Institut teccart

420-TT-IA1: Intelligence Artificielle et Apprentissage Automatique

Projet de Machine Learning : Analyse et Prédiction des Données

Objectif Général

Ce projet vise à appliquer des concepts et des techniques de machine learning pour résoudre un problème de classification. Les étudiants seront amenés à analyser des ensembles de données spécifiques, à appliquer différents algorithmes de machine learning, et à évaluer leurs performances. En plus de l'analyse et de l'apprentissage automatique, les étudiants développeront une application web avec Streamlit pour visualiser et interagir avec les résultats de leurs modèles.

Institut teccart

Partie 1: Problème de Classification

Dataset: Heart Disease UCI (https://archive.ics.uci.edu/dataset/45/heart+disease)

Description: Ce dataset contient des informations sur des patients et indique si une personne est atteinte d'une maladie cardiaque. Les caractéristiques incluent l'âge, le sexe, le type de douleur thoracique, la pression artérielle au repos, le cholestérol, etc.

Le jeu de données contient des informations liées aux maladies cardiaques chez les individus, avec les colonnes suivantes :

- age: Age of the individual in years.
- sex: Sex of the individual (1 = male; 0 = female).
- cp: Chest pain type (values range from 0 to 3, representing different types of chest pain).
- trestbps: Resting blood pressure (in mm Hg on admission to the hospital).
- chol: Serum cholesterol in mg/dl.
- fbs: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false).
- restecg: Resting electrocardiographic results (values 0,1,2).
- thalach: Maximum heart rate achieved.
- exang: Exercise-induced angina (1 = yes; 0 = no).
- oldpeak: ST depression induced by exercise relative to rest.
- slope: The slope of the peak exercise ST segment (values 0,1,2).
- ca: Number of major vessels (0-3) colored by fluoroscopy.
- thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect).
- target: Heart disease (1 = disease; 0 = no disease).

En français:

- âge : Âge de l'individu en années.
- sexe : Sexe de l'individu (1 = homme ; 0 = femme).
- cp : Type de douleur thoracique (les valeurs vont de 0 à 3, représentant différents types de douleurs thoraciques).
- trestbps : Pression artérielle au repos (en mm Hg à l'admission à l'hôpital).
- chol : Cholestérol sérique en mg/dl.
- fbs : Glycémie à jeun > 120 mg/dl (1 = vrai ; 0 = faux).
- restecg : Résultats électrocardiographiques au repos (valeurs 0,1,2).
- thalach : Fréquence cardiaque maximale atteinte.
- exang : Angine induite par l'exercice (1 = oui ; 0 = non).
- oldpeak : Dépression du segment ST induite par l'exercice par rapport au repos.
- slope : La pente du segment ST au pic de l'exercice (valeurs 0,1,2).
- ca : Nombre de gros vaisseaux (0-3) colorés par fluoroscopie.
- thal : Un trouble sanguin appelé thalassémie (3 = normal ; 6 = défaut fixé ; 7 = défaut réversible).

- target : Maladie cardiaque (1 = maladie ; 0 = pas de maladie).

Sur la base de cette structure, nous pouvons poser plusieurs questions pour l'analyse des données :

- ☐ Quelle est la distribution de l'âge des individus dans le jeu de données ?
- ☐ Y a-t-il une différence dans la présence de maladie cardiaque entre les sexes ?
- ☐ Comment le type de douleur thoracique (cp) est-il lié à la présence de maladie cardiaque (cible) ?
- ☐ Quelles sont les valeurs moyennes de la pression artérielle au repos (trestbps), du cholestérol (chol) et de la fréquence cardiaque maximale (thalach) pour les individus avec et sans maladie cardiaque ?
- ☐ La glycémie à jeun au-dessus de 120 mg/dl (fbs) est-elle associée à une présence accrue de maladie cardiaque ?
- ☐ Comment l'angine induite par l'exercice (exang) est-elle corrélée à la maladie cardiaque ?

Algorithmes de Classification à Utiliser:

1. LogisticRegression
2. K-Nearest Neighbors (KNN)
3. Support Vector Machine (SVM)
4. DecisitionTreeClasifier
5. Random Forest
6. Adaboost

Métriques d'Évaluation:

- Accuracy
- Précision
- Rappel
- F1-Score
- AUC-ROC

Travail: comparez les résultats des algorithmes pour chaque métrique.

Livrables: un dossier zip contenant: jupyter notebook + codes python de API et Serveur.