Rapport technique : Prétraitement de données et Modélisation du Churn

1. Objectif du projet

L'objectif est de transformer une base de données client brute en un dataset exploitable pour prédire le départ des clients (**Churn**). Le pipeline garantit la suppression des biais et l'optimisation des performances pour un futur déploiement.

2.Architecture Technique (Modularité) :

`utils.py` : Stockage des fonctions de calcul (parsing d'IP, calcul d'âge).

`preprocessing.py` : Script principal orchestrant le chargement, le nettoyage, l'entraînement et l'optimisation.

3. Méthodologie de Preprocessing

Problème identifié	Solution appliquée
Valeurs manquantes (Âge)	Imputation croisée utilisant AgeCategory, complétée par un algorithme KNNImputer (5 voisins).
Données aberrantes	SupportTicketsCount et SatisfactionScore corrigés (remplacement des -1 par 0/médiane) et plafonnés par clipping.

Formats de dates	Parsing robuste de `RegistrationDate` et conversion en variables de durée (`FirstPurchaseDaysAgo`, `CustomerTenureDays`).
Données brutes (IP)	Extraction de sous-features via `LastLoginIP` (version, premier octet, statut privé/public).
Variables inutiles	Suppression des constantes (`NewsletterSubscribed`) et des identifiants (`CustomerID`).
Data Leakage (Biais)	**Suppression critique** des variables `Recency`, `TenureRatio`, `CustomerType` et `RFMSegment` qui causaient un score artificiel de 100%.
Déséquilibre de classe	Application de la technique **SMOTE** pour équilibrer la classe minoritaire (Churners) après le split.

4. Architecture du Modèle et Optimisation

- **Algorithme :** Random Forest Classifier.
- **Encodage :** One-Hot Encoding pour transformer les variables catégorielles en vecteurs numériques.
- **Mise à l'échelle :** `StandardScaler` appliqué sur les features pour normaliser les données avant l'entraînement.
- **Optimisation :** Utilisation d'**Optuna** pour la recherche d'hyperparamètres (`n_estimators`, `max_depth`) sur 10 itérations (trials).
- **Répartition :** 80% Entraînement / 20% Test avec **stratification** pour conserver la proportion de Churn.

5. Analyse des Résultats

Après correction du Data Leakage (fuite de données), les performances sont passées d'un score suspect de 1.00 à un score réaliste et robuste :

- **Meilleur Score F1 (Optuna) : 0.9520 (95,2%)**.
- **Variables Clés :** L'importance des variables montre que `FirstPurchaseDaysAgo` et `PreferredMonth` sont les principaux moteurs de la prédiction.
- **Diagnostic :** Le modèle est désormais capable de généraliser sur de nouvelles données sans se baser sur des variables "triches".

  .