Devoir : Modélisation sur un jeu de données de votre choix

Objectif du devoir

Dans ce devoir, vous utiliserez un jeu de données de votre choix afin de réaliser une analyse de modélisation prédictive en utilisant des algorithmes de machine learning. Vous devrez synthétiser vos résultats dans un rapport détaillé.

Consignes Générales

- Travail en groupe
- Durée : 1 à 2 semaines
- Livrables :
  - Un fichier Python contenant tout le code exécuté pour l'analyse et la modélisation
  - Un rapport détaillé comprenant :
    - Une présentation du dataset
    - Les étapes de préparation des données
    - Une analyse des performances du modèle
    - Une interprétation des résultats

Choix du Dataset

Vous êtes libre de choisir le jeu de données qui vous intéresse, en veillant à ce qu'il soit adapté à une tâche de modélisation (classification ou régression).

Choix des Variables pour la Modélisation

Une étape cruciale dans ce processus est le choix des variables pertinentes pour la modélisation. Voici des suggestions sur comment procéder :

1. Identifier la variable cible (dépendante) :

Variable cible c'est quoi ? : La **variable cible** (ou variable dépendante) est la variable que vous souhaitez prédire ou expliquer à l'aide de vos autres variables. Le choix de la variable cible dépend du problème que vous souhaitez résoudre et des objectifs de votre analyse.

- **Classification :** Choisissez une variable binaire (ex. : acheter ou non).
- **Régression** : Choisissez une variable continue (ex. : montant d'achat, durée d'appel, etc.).

2. Sélectionner les variables explicatives (indépendantes) :
- Examen des variables disponibles et choix des plus pertinentes. Inclure des variables numériques (ex. : âge, montant d'achat, etc.) et des variables catégorielles (ex. : pays, genre).

3. Vérification de la qualité des données :
- Gestion des valeurs manquantes.
- Identification et traitement des valeurs aberrantes.
- Normalisation/standardisation des variables si nécessaire.

Modélisation

1. Diviser les données: Divisez votre dataset en un ensemble d'entraînement (80%) et un ensemble de test (20%).

2. Choisir un modèle : Vous pouvez utiliser la régression logistique ou la forêt aléatoire. Le choix dépendra du type de variable cible (binaire, continue).

3. Entraînement du modèle : Utilisez `LogisticRegression` ou `RandomForestClassifier`/`RandomForestRegressor` de `sklearn`.

4. Évaluation du modèle : Sur l'ensemble de test, évaluez votre modèle avec des métriques appropriées. Performance du modèle avec la courbe ROC et le calcul de l'AUC (air sous la courbe compris entre 0 et 1).

Interprétation des Résultats

Analyse des performances et discussion des résultats obtenus. Évaluez l'impact des variables et explorez des pistes pour améliorer le modèle.

Critères d'Évaluation

- Pertinence des choix de variables et des résultats de modélisation – 40%
- Clarté et structuration du rapport – 30%
- Interprétation des résultats et des performances du modèle – 20%
- Originalité et rigueur dans l'approche – 10%