Projet : Conception et Évaluation de Modèles Prédictifs à partir de Données Réelles

Objectifs pédagogiques

- Choisir ou collecter un jeu de données adapté
- Nettoyer et prétraiter les données (gestion des valeurs manquantes, encodage, normalisation)
- Construire et évaluer :
  - Un modèle de régression
  - Un modèle de classification binaire
  - Un modèle de classification multiclasse
- Créer des visualisations claires et pertinentes
- Présenter les résultats et les interprétations

Tâches du projet (par phases)

Phase 1 — Choix et préparation du jeu de données

- Choisir un jeu de données contenant :
  - Au moins une variable cible continue (pour la régression)
  - Une variable cible binaire (0/1)
  - Une variable cible multiclasse (≥ 3 classes)

  (Ou 3 jeux de données séparés si nécessaire — Kaggle, UCI, etc.)
- Réaliser :
  - Analyse exploratoire (EDA)
  - Traitement des valeurs manquantes
  - Encodage des variables catégorielles
  - Mise à l'échelle des variables numériques si nécessaire
- Séparer les données en jeu d'entraînement et de test

Phase 2 — Régression

- Choisir 2 ou 3 modèles (ex. : Régression Linéaire, Arbre de Décision, Random Forest)
- Évaluer les modèles avec : MAE, MSE, RMSE et R^2
- Produire :
  - Des courbes d'apprentissage
  - Un graphique valeurs prédites vs valeurs réelles

Phase 3 — Classification binaire

- Sélectionner une variable cible binaire (ex. : « malade » oui/non)
- Entraîner 2 ou 3 modèles (ex. : Régression Logistique, Random Forest, SVM)
- Évaluer avec : Accuracy, Précision, Rappel, F1-score, ROC AUC
- Produire :

  - Matrice de confusion
  - Courbe ROC et courbe Précision-Rappel

Phase 4 — Classification multiclasse

- Choisir une variable cible ayant 3 classes ou plus (ex. : espèce de fleur, type d'animal)
- Entraîner 2 ou 3 modèles (ex. : Arbre de Décision, Random Forest, KNN)
- Évaluer avec : Accuracy, Précision/Rappel/F1 par classe, rapport de classification
- Produire :
  - Matrice de confusion (3×3 ou plus)
  - Diagrammes en barres des métriques par classe

Phase 5 — Visualisation et Présentation

- Créer des visualisations claires avec Matplotlib / Seaborn / Plotly :
  - Carte de corrélation des variables
  - Pair plots / scatter plots pour l'EDA
  - Courbes apprentissage / validation
  - Graphique d'importance des variables
  - Graphiques comparant les performances des modèles
- Rédiger le code proprement dans un Notebook Jupyter ou script Python
- Préparer une présentation orale de 5 à 10 minutes incluant :
  - Jeu de données et prétraitement
  - Modèles choisis et justification
  - Résultats et comparaisons
  - Visualisations et interprétation des résultats

Livrables attendus

- Code source commenté et bien structuré (Notebook ou .py)
- Visualisations et graphiques
- Rapport final (3–5 pages) résumant la méthode, les résultats et la discussion