DIAGNOSTIC D'UNE MALADIE CARDIAQUE

MACHINE LEARNING POUR TOUS- ML01, Groupe 21

FIORITI Federico Pascual, MAZZANTE Lorenzo, PORTELA Juana, SCORZA Martin

Pourquoi ce sujet est-t-il important ?

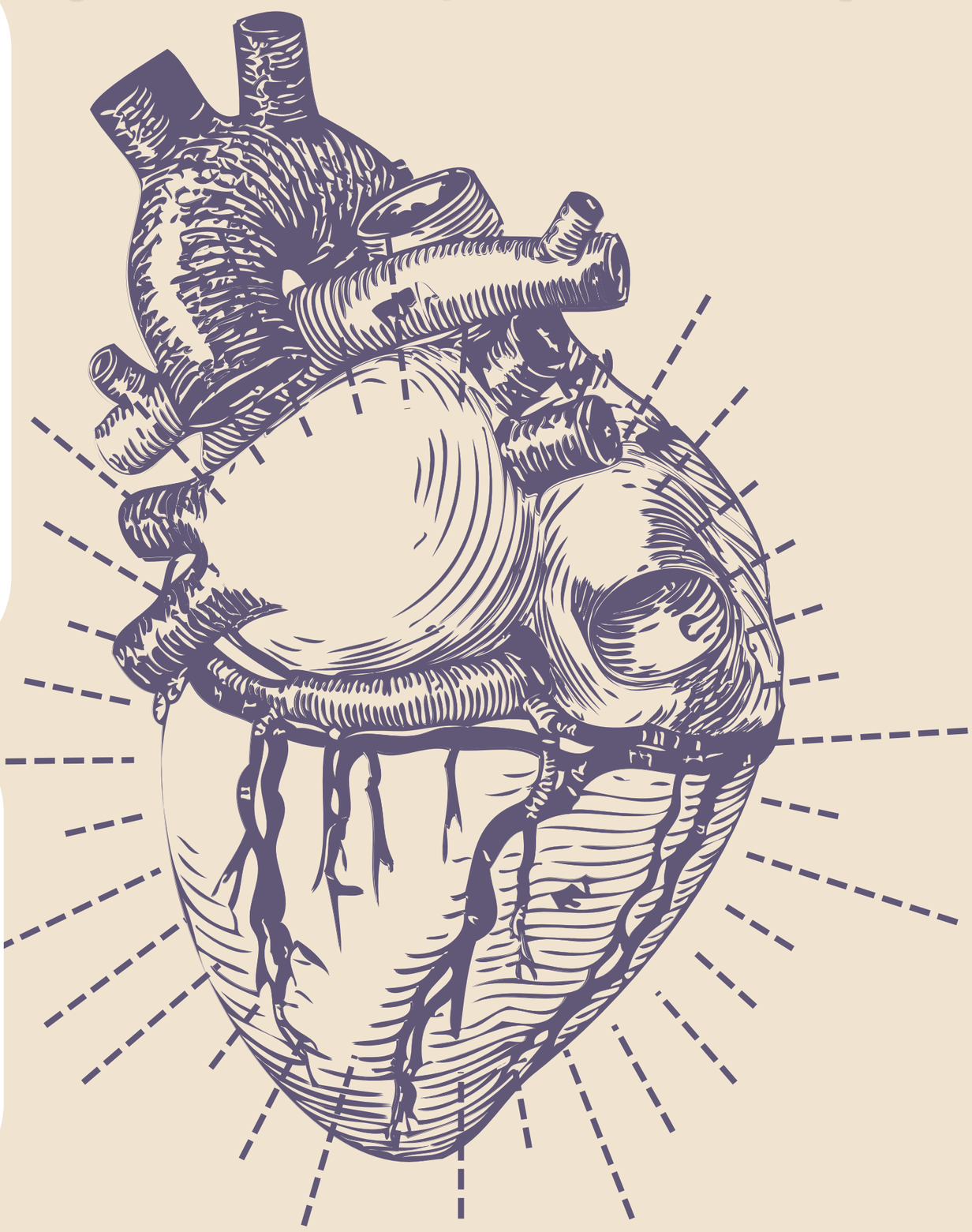
1ère cause de mortalité chez la femme

1ère cause de mortalité dans le monde

Coût des maladies cardiovasculaires: 19 milliards / an

1,2 million d'hospitalisations par an

400 décès par jour chez les adultes



Problématique

Malgré les progrès médicaux, le diagnostic d'une maladie cardiaque reste complexe. Il faut analyser de nombreux paramètres, et certaines combinaisons restent difficiles à interpréter.

Comment identifier rapidement et avec précision les patients susceptibles de développer une maladie cardiaque ?

C'est la question à laquelle essaie de répondre notre projet.

Objectif

Le but de notre travail est de développer un modèle de machine learning capable de prédire la présence ou l'absence d'une maladie cardiaque chez un patient à partir de ses données cliniques.

ÉTUDE DE LA BASE DE DONNÉS

La base de données a été récupérée sur le site web d'OpenML.

Nombre d'observations:	**70.000**
Classes:	**0 (patient en bonne santé), 1 (patient malade)**
5 variables numériques:	**âge, taille, poids, pression systolique et diastolique**
6 variables catégorielles:	**sexe, cholestérol, glucose, tabaquisme, alcool, activité physique**

Nettoyage et préparation des données

Avant l'entraînement des modèles, plusieurs étapes de pré-traitement ont été réalisées:

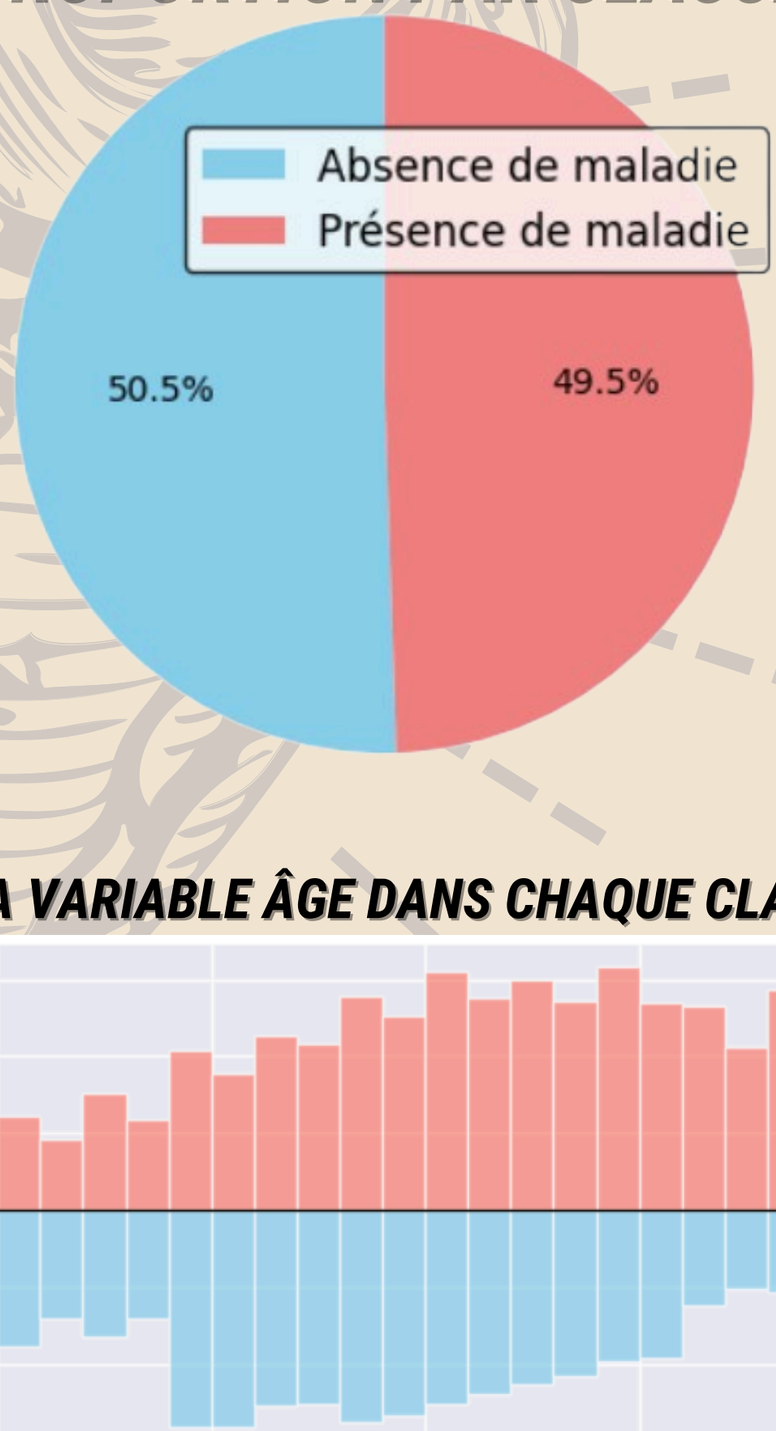
- suppression des valeurs illogiques;
- élimination des variables non pertinentes (ex. Identifiant (ID))

MATRICE DE CORRÉLATION

ON TE RÉSUME : **Prends soin de toi !**

- Plus de poids = plus de pression dans les artères
- Mauvaise alimentation = cholestérol élevé et glucose élevé
- Attention au combo cigarette + alcool
- Avec l'âge, la tension et le cholestérol ont tendance à monter
- Activité physique = bonne tension, un meilleur cholestérol et une glycémie plus stable.

PROPORTION PAR CLASSE



CARDIOPATIE SELON ACTIVITÉ PHYSIQUE

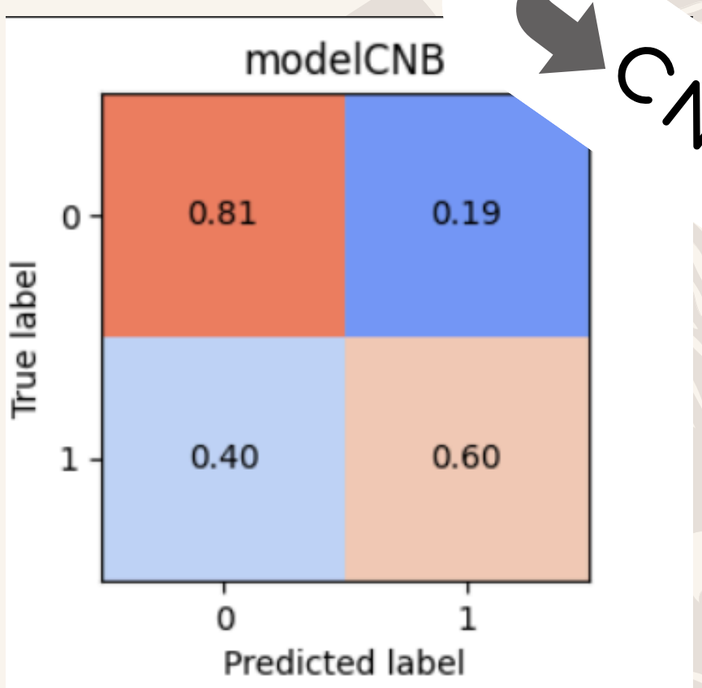
HISTOGRAMME DE LA VARIABLE ÂGE DANS CHAQUE CLASSE

Les mêmes graphiques ont été faits pour d'autres variables afin de visualiser leur distribution.

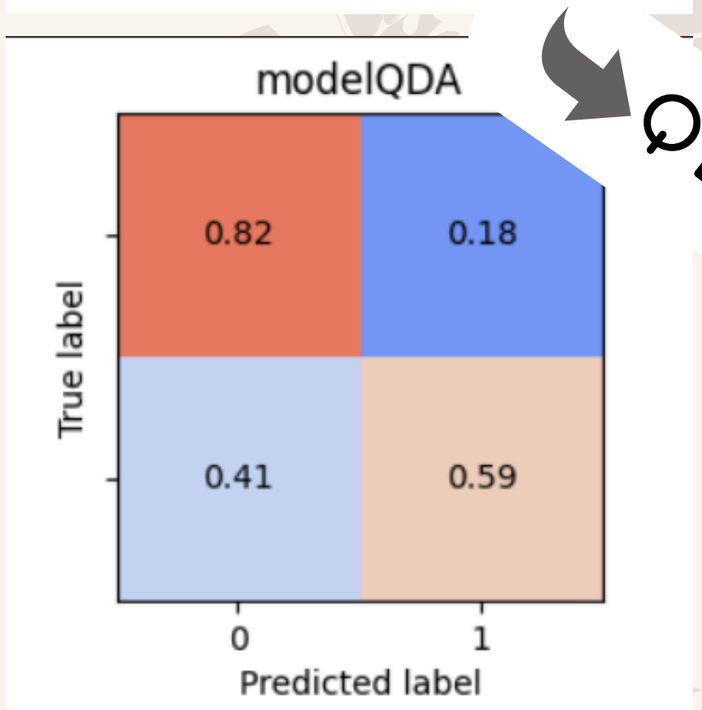
COMMENT LE PRÉDIRE ? → Méthodes de MACHINE LEARNING

MODÈLES UTILISÉS

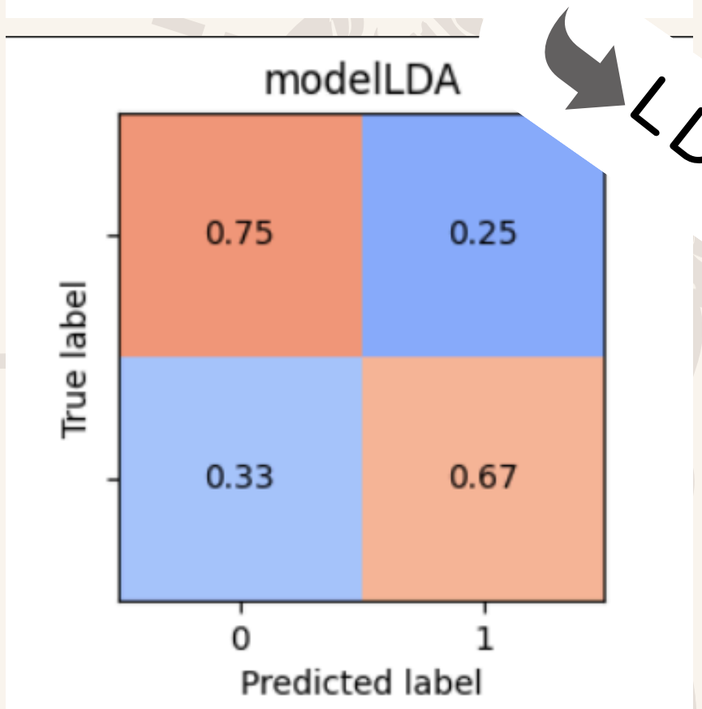
CNB



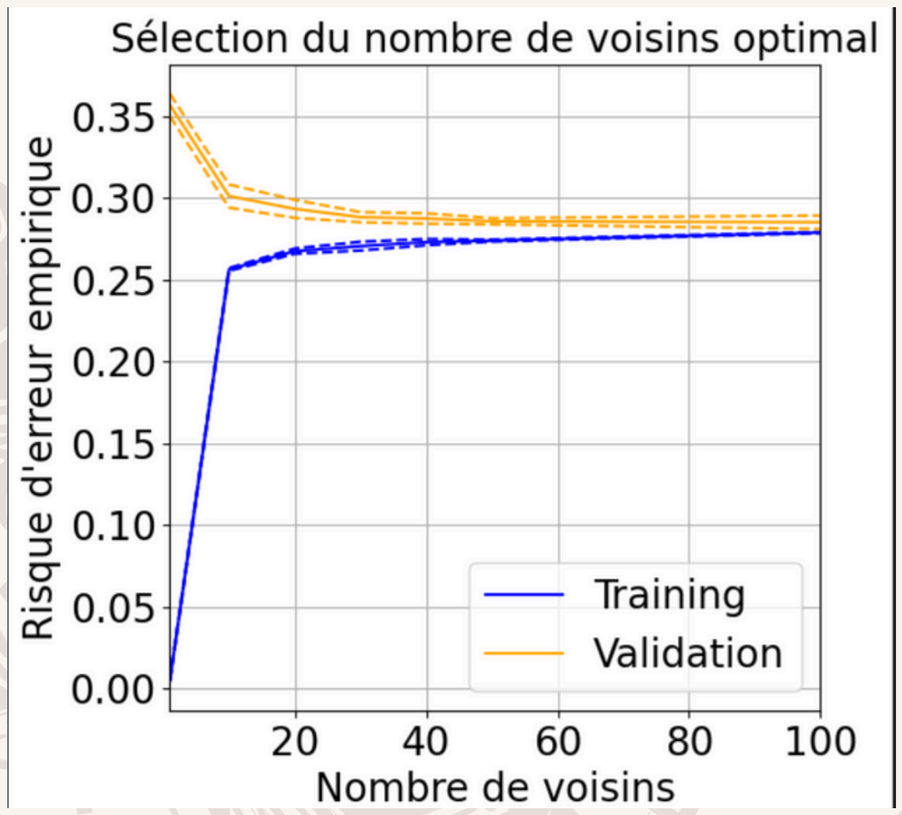
QDA



LDA



PPV

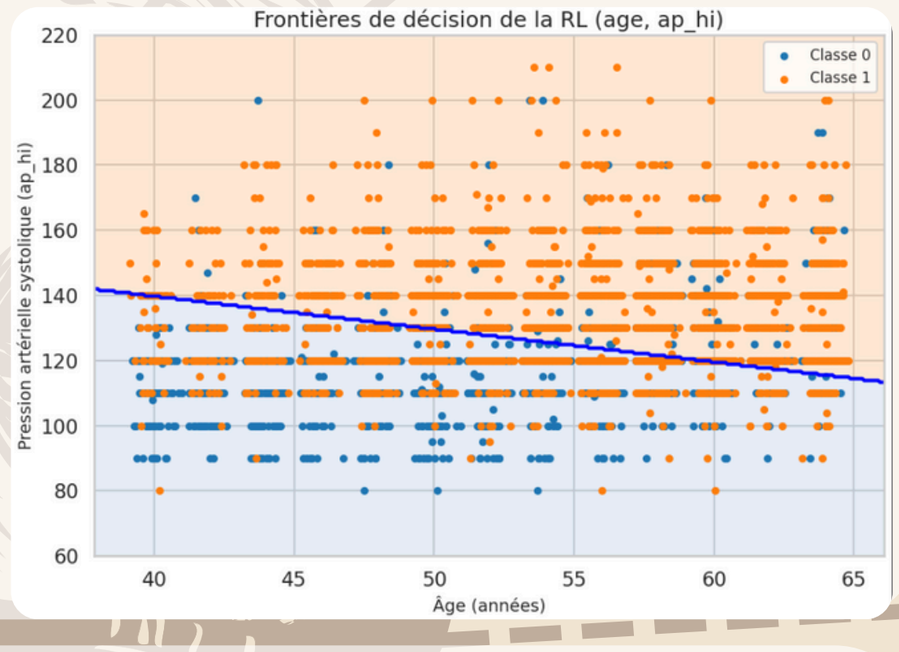


Q OPTIMAL = 40 voisins

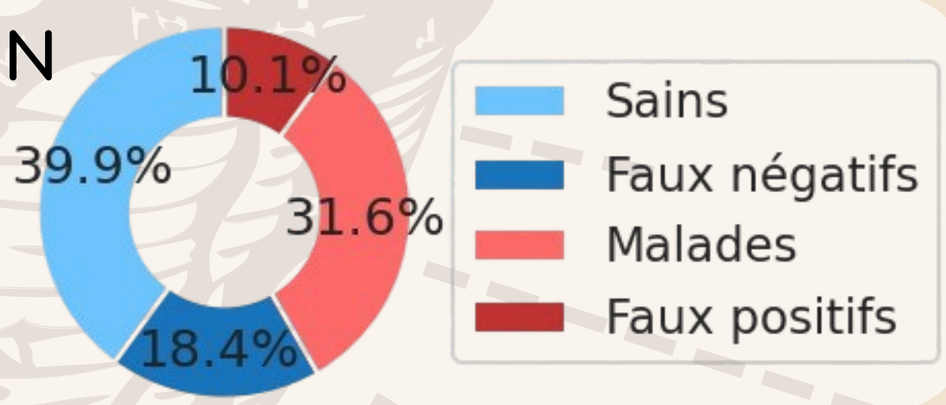
↓ *Q* → *Erreur train* ↓
Erreur test ↑
↑ *Q* → *Erreurs stables*
Erreur test ↓

REGRESSION LOGISTIQUE

Frontière de décision en prenant en compte les variables les plus pertinentes : **Âge et pression systolique**

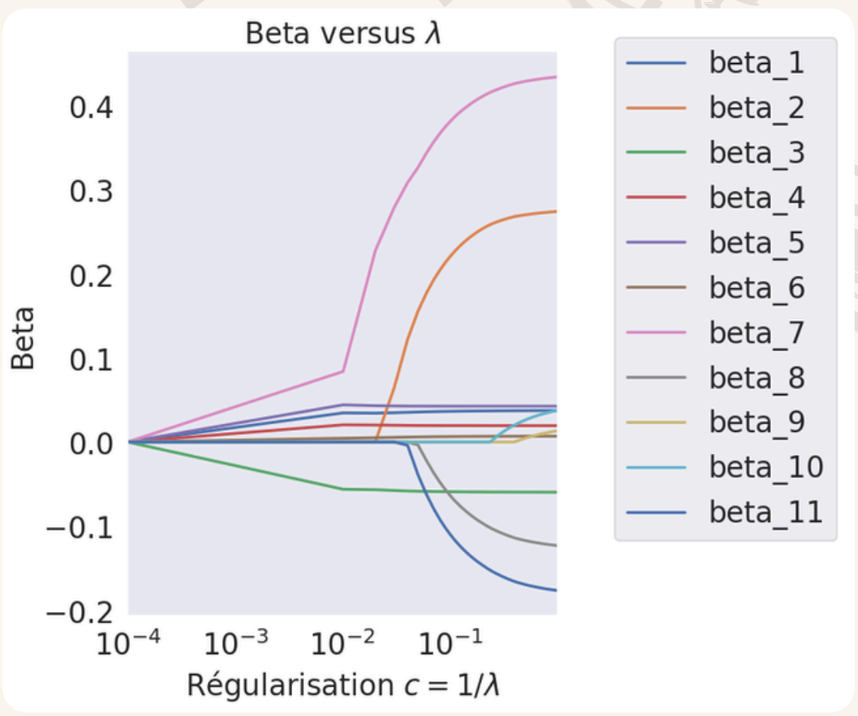


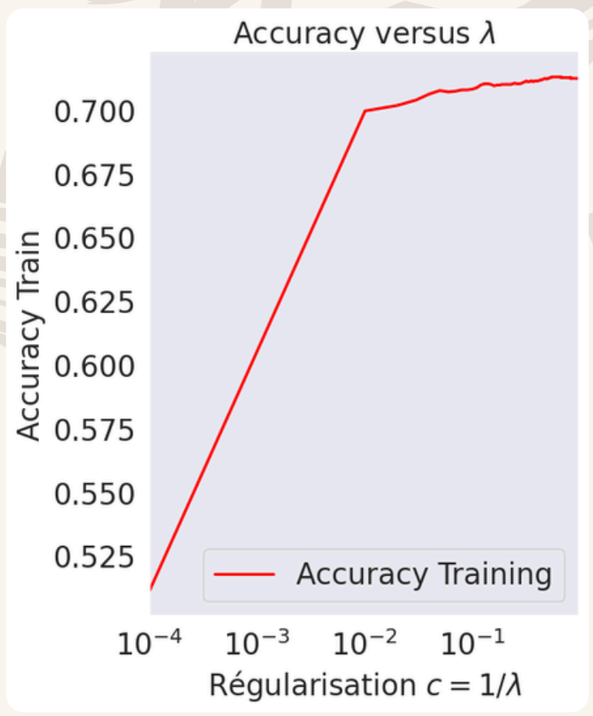
EUCLEDIAN



REGRESSION LOGISTIQUE OPTIMISÉE:

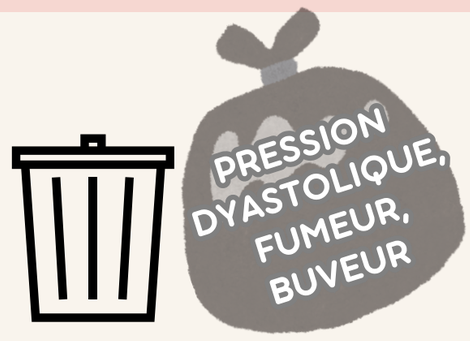
Y A-T-IL DES VARIABLES QUI N'ONT PAS UN RÔLE DÉCISIF ? : MÉTHODE LASSO





COMPROMIS CHERCHÉ :

ACCURACY VS QUANTITÉ DE VARIABLES UTILISÉE



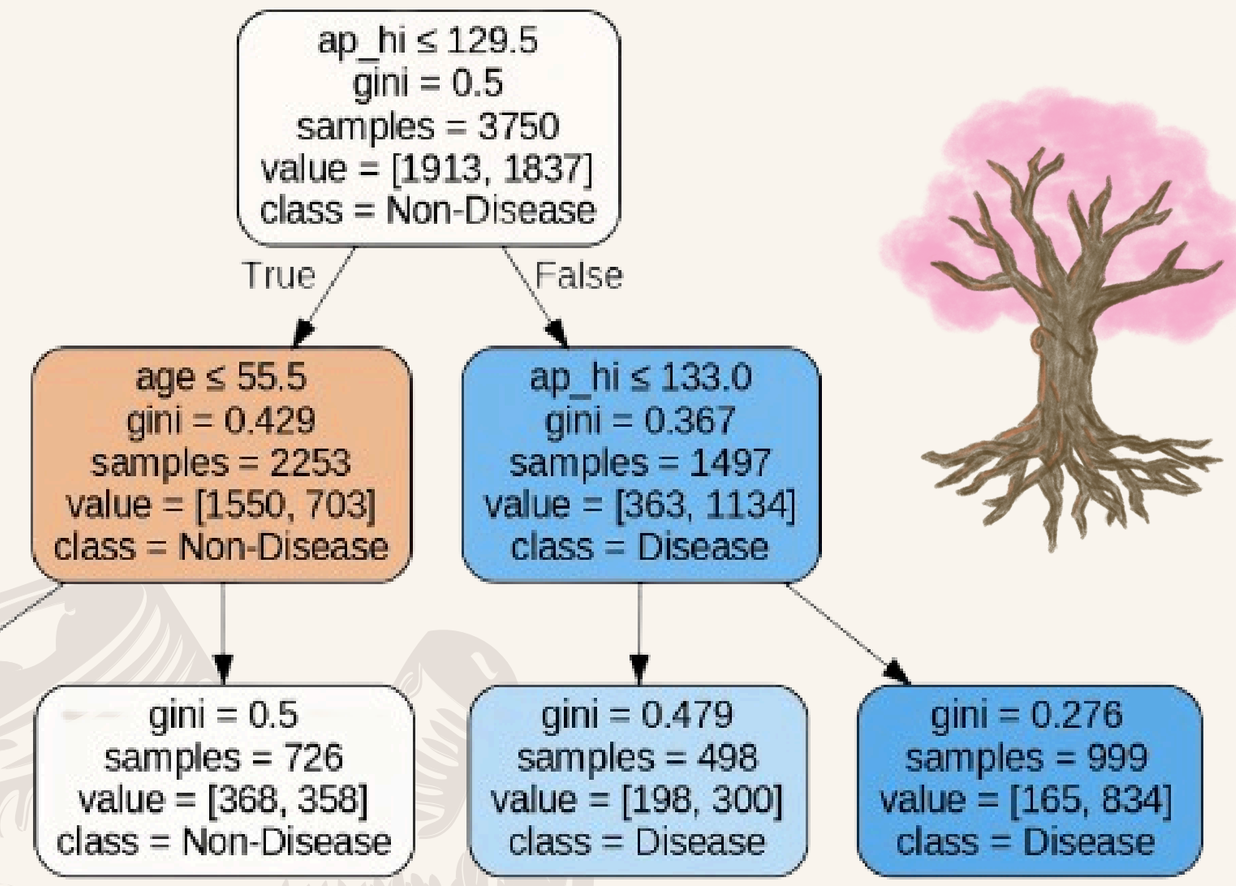
DESICION TREE

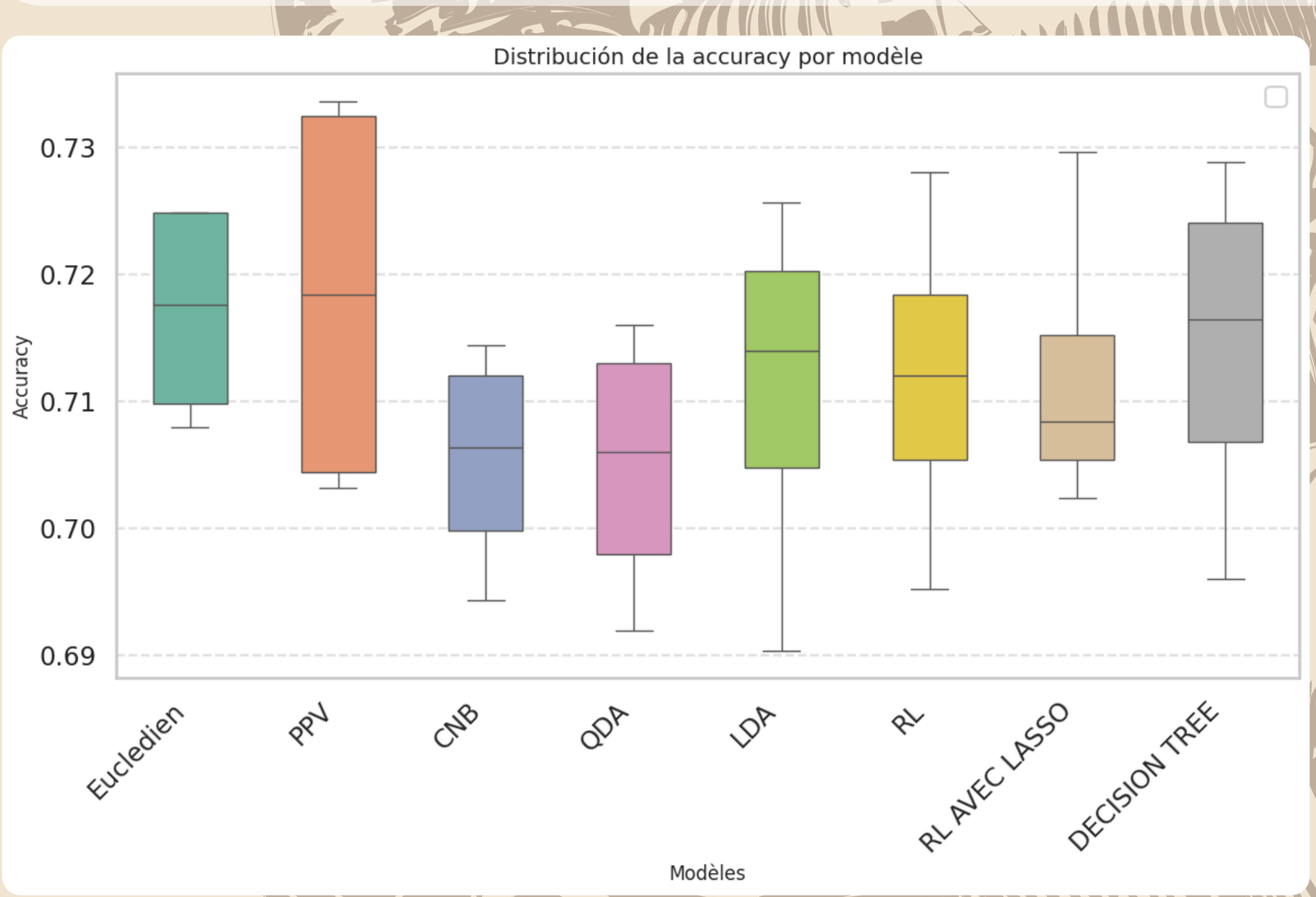
IMPORTANCE DES VARIABLES selon notre arbre:

1) pression artérielle systolique

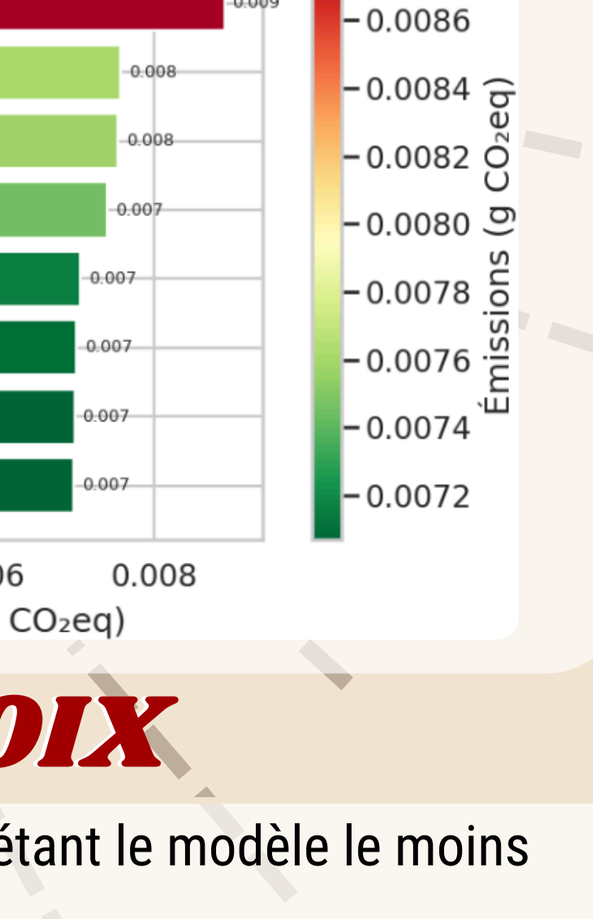
2) l'âge.

Il utilise ces deux variables pour séparer les personnes avec ou sans maladie.





Information statistique sur la distribution de l'accuracy dans la base test pour chaque modèle (médiane et variabilité entre folds). Toutes les accuracies restent proches, autour de 70–72 %.



L'EFFET SUR NOTRE PLANÈTE

Les barres montrent les émissions moyennes de CO_2 de chaque modèle.

CONCLUSION: NOS CHOIX

Eucledian	Il présente une très haute précision, tout en étant le modèle le moins consommateur.
Plus Proches Voisins	C'est le modèle qui offre la meilleure précision, mais aussi celui qui consomme le plus de ressources.
Arbre de décision	Il présente une très haute précision, tout en affichant une émissions de carbone moyenne