ml-brief

0.0.1 Brief : Projet de Machine Learning - Régression ou Classification

Pour ce brief, vous avez le choix entre réaliser un modèle de régression ou un modèle de classification. Voici une description détaillée des deux options avec des objectifs et des questions pour vous guider.

Option 1 : Régression **Objectif :** Prédire une variable continue à partir de plusieurs variables explicatives.

Étapes et questions guidantes : 1. **Comprendre les données** - Quelles sont les variables explicatives et la variable cible ? - Y a-t-il des valeurs manquantes dans le dataset ? Comment allez-vous les traiter ?

2. **Exploration des données**
   - Quels sont les résumés statistiques des variables (moyenne, médiane, écart-type, etc.) ?
   - Y a-t-il des corrélations entre les variables explicatives et la variable cible ?
   - Y a-t-il des variables qui nécessitent une transformation (ex. : normalisation) ?
3. **Préparation des données**
   - Comment allez-vous diviser vos données entre un ensemble d'entraînement et un ensemble de test ?
   - Quelles sont les techniques de prétraitement des données que vous allez utiliser (imputation, encodage des variables catégorielles, etc.) ?
4. **Modélisation**
   - Quel modèle de régression allez-vous utiliser (régression linéaire, forêts aléatoires, etc.) ?
   - Comment allez-vous évaluer la performance de votre modèle (MSE, RMSE, MAE, etc.) ?
   - Avez-vous besoin de réaliser une validation croisée pour optimiser les hyperparamètres ?
5. **Interprétation des résultats**
   - Quels sont les coefficients des variables dans votre modèle (si applicable) ?
   - Quelles variables ont le plus d'influence sur la variable cible ?

Option 2 : Classification **Objectif :** Prédire une classe ou une catégorie à partir de plusieurs variables explicatives.

Étapes et questions guidantes : 1. **Comprendre les données** - Quelles sont les variables explicatives et la variable cible ? - Y a-t-il des valeurs manquantes dans le dataset ? Comment allez-vous les traiter ?

2. **Exploration des données**
   - Quels sont les résumés statistiques des variables ?

- Y a-t-il des corrélations entre les variables explicatives et la variable cible ?
- Y a-t-il des classes déséquilibrées ? Si oui, comment allez-vous les gérer ?

3. **Préparation des données**
   - Comment allez-vous diviser vos données entre un ensemble d'entraînement et un ensemble de test ?
   - Quelles sont les techniques de prétraitement des données que vous allez utiliser (imputation, encodage des variables catégorielles, etc.) ?
4. **Modélisation**
   - Quel modèle de classification allez-vous utiliser (régression logistique, SVM, kNN, forêts aléatoires, etc.) ?
   - Comment allez-vous évaluer la performance de votre modèle (accuracy, précision, rappel, F1-score, AUC, etc.) ?
   - Avez-vous besoin de réaliser une validation croisée pour optimiser les hyperparamètres ?
5. **Interprétation des résultats**
   - Quelle est la matrice de confusion de votre modèle ?
   - Quelles variables ont le plus d'influence sur la variable cible ?

0.0.2 Livrables attendus

1. **Rapport écrit :** Un document détaillant chaque étape du projet, incluant l'analyse exploratoire des données, la préparation des données, la modélisation et l'évaluation du modèle, ainsi que l'interprétation des résultats.
2. **Code :** Le code utilisé pour réaliser l'analyse et entraîner le modèle, idéalement dans un notebook Jupyter.