Institut Francophone pour l'Innovation

Université National du Vietnam, Hanoi

Rapport final Analyse Exploratoire

Sohone Bi Landry-Ange
Oumourou Altine M.Aminou

Professeur : NGUYEN Thi Minh

Master 1 : Système Intelligent et Multimédia | Fouille de Données

Table des matières

Table des figures

1 Introduction

Pour les besoins de notre formation en Fouille de données, il nous a été soumis des TPs (Travaux Pratiques) dont le but est de procéder a l'initiation des étudiants de la promotion 22 de l'IFI a la fouille de données.
Pour la réalisation de cette initiation, nous avons choisit le jeu de données " Abalone" dont le probleme est de predire l'âge d'un ormeau (oreille de mer) a partir des mesures physiques.

2 TP1 : Description des données

2.1 Description détaillée des attributs 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 .

Notre jeu de données ne comporte pas de valeurs manquantes.

ATTRIBUTS D'ENTRÉES

Attribut qualitatif

- **Sexe** : 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

- **Longueur** : qui détermine en mm la longueur de la coquille. elle prend des valeurs comprise entre 0.075 et 0.815 donc de type continu.
- **Diamètre** : qui détermine en mm le diamètre de la coquille. il prend les valeurs comprise entre 0.055 et 0.650 donc de type continu.
- **Hauteur** : 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.
- **Poids total** : 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.
- **Poids concasse** : le poids en grammes de la quantité de viande recueilli dans la coque. ce poids est compris entre 0.001 et 1.488.
- **Poids viscérale** : le poids après saignement. ce poids est compris entre 0.001 et 0.760.
- **Poids de la coquille** : le poids de la coquille de l'ormeau en grammes.compris entre 0.002 et 0.005.

ATTRIBUTS DE SORTIE

- **Anneau** : 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 Analyse exploratoire

2.2.1 Analyse des variables qualitatives

Nous disposons que d'une seule variable qualitative, donc nous procéderons a un tri a plat en vue de voir la distribution des modalités de cet attribut dont le mode est " **M** " (voir Figure 1)

Univariate discrete stat 1

Parameters

Attributes : 1
Examples : 4177

Results

Attribute	Gini	Distribution			
		Values	Count	Percent	Histogram
Sexe	0.6650	M	1528	36.58 %	
		F	1307	31.29 %	
		I	1342	32.13 %	

FIGURE 1 – Proportion de l'attribut sexe

2.2.2 Analyse des variables quantitatives

Nous avons juger utile d'étudier les variables quantitatives continues (Poids, Longueur) car c'est a partir d'elles que nous pouvons répondre a notre problématique, celle de prédire l'âge de l'ormeau. Les informations de cette étude sont observables sur la figure ci-dessous.

More Univariate cont stat 1

Parameters

Attributes : 4
Examples : 4177

Results

Attribute	Stats		Histogram			
	Statistics		Values	Count	Percent	Histogram
Longueur	Average	0.5240	x_<_0.1490	7	0.17%	
	Median	0.5450	0.1490_=<_x_<_0.2230	60	1.44%	
	Std dev. [Coef of variation]	0.1201 [0.2292]	0.2230_=<_x_<_0.2970	147	3.52%	
	MAD [MAD/STDDEV]	0.0967 [0.8050]	0.2970_=<_x_<_0.3710	304	7.28%	
	Min * Max [Full range]	0.08 * 0.81 [0.74]	0.3710_=<_x_<_0.4450	489	11.71%	
	1st * 3rd quartile [Range]	0.45 * 0.62 [0.17]	0.4450_=<_x_<_0.5190	749	17.93%	
	Skewness (std-dev)	-0.6399 (0.0379)	0.5190_=<_x_<_0.5930	1051	25.16%	
	Kurtosis (std-dev)	0.0646 (0.0758)	0.5930_=<_x_<_0.6670	1017	24.35%	
			0.6670_=<_x_<_0.7410	324	7.76%	
			x>=_0.7410	29	0.69%	

FIGURE 2 – Étude statistique de la variable longueur

Pour ce qui est de la variable longueur on peut observer les informations ci-dessous :

— la moyenne (Average) est de 0.5240
— la médiane (Médiane) est de 0.5450
— la classe modale [0.5190 ;0.5930]
— l'écart-type est de 0.1201

On peut observer la dispersion de la longueur
— l'étendue qui représente l'intervalle de valeur entre min et max est de 0.74
— la dispersion inter-quartile est de 0.17
Pour ce qui est de la représentation, on observe le coefficient d'asymétrie qui est negatif (Skewness = - 0.6399) et le coefficient d'aplanissement lui qui est positif (Kurtosis = 0.0646).

Poids

Statistics	
Average	0.8287
Median	0.7995
Std dev. [Coef of variation]	0.4904 [0.5917]
MAD [MAD/STDDEV]	0.4005 [0.8166]
Min * Max [Full range]	0.00 * 2.83 [2.82]
1st * 3rd quartile [Range]	0.44 * 1.15 [0.71]
Skewness (std-dev)	0.5310 (0.0379)
Kurtosis (std-dev)	-0.0236 (0.0758)

Values	Count	Percent	Histogram
x_<_0.2844	632	15.13%	
0.2844_=<_x_<_0.5667	783	18.75%	
0.5667_=<_x_<_0.8491	827	19.80%	
0.8491_=<_x_<_1.1314	824	19.73%	
1.1314_=<_x_<_1.4138	616	14.75%	
1.4138_=<_x_<_1.6961	286	6.85%	
1.6961_=<_x_<_1.9785	129	3.09%	
1.9785_=<_x_<_2.2608	58	1.39%	
2.2608_=<_x_<_2.5432	16	0.38%	
x>=_2.5432	6	0.14%	

FIGURE 3 – Étude statistique de la variable poids

Pour ce qui est de la variable poids on peut observer les informations ci-dessous :
— la moyenne (Average) est de 0.8287
— la médiane (Médiane) est de 0.7995
— la classe modale [0.5667 ;0.8491]
— l'écart-type est de 0.4904

On peut observer la dispersion du poids
— l'étendue qui représente l'intervalle de valeur entre min et max est de 2.82
— la dispersion inter-quartile est de 0.71
Pour ce qui est de la représentation, on observe le coefficient d'asymétrie qui est positif (Skewness = 0.5310) et le coefficient d'aplanissement lui qui est négatif (Kurtosis = -0.0236).

2.3 Analyse de lien entre chaque paire d'attribut

Dans le cadre de notre étude, nous ne possédons qu'un seul attribut qualitatif qui est discret, nous ne pouvons évaluer son lien avec d'autres attributs. ainsi nous présenterons celle des attributs quantitatifs.

2.3.1 Corrélation linéaire : variables quantitatives continue

Suite a la corrélation linéaire, nous observons une forte corrélation positive entre toutes les paires d'attributs continue. avec les valeur de r très proches de "1"

Linear correlation 1

Parameters

Cross-tab parameters	
Sort results	yes
Sort criterion	Y attribute name
Input list	Target (Y) and input (X)

Results

Y	X	r	r²	t	Pr(>|t|)
Diametre	Poids	0.9255	0.8565	157.8331	0.0000
Diametre	Hauteur	0.8337	0.6950	97.5439	0.0000
Diametre	Longueur	0.9868	0.9738	393.9017	0.0000
Diametre	Poids-co	0.9053	0.8196	137.7347	0.0000
Diametre	Poids-v	0.8997	0.8095	133.1971	0.0000
Diametre	Poids-c	0.8932	0.7977	128.3225	0.0000
Hauteur	Poids	0.8192	0.6711	92.3022	0.0000
Hauteur	Diametre	0.8337	0.6950	97.5439	0.0000
Hauteur	Longueur	0.8276	0.6848	95.2494	0.0000
Hauteur	Poids-co	0.8173	0.6680	91.6617	0.0000
Hauteur	Poids-v	0.7983	0.6373	85.6524	0.0000
Hauteur	Poids-c	0.7750	0.6006	79.2320	0.0000
Longueur	Poids-c	0.8979	0.8062	131.8077	0.0000
Longueur	Poids-v	0.9030	0.8154	135.8178	0.0000
Longueur	Poids-co	0.8977	0.8059	131.6503	0.0000
Longueur	Diametre	0.9868	0.9738	393.9017	0.0000
Longueur	Hauteur	0.8276	0.6848	95.2494	0.0000
Longueur	Poids	0.9253	0.8561	157.6067	0.0000
Poids	Poids-c	0.9694	0.9397	255.1785	0.0000
Poids	Poids-v	0.9664	0.9339	242.8344	0.0000
Poids	Poids-co	0.9554	0.9127	208.9277	0.0000
Poids	Longueur	0.9253	0.8561	157.6067	0.0000
Poids	Diametre	0.9255	0.8565	157.8331	0.0000
Poids	Hauteur	0.8192	0.6711	92.3022	0.0000
Poids-c	Poids	0.9694	0.9397	255.1785	0.0000
Poids-c	Poids-v	0.9320	0.8686	166.0921	0.0000
Poids-c	Poids-co	0.8826	0.7790	121.3157	0.0000
Poids-c	Longueur	0.8979	0.8062	131.8077	0.0000
Poids-c	Diametre	0.8932	0.7977	128.3225	0.0000
Poids-c	Hauteur	0.7750	0.6006	79.2320	0.0000
Poids-co	Hauteur	0.8173	0.6680	91.6617	0.0000
Poids-co	Diametre	0.9053	0.8196	137.7347	0.0000
Poids-co	Longueur	0.8977	0.8059	131.6503	0.0000
Poids-co	Poids-v	0.9077	0.8238	139.7320	0.0000
Poids-co	Poids-c	0.8826	0.7790	121.3157	0.0000
Poids-co	Poids	0.9554	0.9127	208.9277	0.0000
Poids-v	Hauteur	0.7983	0.6373	85.6524	0.0000
Poids-v	Diametre	0.8997	0.8095	133.1971	0.0000
Poids-v	Longueur	0.9030	0.8154	135.8178	0.0000
Poids-v	Poids-co	0.9077	0.8238	139.7320	0.0000
Poids-v	Poids-c	0.9320	0.8686	166.0921	0.0000
Poids-v	Poids	0.9664	0.9339	242.8344	0.0000

Computation time : 0 ms.
Created at 19/03/2018 8:23:42 AM

FIGURE 4 – Étude statistique de la variable poids

3 TP2 : Analyse Factorielle

3.1 Analyses Factorielle des données mixtes

Diverses méthodes au niveau de l'analyse factorielle existent telle que l'ACP, AFC et l' ACM. qui sont utiles en fonction de type de données du jeu de données.
Dans le cadre de notre étude, nous opterons pour une méthode peu connues au niveau de la littérature pour l'analyse factorielle, il s'agit de l'AFDM l'analyse factorielle des données mixtes pour les jeux de données comportant a la fois les attributs quantitatifs et qualitatifs. Ci-dessous l'**AFDM** réalise avec TANAGRA.

Result

Y	X	r	r²	t	Pr(>|t|)
Diametre	Poids	0.9255	0.8565	157.8331	0.0000
Diametre	Hauteur	0.8337	0.6950	97.5439	0.0000
Diametre	Longueur	0.9868	0.9738	393.9017	0.0000
Diametre	Poids-co	0.9053	0.8196	137.7347	0.0000
Diametre	Poids-v	0.8997	0.8095	133.1971	0.0000
Diametre	Poids-c	0.8932	0.7977	128.3225	0.0000
Hauteur	Poids	0.8192	0.6711	92.3022	0.0000
Hauteur	Diametre	0.8337	0.6950	97.5439	0.0000
Hauteur	Longueur	0.8276	0.6848	95.2494	0.0000
Hauteur	Poids-co	0.8173	0.6680	91.6617	0.0000
Hauteur	Poids-v	0.7983	0.6373	85.6524	0.0000
Hauteur	Poids-c	0.7750	0.6006	79.2320	0.0000
Poids-v	Hauteur	0.7983	0.6373	85.6524	0.0000
Poids-v	Diametre	0.8997	0.8095	133.1971	0.0000
Poids-v	Longueur	0.9030	0.8154	135.8178	0.0000
Poids-v	Poids-co	0.9077	0.8238	139.7320	0.0000
Poids-v	Poids-c	0.9320	0.8686	166.0921	0.0000
Poids-v	Poids	0.9664	0.9339	242.8344	0.0000
Longueur	Poids-c	0.8979	0.8062	131.8077	0.0000
Longueur	Poids-v	0.9030	0.8154	135.8178	0.0000
Longueur	Poids-co	0.8977	0.8059	131.6503	0.0000
Longueur	Diametre	0.9868	0.9738	393.9017	0.0000
Longueur	Hauteur	0.8276	0.6848	95.2494	0.0000
Longueur	Poids	0.9253	0.8561	157.6067	0.0000
Poids	Poids-c	0.9694	0.9397	255.1785	0.0000
Poids	Poids-v	0.9664	0.9339	242.8344	0.0000
Poids	Poids-co	0.9554	0.9127	208.9277	0.0000
Poids	Longueur	0.9253	0.8561	157.6067	0.0000
Poids	Diametre	0.9255	0.8565	157.8331	0.0000
Poids	Hauteur	0.8192	0.6711	92.3022	0.0000
Poids-c	Poids	0.9694	0.9397	255.1785	0.0000
Poids-c	Poids-v	0.9320	0.8686	166.0921	0.0000
Poids-c	Poids-co	0.8826	0.7790	121.3157	0.0000
Poids-c	Longueur	0.8979	0.8062	131.8077	0.0000
Poids-c	Diametre	0.8932	0.7977	128.3225	0.0000
Poids-c	Hauteur	0.7750	0.6006	79.2320	0.0000
Poids-co	Hauteur	0.8173	0.6680	91.6617	0.0000
Poids-co	Diametre	0.9053	0.8196	137.7347	0.0000
Poids-co	Longueur	0.8977	0.8059	131.6503	0.0000
Poids-co	Poids-v	0.9077	0.8238	139.7320	0.0000
Poids-co	Poids-c	0.8826	0.7790	121.3157	0.0000
Poids-co	Poids	0.9554	0.9127	208.9277	0.0000

FIGURE 5 – Analyse Factorielle des Données Mixtes

3.2 Interprétations des résultats obtenues

Ce tableau indique la variance expliquée par les axes,ainsi nous avons P=36 facteurs avec 7 variables quantitatives et 1 variable qualitative.Ainsi la somme des valeurs propres est égale a 36,et les 36 axes démontrent 100 pour cent de l'inertie totale.nous constatons également que Tanagra a mis en subrillance la première valeur du premier axe qui est largement supérieur au seuil dont la valeur est calculer comme suit : seuil=1+1,65 racine carrée de(P-1)sur(N-1) avec P=36 et N=4177 d'où le seuil =0,2414. Ainsi le premier axe(7,26) dépasse le seuil et il représente a lui seul 20 % de l'information disponible.Pour le choix des axes, l'on prendra les deux premiers axes.

3.3 Le Tableau des variables propres

3.3.1 Corrélation des attributs continus

Ce tableau précise le sens des relations entre les variables quantitatives et les facteurs. Le tableau

Continuous Attributes - Correlation (Factor Loadings)

Attribute	Axis_1	Axis_2	Axis_3
Longueur	-0.965292	0.073153	-0.114334
Diametre	-0.968188	0.059110	-0.101249
Hauteur	-0.878682	-0.015010	-0.066290
Poids	-0.972845	-0.034610	0.048218
Poids-c	-0.934746	0.050123	0.032324
Poids-v	-0.950893	-0.016817	0.052260
Poids-co	-0.950100	-0.111189	0.054585

FIGURE 6 – Tableau des corrélations des attributs continus

ci-dessus nous montre sur l'axe 1 que les ormeaux de longue taille ont également un diamètre grand et aussi la forte dépendance du poids aux poids concasse, poids viscéral et poids de la coquille.

3.3.2 Corrélation des attributs discrets

Ci-dessous le tableau des corrélations des variables discrètes

Discrete Attributes - Conditional means and contributions

Attribute		Axis_1			Axis_2			Axis_3		
-		Mean	CTR (%)	v.test	Mean	CTR (%)	v.test	Mean	CTR (%)	v.test
Sexe	M	-1.0032	0.70	-18.262	0.7173	16.51	34.069	0.4884	7.92	23.395
	F	-1.4146	1.18	-22.880	-0.7764	16.54	-32.766	-0.2984	2.53	-12.701
	I	2.5199	3.86	41.555	-0.0606	0.10	-2.606	-0.2655	2.05	-11.520
	Tot.	-	5.74	-	-	33.15	-	-	12.50	-
Anneau	15	-1.3883	0.09	-5.291	0.2981	0.19	2.964	0.9968	2.22	9.997
	7	2.6496	1.24	20.408	0.3448	0.98	6.930	-0.7615	4.93	-15.436
	9	-0.3141	0.03	-3.346	0.5256	4.00	14.608	-0.1836	0.50	-5.147
	10	-1.1857	0.40	-12.022	0.5350	3.81	14.152	0.4959	3.39	13.231
	8	0.9931	0.25	9.443	0.7373	6.48	18.292	-1.0311	13.12	-25.802
	20	-2.3971	0.07	-4.547	-1.8004	1.77	-8.911	0.4624	0.12	2.308
	16	-1.8986	0.11	-5.810	-1.4646	3.02	-11.694	0.0861	0.01	0.693
	19	-2.0072	0.06	-4.227	-1.1799	0.94	-6.483	0.3457	0.08	1.916
	14	-1.5587	0.14	-6.589	-0.9126	2.20	-10.065	0.1429	0.06	1.590
	11	-1.8711	0.77	-16.292	0.2273	0.53	5.164	0.3167	1.06	7.257
	12	-1.7004	0.35	-10.650	-0.9449	5.01	-15.442	0.0421	0.01	0.694
	18	-2.0642	0.08	-4.986	-1.7209	2.61	-10.846	-0.1255	0.01	-0.798
	13	-1.5009	0.21	-8.131	-0.6260	1.67	-8.848	0.3014	0.40	4.297
	5	4.9107	1.26	19.804	-1.2338	3.68	-12.982	1.7268	7.45	18.327
	4	5.7315	0.85	16.158	-1.5056	2.71	-11.074	4.1755	21.59	30.978
	6	3.6409	1.56	22.437	-0.6713	2.45	-10.793	-0.8264	3.84	-13.402
	21	-2.4587	0.04	-3.417	-2.6596	2.08	-9.645	-0.2500	0.02	-0.914
	17	-2.2408	0.13	-6.373	-1.6540	3.33	-12.274	-0.0678	0.01	-0.508
	22	-2.1559	0.01	-1.960	-1.6522	0.34	-3.919	0.0503	0.00	0.120
	1	7.3285	0.02	2.718	-4.5892	0.44	-4.441	8.6911	1.64	8.483
	3	6.1863	0.26	8.902	-0.9579	0.29	-3.596	7.3596	17.65	27.868
	26	-1.9643	0.00	-0.729	9.9282	2.07	9.608	6.4241	0.90	6.270
	23	-1.9857	0.02	-2.212	-5.1993	5.11	-15.109	-1.6265	0.52	-4.767
	29	-4.6695	0.01	-1.732	-11.5205	2.79	-11.149	-8.3579	1.52	-8.158
	2	6.7041	0.02	2.487	-3.6194	0.28	-3.503	5.8801	0.75	5.739
	27	-4.3109	0.02	-2.261	-5.4340	1.24	-7.438	3.5029	0.53	4.836

FIGURE 7 – Tableau des corrélations des attributs discrets

3.4 Représentation graphique

3.4.1 Tableau de corrélation

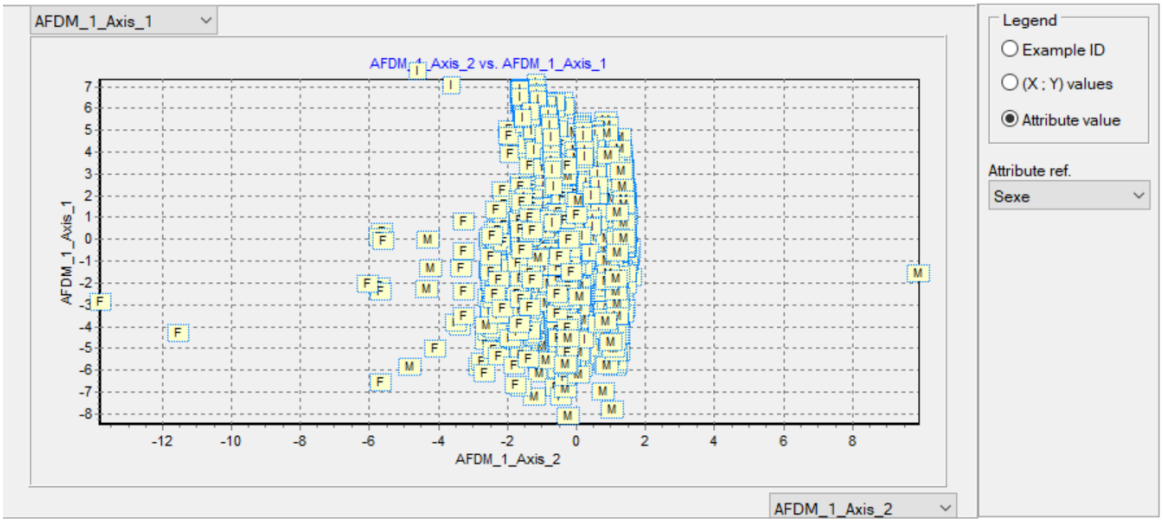


Figure 8 – Tableau des corrélations

3.4.2 Cercle de corrélation

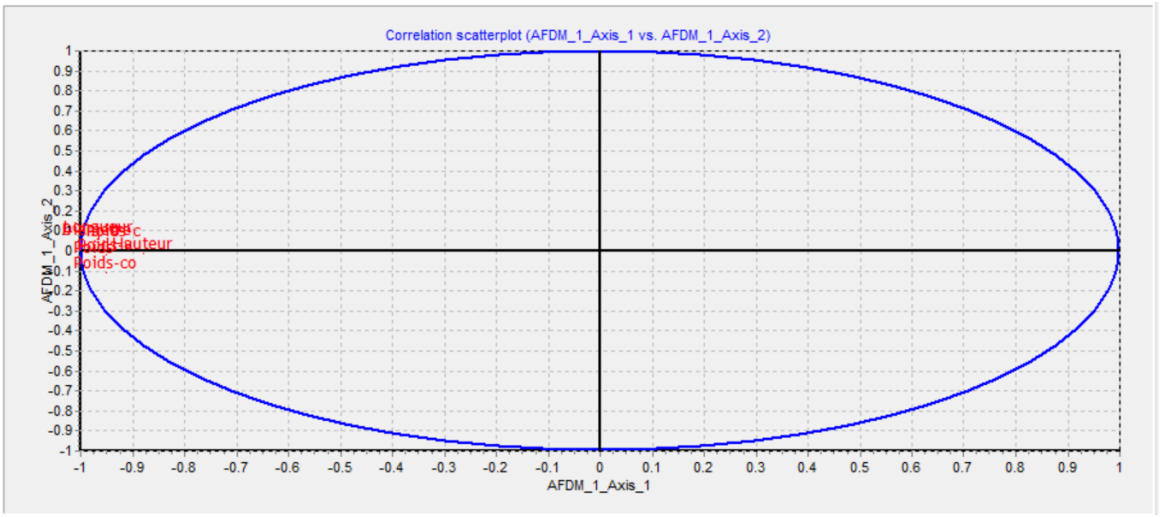


Figure 9 – Cercle de corrélations

Ici le cercle de corrélation ne nous apporte rien de nouveau car précédemment nous avons pu observer la corrélation positive entre les Poids, Poids concasse, Poids viseral et Poids de la coquille.

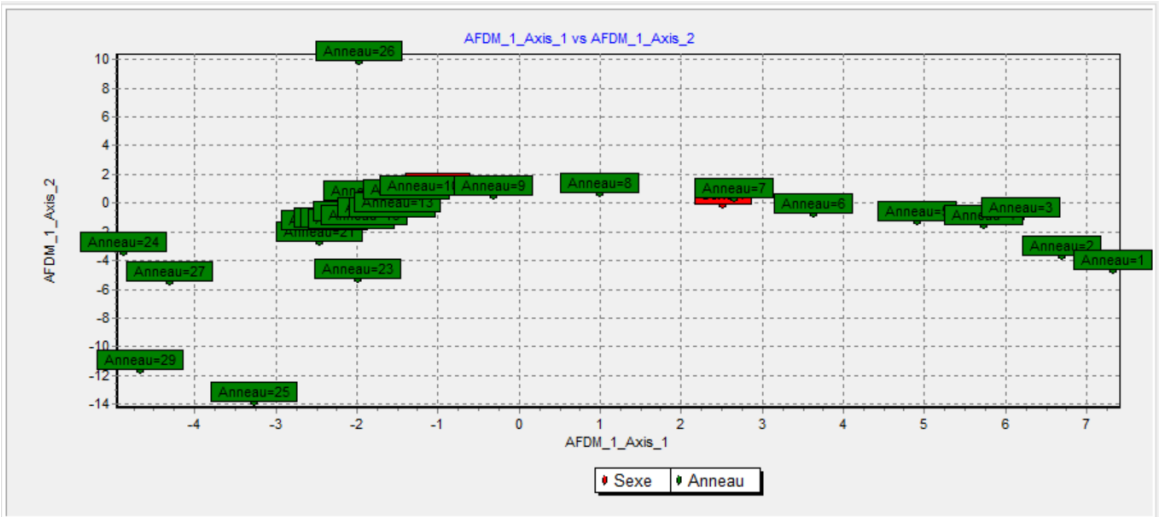


FIGURE 10 – Cercle de corrélations

4 TP3 : Clustering

4.1 Classification Automatique CAH

Report | Dendrogram

Results

Clustering results

Clusters	From the dendrogram	After one-pass relocation
cluster n°1	839	839
cluster n°2	2081	2081
cluster n°3	1257	1257

FIGURE 11 – Résultat de Clustering

Report | Dendrogram

Best cluster selection

Clusters	BSS ratio	Gap
1	0.0000	0.0000
2	0.3355	1.8515
3	0.4065	0.3329
4	0.4299	0.0760
5	0.4425	0.0513
6	0.4478	0.0089
7	0.4518	0.0072
8	0.4547	0.0093
9	0.4564	0.0016
10	0.4578	0.0026

FIGURE 12 – Selection du meilleur cluster

Report | Dendrogram

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
Longueur	0.368331	0.586060	0.525135
Diametre	0.278272	0.460697	0.406953
Hauteur	0.091728	0.160714	0.136321
Poids	0.291303	1.089029	0.756551
Poids-c	0.133496	0.458877	0.345387
Poids-v	0.062794	0.237189	0.165525
Poids-co	0.082660	0.319000	0.210346

Use GROUP CHARACTERIZATION for detailed comparisons

FIGURE 13 – Tableau des valeurs du centre des clusters

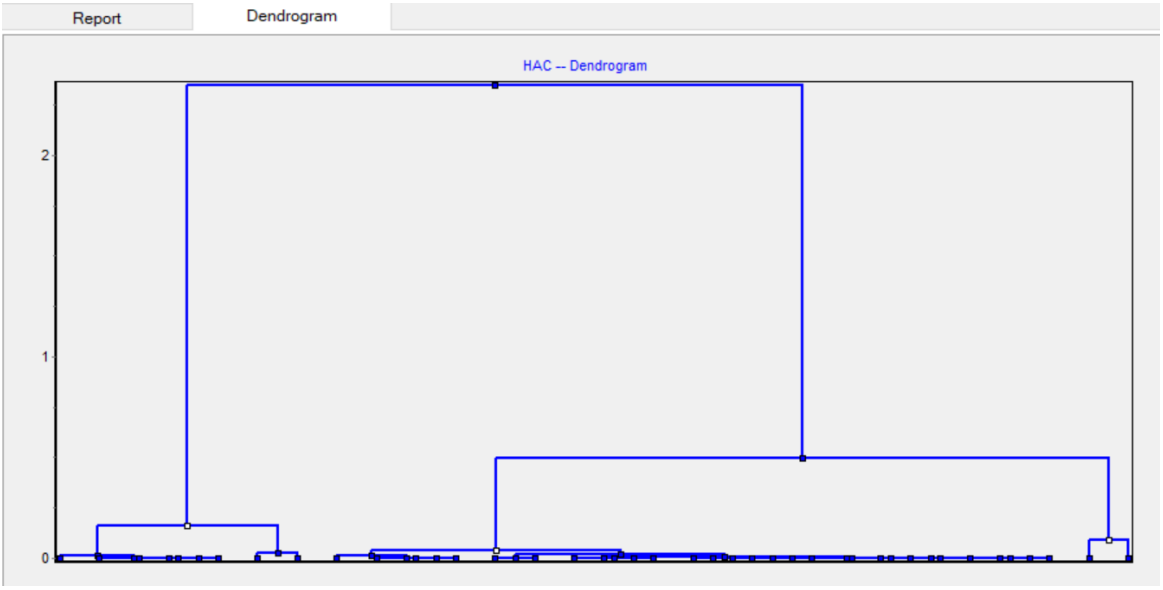


FIGURE 14 – Diagramme de clustering

4.2 Description et Interprétation des clusters

Pour notre classification, nous avons utilisé la variable qualitative (Anneau) en Target et les variables quantitatives (Longueur, Poids, Hauteur, Diametres ..) Suite a notre classification, Tanagra a détectée 3 clusters dont :

- ***Cluster 1 (valeur de l'anneau comprise entre 1 et 7)***
- ***Cluster 2 (valeur de l'anneau comprise entre 10 et 29)***
- ***Cluster 3 (valeur de l'anneau comprise entre 8 et 9)***

Cluster 1 représente 20,1 pourcent des individus (839) de notre Dataset.nous observons a partir de ***la figure 13*** que le centre du cluster 1 est assez éloigné du centre des Clusters 1 et 2 ce qui implique que le cluster 1 est stable.

Cluster 2 représente 49,8 pourcent des individus (2081) de notre Dataset.nous observons a partir de ***la figure 13*** ci-dessus que le centre du cluster 2 est assez proche du centre du Cluster 3 ce qui implique que le cluster 2 n'est pas stable et peut être regrouper avec le cluster 3 tel que nous l'observons sur le diagramme de Clustering de ***la figure 14*** ci-dessus pour donner le Cluster 4. Le Cluster 5 sera donc l'ensemble de nos individus du Cluster 1 et du Cluster 4.

Cluster 3 représente 30,1 pourcent des individus (1257) de notre Dataset.Tanagra l'a mis en surbrillance comme étant le meilleur cluster comme le montre ***la figure 12*** ci-dessus . ce qui implique qu'il est stable.

5 TP4 : Modèle de prédiction

5.1 Rappel du problème

Nous disposons d'un jeu de données de 4177 observations reparties en 08 attributs d'entrées dont 07 quantitatives et 01 qualitatif et 01 attribut de sortie qui est la longueur d'anneau (**Ring**). Celles-ci sont nos variables explicatives de la variable expliquée (Variable à prédire)

5.2 Choix de l'algorithme d'apprentissage automatique

Pour la construction de notre modèle, différents algorithme d'apprentissage automatique existent en fonction du problème à résoudre. Dans notre cas, il s'agit à partir de variables explicatives pouvoir prédire l'anneau (Ring) qui nous permettra de déterminer l'âge en ajoutant 1.5.
Nous utiliserons donc **la régression linéaire multiple** car elle permet de traiter ce type de problème.

5.3 Pré-Traitement des données

Dans le cadre de la régression linéaire multiple, il s'agira pour nous dans un premier temps de s'assurer que notre jeu de données ne comporte pas de valeurs manquantes, par la suite nous procéderons à la sélection des variables explicatives pour la création de notre modèle.

5.3.1 Traitements des valeurs manquantes

Le TP1 relatif à la description des données nous a permit de savoir que notre jeu de données ne comporte pas de valeurs manquantes.

5.3.2 Choix des variables explicatives

Nous avons procéder à une dichotomie de la variable qualitative sex afin de la rendre quantitative comme l'exige la régression linéaire. Pouvant avoir trois valeur, elle sera remplacer par MALE, FEMALE et INFANT.

	LENGTH	DIAMETER	HEIGHT	WHOLE WEIGHT	SHUCKED WEIGHT	VISCERA WEIGHT	SHELL WEIGHT	RINGS	MALE	FEMALE	INFANT
COUNT	4177.000000	4177.000000	4177.000000	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	0.365813	0.312904	0.321283
STD	0.120093	0.099240	0.041827	0.490389	0.221963	0.109614	0.139203	3.224169	0.481710	0.463729	0.467017
MIN	0.075000	0.055000	0.000000	0.002000	0.001000	0.000500	0.001500	1.000000	0.000000	0.000000	0.000000
25%	0.450000	0.350000	0.115000	0.441500	0.186000	0.093500	0.130000	8.000000	0.000000	0.000000	0.000000
50%	0.545000	0.425000	0.140000	0.799500	0.336000	0.171000	0.234000	9.000000	0.000000	0.000000	0.000000
75%	0.615000	0.480000	0.165000	1.153000	0.502000	0.253000	0.329000	11.000000	1.000000	1.000000	1.000000
MAX	0.815000	0.650000	1.130000	2.825500	1.488000	0.760000	1.005000	29.000000	1.000000	1.000000	1.000000

FIGURE 15 – Structure des données après dichotomie

La figure ci-dessous nous présente la très-forte corrélation des variables **Shucked Weigth, Viscera Weight et Shell Weight** avec **Whole Weigth**. Pour la suite nous ne les prendrons pas en compte pour la réalisation du modèle

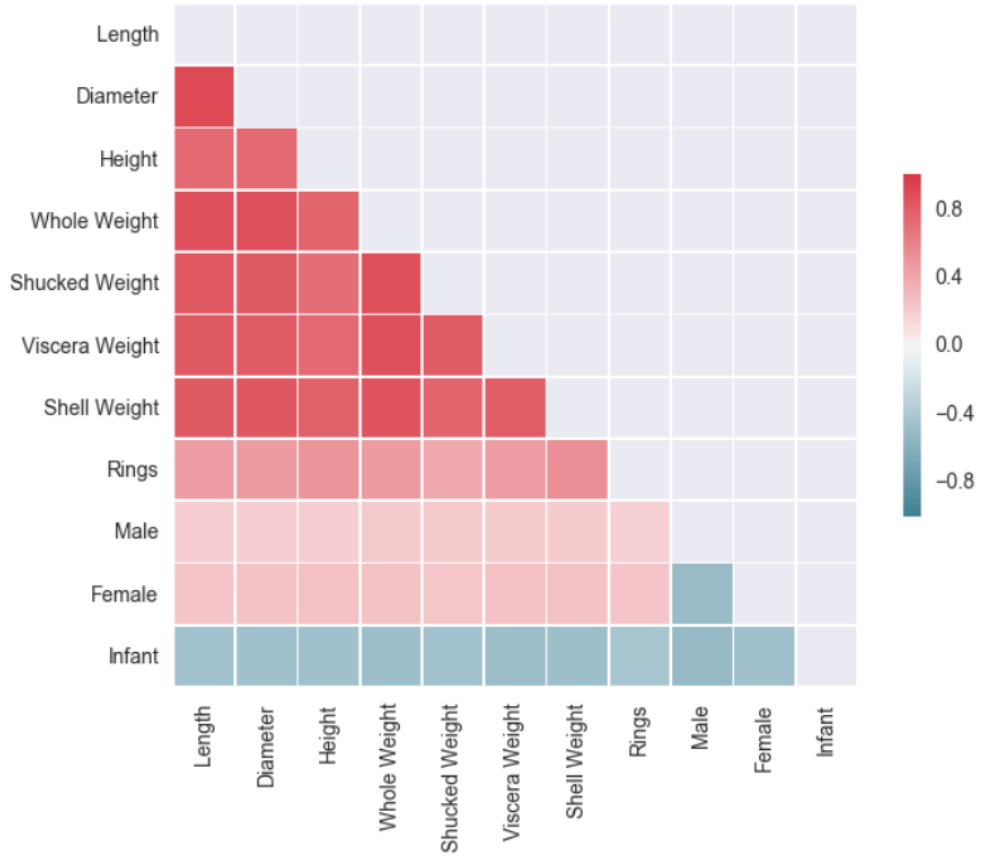


FIGURE 16 – Corrélation des variables explicatives

5.4 Initialisation du modèle

5.4.1 Paramètres

Pour commencer nous choisirons toutes les variables explicatives afin de pouvoir déterminer la limite de la précision de la prédiction que l'on obtiendra. Par la suite, nous réduirons le nombre de variables explicatives pour améliorer la précision du modèle.
Au niveau du partitionnement du jeu de données, nous avons scindes le jeu de données en deux, 70% (2923 observations) pour l'apprentissage et 30% (1254 observations) pour le test. Ces valeurs sont les plus optimales suites aux différentes expérimentations menées.

```
{'Diameter': 10.891124440244319,
 'Female': 0.2092393007306062,
 'Height': 9.7795473302049594,
 'Infant': -0.55766071653771909,
 'Length': 0.11442193544956823,
 'Male': 0.34842141580715397,
 'Shell Weight': 7.0343534592691546,
 'Shucked Weight': -21.087688081427075,
 'Viscera Weight': -11.145502003607316,
 'Whole Weight': 10.142464126255723}
```

FIGURE 17 – Prédiction obtenue

Ci-dessous la figure du MAE pour l'évaluation de notre modèle Le MAE ne semble pas adéquate.Nous

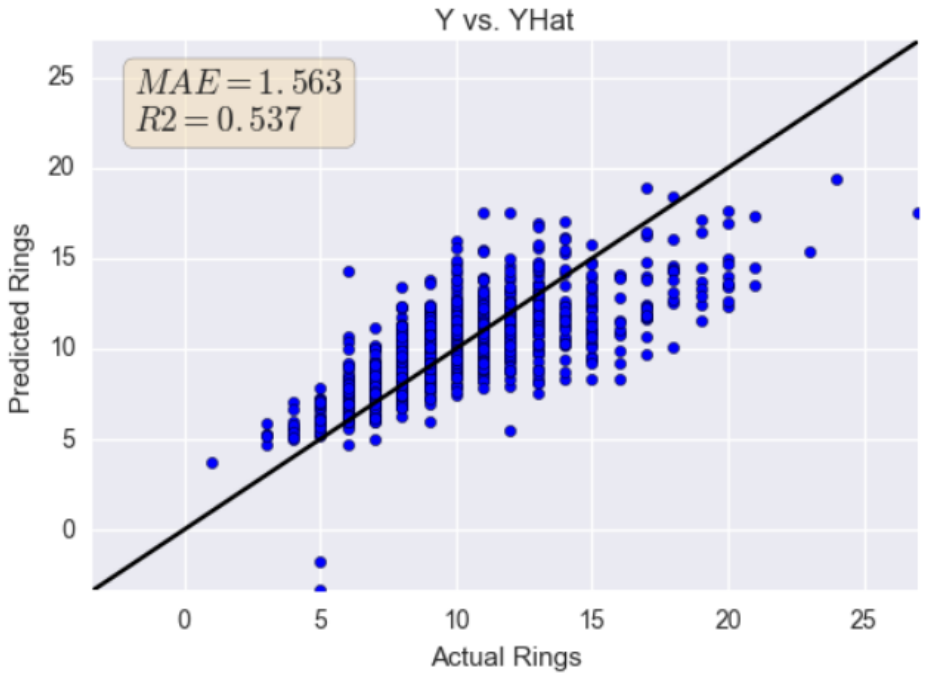
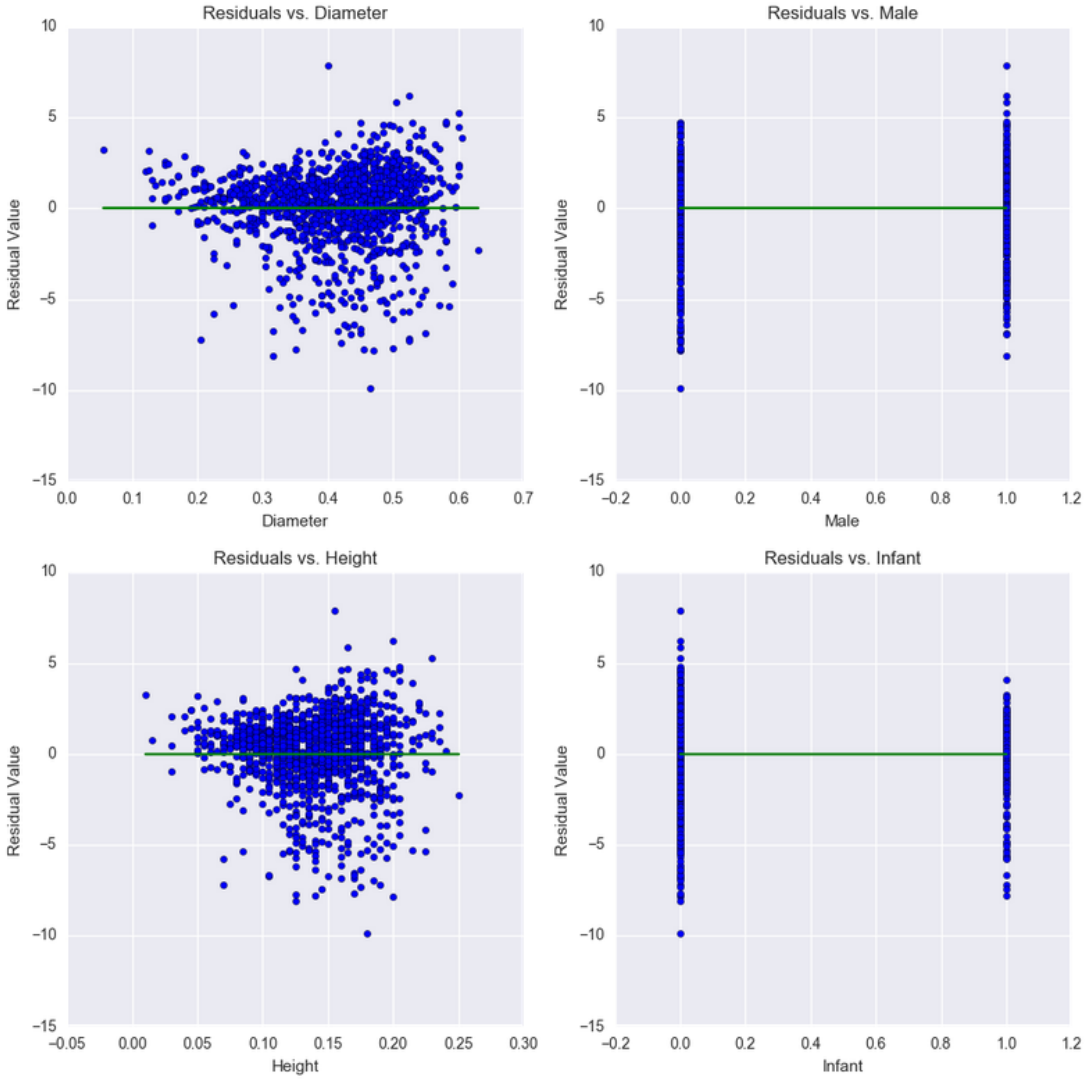


FIGURE 18 – résultat obtenu

avons donc applique une méthode de régulation (**le coefficient de Kendall**) en espérant la réduction de facteur explicatif dans notre modèle.

5.5 Validation du modèle

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. Ci-dessous la figure présentant la comparaison avec les différents résidus.



Dans l'ensemble, la performance du modèle de régression est adéquate,nous essayerons un modèle afin de comparer les performances.

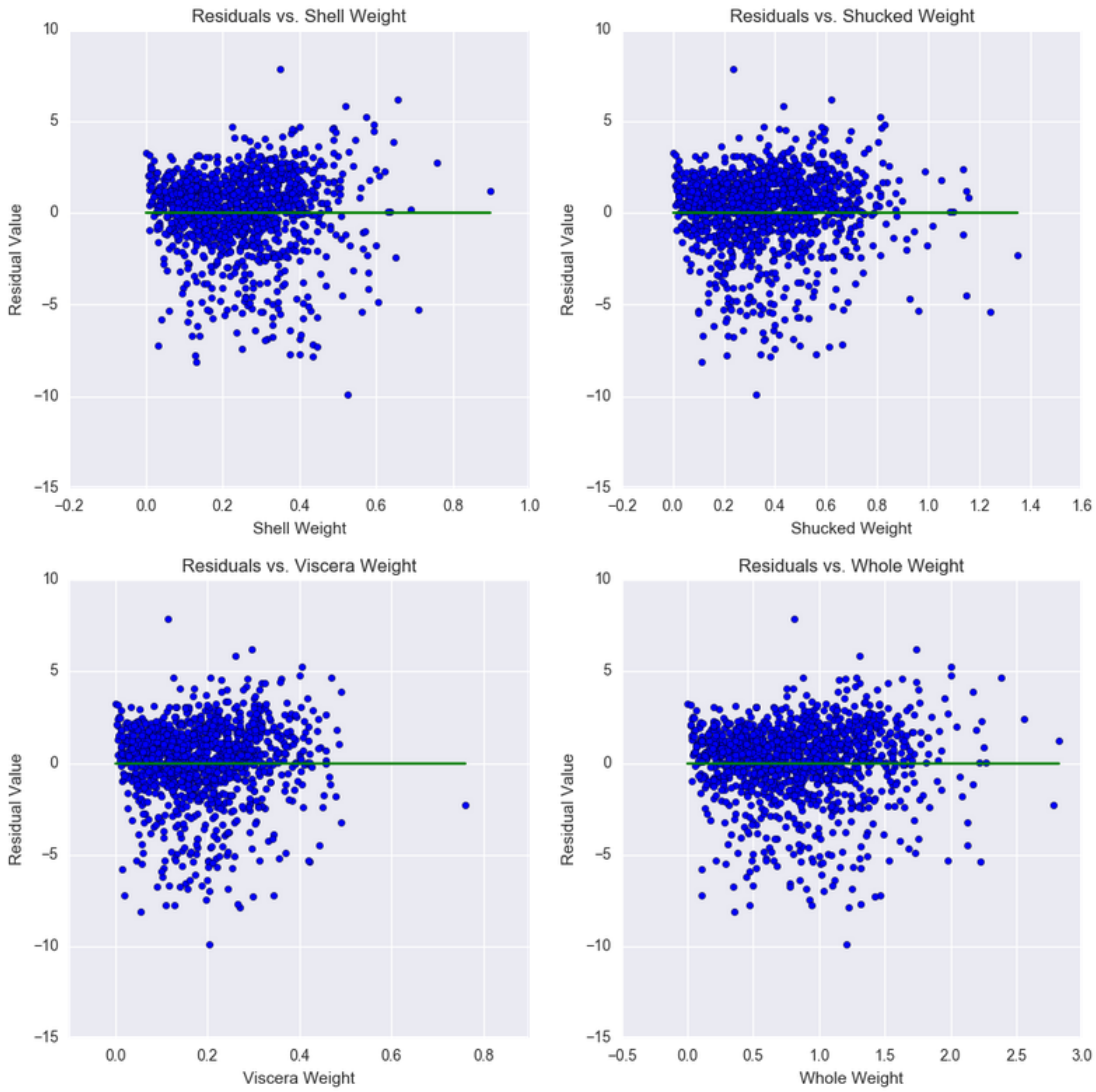


Figure 19 – Corrélation des variables explicatives

5.6 Comparaison avec un autre algorithme d'apprentissage

Notons que les données ont été remises à l'échelle pour obtenir une moyenne nulle et une variance unitaire, ce traitement afin de s'assurer que chaque variable a la même importance.

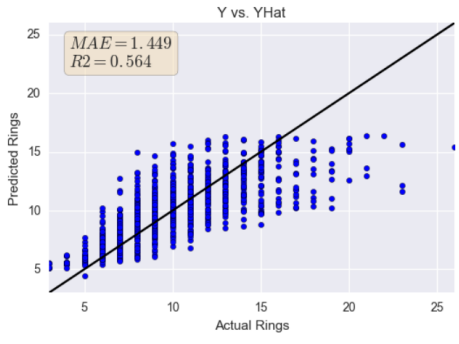


Figure 20 – MAE : Perceptron

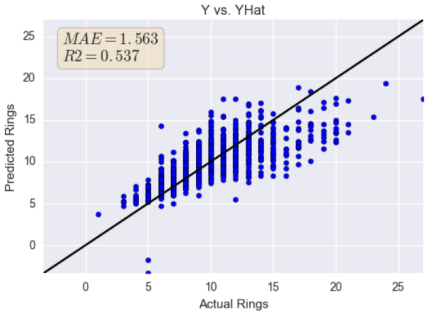


Figure 21 – MAE : Régression Linéaire

6 Conclusion

L'erreur MAE la mieux réalisée était de 1,42, 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 à ceux obtenus lors de la première partie, qui a atteint un MAE de 1,568. Malgré la grande précision du réseau de neurones, la modélisation de réseaux neuronaux présente plusieurs inconvénients, notamment la difficulté 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 réseaux de neurones pourraient être un excellent choix. Suite à notre travail, avec un MAE de 1,4, nous ne pourrons affirmer avoir répondu à la problématique car le taux d'erreur moyen est trop élevée pour pouvoir prendre en compte dans la détermination de l'âge de l'ormeau. Un ACP suite à la méthode de régression linéaire multiple pourrait nous permettre d'améliorer les performances.