Test pour Stagiaire Data Scientist

Durée totale : 2 heures

Instruction :

- Ce test doit être réalisé en utilisant **Python** dans un notebook **Jupyter**. Assurez-vous de bien documenter votre code, vos choix méthodologiques, et d'inclure des commentaires explicatifs pour chaque étape importante du test.
- Veillez à structurer votre notebook de manière claire et logique, en séparant les différentes parties du test en sections distinctes. Cela facilitera la lecture et l'évaluation de votre travail.

Partie 1 : Théorie Rapide (10 minutes)

Validation croisée vs Split train/test :

- Expliquez brièvement la différence entre la validation croisée et le split train/test.

Modèles ensemblistes :

- Définissez ce qu'est un modèle ensembliste et donnez un exemple.

Partie 2 : Analyse de Données (20 minutes)

- Dataset : Titanic
- Tâches :

  Réalisez une visualisation pour examiner la distribution de l'âge des passagers.

  Présentez une visualisation montrant la relation entre la classe de passager et le taux de survie.

Partie 3 : Prétraitement des Données (20 minutes)

Gestion des valeurs manquantes :

- Gérez les valeurs manquantes pour l'âge et le tarif. Documentez la méthode choisie et justifiez votre choix.

Encodage des variables catégorielles :

- Choisissez et appliquez une méthode d'encodage pour la variable 'Sexe'.

Partie 4 : Modélisation (30 minutes)

Création d'un modèle de classification :

- Construisez un modèle de classification pour prédire la survie des passagers. Justifiez le choix du modèle.

Validation croisée :

- Effectuez une validation croisée pour évaluer la performance du modèle. Expliquez votre approche.

Partie 5 : Optimisation (30 minutes)

Réglage des hyperparamètres :

- Réalisez le réglage des hyperparamètres de votre modèle de classification. Expliquez votre démarche et les critères de sélection.