



**UEH- Analyse des données de santé :  
Epidémiologie et aide à la décision**

# **ANALYSE DES MALADIES CARDIAQUES**

**Par AZEMA Camille et LIN  
Mengting**

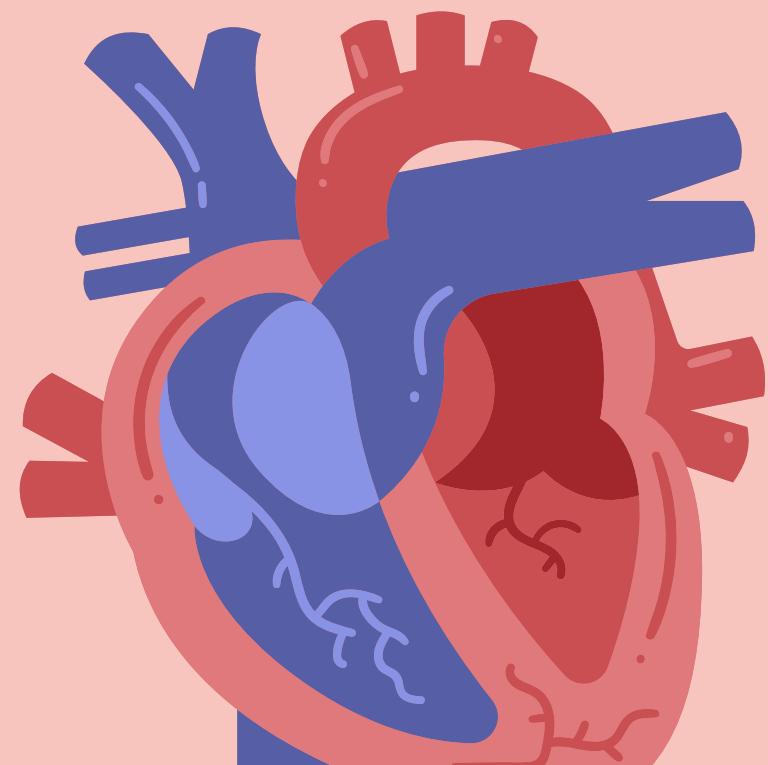
# SOMMAIRE

1. Contexte

2. Exploration du jeu de données

3. Approche supervisée

4. Conclusion



# CONTEXTE

- Les **maladies cardiaques** : 2ème cause de mortalité en France d'après l'Agence Nationale de Santé Publique.
- Préoccupation majeure de santé publique à toutes les échelles en raison de sa **prévalence croissante**.
- Souvent liées à des modes de vie incluant un **manque d'exercice, du tabagisme, consommation d'alcool, mauvaise alimentation**.
- Mais ces modes de vie ne sont pas les seules causes.



# PROBLÉMATIQUE

Source :

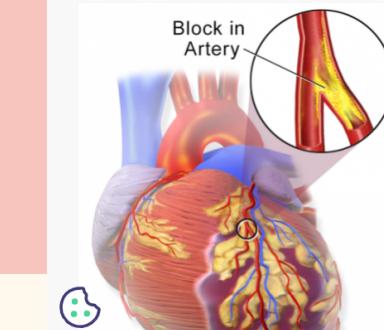
IEEEDataPort™

DATASETS SUBMIT A DATASET COMPETITIONS Q SEARCH

Open Access

## Datasets

### HEART DISEASE DATASET (COMPREHENSIVE)



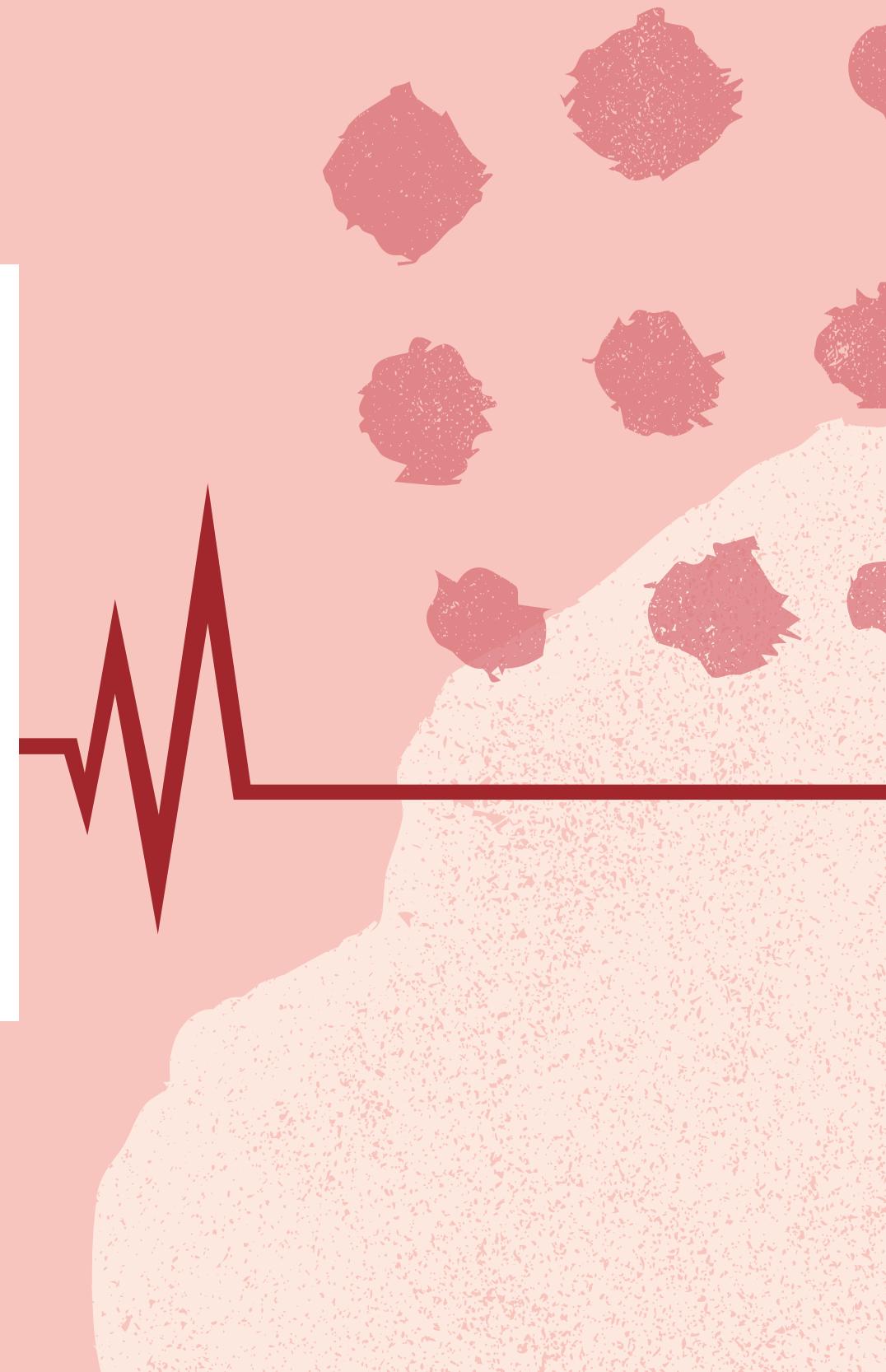
Citation Author(s): Manu Siddhartha (Liverpool John Moore's University)  
Submitted by: MANU SIDDHARTH  
Last updated: Fri, 11/06/2020 - 04:17  
DOI: 10.21227/dz4t-cm36  
Data Format: \*.csv  
Links: A database for using machine learning and data mining techniques for coronary artery disease diagnosis  
License: Creative Commons Attribution

41750 Views  
2 Citations  
Machine Learning, Health, Biomedical and Health Sciences, Heart Disease, Coronary artery disease, Cardiovascular disease, heart disease dataset

- Identification des facteurs qui pourraient favoriser l'apparition de ces maladies
- Jeu de données de 11 caractéristiques quantitatives et qualitatives de 1190 instances
- Combinaison de 5 ensembles de données populaires sur les maladies cardiaques déjà disponibles indépendamment : un des plus grands jeu de données sur les maladies cardiaques disponible à des fins de recherche
- Les 5 ensembles de données proviennent de : Cleveland, Suisse, Hongrie, Virginie et données de Statlog.

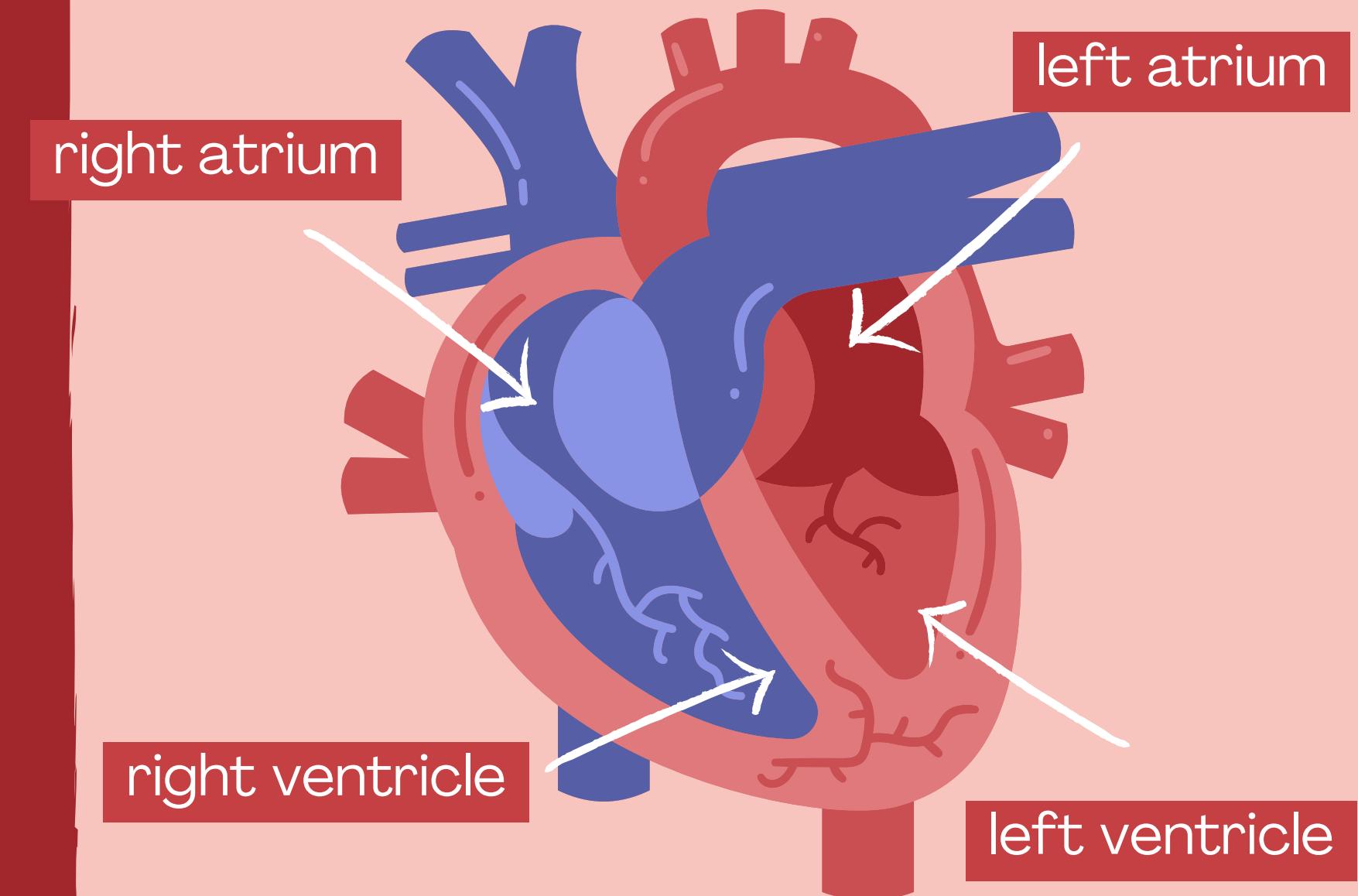
# BASE DE DONNÉES

S.No.	Attribute	Code given	Unit	Data type
1	age	Age	in years	Numeric
2	sex	Sex	1, 0	Binary
3	chest pain type	chest pain type	1,2,3,4	Nominal
4	resting blood pressure	resting bp s	in mm Hg	Numeric
5	serum cholesterol	cholesterol	in mg/dl	Numeric
6	fasting blood sugar	fasting blood sugar	1,0 > 120 mg/dl	Binary
7	resting electrocardiogram results	resting ecg	0,1,2	Nominal
8	maximum heart rate achieved	max heart rate	71–202	Numeric
9	exercise induced angina	exercise angina	0,1	Binary
10	oldpeak =ST	oldpeak	depression	Numeric
11	the slope of the peak exercise ST segment	ST slope	0,1,2	Nominal
12	class	target	0,1	Binary



# EXPLORATION DE LA BASE DE DONNÉES

Dans cette partie nous allons décrire les différents outils de statistiques utilisés pour décrire notre jeu de données.



# EXPLORATION

## Vérification de l'équilibre du jeu de données et données manquantes

```
> round(prop.table(table(heart$target)), 2)
```

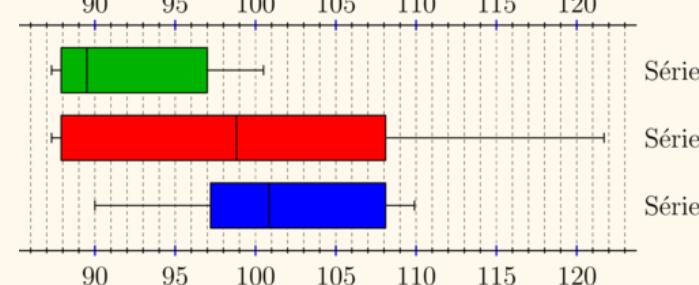
Health	Disease
0.47	0.53

- Variable d'intérêt “target” : variable binaire avec comme valeur “Healthy” ou “Heart disease”
- Equilibrée puisqu'il y a **53% de personnes atteintes de maladies cardiaques** dans le jeu de données et **47% de personnes en bonne santé**.
- Dans notre cas nous n'avions pas de données manquantes.

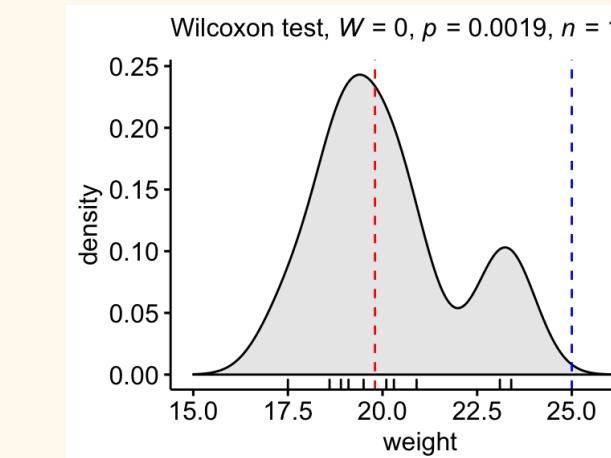
# EXPLORATION

Outils statistiques utilisés

Boîtes à  
moustache/histogram  
mes

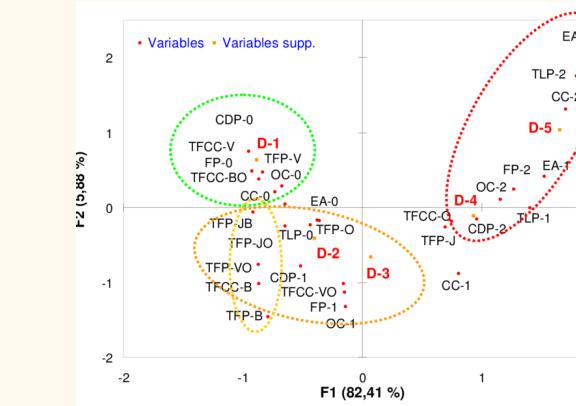


Test de  
Wilcoxon



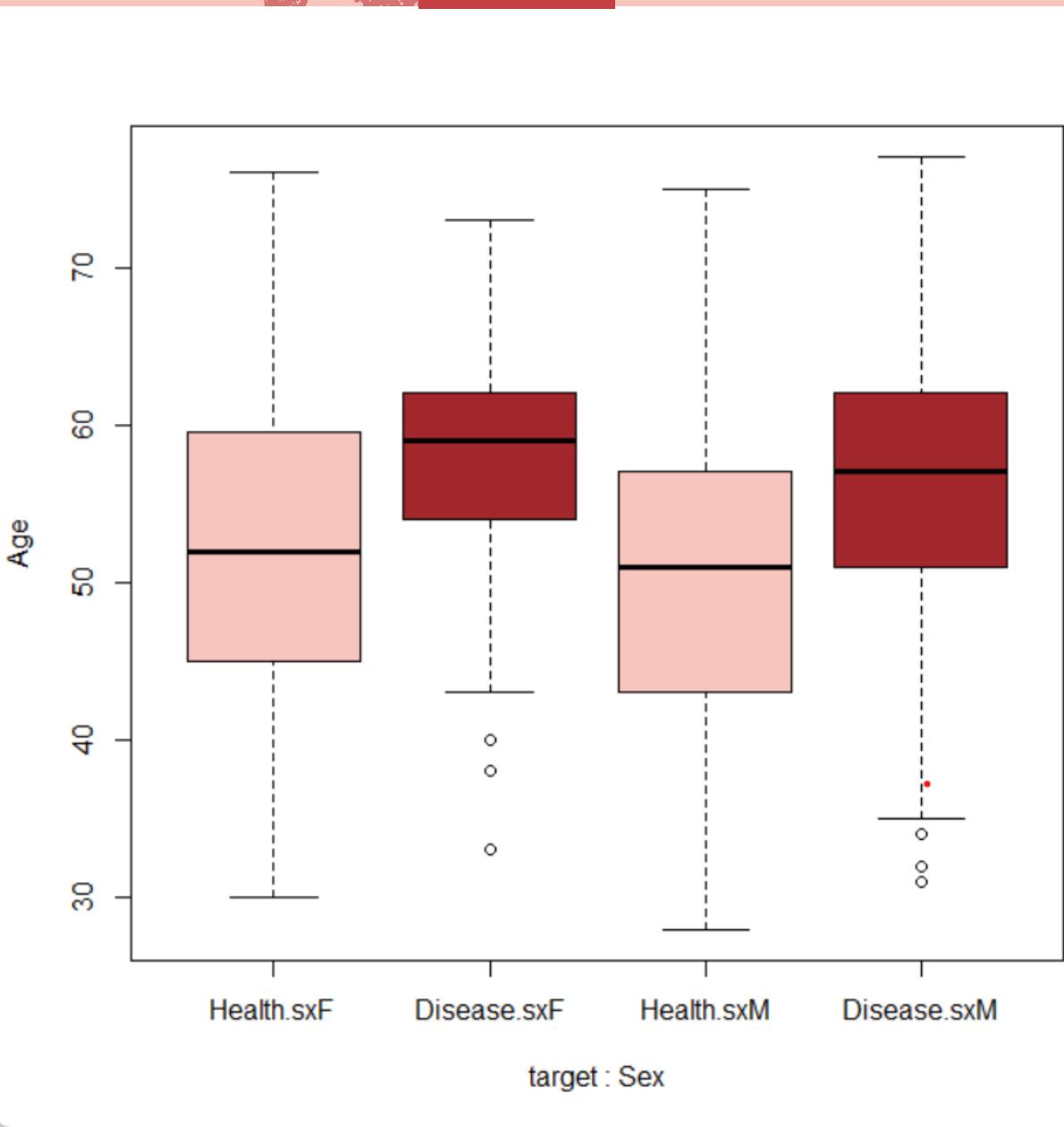
ACM

Analyse des  
correspondances multiples

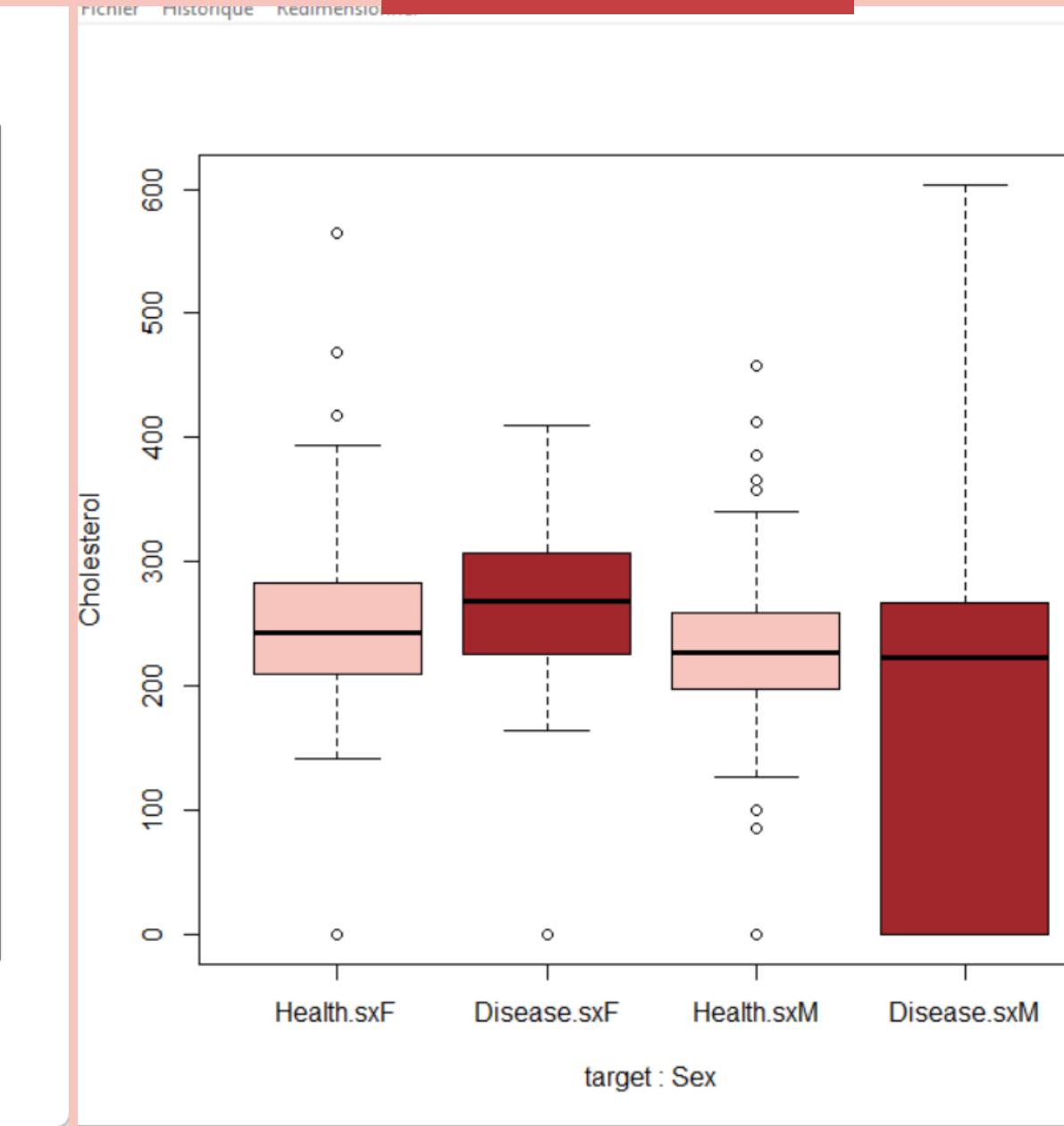


# BOÎTES À MOUSTACHE

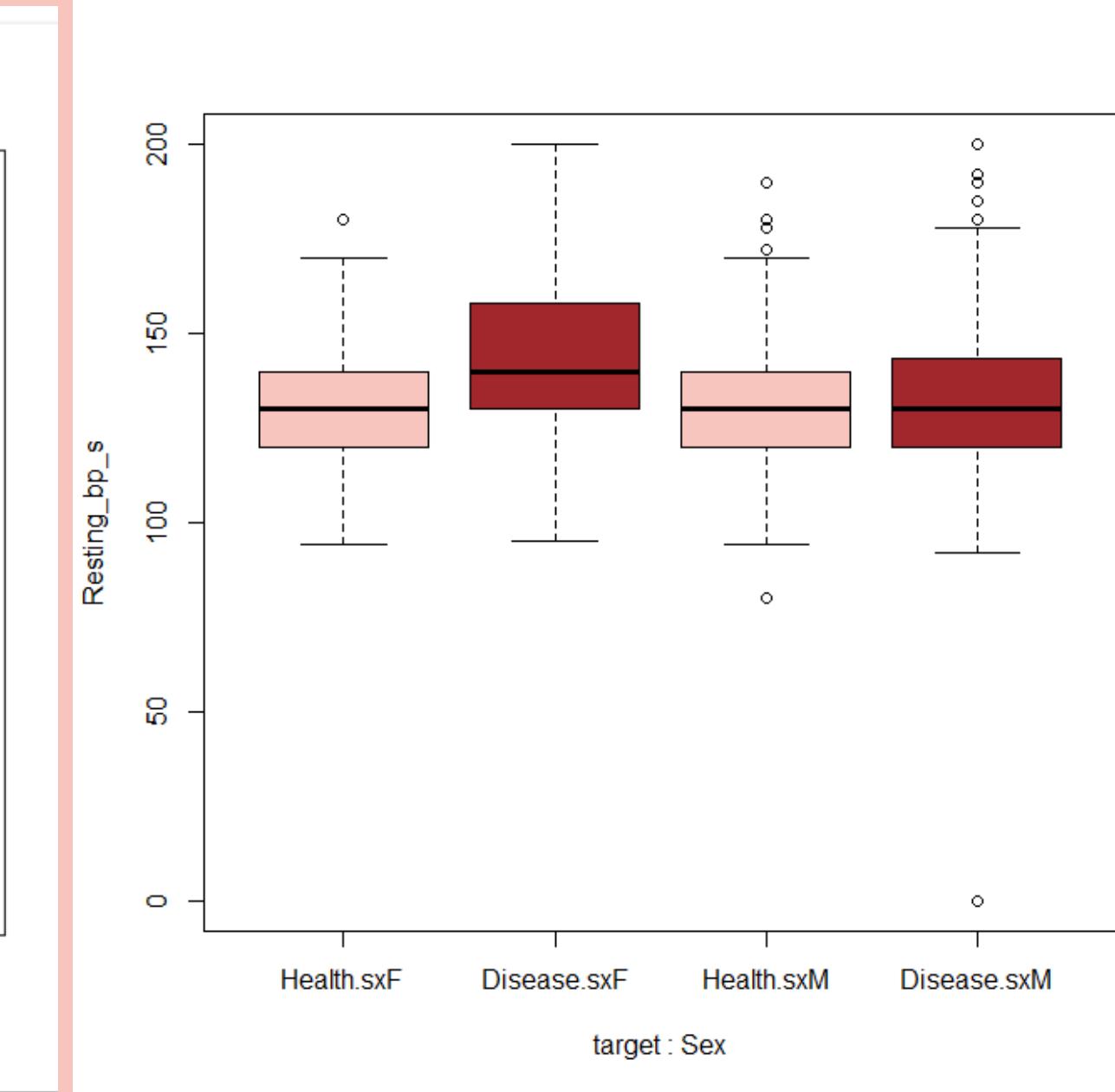
Age



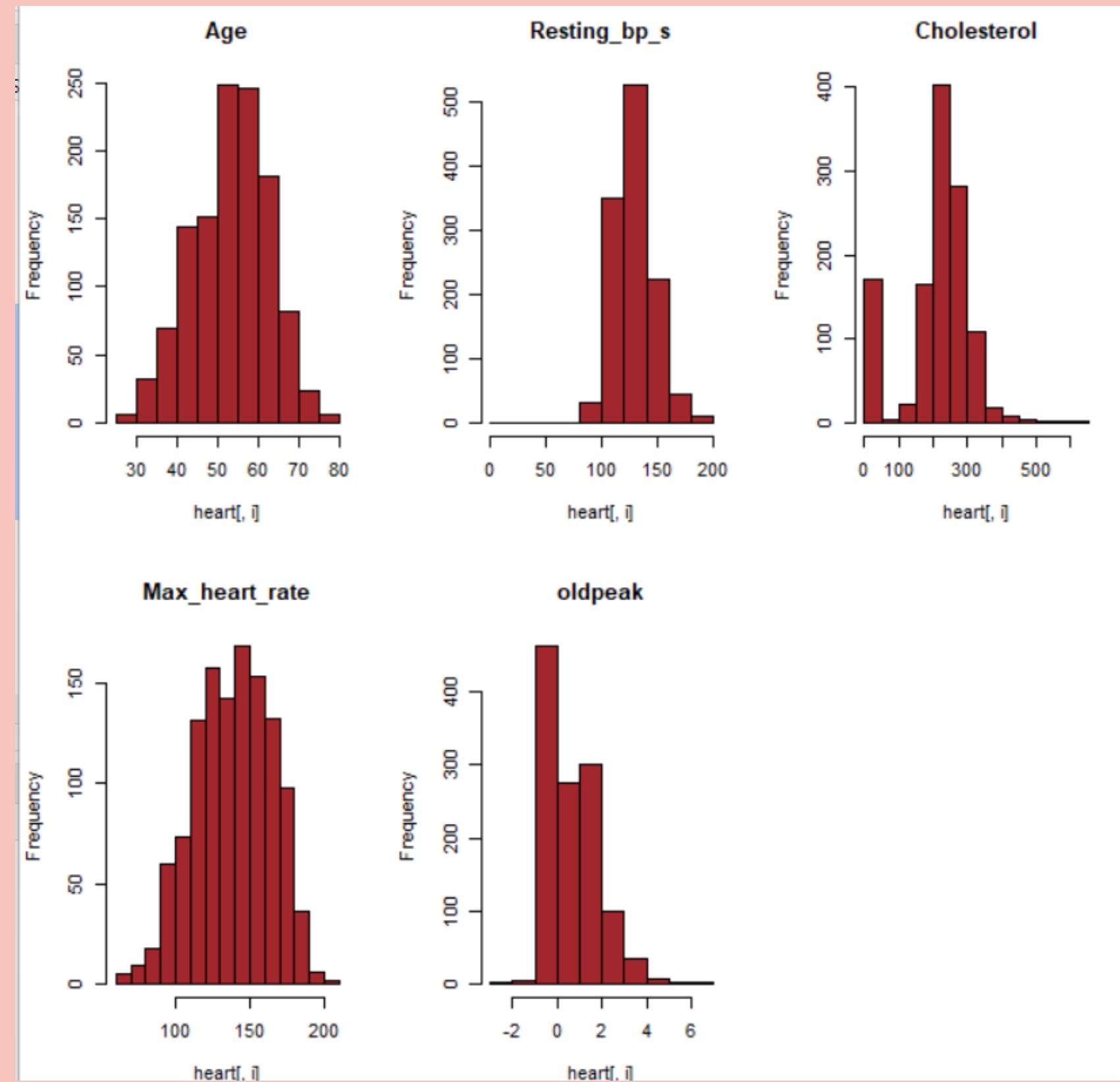
Cholestérol



BPM au repos



# HISTOGRAMMES



# TEST DE WILCOXON

## Principe

- Test **non paramétrique** utilisé pour déterminer s'il existe une différence significative entre deux **échantillons indépendants.**
- **Pas de supposition** que les données suivent une **distribution normale** contrairement aux tests paramétriques comme Student ou Chi2.

## Résultats

```
> wilcox.test(Age~target,data=heart)

Wilcoxon rank sum test with continuity correction

data: Age by target
W = 121014, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(Resting_bp_s~target,data=heart)

Wilcoxon rank sum test with continuity correction

data: Resting_bp_s by target
W = 151646, p-value = 2.63e-05
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(Cholesterol~target,data=heart)

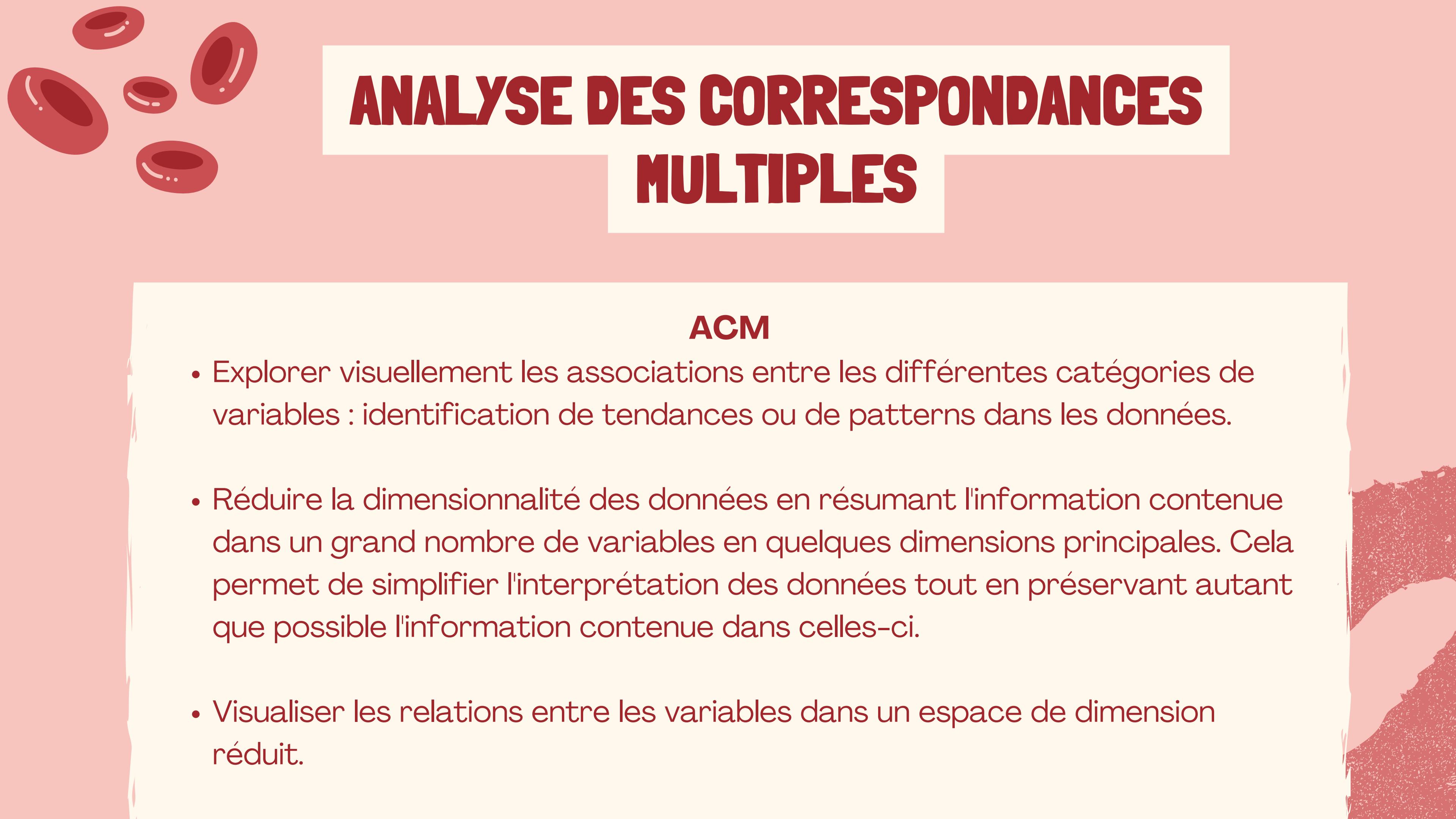
Wilcoxon rank sum test with continuity correction

data: Cholesterol by target
W = 194094, p-value = 0.002804
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(Max_heart_rate~target,data=heart)

Wilcoxon rank sum test with continuity correction

data: Max_heart_rate by target
W = 261978, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

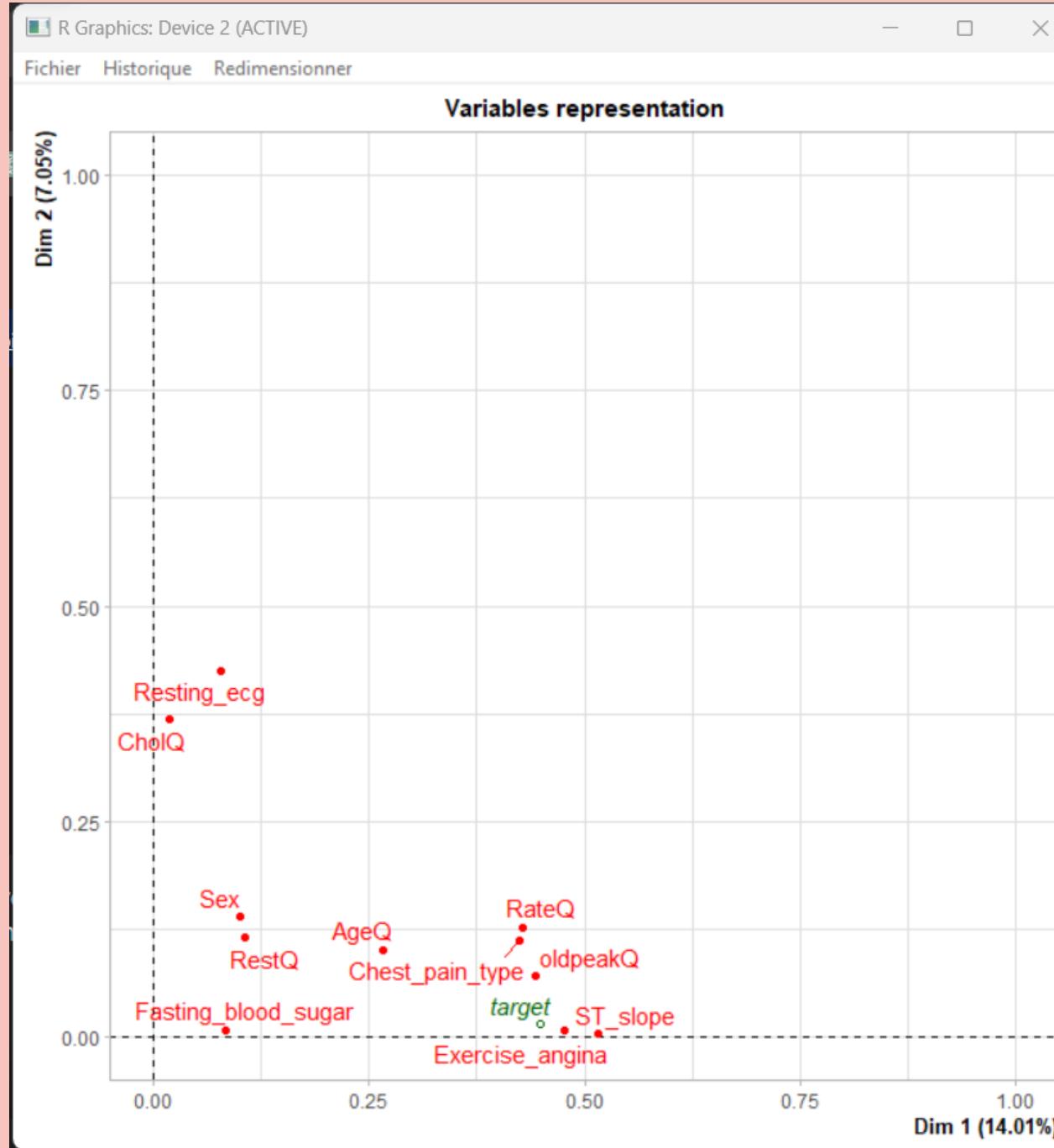


# **ANALYSE DES CORRESPONDANCES MULTIPLES**

## **ACM**

- Explorer visuellement les associations entre les différentes catégories de variables : identification de tendances ou de patterns dans les données.
- Réduire la dimensionnalité des données en résumant l'information contenue dans un grand nombre de variables en quelques dimensions principales. Cela permet de simplifier l'interprétation des données tout en préservant autant que possible l'information contenue dans celles-ci.
- Visualiser les relations entre les variables dans un espace de dimension réduit.

# ANALYSE DES CORRESPONDANCES MULTIPLES



## Résultats

- Dim1 et dim 2 : combinaisons linéaires des variables originales qui expliquent la plus grande part de la variance dans les données.
- L'étiquette de chaque axe indique le pourcentage de variance qu'il explique

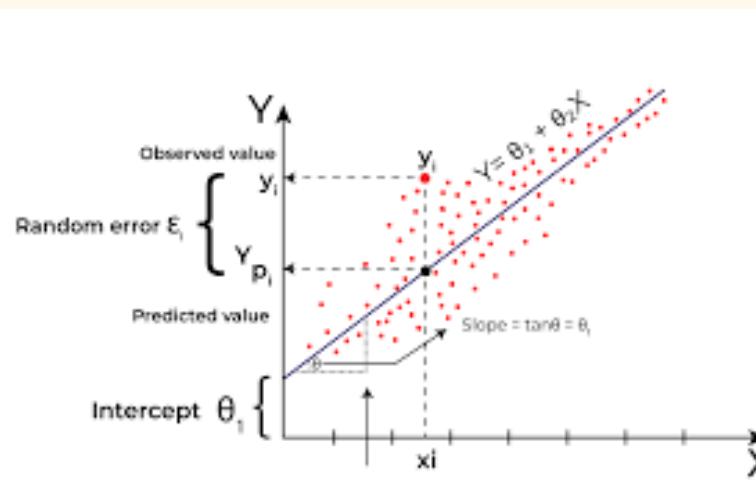
# APPROCHE SUPERVISÉE

Dans cette partie nous allons commenter les 3 méthodes de Machine Learning que nous avons utilisées.

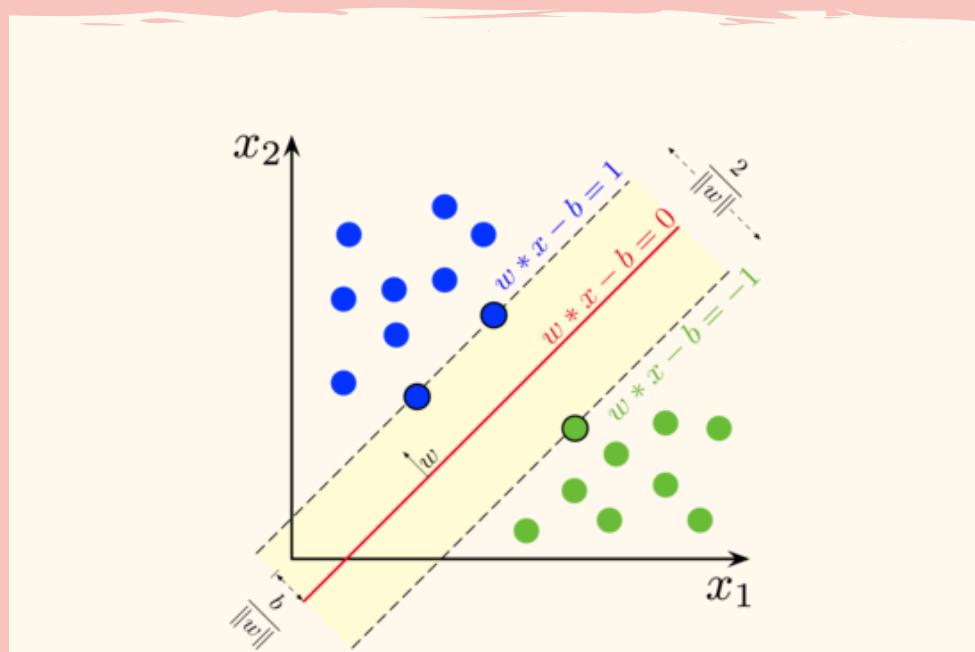


# APPROCHE SUPERVISÉE

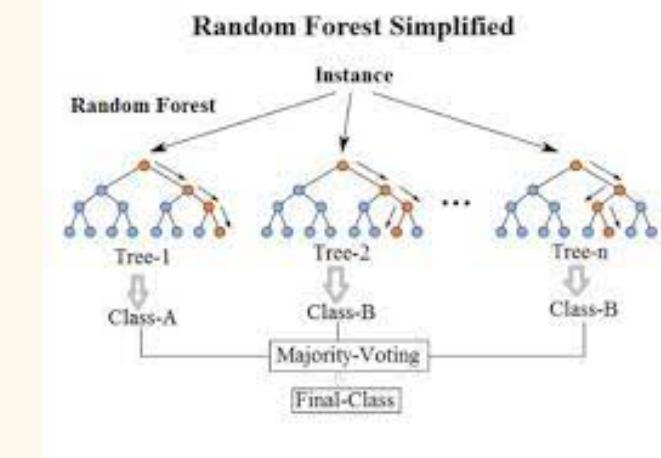
Méthodes de Machine Learning utilisées



Linear  
Regression



SVM



Random  
Forest

# PRÉPARATION DES DONNÉES POUR LA CLASSIFICATION BINAIRE

Objectif : Prédire la présence de maladie cardiaque.

- Séparation des données : **70% pour l'apprentissage, 30% pour le test**, assurant une évaluation robuste.
- **Reproductibilité** : Fixation de la graine aléatoire pour des résultats constants.
- Distribution de la cible : Vérification de l'équilibre entre les ensembles pour éviter les biais.
- **Validation croisée** : 5 plis pour une estimation précise de la performance du modèle.

# LINEAR REGRESSION

- La régression linéaire est un modèle statistique applicable aux problèmes de régression et de classification, capable de traiter à la fois les variables de réponse continues et les distributions binomiales.
- L'AIC, ou critère d'information d'Akaike, est un indicateur mesurant à la fois l'ajustement et la complexité du modèle ; plus la valeur de l'AIC est faible, meilleur est le modèle.
- La fonction `glmStepAIC` effectue automatiquement l'ajustement du modèle, le calcul de l'AIC et l'ajout ou la suppression de variables jusqu'à ce que la meilleure valeur d'AIC soit atteinte.

# LINEAR REGRESSION

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.217006	1.160653	-1.049	0.294384
Age	0.024983	0.012724	1.963	0.049598 *
SexsxM	1.570204	0.279974	5.608	2.04e-08 ***
Chest_pain_typeV4	1.735794	0.226928	7.649	2.02e-14 ***
Cholesterol	-0.003502	0.001188	-2.948	0.003203 **
Fasting_blood_sugarF	0.954620	0.280024	3.409	0.000652 ***
Max_heart_rate	-0.009750	0.005035	-1.936	0.052817 .
Exercise_anginaoui	0.783773	0.239702	3.270	0.001076 **
oldpeak	0.453442	0.118147	3.838	0.000124 ***
ST_slopeUp	-1.907823	0.249505	-7.646	2.07e-14 ***
ST_slopeDown	-1.478676	0.434324	-3.405	0.000663 ***
---				

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1153.41 on 833 degrees of freedom  
Residual deviance: 590.97 on 823 degrees of freedom  
AIC: 612.97

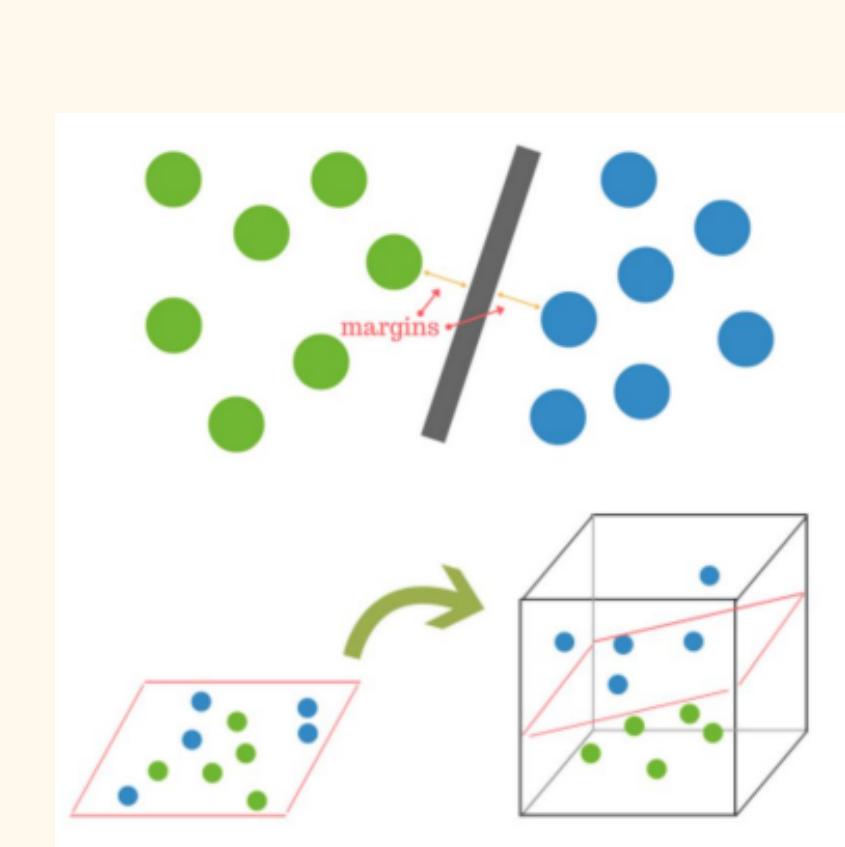
```
prediction_glm_tuned Health Disease
                           Health      144      23
                           Disease      24     165
>
> err_glm_tuned <- 1-(sum(diag(matconfu
> err_glm_tuned
[1] 0.1320225
```

**Accuracy : 86.8%**

**C'est un modèle qui est satisfaisant**

# SUPPORT VECTOR MACHINE

- Trouver l'hyperplan linéaire séparant les observations tel qu'il maximise la marge avec les observations les plus proches (vecteurs supports).
- utiliser une projection dans un espace de dimension supérieure en utilisant un noyau (linéaire, polynomial, radial) pour trouver une séparation linéaire optimale.
- Paramètres à optimiser :
- le type de noyau
  - C : la constante de tolérance, qui équilibre le nombre d'erreurs de classification et la largeur de la marge
  - sigma pour un noyau radial
  - degré pour un noyau polynomial



Dans R : fonction générique train avec  
method = “svmLinear”, “svmRadial”, etc.  
ou package e1071, fonction svm

# SUPPORT VECTOR MACHINE

```
Parameters:  
  SVM-Type: C-classification  
  SVM-Kernel: sigmoid  
    cost: 1  
   coef.0: 0
```

```
Number of Support Vectors: 345  
( 173 172 )
```

```
Number of Classes: 2
```

```
Levels:  
  Health Disease
```

```
> err_svm <- 1 - (sum(diag(conf_matrix)) / sum(conf_matrix))  
> print(err_svm)  
[1] 0.8342697
```

Nous avons utilisé les fonctions de noyau **linéaire, radial et sigmoïde**, mais les résultats des données sont presque les mêmes, avec une erreur très élevée.

**Ce n'est pas le résultat que nous attendions**

# RANDOM FOREST

## Principe de la forêt aléatoire :

- La forêt aléatoire améliore la précision prédictive en combinant les résultats de plusieurs arbres décisionnels simples, où chaque arbre prédit indépendamment et un vote majoritaire détermine le résultat final.

## Étapes de création de la forêt aléatoire :

- Préparation des données : Créer  $B$  échantillons d'entraînement par échantillonnage avec remplacement.
- Entraînement des arbres : Former un arbre sur chaque échantillon en sélectionnant un sous-ensemble de prédicteurs.
- Enregistrement des résultats : Sauvegarder les prédictions de chaque arbre.
- Décision par vote : Les prédictions de tous les arbres sont agrégées par vote pour obtenir la prédition finale.

Utiliser la fonction `train` avec `method="rf"`, ou le package **randomForest** pour entraîner un modèle de forêt aléatoire.

# RANDOM FOREST

Resampling results across tuning parameters:

mtry	ROC	Sens	Spec
2	0.9504658	0.8675430	0.9001021
9	0.9523851	0.8878286	0.9296731
16	0.9463262	0.8802986	0.9274259

```
> print(matconfus_rf)
```

prediction_rf	Health	Disease
	Health	Disease
Health	151	9
Disease	17	179

ROC was used to select the optimal model using the largest value.  
The final value used for the model was mtry = 9.

```
> err_rf <- 1 - (sum(diag(matconfus_rf)) / sum(matconfus_rf))
> print(err_rf)
[1] 0.07303371
```

**Accuracy : 92.7%**  
**C'est un modèle très satisfaisant**

# CONCLUSION

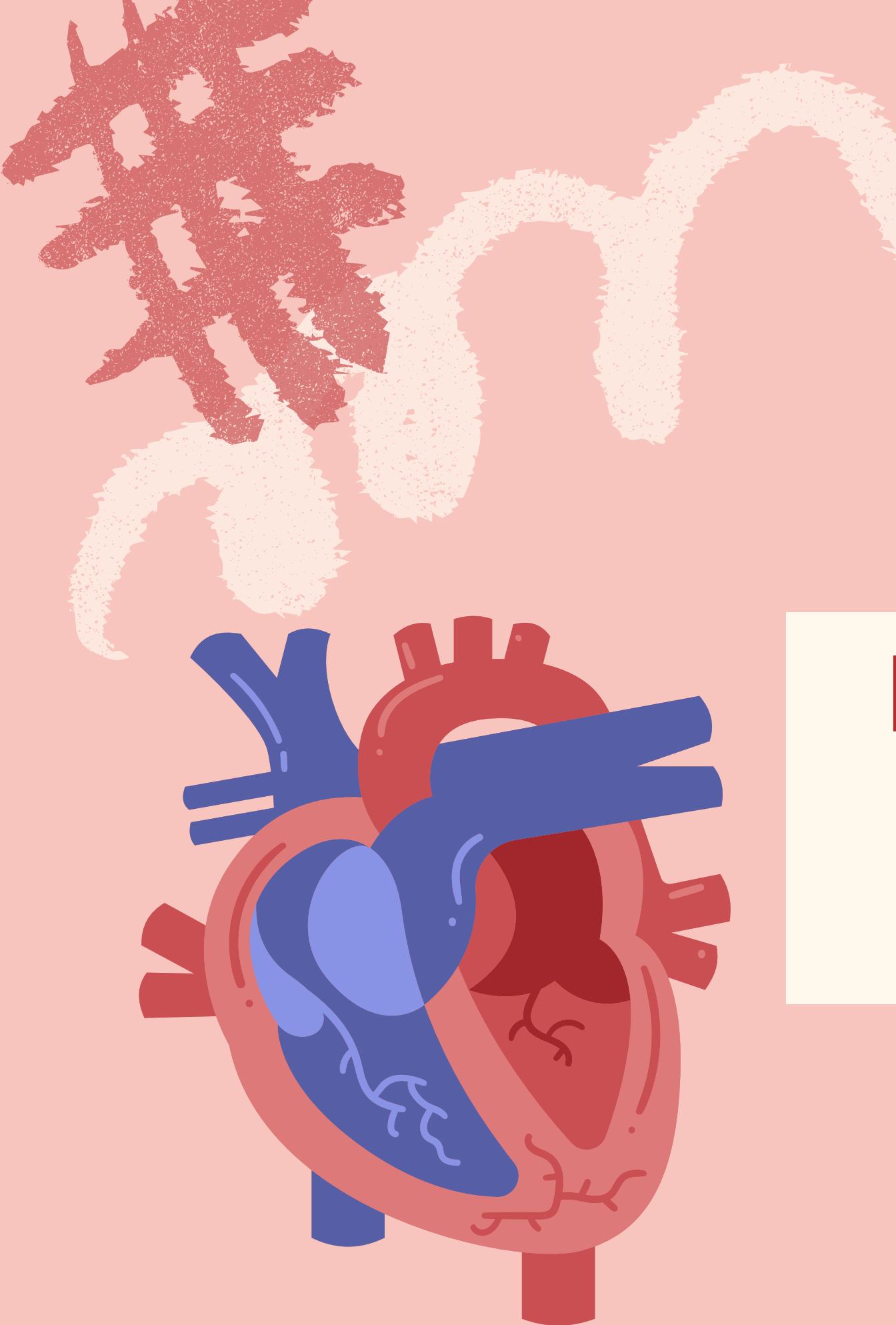
Toutes nos variables sont **significatives** et ont un rôle dans l'apparition de maladies cardiaques

Deux modèles prédictifs sortent du lot :

Les **forêts aléatoires** et la régression linéaire avec des erreurs de prédiction respectives de 13% et 7% nous pouvons les considérer comme exploitables dans un cadre de santé.

Cependant nous n'avons pas obtenu de résultats concluant avec la méthode SVM nous l'écartons.





**Merci de votre  
attention !**

Questions ?

