# Description du jeu de données

(<https://plu.mx/plum/a?mendeley_data_id=wmhctcrt5v&theme=plum-bigben-theme> )

Notre jeu de données contient 1319 observations et 9 variables, dont la variables cibles.

Une image contenant texte, reçu, capture d’écran, Police

Description générée automatiquement

*Signification de chaque variable \**

Une image contenant texte, capture d’écran, Police, reçu

Description générée automatiquement

Une étude du dataset nous révèle que les individus qui le composent ont en moyenne 59 ans.

66% d’hommes et 34% de femmes

*Etc : a dvpt avec la signification de chaque variable*

Une image contenant texte, capture d’écran, Rectangle, diagramme

Description générée automatiquement

61% de cas positifs et 39% de cas négatifs

Une image contenant texte, capture d’écran, diagramme, carré

Description générée automatiquement

# Preprocessing

Le preprocessing , ou le pré-traitement des données, est une étape cruciale. L’objectif est de rendre les données utilisables pour l’apprentissage et l’implémentation des modèles.

Le prétraitement des données inclue :

* La suppression des valeurs aberrantes
* La gestion des valeurs manquantes
* La normalisation ou la standardisation des variables numériques
* L’encodage des variables qualitatives / catégorielles

Il s’agit donc de corriger les erreurs (outliers, bruits…) pour obtenir des données précises et cohérentes pour des modèles fiables.

Dans notre cas voici les traitements opérés :

* Nettoyage des valeurs aberrantes

*Thomas je te laisse expliquer ce que t’as fait et pour quoi on garde que preprocess2*

* Encodage

L’encodage consiste à convertir la façon dont l’information est présentée. En claire, c’est la façon de passer d’une variable qualitative à une variable de type numérique, tout en préservant l’information.

Les variables dont nous avons changé l’encodage sont notre variables cibles « Résult » et la variables « Gender ».

* Train-test-split

# Analyse comparative

On cherche le modèle qui obtiendra les meilleurs résultats. Etant donné que l’on veut pouvoir prédire la présence ou l’absence de maladie cardiaque, il faut définir le type d’erreur que l’on cherche à minimiser :

* Les faux positifs : prédire la présence de maladie cardiaque chez une personne à tort.
* Les faux négatifs : prédire l’absence de maladie cardiaque chez une personnes à tort.

Dans notre cas l’erreur la plus grave serait de prédire l’absence de maladie cardiaque à tort, donc on va chercher à minimiser le nombre de faux négatifs.

~~Ainsi, pour évaluer nos modèls nous utiliserons le recall, une métrique qui tient compte de ce type d’erreur : c’est la proportion de cas positifs, bien prédites par le modèle~~

Ainsi pour évaluer nos modèles nous utiliserons le f1-score, une métrique qui fait la moyenne entre les valeurs du recall et de la précision.

## Apprentissage-supervisé

Les modèles d’apprentissage supervisé se base sur des données étiquetées

* Arbre de décision

*Explications rapides*

* Forêt aléatoire
* Adaboost
* Machine à vecteur de support
* K-nearest neighbors (KNN)
* Naïve Bayes

## Apprentissage non-supervisé

* K-Means Clusturing
* Clusturing Hierarchique Agglomératif
* Clusturing DBSCAN
* Clusturing HDBSCAN

# Sélection du modèle

## Apprentissage supervisé

## Apprentissage non-supervisé