

PRÉDICTION DE LA QUALITÉ DU VIN ROUGE

EN UTILISANT DES
TECHNIQUES DE
MACHINE LEARNING

Prepared by
Aourik khalid

30/05/2022



TABLE OF CONTENTS

01 OVERVIEW

about us	03
our clients	04
Business understanding	05

02 DATA

data understanding	06
data preparation	07

03 STRATEGIC REPORT

EDA	08
modeling	23
evaluation	28
conclusion	29

QUI SOMME NOUS ?



Anacorp

Nous sommes une agence de conseil pour les entreprises. Grâce aux données fournies par les entreprises, nous faisons des analyses et des prédictions affinées pour mieux les conseiller pour divers projets.

NOS CLIENTS



Best companies of 2021



BUSINESS UNDERSTANDING

Le secteur du vin rouge connaît une croissance exponentielle depuis peu, en raison de l'augmentation de la consommation sociale. Aujourd'hui, les acteurs du secteur utilisent les certifications de qualité des produits pour promouvoir leurs produits. Il s'agit d'un processus qui prend du temps et nécessite l'évaluation d'experts humains, ce qui rend ce processus très coûteux. En outre, le prix du vin rouge dépend d'un concept plutôt abstrait d'appréciation du vin par les dégustateurs, dont l'opinion peut varier fortement. Un autre facteur essentiel de la certification et de l'évaluation de la qualité du vin rouge est constitué par les tests physico-chimiques, qui sont effectués en laboratoire et tiennent compte de facteurs tels que l'acidité, le niveau de pH, le sucre et d'autres propriétés chimiques. Il serait intéressant pour le marché du vin rouge que la qualité humaine de la dégustation puisse être liée aux propriétés chimiques du vin afin que les processus de certification et d'évaluation de la qualité soient mieux contrôlés. Ce projet vise à déterminer les caractéristiques qui sont les meilleurs indicateurs de la qualité du vin rouge et à générer un aperçu de chacun de ces facteurs pour la qualité du vin rouge de notre modèle.

DATA UNDERSTANDING

Mon analyse utilisera le Red Wine Quality Data Set, disponible sur le dépôt de l'UCI machine learning : (<https://archive.ics.uci.edu/ml/datasets/wine+quality>).

J'ai obtenu les échantillons de vin rouge du nord du Portugal pour modéliser la qualité du vin rouge sur la base de tests physico-chimiques. L'ensemble de données contient un total de 12 variables, qui ont été enregistrées pour 1 599 observations. Ces données nous permettront de créer différents modèles de régression pour déterminer comment les différentes variables indépendantes aident à prédire notre variable dépendante.

Connaître l'impact de chaque variable sur la qualité du vin rouge aidera les producteurs, les distributeurs et les entreprises de l'industrie du vin rouge à mieux évalué leur stratégie de production, de distribution et de tarification.

DATA PREPARATION

Nettoyage des données

La première étape a été de nettoyer et de préparer les données pour l'analyse.

Je suis passé par différentes étapes de nettoyage des données. Tout d'abord, j'ai vérifié les types de données en me concentrant sur les données numériques et catégorielles afin de simplifier le calcul et la visualisation de la corrélation.

Ensuite, j'ai essayé d'identifier toutes les valeurs manquantes existant dans notre ensemble de données.

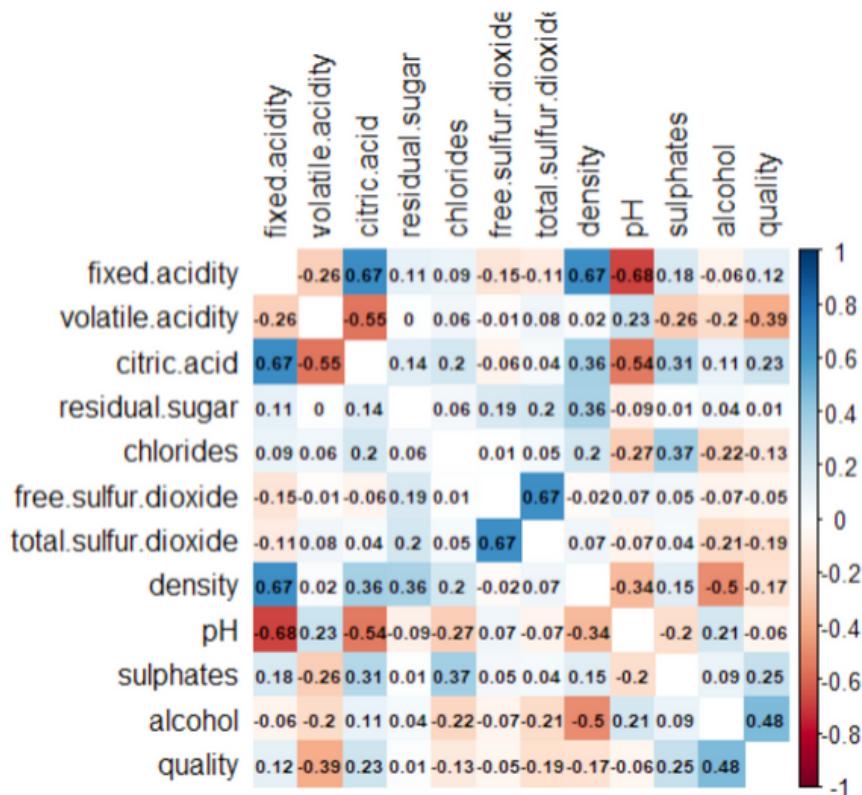
Enfin, j'ai recherché le résumé statistique de chaque colonne/caractéristique pour détecter tout problème tel que les valeurs aberrantes et les distributions anormales.

EDA

Exploration et transformation des données 1/15

Pour voir quelles variables sont susceptibles d'affecter le plus la qualité du vin rouge, j'ai effectué une analyse de corrélation de nos variables indépendantes avec notre variable dépendante (quality). Cette analyse a permis d'établir une liste de variables d'intérêt présentant les corrélations les plus élevées avec la qualité (Figure 1).

Figure 1. Correlation Matrix



EDA

Exploration et transformation des données 2/15

Dans l'ordre de la plus forte corrélation, ces variables sont :

1. Alcohol : la quantité d'alcool dans le vin
2. Volatile acidity: l'acide acétique est élevé dans le vin, ce qui donne un goût de vinaigre désagréable.
3. Sulphates : un additif du vin qui contribue aux niveaux de SO₂ et qui agit comme un antimicrobien et un antioxydant
4. Citric Acid : agit comme un conservateur pour augmenter l'acidité (de petites quantités ajoutent de la fraîcheur et de la saveur aux vins).

EDA

Exploration et transformation des données 3/15

5. Total Sulfur Dioxide : c'est la quantité de formes libres + liées de SO₂.

6. Density : les vins les plus doux ont une densité plus élevée.

7. Chlorides : la quantité de sel dans le vin.

8. Fixed acidity : acides non volatils qui ne s'évaporent pas facilement.

9. pH : le niveau d'acidité

10. Free sulfur Dioxide : il empêche la croissance microbienne et l'oxydation du vin.

EDA

Exploration et transformation des données 4/15

11. Residual sugar : c'est la quantité de sucre qui reste après l'arrêt de la fermentation. La clé est d'avoir d'avoir un équilibre parfait entre la douceur et l'acidité (les vins > 45g/ltrs sont sucrés).

En commençant par notre variable dépendante, la qualité, j'ai pu observer les valeurs moyennes de la qualité : 5 et 6. En considérant la transformation de la variable dépendante, j'ai constaté que nos données sont normalement distribuées (figure 2). Cette conclusion peut être vérifiée en effectuant un QQ plot, qui montre qu'il n'est pas nécessaire de transformer nos données (Figure 3)

EDA

Exploration et transformation des données 5/15

Figure 2.

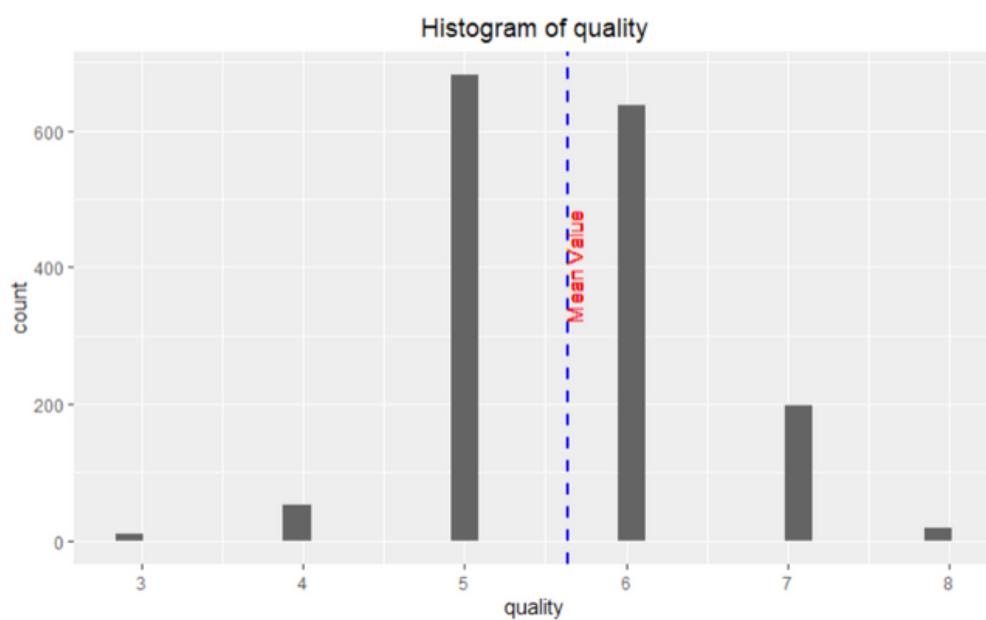
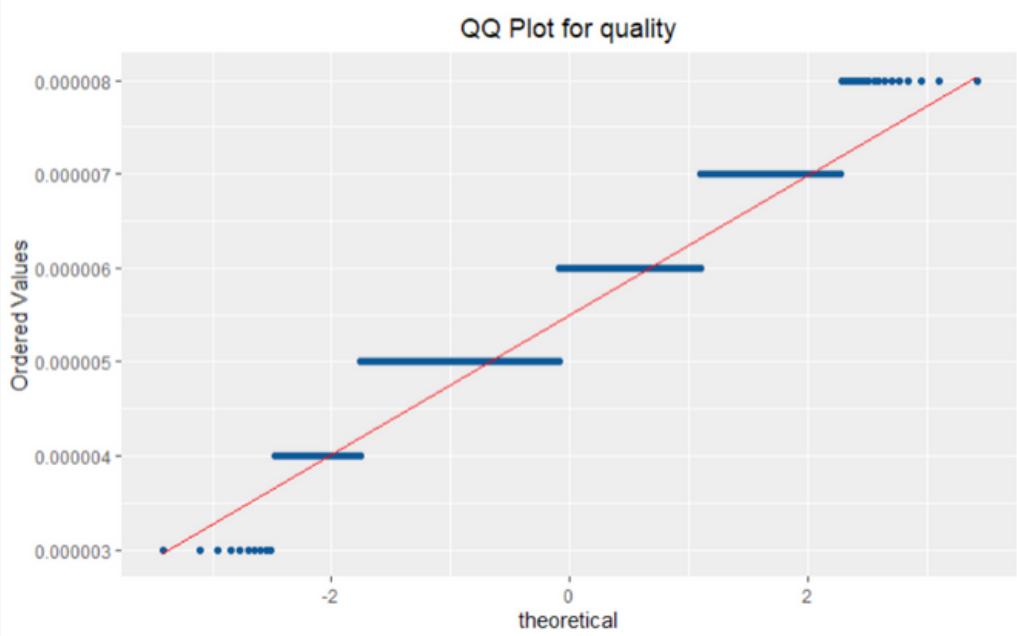


Figure 3.



EDA

Exploration et transformation des données 6/15

Ensuite, pour les variables numériques indépendantes, la première étape pour analyser plus en profondeur la relation avec notre variable dépendante a été de créer des diagrammes de densité visualisant la dispersion des données (Figure 4, 5, 6).

EDA

Exploration et transformation des données 7/15

Figure 4.

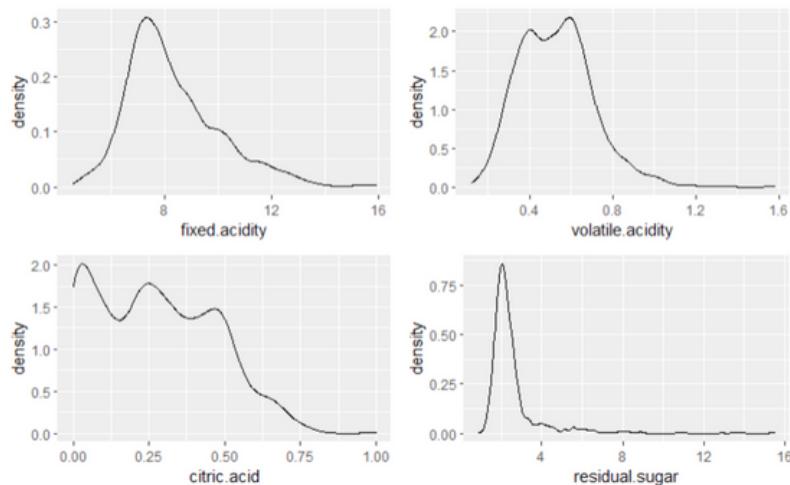


Figure 5.

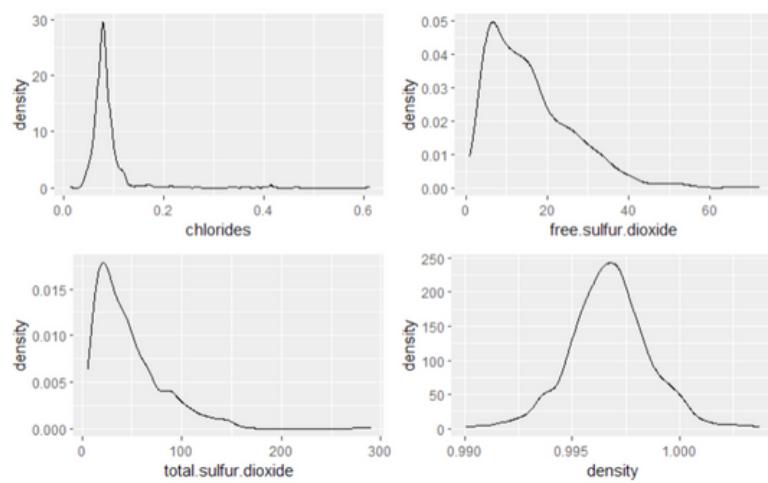
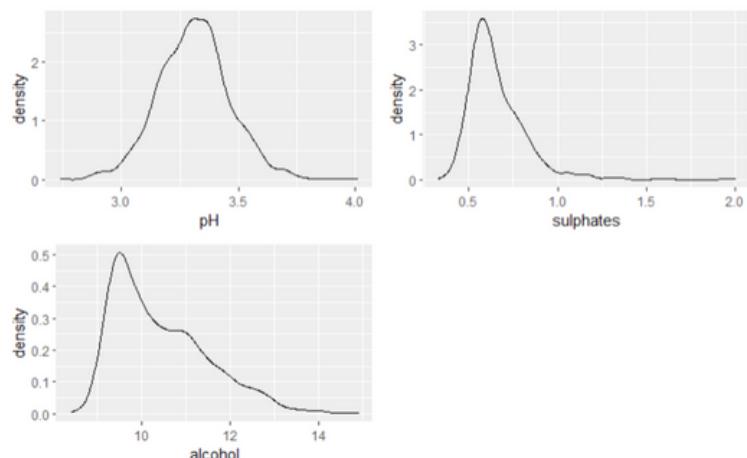


Figure 6.



EDA

Exploration et transformation des données 8/15

On peut constater que les niveaux de pH de la plupart des vins rouges se situent toujours entre 3 et 4 et que les chlorures - la quantité de sel est la plus répandue au niveau 0,1. Après avoir analysé les diagrammes de densité, j'ai tracé l'interaction entre nos variables numériques et notre variable dépendante de qualité (Figure 7, 8, 9).

EDA

Exploration et transformation des données 9/15

Figure 7.

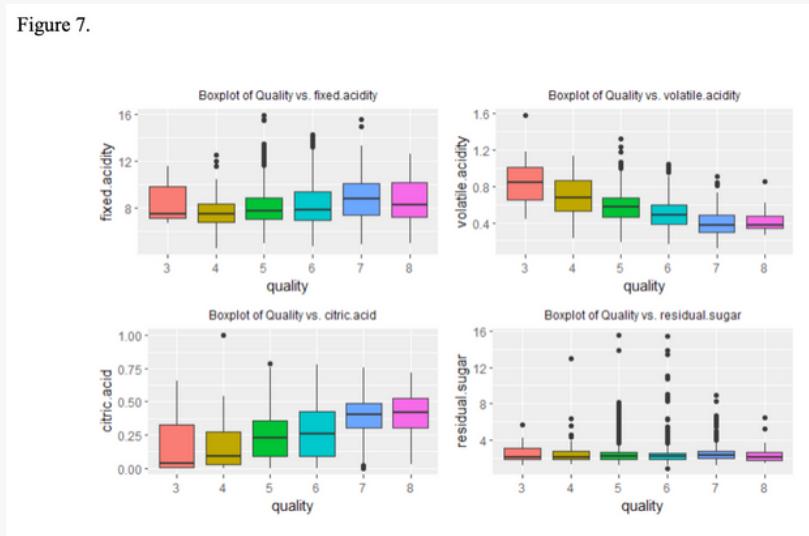


Figure 8.

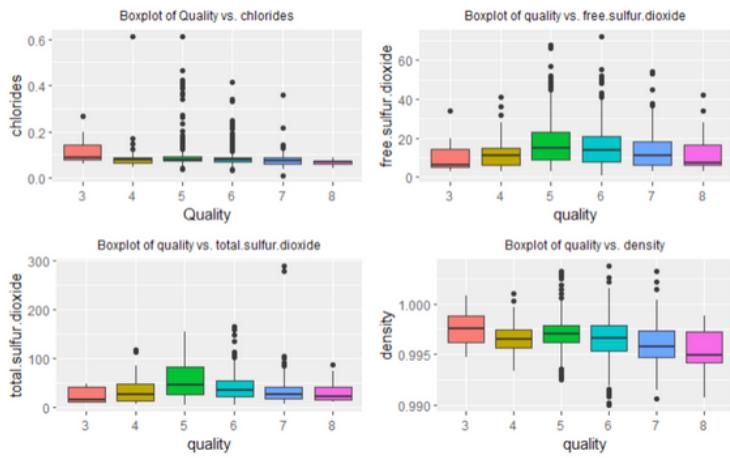
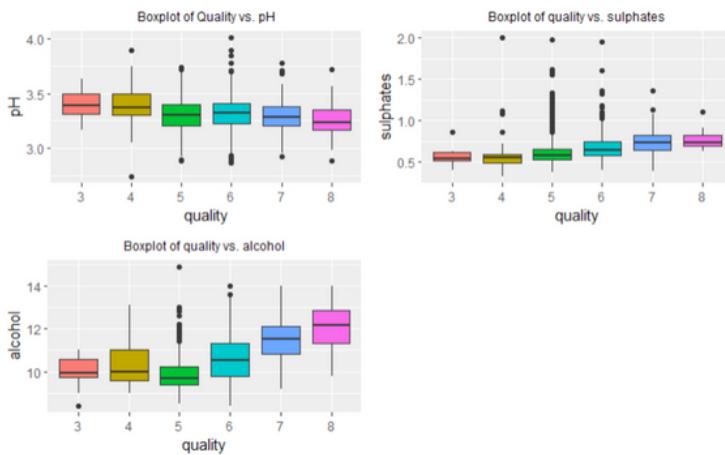


Figure 9.



EDA

Exploration et transformation des données 10/15

Trois modèles différents peuvent être observés. Premièrement, il existe des relations positives entre la qualité et les critiques acides, alcool et sulfates. Même si les vins ayant un niveau d'alcool plus élevé peuvent les rendre moins populaires, ils devraient être bien notés en termes de qualité. Deuxièmement, il existe des relations négatives entre la qualité et l'acidité volatile, la densité et le pH. Il est raisonnable que les vins moins sucrés et ayant un niveau d'acidité plus faible soient favorisés dans les tests de qualité.

EDA

Exploration et transformation des données 11/15

Enfin, les variables indépendantes suivantes ne présentent aucune relation significative avec la qualité : sucre résiduel, chlorures et dioxyde de soufre total.

Pour approfondir les relations au sein des variables indépendantes et avec la qualité, j'ai construit différents graphiques tridimensionnels. En inspectant les deux variables, l'alcool et l'acidité volatile avec la qualité (Figures 10), nous pouvons voir qu'avec un taux d'alcool des vins rouges entre 9% et 12%, le niveau d'acidité volatile diminue à mesure que le taux d'alcool des vins augmente. Pour un taux d'alcool plus élevé (>12%), la tendance s'inverse, ce qui implique la popularité des vins de haute qualité. En continuant à rechercher la variable alcool, j'ai sélectionné l'acide citrique et visualisé leurs interactions avec la qualité (figure 11).

EDA

Exploration et transformation des données 12/15

Figure 10.

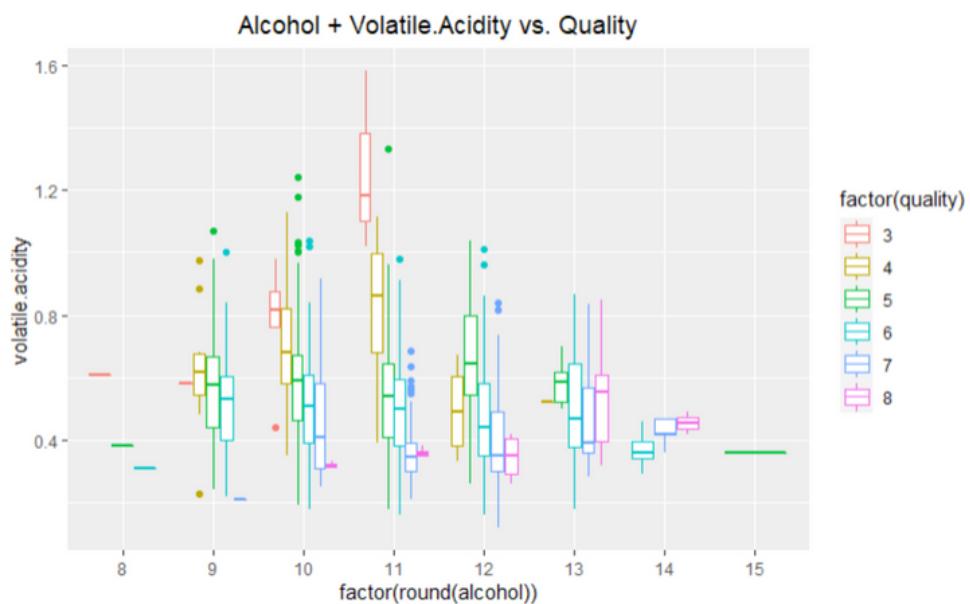
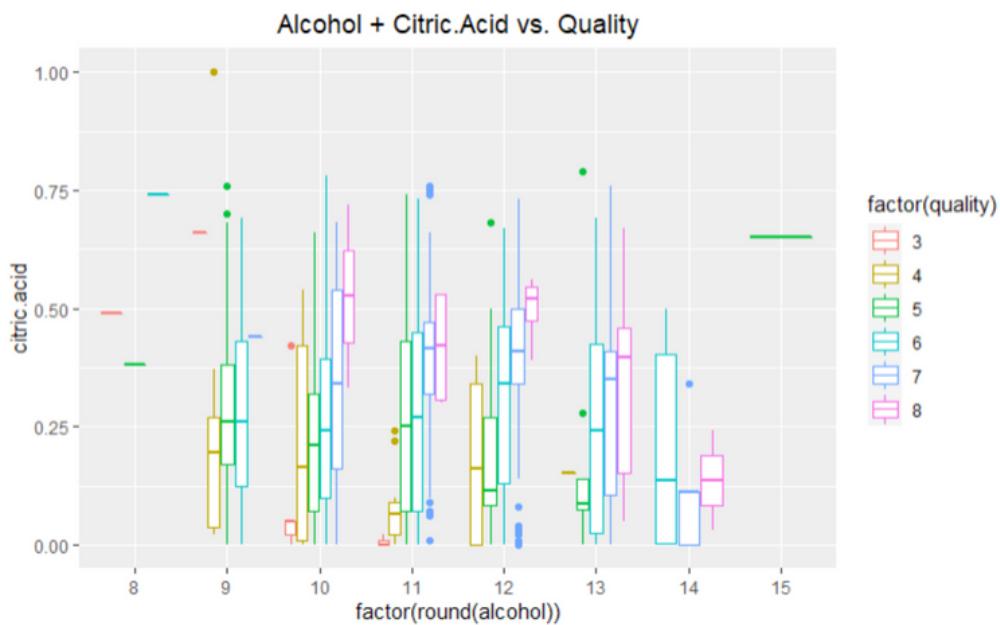


Figure 11.



EDA

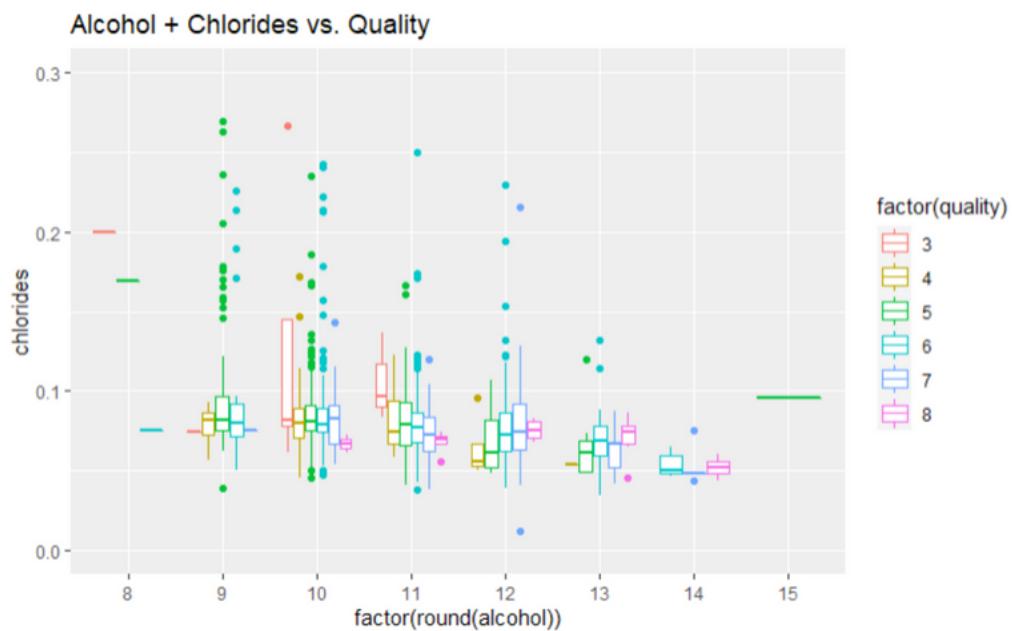
Exploration et transformation des données 13/15

Il est intéressant de noter que pour les vins dont le pourcentage d'alcool est inférieur à 14, plus le niveau d'acide citrique augmente, plus la qualité des vins rouges augmente. La seule exception est à 14% d'alcool, où le niveau d'acide citrique diminue à mesure que la qualité du vin augmente. Enfin, une analyse d'interaction utilisant les chlorures dans les relations avec l'alcool et la qualité montre que la qualité des vins diminue lorsque le niveau de chlorure diminue à l'alcool avant 12%. Cependant, la qualité du vin rouge augmente lorsque le niveau de chlorure augmente au niveau de l'alcool à partir de 12% (Figure 12).

EDA

Exploration et transformation des données 14/15

Figure 12.



EDA

Exploration et transformation des données 15/15

Enfin, nous avons examiné si le problème de colinéarité existait dans nos données. Suite à l'analyse des corrélations et à la vérification du VIF, nous avons découvert certaines variables présentant des corrélations légèrement élevées. Pour traiter ce problème potentiel, nous allons tirer parti de la technique de régularisation LASSO dans la prochaine partie de la modélisation.

MODELING

Sur la base de l'EDA et de l'analyse de corrélation, trois modèles potentiels ont été utilisés dans la partie modélisation.

Modèle 1 : L'analyse de corrélation montrant que la qualité est fortement corrélée avec un sous-ensemble de variables (notre "Top 5"), j'ai utilisé la régression multi-linéaire pour construire un modèle optimal.

modèle de prédiction de la qualité du vin rouge. En retirant une variable indépendante non significative du modèle initial, nous obtenons le "Modèle 1", qui inclut nos "4 principales" variables explicatives. En utilisant la validation croisée K-Fold, nous avons le résumé du modèle 1 comme ci-dessous :

```
Residuals:
    Min      1Q   Median      3Q      Max 
-2.72716 -0.38486 -0.06503  0.44980  2.13257 

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)    
(Intercept) 2.8258128  0.2006892 14.081 < 0.0000000000000002 *** 
alcohol       0.2953105  0.0160331 18.419 < 0.0000000000000002 *** 
volatile.acidity -1.1985632  0.0966011 -12.407 < 0.0000000000000002 *** 
sulphates      0.7121396  0.1005146   7.085  0.000000000000208 *** 
total.sulfur.dioxide -0.0022354  0.0005108  -4.376  0.00001284518270 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.655 on 1594 degrees of freedom

1599 samples
 4 predictor

No pre-processing
Resampling: cross-validated (10 fold)
Summary of sample sizes: 1438, 1439, 1440, 1438, 1439, 1439, ...
Resampling results:

RMSE      Rsquared      MAE
0.6549281  0.3479475  0.5092899
```

MODELING

Dans le modèle 1, toutes les variables identifiées sont fortement corrélées avec notre variable cible (la qualité) et présentent une signification statistique. L'alcool et les sulfates ont des relations positives avec la qualité, ce qui implique que plus le niveau d'alcool et de sulfates est élevé, plus la qualité du vin rouge est élevée. Inversement, il existe des relations négatives entre l'acidité volatile et le dioxyde de soufre total et la qualité, ce qui montre que les gens s'attendent à un faible niveau d'acide acétique et de SO₂ dans un vin de qualité. Une grande quantité d'acide acétique peut entraîner un goût de vinaigre désagréable, par exemple.

MODELING

Modèle 2 : Ensuite, à l'aide de la méthode LASSO, j'ai créé le deuxième modèle ("Modèle 2") qui effectue à la fois une sélection de variables et une régularisation. Il en résulte un sous-ensemble de prédicteurs (notre "Top 6") qui minimise l'erreur de prédiction pour une variable de réponse quantitative - la qualité. Ce sous-ensemble comprend six variables : fixed.acidity, volatile.acidity, chlorides, total.sulfur.dioxide, sulphates, and alcohol.

En appliquant à nouveau la validation croisée K-Fold, nous avons obtenu le résumé du modèle 2 comme ci-dessous :

```
Residuals:
    Min      1Q   Median      3Q     Max 
-2.70812 -0.37181 -0.06238  0.45933  1.99472 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.7365412  0.2325021 11.770 < 0.000000000000002 *** 
fixed.acidity 0.0236576  0.0099187  2.385   0.0172 *    
volatile.acidity -1.0856214  0.0996323 -10.896 < 0.000000000000002 *** 
chlorides     -1.7376885  0.3913566 -4.440   0.0000960779597327 *** 
total.sulfur.dioxide -0.0021460  0.0005121 -4.191   0.00002933553690691 *** 
sulphates      0.8846921  0.1108310  7.982   0.00000000000000272 *** 
alcohol        0.2825603  0.0166180 17.003 < 0.00000000000000002 *** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.6504 on 1592 degrees of freedom

1599 samples
 6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1438, 1439, 1440, 1438, 1439, 1439, ...
Resampling results:

RMSE      Rsquared     MAE
0.6515146  0.3546232  0.5063893
```

MODELING

Ces six variables sont toutes fortement corrélées avec notre variable cible (la qualité) et présentent une forte signification statistique. Par rapport au modèle 1, le nouveau modèle comporte deux variables supplémentaires : l'acidité fixe et les chlorures, dont les impacts marginaux sur la qualité sont dans des directions différentes. Un coefficient d'estimation négatif des chlorures signifie qu'un vin de qualité supérieure devrait avoir une plus petite quantité de sel.

quantité de sel. Parallèlement, il existe une relation légèrement positive entre l'acidité fixe et la qualité, ce qui implique que les acides non volatils qui ne s'évaporent pas facilement devraient être un indicateur de vin de haute qualité.

MODELING

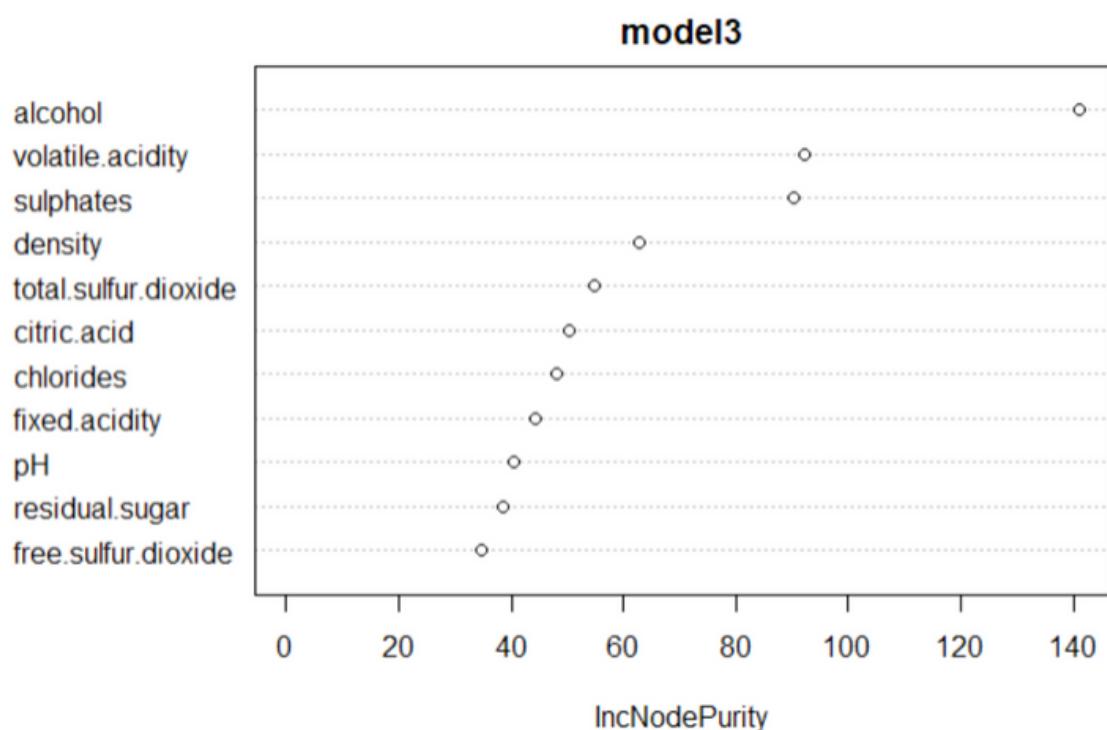
Modèle 3 : Enfin, j'ai exécuté Random Forest, un algorithme d'apprentissage automatique d'arbres de régression utilisé dans le processus de modélisation. Cela permet de créer un échantillon aléatoire de plusieurs arbres de décision de régression et de les fusionner pour obtenir une prédiction plus stable et plus précise par validation croisée. Nous appelons cela le "modèle 3", dont le résumé est présenté ci-dessous :

```
call:  
randomForest(formula = quality ~ ., data = training, mtry = 3,           importance = TRUE,  
na.action = na.omit)  
  Type of random forest: regression  
    Number of trees: 500  
No. of variables tried at each split: 3  
  
  Mean of squared residuals: 0.3414535  
    % var explained: 48.5
```

En plongeant dans la sélection des variables, nous obtenons les 10 prédicteurs les plus importants pour le modèle (figure 13). Pour ce faire, nous utilisons le MDI (Gini Importance or Mean Decrease in Impurity) qui calcule l'importance de chaque caractéristique comme la somme sur le nombre de fractionnements (sur tous les arbres) qui incluent la caractéristique, proportionnellement au nombre d'échantillons qu'elle fractionne. En comparaison avec le modèle 1 et le modèle 2, nous obtenons des informations supplémentaires sur des variables telles que la densité et le pH.

MODELING

Figure 13.



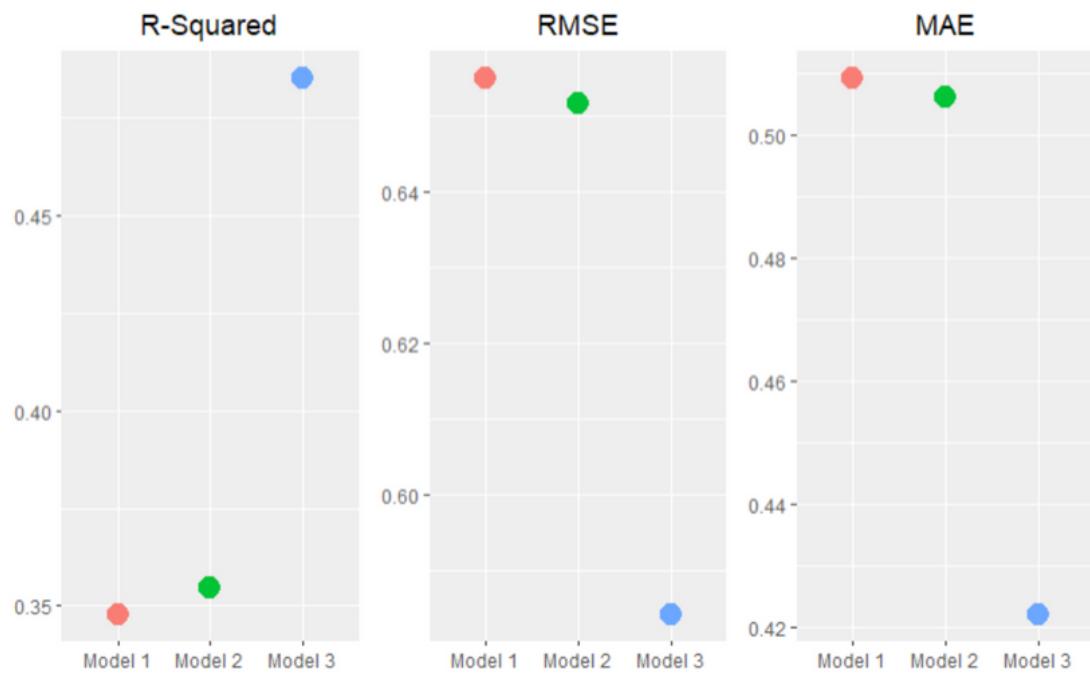
EVALUATION

Après avoir exécuté nos trois modèles, j'ai utilisé trois métriques : R-squared, RMSE et MAE, pour évaluer les performances de prédiction de nos modèles. Comme nous l'attendions de la Figure 14, le modèle 3 est le meilleur en termes de ces trois métriques, avec R-Squared : 48,50 %, RMSE : 0,5843, et MAE : 0,4222. Le modèle 1 et le modèle 2, dont les prédicteurs sont sélectionnés à partir de notre analyse de corrélation et de nos techniques de régularisation, n'enregistrent pas de grande différence en termes de ces métriques de performance.

Il est raisonnable que le Random Forest dans le modèle 3 nous donne des "prédictions" supérieures. Cependant, du point de vue de l'interprétation de "l'impact marginal", les modèles 1 et 2 peuvent être les gagnants. même si leurs mesures de performance sont en retard. Dans le contexte de notre question commerciale axée sur la prédiction de la qualité du vin rouge, le modèle 3 sera le meilleur choix.

EVALUATION

Figure 14.



CONCLUSION

En analysant les données des échantillons de tests physico-chimiques des vins rouges du nord du Portugal, j'ai pu créer un modèle qui peut aider les producteurs, les distributeurs et les vendeurs de l'industrie à prédire la qualité des produits à base de vin rouge et à avoir une meilleure compréhension de chaque caractéristique critique et actualisée. J'ai constaté que les ensembles de caractéristiques basés sur le modèle 3 - Random Forest ont donné de meilleurs résultats que les autres. En général, en utilisant le modèle 3 comme meilleur modèle de prédiction, j'ai déterminé que quatre des caractéristiques étaient les plus influentes : l'acidité volatile, l'acide citrique, les sulfates et l'alcool. Pour être plus précis, les vins de haute qualité semblent avoir une acidité volatile plus faible, un taux d'alcool plus élevé et des valeurs de sulfates moyennement élevées. En revanche, les vins de moindre qualité ont tendance à présenter de faibles valeurs d'acide citrique.

Toutefois, cette analyse présente certaines limites. Tout d'abord, le principal problème vient du fait que notre ensemble de données était déséquilibré. La majorité des valeurs de qualité étaient "régulières" (5 et 6), ce qui n'a pas contribué de manière significative à la recherche d'un modèle optimal. Ces valeurs ont rendu plus difficile l'identification de l'influence différente de chaque facteur sur une qualité "élevée" ou "faible" du vin, ce qui était l'objectif principal de cette analyse. Afin d'améliorer notre modèle prédictif, nous avons besoin de données plus équilibrées. Une autre limite de l'ensemble de données est qu'il ne comporte que 12 attributs, ce qui peut réduire la précision de notre prédiction de la qualité du vin rouge. La solution à ce problème est d'inclure des caractéristiques de données plus pertinentes, comme l'année de récolte, le temps de brassage, le lieu ou le type de vin. À l'avenir, nous pourrons également essayer d'autres mesures de performance et d'autres techniques d'apprentissage automatique pour améliorer les performances et la comparaison des résultats. Cette analyse aidera les entreprises viticoles à prédire la qualité des vins rouges en fonction de certains attributs et à fabriquer et vendre de bons produits associés.

PC
W
Σ



FIN DU PROJETS

Merci.



Clichy - 92110 - FRANCE

Email: aourik.khalid@hotmail.com, Web: www.khalidcode.com