

Projet : Classification du Diagnostic du Cancer du Sein à l'Aide de Modèles de Machine Learning

Description du Projet :

Ce projet vise à développer un système de classification des patients en fonction de leur diagnostic médical à partir de données médicales anonymisées. Permettant de prédire si une tumeur mammaire est bénigne ou maligne en utilisant des techniques de machine learning. Les données utilisées pour former et évaluer les modèles comprendront des mesures cliniques des cellules présentes dans les biopsies de tumeurs.

Dataset :

Le dataset "Breast Cancer Wisconsin (Diagnostic)" contient des mesures cliniques de cellules présentes dans les biopsies de tumeurs mammaires. Il est composé de 569 exemples, avec 30 caractéristiques numériques et une variable cible indiquant si la tumeur est bénigne (B) ou maligne (M).

Lien : <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

Voici un aperçu des caractéristiques du dataset :

- Nombre de caractéristiques : 30
- Nombre de classes : 2 (Bénigne, Maligne)
- Nombre total d'exemples : 569

Étapes du Projet :

1. Collecte et Exploration des Données :
 - Charger le dataset et explorer ses caractéristiques.
 - Visualiser les relations entre les caractéristiques et les classes de tumeurs.
2. Prétraitement des Données :
 - Vérifier s'il y a des valeurs manquantes ou des valeurs aberrantes.
 - Normaliser les données si nécessaire.
3. Construction des Modèles :
 - Choix de différents algorithmes de classification tels que la regression logistique, les k-plus proches voisins (KNN), les arbres de décision, ou les machines à vecteurs de support (SVM).
 - Division des données en ensembles d'entraînement et de test.
4. Entraînement des Modèles :
 - Entraînement des modèles sur l'ensemble d'entraînement.
 - Optimisation des hyperparamètres pour améliorer les performances des modèles.
5. Évaluation des Modèles :

- Évaluation des performances des modèles sur l'ensemble de test à l'aide de mesures telles que la précision, le rappel, le score F1, etc.
- Comparaison des performances des différents modèles.

6. Interprétation des Résultats :

- Analyse des caractéristiques qui influent le plus sur la prédiction du diagnostic du cancer du sein selon les modèles développés.
- Identification des caractéristiques les plus importantes pour la classification des tumeurs.

7. Déploiement et Utilisation :

- Création d'une interface utilisateur simple permettant à un médecin de saisir les mesures cliniques d'une biopsie de tumeur mammaire et d'obtenir une prédiction sur son diagnostic.

Ressources :

- Langages de programmation comme Python avec des bibliothèques telles que scikit-learn pour l'implémentation des modèles de machine learning.
- Le dataset "Breast Cancer Wisconsin (Diagnostic)" est disponible sur UCI Machine Learning Repository.
- Documentation en ligne et tutoriels sur les techniques de prétraitement des données, la construction de modèles de machine learning et l'évaluation des performances.

Ce projet offre une expérience pratique dans le domaine du machine learning appliqué à la médecine, en utilisant des données médicales pour résoudre un problème de classification du diagnostic du cancer du sein. Il vous permettra de se familiariser avec les étapes typiques de développement de modèles de machine learning pour des applications médicales.