## **Choisir un des 2 projets**

## **Sujet 1 : Prédiction du risque d’abandon scolaire**

### **Contexte**

Vous travaillez dans une structure éducative qui souhaite identifier les étudiants à risque d’abandon afin de mettre en place des actions de soutien. Un jeu de données a été constitué à partir des informations académiques et sociales des étudiants.

### **Objectif**

Construire un modèle de classification permettant de prédire si un étudiant risque ou non d’abandonner ses études.

### **Jeu de données fourni**

Chaque ligne correspond à un étudiant, avec les colonnes suivantes :

* Age
* Sexe
* Taux\_presence (en %)
* Nombre\_retards
* Note\_moyenne (sur 20)
* Situation\_familiale (Célibataire, Marié, Enfants à charge, Divorcé)
* Abandon (1 = a abandonné, 0 = n’a pas abandonné)

## **Sujet 2 : Prédiction de réponse à une campagne marketing**

### **Contexte**

Une entreprise de e-commerce cherche à optimiser ses campagnes marketing en ciblant les clients les plus susceptibles d’y répondre.

### **Objectif**

Développer un modèle permettant de prédire si un client répondra positivement à une campagne marketing à partir de ses comportements d’achat et ses données socio-démographiques.

### **Jeu de données fourni**

Chaque ligne représente un client avec :

* Age
* Sexe
* Revenu\_annuel (en euros)
* Temps\_passe\_sur\_site (en minutes)
* Achats\_en\_ligne (au cours des 6 derniers mois)
* Reponse\_campagne (1 = a répondu, 0 = n’a pas répondu)

## **Objectifs pédagogiques**

* Maîtriser les étapes d’un projet de classification supervisée.
* Comprendre et appliquer les métriques de performance (accuracy, precision, recall, f1-score).
* Utiliser la validation croisée et le GridSearch pour ajuster les hyperparamètres.
* Expérimenter plusieurs algorithmes vus en cours (KNN, arbre de décision, régression logistique).
* **Interpréter les résultats obtenus de façon critique.**

## **Consignes Data-science**

Appliquer et faire apparaître avec soin les 4 grandes étapes de la méthode CRISP-DM tel que vu en cours:

1. Compréhension et explication du besoin (mesures et métriques)
2. Compréhension de la donnée et nettoyage
3. Modélisation
4. Validation (mesures et métriques)

## **Approche technique minimale attendue**

1. **Prétraitement des données** : gérer les variables catégorielles, vérifier les valeurs manquantes, explorer la distribution des variables.
2. **Exploration des corrélations** et visualisation des données (PCA ?).
3. **Choix d’un ou plusieurs modèles de classification** parmi ceux étudiés (KNN, arbre de décision, régression logistique, KMeans).
4. **Évaluation des performances** : utilisez accuracy, recall, precision, f1-score et matrice de confusion
5. **Recherche des meilleurs hyperparamètres** avec GridSearch et validation croisée.
6. **Analyse critique des résultats** :  
   * Quels modèles fonctionnent le mieux ? Pourquoi ?
   * Quelles sont les limites du modèle ? Des données ?
   * Quelle interprétation peut-on faire des métriques obtenues ?
   * Que signifie une bonne ou mauvaise précision dans ce contexte ?

**Attention : la partie interprétation comptera pour 50% de la note finale.** Il ne s’agit pas uniquement d’écrire du code mais de démontrer votre compréhension du problème.

## **Livrable attendu**

Un **notebook Jupyter unique** contenant :

* Le code complet, propre et commenté.
* Des visualisations claires et lisibles.
* Les résultats des modèles testés.
* Vos **interprétations argumentées**.

Le travail est à faire en binôme et présenté sur un lien github

Nom du fichier {nomDuProjet}\_<NomsDuBinomes>.ipynb.