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.

Traitement des données

Pour traiter les données, nous avons séparé les features en plusieurs catégories en fonction de l'impact estimé des features par rapport au risque d'un avc :

Les features majeures : Hypertension + maladies cardiaques + âge

- Les features moyennes : Taux de glucose moyen + IMC + statut tabagique

- Les features mineures : Type de résidence

- Les features non pertinentes : sexe, identité, type de travail, jamais marié.

Ensuite, nous avons attribué des valeurs pour les features nominal du dataset. Nous avons attribué les valeurs avec le dictionnaire suivant :

```
cleanup_nums = {"Residence_type":     {"Rural": 1, "Urban": 0},
                "smoking_status":     {"never smoked": 0, "Unknown": 0,
"smokes": 1.5, "formerly smoked": 1.25 }}`
```

Nous avons éliminé toutes les features appartenant à la dernière catégorie, et les avons extrait du dataset. Toutes les features restantes ont été normalisées de manière à avoir une meilleure approche et comparer leurs valeurs entre elles de manière plus pertinente.

Enfin, nous avons attribué un poids d'importance en fonction de la catégorie :

Features majeurs : 0.8

Features moyennes : 0.5

Features mineurs : 0.2

Features non pertinentes : 0.0

Clustering Méthode : K-means ++

Nous avons implémenté la méthode de clustering de K-mean++. Le partitionnement en k-mean est une méthode de partitionnement de données et un problème d'optimisation combinatoire. Étant donnés des points et un entier k, le problème est de diviser les points en k groupes, souvent appelés clusters, de façon à minimiser une certaine fonction. On considère la distance d'un point à la moyenne des points de son cluster ; la fonction à minimiser est la somme des carrés de ces distances.

K-mean a un problème d'optimisation dans la création de ses clusters par défaut. Son initialisation est cruciale dans son optimisation et ses performances.

Nous avons donc orienté notre choix sur k-mean ++.

K-moyennes++ est un algorithme d'initialisation des k points qui propose une initialisation améliorant la probabilité d'obtenir la solution optimale (minimum global). L'intuition derrière cette approche consiste à répartir les k points des moyennes initiales. Le point de moyenne initial du premier cluster est choisi aléatoirement parmi les données. Puis chaque point de moyenne initiale est choisi parmi les points restants, avec une probabilité proportionnelle au carré de la distance entre le point et le cluster le plus proche. Celle-ci permet de choisir parmi une liste de valeurs avec une loi de probabilité donnée.

Ici, chaque point a une probabilité d'être choisi comme centroïde proportionnelle à sa distance avec le centroïde le plus proche ; plus le centroïde est proche, plus la probabilité que le point soit choisi est faible et inversement.

Evaluation du modèle : Elbow méthode et Silhouette

Pour évaluer de manière pertinente notre modèle k-means++, nous avons implémenté deux méthodes disctinces pour comparer leurs résultats et augmenter la fiabilité de ses sources.

Tout d'abord, La méthode Elbow ou méthode du coude : la méthode du coude est une heuristique utilisée pour déterminer le nombre de grappes dans un ensemble de données. La méthode consiste à tracer la variation expliquée en fonction du nombre de clusters, et à choisir le coude de la courbe comme nombre de clusters à utiliser. La même méthode peut être utilisée pour choisir le nombre de paramètres dans d'autres modèles basés sur les données, comme le nombre de composantes principales pour décrire un ensemble de données.

Ensuite, la silhouette méthode : le coefficient de silhouette est une mesure de qualité d'une partition d'un ensemble de données en classification automatique. Pour chaque point, son coefficient de silhouette est la différence entre la distance moyenne avec les points du même groupe que lui (cohésion) et la distance moyenne avec les points des autres groupes voisins (séparation). Si cette différence est négative, le point est en moyenne plus proche du groupe voisin que du sien : il est donc mal classé. À l'inverse, si cette différence est positive, le point est en moyenne plus proche de son groupe que du groupe voisin : il est donc bien classé.

L'objectif était de comparer ces deux heuristiques et de vérifier si elles nous donnaient le même nombre optimal de clusters pour la méthode de clustering k-means. Cette expérience était intéressante car le dataset étudié est constitué de données réelles et il n'y a pas de clusters artificiels dans la distribution des points. Si les deux heuristiques nous donnent un même nombre de clusters optimal, cela conforterait le choix du nombre de clusters optimal. Dans le cas inverse, il faudrait nuancer le résultat.

Commentaire sur les résultats

La méthode du coude nous donne un nombre optimal de cluster de 3.

Malheureusement, la méthode silhouette n'a pas donné les résultats espérés. Le coefficient de silhouette est toujours aux alentours de 0.6 quel que soit le nombre de clusters essayé ; il y a là surement une erreur d'implémentation qui nous empêche de comparer les résultats des deux méthodes. Nous admettons une faille dans l’implémentation de notre méthode.