**Projet d’IAS :**

****

**DATASET** : Heart Failure Prediction

**Lien :** <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data?select=heart_failure_clinical_records_dataset.csv>

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

**Sommaire :**

# Introduction

## Contexte

## Dataset

# Visualisation

## Corrélations

## Quelques données

## Données à forte corrélation

# Modèle

## Pré-processing

## Algorithme

## Optimisation du modèle

# Conclusion

# Annexe

1. [**Introduction**](#_vb6yy5jfgae9) **:**
2. Contexte :

Les maladies cardio-vasculaires sont des maladies qui agissent sur le cœur et le sang. Ce sont des grands facteurs de décès responsables de plusieurs millions de morts chaque année, il s’agit aussi des deuxièmes plus grandes causes de décès chez l’homme et les premières chez la femme d’après des données 2008 de Wikipédia. Pouvons-nous éviter certains décès en analysant les caractéristiques qui influent sur le risque de décès tels qu’une mauvaise alimentation, l’obésité, l’alcool, le tabagisme, le sexe, etc ?

1. Dataset :

Notre dataset contient 12 caractéristiques (features) contenant chacune 299 lignes qui vont nous permettre de prédire si une personne va mourir ou non d’une insuffisance cardiaque :

-l’âge,

-l’anémie (booléen),

-la créatine phosphokinase (normalement pour l’homme de 0 à 195 UI /l et la femme de 0 à 170 UI /l),

-le diabète (diabétique ou pas),

-la fraction d’éjection (valeur normale de 60 à 70%),

-l’hypertension (booléen),

-le nombre de thrombocytes (150 000 et 400 000 par mm normalement),

-le taux de créatinine (0.6 à 1.1 mg/dl pour la femme et 0.7 à 1.4 mg/dl chez l’homme normalement),

-le taux de sodium (135 à 145 mmol/l normalement),

-le sexe,

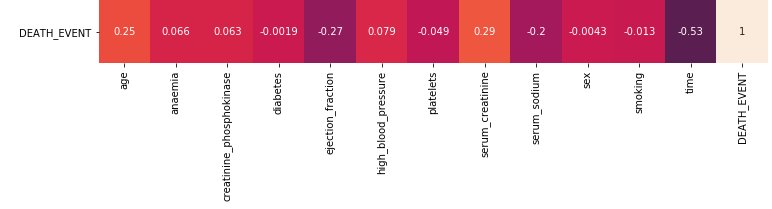
-fumeur (booléen),

-time (durée de suivi,en jour),

-mort (booléen si le patient meurt durant la période de suivi).

1. **Visualisations:**
2. Corrélation

Parmi les données du dataset, certaines sont plus importantes que d’autres dans notre interprétation, car elles influent fortement sur la valeur de la mort (death event) : elles ont donc une forte corrélation avec. D’autres ont une faible corrélation et ont donc une importance vraiment faible dans notre cas. En supprimer certaines n’impacte quasiment pas nos résultats.

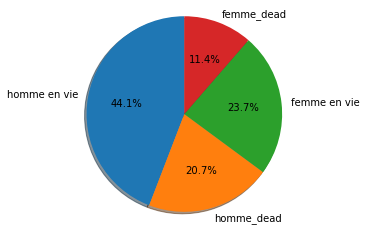


Plus on obtient une valeur absolue proche de 1 et plus la colonne est importante.

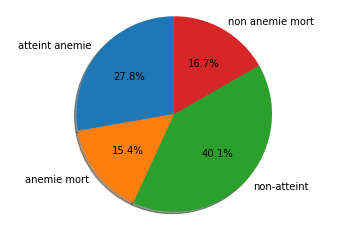
Les variables qui influent le plus sur la mort mis-à-part time sont au nombre de 4, du plus au moins impactant on a le taux de créatine (0.29), la fraction d’éjection (-0.27), l’âge (0.25), et le taux de sodium (-0.2).

Les variables les moins importantes sont le sexe (-0.0043) et le diabète (-0.0019).

1. Quelques données

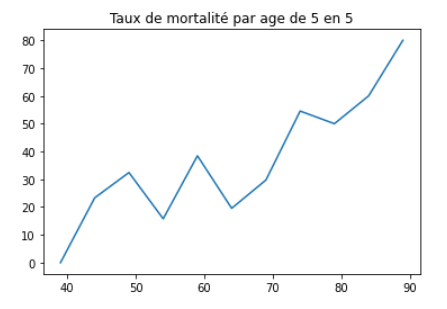


64,8% de notre dataset est composé d’hommes, et 35,1% de femmes. 20,7% des patients masculins sont morts à la suite de maladies cardiovasculaires et 11,4 % pour les femmes suivies. Ce qui représente 32% parmi les hommes et 33% parmi les femmes suivies. Ces résultats confirme bien la faible corrélation entre sexe et death\_event



43.2% des patients sont atteints d'anémie contre 56.8% de non-anémie. 15.4% des patients atteints d’anémies sont morts, ce qui représente 35% de décès parmi les patients atteints d’anémie. Tandis que nous avons 16.7% de décès qui n'avaient pas d’anémie, ce qui représente 29% de décès parmi les non-anémies. Donc les patients atteints d’anémies ont un peu plus de risques de décès par rapport à ceux qui ne sont pas atteints. L’ anémie peut être une donnée intéressante.

1. Données à forte corrélation

Nous avons choisi de calculer les taux par groupe de 5 ans pour lisser les résultats, notre dataset étant plutôt maigre certaines tranche d'âge sont vide ou presque 

Parmi les patients suivis, on peut remarquer des piques à l'âge de 60 ans et à partir de 70 ans le taux de décès de ne cesse d’augmenter.

La visualisation de de l’effet du taux de sérum créatine pour les femmes dans le notebook est aussi un bon exemple de corrélation forte .

1. **Modèle:**
2. Pré-processing

On a choisi d'enlever le feature : "time" qui représente la période de suivi du patient finissant le jour où le patient est mort, soit le jour où on a perdu contact avec lui.

On pense que ce paramètre est fortement corrélé avec notre label "death\_event" car la mort d'un patient à un impact direct sur le "time".

De plus, un patient peut bien mourir en dehors de la période de suivi donc le feature "time" ne nous intéresse pas.

L’algorithme demande d’avoir une moyenne de 0 ainsi qu’une variance de 1 afin d’obtenir le meilleur score possible. Nous devons donc faire un choix de pré-processing. Nous avons choisi la standardisation, car elle nous permet d’avoir ce résultat.

1. Algorithme:

Rappelons notre but : déterminer si une personne atteinte de maladie cardiovasculaire va mourir ou non en fonction de ses antécédents. C’est un problème de classification binaire.

Le modèle à étudier est le suivant : SGDClassifier.

Quelques informations sur le SGDClassifier :

- SGDClassifier pour Stochastic Gradient Descent Classifier.

Avantage d’un tel modèle :

-efficace sur des grandes quantités de données.

Désavantage :

-nombreux hypers paramètres

-sensible lorsque les données ne sont pas “à la même échelle”

Comment fonctionne ce modèle ?

On souhaite apprendre un modèle linéaire de la forme :



avec w et b, les paramètres de notre modèle.

Pour déterminer les paramètres de notre modèle, il faut minimiser les erreurs d'entraînement E :



-L qui est notre fonction de perte,

- α >0 , est un hyper-paramètre qui influe sur le terme de régularisation.

-R est le terme de régularisation qui pénalise la complexité de notre modèle. Il existe plusieurs choix pour le terme de régularisation :





En réalité, on pénalise les valeurs extrêmes de qui correspondent souvent à un cas de sur-apprentissage

Quelle est la différence entre la Stochastic Gradient Descent et la descente de gradient vue en cours ? SGD est un algorithme d'optimisation permettant de trouver un minima d’une fonction concave/convexe tout comme la descente de gradient.

Dans cette dernière à chaque époque de l’algorithme, l’ensemble du dataset doit être traversé pour calculer la nouvelle valeur de W tandis que pour un SGD, seulement un échantillon ou un sous-ensemble du dataset (On parle de minibatch dans ce cas-là) est utilisé.

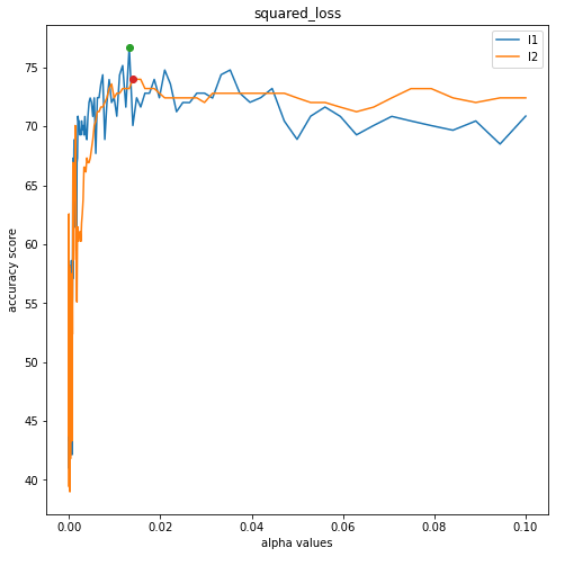
Le choix des données utilisées induit de l’aléatoire, d’où le terme stochastique.

Cette différence fait qu’en général un SGD converge beaucoup plus vite , d’autant plus si le dataset est grand, au prix d’une légère perte en précision du minimum de la fonction.

1. Optimisation du modèle

Afin d'obtenir le meilleur modèle qui maximise sa précision, nous optimisons 3 hypers paramètres en même temps.

* loss, la fonction de coût L dans la formule E
* penalty, le terme de régularisation R dans la formule E
* alpha, le coefficient de R dans la formule E

Nous utilisons ‘accuracy’ comme métrique d’évaluation et non pas une méthode avec des poids, car nous faisons une classification binaire et que nous accordons autant d'importance aux faux positifs qu’au faux négatifs. 

Pour chaque combinaison possible, le modèle est entraîné et évalué avec une cross-validation en 5 parties.

Sur ce graphique, on peut voir les résultats de la L1 norm et la L2 norm avec squared loss comme fonction de coût pour différentes valeurs d’alpha.

La meilleure combinaison retenue est

* loss = squared\_loss
* penalty = l1
* alpha = 0.013200884008314194

On obtient avec ces paramètres sur notre SGD Classifier un score d'accuracy de 76.7 sur l’ensemble d'entraînement/validation et 71.1 sur l’ensemble de test.

1. **Conclusion**

Bien que notre résultat soit bien en dessous de ceux affichés par d’autres sur Kaggle (qui montent jusqu'à 97%), on peut l’expliquer notamment par le fait d’avoir retiré ‘time’ notre feature avec le plus de corrélation et que SGDClassifier soit plus optimisé pour des centaines de milliers de lignes.

Malgré tout, notre modèle a produit un résultat mettant en évidence l'existence de facteurs permettant de prédire le fait de décéder ou non d’une maladie cardio-vasculaire.

**Annexe**

Liens utilisés : [wikipédia](https://fr.wikipedia.org/wiki/Maladie_cardiovasculaire), [créatinine](https://www.lamutuellegenerale.fr/le-mag-sante/sante-au-quotidien/la-creatinine-c-est-quoi.html), [sodium](https://sante.journaldesfemmes.fr/fiches-sante-du-quotidien/2550126-hyponatremie-definition-carence-sodium-cause-symptome-traitement/#:~:text=La%20mesure%20du%20taux%20de,inf%C3%A9rieure%20%C3%A0%20135%20mmol%2Fl.), [fraction d’éjection](https://www.rythmo.fr/glossary/fraction-dejection/#:~:text=S'exprime%20en%20%25%20du%20volume,ne%20se%20vidangent%20jamais%20totalement.), [créatine phosphokinase](https://www.doctissimo.fr/html/sante/analyses/ana_proteines05.htm#:~:text=Le%20taux%20normal%20dans%20le,%3A%200%20%2D%20170%20UI%20%2Fl), [SGD maths](https://scikit-learn.org/stable/modules/sgd.html#implementation-details), [SGD Classifier model](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier), [métrique de score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score).