**Projet de prédiction des maladies cardiaques : Régression logistique**

# the basics on heart diseaseIntroduction :

Les maladies cardiovasculaires sont la cause la plus fréquente de décès dans le monde entier au cours des dernières décennies, tant dans les pays développés que dans les pays en développement.

Il existe de nombreux types de maladies cardiaques, et chacune a ses propres symptômes et son propre traitement.

* Pour certains, des changements de mode de vie et des médicaments peuvent faire une énorme différence pour améliorer leur santé.
* Pour d'autres, une intervention chirurgicale peut s'avérer nécessaire pour que leur cœur fonctionne à nouveau correctement.

Découvrez quelques-uns des types courants de maladies cardiaques, comment les prévenir et comment les traiter. [ici](https://www.webmd.com/heart-disease/heart-disease-types-causes-symptoms)

Les maladies cardiovasculaires sont la première cause de mortalité dans le monde, avec 17,3 millions de décès par an, un chiffre qui devrait atteindre plus de 23,6 millions d'ici 2030.

La détection précoce La détection précoce des maladies cardiaques et la surveillance continue des cliniciens peuvent réduire le taux de mortalité. Cependant, il n'est pas possible de surveiller les patients tous les jours et dans tous les cas de manière précise, ainsi la consultation d'un patient pendant 24 heures par un médecin n'est pas disponible car elle nécessite plus de savoir de temps et d'expertise.

Dans ce projet, nous avons développé et recherché des modèles pour la prédiction des maladies cardiaques à travers les différents attributs cardiaques du patient et de détecter les maladies cardiaques imminentes, en utilisant des techniques d'apprentissage automatique comme l'algorithme d'élimination à rebours, la régression logistique et le REFCV sur le jeu de données disponible publiquement sur le site Web [Kaggle](https://www.kaggle.com/ronitf/heart-disease-uci), en évaluant ensuite les résultats en utilisant confusion matrix et cross validation.

## Définition du problème :

Le défi majeur des maladies cardiaques est leur détection. Il existe des instruments qui peuvent prédire les maladies cardiaques, mais ils sont soit coûteux, soit inefficaces pour calculer le risque de maladie cardiaque chez l'homme.

La détection précoce des maladies cardiaques permet de réduire le taux de mortalité et les complications. Cependant, il n'est pas possible de surveiller les patients tous les jours et dans tous les cas de manière précise.

Puisque nous disposons d'une grande quantité de données dans le monde d'aujourd'hui, nous pouvons utiliser divers algorithmes d'apprentissage automatique pour analyser les données à la recherche de modèles cachés qui peuvent être utilisés pour le diagnostic de santé dans les données médicales.

## Motivation :

Les techniques d'apprentissage automatique nous entourent et ont été comparées et utilisées pour l'analyse de nombreux types d'applications de science des données. La motivation principale de ce projet de recherche était d'explorer les méthodes de sélection des caractéristiques, la préparation et le traitement des données derrière les modèles de formation dans l'apprentissage automatique. Avec des modèles et des bibliothèques de première main, le défi auquel nous sommes confrontés aujourd'hui est celui des données où, à côté de leur abondance, et nos modèles, la précision que nous voyons pendant la formation, le test et la validation réelle a une variance plus élevée.

Ce projet est donc mené avec la motivation d'explorer les modèles et de mettre en œuvre le modèle de régression logistique pour entraîner les données obtenues. En outre, comme l'apprentissage automatique est motivé pour développer un système informatique approprié et une aide à la décision qui peut aider à la détection précoce des maladies cardiaques, dans ce projet, nous avons développé un modèle qui classe si le patient aura une maladie cardiaque dans dix ans ou non en fonction de diverses caractéristiques (c'est-à-dire les facteurs de risque potentiels qui peuvent causer des maladies cardiaques) en utilisant la régression logistique. Par conséquent, le pronostic précoce des maladies cardiovasculaires peut aider à prendre des décisions sur les changements de style de vie chez les patients à haut risque et à réduire les complications, ce qui peut être un grand pas en avant dans le domaine de la médecine.

## Objectifs :

Les principaux objectifs du développement de ce projet sont :

1. Développer un modèle d'apprentissage automatique pour prédire la possibilité future de maladie cardiaque en mettant en œuvre la régression logistique.

2. Déterminer les facteurs de risque significatifs basés sur un ensemble de données médicales qui peuvent conduire à une maladie cardiaque.

3. Analyser les méthodes de sélection des caractéristiques et comprendre leur principe de fonctionnement.

4. Prédire si le patient présente un **risque de maladie coronarienne (CHD) sur 10 ans.**

# Datasets :

L'ensemble de données fournit les informations sur les patients. Il comprend plus de 4 000 enregistrements et 15 attributs. Chaque attribut est un facteur de risque potentiel. Il existe des facteurs de risque démographiques, comportementaux et médicaux.

Les facteurs de risque démographiques :

* Sex : mâle ou femelle(Nominal)
* Age: Âge du patient;(Continu — Bien que les âges enregistrés aient été tronqués en nombres entiers, le concept d’âge est continu)

Les facteurs de risque comportementaux :

* Current Smoker : si le patient est ou non un fumeur actuel (Nominal)
* Cigs Per Day: le nombre de cigarettes que la personne a fumé en moyenne en une journée. (peut être considéré comme continu car on peut avoir n’importe quel nombre de cigarettes, même une demi-cigarette.)

Les facteurs de risque médicaux antécédents:

* BP Meds: si le patient était ou non sous médicament pour la pression artérielle (nominal)
* Prevalent Stroke: si le patient avait déjà eu un AVC (nominal)
* Prevalent Hyp: si le patient était hypertensif ou non (nominal)
* Diabetes: si le patient souffrait ou non de diabète (nominal)

Les facteurs de risque médicaux courant:

* Tot Chol: taux de cholestérol total (continu)
* Sys BP: pression artérielle systolique (continue)
* Dia BP : pression artérielle diastolique (continue)
* IMC : Indice de masse corporelle (continu)
* Heart Rate: fréquence cardiaque (continue - Dans la recherche médicale, des variables telles que la fréquence cardiaque sont en fait discrètes, mais sont considérées comme continues en raison du grand nombre de valeurs possibles.)
* Glucose: taux de glucose (continu)

Prédire la variable (cible souhaitée)

* TenYearCHD: 10 ans de risque de coronaropathie (binaire: « 1 », signifie « Oui », « 0 » signifie « Non »)



Figure 1 Capture de l'originale Data

Les données relatives à l'éducation ne sont pas pertinentes pour les maladies cardiaques d'un individu, elles sont donc abandonnées. cet ensemble de données procèdera à un prétraitement .

# Expérimentations :

## Préparations des données :

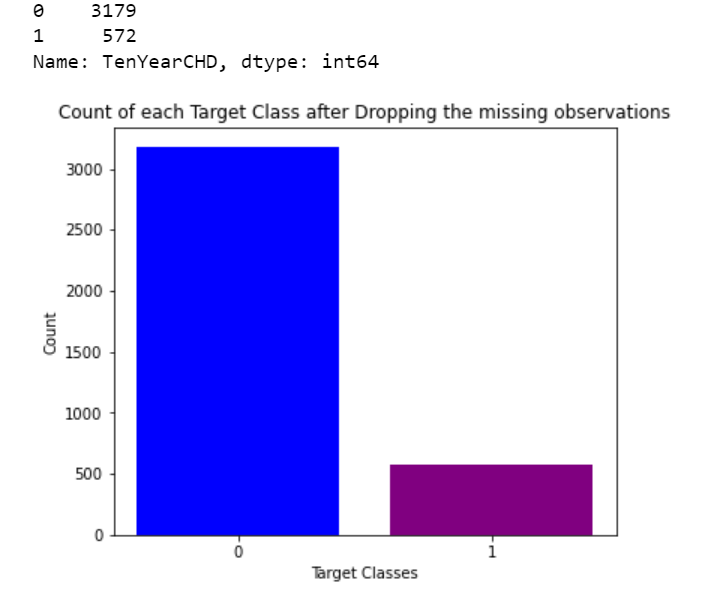
Comme l'ensemble de données se compose de 4240 observations avec 388 données manquantes et 644 observations à risque pour les maladies cardiaques, deux expériences différentes ont été réalisées pour la préparation des données. Tout d'abord, nous avons vérifié en supprimant les données manquantes, ce qui laisse seulement 3179 données et seulement 572 observations risquées pour les maladies cardiaques.

Figure 3: Bar Graph of the Target Classes After Dropping

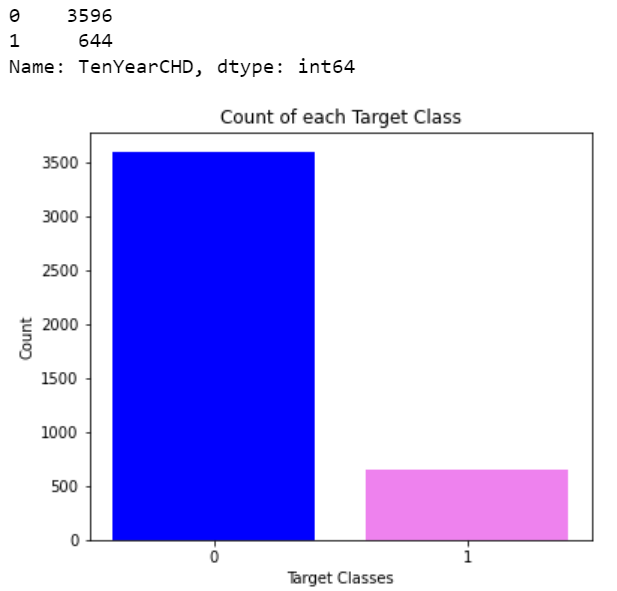


Figure 2: Bar Graph of the Target Classes before Dropping

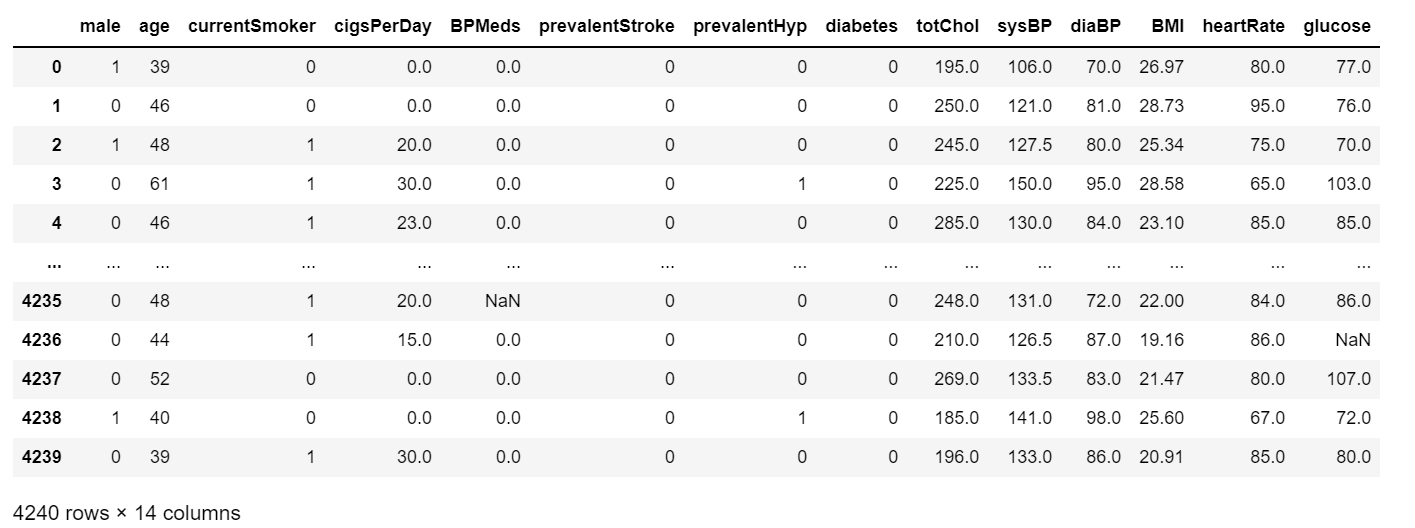
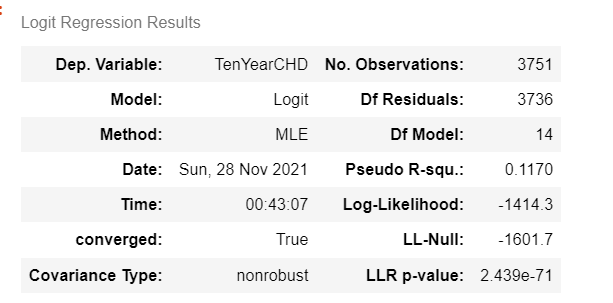
Cela conduit à une réduction du nombre d'observations fournissant une formation non pertinente à notre modèle. Nous avons donc procédé à l'imputation des données avec la valeur moyenne des observations et à leur mise à l'échelle en utilisant les modules SimpleImputer et StandardScaler de Sklearn.

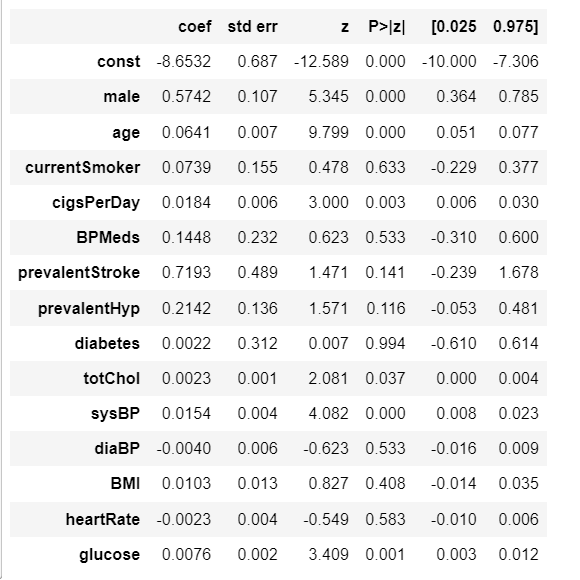
Figure 4 : Dataset après Scaling et Imputng

# Logistique régression :

La régression logistique est un type d'analyse de régression en statistique utilisé pour prédire le résultat d'une variable dépendante catégorique à partir d'un ensemble de variables prédictives ou indépendantes.

Dans la régression logistique, la variable dépendante est toujours binaire. La régression logistique est principalement utilisée pour la prédiction et le calcul de la probabilité de succès.





Les résultats ci-dessus montrent que certains attributs ont une valeur P supérieure à l'alpha préféré (5 %) et présentent donc une faible relation statistiquement significative avec la probabilité de maladie cardiaque. L'approche Backward elimination est utilisée ici pour éliminer les attributs ayant la valeur P la plus élevée, un par un, puis en exécutant la régression à plusieurs reprises jusqu'à ce que tous les attributs aient des valeurs P inférieures à 0,05.

## Feature selection :

Les données ont ensuite été soumises à la fonction the backward elimination afin de sélectionner les caractéristiques les plus pertinentes, ce qui a donné le résultat suivant :

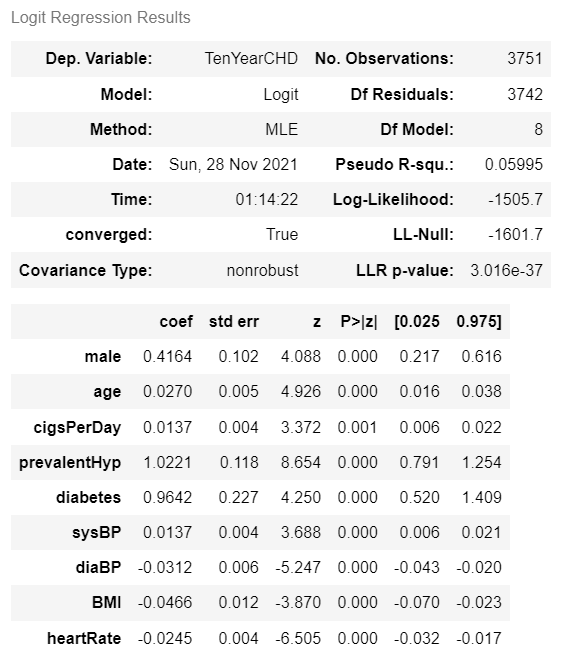


Figure 5 : Result from Feature Selection using Backward Elimination Method

Selon le résultat ci-dessus, les colonnes ont été abandonnées.

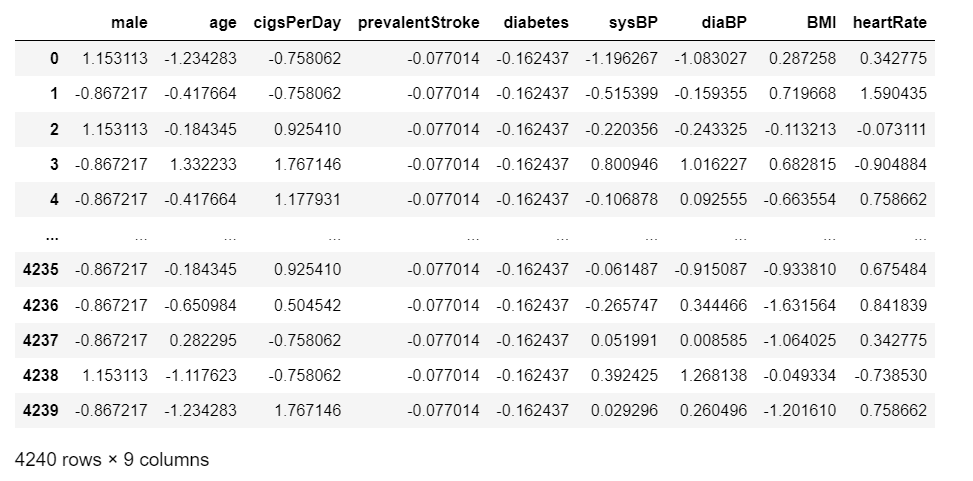
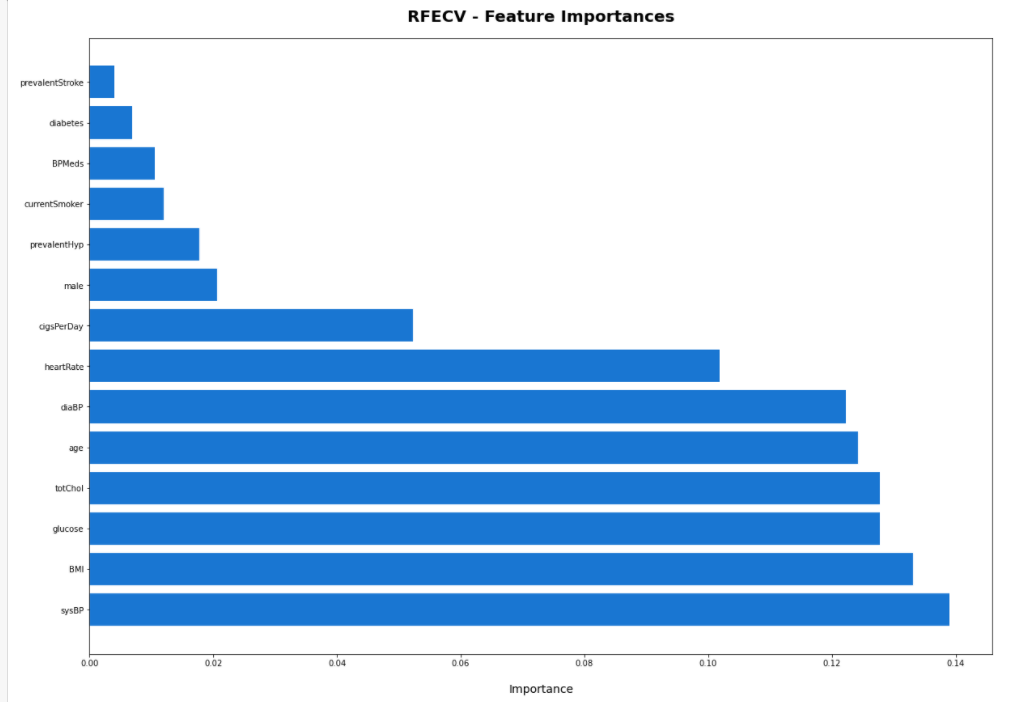


Figure 6: Dataset After Dropping Columns after Feature Selection

## Feature Selection using Recursive Feature Elimination and Cross-Validated selection method:

Figure 7:Top 10 important features supported by RFECV