Rapport Analyse de Données
M1 – CNS - MIAGE

Groupe : Nesrine Zid & Jean-Charles Fournier

Dataset : http://archive.ics.uci.edu/ml/datasets/Wine

Tables des matières :

Introduction :

Dans ce rapport on va analyser un dataset en utilisant les méthodes d'analyses de données vu en cours.

Tout d'abord par l'Analyse en Composantes Principales qui consiste à réduire la dimension de notre tableau de données pour faciliter l'observation des individus par rapport aux autres. Cela permettra de corréler les différentes variables entre elles.

Ensuite on utilisera ensuite le Clustering, qui consiste à regrouper les points dans différents groupes en fonction de leurs attributs. Nous utiliserons pour cela l'algorithme K-means, d'abord sur les données brutes, puis sur les données obtenues par l'ACP pour observer les différences entre les deux en testant les clusters trouvés par rapport aux labels présents dans les données.

Enfin l'Analyse des Corrélations Canoniques qui permettra d'étudier plus précisément la corrélation entre des groupes de variables et permet d'explorer les relations pouvant exister entre ces groupes de variables quantitatives observées sur le même ensemble d'individus, et qui sera faite en 5 étapes et on va également interpréter les résultats à l'aide du cercle de corrélation.

On va introduire en premier lieu le dataset qu'on va utiliser. Après quoi on va définir l'ACP et on va interpréter les résultats. Puis on va développer la partie clustering, et faire une comparaison entre les 2 résultats. Et pour finir, on va finir avec l'Analyse en Composantes Canoniques.

Dataset :

Notre dataset contient des données qui sont le résultat d'une analyse chimique des vins cultivés dans la même région en Italie, mais dérivés de trois cultivars différents. Nous chercherons dans ce projet à vérifier si les différences des cultivars sont vérifiables à l'aide des données fournies.

La page du dataset indique 178 individus représentés par une variable qualitative (a) et 13 variables quantitatives (b, c, ..., n)

Pour une bonne partie des opérations sur les données, on utilisera des composants de la bibliothèque Scikit-Learn, abrégée S-L.

S-L fournit une sélection d'outils efficaces pour l'apprentissage automatique et la modélisation statistique, notamment la classification, la régression et le clustering via une interface cohérente en Python.

Nous n'avons aucune colonne de type catégorie, donc on n'aura pas recours à OneHotEncoder de S-L et on a aucune donnée manquante, donc on n'a pas à les remplacer par la moyenne de la colonne.

On analysera les données au début du jupyter Clustering.

I Réduction de dimension (jupyter ACP)

L'analyse en composantes principales (ACP)

• Objectifs :

Extraire l'essentiel de l'information contenue dans le tableau de données et d'en fournir une représentation se prêtant plus aisément à l'interprétation.

• Principe de l'ACP :

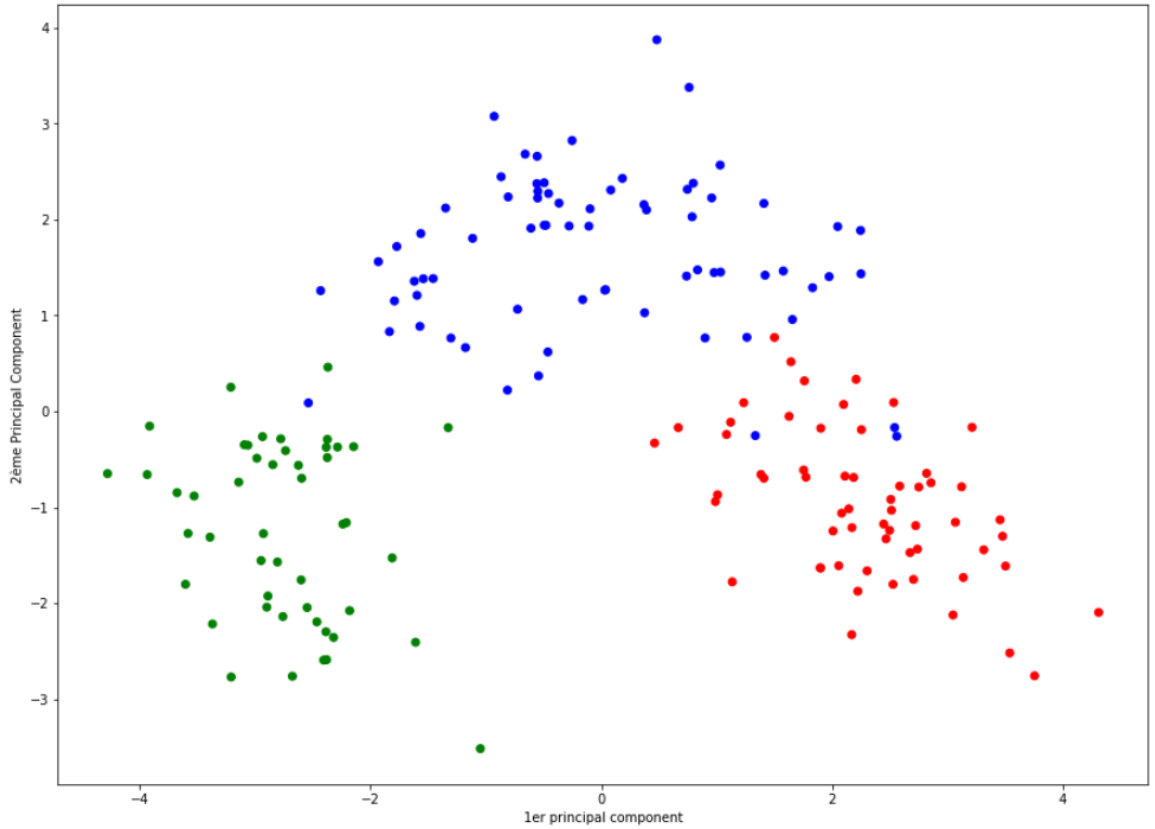
- À partir de n variables initiales continues, construire m (≤ n) autres variables, appelées composantes principales, combinaisons linéaires des variables initiales, telles que :

  Les CP sont ordonnés selon l'information (variance) qu'elles restituent, la 1ère étant celle qui restitue le plus d'information
  Les CP sont des vecteurs indépendants, c'est-à-dire des variables non corrélées entre elles

Dans un premier temps, on retire la colonne des labels qu'on cherchera à déduire, puis on utilise StandardScaler() de S-L pour centrer et réduire les données.

Pour effectuer l'analyse en composantes principales, on utilise le package PCA de S-L qu'on applique à nos données standardisées.

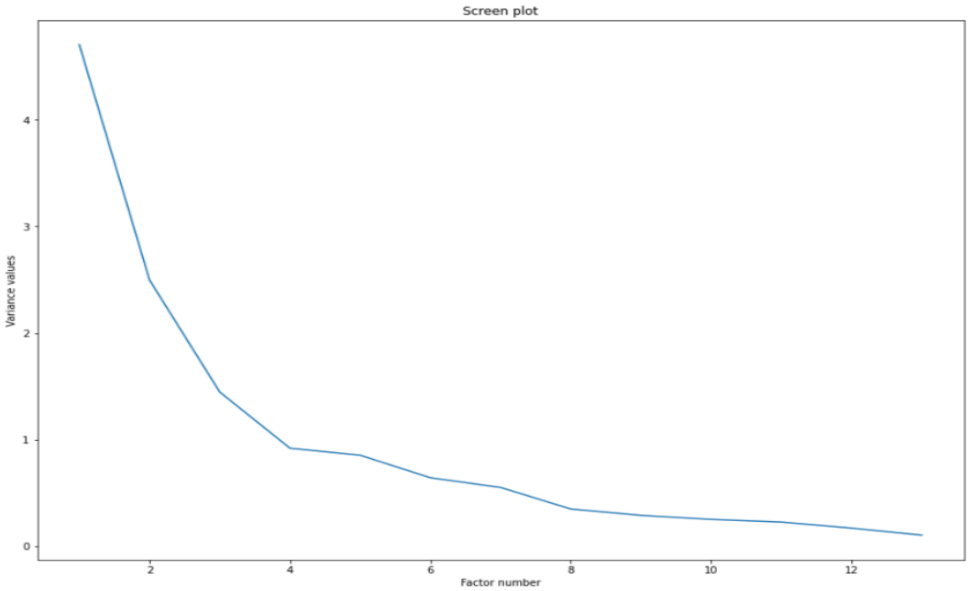
On projette ensuite les individus sur un plan à deux dimensions qui correspondent au premier et au second axe factoriel. On utilise les labels pour colorier et mettre en évidence leur catégorie par rapport à leur placement. Cela donne ce plan :



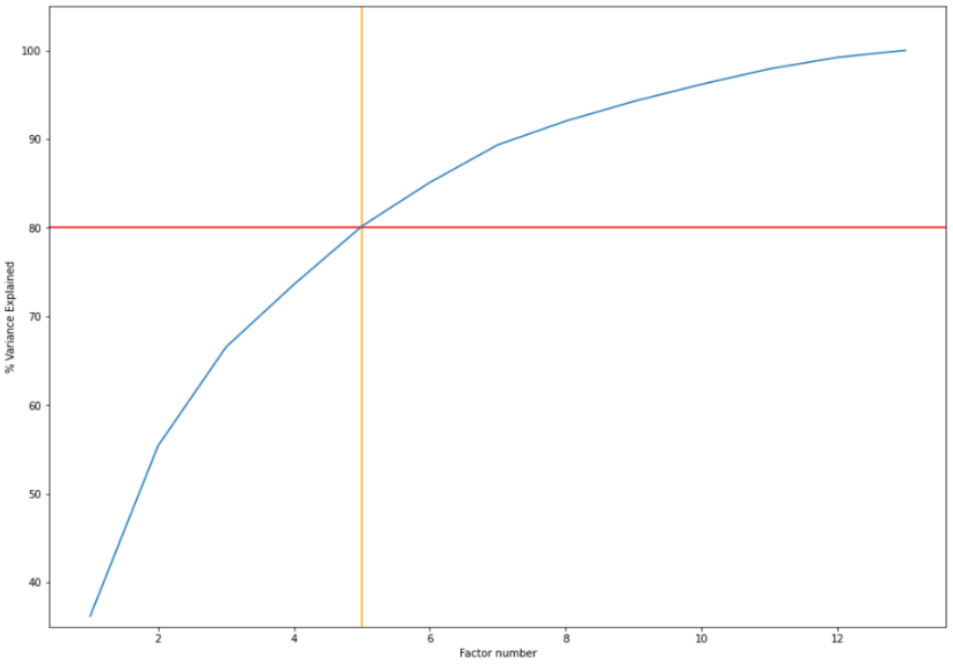
Plan des données sur 2 axes

On constate que dans ce nouvel espace réduit que les individus sont assez distinctement regroupés selon leur label, à quelques exceptions près. On peut en dire qu'à priori les valeurs prises par les variables d'origine sont différentes entre les 3 classes, à l'origine d'une variabilité au sein des variables. L'ACP a tenu en compte de cette information.

On utilise l'attribut « explained_variance_ » qui est la quantité de variance expliquée par chacune des composantes sélectionnées. Il nous permet d'obtenir 2 figures (en plus grandes dans le jupyter) :



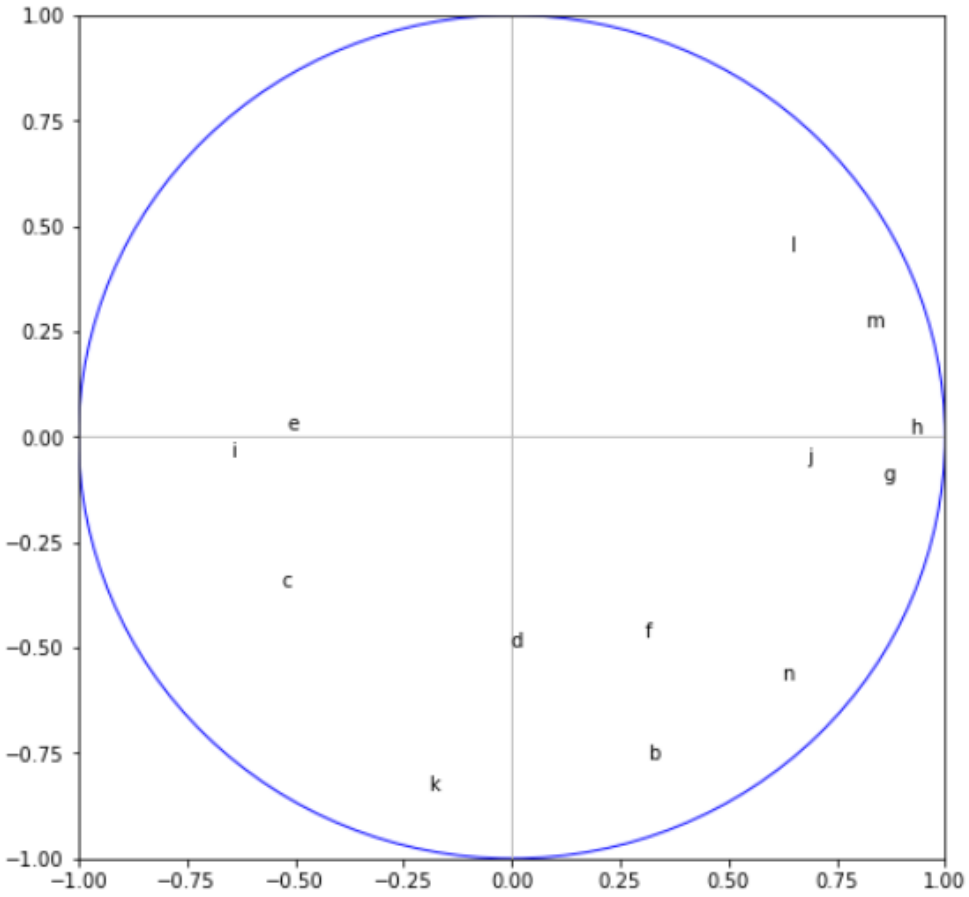
On voit sur le premier graphique que les 4 premiers facteurs retiennent une grande partie de la variabilité des données. La variabilité retenue par les facteurs suivants est décroissante.



Ici, on voit que la courbe monte assez vite au début et elle commence à monter doucement et se stabilise au fur et à mesure, ce qui montre que la variance est expliquée par les premières variables.

Cette observation est confirmée car on voit aussi que 80% de la variance cumulée est obtenue avec les 5 premières variables.

On trace ensuite le cercle de corrélation suivant :



On sait que les variables sont positivement corrélées si elles sont proches les unes des autres, négativement corrélées si elles forment un angle de 180 degrés et ne sont pas corrélées si elles forment un angle de 90 degrés.

L'interprétation des axes à partir d'un cercle des corrélations se fait de la manière suivante :

- L'axe 1, qui est corrélé positivement avec g,h, j, m, l et négativement avec i et e, et c'est un axe qui oppose les individus meilleurs en variables g,h,j, m,l et moins meilleurs en variables e et i.
- L'axe 2, qui est corrélé négativement avec les variables k,b ordonne les individus selon leur importance pour ces variable.

Axe 1 :

-	+
e, i	j, g, h, m, l

Axe 2 :

-	+
k, b	

Contribution de la variable à l'inertie de l'axe factoriel :

En pratique :

- On retient pour l'interprétation les variables dont la contribution est > à la contribution moyenne (>1/p)

- En ACP normée, ce sont les variables qui sont proches du bord du cercle qui contribuent le plus. Dans notre cas c'est les variables g,h,m

Etude de proximité entre les points :

Une fois les axes interprétés, on peut regarder les graphiques et analyser plus finement les proximités entre points.

• Un point est dit bien représenté sur un axe ou un plan factoriel s'il est proche de sa projection sur l'axe ou le plan.

S'il est éloigné, on dit qu'il est mal représenté.

• Indicateur =angle formé entre le point et sa projection sur l'axe : au plus il est proche de 90 degrés, au moins le point est bien représenté

Qualité de représentation d'une variable i sur l'axe k :

Une variable est d'autant mieux représentée sur un axe qu'elle est proche du bord du cercle des corrélations et de l'axe, d'autant plus mal représentée qu'elle est proche de l'origine.

Dans notre cas : nous avons m, h, j qui sont les mieux représentés

Remarque :

En ACP normée, les variables qui contribuent le plus à l'axe sont aussi celles qui sont le mieux représentées et inversement.

II Clustering (jupyter Clustering)

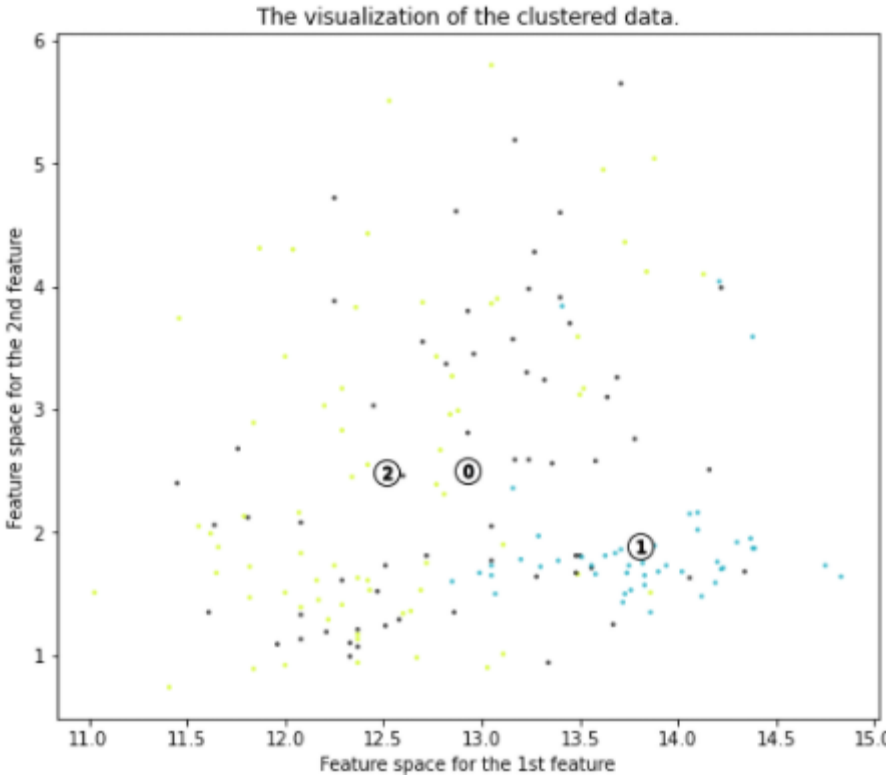
Une rapide analyse du dataset est faite au début du jupyter, pour s'assurer qu'il n'y a rien à éditer. Il n'y a pas de données catégoriques, ni de données manquantes. On sépare également l'étiquette du reste des données, comme précédemment.

Nous utilisons le composant silhouette_score de S-L qui nous donne une mesure de qualité pour une partition pour nos données. On teste plusieurs partitionnements pour prendre le meilleur score possible (proche de 1).

```
For n_clusters = 2 The average silhouette_score is : 0.6568490946514269
For n_clusters = 3 The average silhouette_score is : 0.5711220218931753
For n_clusters = 4 The average silhouette_score is : 0.562013637082329
For n_clusters = 5 The average silhouette_score is : 0.548969124044004
```

On a la valeur de silhouette la plus élevée pour 2 clusters, puis pour 3 et 4, et la valeur décroit pour plus de clusters. Pour chaque point, ce score est la différence entre la distance moyenne avec les points du même groupe que lui (cohésion) et la distance moyenne avec les points des autres groupes voisins (séparation).

Nous avons 3 labels donc on utilisera 3 clusters, même si le score est un peu plus faible.



Ici, la séparation entre les clusters n'est pas vraiment visible et vu le résultat du silhouette_score, on comprend que le nombre d'attributs pose quelques problèmes pour faire de bon clusters.

À l'aide de S-L et des labels donnés dans le dataset, nous allons évaluer la qualité de ces clusters.

On trace la matrice de confusion pour évaluer comment nos clusters classent chacun de nos individus.

La quasi-totalité des individus labelisé 1 est vraiment labelisé 1. Un peu moins d'un quart des 1 sont labelisés 3.

La classification du label 2 est un petit peu plus diffus mais reste correcte. Le label 3 est très peu réussi puisque plus de la moitié des individus classés 3 ne le sont pas.

```
Kmeans results:
[[46  0 13]
 [ 1 50 20]
 [ 0 19 29]]
```

On utilise ensuite « classification_report » pour obtenir des chiffres sur ces classements et pouvoir les comparer par la suite.

K-means sur ACP

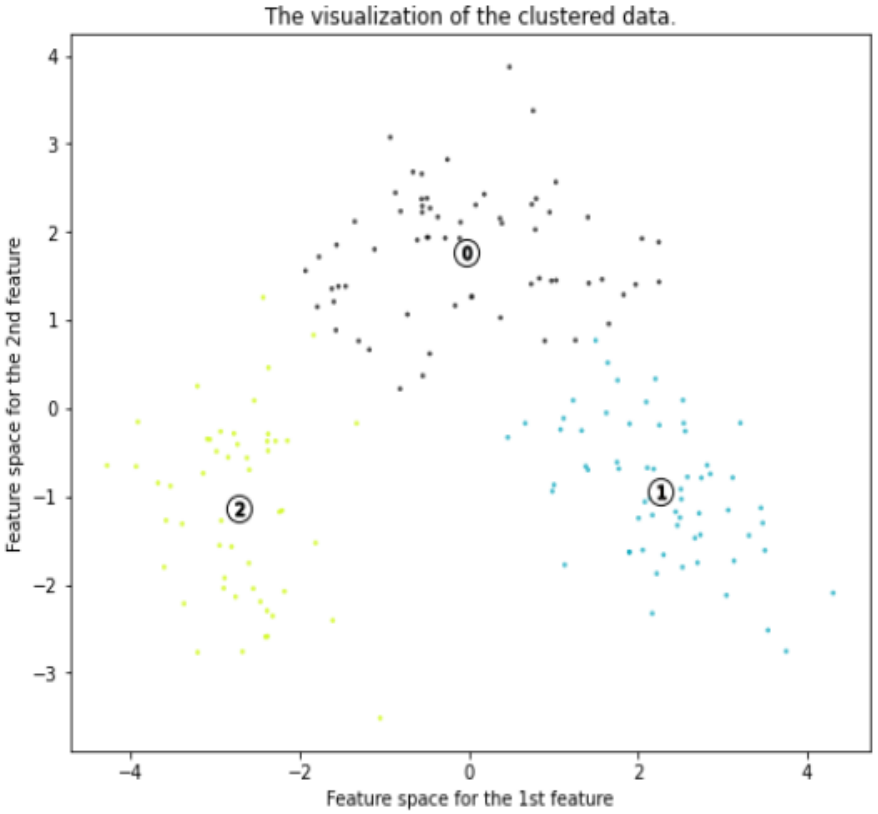
L'application de kmeans sur les données ACP se fait sur le jupyter ACP.

Cette fois-ci, le meilleur silhouette-score est pour un nombre de cluster égal à 3, ce qui correspond à notre cas.

```
For n_clusters = 2 The average silhouette_score is : 0.2593169555318254
For n_clusters = 3 The average silhouette_score is : 0.28485891918989864
For n_clusters = 4 The average silhouette_score is : 0.25558188192125253
For n_clusters = 5 The average silhouette_score is : 0.2001844376105971
```

On voit bien sur le graphe avec les 3 clusters, contrairement au nuage de points vu précédemment, que les clusters regroupent trois différentes classes.

Les trois zones sont parfaitement distinctes par rapport au premier cas.



La matrice de confusion montre que les points sont classés très précisément.

```
Kmeans results:
[[65  3  3]
 [ 0 59  0]
 [ 0  0 48]]
```

On peut à présent comparer les deux résultats obtenus grâce à classification_report de S-L. (Les labels sont réduits de 1, c'est-à-dire que le label indiqué 0 est en réalité le 1, le label 1 est le 2…)

Avec ACP

	precision	recall	f1-score	support
0	1.00	0.92	0.96	71
1	0.95	1.00	0.98	59
2	0.94	1.00	0.97	48
accuracy			0.97	178
macro avg	0.96	0.97	0.97	178
weighted avg	0.97	0.97	0.97	178

Sans ACP

	precision	recall	f1-score	support
0	0.98	0.78	0.87	59
1	0.72	0.70	0.71	71
2	0.47	0.60	0.53	48
accuracy			0.70	178
macro avg	0.72	0.70	0.70	178
weighted avg	0.74	0.70	0.71	178

La comparaison nous montre que l'application de kmeans sur les données où nous avons réalisé une analyse en composantes principales donne d'excellents résultats, bien meilleurs que l'application sans.

On peut remarquer que dans les deux cas, les individus labelisés 0 sont les plus classés sans erreur. Cependant, le recall dans le premier cas indique que c'est le seul label attribué à des individus non-concernés. Dans le second cas également, le recall est bien en deçà de la précision.

Le f1-score est calculé grâce à la précision et au recall, qui donne pour ce premier label un score correct dans les deux cas.

Les clusters classent presque parfaitement les autres labels avec l'ACP.

Sans ACP, les clusters donnent une précision qui est moyenne pour le 2 et mauvaise pour le 3. Cela était prévisible avec la silhouette score qui indiquait un meilleur score pour 2 labels, cela était une indication que les données étaient difficilement séparables en 3 groupes, mais plutôt en 2.

III ACC

L’analyse canonique (A.C.) est une méthode de statistique descriptive multidimensionnelle qui présente des analogies à la fois avec l’analyse en composantes principales (A.C.P.), pour la construction et l’interprétation de graphiques, et avec la régression linéaire, pour la nature des données. L’objectif général de l’A.C. est d’explorer les relations pouvant exister entre deux groupes de variables quantitatives observées sur le même ensemble d’individus. L’étude des relations entre deux groupes de variables constitue la principale particularité de l’A.C. par rapport à l’A.C.P.[1]

Étant donné que CCA ne peut évaluer que les corrélations r entre les variables numériques, pour notre dataset l'analyse omettra que les données numériques après le nettoyage des données comme on a fait au début.

Caractéristiques de corrélations :

• plage de valeurs : [-1,1]

• si r >0 : relation positive => les deux variables évoluent en tandem, c'est-à-dire dans le même sens. Cela signifie que lorsque l'un d'eux grandit, l'autre grandit aussi, et lorsque l'un d'eux diminue, l'autre fait de même.

• si r <0 : relation négative => en sens inverse.

• si |r| proche de 1 : plus conforme à un modèle linéaire (forte corrélation).

• si |r| proche de 0 : corrélation faible.

Cas spéciaux :

- Si r=0 : pas de relation entre les deux variables. MAIS nous ne pouvons pas conclure s'ils sont indépendants !
- Si |r|=1 : corrélation parfaite positive/négative.

Objectif :

Rechercher des combinaisons linéaires (variables canoniques) des caractères des deux ensembles de variables les plus corrélées possible.

---

[1] **Saporta, G.**, (1990). Analyse en composantes canoniques.

2 méthodes possibles :

On peut appliquer l'ACC avec ou sans utiliser S-L. Dans notre cas, on va utiliser la méthode sans sklearn (avec variance et covariance)

On fait l'analyse et la préparation des données comme on a fait avant puis on peut commençait l'ACC :

Au départ on détermine les groupe p et q et puis on commence l'ACC par les étapes suivantes :

- **Etape 1** : On standardise les données en important le package StandardScaler c'est-à-dire qu'on fait le centrage et réduction des données.

- **Étape 2** : On calcule les covariances ou les corrélations intra et inter groupes

  VX (Dim: p x p), VY (Dim: q x q),
  cross-covariance: VXY (Dim: p x q) and VYX (Dim: q x p).
  Cela va nous permettre de calculer les projections

- **Etape 3** : On calcule les Rx (Dim : pxp) et Ry (Dim : qxq) respectivement afin de récupérer les valeurs propres

  $R_X = P_X P_Y$ *et* $R_Y = P_Y P_X$ *car* $P_X P_Y a_1 = \lambda_1 a_1$ *et* $P_Y P_X b_1 = \lambda_1 b_1$

- **Etape 4** : On calcule les valeurs propres λk et les vecteurs propres ak, bk respectivement pour Rx et Ry depuis les résultats trouvés dans l'étape 3

On trouvera autant de k facteurs que la dimension originale des matrices X et Y. On ne prend que les valeurs propres communes si p<q: k=p.

$a_k^T X$ permet d'obtenir une kème-combinaison linéaire du 1er ensemble.

$b_k^T Y$ permet d'obtenir une kème-combinaison linéaire du 2nd ensemble.

où a_k et b_k correspondent aux coefficient des combinaisons linéaires ou facteurs canoniques qui maximisent les corrélations entre X et Y

Les valeurs propres et leurs vecteurs propres correspondants doivent être triés par ordre décroissant.

- **Étape 5** : Calculer les composantes canoniques Uk et Vk

$U_k = a_k^T X$ *Résulte de la combinaison linéaire du 1er ensemble*

$V_k = b_k^T Y$ *Résulte de la combinaison linéaire du 2nd ensemble*

Objectif : Maximiser la corrélation entre les composantes canoniques : max(r(Uk,Vk)). On sait que Var(U1)=Var(V1)=1 et λ1=r2(U1,V1). Algébriquement, maximiser la corrélation entre les combinaisons linéaires revient à minimiser les projections de X et Y (angle minimal). En général, le premier couple a une corrélation plus forte. On choisis un nombre limite de facteurs qui expliquent suffisamment la corrélation entre les deux groupes.

	U_1	U_2	V_1	V_2
U_1	1.000000	-0.000000	0.924994	-0.000000
U_2	-0.000000	1.000000	-0.000000	-0.000000
V_1	0.924994	-0.000000	1.000000	-0.000000
V_2	-0.000000	-0.000000	-0.000000	1.000000

Figure : Résultat de notre dataset

On voit que la corrélation est positive en plus d'être forte car les valeurs sont proches de 0.

De plus on a des Uk qui ne sont pas corrélés comme U2 V2 et U2 V1 car c'est (==0)

Nous pouvons maintenant visualiser l'observation et les variables en utilisant ses 2 premières paires de variables canoniques

Calculant la corrélation des variables avec les axes

	C1	C2	X1	X2	X3	X4	X5	X6	X7	Y1	Y2	Y3	Y4	Y5	Y6
C1	1.000000	-0.000000	-0.249380	0.451378	-0.138647	-0.913325	-0.683502	-0.626991	-0.858620	0.340739	-0.227780	-0.963910	0.521218	0.228560	-0.534268
C2	-0.000000	1.000000	0.756605	0.063127	0.016136	0.140094	0.085971	-0.323517	-0.225628	-0.559361	0.136625	0.034108	-0.196146	0.676143	0.518112
X1	-0.249380	0.756605	1.000000	0.094397	0.211545	0.289101	0.136698	-0.071747	0.072343	-0.310235	0.270798	0.236815	-0.155929	0.546364	0.643720
X2	0.451378	0.063127	0.094397	1.000000	0.164045	-0.335167	-0.220746	-0.561296	-0.368710	0.288500	-0.054575	-0.411007	0.292977	0.248985	-0.192011
X3	-0.138647	0.016136	0.211545	0.164045	1.000000	0.128980	0.009652	-0.074667	0.003911	0.443367	0.286587	0.115077	0.186230	0.258887	0.223626
X4	-0.913325	0.140094	0.289101	-0.335167	0.128980	1.000000	0.612413	0.433681	0.699949	-0.321113	0.214401	0.864564	-0.449935	-0.055136	0.498115
X5	-0.683502	0.085971	0.136698	-0.220746	0.009652	0.612413	1.000000	0.295544	0.519067	-0.197327	0.236441	0.652692	-0.365845	-0.025250	0.330417
X6	-0.626991	-0.323517	-0.071747	-0.561296	-0.074667	0.433681	0.295544	1.000000	0.565468	-0.273955	0.055398	0.543479	-0.262640	-0.521813	0.236183
X7	-0.858620	-0.225628	0.072343	-0.368710	0.003911	0.699949	0.519067	0.565468	1.000000	-0.276769	0.066004	0.787194	-0.503270	-0.428815	0.312761
Y1	0.340739	-0.559361	-0.310235	0.288500	0.443367	-0.321113	-0.197327	-0.273955	-0.276769	1.000000	-0.083333	-0.351370	0.361922	0.018732	-0.440597
Y2	-0.227780	0.136625	0.270798	-0.054575	0.286587	0.214401	0.236441	0.055398	0.066004	-0.083333	1.000000	0.195784	-0.256294	0.199950	0.393351
Y3	-0.963910	0.034108	0.236815	-0.411007	0.115077	0.864564	0.652692	0.543479	0.787194	-0.351370	0.195784	1.000000	-0.537900	-0.172379	0.494193
Y4	0.521218	-0.196146	-0.155929	0.292977	0.186230	-0.449935	-0.365845	-0.262640	-0.503270	0.361922	-0.256294	-0.537900	1.000000	0.139057	-0.311385
Y5	0.228560	0.676143	0.546364	0.248985	0.258887	-0.055136	-0.025250	-0.521813	-0.428815	0.018732	0.199950	-0.172379	0.139057	1.000000	0.316100
Y6	-0.534268	0.518112	0.643720	-0.192011	0.223626	0.498115	0.330417	0.236183	0.312761	-0.440597	0.393351	0.494193	-0.311385	0.316100	1.000000

Cercle de corrélation :

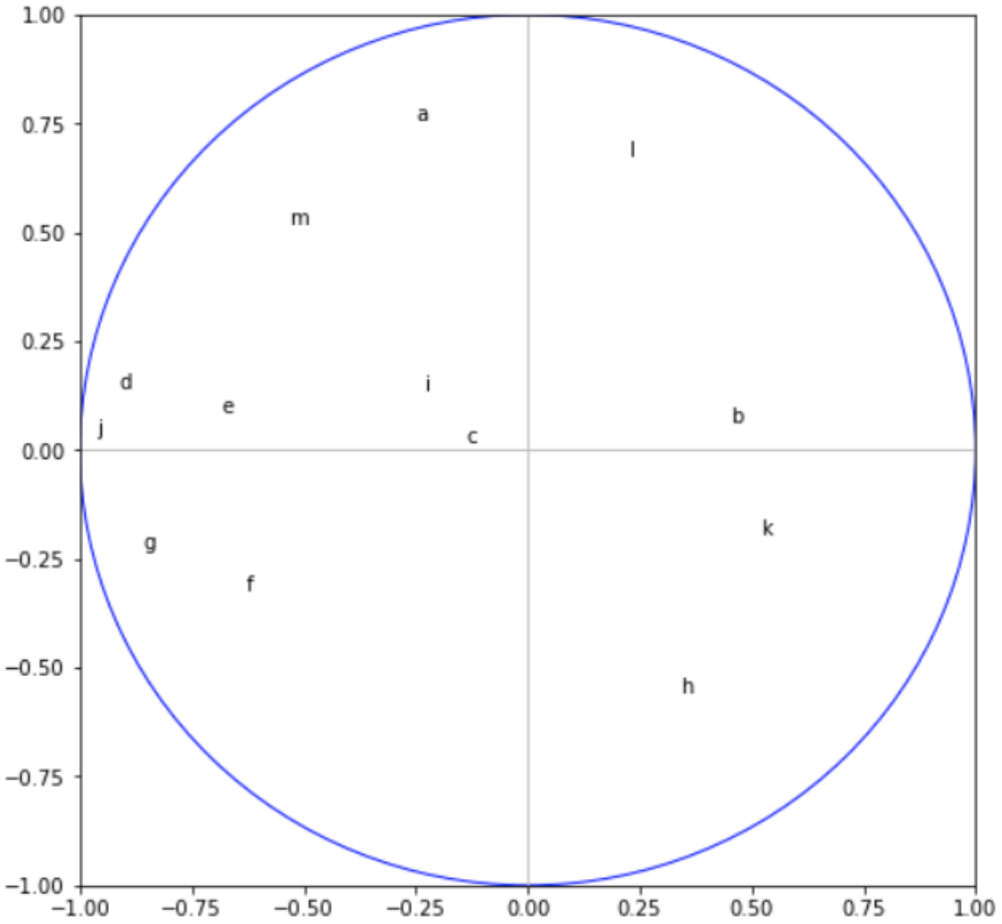


Figure : Cercle de corrélation ACC

Ce cercle de corrélation nous montre que la plupart des variables sont assez bien représentés mis à part i et c car ils sont proches de centre 0

Il nous montre également qu'il y a 2 groupes de variables corrélées à l'axe1 et 2 variables corrélé à l'axe 2 :

Axe 1 :

-	+
j, d, g, e,f	b, k

Axe 2 :

-	+
	l, m

- Groupe 1 :j, d , e, g et f
- Groupe 2 : b, k

et donc

- Groupe 1: Proanthocyanins, Ash et Alcalinity of ash, Total phenols, Magnesium
- Groupe 2: Alcohol, Color intensity

Les corrélations les plus fortes sont celles correspondant aux coefficients de corrélation les plus proches de 1 ou -1. Ici, le coefficient de corrélation dont la valeur absolue est la plus proche de 1 est celui qui relie j:Proanthocyanins et d:Ash

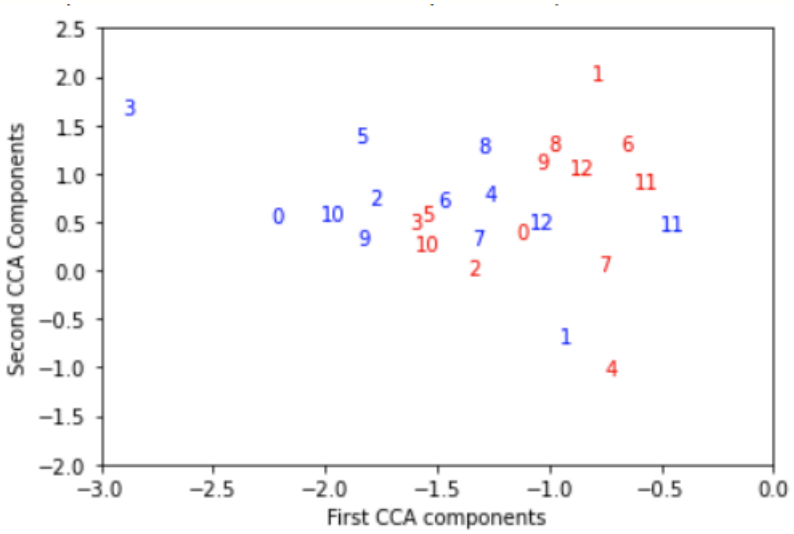
On a une forte corrélation canonique selon le y1(lambda) qui nous indique qu'il existe une assez forte relation entre les variables canoniques U1 et V1.

U1 possède une forte relation négative avec j,d,g,e,f cela correspond aux premières lignes du bloc C1 de la matrice de corrélations.

La relation des Y avec Vj est aussi forte positivement avec m, l.

Dans le 2ème exemple, on a une corrélation canonique y2(lambda) qui nous indique qu'il n'existe pas une relation entre les variables canoniques U2 et V2 car c'est 0.

Visualisation des observations



Représentation de graphe des individus

Les individus ont des caractéristiques différentes. Il est difficile de juger si les points ne sont pas confondus. On voit que pour le vin numéroté 11 dont les caractéristiques b et k ont une valeur plus élevée que pour les autres individus sur la variable b, les

deux points qui le représentent sont plus proches.
Les vins indexées 8 et 9 ont des propriétés proches mais ont des variables j, d, g, e,f différentes.

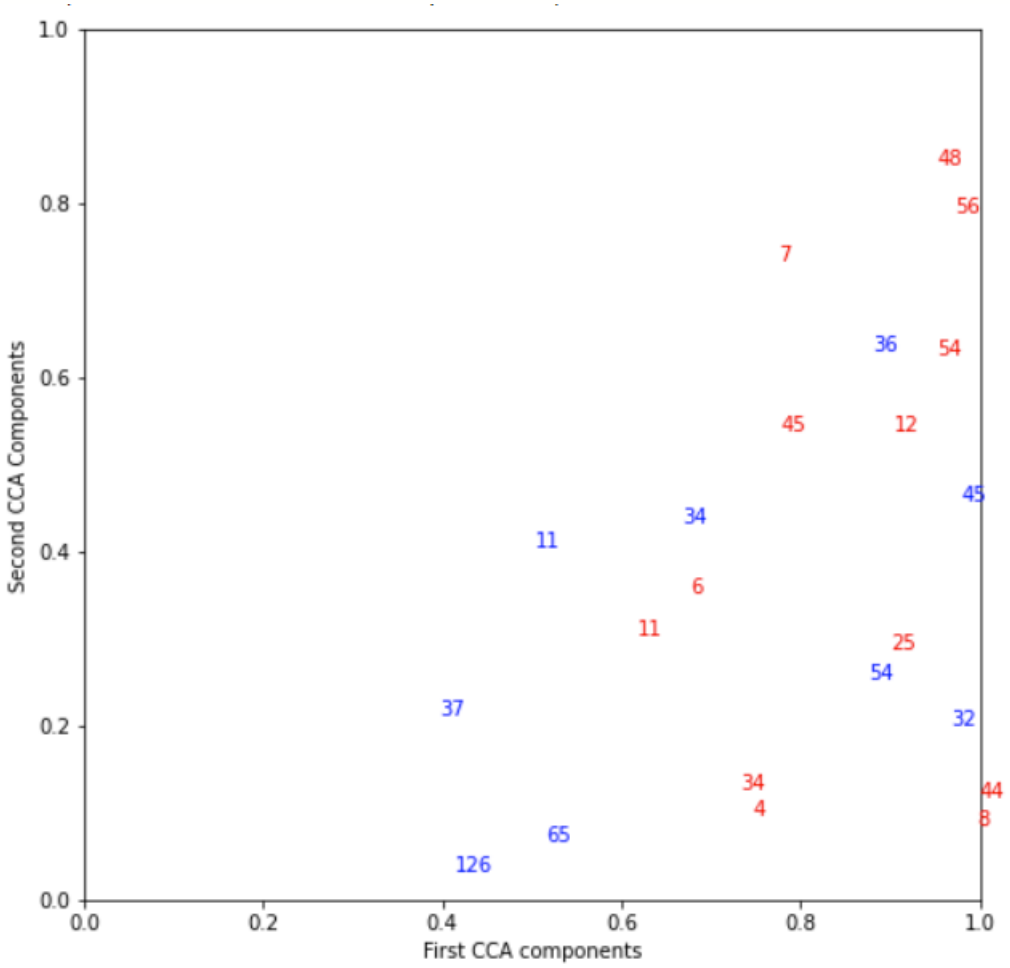


Figure : Représentation des individus sur le plan

Conclusion

Au travers de ce projet, on a mis en évidence l'intérêt de l'ACP. Tout d'abord pour la visualisation des données puisque la projection des données sur uniquement deux axes est bien plus évocatrice que lorsqu'on a essayé avec les données brutes. Mais elle permet également une bien meilleure utilisation des données. La comparaison entre les deux matrices de confusion, ainsi que la précision et le f_score qu'on peut en déduire montre non seulement une différence flagrante entre la présence et l'absence d'ACP, mais également une précision remarquable grâce à celle-ci. Cependant l'ACP a été particulièrement efficace dans notre cas puisque nous n'avions pas de valeurs aberrantes. Dans ce cas de figure, l'ACP pourrait être faussé car trop influencé par quelques valeurs. Cela peut néanmoins être utile pour identifier ces valeurs.

Concernant la partie de l'analyse en composante canonique on conclut que L'AC décrit les relations linéaires existant entre 2 ensembles de variables, les premières étapes mettent en évidence les directions de l'espace des variables selon lesquelles les deux ensembles sont les plus proches. Mais il est possible que les variables canoniques soient faiblement corrélées aux variables des tableaux X et Y. Donc elles sont difficilement interprétables. En effet, les variables d'origine n'interviennent pas dans les calculs de détermination des composantes canoniques, seuls interviennent les projecteurs sur les espaces engendrés par ces variables.