Projet Data Science 2024-2025

Licence 3 SDN

Le sujet

Les maladies cardiaques restent une des principales causes de mortalité dans le monde. Détecter précocement les facteurs de risque et prédire la probabilité qu'un patient développe une maladie cardiaque pourrait sauver de nombreuses vies grâce à une intervention préventive.

Vous êtes recrutés par une entreprise de R&D spécialisée dans les solutions de santé, qui collecte des données médicales anonymisées auprès de plusieurs hôpitaux. Votre mission est d'explorer ces données, de comprendre les facteurs de risque associés aux maladies cardiaques, et de développer un modèle prédictif capable d'évaluer le risque qu'un nouveau patient soit atteint d'une maladie cardiaque.

Les données

Les données sont disponibles sur **I-Campus** sous le fichier `heart_disease_data.csv`. Les attributs de ces données sont les suivants :

- **Age** : âge du patient (en années).
- **Sex** : sexe du patient (`M` : Homme, `F` : Femme).
- **ChestPainType** : type de douleur thoracique (`TA` : Angine typique, `ATA` : Angine atypique, `NAP` : Douleur non angineuse, `ASY` : Asymptomatique).
- **RestingBP** : tension artérielle au repos (en mm Hg).
- **Cholesterol** : cholestérol sérique (en mg/dl).
- **FastingBS** : glycémie à jeun (`1` : si > 120 mg/dl, `0` : sinon).
- **RestingECG** : résultats de l'électrocardiogramme au repos (`Normal` : Normal, `ST` : anomalies ST-T, `LVH` : Hypertrophie ventriculaire gauche).
- **MaxHR** : fréquence cardiaque maximale atteinte.
- **ExerciseAngina** : angine provoquée par l'effort (`Y` : Oui, `N` : Non).
- **Oldpeak** : dépression du segment ST mesurée (valeur numérique).
- **ST_Slope** : pente du segment ST à l'effort (`Up` : ascendante, `Flat` : plate, `Down` : descendante).
- **HeartDisease** : présence ou absence de maladie cardiaque (`1` : Oui, `0` : Non).

Objectifs du projet

1. Exploration et analyse des données

Vous commencerez par une phase exploratoire des données pour identifier les informations pertinentes. Vous utiliserez les techniques de visualisation des données vues au premier semestre pour mettre en lumière les facteurs de risque potentiels.

- Analyser la répartition des maladies en fonction de l'âge, du sexe et des autres attributs.
- Visualiser les corrélations entre les différentes variables.
- Identifier les tendances significatives et les valeurs aberrantes.

2. Nettoyage et préparation des données

- Traiter les valeurs manquantes ou aberrantes.
- Transformer les variables catégoriques (ex. : `ChestPainType`) en variables numériques exploitables par les modèles.
- Normaliser ou standardiser les variables si nécessaire (ex. : `Cholesterol`).

3. Construction du modèle prédictif

Vous développerez un modèle de machine learning pour prédire si un patient est susceptible d'avoir une maladie cardiaque.

- Vous êtes libres de choisir les outils et algorithmes que vous jugez appropriés (ex. : KNN, Random Forest, Logistic Regression).
- Vous évaluerez la performance de votre modèle à l'aide des métriques vues en cours (précision, rappel, F1-score, etc.).
- Attention : minimisez les faux négatifs, car une mauvaise classification pourrait avoir des conséquences graves sur la santé des patients.

Organisation

Vous vous répartirez en groupe de 2-3 étudiants. Vous disposez de 9 séances de 3h de travail pour réaliser le projet en autonomie.

Il n'y a pas d'enseignant durant les séances en autonomie. Cependant, n'hésitez pas à solliciter vos enseignants si vous avez des questions sur les différents aspects du projet :

- **M. Almaksour** pour Data Science.
- **M. Demoli** pour Data Visualisation.

Rendu attendu

Vous devrez réaliser un rapport de 10 pages maximum qui décrira votre démarche, vos résultats, et vos conclusions. Voici les points que vous devez aborder et faire attention dans votre rapport (liste non exhaustive) :

- Une introduction qui décrit les données récoltées et vos questions d'intérêts.
- La qualité de justifications de choix d'analyse afin de répondre aux questions.
- La qualité de visualisations (lisibilité, clarté, auto suffisance, etc.).
- Le choix d'une visualisation vis-à-vis de l'analyse effectuée.
- L'originalité et la validité des conclusions tirées des visualisations.
- La bonne utilisation des techniques statistiques.
- Le choix de la méthode de résolution à utiliser pour la prédiction et sa calibration.
- L'évaluation de la qualité de vos résultats :
  - Votre protocole : comment vous vous êtes assurés que votre méthode donne de bons résultats (sur quels jeux de données vous avez testé, en les découpant comment, en évaluant combien de fois, séparation apprentissage/test, etc.).
  - Votre protocole est complet s'il est suffisant pour que vos résultats soient reproductibles par une autre personne à partir du même jeu de données.

- Les tableaux de résultats obtenus.
- L'interprétation des résultats obtenus.

Vous devrez également présenter votre projet en 10 minutes lors de la soutenance orale. Vous devrez présenter votre démarche, vos résultats, et vos conclusions. Vous devrez également répondre aux questions de vos enseignants.

Bonus

Tester et évaluer plusieurs algorithmes de classification pour identifier le plus performant dans ce contexte.

Sujet alternatif

Vous êtes également libres de travailler sur un autre sujet de votre choix à condition que cela soit validé par vos deux professeurs.