Wine Quality

Alexandre SALMON, Thomas FRION

13/01/2021

1 Objectif de l'étude

Notre étude va porter sur la qualité des vins portugais, de la région Vinho Verde. Nos objectifs via cette étude sont :

- Déterminer un modèle de prédiction de la qualité d'un vin rouge et d'un vin blanc. Ainsi, connaitre le critère physionomique le plus important dans la détermination de la qualité pour chaque type de vin.
- Comparer les deux modèles (blanc et rouge) pour savoir si ce qui vait un bon vin blanc, fait un bon vin rouge

Les conclusions de cette étude pourront permettre aux vignerons d'améliorer la qualité de leurs vins.

2 Analyse descriptive

Pour cette étude nous disposons de deux jeux de données : un pour le vin rouge et 1 pour le vin blanc. Nous avons obtenu ces données sur la page UCI: Wine Quality Data Set

Pour les deux jeux de données nous avons douze variables: onze variables d'entrée qui correspondent aux critères physionomiques du vin et une variable de sortie qui correspond à la qualité du vin en 0 et 10. Les onze critères physionomiques sont les suivantes :

1. Acidité fixe (ou acidité naturelle du raisin)
2. Acidité volatile (teneur d'acides gras)
3. Acide citrique (teneur d'acide citrique)
4. Sucre résiduel (sucres encore présents après la fermentation)
5. Chlorures (teneur des différents chlorures)
6. Dioxyde de soufre (SO_2) libre (teneur du principe actif du SO_2)
7. Total du dioxyde de soufre (SO_2) (teneur de toutes les formes du SO_2)
8. Densité
9. pH
10. Sulfates (teneur du fongicide)
11. Alcool

La variable de sortie correspond à la médianne des notes données par des experts (au minimun trois notes). Si la variable vaut 0 sela signifie que le vin est de très mauvaise qualité. Si la variable vaut 10 alors le vin est de bonne qualité.

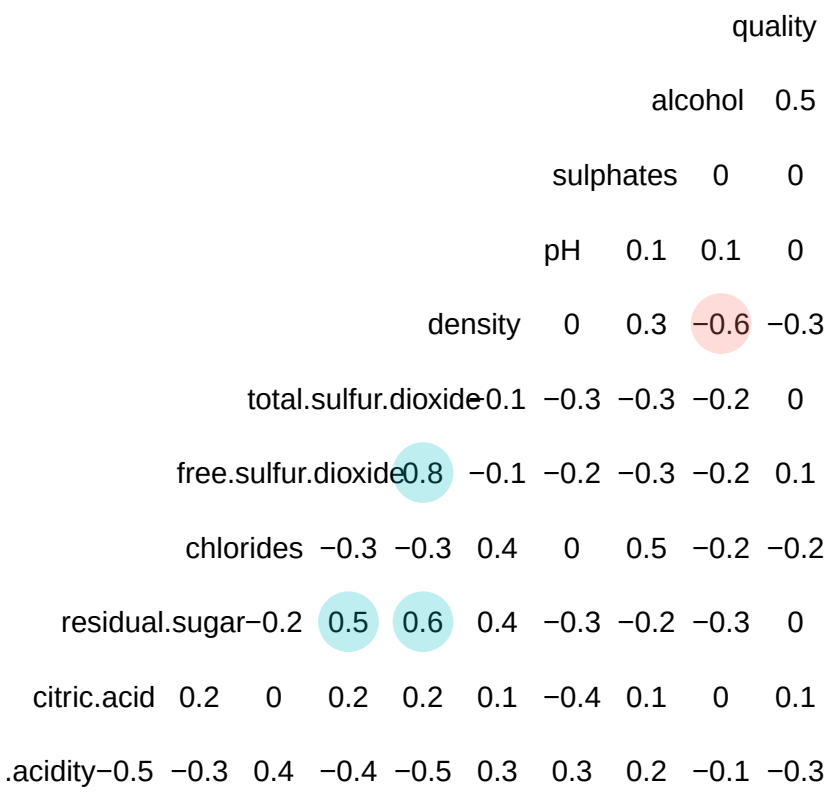
Nous avons 1599 vins rouges et 4898 vins blancs.

Nous allons maintenant décrire les liens entre les différentes variables.

Nous pouvons déjà annoncer un lien entre deux variables : SO_2 libre et SO_2 total. Ce lien est le suivant : plus la teneur en SO_2 libre augmente et la teneur totale en SO_2 augmente.

Afin de connaître d'autres liens entre les variables d'entrée, nous avons décidé de calculer la corrélation entre les différentes variables sur trois jeux de données différents. Le premier calcul de corrélation a été fait sur un

ensemble de données composé de 50% de vins rouges et 50% de vins blancs. Nous avons créer cette ensemble de données afin de voir de façon globale les liens entre les différentes variables. Voici les résultats de ce premier calcul :



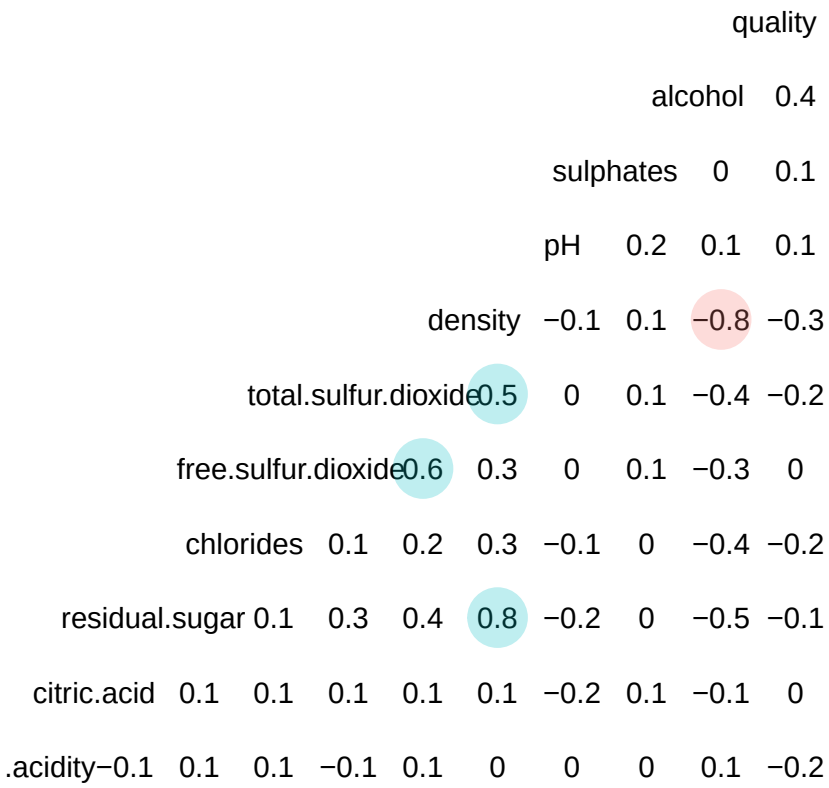
Nous pouvons voir que le lien de précédement annoncé : le lien entre le SO_2 libre et le SO_2 total. Mais nous constatons également que, dans l'ensemble, il la variable de densité et d'alcool sont liées. Les variables de sucre résiduel et SO_2 libre sont également liées.

Ensuite nous avons réalisé le calcul de corrélation uniquement sur les vins rouges, afin de voir s'il y a des liens entre les variables qui seraient spécifiques aux vins rouges.

									quality
								alcohol	0.5
							sulphates	0.1	0.3
						pH	−0.2	0.2	−0.1
					density	−0.3	0.1	−0.5	−0.2
				total.sulfur.dioxide	0.1	−0.1	0	−0.2	−0.2
			free.sulfur.dioxide	0.7	0	0.1	0.1	−0.1	−0.1
		chlorides	0	0	0.2	−0.3	0.4	−0.2	−0.1
	residual.sugar	0.1	0.2	0.2	0.4	−0.1	0	0	0
citric.acid	0.1	0.2	−0.1	0	0.4	−0.5	0.3	0.1	0.2
.acidity	−0.6	0	0.1	0	0.1	0	0.2	−0.3	−0.2	−0.4

Dans le cas du vin rouge, nous pouvons voir qu'il y a un lien entre le la teneur d'acide citrique et le pH du vin. Nous constatons également que l'acidité volatile et l'acide citrique sont liées.

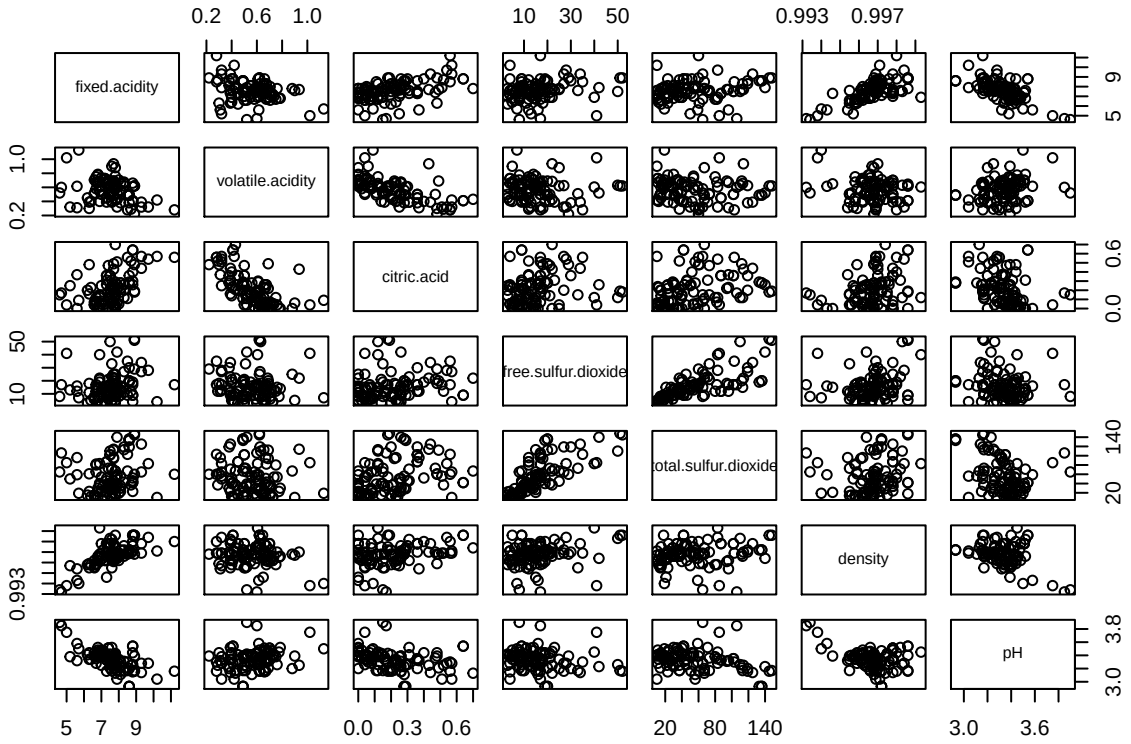
Le troisième et dernier calcul de corrélation s'est effectué sur les vins blancs. Comme pour le précédent calcul, nous avons réalisé se calcul afin de connaître l'existence de liens entre les variables qui seraient spécifiques aux vins blancs.



Nous pouvons voir avec ces résultats, nous pouvons voir que la variable de la densité du vin est liée aux variables de sucre résiduel et de SO_2 total.

Nous pouvons observer dans les résultats ci-desssus, qu'il y a des corrélations positives et négatives. Par exemple, la densité et l'alcool ont une corrélation négative : cela signifie que plus il y a d'alcool moins le vin est dense. À l'inverse, le SO_2 libre et le SO_2 total ont une corrélation positive. Ce qui signifie que plus il y a de SO_2 libre plus la teneur totale de SO_2 sera importante.

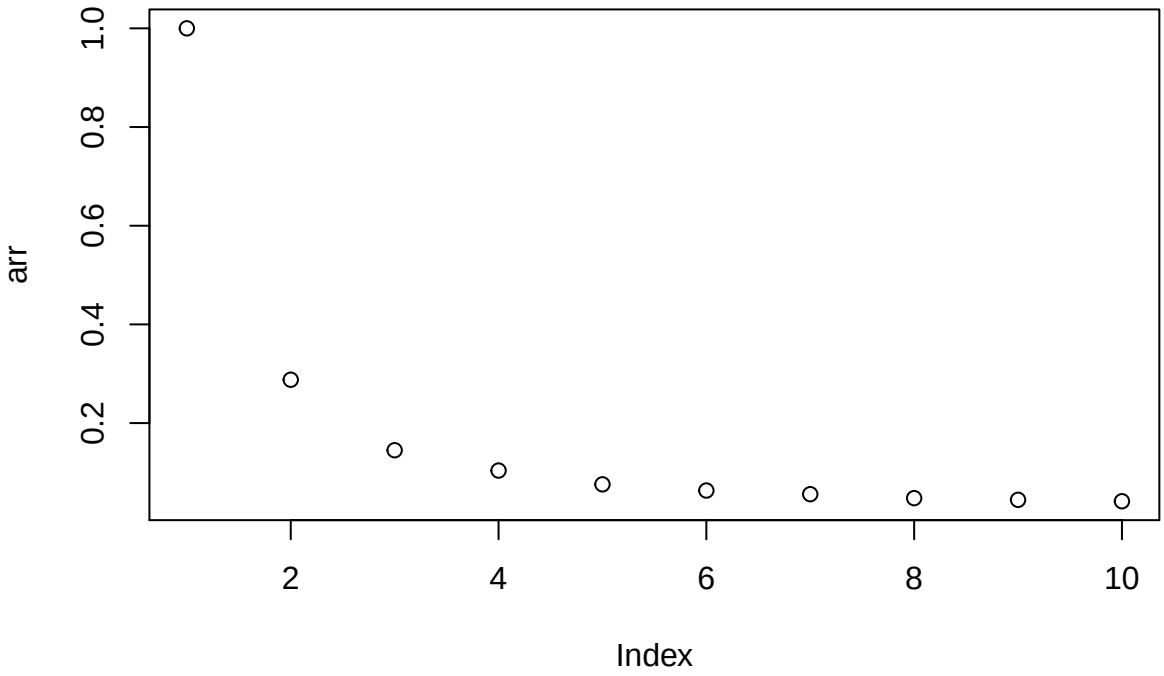
Nous avons choisis de faire le calcul des corrélations afin d'avoir une vision d'ensemble des liens entre les variables. Ainsi nous pouvons faire un plot plus lisible des données pour mieux apprécier les liens. Ainsi nous avons pu établir le diagramme suivant :



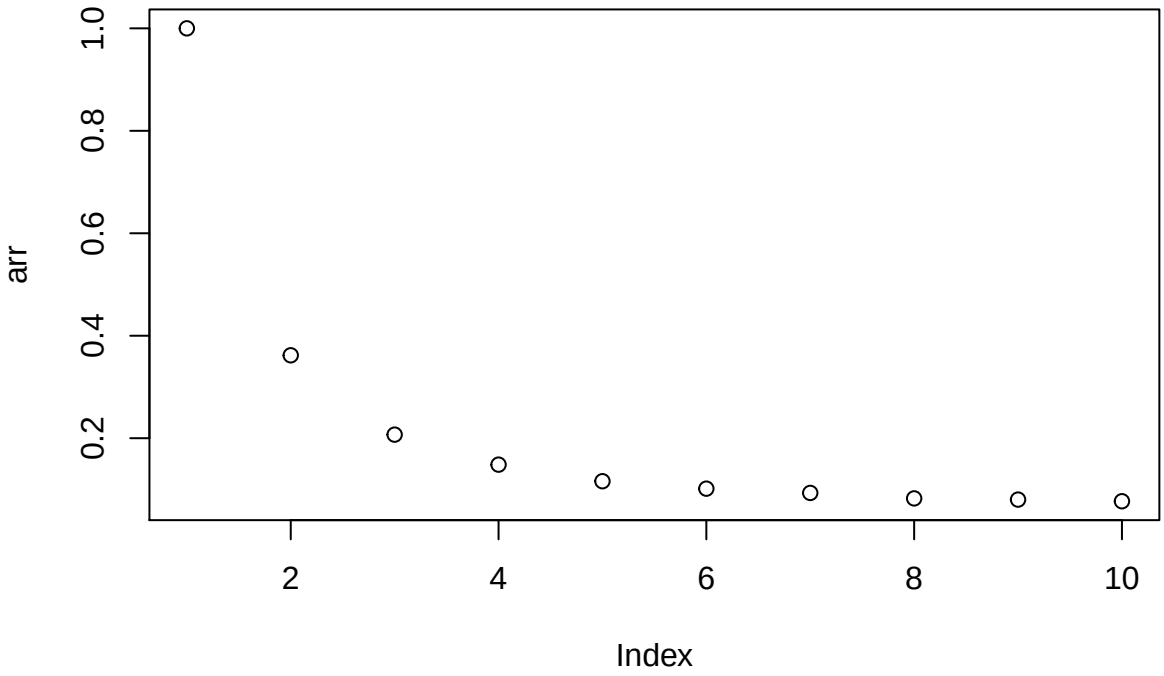
Sur le diagramme ci-dessus, nous pouvons voir qu'il existe un lien entre le pH du vin et son acidité fixe. Ce lien s'explique par le fait que la mesure du pH permet de mesurer l'acidité ou la basicité d'une solution. Nous pouvons constater qu'il y a un lien entre le pH et la teneur en acide volatide, et entre le pH et la teneur en acide citrique.

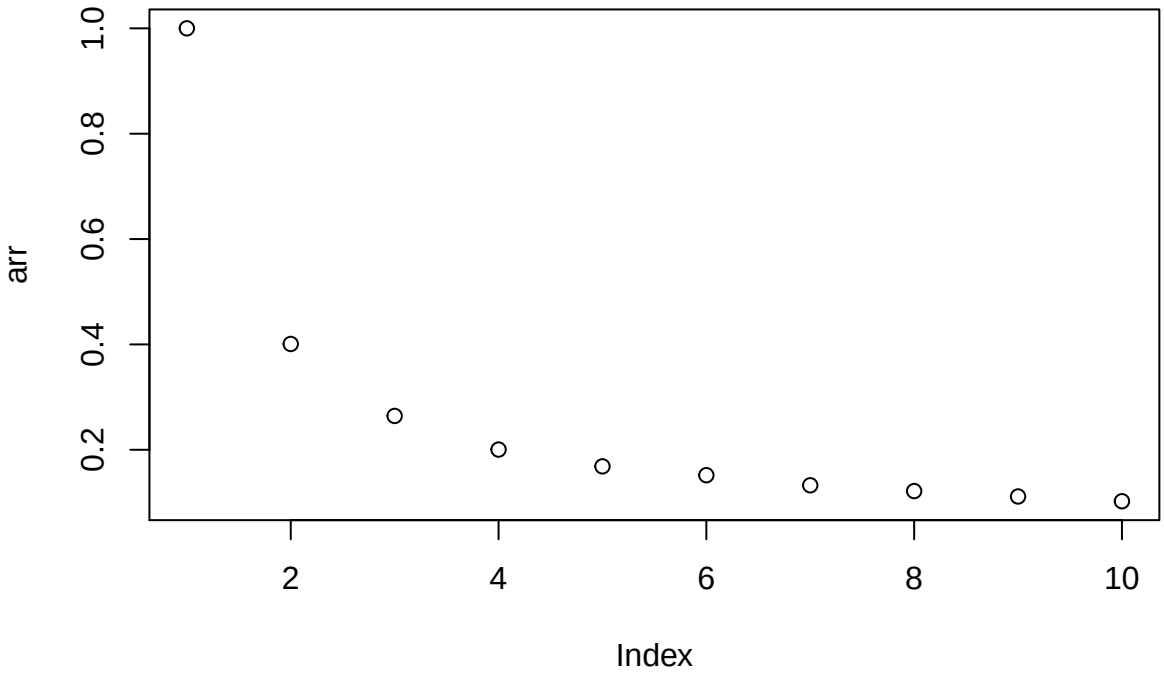
3 Classification Non Supervisée

On cherche à determiner le nombre optimal de cluster pour notre jeu de données. Pour cela, on applique la fonction coude avec 50% de vin blanc et 50% de vin rouge :

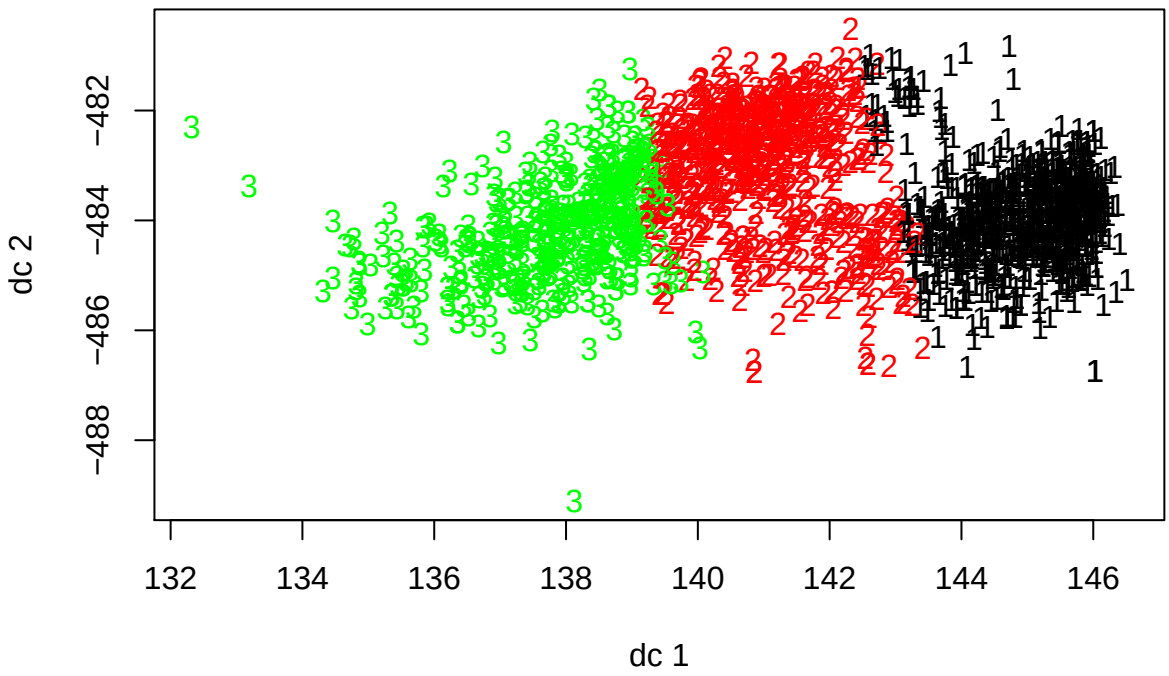


On constate donc que le découpage 3 classes semble le plus pertinent. En appliquant la même méthode séparément aux vins rouges et aux vins blancs, on obtient le ême nombre de classe recommandé.

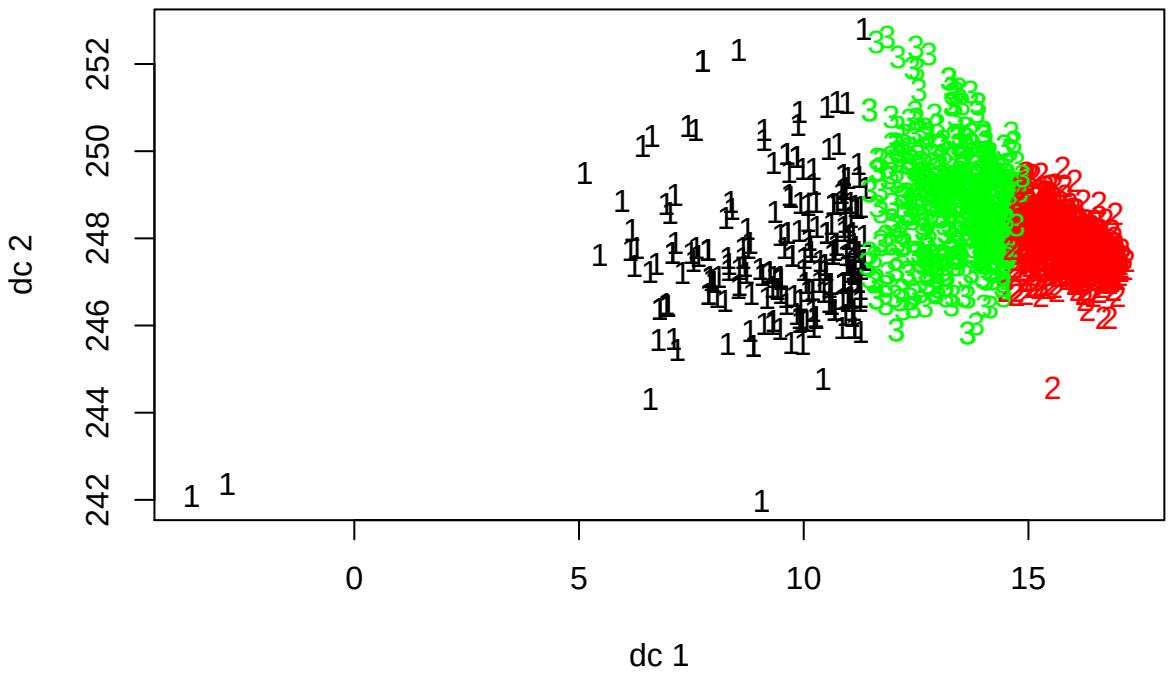

arr
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
Index



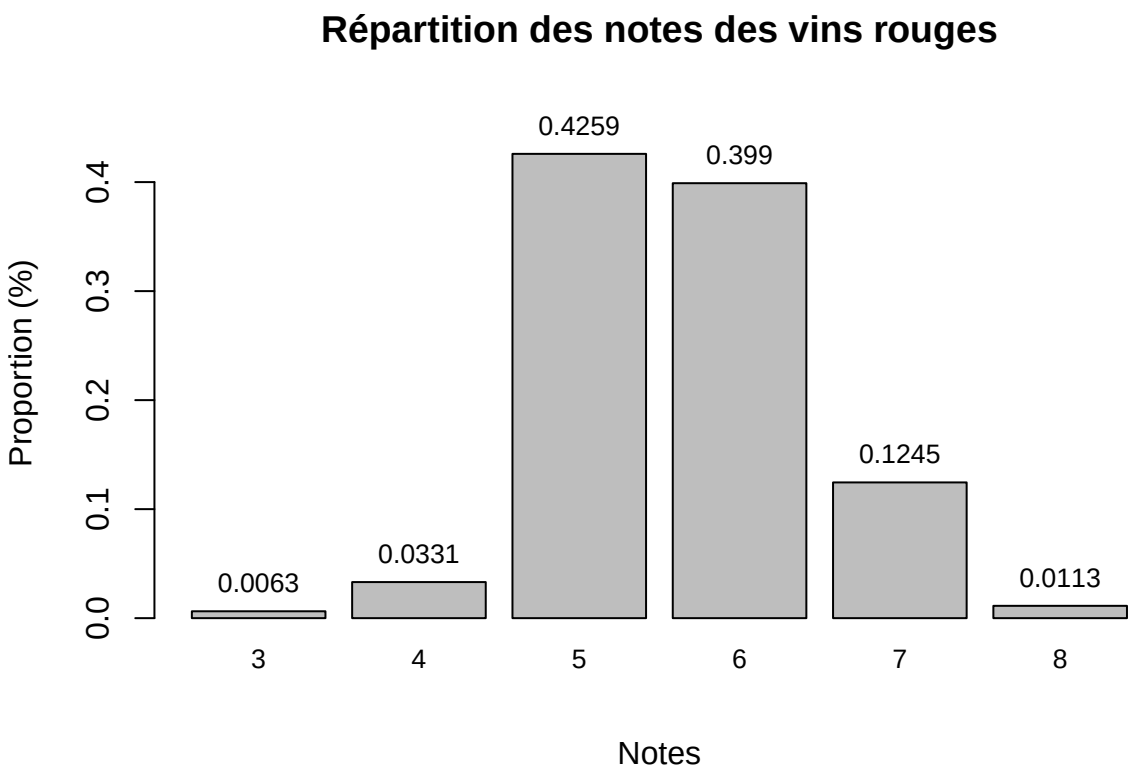
Maintenant que nous connaissons le nombre de classes optimal pour du clustering, nous commençons par l'algorithme des kmeans. Nous applicons l'algorithme comme précédemment, c'est-à-dire d'abord un ensemble de données composé à 50% de vins blancs et 50% de vins rouges, puis nous distingons les vins rouges et les vins blancs.

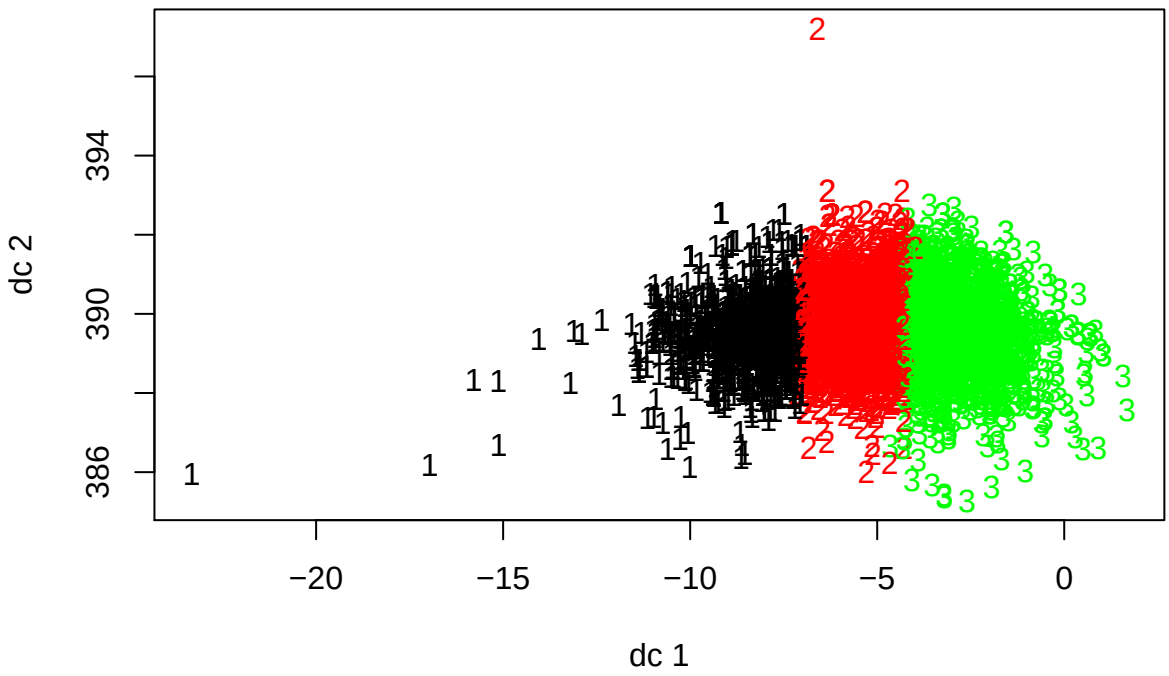


Sur le résultat ci-dessus nous constatons que la séparation entre chaque classe est nette : il n'y a pas d'élément de la classe 3 dans la classe 1 ou 2. Ce qui confirme nos résultats précédents avec la méthode du coude. Nous pouvons dire également que la classe la plus à gauche correspond aux vins ayant une qualité médiocre, que la classe centrale correspond aux vins moyens et que la classe la plus à droite représente les meilleurs vins.

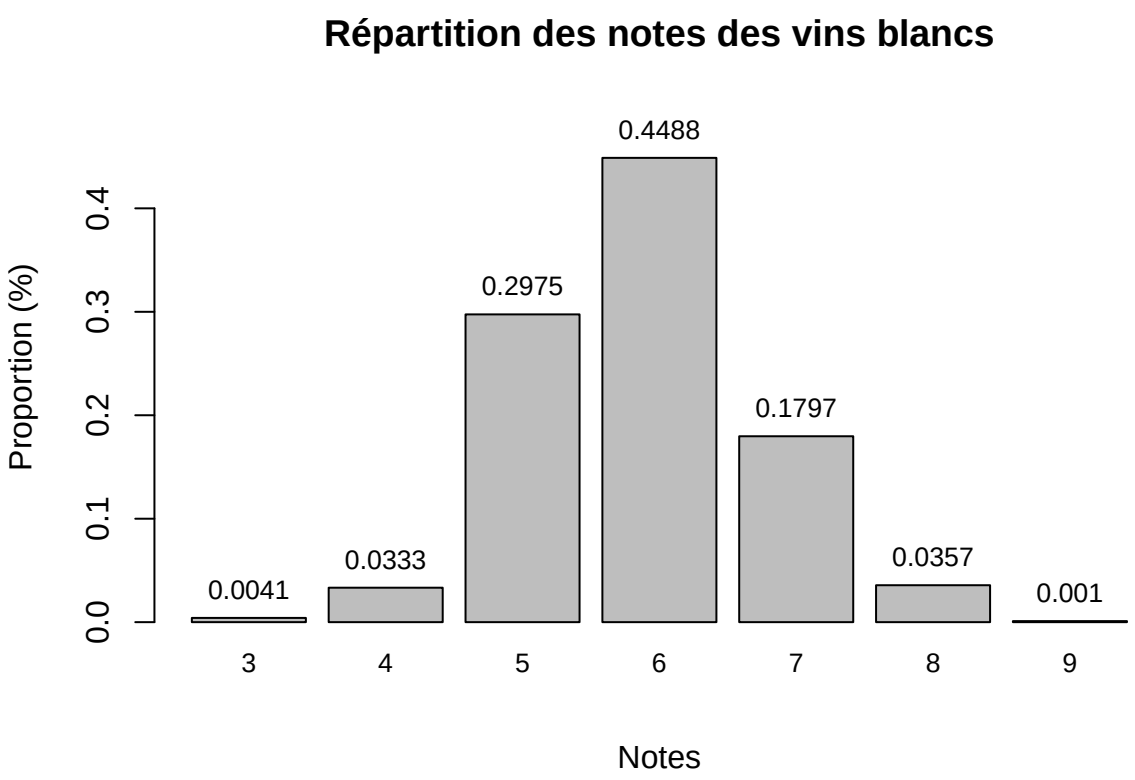


Nous constatons que les classes, dans l'ensemble, sont compactes et très proches. Cela nous indique qu'il y a un nombre important de vins de qualité moyenne et qu'il y a quelques vins qui ont obtenu une notes très petite ou très grande. Cette analyse est confirmée par l'histogramme ci-dessous représentant la proportion de chaque note dans les vins blancs.


Répartition des notes des vins rouges
Proportion (%)
0.0
0.1
0.2
0.3
0.4
0.0063
0.0331
0.4259
0.399
0.1245
0.0113
3
4
5
6
7
8
Notes

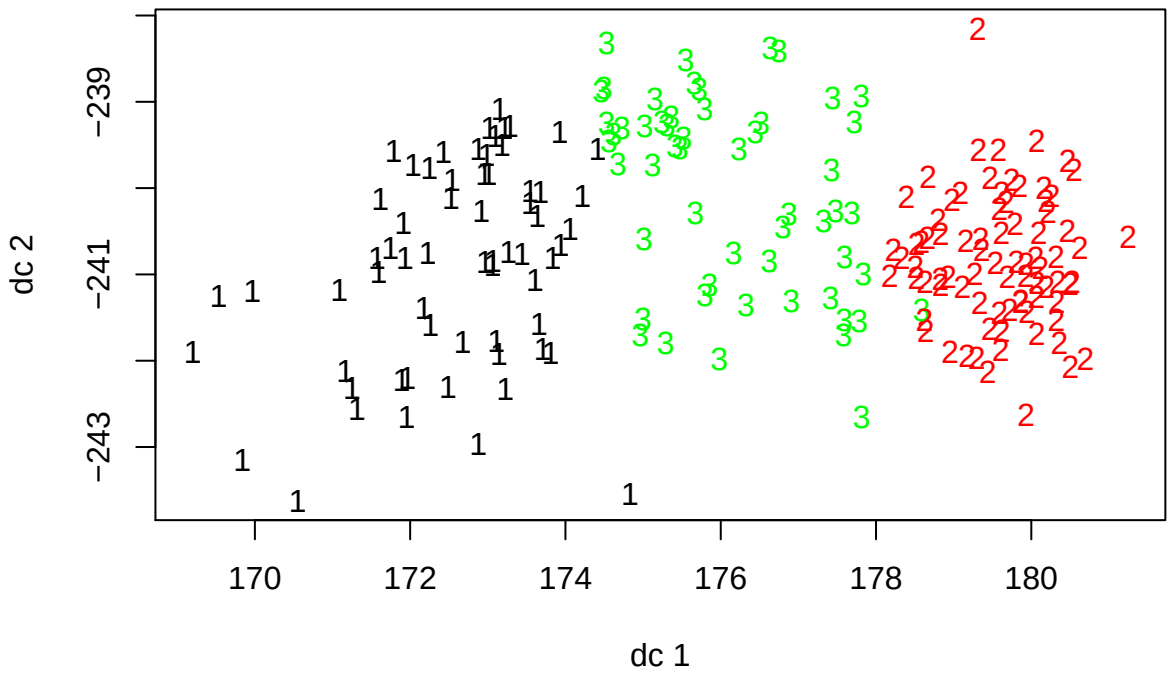


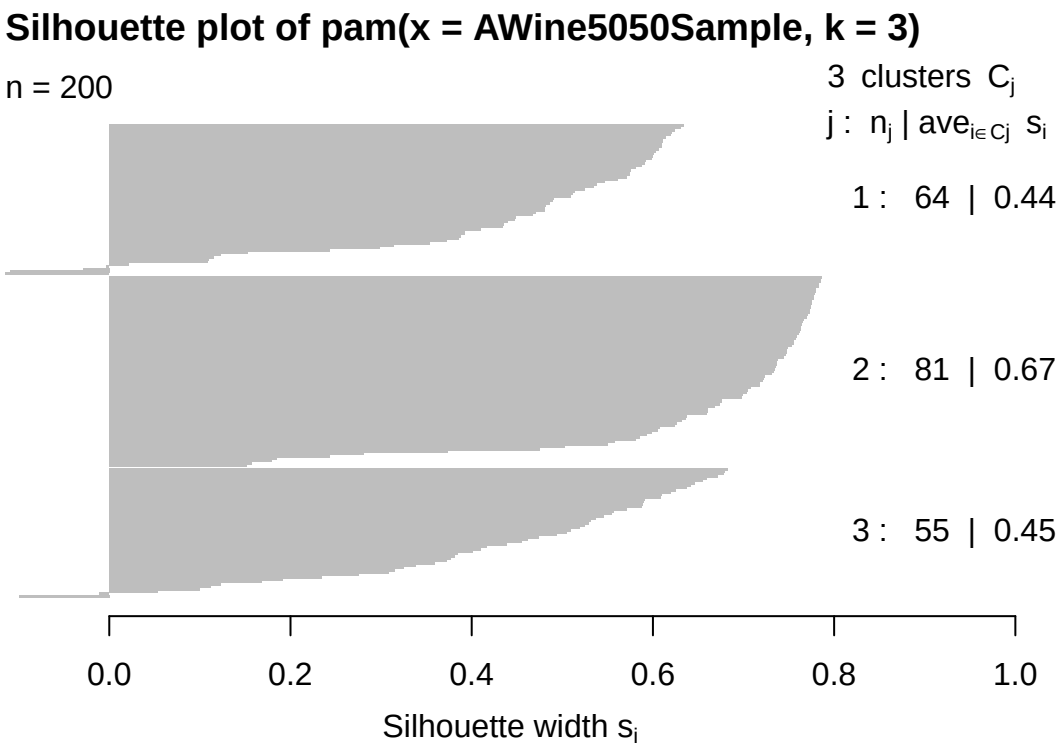
Nous constatons, comme pour les vins rouges, sur le graphique ci-dessus que les classes, dans l'ensemble, sont compactes et très proches des unes et des autres. Cela nous informe qu'il y a un nombre important de vins de qualité moyenne (au sein des vins blancs) et qu'il y a quelques vins qui ont obtenu une notes très petite ou très grande. Cette analyse est confirmée par l'histogramme ci-dessous représentant la proportion de chaque note dans les vins blancs.



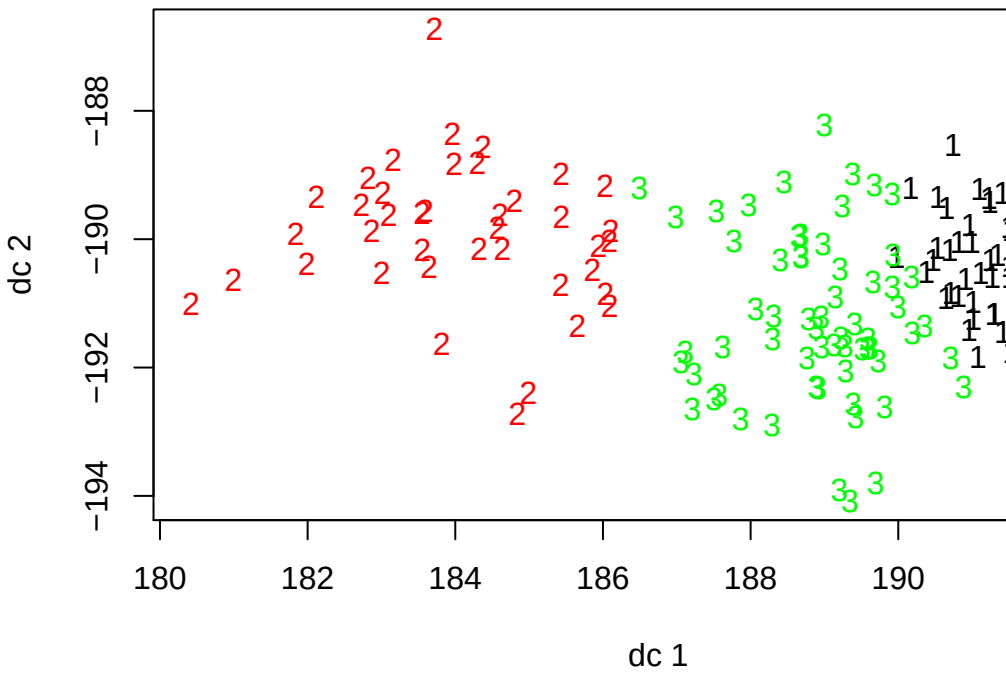
Maintenant que nous avons “clusterisé” nos données, nous allons cérifier la qualité des résultats obtenus grace aux silouhettes.

Nous commençons par le jeu de données contenant 50%de vins rouges et 50% de vins blancs.

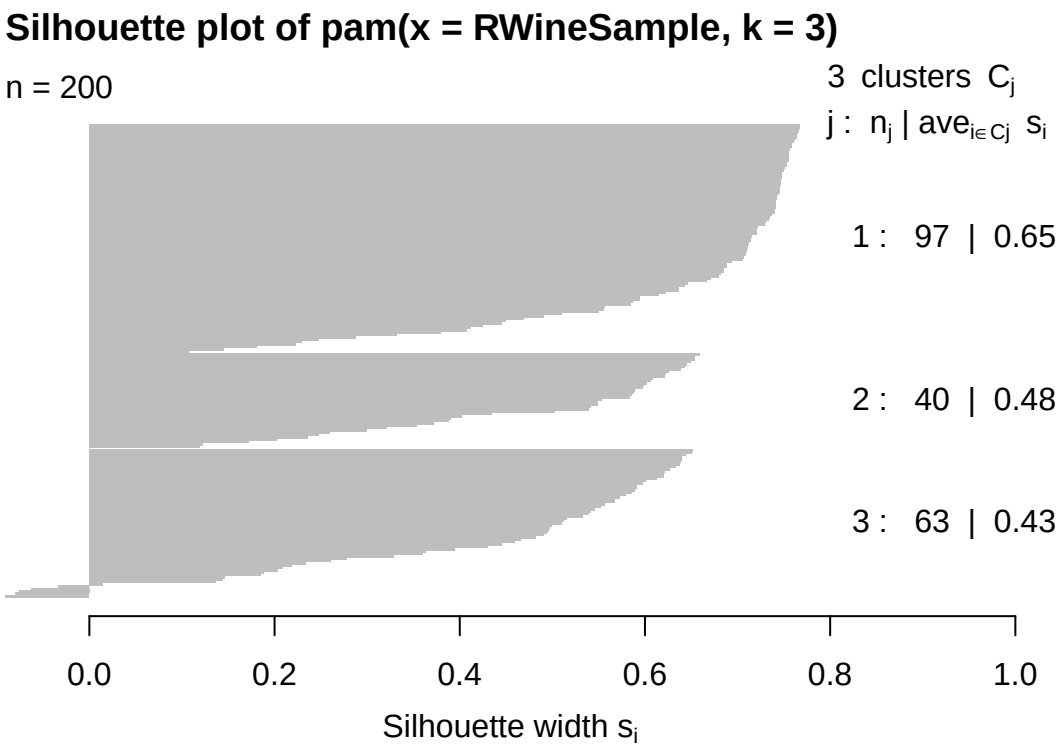

dc 2
−239
−241
−243
170
172
174
176
178
180
dc 1



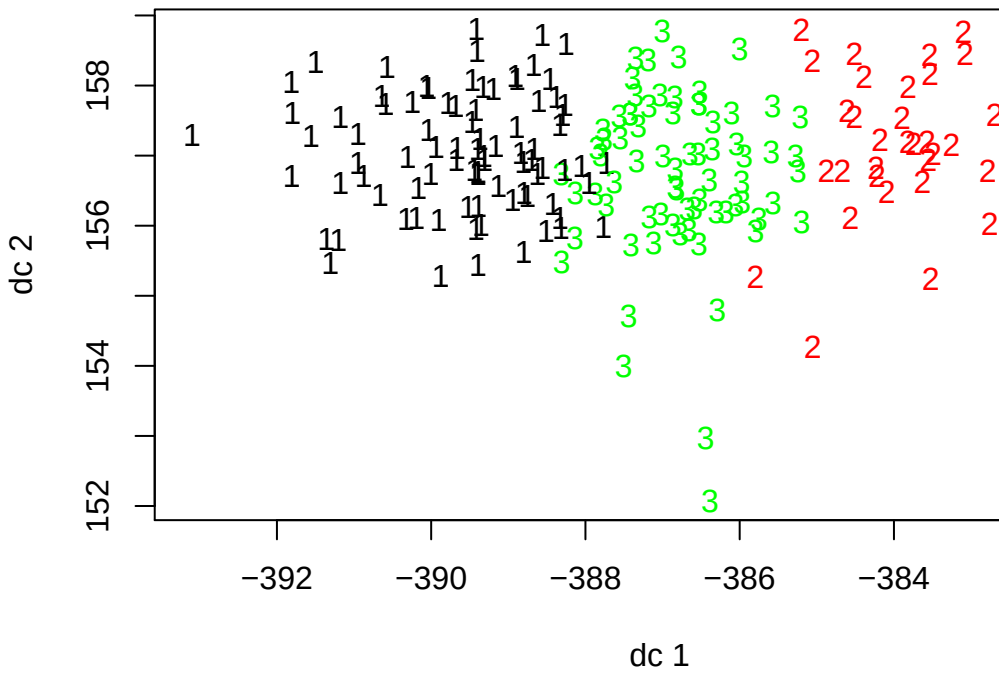
Nous pouvons voir sur la silouhette obtenue, qu'il y a quelques éléments qui ne sont pas satisfaits de leur classe. Mais si nous regardons de manière générale : nous avons obtenue une moyenne qui est proche de 0.5, ce qui veut dire que nous avons un découpage de qualité moyenne.



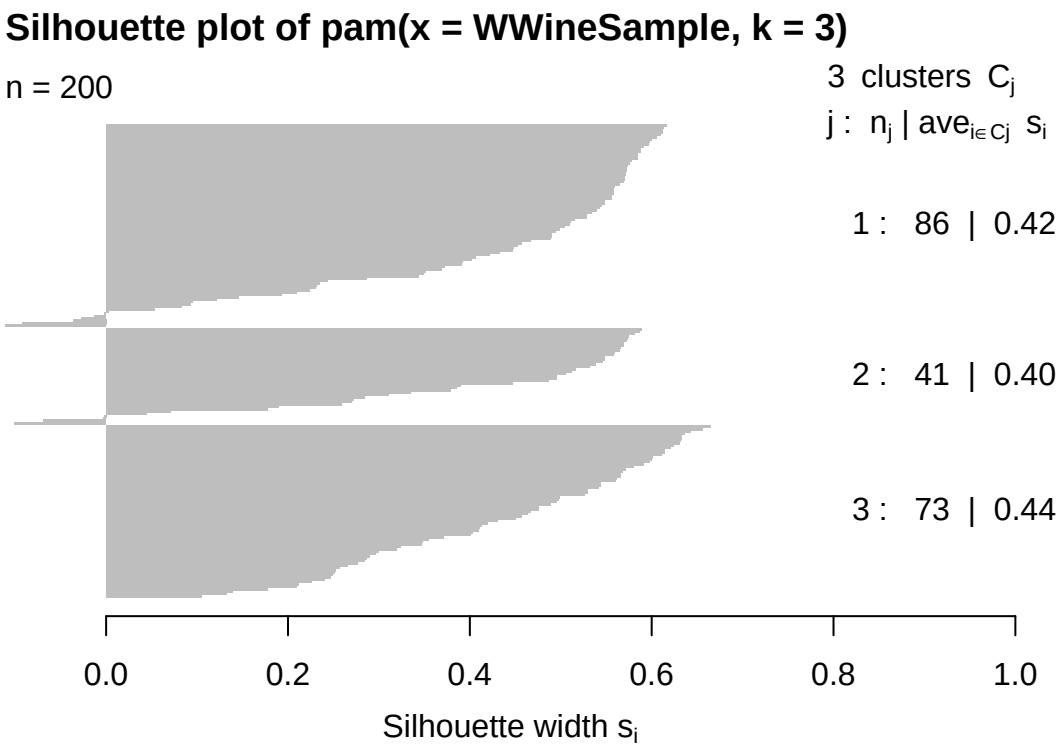
Maintenant regardons les vins rouges.



De manière générale, la qualité de notre découpage est moyenn (d'après la silouhette obtenue), car la qualité moyenne obtenue est proche de 0.5. Nous notons toutefois qu'il y a quelques éléments mal classés.

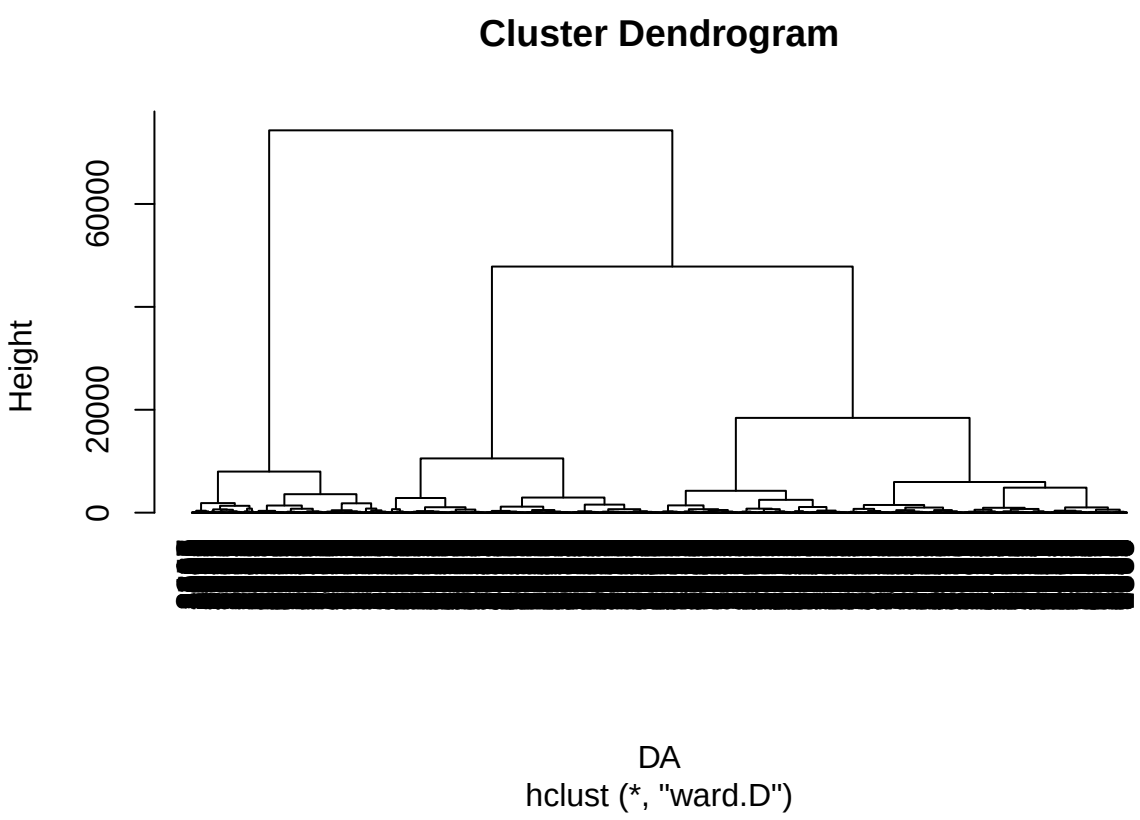


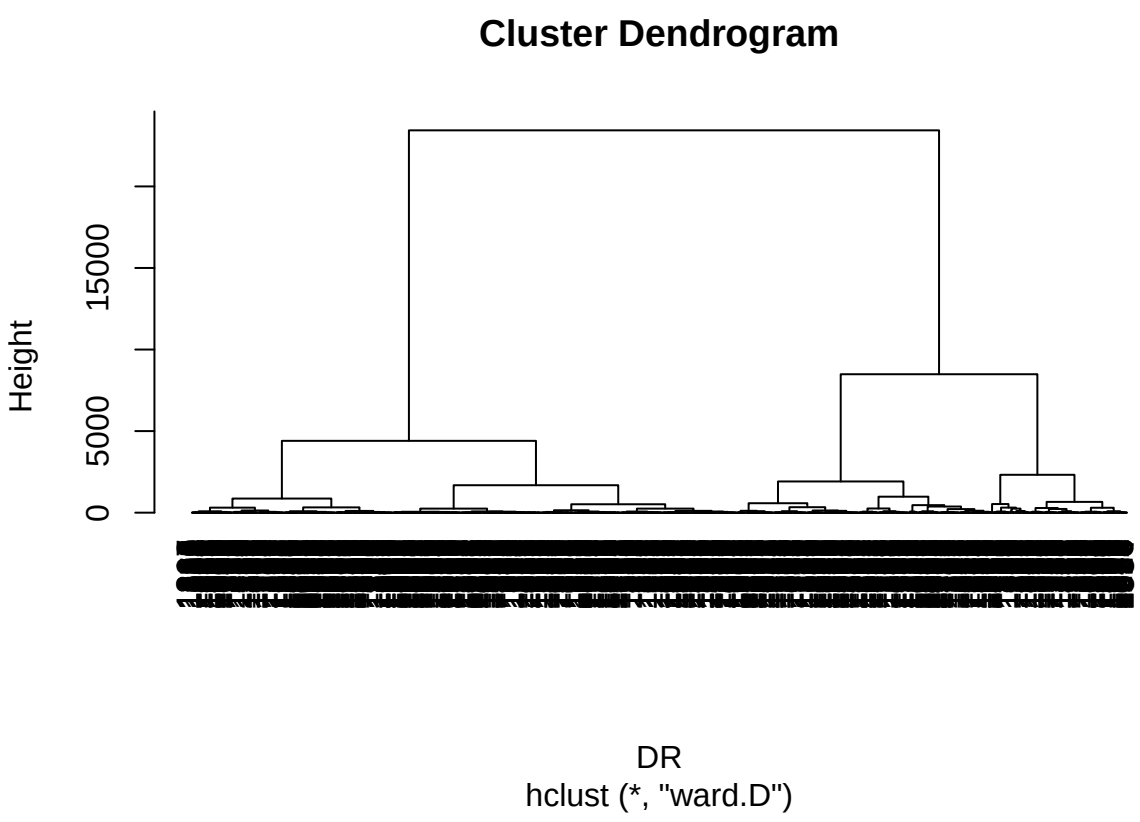
Pour finir nous passons aux vins blancs

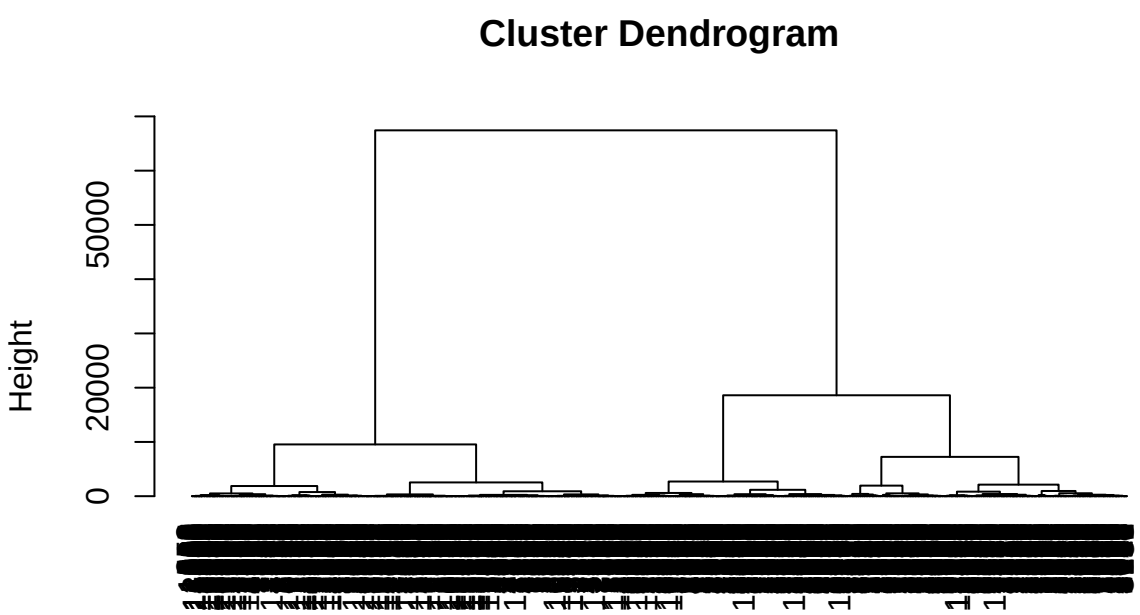


Pour les vins blancs, la qualité de notre découpage est légèrement moins satVins rougesisfaisante. En effet, la qualité moyenne est plus proches de 0.4 que de 0.5. Cela peut s'expliquer par la répartition des notes pour les vins blancs que nous avons observé précédement.

Maintenant que nous avons étudié les silouhettes, nous allons étudier les dendogrammes des trois jeux de données.


Cluster Dendrogram
Height
0
20000
60000
DA
hclust (*, "ward.D")


Cluster Dendrogram
Height
0
5000
15000
DR
hclust (*, "ward.D")



Avec les trois dendogrammes obtenus, nous constatons à chaque fois qu'il y a trois classes possibles. Ce qui confirme les résultats de la méthode du coude.

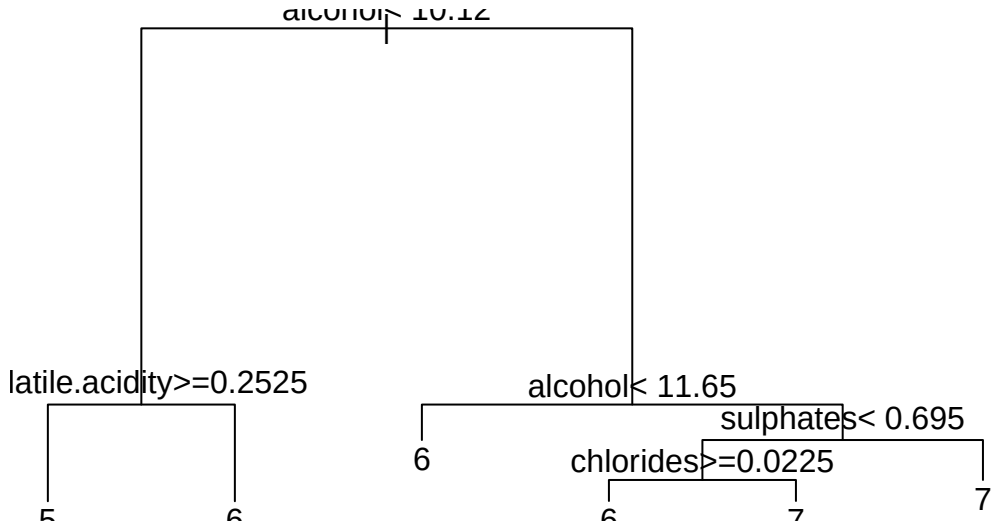
4 Classification supervisée

Comme pour les paries précédentes nous allons commencer par étudier le jeu de données contenant 50% de vins rouges et 50% de vins blancs, afin d'avoir une idée générale. Puis nous fairons l'analyse des jeux de données des vins rouges et blancs.

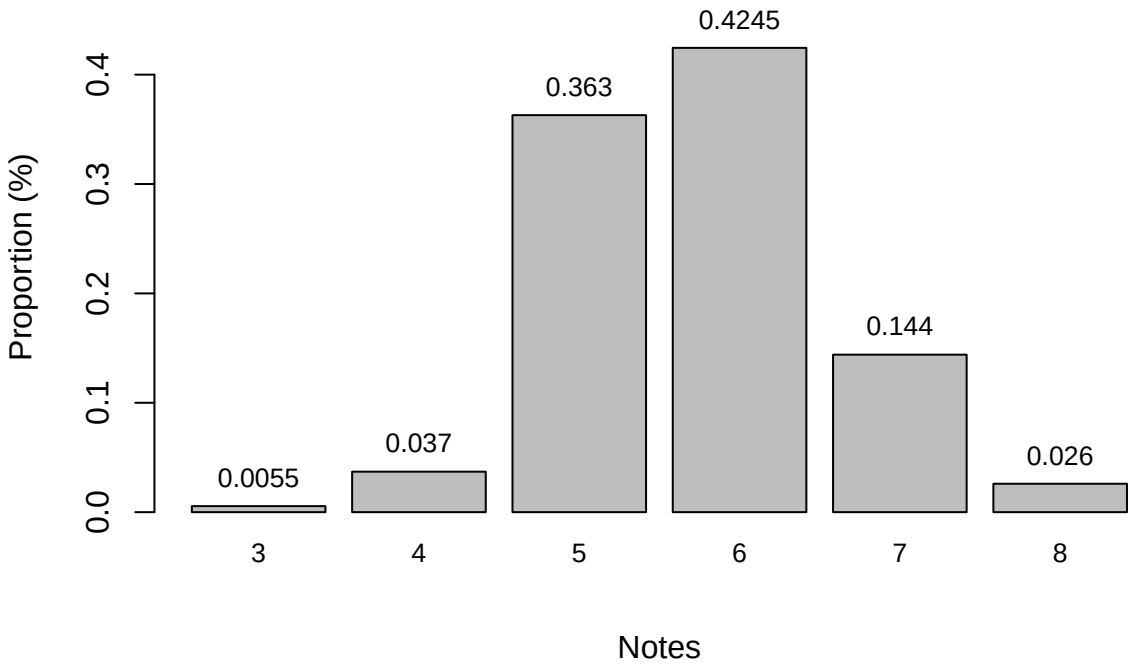
Jeu de données contenant 50% de vins rouges et 50% de vins blancs :

```
## [1] 0.4325
```

Avec 10 modèles différens nous obtenons un taux d'erreur moyen de 46%, et nous obtenons le modèle suivant.



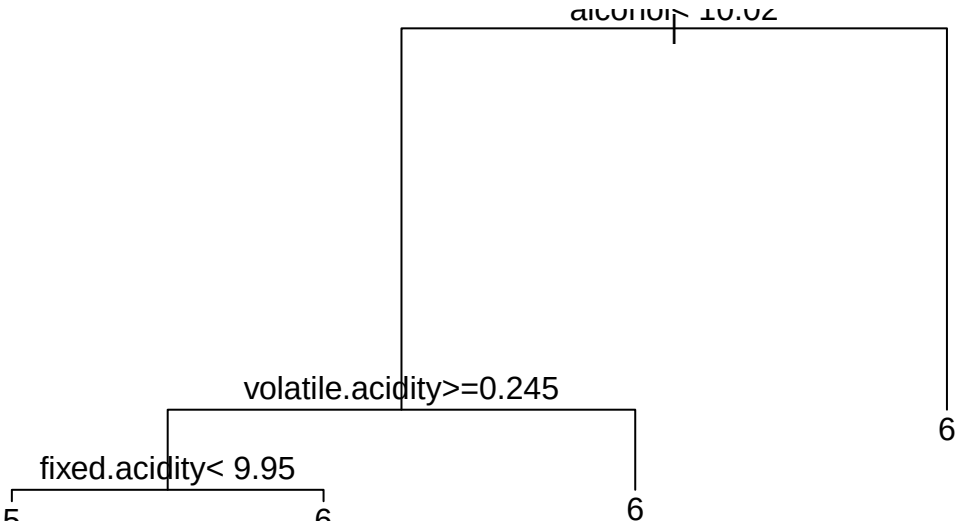
Le taux d'erreur obtenue peut s'expliquer par la répartition des notes pour ce jeu de données (graphique ci-dessous). Nous avons constaté qu'il y a majoritairement des vins de qualité moyenne (note entre 5 et 6). Cela a pour conséquence qu'il n'y a pas assez de données pour les autres notes et ainsi l'entrainement du modèle se concentre uniquement sur les vins de qualité moyenne.



Maintenant, nous allons refaire la même opération sur le même jeu de données, mais cette nous garderons uniquement les vins de qualité moyenne (note entre 5 et 6).

```
## [1] 0.2920635
```

En considérant, uniquement les vins de qualité moyenne nous avons un taux d'erreur moyen (pour 10 modèles différents) de 30%. Ainsi nous obtenons donc le modèle suivant :

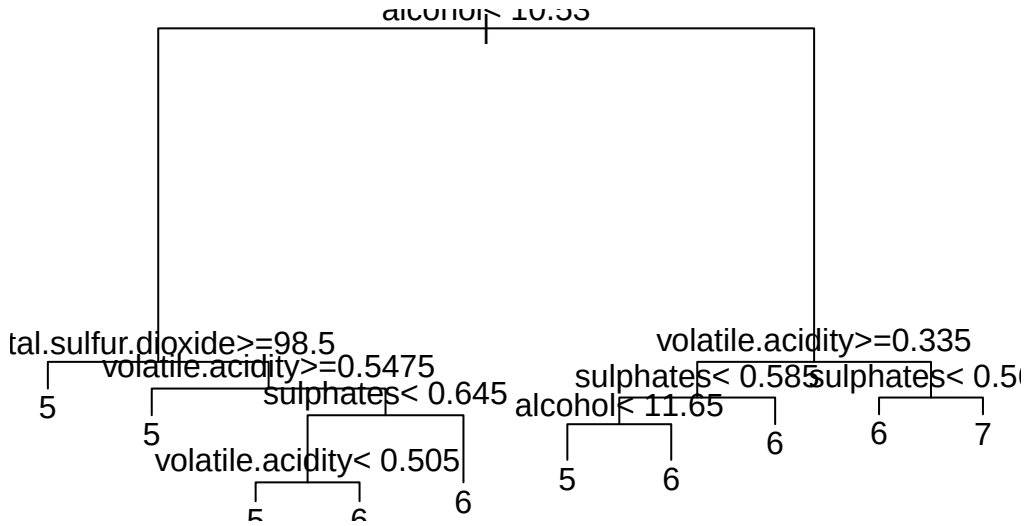


Avec ce modèle nou pouvons voir que, de façon général, le paramètre le plus important pour déterminer la qualité du vin est sa teneur en alcool. Le second paramètre est l'acidité volatile.

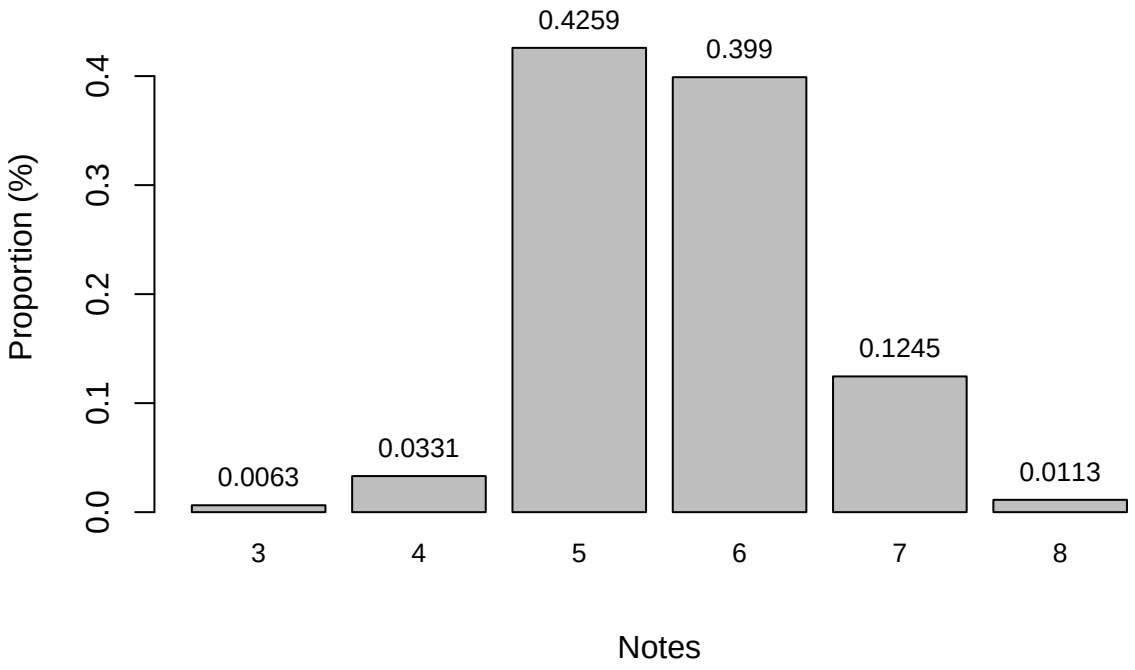
Maintenant voyons qu'est-ce qui fait un bon vin rouge. Pour cela, nous allons procéder de la même manière que précédement.

```
## [1] 0.4367601
```

Avec 10 modèles différens nous obtenons un taux d'erreur moyen de 43%, et nous obtenons le modèle suivant.



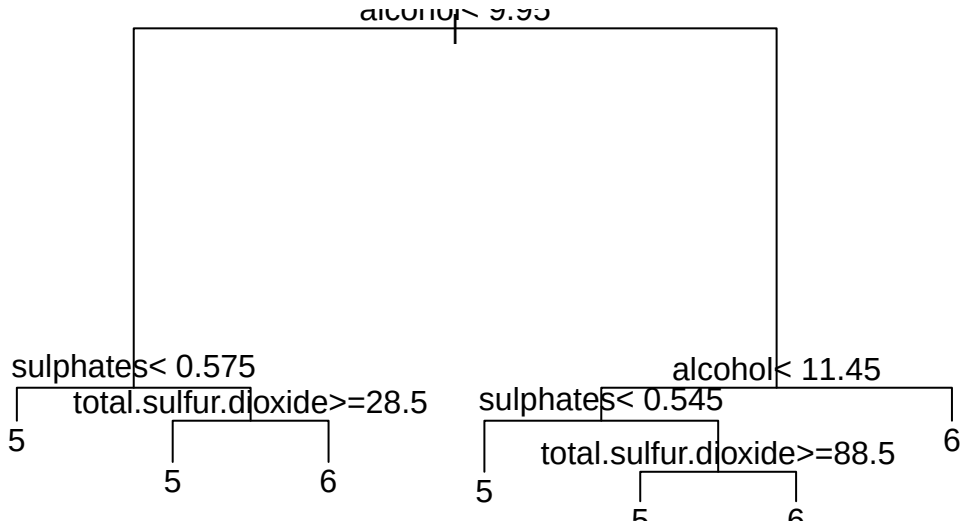
Le taux d'erreur obtenue peut s'expliquer par la répartition des notes pour ce jeu de données (graphique ci-dessous). Nous avons constaté qu'il y a majoritairement des vins de qualité moyenne (note entre 5 et 6). Cela a pour conséquence qu'il n'y a pas assez de données pour les autres notes et ainsi l'entrainement du modèle se concentre uniquement sur les vins de qualité moyenne.



Maintenant, nous allons refaire la même opération sur le même jeu de données, mais cette nous garderons uniquement les vins de qualité moyenne (note entre 5 et 6).

```
## [1] 0.3238636
```

En considérant, uniquement les vins de qualité moyenne nous avons un taux d'erreur moyen (pour 10 modèles différents) de 30%. Ainsi nous obtenons donc le modèle suivant :

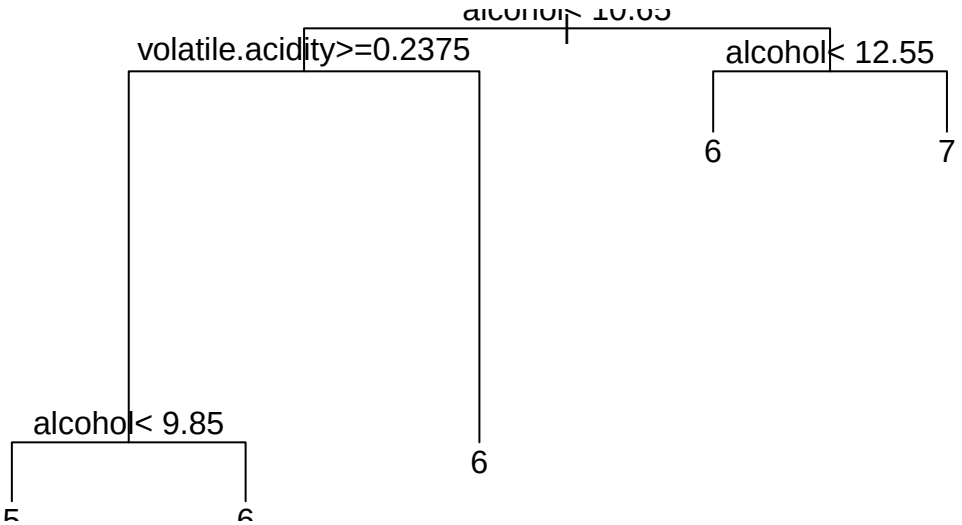


Avec ce modèle nous pouvons noter que le critère principal pour un bon vin rouge c'est sa teneur en alcool. Le second paramètre (moins impactant que le premier) est la teneur de sulfates dans le vin. Ce second paramètre ne fait qu'affiner la prédiction.

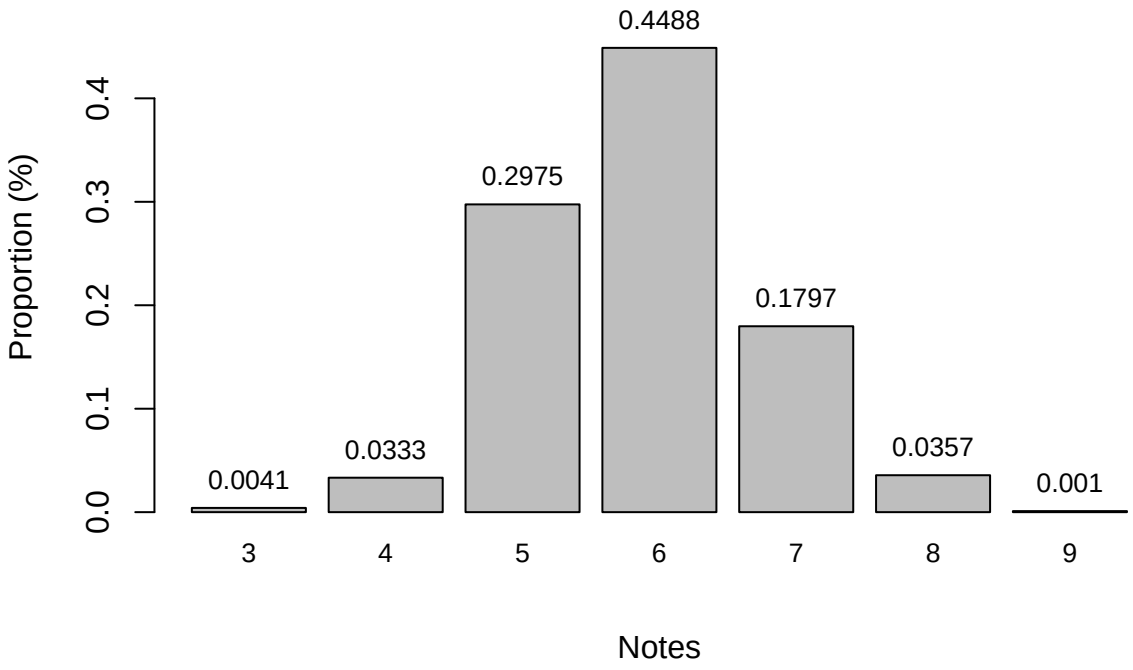
Passons maintenant au cas des vins blancs.

```
## [1] 0.4748216
```

Avec 10 modèles différens nous obtenons un taux d'erreur moyen de 47%, et nous obtenons le modèle suivant.



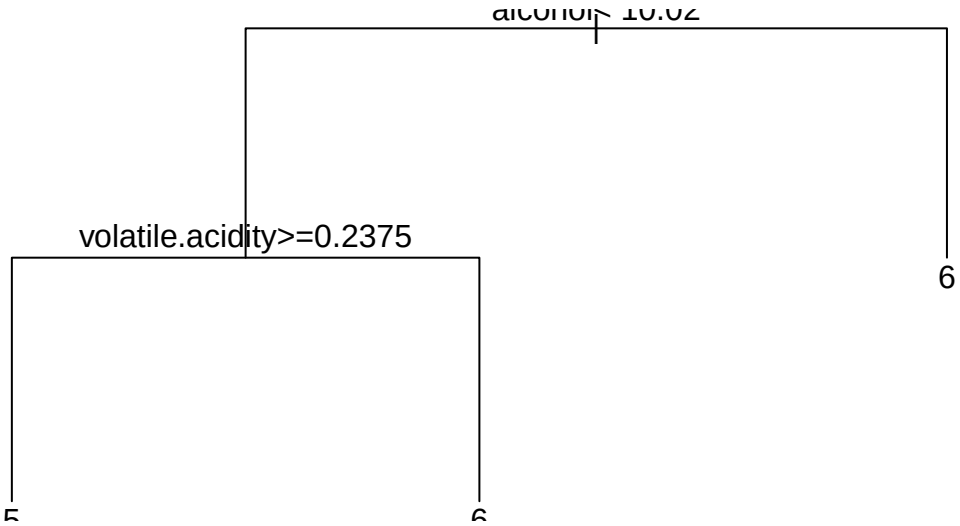
Le taux d'erreur obtenue peut s'expliquer par la répartition des notes pour ce jeu de données (graphique ci-dessous). Nous avons constaté qu'il y a majoritairement des vins de qualité moyenne (note entre 5 et 6). Cela a pour conséquence qu'il n'y a pas assez de données pour les autres notes et ainsi l'entrainement du modèle se concentre uniquement sur les vins de qualité moyenne.



Maintenant, nous allons refaire la même opération sur le même jeu de données, mais cette nous garderons uniquement les vins de qualité moyenne (note entre 5 et 6).

```
## [1] 0.304104
```

En considérant, uniquement les vins de qualité moyenne nous avons un taux d'erreur moyen (pour 10 modèles différents) de 30%. Ainsi nous obtenons donc le modèle suivant :



Avec le modèle obtenu, nous pouvons dire que ce qui fait un bon vin blanc c'est la teneur en alcool et l'acidité volatile.

5 Conclusion

Au cours de cette étude, nous avons travailler avec des données sur des vins de la région Vinho Verde, afin d'aider les viticulteurs à améliorer leurs vins. Nous avons vu, lors de la phase "Classification non supervisée", que de manière générale: qu'il est possible de classifier les vins en trois classes : médiocre, moyen, bon. Nous avons également noté que la majorité des vins de la région Vinho Verde sont de qulaité moyenne et qu'il y a peu de vins médiocres et peu de bons vins. Ainsi, les vins peuvent être améliorés de manière significative. Lors de la phase "Classification supervisée", nous avons montré que ce qui fait un bon vin est principalement la teneur en alcool de ce dernier. L'une des pistes d'amélioration que nous pouvons ainsi proposer aux viticulteurs, est d'allonger le temps de fermentation du vin. Une autre piste d'amélioration que nous pouvons donner (suite à cette étude), est qu'il pourrait être intéressant d'avoir une plus grande teneur d'acide volatile (ou acide gras).