université
PARIS-SACLAY

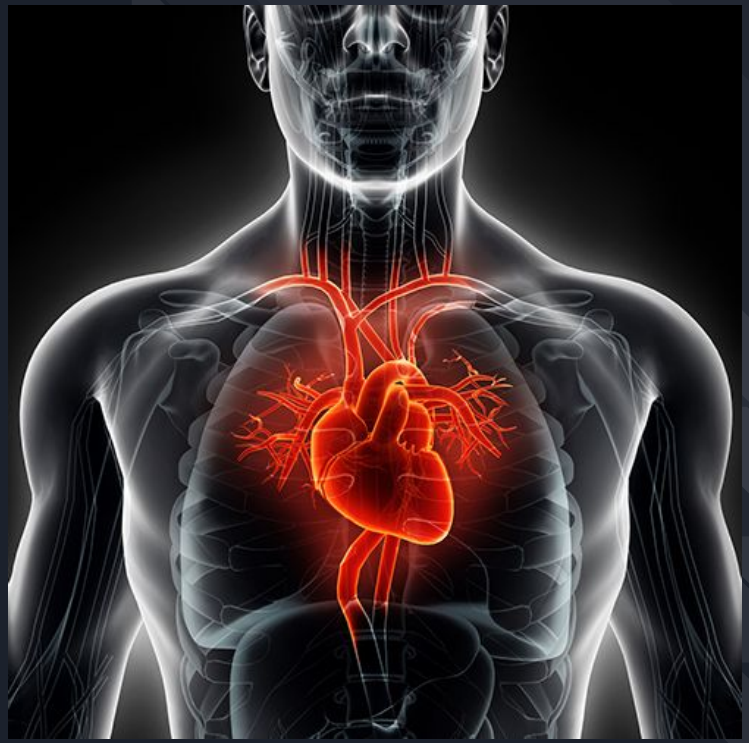
L3 INFO
Référent : Caio
2020-2021

Projet D'IAS

Markandu Jeyanthan
De Souza Albanio
Dahalani Luqman
Bourdet Jeremie
Renouard Gwenn

Sujet : Déterminer si une personne atteinte de maladies cardiovasculaire va mourir en fonction de ses antécédents (pré-processing & classification).

TABLE DES MATIÈRES



Introduction

Les Maladies Cardiovasculaires :

- agissent sur le cœur et le sang
- 140 000 décès chaque années en France (~ 400 /jour)
- De multiples facteurs les causes

Introduction

Le Dataset :

- 13 features
- 299 lignes

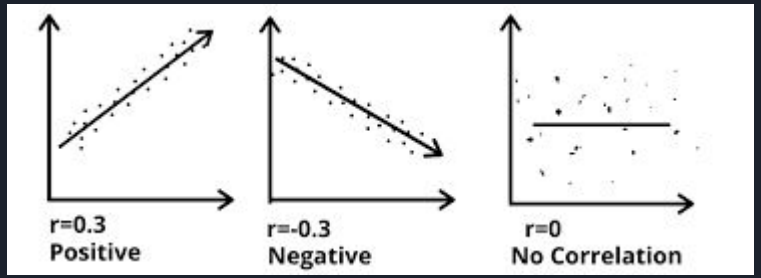
age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4	1
55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7	1
50.0	1	111	0	20	0	210000.00	1.9	137	1	0	7	1
65.0	1	160	1	20	0	327000.00	2.7	116	0	0	8	1

2. Visualisations

- Corrélation

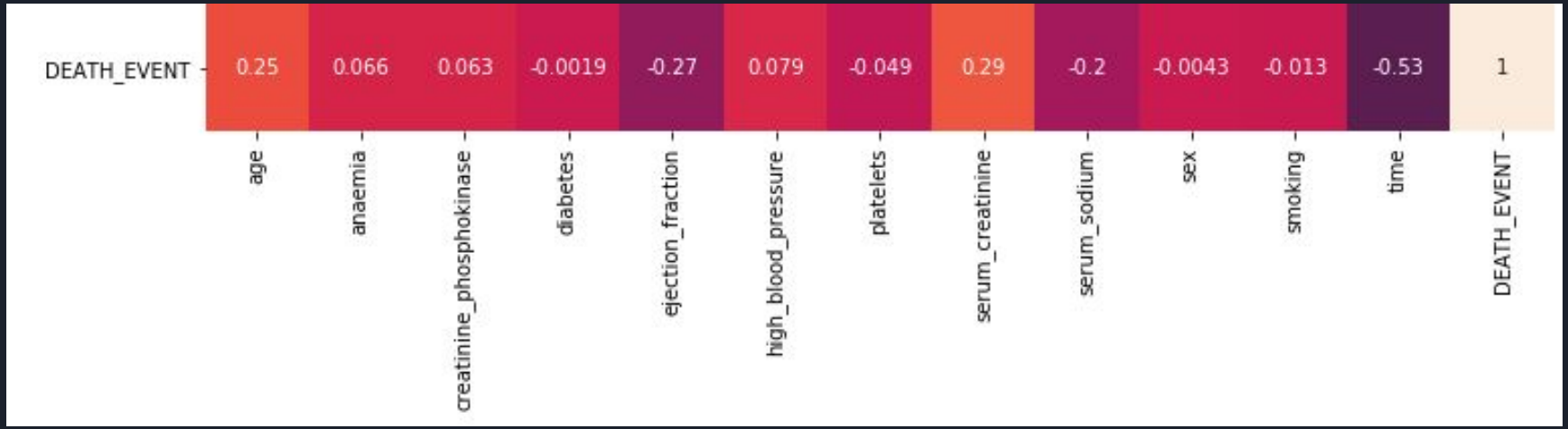
- Données à faible corrélation

- Données à forte corrélation



Corrélation

valeur absolue proche de 1 ➡ colonne importante



Quelques données

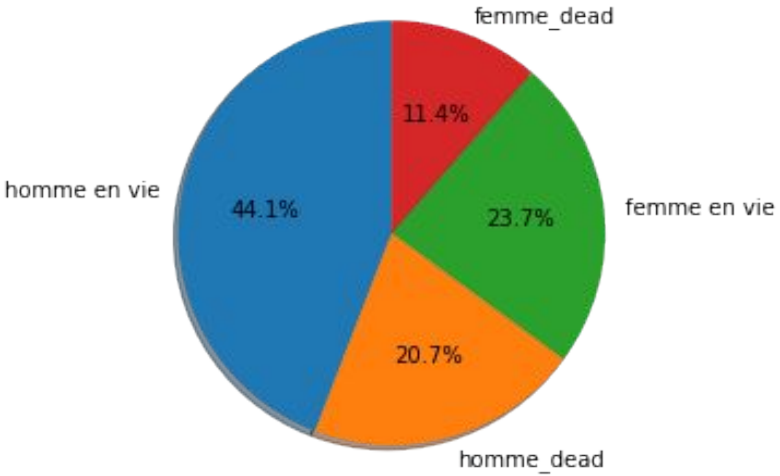
Faibles corrélations :

- Sexe

  ~ 32% de décès parmi les hommes

  ~ 33% de décès parmi les femmes

197 patients hommes
63 décès parmi les hommes



102 patients femmes
34 décès parmi les femmes

Quelques données

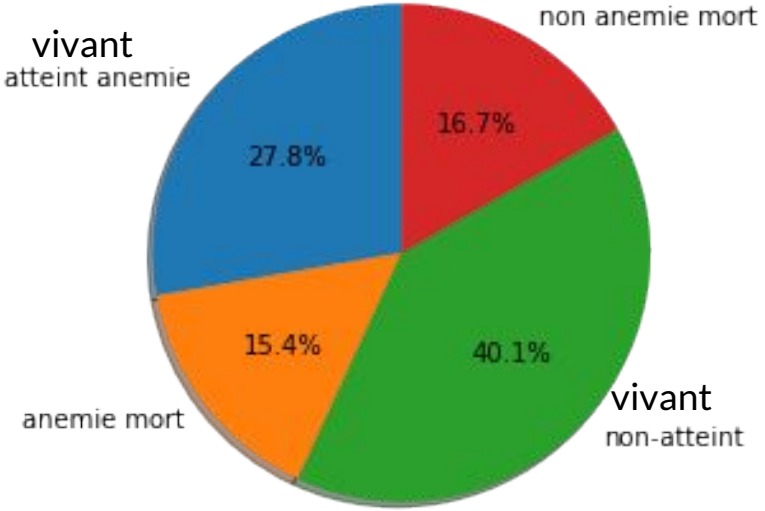
Faibles corrélations :

- Anémie

  ~ 35% de décès parmi les patients atteints d'anémie

  ~ 29% de décès : non-anémies

129 patients atteints d'anémie
45 décès parmi les patients atteints d'anémies



170 patients non atteints d'anémie
50 décès parmi les patients non atteints d'anémies

Quelques données

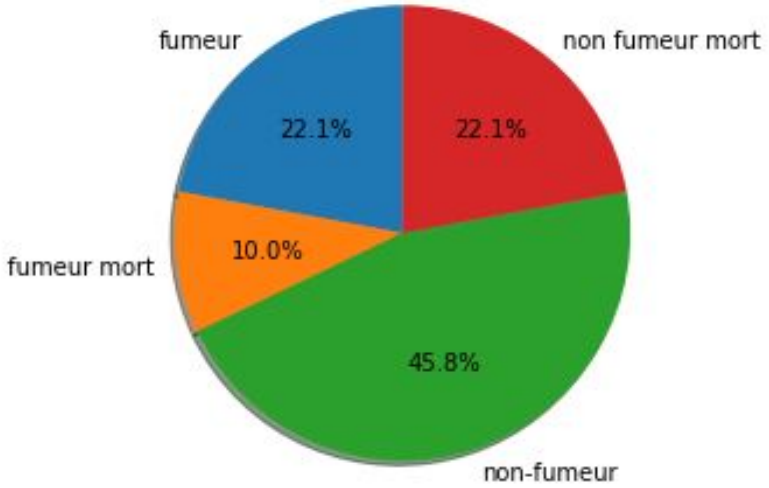
Donnée qui semble absurde : smoking

~ 31% de mort parmi les fumeurs

~ 32% de mort parmi les non fumeurs

96 patients fumeurs

30 décès parmi les fumeurs



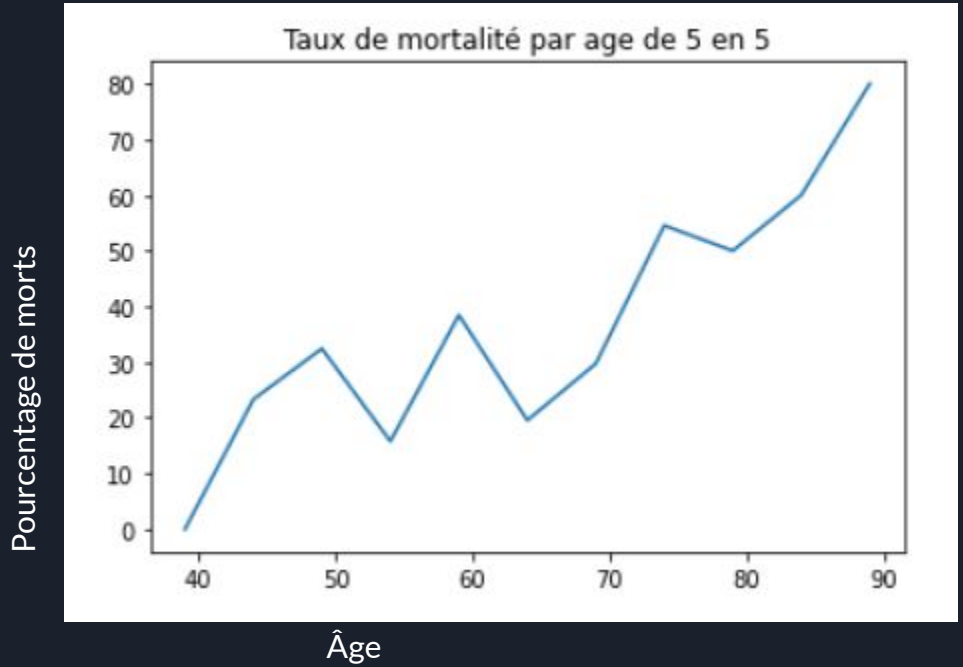
203 patients non fumeurs

65 décès parmi les non fumeurs

Données à forte corrélation : âge

Taux de mortalité :

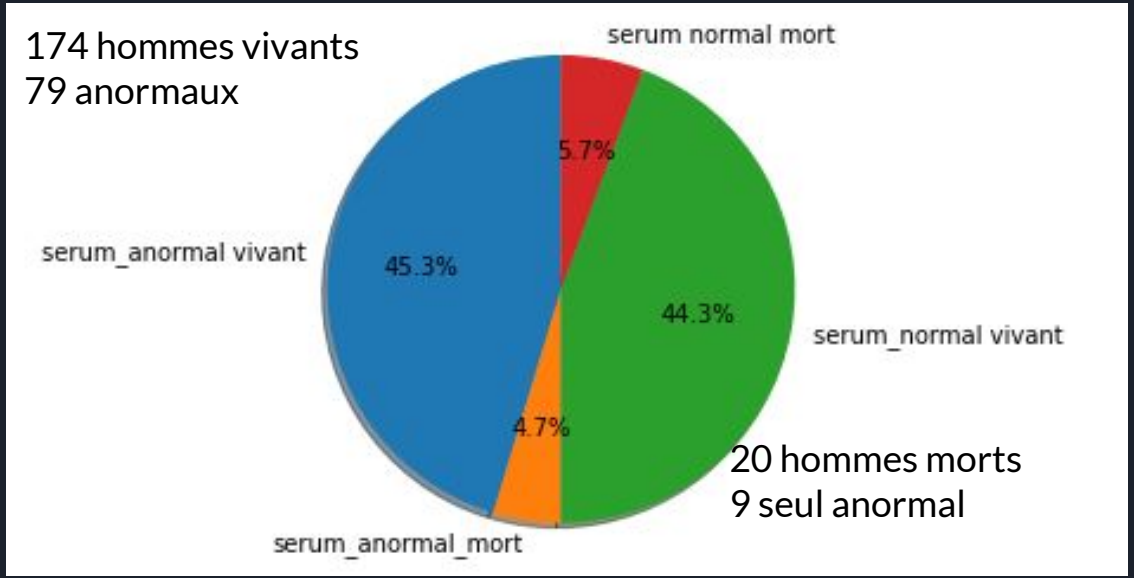
- 3 pics
- croissant



Données à forte corrélation : créatinine H

Taux de créatinine hommes :

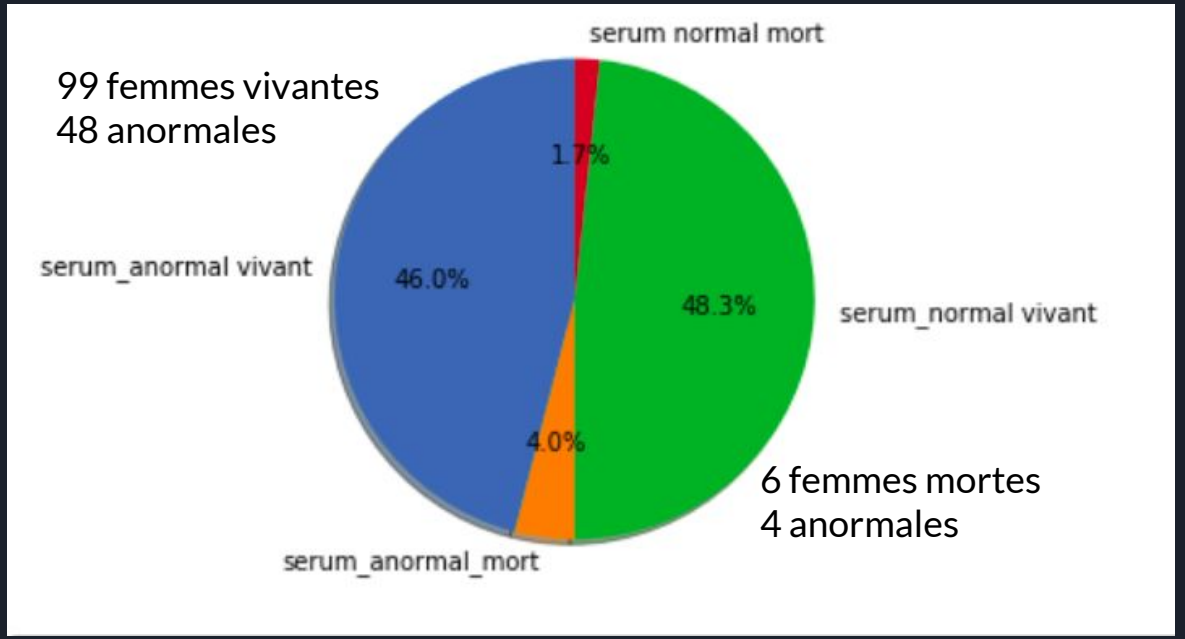
- vivants anormaux : ~50.6%
- morts anormaux : ~45.2%



Données à forte corrélation : créatinine F

Taux de créatinine femmes :

- vivants anormaux : ~48.8%
- morts anormaux : ~70.2%



Données à forte corrélation : conclusion

- L’âge

- Le taux de créatinine n’impacte pas chez l’homme

- Un taux de corrélation de 0.29 faux ?

3. Modèle

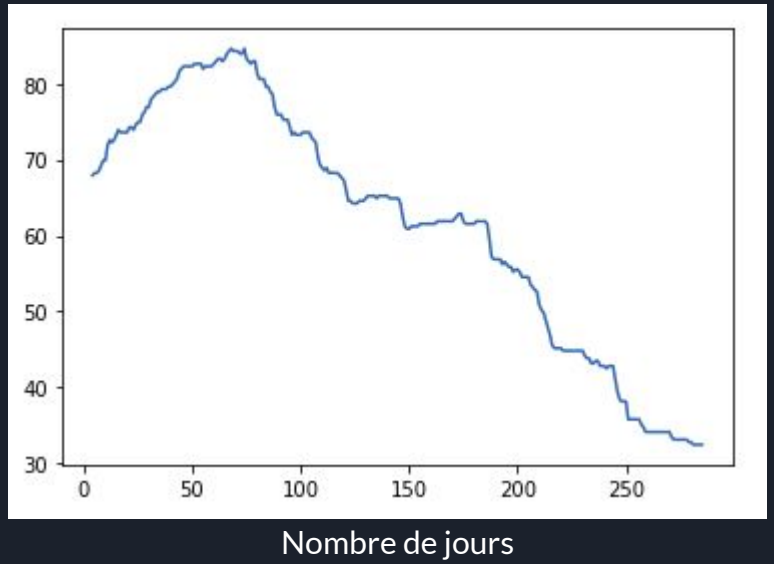
- Pré-processing

- SGDClassifier

- Optimisation du modèle

Pré-processing

-Forte corrélation entre ’time’ et ‘DEATH_EVENT’



Nombre de jour en fonction du pourcentage de mort

time	DEATH_EVENT
4	1
6	1
7	1
7	1
8	1
...	...
270	0
271	0
278	0
280	0
285	0

SGDClassifier et Standardisation

X

age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking

Y

DEATH_EVENT

Standardisation :

-moyenne: 0

- variance: 1

méthode : StandardScaler()

ex avant standardisation :

```
array([7.50e+01, 0.00e+00, 5.82e+02, 0.00e+00, 2.00e+01, 1.00e+00,
       2.65e+05, 1.90e+00, 1.30e+02, 1.00e+00, 0.00e+00])
```

ex après standardisation :

```
array([-6.98463369e-02,  1.14796753e+00,  1.65728387e-04, -8.47579380e-01,
       -6.84180207e-01,  1.35927151e+00, -1.39653077e+00, -4.78204687e-01,
        1.90111381e+00, -1.35927151e+00, -6.87681906e-01])
```

SGDClassifier: algorithme

Notre but :

$$f(x) = w^T x + b$$

w = pente
b = ordonnée à l'origine

Minimiser l'erreur d'entraînement:

$$E(w,b) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \alpha R(w)$$

hyper-paramètres :
L = fonction de perte
α >0
R = terme de régularisation

Pénalité R:

L1 norm: $R(w) := \sum_{j=1}^{m} |w_j|$

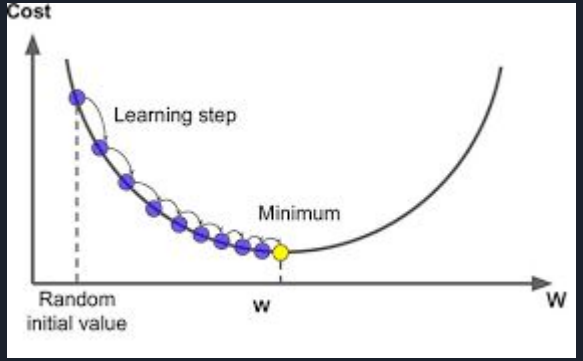
L2 norm: $R(w) := \left(\sum_{j=1}^{m} w_j^2\right)^{\frac{1}{2}} = ||w||_2^2$

SGD vs GD

Stochastic Gradient Descent/ Gradient Descent

Quelle est leur point commun ?

Quelle est leur différence ?



Optimisation du modèle

Découpage du dataset: 85% train/validation 15 % test
Kfold avec 5 folds sur train/validation → 80% train et 20% validation à chaque étape

But : maximiser la précision

3 hypers paramètres dans la formule E :

- loss

- penalty

- alpha

$$E(w,b) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \alpha R(w)$$

Optimisation du modèle : résultats

Meilleure combinaison :

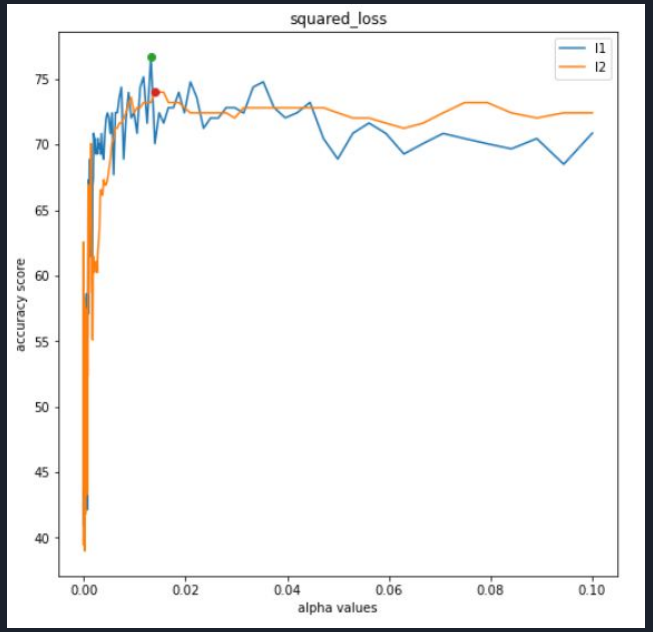
loss = squared_loss

penalty = l1

alpha = 0.013200884008314194

Score d'accuracy ensemble d'entraînement : 76.7

Score de validation ensemble de test : 71.1



Conclusion

Critiques de notre score ?

Explication ?

Peut-on prédire la mort d’un patient selon ses antécédents ?