La Plateforme_

Job Risques Cardio-Vasculaires

Ce job va vous permettre de mettre en œuvre une régression logistique. Pour analyser une variable binaire (dont les valeurs seraient VRAI/FAUX, 0/1, ou encore OUI/NON) en fonction d'une variable explicative quantitative, on peut utiliser une régression logistique. La régression logistique consiste à prédire des variables binaires et non à prédire des variables continues.

Vous travaillez dans le domaine de la médecine préventive. Votre métier est donc de donner des conseils d'hygiène de vie (propreté, mais aussi diététique, encouragement à un sport ou une activité physique, ergonomie et manière de faire des efforts, prévention des comportements à risque, etc.) ainsi que de proposer un accompagnement dans le dépistage de maladies et plus spécifiquement dans la prévention des risques cardio-vasculaire.

Entre 300 000 et 400 000 accidents cardiovasculaires surviennent chaque année en France, dont un tiers sont mortels. Comment mieux prédire le risque cardiovasculaire ? Si plusieurs facteurs de risque sont identifiés, quelles sont les interactions entre ces facteurs ? Les maladies cardiovasculaires, principalement les accidents vasculaires cérébraux (AVC) et les infarctus du myocarde, sont la deuxième cause de mortalité en France. La liste des facteurs de risque cardiovasculaire est malheureusement longue :

Les 12 facteurs de risque constituent ainsi un véritable réseau de facteurs de risque cardiovasculaire.
En fonction de ces interactions, des chercheurs ont pu mettre en évidence 4 groupes de facteurs de risque :

1. Des «**facteurs non modifiables**» (le sexe, l'âge et les antécédents familiaux) : ils prédisent d'autres facteurs, mais ne peuvent pas être prédits par d'autres facteurs.
2. Des «**facteurs liés au mode de vie**» (le tabagisme, la sédentarité, l'alcoolisme) : ils prédisent beaucoup d'autres facteurs (sauf les facteurs non modifiables), mais sont très peu prédits par d'autres facteurs.
3. Des «**facteurs cliniques en amont**» (les troubles du sommeil, l'obésité, la dépression) : ils prédisent beaucoup d'autres facteurs et sont eux-mêmes prédits par de nombreux facteurs.
4. Des «**facteurs cliniques en aval**» (l'hypertension artérielle, les dyslipidémies, le diabète) : ils prédisent très peu de facteurs, mais sont en revanche prédits par beaucoup de facteurs.

Pour vous accompagner au mieux dans votre démarche de prévention de ces risques cardio-vasculaire, vous avez décidé de développer un outil permettant de poser un diagnostic rapide de risques cardio-vasculaire. Cet outil mettra en œuvre un algorithme de machine learning (de classification binaire : prédiction binaire : 0 ou 1) permettant de prédire s'il y a un risque cardio-vasculaire ou s'il n'y en a pas.

Présentation des données

Pour pouvoir entraîner votre algorithme, vous avez monté un partenariat avec des médecins généralistes de votre ville et ainsi pu récolter des données de patients. Ces données sont stockées dans un fichier .csv. Ce fichier comporte 12 colonnes :

AGE: integer (number of days)
HEIGHT: integer (cm)
WEIGHT: integer (kg)
GENDER: categorical (1: female, 2: male)
AP_HIGH: systolic blood pressure, integer
AP_LOW: diastolic blood pressure, integer
CHOLESTEROL: categorical (1: normal, 2: above normal, 3: well above normal)
GLUCOSE: categorical (1: normal, 2: above normal, 3: well above normal)
SMOKE: categorical (0: no, 1: yes)
ALCOHOL: categorical (0: no, 1: yes)
PHYSICAL_ACTIVITY: categorical (0: no, 1: yes)

et la variable cible :

CARDIO_DISEASE: categorical (0: no, 1: yes)

To-Do

En vous appuyant sur ces données, Construisez un modèle de régression logistique permettant de prédire qui sont les sujets à risque !

1. Réaliser une veille sur la régression logistique.
2. Utiliser un jupyter-notebook pour le travail qui suit.
3. Visualiser et analyser les données avec les librairies Matplotlib et Seaborn.
4. Résoudre le cas d'étude présenté ci-dessus avec la librairie Scikit-Learn.
5. Résoudre le cas d'étude présenté ci-dessus avec votre propre classe python sans utiliser la librairie Scikit-Learn.
6. **Prédire si Arthur 53 ans, fumeur, sportif, 175 cm, 85 kg, avec un taux de cholestérol au dessus de la normal et un taux de glucose normal, une tension artérielle systolique dans la moyenne et une pression sanguine diastolique correspondant à la moyenne du 3e quartile (50%-75%) du jeu de données, est un sujet à risques cardio-vasculaires.**
7. Rendre accessible votre notebook via Github.
8. Partager votre lien github.