



Cahier des Charges du Projet

Projet : Customer Segmentation & Churn Prediction Using Machine Learning

1. Contexte

*.zip

Les entreprises doivent mieux comprendre et anticiper les comportements de leurs clients pour améliorer leur fidélisation et réduire le churn (départ client). Ce projet vise à :

- Regrouper les clients en segments homogènes via l'analyse de leurs comportements.
- Prédire le risque de churn grâce à des algorithmes classiques de machine learning.

Le travail portera exclusivement sur des méthodes **classiques** de machine learning supervisé et non supervisé.

2. Objectifs

- **Customer Segmentation** : Segmenter les clients en groupes similaires selon leurs données comportementales en utilisant des algorithmes de clustering (K-Means, Hierarchical Clustering, DBSCAN).
- **Churn Prediction** : Construire des modèles de classification pour prédire si un client est susceptible de partir ou de rester (Logistic Regression, Decision Tree, Random Forest, SVM, etc.).

3. Données

- **Sources** : Datasets publics (exemples : Mall Customer Dataset, Telco Customer Churn Dataset).
- **Variables attendues** :
 - Informations clients (âge, sexe, revenu, profession, type d'abonnement, etc.)
 - Historique d'achats ou d'utilisation de service
 - Indicateur de churn (pour la partie classification)

4. Méthodologie du Projet

Le projet se déroulera selon 4 grandes étapes réparties sur 4 semaines.



Plan des 4 Semaines :

Semaine 1 : Exploration et Préparation des Données

- Recherche, choix et téléchargement des datasets.
- Analyse exploratoire des données (EDA) pour comprendre les variables.
- Nettoyage des données : traitement des valeurs manquantes, doublons, incohérences.
- Transformation des variables (encodage des catégories, normalisation/standardisation).
- Premiers graphiques descriptifs pour identifier les patterns.

Livrable : Jeux de données propres et un premier rapport d'exploration.

Semaine 2 : Segmentation Clients (Clustering)

- Application des techniques de clustering :
 - K-Means avec choix optimal de K (méthode du coude, silhouette score).
 - Hierarchical Clustering avec dendrogramme.
 - DBSCAN pour tester des clusters de formes complexes.
- Visualisation 2D/3D des clusters (PCA, t-SNE).

- Interprétation des segments clients : caractéristiques dominantes de chaque groupe.

Livrable : Visualisations des clusters et analyse des groupes clients.

Semaine 3 : Prédiction du Churn (Classification)

- Séparation des données en train/test sets.
- Implémentation de plusieurs modèles de classification :
 - Logistic Regression, Decision Tree, Random Forest, SVM, k-NN.
- Validation croisée et recherche d'hyperparamètres optimaux.
- Évaluation des modèles avec des métriques :
 - Accuracy, Precision, Recall, F1-score, ROC-AUC.
- Analyse de l'importance des variables pour la prédiction de churn.




Livrable : Modèle final de prédiction du churn avec évaluation complète.



Semaine 4 : Synthèse, Rapport et Présentation

- Finalisation des notebooks (mise au propre, ajout de commentaires clairs).
- Rédaction du rapport final structuré comprenant :
 - Introduction, méthodologie, résultats, discussion, conclusion.
- Proposition de recommandations basées sur l'analyse (ex : ciblage marketing des segments sensibles au churn).
- Préparation d'une présentation orale (si demandée).

Livrable : Rapport écrit du projet, fichiers code complets et présentation.

5. Livrables attendus

-  Jeux de données nettoyés.
-  Notebooks Python propres et commentés.
-  Rapport final détaillé.

-  Graphiques de visualisation pour la segmentation et l'évaluation des modèles.
-  Présentation synthétique des résultats (si exigée).

6. Contraintes

- Utilisation exclusive d'algorithmes **classiques** de machine learning (pas de deep learning).
- Respect des bonnes pratiques : validation croisée, tuning d'hyperparamètres, analyse critique des résultats.
- Travail régulier réparti sur les 4 semaines.

7. Outils et Technologies

- **Langage** : Python
- **Librairies** :
 - Pandas, NumPy pour la manipulation des données
 - Scikit-learn pour le machine learning
 - Seaborn, Matplotlib pour la visualisation
 - Imbalanced-learn (en cas de classes déséquilibrées)