Algorithmes avancés
Clustering et classification supervisée

Claudio Sousa, David Gonzalez

10/06/2018

1 Introduction

Le but de ce travail pratique pour le cours d'algorithmes avancés est de mettre en pratique les différentes techniques en science des données, plus particulièrement, le clustering ainsi que la classification des données.

Ce travail est divisé en 3 parties :
- clustering de données imposées ;
- classification supervisée avec algorithmes d'apprentissage et données imposés ;
- classification supervisée avec algorithmes d'apprentissage et données libres.

1.1 Normalisation des données

Les données utilisées durant ce travail pratique peuvent être de différentes natures et requièrent d'être normalisées.
En effet, certaines méthodes qui sont utilisées pour ce travail souffrent beaucoup des facteurs d'échelle, ce qui réduit les performances générales.

Donc, les données sont normalisées avec une procédure de *MinMaxScaler*[1].

1. `http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`

2 Clustering

La méthode de clustering utilisée est le *K-Means* et est testée sur les données du titanic.

Les buts de cet exercice sont :
- trouver la valeur de K adéquate ;
- afficher les clusters sur un graphique 3D ;
- commenter les résultats.

2.1 Normalization des données

Les différentes variables (*features*) avaient initialement des valeurs qui variaient beaucoup et que nous avons normalisé dans la plage $[0, 1]$. Nous avons fait le choix de mettre un poids pour la classe (*Survived*) supérieur aux variables de manière à ce que une différence de classe ait plus d'importance qu'une différence de valeur pour une variable. Après nos tests, nous avons fixé la valeur de la classe dans la plage $[0, 3]$.
Nous avons aussi transformé les données afin de remplacer tous les points qui se trouvent au même endroit dans l'espace par un seul point qui représente la moyenne des personnes qui ont survécu pour cette combinaison de paramètres.

2.2 Trouver K

Dans ce chapitre le but est de trouver la valeur de K qui minimise la moyenne des distances moyennes entre les différents clusters.

Nous avons fait varier K et avons utilisé comme mesure de distance la propriété *inertia_* de l'object KMeans. Cette measure est calculée en faisant la somme des carrés des distances ($\sum_{i=0}^{n}(x_i - \mu_i)^2$) et est proportionnelle à la "moyenne des distances moyennes".

Voici le graphique de la distance en fonction de K :

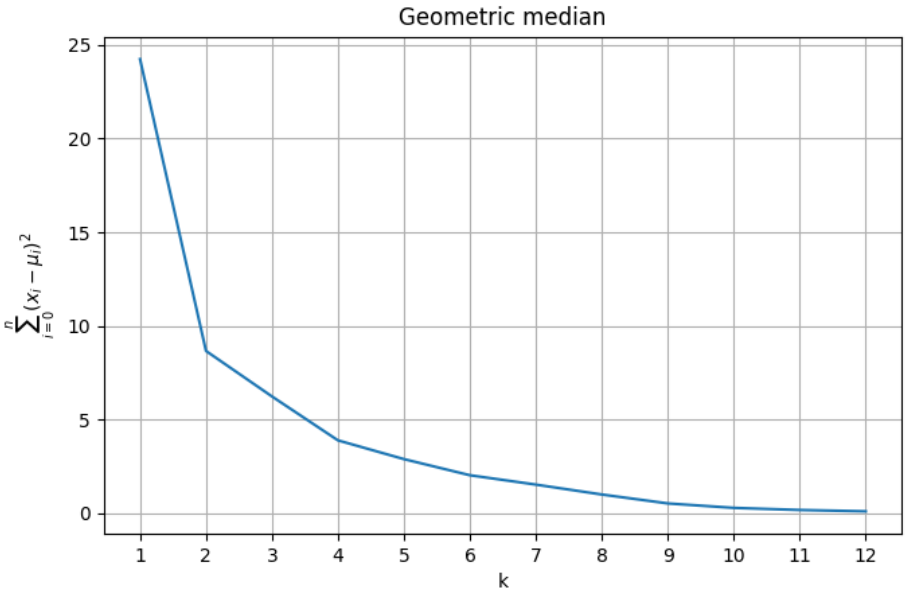

FIGURE 1 – Distance en fonction de K

Ce graphique nous montre que nous trouvons des variations importantes pour $K \in \{2, 4\}$. Après quelques tests empiriques nous avons fixé $K = 4$.

2.3 Affichage des données *K*

Les données normalisées et transformées sont affichées :

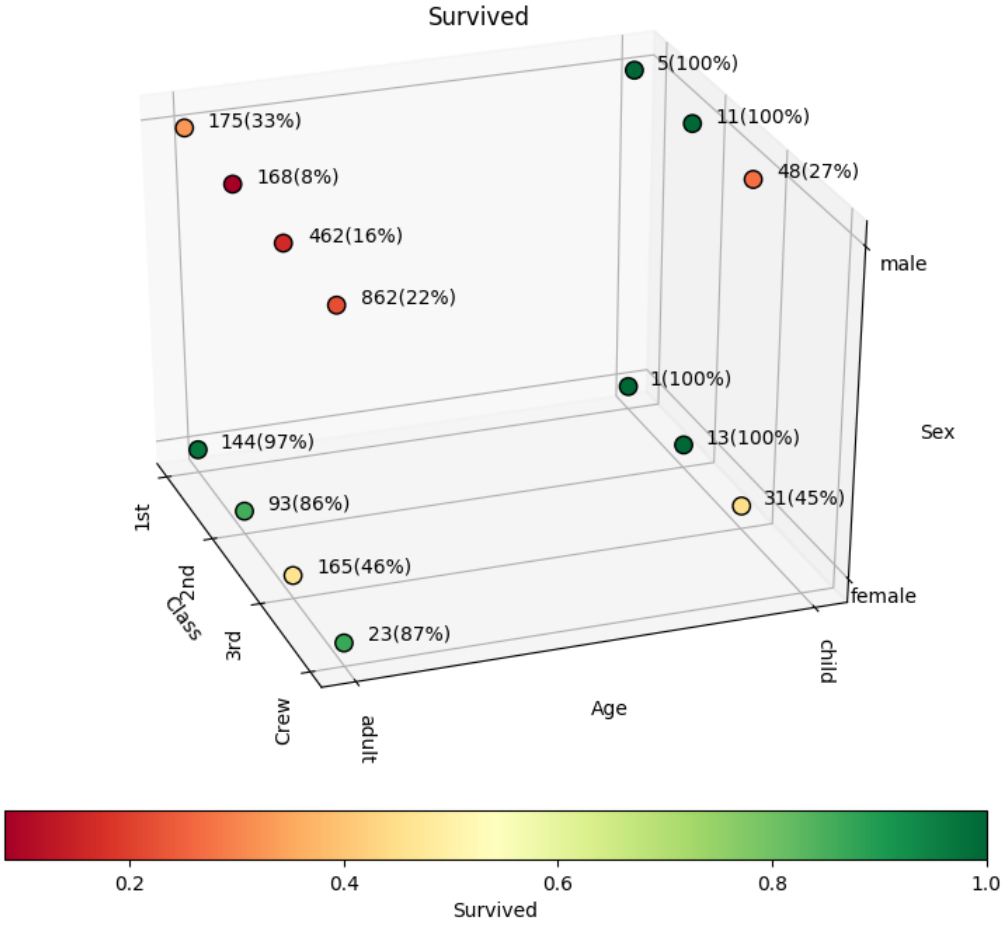

FIGURE 2 – Survie selon paramètres

Chaque point est accompagné de l'information du nombre de personnes ayant ces paramètres et le pourcentage de personnes ayant survécu. L'échelle de couleur nous renseigne sur la quantité de personnes ayant survécu, allant du rouge pour 0% jusqu'au vert pour 100%.

2.4 Clustering

Voici le résultat du clustering :

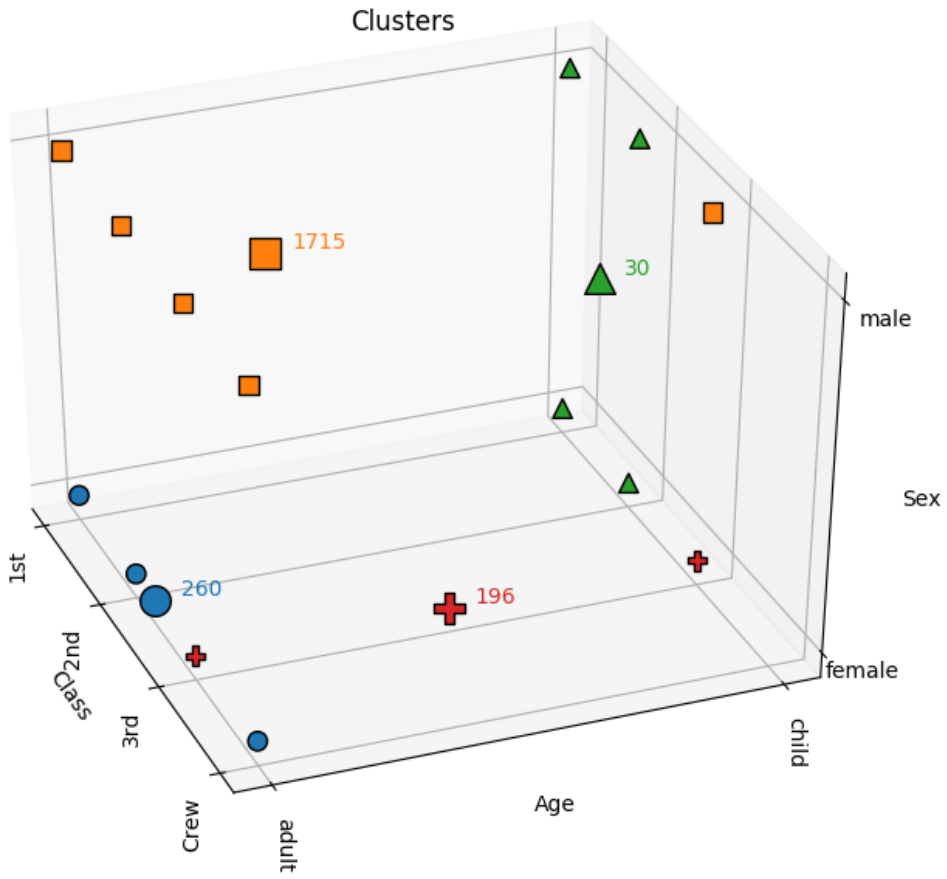

FIGURE 3 – Clustering

Chaque point est représenté par un symbole et une couleur selon le cluster auquel le point appartient. Un point supplémentaire du même symbole et couleur, mais de taille plus large symbolise le centre du cluster et est inscrit le nombre de personnes de ce cluster.

2.5 Commentaires

Le cluster *Triangle Vert* forme un groupe de personnes qui sont caractérisées par :
- leur age : ce sont des enfants ;
- leur classe : 1ème ou 2ème classe ;
- leur taux de survie : 100%.

Le cluster *Croix Rouge* forme un groupe de personnes qui sont caractérisées par :
- leur sexe : ce sont des femmes ;
- leur classe : 3ème classe ;
- leur taux de survie : autour des 50%.

Le cluster *Cercle Bleu* forme un groupe de personnes qui sont caractérisées par :
- leur age : ce sont des adultes ;
- leur sexe : ce sont des femmes ;
- leur classe : toutes sauf la 3ème ;

— leur taux de survie : au dessus de 86%.

Le cluster *Carré Orange* forme un groupe de personnes qui sont caractérisées par :
— leur sexe : ce sont des hommes ;
— leur age/classe sont :
  — des hommes toutes classes confondues ;
  — des garçons en 3ème classe.

Nous pouvons faire quelques remarques :
— les hommes adultes ont très peu survécu ;
— les hommes en 3ème classe ont eu une mortalité bien plus importante ;
— les enfants ont été épargnés, sauf ceux de 3ème classe ;
— les femmes ont une une mortalité relativement basse, sauf celles de 3ème classe.

3 Classification supervisée imposée

3.1 Introduction

Le but de cette deuxième partie est de trouver par *validation croisée* la meilleure méthode de classification sur les données suivantes :

— cancer du sein [2] ;
— vins [3].

Les méthodes de classification supervisée sont imposées et les voici :

— méthode des k-plus proches voisins ;
— arbres de décision ;
— perceptron multi-couche.

Pour la méthode des k-plus proches voisins, le paramètre K est fait varier de 1 à 10 compris.
Pour la méthode des arbres de décision, le nombre minimum d'échantillon dans une feuille est fait varier de 1 à 10 compris.
Concernant la méthode du Perceptron multi-couche, toutes les combinaisons de solveurs et de fonctions d'activation possibles sont testées, avec 1 couches cachées possédant 2 neurones.

La validation croisée divise les données en 5 parties et est répétée 10 fois. Cela implique qu'il y aura au total 50 scores de performance pour une méthode donnée. Les résultats de cette validation croisée sont ensuite tracés sur un graphe, permettant de comparer les performances des différentes méthodes.

Chaque graphe est tracée selon la légende suivante :
En abscisse, nous avons la variation de ou des paramètres intéressants pour la méthode concernée.
En ordonnée, nous avons plusieurs choses. Chaque point représente un score. Ces points sont regroupés par couleur, qui représente les variations de la méthode. Les barres horizontales noires représentent la performance moyenne pour la variation de la méthode concernée. La barre horizontale rouge représente la meilleure performance moyenne parmi toutes les variations d'une même méthode.

Un tableau de données accompagne chaque graphe et donne les moyennes de chaque variations dans le même ordre que le graphe.
En abscisse, nous avons la variation de ou des paramètres intéressants pour la méthode concernée.
En ordonnée, nous avons les méthodes.

2. `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`
3. `https://archive.ics.uci.edu/ml/datasets/Wine`

3.2 Données du cancer du sein

Voici ci-dessous le résultat de la validation croisée pour les méthodes citées sur les données du cancer du sein :

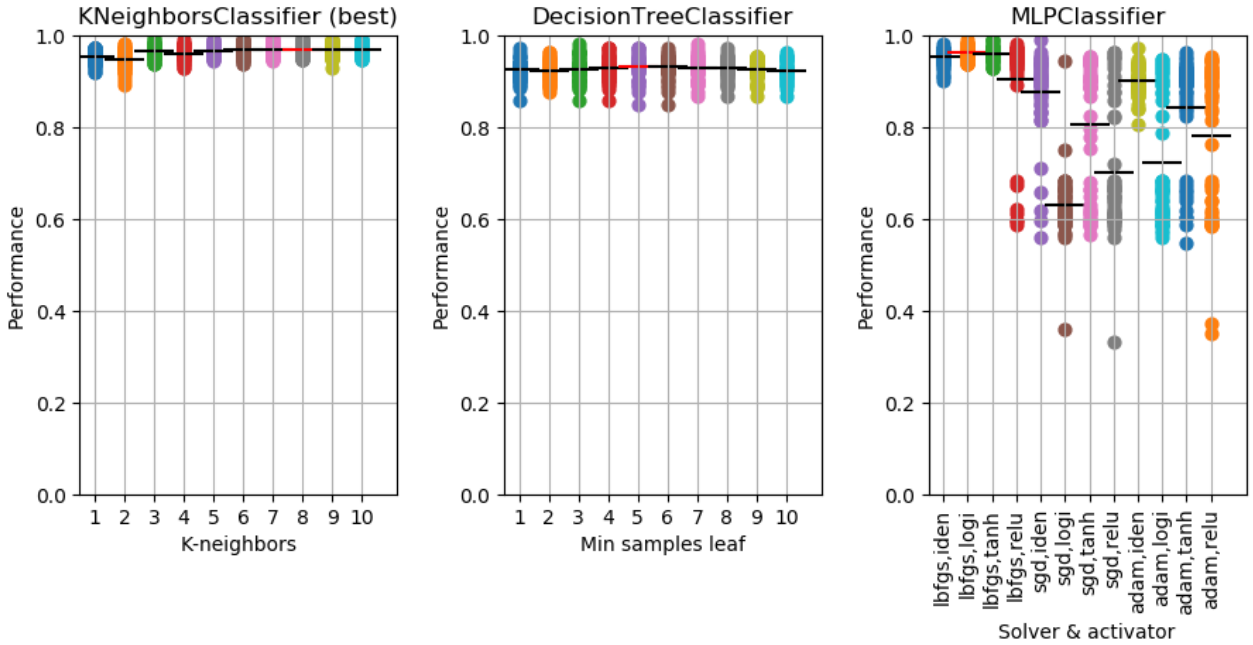

FIGURE 4 – Résultat de la validation croisée sur les données du cancer du sein

M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
0.955	0.950	0.968	0.962	0.968	0.969	0.970	0.970	0.969	0.970		
0.924	0.924	0.930	0.930	0.933	0.933	0.931	0.929	0.926	0.927		
0.953	0.964	0.954	0.873	0.881	0.630	0.831	0.740	0.887	0.716	0.834	0.747

TABLE 1 – Résultat de la validation croisée sur les données du cancer du sein

De manière générale, les 3 méthodes, avec les paramètres adéquats, donnent des résultats satisfaisants. En effet, la plupart des variations donne un score moyen qui dépasse les 90%.

La méthode des k-plus proches voisins est ici pour peu la meilleure méthode. La variation du paramètre K n'affecte que de quelques pourcents le score moyen. Par ailleurs, on peut observer qu'après $K = 6$, le score moyen est stable et varie très peu.

Concernant la méthode des arbres de décision, on remarque que la courbe construite à partir des scores moyens produit un pic lorsque le nombre d'éléments minimums dans un nœud feuille est de 5. On peut donc en déduire que les données peuvent facilement être regroupé par 5. Ceci explique également le K obtenu pour la méthode des k-plus proches voisins.

Quant au Perceptron multi-couche, les différents solveurs donnent des résultats bien différents, alors que changer la fonction d'activation ne produit que peu d'effet, excepté certaines combinaisons, comme le solveur *sdg (stochastic gradient descent)* et la fonction d'activation *logistic (fonction sigmoid)* qui produit un effet désastreux. Cependant, cette même fonction d'activation marche très bien avec le solveur *lbfgs* (*Limited-memory Broyden–Fletcher–Goldfarb–Shanno*), produisant un résultat similaire à la méthode des k-plus proches voisins. Le solveur présentant les meilleurs résultats est le *lbfgs*.

3.3 Données des vins

Voici ci-dessous le résultat de la validation croisée pour les méthodes citées sur les données des vins :

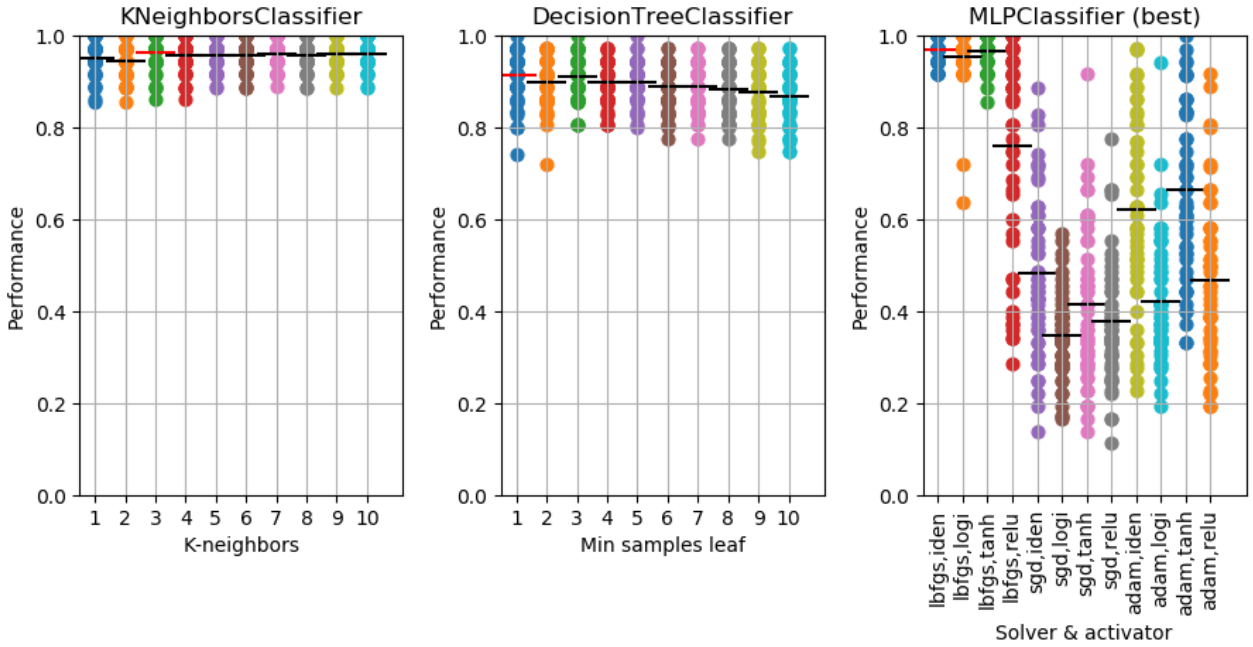

FIGURE 5 – Résultat de la validation croisée sur les données des vins

M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
0.952	0.945	0.963	0.958	0.958	0.958	0.961	0.957	0.961	0.962		
0.913	0.901	0.912	0.899	0.897	0.890	0.890	0.883	0.880	0.870		
0.960	0.973	0.959	0.775	0.428	0.331	0.425	0.422	0.652	0.464	0.586	0.498

TABLE 2 – Résultat de la validation croisée sur les données des vins

Par rapport au données du cancer du sein, on remarque immédiatement que les scores sont de manière générale plus étalés. On peut donc en déduire que ces données se regroupent moins facilement. On peut par ailleurs bien l'observer avec la méthode des arbres de décision, où l'on voit que la meilleure variation est lorsqu'il n'y a qu'un seul élément dans un nœud feuille.

Pour ces données, la meilleure méthode est le Perceptron multi-couche. Nous retrouvons encore le solveur *lbfgs*, mais cette fois avec la fonction d'activation *identité* au lieu de la *sigmoid*. On voit également que les autres solveurs, quelque soit la fonction d'activation, sont mauvais avec ces données.

Pour peu, la méthode des k-plus proches voisins n'est pas la meilleure. Ceci dit, toutes les variations de cette méthode sont bonnes. Le paramètre K n'a que peu d'effet.

4 Classification supervisée libre

La partie 3 de ce travail suit le même principe que la partie 2 (voir section 3.1), mais les 3 méthodes ainsi que les données sont à choix.

Les données choisies sont celles caractérisant les feuilles de plantes [4]. Chaque feuille est accompagnée du nom de l'espèce auquel elle appartient, représentant ici la classe recherchée. Il y a 340 instances, 16 attributs et 30 classes.

Les 3 méthodes choisies sont les suivantes :
- support vector machines ;
- forêts aléatoires ;
- arbres de décision.

Pour la méthode des support vector machines, tous les noyaux sont testés.
Pour la méthode des forêts aléatoires ainsi que celle des arbres de décision, le nombre minimum d'échantillon dans une feuille est fait varier de 1 à 10 compris.

4. `https://archive.ics.uci.edu/ml/datasets/Leaf`

Voici ci-dessous le résultat de la validation croisée pour les méthodes citées sur les données choisies :

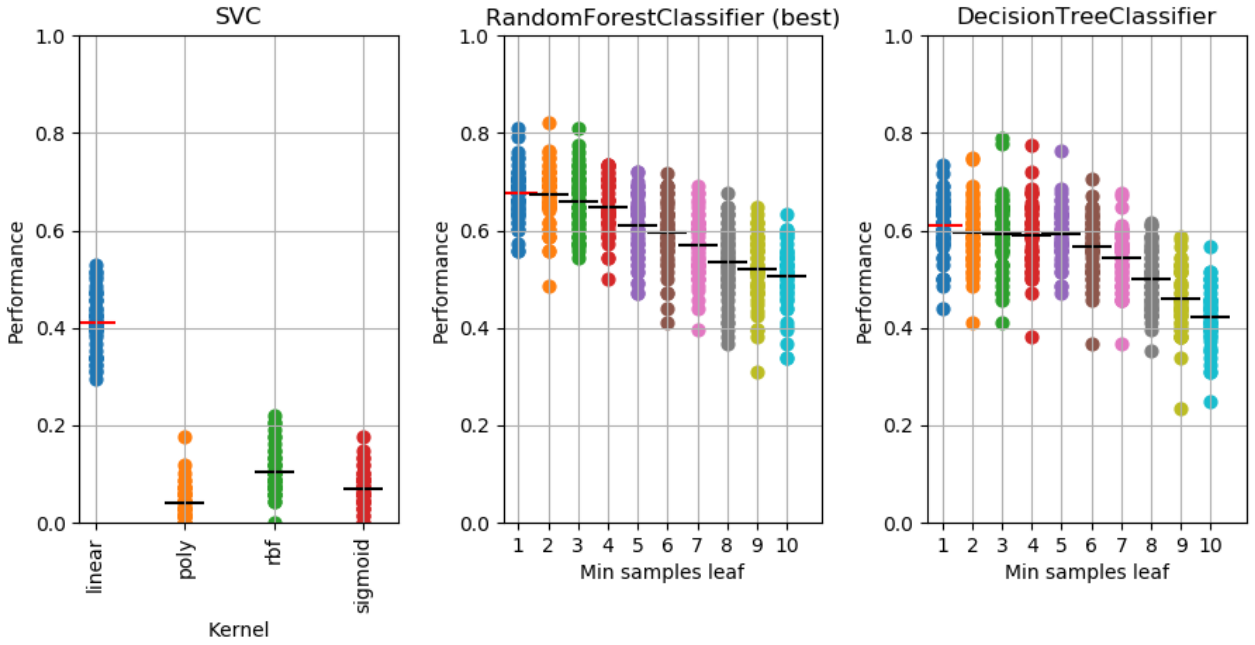

FIGURE 6 – Résultat de la validation croisée sur les données des feuilles de plantes

M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
0.410	0.042	0.104	0.070						
0.682	0.667	0.658	0.635	0.621	0.582	0.574	0.528	0.506	0.487
0.614	0.600	0.592	0.589	0.594	0.573	0.541	0.502	0.458	0.422

TABLE 3 – Résultat de la validation croisée sur les données des feuilles de plantes

La première méthode, le SVC, est simplement mauvaise, quelque soit le noyau choisi. Le noyau *linéaire* est le seul à donner un résultat décent. Ceci dit, les paramètres ont tous été laissés par défaut.

La meilleure méthode est celle des forêts aléatoires. On remarque que de forcer plus d'un élément dans une feuille ne fait qu'empirer le score. On peut en déduire que les données ne peuvent être regroupées facilement.

La méthode des arbres de décision a ici été utilisée pour comparer ses performances à la méthode des forêts aléatoires. On remarque le même effet concernant le nombre d'éléments minimums. De manière générale, cette méthode produit un score un peu inférieur à la méthode des forêts aléatoires.

5 Conclusion

Nous avons vu durant le clustering que l'analyse de données est largement laissé à l'intuition.
La manipulation de données est parfois nécessaire afin d'obtenir des résultats plus parlants. En effet, un attribut a été modifié pour lui donner plus de poids, créant ainsi des clusters qui correspondent mieux aux résultats désirés.

Durant les différentes classifications, nous avons vu que le choix de la méthode et de ces paramètres est crucial pour arriver à un bon apprentissage. Par ailleurs, ce choix est également dépendant des données sur lesquelles on travail. En effet, chaque méthode applique des techniques différentes qui sont plus ou moins efficaces suivant la structure et la répartition des données.

En conclusion, le data science est une branche qui requière beaucoup de tests empiriques, d'instinct et d'expérience pour arriver à des résultats intéressants.