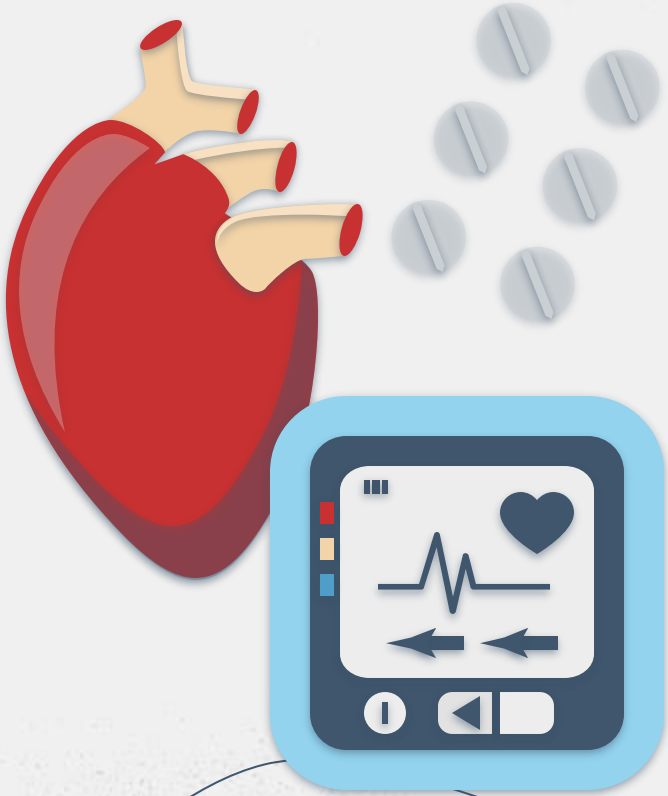
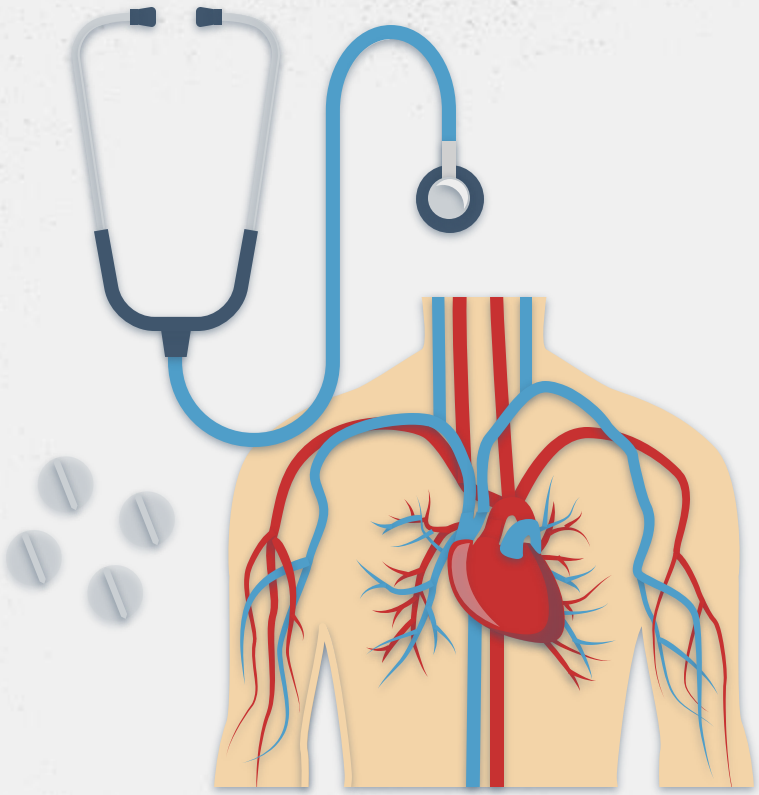
Prédiction sur le développement d'une maladie cardiaque



Projet basé sur un jeu de données du CDC
US Centers for Disease Control and Prevention

Maud Gesbert

TABLE DES MATIERES



01
Le jeu de données

Présentation du jeu de données

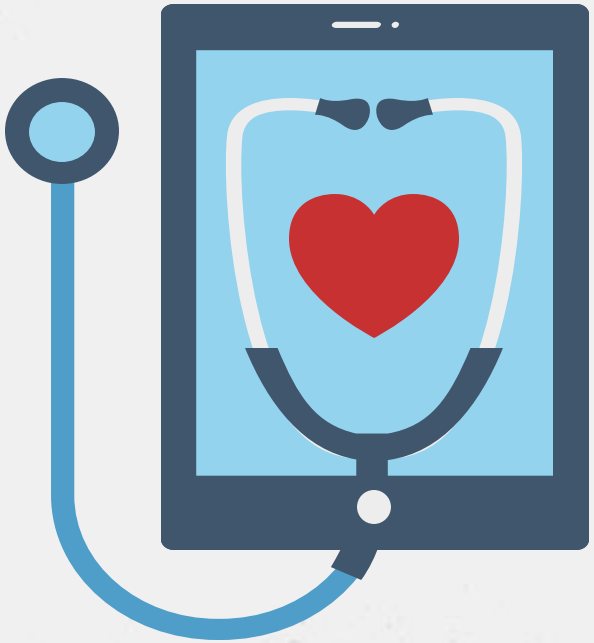
Jeu de données disponible sur Kaggle, trouvé sur le site du CDC :

- Contient des informations anonymisées sur 253 680 patients
- Contient une colonne cible : indique si la personne en question a développé une maladie cardiaque, ou fait une crise cardiaque, information que l'algorithme devra être capable de prédire après avoir été entrainé
- Contient 21 autres colonnes : indiquent les réponses données par les patients, vont permettre à l'algorithme d'identifier quel type de patient ont développé une pathologie

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	18.0	15.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	0.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	30.0	30.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	3.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
253675	0.0	1.0	1.0	1.0	45.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0	5.0
253676	0.0	1.0	1.0	1.0	18.0	0.0	0.0	2.0	0.0	0.0	...	1.0	0.0	4.0	0.0	0.0
253677	0.0	0.0	0.0	1.0	28.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0	0.0
253678	0.0	1.0	0.0	1.0	23.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0	0.0
253679	1.0	1.0	1.0	1.0	25.0	0.0	0.0	2.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0

253680 rows × 22 columns

- 253 680 personnes interrogées
- 22 colonnes en tout



02
Analyses préliminaires

* Voir fichier Heart_Disease_EDA

Etude de la variable cible et des distributions

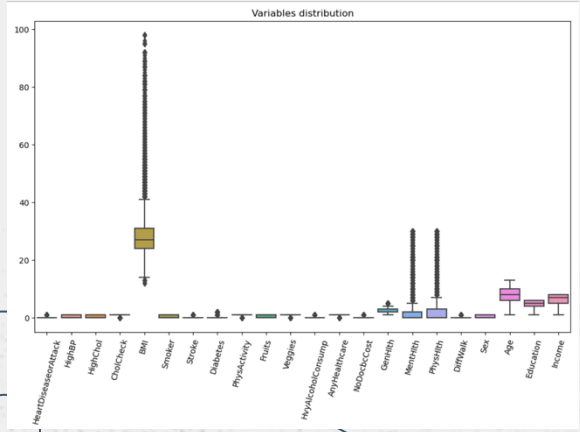
La première étape consiste à étudier le jeu de données, étudier la colonne cible, ainsi que les différentes variables (21 colonnes)

- Lorsque la colonne cible contient un 0 : la personne n'a pas développé de maladie cardiaque
- Lorsque la colonne cible contient un 1 : la personne a développé de maladie cardiaque

HeartDiseaseorAttack	
0.0	229787
1.0	23893

On note directement un déséquilibre : le nombre de personne n'ayant pas de maladie est très supérieur, il faudra prendre cela en compte pour éviter d'induire le modèle en erreur lors de l'entrainement

Puis,

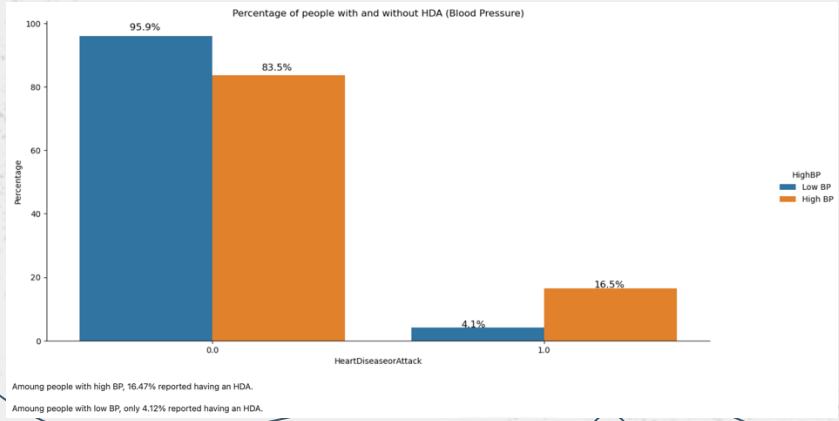


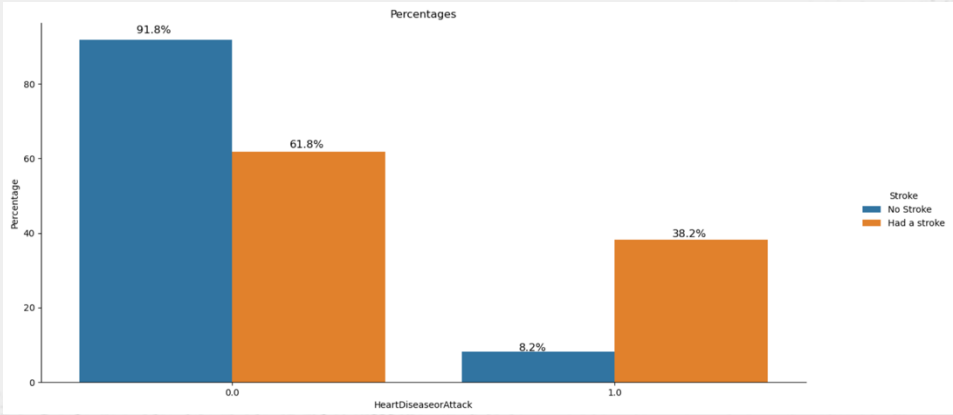
- Etude de la distribution des variables

Etudes des variables non-cibles

Etude de la répartition des variables et de leur lien avec la possibilité de développer une maladie cardiaque, par exemple :

- Quel pourcentage de personne ayant une tension élevée ont développé une maladie cardiaque ? (16.5%)
- Quel pourcentage de personne ayant fait un AVC ont développé une maladie cardiaque ? (38.2%)

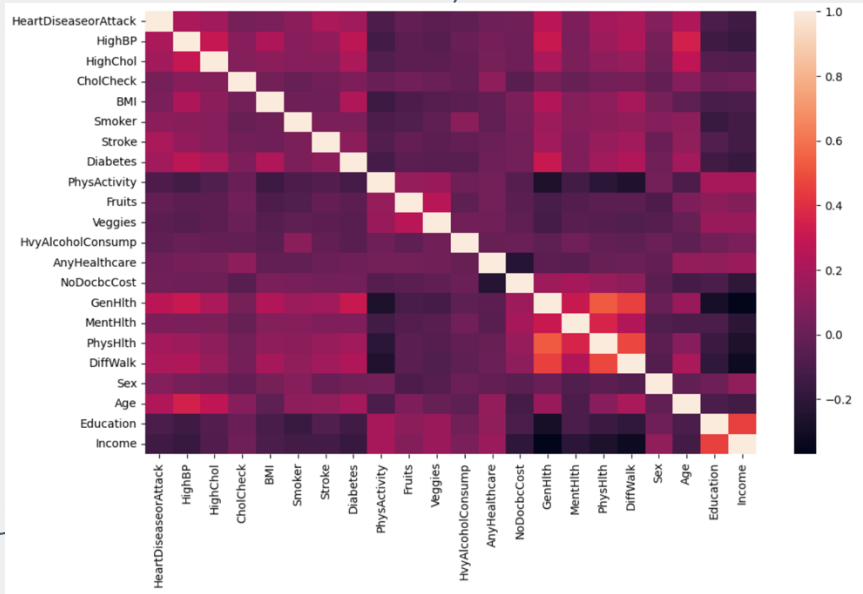




Etude des corrélations entre variables

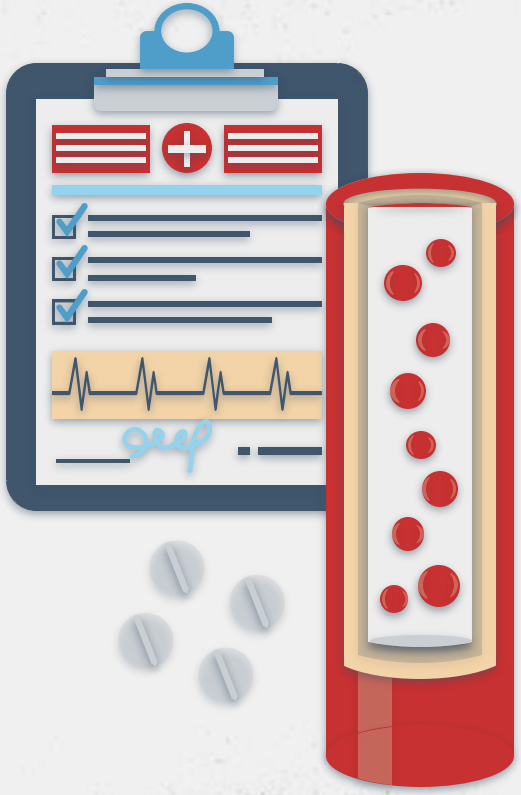
Rapide étude de la corrélation de Pearson entre les variables :

- Avoir un premier avis sur les variables qui influenceront le plus la prédiction
- Vérifier qu'il n'y a pas de forte corrélation linéaire entre deux variables (en dehors d'une corrélation avec la variable cible)



Pas de variable qui semblent fortement corrélées linéairement à la cible

Pas de variables fortement corrélées entre elles (>0,7)



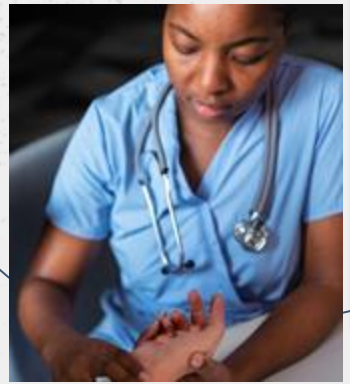
03

Tests et Sélection du modèle final

* Voir fichier ML_models_creation_with_all_variables

Tests sur plusieurs types de modèles

	KNN	LightGBM	CART	SVM
Type de modèle	Classification	Classification	Classification	Classification
Validation croisée	5 plis Stratifiée	NA	5 plis Stratifiée	NA
Temps d'entrainement	1 187 secondes (= 19 min)	Arrêté à plus de 83 minutes sans résultat	7 secondes	Arrêté à plus de 90 minutes sans résultat
Recherche des meilleurs hyperparamètres	Aléatoire (Randomized)	NA	Aléatoire (Randomized)	NA
Score Personnalisé	22 598.4	NA	22 787.2	NA
Précision	0,91	NA	0,91	NA



Précision et Score personnalisé

- La précision permet d'évaluer le modèle de façon simple : sur 100 individus, combien sont correctement classés après avoir entrainé le modèle ?

- Cependant, dans certains cas, il peut être utile de mettre en place un score personnalisé basé sur le contexte

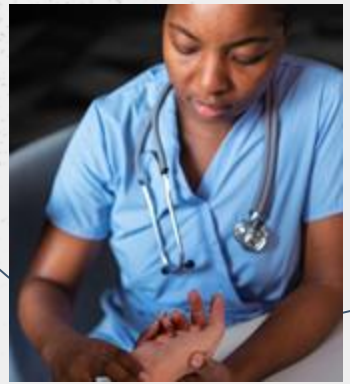
- Pourquoi un score personnalisé ? Un modèle de classification a quatre possibilités :

    - Vrai Positif (TP): Individu correctement classifié comme étant à risque (+1 point pour le modèle)
    - Vrai Négatif (TN): Individu correctement classifié comme n'étant **pas** à risque (+1 point pour le modèle)
    - Faux Positifs (FP): Individu incorrectement classifié comme à risque (-1 point pour le modèle).
    - Faux Négatif (FN): Individu incorrectement classifié comme n'étant **pas** à risque (-3 points pour le modèle)

Ici, la situation la plus grave est que le modèle indique qu'une personne n'est pas à risque alors qu'elle l'est (Faux Négatif), ce qui peut mener quelqu'un qui est à risque à ne pas consulter de médecin. J'ai donc créé un score personnalisé qui pénalise plus fortement le modèle en cas de FN
Les Faux Positifs sont aussi pénalisés, mais il est moins grave de suggérer à quelqu'un de consulter alors qu'il n'est pas à risque

Modèle retenu

	CART
Type de modèle	Classification
Validation croisée	5 plis Stratifiée
Temps d'entrainement	7 secondes
Recherche des meilleurs hyperparamètres	Aléatoire (Randomized)
Score Personnalisé	22 787.2
Précision	0,91



- Modèle Retenu : Classification And Regression Trees - CART

- Meilleur temps d'entrainement
- Meilleure score personnalisé
- Précision similaire au modèle KNN

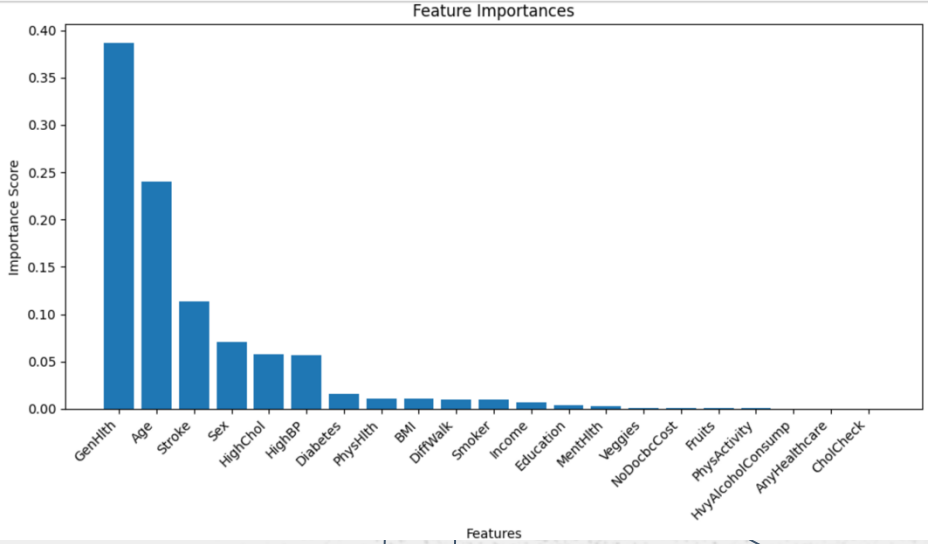
04

Sélection des champs à faire remplir

Choix des champs à faire remplir par les utilisateurs du modèle

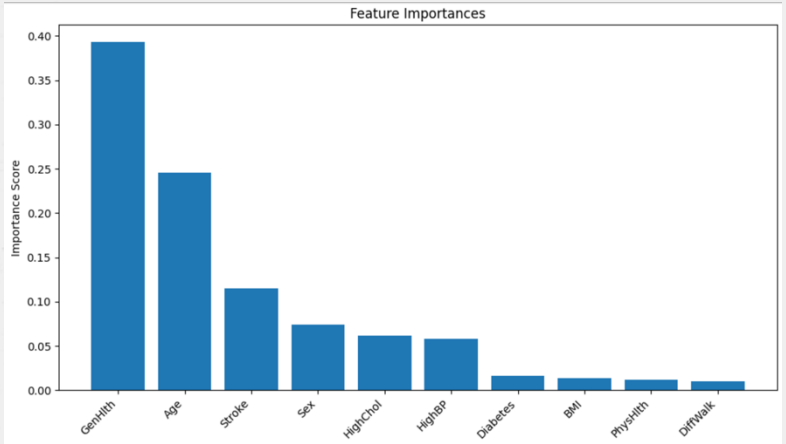
Suite à la sélection du modèle, j'ai choisi de regarder quelles variables influençaient le plus la « décision » pour éviter que les utilisateurs du modèle ne doivent répondre à 21 questions

J'ai donc identifié les variables considérées comme les plus importantes par le modèle et retenu les 10 premières



4. Sélection des champs à faire remplir

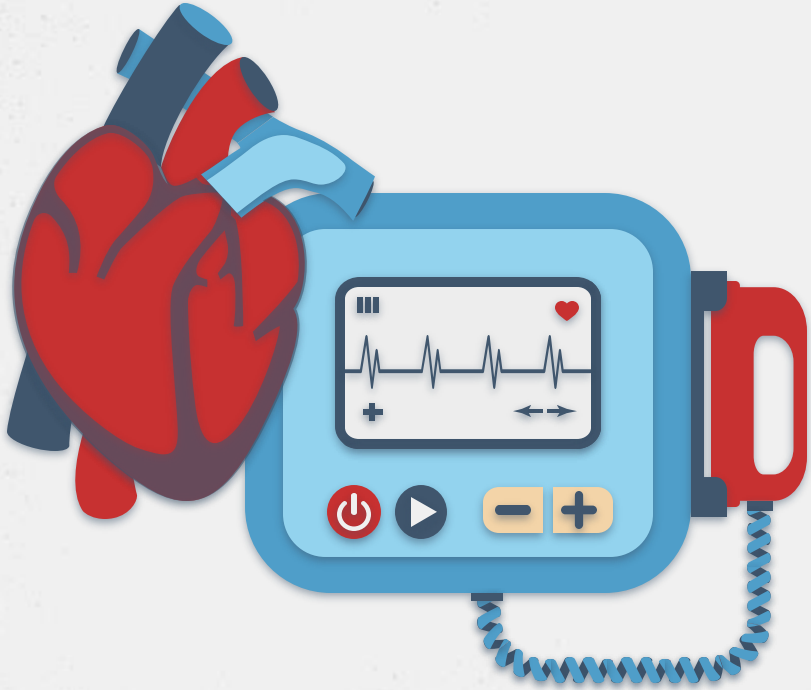
Choix des champs à faire remplir par les utilisateurs du modèle



En ne prenant en compte que ces 10 variables :

- La précision reste de 0,91
- Le score personnalisé monte à 22 826

4. Sélection des champs à faire remplir



05
Conclusion

Etapes finales

Le modèle choisi et ses hyperparamètres sont ensuite enregistrés dans un fichier pickle

Il est ensuite possible d'envoyer des données au modèle via le site web grâce à une méthode POST

La pipeline associée au modèle lors de son enregistrement :

- Va effectuer les modifications choisies en recevant ces données (par exemple la normalisation des valeurs lorsque c'est nécessaire)
- Va faire passer les données dans le modèle pour que celui-ci retourne une prédiction