Projet Machine Learning

Classification & Régression

Nom : **Baati Yossri**
Travail : **Individuel**
Date : ____ / ____ / _____

1. Introduction Générale

Ce projet a pour objectif d'appliquer les concepts fondamentaux du Machine Learning vus en cours : exploration des données, prétraitement, classification supervisée, régression, évaluation des modèles et mise en place de fonctions de prédiction. Deux jeux de données ont été étudiés : le dataset Breast Cancer (classification) et le dataset Weather (régression). L'objectif principal est de comprendre et de maîtriser les différentes étapes d'un pipeline complet d'apprentissage automatique.

2. Partie 1 — Classification : Breast Cancer

Exploration et Visualisation

L'exploration du dataset Breast Cancer a permis d'analyser la distribution des variables, d'identifier les valeurs manquantes et de visualiser les corrélations entre les attributs. Des outils comme les heatmaps, histogrammes et scatterplots ont été utilisés pour comprendre la structure des données.

Prétraitement

Les étapes majeures du prétraitement ont inclus : la normalisation (StandardScaler), la séparation des données en ensembles d'entraînement et de test, et l'encodage éventuel des variables catégorielles. Ces étapes garantissent une meilleure qualité d'apprentissage.

Modélisation et Évaluation

Plusieurs modèles ont été testés : KNN, SVM, et Régression Logistique. Les performances ont été évaluées via les métriques classiques : Accuracy, Precision, Recall et F1■Score. Le meilleur modèle a été sélectionné en fonction de ses résultats sur les données de test.

Fonction de Test

Une fonction a été développée pour prédire automatiquement, à partir des caractéristiques d'une nouvelle tumeur, si celle■ci est bénigne ou maligne en utilisant le modèle entraîné.

Conclusion Partielle

Le modèle final présente une bonne capacité de généralisation. Cette partie a montré l'importance du prétraitement et du choix du modèle pour obtenir de bonnes performances.

3. Partie 2 — Régression : Weather

Exploration et Visualisation

L'analyse exploratoire du dataset Weather a permis d'identifier les colonnes inutilisables, les valeurs manquantes et les relations entre variables climatiques. Des visualisations telles que les matrices de corrélation et les scatterplots ont été réalisées.

Prétraitement

Plusieurs colonnes ont été supprimées (colonnes vides, texte non convertible, colonnes météorologiques peu utiles). Ensuite, toutes les valeurs ont été converties en format numérique, puis les lignes invalides ont été supprimées. Enfin, les données ont été normalisées et séparées en ensembles d'apprentissage et de test.

Modélisation et Évaluation

Deux modèles régressifs ont été utilisés : la Régression Linéaire et un Arbre de Décision. Les métriques utilisées pour l'évaluation sont le MSE, RMSE, MAE et R^2. Les résultats ont montré que :
- L'arbre de décision obtient des métriques légèrement meilleures,
- Cependant, la régression linéaire produit des prédictions plus cohérentes et continues, mieux adaptées à une variable physique comme la température.

Fonction de Prédiction

Une fonction de prédiction a été conçue : elle prend les caractéristiques météorologiques d'un jour donné, applique la normalisation via le scaler entraîné, charge le modèle final et retourne une estimation de la température moyenne.

Conclusion Partielle

Bien que l'arbre de décision obtienne de meilleures performances numériques, la régression linéaire est plus stable et plus réaliste. Elle constitue donc un choix recommandé pour une application pratique.

4. Conclusion Générale

Ce projet a permis de comparer deux grandes tâches du Machine Learning : la classification et la régression. Les enseignements principaux incluent :
- L’importance du prétraitement et du nettoyage des données,
- La nécessité de choisir un modèle adapté à la nature de la cible,
- La compréhension des métriques d’évaluation selon la tâche,
- La mise en place de fonctions de test ou de prédiction réutilisables dans un système réel.

Ce travail a renforcé les compétences en modélisation, en analyse de données, et en construction de pipelines ML complets.