Une image contenant texte

Description générée automatiquement

Effectif = 59+71+48 = 178

On a 13 features.

L’objectif du canvas est réduire la dimensionnalité.

La dimensionnalité actuelle est : (178 x 13) [59,71,48]

Orange affiche un « ouputting » (sortie) de 9 instances

L’échantillon aléatoire n comprend 5% des données.

n = effectif \* % de l’échantillon

n = 178 \*0,05

n = 8,9 => on arrondi à 9

Donc la nouvelle dimensionnalité est : (9 x 5)

Pour avoir les nouveaux effectifs, on fait :

Effectif \* % de l’échantillon

59 \* 0,05 = 2,95 => on arrondi à 3

71 \* 0,05 = 3,55 => on arrondi à 4

48 \* 0,05 = 2,4 => on arrondi à 2

Dimensionnalité : (n,p)

Donc nouvelle dimensionnalité : (9,5)

Le p=5 représente les variables les plus discriminantes pour la statistique de Fisher.

Une image contenant texte

Description générée automatiquement

Pour calculer le boxplot, on a besoin des quartiles (Q1 et Q3), de la médiane (Q2), x1 (la plus grande valeur et xn la plus petite valeur), et l’écart inter-quartiles (EIQ).

Mais pour calculer ça, on doit ranger les données en ordre croissant.

Tableau données croissantes :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Wine | | Hue | Proline | Color | Alcohol | OD |
| 1 | 2 | 0.61 | 425 | 2.8 | 11.56 | 1.33 |
| 2 | 2 | 0.61 | 428 | 3.05 | 11.64 | 1.6 |
| 3 | 1 | 0.93 | 465 | 3.7 | 11.66 | 2.06 |
| 4 | 2 | 0.96 | 495 | 3.8 | 12.69 | 2.14 |
| 5 | 2 | 1 | 560 | 4.5 | 12.93 | 2.69 |
| 6 | 1 | 1.02 | 680 | 5.25 | 13.07 | 2.75 |
| 7 | 1 | 1.03 | 770 | 6 | 13.11 | 3.52 |
| 8 | 3 | 1.18 | 1020 | 7.1 | 14.13 | 3.58 |
| 9 | 3 | 1.23 | 1290 | 9.2 | 14.39 | 3.69 |
| X̅ | | 0.925 | 681.4 | 5.044 | 12.8 | 2.6 |
| s | | 0.205 | 282.7 | 1.975 | 0.982 | 0.826 |

1. Pour Proline :

 Q1 = ¼ \* n Q3 = ¾ \* n

Q2 = x(9+1/2) Q1 = ¼ \* 9 = 2.25 Q3 = ¾ \* 9 = 6.75

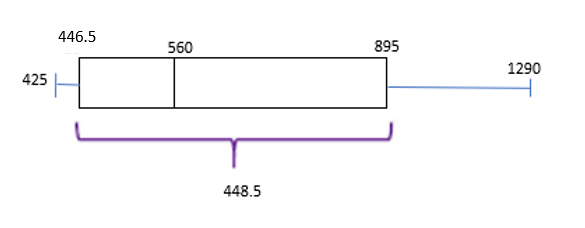
Q2 = x5 Q1 = ( x2 + x3 )/ 2 Q3 = ( x7 + x8)/ 2

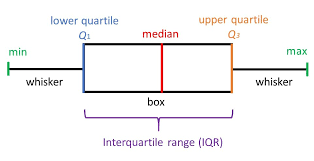
Q2 = 560 Q1 = 446.5 Q3 = 895

EIQ = |Q3 -Q1|

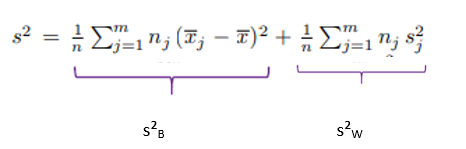
EIQ = |895-428|

EIQ = 448.5









S²B = 3/9 (1026.67 − 681.44) ² + 4/9 (517 − 681.44) ² + 2/9 (492.5 − 681.44) ² = 59 678.88

s²W = 3/9 x(212.342) ² + 4/9 x(97.052) ² + 2/9 x(67.5) ² = 20 228.01



F = ( (59 678.88)/2 ) / ( (20 228.01)/2 ) = 8,85





Une image contenant texte, montre, horloge, jauge

Description générée automatiquement

s²CH = 40.28/9 – 5.044\*0.952

s²CH = -0.369

rCH = -0.369 / 0.205\*1.975

rCH = -0.811

On remarque que le coefficient de corrélation est négatif, ce qui signifie que les valeurs sont très dispersées.

On pourrait ajuster avec un modèle affine et décroissant.



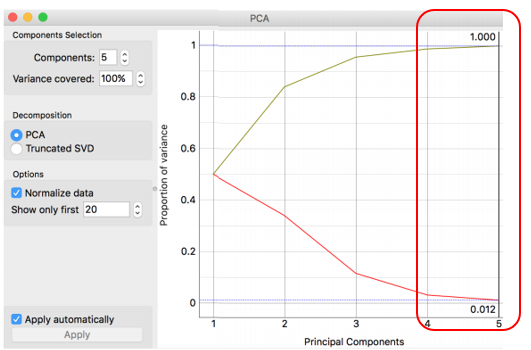
Il est préférable de choisir le modèle (2) car il est moins précis que le (3) et permet donc de ne pas passer à côté de certains points plus éloignés.

Modèle : C = aX² + bX + c



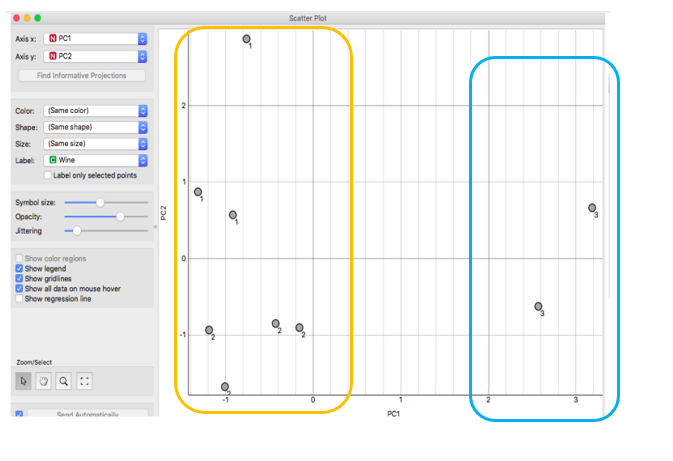
L’analyste à eut raison de normer les données lors de l’ACP car les données sont vraiment dispersées, et faire une ACP (analyse de composantes principales) sert à donner un nombre de composantes qui représente bien le jeu de données.

Si on devait donner le nombre de composantes à retenir après l’ACP, ce serait 3 ou 4 composantes principales (car sur orange, les « coudes » situés entre 3 et 5 montrent que la dispersion est trop élevée pour avoir 5 composantes, le cadre rouge montre bien que les valeurs s’éloignent trop pour pouvoir ajouter une 5e composante => variance trop élevée).



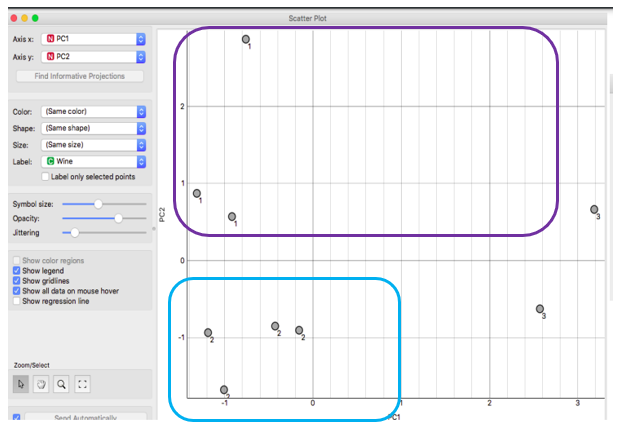


**Pour l’axe PC1 (abscisses) :** il oppose les vins 1 et 2 (avec fort « Hue » et « OD », et faible « Color ») aux vins 3 (fort « Color », mais faible « Hue » et « OD »).

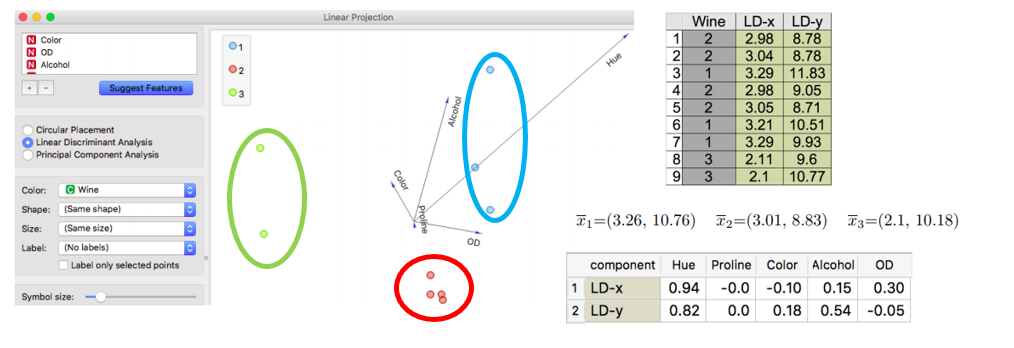


On voit bien que les vins 1 et 2 sont proches sur l’axe PC1, et s’opposent au vin 3.

**Pour l’axe PC2 (ordonnées):** oppose les vins 1 (fort  « Alcohol », « Proline ») aux vins 2 (faible « Alcohol » et « Proline »)





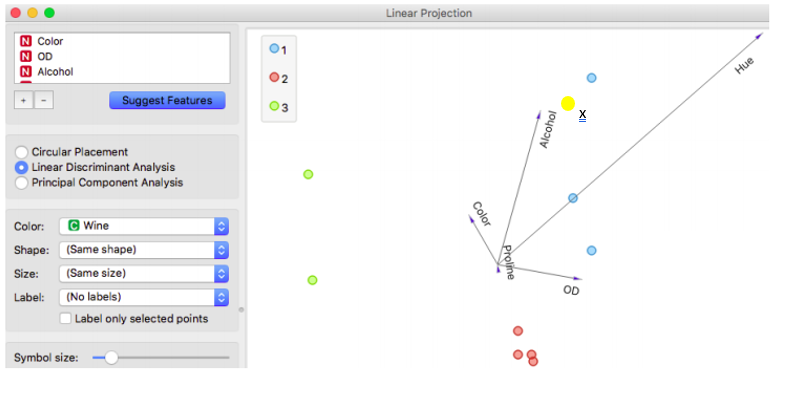


L’axe LDx donne les vins selon leur qualité :

* vins 3 (forte « Color », faible « Hue » et « OD ») ;
* vins 2 et 1 (forte « Alcohol » et « Hue » et « OD », faible « Color ») ;

L’axe LDy oppose les vins 1 (fort « Alcohol ») aux vins 2 (faible « Alcohol »).





(Pour placer x, on utilise les coordonnées par rapports aux axes => x = (Hue, Proline, Color, Alcohol, OD).

C’est un placement approximatif 😊

Qualité de x : fort « Alcohol », « Proline », « Hue » et faible « OD » , « Color »

Une image contenant table

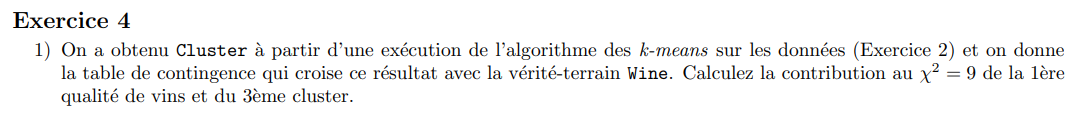
Description générée automatiquement

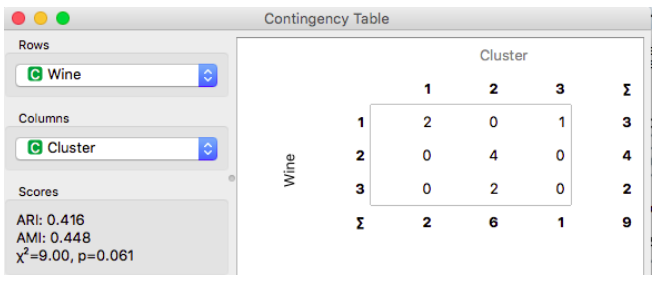


**Distance euclidienne :**









On regarde sur la table de contingence ou il y a un « 1 ».

Dans cluster, c’est la colonne 3 ; dans wine c’est la ligne 1.

***Contribution 1ère qualité de vin :***

t ligne colonne = (Colonne x ligne)/∑

t1 3 = (3x1)/9 = 1/3

***Contribution 3ème cluster :***

e 13 = ((ligne /colonne)²)/t ligne colonne

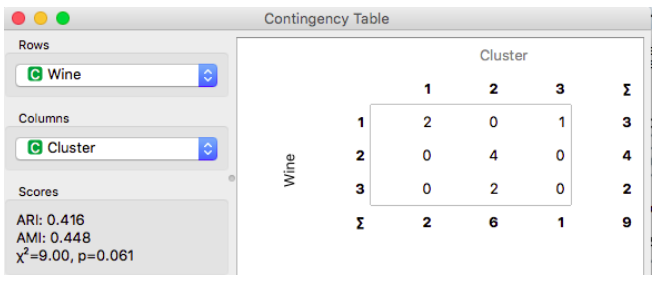
e 13 = (1-/3)² / (1/3) = 4/3

% du X² : e 13 /∑

% du X² = ( (4/3)/9 )\*100

% du X² = 14,81 (environ)

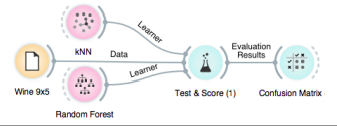




La p value vaut environ 6%, donc il y a peu de risques que l’hypothèse d’indépendance soit rejetée entre Cluster et Wine (le % pour que l’hypothèse ne soit pas rejetée est < à 5%)

Les variables Cluster et Wine sont donc plutôt dépendantes (taux à 6%), et Cluster est bien conforme à la vérité-terrain.





Wine 9x5 correspond au fichier avec les données (avec l’ACP) ;

kNN correspond aux PPV

Random Forest correspond

Test & Score sert à prédire les ensembles de décisions (appartenance à un groupe) ;

Confusion Matrix permet d’avoir la table de contigence





Confusion vins 1 :

|  |  |
| --- | --- |
| 2 | 1 |
| 1 | 5 |

R1 = 2/2+1

R1 = 2/3

R1 = 0.67

Confusion vins 2 :

|  |  |
| --- | --- |
| 2 | 2 |
| 2 | 3 |

P2 = 2/2+3

R2 = 2/5

R2 = 0.4



P3= 0

Pour trouver le score global manquant, on doit additionner les 3 valeurs de précision (P1, P2, P3) et les diviser par l’effectif des vins.

P= (3\*2/3 + 3\*2/5 + 2\*0) /9 = 0.40