RAPPORT

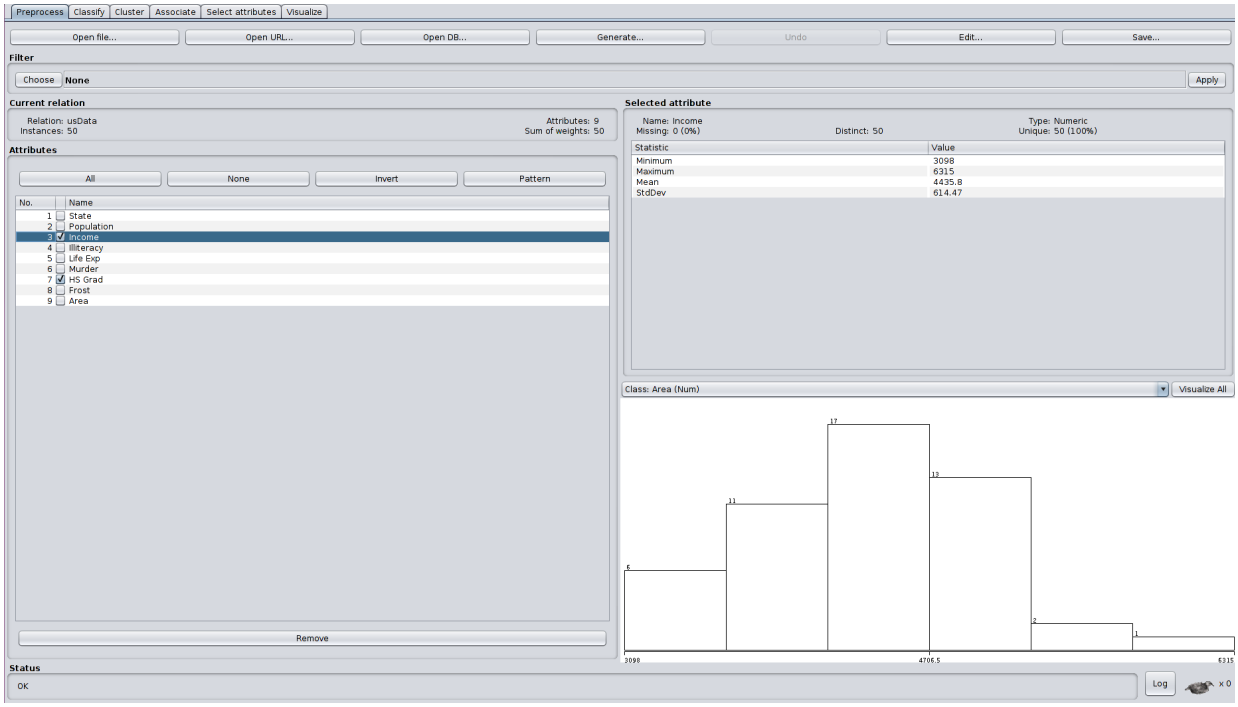
TP FOUILLE DE DONNÉES

Youssouf IBRAHIM ELMI
Mohamed Amine TOUCHENE
Mohamadou LO

EXERCICE 1

Question 1

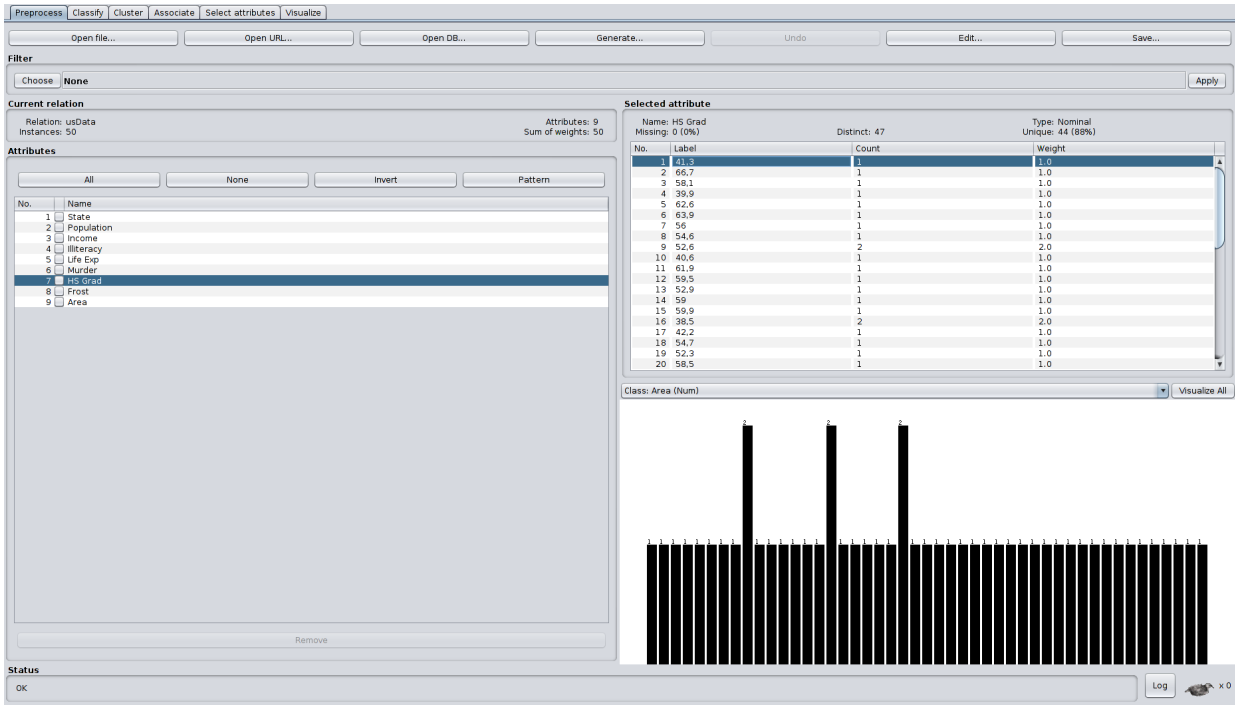
Dans un premier temps nous devrons sélectionner les deux attributs sur les quels on va faire l'analyse avec les différents algorithmes demander. Voici le statistique des deux attributs qui sont : **Income** et **HS Grad**.



La population " **Income** " <u>minimum de 3098 et un maximum de 6315</u>.

- **6** Etats sur 50 sont dans l'intervalle de [3098 ;3635].
- **11** Etats sur 50 sont dans l'intervalle de [3635 ;4171].
- **17** Etats sur 50 sont dans l'intervalle de [4171 ;4707].
- **13** Etats sur 50 sont dans l'intervalle de [4707 ;5243].
- **2** Etats sur 50 sont dans l'intervalle de [524 ;5779].
- **1** Etat sur 50 est a une seuil de PIB supérieur à 5779.

La moyenne de PIB par habitants est de **4435.8**.



Le pourcentage de " HS Grad " est d'un minimum de 37.8% et d'un maximum de 67.3% .

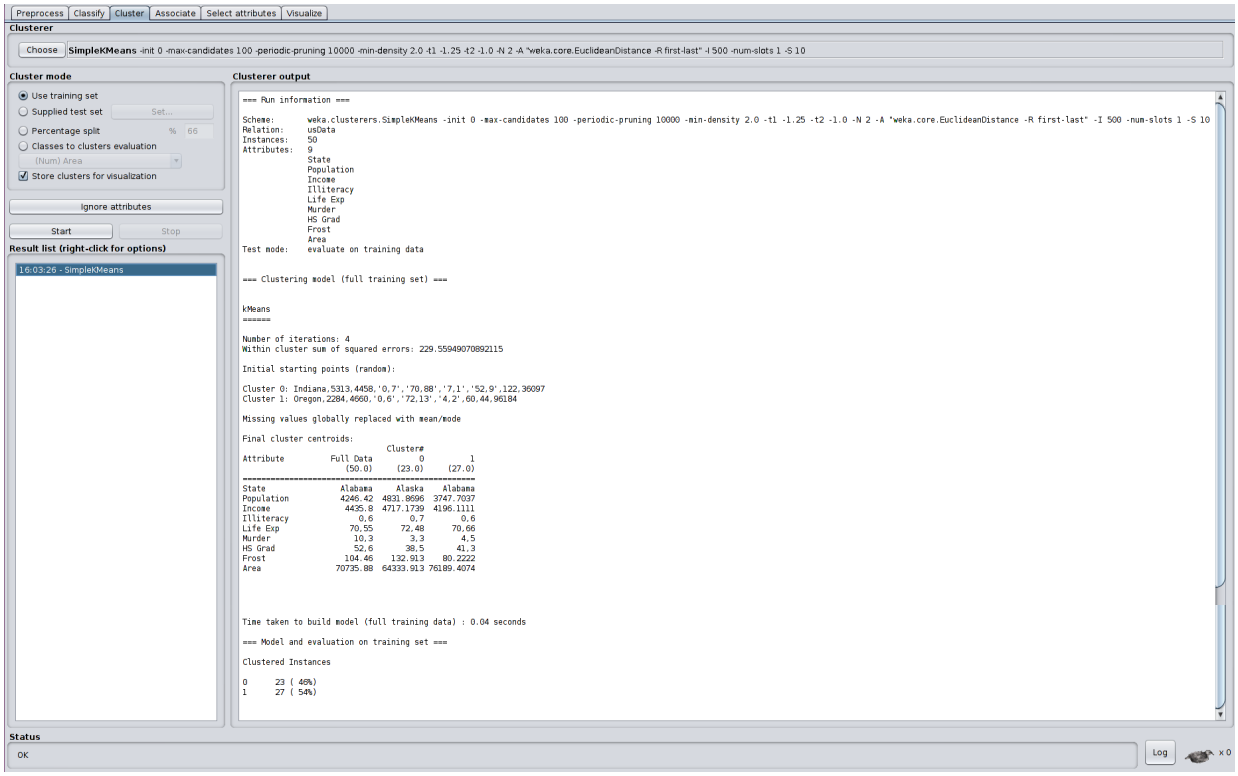
- On a 44 Etats respectives sur 47.
- On a 3 Etats respectives (52,6 ; 38,5 ; 52,7 ; 57,6) sur 47.

Les résultats des algorithmes de clustering SimpleKMeans, EM et Canopy.

1. **SimpleKMeans**

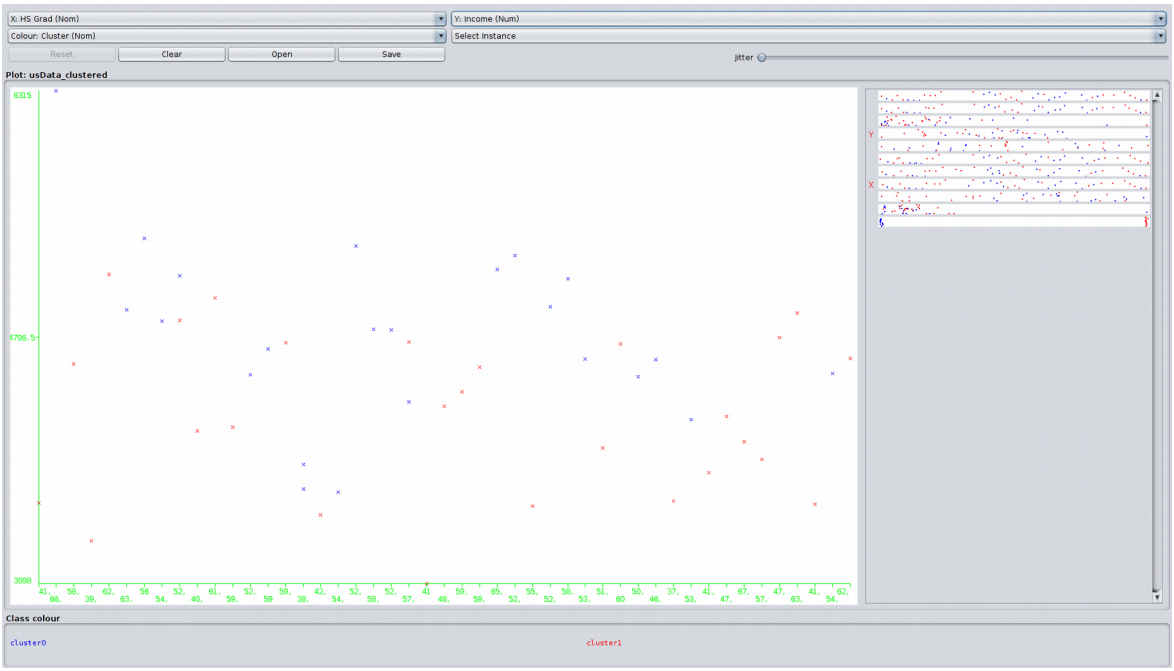
On a lancer l'algorithme de cluster SimpleKMeans, le résultat est la suivante :

Pour le premier cluster, **23** états sur 50 sont classer, soit **46%** des états.
Pour le second cluster, **27** états sur 50, soit **54%** des états.
Avec un taux d'erreur de 61% on a la répartition suivante :



Quand on visualise l'affectation du cluster dans un graphe et qu'on choisie les deux axes (X;Y)

respectivement le "HS Grad" et "Income", on remarque que le premier cluster (c’est-à-dire 10/23 états sont à un PIB par habitants faible (inférieure à la moyenne). Ce dernier analyse montre que majoritairement, plus le PIB est petit et plus le taux des diplômé du supérieur est faible.



2. EM

Dans cet algorithme le résultat est légèrement inverser à savoir que dans un premier cluster on a 14 états sur 50 sont classer dans un premier cluster (soit 28% des états).
Le deuxième cluster, on a 14 états sur 50 sont classer en cluster (soit 28% des états).
Le troisième cluster, on a 13 états sur 50 sont classer dans en (soit 26% des états).
Le quatrième, on a 09 états sur 50 sont classer en cluster (soit 18% des états).

EM
==

Number of clusters selected by cross validation: 4
Number of iterations performed: 16

Attribute	Cluster 0 (0.28)	1 (0.28)	2 (0.26)	3 (0.18)
Income				
mean	4809.223	3863.8558	4612.2507	4471.6504
std. dev.	305.4583	460.2249	422.252	751.4086
HS Grad				
41,3	1	2	1	1
66,7	1	1	1	2
58,1	2	1	1	1
39,9	1	2	1	1
62,6	2	1	1	1
63,9	1.0002	1	1.9998	1
56	1.0067	1	1.9933	1
54,6	1.0814	1	1.0031	1.9155
52,6	3	1	1	1
40,6	1.0002	1.9998	1	1
61,9	2	1	1	1
59,5	1.0001	1	1.0117	1.9882
52,9	1.0057	1.009	1.9853	1
59	1.0002	1.0002	1.9996	1
59,9	1.0057	1.0005	1.9938	1
38,5	1.0001	2.9999	1	1
42,2	1	2	1	1
54,7	1	1.0001	1.9633	1.0366
52,3	1.9962	1.0002	1.0036	1
58,5	1.9978	1.0001	1.002	1
52,8	2	1	1	1
57,6	1.0004	1	2.9878	1.0118
41	1	2	1	1
48,8	1.0064	1.9783	1.0153	1
59,2	1.0001	1	1.0013	1.9987
59,3	1.0008	1	1.9828	1.0163
65,2	1.0005	1	1.9791	1.0204
52,5	1.9998	1	1.0002	1
55,2	1	1	1.0007	1.9993
52,7	2	1	1	1
50,3	1.0007	1	1.9877	1.0116
53,2	2	1	1	1
51,6	1.0003	1.9995	1.0002	1
60	1.9994	1.0005	1	1
50,2	2	1	1	1
46,4	1.0015	1	1.0055	1.993
37,8	1	2	1	1
53,3	1	1	1.9866	1.0134
41,8	1	2	1	1
47,4	2	1	1	1
67,3	1.0001	1	1.0974	1.9025
57,1	1	1	1.0122	1.9878
47,8	1.018	1.9818	1.0002	1
63,5	1.0873	1.9127	1	1
41,6	1	1.9992	1.0007	1.0001
54,5	1.0002	1.0002	1.9996	1
62,9	1.0005	1	1.1307	1.8687
[total]	61.2102	60.8822	60.1438	55.7639
Frost				
mean	73.4582	66.1617	154.4951	140.3521
std. dev.	47.8236	28.0961	22.2363	21.9589
Area				
mean	68137.5863	48515.4745	61101.2593	124596.134
std. dev.	67801.9615	13097.3085	31593.862	166288.0051

Time taken to build model (full training data) : 0.12 seconds

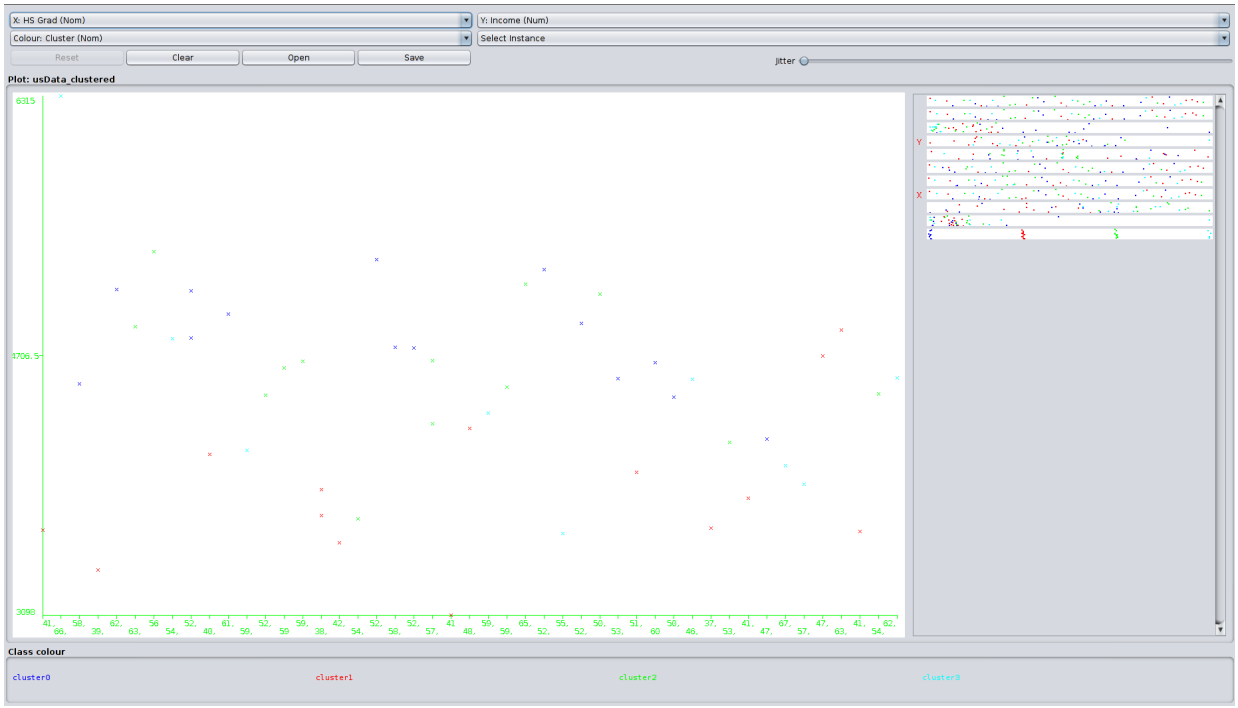
=== Model and evaluation on training set ===

Clustered Instances

0 14 (28%)
1 14 (28%)
2 13 (26%)
3 9 (18%)

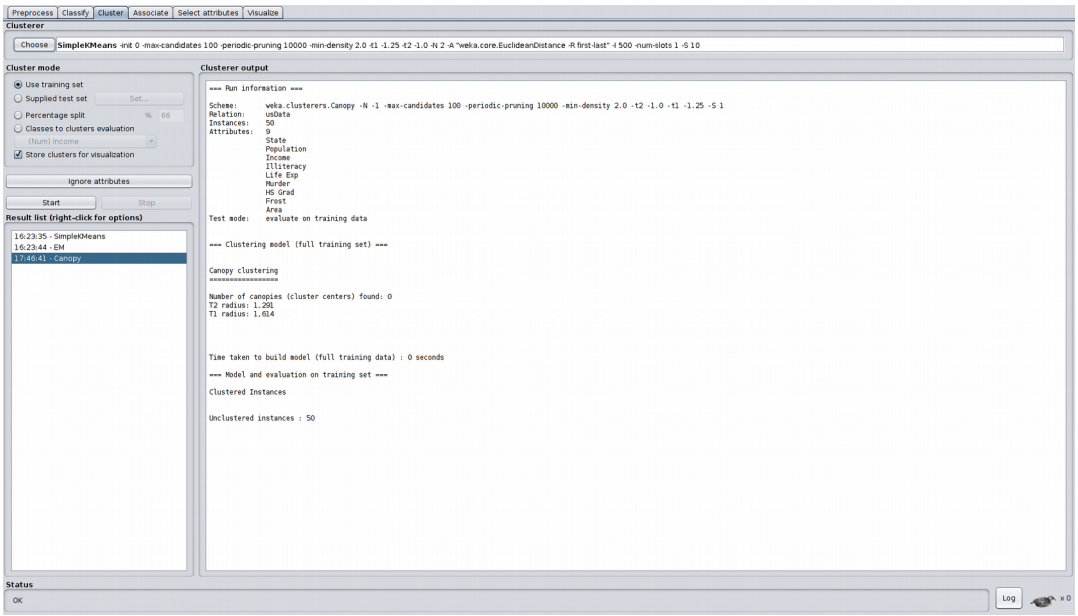
Log likelihood: -50.34243

La précédente répartition montre qu'avec l'algorithme EM le dernier cluster regroupe une bonne majorité des états qui ont une " Income " faible avec un " HS Grad " inférieur à la moyenne. Le cluster 0,1 et 2 (c’est-à-dire le trois premiers cluster) regroupe des états repartie sur un PIB plus ou moins dans la moyenne et un taux des diplômé du supérieur largement bien et même supérieurs à la moyenne. La visualisation de ce dernier résultat dans un graphe avec comme abscisse les " HS Grad " et comme ordonnées les " Income ".



3. CANOPY

Pour l’algorithme canopy, on obtient aucun cluster, c’est à dire qu’on a eu 0 états de clusters sur 50.

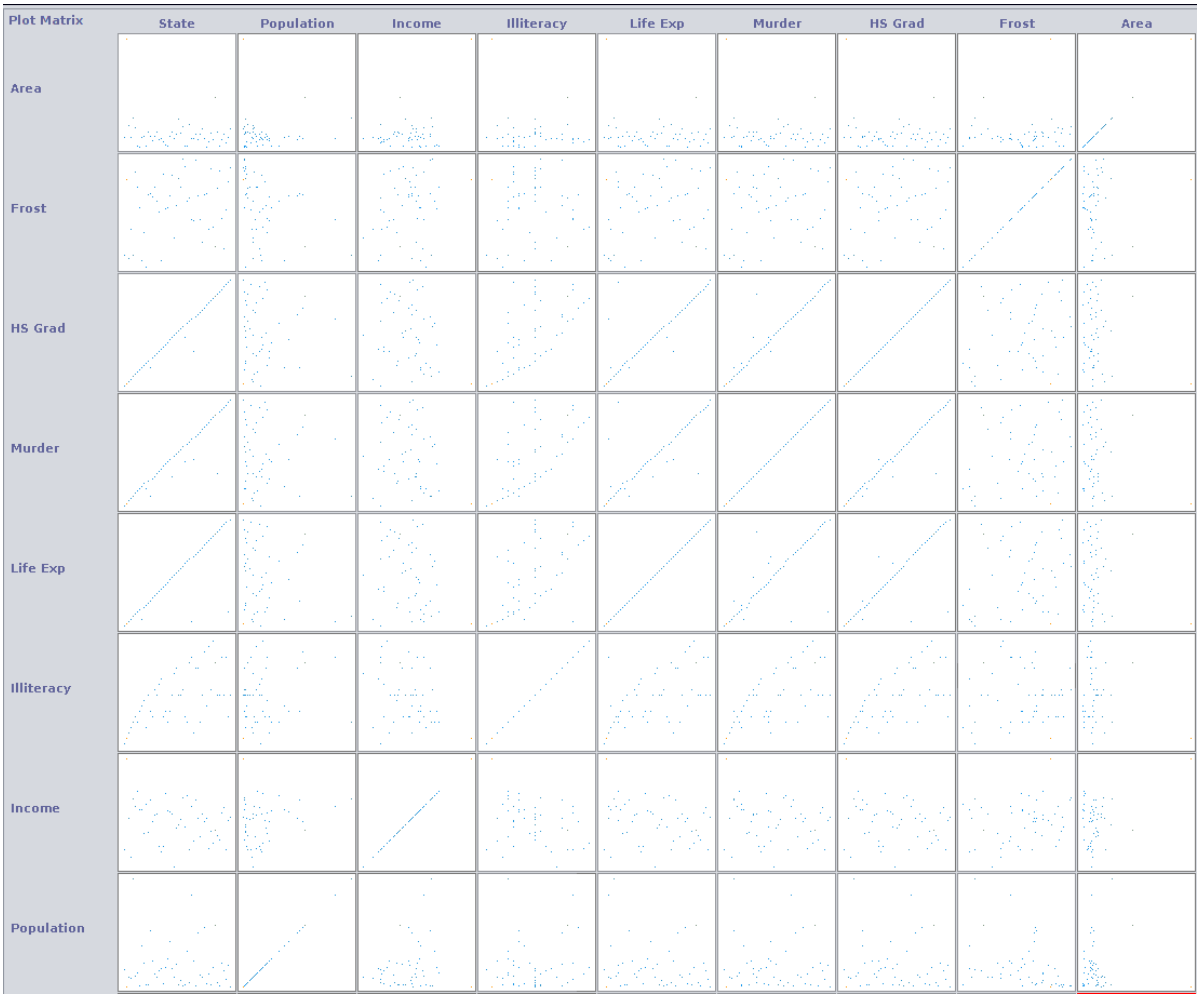


Comme on a eu aucun cluster, la répartition avec l’algorithme Canopy affiche aucun graphe puisqu’on a eu « 50 états unclustered »

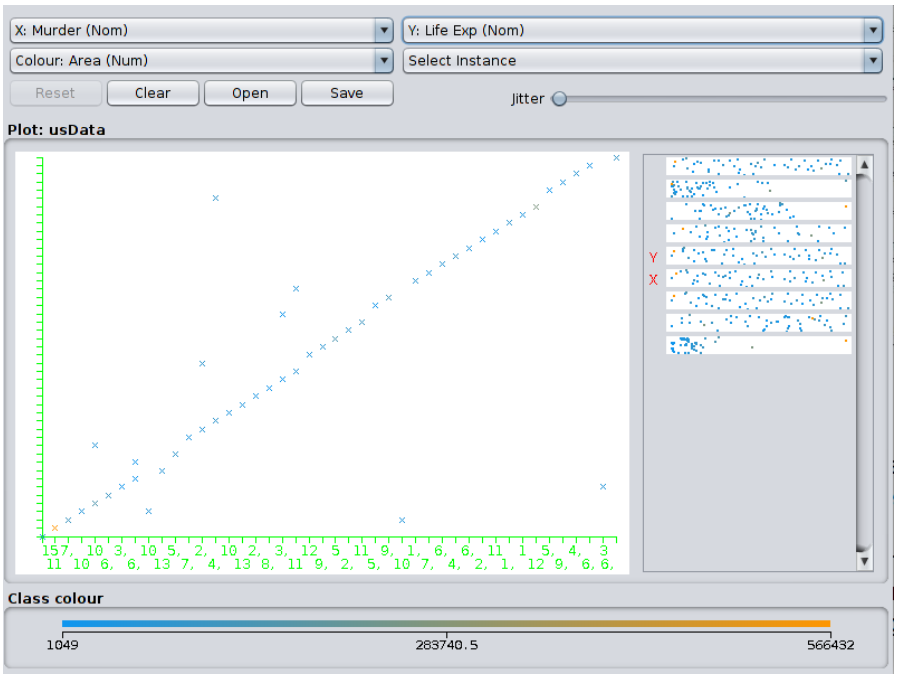
Question 2 :

Dans un premier temps, on réalise une visualisation des ensembles des attributs chacune en fonction des autres.

Nous aurons une visualisation qui est :



Prenons l'exemple le « Murder » en fonction « Life Exp ».
Nous remarquons que l'espérance de vie est plus ou moins bonne mais avec le taux de meurtre, nous observons que l'espérance de vie diminue en se concentrant sur un taux inférieure à la moyenne d'après le graphe ci-dessous.

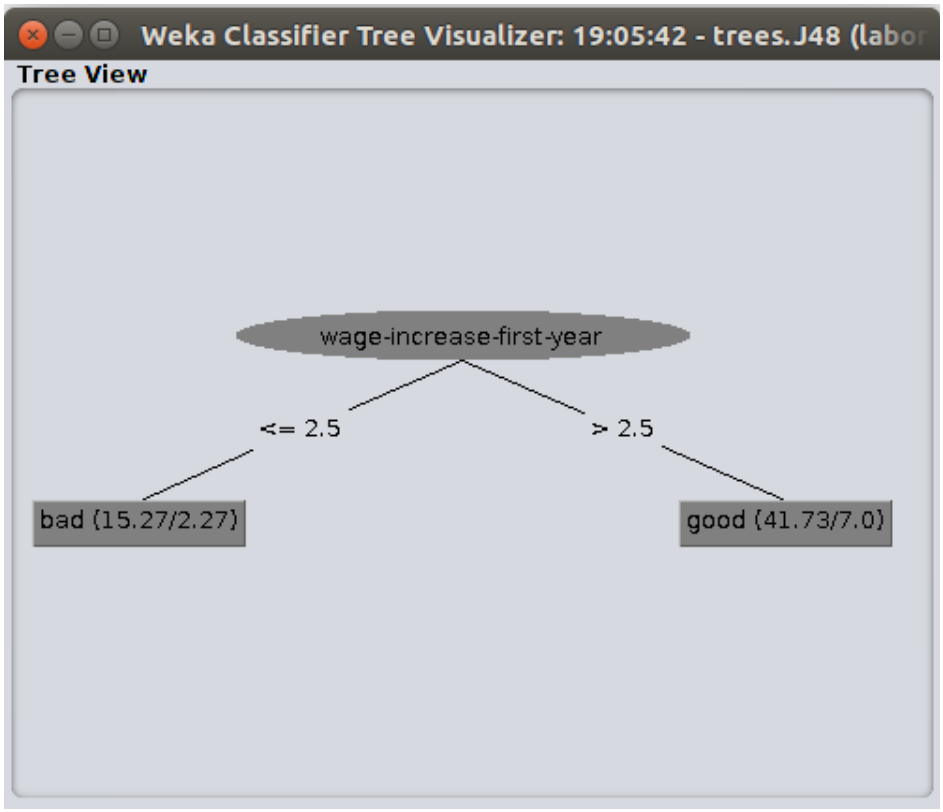

X: Murder (Nom)
Y: Life Exp (Nom)
Colour: Area (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: usData
Y
X
Class colour
1049
283740.5
566432

EXERCICE 2 :

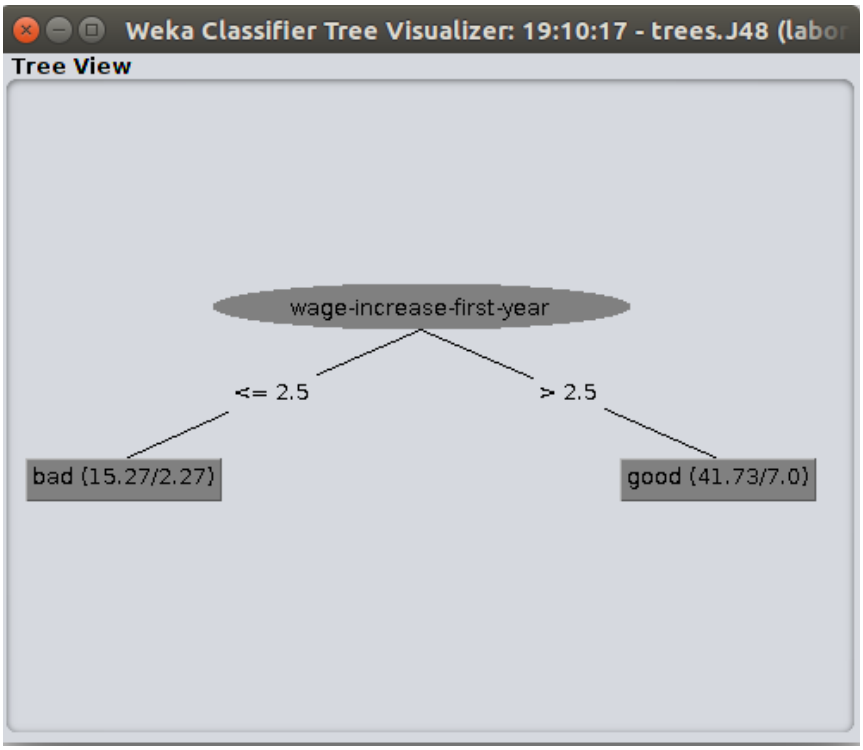
Afin de faire la classification et pouvoir tracer l'arbre de décision, on va devoir attribuer plusieurs valeurs à l'indice de confiance **(confidenceFactor)**.

Initialement on lui attribue la valeur : **0.10**.

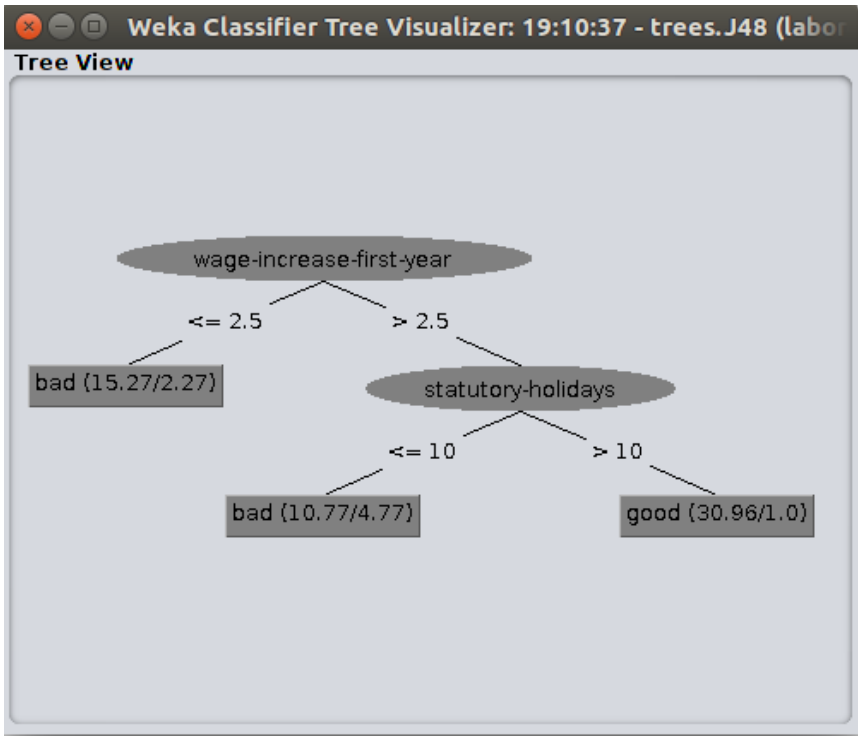
l'arbre obtenu est le suivant :



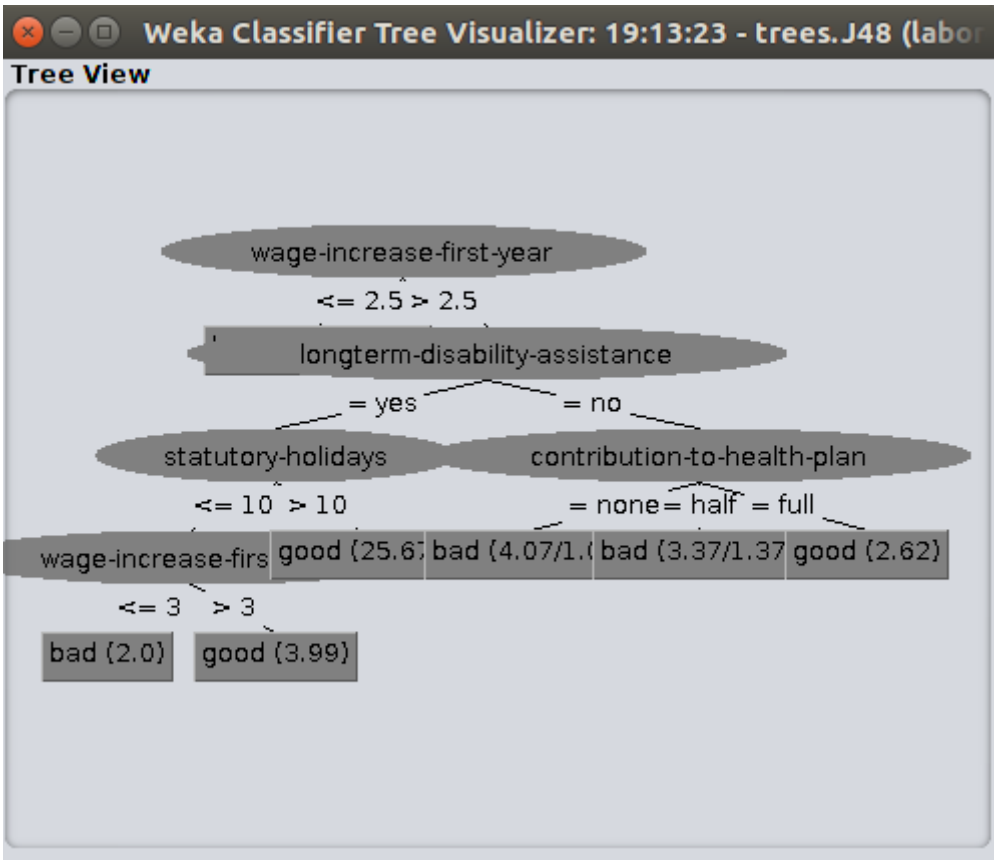
Il est de meme pour un indice de confiance à **0.20** ou l'arbre n'est toujours pas bien exploitable



Avec une valeur augmenter à **0.25**, on obtient un arbre ayant plus de nœuds que le précédent, le resultat est illustré dans la figure suivante :



En vue d’avoir un arbre avec plus d’information, la valeur **0.50** est attribuée pour le confidenceFactor. La figure suivante montre que l’arbre est bien exploitable avec un taux d’erreur moins considérable par rapport aux autres cas.



On assigne respectivement la valeur **0 .80** et **1.0** pour le facteur de confidence, on remarque que l'arbre généré n'est pas exploitable (voir les deux figures suivantes). De ce fait la valeur idéal pour le facteur est égale à **0.5** :