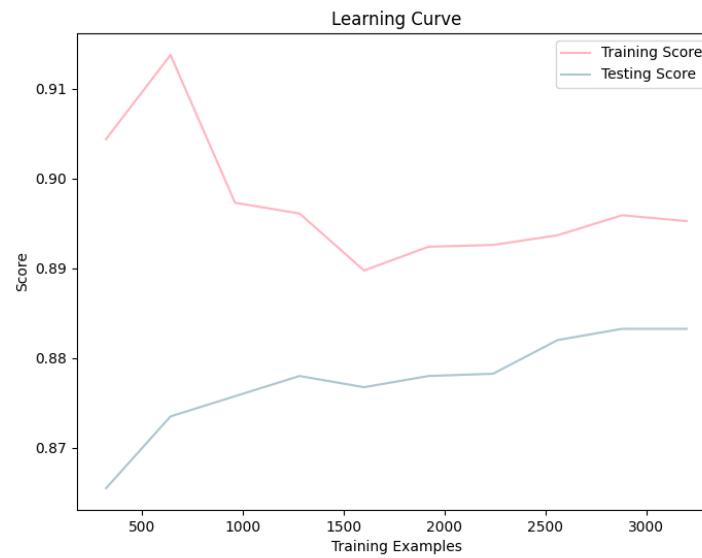
Voici la représentation graphique de la matrice de confusion obtenue pour le modèle XGBoost :



**Figure 5.8:** Matrice de confusion de XGBoost

L'algorithme a prédit :

- 850 vrais négatifs et 74 vrais positifs.
- 68 faux négatifs et 8 faux positifs.



**Figure 5.9:** Learning Curve de XGBoost

Le modèle XGBoost a une capacité de généralisation des données efficace, ce qui signifie qu'il peut améliorer sa précision sur l'ensemble de test à mesure que la taille de l'ensemble d'entraînement augmente. Cependant, il atteint un point de saturation où l'ajout de nouvelles données n'entraîne plus d'amélioration significative de sa précision sur le test. Lorsque le score de test se stabilise, indiquant que le modèle ne peut plus tirer de nouvelles informations des données d'entraînement supplémentaires.

### 5.2.5 K-Nearest Neighbors : KNN

La matrice de confusion du modèle « KNN » est illustrée dans la figure suivante :

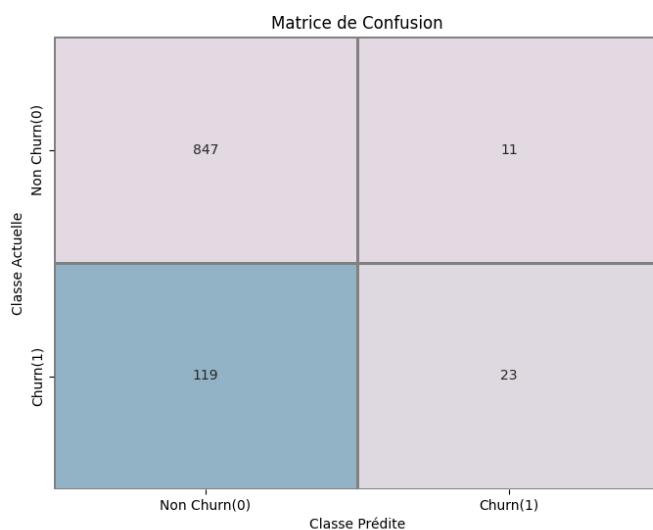
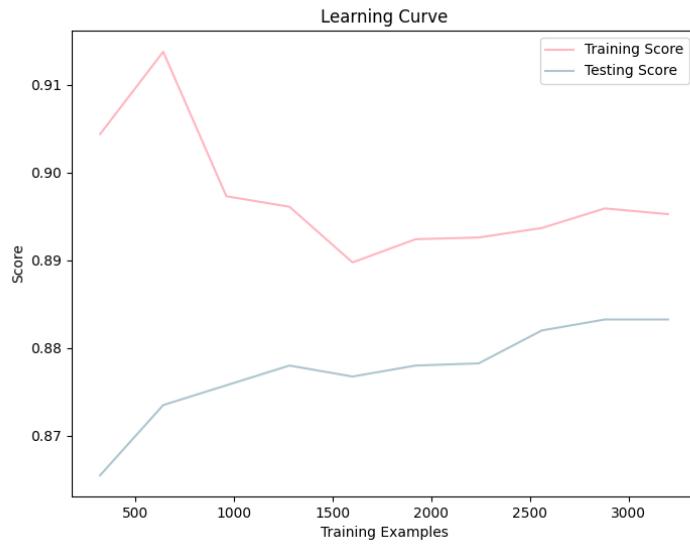


Figure 5.10: Matrice de confusion de KNN

L'algorithme a prédit :

- 847 vrais négatifs et 23 vrais positifs.
- 119 faux négatifs et 11 faux positifs.

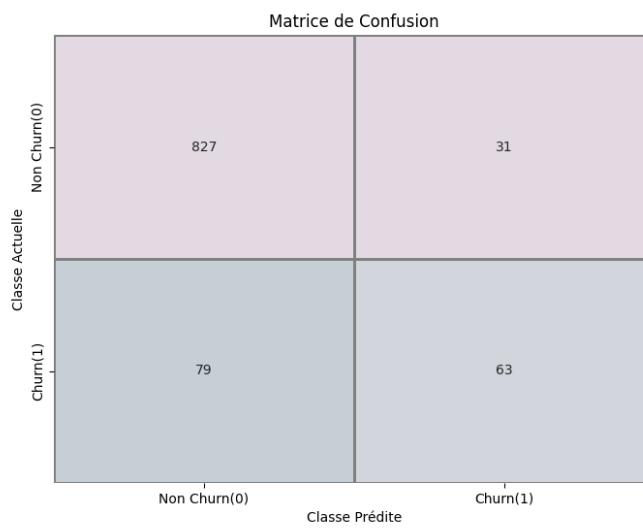


**Figure 5.11:** Learning Curve de KNN

Au début de l'apprentissage, le modèle KNN commence à apprendre et à s'adapter aux données d'entraînement. Cependant, il peut arriver un moment où le modèle devient trop adapté aux données d'entraînement spécifiques. Cela peut se produire lorsque la valeur de  $k$  (le nombre de voisins les plus proches) est faible, ce qui rend le modèle excessivement sensible aux données locales de l'ensemble d'entraînement. Lorsque le modèle est trop adapté aux données d'entraînement, il peut ne pas généraliser correctement sur de nouvelles données qui n'ont pas été utilisées pendant l'entraînement.

### 5.2.6 Naïve Bayes

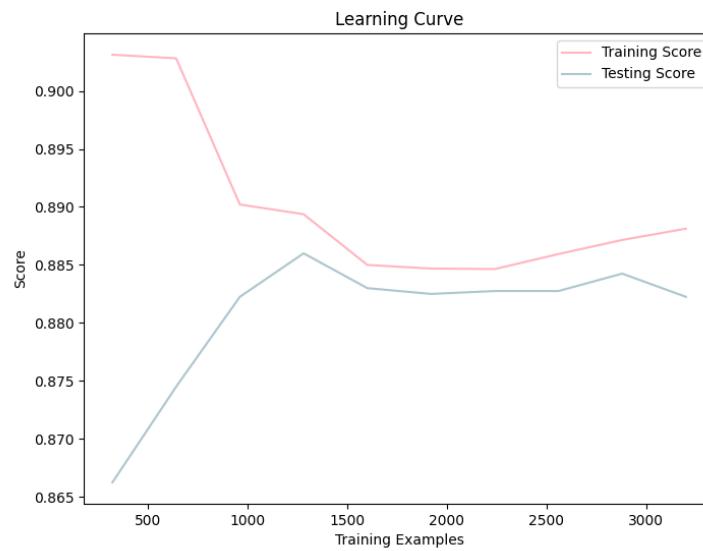
La figure ci-dessous illustre la matrice de confusion pour le modèle naïve bayes :



**Figure 5.12:** Matrice de confusion de Naïve Bayes

L'algorithme a prédit :

- 827 vrais négatifs et 63 vrais positifs.
- 79 faux négatifs et 31 faux positifs.



**Figure 5.13:** Learning Curve de Naïve Bayes

### 5.3 Comparaison et évaluation des algorithmes utilisés



**Figure 5.14:** Comparaison des performances des différents modèles

Après comparaison des différents modèles sur diverses métriques de performance, le modèle Random Forest affiche la meilleure précision globale (Accuracy) à 0,927. De plus, les modèles XGBoost, KNN et Logistic Regression se démarquent avec des performances similaires en termes de rappel (Recall) autour de 0,99, indiquant leur efficacité dans la détection des vrais positifs (True Positives). Le modèle XGBoost obtient le meilleur score en précision pour la classe positive à 0,93. En termes de F1-score, XGBoost domine avec une valeur de 0,96. Enfin, en ce qui concerne l'aire sous la courbe ROC (AUC), Random Forest et Decision Tree affichent des performances autour de 0,92.

## 5.4 L'Ensemble Learning

L'apprentissage ensembliste (ou ensemble learning) est une technique qui repose sur la combinaison de multiples algorithmes de machine learning pour accroître les performances du modèle d'apprentissage, et parvenir à un niveau de précision supérieur à celui qui serait réalisé si on utilisait un de ces algorithmes pris séparément. [54]

En termes simples, il existe deux approches principales en apprentissage par ensemble : le vote de classificateurs et le stacking. Le vote combine les prédictions de plusieurs modèles pour choisir la classe la plus fréquemment prédite, tandis que le stacking utilise les prédictions comme caractéristiques pour entraîner un autre modèle. Nous explorerons ces deux méthodes pour nos trois meilleurs modèles.

- Voting Classifier

```
Voting Classifier Accuracy: 0.923
Classification Report for Voting Classifier:
    precision    recall    f1-score   support
          0         0.92      0.99      0.96      858
```

Figure 5.15: Résultat de la méthode Voting Classifier

Pour notre étude, nous avons utilisé le vote de classificateurs pour combiner les prédictions de nos trois modèles les plus performants : Random Forest, XGBoost et Decision Tree. Le vote de classificateurs a été utilisé dans l'espoir d'améliorer encore davantage les performances en agrégeant les prédictions de ces modèles. Cependant, la précision du Voting Classifier s'est avérée légèrement inférieure à celle du Random Forest, avec 0,923 contre 0,927.

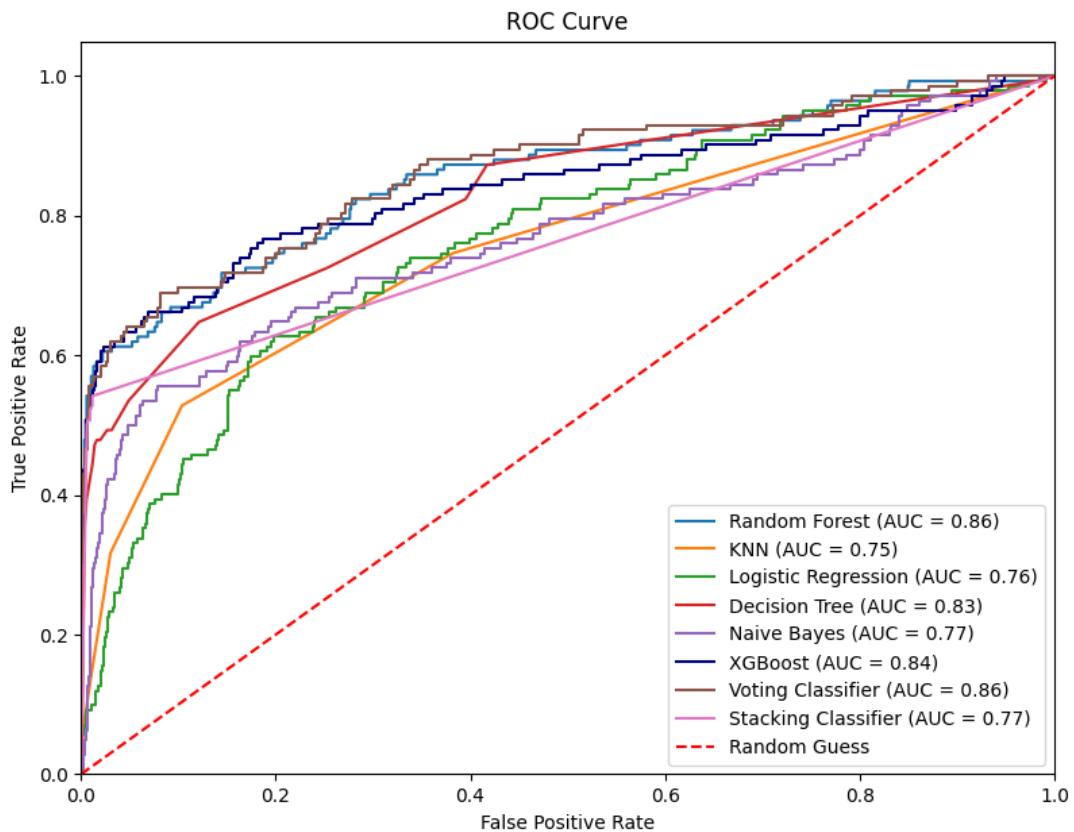
- Stacking Classifier

```
Stacking Classifier Accuracy: 0.93
Classification Report for Voting Classifier:
    precision    recall    f1-score   support
          0         0.92      0.99      0.96      858
```

Figure 5.16: Résultat de la méthode Stacking Classifier

En revanche, nous avons utilisé l'approche du Stacking en utilisant les prédictions de nos trois modèles les plus performants : Random Forest, XGBoost et Decision Tree. Cette fois-ci, les résultats ont dépassé nos attentes, avec une précision atteignant 0,93, dépassant ainsi celle du Voting Classifier et même celle du Random Forest, qui affichait une précision de 0,927.

La figure 5.15 montre un diagramme ROC/AUC de tous les modèles.



**Figure 5.17 :** Diagramme ROC/AUC de tous les modèles

Dans notre situation, le meilleur algorithme est celui avec le F1-score le plus élevé. Comme nous l'avons mentionné précédemment, nous ne sélectionnerons pas la mesure "Accuracy" pour l'évaluation en raison de notre jeu de données déséquilibré. Après avoir examiné plusieurs mesures et scores, nous avons conclu que l'algorithme de Stacking Classifier est le meilleur choix. Il atteint non seulement le plus haut F1-score de 0.96, mais démontre également la meilleure précision avec une valeur de 0.93.

Par conséquent, nous avons décidé de le déployer dans notre application pour prédire si un client résiliera son abonnement ou non.

## 5.5 Ajustement du modèle

L'ajustement du modèle est essentiel pour identifier la cause de la faible précision du modèle. En évaluant les erreurs de prédition sur les données de test, nous pouvons déterminer si un modèle prédictif présente un 'underfitting' ou un 'overfitting'. Dans notre projet, pour évaluer l'ajustement de notre modèle, nous allons comparer les performances sur les données d'entraînement et les données de test en utilisant leurs scores.

```
#tester si on a :underfitting /overfitting /bestfit
print("Le score de train est: ", round(stacking_classifier.score(x_train,y_train),2))
print("Le score de test est: ", round(stacking_classifier.score(x_test,y_test),2))
✓ 0.9s
Le score de train est:  1.0
Le score de test est:  0.92
```

**Figure 5.18 :** Comparaison entre le score du partie Train et Test

Si le score d'entraînement est presque égal au score de test (1.0 / 0.92), nous pouvons conclure que le modèle présente le meilleur ajustement, également connu sous le nom de "best-fitting".

## 5.6 Conclusion

Au cours de ce chapitre, nous avons évalué nos divers modèles en utilisant des métriques de performance. Ensuite, nous avons effectué une comparaison entre ces modèles afin de déterminer le meilleur d'entre eux. Ce modèle choisi sera ensuite déployé dans notre application.

# Chapitre 6 : Déploiement

## Plan

1.	Introduction . . . . .	90
2.	Déploiement. . . . .	90
3.	Extraction du modèle . . . . .	90
4.	Développement de l'interface Web. . . . .	91
5.	Construction de tableau de bord. . . . .	97
6.	Diagramme de Gantt. . . . .	99
7.	Conclusion . . . . .	100

## 6.1 Introduction

Dans cette étape, nous allons déployer le modèle choisi en créant une application web simple. Par la suite, nous allons présenter les tableaux de bord réalisés par l'outil Power BI Desktop.

## 6.2 Déploiement

Dans le chapitre précédent, nous avons sélectionné le modèle "XGBoost" en raison de ses performances supérieures par rapport aux autres modèles. Maintenant, nous allons passer à l'étape de déploiement, où nous intégrerons ce modèle dans une interface afin de démontrer son utilité.

La figure 6.1 représente l'architecture de déploiement de notre modèle ML.

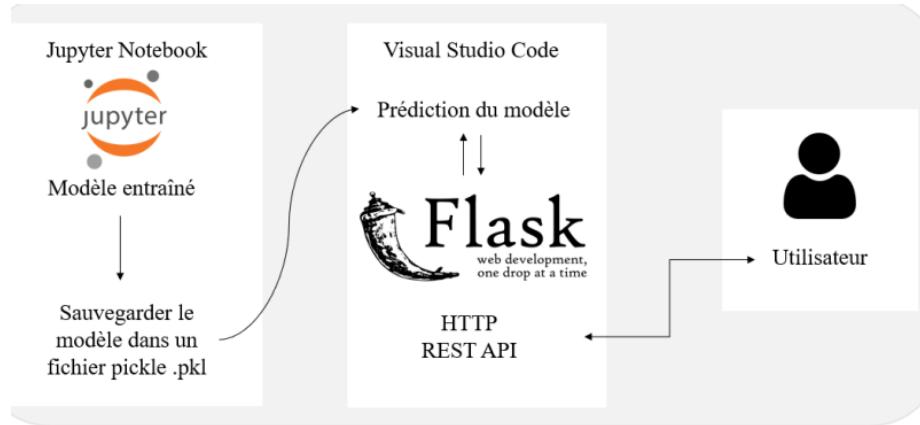


Figure 6.1 : Architecture de déploiement du modèle

## 6.3 Extraction du modèle

En tant que data scientist, vous manipulez souvent des datasets représentés sous forme de dictionnaires, de Data Frames ou d'autres types de données. Vous souhaiterez peut-être les enregistrer dans un fichier pour une utilisation ultérieure ou pour les partager avec d'autres.

Pour cela, le module "pickle" de Python est souvent utilisé. Pour notre projet, nous utiliserons ce module pour récupérer notre modèle afin de pouvoir l'utiliser dans notre interface web. La figure ci-dessous illustre la méthode utilisée pour enregistrer le modèle.

```
#Enregistrer le modèle dans un fichier
import pickle
with open ("final_modelxgb.pkl","wb") as fichier:
    pickle.dump(xgb_model,fichier)
```

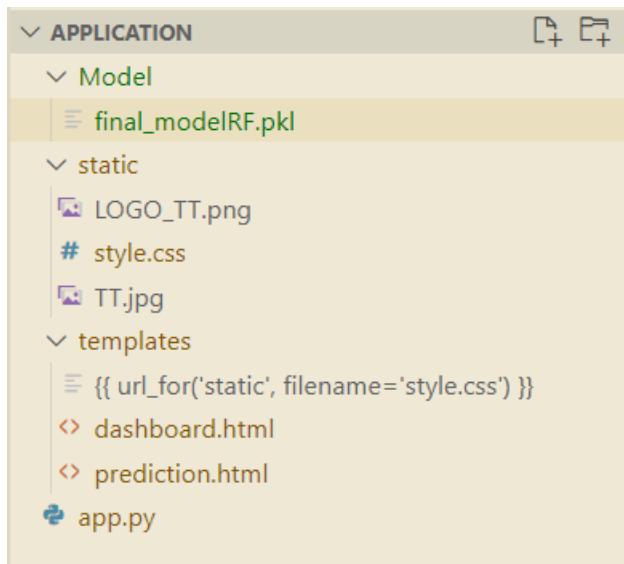
**Figure 6.2 :** Méthode d'enregistrement le modèle

## 6.4 Développement de l'interface Web

Lors de cette étape, nous utiliserons l'environnement de développement Visual Studio Code pour créer la partie front-end en utilisant les langages HTML et CSS, et pour le back-end, nous utiliserons Flask pour développer une API.

- **Structure de l'application:**

Nous avons divisé notre application en deux dossiers pour faciliter sa structure.



**Figure 6.3 :** La structure de la partie développement

Le répertoire "Templates" rassemble l'ensemble des fichiers de la partie front-end, notamment les fichiers HTML, CSS, JavaScript, ainsi que les images utilisées. Le répertoire "model" contient le fichier du modèle utilisé par l'application. Enfin, le fichier "app.py" correspond à la partie back-end de l'application.

- Code de chargement modèle Flask :

```
import pickle
model =pickle.load(open("Model/final_modelxgb.pkl", "rb"))
```

**Figure 6.4** : Chargement du modèle Flask

Le fichier "final\_modelxgb.pkl", contenant notre modèle, a été chargé en utilisant la bibliothèque "Pickle".

- Code de fonction de prédiction :

```
app.py > predict
1  from flask import Flask, render_template, request, redirect, url_for
2  import numpy as np
3  import logging
4  import pickle
5
6  model =pickle.load(open("Model/stacked_classifier.pkl", "rb"))
7  app = Flask(__name__, static_folder='static', static_url_path='/static')
8  logging.basicConfig(level=logging.INFO)
9
10 @app.route('/')
11 def hello_world():
12     |     return render_template('prediction.html')
13
14 @app.route('/predict', methods=['POST'])
15 def predict():
16     |     field_data_types = {
17     |         'nb_jours_abonne': int,
18     |         'genre': str,
19     |         'age': int,
20     |         'marie': str,
21     |         'nb_appel_jour': int,
22     |         'duree_appel_jour': float,
23     |         'cout_appel_jour': float,
24     |         'nb_appel_soiree': int,
25     |         'duree_appel_soiree': float,
26     |         'cout_appel_soiree': float,
27     |         'nb_appel_nuit': int,
28     |         'duree_appel_nuit': float,
29     |         'cout_appel_nuit': float,
30     |         'nb_appel_inter': int,
31     |         'duree_appel_inter': float,
32     |         'cout_appel_inter': float,
33     |         'active_msg_vocaux': str,
34     |         'nb_msg_vocaux': int,
35     |         'nb_reclamation': int,
36     |         'offer_type': str
37     |     }
```

```

39     init_features = {}
40     for field, data_type in field_data_types.items():
41         value = request.form.get(field)
42
43         app.logger.debug("Field: %s, Value: %s", field, value)
44         if value is None:
45             return f"Entrée invalide pour '{field}'"
46         if data_type == int:
47             if value.isdigit():
48                 init_features[field] = int(value)
49             else:
50                 return f"Entrée invalide pour '{field}'"
51         elif data_type == float:
52             try:
53                 init_features[field] = float(value)
54             except ValueError:
55                 return f"Entrée invalide pour '{field}'"
56         else:
57             init_features[field] = value
58
59
60     if init_features['active_msg_vocaux'] == "oui":
61         init_features['active_msg_vocaux'] = 1
62     else:
63         init_features['active_msg_vocaux'] = 0
64
65     if init_features['marie'] == "oui":
66         init_features['marie'] = 1
67     else:
68         init_features['marie'] = 0
69
70
71     if init_features['genre'] == "homme":
72         init_features['genre'] = 0
73     else:
74         init_features['genre'] = 1
75
76     option_map = ...
77     offer_type = init_features['offer_type']
78     if offer_type in option_map:
79         init_features['offer_type'] = float(option_map[offer_type])
80     else:
81         raise ValueError("Type d'offre non trouvé dans la carte d'options.")
82
83     # Faire une prédiction
84
85     # Convertir en tableau numpy et remodeler
86     final_features = np.array(list(init_features.values())).reshape(1, -1)
87
88     prediction = model.predict(final_features)
89
90     # Déterminez le message de prédiction
91     if prediction == 1:
92         message = " Le client résilierai son abonnement"
93     else:
94         message = " Le client ne résilierai pas son abonnement"
95
96     return render_template('prediction.html',
97                           prediction_text='Résultat de prédiction: {}'.format(message))
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138

```

Figure 6.5 : Code source de fonction "predict"

La fonction predict () est utilisée pour effectuer des prédictions basées sur un modèle. Lorsqu'une requête est reçue depuis le frontend, les données sont extraites du formulaire et utilisées comme entrées pour effectuer une prédiction à l'aide d'un modèle préalablement entraîné. Les données d'entrée comprennent des informations telles que le nombre de jours d'abonnement, la durée des appels, le nombre d'appels, etc. Chaque valeur d'entrée est extraite et utilisée comme entrée pour la fonction model.predict(). Les résultats de prédiction sont renvoyés au frontend et affichés pour indiquer si le client résiliera ou non son abonnement.

La fonction est ajoutée en tant que route dans l'application Flask avec un chemin spécifique ("/predict").

- **Exécution de l'application:**

Finalement, nous avons démarré le serveur Flask en utilisant le script fourni dans la figure ci-dessous.

```
if __name__ == '__main__':
    app.run(debug=True)
```

**Figure 6.6 :** Script de l'exécution de l'application

#### 6.4.1 Les interfaces de l'application

Dans cette partie, nous présentons l'interface graphique de notre application. Sur le port 5000, qui est le port par défaut de Flask, nous allons lancer notre application finale.

La figure 6.7 présente l'interface de prédiction.

The screenshot shows a web form titled "PRÉDICTION DU CHURN DES CLIENTS TUNISIE TELECOM". The form contains 20 input fields arranged in a grid. The fields include: Nombre de jours d'abonnement (text input), Genre (dropdown: Homme, Femme), Âge (text input), État civil (dropdown: Marié(e), Veuf(e), Divorcé(e), Séparé(e), Célibataire), Nombre d'appels par jour (text input), Durée de l'appel par jour en minutes (text input), Coût des appels par jour (text input), Nombre d'appels par soirée (text input), Durée de l'appel par soirée en minutes (text input), Coût des appels par soirée (text input), Nombre d'appels par nuit (text input), Durée de l'appel par nuit en minutes (text input), Coût des appels par nuit (text input), Nombre d'appels international (text input), Durée de l'appel international en minutes (text input), Coût des appels international (text input), Activer option message vocaux (dropdown: Oui, Non), Nombre des messages vocaux (text input), Nombre des réclamations (text input), Le nom d'offre tarifaire suivie (dropdown: Hayya, Free, Djerba, Kairouan, Sousse), and a "Prédire" button.

**Figure 6.7 :** L'interface de prédiction

L'interface de notre application web comprend un formulaire avec 20 champs. Une fois que l'utilisateur a rempli ces champs, nous utilisons les données fournies pour prédire la situation de churn. De plus, nous allons créer des tableaux de bord afin de faciliter la visualisation et d'améliorer la compréhension des données.

Le résultat de la prédiction est illustré dans les figures 6.8 et 6.9.

**PRÉDICTION DU CHURN DES CLIENTS TUNISIE TELECOM**

Nombre de jours d'abonnement:

Genre:  Âge:  État civil:

Nombre d'appels par jour:  Durée de l'appel par jour en minutes:  Coût des appels par jour:

Nombre d'appels par soirée:  Durée de l'appel par soirée en minutes:  Coût des appels par soirée:

Nombre d'appels par nuit:  Durée de l'appel par nuit en minutes:  Coût des appels par nuit:

Nombre d'appels international:  Durée de l'appel international en minutes:  Coût des appels international:

Activer option message vocaux:  Nombre des messages vocaux:

Nombre des réclamations:  Le nom d'offre tarifaire suivie:

**Résultat de prédiction: Le client résiliera son abonnement**

**Prédire**      **Visualisez les tableaux de bord**

Figure 6.8 : Résultat de prédiction positive

**PRÉDICTION DU CHURN DES CLIENTS TUNISIE TELECOM**

Nombre de jours d'abonnement:

Genre:  Âge:  État civil:

Nombre d'appels par jour:  Durée de l'appel par jour en minutes:  Coût des appels par jour:

Nombre d'appels par soirée:  Durée de l'appel par soirée en minutes:  Coût des appels par soirée:

Nombre d'appels par nuit:  Durée de l'appel par nuit en minutes:  Coût des appels par nuit:

Nombre d'appels international:  Durée de l'appel international en minutes:  Coût des appels international:

Activer option message vocaux:  Nombre des messages vocaux:

Nombre des réclamations:  Le nom d'offre tarifaire suivie:

**Résultat de prédiction: Le client ne résiliera pas son abonnement**

**Prédire**      **Visualisez les tableaux de bord**

Figure 6.9 : Résultat de prédiction négative

## 6.5 Construction de tableau de bord

Nous passons maintenant à l'étape de création du tableau de bord dédié au taux d'attrition des clients au sein de l'entreprise Tunisie Télécom. Ce tableau de bord nous permettra de visualiser les données et d'extraire les informations importantes concernant le volume des appels et la consommation des abonnés en termes de coûts.

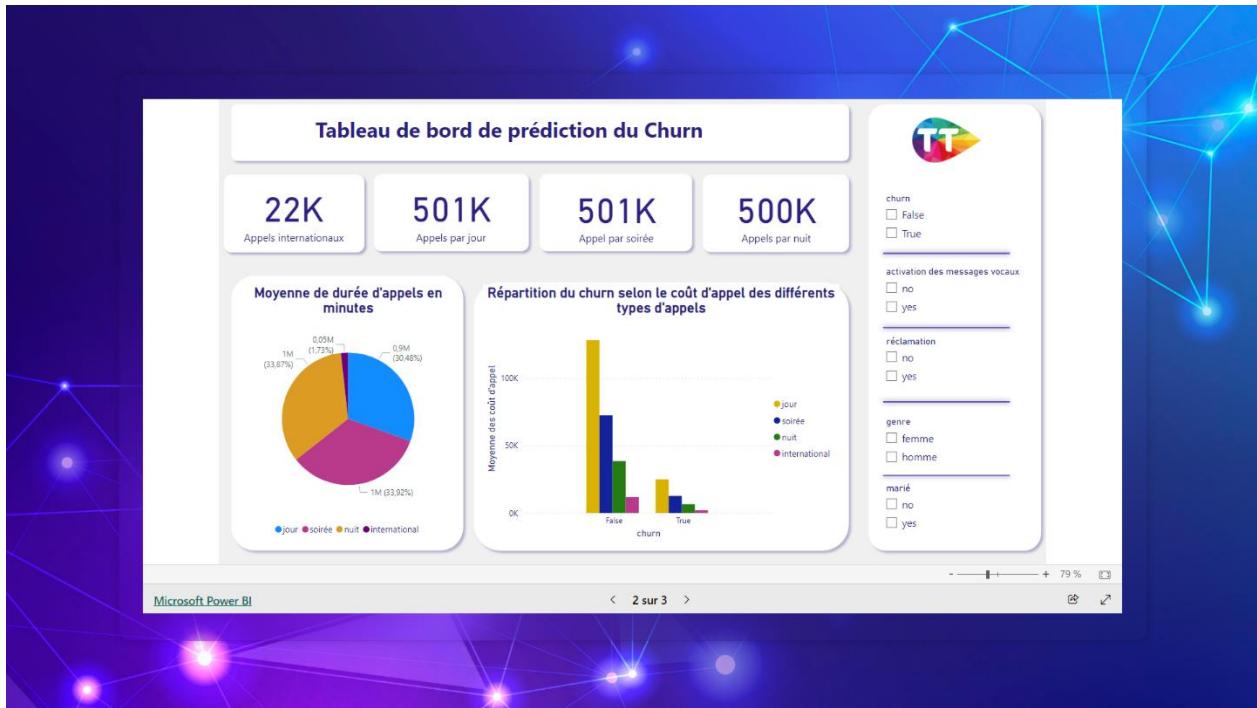
La figure 6.10 représente le premier tableau de bord.



Figure 6.10 : Interface de tableau de bord 1

Le tableau de bord fournit une visualisation du pourcentage de résiliation des abonnements des clients, ainsi que leur choix d'activation des messages vocaux en fonction du churn. Il affiche également le nombre total de réclamations. Nous représentons en outre le nombre de jours d'abonnement répartis en cinq catégories distinctes. Nous avons un total de 5 000 clients dans notre base de données, répartis en deux catégories : 4 292 clients qui ne résilient pas leur abonnement représentant 85,84% des données et 708 qui résilient représentant 14,16% des données.

La figure 6.11 présente le deuxième tableau de bord.



**Figure 6.11 :** Interface de tableau de bord 2

Ce tableau de bord illustre la répartition du churn en fonction du coût moyen des différents types d'appels (jour, soirée, nuit, international). Une observation essentielle réside dans la constatation que le coût des appels durant la journée dépasse celui des autres périodes. De plus, il présente le nombre total des appels par jour, soirée, nuit et internationaux. De plus, il fournit la durée moyenne des appels de tous les types d'appels en minutes. Il est possible de filtrer les données en utilisant les variables "churn", "active\_msg\_vocaux", "réclamation", "genre" et "marié".

La figure 6.12 présente le troisième tableau de bord.



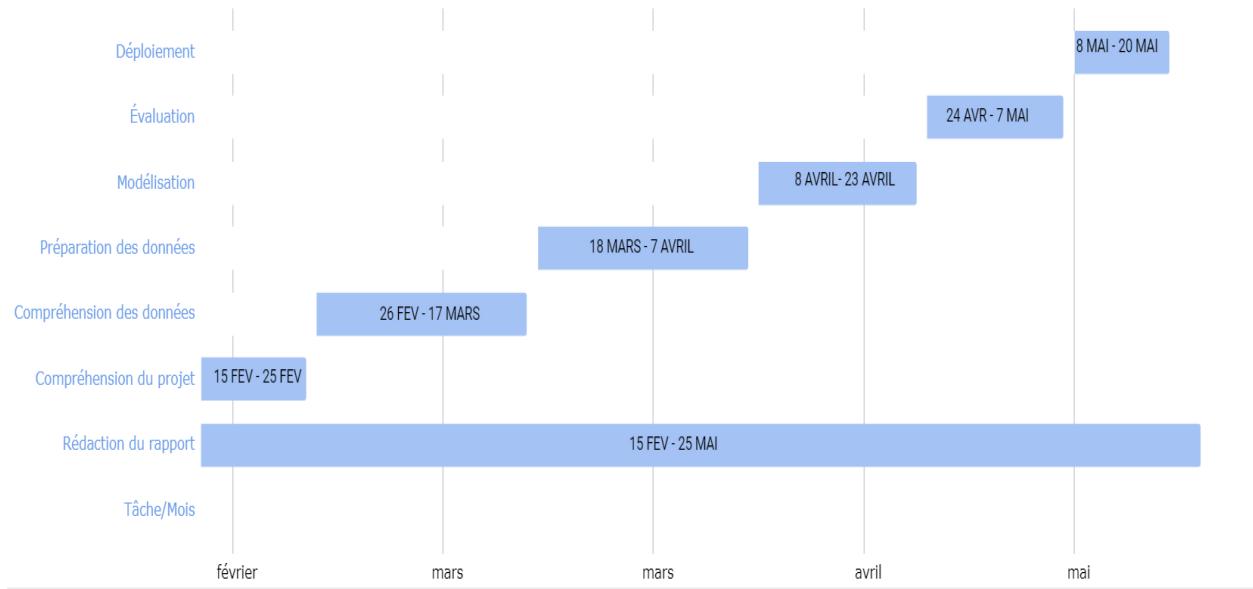
**Figure 6.12 :** Interface de tableau de bord 3

Ce tableau de bord illustre une représentation visuelle des données démographiques des clients de Tunisie Télécom, en relation avec le churn. En fait, on trouve le nombre de chaque genres (2593 femmes et 2407 hommes), la répartition de l'état civile (48.96% marié et 51.04% non marié) et l'âge moyenne des clients, ainsi que la répartition des offres tarifaires des clients. De plus, le tableau de bord offre une fonctionnalité de filtrage des données en utilisant la variable "churn"

## 6.6 Diagramme de Gantt

C'est un outil permettant de modéliser la planification de tâches nécessaires à la réalisation d'un projet.

La figure 6.13 présente la planification de tâches pour la réalisation du notre projet.



**Figure 6.13 :** Diagramme de Gantt

## 6.7 Conclusion

Dans ce dernier chapitre, nous avons présenté en détail notre architecture de déploiement. Puis nous avons créé une application pour mettre en évidence les fonctionnalités de notre modèle de prédiction, ainsi que quelques tableaux de bord et le diagramme de Gantt.

# Conclusion générale

Ce rapport représente le travail réalisé lors de notre stage de fin d'études pendant quatre mois chez l'entreprise Tunisie Télécom dans le cadre du diplôme en Licence en Big Data et Analyse de données.

Au cours de cette expérience, nous avons appliquée nos connaissances théoriques acquises lors de notre formation à l'ISAMM et les avons enrichies grâce à notre participation à la réalisation d'un projet qui consiste à développer un modèle de prédiction du churn des clients de Tunisie Télécom.

Actuellement, Tunisie Télécom compte plus de trois millions d'abonnés mobiles, et la perte de clients représente une diminution de revenus. Afin de prévoir cette situation, nous avons utilisé des techniques d'apprentissage automatique pour réaliser une modélisation prédictive. Cette méthode permet de déterminer avec une certaine probabilité si un client est susceptible de résilier son abonnement ou de rester actif.

En premier lieu, Nous avons commencé par comprendre le contexte général de notre projet, Par la suite nous avons suivi les étapes CRISP-DM pour construire ce rapport, notamment le nettoyage des données, la modélisation et l'évaluation. Enfin, nous avons développé une interface web qui aide à prédire la situation de churn. Et enfin, nous avons créé un tableau de bord pour faciliter l'accès aux informations pertinentes.

Lors de notre stage, nous a non seulement découvrir le monde professionnel mais aussi nous a permis d'explorer le domaine de l'intelligence artificielle. Nous avons implémenté en pratique plusieurs algorithmes d'apprentissage automatique supervisé ce qui nous a permis de développer nos connaissances en ce domaine.

Finalement, notre travail ne se termine pas à ce niveau. Des fonctionnalités supplémentaires peuvent être ajoutées au projet afin d'améliorer sa pertinence. Par exemple :

- Utilisation du traitement du langage naturel pour analyser les mots et les phrases répétés dans les appels au service client et d'organiser les réclamations des clients en différents thèmes tels que les problèmes de facturation, les problèmes liés aux offres, les problèmes de couverture réseau, etc...
- Utilisation du traitement du langage naturel pour analyser les sentiments à partir des enquêtes de satisfaction, des commentaires en ligne ou des réseaux sociaux et évaluer leur satisfaction.

# Bibliographie

- [1] <https://www.leaders.com.tn/article/2313-le-groupe-tunisie-telecom>, consulté le 15/02/2024.
- [2] <https://www.tunisetelecom.tn/Fr/Particulier/A-Propos/Organisation>, consulté le 15/02/2024.
- [3] <http://documentation.sas.com/doc/en/emcs/14.3/n0pejm83csbj4n1xueveo2uoujy.html>, consulté le 16/02/2024.
- [4] <https://subscription.packtpub.com/book/data/9781789955248/2/ch02lvl1sec06/semma>, consulté le 16/02/2024.
- [5] <https://www.datascience-pm.com/tdsp/>, consulté le 16/02/2024.
- [6] <https://www.datascience-pm.com/crisp-dm-2/>, consulté le 17/02/2024.
- [7] [https://www.researchgate.net/figure/CRISP-DM-data-mining-framework\\_fig1\\_341627969](https://www.researchgate.net/figure/CRISP-DM-data-mining-framework_fig1_341627969), consulté le 17/02/2024.
- [8] [https://fr.wikipedia.org/wiki/Anaconda\\_\(distribution\\_Python\)](https://fr.wikipedia.org/wiki/Anaconda_(distribution_Python)), consulté le 19/02/2024.
- [9] <https://www.lebigdata.fr/jupyter-notebook>, consulté le 19/02/2024.
- [10] <https://www.lebigdata.fr/power-bi-microsoft>, consulté le 19/02/2024.
- [11] [https://fr.wikipedia.org/wiki/Visual\\_Studio\\_Code](https://fr.wikipedia.org/wiki/Visual_Studio_Code), consulté le 19/02/2024.
- [12]  
[https://data.sigea.educagri.fr/download/sigea/supports/PostGIS/distance/perfectionnement/M02\\_administration/co/30\\_pgAdmin.html](https://data.sigea.educagri.fr/download/sigea/supports/PostGIS/distance/perfectionnement/M02_administration/co/30_pgAdmin.html), consulté le 19/02/2024.
- [13] <https://fr.wikipedia.org/wiki/Pandas>, consulté le 19/02/2024.
- [14] <https://fr.wikipedia.org/wiki/NumPy>, consulté le 19/02/2024.
- [15] <https://datascientest.com/seaborn-tout-savoir>, consulté le 19/02/2024.
- [16] <https://fr.wikipedia.org/wiki/Matplotlib>, consulté le 19/02/2024.
- [17] <https://www.data-transitionnumerique.com/scikit-learn-python/>, consulté le 19/02/2024.
- [18] [https://fr.wikipedia.org/wiki/Flask\\_\(framework\)](https://fr.wikipedia.org/wiki/Flask_(framework)), consulté le 19/02/2024.

- [19] <https://docs.python.org/fr/3/library/pickle.html>, consulté le 19/02/2024.
- [20] <https://www.logilab.fr/blogentry/13252264>, consulté le 19/02/2024.
- [21] <https://blueorange.digital/essential-python-libraries-for-machine-learning-projects/>, consulté le 19/02/2024.
- [22] <https://www.techno-science.net/glossaire-definition/Python-langage.html>, consulté le 21/02/2024.
- [23] <https://developer.mozilla.org/fr/docs/Web/HTML>, consulté le 21/02/2024.
- [24] <http://users.polytech.unice.fr/hermenie/adw/TPs/html-css/>, consulté le 21/02/2024.
- [25] <https://fr.wikipedia.org/wiki/JavaScript>, consulté le 21/02/2024.
- [26] <https://www.codeur.com/blog/front-end-framework/>, consulté le 21/02/2024.
- [27] <https://www.definitions-marketing.com/definition/churn-rate-2/>, consulté le 01/03/2024.
- [28] <https://www.lebigdata.fr/machine-learning-et-big-data>, consulté le 01/03/2024.
- [29] [https://fr.wikipedia.org/wiki/Apprentissage\\_non\\_supervis%C3%A9](https://fr.wikipedia.org/wiki/Apprentissage_non_supervis%C3%A9), consulté le 01/03/2024.
- [30] <https://deeplearning.fr/cours-theoriques-deep-learning/comprendre-overfitting-et-underfitting/>, consulté le 01/03/2024.
- [31] <https://www.superannotate.com/blog/overfitting-and-underfitting-in-machine-learning>, consulté le 01/03/2024.
- [32] <https://management-datasceince.org/articles/14442/>, consulté le 12/03/2024.
- [33] <https://dataanalyticspost.com/Lexique/normalisation/>, consulté le 17/04/2023.
- [34] <http://www.sthda.com/french/wiki/matrice-de-correlation-guide-simple-pour-analyser-formater-et-visualiser>, consulté le 17/04/2023.
- [35] <https://guidesurvie.com/techniques-survie/bagging-et-random-forest-en-machine-learning-comment-ca-marche/>, consulté le 19/04/2023.
- [36] <https://larevueia.fr/random-forest/>, consulté le 29/04/2023.
- [37] [https://fr.wikipedia.org/wiki/Arbre\\_de\\_d%C3%A9cision\\_\(apprentissage\)](https://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision_(apprentissage)), consulté le 20/04/2023.
- [38] <https://fr.linedata.com/principaux-algorithmes-de-classification-partie-2>, consulté le 20/04/2023.

- [39] <https://minhdq99hp.github.io/machinelearning/logistic-regression/>, consulté le 20/04/2023.
- [40] <https://datascientest.com/knn>, consulté le 23/04/2023.
- [41] <https://intuitivetutorial.com/2023/04/07/k-nearest-neighbors-algorithm/>, consulté le 23/04/2023.
- [42] <https://www.jedha.co/formation-ia/algorithme-xgboost>, consulté le 23/04/2023.
- [43] [https://www.researchgate.net/figure/Flow-chart-of-XGBoost-framework\\_fig4\\_357640947](https://www.researchgate.net/figure/Flow-chart-of-XGBoost-framework_fig4_357640947), consulté le 23/04/2023.
- [44] <https://mrmint.fr/naive-bayes-classifier>, consulté le 23/04/2023.
- [45] <https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/>, consulté le 23/04/2023.
- [46] <https://datascientest.com/cross-validation>, consulté le 24/04/2023.
- [47] <https://www.lebigdata.fr/confusion-matrix-definition>, consulté le 26/04/2023.
- [48] <https://www.jedha.co/formation-ia/matrice-confusion>, consulté le 26/04/2023.
- [49] <https://kobia.fr/classification-metrics-accuracy/>, consulté le 26/04/2023.
- [50] <https://beranger.medium.com/ml-accuracy-pr%C3%A9cision-f1-score-courbe-roc-que-choisir-5d4940b854d7>, consulté le 27/04/2023.
- [51] <https://datascience.eu/fr/apprentissage-automatique/comprendre-la-courbe-cua-roc/>, consulté le 27/04/2023.
- [52] [https://en.wikipedia.org/wiki/Learning\\_curve](https://en.wikipedia.org/wiki/Learning_curve), consulté le 02/05/2024.
- [53] <https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/>, consulté le 07/05/2024.
- [54] <https://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligence-artificielle/1501883-methode-ensembliste-ensemble-learning/>, consulté le 10/05/2024.