TEST DATA SCIENCE

Ce test vise à évaluer vos compétences en analyse de données et votre capacité à interpréter et tirer des insights à partir d'ensembles de données complexes. Vous trouverez ci-dessous quatre questions qui couvrent divers aspects du processus d'analyse de données, allant de la description initiale d'un ensemble de données à des questions plus avancées liées à la modélisation prédictive.

Le jeu de données est fourni en annexe.

Data	
ClientID	Identifiant unique du client
taxable income amount	Indique si le client a déclaré un revenu supérieur à 50 000 $
Les autres colonnes contiennent des informations relatives au client	

Présentez vos résultats de manière claire et concise et expliquez vos choix de méthodes et d'approches.

QUESTIONS

0. Fournissez une analyse descriptive de l'ensemble de données fourni, en mettant en évidence ses caractéristiques clés. Exemples :

- Caractéristiques statistiques : Donnez un aperçu des statistiques de base telles que la moyenne, la médiane, l'écart type, etc., pour les variables numériques.
- Caractéristiques catégorielles : Pour les variables catégorielles, décrivez les différentes catégories et leur répartition dans l'ensemble de données.
- Présence de données manquantes : vérifiez la présence de données manquantes dans certaines colonnes
- Corrélations intéressantes : corrélations qui pourraient être pertinentes ou intéressantes du point de vue de l'objectif de l'analyse

1. Quelle est la distribution de l'âge par rapport à la valeur du montant du revenu imposable (taxable income amount) ?

2. Utilisez des techniques de regroupement (clustering) pour identifier et décrire différents segments de clients dans l'ensemble de données.

3. Mettez en œuvre un modèle prédictif pour identifier des clients qui n'ont pas déclaré un revenu supérieur supérieurs à 50 000 $ (taxable income amount=0), mais qui ont des caractéristiques très similaires à ceux qui ont déclaré gagner plus de 50 000 $ (taxable income amount=1).