Commentaires

Description du dataset

Le dataset présente des données médicales sur des patients qui ont eu ou non des accidents vasculaires-cérébraux (AVC). Il provient de la plateforme kaggle. Le but de ce projet sera de prédire les AVC chez les patients à partir des différentes mesures afin de pouvoir les prévenir.

Le dataset contient des informations sur 5110 patients ; certaines colonnes ont des données manquantes (4909 lignes complètes).

Les 12 propriétés présentes dans le dataset sont : - un id unique pour chaque patient - l'âge en années - le genre - la présence ou non d'hypertension - la présence ou non de maladies cardiaques - si l'individu a déjà été marié - le type de travail (public, privé, indépendant, etc.) - le type de résidence - la glycémie moyenne (en g/cL – unité de mesure non précisée dans le dataset mais déduite en comparant les valeurs du dataset avec les valeurs normales de glycémie - l'IMC, mesure de la corpulence d'une personne à partir de sa taille et de son poids (techniquement en kg/m2 mais en pratique sans unité) - le tabagisme - le fait que l'individu aie déjà eu un AVC

Analyse préliminaire du dataset

De premier abord, on peut supposer que des mesures qui correspondent à conditions propices aux maladies en général (comme l'IMC, indicateur du surpoids ; la glycémie, indicative du diabète ; l'âge avancé) seront positivement corrélés avec la présence d'AVC. On peut aussi supposer que des facteurs extérieurs qui sont connus pour avoir un effet négatif sur la santé (le tabagisme, la pollution de l'air quand on habite en ville) présenteront aussi une corrélation avec la présence d'AVC.

La visualisation par histogramme des différentes propriétés du dataset montrent que pour les variables catégoriques, toutes les catégories sont assez équitablement représentées (âge, genre, présence de maladies cardiaques, type d'emploi, type de résidence). Seul la variable prédite, la présence ou non d'AVC chez un patient, est fortement déséquilibré en faveur des patients non-victimes d'AVC.

Les taux de glycémies présentent une distribution bi-modale.

Explication de l'analyse quantitative et de la visualisation

Visualisation 1 en nuage de points

Le graphique en nuage de points est une technique de visualisation basique qui permet de d'avoir une vue d'ensemble du jeu de données. Elle convient mieux aux données numériques qu'aux données catégoriques. Pour cette raison, nous avons choisi de représenter toutes les propriétés numériques du dataset :

l'âge, l'IMC, la glycémie. Chaque propriété correspond à un axe d'un repère à 3 dimensions. La couleur des points indique la présence ou non d'un AVC.

Par défaut, la visualisation ne montre que 400 points avec un nombre à peu près égal de patients victimes d'AVC et de patients qui ne le sont pas (la surreprésentation des patients sans AVC dans le jeu de données rend le graphique illisible sinon). Il est possible de sélectionner manuellement les index des données à utiliser avec les options de ligne de commande.

Par rapport aux hypothèses initiales, on observe que les AVC sont bien corrélés avec l'âge, mais ils ne semblent pas corrélés avec l'IMC ou la glycémie d'après ce graphique.

Visualisation 2 : scatter matrix et clustering

Afin de visualiser différemment et de faire apparaître la structure des données, nous avons effectué une scatter matrix des données. Dans ce cas, on ne peut que travailler avec des valeurs numériques (pas de catégories) on n'utilise donc que les colones "age", "bmi" et "avg_glucose_level".

Les *scatterplots* de chaque combinaison de paramètres semblent confirmer un lien entre l'âge et la présence d'AVC; de plus, ils montrent que la combinaison glucose/bmi semble contenir 2 clusters de points.

Afin de tester la validité des clusters, nous avons exécuté l'algorithme de *k-means clustering* sur ces paramètres.

Pour avoir de l'aide sur comment exécuter l'algorithme de clustering, choisir les colonnes et les index de points à utiliser, appeler le script `visualisation_2.py` avec l'option `-h`

Traitement des données

Afin de traiter les données avec des méthodes supervisées et non-supervisées, nous les traitons au préalable de la manière suivante: * one-hot encoding des variables catégoriques * conversion en types numériques des variables catégoriques binaires * normalisation des variables numériques pour arriver à une moyenne de 0 et un écart type de 1

Analyse supervisée : Multi-Layer Perceptron

Afin de prédire les AVC chez les patients, nous avons choisi d'entraîner un réseau neuronal de type perceptron multi-couche le jeu de données.

Le problème de prédiction d'AVC correspond à un problème de classification binaire, avec comme classes prédites `présence d'AVC` ou `Absence d'AVC`. En conséquence, la fonction de coût choisie est *binary cross-entropy*, qui convient aux problèmes de classification binaire.

Pour évaluer notre algorithme, nous choisissons l'exactitude (*accuracy*) et le rappel (*recall*). L'exactitude est une mesure générale de performance du modèle ; le rappel est important pour notre problème car il correspond au taux de cas positif correctement identifiés. Dans le cas de problèmes de santé comme les AVC, il est plus intéressant d'identifier correctement tous les cas positifs, quitte à avoir quelques faux-positifs, que de se concentrer sur la précision.

Pour l'architecture de notre MLP, nous avons commencé avec 3 couches, avec 8, 16 et 32 neurones respectivement. Nous avons choisi comme fonction d'activation ReLU car elle est standard et a l'avantage d'éviter le problème de disparition des gradients. Les performances de ce premier modèle sont très bonnes : ~95% de précision et autant de rappel.

Pour voir si on peut alléger le modèle, on enlève une couche et on se rend compte que les mesures d'évaluation ne diminuent pas. On enlève une autre couche, puis on laisse une seule couche avec un seul neurone avec une fonction d'activation linéaire : les performances sont toujours au même niveau.

Un unique perceptron avec une fonction d'activation linéaire est en fait un modèle linéaire. Nous avons d'abord conclu que les données sont séparables linéairement et qu'un modèle linéaire suffit pour prédire la présence d'un ACV chez un patient avec 95% d'exactitude.

Mais nous avons vu lors de l'analyse préliminaire du dataset que la classe "AVC" était largement sous-représentée. Nous en concluons que le modèle a simplement appris à prédire systématiquement une absence d'AVC. En testant le modèle sur un set de test qui ne contient que des patients victimes d'AVC, on obtient bien une exactitude et un rappel de 0.

Pour pallier à ce problème il faudrait soit rééquilibrer le dataset en enrichissant par exemple avec la méthode SMOTE, soit utiliser une fonction de coût qui pénalise plus fortement une mauvaise prédiction sur un patient victime d'AVC.

Commentaire sur les résultats obtenus

En conclusion, nous avons utilisé deux méthodes pour visualiser le dataset : le nuage de points en 3d nous a permis de représenter 4 dimensions des données efficacement ; la scatter matrix nous permet de représenter théoriquement une infinité de dimensions. Ces deux représentations nous ont permis d'avoir des données préliminaires sur la répartition des données.

Les performances de l'algorithme de classification supervisé MLP ne sont pas satisfaisantes car les classes de la variable prédite ne sont pas balancées.