Institut Francophone pour l'Innovation

Université National du Vietnam, Hanoi

Rapport final

Sohone Bi Landry-Ange
Oumourou Altine M.Aminou

Professeur : NGUYEN Thi Minh

Master 1 : Système Intelligent et Multimédia | Fouille de Données

Table des matières

1 Introduction

Avec l'apparition d'internet, l'on assiste a une croissance exponentielle des données appelle Big Data. Cela a donnée naissance a de nouvelles techniques et méthodes permettant a partir de données brutes de pouvoir extraire de la connaissance pour la prise de décision...

L'objet de notre étude est de pouvoir appliquer ses méthodes et techniques pour la conception d'un modèle de prédiction. Pour cette étude notre choix s'est porte sur le **Dataset Abalone** accessible a l'adresse `https://archive.ics.uci.edu/ml/datasets/abalone` qui a pour objectif de prédire l'âge d'un ormeau a partir de ses mesures, ce qui permettra de réduire la charge de travail dans la recherche de l'âge de l'ormeau qui se fait a l'aide de microscope.

Pour la réalisation de notre travail, nous avons choisie **python (2.7)**. qui nous fournit **Jupyter Notebook** comme environnement pour les Datas Sciences. Les librairies utilisées sont : **Pandas, matplotlib, Numpy et sklearn**

2 Description détaillée du Jeu de données

2.1 Description du Jeu de données

Le jeu de données est constitue de 08 attributs d'entrées parmi lesquelles l'on a (1) un qualitatif et huit (7) quantitatifs et un (1) attribut de sortie qui est l'attribut a prédire a partir duquel l'on pourra avoir l'âge de l'ormeau . Notons que le jeu de données ne comporte pas de valeurs manquantes.

	length	diameter	height	weight.w	weight.s	weight.v	weight.sh	rings
count	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000
mean	0.523992	0.407881	0.139516	0.828742	0.359367	0.180594	0.238831	9.933684
std	0.120093	0.099240	0.041827	0.490389	0.221963	0.109614	0.139203	3.224169
min	0.075000	0.055000	0.000000	0.002000	0.001000	0.000500	0.001500	1.000000
25%	0.450000	0.350000	0.115000	0.441500	0.186000	0.093500	0.130000	8.000000
50%	0.545000	0.425000	0.140000	0.799500	0.336000	0.171000	0.234000	9.000000
75%	0.615000	0.480000	0.165000	1.153000	0.502000	0.253000	0.329000	11.000000
max	0.815000	0.650000	1.130000	2.825500	1.488000	0.760000	1.005000	29.000000

FIGURE 1 – Description du jeu de données

ATTRIBUTS D'ENTRÉES

Attribut qualitatif

- **Sex** : qui détermine le sexe de l'ormeau et peut avoir les valeurs suivantes : **M** pour masculin, **F** pour Féminin, et **I** pour nourrisson.

Attributs quantitatifs

- **Lenght** : qui détermine en mm la longueur de la coquille. elle prend des valeurs comprise entre **0.075** et **0.815** donc de type continu.
- **Diameter** : qui détermine en mm le diamètre de la coquille. il prend les valeurs comprise entre **0.055** et **0.650** donc de type continu.
- **Height** : qui détermine en mm l épaisseur de la viande contenu dans la coque. il prend les valeurs comprise entre **0.000** et **1.130** donc de type continu.
- **Whole Weight** : le poids de l'ormeau en grammes qui est en additionnant son poids après saignement, son poids après séchage et le poids de sa viande. ce poids est compris entre **0.002** et **2.826**.
- **Shucked weight** : le poids en grammes de la quantité de viande recueilli dans la coque. ce poids est compris entre **0.001** et **1.488**.
- **Visceral Weight** : le poids après saignement. ce poids est compris entre **0.001** et **0.760**.
- **Shell Weight** : le poids de la coquille de l'ormeau en grammes.compris entre **0.002** et **0.005**.

ATTRIBUTS DE SORTIE

- **Rings** : anneau d arrêt de croissance qui permet de déterminer l'âge en ajoutant **1.5** a sa valeur initiale. compris entre **1** et **29**.

2.2 Analyses des variables

A ce niveau, nous avons juge utile de montrer la distribution de chacune des variable de notre jeu de données.

2.2.1 Variable qualitative

Dans le cadre de l'analyse de la variable qualitative, nous en disposons qu'une seule dans notre jeu de données. Il s'agira donc de faire un tri a plat sur la variable **Sex** afin d'observer la distribution des modalités dont le mode est **M**

Tri a plat de sex

```
df['sex'].describe()
```

```
count     4177
unique       3
top          M
freq      1528
Name: sex, dtype: object
```

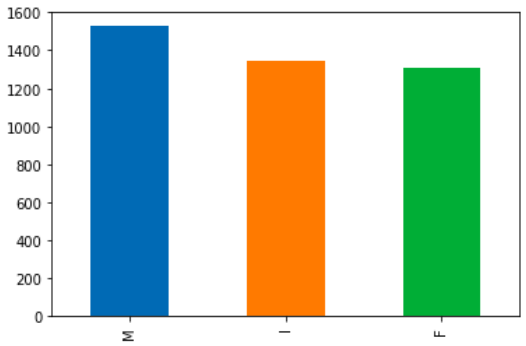

FIGURE 2 – Tri a plat de la variable sex

2.2.2 Variables quantitatives

Au niveau des variables quantitatives, nous présenterons que **Whole Weight**, et **Lenght** en supposons qu'a elles seules elles peuvent expliquer au mieux la variable a prédire.

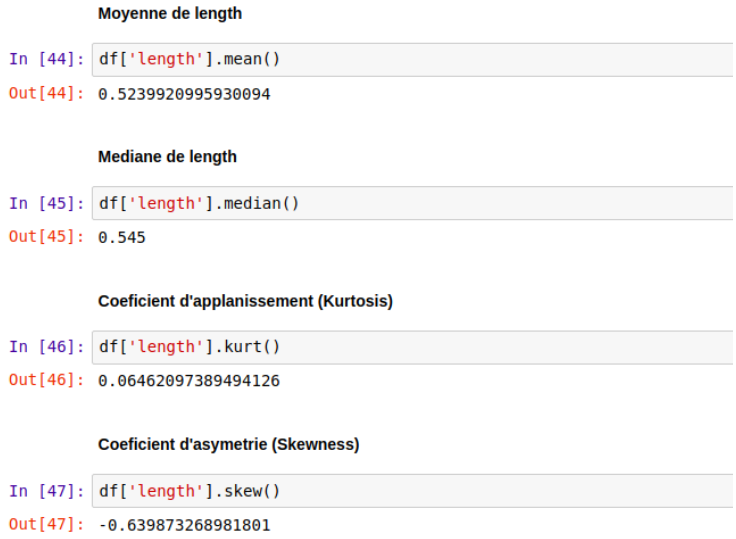

FIGURE 3 – Analyse de la variable lenght

Pour ce qui est de la variable **lenght**, nous pouvons observer les informations suivantes :

- *la moyenne : 0.5239*
- *la médiane : 0.545*

Pour ce qui est de la représentation, on observe le coefficient d'asymétrie (**skewness**) qui est négatif et celui d'aplanissement (**kurtosis**) qui est positif

FIGURE 4 – Analyse de la variable whole weigth.w

Pour ce qui est de la variable **weight.w**, nous pouvons observer les informations suivantes :

- *la moyenne : 0.828*
- *la médiane : 0.799*

Pour ce qui est de la représentation, on observe le coefficient d'asymétrie (**skewness**) qui est positif et celui d'aplanissement (**kurtosis**) qui est négatif

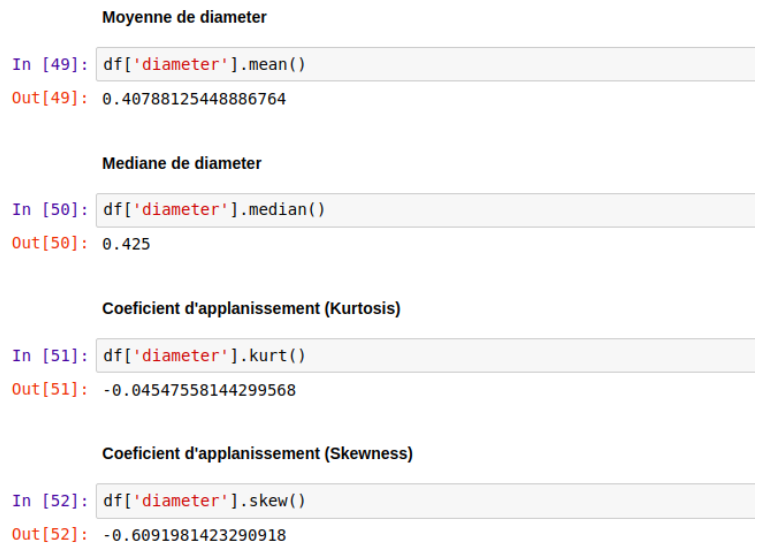

FIGURE 5 – Analyse de la variable diameter

Pour ce qui est de la variable **diameter**, nous pouvons observer les informations suivantes :

- *la moyenne : 0.407*
- *la médiane : 0.425*

Pour ce qui est de la représentation, on observe le coefficient d'asymétrie (**skewness**) celui d'aplanissement (**kurtosis**) sont négatif.

3 Énoncé du problème et Choix de la méthode

3.1 Énoncé du Problème

Comme énonce dans l'introduction, il s'agit de pouvoir prédire un variable a partir d'autres variables. dans ce cas de figure, il s'agit d'un **problème de régression**. Pour donc valider cette hypothèse, nous avons choisir de vérifier si la relation entre les variables que nous supposons explicatives sont linéaires par rapport a la variable a prédire **rings**. Ci dessous le résultat de l'analyse des liens :

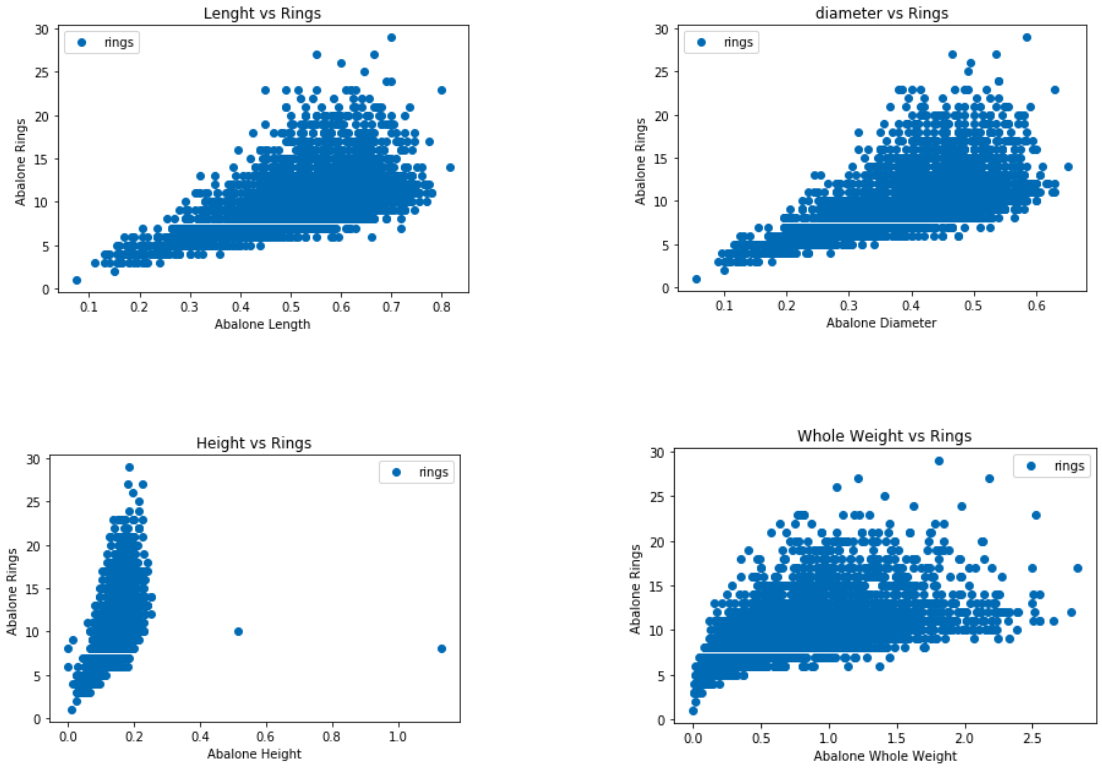

FIGURE 6 – Liens des variables explicatives avec la variable a prédire rings

Ces figures montre la relation linéaire qui existe entre ces variables, ce qui vient corroborer notre hypothèse de départ.

3.2 Justification de la méthode d'apprentissage automatique

Nous avons une variable a prédire de type quantitative continu et plusieurs variables quantitatives qui sont les variables explicatives, notre choix s'est porte sur l'algorithme de régression linéaire multiple car il correspond plus a ce type de problème. Dans la section ci=dessous nous présenterons de manière détaillée cette algorithmes.

4 Présentation de la Regression Linéaire Multiple

4.1 Généralités

On dispose d'un échantillon de n individus pour chacun desquels on a observé – yi , la valeur de la variable réponse y quantitative, – x 1 i ,...,x p i , les valeurs de p autres variables quantitatives x 1 ,..., x p , pour i = 1,...,n. On veut expliquer une variable quantitative y par p variables quantitatives x 1 ,..., x p . Le modèle s'écrit :

$$\boxed{y_i = \beta_0 + \beta_1 x_i^1 + ... + \beta_p x_i^p + e_i} \qquad \forall i = \{1, ..., n\}$$

FIGURE 7 – Squelette du model linéaire

où ei est une réalisation de Ei N(0,2) et où les n v.a. Ei sont indépendantes

4.2 Paramètres de la méthode

Les paramètres du modèle de régression linéaire sont estimés par :

$$\widehat{\beta}(y) = (X'X)^{-1}X'\mathbf{y}$$

4.2.1 Proprietes

1. *eb = 0,*
2. *yb = y,*
3. *La droite de régression passe par le point de coordonnées (x, y)*
4. *Le vecteur des résidus n'est pas corrélé avec la variable explicative : cov(x, be) = 0*
5. *Le vecteur des résidus n'est pas corrélé avec la variable ajustée Y : cov(yb, be) = 0*
6. *La variance de Y admet la décomposition :*

$$var(\mathbf{y}) = var(\widehat{\mathbf{y}}) + var(\widehat{\mathbf{e}}).$$

4.2.2 Le coefficient R2

On déduit de cette décomposition que le coefficient R2 , défini comme le carré du coefficient de corrélation de x et y est une mesure de qualité de l'ajustement, égale au rapport de la variance effectivement expliquée sur la variance à expliquer :

4.2.3 Choix des variables explicatives

En présence de p variables explicatives dont on ignore celles qui sont réellement influentes, on doit rechercher un modèle d'explication de Y à la fois performant (résidus les plus petits possibles) et économique (le moins possible de variables explicatives). Nous avons plusieurs methodes de criteres et methodes dont les methodes ascendantes et descendantes.

1. Les méthodes ascendantes : On cherche d'abord la variable qui explique le mieux y au sens du R2 (R2 maximum), puis on cherche celle qui, ajoutée à la première, augmente le plus le R2 , etc. Un critère d'arrêt de la procédure peut-être obtenu en utilisant des critères du type R2 ajusté, Cp de Mallows ou critère AIC : par exemple, on arrête le processus lorsque le R2 ajusté commence à décroître.
2. Les méthodes descendantes : On part du modèle utilisant les p variables explicatives et on cherche, parmi les p variables, celle qui peut être supprimée en occasionnant la plus forte croissance du critère. Cette variable étant supprimée, on itère le processus tant que le R2 ajusté ne décroît pas.

4.3 Caractéristiques

4.3.1 Introduction de variables explicatives qualitatives

En régression linéaire, les variables explicatives ne sont que quantitatives, donc si vous disposez de variable qualitatif, elles doivent être transformée en m-1 variables binaires (on parle de dichotomisation) correspondant aux modalités de la variable.

4.3.2 La demarche de modelisation

1. *estimer les paramètres « a » en exploitant les données*
2. *évaluer la précision de ces estimateurs (biais, variance, convergence*
3. *mesurer le pouvoir explicatif global du modèle*
4. *évaluer l'influence des variables dans le modèle*
5. *sélectionner les variables les plus « pertinentes »*
6. *évaluer la qualité du modèle lors de la prédiction*
7. *détecter les observations qui peuvent fausser ou influencer exagérément les résultats*

4.3.3 Validation du modèle

Effectuer la régression linéaire de Y sur X1, X2,..., Xp consiste à déterminer 0, 1, 2,..., p. C'est en testant si i = 0 que l'on teste l'association entre la covariable Xi et Y. Le résultat du test n'est valide que si les résidus, c'est-à-dire les erreurs entre les valeurs observées de Y et leur estimation dérivée du modèle, suivent une distribution normale de moyenne nulle, de même variance (hypothèse d'homoscédasticité) et s'ils ne sont pas corrélés entre eux (hypothèse d'indépendance).
Ces hypothèses peuvent être vérifiées par des tests ou de manière plus pratique à l'aide de graphiques :

1. distribution des résidus et graphique des résidus en fonction des covariables (la dispersion des résidus doit être homogène autour de zéro)
2. QQplot (ou diagramme quantile-quantile) représentant les quantiles de la distribution de l'échantillon en fonction des quantiles de la distribution normale (gaussienne) (les points doivent être quasiment alignés sur la première bissectrice y = x)

4.3.4 Intervalle de confiance

sert a prédire une réponse moyenne correspondant aux variables explicatives

4.3.5 Intervalle de prédiction

sert à prédire une nouvelle valeur“individuelle”. Par exemple, si on étudie la liaison entre le poids et l’âge d’un animal, on peut prédire la valeur du poids à 20 jours soit comme le poids moyen d’animaux à 20 jours, soit comme le poids à 20 jours d’un nouvel animal. Pour le nouvel animal, on doit prendre en compte la variabilité individuelle, ce qui augmente la variance de l’estimateur et donc la largeur de l’intervalle

4.4 Limites

Bien qu’elle soit adaptée pour montre la dépendance d’une variable par rapport a d’autres, elle ne permet pas dans tous les cas de figure de pouvoir déterminer le facteur de cause a effet. comme exemple illustratif, il a y’a beaucoup de pompiers reviens a dire qu’il y a beaucoup de feu. ce qui est juste mais l’interprétation de variable explicatives perd son sens car on dira que la quantité importante de feu se justifie par le nombre de pompiers.

5 Application de la méthode

5.1 Paramètres choisis

Au niveau des paramètres choisit pour notre méthode, nous nous sommes focaliser sur le coefficient de corrélation car nous avons juge utile du fait qu’il nous permet de voir clairement l’effet de la variation des variable explicatives sur le modèle.

	Coefficient
length	-11.619835
diameter	25.992209
height	15.742263
weight.w	0.178434

	Coefficient
length	-1.268301
diameter	13.435928
height	9.165784
weight.w	9.662898
weight.s	-20.626437
weight.v	-9.947989
weight.sh	8.152597

FIGURE 8 – Coefficients optimaux de nos variables

Ci-dessus les coefficient de corrélation optimaux pour notre modèle.

5.2 Choix des variables explicatives

Dans un premier temps nous avions, nous avions utilise que les trois variables explicatives au niveau de la section de l'analyse des variables explicatives. Mais suite au calcul des différents coefficient de régression pour nos variables lors de la phase de préparation de données, nous avons eu les résultats exprimées par la figure ci-dessus

Cas 1
la figure de gauche montre que l'augmentation de la variable **length** a un effet et réduire de -11 la valeur de la variable **rings**. **diameter** et **height** quand a elle permettrait de faire augmenter de 25 et 15 lorsqu'elles augmentent. On voit que l'effet de **weigth.w** est assez minime donc devrait au préalable être enlevé pour la création du model.

Cas 2
la figure de droite (*après consideration de toutes nos variables explicatives*) montre que les variations de toutes nos variables ont un effet sur la variable a predire. Fort de cette analyse que le choix de nos variables explicatives s'est porte sur l'ensemble de nos variables quantitatives.

5.3 Taille des données d'apprentissage

Dans le cadre de l'apprentissage, nous avons effectues plusieures expérimentations avec les partitionnements suivants. en gras le partitionnement qui nous a fournit les meilleurs resultats

1. ***Apprentissage : 80 et Test 20***
2. *Apprentissage : 70 et Test 30*
3. *Apprentissage : 60 et Test 40*
4. *Apprentissage : 50 et Test 50*

6 Expérimentations et Évaluation

6.1 Expérimentations avec les 3 variables explicatives de départ

	Donnees actuelles	Donnees predites
668	13	10.330322
1580	8	9.656731
3784	11	11.006815
463	5	5.678679
2615	12	11.596509

```
('MAE:', 1.8338934202804407)
('MSE:', 6.601150083679482)
('RMSE:', 2.569270340715333)
```

FIGURE 9 – Expérimentation 1 Training : 80 Test : 20

	Donnees actuelles	Donnees predites
668	13	10.311972
1580	8	9.640498
3784	11	11.045866
463	5	5.704825
2615	12	11.654663

```
('MAE:', 1.8373056139125932)
('MSE:', 6.568756187172609)
('RMSE:', 2.562958483310373)
```

FIGURE 10 – Expérimentation 2 Training : 70 Test : 30

	Donnees actuelles	Donnees predites
668	13	10.263538
1580	8	9.625892
3784	11	11.041889
463	5	5.621378
2615	12	11.634777

```
('MAE:', 1.8333553085333796)
('MSE:', 6.601375313641947)
('RMSE:', 2.5693141718446864)
```

FIGURE 11 – Expérimentation 3 Training : 60 Test : 40

	Donnees actuelles	Donnees predites
668	13	10.267167
1580	8	9.705575
3784	11	11.087024
463	5	5.611297
2615	12	11.633155

```
('MAE:', 1.8361110156763176)
('MSE:', 6.651577924470357)
('RMSE:', 2.579065319930916)
```

FIGURE 12 – Expérimentation 4 Training : 50 Test : 50

6.2 Expérimentation avec toutes les variables explicatives

	Donnees actuelles	Donnees predites
668	13	12.976738
1580	8	9.651631
3784	11	10.300840
463	5	5.656562
2615	12	10.637666

```
('MAE:', 1.6151862192084647)
('MSE:', 5.103381714160202)
('RMSE:', 2.2590665581518845)
```

FIGURE 13 – Expérimentation 5 Training : 80 Test : 20

Au terme de nombreuses expérimentations, nous avons pu constater qu'il était judicieux pour nous de prendre en compte toutes les variables explicatives pour la conception du modèle. Les expériences effectuées précédemment nous on fait perdre des données en laissant délibérément certaines de nos variables explicatives.

Cela veut dire que toutes nos variables explicatives utilisées sont pertinentes.

6.3 Comparaison avec le Perception

Afin de s'assurer que nous avons effectue un bon choix d'algorithme d'apprentissage automatique pour la résolution de notre problème, nous le comparerons au réseaux de neurone, plus précisément le plus simple d'entre eux, **le perception**. ci dessous l'étude comparée.

	Donnees actuelles	Donnees predites
668	13	12.976738
1580	8	9.651631
3784	11	10.300840
463	5	5.656562
2615	12	10.637666

```
('MAE:', 1.6151862192084647)
('MSE:', 5.103381714160202)
('RMSE:', 2.2590665581518845)
```

FIGURE 14 – Expérimentation 5 Training : 80 Test : 20

	Donnees actuelles	Donnees predites
668	13	14.166701
1580	8	9.978967
3784	11	10.856909
463	5	5.751406
2615	12	10.843657

```
('MAE:', 1.5674162296227565)
('MSE:', 4.803452085647091)
('RMSE:', 2.1916779155813684)
```

FIGURE 15 – Expérimentation 5 Training : 80 Test : 20

7 Conclusion

L'erreur MAE la mieux réalisée était de 1,56, utilisant une architecture perceptron avec 2 couches-cachées ([20,5]), un alpha de 0,01, un taux d'apprentissage de 0,01, et une fonction d'activation logistique.
Comparez ces résultats au meilleure de notre methode, qui a atteint un MAE de 1,61. Malgré la grande précision du réseau de neurones, la modélisation de réseaux neuronaux présente plusieurs inconvénients, notamment la dicculté de l'optimisation de l'hyperparamètre. Notez que nous avons simplement utilisé un processus d'essai et d'erreur pour sélectionner les hyperparamètres. Pouvoir interpréter rapidement et facilement des modèles peut constituer un avantage important dans de nombreux cas. Bien sûr, si tout ce dont vous avez besoin est la précision, les reseaux de neurones pourraient être un excellent choix.

Suite à notre travail, avec un MAE de 1,61, nous pouvons affirmer avoir répondu à la problématique par l'élaboration de ce modèle qui est a meme de predire l'age avec un taux d'erreur plus d'acceptable. Aussi, en prenons en compte la variable **sex** par une dichotomie et un ACP suite à la méthode de régression linéaire multiple pourrait nous permettre d'améliorer les performances.

Références

[1] Debuter un projet de Machine Learning.
http ://www.xavierdupre.fr/app/ensae$_t eaching_c s/helpsphinx/debutermlprojet.html$

[2] Initiation au Machine Learning avec Python
https ://makina-corpus.com/blog/metier/2017/initiation-au-machine-learning-avec-python-pratique.

[3] Faire une regression lineaire en Python
https ://www.stat4decision.com/fr/faire-regression-lineaire-r-python/.