ensiie
école nationale supérieure d'informatique
pour l'industrie et l'entreprise

Rapport du

Projet de modèle linéaire généralisé

Malek BEN NEYA

Karim ASSAAD

2016

Sommaire

INTRODUCTION 3
Analyse descriptive 4
Diagrammes en secteur et en bar 4
Boxplots 4
Boxplots unidimensionnels 4
Nettoyage des données 5
Boxplots bidimensionnels 5
Diagramme de chaleur 6
Modélisation 6
Régression binomiale 6
Modèle Full 7
Modèle des variables significatives 7
Modèle de STEPWISE 8
AIC 8
BIC 9
Calcule des performances 10
Régression multinomiale 10
Modèle Full 11
Modèle de STEPWISE 11
AIC 11
BIC 11
Calcule des performances 12
Analyse des résultats sur la qualité du vin 12
Conclusion 12

INTRODUCTION

Le jeu de donnée que nous avons utilisé est issu d'un article d'étude statistique sur le vin produit au nord du Portugal dans la région de *vinho* verde publié en 2009.

Il regroupe des données basées sur des tests physicochimiques qui sont basés sur des variables d'entrée et une donnée sensorielle qui représente la variable à expliquer.

Le jeu de donnée est une matrice de 12 variables et 1599 individus qui sont présentés comme suit :

1 - Fixed acidity : Le taux d'acide présent dans le vin
2 - Volatile acidity : La volatilité de l'acide
3 - Citric acid : La concentration de l'acide citrique
4 - Residual sugar : Le taux de sucre présent dans le vin
5 - Chlorides : Le taux de chlorure présent dans le vin
6 - Free sulfur dioxide : La quantité d'antioxydant libre dans le vin
7 - Total sulfur dioxide : La quantité totale de dioxyde de soufre ajoutée dans le vin
8 - Density : Le volume d'eau présent dans le vin
9 - PH : La mesure de la force de l'acidité du vin
10 - Sulphates : Le sel de l'acide sulférique
11 - Alcohol :le taux d'alcool présent dans le vin
12 - Quality : des scores entre 0 et 10 pour estimer la qualité du vin

En observant le jeu de données, nous avons constaté qu'ils n'existent pas de valeurs manquantes qui nécessitent un traitement spécial.

L'objectif de notre étude est donc d'étudier les variables explicatives et leurs impacts sur la qualité du vin en utilisant des études analytiques de régression afin de réaliser des modèles prédictifs.

Pour ce fait nous avons visualisé le comportement des différentes variables par le biais d'une analyse descriptive détaillée. Apres avoir étudié et analysé nos variables prédictives et spécifier leurs liaison avec la variable réponse, nous avons essayé d'employer les modèles vue en cours pour effectuer de nombreux modèles prédictifs, nous avons conclus qu'aucun modèle ne correspond parfaitement à notre jeu de données et n'a atteint la performance attendue.

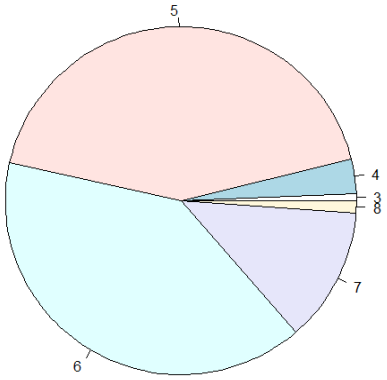
Nous nous sommes donc approfondie dans la compréhension de notre projet et réaliser d'autres modèles qui seront détaillés par la suite.

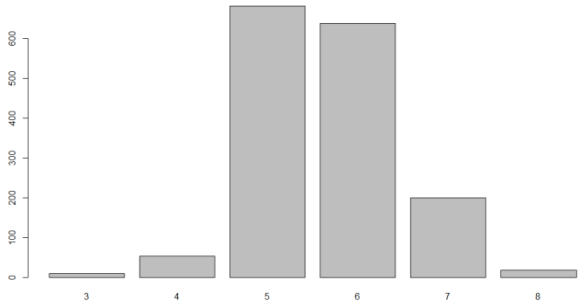
Dans un premier temps, nous appliquerons le modèle logistique binomiale, ensuite nous aurons recours a un modèle multinomiale.

Cette partie est une suite de l'étude présentée dans le premier rapport.

Analyse descriptive

Diagrammes en secteur et en bar



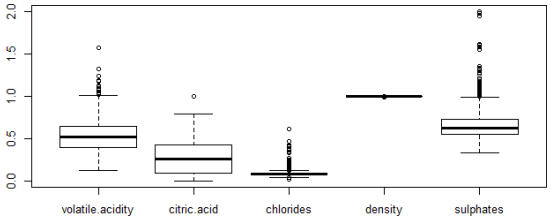


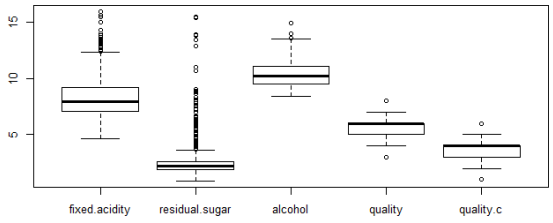
Le diagramme en bar a une allure de loi binomiale. D'après les deuxx diagrammes,nous observons une concentration aux niveaux des scores 5 et 6 donc la qualité d'une grande partie du vin est une qualité moyenne.

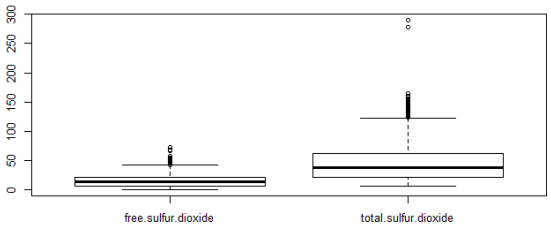
Boxplots

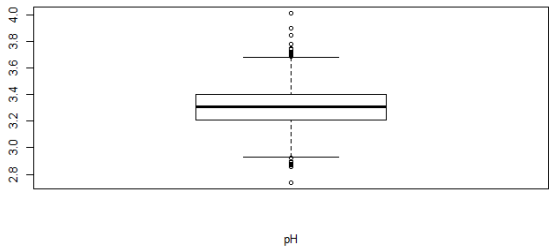
Boxplots unidimensionnels

Nous avons visualisé les boxplot s des variables qui sont presque sur la même échelle pour bien mettre en évidence leurs valeurs caractéristiques (quartiles, médianes, valeurs aberrantes) :









Pour les variables ' volatile acidity,' et 'sulphates ' les allures des boites à moustache sont à peu pré similaires, en effet les médianes pour ces deux variables sont aux alentours de 5, les valeurs maximales aux alentours de 10 et minimales aux environs de 3.

Les varibles 'fixced acidity' et 'quality ' se ressemblent, leurs médianes se rapprochent de 7.Nous détectons L'existence de valeurs aberrantes donc nous allons les détruire

Nettoyage des données

Avant toute manipulation des données, nous avons pris le soin de nettoyer les valeurs aberrantes les plus apparentes. Nous avons donc visualisé nos box plots pour les variables et précisé des seuils pour chaque variable avec une valeur minimal et une valeur maximale.

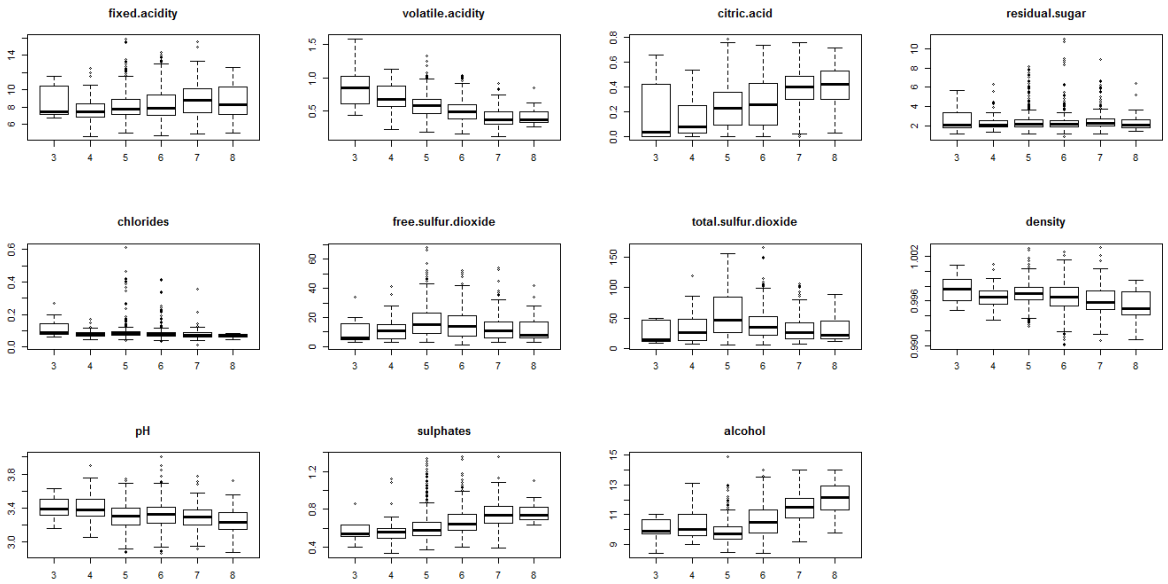
De ce fait, les point qui seront au delà ou au dessus de l'intervalle établie seront supprimés et ce qui contribuera à l'amélioration de notre prédiction par la suite.

Naturellement, il est impossible de détruire toutes les valeurs aberrantes qui existantes dans la base d'autant plus que le risque de perte d'information si jamais nous supprimons plusieurs lignes d'individus ou une variable à effet.

Après le nettoyage nous avons eu une légère modification de l'apparence des box plots avec l'apparition de nouvelles valeurs aberrantes.

Nous ne pouvons donc pas les supprimer infiniment donc nous nous contentons par ce résultat pour l'instant.

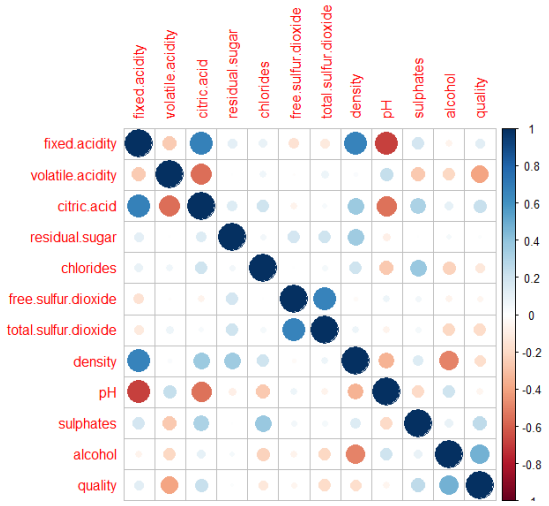
Boxplots bidimensionnels



Nous avons effectué la présentation des boites à moustaches de la variable réponse 'quality ' sous sa forme factorielle en fonction de chaque variable explicative afin de mieux comprendre l'effet de ces dernières sur 'quality'. Après le nettoyage des valeurs aberrantes nos figures se sont légèrement modifiées. Nous remarquons que dans le cas de la 'volatile acidity' les boxplot sont linéaires par médianes et plus la volatilité de l'acide diminue plus la qualité du vin augmente.

Pour la variable 'citric acid' la variation et proportionnelle, une augmentation de cette variable entraine l'augmentation de la qualité donc un vin de bon qualité implique une quantité plus au moins élevée de cet acide.

Diagramme de chaleur



Le diagramme de chaleur démontre la corrélation entre les différentes varibles, ici nous apercevons une forte corrélation positive de l'ordre de 0.8 entre le 'cirtric acid' et la 'fixed acidity'.
La 'density' est fortement corrélée avec la 'fixed acidity' avec une valeur de 0.7 donc ces deux varibles peuvent apporter la même information. Le 'ph 'est anti corrélé avec la 'fixed acidity', nous constatons aussi que la qualité dépond fortement du taux de l'alcool avec une corrélation de 0.6.

Modélisation

Régression binomiale

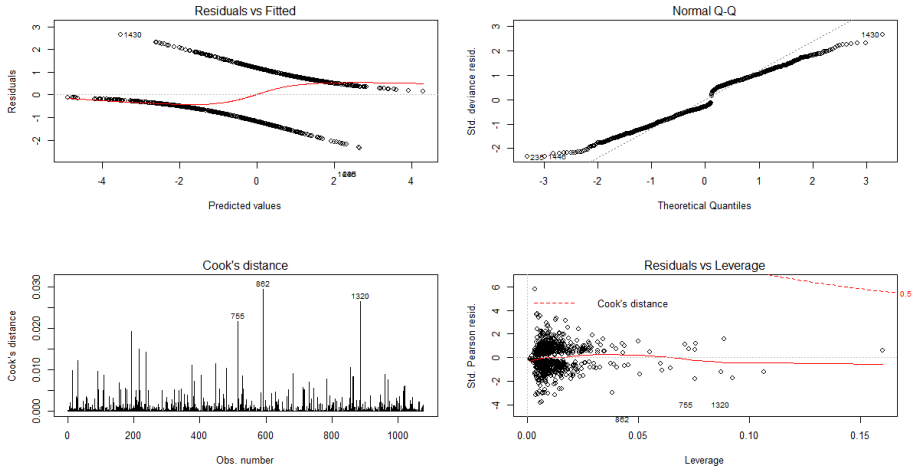
Nous avons choisi en premier temps d'appliquer le modèle logistique binomiale afin de voir le résultat qu'il donne et mieux assimiler le principe de cette modélisation.

Pour ce but, nous avons transformé notre variable réponse 'quality' qui est définie par des scores entre 3 et 8 par une varible qualitative binaire dont les valeurs sont 0 et 1.

Si la qualité du vin est inférieure à 6, elle se voit attribuer la valeur 0, sinon la valeur 1.

Nous avons tout d'abord essayé le modèle full pour pouvoir en déduire les variables les plus significatives.

Modèle Full



La courbe des valeurs prédites en fonctions des résidus, indique que le modèle ne satisfait pas nos jeux de données et qu'il faudra spécifier les variables pertinentes.

Matrice de confusion

```
      0   1
  0 201  57
  1  74 172
```

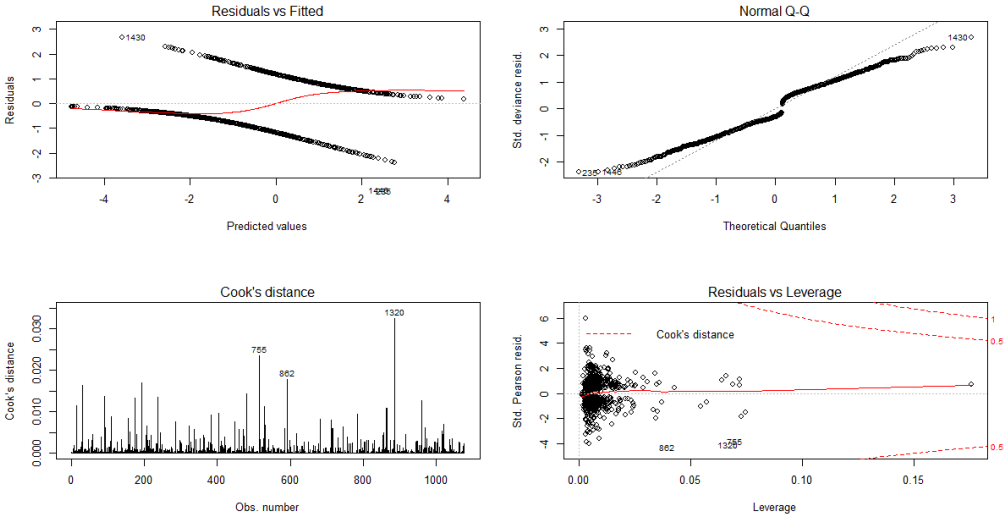
La matrice de confusion de ce modèle indique qu'il a réussi à prédire 201 valeur de 0 qui était correct et 172 valeurs de 1 mais qu'il à échouer prédire 74 valeurs de 1 qu'il à prédite 0 et 57 valeurs de 0 qu'il a prédite 1 ce qui ne le rend pas très performant en terme de prédiction.

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)          -79.671717 102.502262  -0.777   0.4370
fixed.acidity         -0.176570   0.126878  -1.392   0.1640
volatile.acidity       3.393881   0.604114   5.618 1.93e-08 ***
citric.acid            1.738489   0.701540   2.478   0.0132 *
residual.sugar        -0.020277   0.088547  -0.229   0.8189
chlorides              4.810671   2.086756   2.305   0.0211 *
free.sulfur.dioxide   -0.023788   0.011071  -2.149   0.0317 *
total.sulfur.dioxide   0.017760   0.003872   4.587 4.50e-06 ***
density               88.222133 104.726362   0.842   0.3996
pH                     0.573239   0.930452   0.616   0.5378
sulphates             -4.117024   0.652136  -6.313 2.73e-10 ***
alcohol               -0.901435   0.132564  -6.800 1.05e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En étudiant le summary de ce modèle nous avons constaté que ces varibles sont 'volatile.acidity', 'total.sulfur.dioxide' ,'sulphates', 'alcohol' , 'free.sulfur.dioxide' , 'chlorides' et le ' citric.acid' **.**

Nous avons donc construit un deuxième modèle à partir de ces données.

Modèle des variables significatives



Notre deuxième modèle est légèrement moins performant que le premier en terme de prédiction mais il est moins complexe ce qui le rend le meilleur jusqu' présent.

Matrice de confusion

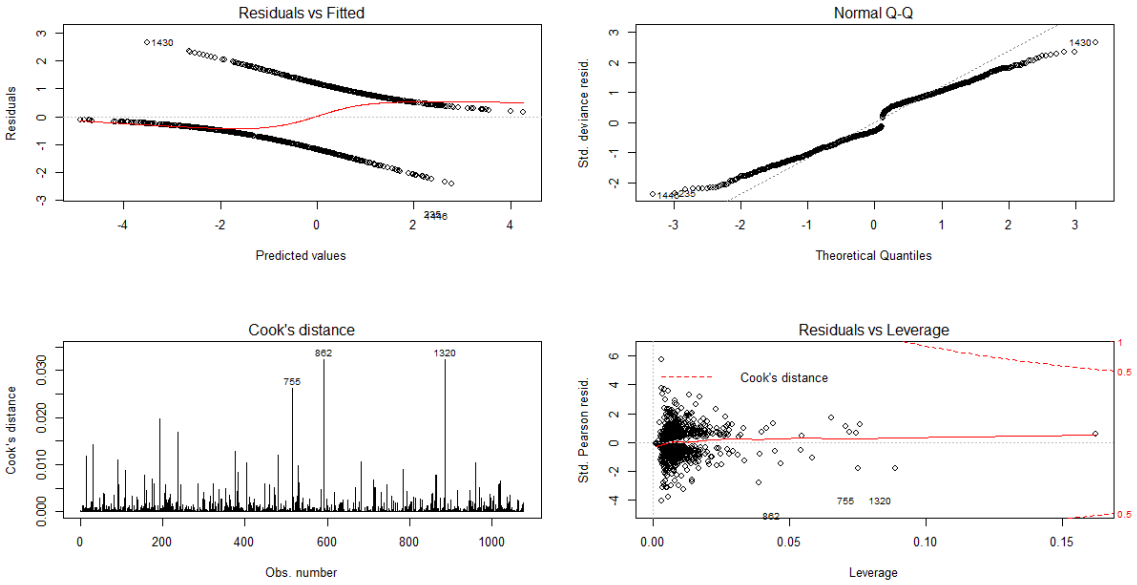
```
     0   1
0  199  59
1   77 169
```

Les valeurs présentes sur la diagonale de la matrice de confusion indiquent que ce modèle réussi prédire 199 valeurs correctes de 0 contre 59 fausses et 169 valeurs correctes de 1 contre 77 fausses.

Modèle de STEPWISE

Afin de mieux choisir les variables prédictives nous avons eu recours à la modélisation de stepwise qui nous a permis par la suite de choisir le modèle le plus robuste.

AIC



La figure ci-dessous décrit les résultats tournés par le modèle AIC, selon la distance de cook ils existent trois valeurs aberrantes qu'il faut supprimer qui sont les individus 755,862 et 1320.

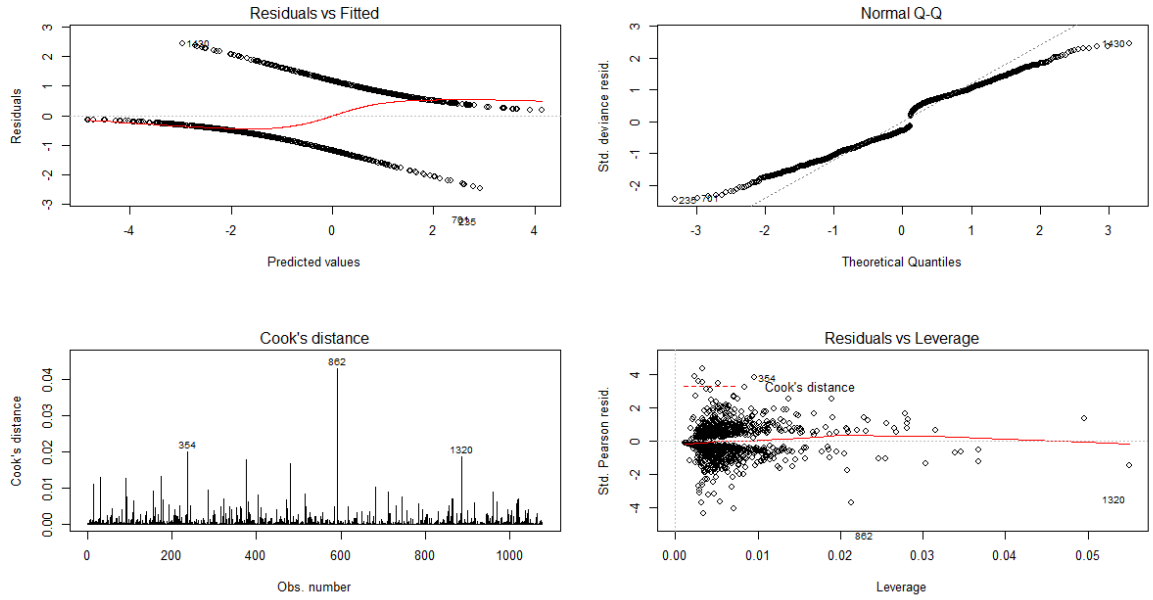
Les variables choisies par l'AIC sont la'fixed.acidity' ,'volatile.acidity', 'citric.acid','chlorides', 'free.sulfur.dioxide',' total.sulfur.dioxide' , 'density','sulphates' et 'alcohol'.

Matrice de confusion

```
     0   1
0  202  56
1   78 168
```

La matrice de confusion de ce modèle a donnée des résultats plutôt bons, en effet sa diagonale démontre que la prédiction à donné 202 valeurs correctes de 0 contre 56 fausses et 168 valeurs correcte de 1 contre 78 fausses.

BIC



D'après la figure, le modèle n'est pas tout à fait au point, mais il s'approche de plus de la performance attendue. Les variables les plus pertinentes selon ce modèle sont donc 'fixed.acidity','volatile.acidity' 'citric.acid','total.sulfur.dioxide', 'sulphates' et 'alcohol',

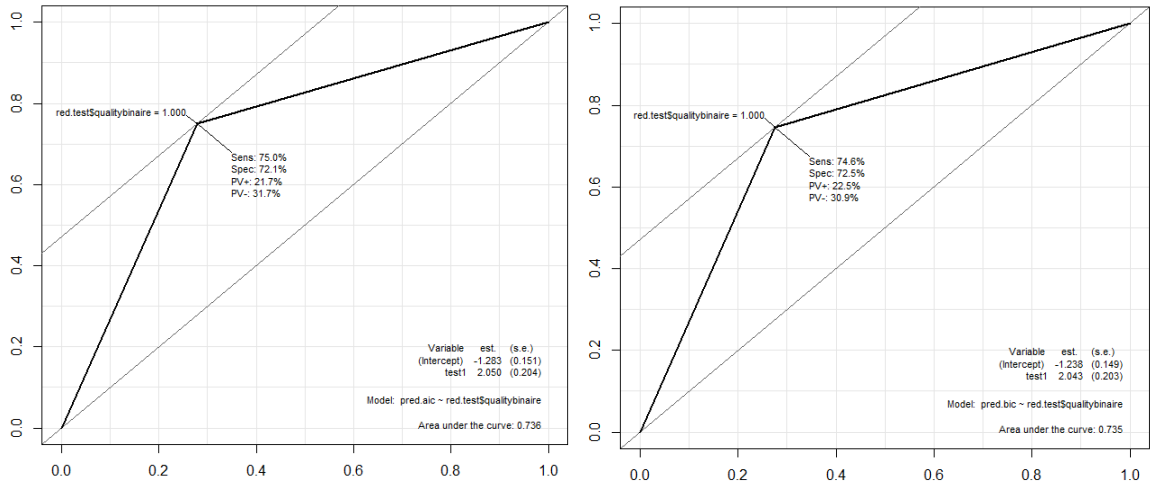
Matrice de confusion

```
    0   1
0  200  58
1  76  170
```

Nous avons vérifié par la matrice de confusion que les résultats prédites restent aussi bonnes que celles prédites par l'AIC, pour ce faite nous allons nous appuyer sur la courbe ROC pour essayer de choisir le modèle qui convient au mieux à notre jeu de donnée.

Courbe ROC

La courbe ROC est une bonne méthode pour évaluer la qualité du modèle. En effet plus l'aire sous la courbe ROC est proche de 1, meilleur est le modèle.



L'AUC du AIC est de 0.736 tandis que celui du BIC est de 0.735 ce qui est une valeur très proche de 1 donc nous confirmons que sont les deux de bons modèles en terme de performance et taux d'erreur minimale.

Calcule des performances

Pour chaque modèle nous avons calculé la performance en divisant la somme des valeurs de la diagonales des matrices de confusions obtenues par chaque modèle par la sommes de toutes les valeurs de la matrices correspondante, ainsi nous avons obtenus le tableau suivant qui nous permet de comparer les performances.

	MODELE FULL	MODELE 2	MODELE AIC	MODELE BIC
PERFORMANCE	0.7400794	0.7301587	0.734127	0.734127

Comparaison faite, et en tenant comptes des différentes matrices de confusions, nous avons conclu que le BIC est un bon compromis puisqu'il est plus simple à étudier vu le nombre de prédicateurs qu'il utilise ainsi qu'une performance assez élevée.

Cette régression logistique binaire nous a permis d'améliorer notre performance mais elle nous a fait perdre de l'information puisque les six scores de la qualité sont présentés par seulement deux valeurs.

Régression multinomiale

La régression multinomiale vise à construire un modèle permettant d'expliquer les valeurs prises par une variable cible qualitative qui possède plus de 2 modalités ce qui correspond parfaitement notre cas puisque notre variable 'quality' prend 6 valeurs.

Modèle Full

Nous avons dans un premier temps effectué un modèle qui englobe toutes les varibles explicatives pour pouvoir par la suite choisir les plus pertinentes en utilisant les critères vues en cours.

Matrice de confusion

La matrice de confusion de ce modèle nous a retourné le résultat suivant :

	3	4	5	6	7	8
3	0	1	2	1	0	0
4	0	0	9	6	0	0
5	0	1	159	66	1	0
6	0	0	57	119	18	4
7	0	0	4	29	24	0
8	0	0	0	2	1	0

Pour avoir la performance de ce modèle à partir de la matrice de confusion il suffit de faire la somme des valeurs obtenues sur la diagonale divisées par la sommes de toutes les valeurs de la matrice, notre performance est donc de 0.5992063, qui n'est pas une valeur très satisfaisante mais c'est normal puisque nous utilisons le modèle full.

Modèle de STEPWISE

AIC

Pour choisir nos variables nous avons appliqué le critère de l'Aic sur notre modèle de toutes les varibles et celui la a choisi les prédicteurs suivants : 'volatile.acidity', 'chlorides','total.sulfur.dioxide', 'pH', 'sulphates' ,'alcohol' et 'free.sulfur.dioxide'.

Matrice de confusion

La matrice de confusion de l'AIC nous a retourné le résultat présenté ci-dessous :

	3	4	5	6	7	8
3	0	1	2	1	0	0
4	0	0	9	6	0	0
5	0	1	162	63	0	1
6	0	0	57	121	19	1
7	0	0	4	31	22	0
8	0	0	0	2	1	0

Nous remarquons que les qualités "5" et "6" sont les plus qui ont été prédis. Il ya 162 5 prédis 5 en réalités contre 63 "6" prédis "5" en réalité .Il ya aussi 121 "6" prédis "6" en réalité contre 57 "5" prédis 6 en réalité. Ce qui semble normale vu nos données puisque nous avons plus d'échantillon de qualités moyennes que de qualité mauvaises et bonnes

BIC

Le modèle BIC a permis d'obtenir les varibles 'volatile.acidity', 'chlorides','total.sulfur.dioxide', 'pH', 'sulphates' ,'alcohol' et 'free.sulfur.dioxide' qui sont les mêmes choisies par le AIC.

Matrice de confusion

La matrice de confusion de l'BIC nous a retourné le résultat présenté ci-dessous :

	3	4	5	6	7	8
3	0	1	2	1	0	0
4	0	0	9	6	0	0
5	0	1	162	63	0	1

```
  6    0    0   57  121   19    1
  7    0    0    4   31   22    0
  8    0    0    0    2    1    0
```

La matrice de confusion nous renvoi à peu prés les mêmes prédictions que le modèle précédent et c'est tout à fait attendu puisque les deux modèles ont choisi les mêmes variables prédictives.

Calcule des performances

	MODELE FULL	MODELE AIC	MODELE BIC
PERFORMANCE	0.5992063	0.6051587	0.6051587

Nous remarquons que les critères de sélections de variables AIC et BIC nous renvoi les mêmes variables : l'acide volatile, les chlorites, le total dioxyde de sulfure, le pH, les sulfates, l'alcool et le free dioxyde de sulfure. Donc nous pouvons conclure que ces variables sélectionnées sont les plus influentes sur la qualité du vin.

Analyse des résultats sur la qualité du vin

Nous remarquons d'après notre étude que la 'volatile acidity' a donc un effet inverse sur la qualité du vin, en effet plus la volatilité de l'acide diminue plus la qualité du vin augmente. Pour la variable l'alcool, la variation et linéaire, quand cette variable s'élève, cela entraine l'augmentation de la qualité donc un vin de bon qualité nécessite une bonne quantité d'alcool et de même pour l'acide citrique.

Conclusion

Nous avons employés les modèles logistique binaire et multinomiales et avons obtenue la meilleure performance pour le modèle BIC de la régression logistique mais étant donnée la nature de notre variable réponse et les différente scores qu'elle présente, il est plus judicieux de choisir le modèle BIC de la régression multinomiale même si il est moins élevé en terme de performance.

En effet il représente un bon compromis en tenant comptes des critères d'un bon modèle puisqu'il offre une performance de plus de 60% et prends en considération le moindre nombre de variables.

Ces modèles ont aboutis à des résultats plus ou moins satisfaisant, mais si nous voulions avoir le meilleur modèle nous pouvons avoir recours à la puissante modélisation de maching learning tels que le SVM ou les Random Forest.