TI508A-Introduction au Machine Learning

Projet : analyse des impacts environnementaux des services numériques

Consignes :

⇒ Ce projet doit être réalisé en groupes.
⇒ Il est impératif de répondre à toutes les questions du projet.
⇒ Les notions non vues en cours seront notées en **BONUS**

Acquis d'apprentissage :

- Comprendre les algorithmes de ML
- Apprendre à évaluer des modèles
- Apprendre à interpréter et communiquer les résultats
- Comprendre les enjeux énergétiques associés à l'utilisation de différents services numériques

Partie 1 : Génération des données artificielles

Dans cette partie, vous allez générer artificiellement les données sur lesquelles vous allez appliquer les algorithmes de ML. Pour ce faire, voici le prompt à utiliser dans ChatGPT pour générer le code Python vous permettant de générer votre jeu de données.
« Vous allez générer un code Python qui permet de générer un jeu de données artificielles sur l'empreinte carbone de 25 services numériques caractérisés par un ensemble de variables pour mesurer leur impact environnemental. Le jeu de données est décrit par le tableau suivant.

Service numérique : Nom du service (ex : "Streaming vidéo", "Recherche sur internet")
Consommation énergétique (kWh) : Énergie consommée par le service (en kilowattheure)

Structure des variables

Variable	Description
Service numérique	Nom du service (ex : "Streaming vidéo", "Recherche sur internet")
Consommation énergétique (kWh)	Énergie consommée par le service (en kilowattheure)
Nombre d'utilisateurs	Nombre moyen d'utilisateurs du service
Type d'énergie utilisée	Renouvelable ou Non-renouvelable

Variable	Description
Émissions de CO2 (kg)	Émissions de CO2 associées (calculées ou estimées)
Pays	Pays d'opération principal (ex : France, USA, Allemagne)

Voici la liste des 25 services numériques proposés

1. Streaming vidéo
2. Recherche sur internet
3. Services de messagerie instantanée
4. Réseaux sociaux
5. Hébergement de sites web
6. E-commerce
7. Plateformes d'apprentissage en ligne
8. Transactions blockchain
9. Cloud computing
10. Stockage de données en ligne
11. Jeux en ligne
12. Systèmes de navigation GPS
13. Télétravail (visioconférences)
14. Streaming de musique
15. Services bancaires en ligne
16. IoT (Internet des objets)
17. Sauvegarde automatique de données
18. Partage de fichiers
19. Applications mobiles de livraison
20. Édition collaborative de documents (ex : Google Docs)
21. Email
22. Téléchargement de logiciels
23. Plateformes de vidéos courtes (ex : TikTok)
24. Réalité virtuelle en ligne
25. Services de santé numérique

Le jeu de données devrait être enregistrés dans un fichier .csv »

Partie 2 : Analyse des données

1. Régression (Prédiction de la quantité de CO2 émise)

Objectif : Prédire la quantité de CO2 émise en fonction de la consommation énergétique totale (Consommation totale (kWh)) et du nombre d'utilisateurs.
Modèle à utiliser : Régression linéaire.

Étapes :

- **Étape 1** : **Préparation des données**
  - Sélectionner les variables explicatives (features) : Consommation énergétique (kWh) et Nombre d'utilisateurs. La variable cible (label) sera Emissions CO2 (kg).
- **Étape 2** : **Nettoyage des données**
  - Vérifier les données manquantes ou aberrantes (outliers) et les traiter (soit en les supprimant, soit en les remplaçant par des valeurs moyennes ou médianes).
- **Étape 3** : **Séparation des données**
  - Diviser le jeu de données en ensemble d'entraînement (70 %) et ensemble de test (30 %).
- **Étape 4** : **Choix du modèle de régression**
  - Utiliser un modèle de régression linéaire pour modéliser la relation entre la consommation énergétique, le nombre d'utilisateurs et les émissions de CO2.
- **Étape 5** : **Entraînement du modèle**
  - Entraîner le modèle sur l'ensemble d'entraînement pour ajuster les poids en fonction des variables explicatives.
- **Étape 6** : **Évaluation du modèle**
  - Utiliser l'ensemble de test pour évaluer les performances du modèle avec des métriques telles que l'Erreur Quadratique Moyenne (MSE) ou le Coefficient de Détermination (R^2).

2. KNN (Classification de l'efficacité énergétique)

Objectif :
Classer les services numériques selon leur empreinte carbone (faible, moyenne, élevée) sur la base des émissions de CO2.

Étapes :

- Étape 1 : Préparation des données
  - Diviser la variable Emissions CO2 (kg) en catégories : "faible", "moyenne", "élevée". Par exemple, vous pouvez définir des seuils en fonction de quantiles.

- Étape 2 : Séparation des données
    - Diviser les données en ensemble d'entraînement et de test.
- Étape 3 : Choix du modèle KNN
    - Utiliser l'algorithme K-Nearest Neighbors (KNN) pour classifier les services. Sélectionner une valeur de K (nombre de voisins) adaptée. Essayer plusieurs valeurs de K pour optimiser la performance.
- Étape 4 : Entraînement du modèle
    - Entraîner le modèle KNN sur les variables explicatives : Consommation énergétique (kWh) et Nombre d'utilisateurs. La variable cible sera la catégorie d'empreinte carbone (faible, moyenne, élevée).
- Étape 5 : Évaluation du modèle
    - Utiliser la matrice de confusion et d'autres métriques de classification telles que la précision (accuracy), le rappel (recall), et la précision (precision) pour mesurer la qualité du modèle.

3. K-means et CAH (Segmentation des services numériques)

Objectif :

Segmenter les services numériques en groupes ayant des profils similaires en termes de consommation énergétique et de nombre d'utilisateurs.

En utilisant K-means

- **Étape 1 : Préparation des données**
    - Utiliser les variables Consommation énergétique (kWh) et Nombre d'utilisateurs.
- **Étape 2 : Normalisation des données**
    - Appliquer une normalisation ou une standardisation des variables pour s'assurer qu'elles ont des échelles comparables.
- **Étape 3 : Choix du nombre de clusters K**
    - Utiliser la méthode du "coude" (Elbow method) pour déterminer le nombre optimal de clusters K en traçant l'inertie (somme des distances intra-cluster) en fonction de K.
    - Vous aurez un bonus si vous utilisez d'autres méthodes.

- **Étape 4 : Exécution de K-means**
    - Appliquer l'algorithme K-means pour former des clusters en fonction des similarités entre les services numériques.
- **Étape 5 : Interprétation des clusters**
    - Analyser les groupes formés et interpréter les résultats pour comprendre quels types de services numériques partagent des caractéristiques similaires.

En utilisant CAH (Classification Ascendante Hiérarchique)

- **Étape 1 : Préparation des données**
    - Sélectionner les variables Consommation énergétique (kWh) et Nombre d'utilisateurs.
- **Étape 2 : Normalisation des données**
    - Normaliser les variables comme pour K-means.
- **Étape 3 : Construction du dendrogramme**
    - Utiliser un dendrogramme pour visualiser la hiérarchie des regroupements, et déterminer un nombre raisonnable de clusters en fonction de la coupure du dendrogramme.
- **Étape 4 : Formation des clusters**
    - Appliquer la CAH pour former des clusters basés sur la hiérarchie.
- **Étape 5 : Interprétation des clusters**
    - Comparer les clusters obtenus par la CAH avec ceux de K-means et analyser les différences.

4. Validation des modèles

Objectif :

Valider la qualité des modèles de régression, de classification et de clustering pour s'assurer qu'ils sont généralisables aux nouvelles données.

Étapes :

- Validation pour la régression :

- Cross-validation : Utiliser la validation croisée (K-fold cross-validation) pour évaluer les performances de votre modèle de régression en testant plusieurs découpages de données.
- Métriques : Évaluer les performances avec des métriques comme l'Erreur Quadratique Moyenne (MSE), l'Erreur Absolue Moyenne (MAE) ou le Coefficient de Détermination (R^2).

- Validation pour KNN :
  - Matrice de confusion : Analyser la matrice de confusion pour identifier les erreurs de classification (faux positifs, faux négatifs).
  - Cross-validation : Effectuer une validation croisée pour s'assurer de la robustesse du modèle KNN.

- Validation pour K-means et CAH :
  - Faire une recherche sur les différents indices de validation et les appliquer pour analyser la qualité des clusters obtenus,
    ex : l'indice de Silhouette : calculer le score de silhouette pour mesurer la cohésion et la séparation des clusters (score de silhouette entre -1 et 1, avec des valeurs proches de 1 indiquant une meilleure séparation des clusters).

A vous de jouer !