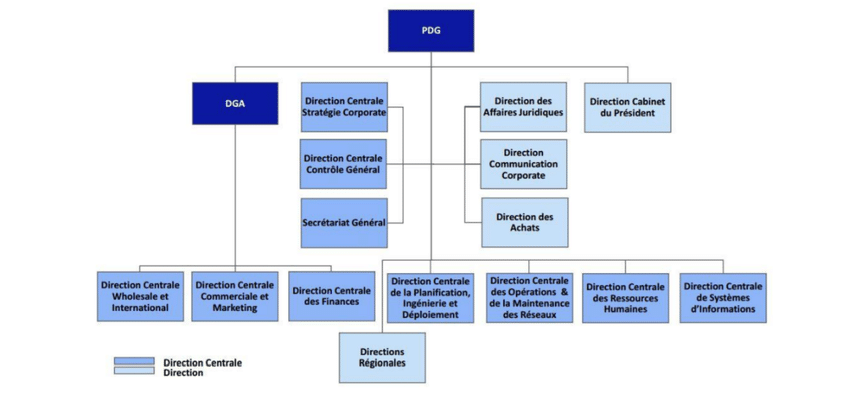
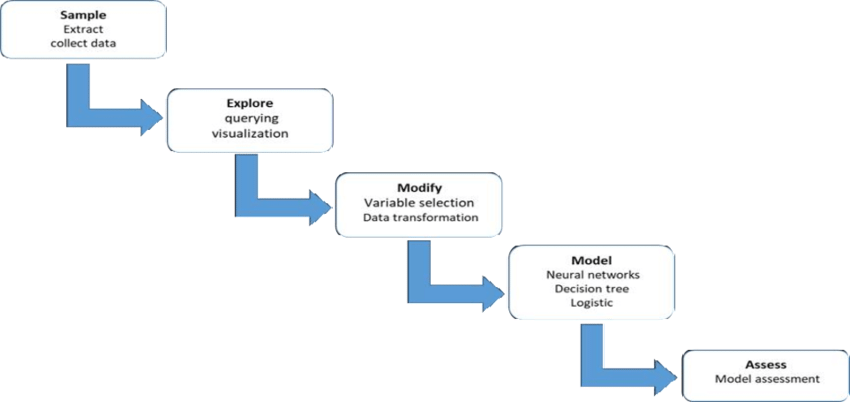
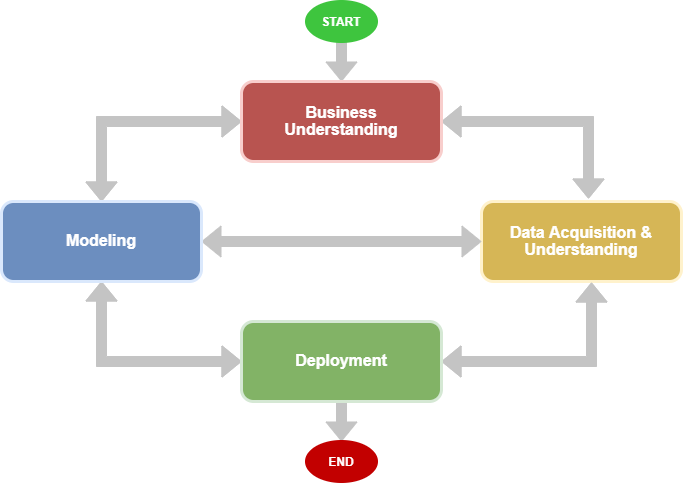
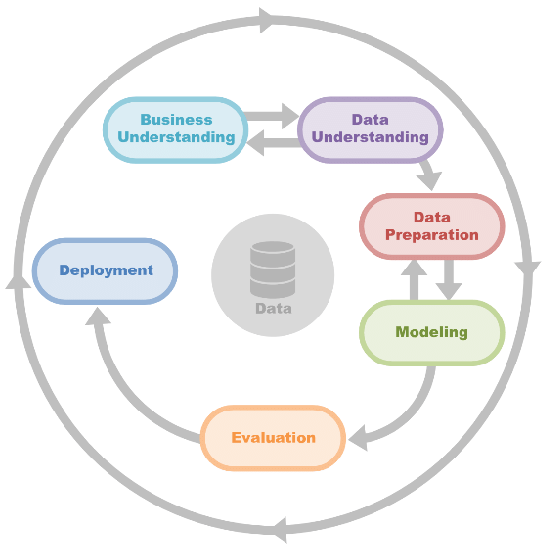
**Ministère de l’Enseignement Supérieur et de la Recherche Scientifique**

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 Benabdallah**  
 Organisme d’accueil : **Tunisie Télécom  
  
  
  
  
  
  
  
 Encadrant Universitaire : Mr Bechikh Chedi   
Encadrant Professionnel : Mr Jemai Hatem   
  
  
 Année Universitaire 2023 / 2024  
  
  
  
Dédicace  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
Remerciements  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
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

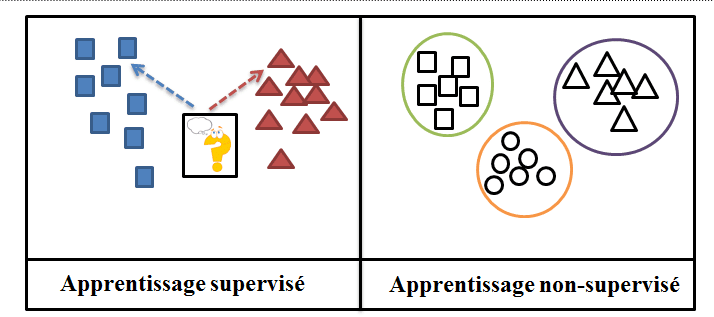
Abstract  
  
  
 The main objective of this internship report is to develop a churn (churn rate) prediction model that will identify the customers closest to cancel their subscriptions. We will use machine learning techniques to predict customer churn and embed the model in a web application.  
 This 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  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
Liste des figures  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
Liste des Tableaux  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
Liste des abréviations  
  
  
  
  
  
  
  
  
  
  
  
  
  
Introduction générale  
 De nos jours, le secteur des télécommunications s’améliore de jour en jour, qui 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 efface 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 trouvent pas sa satisfaction.  
 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 pourquoi elle a décidé de faire une première étape en cherchant une solution au problème de l’attrition 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 proches à arrêter leurs lignes téléphoniques avec Tunisie Télécom on utilise les 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 concentré 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 de projet et compréhension du métier  
  
  
  
  
**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.  
  
   
 **1.2.2 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 ouvert 35% au profit du 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.  
  
  
**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’École Supérieure d’Économie Numérique 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, Tunisie Telecom met en place une démarche qualité visant à mesurer la satisfaction de ses clients et à fournir un diagnostic précis de leurs satisfactions et 7 insatisfactions. 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.  
 *1. Échantillonner les données (Sample) :* Cette étape consiste à choisir les données pertinentes qui sera utilisé pour l’analyse et 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.  
  
   
 **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. 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.  
   
 **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.   
 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 9 Chapitre 1. Étude du projet 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.  
  
  
  
   
  
  
  
  
  
  
  
  
  
 **1.4.8 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.

|  |  |  |  |
| --- | --- | --- | --- |
| 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** 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.

**1.7 Conclusion**  
 Ce premier chapitre a inclus une présentation de l’organisme d’accueil, ensuite la problématique et la solution proposée. Puis la méthodologie de gestion de projet que nous allons adopter tout au long du projet. Et enfin, nous avons décrit l’environnement de travail.

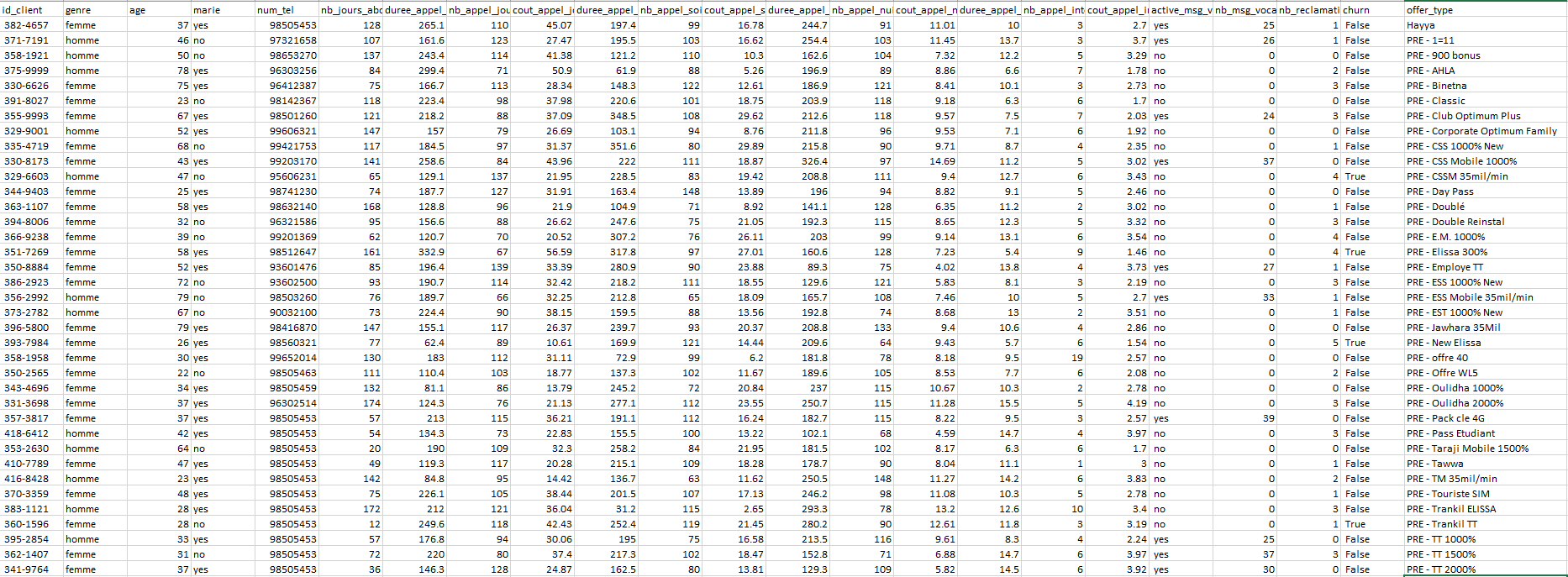
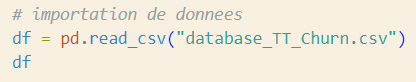
Chapitre 2 : Compréhension du problème métier et des données

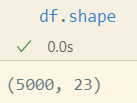
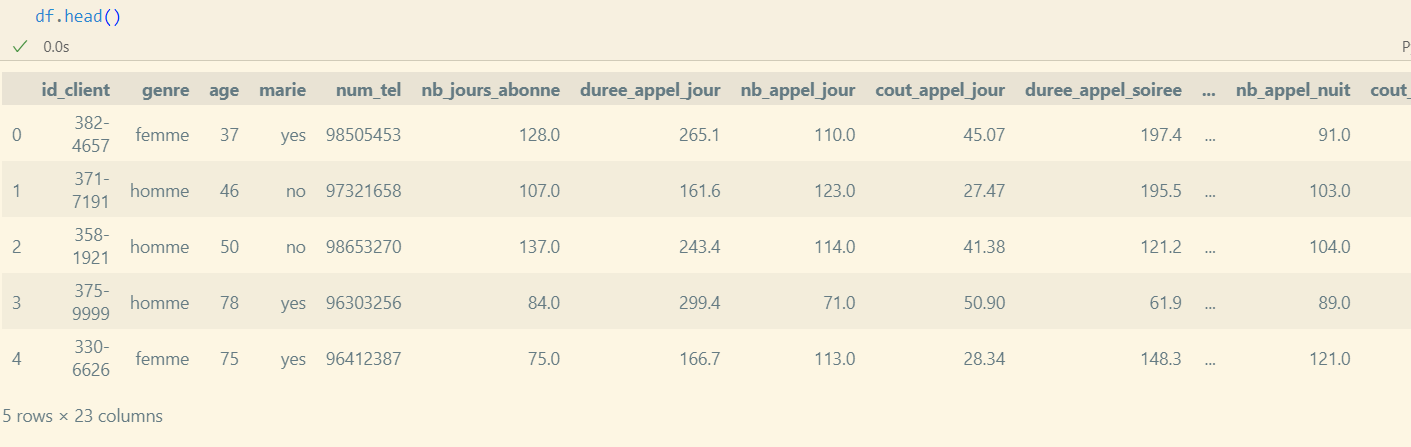
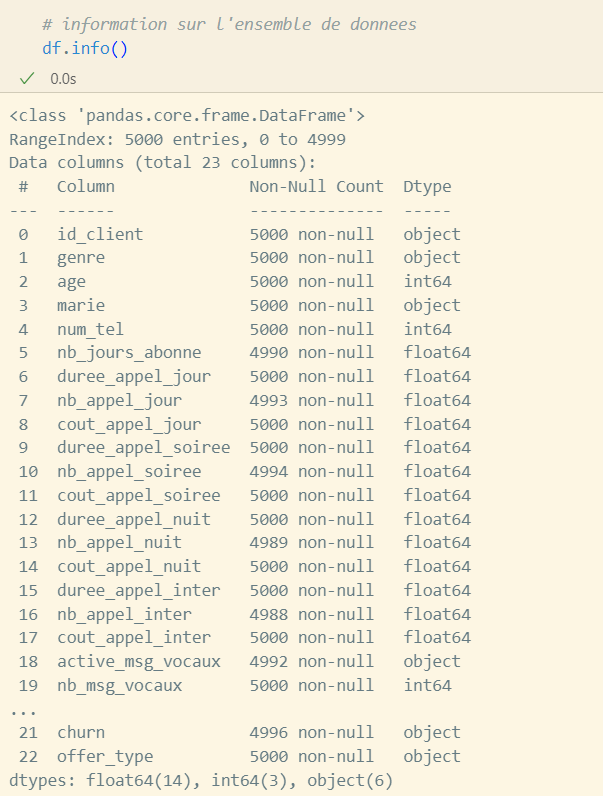
**2.1 Introduction** Au cours de ce chapitre, nous allons présenter les concepts fondamentaux de notre projet, puis analyser en détails les données collectées. Nous mettrons en œuvre des techniques de visualisation pour explorer et interpréter les données, dans le but d’enrichir leur compréhension.  
  
**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, 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.  
  
 Taux de churn = Nombre de clients perdus / 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, également connu sous le nom de Machine Learning en anglais, est un sous-domaine de l’intelligence artificielle qui se concentre sur la création de modèles permettant  
Compréhension du problème métier et des données à une machine d’apprendre à partir de données et de s’améliorer automatiquement en les exploitant.  
  
 Les techniques d’apprentissage automatique peuvent être regroupés en deux catégories :   
• Supervisé   
• Non-supervisé  
 La figure présentée ci-dessous représente la différence entre les 2 catégories  
****

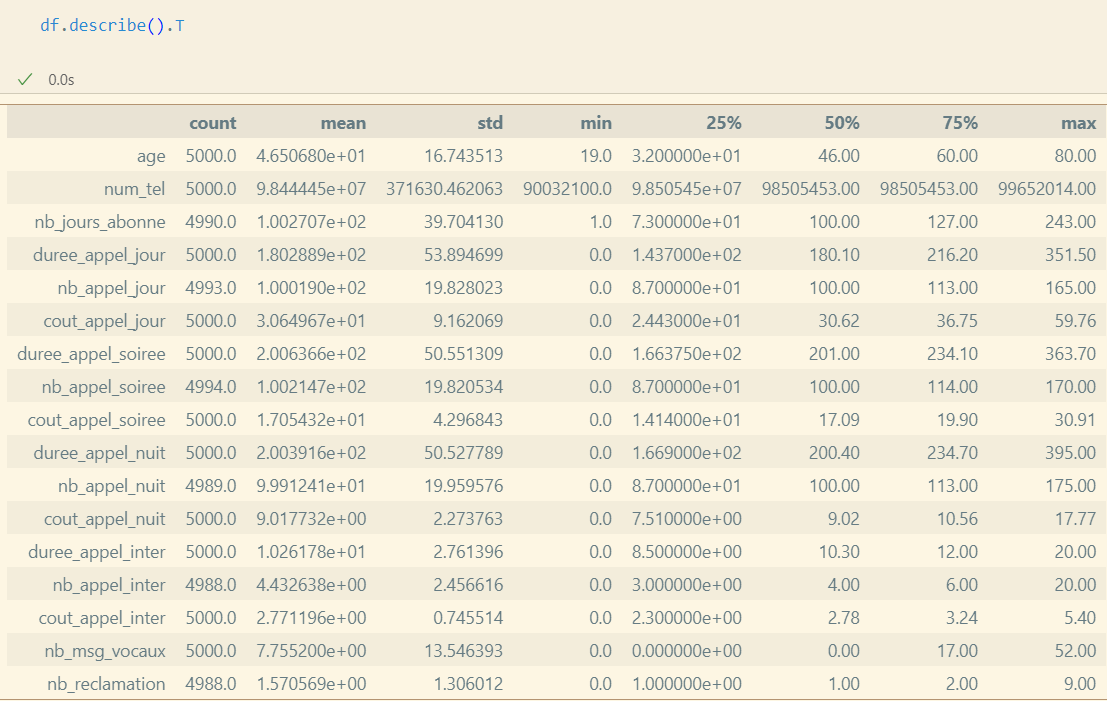
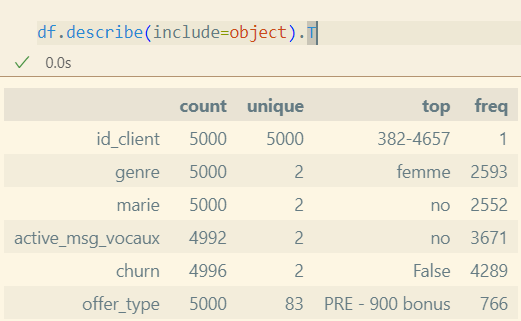
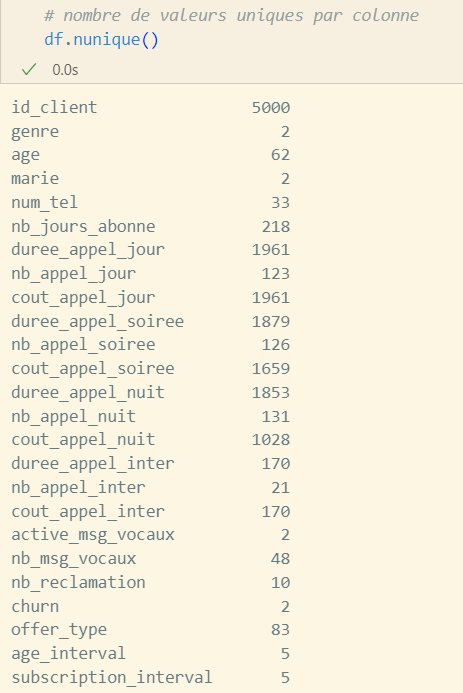
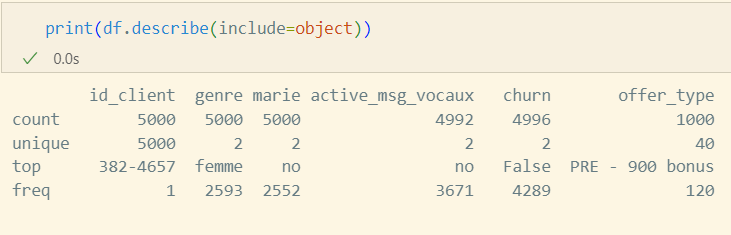
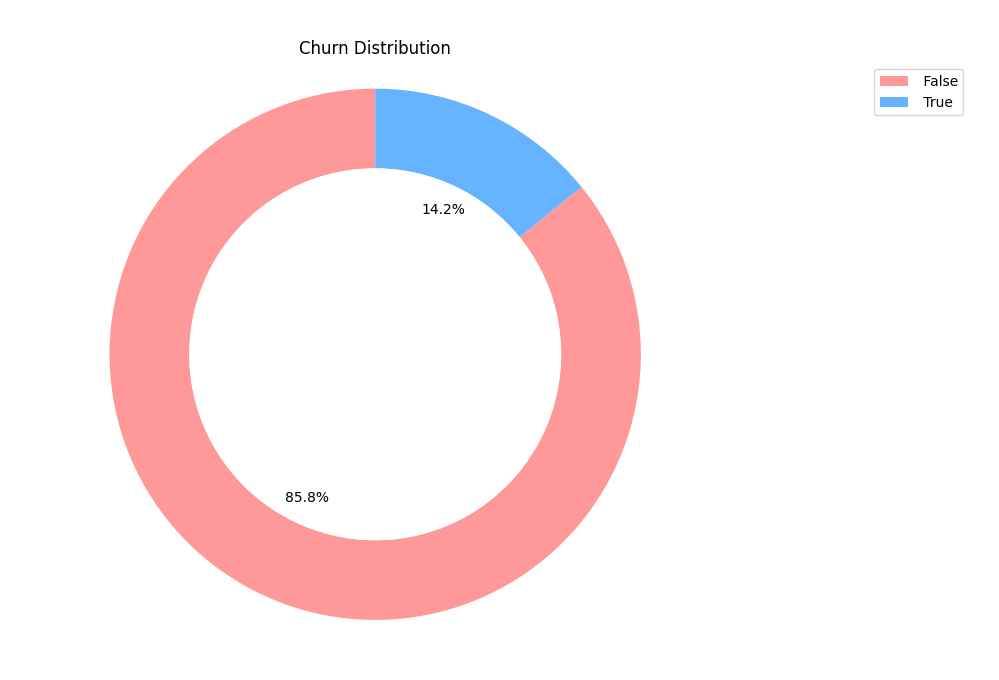
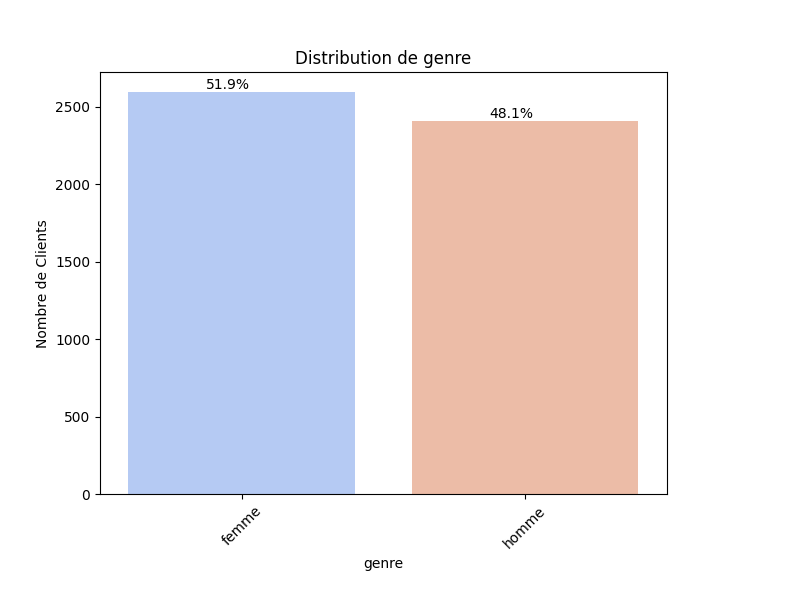
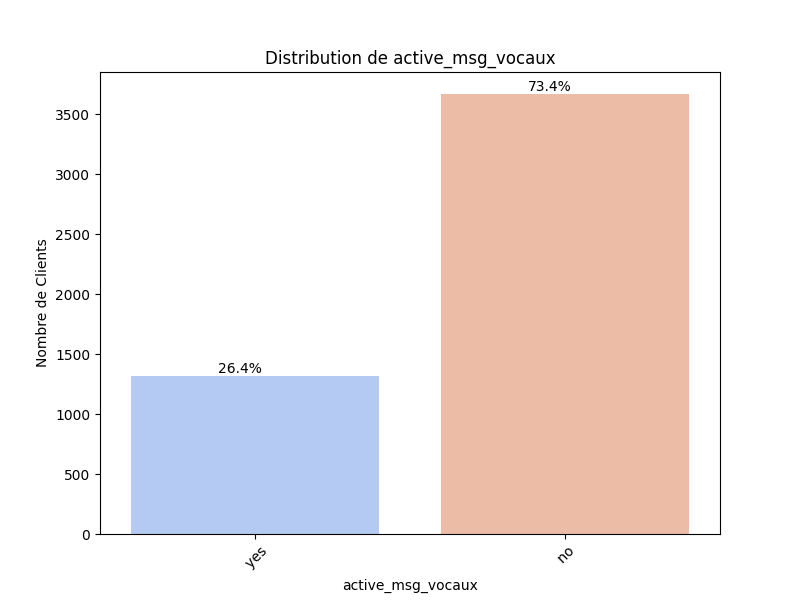
**2.3 Choix de la base de données** Notre dataframe disponible pour cette recherche comprend des renseignements sur l'historique des opérations d'appels nationaux et internationaux, les messages vocaux et les réclamations des clients de Tunisie Télécom, ainsi que leurs données démographiques et leur numéro téléphonique.

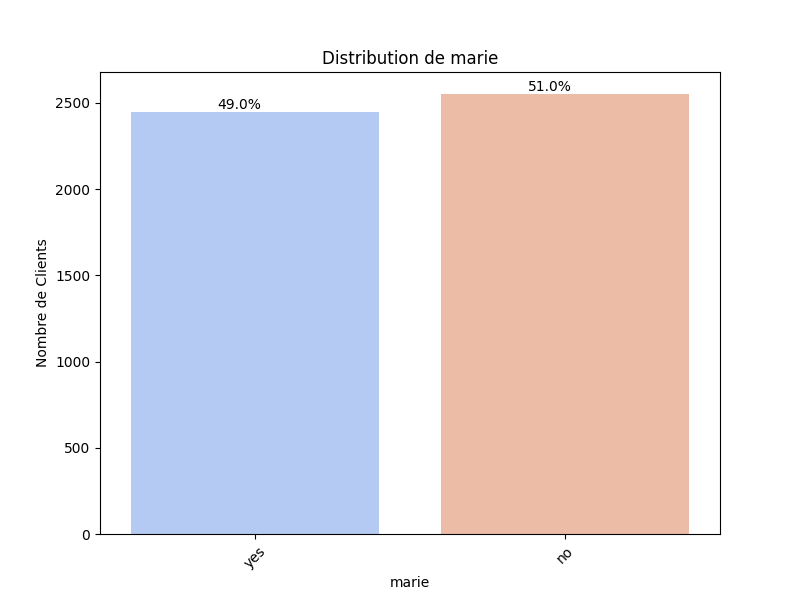
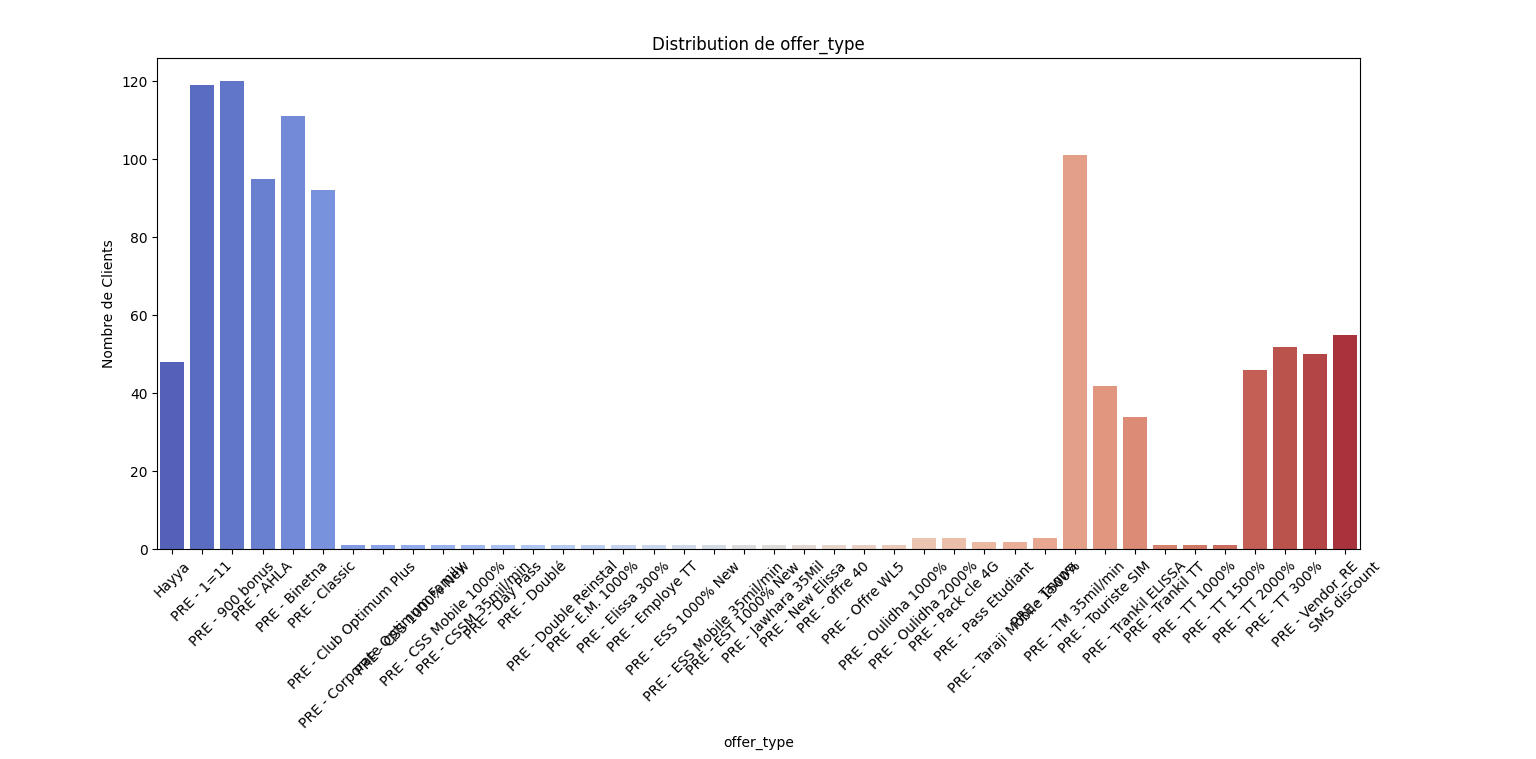
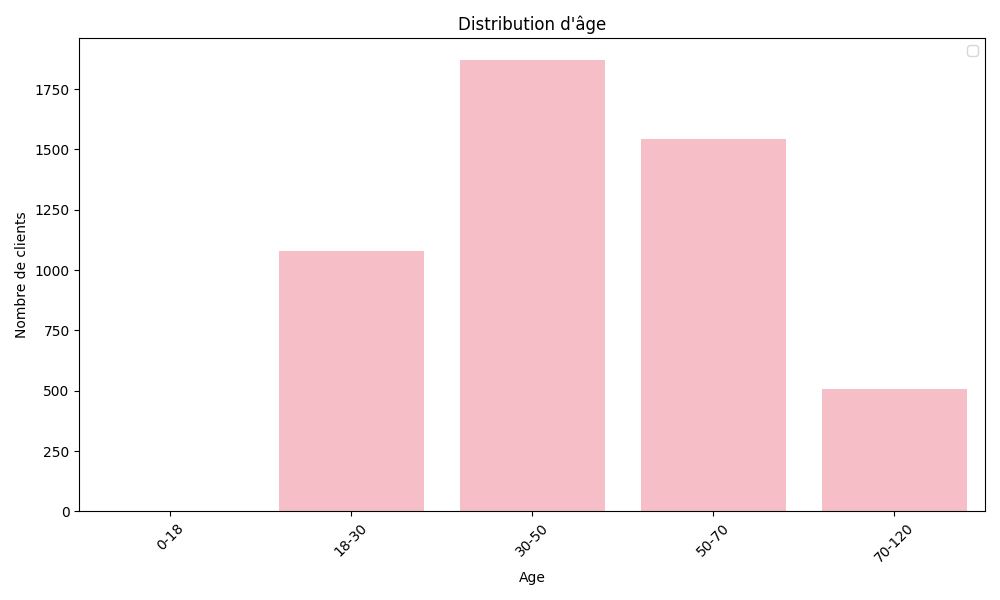
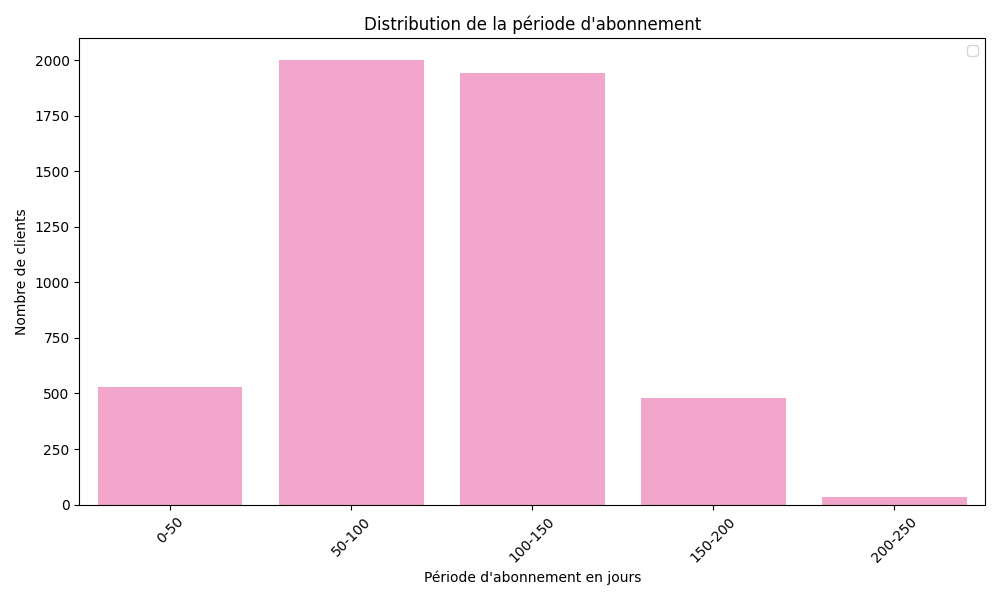
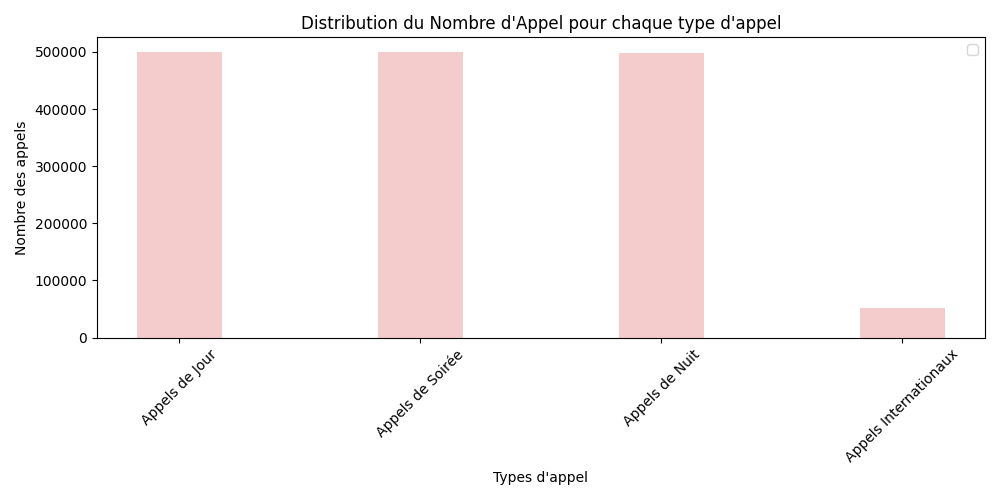
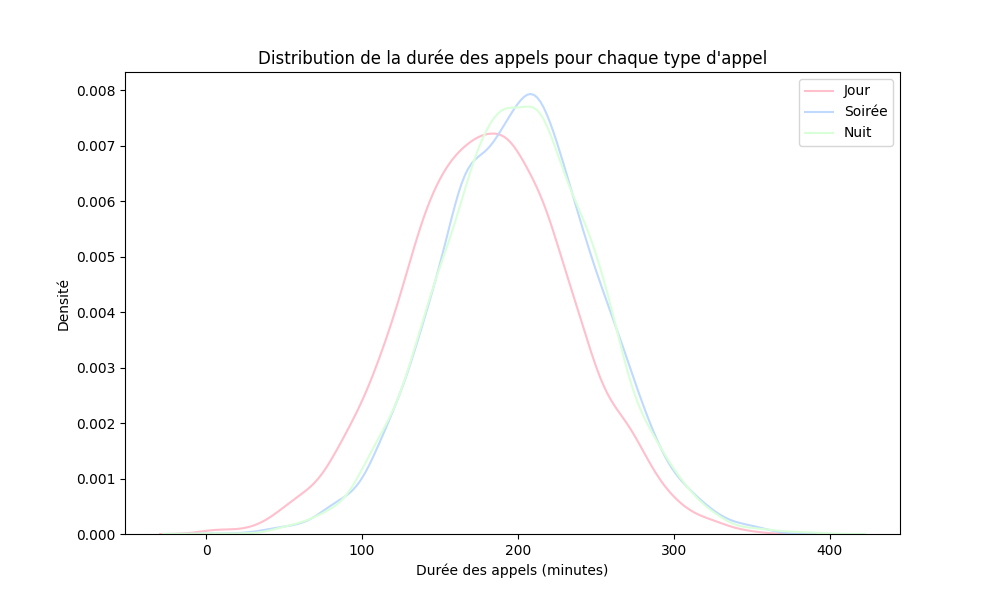
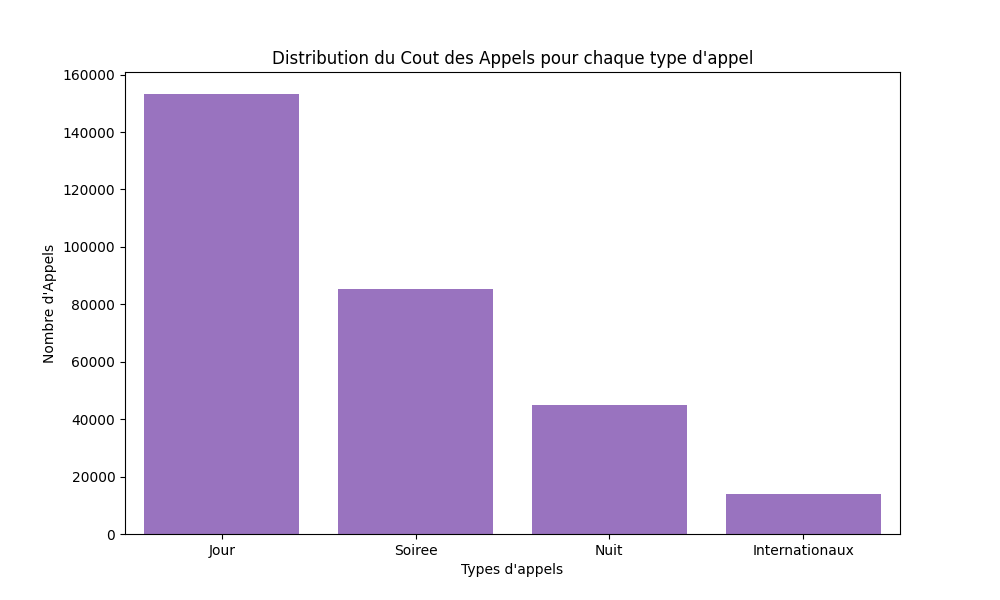
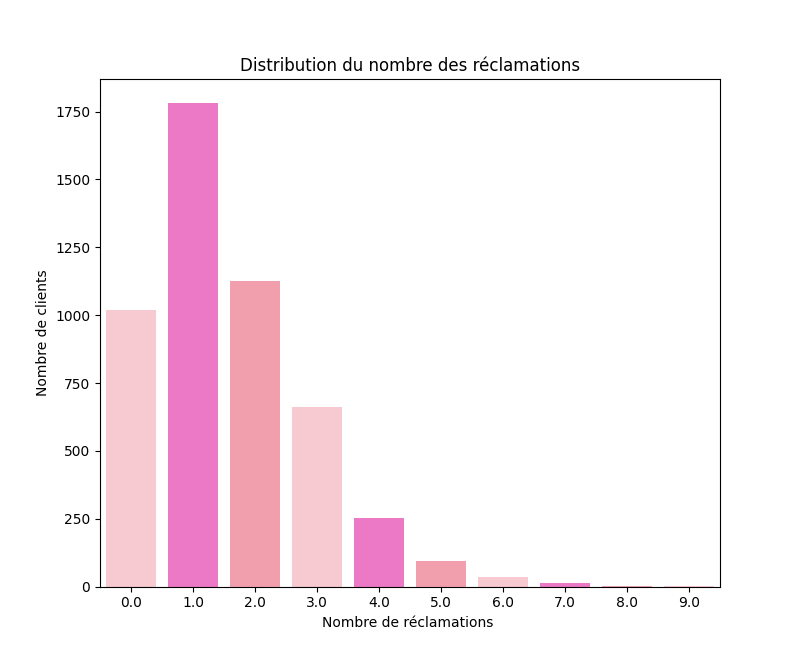
Le tableau 2.1 donne une vue d’ensemble des données. Il identifie les variables, leurs types et une simple description de leur signification.

|  |  |  |
| --- | --- | --- |
| **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 |

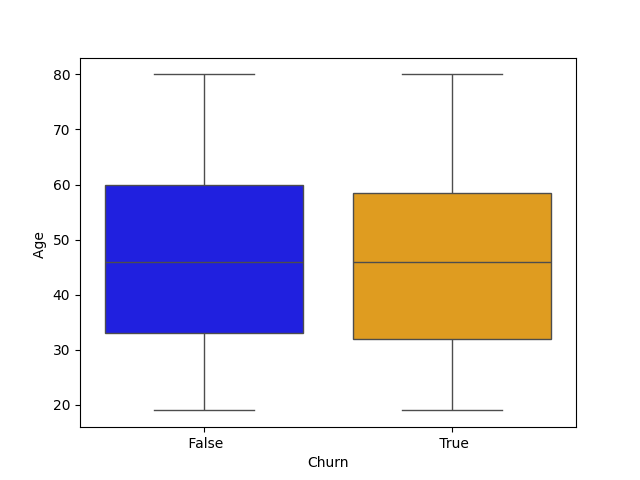
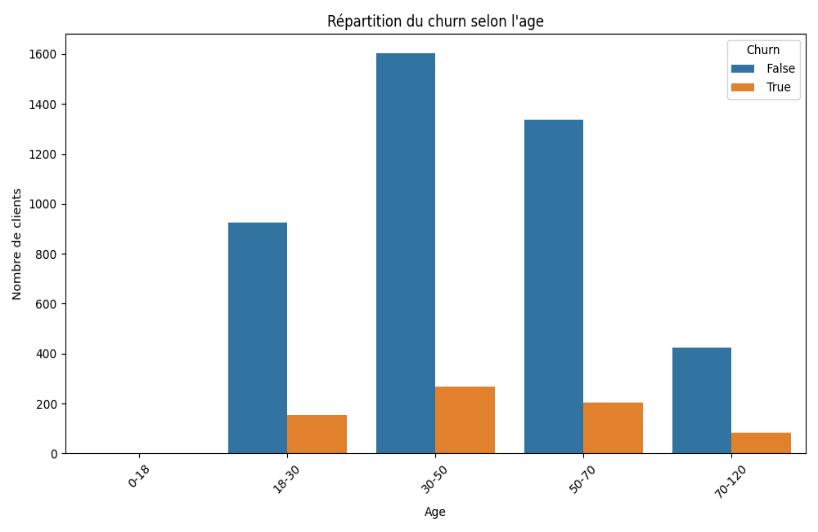
Notre base de données contient 23 variables dont 5 variables "id\_client", "marie" "offer\_type”, “active\_msg\_vocaux" et "churn" sont des variables de type "Chaîne de caractères".  
  
  
 Afin de réaliser notre étude, nous disposons d’un échantillon de 5000 clients qui ont été répartis de la manière 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.  
 

**2.4 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 interprétant 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   
  
**• data.head() :** Permet d’afficher les premières lignes d’un DataFrame. Par défaut, renvoie les 5 premières lignes.  
  
  
**• 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)  
  
**• 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)

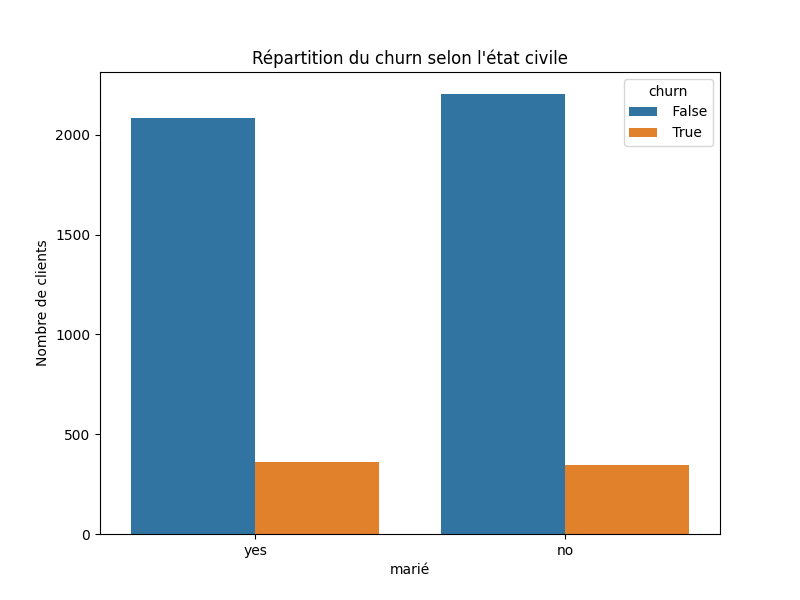
  
  
   
 Nous allons afficher des statistiques pour les colonnes qui contiennent des données de type "Object".  
  
  
**• data.nunique() :** renvoie le nombre de valeurs uniques pour chaque colonne.  
  
 **2.5 Analyse des données   
 2.5.1 Analyse univarié**   
 L’analyse univariée est une méthode d’analyse 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 :  
  
 • Churn : La répartition des clients en fonction de leur statut de churn est représentée dans la figure suivante   
   
 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 :  
   
 D’après le graphique, nous pouvons observer que notre jeu de données a 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.  
 

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 :   
   
  
 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 :   
   
  
☆ Variables numériques :  
 • Age :  
   
  
  
  
 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 :  
   
 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:  
   
  
   
 • Durée des Appels pour chaque type d’appel:  
   
  
  
  
 • Cout des Appels pour chaque type d’appel:  
   
 • Nombre de réclamations :  
   
  
 **2.5.2 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.

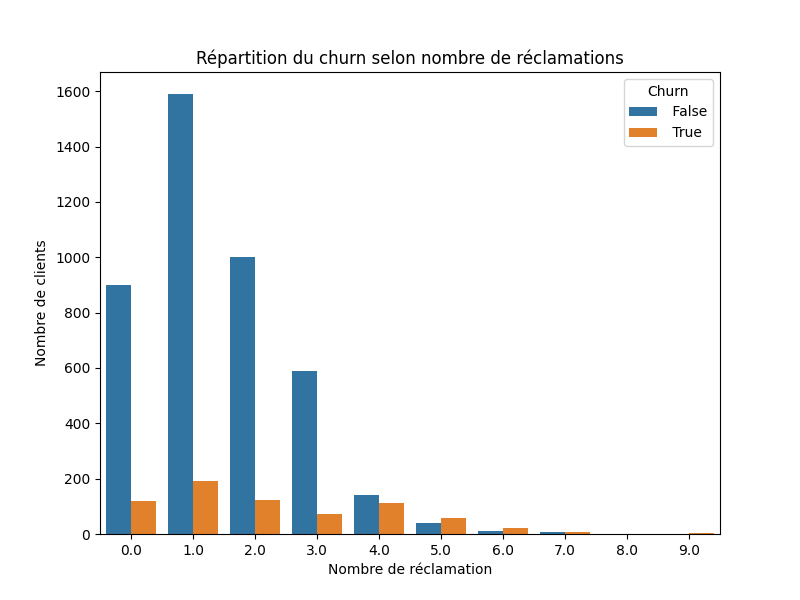
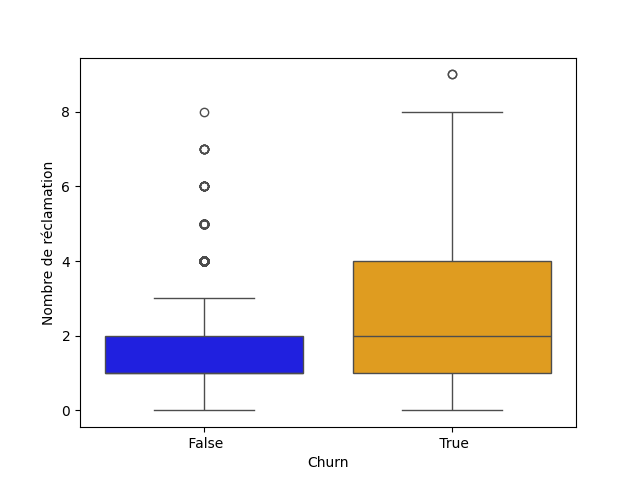
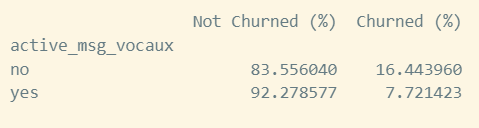
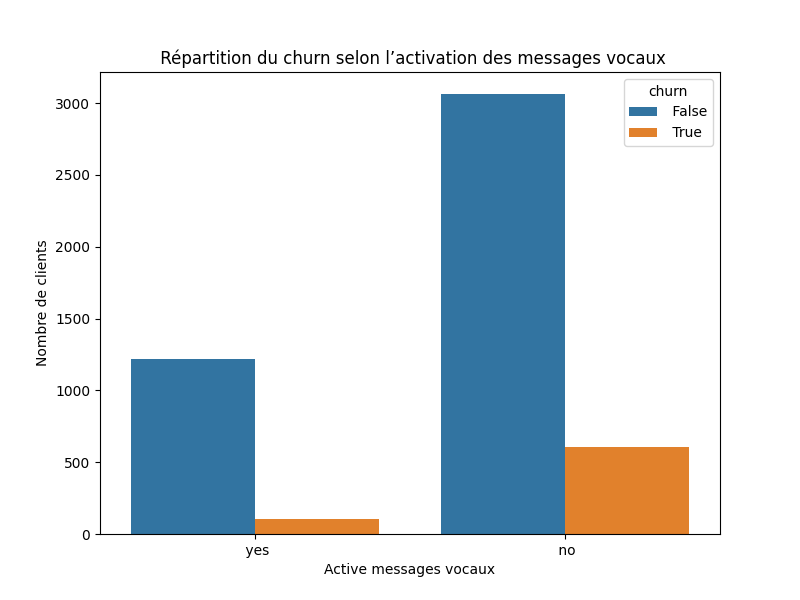
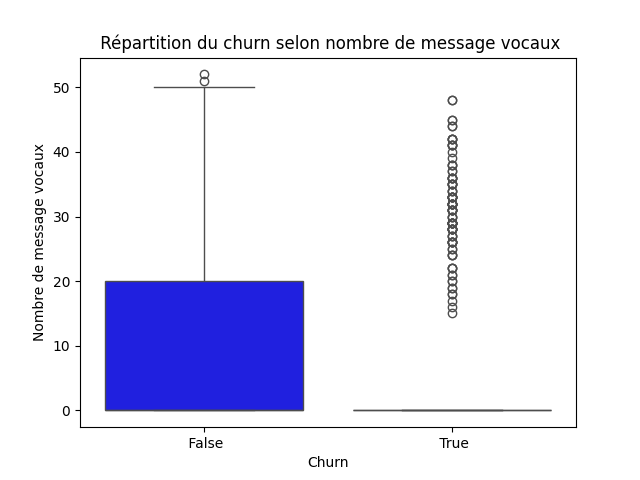
 • La figure illustre la distribution du churn selon l’age.  
   
 

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.

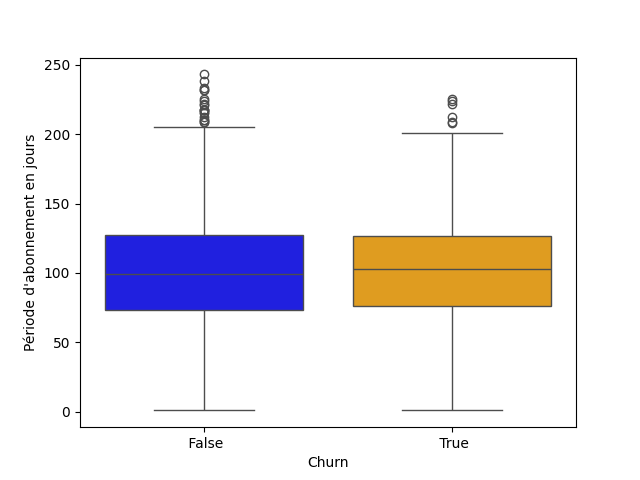
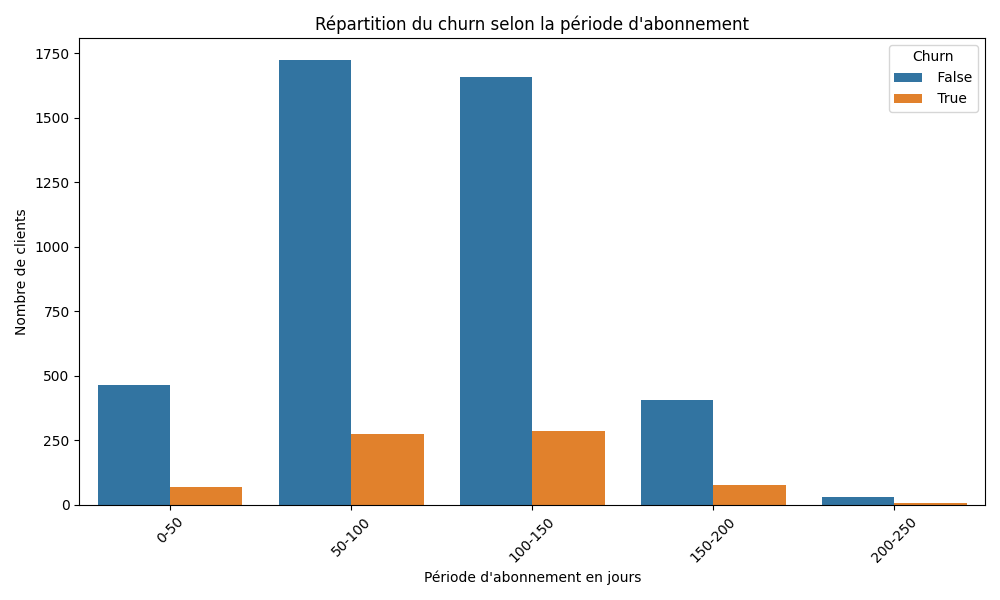
• La figure illustre la distribution du churn selon l’état civile.  
 

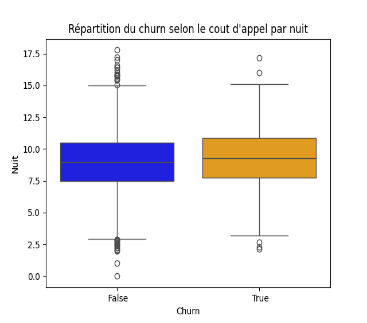
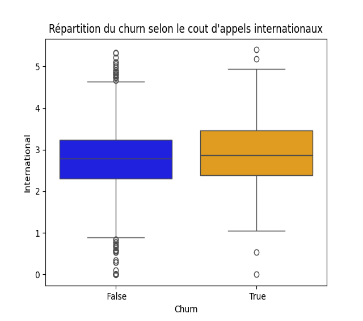
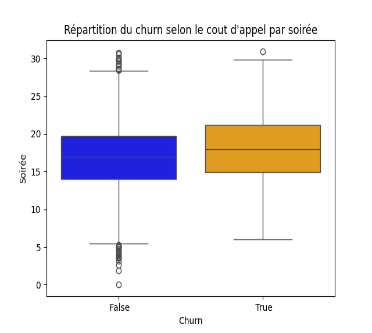
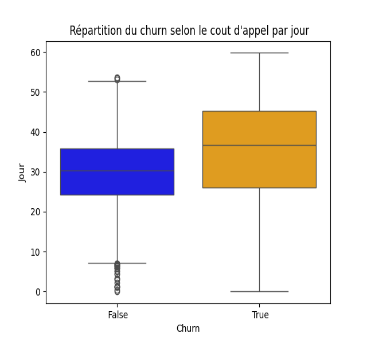
L'analyse de Tunisie Telecom révèle une corrélation entre l'état civil des clients et leur taux de churn. Les non-mariés ont un taux de churn plus élevé (environ 11%) par rapport aux mariés (environ 7%). Cette différence suggère que l'état civil influence la résiliation d'abonnement. Une exploration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

  
 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.  
  
 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.  
 

En analysant les données des messages vocaux, nous pouvons conclure que les clients qui reçoivent plus de 15 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.  


 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.  


Les coûts totaux qui correspondent aux clients churneurs 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.

Chapitre 3 : Prétraitement de données  
  
  
  
  
  
  
  
  
  
  
 Chapitre 4 : La modélisation

**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 :**

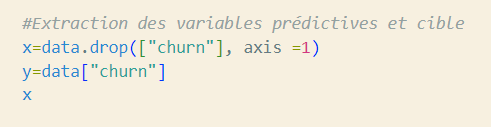
• Variables prédictives : « genre », « âge », « marié », « nb\_jours\_abonne », « nb\_appels\_jour »,

« 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 ».



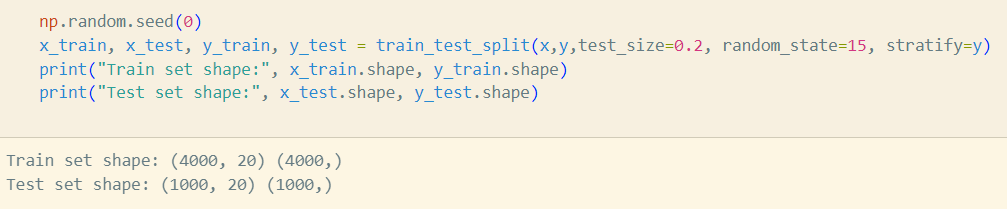
**Figure 4.1 :** Extraction de Feature et Target

**4.2.2 Données d’entrainement 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.



**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 utilisions 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 in 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 Aggregging) 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’entrainement. 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.

Random Forest utilise cette technique pour améliorer les performances prédictives qui est une méthode permet de créer plusieurs ensembles d’entrainement 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’entrainement.

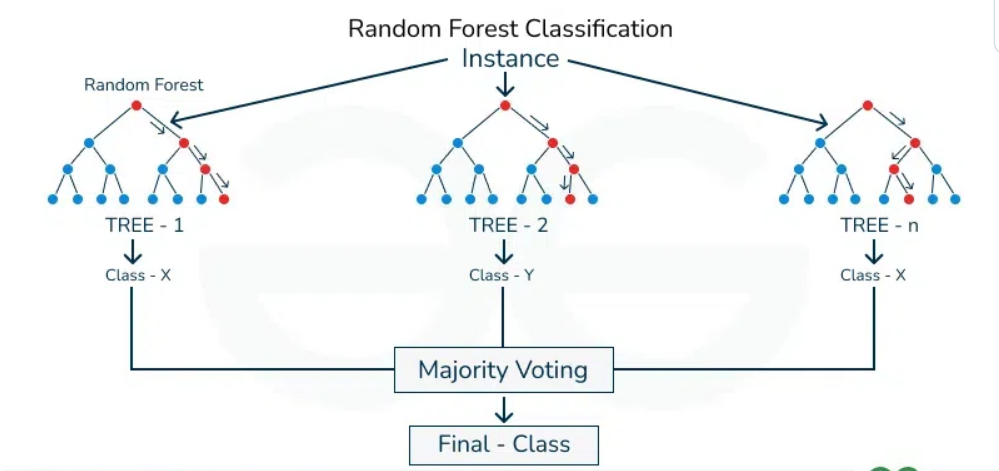
Nous allons décrire le fonctionnement de l’algorithme Random Forest en détaillant les étapes suivantes :

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

**• Les avantages de Random Forest :**

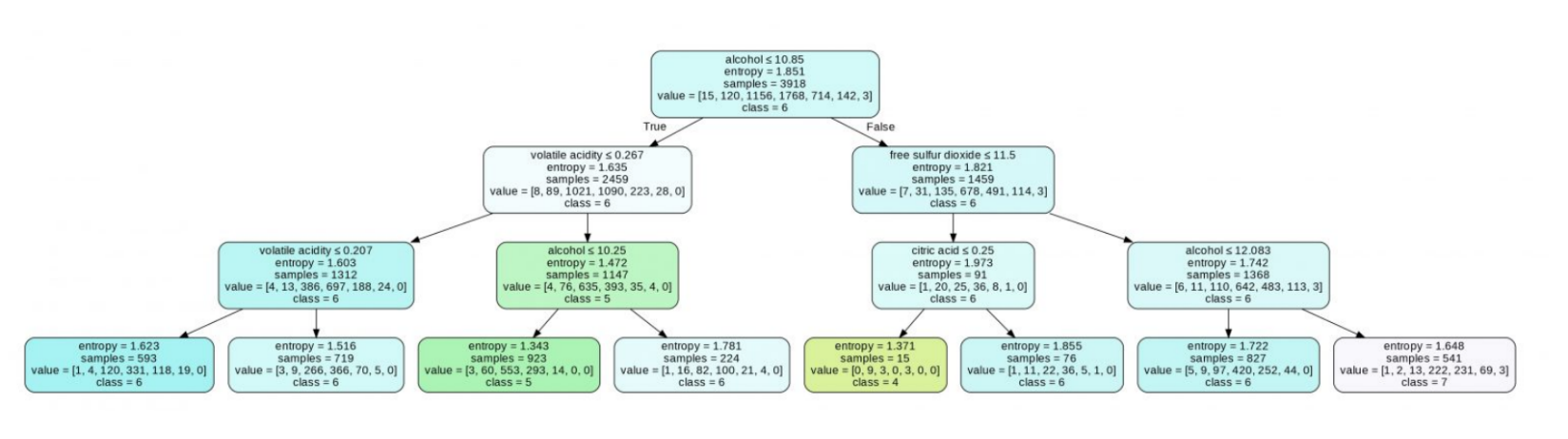
* Offre une estimation de l’importance des variables pour identifier les caractéristiques les plus pertinentes pour la prédiction.
* Réduit le risque de surajustement grâce à la méthode d’agrégation.
* Fournit des prédictions plus précises que l’algorithme de l’arbre de décision.
* Capable de gérer efficacement 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 supervise qui aide à la prise de décision en construisant un modèle sous forme d’un arbre compose 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.

**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.

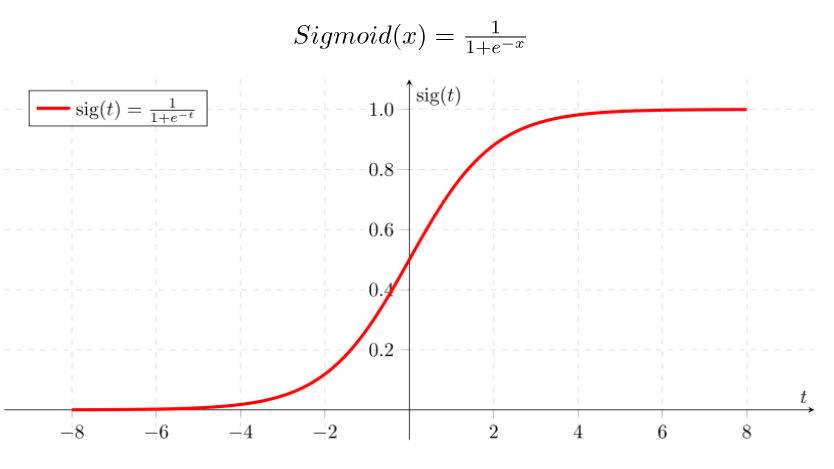
**4.3.3 Différence entre arbre de décision et Random Forest**

La Random Forest est compose 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

|  |  |
| --- | --- |
| Arbre de décision | 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 les arbres |

**4.3.4 Régression Logistique**

La régression logistique est un algorithme d’apprentissage supervise utilisée 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.

**Figure 4.5 :** Le fonctionnement d'Arbre de décision

**• Les avantages de la régression logistique**

* Simple à utiliser et à interpréter
* Moins sensible aux valeurs manquantes
* Possibilité de régularisation pour éviter le surapprentissage

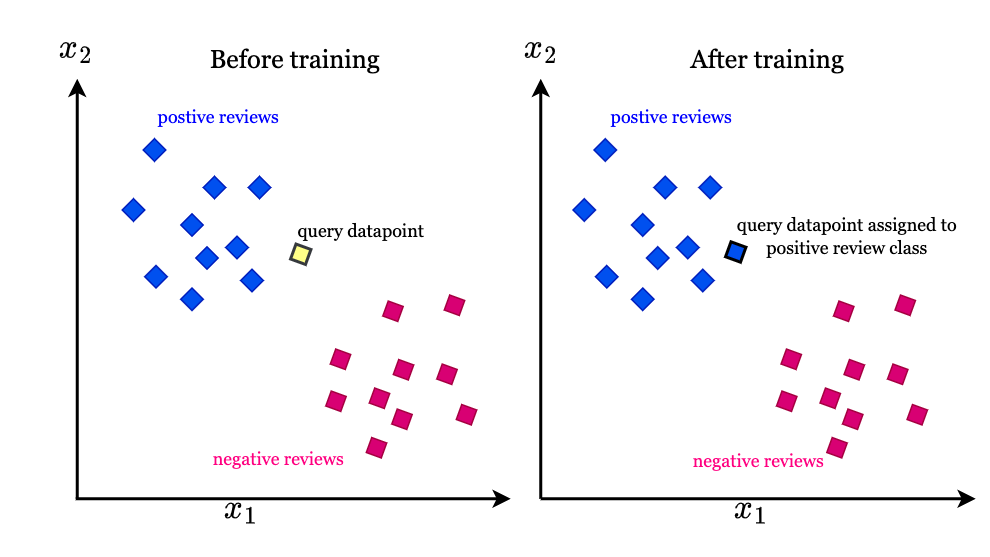
**• Les inconvénients de la régression logistique**

* 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 u algorithme d’apprentissage automatique supervisé simple qui peut être utilise 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.

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

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 |
|  |  |

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.

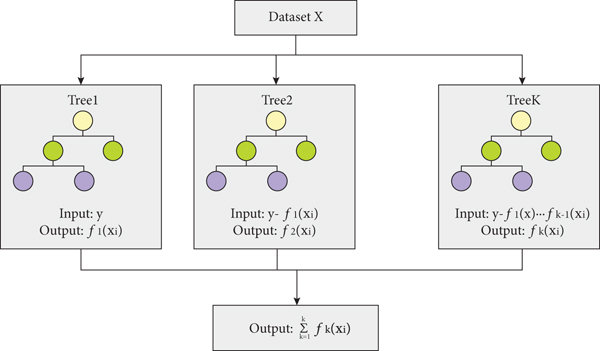
**4.3.6 eXtreme Gradient Boosting: XGboost**

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.

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

**• 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 a une classe donnée. Il est base sur le théorème de Bayes.

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

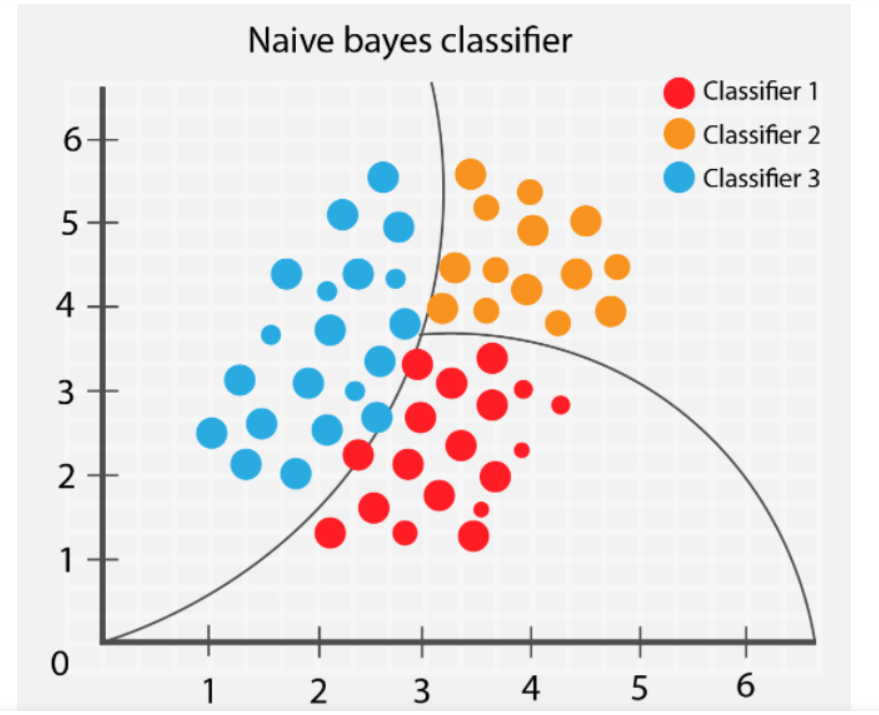
Le théorème de Bayes est formulé comme suit :

**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

**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.

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.

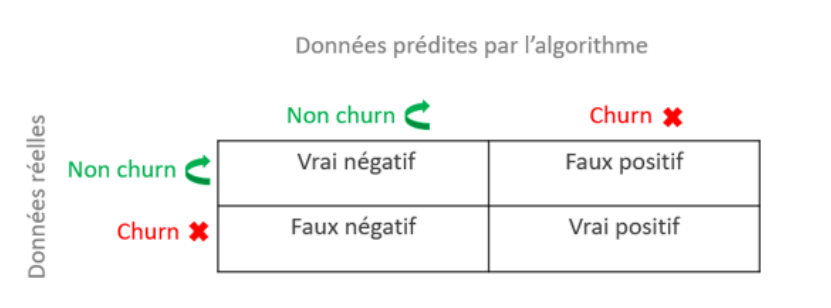
Dans note étude, nous avons étudié les problèmes de classification, pour lesquels nous avons utilise un ensemble de métriques telles que l’exactitude (Accuracy), la précision (Precision), le rappel (Recall) et le score F1 (F1-score).

**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. 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.



**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.

**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.

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.

La formule suivante est utilisée pour calculer le rappel :

**4.6.1.4 F1-Score**

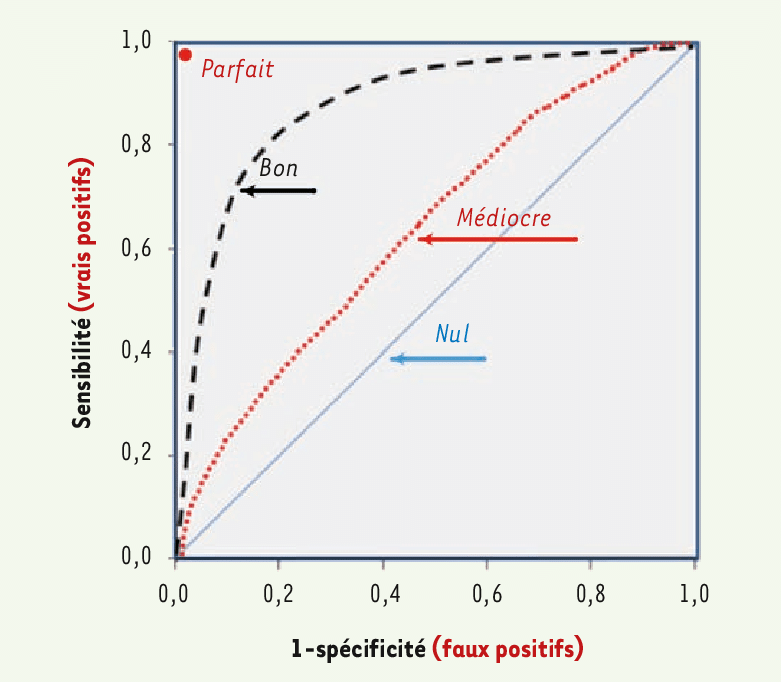
Le score F1 est une métrique de performance qui combine la précision et le rappel en une moyenne harmonique, ce qui permet d’obtenir une vue d’ensemble de la qualité des prédictions d’un modèle.

Nous calculons le F1 score avec la formule suivante :

**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.

La figure 4.10 illustre la courbe ROC-AUC.



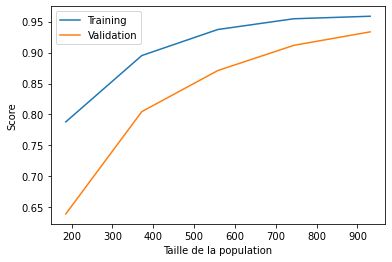
**Figure 4.10:** Lecture de courbe ROC

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.

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.
  + **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.

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

|  |  |  |
| --- | --- | --- |
| **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: 4  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: 4 |  |
| Decision Tree | criterion: entropy  max\_depth: 7 | La fonction de mesure de qualité de la division d’un nœud 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

**5.1 Introduction**

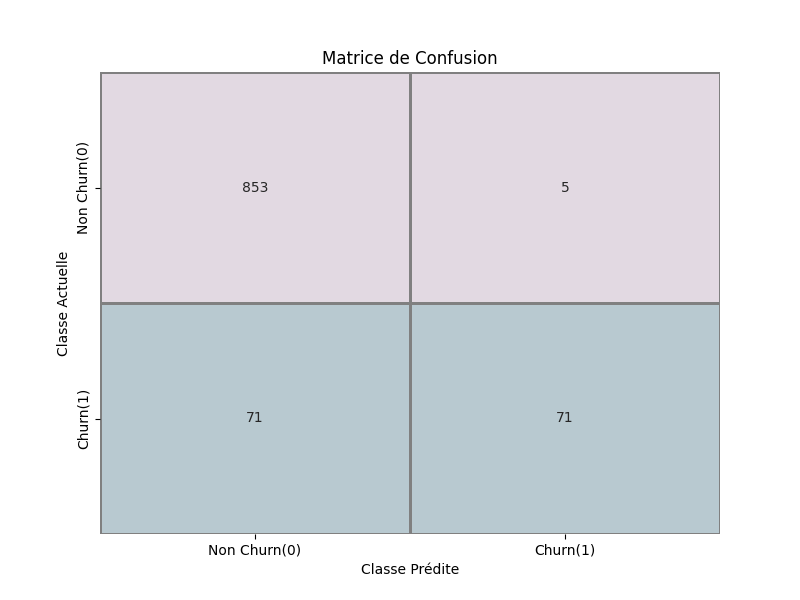
Apres avoir développé les modèles d’apprentissage automatique, nous comparons leur performance individuelle pour identifier le modèle le plus performant. Ce chapitre présente les évaluations de chacun des modèles, basées sur des indicateurs de performances spécifiques. En se basant sur ces résultats, 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**

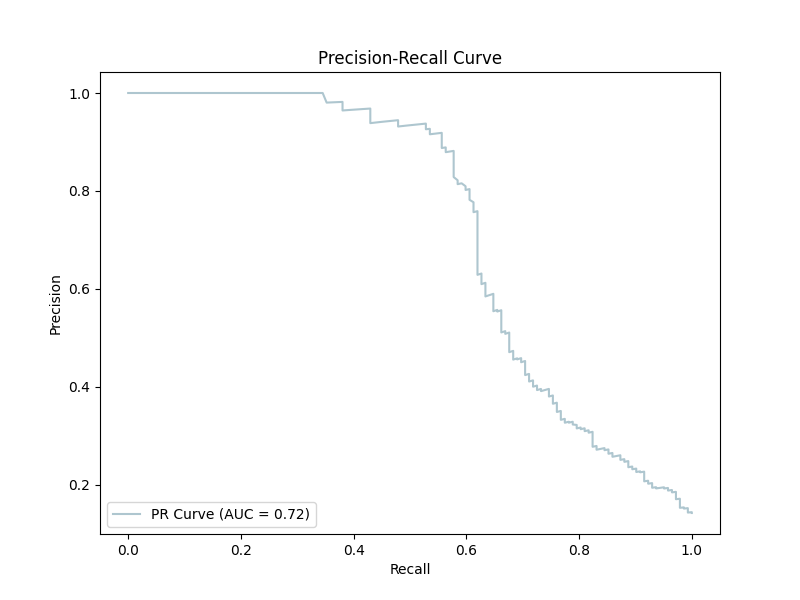
La matrice de confusion du modèle « Random Forest » est illustrée dans la figure suivante :



**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.

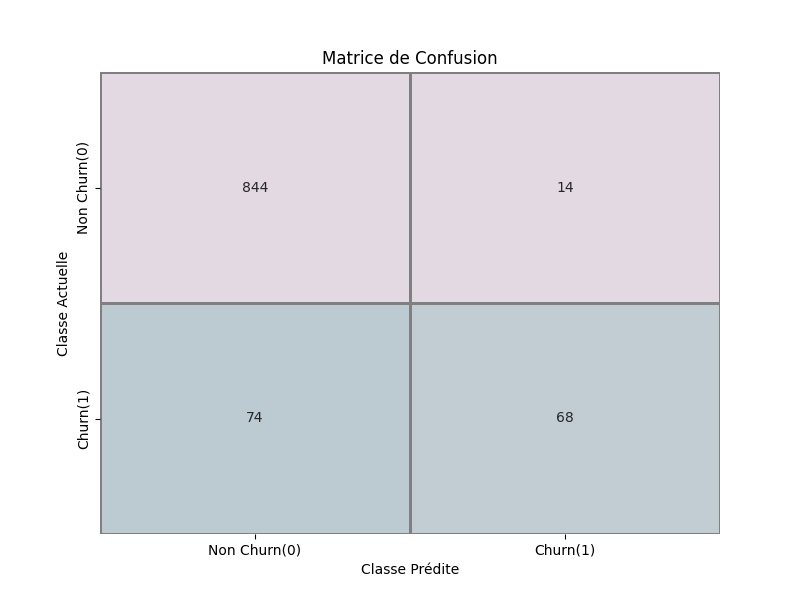


**Figure 5.3:** Learning Curve de Random Forest

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.

**5.2.1 Decision Tree**

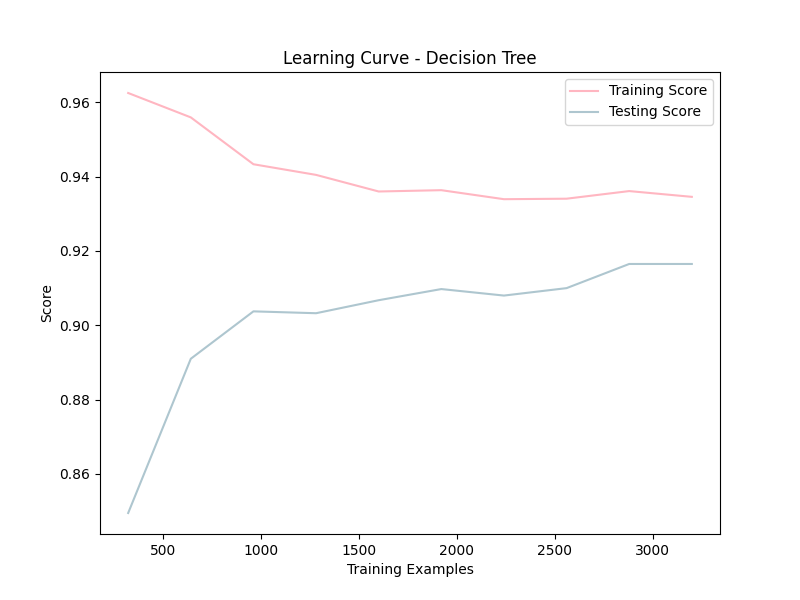
La matrice de confusion du modèle « Decision Tree » est illustrée dans la figure suivante :



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

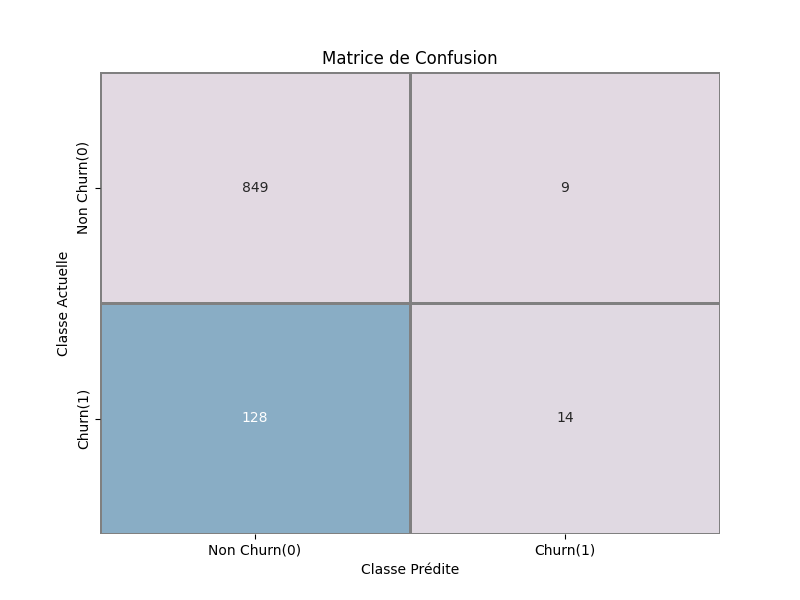
L’algorithme a prédit :

* + 844 vrais négatifs et 68 vrais positifs.
  + 74 faux négatifs et 14 faux positifs.



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

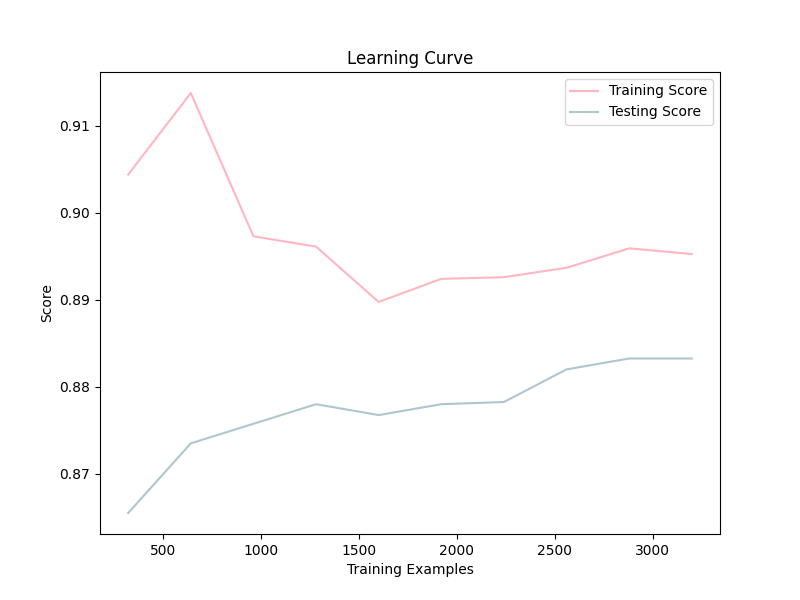
L'arbre de décision peut bien généraliser les données, car sa précision sur les données de test s'améliore avec la taille des données d'entraînement. Mais à un moment donné, ajouter plus de données ne conduit plus à une amélioration significative de la précision du test. Le score de test se stabilise 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.

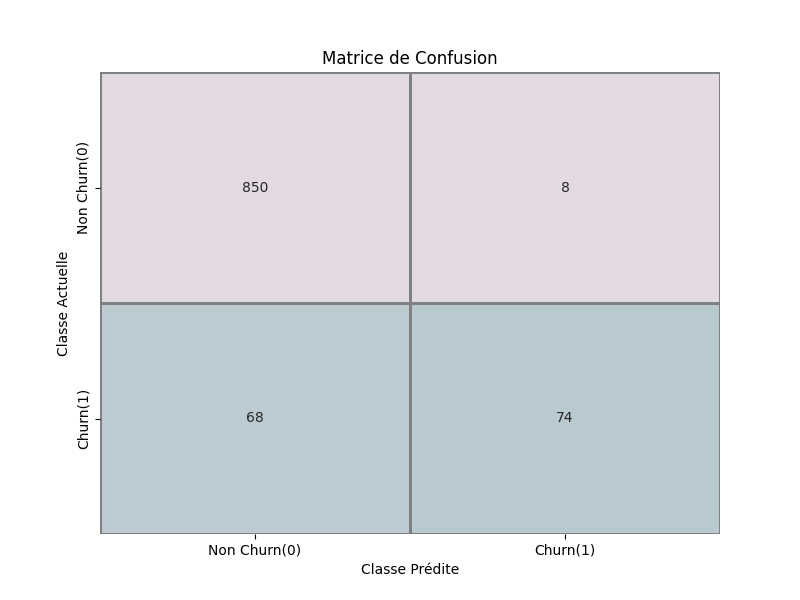


**Figure 5.7:** Learning Curve de Logistic Regression

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 les 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.

**5.2.3 XGBoost**

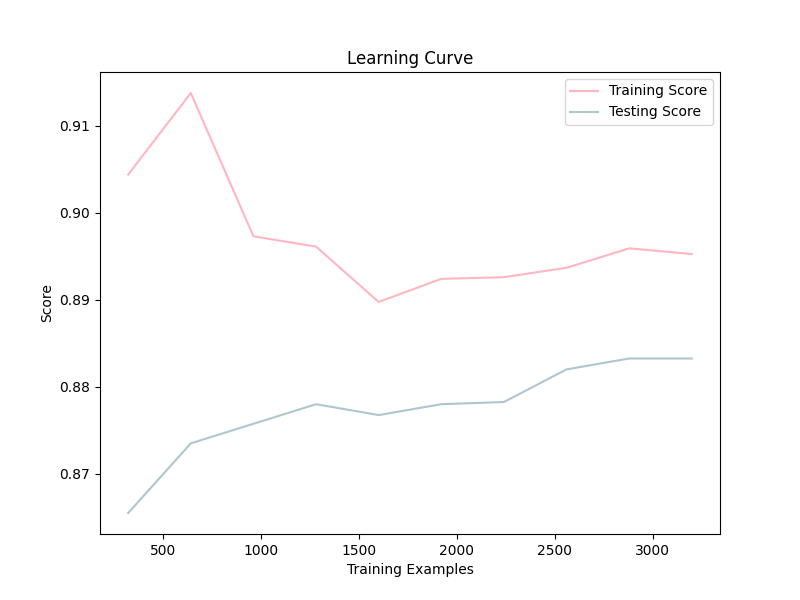
Voici la représentation graphique de la matrice de confusion obtenue pour le modèle XGBoost :



**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.

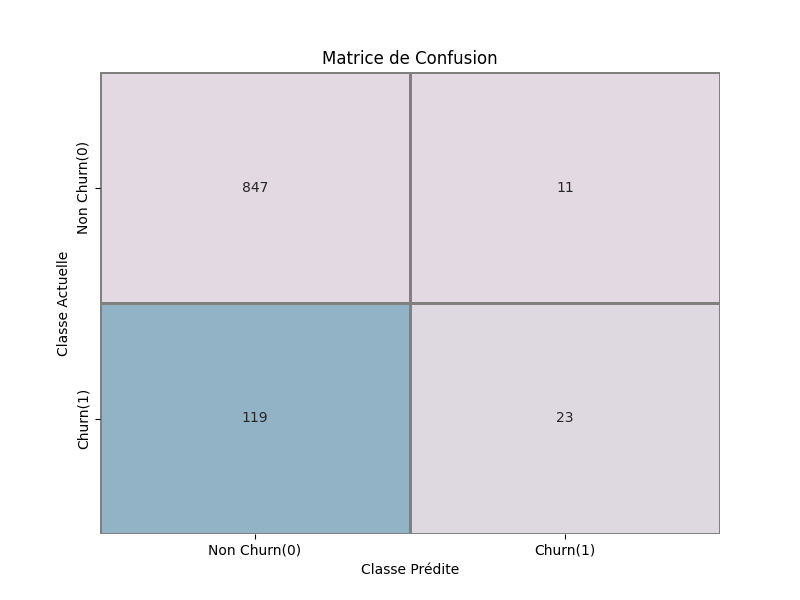


**Figure 5.9:** Learning Curve de XGBoost

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.

**5.2.4 K-Nearest Neighbors : KNN**

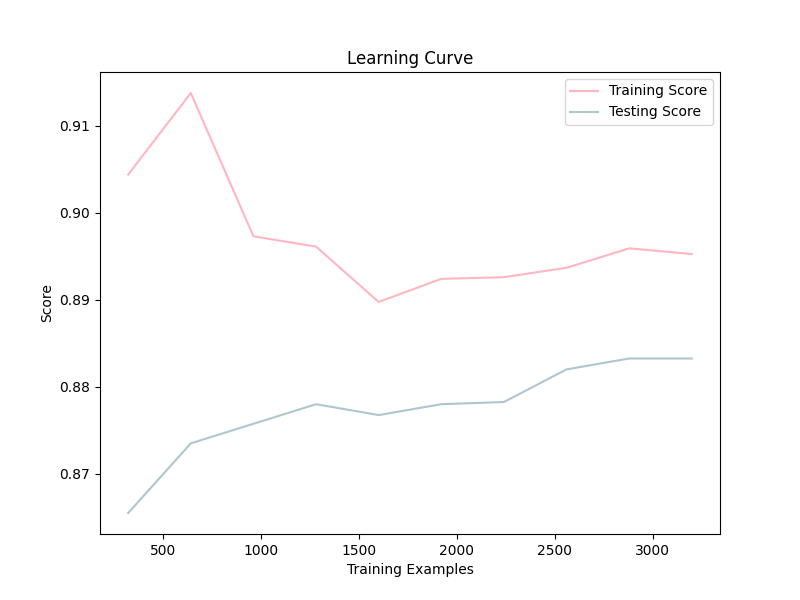
La matrice de confusion du modèle « KNN » est illustrée dans la figure suivante :



**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.

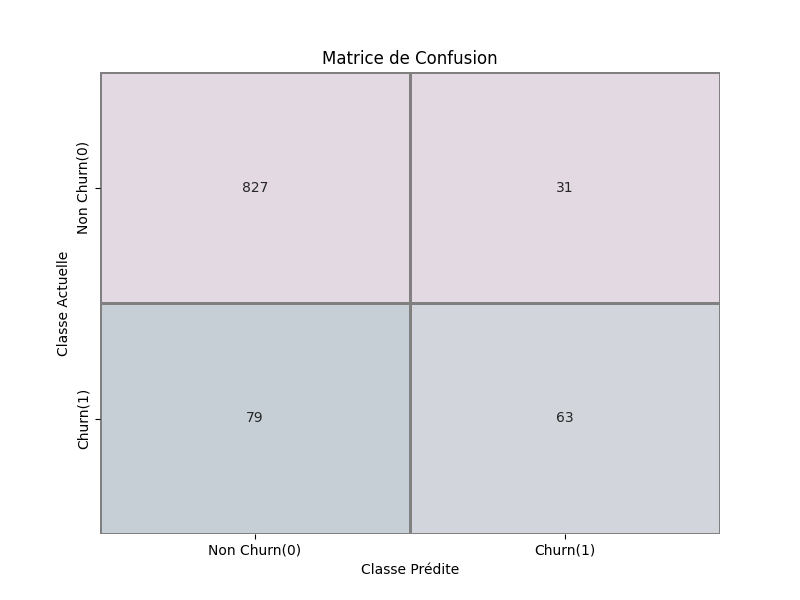


**Figure 5.11:** Learning Curve de KNN

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.

**5.2.5 Naïve Bayes**

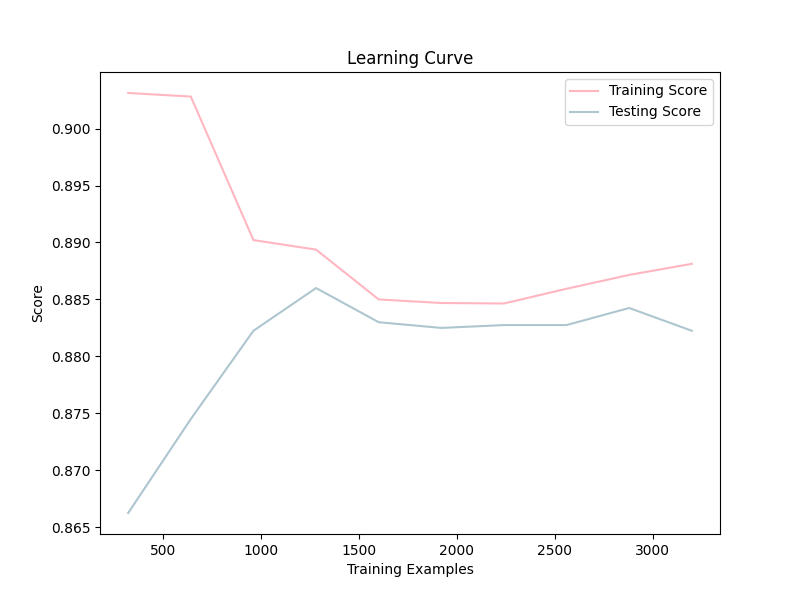
La figure ci-dessous illustre la matrice de confusion pour le modèle naïve bayes :



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

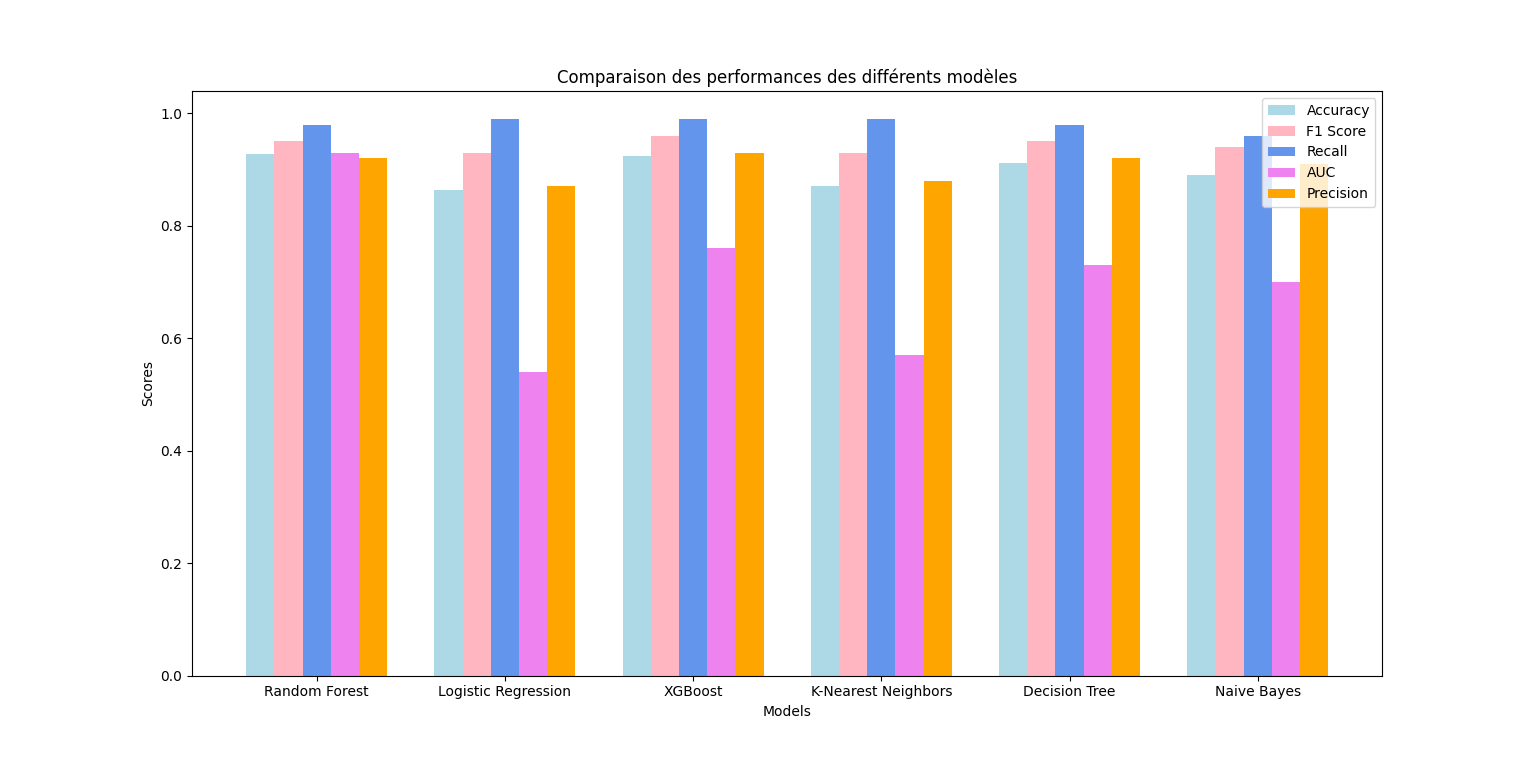
L’algorithme a prédit :

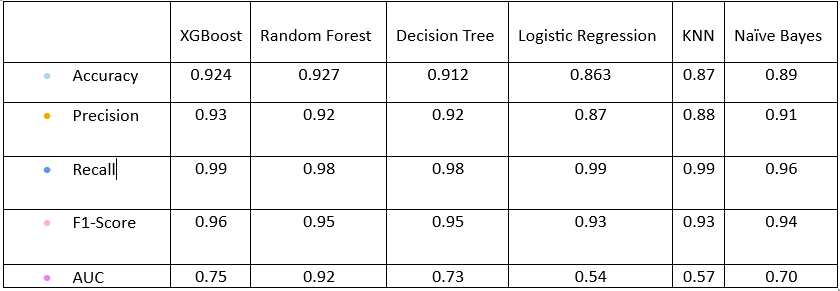
* + 827 vrais négatifs et 63 vrais positifs.
  + 79 faux négatifs et 31 faux positifs.



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

**5.3 Comparaison et évaluation des algorithmes utilisés**

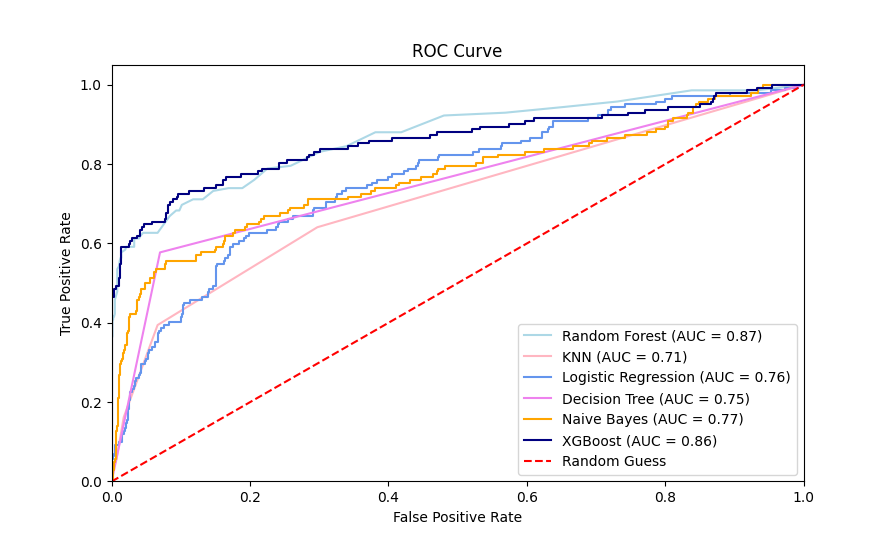




**Figure 5.14:** Comparaison des performances des différents modèles

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 affichent des performances autour de 0,92.

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

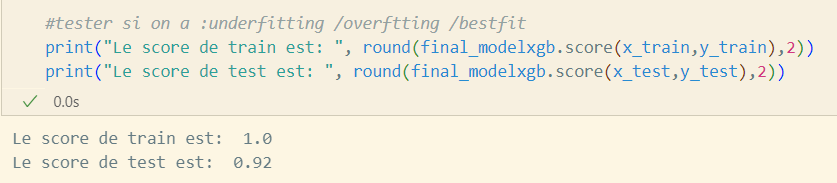


**Figure 5.15 :** Diagramme ROC/AUC de tous les modèles

Dans notre situation, le meilleur algorithme est celui avec le F1-score le plus élevé. Comme nous avons expliqué précédemment, nous ne choisirons pas la mesure de performance "Accuracy" pour l’évaluation, car notre ensemble de données est déséquilibré. Après avoir examiné plusieurs mesures et scores, nous avons conclu que l’algorithme Random Forest est le meilleur choix car il affiche le F1-score le plus élevé avec la valeur 0.958. Par conséquent, nous avons choisi de le déployer dans notre application afin de prédire si un client résiliera ou non son abonnement.

**5.4 Ajustement du modèle**

L’ajustement du modèle est essentiel pour identifier la cause de la faible précision du modèle. En évaluant les erreurs de prédiction 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



**Figure 5.16 :** 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".

**5.5 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

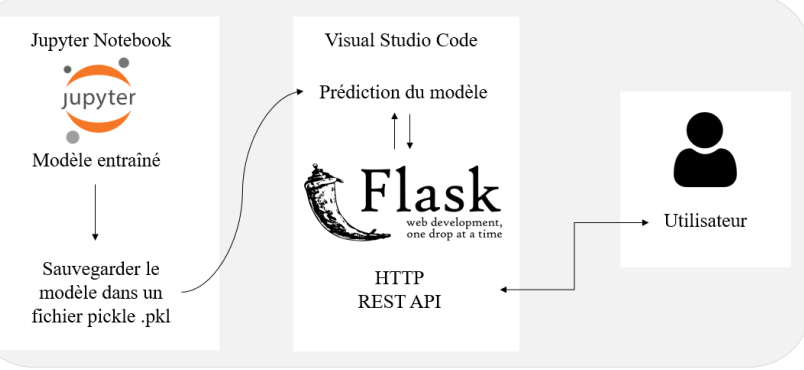
**6.1 Introduction**

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.

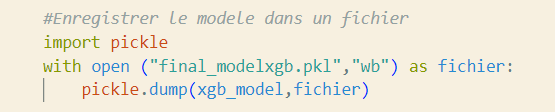


**Figure 6.1 :** Architecture de déploiement du modèle

**6.3 Extraction du modèle**

En tant que data scientist, vous manipulerez 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.



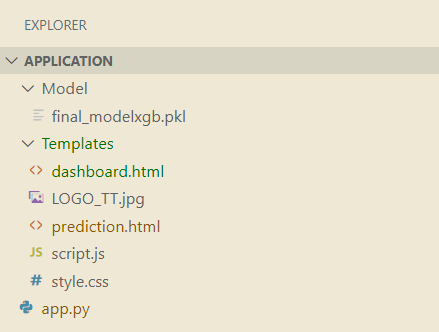
**Figure 6.2 :** Méthode d'enregistrement le modèle

**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:**

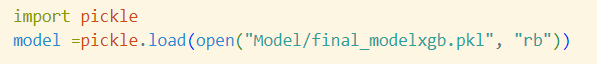
Nous avons divisé notre application en deux dossiers pour faciliter sa structure.



**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 :**



**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 :**
* **Exécution de l’application:**

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



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

**6.4.1 Les interfaces de l’application**

**6.5 Construction de tableau de bord**

**6.6 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.