

**Modèle de dossier du candidat**

Titre RNCP “Expert en Informatique et Systèmes d’Information”

N° 36286 – Niveau 7 (Bac+5)

Ce modèle est basé sur le référentiel RNCP 36286. Il contient le sommaire minimal exigé pour cet examen. Il reprend tous les points du référentiel encadrés dans un projet unique de fn de cycle.

**1. Page de garde**





Sujet sous forme de

question

COUDRIER Georges

Référent Société:

PLESSIS Hubert

Patron

ESI-03 2023-2025

​ ​ ​ ​ ​ ​

1. **Remerciements**

Je tiens à remercier Hubert Plessis, mon tuteur en entreprise, pour son accompagnement et ses conseils tout au long de cette année. Sa disponibilité et son expérience ont été d’une grande aide pour moi dans ce projet.

Un grand merci également à mes collègues, Vivien et Sonnarin, qui ont toujours été là pour répondre à mes questions, m’encourager et partager leur savoir-faire. Travailler avec eux a été un vrai plaisir.

Enfin, je remercie tous mes professeurs pour leur enseignement et leur soutien pendant ma formation. Grâce à eux, j’ai pu développer les compétences nécessaires pour mener à bien ce projet.

Merci à vous tous !

1. **Résumé**
2. **Abstract**
3. **Sommaire**
4. **Liste des abréviations**
5. **Glossaire (défnition des termes utilisés)**
6. **Présentation personnelle**

**8.1 Introduction**

**8.1 Parcours académique et professionnel**

**8.1 Objectifs de la présentation**

**9. Présentation de la problématique**

**9.1 Contexte du projet**

**9.2 Description de la problématique**

**9.3 Objectifs du projet**

**9.4 Importance et impact de la problématique**

**9.5 Questions de recherche et hypothèses**

1. **Diagnostic général du Système d’Information et Veille technologique**

* **10.1 Diagnostic du Système d’Information**
* ​10.1.1 Observations générales
* ​10.1.2 Recommandations
  + - **10.2 Veille technologique**
      * ​10.2.1 Introduction
      * ​10.2.2 Méthodologie de la veille technologique
      * ​10.2.3 Résultats de la veille technologique
  + ​10.2.4 Recommandations et perspectives
  + ​10.2.5 Conclusion

1. **Besoins fonctionnels et Cahier des charges fonctionnel**

**11.1 Collecte et analyse des besoins**

* 11.1.1 Méthodologie de collecte des besoins
* 11.1.2 Analyse approfondie des besoins
* 11.1.3 Identifcation des contraintes et des exigences
* 11.1.4 Validation des besoins avec les parties prenantes

**11.2 Description des fonctionnalités requises**

* 11.2.1 Fonctionnalités essentielles
* 11.2.2 Fonctionnalités secondaires

**11.3 Analyse stratégique d’implémentation**

* 11.3.1 Introduction
* 11.3.3 Gestion des risques
* 11.3.4 Élaboration de la stratégie d’implémentation
* 11.3.5 Présentation des préconisations
* 11.3.6 Conclusion

**12. Implémentations techniques**

**12.1 Infrastructure et environnement de déploiement**

* 12.1.1 Infrastructure matérielle
* 12.1.2 Environnement de déploiement (cloud, on-premise, hybride)

**12.2 Technologies et outils utilisés**

* + 12.2.1 Algorithmes et frameworks
* 12.2.2 Outils de développement et de gestion de projet

**12.3 Surveillance et monitoring de la sécurité**

* 12.3.1 Outils de monitoring
* 12.3.2 Techniques de détection et de réponse aux incidents

**12.4 Plans de sauvegarde et de récupération**

* 12.4.1 Stratégies de sauvegarde des données
* 12.4.2 Plans de reprise après sinistre

**13. Gestion du projet et management**

**13.1 Élaboration du planning et gestion des ressources**

* 13.1.1 Création d'un planning détaillé
* 13.1.2 Allocation des ressources

**13.2 Suivi de l'avancement et gestion des risques**

* 13.2.1 Indicateurs de performance
* 13.2.2 Gestion des risques

**13.3 Management de l'équipe projet**

* 13.3.1 Encadrement et motivation
* 13.3.2 Développement des compétences

**13.4 Communication et coordination avec les parties prenantes**

* 13.4.1 Plan de communication
* 13.4.2 Engagement des parties prenantes

**13.5 Planifcation**

* 13.5.1 Introduction
* 13.5.2 Méthodologie de projet
* 13.5.3 Phases du projet
* 13.5.4 Planning détaillé (diagramme inclus)
* 13.5.5 Ressources et budget

**13.6 Note de cadrage du projet**

13.6.1 Contexte du projet

13.6.2 Objectifs et périmètre

13.6.3 Organisation et ressources

13.6.4 Budget et délais

13.6.5 Contraintes et risques

13.6.6 Modalités de pilotage et de communication

1. **Conception et Développement de l’Application Informatique**

**14.1 Conception**

* 14.1.1 Introduction
* 14.1.2 Décomposition de l’application
* 14.1.3 Fonctionnement de l’application
* 14.1.4 Données utilisées
* 14.1.5 Environnement informatique proposé
* 14.1.6 Spécifcations techniques

**14.2 Développement et Stratégie de Tests**

* + 14.2.1 Objectifs de la stratégie de tests
  + 14.2.2 Types de tests et outils utilisés
* 14.2.3 Plan de test détaillé
  + 14.2.4 Environnement de test

**14.3 Modalités de livraison**

14.3.1 Mise en service opérationnelle et vérifcation de service régulier (VSR)

14.3.2 Conclusion

1. **Construction et Développement d’un Modèle de « Big Data » (Option Big Data)**

**15.1 Enjeux du Big Data et l’analyse des données**

* 15.1.1 Introduction au Big Data

Le Big Data et l’IA apportent des bénéfices majeurs à NexaCore pour répondre efficacement à la problématique de churn et améliorer l’expérience utilisateur.

1. **Prise de Décision Plus Rapide** :  
   En exploitant les capacités analytiques des Big Data, NexaCore peut collecter et traiter en temps réel des données comportementales des clients, permettant des ajustements immédiats aux campagnes marketing et aux processus de fidélisation.
2. **Analyse Plus Exhaustive** :  
   Grâce au Big Data, NexaCore peut centraliser les interactions clients, les performances des campagnes et les tendances d’utilisation, offrant une vision complète et holistique pour des actions mieux ciblées.
3. **Précision des Recommandations** :  
   Les algorithmes d’IA appliqués aux données massives permettent de proposer des recommandations adaptées aux besoins spécifiques de chaque utilisateur, augmentant la pertinence des actions entreprises.
4. **Identification Proactive des Risques** :  
   Le traitement des Big Data permet de détecter précocement les signaux de désengagement client, grâce à des modèles prédictifs qui alertent sur les risques de churn.
5. **Personnalisation des Expériences** :  
   En combinant IA et Big Data, NexaCore peut segmenter les clients de manière précise et offrir des parcours utilisateur personnalisés, renforçant ainsi leur satisfaction et leur engagement.

* 15.1.2 Objectifs

L'objectif principal est de réduire le taux de churn en fidélisant les clients existants, car conserver un client coûte généralement moins cher que d'en acquérir un nouveau. Cette réduction a un impact direct sur :

* **L'augmentation des revenus récurrents** : Chaque client fidélisé contribue à la stabilité financière de l’entreprise.
* **L'amélioration de la réputation** : Un churn réduit reflète une satisfaction client accrue, attirant de nouveaux prospects grâce au bouche-à-oreille et aux témoignages positifs.
* **La diminution du coût d'acquisition client (CAC)** : Fidéliser les clients réduit la nécessité d'investir massivement dans la conquête de nouveaux clients pour compenser les pertes.

**Pourquoi repérer la probabilité de churn ?**

Identifier la probabilité de churn permet d’agir de manière préventive en :

* **Ciblant les clients à risque** : Une action spécifique peut être déclenchée (offres personnalisées, support renforcé) pour inverser leur désengagement.
* **Optimisant les ressources** : Au lieu de déployer des actions marketing générales, les efforts se concentrent sur les segments réellement à risque.
* **Améliorant l’expérience utilisateur** : Les clients identifiés comme insatisfaits peuvent bénéficier d’un accompagnement adapté, renforçant leur perception positive de NexaCore.

**Ce que cela apporte à l’entreprise**

* **Augmentation de la valeur vie client (CLV)** : Un client fidèle reste plus longtemps, générant un chiffre d’affaires accru.
* **Réduction des pertes de revenus** : Chaque client retenu contribue à stabiliser les revenus mensuels.
* **Insights stratégiques** : Les données collectées pour analyser le churn peuvent aussi être utilisées pour identifier des tendances et ajuster les offres ou services.

**Pourquoi utiliser l’IA plutôt qu’une personne ?**

1. **Scalabilité** : L’IA peut analyser des millions de données simultanément, une tâche impossible pour une personne.
2. **Précision** : Les algorithmes d’IA détectent des corrélations complexes et des patterns subtils dans les données, souvent invisibles à l'œil humain.
3. **Temps de réponse** : L’IA travaille en temps réel, permettant des actions immédiates pour contrer les risques de churn.
4. **Automatisation** : Une fois les modèles prédictifs en place, l’IA peut fonctionner de manière autonome, réduisant les coûts opérationnels liés à l’analyse manuelle.

* 15.1.3 Données disponibles

Dans le cadre du projet NexaCore, le dataset utilisé offre une vue complète des comportements d'achat et des insights sur les ventes, adapté aux besoins de prédiction de churn et d'optimisation des stratégies de fidélisation. Les données disponibles sont structurées autour des dimensions suivantes :

**1. Détails des Clients**

* **Identifiants Clients** : Chaque client possède un ID unique permettant un suivi précis de ses interactions.
* **Attributs d’Achat** : Fréquence d'achat, panier moyen et valeur client à vie (Lifetime Value).

**2. Informations Produits**

* **Identifiants Produits** : ID unique pour chaque produit.
* **Catégories Populaires** : Liste des catégories les plus achetées par les clients, permettant d'identifier les préférences générales.

**3. Métriques de Ventes**

* **Valeur Moyenne des Commandes** : Indicateur de la dépense typique des clients.
* **Valeur Client à Vie (CLV)** : Montant total attendu qu’un client dépensera sur toute la durée de sa relation avec l’entreprise.
* **Fréquence d’Achat** : Nombre de transactions réalisées par client sur une période donnée.

**4. Insights Comportementaux**

* **Temps entre Achats** : Intervalle moyen entre deux transactions d’un même client.
* **Moments d’Achat Préférés** : Analyse des heures et des jours où les achats sont les plus fréquents.
* **Catégories Fréquemment Achetées** : Produits ou services les plus populaires par client.

**5. Insights Géographiques**

* **Régions** : Les données couvrent différentes zones géographiques, notamment :
  + Amérique du Nord
  + Europe
  + Asie
  + Amérique du Sud
* Ces informations aident à adapter les stratégies marketing en fonction des spécificités locales.

**6. Saisonnalité**

* **Dates de Vente en Pic** : Identification des périodes de vente les plus intenses, souvent liées aux fêtes ou aux événements promotionnels.
* **Saisons Associées** : Correspondance des tendances de vente avec des périodes spécifiques, comme Noël ou la rentrée.

**7. Stratégies de Rétention**

* **Méthodes Appliquées** : Informations sur les campagnes de fidélisation déjà en place, telles que :
  + Remises ou promotions personnalisées.
  + Programmes de fidélité.
  + Campagnes d’e-mails ciblés pour les clients à risque.

**8. Probabilité de Churn**

* **Métriques de Probabilité** : Score calculé indiquant le risque d’attrition pour chaque client, basé sur des modèles prédictifs.

**Description Globale**

Le dataset de NexaCRM est conçu pour des analyses détaillées et des modèles prédictifs, en combinant des données démographiques, comportementales et transactionnelles. Il offre une base solide pour :

* **L’exploration des données (EDA)** : Identifier les tendances et segments clients clés.
* **Les modèles de segmentation** : Créer des groupes homogènes pour des actions marketing ciblées.
* **La prédiction du churn** : Anticiper les comportements à risque et agir en conséquence.

Ces données permettent à NexaCore de fournir une vision précise et contextualisée des clients, renforçant les capacités de personnalisation et d’engagement pour réduire le churn efficacement.

\* NexaCRM est une entreprise fictive, les données ont été récupéré sur Kaggle  
- <https://www.kaggle.com/datasets/imranalishahh/sales-and-customer-insights/data>

* 15.1.4 Analyse des données

Les données fournies permettent de réaliser une exploration approfondie afin de dégager des tendances et des corrélations entre différents comportements clients et leur risque de churn. Voici les principaux éléments d’analyse :

1. **Comportement d'Achat** :
   * Les **fréquences d’achat** et le **temps moyen entre les transactions** sont des indicateurs clés pour détecter les clients en phase de désengagement. Une augmentation des intervalles entre les achats peut signaler une baisse d'intérêt.
   * La **valeur client à vie (CLV)** peut révéler les clients stratégiques pour lesquels des actions spécifiques de rétention doivent être mises en œuvre.
2. **Données Comportementales** :
   * Les **moments d'achat préférés** et les catégories de produits régulièrement achetées permettent d’identifier des habitudes spécifiques. Un changement soudain dans ces habitudes peut indiquer une probabilité accrue de churn.
   * L’analyse des **données saisonnières** peut montrer si certains clients ont un comportement d’achat lié à des périodes spécifiques. Une absence d’activité durant ces périodes pourrait être un signal d’alerte.
3. **Interactions avec les Stratégies de Rétention** :
   * Les réponses aux campagnes de fidélisation, comme l’engagement avec les e-mails ou les remises, fournissent des insights sur la probabilité de réengagement. Les clients qui n'interagissent pas avec ces initiatives sont plus à risque.
4. **Insights Géographiques** :
   * Les comportements d'achat varient souvent selon les régions. L’analyse des performances par région peut révéler des tendances spécifiques (ex. : zones où le churn est plus élevé).
5. **Score de Probabilité de Churn** :
   * Les scores préexistants calculés à partir des données comportementales et transactionnelles permettent d’identifier directement les clients à risque pour des interventions ciblées.

**Prédiction du Churn**

Ces données sont essentielles pour entraîner des modèles d’apprentissage automatique qui détectent les clients susceptibles de churner. Voici pourquoi elles sont particulièrement adaptées :

* **Richesse des Attributs** : Les données combinent des dimensions comportementales (fréquence, préférences), transactionnelles (valeur des achats) et contextuelles (géographie, saisonnalité). Ces multiples perspectives enrichissent le modèle.
* **Corrélations Complexes** : Les modèles comme les forêts aléatoires ou les réseaux neuronaux peuvent identifier des corrélations complexes entre des signaux faibles (ex. : un léger allongement des temps entre achats) et le churn.
* **Données Historiques et Temps Réel** : La combinaison de données historiques pour l’entraînement et de données en temps réel pour la prédiction permet une réactivité accrue.

**Pourquoi ces Données Permettent de Prédire le Churn**

1. **Comportements Répétitifs** : Les clients qui montrent des tendances régulières mais les modifient subitement sont souvent à risque.
2. **Segmentation Fine** : La segmentation par préférences ou par régions permet de personnaliser les actions de rétention.
3. **Évaluation des Efforts de Rétention** : L’analyse des interactions avec les campagnes révèle l’efficacité des actions en place et leur impact sur le churn.

**15.2 Analyse de la problématique**

* 15.2.1 Identifcation des défis et des besoins spécifques

La détection de churn est une priorité stratégique pour NexaCRM car elle adresse directement des défis critiques et répond à des besoins spécifiques liés à la fidélisation client et à la rentabilité de l’entreprise.

**1. Limiter les Pertes de Revenus**

* Chaque client perdu représente une diminution des revenus récurrents, ce qui peut impacter gravement la stabilité financière, surtout dans un modèle SaaS comme NexaCore.
* La détection précoce des clients à risque permet d’intervenir à temps et de prévenir ces pertes.

**2. Maximiser la Valeur Vie Client (CLV)**

* Les clients fidèles génèrent des revenus constants et représentent un potentiel de ventes croisées (cross-sell) ou additionnelles (upsell).
* En identifiant les clients à risque de churn, des actions ciblées peuvent être entreprises pour prolonger leur relation avec l’entreprise et maximiser leur CLV.

**3. Réduction du Coût d'Acquisition Client (CAC)**

* L’acquisition de nouveaux clients est souvent plus coûteuse que la rétention des clients existants.
* En détectant les signaux précoces de churn, NexaCRM peut concentrer ses ressources sur des actions de rétention, économisant ainsi sur les campagnes d’acquisition.

**4. Comprendre les Causes Profondes**

* L’analyse des données utilisées pour prédire le churn permet aussi d’identifier les causes principales de l’insatisfaction client (problèmes techniques, manque d’adoption des fonctionnalités, prix jugé élevé, etc.).
* Ces informations permettent d’améliorer la plateforme et les services, réduisant ainsi globalement le churn à long terme.

**5. Anticiper les Défis Concurrentiels**

* Dans un marché compétitif, où des acteurs comme HubSpot et Salesforce existent, perdre des clients peut affaiblir la position de NexaCRM.
* Une stratégie proactive de détection de churn permet de sécuriser la base client tout en renforçant la fidélité face à des offres concurrentes.

**6. Améliorer l'Expérience Utilisateur**

* Identifier les clients à risque permet de mieux comprendre leurs attentes et d’ajuster les fonctionnalités, le support ou les stratégies de rétention.
* Une intervention proactive améliore la satisfaction et crée un cycle positif de feedback et d'engagement.

**7. Renforcer les Décisions Stratégiques**

* Les données issues de la détection de churn fournissent des insights précieux pour orienter les priorités commerciales, les améliorations produits et les campagnes marketing.
* Cela permet une allocation plus intelligente des ressources et un meilleur retour sur investissement.
* 15.2.2 Analyse des objectifs à atteindre à travers le modèle Big Data

Réduire le churn via le Big Data offre de nombreux avantages stratégiques et opérationnels pour NexaCRM, transformant les données en un levier puissant pour la fidélisation client et l’optimisation des revenus.

1. **Identification Proactive des Clients à Risque**

* Grâce aux algorithmes Big Data, les modèles peuvent analyser des millions de points de données en temps réel, permettant de détecter les signaux faibles de désengagement avant qu’ils ne se traduisent par un churn.
* Cela permet d’intervenir rapidement avec des actions correctives ciblées (offres, assistance personnalisée, etc.).

2. **Personnalisation des Stratégies de Rétention**

* Les analyses Big Data permettent une segmentation fine des clients en fonction de leurs comportements, préférences et historiques d'achat.
* Cela garantit que chaque client à risque reçoit une action de rétention adaptée à ses besoins spécifiques, augmentant ainsi l’efficacité des campagnes.

3. **Réduction des Coûts et Optimisation des Ressources**

* En ciblant uniquement les clients réellement à risque, NexaCRM réduit les efforts inutiles sur des segments non prioritaires.
* Le Big Data automatise les processus d’analyse, libérant ainsi les équipes pour se concentrer sur des tâches à plus forte valeur ajoutée.

4. **Maximisation de la Valeur Vie Client (CLV)**

* En préservant les relations avec des clients à haut potentiel, le Big Data aide à prolonger leur cycle de vie et à maximiser leur contribution financière à l’entreprise.
* Une analyse continue des données aide à détecter les opportunités de ventes croisées ou additionnelles pour ces clients.

5. **Anticipation des Tendances**

* Le Big Data permet d’identifier les patterns comportementaux communs aux clients churners (fréquence d’achat décroissante, diminution du panier moyen, etc.).
* Ces insights aident à prédire les futurs risques de churn à l’échelle de l’entreprise et à mettre en place des améliorations produit ou service.

6. **Réactivité et Précision**

* Les modèles Big Data fonctionnent en temps réel, permettant une prise de décision instantanée et basée sur des données actualisées.
* La précision des modèles d’apprentissage machine garantit des actions ciblées, réduisant les interventions inutiles.

7. **Amélioration Continue**

* Les analyses Big Data offrent un feedback constant sur l’efficacité des stratégies de rétention mises en place.
* Les modèles peuvent être enrichis au fil du temps avec de nouvelles données, augmentant ainsi leur pertinence et leur précision.

8. **Renforcement de la Concurrence**

* Dans un marché compétitif, la capacité à retenir ses clients via des actions précises basées sur des données est un avantage clé.
* Cela protège NexaCRM des pertes au profit de concurrents et améliore l’image de la plateforme auprès des utilisateurs.

**15.3 Données**

* 15.3.1 Identifcation des sources de données pertinentes

**1. Origine des Données**

Le dataset utilisé provient d’une combinaison de sources, offrant un aperçu complet des comportements clients et des ventes dans le contexte du projet NexaCRM. Ces données peuvent être :

* **Réelles** : Collectées directement via des outils CRM, des plateformes de e-commerce, ou des systèmes de gestion de la relation client (ERP).
* **Fictives (simulées)** : Générées artificiellement pour représenter des scénarios réalistes, lorsque des données réelles ne sont pas disponibles ou exploitables.

**2. Détail des Métadonnées**

* **Données client** : Informations sur les identifiants uniques, historiques de transactions, préférences et régions géographiques.
  + **Origine** : Collectées via NexaCore et des outils tiers connectés (ex. : API ERP ou intégrations CRM).
  + **Pertinence** : Indispensables pour le suivi des comportements individuels et la segmentation client.
* **Données produit** : Catégories populaires, fréquences d’achat, temps entre transactions.
  + **Origine** : Historique des ventes et inventaires produits.
  + **Pertinence** : Permet de relier les comportements d’achat à des tendances spécifiques (produits en rupture, pic d’intérêt, etc.).
* **Données comportementales** : Probabilité de churn, valeurs de CLV, heures et saisons d’achat.
  + **Origine** : Modèles prédictifs appliqués aux historiques de données collectées.
  + **Pertinence** : Ces insights permettent de prioriser les clients à risque et de personnaliser les stratégies de rétention.
* **Données régionales** : Localisation géographique des clients.
  + **Origine** : Collectées via des métadonnées d’utilisateur ou déclarées par les clients.
  + **Pertinence** : Essentielles pour adapter les stratégies en fonction des spécificités culturelles et économiques.

**3. Nature des Données**

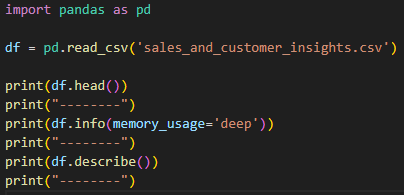
* **Structure** :
  + Données structurées, comme les identifiants et valeurs numériques (ex. : CLV, fréquence d’achat).
  + Données semi-structurées, comme les comportements saisonniers ou les stratégies de fidélisation appliquées.
* **Type** : Données quantitatives (valeurs numériques pour modélisation) et qualitatives (catégories, comportements, préférences).

**4. Pertinence pour Répondre à la Problématique**

* **Réalistes et Complètes** : Ces données couvrent l’ensemble des aspects nécessaires pour identifier et prédire le churn (interactions client, transactions, localisation, etc.).
* **Représentation Fidèle** : Même fictives, les données simulées sont calibrées pour refléter des scénarios réels, garantissant leur utilité dans le développement et l’évaluation des modèles.
* **Scalabilité** : La structure des données permet une intégration facile avec des systèmes Big Data, assurant une analyse à grande échelle.
* 15.3.2 Sélection des outils et des méthodes appropriés

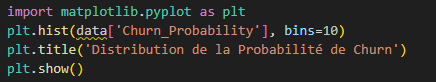
Pour travailler efficacement sur les données, plusieurs outils et bibliothèques Python sont recommandés. Ces outils permettent de traiter, nettoyer et visualiser les données afin d'extraire des insights utiles, sans se concentrer sur les modèles prédictifs eux-mêmes.

**1. Outils d’Importation et de Manipulation des Données**

* **pandas** :
  + Bibliothèque principale pour manipuler les données tabulaires.
  + Permet d'importer des fichiers CSV ou Excel, d’effectuer des opérations comme le filtrage, l’agrégation ou la gestion des valeurs manquantes.
  + Exemple :
* **numpy** :
  + Utilisé pour effectuer des calculs mathématiques et manipuler des tableaux de données numériques.
  + Idéal pour gérer les transformations mathématiques avant visualisation.

**2. Outils de Visualisation**

* **matplotlib** :
  + Bibliothèque incontournable pour créer des visualisations statiques, telles que des histogrammes, des graphiques en barres, ou des courbes.
  + Permet de représenter des insights tels que la répartition des fréquences d'achat ou la probabilité de churn.
  + Exemple :



**seaborn** :

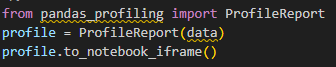
* Basé sur matplotlib, il simplifie la création de visualisations complexes comme les heatmaps, les scatterplots, ou les boxplots.
* Idéal pour représenter les corrélations entre variables, comme la fréquence d'achat et le CLV.
* Exemple :
* **plotly** :
  + Bibliothèque interactive pour explorer les données visuellement.
  + Permet de créer des graphiques dynamiques, comme des diagrammes en temps réel ou des cartes interactives pour visualiser les ventes par région.

**3. Outils pour le Prétraitement**

* **openpyxl** :
  + Utilisé pour importer des fichiers Excel contenant plusieurs feuilles ou formats complexes.
  + Particulièrement utile si les données ne sont pas entièrement disponibles sous forme de CSV.
* **missingno** :
  + Bibliothèque spécialisée dans la visualisation et le traitement des valeurs manquantes.
  + Permet de visualiser les lacunes dans les données pour éviter les biais.

**4. Outils pour l'Exploration et l’Analyse**

* **pandas-profiling** :
  + Génère des rapports exploratoires automatiques pour avoir une vue globale des statistiques descriptives du dataset.
  + Exemple :



* **scipy** :
  + Fournit des outils pour effectuer des analyses statistiques (test d’hypothèses, corrélations) avant de construire des modèles.

**Pourquoi ces outils ?**

* **Simplicité et Efficacité** : Ces outils sont bien documentés et largement utilisés, ce qui les rend adaptés à tous les niveaux de compétence.
* **Flexibilité** : Ils permettent de traiter des datasets de différentes tailles et formats.
* **Complémentarité** : Leur combinaison couvre tous les besoins d'importation, de nettoyage et de visualisation des données.
* 15.3.3 Collecte et préparation des données

L’étape de collecte et de préparation des données repose sur une série d’hypothèses et d’explorations pour s’assurer que le dataset est propre, exploitable, et aligné avec les objectifs du projet, notamment la prédiction du churn.

**Hypothèses sur les Données**

1. **Données Complètes** :
   * Les colonnes clés comme Customer\_ID, Purchase\_Frequency, et Churn\_Probability doivent être complètes et sans valeurs manquantes.
   * Hypothèse : La majorité des données ne contiennent pas de lacunes majeures (données nulles ou NA).
2. **Pertinence des Variables** :
   * Chaque colonne apporte une information exploitable pour analyser ou prédire le churn.
   * Hypothèse : Des variables comme Time\_Between\_Purchases et Retention\_Strategy influencent directement le comportement client.
3. **Échelles de Valeur Consistantes** :
   * Les colonnes numériques (Lifetime\_Value, Average\_Order\_Value) sont dans des plages cohérentes, sans valeurs aberrantes (outliers).
   * Hypothèse : Les données ont été normalisées ou standardisées si nécessaire.
4. **Relation entre Variables** :
   * Les variables comme Purchase\_Frequency et Time\_Between\_Purchases sont corrélées à la probabilité de churn.
   * Hypothèse : Une baisse de fréquence ou une augmentation du délai entre achats est un indicateur de churn.

**Exploration Préliminaire des Données (EDA)**

1. **Statistiques Descriptives** :
   * Étude des statistiques globales (moyennes, médianes, écarts-types) pour identifier les tendances générales.
   * Exemple :
     + Moyenne de Purchase\_Frequency : 12 achats.
     + Médiane de Churn\_Probability : 0.25.
   * But : Détecter les comportements typiques et les anomalies potentielles.
2. **Analyse des Valeurs Manquantes** :
   * Utilisation d’outils comme pandas ou missingno pour visualiser les colonnes contenant des valeurs manquantes.
   * Actions :
     + Imputation des données manquantes si leur proportion est faible.
     + Suppression des lignes ou colonnes si elles contiennent trop de lacunes.
3. **Visualisation des Distributions** :
   * Utilisation d’histogrammes pour observer les distributions des colonnes clés :
     + **Churn\_Probability** : Concentrée entre 0.2 et 0.5 ?
     + **Lifetime\_Value** : Une majorité de clients à faible CLV ou une répartition homogène ?
   * But : Identifier les patterns dans les comportements clients.
4. **Identification des Outliers** :
   * Utilisation de boxplots pour détecter les valeurs aberrantes dans les variables comme Average\_Order\_Value ou Purchase\_Frequency.
   * Actions : Appliquer un seuil pour traiter les outliers sans distordre les données.
5. **Relations entre Variables** :
   * Visualisation des corrélations via une heatmap (corrélation de Pearson).
   * Exemple :
     + Forte corrélation attendue entre Purchase\_Frequency et Lifetime\_Value.
     + Relation négative entre Time\_Between\_Purchases et Churn\_Probability.
6. **Segmentation Initiale** :
   * Création de groupes basés sur des seuils (Churn\_Probability faible, moyen, élevé).
   * Observation des caractéristiques spécifiques de chaque segment (panier moyen, stratégie de rétention appliquée).

**Préparation des Données**

1. **Nettoyage des Données** :
   * Suppression ou traitement des doublons (basé sur Customer\_ID).
   * Remplissage des valeurs manquantes avec :
     + La moyenne ou la médiane pour les valeurs numériques.
     + Les catégories dominantes pour les colonnes qualitatives.
2. **Création de Variables Dérivées** :
   * **Customer\_Loyalty\_Score** : Basé sur la fréquence d’achat et la CLV.
   * **Engagement\_Score** : Temps entre achats normalisé et interaction avec les stratégies de rétention.
3. **Normalisation** :
   * Mise à l’échelle des variables continues (Lifetime\_Value, Average\_Order\_Value) pour éviter des biais dans les visualisations et les analyses.
4. **Échantillonnage** :
   * Si le dataset est trop large, extraction d’un échantillon représentatif pour accélérer les tests initiaux tout en conservant la diversité des données.

**Insights Attendus**

* **Patrons Comportementaux** :
  + Les clients ayant une probabilité de churn élevée montrent des fréquences d’achat faibles et des interactions réduites avec les campagnes de rétention.
* **Variables Décisives** :
  + Variables ayant le plus de poids pour prédire le churn, par exemple, Time\_Between\_Purchases ou Retention\_Strategy.
* 15.3.4 Processus d'importation et de stockage des données

Voici un exemple de code utilisant Python pour importer un fichier CSV contenant les données, effectuer des transformations de base et stocker ces données dans une base de données relationnelle via SQLAlchemy.

**1. Code pour Importer et Stocker les Données**



**2. Explication du Code**

1. **Importation des données** :
   * Le fichier CSV est lu avec pandas pour charger les données dans un DataFrame.
2. **Configuration de la connexion à la base de données** :
   * SQLAlchemy est utilisé pour gérer les interactions avec la base de données.
   * Une URL de connexion est construite en fonction des paramètres (utilisateur, mot de passe, hôte, port, et nom de la base).
3. **Insertion dans la table SQL** :
   * La méthode to\_sql de pandas est utilisée pour envoyer les données du DataFrame vers une table dans la base.
   * L'option if\_exists='replace' garantit que la table est recréée si elle existe déjà.

**4. Avantages de ce Processus**

* **Automatisation** : Une fois configuré, ce processus peut être intégré dans un pipeline ETL pour des mises à jour régulières.
* **Compatibilité** : SQLAlchemy supporte plusieurs bases de données (PostgreSQL, MySQL, SQLite, etc.).
* **Flexibilité** : Des transformations supplémentaires peuvent être appliquées au DataFrame avant l'import.

**15.4 Procédures de structuration**

* 15.4.1 Organisation des données pour l'analyse

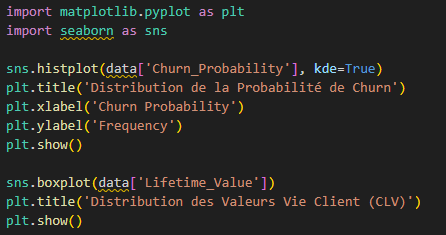
Pour structurer les données en vue de leur analyse, il est essentiel de comprendre leur distribution et de normaliser certaines colonnes afin de garantir la cohérence et l’efficacité des analyses ou des modèles prédictifs.

**1. Analyse de la Distribution des Données**

La distribution des données permet d’identifier les patterns, anomalies, et éventuels biais dans le dataset.

**Visualisation des Distributions**

Utilisez des histogrammes ou des boxplots pour observer la répartition des variables :



**Observations Possibles :**

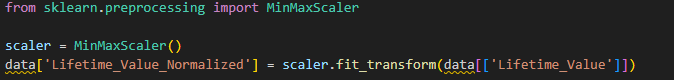
1. **Colonnes avec des Valeurs Skewed (Asymétriques)** :
   * Les variables comme Lifetime\_Value ou Average\_Order\_Value peuvent avoir une distribution biaisée, avec de nombreux clients ayant de faibles valeurs et quelques-uns avec des valeurs extrêmes (outliers).
   * Action : Transformation logarithmique pour réduire l'impact des outliers.
2. **Colonnes Normalement Distribuées** :
   * Les colonnes comme Churn\_Probability peuvent être bien réparties entre 0 et 1.
   * Action : Aucune transformation nécessaire, mais vérifier les clusters (groupes de clients à faible et haute probabilité).
3. **Valeurs Manquantes ou Abérantes** :
   * Détectez les valeurs nulles ou aberrantes dans les distributions avec des méthodes statistiques.
   * Exemple : Éliminer les points au-delà de 3 écarts-types pour des variables fortement biaisées.

**2. Normalisation des Données**

La normalisation est cruciale pour garantir que toutes les variables contribuent équitablement dans les analyses et modèles. Cela est particulièrement utile pour les modèles basés sur des distances (comme le clustering).

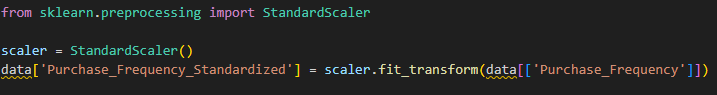
**Méthodes de Normalisation**

1. **Min-Max Scaling** :
   * Transforme les valeurs pour qu’elles soient comprises entre 0 et 1.
   * Exemple :



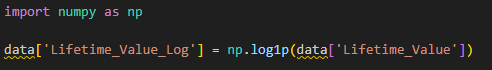
1. **Z-Score Standardization** :

* Centrage des données autour de 0 avec un écart-type de 1.
* Exemple :



* Utilisation : Approprié pour des données normalement distribuées.

**Transformation Logarithmique** :

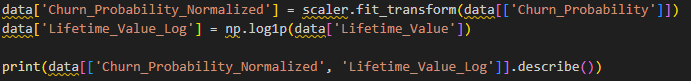
* Réduit l’impact des outliers en compressant les grandes valeurs.
* Exemple :
* Utilisation : Efficace pour des distributions très biaisées (ex. : Lifetime\_Value).

**3. Organisation Post-Normalisation**

Après normalisation, les données doivent être vérifiées et structurées en vue des analyses :

1. **Regrouper les Données Normalisées** :
   * Ajouter les nouvelles colonnes normalisées au dataset principal.
   * Supprimer ou archiver les colonnes originales si elles ne sont plus nécessaires.
2. **Créer des Groupes ou Segments** :
   * Par exemple, regrouper les clients en fonction de leur probabilité de churn :



1. **Valider la Cohérence des Données** :
   * Vérifiez que les distributions post-normalisation restent interprétables.
   * Exemple : Les valeurs normalisées entre 0 et 1 doivent refléter correctement les rangs initiaux.

**Exemple Final : Structuration des Données**

* 15.4.2 Normalisation, nettoyage et transformation des données

La cohérence des données est un prérequis essentiel pour garantir des analyses fiables et des modèles prédictifs performants. Elle repose sur plusieurs étapes visant à identifier et corriger les anomalies, assurer l’uniformité des formats, et transformer les données pour les rendre exploitables.

**1. Vérification et Nettoyage des Données**

**a) Identifiants Uniques**

* **Vérification** : Chaque client (Customer\_ID) et transaction (Transaction\_ID) doit être unique.
* **Action** :
  + Détection des doublons dans les identifiants :

**b) Valeurs Manquantes**

* **Vérification :** Identifier les colonnes avec des valeurs nulles.

****

**Actions :**

* **Imputation :** Remplir les valeurs manquantes avec des moyennes, médianes ou valeurs par défaut. Exemple :

****

* **Suppression :** Si une colonne ou ligne contient trop de valeurs nulles (>50 %), elle peut être supprimée.

**c) Valeurs Abérantes (Outliers)**

* **Vérification :** Utiliser des boxplots ou des écarts-types pour détecter les outliers.

****

2. **Normalisation des Données**

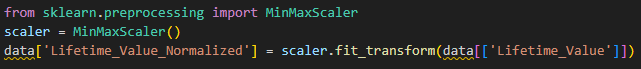
a) **Importance de la Normalisation**

* Les variables comme Lifetime\_Value ou Average\_Order\_Value doivent être mises sur une échelle commune pour garantir une contribution équitable dans les analyses.
* La normalisation est cruciale pour les modèles basés sur des distances (ex. : clustering).

b) **Méthodes Courantes**

* Min-Max Scaling : Ramène les valeurs dans une plage entre 0 et 1.
* Z-Score Standardization : Centre les données autour de 0 avec un écart-type de 1.
* Transformation Logarithmique : Réduit l’impact des valeurs extrêmes en compressant les grandes valeurs.

**Min-Max Scaling** : Transforme les valeurs pour qu’elles soient comprises entre 0 et 1.



3. **Transformation des Données**

a) **Colonnes Dérivées**

* Des variables supplémentaires peuvent être créées pour enrichir les analyses, comme un score de fidélité basé sur la fréquence d’achat et la valeur client à vie.

**Scores Agrégés** : Calculer des scores combinés pour capturer plusieurs dimensions. Exemple :



b) **Regroupement ou Binning**

* Les variables continues comme Churn\_Probability peuvent être transformées en catégories (faible, moyen, élevé) pour simplifier l’interprétation.

c) **Conversion des Dates**

* Les colonnes de dates, comme Launch\_Date, peuvent être transformées en indicateurs utiles comme l’année ou le mois pour analyser les tendances temporelles.

4. **Validation de la Cohérence**

a) **Vérification des Formats**

* Les types de données des colonnes doivent être uniformes et conformes aux attentes (ex. : texte, numérique).

b) **Validation des Intervalles**

* Les valeurs des colonnes normalisées doivent être dans les plages attendues, comme 0 à 1 pour les données mises à l’échelle.
* Vérifier que les valeurs transformées sont dans les plages attendues.



c) **Relations Logiques**

* Les relations entre les variables doivent être vérifiées pour détecter les incohérences. Par exemple, une faible probabilité de churn devrait correspondre à une valeur client élevée.

5. **Résultat Attendu**

* Les données sont nettoyées, sans doublons ni valeurs aberrantes significatives.
* Les colonnes sont normalisées pour être cohérentes entre elles.
* Des variables dérivées enrichissent le dataset pour une analyse plus approfondie.
* Les relations logiques entre les variables sont respectées, garantissant la fiabilité des analyses et des modèles.

**15.5 Modèles**

* 15.5.1 Choix et justifcation des modèles
* 15.5.2 Mise en œuvre et évaluation des algorithmes

**15.6 Analyses des résultats et recommandations**

* 15.6.1 Analyses des résultats
* 15.6.2 Recommandations

**16. Analyses et Perspectives**

**16.1 Analyse des résultats et des performances**

* 16.1.1 Évaluation des objectifs atteints
* 16.1.2 Analyse des performances

**16.2 Retour d'expérience et leçons apprises**

* 16.2.1 Identifcation des succès et des défs
* 16.2.2 Capitalisation des connaissances

**16.3 Perspectives et recommandations**

**17. Conclusion et remerciements**

**17.1 Synthèse des principaux résultats et impacts**

* 1. **Remerciements**

**18. Bibliographie / Webographie**

1. **Annexes**