# Description du jeu de données

# Preprocessing

* Nettoyage
* Encodage
* 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