TP2 - Classification automatique

Chang QI (chang.qi@etu.utc.fr)

Valentin MONTUPET (valentin.montupet@etu.utc.fr)

UV : **SY09**

Semestre : **P17**

1. Visualisation des données

Dans cette première partie, nous nous contenterons de visualiser les données que nous serons amenés à utiliser par la suite. Pour ce faire, nous utiliserons l'ACP pour les données *Iris* et *Crabs* et l'AFTD pour les données *Mutations*.

Données Iris

Rappelons que le dataset Iris est composé de de 150 individus de 5 variables : 4 quantitatives correspondant à des mesures et une qualitative correspondant à l'espèce. Pour effectuer une ACP, nous excluons cette dernière.

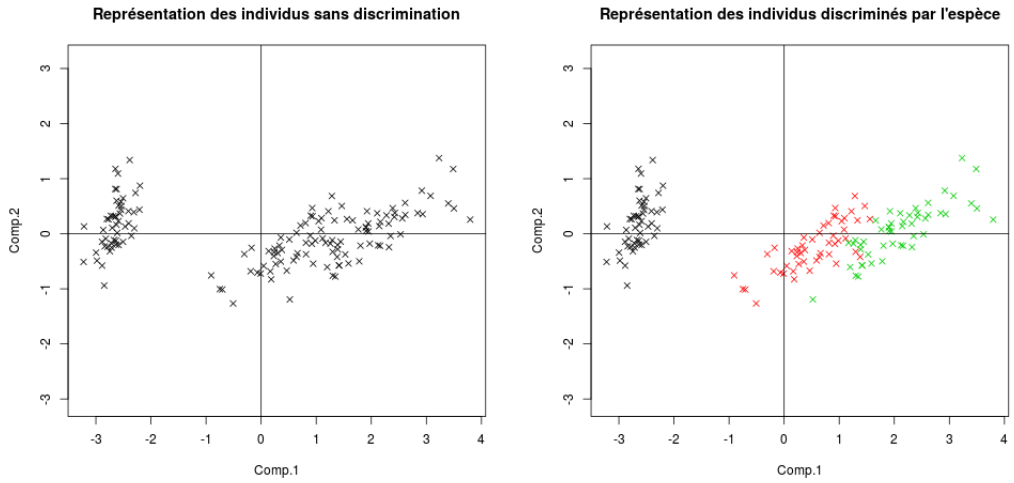

FIGURE 1.1 – Visualisation des données IRIS dans le premier plan factoriel

A première vue, sans tenir compte de l'espèce, il semble qu'il y ait 2 groupes de points distincts et très clairement séparés alors que 3 valeurs d'espèces sont présentes dans le dataset. Il semblerait donc que deux espèces soient mélangés, ou bien qu'une espèce soit « dispatchée » entre les deux autres. Nous décidons donc de reproduire la visualisation en discriminant par espèce. Nous distinguons sur le graphique de droite de la figure 1.1 que nous pouvons former trois groupes d'individus : Un très clairement identifié en noir (correspondant à l'espèce *setosa*) et deux groupes se chevauchant légèrement (le cluster

rouge pour *versicolor* et le vert pour *virginica*).

Données Crabs

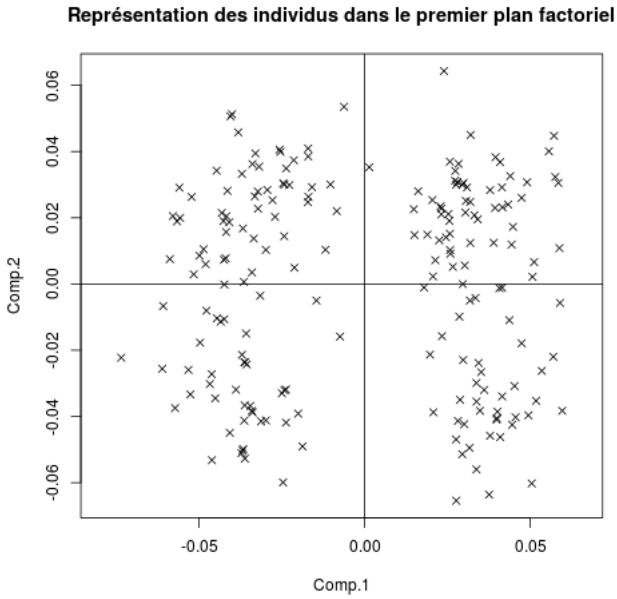

FIGURE 1.2 – Visualisation des données Crabs sans discrimination dans le premier plan factoriel

Nous avons effectué une ACP sur les 250 individus crabes sans se soucier du sexe ou de l'espèce. Comme nous le constatons figure 1.2, on pourrait deviner 2 groupes d'individus différents : Un groupe d'individus dont la valeur sur la première composante est négative et un autre où cette valeur est positive ; ou bien nous pouvons imaginer trois groupes : un « grand » avec la valeur sur la première composante négative, un deuxième plus petit avec Comp.1 > 0 et Comp.2 > 0 et un troisième avec Comp.1 > 0 et Comp.2 < 0. Sans discrimination, les hypothèses sont nombreuses. Nous décidons donc de discriminer une première fois par le sexe et une seconde fois par l'espèce, comme nous pouvons le constater figure 1.3

La discrimination par espèce nous confirme bien la première hypothèse émise, à savoir qu'il y avait 2 clusters différents, l'un avec Comp.1 négative, l'autre positive, tandis que la discrimination par sexe nous confirme partiellement l'autre hypothèse, à savoir que les individus ayant une Comp.1 > 0 pouvait se décomposer en 2 groupes. Ainsi, lorsque nous discriminons par sexe et espèce en même temps, nous apercevons 4 groupes, correspondant bien aux différentes possibilités {OM, OF, BM, BF}

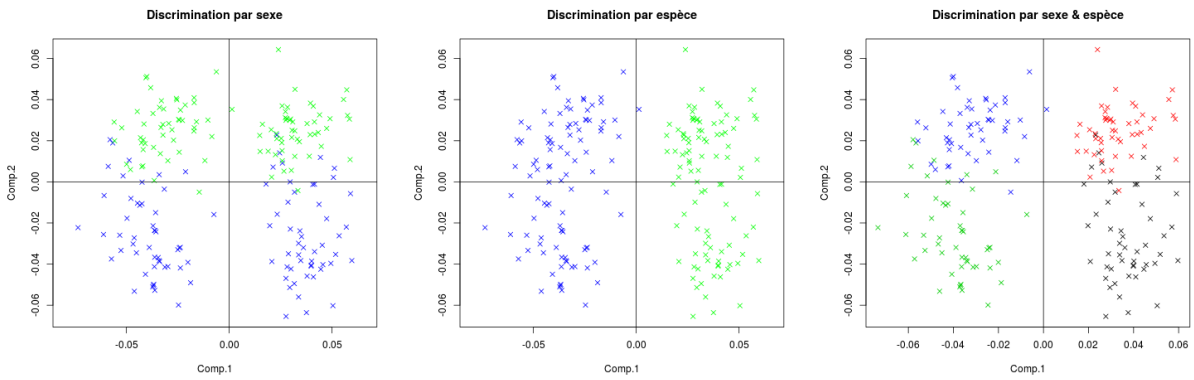

FIGURE 1.3 – Visualisation des données Crabs discriminées dans le premier plan factoriel

Mutations

Sur ce jeu de données, il est question de réaliser une AFTD. Cette méthode est similaire à l'analyse ACP sauf qu'elle ne prend pas en entrée des données centrées mais une matrice de dissimilarités. Quand il y a deux variables (ie k=2), on obtient la représentation euclidienne suivante :

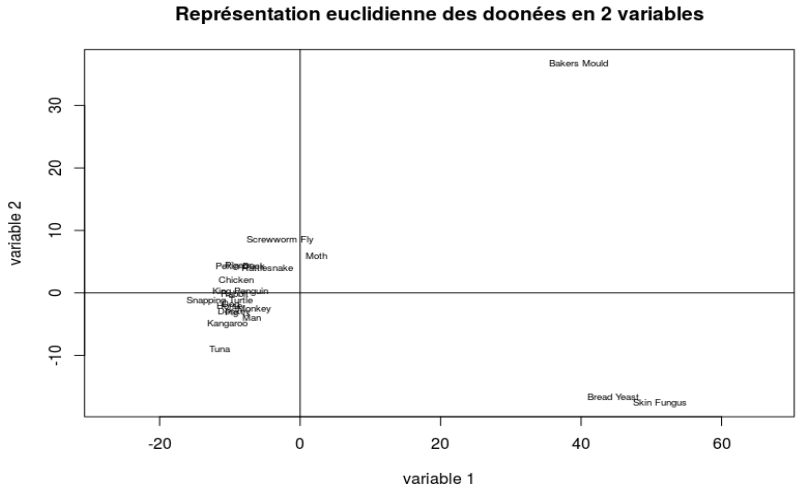

FIGURE 1.4 – Représentation de résultat de l'AFTD avec k=2

Pour calculer la qualité de la représentation, il faut calculer le pourcentage de la valeur propre de la matrice de distance. Plus cette valeur est grande, plus la qualité de la représentation est bonne. Nous avons calculé avec un nombre de variable variant entre 2 et 5.

Par la suite, on se décide de représenter ces points sur le diagramme de Shepard en choisissant le nombre de variables de représentation allant de de 2 à 5 (le nombre d'axe). En observant la figure 1.5, on peut constater que lorsque l'on augmente le nombre de variables k, les points semblent s'aligner sur la bissectrice y = x, ce qui traduit visuellement que la

nb de variables	2	3	4	5
Pourcentage	70.6%	81.9%	88.8%	93.7%

TABLE 1.1 – Qualité de la représentation

qualité de la représentation s'améliore. En effet, sur le diagramme de Shepard, l'axe des abscisses représente la dissimilarité initiale δ et les ordonnées la distance calculée d. Si les points sont parfaitement alignés, cela veut dire que l'on retrouve $d_{ij} = \delta_{ij}$. En revanche, si un point se situe au dessus de la bissectrice, cela traduit qu'on a surestimé la distance de départ. De la même manière, un point en dessous de la bissectrice traduit une sous-estimation de la distance initiale. Nous pouvons donc conclure que la représentation est plus fidèle lorsque k=5.

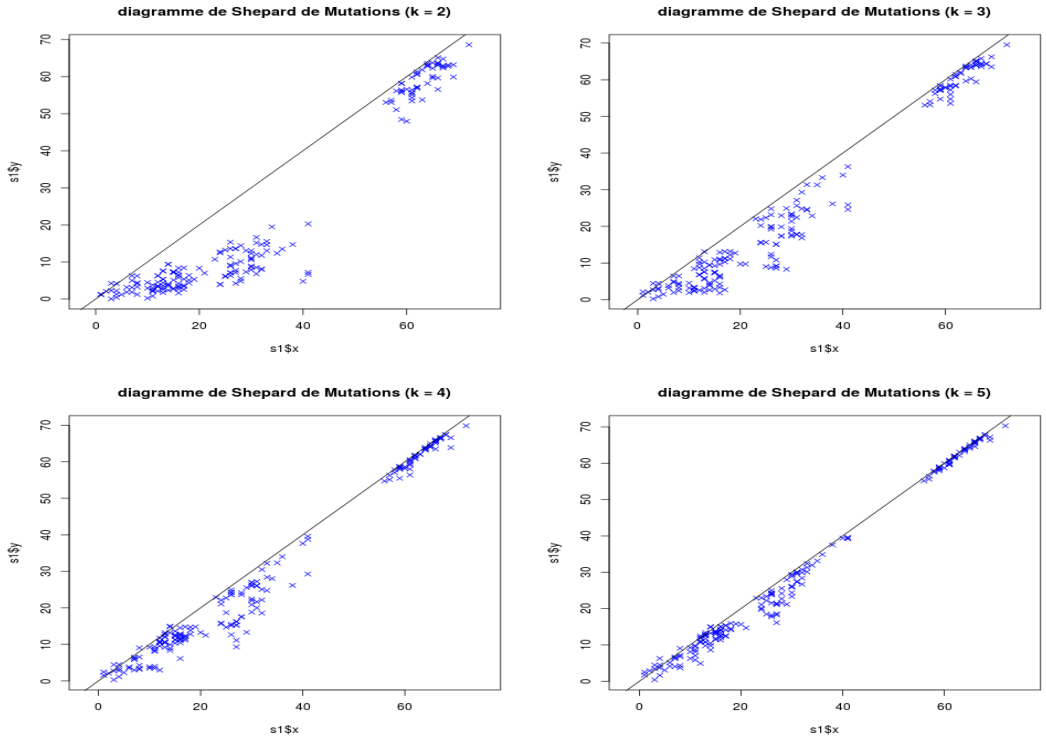

FIGURE 1.5 – Diagramme de Shepard quand le nombre de variables varie entre 2 et 5

2. Classification hiérarchique

Dans cette partie, nous réaliserons plusieurs classifications hiérarchiques ascendantes (CAH) avec différents critères d'agrégation possibles sur les données Mutations et Iris, le but étant d'observer les différences selon le critères d'agrégation choisi.

Mutations

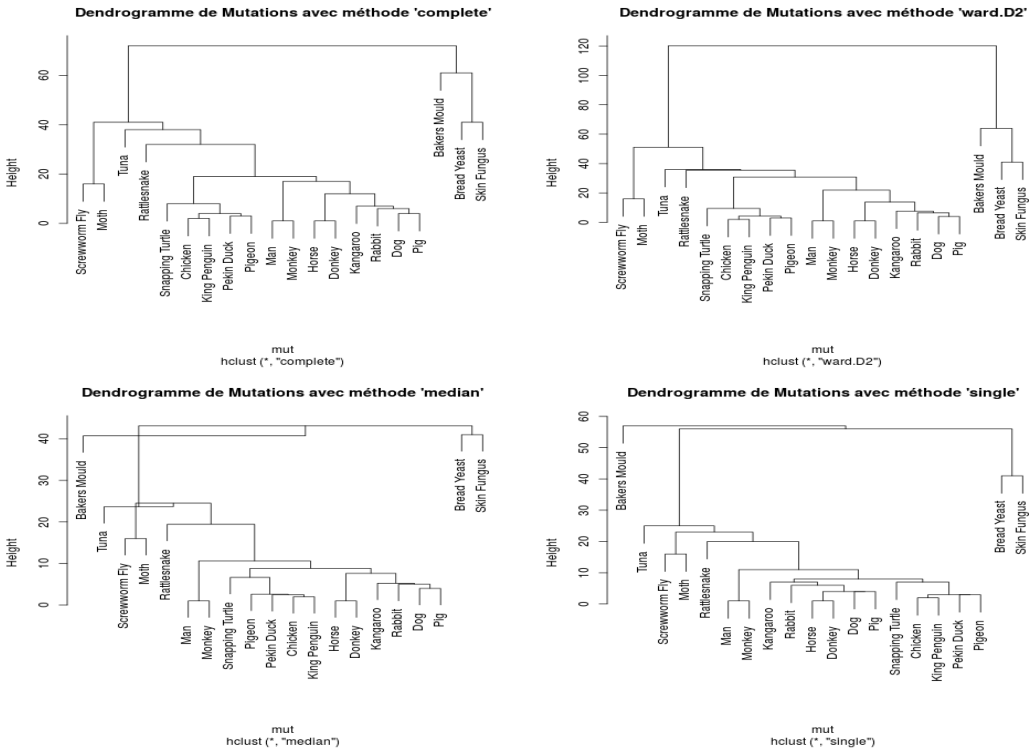

FIGURE 2.1 – Dendrogrammes des données Mutations

Nous avons décidé de représenter 4 dendrogrammes selon 4 critères différents : *complete*, *ward*, *median* et *single*. Il est à noter qu'il en existe d'autres : à la hauteur près, les critères *average* et *mcquitty* produisent le même résultat que *complete*. On observe un phénomène d'inversion avec le critère *median*.

Si l'on fait le parallèle entre la figure 1.4 et les dendrogrammes, on retrouve bien la grande majorité des espèces relativement proches les unes des autres et les 3 espèces Skin Fugus, Bread Yeast et Bakers Mould se trouvant à l'écart du reste. Alors que sur la représentation euclidienne cette forte dissimilarité se traduit par une grande distance, elle se traduit par un indice très élevé pour relier les deux groupes sur les dendrogrammes. Cet indice (appelé « height » sur la figure 2.1) n'est autre que la distance D qui séparaient les deux classes fusionnées pour former cette nouvelle classe.

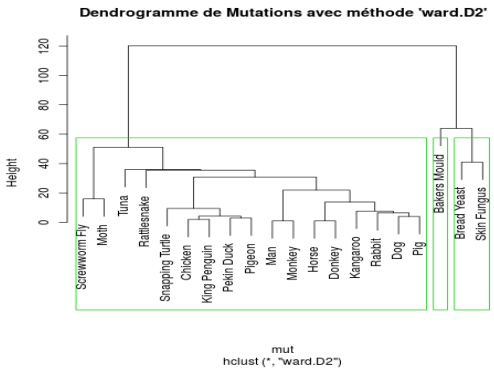

FIGURE 2.2 – En vert, les 3 « clusters » que l'on peut identifier sur la représentation euclidienne des données

Iris

Nous nous proposons de réaliser la classification ascendante hiérarchique (CAH) des données *Iris* et de l'analyser en nous appuyant sur nos connaissances de ce jeu de données, puis nous effectuons la classification descendante hiérarchique (CDH) pour comparer les résultats avec ceux obtenus avec la CAH.

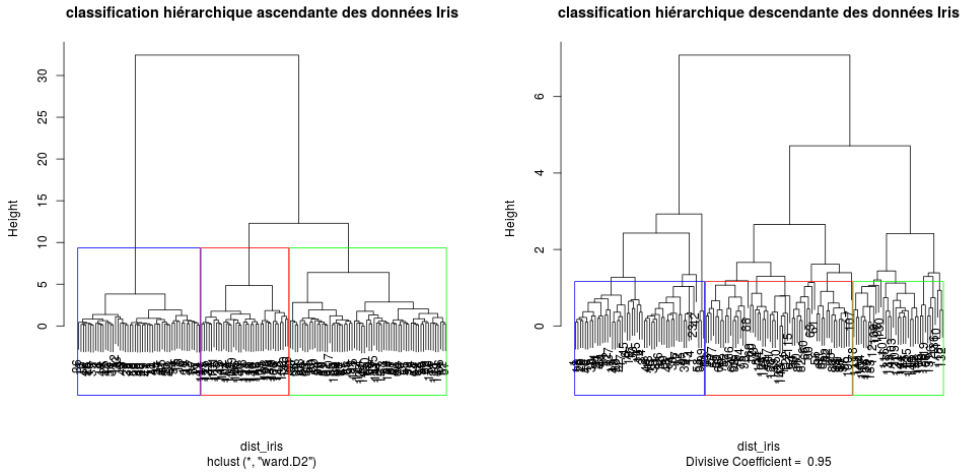

FIGURE 2.3 – Classification ascendante (gauche) et descendante (droite) des données Iris

La CAH semble cohérente avec la représentation réalisée figure 1.1. En effet, nous retrouvons trois classes différentes (de hauteurs à peu près toutes égales à 5) qui sembleraient correspondre aux 3 espèces que nous avions identifiées. De plus, on constate que deux de ces classes sont suffisamment proches pour être fusionnées (donnant une distance d'environ 13) alors que la troisième semble beaucoup plus éloignée des deux autres puisque sa fusion à pour conséquence de doubler la taille du dendrogramme. Cette observation rejoint la représentation faite figure 1.1.

En comparaison, la CDH va chaque fois distinguer un individu parmi tous les autres individus de la classe, on peut constater qu'à chaque niveau du dendrogramme, le découpage se fait par un groupe de petite taille et un groupe de taille plus grande. Mais en vue globale, on peut voir que ces deux méthodes permettent de distinguer les 3 espèces.

3. Méthode des centres mobiles

Le but de cette partie est d'appliquer l'algorithme des k-means aux 3 jeux de données utilisés jusqu'alors et de constater les résultats et les performances de ce moyen de classification.

Iris

Dans un premier temps, nous appliquons l'algorithme des k-means sur les données Iris, où l'on fait varier le nombre de classe de classification entre 2 et 4. On obtient le résultat comme suit :

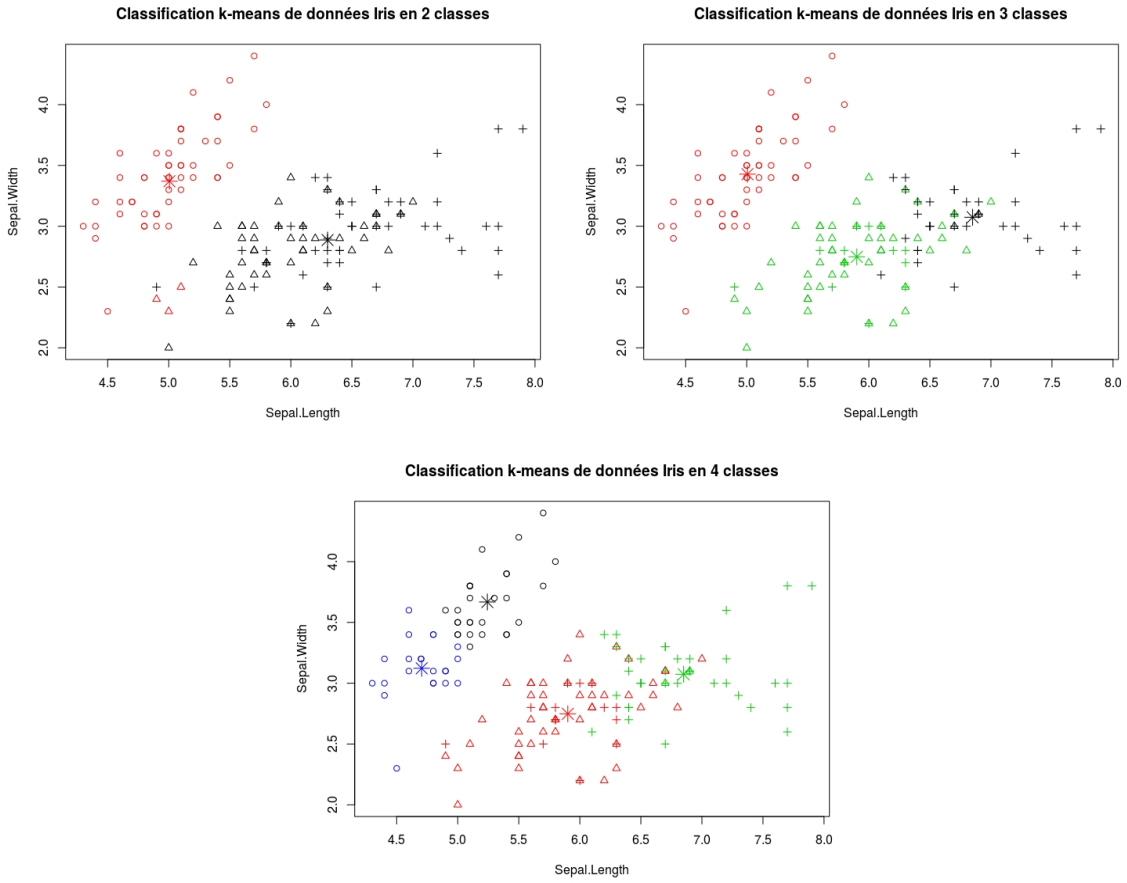

FIGURE 3.1 – Classification k-means de 2,3 et 4 classes de Iris

On peut constater que si on classe ces fleurs en 2 classes, les espèces *versicolor* et *virginica* sont regroupées dans une même classe, et l'espèce *setosa* est classée dans une autre classe. Le résultat est cohérent avec l'ACP, la CAH et la CDH. Cela veut dire qu'il y a moins de différences entre l'espèce *versicolor* et l'espèce *virginica* qu'avec *setosa*. Si on fait en 3 classes, le résultat est exactement comme la réalité et les trois espèces sont bien classifiées. Si on fait en 4 classes, on peut voir que l'espèce *setosa* est découpée en deux sous-espèces, mais dans la réalité, nous savons qu'il n'y a que 3 espèces. Le nombre de classes à fixer et donc très important.

Ensuite, on cherche à déterminer si le résultat de l'algorithme est toujours le même. Pour ce faire, on exécute 10 fois l'algorithme k-means avec K=3 classes. On calcule ensuite la somme des inerties intra-classes.

k-means n°	1	2	3	4	5	6	7	8	9	10
Inertie	78.85	78.85	78.85	142.75	142.75	78.85	78.85	78.85	78.85	78.85

TABLE 3.1 – Inertie intra-classe à la fin de chaque exécution de l'algorithme

On constate valeurs différentes : 78.85 et 142.75, En théorie, l'inertie totale intra-classe devrait être de plus en plus petite a chaque itération, jusqu'à une valeur constante. Puisqu'on constate deux valeurs, cela signifie que l'inertie intra-classe reste « bloquer » dans un minimum local (ici de valeur 142.75). Ce phénomène est lié au choix aléatoire et arbitraire des centroïdes au début de l'algorithme, et est une des faiblesses des k-means. Pour contourner ce problème, il suffit comme on vient de le faire d'exécuter plusieurs fois l'algorithme et s'il y a plusieurs valeurs, choisir la plus cohérente et/ou celle qui revient le plus souvent. Si l'on trace le graphe quand l'inertie est égale a 142.75, on observe que les 3 classes ne sont pas correctement classées :

Puis on fait une itération de nombre de classe de 2 a 10, et pour chaque classe on fait 100 fois la classification k-means. on obtient le resultat suivant et on fait aussi un barplot sur ces inerties minimales, comme suivant :

k-means	2	3	4	5	6	7	8	9	10
Inertie	152.35	78.85	57.27	69.24	39.05	34.30	30.18	35.68	31.14

TABLE 3.2 – Inertie totale

Selon le graphe 3.3, on peut affirmer que plus le nombre de classes augmente, plus l'inertie intra-classe diminue jusqu'à se stabiliser à une certaine valeur. Grâce à la méthode du coude, on peut déterminer le nombre de classes à choisir : il suffit de prendre K à

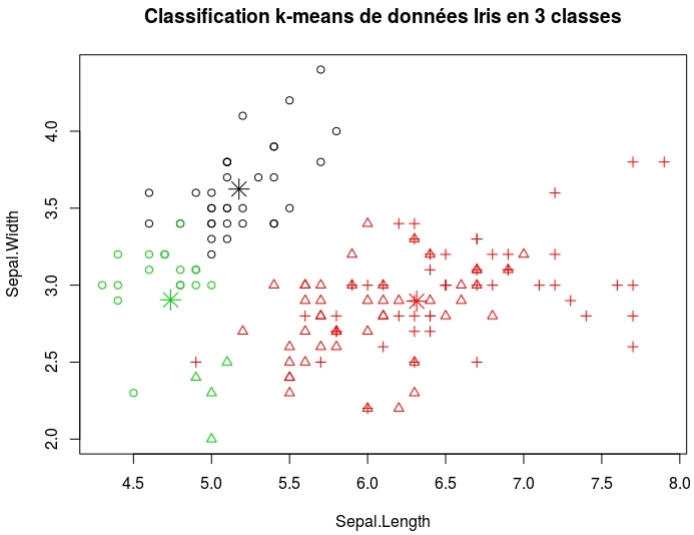

FIGURE 3.2 – Classification k-means de 3 classes de Iris

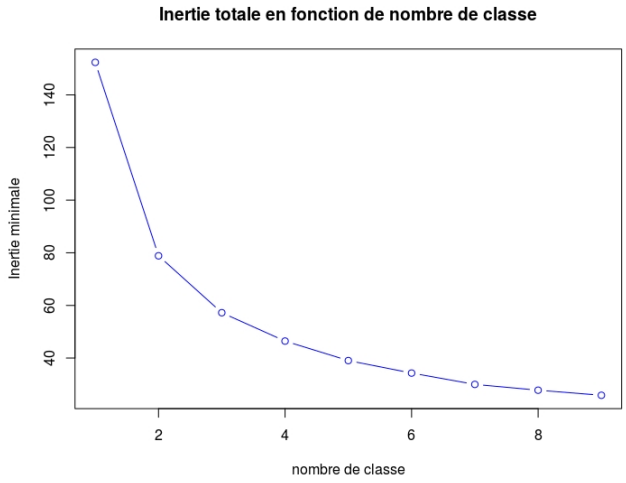

FIGURE 3.3 – Inertie totale de k-means de Iris

l'endroit de la rupture de la décroissance du graphe, lorsque l'on passe du très forte décroissance à une décroissance modérée : ici, on a le choix entre K=2 et K=3. A l'aide de nos connaissances, nous choisissons la deuxième option. La méthode du coude n'est pas un théorème et ne donne pas une solution toute faite, elle nous aide juste à choisir une à deux valeurs possibles de K.

Crabs

Premièrement, on a effectué 10 fois la classification k-means en 2 classes sur les données Crabs. On a enregistré l'inertie totale de intra-classe et on a constaté deux résultats différents.

k-means n°	1	2	3	4	5	6	7	8	9	10
Inertie	0.259	0.259	0.356	0.259	0.259	0.259	0.356	0.259	0.259	0.259

TABLE 3.3 – Inertie totale intra-classe

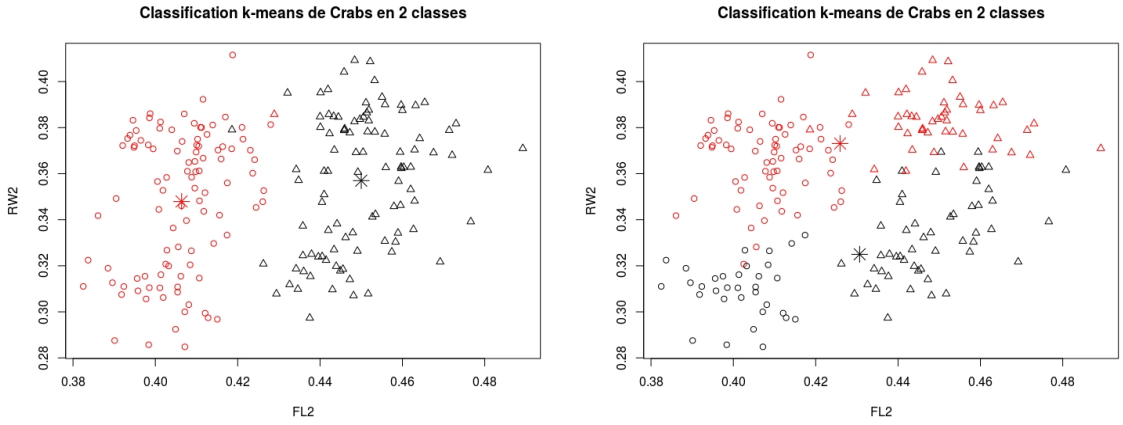

FIGURE 3.4 – Classification finale de Crabs selon K-means en fonction de l'espèce(gauche) et du sexe(droite)

Les inerties 0.259 et 0.356 correspondent respectivement à une classification selon l'espèce et sexe. Cela veut dire qu'il y a deux façons différentes pour la classification de données Crabs, soit par l'espèce, soit par sexe, confirmant nos observations de la partie précédente On peut aussi effectuer une classification avec K=4 classes sur des données Crabs en utilisant la méthode de k-means. On a obtenu le résultat comme suivant :

On peut voir que quand on fait k-means de 4 classes, le résultat est comme le vrai résultat, cf 1.3, Crabs est bien classe en fonction de l'espèce et sexe en 4 groupes. Comme il y a deux espèces de crabs et seulement 2 différentes sexes dans le monde, on peut pré-calculer le nombre total de groupes 2*2=4.

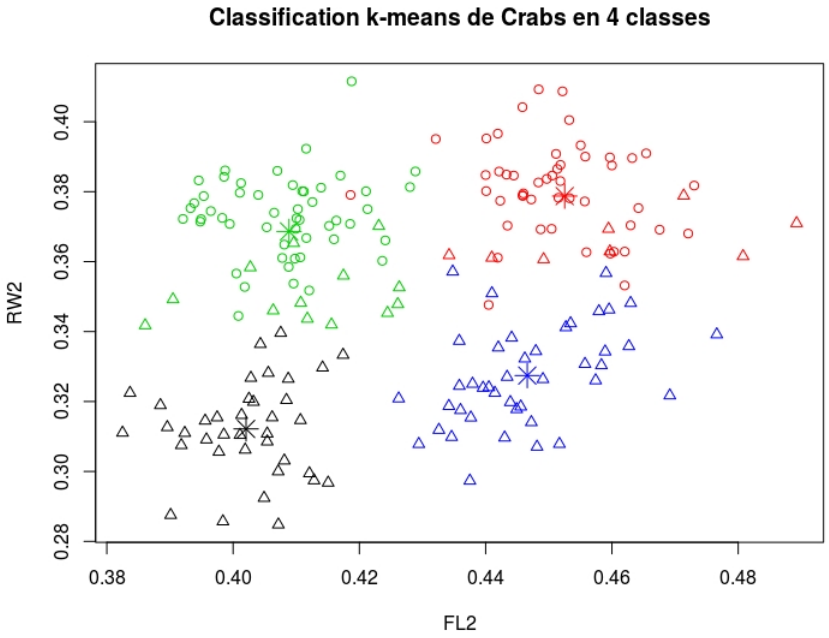

FIGURE 3.5 – Classification k-means de Crabs en 4 classes

Mutations

On a fait d'abord une représentation de données de Mutation dans un espèce de dimension 5, comme on fait dans le premier exercice. Ensuite on a fait plusieurs fois de classification de k-means, on a obtenu deux résultats différents. On a tracé le résultat dans le premier plan factoriel, comme suivant : Concernant la stabilité des résultats, on constate plusieurs possibilités de regroupements (au moins 6). Nous en avons exposé deux figure 3.6. Grâce à nos connaissances sur les données, on devine que la première classification est la bonne. Mathématiquement, cette affirmation se confirme car l'inertie intra-classe de la seconde classification est très supérieure à celle de la première classification. En globalité, la « bonne » classification est celle qui possède l'inertie intra-classe la plus faible.

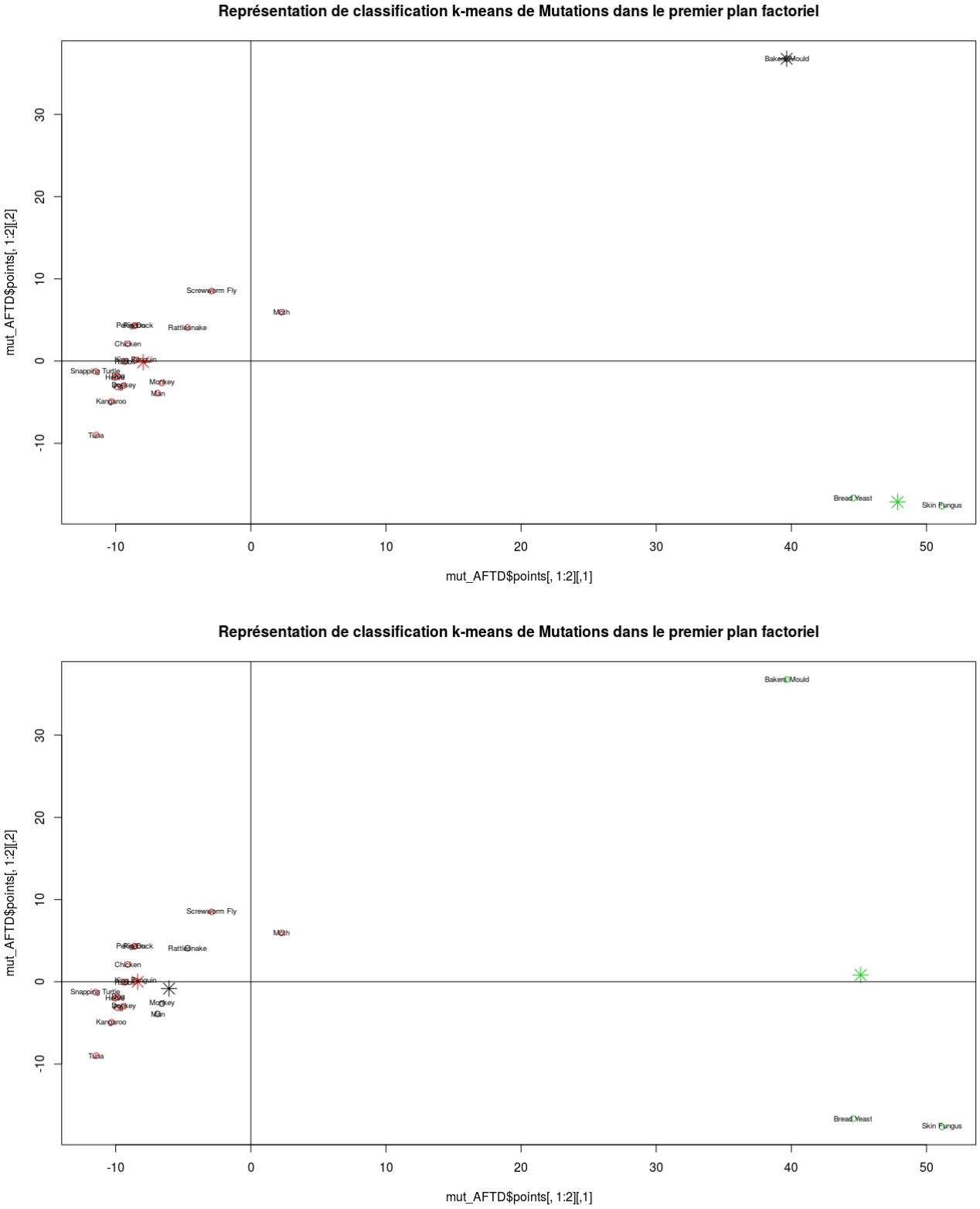

FIGURE 3.6 – Classification k-means de Mutations

Conclusion

Dans ce TP, nous avons pu nous familiariser avec différentes méthodes de classification automatique. Dans un premier temps, nous avons visualiser les données afin d'en prendre connaissance pour pouvoir appréhender la classification sereinement et pouvoir prendre du recul. Puis nous avons réaliser des classifications hiérarchiques et enfin nous avons mis en place l'algorithme de classification non-supervisé « k-means » afin d'évaluer son comportement, ses forces et ses faiblesses (notamment le problème de minimum local).