

TP Prédiction d'un risque médical à l'aide de l'apprentissage automatique

Problématique

Construire un système de classification permettant de prédire si une personne présente un risque de maladie à partir de données médicales simulées (âge, cholestérol, tension artérielle).

L'objectif est de proposer :

- Une approche classique avec Scikit-learn (régression logistique)

En suivant le cycle de vie de la data, vous développerez étape par étape le modèle, depuis la création des données jusqu'à l'entraînement et l'évaluation du modèle, en passant par la visualisation et l'analyse exploratoire.

- Manipuler des données tabulaires avec Pandas et NumPy
- Visualiser et interpréter les relations entre variables avec Matplotlib
- Appliquer un modèle de classification avec Scikit-learn
- Représenter l'évolution des performances d'un modèle d'apprentissage

Données utilisées

Les données sont générées aléatoirement pour simuler un fichier patient avec :

- Âge (age)
- Taux de cholestérol (cholesterol)
- Pression artérielle (blood_pressure)
- Présence ou non d'une maladie (has_disease) – variable cible

Etapes du TP

1. Importation des bibliothèques

Initialiser correctement l'environnement Python avec les outils essentiels, importer numpy, pandas, matplotlib.pyplot, scikit-learn, tensorflow

- Vérifier les versions installées

2. Création ou importation des données

Générer un jeu de données médical simulé avec un DataFrame

3. Analyse exploratoire

Comprendre les relations entre les variables.

4. Préparation des données

Structurer les données pour l'entraînement en séparant les données en X (features) et y (cible)

- Diviser en données d'apprentissage et de test (80% / 20%)

5. Modélisation avec Scikit-learn

Entraîner une régression logistique et mesurer sa précision.

- Créer un modèle avec LogisticRegression
- Entraîner avec .fit()
- Évaluer la précision avec accuracy_score

6. Visualisation des performances

Suivre visuellement l'évolution de l'entraînement du modèle.
