Projet de M8 :
Peut-on prédire la qualité du vin ?

Coralie Farges, Charlotte Le Grand et Claire Rocchisani

2 juillet 2010

Table des matières

Introduction

Nous avons choisi ce projet d'un commun accord, après avoir rapidement étudié ses composantes. En effet, nous trouvions très intéressant d'étudier la relation entre différentes quantités et la qualité du vin, représentée par une note entre 3 et 8.

Nous avons pensé que ce projet consistait en quelque chose de très concret, et que peut-être nous pourrions envisager de trouver le vin parfait grâce à ces quantités ! De plus, c'est un sujet que nous jugeons très original, il n'a rien de très scientifique, ni de très environnemental, on pourrait l'appliquer à notre vie de tous les jours. Nous pourrions donc prendre ce projet sous un angle assez ludique et malgré tout garder en tête l'objectif initial : nos données sont-elles liées ?

Si notre sujet n'a pas l'air très scientifique, nos données le sont ! En effet, elles consistent en : le pH, le taux de dioxyde de soufre, l'acidité volatile ou encore la densité. Si l'on suppose que nos données sont liées à la qualité, le vin serait bon en fonction de ces caractères, en oubliant le goût personnel du consommateur. Connaissant ses caractéristiques telles que le pH ou le degré d'alcool, peut-on alors prédire si un vin sera jugé bon ou non ? Nous nous sommes penchées sur cette question.

Tout d'abord, nous avons commencé par de simples calculs sur nos variables, puis nous avons appliqué des méthodes vues en cours sur celles-ci, en fonction de la qualité, dernière variable principale.

Enfin, pour trouver un résultat optimal, nous avons également opté pour d'autres tests, non vus en cours, ainsi que pour l'utilisation d'un arbre de décision réalisé sous Weka. Ces outils enrichissent considérablement le contenu de notre rapport, ils sont très utilisés dans le domaine scientifique

1 Description

Nos données sont réparties en 12 variables. Ces dernières correspondent aux différentes quantités étudiés dans un vin : l'acidité, le pH, l'alcool, etc. Bien évidemment, la variable de la qualité est celle qui est prépondérante. Nous avons donc étudié ces différentes variables selon leur impact sur la qualité du vin.

Nous allons donc tenter d'établir une relation entre les variables et celle de la qualité. Mais tout d'abord, décrivons nos variables. Chaque variable contient 1599 observations. Vous pouvez voir un échantillon de nos données dans le tableau 1.

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5

TAB. 1 – Visualisation des dix premières lignes de nos données

Après avoir effectué des calculs, nous avons trouvé ces résultats pour chaque variable : la moyenne, la médiane, l'écart-type et la variance. Ces résultats sont réunis dans le tableau 2.

	Moyenne	Mediane	Ecart Type	Variance
Alcohol	10.42298	10.20000	1.06567	1.13565
Sulphates	0.65815	0.62000	0.16951	0.02873
PH	3.31111	3.31000	0.15439	0.02384
Density	0.99675	0.99675	0.00189	0.00000
TotalSulfurDioxide	46.46779	38.00000	32.89532	1082.10200
Free SulfurDioxide	15.87492	14.00000	10.46016	109.41490
Chlorides	0.08747	0.07900	0.04707	0.00222
ResidualSugar	2.53881	2.20000	1.40993	1.98790
CitricAcid	0.27098	0.26000	0.19480	0.03795
VolatileAcidity	0.52782	0.52000	0.17906	0.03206
FixedAcidity	8.31964	7.90000	1.74110	3.03142

TAB. 2 – Statistiques descriptives sur les données

Nous pouvons remarquer que certaines de ces données sont plus ou moins similaires, d'autres sont très différentes. Par exemple, les différentes sortes de dioxide ont une variance élevée. On peut donc dire qu'elle sont plus dispersées que les autres variables. Pour comparer, nous utilisons les résultats faits sur la qualité :

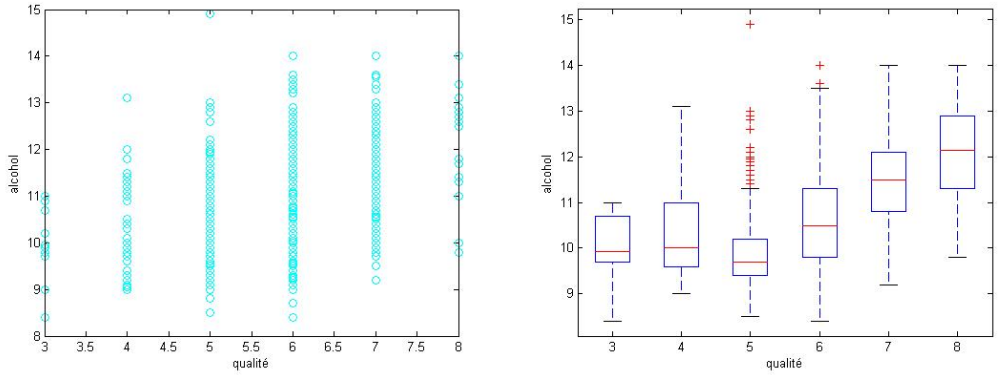
- Moyenne : 5.636023
- Ecartype : 8.075694E-1
- Variance : 6.521684E-1

La variance de la qualité est bien inférieure à celle des dioxides. On peut supposer que les données n'ont pas un fort coefficient de corrélation.

Ensuite, nous avons visualisé chacune de nos variables, en relation avec la qualité.

1.1 L'alcool

L'alcool dans le vin procure une sensation d'onctuosité ou de brûlure suivant le dosage. Les substances sucrées sont le plus souvent sur les corps possédant la fonction alcool, ce qui implique que la qualité du vin s'améliore quand le degré d'alcool augmente, comme le montre la figure 1. Cependant, un fort degré d'alcool entraîne aussi des brûlures, on peut remarquer sur la figure 1 que les valeurs extrêmes d'alcool n'ont qu'une qualité de 5 ou 6.



(a) Nuage de points (b) Boîtes à moustaches

FIG. 1 – Visualisation de la variable Alcool

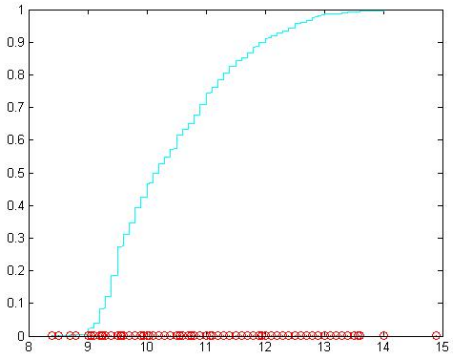
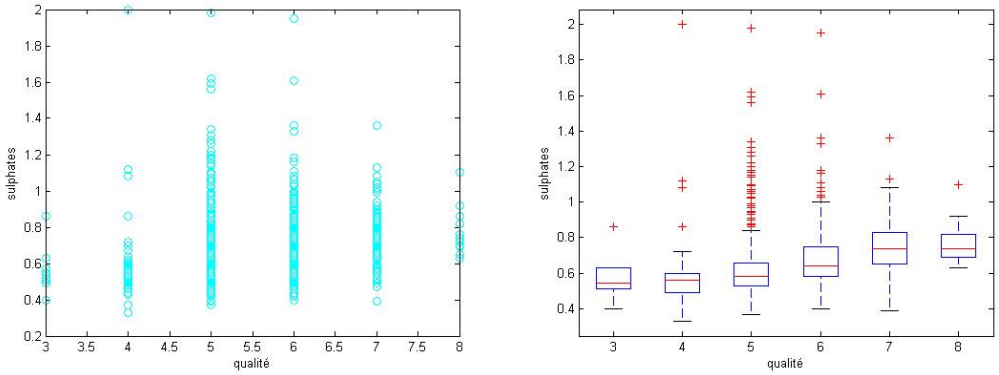


FIG. 2 – Visualisation de la fonction empirique de répartition de la variable Alcool

1.2 Les sulfates

Les sulfates sont des ions particuliers présents un peu partout dans notre vie : en agriculture, dans la fabrication d'encres. Ils sont également présents dans le vin, sous la forme particulière des sulfates de potassium. Ils ont un rôle bénéfique dans la formation de sucre et des composants organoleptiques dont les teneurs déterminent la qualité des vins. Cependant, on peut voir sur la figure 3 que les valeurs les plus fortes de sulfates ne correspondent pas à des valeurs fortes de qualité, l'influence des sulfates à des limites.



(a) Nuage de points (b) Boîtes à moustaches

FIG. 3 – Visualisation de la variable Sulfate

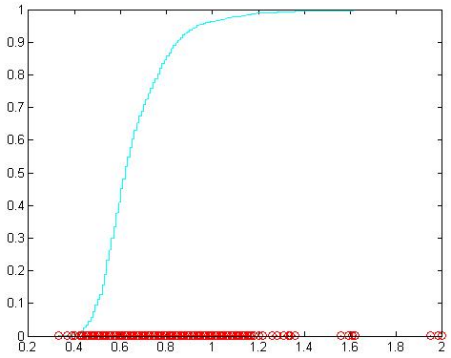
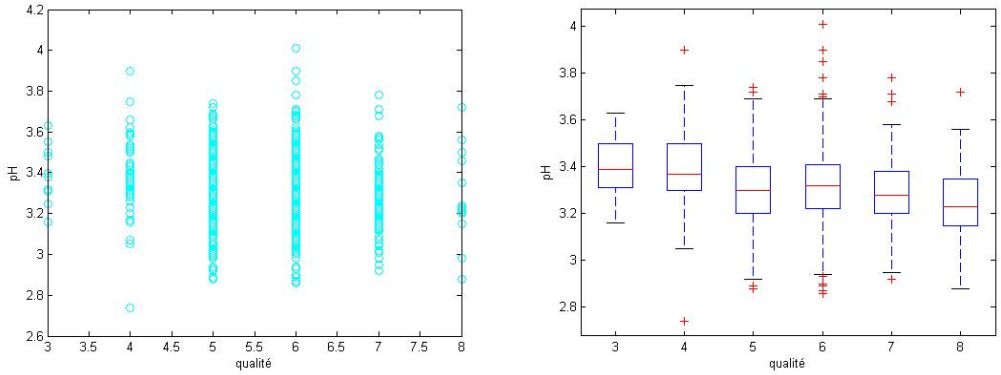


FIG. 4 – Visualisation de la fonction empirique de répartition de la variable Sulfate

1.3 Le pH

De nos jours, tout le monde sait ce que représente le pH : si une solution est plus ou moins acide, d'un point de vue chimique. De plus, les réactions chimiques de la fermentation du vin nécessite une solution acide, c'est pourquoi les valeurs du pH de la figure 5 reste dans une certaine gamme. On peut aussi remarquer que la qualité semble augmenter avec des valeurs de pH légèrement plus faible.



(a) Nuage de points (b) Boîtes à moustaches

FIG. 5 – Visualisation de la variable pH

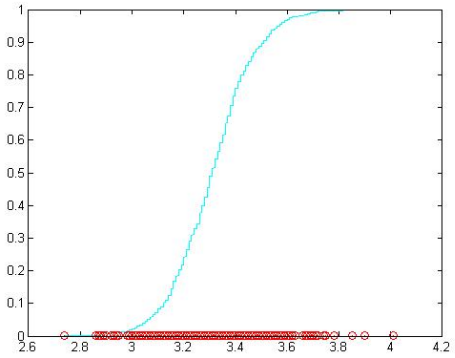
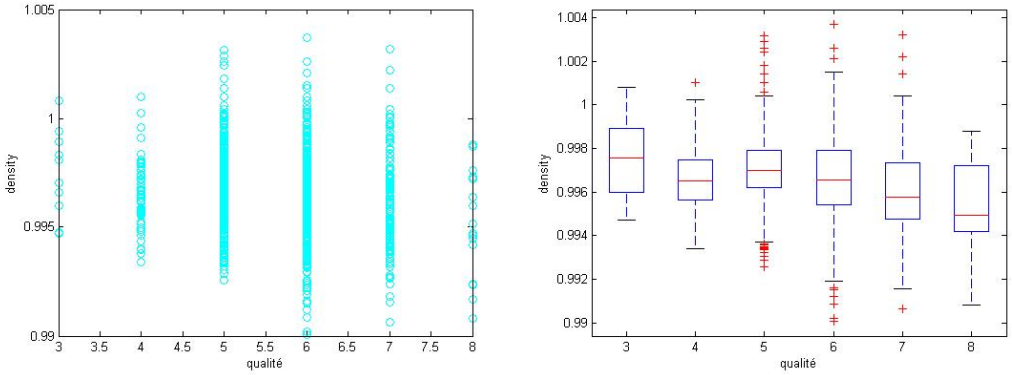


FIG. 6 – Visualisation de la fonction empirique de répartition de la variable pH

1.4 La densité

La densité est une mesure de masse volumique du vin sur la masse volumique de l'eau. Le résultat permet de savoir si un liquide ou un solide est plus ou moins "lourd" ou "léger" à volume égal. Le vin a une densité de 0.994 et l'eau une densité de 1.000. La densité inférieure du vin fait qu'il est plus "léger" que l'eau. On peut remarquer sur la figure 7 que les vins les plus appréciés ont une densité de l'ordre de 0.995.



(a) Nuage de points

(b) Boîtes à moustaches

FIG. 7 – Visualisation de la variable Densité

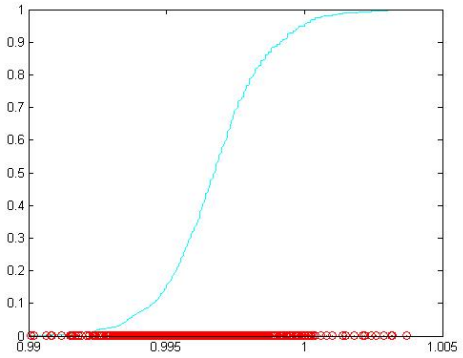
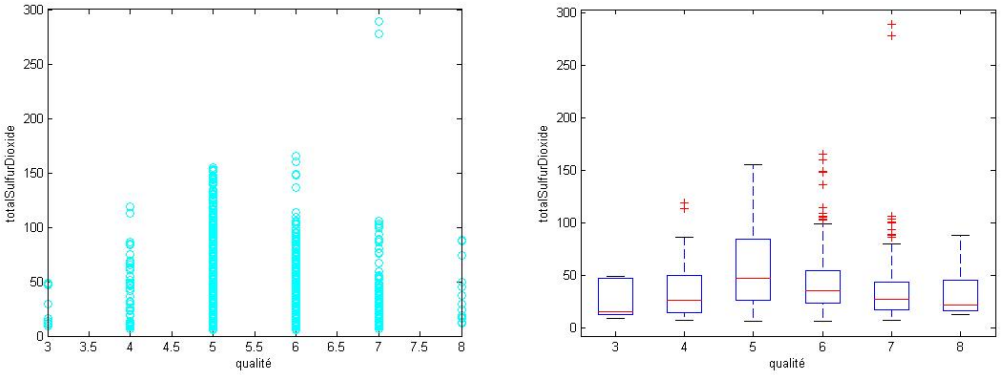


FIG. 8 – Visualisation de la fonction empirique de répartition de la variable Densité

1.5 Le total de dioxyde de soufre

Le dioxyde de Soufre est très utilisé comme désinfectant, antiseptique et antibactérien. De plus, il sert souvent de conservateur de produits alimentaires et est très utilisé en oenologie. Lors de son ajout, il évite le développement des levures et bactéries, arrêtant ainsi la fermentation dès que l'on le souhaite. Il inhibe certaines levures, mais laisse passer celles nécessaires à la vinification. On peut donc affirmer que ce taux est directement lié à la qualité du vin, car un vin débordant de bactéries n'est pas bon, ni au goût, ni pour la santé.



(a) Nuage de points (b) Boîtes à moustaches

FIG. 9 – Visualisation de la variable total de dioxyde de soufre

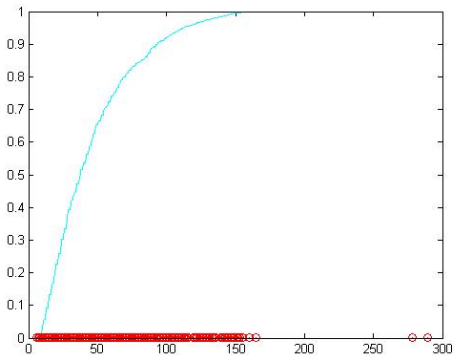
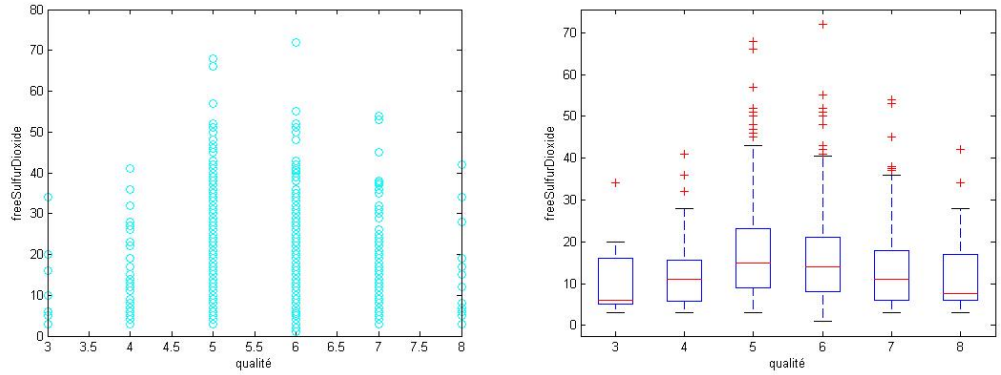


FIG. 10 – Visualisation de la fonction empirique de répartition de la variable total de dioxyde de soufre

1.6 Le dioxyde de soufre libre

Le dioxyde de soufre libre a les mêmes propriété que le dioxyde de souffre total, mais semble moins présent dans le vin .



(a) Nuage de points (b) Boîtes à moustaches

FIG. 11 – Visualisation de la variable dioxyde de soufre libre

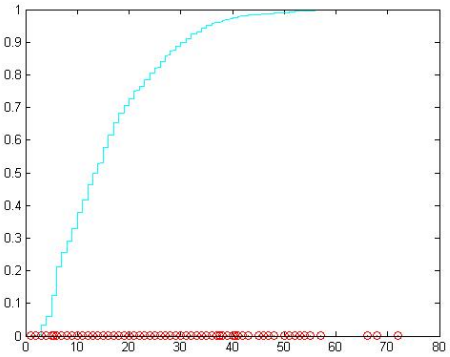
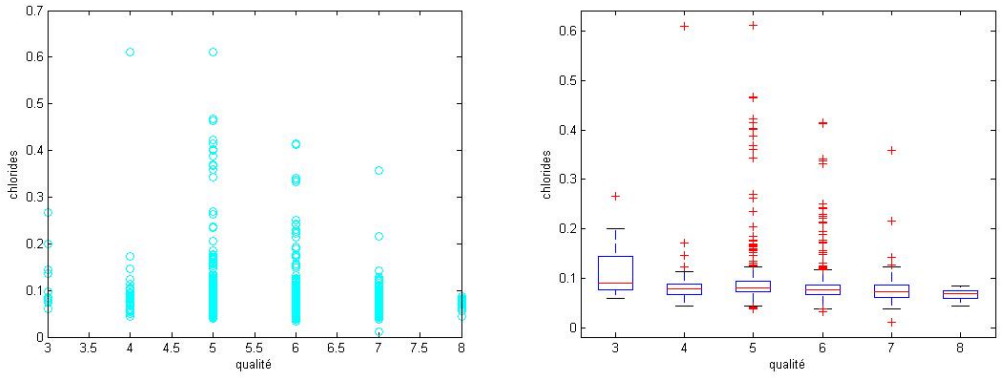


FIG. 12 – Visualisation de la fonction empirique de répartition de la variable dioxyde de soufre libre

1.7 Les chlorures

Les chlorures sont ici au même titre que le dioxyde de soufre, en tant que composé chimique. Ils peuvent agir sur le goût, mais surtout ils agissent sur la cohésion des molécules entre elles pour donner un liquide homogène, dans lequel chaque particule a sa place. Sur la figure 13 et la figure 14, le taux de chlorures semblent centré autour d'une valeur.



(a) Nuage de points

(b) Boîtes à moustaches

FIG. 13 – Visualisation de la variable chlorure

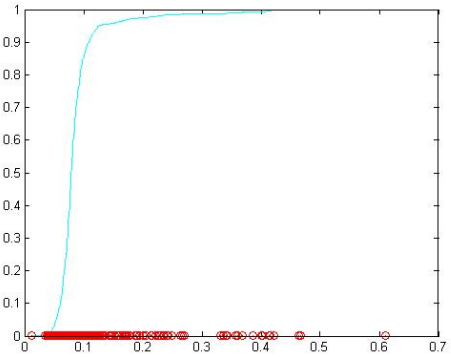
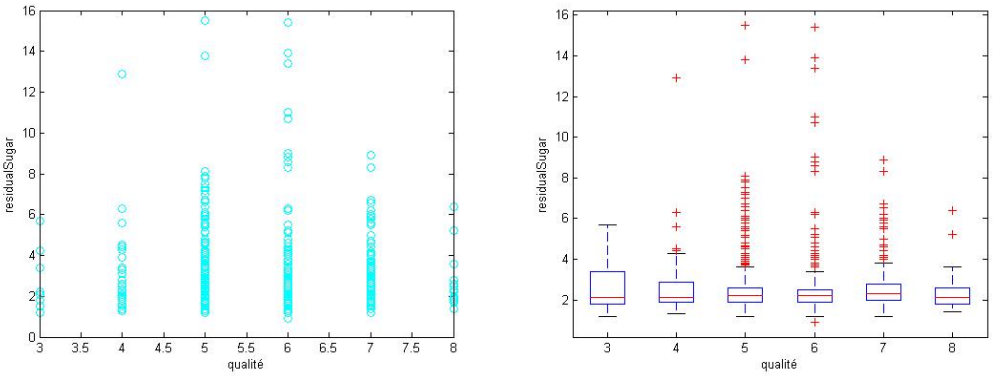


FIG. 14 – Visualisation de la fonction empirique de répartition de la variable chlorure

1.8 Les résidus de sucre

Tout comme le chlorure, les résidus de sucre semble toujours autour de la même valeur, d'après les figures 15 et 16.



(a) Nuage de points

(b) Boîtes à moustaches

FIG. 15 – Visualisation de la variable résidu de sucre

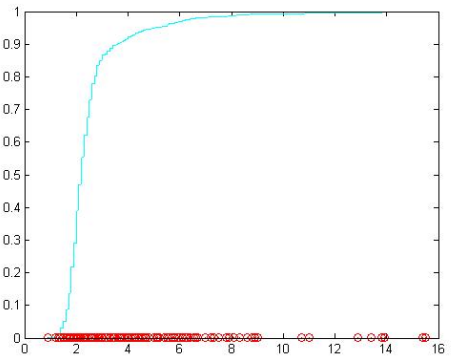
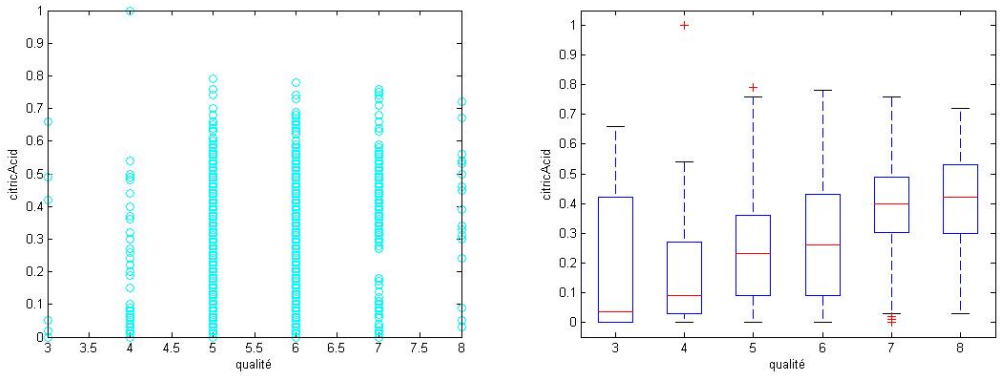


FIG. 16 – Visualisation de la fonction empirique de répartition de la variable résidus de sucre

1.9 L'acide citrique

La qualité du vin augment avec l'acide citrique d'après la figure 17. La fonction de répartition, figure 18, est presque uniforme.



(a) Nuage de points (b) Boîtes à moustaches

FIG. 17 – Visualisation de la variable acide citrique

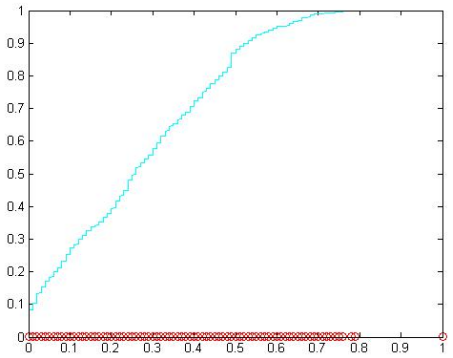
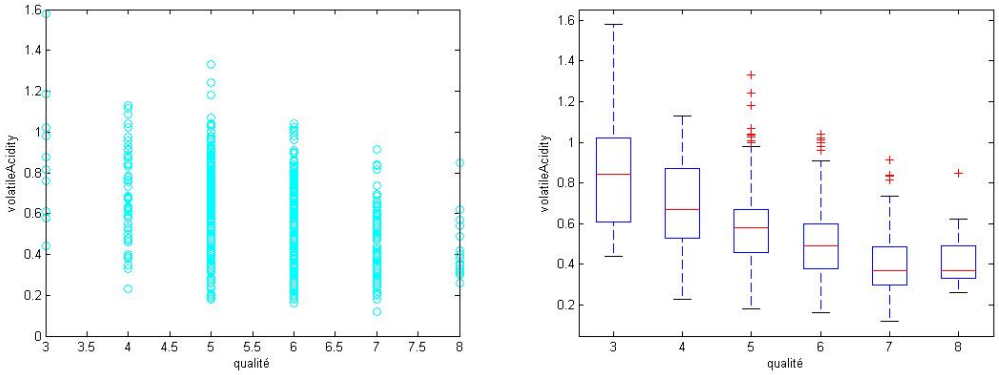


FIG. 18 – Visualisation de la fonction empirique de répartition de la variable acide citrique

1.10 L'acidité volatile

D'après la figure 19, la qualité du vin nécessiterait une acidité volatile plutôt faible.



(a) Nuage de points (b) Boîtes à moustaches

FIG. 19 – Visualisation de la variable acidité volatile

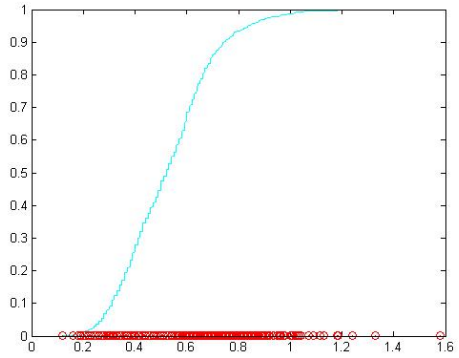
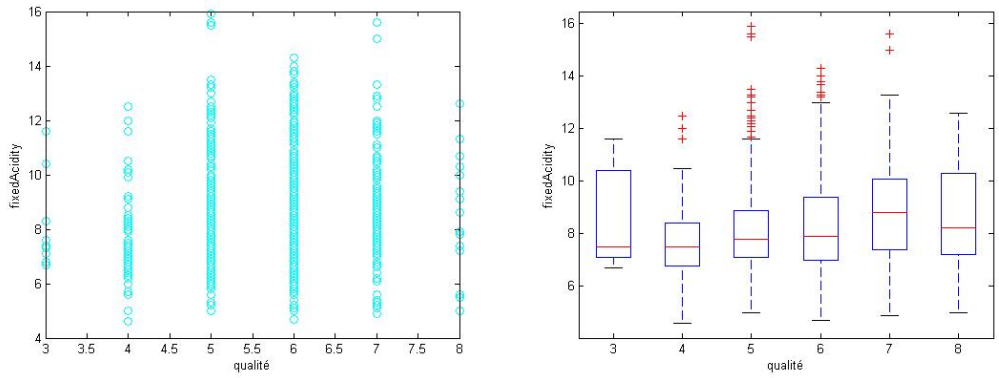


FIG. 20 – Visualisation de la fonction empirique de répartition de la variable acidité volatile

1.11 L'acidité fixe

Par contre, l'acidité fixe de semble pas trop influencer la qualité, figure 21.



(a) Nuage de points (b) Boîtes à moustaches

FIG. 21 – Visualisation de la variable acidité fixe

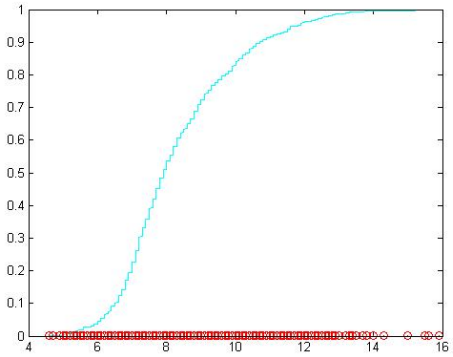


FIG. 22 – Visualisation de la fonction empirique de répartition de la variable acidité fixe

2 Regression linéaire

2.1 Regression multiple avec toutes les variables

Nous cherchons à savoir s'il est possible de prévoir la qualité du vin en se fiant à différents critères tels que l'alcool, le Ph, etc... Pour cela nous avons fait la régression de toutes nos variables en fonction de la qualité. Nous avons suivi le modèle suivant où y représente la qualité :

$$\begin{aligned} y =& a_0 + a_1\, x_{alcohol} + a_2\, x_{chlorides} + a_3\, x_{citric\_acid} + a_4\, x_{density} \\ &+ a_5\, x_{fixed\_acidity} + a_6\, x_{free\_sulfur\_dioxide} + a_7\, x_{pH} + a_8\, x_{residual\_sugar} \\ &+ a_9\, x_{sulfates} + a_{10}\, x_{total\_sulfur\_dioxide} + a_{11}\, x_{volatile\_acidity} \end{aligned}$$

Écrivons notre modèle sous la forme matricielle pour simplifier l'écriture :

$$y = X\, a$$

$$\text{avec } y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ et } a = \begin{pmatrix} a_0 \\ \vdots \\ a_{11} \end{pmatrix} \text{ et } X = \begin{pmatrix} 1 & x_{alcohol,1} & \cdots & x_{volatile\_acidity,1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{alcohol,n} & \cdots & x_{volatile\_acidity,n} \end{pmatrix}$$

Pour faire notre régression, nous avons créé un script sous Matlab :

```
Breg=[ones(n,1) X]; % X la matrice de nos variables
yn=y; % y la qualité du vin
coef=(inv(Breg'*Breg))*(Breg'*yn); % coefficient de ma regression
z=Breg*coef;% modèle
SCE=(yn-z)'*(yn-z); % variance résiduelle
SCM=(z-(ones(n,1)*mean(yn)))'*(z-(ones(n,1)*mean(yn))); % variance au  modèle
SCT = SCM + SCE; % variance totale
R2 = SCM/SCT; % coefficient de corrélation
```

Si nous prenons toutes les variables, nous obtenons :

$$R^2 = 0.3606 \quad \text{et} \quad a = \begin{pmatrix} 21.9652 \\ 0.2762 \\ -1.8742 \\ -0.1826 \\ -17.8812 \\ 0.0250 \\ 0.0044 \\ -0.4137 \\ 0.0163 \\ 0.9163 \\ -0.0033 \\ -1.0836 \end{pmatrix}$$

On peut remarquer que notre coefficient de corrélation R^2 est éloigné de 1. On peut en conclure que notre régression n'est pas satisfaisante. De plus, nous pouvons noter que certains de nos coefficients sont relativement faibles donc certaines variables n'interviennent presque pas dans le modèle. Nous allons essayer de faire une régression en ajoutant une à une les variables.

2.2 Regression multiple en ajoutant une à une les variables

Nous allons commencer par faire une régression pour chaque variable en fonction de la qualité et calculer le R^2, nous saurons alors quelle variable est plus importante pour le modèle.

Variables	Modèle	R^2
alcohol	$y = a_0 + a_1 x_{alcohol}$	0.2267
chlorides	$y = a_0 + a_1 x_{chlorides}$	0.0166
citric acid	$y = a_0 + a_1 x_{citric\,acide}$	0.0512
density	$y = a_0 + a_1 x_{density}$	0.0306
fixed acidity	$y = a_0 + a_1 x_{acidity}$	0.0154
free sulfur dioxide	$y = a_0 + a_1 x_{free\,sulfur\,dioxide}$	0.0026
pH	$y = a_0 + a_1 x_{pH}$	0.0033
residual sugar	$y = a_0 + a_1 x_{residual\,sugar}$	1.8856E-4
sulfates	$y = a_0 + a_1 x_{sulfates}$	0.0632
total sulfur dioxide	$y = a_0 + a_1 x_{total\,sulfur\,dioxide}$	0.0343
volatile acidity	$y = a_0 + a_1 x_{volatile\,acidity}$	0.1525

TAB. 3 – Coefficient R^2 pour chaque variable

On remarque que l'alcool est la variable avec le plus grand R^2 suivie par l'acidité volatile. Nous allons essayer de faire la régression linéaire de ces deux variables uniquement. Nous allons suivre le modèle suivant :

$$y = a_0 + a_1\, x_{alcohol} + a_2\, x_{volatile\,acidity}$$

Sous forme matricielle :

$$y = X\, a$$

$$\text{avec } y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ et } a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \text{ et } X = \begin{pmatrix} 1 & x_{alcohol,1} & x_{volatile\_acidity,1} \\ \vdots & \vdots & \vdots \\ 1 & x_{alcohol,n} & x_{volatile\_acidity,n} \end{pmatrix}$$

On obtient :

$$R^2 = 0.3170 \quad \text{et} \quad a = \begin{pmatrix} 3.0955 \\ 0.3138 \\ -1.3836 \end{pmatrix}$$

Notre coefficient de corrélation n'a pas augmenté : nous avions $R^2 = 0.3606$ avec toutes les variables et maintenant avec les deux meilleures nous avons $R^2 = 0.3170$. Essayons d'ajouter le sulfate :

$$R^2 = 0.3359 \quad \text{et} \quad a = \begin{pmatrix} 2.6108 \\ 0.3092 \\ -1.2214 \\ 0.6790 \end{pmatrix}$$

Nous supposons donc que nos données ne sont pas très bonnes et nous décidons d'enlever des points aberrants.

2.3 Regression multiple en enlevant des points aberrants

Nous enlevons les points aberrants grâce à un script Matlab :

```
Xn=B;
yn=y;
indn=(1:size(B,1))';
% Points aberrants
p_ab=[]; % exemple: [1,5] on enlève le point 1 et le 5
Xn(p_ab,:)=[];
yn(p_ab)=[];
indn(p_ab)=[];
```

Pour savoir quels points enlever nous faisons un diagnostic. Nous allons calculer la contribution de chaque point au modèle, nous pourrons ainsi enlever les points ayant une forte contribution car

ils ont un effet de levier. Les points ayant une forte erreur par rapport au modèle peuvent eux aussi être retirés. Étant donné le nombre de données dont nous disposons nous avons créé un programme Matlab qui effectue le diagnostic.

```
% Prédictions:
z = Xn*a;
% Erreurs (Résidus):
e  = yn-z;
% Calcul du coefficient de détermination R2
% Variance totale
SCT=(yn-mean(y))'*(yn-mean(y));
% Variance Expliquée
SCM=(z-mean(yn))'*(z-mean(yn));
% coefficient de détermination R2 :
R2=SCM/SCT
% Matrice d'influence
H = Xn*inv(Xn'*Xn)*Xn';
h = diag(H); % contient les Hii
% Calculs des résidus standardisés
nn = n-1;
err= e'*e/(nn-p);
r = e./sqrt(err*(1-h));
% erreur de validation croisée
ei = e./(1-h);
% Contribution
c=e.*e.*h./(p*(1-h).^2.*err);
```

Nous enlevons tous les points ayant une contribution c forte et une erreur de validation croisée e_i importante. En faisant la régression linéaire, nous obtenons :

$$R^2 = 0.4017 \quad \text{et} \quad a = \begin{pmatrix} 33.4991 \\ 0.2768 \\ -1.9506 \\ -0.2450 \\ -29.7704 \\ 0.0362 \\ 0.0046 \\ -0.4335 \\ 0.0204 \\ 1.2604 \\ -0.0032 \\ -0.9434 \end{pmatrix}$$

Ce résultat est meilleur mais ne nous semble pas satisfaisant. Nous avons essayé de reprendre les variables une à une mais le coefficient de corrélation ne change pas. Nos données pourrait ne pas être linéaire, nous allons faire une Analyse en Composante Principale (ACP) pour éliminer le bruit de nos variables et voir si la régression s'améliore.

2.4 Regression en utilisant les données de l'ACP

2.4.1 L'ACP

L'ACP est une méthode qui consiste à transformer des variables corrélées en nouvelles variables indépendantes les unes des autres. Ces nouvelles variables sont appelées les composantes principales ou les axes. Pour faire une ACP il faut d'abord réduire et centrer notre matrice de données :

$$X_{centre\,rduit} = \frac{X - \mu}{\sigma}$$

avec μ la moyenne de X et σ son écart-type.

Nous avons ensuite besoin d'une matrice symétrique définie positive S telle que :

$$S = \frac{X^T X}{n}$$

Nous allons calculer les valeurs propres et les vecteurs propres de cette nouvelle matrice S, nous pourrions très bien utiliser ses valeurs singulières. Pour faire ces différents calculs sur nos données nous avons fait un script Matlab :

```
B=X; %donnée
n=size(B,1);
Bn=(B-(ones(n,1)*mean(B)))./(ones(n,1)*std(B)); %on centre et on reduit nos données
p=size(B,2);
S=1/n*(Bn'*Bn); % matrice des corrélations (Bn doit être centré et réduit)
[V, D]= eig(S); % Calcul des valeurs propres et des axes
```

On obtient les valeurs propres (les composantes de $diag(D)$) et les vecteurs propres (colonnes de V) suivants :

$$\texttt{diag(D)} = \begin{pmatrix} 0.0595 \\ 0.1812 \\ 0.3444 \\ 0.4227 \\ 0.5834 \\ 0.6592 \\ 0.9587 \\ 1.2125 \\ 1.5496 \\ 1.9247 \\ 3.0972 \end{pmatrix}$$

et

$$V = \begin{pmatrix} -0.3145 & 0.3030 & -0.0376 & -0.2176 & 0.3277 & -0.3611 & 0.3507 & 0.1222 & -0.4717 & -0.3862 & 0.1132 \\ 0.0531 & 0.2177 & -0.1113 & -0.3570 & -0.3704 & -0.3043 & 0.2465 & -0.6662 & 0.0926 & 0.1481 & -0.2122 \\ -0.0709 & -0.6217 & 0.3814 & -0.3775 & -0.1055 & -0.0696 & -0.0586 & 0.0794 & -0.238 & -0.1518 & -0.4636 \\ -0.5673 & 0.2400 & -0.0207 & -0.2392 & 0.1705 & 0.3912 & 0.1571 & 0.1745 & 0.3389 & 0.2336 & -0.3954 \\ 0.6397 & 0.2495 & -0.1940 & -0.1776 & 0.3502 & -0.1015 & -0.0826 & 0.2296 & 0.1233 & -0.1105 & -0.4893 \\ -0.0514 & -0.2485 & -0.6354 & -0.2048 & 0.1166 & 0.0140 & -0.1592 & 0.0435 & -0.4288 & 0.5136 & 0.0362 \\ 0.3407 & 0.0110 & 0.1677 & -0.5614 & 0.0251 & 0.5221 & 0.2675 & 0.0038 & -0.0577 & 0.0067 & 0.4385 \\ 0.1840 & -0.0929 & -0.0075 & 0.2998 & -0.2907 & -0.0492 & 0.7321 & 0.3728 & -0.1013 & 0.2721 & -0.1461 \\ 0.0696 & -0.1123 & 0.0584 & 0.3746 & 0.4475 & 0.3813 & 0.2260 & -0.5509 & -0.2798 & -0.0376 & -0.2429 \\ 0.0687 & 0.3708 & 0.5921 & 0.0190 & 0.0937 & -0.1363 & -0.2225 & 0.0346 & -0.3224 & 0.5695 & -0.0236 \\ 0.0024 & -0.3659 & 0.1291 & -0.0788 & 0.5337 & -0.4114 & 0.2187 & -0.0790 & 0.4500 & 0.2749 & 0.2380 \end{pmatrix}$$

Afin de donner plus de sens à nos valeurs propres, nous allons regarder leur somme cumulée et ainsi voir leur influence sur nos données.

Valeurs propres	Influence cummulée (%)
3.0972	28,2
1.9247	45,7
1.5496	59,8
1.2125	70,8
0.9587	79,6
0.6592	85,5
0.5834	90,8
0.4227	94,7
0.3444	97,8
0.1812	99,5
0.0595	100,0

TAB. 4 – Visualisation de l'influence des valeurs propres

On peut remarquer que pour obtenir 70% de nos données il faut 4 valeurs propres. Il est difficile de visualiser une projection en 4 dimensions. Nous allons donc essayer de projeter nos variables sur nos 3 axes qui ont le plus d'influence et ainsi avoir 60% d'influence.

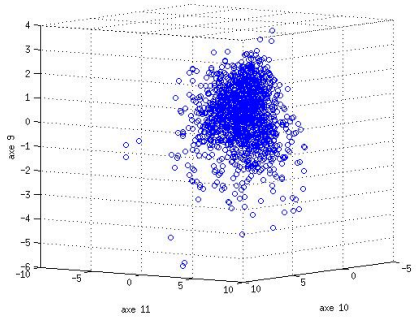


FIG. 23 – Visualisation des données projetées (1^e vue)

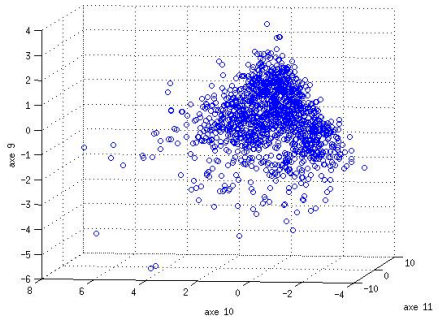


FIG. 24 – Visualisation des données projetées (2^e vue)

Nous avons projeté la matrice selon différents axes principaux. Dans le nuage d'individus, on doit discerner la formation de plusieurs groupes d'individus. Les proximités entre individus s'interprètent en termes de similitudes de comportement vis à vis des axes principaux. Cependant dans notre projection, les groupes d'individus ne sont pas visibles. Il semblerait qu'on ne puisse pas trouver de similitude distincte entre les individus.

Nous allons maintenant projeter les variables en fonction de nos axes principaux :

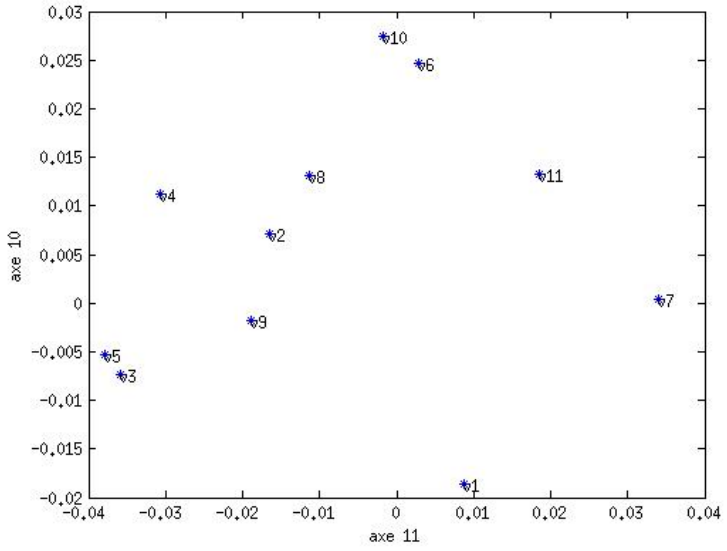


FIG. 25 – Visualisation des variables

Cette visualisation permet de savoir si des variables sont corrélées : si 2 variables sont proches, on dit qu'elles sont fortement corrélées, par contre si elles sont éloignées, on dit qu'elles sont faiblement corrélées. Dans notre cas, les variables sont relativement proches les unes de autres vu que notre échelle va de -0.02 à 0.03 en ordonnée et de -0.04 à 0.04 en abscisse. Il semblerait que nos variables soient toutes corrélées entre elles.

Essayons tout de même une régression linéaire pour voir le changement sur le coefficient de corrélation.

2.4.2 La regression avec les données obtenues

On crée une nouvelle matrice C qui calcule la composante principale des individus :

$$C = Xn\,V$$

où Xn est notre matrice de données centrées et réduites et V la matrice des vecteurs propres.
On obtient alors la meilleure approximation de nos données A en faisant :

$$A = C\,Vn^T$$

où Vn est la matrice des vecteurs propres centrés et réduits.
En appliquant notre programme Matlab de régression on obtient :

$$R^2 = 0.3606 \quad \text{et} \quad a = \begin{pmatrix} 5.6360 \\ 26.6514 \\ -2.7459 \\ -17.5430 \\ 27.3844 \\ -18.7353 \\ 6.5785 \\ -19.9419 \\ -6.8822 \\ 2.6327 \\ -5.6963 \\ -11.4082 \end{pmatrix}$$

On note que le coefficient de corrélation est le même que sans l'ACP. Notre ACP est donc inutile, cela semble normal vu que nos variables sont toutes corrélées les unes avec les autres.

Nous avons essayé de faire une régression linéaire avec toutes nos variables, puis en ajoutant une à une nos variables et enfin en faisant une ACP. Cependant, nos coefficients de corrélation sont toujours faibles. Il semblerait que nos variables ne soient pas linéaires. Nous allons essayer de faire différents tests pour le prouver.

3 Tests

Les tests en statistique permettent de rejeter ou ne pas rejeter une hypothèse en se basant sur des données. Il existe différents types de tests, le plus souvent classés selon leurs finalités :

- « le **test de conformité** consiste à confronter un paramètre calculé sur l' échantillon à une valeur pré-établie ;
- le **test d'adéquation** consiste à vérifier la comptabilité des données avec une distribution choisie a priori ;
- le **test d'homogénéité** (ou de comparaison) consiste à vérifier que K (K >= 2) échantillons (groupes) proviennent de la même population ou, cela revient à la même chose, que la distribution de la variable d'intérêt est la même dans les K échantillons ;
- le **test d'association** (ou indépendance) consiste à éprouver l'existence d'une liaison entre 2 variables. » (Source : Wikipédia)

Nos données ne semblent pas linéaires. Pour le prouver nous allons faire des tests. Nous proposons de diviser nos données selon les valeurs de la qualité, ainsi nous allons créer des échantillons où la qualité est égale à 5 par exemple. Nous pourrons ainsi comparer les échantillons avec des tests d'homogénéité (Test de Kolmogorov-Smirnov, Test de la somme des rangs de Wilcoxon) et voir s'ils sont en adéquation (test du χ^2).

Nous avons divisé nos données suivant leur qualité, pour simplifier nous avons juste pris la 11^e variable de l'ACP (celle qui représente le plus d'information) et nous les avons représentées sous la forme d'histogramme. Nous avons créé le script Matlab suivant pour faire nos histogrammes :

```
w=max(C(:,11))-min(C(:,11));
k=10;
h=w/k;
H=ones(1,k);
for␣i=1:k+1
%i=0,
H(i)=length(find(␣((i-8)*h)␣<=␣C(y==4,11)␣&␣C(y==4,11)␣<((i-7)*h)));
(i-8)*h,
end
bar(H)
```

On obtient :

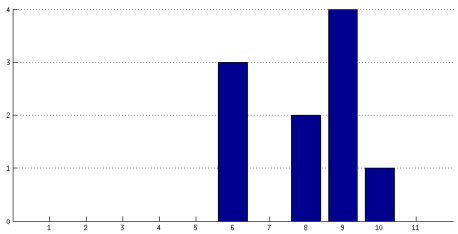


FIG. 26 – Histogramme de la qualité égale à 3

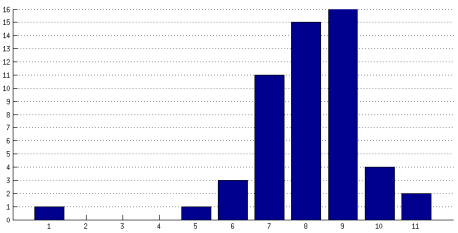


FIG. 27 – Histogramme de la qualité égale à 4

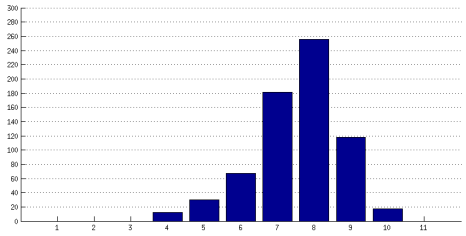


FIG. 28 – Histogramme de la qualité égale à 5

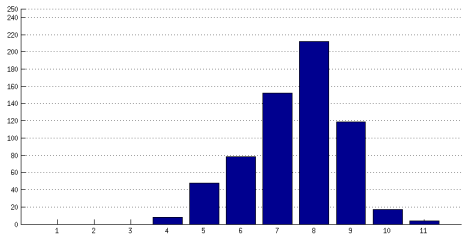


FIG. 29 – Histogramme de la qualité égale à 6

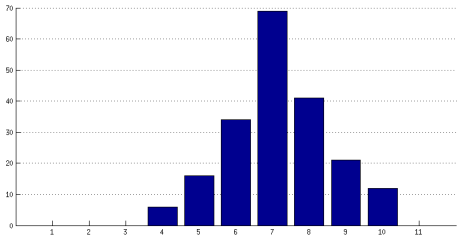


FIG. 30 – Histogramme de la qualité égale à 7

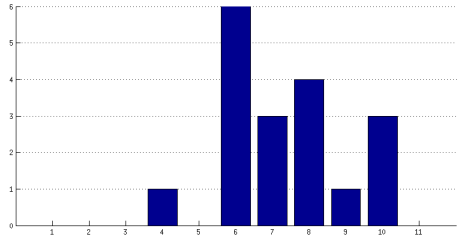


FIG. 31 – Histogramme de la qualité égale à 8

3.1 Le test de Kolmogorov-Smirnov

3.1.1 Le principe

Le test de Kolmogorov-Smirnov est un test de comparaison ou d'homogénéité, il vérifie que nos échantillons viennent de la même distribution. En d'autres termes, ce test permet de voir si les deux échantillons viennent de la même distribution en déterminant si la distance entre les deux fonctions de répartition est grande ou pas. Ce test nécessite différentes étapes avant d'arriver à une conclusion :

1. on définit les hypothèses H_0 (les données viennent de la même distribution) et H_1 (les données viennent de distributions différentes) ;
2. on dispose de deux échantillons x et y, on calcule les distances $D^+ = \max\{x_i - y_i\}$, $D^- = \max\{y_i - x_i\}$ et on pose $D = \max\{D^+, D^-\}$;
3. on détermine le seuil critique $D_a(n)$ grâce au table ci-dessous ;
4. on calcule la $p - val$, grâce à $p - val = P(D \geq D_a(n))$
5. si la $p - val \geq 0.05$ (avec 0.05 le risque de première espèce) alors les échantillons ne sont pas homogènes (H_1).

Taille de l'échantillon (n)	Seuils critiques $D_a(n)$				
	$a = .20$	$a = .15$	$a = .10$	$a = .05$	$a = .01$
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	.356
25	.210	.220	.240	.270	.320
30	.190	.200	.220	.240	.290
35	.180	.190	.210	.230	.270
> 35	$1.07/\sqrt{n}$	$1.14/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

FIG. 32 – Table des seuils $D_a(n)$

3.1.2 Le test sur nos données

Nous avons séparé notre distribution suivant la qualité, nous allons donc comparer deux à deux chaque échantillon.

1. On pose :
– H_0 : les données sont homogènes
– H_1 : les données ne sont pas homogènes.
2. On calcule les distances : D^+, D^- et D.
3.$n = 1599$ on calcule donc $D_a(n)$ avec 0.05 :

$$D_a(1599) = \frac{1.36}{\sqrt{1599}} = 0.034$$

4.On calcule les $p - val$ on obtient :

$p - val$	y=4	y=5	y=6	y=7	y=8
y=3	0.7333	0.0353	0.508	0.0086	0.0719
y=4	-	0.0007	0.0016	0.0000	0.0263
y=5	-	-	0.1984	0.0000	0.0263
y=6	-	-	-	0.0000	0.3849
y=7	-	-	-	-	0.8697

TAB. 5 – Tableau des $p - val$ pour le test de Kolmogorov-Smirnov

5. On peut en conclure les hypothèses :

H	y=4	y=5	y=6	y=7	y=8
y=3	0	1	0	1	0
y=4	-	1	1	1	1
y=5	-	-	0	1	0
y=6	-	-	-	1	0
y=7	-	-	-	-	0

TAB. 6 – Tableau des hypothèses pour le test de Kolmogorov-Smirnov

Afin de donner un aperçu visuel du test, dessinons les courbes de deux échantillons homogènes et de deux échantillons non-homogènes :

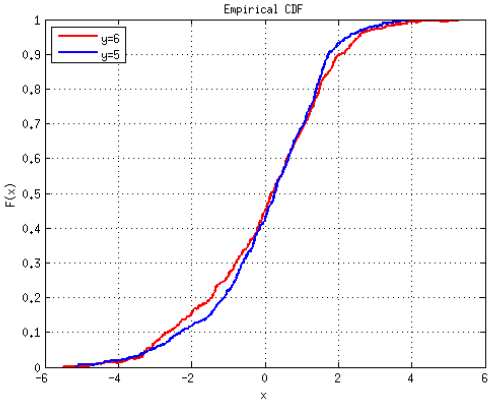


FIG. 33 – Fonctions de répartition homogènes

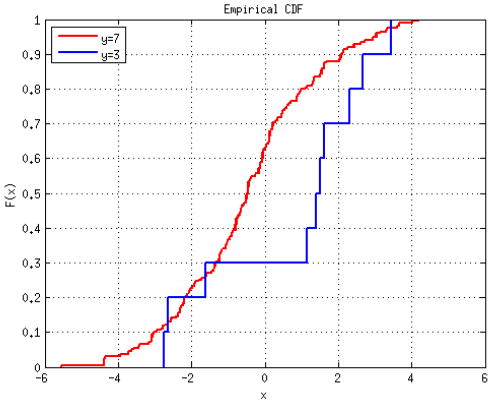


FIG. 34 – Fonctions de répartition non-homogènes

On peut en conclure que dans 8 cas les échantillons ne sont pas homogènes et dans 7 ils le sont. Plus de la moitié n'est pas homogène, donc l'ensemble des échantillons n'est pas homogène. Cependant, pour s'en assurer faisons un autre test.

3.2 Le test de la somme des rangs de Wilcoxon

3.2.1 Le principe

Le test de la somme des rangs de Wilcoxon est un test de comparaison ou d'homogénéité, il vérifie si les 2 échantillons viennent de la même distribution. En d'autres termes, il regarde si les échantillons ont la même loi, pour cela il se sert de la somme des rangs des échantillons. Ce test se déroule ainsi :

1. on pose les hypothèses H_0 (les échantillons suivent la même loi) et H_1 (le contraire) ;
2. on crée un nouvel échantillon Z en combinant nos deux échantillons, X et Y, Z sera rangé par ordre croissant ;
3. on fait la somme des rangs de l'échantillon X dans Z, que l'on note W_x ;
4. on calcule la $p-val$ grâce à $p-val = P(Wx)$;
5. on conclut, si la $p-val \geq 0.05$ alors on ne peut pas rejeter H_0, les deux échantillons suivent la même loi.

3.2.2 Le test sur nos données

Nous possédons plus de deux échantillons, donc comme pour le test de Kolmogorov-Smirnov, nous allons les comparer deux à deux.

1.On pose :

– H_0 : les deux échantillons suivent la même loi
– H_1 : les deux échantillons suivent des lois différentes.

2. Prenons l'exemple de $y = 3$ et $y = 8$.

```
X = -2.7507   -2.6339    2.6796   -1.6254    1.5030    3.4338    1.3947    1.6037
     1.1546    2.2967
Y = -0.6746   -2.0178    3.6346   -4.4105   -2.3663   -1.9261   -1.9453   -1.9453
     3.5323    1.9100   -0.4594   -1.5978    0.8207    0.1820    3.5313   -0.6558
     0.3734    0.5005
Z = -4.4105   -2.3663   -2.7507   -2.6339   -2.0178   -1.9453   -1.9453   -1.9261
    -1.6254   -1.5978   -0.6746   -0.6558   -0.4594    0.1820    0.3734    0.5005
     0.8207    1.1546    1.3947    1.5030    1.6037    1.9100    2.2967    2.6796
     3.4338    3.5313    3.5323    3.6346.
```

3. On calcule W_x :

$$W_x = \sum rg(X) = 164$$

4. On calcule les $p - val$:

$p - val$	y=4	y=5	y=6	y=7	y=8
y=3	0.9775	0.1246	0.1393	0.0722	0.3750
y=4	-	0.0001	0.0002	0.0000	0.0307
y=5	-	-	0.6686	0.0000	0.3249
y=6	-	-	-	0.0002	0.4436
y=7	-	-	-	-	0.7257

TAB. 7 – Tableau des $p - val$ pour le test de Wilcoxon

5. On peut en conclure les hypothèses :

H	y=4	y=5	y=6	y=7	y=8
y=3	0	0	0	0	0
y=4	-	1	1	1	1
y=5	-	-	0	1	0
y=6	-	-	-	1	0
y=7	-	-	-	-	0

TAB. 8 – Tableau des hypothèses pour le test de Wilcoxon

Il y a 5 cas où les variables ne suivent pas la même loi et 9 où elles suivent la même loi. On peut donc conclure que dans l'ensemble nos variables ne sont pas homogènes.

3.3 Le test du χ^2

3.3.1 Le principe

Le test du chi^2 est un test d'adéquation, il permet de déterminer si un ensemble est compatible. Ce test nécessite différentes étapes avant d'arriver à une conclusion :

1. on définit les hypothèses H_0 (les données sont indépendantes) et H_1 (les données ne sont pas indépendantes) ;
2. on construit le tableau de contingence des observations O (2 variables qualitatives de I et J modalités) ;
3. on calcule les marginales grâce à la formule $p_i = \frac{1}{n} \sum_{j=1}^{J} O_{ij}$;
4. on calcule le tableau de contingence théorique T en se basant sur les lois marginales ;
5. on calcule la distance du χ^2 avec la formule $D(O,T) = \sum_{i=1}^{I} \sum_{j=i}^{J} \frac{(0_{ij}-T_{ij})^2}{T_{ij}}$;

6. on calcule le degré de liberté grâce à $d = (I-1)(J-1)$
7. on détermine la $p-val$ grâce aux tables du chi^2 sachant que $p-val = P(X \geq D(O,T))$;
8. on compare la $p-val$ à notre risque de première espèce (le plus souvent 5%), si la $p-val \geq 0,05$ alors on ne peut pas rejeter l'hypothèse d'indépendance des données.

Ce test ne peut pas être utilisé avec n'importe quelle distribution, il faut que :
- les observations soient triées au hasard ;
- les observations soient indépendantes ;
- l'effectif des observations soit suffisamment grand ;
- l'effectif de chaque élément du tableau soit supérieur à 5.

3.3.2 Le test sur nos données

1. On pose :
- H_0 : les données sont indépendantes, elles ne suivent pas la même loi ;
- H_1 : les données suivent la même loi.

2. Grâce à nos histogrammes, on obtient un tableau le tableau de contingence suivant :

O	1	2	3	4	5	6	7	8	9	10	11
y=3	0	0	0	0	0	3	0	2	4	1	0
y=4	1	0	0	0	1	3	11	15	16	4	2
y=5	0	0	0	12	30	67	181	256	118	17	0
y=6	0	0	0	8	48	78	152	212	119	17	4
y=7	0	0	0	6	16	34	69	41	21	12	0
y=8	0	0	0	1	0	6	3	4	1	3	0

TAB. 9 – Le tableau de contingence de nos histogrammes

On remarque que ce tableau n'est pas conforme à l'utilisation du test du χ^2 car il y a des effectifs inférieurs à 5. Nous allons devoir fusionner des colonnes afin d'obtenir un nouveau tableau.

O	1 à 8	9 à 11
y=3	5	5
y=4	31	22
y=5	546	135
y=6	498	140
y=7	166	33
y=8	14	4

TAB. 10 – Le tableau nouveau de contingence

3. Nous allons effectuer le test sur ce tableau malgré le 4. Nous obtenons les lois marginales suivantes.

O	1 à 8	9 à 11	loi marginale
y=3	5	5	0.0063
y=4	31	22	0.0331
y=5	546	135	0.4259
y=6	498	140	0.3990
y=7	166	33	0.1245
y=8	14	4	0.0113
loi marginale	0.7880	0.2120	1

TAB. 11 – Le tableau nouveau de contingence avec les lois marginales

4. On calcule le tableau théorique grâce à la formule suivante :

$$p_{ij} = p_{.i}\, p_{.j}\, n$$

On obtient le tableau suivant :

T	1 à 8	9 à 11	loi marginale
y=3	7.8799	2.1201	0.0063
y=4	41.764	11.238	0.0331
y=5	536.62	144.38	0.4259
y=6	502.74	135.26	0.3990
y=7	156.81	42.186	0.1245
y=8	14.184	3.8161	0.0113
loi marginale	0.7880	0.2120	1

TAB. 12 – Le tableau théorique

5. On calcule la distance du χ^2 , on obtient 21.5844.
6. On calcule le degré de liberté :

$$d = (2-1)(6-1) = 5$$

7. On regarde dans la table la probabilité $P(X \geq 21.5844)$ avec un degré de 5 et on trouve que la $p - val = 6.2791$.
8. La $p - val$ est supérieure à 0.05, la probabilité d'obtenir une distance du χ^2 plus grande est importante, on peut donc conclure que nos données sont indépendantes et non-compatibles.

L'ensemble des tests nous a permis de montrer que nos données sont ni homogènes ni compatibles, elles ne sont donc pas linéaires. Pour traiter nos données, nous allons donc faire de la classification.

4 Arbre de décision sous Weka

Weka est un logiciel libre de **Data Mining**.

Le **Data Mining**, ou exploration des données, est un processus d'extraction de connaissances ou de savoirs à partir de grandes quantités de données grâce à des méthodes automatiques ou semi-automatiques. Contrairement à l'analyse de données et de statistiques, le data mining n'exige pas que l'on pose une hypothèse initiale qu'il faut ensuite vérifier. En effet, le logiciel de Data Mining déduit les corrélations intéressantes directement des données.

Weka permet notamment d'établir des arbres de décision. Un **arbre de décision** peut être défini comme un outil d'aide à la décision et à l'exploration des données. Il peut ainsi modéliser simplement, graphiquement et rapidement des phénomènes mesurés plus ou moins complexes.

Nous avons donc téléchargé ce logiciel à partir du site officiel de Weka :
`http://prdownloads.sourceforge.net/weka/weka-3-4-12jre.exe`

4.1 Conversion du fichier de données au format arff

Le format ARFF de Weka correspond à un fichier texte. Il est subdivisé en deux parties : la première correspond au dictionnaire de données, la seconde à la description des valeurs. Notre fichier étant au format csv, nous l'avons exporté au format arff grâce à un outil de conversion disponible sur internet au lien suivant :
`http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php`

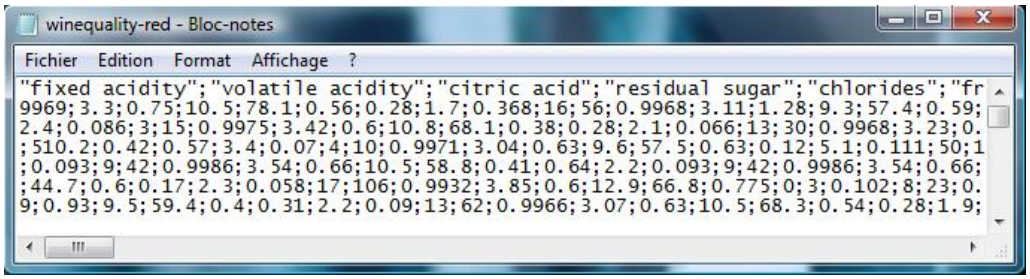


FIG. 35 – Aperçu du fichier au format csv

Lors de la conversion de nos données, nous devions choisir entre deux types pour chacune de nos variables : "numeric" et "nominal". Le type "numeric" désigne les réels, les entiers ou simplement des variables numériques. Ce type peut donc être utilisé pour des variables quantitatives continues telles que sulfates ou alcohol. Le type "nominal" désigne les variables discrètes ou qualitatives. Dans notre cas, seule la variable quality est de type nominal. Comme on peut le voir dans la figure ci-dessous, l'ensemble des valeurs que peut prendre la qualité est mis entre accolades.

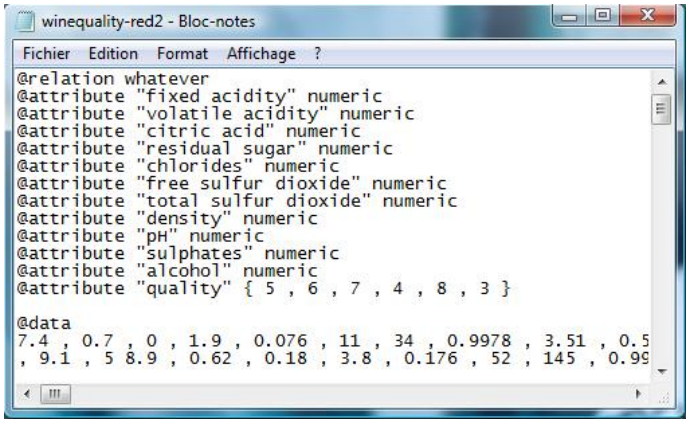


FIG. 36 – Aperçu du fichier au format arff

4.2 Visualisation des données

En ouvrant notre fichier avec le logiciel Weka, nous obtenons l'aperçu suivant :

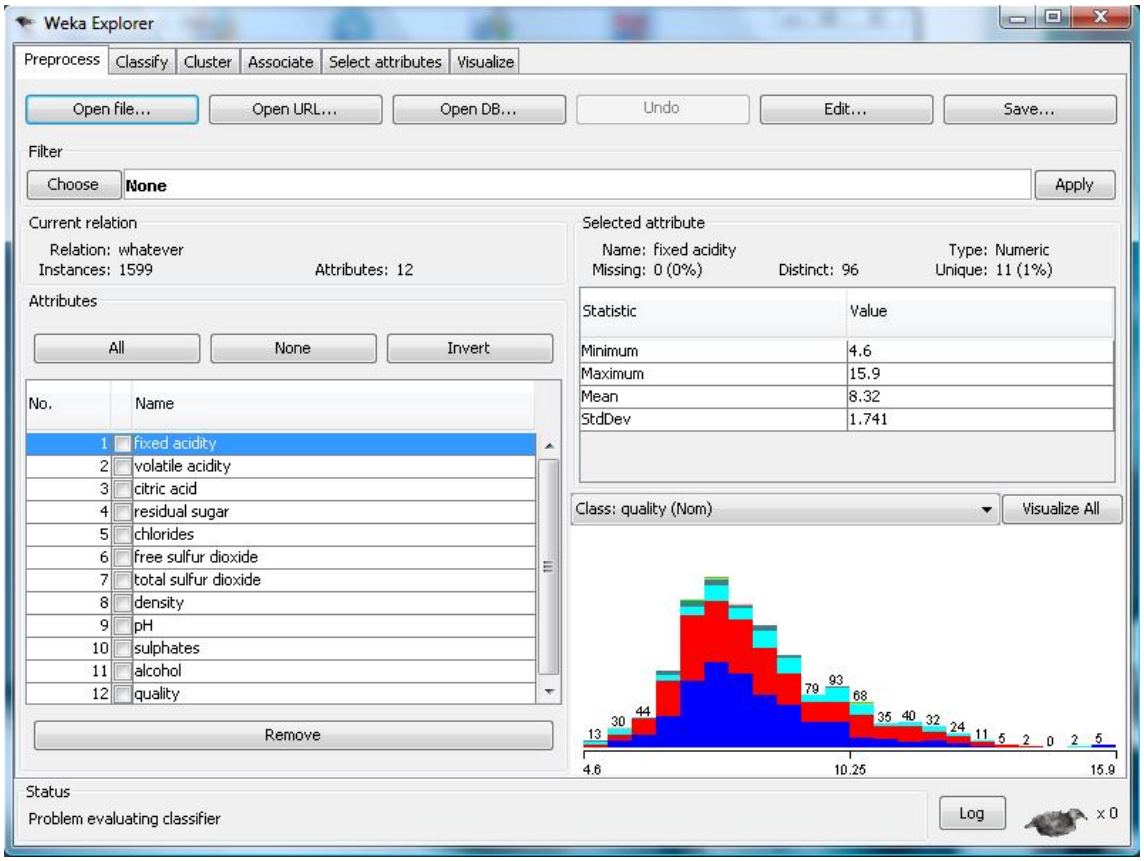


FIG. 37 – Onglet Preprocess

L'onglet Preprocess est composé de plusieurs parties.

Filter propose différents filtres modifiant les variables ou les données.

Current relation comporte 3 indications :
– Relation : nom du fichier utilisé

– Instances : nombre de données
– Attributes : nombre de variables.

Attributes énonce les variables figurant dans le fichier à traiter. De plus, on peut supprimer des données de notre analyse sans modifier notre fichier d'origine.

Selected Attribute donne les caractéristiques de la variable sélectionnée dans la partie Attributes : nom, type, nombre d'occurrences distinctes, minimum, maximum, moyenne et écart-type.

Cette partie comporte aussi un graphe composé de plusieurs couleurs. Chaque couleur représente l'effectif d'une valeur précise de la variable qualité. Pour savoir la signification de chaque couleur, il suffit de sélectionner la variable qualité dans la partie Attributes.

Dans la partie Selected Attributes, le graphe suivant s'affiche :

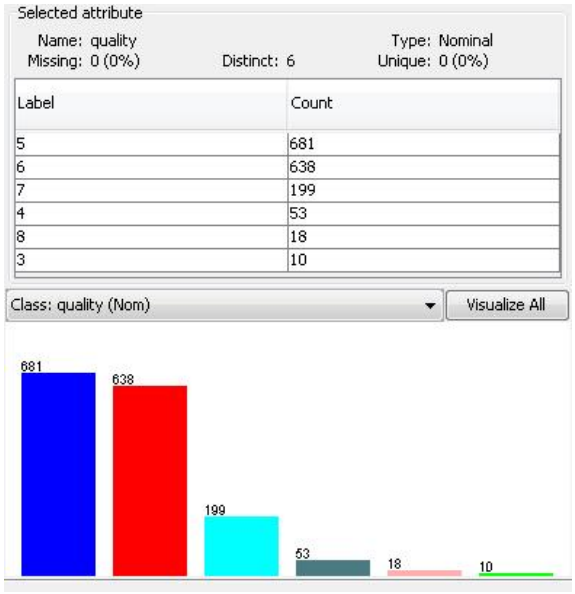


FIG. 38 – Graphe de la variable qualité

On peut ainsi savoir le nombre d'occurrences de chaque valeur, ainsi que la couleur associée : 3 vert, 4 gris, 5 bleu, 6 rouge, 7 cyan et 8 rose.

4.3 Etablissement de l'arbre de décision

Pour établir l'arbre de décision, on clique sur l'onglet **Classify**. Puis on choisit l'algorithme **J48** et l'option **Use training set**. Cette option utilise l'ensemble d'entraînement pour son évaluation. Après avoir appuyé sur **Start**, nous obtenons l'aperçu suivant :

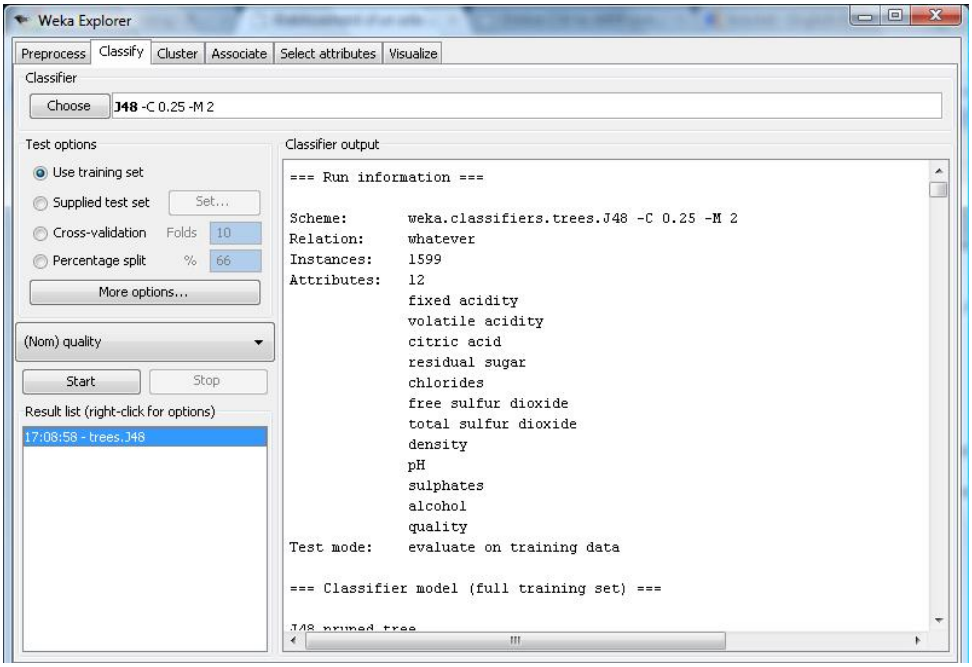


FIG. 39 – Onglet Classify

La partie **Classifier output** est composée des éléments suivants :

Classifier model qui présente l'arbre de décision obtenu (voir Annexe1) en précisant son nombre de branches et sa taille. Nous obtenons un arbre de taille 455 avec 228 branches.

Summary qui nous indique le nombre de données correctement classées. Comme on peut le voir dans l'aperçu ci-dessous, sur 1599 données, 1455 sont correctement classées. L'arbre de décision semble donc correspondre aux observations.

```
=== Summary ===

Correctly Classified Instances        1455               90.9944 %
Incorrectly Classified Instances       144                9.0056 %
Kappa statistic                          0.8584
Mean absolute error                      0.0465
Root mean squared error                  0.1525
Relative absolute error                 21.6818 %
Root relative squared error             46.5894 %
Total Number of Instances             1599
```

Confusion Matrix est la matrice de confusion, c'est-à-dire un outil servant à mesurer la qualité d'un système de classification. Chaque colonne de la matrice représente le nombre d'occurrences d'une classe estimée, tandis que chaque ligne représente le nombre d'occurrences d'une classe réelle. Dans l'aperçu ci-dessous, on peut remarquer que la majorité des variables mal classées n'est pas très éloignée de la réalité.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 634  43   1   3   0   0 |   a = 5
  29 598   7   2   2   0 |   b = 6
   7  16 173   1   2   0 |   c = 7
   7  10   0  35   0   1 |   d = 4
   1   2   3   0  12   0 |   e = 8
   2   1   2   2   0   3 |   f = 3
```

Avec 91 % de données correctement classées et un faible écart entre l'estimation et la réalité, nous avons conclu que la qualité de notre modèle est très bonne.

Nous avons ensuite testé notre modèle en prenant les valeurs prises par une de nos données.

	A	B	C	D	E	F	G	H	I	J	K	L
1	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5

FIG. 40 – Valeurs de la première donnée

```
alcohol <= 10.5
|   total sulfur dioxide <= 81
|   |   sulphates <= 0.57
|   |   |   alcohol <= 9.7
|   |   |   |   alcohol > 9.05
|   |   |   |   |   alcohol <= 9.55
|   |   |   |   |   |   alcohol > 9.3
|   |   |   |   |   |   |   sulphates > 0.52
|   |   |   |   |   |   |   |   chlorides <= 0.079
|   |   |   |   |   |   |   |   |   alcohol <= 9.4
|   |   |   |   |   |   |   |   |   |   volatile acidity > 0.52: 5 (6.0)
```

Grâce au modèle nous obtenons une qualité de 5 qui correspond à la qualité réelle de l'observation.

Cependant, l'arbre de décision obtenu est de taille importante. Une des méthodes pour le réduire serait donc d'utiliser un filtre sur nos variables ou sur nos données.

4.4 Utilisation d'un filtre

Précédemment, nous avons pu remarquer que l'onglet **Preprocess** mettait en évidence les caractéristiques principales de chaque variable. Cependant, Preprocess a une autre fonctionnalité : celle d'effectuer des traitements préliminaires sur les données ou sur les variables grâces au filtres.

En effet, les filtres proposés transforment l'ensemble de données en enlevant ou ajoutant des variables, en rééchantillonnant l'ensemble de données, en supprimant des cas... Ces filtres sont organisés entre les filtres **supervised** et ceux **unsupervised**. Les filtres **supervised** prennent en compte les informations de l'ensemble des données. Ces deux types de filtres sont ensuite divisés entre les filtres concernant les variables et ceux concernant les données.

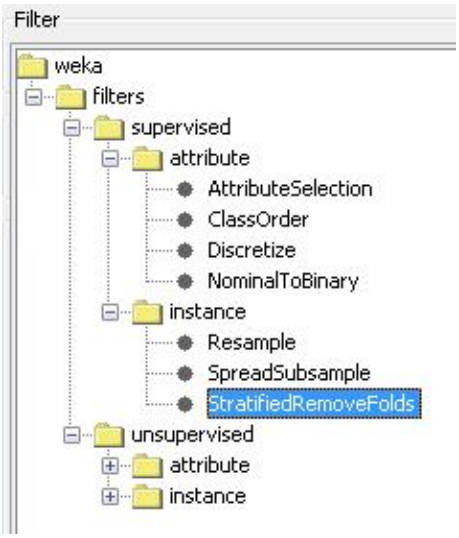


FIG. 41 – Filtres proposés

4.4.1 Filtres sur les variables

Il existe diverses possibilités de filtres agissant sur les variables : certains filtres les discrétisent d'autres les normalisent. Nous avons choisi d'utiliser le filtre **AttributeSelection**. Ce dernier sélectionne les variables les plus importantes et enlève les autres.

Après l'avoir appliqué à notre ensemble de données, il nous reste les variables suivantes :
- volatile acidity
- total sulfur acidity
- sulfates
- alcohol
- quality

Nous avons ensuite établit un nouvel arbre de décision en reprenant la même méthode qu'auparavant.

Nous obtenons alors un arbre décision de taille 361 avec 181 feuillages, ainsi que des informations sur la qualité de notre nouveau modèle.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1294               80.9256 %
Incorrectly Classified Instances       305               19.0744 %
Kappa statistic                          0.6973
Mean absolute error                      0.0918
Root mean squared error                  0.2142
Relative absolute error                 42.7914 %
Root relative squared error             65.4513 %
Total Number of Instances             1599

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 605  64   8   3   1   0 |   a = 5
  93 513  29   1   2   0 |   b = 6
  12  45 142   0   0   0 |   c = 7
```

```
 19  10   2  22   0   0 |   d = 4
  2   3   3   0  10   0 |   e = 8
  4   2   1   1   0   2 |   f = 3
```

En observant les données précédentes, on peut remarquer que la qualité de notre modèle a baissé. En effet, le nombre de données mal classées a augmenté de 10 % par rapport à celui du modèle antérieur. La taille de notre arbre n'ayant que très faiblement diminué, le filtre utilisé ne semble pas adapté à notre cas.

Nous avons repris notre fichier de données initial pour lui appliqué un autre de type de filtres.

4.4.2 Filtre sur les données

Afin de réduire la taille de notre modèle, nous avons choisi de diminuer le nombre de nos données. Nous avions donc le choix entre le filtre **StratifiedRemoveFolds** de la catégorie supervised et le filtre **RemoveFolds** de la catégorie unsupervised.

Ces deux filtres réduisent la taille de notre échantillon à 160 données. Toutefois, en établissant l'arbre de décision et en observant les caractéristiques des variables, nous avons pu discerner des différences .

Avec le filtre **StratifiedRemoveFolds**, nous obtenons un arbre de taille 49 avec 25 feuillages ainsi que les informations sur la qualité de notre modèle :

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         139               86.875  %
Incorrectly Classified Instances        21               13.125  %
Kappa statistic                          0.7943
Mean absolute error                      0.0661
Root mean squared error                  0.1818
Relative absolute error                 30.5814 %
Root relative squared error             55.5869 %
Total Number of Instances              160

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 64  4  0  1  0  0 |  a = 5
  8 53  2  0  0  0 |  b = 6
  1  2 17  0  0  0 |  c = 7
  1  0  0  5  0  0 |  d = 4
  0  0  1  0  0  0 |  e = 8
  0  0  0  1  0  0 |  f = 3
```

Avec le filtre **RemoveFolds**, nous obtenons un arbre de taille 49 avec 25 feuillages ainsi que les informations sur la qualité de notre modèle :

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         148               92.5    %
Incorrectly Classified Instances        12                7.5    %
Kappa statistic                          0.8464
Mean absolute error                      0.0396
Root mean squared error                  0.1407
```

```
Relative absolute error                 23.1968 %
Root relative squared error             48.7193 %
Total Number of Instances              160

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 102   2   1   0   0   0 |   a = 5
   3  38   0   0   0   0 |   b = 6
   1   1   4   0   0   0 |   c = 7
   1   3   0   4   0   0 |   d = 4
   0   0   0   0   0   0 |   e = 8
   0   0   0   0   0   0 |   f = 3
```

Avec 92.5 % de données correctement classées, la qualité de notre second modèle semble supérieure à celle du premier. Cependant, il serait judicieux de se poser la question suivante : lors de la réduction de notre échantillon n'a-t-on pas perdu des informations sur notre ensemble de données ?

Nous avons donc comparé les histogrammes de notre échantillon de 1599 données avec ceux de nos échantillons réduits.

Prenons d'abord l'exemple de la répartition de la variable qualité.

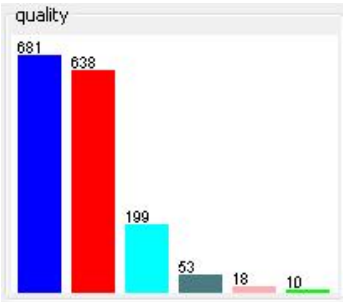


FIG. 42 – Histogramme initial

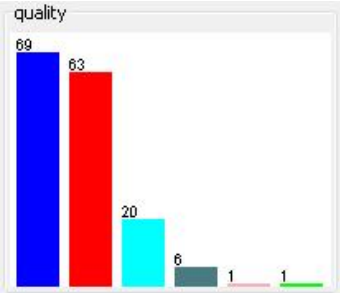


FIG. 43 – StratifiedRemoveFolds

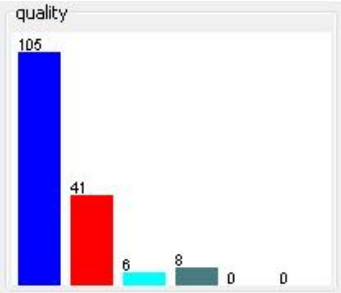


FIG. 44 – RemoveFolds

Lorsqu'on observe les histogrammes précédents, on remarque facilement que la ressemblance entre l'échantillon initial et celui réduit grâce au filtre Stratified RemoveFolds.

Cette impression peut être confirmée en calculant la répartition des différentes valeurs de la qualité :

	A	B	C	D	E	F	G
1	**qualité**	**5**	**6**	**7**	**4**	**8**	**3**
2	**Echantillon initial**	42,6%	39,9%	12,4%	3,3%	1,1%	0,6%
3	**StratifiedRemoveFolds**	43,1%	39,4%	12,5%	3,8%	0,6%	0,6%
4	**RemoveFolds**	65,6%	25,6%	3,8%	5,0%	0,0%	0,0%

FIG. 45 – Répartition des valeurs que prend la qualité

Bien que la qualité du second modèle réduit soit supérieure à celle du premier, le premier échantillon diminué est beaucoup plus similaire à notre échantillon initial. En le choisissant, nous pouvons ainsi réduire la taille de notre arbre tout en gardant le maximum d'informations sur notre échantillon.

En modifiant les paramètres du filtre StratifiedRemoveFolds, nous pouvons réduire de nouveau la taille de notre échantillon de 2 et obtenir l'arbre suivant :

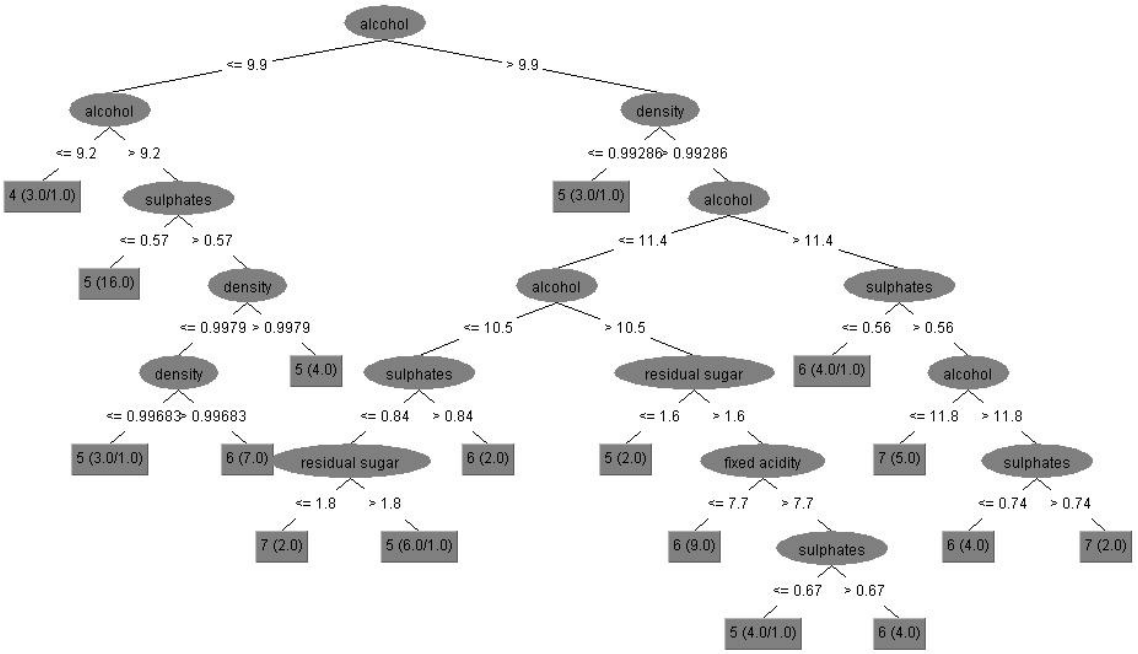


FIG. 46 – Arbre final

5 SVM : Des méthodes à noyaux

Introduite par Vladimir Vapnik en 1995, **SVM**, Machine à Vecteurs de Support ou Séparateur à Vastes Marges, désigne la forme la plus connue des méthodes à noyaux. Puisqu'elle est un problème de classification à deux classes et qu'elle fait appel à un jeu de données d'apprentissage supervisé, SVM peut être définie comme une méthode de classification binaire par apprentissage supervisé. Elle repose sur l'utilisation de fonctions de Kernel (noyau) qui permettent une séparation optimale des données.

L'**apprentissage automatique** un des sous-domaines de l'intelligence artificielle a pour objectif d'extraire et d'exploiter automatiquement l'information présente dans un jeu de données. Il existe deux types d'apprentissage automatique : l'apprentissage supervisé et celui non-supervisé. L'**apprentissage supervisé** cherche à produire automatiquement des règles à partir d'une base de données d'apprentissage contenant des « exemples » , le plus souvent des cas déjà traités et validés. Les exemples déjà traités sont représentés par un ensemble de couples d'entrée/sortie. Le but est d'apprendre une fonction qui correspond aux exemples vus et qui prédit les sorties pour les entrées qui n'ont pas encore été vues. Les sorties sont appelées les classes d'objets donnés en entrée.

5.1 Quelques principes

L'objectif de SVM est de trouver un classificateur qui va séparer les données et maximiser la distance entre deux classes d'exemple donnés. Ce classificateur linéaire est appelé **hyperplan**.

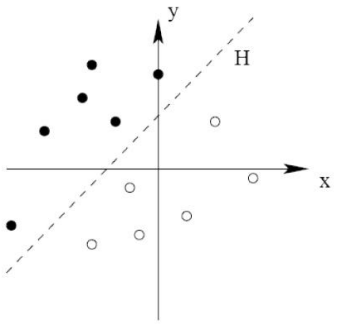


FIG. 47 – Hyperplan H

Les échantillons les plus proches de l'hyperplan sont appelés **vecteurs de support**. Ceux sont eux qui déterminent l'hyperplan.

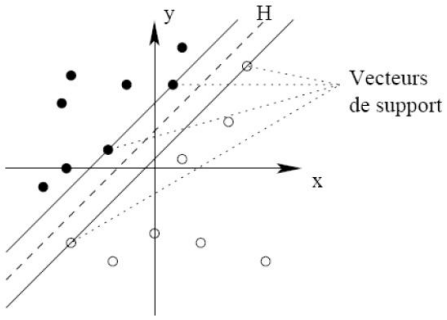


FIG. 48 – Vecteurs de support

Parmi la multitude d'hyperplans possibles, SVM recherche l'hyperplan optimal. La distance entre les observations et l'hyperplan est appelée **marge**. Pour qu'un hyperplan soit optimal, il faut

que la marge minimale aux observations soit maximale. Ce critère de sélection explique l'appellation Séparateur à Vastes Marges pour SVM.

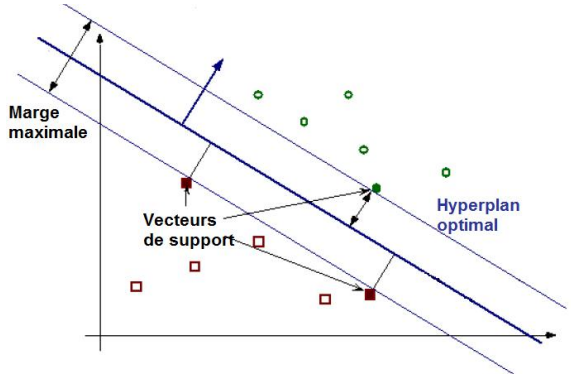


FIG. 49 – Hyperplan optimal

Avoir une marge maximale permet d'avoir plus de sécurité lorsqu'on rajoute une nouvelle observation. La classification d'un nouvel exemple est donné par sa position par rapport à l'hyperplan. Dans l'exemple ci-dessous, les observations à gauche de l'hyperplan sont classées dans la catégorie des "+", celles à droite dans la catégorie des "o". On rajoute alors une nouvelle observation. On remarque alors qu'avec une marge faible l'exemple est mal classé (du côté des croix alors que c'est un rond). Au contraire avec une marge maximale, comme dans le cas gauche, l'exemple est bien classé.

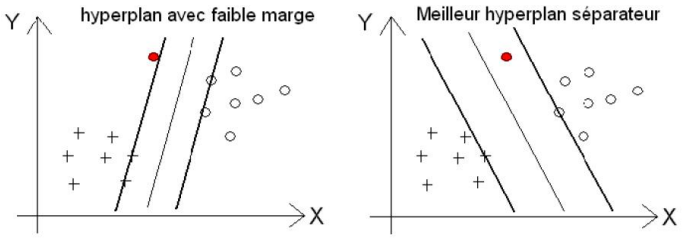


FIG. 50 – Ajout d'une observation pour deux types de marges

Les modèles de SVM sont divisés entre les cas linéairement séparable et ceux non linéairement séparable. Pour les premiers, il est facile de trouver un classificateur linéaire. Cependant, les seconds sont les plus fréquents. L'hyperplan à marge maximale ne peut pas être utilisé directement dans les cas non linéairement séparable.

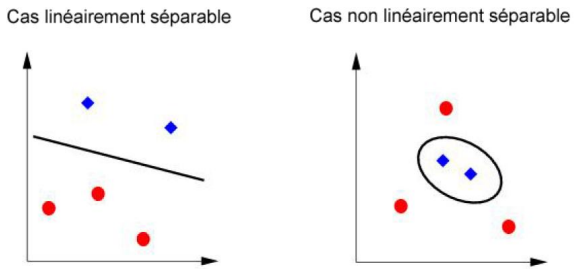


FIG. 51 – Cas linéairement séparable et cas non linéairement séparable

Pour résoudre ces derniers, SVM transforme l'espace de données en un espace où les classes seront linéairement séparables. La dimension de ce nouvel espace est plus grande. En effet, plus la dimension de cet espace, appelé "espace de re-description" est grande, plus la probabilité de pouvoir

trouver un hyperplan séparateur entre les exemples est élevée. On peut le voir dans l'exemple suivant.

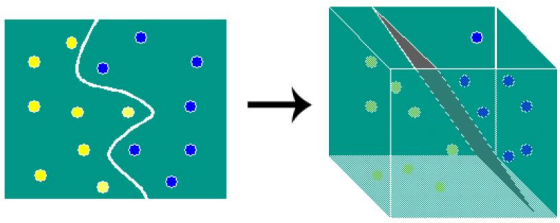


FIG. 52 – Transformation de l'espace initial en un espace de re-description

Cette transformation non-linéaire s'effectue grâce à une fonction Kernel (noyau) dont voici quelques exemples : polynomiale, gaussienne, sigmoïde et laplacienne.

5.2 Problème du Multi-classe

Les SVM sont bien adaptés aux problèmes binaires. Toutefois, dans notre cas, la qualité possède plus de 2 valeurs. Il faut donc des méthodes qui étendent l'utilisation des SVM à plus de deux classes.

One versus Rest : cette méthode consiste à construire autant de classifieurs que de classes. Chaque classifieur renvoie 1 si la forme à reconnaître appartient à la classer, -1 sinon. Pour reconnaître une forme il faut donc la soumettre à tous les classifieurs, la plus grande valeur étant alors retenue. Pour N classes, il faut donc construire N classifieurs et pour chaque décision il faut effectuer N comparaisons.

One versus One : cette méthode consiste à concevoir des classifieurs spécialisés dans la comparaison classe à classe. Pour N classes il faut donc N(N-1)/2 classifieurs.On soumet la forme à reconnaitre à tous ces classifieurs 1-vs-1, la classe remportant le plus de suffrage remporte la décision. Le gros inconvénient de cette méthode est alors évident : si N augmente, sa complexité augmentera très rapidement.

Afin de réaliser des SVM dans le cas multi-classe sur Weka, un logiciel LIBSVM est disponible sur internet, fournissant les classes nécessaires en Java.

Conclusion

Ce projet a été pour nous très avantageux : nous avons vraiment appliqué les méthodes vues en cours sur des échantillons à grand nombre de modalités. Nous avons également utilisé de nouvelles méthodes, ce qui nous a permis de comprendre de nouvelles choses, de s'intéresser à des tests toujours nouveaux. C'est un plus pour notre culture générale, et notre culture d'ingénieur. De plus, nous avons découvert le logiciel Weka, qui est extrêmement utile dans la pratique.

Ce projet nous a donné une idée significative de comment seront les projets dans nos futures années d'études. Ce projet n'a pas été simple à réaliser. Cependant, nous avons pris beaucoup de plaisir à le mener et notre forte implication nous a permis de surmonter les difficultés.

Nous avons essayé différentes régressions (toutes les variables, seulement quelques unes, avec une ACP) mais nous n'avons obtenu aucun résultat satisfaisant. Nous avons donc émis l'hypothèse que nos données n'étaient pas linéaires.

Pour le prouver nous avons fait différents tests (χ^2, Wilcoxon, Kolmogorov-Smirnov, tests linéaires), nous pouvons émettre une conclusion réellement vérifiée : nos données ne sont ni linéaires, ni homogènes, ni compatibles.

Nous avons ensuite travaillé sur d'autres méthodes d'estimation de la qualité du vin. Sous le logiciel Weka, nous avons établi un arbre de décision qui, grâce aux caractéristiques du vin, nous donnait une bonne prédiction de sa qualité.

Bien que n'ayant pas de résultats concrets, nous avons pris aussi conscience du fort potentiel de la méthode SVM.

Afin d'établir un modèle linéaire prédisant la qualité d'un vin, il nous reste la question de la façon de fabriquer son vin, ou de la qualité du raisin. Il serait intéressant d'étudier des données sur la qualité du raisin, tout comme nous venons de le faire pour le vin.

A Arbre de décision avec toutes les variables

```
alcohol <= 10.5
|   total sulfur dioxide <= 81
|   |   sulphates <= 0.57
|   |   |   alcohol <= 9.7
|   |   |   |   alcohol <= 9.05
|   |   |   |   |   fixed acidity <= 7.8: 4 (4.0/1.0)
|   |   |   |   |   fixed acidity > 7.8
|   |   |   |   |   |   citric acid <= 0.55: 6 (5.0)
|   |   |   |   |   |   citric acid > 0.55: 5 (2.0/1.0)
|   |   |   |   alcohol > 9.05
|   |   |   |   |   alcohol <= 9.55
|   |   |   |   |   |   alcohol <= 9.3
|   |   |   |   |   |   |   chlorides <= 0.062: 4 (3.0/1.0)
|   |   |   |   |   |   |   chlorides > 0.062
|   |   |   |   |   |   |   |   fixed acidity <= 8.2: 5 (31.0/2.0)
|   |   |   |   |   |   |   |   fixed acidity > 8.2
|   |   |   |   |   |   |   |   |   volatile acidity <= 0.605: 5 (4.0)
|   |   |   |   |   |   |   |   |   volatile acidity > 0.605: 4 (2.0)
|   |   |   |   |   |   alcohol > 9.3
|   |   |   |   |   |   |   sulphates <= 0.52: 5 (44.0/2.0)
|   |   |   |   |   |   |   sulphates > 0.52
|   |   |   |   |   |   |   |   chlorides <= 0.079
|   |   |   |   |   |   |   |   |   alcohol <= 9.4
|   |   |   |   |   |   |   |   |   |   volatile acidity <= 0.52: 6 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   volatile acidity > 0.52: 5 (6.0)
|   |   |   |   |   |   |   |   |   alcohol > 9.4: 6 (7.0/1.0)
|   |   |   |   |   |   |   |   chlorides > 0.079
|   |   |   |   |   |   |   |   |   density <= 0.9988: 5 (19.0)
|   |   |   |   |   |   |   |   |   density > 0.9988: 6 (3.0/1.0)
|   |   |   |   |   alcohol > 9.55
|   |   |   |   |   |   residual sugar <= 4.3
|   |   |   |   |   |   |   pH <= 3.28: 5 (21.0/2.0)
|   |   |   |   |   |   |   pH > 3.28
|   |   |   |   |   |   |   |   pH <= 3.34
|   |   |   |   |   |   |   |   |   free sulfur dioxide <= 15: 4 (4.0)
|   |   |   |   |   |   |   |   |   free sulfur dioxide > 15: 5 (3.0/2.0)
|   |   |   |   |   |   |   |   pH > 3.34
|   |   |   |   |   |   |   |   |   sulphates <= 0.46: 4 (3.0/1.0)
|   |   |   |   |   |   |   |   |   sulphates > 0.46: 5 (11.0)
|   |   |   |   |   |   residual sugar > 4.3: 6 (2.0/1.0)
|   |   |   alcohol > 9.7
|   |   |   |   residual sugar <= 1.2: 7 (2.0/1.0)
|   |   |   |   residual sugar > 1.2
|   |   |   |   |   sulphates <= 0.47
|   |   |   |   |   |   sulphates <= 0.42: 6 (2.0/1.0)
|   |   |   |   |   |   sulphates > 0.42: 5 (19.0/1.0)
|   |   |   |   |   sulphates > 0.47
|   |   |   |   |   |   free sulfur dioxide <= 3: 4 (2.0/1.0)
|   |   |   |   |   |   free sulfur dioxide > 3
|   |   |   |   |   |   |   citric acid <= 0.01
|   |   |   |   |   |   |   |   citric acid <= 0: 5 (5.0/1.0)
|   |   |   |   |   |   |   |   citric acid > 0: 7 (2.0)
|   |   |   |   |   |   |   citric acid > 0.01
|   |   |   |   |   |   |   |   chlorides <= 0.102
```

```
|   |   |   |   |   |   |   |   |   volatile acidity <= 0.575: 6 (45.0/14.0)
|   |   |   |   |   |   |   |   |   volatile acidity > 0.575
|   |   |   |   |   |   |   |   |   |   total sulfur dioxide <= 28: 5 (10.0)
|   |   |   |   |   |   |   |   |   |   total sulfur dioxide > 28
|   |   |   |   |   |   |   |   |   |   |   alcohol <= 10.1
|   |   |   |   |   |   |   |   |   |   |   |   residual sugar <= 2.05: 5 (6.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   residual sugar > 2.05
|   |   |   |   |   |   |   |   |   |   |   |   |   pH <= 3.35
|   |   |   |   |   |   |   |   |   |   |   |   |   |   fixed acidity <= 9.7
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   fixed acidity <= 7.8: 5 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   fixed acidity > 7.8: 6 (7.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   fixed acidity > 9.7: 5 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   pH > 3.35: 5 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   alcohol > 10.1: 6 (3.0/1.0)
|   |   |   |   |   |   |   |   |   chlorides > 0.102
|   |   |   |   |   |   |   |   |   |   citric acid <= 0.24: 4 (2.0)
|   |   |   |   |   |   |   |   |   |   citric acid > 0.24: 5 (2.0)
|   |   sulphates > 0.57
|   |   |   fixed acidity <= 11.4
|   |   |   |   sulphates <= 1.08
|   |   |   |   |   pH <= 3.53
|   |   |   |   |   |   alcohol <= 9.8
|   |   |   |   |   |   |   chlorides <= 0.168
|   |   |   |   |   |   |   |   volatile acidity <= 0.55
|   |   |   |   |   |   |   |   |   volatile acidity <= 0.26: 6 (8.0)
|   |   |   |   |   |   |   |   |   volatile acidity > 0.26
|   |   |   |   |   |   |   |   |   |   pH <= 3.39
|   |   |   |   |   |   |   |   |   |   |   alcohol <= 9.3
|   |   |   |   |   |   |   |   |   |   |   |   fixed acidity <= 7.9: 5 (7.0)
|   |   |   |   |   |   |   |   |   |   |   |   fixed acidity > 7.9
|   |   |   |   |   |   |   |   |   |   |   |   |   citric acid <= 0.29: 6 (9.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   citric acid > 0.29
|   |   |   |   |   |   |   |   |   |   |   |   |   |   citric acid <= 0.4: 5 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   citric acid > 0.4
|   |   |   |   |   |   |   |   |   |   |   |   |   |   chlorides <= 0.086: 6 (7.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   chlorides > 0.086: 5 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   alcohol > 9.3
|   |   |   |   |   |   |   |   |   |   |   |   alcohol <= 9.6
|   |   |   |   |   |   |   |   |   |   |   |   |   density <= 0.9984
|   |   |   |   |   |   |   |   |   |   |   |   |   |   pH <= 3.19: 5 (6.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   pH > 3.19
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   sulphates <= 0.63: 5 (7.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   sulphates > 0.63
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   pH <= 3.34: 6 (9.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   pH > 3.34
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   residual sugar <= 2: 5 (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   residual sugar > 2: 6 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   density > 0.9984: 5 (12.0)
|   |   |   |   |   |   |   |   |   |   |   alcohol > 9.6
|   |   |   |   |   |   |   |   |   |   |   |   fixed acidity <= 9.9
|   |   |   |   |   |   |   |   |   |   |   |   |   pH <= 3.34
|   |   |   |   |   |   |   |   |   |   |   |   |   |   alcohol <= 9.7
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   pH <= 3.26: 7 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   pH > 3.26: 5 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   alcohol > 9.7: 5 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   pH > 3.34: 6 (2.0)
```

```
|   |   |   |   |   |   |   |   |   |   |   |   |   fixed acidity > 9.9: 6 (5.0)
|   |   |   |   |   |   |   |   |   |   |   pH > 3.39
|   |   |   |   |   |   |   |   |   |   |   |   citric acid <= 0.33
|   |   |   |   |   |   |   |   |   |   |   |   |   fixed acidity <= 7.4: 6 (12.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   fixed acidity > 7.4
|   |   |   |   |   |   |   |   |   |   |   |   |   |   residual sugar <= 2.05: 6 (9.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   residual sugar > 2.05: 5 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   citric acid > 0.33: 5 (4.0)
|   |   |   |   |   |   |   |   volatile acidity > 0.55
|   |   |   |   |   |   |   |   |   pH <= 3.49
|   |   |   |   |   |   |   |   |   |   free sulfur dioxide <= 15
|   |   |   |   |   |   |   |   |   |   |   fixed acidity <= 6.9: 5 (7.0)
|   |   |   |   |   |   |   |   |   |   |   fixed acidity > 6.9
|   |   |   |   |   |   |   |   |   |   |   |   citric acid <= 0.13
|   |   |   |   |   |   |   |   |   |   |   |   |   density <= 0.99631
|   |   |   |   |   |   |   |   |   |   |   |   |   |   free sulfur dioxide <= 10: 5 (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   free sulfur dioxide > 10: 6 (2.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   density > 0.99631
|   |   |   |   |   |   |   |   |   |   |   |   |   |   fixed acidity <= 7.3: 6 (10.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   fixed acidity > 7.3
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chlorides <= 0.08: 5 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chlorides > 0.08: 6 (12.0/3.0)
|   |   |   |   |   |   |   |   |   |   |   |   citric acid > 0.13
|   |   |   |   |   |   |   |   |   |   |   |   |   citric acid <= 0.26: 5 (19.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   citric acid > 0.26: 6 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   free sulfur dioxide > 15
|   |   |   |   |   |   |   |   |   |   |   |   alcohol <= 9.55
|   |   |   |   |   |   |   |   |   |   |   |   |   sulphates <= 0.66: 5 (7.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   sulphates > 0.66: 6 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   alcohol > 9.55: 5 (16.0)
|   |   |   |   |   |   |   |   |   |   pH > 3.49: 6 (4.0)
|   |   |   |   |   |   |   |   chlorides > 0.168: 6 (13.0/2.0)
|   |   |   |   |   |   |   alcohol > 9.8
|   |   |   |   |   |   |   |   residual sugar <= 1.65
|   |   |   |   |   |   |   |   |   sulphates <= 0.8: 6 (9.0/1.0)
|   |   |   |   |   |   |   |   |   sulphates > 0.8: 7 (2.0/1.0)
|   |   |   |   |   |   |   |   residual sugar > 1.65
|   |   |   |   |   |   |   |   |   chlorides <= 0.063
|   |   |   |   |   |   |   |   |   |   fixed acidity <= 7.2: 6 (8.0)
|   |   |   |   |   |   |   |   |   |   fixed acidity > 7.2
|   |   |   |   |   |   |   |   |   |   |   residual sugar <= 2: 6 (2.0)
|   |   |   |   |   |   |   |   |   |   |   residual sugar > 2: 7 (5.0/1.0)
|   |   |   |   |   |   |   |   |   chlorides > 0.063
|   |   |   |   |   |   |   |   |   |   sulphates <= 0.68
|   |   |   |   |   |   |   |   |   |   |   chlorides <= 0.072: 5 (8.0)
|   |   |   |   |   |   |   |   |   |   |   chlorides > 0.072
|   |   |   |   |   |   |   |   |   |   |   |   total sulfur dioxide <= 61
|   |   |   |   |   |   |   |   |   |   |   |   |   free sulfur dioxide <= 28
|   |   |   |   |   |   |   |   |   |   |   |   |   |   density <= 0.996: 6 (9.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   density > 0.996
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chlorides <= 0.082
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   citric acid <= 0.07: 6 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   citric acid > 0.07
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   density <= 0.99765: 5 (12.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   density > 0.99765
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   volatile acidity <= 0.47: 5 (2
```

```
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   volatile acidity > 0.47: 6 (2.
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chlorides > 0.082
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   total sulfur dioxide <= 33
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chlorides <= 0.084: 6 (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chlorides > 0.084
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   total sulfur dioxide <= 19: 6
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   total sulfur dioxide > 19: 5 (
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   total sulfur dioxide > 33: 6 (15.0/4.0
|   |   |   |   |   |   |   |   |   |   |   |   |   free sulfur dioxide > 28: 5 (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   total sulfur dioxide > 61: 6 (11.0)
|   |   |   |   |   |   |   |   |   |   sulphates > 0.68
|   |   |   |   |   |   |   |   |   |   |   residual sugar <= 3.4
|   |   |   |   |   |   |   |   |   |   |   |   chlorides <= 0.109: 6 (33.0/7.0)
|   |   |   |   |   |   |   |   |   |   |   |   chlorides > 0.109: 5 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   residual sugar > 3.4
|   |   |   |   |   |   |   |   |   |   |   |   total sulfur dioxide <= 52: 7 (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   total sulfur dioxide > 52: 5 (3.0)
|   |   |   |   |   pH > 3.53
|   |   |   |   |   |   total sulfur dioxide <= 15: 4 (3.0/1.0)
|   |   |   |   |   |   total sulfur dioxide > 15
|   |   |   |   |   |   |   sulphates <= 0.9
|   |   |   |   |   |   |   |   volatile acidity <= 0.53
|   |   |   |   |   |   |   |   |   residual sugar <= 2.05: 6 (3.0)
|   |   |   |   |   |   |   |   |   residual sugar > 2.05: 5 (6.0/1.0)
|   |   |   |   |   |   |   |   volatile acidity > 0.53: 5 (20.0/1.0)
|   |   |   |   |   |   |   sulphates > 0.9: 6 (4.0)
|   |   |   |   |   sulphates > 1.08
|   |   |   |   |   |   free sulfur dioxide <= 23: 5 (18.0/2.0)
|   |   |   |   |   |   free sulfur dioxide > 23: 6 (3.0/1.0)
|   |   |   |   fixed acidity > 11.4
|   |   |   |   |   sulphates <= 0.64
|   |   |   |   |   |   citric acid <= 0.49: 5 (5.0)
|   |   |   |   |   |   citric acid > 0.49: 6 (5.0/1.0)
|   |   |   |   |   sulphates > 0.64
|   |   |   |   |   |   free sulfur dioxide <= 29
|   |   |   |   |   |   |   pH <= 3.04
|   |   |   |   |   |   |   |   alcohol <= 9.8: 6 (5.0/1.0)
|   |   |   |   |   |   |   |   alcohol > 9.8
|   |   |   |   |   |   |   |   |   fixed acidity <= 12: 7 (4.0)
|   |   |   |   |   |   |   |   |   fixed acidity > 12: 5 (2.0)
|   |   |   |   |   |   |   pH > 3.04
|   |   |   |   |   |   |   |   residual sugar <= 4.8
|   |   |   |   |   |   |   |   |   fixed acidity <= 14.3: 6 (23.0/3.0)
|   |   |   |   |   |   |   |   |   fixed acidity > 14.3: 7 (2.0)
|   |   |   |   |   |   |   |   residual sugar > 4.8: 7 (2.0)
|   |   |   |   |   free sulfur dioxide > 29: 5 (2.0/1.0)
|   total sulfur dioxide > 81
|   |   residual sugar <= 1.65
|   |   |   volatile acidity <= 0.56: 5 (4.0)
|   |   |   volatile acidity > 0.56: 4 (2.0)
|   |   residual sugar > 1.65
|   |   |   total sulfur dioxide <= 98
|   |   |   |   free sulfur dioxide <= 15: 5 (15.0)
|   |   |   |   free sulfur dioxide > 15
|   |   |   |   |   total sulfur dioxide <= 85
|   |   |   |   |   |   chlorides <= 0.073: 6 (2.0/1.0)
```

```
|   |   |   |   |   |   chlorides > 0.073: 5 (15.0/1.0)
|   |   |   |   |   total sulfur dioxide > 85
|   |   |   |   |   |   chlorides <= 0.083
|   |   |   |   |   |   |   pH <= 3.29: 6 (8.0)
|   |   |   |   |   |   |   pH > 3.29
|   |   |   |   |   |   |   |   chlorides <= 0.078: 5 (10.0/1.0)
|   |   |   |   |   |   |   |   chlorides > 0.078
|   |   |   |   |   |   |   |   |   alcohol <= 9.9: 6 (7.0)
|   |   |   |   |   |   |   |   |   alcohol > 9.9: 5 (2.0)
|   |   |   |   |   |   chlorides > 0.083
|   |   |   |   |   |   |   alcohol <= 9.05: 6 (2.0)
|   |   |   |   |   |   |   alcohol > 9.05
|   |   |   |   |   |   |   |   volatile acidity <= 0.775
|   |   |   |   |   |   |   |   |   pH <= 3.31: 5 (17.0)
|   |   |   |   |   |   |   |   |   pH > 3.31
|   |   |   |   |   |   |   |   |   |   chlorides <= 0.09: 5 (3.0)
|   |   |   |   |   |   |   |   |   |   chlorides > 0.09: 6 (2.0)
|   |   |   |   |   |   |   |   volatile acidity > 0.775: 6 (2.0)
|   |   |   total sulfur dioxide > 98
|   |   |   |   pH <= 2.93: 6 (4.0/1.0)
|   |   |   |   pH > 2.93: 5 (107.0/6.0)
alcohol > 10.5
|   sulphates <= 0.64
|   |   volatile acidity <= 1.01
|   |   |   pH <= 3.27
|   |   |   |   volatile acidity <= 0.49
|   |   |   |   |   free sulfur dioxide <= 39
|   |   |   |   |   |   alcohol <= 11.8
|   |   |   |   |   |   |   citric acid <= 0.39: 7 (8.0/1.0)
|   |   |   |   |   |   |   citric acid > 0.39
|   |   |   |   |   |   |   |   volatile acidity <= 0.4
|   |   |   |   |   |   |   |   |   fixed acidity <= 11.9
|   |   |   |   |   |   |   |   |   |   pH <= 3.17: 6 (6.0)
|   |   |   |   |   |   |   |   |   |   pH > 3.17
|   |   |   |   |   |   |   |   |   |   |   volatile acidity <= 0.315: 6 (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   volatile acidity > 0.315: 5 (4.0)
|   |   |   |   |   |   |   |   |   fixed acidity > 11.9: 5 (2.0/1.0)
|   |   |   |   |   |   |   |   volatile acidity > 0.4: 7 (3.0/1.0)
|   |   |   |   |   |   alcohol > 11.8
|   |   |   |   |   |   |   residual sugar <= 3.9
|   |   |   |   |   |   |   |   sulphates <= 0.54: 6 (8.0)
|   |   |   |   |   |   |   |   sulphates > 0.54
|   |   |   |   |   |   |   |   |   volatile acidity <= 0.315: 7 (5.0)
|   |   |   |   |   |   |   |   |   volatile acidity > 0.315: 6 (6.0/1.0)
|   |   |   |   |   |   |   residual sugar > 3.9: 7 (7.0)
|   |   |   |   |   free sulfur dioxide > 39: 6 (2.0/1.0)
|   |   |   |   volatile acidity > 0.49
|   |   |   |   |   chlorides <= 0.115
|   |   |   |   |   |   pH <= 3.23: 5 (7.0)
|   |   |   |   |   |   pH > 3.23: 6 (4.0/1.0)
|   |   |   |   |   chlorides > 0.115: 6 (5.0)
|   |   |   pH > 3.27
|   |   |   |   free sulfur dioxide <= 4
|   |   |   |   |   sulphates <= 0.53: 5 (3.0)
|   |   |   |   |   sulphates > 0.53
|   |   |   |   |   |   alcohol <= 12.1: 4 (6.0/2.0)
```

```
|   |   |   |   |   |   alcohol > 12.1
|   |   |   |   |   |   |   chlorides <= 0.066: 6 (2.0)
|   |   |   |   |   |   |   chlorides > 0.066: 7 (4.0)
|   |   |   |   free sulfur dioxide > 4
|   |   |   |   |   residual sugar <= 3.3
|   |   |   |   |   |   free sulfur dioxide <= 8
|   |   |   |   |   |   |   alcohol <= 11.2
|   |   |   |   |   |   |   |   chlorides <= 0.073
|   |   |   |   |   |   |   |   |   alcohol <= 10.8: 5 (3.0/1.0)
|   |   |   |   |   |   |   |   |   alcohol > 10.8: 4 (3.0)
|   |   |   |   |   |   |   |   chlorides > 0.073: 5 (9.0)
|   |   |   |   |   |   |   alcohol > 11.2
|   |   |   |   |   |   |   |   fixed acidity <= 6.3: 5 (8.0/2.0)
|   |   |   |   |   |   |   |   fixed acidity > 6.3
|   |   |   |   |   |   |   |   |   density <= 0.99652: 6 (20.0/1.0)
|   |   |   |   |   |   |   |   |   density > 0.99652: 5 (2.0)
|   |   |   |   |   |   free sulfur dioxide > 8
|   |   |   |   |   |   |   total sulfur dioxide <= 15: 7 (2.0)
|   |   |   |   |   |   |   total sulfur dioxide > 15
|   |   |   |   |   |   |   |   sulphates <= 0.55
|   |   |   |   |   |   |   |   |   pH <= 3.35: 6 (5.0)
|   |   |   |   |   |   |   |   |   pH > 3.35
|   |   |   |   |   |   |   |   |   |   citric acid <= 0.09
|   |   |   |   |   |   |   |   |   |   |   chlorides <= 0.063: 6 (7.0)
|   |   |   |   |   |   |   |   |   |   |   chlorides > 0.063
|   |   |   |   |   |   |   |   |   |   |   |   density <= 0.99392: 5 (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   density > 0.99392: 6 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   citric acid > 0.09: 5 (9.0/1.0)
|   |   |   |   |   |   |   |   sulphates > 0.55
|   |   |   |   |   |   |   |   |   total sulfur dioxide <= 47: 6 (55.0/6.0)
|   |   |   |   |   |   |   |   |   total sulfur dioxide > 47
|   |   |   |   |   |   |   |   |   |   sulphates <= 0.62
|   |   |   |   |   |   |   |   |   |   |   total sulfur dioxide <= 48: 7 (2.0)
|   |   |   |   |   |   |   |   |   |   |   total sulfur dioxide > 48
|   |   |   |   |   |   |   |   |   |   |   |   sulphates <= 0.61: 6 (10.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   sulphates > 0.61
|   |   |   |   |   |   |   |   |   |   |   |   |   alcohol <= 11.5: 5 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   alcohol > 11.5: 6 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   sulphates > 0.62: 5 (5.0/1.0)
|   |   |   |   |   residual sugar > 3.3
|   |   |   |   |   |   volatile acidity <= 0.63
|   |   |   |   |   |   |   free sulfur dioxide <= 6: 6 (4.0/1.0)
|   |   |   |   |   |   |   free sulfur dioxide > 6
|   |   |   |   |   |   |   |   fixed acidity <= 6.6: 6 (2.0)
|   |   |   |   |   |   |   |   fixed acidity > 6.6: 7 (4.0)
|   |   |   |   |   |   volatile acidity > 0.63
|   |   |   |   |   |   |   pH <= 3.38: 6 (2.0)
|   |   |   |   |   |   |   pH > 3.38
|   |   |   |   |   |   |   |   alcohol <= 11.8: 4 (6.0/1.0)
|   |   |   |   |   |   |   |   alcohol > 11.8: 5 (3.0/1.0)
|   |   volatile acidity > 1.01
|   |   |   residual sugar <= 1.9: 5 (3.0)
|   |   |   residual sugar > 1.9
|   |   |   |   alcohol <= 11.1: 3 (4.0/1.0)
|   |   |   |   alcohol > 11.1: 4 (3.0)
|   sulphates > 0.64
```

```
|   |   total sulfur dioxide <= 104
|   |   |   alcohol <= 11.5
|   |   |   |   volatile acidity <= 0.395
|   |   |   |   |   pH <= 3.25
|   |   |   |   |   |   citric acid <= 0.31: 8 (3.0/1.0)
|   |   |   |   |   |   citric acid > 0.31
|   |   |   |   |   |   |   sulphates <= 0.66: 8 (3.0/1.0)
|   |   |   |   |   |   |   sulphates > 0.66
|   |   |   |   |   |   |   |   residual sugar <= 3.2
|   |   |   |   |   |   |   |   |   volatile acidity <= 0.37
|   |   |   |   |   |   |   |   |   |   fixed acidity <= 10.4
|   |   |   |   |   |   |   |   |   |   |   volatile acidity <= 0.32
|   |   |   |   |   |   |   |   |   |   |   |   pH <= 3.21: 6 (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   pH > 3.21: 7 (2.0)
|   |   |   |   |   |   |   |   |   |   |   volatile acidity > 0.32: 7 (6.0)
|   |   |   |   |   |   |   |   |   |   fixed acidity > 10.4: 7 (12.0)
|   |   |   |   |   |   |   |   |   volatile acidity > 0.37: 6 (3.0)
|   |   |   |   |   |   |   |   residual sugar > 3.2: 6 (3.0)
|   |   |   |   |   pH > 3.25
|   |   |   |   |   |   alcohol <= 10.75
|   |   |   |   |   |   |   free sulfur dioxide <= 29
|   |   |   |   |   |   |   |   volatile acidity <= 0.315
|   |   |   |   |   |   |   |   |   residual sugar <= 2.7: 6 (3.0)
|   |   |   |   |   |   |   |   |   residual sugar > 2.7: 5 (2.0)
|   |   |   |   |   |   |   |   volatile acidity > 0.315: 5 (4.0)
|   |   |   |   |   |   |   free sulfur dioxide > 29: 7 (2.0)
|   |   |   |   |   |   alcohol > 10.75
|   |   |   |   |   |   |   volatile acidity <= 0.18: 5 (3.0/1.0)
|   |   |   |   |   |   |   volatile acidity > 0.18
|   |   |   |   |   |   |   |   residual sugar <= 1.8: 6 (9.0/1.0)
|   |   |   |   |   |   |   |   residual sugar > 1.8
|   |   |   |   |   |   |   |   |   pH <= 3.39
|   |   |   |   |   |   |   |   |   |   citric acid <= 0.31: 5 (2.0/1.0)
|   |   |   |   |   |   |   |   |   |   citric acid > 0.31
|   |   |   |   |   |   |   |   |   |   |   density <= 0.99728: 6 (15.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   density > 0.99728: 7 (6.0/1.0)
|   |   |   |   |   |   |   |   |   pH > 3.39
|   |   |   |   |   |   |   |   |   |   sulphates <= 0.7: 6 (2.0)
|   |   |   |   |   |   |   |   |   |   sulphates > 0.7: 7 (8.0)
|   |   |   |   volatile acidity > 0.395
|   |   |   |   |   fixed acidity <= 14.3
|   |   |   |   |   |   chlorides <= 0.097
|   |   |   |   |   |   |   chlorides <= 0.056
|   |   |   |   |   |   |   |   free sulfur dioxide <= 10: 6 (2.0)
|   |   |   |   |   |   |   |   free sulfur dioxide > 10: 5 (3.0)
|   |   |   |   |   |   |   chlorides > 0.056
|   |   |   |   |   |   |   |   alcohol <= 11
|   |   |   |   |   |   |   |   |   pH <= 3.57
|   |   |   |   |   |   |   |   |   |   alcohol <= 10.6
|   |   |   |   |   |   |   |   |   |   |   volatile acidity <= 0.49: 6 (4.0)
|   |   |   |   |   |   |   |   |   |   |   volatile acidity > 0.49
|   |   |   |   |   |   |   |   |   |   |   |   volatile acidity <= 0.6: 7 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   volatile acidity > 0.6: 5 (2.0/1.0)
|   |   |   |   |   |   |   |   |   |   alcohol > 10.6: 6 (34.0/4.0)
|   |   |   |   |   |   |   |   |   pH > 3.57: 7 (2.0)
|   |   |   |   |   |   |   |   alcohol > 11
```

```
|   |   |   |   |   |   |   |   |   residual sugar <= 2.7
|   |   |   |   |   |   |   |   |   |   free sulfur dioxide <= 8
|   |   |   |   |   |   |   |   |   |   |   fixed acidity <= 10.3: 7 (2.0)
|   |   |   |   |   |   |   |   |   |   |   fixed acidity > 10.3: 6 (4.0)
|   |   |   |   |   |   |   |   |   |   free sulfur dioxide > 8: 6 (24.0)
|   |   |   |   |   |   |   |   |   residual sugar > 2.7
|   |   |   |   |   |   |   |   |   |   volatile acidity <= 0.69: 7 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   volatile acidity > 0.69: 6 (2.0)
|   |   |   |   |   |   chlorides > 0.097
|   |   |   |   |   |   |   volatile acidity <= 0.51
|   |   |   |   |   |   |   |   density <= 0.99818: 5 (7.0/1.0)
|   |   |   |   |   |   |   |   density > 0.99818: 6 (4.0)
|   |   |   |   |   |   |   volatile acidity > 0.51
|   |   |   |   |   |   |   |   sulphates <= 0.69: 6 (4.0/1.0)
|   |   |   |   |   |   |   |   sulphates > 0.69: 7 (2.0)
|   |   |   |   |   fixed acidity > 14.3: 5 (4.0/1.0)
|   |   |   alcohol > 11.5
|   |   |   |   fixed acidity <= 12.9
|   |   |   |   |   sulphates <= 0.68
|   |   |   |   |   |   free sulfur dioxide <= 7
|   |   |   |   |   |   |   chlorides <= 0.109: 7 (8.0/1.0)
|   |   |   |   |   |   |   chlorides > 0.109: 6 (2.0)
|   |   |   |   |   |   free sulfur dioxide > 7: 6 (12.0)
|   |   |   |   |   sulphates > 0.68
|   |   |   |   |   |   sulphates <= 0.69
|   |   |   |   |   |   |   free sulfur dioxide <= 8: 8 (4.0/1.0)
|   |   |   |   |   |   |   free sulfur dioxide > 8: 5 (2.0/1.0)
|   |   |   |   |   |   sulphates > 0.69
|   |   |   |   |   |   |   free sulfur dioxide <= 18
|   |   |   |   |   |   |   |   free sulfur dioxide <= 9
|   |   |   |   |   |   |   |   |   volatile acidity <= 0.34
|   |   |   |   |   |   |   |   |   |   density <= 0.99625: 7 (11.0)
|   |   |   |   |   |   |   |   |   |   density > 0.99625: 8 (3.0/1.0)
|   |   |   |   |   |   |   |   |   volatile acidity > 0.34
|   |   |   |   |   |   |   |   |   |   density <= 0.99388: 7 (2.0/1.0)
|   |   |   |   |   |   |   |   |   |   density > 0.99388
|   |   |   |   |   |   |   |   |   |   |   density <= 0.99692: 6 (7.0)
|   |   |   |   |   |   |   |   |   |   |   density > 0.99692
|   |   |   |   |   |   |   |   |   |   |   |   sulphates <= 0.88: 7 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   sulphates > 0.88: 6 (2.0)
|   |   |   |   |   |   |   |   free sulfur dioxide > 9
|   |   |   |   |   |   |   |   |   residual sugar <= 2.65: 7 (28.0/1.0)
|   |   |   |   |   |   |   |   |   residual sugar > 2.65
|   |   |   |   |   |   |   |   |   |   volatile acidity <= 0.32: 6 (2.0)
|   |   |   |   |   |   |   |   |   |   volatile acidity > 0.32: 7 (8.0/1.0)
|   |   |   |   |   |   |   free sulfur dioxide > 18
|   |   |   |   |   |   |   |   pH <= 3.36
|   |   |   |   |   |   |   |   |   residual sugar <= 2.15
|   |   |   |   |   |   |   |   |   |   sulphates <= 0.77: 7 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   sulphates > 0.77: 6 (4.0)
|   |   |   |   |   |   |   |   |   residual sugar > 2.15
|   |   |   |   |   |   |   |   |   |   total sulfur dioxide <= 57: 7 (8.0)
|   |   |   |   |   |   |   |   |   |   total sulfur dioxide > 57
|   |   |   |   |   |   |   |   |   |   |   chlorides <= 0.073: 6 (2.0)
|   |   |   |   |   |   |   |   |   |   |   chlorides > 0.073: 7 (3.0)
|   |   |   |   |   |   |   |   pH > 3.36
```

```
|   |   |   |   |   |   |   |   |   alcohol <= 12.5: 6 (11.0)
|   |   |   |   |   |   |   |   |   alcohol > 12.5
|   |   |   |   |   |   |   |   |   |   volatile acidity <= 0.4: 6 (3.0)
|   |   |   |   |   |   |   |   |   |   volatile acidity > 0.4: 8 (3.0)
|   |   |   |   |   fixed acidity > 12.9
|   |   |   |   |   |   residual sugar <= 4.5: 6 (3.0)
|   |   |   |   |   |   residual sugar > 4.5: 5 (2.0)
|   |   total sulfur dioxide > 104: 5 (6.0)

Number of Leaves  :  228

Size of the tree :  455
```

Références

[1] La définition de test en statistique dans l'encyclopédie virtuelle Wikipédia :
http ://fr.wikipedia.org/wiki/Test_(statistique)

[2] Le cours sur le test de Kolmogorov-Smirnov :
http ://www.apprendre-en-ligne.net/random/KS.html

[3] Le cours sur le test de Wilcoxon :
http ://www.math-info.univ-paris5.fr/smel/cours/ts/node10.html

[4] Le cours sur le test du χ^2 et de la régression de M. Canu sur moodle :
https ://moodle.insa-rouen.fr/course/view.php ?id=169

[5] Un tutoriel sur l'établissement d'un arbre de décision sous Weka :
http ://jaub.developpez.com/tutoriels/weka/weka/

[6] La définition du Data mining donnée par Wikipédia :
http ://fr.wikipedia.org/wiki/Data_mining

[7] La définition de SVM dans Wikipédia :
http ://fr.wikipedia.org/wiki/Machine_%C3%A0_vecteurs_de_support