Sommaire exécutif

Ce rapport présente une analyse approfondie des données de vin pour prédire la qualité du vin en utilisant différentes variables comme prédicteurs. Les analyses ont débuté par des tests statistiques tels que Anova et Tukey pour déterminer si les 3 sous-ensembles de données étaient homogènes. Ensuite, un modèle de régression stepwise a été développé pour les 11 variables pour l'ensemble de tous les vins, identifiant les variables critiques pour la qualité du vin.

Différentes méthodes de modélisation ont été testées, et le modèle retenu a été comparé à deux autres modèles spécifiques à la couleur du vin. Ensuite, une analyse DMR a été effectuée, comparant les performances de trois modèles différents (Boosted Trees, C&RT, Neural Network) en termes d'erreur quadratique moyenne pour les données de test. Les résultats ont montré que le modèle Boosted Trees a donné les meilleures performances pour prédire la qualité du vin.

Enfin, une analyse a été menée pour répondre à une question posée par un viticulteur lors d'une présentation: les conditions optimales pour obtenir les meilleurs vins sont-elles différentes pour les vins blancs et les vins rouges ? Des résultats ont été présentés pour répondre à cette question, montrant que les conditions optimales sont effectivement différentes pour les deux types de vin.

TABLE DES MATIERES

LISTE DES TABLEAUX

LISTE DES GRAPHIQUES

Étude exploratoire sur la classification des vins du Portugal

30a-Comparaison des sous-ensembles de vins blancs et rouges pour la qualité du vin

Le problème consiste à déterminer si les sous-ensembles Train et Test sont homogènes en termes de qualité du vin. Nous allons utiliser des tests statistiques appropriés pour comparer les moyennes et les variances entre les deux sous-ensembles. Cette étape est importante pour assurer la validité des résultats et l'efficacité des modèles de prédiction développés par la suite. Nous avons utilisé un test d'ANOVA univarié pour comparer les moyennes de la qualité du vin entre les sous-ensembles Train et Test. Nous allons également effectué une analyse de Tukey (HSD) pour comparer les paires de moyennes entre les sous-ensembles, ainsi qu'un test d'ANOVA de Levene pour évaluer l'homogénéité des variances entre les sous-ensembles.

Tableau 1: Tableau croisé de Y

Summary Frequency Table (Vins (6497cX16v
Marked cells have counts > 10
(Marginal summaries are not marked)
Exclude condition: Groupe="Validate"

	Y_qualité	groupe train	groupe test	Row Totals
Count	3	19	6	25
Row Percent		76,00%	24,00%	
Total Percent		0,37%	0,12%	0,48%
Count	4	117	55	172
Row Percent		68,02%	31,98%	
Total Percent		2,26%	1,06%	3,32%
Count	5	1263	433	1696
Row Percent		74,47%	25,53%	
Total Percent		24,38%	8,36%	32,74%
Count	6	1735	550	2285
Row Percent		75,93%	24,07%	
Total Percent		33,49%	10,62%	44,11%
Count	7	658	189	847
Row Percent		77,69%	22,31%	
Total Percent		12,70%	3,65%	16,35%
Count	8	120	30	150
Row Percent		80,00%	20,00%	
Total Percent		2,32%	0,58%	2,90%
Count	9	4	1	5
Row Percent		80,00%	20,00%	
Total Percent		0,08%	0,02%	0,10%
Count	All Grps	3916	1264	5180
Total Percent		75,60%	24,40%	

Le tableau croisé (Tableau 1) présente la répartition des niveaux de qualité du vin (Y), allant de 3 à 9, pour les sous-ensembles Train et Test. Les pourcentages par ligne montrent que la majorité des observations se situent dans les niveaux de Y 5, 6 et 7 pour les deux groupes, avec des proportions assez similaires. Les niveaux de qualité 5, 6 et 7 représentent respectivement 32,7%, 44% et 16,35% des observations totales. Bien que les proportions ne soient pas exactement identiques, elles sont similaires et peuvent suggérer une distribution comparable des niveaux de qualité entre les sous-ensembles Train et Test. Néanmoins, il est crucial de prendre en considération d'autres analyses qui seront menées par la suite afin d'obtenir une vision plus approfondie de la situation et d'évaluer de manière adéquate l'homogénéité entre les sous-ensembles Train et Test.

Tableau 2: Test univarié

Exclude condition: Groupe='validate'

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	128383,2	1	128383,2	169603,0	0,000000
groupe	6,1	1	6,1	8,1	0,004491
Error	3919,6	5178	0,8		

Le test d'ANOVA univarié compare les moyennes de la qualité des vins entre les sous-ensembles Train et Test. La valeur de p obtenue (0,0045) est inférieure au seuil de 5%, indiquant que les moyennes de la qualité des vins sont significativement différentes entre les sous-ensembles Train et Test.

Tableau 3: Test de Tukey

Tukey HSD test; variable Y_qualité (V
Approximate Probabilities for Post H
Error: Between MS = ,75696, df = 5
Exclude condition: Groupe='validate

Cell No.	groupe	{1} 5,8355	{2} 5,7555
1	train		0,004491
2	test	0,004491	

Tableau 4: Test d'ANOVA de Levene

Levene's Test for Homogeneity of Variances (Vins
Effect: "groupe"
Degrees of freedom for all F's: 1, 5178
Exclude condition: Groupe='validate'

	MS Effect	MS Error	F	p
Y_qualité	0,398111	0,292748	1,359908	0,243607

L'analyse de Tukey (HSD) est un test post-hoc qui permet de comparer les paires de moyennes entre les sous-ensembles. Dans ce cas, elle confirme les résultats du test d'ANOVA univarié, indiquant que les

moyennes de la qualité des vins sont significativement différentes entre les sous-ensembles Train et Test, avec une valeur de p de 0,0045. Le test d'ANOVA de Levene évalue l'homogénéité des variances entre les sous-ensembles Train et Test. La valeur de p obtenue est supérieure à 0,05, indiquant qu'il n'y a pas de différence significative dans la variabilité de la qualité des vins entre les sous-ensembles. Ainsi, les sous-ensembles sont homogènes en termes de variabilité, mais pas en termes de moyenne de qualité.

Tableau 4: Analyse descriptive

Breakdown Table of Descriptive Statistics (Vins r
N=5180 (No missing data in dep. var. list)
Exclude condition: Groupe="Validate"

groupe	Y_qualité Means	Y_qualité N	Y_qualité Std.Dev.
train	5,835546	3916	0,871002
test	5,755538	1264	0,867036
All Grps	5,816023	5180	0,870631

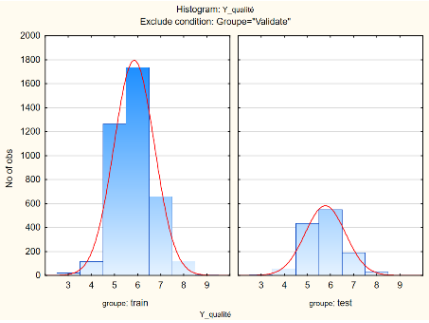


Figure 1: Histogramme par groupe

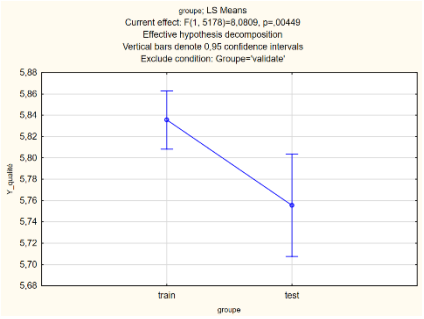


Figure 2: Moyennes des groupes

La Figure 1 présente la distribution des niveaux de qualité du vin pour les sous-ensembles Train et Test, tandis que la Figure 2 illustre les moyennes de Y_qualité pour ces groupes, ainsi que les intervalles de confiance à 95%. Comme mentionné précédemment, la majorité des observations se situent dans les niveaux de Y 5, 6 et 7, confirmant les résultats obtenus dans le tableau croisé. Le Tableau 4 fournit les statistiques descriptives de Y_qualité, telles que la moyenne, l'écart-type, l'erreur-type et l'intervalle de confiance à 95%. Ces statistiques apportent des informations supplémentaires pour évaluer l'homogénéité entre les sous-ensembles.

En somme, les tests d'ANOVA univarié et de Tukey (HSD) montrent que les moyennes de qualité des vins sont significativement différentes entre les sous-ensembles Train et Test. Cependant, le test d'ANOVA de Levene révèle qu'il n'y a pas de différence significative dans la variabilité de la qualité des vins entre les sous-ensembles. Ainsi, les sous-ensembles sont homogènes en termes de variabilité, mais pas en termes de moyenne de qualité. Les histogrammes (Figure 1 et 2) et les statistiques descriptives du Tableau 4 viennent appuyer cette analyse et confirmer la distribution des niveaux de qualité du vin pour les sous-ensembles Train et Test.

En conclusion, les sous-ensembles Train et Test ne sont pas homogènes en termes de moyenne de qualité, comme le montrent les résultats du test d'ANOVA univarié et de l'analyse de Tukey (HSD). Toutefois, ils présentent une homogénéité en termes de variabilité, selon les résultats du test d'ANOVA de Levene, ainsi que les observations des Figures 1 et 2 et les statistiques descriptives du Tableau 4.

30b-Développement d'un modèle de régression pour la qualité du vin en utilisant la méthode stepwise

Le problème consiste à identifier les variables physico-chimiques qui ont le plus d'influence sur la qualité du vin en utilisant la méthode de régression stepwise. Cette méthode permettra de développer un modèle de régression basé sur l'ensemble des vins rouges et blancs en ne retenant que les variables les plus critiques pour prédire la qualité du vin.Nous allors utiliser la mthedode GRM au lieu de mutliple regression car elle permet de donner des modele plus simple pour un meme R2

Nous allons développer un modèle de régression sur l'ensemble d'entraînement pour identifier les facteurs significatifs qui influencent la qualité du vin. Nous explorerons différentes méthodes statistiques pour déterminer les variables les plus importantes et les comparerons pour sélectionner le meilleur modèle.

GRM modèle global :

Parameter Estimates (Vins (6497cX16v).sta in 202
Sigma-restricted parameterization
Include condition: Groupe="train"

Effect	Y_qualité Param.	Y_qualité Std.Err	Y_qualité t	Y_qualité p
Intercept	59,4484	16,48162	3,6070	0,000314
fixed_acidity	0,0711	0,02107	3,3730	0,000751
volatile_acidity	-1,3154	0,10221	-12,8700	0,000000
citric_acid	-0,0777	0,10264	-0,7571	0,449019
residual_sugar	0,0423	0,00689	6,1332	0,000000
chlorides	-0,3454	0,43621	-0,7917	0,428572
free_sulfur_dioxide	0,0061	0,00099	6,2087	0,000000
total_sulfur_dioxide	-0,0025	0,00036	-6,8959	0,000000
density	-58,7631	16,80990	-3,4957	0,000478
pH	0,5007	0,11923	4,1995	0,000027
sulphates	0,6889	0,09882	6,9712	0,000000
alcohol	0,2576	0,02284	11,2777	0,000000

Figure 3: Sommaire de la régression

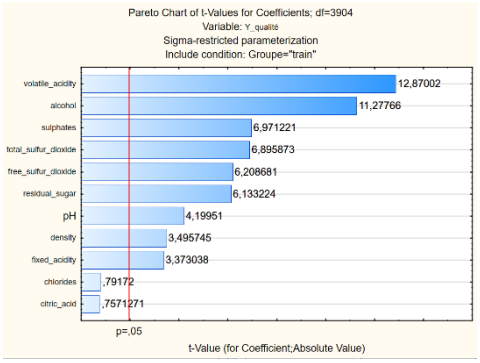


Figure 4:Diagramme de Pareto

Tableau 5: Tableau d'Anova

Test of SS Whole Model vs. SS Residual (Vins (6497cX16v).sta in 2023-MTH8302-Devoir3-data)
Include condition: Groupe="train"

Dependent Variable	Multiple R	Multiple R^2	Adjusted R^2	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual
Y_qualité	0,538830	0,290338	0,288339	862,3314	11	78,39376	2107,761	3904	0,539898

L'analyse de la régression (Figure 3) montre que les variables les plus significatives pour prédire la qualité du vin sont l'alcool, l'acidité volatile, la densité, l'acidité fixe et les sulfates, car elles ont toutes des coefficients de régression significatifs. Cependant, le chlorure et l'acide citrique ne sont pas significatifs. Le diagramme de Pareto (Figure 4) confirme l'importance de l'alcool et de l'acidité volatile, car ils ont les coefficients de régression les plus élevés et sont les seuls à dépasser la ligne de seuil. Le tableau d'ANOVA indique que le modèle de régression est significatif, avec une valeur du p-value très faible. Le R^2 est de 0,29, ce qui signifie que le modèle explique 29% de la variation de la qualité du vin. Le R^2 ajusté est très proche de R^2, ce qui suggère que toutes les variables significatives ont été incluses dans le modèle.

Résultats avec GRM Backward Stepwise :

Tableau 6: Sommaire de la régression

Parameter Estimates (Vins (6497cX16v).sta in 2023-MTH8302-D
Sigma-restricted parameterization
Include condition: Groupe="train"

Effect	Comment (B/Z/P)	Y_qualité Param.	Y_qualité Std.Err	Y_qualité t	Y_qualité p
Intercept		62,5446	16,13878	3,8754	0,000108
fixed_acidity		0,0699	0,02031	3,4416	0,000584
volatile_acidity		-1,2996	0,09360	-13,8853	0,000000
citric_acid	Pooled				
residual_sugar		0,0435	0,00672	6,4747	0,000000
chlorides	Pooled				
free_sulfur_dioxide		0,0061	0,00099	6,1898	0,000000
total_sulfur_dioxide		-0,0025	0,00035	-7,0791	0,000000
density		-61,9768	16,44753	-3,7682	0,000167
pH		0,5270	0,11680	4,5119	0,000007
sulphates		0,6696	0,09716	6,8920	0,000000
alcohol		0,2558	0,02276	11,2359	0,000000

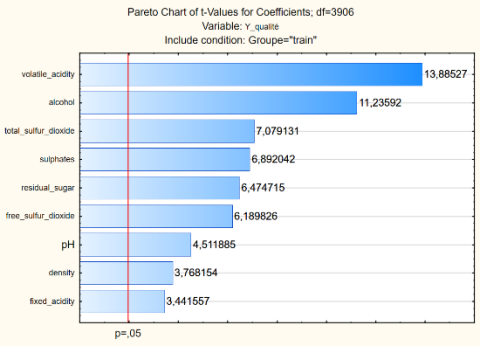


Figure 5: Diagramme de Pareto

Tableau 7: Tableau d'Anova

Test of SS Whole Model vs. SS Residual (Vins (6497cX16v).sta in 2023-MTH8302-Devoir3-data)
Include condition: Groupe="train"

Dependent Variable	Multiple R	Multiple R^2	Adjusted R^2	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual
Y_qualité	0,538588	0,290077	0,288442	861,5567	9	95,72852	2108,535	3906	0,539820

L'analyse de la régression montre que les variables les plus significatives pour prédire la qualité du vin sont l'acidité volatile, l'alcool, la densité, l'acidité fixe et les sulfates, car elles ont tous des coefficients de régression significatifs. Le chlorure et l'acide citrique ne sont pas significatifs et ont été exclus du modèle. Le diagramme de Pareto confirme l'importance de l'acidité volatile et de l'alcool, car ils ont les coefficients de régression les plus élevés et dépassent la ligne de seuil. Le tableau d'ANOVA indique que le modèle de régression est significatif, avec une valeur du p-value très faible. Le R^2 est de 0,29, ce qui signifie que le modèle explique 29% de la variation de la qualité du vin.

Résultats avec GRM Forward Stepwise :

Tableau 8: Sommaire de la régression

Parameter Estimates (Vins (6497cX16v).sta in 2023-MTH8302-D
Sigma-restricted parameterization
Include condition: Groupe = "train"

Effect	Comment (B/Z/P)	Y_qualité Param.	Y_qualité Std.Err	Y_qualité t	Y_qualité p
Intercept		1,85400	0,281643	6,5828	0,000000
fixed_acidity	Pooled				
volatile_acidity		-1,45226	0,083226	-17,4495	0,000000
citric_acid	Pooled				
residual_sugar		0,02090	0,003068	6,8122	0,000000
chlorides	Pooled				
free_sulfur_dioxide		0,00604	0,000988	6,1075	0,000000
total_sulfur_dioxide		-0,00225	0,000337	-6,6935	0,000000
density	Pooled				
pH		0,21354	0,078351	2,7254	0,006451
sulphates		0,52634	0,084442	6,2332	0,000000
alcohol		0,33015	0,010788	30,6028	0,000000

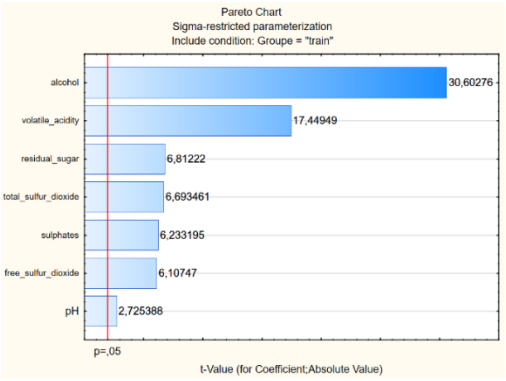


Figure 6: Diagramme de Pareto

Tableau 9: Tableau d'Anova:

Dependent Variable	Test of SS Whole Model vs. SS Residual (Vins (6497cX16v).sta in 2023-MTH8302-Devoir3-data) Include condition: Groupe = "train"								
	Multiple R	Multiple R^2	Adjusted R^2	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual
Y_qualité	0,536144	0,287450	0,286174	853,7527	7	121,9647	2116,339	3908	0,541540

D'après les résultats obtenus après l'application de la méthode GRM, le modèle retenu inclut 7 variables explicatives : volatile_acidity, residual_sugar, free_sulfur_dioxide, total_sulfur_dioxide, pH, sulphates et alcohol.
L'équation du modèle est la suivante :

Y_qualité = 1,854 - 1,452 * volatile_acidity + 0,021 * residual_sugar + 0,006 * free_sulfur_dioxide - 0,002 * total_sulfur_dioxide + 0,214 * pH + 0,526 * sulphates + 0,330 * alcohol

Le diagramme de Pareto montre que l'alcool et l'acidity sont les facteurs les plus significatifs, tandis que le pH est le moins significatif. Le R2 pour ce modèle est de 0,287 et le R2 ajusté est de 0,286, ce qui indique que le modèle explique environ 28,6% de la variance de la variable Y_qualité. Cette performance est similaire à celle du modèle de régression multiple avec forward stepwise. Cependant, le modèle GRM inclut moins de variables explicatives, ce qui le rend plus simple et potentiellement plus facile à interpréter.

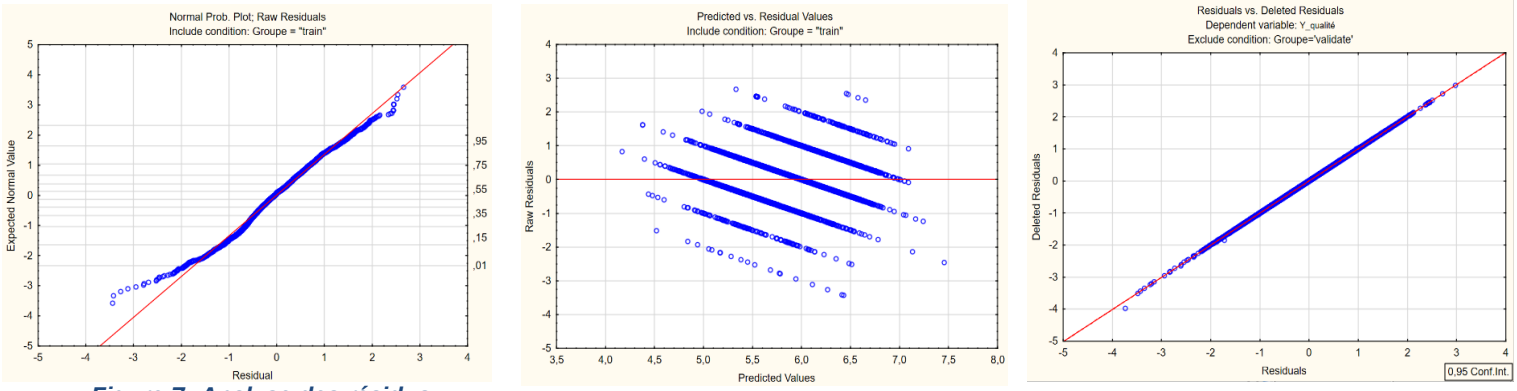


Figure 7: Analyse des résidus

D'après les résultats obtenus après l'application de la méthode forward stepwise, les figures montrent que les résidus présentent des caractéristiques intéressantes. La figure (Predicted vs Residuals) indique que les observations forment plusieurs lignes parallèles, ce qui pourrait suggérer un problème dans le modèle, tel qu'une catégorisation des données de réponse ou une relation non linéaire entre les variables explicatives et la variable réponse. La figure des résidus sur une échelle normale (Normal Plot) montre que les observations sont approximativement alignées avec la droite normale, ce qui indique que la distribution des résidus suit une loi normale, bien que l'ajustement ne soit pas parfait. Enfin, pour la figure (Residuals vs Deleted), on constate qu'il n'y a pas de données aberrantes et que les observations suivent la droite. Cela suggère une homogénéité et une constance de la variance, ce qui est une caractéristique souhaitable pour un modèle de régression.

Tableau 10: Tableau de calcul du BIC pour les 3 modèles

Modèle	SSE	MSE	R2	R2_Ajusté	BIC
Y_GRM_Global	2107.76	0.54	0.29	0.288	8776,4
Y_GRM_Backward	2108.5	0.54	0.29	0.288	8769.5
Y_GRM_Forward	2116.34	0.54	0.287	0.286	8767.5

Le tableau présente les résultats de trois modèles de régression basés sur la méthode stepwise : Global, Forward et Backward. Les valeurs SSE et MSE sont toutes très proches, indiquant une performance similaire pour les trois modèles. Le R2 et le R2 ajusté sont également très similaires entre les trois modèles, avec un R2 légèrement supérieur pour le modèle Global. Cependant, le critère BIC montre des différences significatives entre les trois modèles. Le modèle Forward a le BIC le plus faible, suivi de près par le modèle Backward, tandis que le modèle Global a le BIC le plus élevé. Par conséquent, nous avons choisi le modèle Forward comme le modèle de régression final, car il a la plus faible valeur de BIC et fournit la meilleure balance entre la précision de la prédiction et la complexité du modèle. Nous avons choisi d'utiliser le critère BIC plutôt que l'AICc car le BIC est plus approprié lorsque le nombre de données est important.

En conclusion, le modèle GRM Forward Stepwise a été choisi car il présente une performance similaire à celle du modèle de régression multiple avec backward stepwise, tout en incluant moins de variables explicatives et un meilleur BIC. Cela rend le modèle GRM plus simple et potentiellement plus facile à interpréter, tout en offrant une performance équivalente en termes de prédiction de la qualité du vin. Par rapport au modèle GRM global, le modèle GRM Forward Stepwise a également permis de ne pas inclure les variables non significatives dans le modèle final, améliorant ainsi la qualité de la prédiction.

30c-Comparaison des modèles de prédiction pour les vins blancs et rouges

L'objectif de cette analyse est de déterminer si la conclusion concernant l'importance des variables pour prédire la qualité du vin est la même pour les vins blancs et rouges. Pour ce faire, les modèles de prédiction ont été développés sur l'ensemble d'entraînement uniquement pour les vins blancs et rouges, puis les performances des différents modèles ont été comparées sur l'ensemble de test correspondant. Nous comparerons ensuite les performances des modèles obtenus avec celles du modèle en 30b.

Modèle pour le Vin Blanc

Tableau 11: Sommaire de la régression

Parameter Estimates (Vins (6497cX16v).sta in 2023-MTH8302-D
Sigma-restricted parameterization
Include condition: Groupe = "train"
Exclude condition: couleur= "rouge"

Effect	Comment (B/Z/P)	Y_qualité Param.	Y_qualité Std.Err	Y_qualité t	Y_qualité p
Intercept		199,780	28,76283	6,9458	0,000000
fixed_acidity		0,097	0,03016	3,2277	0,001262
volatile_acidity		-1,982	0,14438	-13,7295	0,000000
citric_acid	Pooled				
residual_sugar		0,094	0,01099	8,5700	0,000000
chlorides	Pooled				
free_sulfur_dioxide		0,004	0,00089	4,6186	0,000004
total_sulfur_dioxide	Pooled				
density		-200,259	29,13003	-6,8746	0,000000
pH		0,849	0,14247	5,9589	0,000000
sulphates		0,538	0,12930	4,1582	0,000033
alcohol		0,127	0,03750	3,3917	0,000704

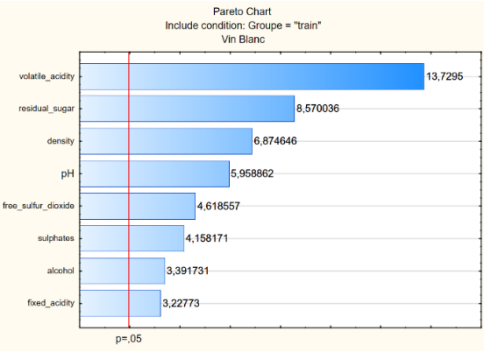


Figure 8: Diagramme de Pareto

Tableau 12: Tableau D'Anova

Dependent Variable	Multiple R	Multiple R^2	Adjusted R^2	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual
Y_qualité	0,534569	0,285764	0,283813	656,2676	8	82,03345	1640,266	2928	0,560200

Pour le modèle de régression développé sur les vins blancs uniquement et basé sur l'ensemble d'entraînement.
l'équation est la suivante :

Y_qualité = 199,780 + 0,097 * fixed_acidity - 1,982 * volatile_acidity + 0,094 * residual_sugar + 0,004 * free_sulfur_dioxide - 200,259 * density + 0,849 * pH + 0,538 * sulphates + 0,127 * alcohol

Le R2 pour ce modèle est de 0,285 et le R2 ajusté est de 0,284, ce qui signifie que le modèle explique environ 28,4% de la variance de la variable Y_qualité pour les vins blancs. I
I est intéressant de noter que, bien que le modèle soit développé uniquement pour les vins blancs, les facteurs les plus significatifs restent similaires à ceux du modèle global (30b). Cela confirme que les facteurs clés pour la qualité des vins blancs sont également l'alcool, l'acidité volatile, les sulfates et le pH, entre autres.

En effet, en observant le diagramme de Pareto pour le modèle de régression des vins blancs, on constate que la variable la plus importante est désormais l'acidité volatile, tandis que l'alcool devient l'une des variables les moins significatives. Cela montre que, bien que les facteurs globaux soient similaires, l'importance relative de ces facteurs peut varier entre les vins blancs et l'ensemble des vins. Cela suggère que l'acidité volatile joue un rôle plus crucial dans la qualité des vins blancs, tandis que l'alcool a un impact moindre par rapport au modèle global.L'analyse des résidus pour ce modèle est similaire à celle de la partie 30b, ce qui indique une bonne homogénéité et une constance de la variance.

Vin Rouge

Tableau 13: Sommaire de la régression

Effect	Comment (B/Z/P)	Y_qualité Param.	Y_qualité Std.Err	Y_qualité t	Y_qualité p
Intercept		4,15718	0,501288	8,29299	0,000000
fixed_acidity	Pooled				
volatile_acidity		-0,90088	0,130485	-6,90405	0,000000
citric_acid	Pooled				
residual_sugar	Pooled				
chlorides		-1,87955	0,503535	-3,73271	0,000200
free_sulfur_dioxide	Pooled				
total_sulfur_dioxide		-0,00341	0,000654	-5,22252	0,000000
density	Pooled				
pH		-0,41862	0,146786	-2,85190	0,004438
sulphates		0,96690	0,141972	6,81047	0,000000
alcohol		0,29152	0,021099	13,81666	0,000000

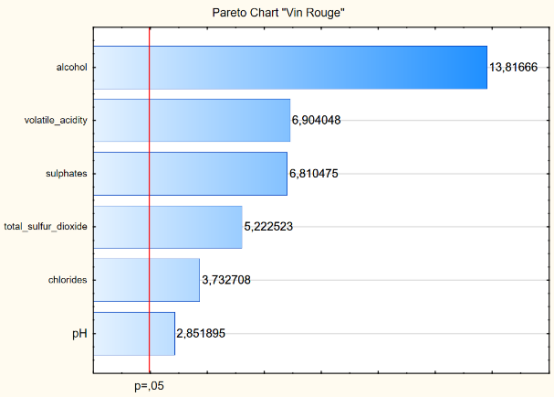


Figure 9: Diagramme de Pareto

Tableau 14: Tableau D'Anova

Dependent Variable	Multiple R	Multiple R^2	Adjusted R^2	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual
Y_qualité	0,602547	0,363063	0,359131	228,7007	6	38,11679	401,2196	972	0,412777

Pour le modèle de régression des vins rouges, le R2 est de 0,363 et le R2 ajusté est de 0,359, ce qui indique que le modèle explique environ 35,9% de la variance de la variable Y_qualité.

L'équation du modèle est la suivante :

Y_qualité = 4,157 - 0,901 * volatile_acidity - 1,880 * chlorides - 0,003 * total_sulfur_dioxide - 0,419 * pH + 0,967 * sulphates + 0,292 * alcohol

Le diagramme de Pareto montre que l'alcool est la variable la plus significative pour les vins rouges, suivi de l'acidité volatile et des sulfates. Le pH est la variable la moins significative. Il est intéressant de constater que, bien que les facteurs globaux soient similaires pour les vins rouges et blancs, l'importance relative de ces facteurs varie. Pour les vins rouges, l'alcool est le facteur le plus important, tandis que l'acidité volatile et les sulfates jouent également un rôle important. Le pH, en revanche, a un impact moindre sur la qualité des vins rouges. Ces résultats soulignent l'importance de prendre en compte les différences entre les vins rouges et blancs lors de l'analyse de la qualité, car cela permet d'identifier des facteurs spécifiques qui influencent davantage la qualité de chaque type de vin.

On peut déjà observer sur le groupe train que l'alcool est le facteur le plus important pour les vins rouges, alors que pour les vins blancs, il devient l'un des facteurs les moins importants. L'acidité volatile est un facteur majeur pour les deux types de vins, mais elle est encore plus importante pour les vins blancs. Le R2 ajusté est également plus élevé pour les vins rouges que pour les vins blancs, suggérant que le modèle explique une plus grande part de la variance de la qualité des vins rouges. Ces résultats soulignent l'importance de considérer séparément les vins rouges et blancs dans l'analyse pour mieux comprendre les éléments clés qui déterminent la qualité des vins rouges et blancs.

À présent, nous procéderons à l'évaluation des trois modèles, développés sur l'ensemble d'entraînement en utilisant la méthode GRM forward stepwise de la question 30b, le modèle spécifique aux vins blancs et le modèle spécifique aux vins rouges sur l'ensemble de test. Ceci nous permettra de déterminer si les modèles présentent des performances similaires ou s'il est avantageux de distinguer les vins par couleur pour obtenir des prédictions plus précises et adaptées.

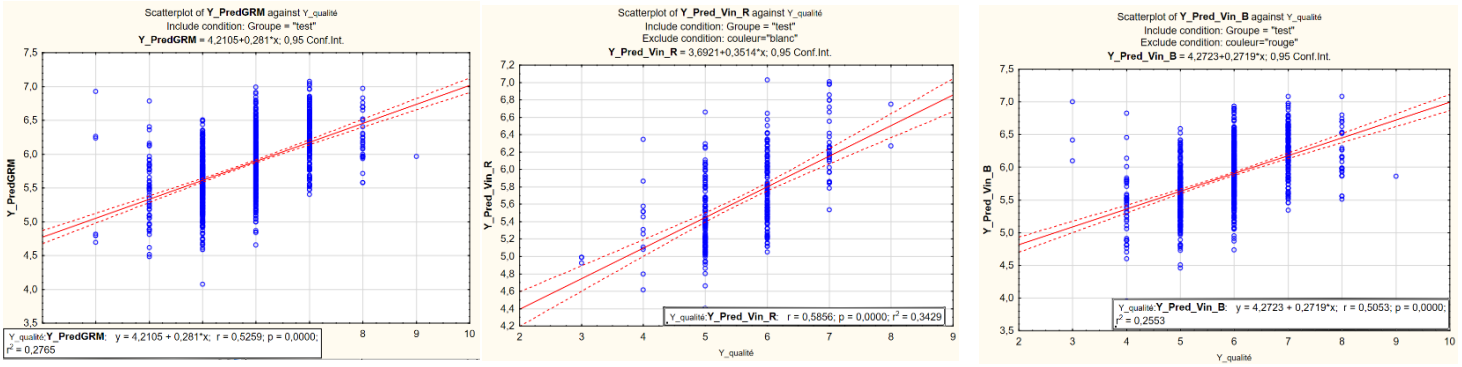
Tableau 15: Comparaison des 3 modèles

Modèle	SSE	MSE	R2	R2_Ajusté	BIC
Y_GRM_Forward	693.5	0.55	0.27	0.265	2884.85
Y_Vin_Blanc	552.2	0.59	0.253	0.246	2244.39
Y_Vin_Rouge	130.3	0.43	0.331	0.316	653.13

Après avoir analysé les modèles sur l'ensemble de test, le tableau 15 montre que le modèle de régression spécifique aux vins rouges a le SSE le plus faible (130,3), suivi du modèle pour les vins blancs (552,2), tandis que le modèle global GRM (693,5) a le SSE le plus élevé. Les valeurs R2 et R2 ajusté indiquent également que le modèle pour les vins rouges explique une plus grande part de la variance de la qualité des vins (R2 ajusté de 0,316) par rapport aux deux autres modèles (R2 ajusté de 0,246 et 0,265 pour les vins blancs et le modèle GRM, respectivement), ce qui concorde avec les resultats obtenues sur le groupe train. Nous avons utilisé le critère BIC pour choisir le meilleur modèle, car il est plus

approprié lorsque le nombre de données est important. Le BIC du modèle pour les vins rouges est le plus faible, suivi de près par le modèle pour les vins blancs, ce qui confirme que la distinction entre les couleurs de vin est importante pour obtenir une prédiction précise de la qualité du vin. En effet, comme observé précédemment, certains facteurs qui sont importants pour les vins blancs ne le sont pas pour les vins rouges, et vice versa.

Figure 10: Scatterplots des 3 modèles de regression



La Figure 10 représente les scatterplots des trois modèles de régression. Les résultats de la figure confirment ceux du tableau 15, montrant que le modèle pour les vins rouges présente une meilleure performance que les deux autres modèles, tandis que les modèles pour les vins blancs et l'ensemble de données combiné ont une performance similaire. Ainsi, les modèles spécifiques aux couleurs de vin permettent une meilleure adaptation aux spécificités de chaque type de vin, offrant ainsi une prédiction plus précise et adaptée.

En conclusion, les résultats obtenus sur le groupe test ont confirmé que la séparation des vins par couleur est importante pour obtenir des prédictions précises de la qualité du vin. Les modèles spécifiques pour les vins rouges et blancs ont des performances supérieures par rapport au modèle qui prend en compte les deux couleurs, ce qui indique que les facteurs qui influencent la qualité des vins diffèrent selon leur couleur. Le critère BIC a été utilisé pour choisir le meilleur modèle, car il tient compte de la complexité du modèle. Ainsi, nous recommandons d'utiliser des modèles distincts pour les vins rouges et blancs afin d'obtenir des prédictions précises de la qualité du vin.

30d-Utilisation du module DMR pour l'analyse de données

L'objectif de cette section consiste à implémenter le module DMR afin d'examiner les diverses méthodes d'analyse proposées (C&RT, Réseau de Neurones, Forêt Aléatoire, Arbres Renforcés et SVM) en utilisant les variables de réponse Y et Z comme sorties continues et catégoriques, respectivement. Plusieurs figures et tableaux en résultent, que nous allons désormais analyser.

Tout d'abords, le graphique de la figure "Summary of Boosted Trees" montre que le nombre optimal d'arbres est de 196, avec une taille maximale d'arbre de 13. L'erreur quadratique moyenne diminue décroît rapidement puis de manière plus progressive pour les données d'entraînement en fonction du nombre d'arbres, ce qui est normal puisqu'on entraîne le modèle sur une plus grande quantité de données.

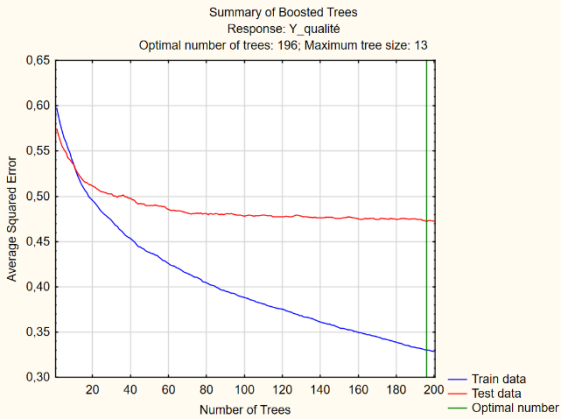


Figure 11: Summary of Boosted Trees

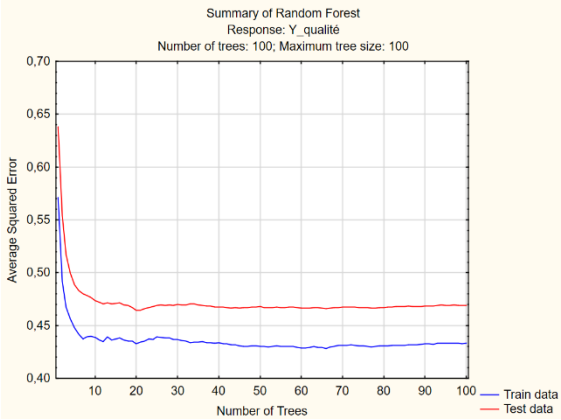


Figure 12: Random Forest

Cependant, pour les données de test, l'erreur quadratique diminue graduellement pour atteindre 0,45 avec 196 arbres, avec une baisse rapide entre 0 et 30 arbres, passant de 0,58 à 0,5, puis continue de diminuer progressivement. Le nombre optimal d'arbres et la taille maximale d'arbre ont été sélectionnés afin de minimiser l'erreur quadratique moyenne pour les données d'entraînement. En comparaison, la figure 12 montre les résultats du modèle de forêt aléatoire. Ce modèle présente une erreur quadratique moyenne très similaire à celle des arbres renforcés. L'erreur quadratique atteint 0,47 après 20 arbres, avec une baisse plus rapide pour les 10 premiers arbres. Cependant, contrairement au modèle d'arbres renforcés, l'erreur quadratique du modèle de forêt aléatoire reste constante après les 20 premiers arbres. En somme, les deux modèles (arbres renforcés et forêt aléatoire) présentent des performances similaires en termes d'erreur quadratique moyenne. Les deux modèles montrent une amélioration rapide de la performance lors de l'ajout des premiers arbres. Cependant, la forêt aléatoire atteint un plateau plus rapidement que les arbres renforcés, avec une erreur quadratique qui ne change pas significativement après 20 arbres.

Tableau 16: Sommaire des réseaux de neurones

Variables: ^
Include cases: 1

	1 Index	2 Net. name	3 Training perf.	4 Test perf.	5 Validation perf.	6 Training error	7 Test error	8 Validation error	9 Training algorithm	10 Error function	11 Hidden activation	12 Output activation
1	5	MLP 13-3-1	0,551782	0,521774		0,260333	0,291646		BFGS 64	SOS	Tanh	Identity

Le tableau 16 présente les résultats de la performance des modèles de réseaux de neurones pour la prédiction de la qualité du vin. Les résultats montrent une performance de prédiction de 0,55 pour le groupe d'entraînement et de 0,52 pour le groupe de test. En ce qui concerne l'erreur de prédiction, les résultats montrent une erreur de 0,26 pour le groupe train et 0,29 pour le groupe test. les valeurs d'erreur relativement faibles ainsi que la performance des deux groupe suggèrent que le modèle de réseau de neurones est assez précis dans la prédiction de la qualité du vin et ne souffre pas de surapprentissage.

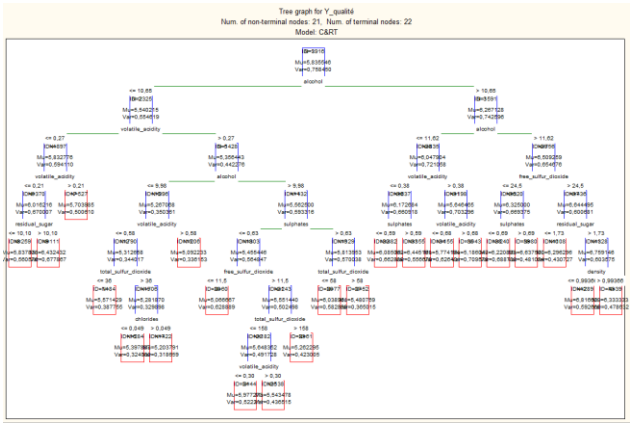


Figure 13: Arbre C&RT

Tableau 17: Importance des predicteurs

	Predictor importance (2023-M Response: Y_qualité Model: C&RT	
	Variable Rank	Importance
alcohol	100	1,000000
volatile_acidity	69	0,687633
density	68	0,680006
chlorides	52	0,518932
citric_acid	50	0,503457
free_sulfur_dioxide	36	0,358853
total_sulfur_dioxide	31	0,312899
residual_sugar	27	0,265325
fixed_acidity	22	0,223339
sulphates	20	0,201350
pH	16	0,163162
couleur	13	0,132668

Le graphique de l'arbre de classification et de régression montre que les variables les plus importantes pour prédire la qualité du vin sont l'alcool, l'acidité volatile, la densité et les chlorures, ce qui est en accord avec les résultats précédents. L'arbre comporte 21 nœuds non terminaux et 22 nœuds terminaux. Le tableau représentant l'importance des prédicteurs confirme le classement des variables les plus importantes, avec l'alcool en tête, suivi de la densité, de l'acidité volatile et des chlorures. Les variables les moins importantes l'acidité fixe et le pH. Les valeurs de rang sont calculées en comparant l'importance de chaque variable à celle des autres variables dans le modèle.
Les résultats de ces deux graphiques confirment que l'alcool, l'acidité volatile, la densité et le chlorure sont les variables les plus importantes pour prédire la qualité du vin, tandis que l'acidité fixe les sulphates et le pH ont une importance relativement faible avec la variable categorique couleur du vin etant le facteur le moins important, ce qui est en concordance avec les modelès precédents.

Tableau 18: Sommaire de deploiment

	Summary of Deployment (Error rates) (2023-MTH8302-Devoir3-data_Validation)				
	2-Random forest	3-Boosted trees	5-Neural	1-C&RT	4-SVM
Mean squared error	0,504531	0,480454	0,544493	0,556765	0,823756

Le tableau 18 résume les erreurs quadratiques moyennes pour les cinq méthodes d'analyse. Le modèle Boosted Trees présente l'erreur quadratique moyenne la plus faible (0,48), suivi du modèle Random Forest (0,50), du Neural Net (0,54) et du C&RT (0,55). Le modèle SVM est le moins performant avec une erreur quadratique de 0,82. Ces résultats montrent que le modèle Boosted Trees offre une performance de prédiction supérieure pour la qualité du vin, avec la plus faible erreur quadratique moyenne pour les données de test.

En conclusion, l'application du module DMR a confirmé l'importance des variables précédemment identifiées pour prédire la qualité du vin, notamment l'alcool, l'acidité volatile, la densité et les chlorures. Les résultats ont également démontré que le modèle Boosted Trees présente la plus faible erreur quadratique moyenne pour les données de test, suggérant une meilleure performance de prédiction pour la qualité du vin. Afin de déterminer le modèle le plus optimal, il est essentiel de comparer les performances des cinq modèles analysés. Les mêmes figures ont été effectuées pour Z_quality en annexe.

30e-Synthèse et comparaison des résultats obtenus avec le module DMR

L'objectif de cette analyse est de prédire la qualité du vin en utilisant différents modèles de modélisation, tels que Boosted Trees, Random Forest, Neural Net, C&RT et SVM. Les modèles ont été développés sur l'ensemble de données d'entraînement et évalués sur l'ensemble de données de test. Le but est de comparer les performances des différents modèles en termes d'erreur quadratique moyenne (MSE) pour le groupe de test et le groupe d'entraînement sur Y_quality.

Tableau 19: Tableau de synthèse pour Y_qualité

Name	Training residual (MSE)	Testing residual (MSE)	Correlation Coefficient (Training)	Correlation Coefficient (Testing)
Boosted trees	0,37	0,48	0,71	0,61
Random forest	0,44	0,5	0,66	0,58
Neural network	0,53	0,54	0,55	0,53
C&RT	0,51	0,56	0,57	0,52
SVM	0,75	0,82	0,56	0,51

En comparant les résultats des cinq modèles, on peut observer que le modèle Boosted Trees présente les erreurs quadratiques moyennes les plus faibles pour les données d'entraînement (0,37) et de test (0,48), suivi du modèle Random Forest avec des erreurs quadratiques moyennes de 0,44 pour les données d'entraînement et 0,5 pour les données de test. À l'inverse, le modèle SVM affiche la plus grande erreur quadratique moyenne pour les données de test (0,82) et d'entraînement (0,75), suggérant une performance de prédiction moins satisfaisante.

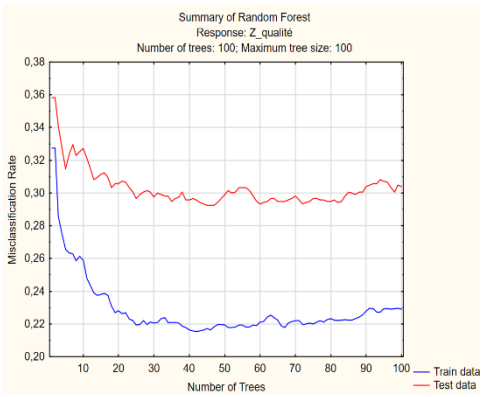
En ce qui concerne les coefficients de corrélation, les résultats montrent que le modèle Boosted Trees a le coefficient de corrélation le plus élevé pour les données d'entraînement (0,71) et de test (0,61), suivi de près par le modèle Random Forest avec des coefficients de corrélation de 0,66 pour les données d'entraînement et 0,58 pour les données de test. Les modèles C&RT, Neural Net et SVM présentent des coefficients de corrélation plus faibles pour les deux groupes de données, avec des valeurs respectives de 0,57 et 0,52 pour le C&RT, 0,55 et 0,53 pour le Neural Net, et 0,56 et 0,51 pour le SVM.

Ces résultats, combinés avec les erreurs quadratiques moyennes, suggèrent que le modèle Boosted Trees est le plus performant pour la prédiction de la qualité du vin pour Y_quality, offrant à la fois une faible erreur quadratique moyenne et un coefficient de corrélation élevé

L'analyse précédente a été centrée sur la prédiction de la variable Y_qualité. Cependant, il est également important de considérer la prédiction de la variable Z_qualité, qui est un recodage de la variable Y_qualité en trois catégories : mauvais, ok et bon. Les figures 14, 15 et le tableau 20 fournissent des informations sur la performance des différents modèles pour la prédiction de Z_qualité.

Predictor importance (2023-M
Response: Z_qualité
Model: C&RT
Important variables selection

	Variable Rank	Importance
volatile_acidity	100	1,000000
citric_acid	93	0,933792
alcohol	82	0,819280
free_sulfur_dioxide	80	0,799139
density	62	0,623148
residual_sugar	52	0,515614
total_sulfur_dioxide	51	0,513025
chlorides	51	0,511272
fixed_acidity	44	0,439761
sulphates	38	0,383003
pH	36	0,364460
couleur	21	0,214236



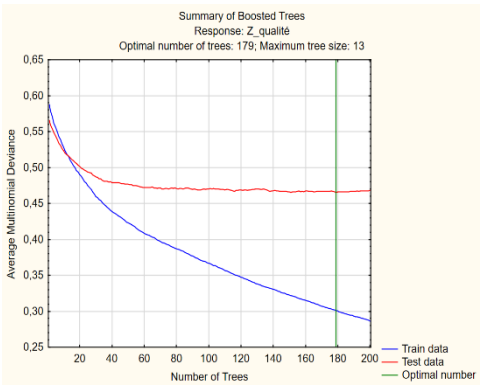
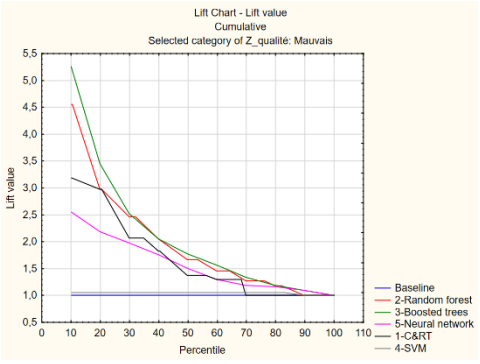
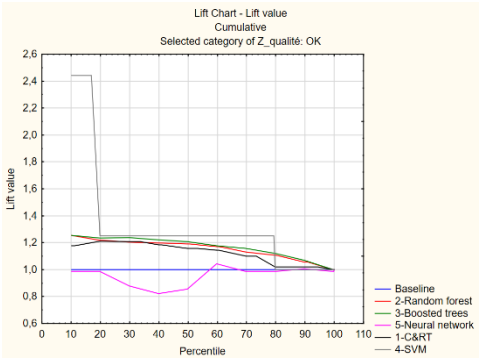


Figure 14: Resultats pour Z_quality

Tout d'abord, la Figure 14 révèle que les facteurs les plus importants pour la prédiction de Z_qualité diffèrent de ceux de Y_qualité. Dans ce cas, la volatilité de l'acide est le facteur prédominant, suivi de l'acide citrique et enfin de l'alcool.





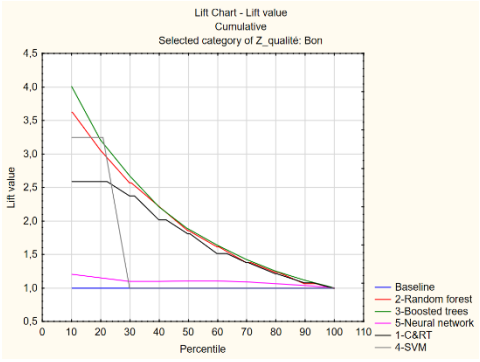


Figure 15: Comparaison des courbes de gain pour les différentes catégories d'évaluation

La Figure 15 compare les courbes de gain pour les différentes catégories d'évaluation (mauvais, ok, bon) en fonction des modèles prédictifs. Cette comparaison met en évidence des différences notables entre les modèles pour chaque catégorie. Par exemple, pour la catégorie "Bon", le modèle Boosted Trees présente la valeur de lift la plus élevée, tandis que pour la catégorie "Ok", c'est le modèle SVM qui obtient la meilleure performance en termes de lift value.

L'analyse du Tableau 20 permet de comparer les performances des différents modèles pour la prédiction de la variable Z_qualité. En se basant sur l'erreur d'entraînement, l'erreur de test et le coût de mauvaise classification, nous pouvons déterminer quel modèle est le plus adapté pour prédire la qualité du vin.

Parmi les modèles étudiés, le modèle Boosted Trees se démarque par ses performances. Il présente la plus faible erreur d'entraînement (13,2 %) et d'erreur de test (17,88 %), ce qui suggère une bonne précision et une capacité à généraliser sur des données nouvelles. De plus, le coût de mauvaise classification du modèle Boosted Trees est le plus bas, avec une valeur de 226, confirmant son efficacité meme pour Z_qualité.

Tableau 20: Tableau de synthèse pour Z_qualité

Name	Training error (%)	Testing error (%)	Missclassification cost
Boosted trees	13,2	17,88	226
Neural network	22,8	21,76	275
SVM	23,21	24,53	310
Random forest	25,23	27,93	353
C&RT	50,49	53,48	676

Le deuxième meilleur modèle est le SVM, qui présente une erreur de test de 24,53 % et un coût de mauvaise classification de 310. Le modèle Neural Network affiche des performances légèrement inférieures, avec une erreur de test de 21,76 % et un coût de mauvaise classification de 275. En revanche, le modèle Random Forest obtient des résultats nettement moins satisfaisants, avec une erreur de test de 27,93 % et un coût de mauvaise classification de 353. Le modèle C&RT est de loin le moins performant, avec une erreur de test de 53,48 % et un coût de mauvaise classification de 676.

En conclusion, nous avons comparé cinq modèles de classification pour prédire la qualité du vin sur les variables Y_qualité et Z_qualité. Nous avons constaté que le modèle de Boosted Trees est le plus performant pour la prédiction du vin, avec la plus faible erreur pour les données de test, la meilleure corrélation et le coût le plus faible de mauvaise classification.

30f-Identification des caractéristiques chimiques importantes pour la qualité du vin

En tenant compte des résultats obtenus à partir de l'ensemble des modèles étudiés, l est possible d'identifier les caractéristiques chimiques du vin qui ont le plus d'impact sur sa qualité. Les deux variables de réponse étudiées, Y_qualite et Z_qualite, ont montré des différences dans l'importance des facteurs.

Tableau 21: Importance des facteurs pour Y_qualité

Predictor importance (2023-M
Response: Y_qualité
Model: C&RT

	Variable Rank	Importance
alcohol	100	1,000000
volatile_acidity	69	0,687633
density	68	0,680006
chlorides	52	0,518932
citric_acid	50	0,503457
free_sulfur_dioxide	36	0,358853
total_sulfur_dioxide	31	0,312899
residual_sugar	27	0,265325
fixed_acidity	22	0,223339
sulphates	20	0,201350
pH	16	0,163162

Tableau 22: Importance des facteurs pour Z_qualité

Predictor importance (2023-M
Response: Z_qualité
Model: C&RT
Important variables selection

	Variable Rank	Importance
volatile_acidity	100	1,000000
citric_acid	93	0,933792
alcohol	82	0,819280
free_sulfur_dioxide	80	0,799139
density	62	0,623148
residual_sugar	52	0,515614
total_sulfur_dioxide	51	0,513025
chlorides	51	0,511272
fixed_acidity	44	0,439761
sulphates	38	0,383003
pH	36	0,364460

Pour Y_qualite, les caractéristiques les plus importantes sont l'alcool, l'acidité volatile, le sucre résiduel, les sulfates, la densité, le sulfure de dioxyde free, le sulfure de dioxyde total, les chlorites, le pH, l'acide citrique et l'acide fixe.

Pour Z_qualite, les caractéristiques les plus importantes sont l'alcool, l'acidité volatile, les chlorites, les sulfates, le sulfure de dioxyde free, le sucre résiduel, la densité, l'acide citrique, le sulfure de dioxyde total et le pH.
Il est intéressant de noter que l'alcool et l'acidité volatile sont les caractéristiques les plus importantes pour les deux variables de réponse, Y_qualite et Z_qualite, ce qui indique qu'elles sont probablement les plus influentes dans l'évaluation globale de la qualité du vin. En revanche, le pH et les sulfates sont les caractéristiques les moins importantes pour les deux variables de réponse.

En conclusion, cette analyse suggère que les viticulteurs devraient se concentrer principalement sur les niveaux d'alcool et d'acidité volatile dans leur production de vin, ainsi que sur d'autres facteurs importants tels que le sucre résiduel, la densité et les sulfures de dioxyde pour obtenir des vins de qualité.

30g-Réponse à la question sur la différence entre les vins blancs et rouges et présentation des résultats supplémentaires.

En réponse à la remarque du participant, il est en effet important de distinguer les vins blancs et rouges dans l'analyse de prédiction de la qualité du vin. Pour répondre à cette préoccupation, nous avons développé des modèles spécifiques pour chaque type de vin en plus du modèle général, qui inclut les vins rouges et blancs ensemble.
En comparant les performances des modèles spécifiques pour les vins rouges et blancs avec le modèle général, nous avons constaté que le modèle spécifique aux vins rouges présente un R^2 ajusté supérieur sur les données de test (0,3429) par rapport au modèle général (0,27), tandis que le modèle spécifique aux vins blancs présente un R^2 ajusté légèrement inférieur (0,25). Cela suggère que le modèle spécifique aux vins rouges est effectivement plus adapté pour prédire la qualité des vins rouges, tandis que le modèle général est plus adapté pour prédire la qualité des vins en général.

Tableau 23: Predicteur important Vin Blanc

	Predictor importance (2023-I Response: Y_qualité Model: C&RT Important variables selectior	
	Variable Rank	Importance
alcohol	100	1,000000
density	69	0,690525
volatile_acidity	52	0,522840
chlorides	50	0,501372
free_sulfur_dioxide	39	0,386877
citric_acid	36	0,357833
total_sulfur_dioxide	33	0,325258
pH	20	0,201475
residual_sugar	17	0,172262
fixed_acidity	15	0,152798
sulphates	8	0,078181

Tableau 24: Predicteur important Vin Rouge

	Predictor importance (2023-M Response: Y_qualité Model: C&RT Important variables selection	
	Variable Rank	Importance
alcohol	100	1,000000
sulphates	92	0,921896
volatile_acidity	79	0,793843
citric_acid	50	0,501353
total_sulfur_dioxide	44	0,440217
fixed_acidity	39	0,386925
density	38	0,380001
chlorides	29	0,291217
pH	27	0,268505
free_sulfur_dioxide	24	0,237763
residual_sugar	11	0,107964

Tableau 25: Sommaire de deploimenet « Vin Blanc »

	Summary of Deployment (Error rates) (2023-MTH8302-Devoir3-data_Validation)			
	2-Random forest	3-Boosted trees	5-Neural	1-C&RT
Mean squared error	0,518303	0,533031	0,593506	0,564719

Tableau 26: Sommaire de deploimenet « Vin Rouge »

	Summary of Deployment (Error rates) (2023-MTH8302-Devoir3-data_Validation)			
	2-Random forest	3-Boosted trees	5-Neural	1-C&RT
Mean squared error	0,394680	0,458784	0,429463	0,436794

En analysant les tableaux 25 et 26, qui résument le déploiement des modèles spécifiques aux vins blancs et rouges respectivement, on observe que le MSE le plus faible est différent pour les deux types de vin. Pour le vin blanc, le modèle Random Forest présente le MSE le plus faible de 0,51, tandis que pour le vin rouge, le même modèle présente un MSE de 0,394. Cela suggère que la séparation des vins en fonction de leur couleur améliore la précision des prédictions, comme cela a été observé pour le modèle GRM.

En résumé, nos résultats indique que le viticulteur avait raison de souligner l'importance de distinguer les vins rouges et blancs dans l'analyse. Les résultats de l'étude ont montré que les conditions optimales pour produire des vins de qualité diffèrent entre les vins blancs et les vins rouges. En utilisant des modèles spécifiques pour chaque type de vin, il a été possible d'améliorer la précision des prédictions de qualité du vin en prenant en compte ces différences. Les modèles spécifiques ont également permis d'identifier des prédicteurs importants différents pour chaque type de vin. En somme, la distinction entre les vins rouges et blancs est un élément crucial à considérer pour prédire la qualité du vin avec précision.

Conclusion

En conclusion, les résultats de notre étude ont confirmé l'importance de distinguer les vins rouges et blancs dans l'analyse pour prédire la qualité du vin. Nos modèles spécifiques aux vins rouges et blancs ont montré des performances supérieures au modèle général pour prédire la qualité du vin correspondant à leur couleur respective. Ces résultats pourraient être utiles aux viticulteurs et aux producteurs de vin pour améliorer la qualité de leur production en prenant en compte les différences entre les vins rouges et blancs dans leur processus de production.

Annexes

Annexe A

Y_GRM	Residus	Residus^2	SSE
5,51908	-0,51908	0,26944	693,52
5,6282	-0,62820	0,39464	
5,54128	-0,54128	0,29298	
4,48422	-0,48422	0,23447	
5,8367	-0,83670	0,70007	
5,0802	-0,08020	0,00643	
5,32414	0,67586	0,45679	
5,3442	0,65580	0,43007	
5,60004	0,39996	0,15997	
5,2221	-0,22210	0,04933	
5,43283	-0,43283	0,18734	
5,44218	-0,44218	0,19552	
5,5597	0,44030	0,19386	
5,93502	-0,93502	0,87426	
5,0971	0,90290	0,81523	
5,68532	0,31468	0,09902	
5,14626	-0,14626	0,02139	
4,94574	0,05426	0,00294	

n	1264
K	8
AICc	2843,8304
n	1264
K	8
BIC	2884,852
SST	949,461234
R2	0,2696
R2_ajusté	0,2649
MSE	0,55

Y_OBS	Y_Vin_B	Residus	Residus^2	SSE
6	5,217274	0,78273	0,61266	552,20
6	5,217274	0,78273	0,61266	
8	5,8684672	2,13153	4,54343	
8	5,928663	2,07134	4,29044	
6	5,786203	0,21380	0,04571	
6	5,7400999	0,25990	0,06755	
5	6,3802705	-1,38027	1,90515	
7	5,7854071	1,21459	1,47524	
4	5,398244	-1,39824	1,95509	
6	6,3461397	-0,34614	0,11981	
7	6,1920274	0,80797	0,65282	
7	6,8761923	0,12381	0,01533	
5	5,850692	-0,85069	0,72368	
5	5,4094678	-0,40947	0,16766	
5	5,6096742	-0,60967	0,37170	
5	5,5726901	-0,57269	0,32797	
6	6,0725689	-0,07257	0,00527	
6	5,1567275	0,84327	0,71111	

n	952
K	9
AICc	2200,8545
n	952
K	9
BIC	2244,3905
SST	738,92437
R2	0,2527
R2_ajusté	0,2456
MSE	0,59

Y_OBS	Y_Vin_R	Residus	Residus^2	SSE
5	6,10585	-1,10585	1,22290	130,34
5	5,42452	-0,42452	0,18022	
5	5,78503	-0,78503	0,61627	
4	4,61777	-0,61777	0,38164	
5	5,82477	-0,82477	0,68025	
5	5,24505	-0,24505	0,06005	
6	5,44317	0,55683	0,31006	
6	5,47993	0,52007	0,27047	
6	5,53107	0,46893	0,21990	
5	5,19251	-0,19251	0,03706	
5	5,440235	-0,44024	0,19381	
5	5,577815	-0,57781	0,33387	
6	5,910865	0,08914	0,00795	
5	6,6583	-1,65830	2,74996	
6	5,226385	0,77362	0,59848	
6	5,61336	0,38664	0,14949	
5	5,30357	-0,30357	0,09215	
5	5,077735	-0,07774	0,00604	

n	312
K	7
AICc	627,30192
n	312
K	7
BIC	653,13452
SST	194,842949
R2	0,3310
R2_ajusté	0,3156
MSE	0,43

17	18	19
Y_PredGRM	Y_Pred_Vin_B	Y_Pred_Vin_R
5,05432	4,010297	5,26149
5,68466	4,939038	5,68214
5,0232	3,9896398	5,09705
5,07918	4,0675198	5,11697
5,03856	4,2724124	5,1166
5,25608	4,5959686	5,34023
5,2257	4,3586688	5,36053
5,7568	4,9475198	5,70297
5,0316	4,3152419	5,09132
5,7568	4,9475198	5,70297
5,27926	4,7776263	5,152185
5,51908	4,5383734	6,10585
5,23068	4,3257026	5,04582
5,24922	4,3664926	5,09226
6,05494	5,2592429	5,83131
5,4684	4,5715388	5,44287
5,02404	4,3093134	5,07201
5,6718	4,8414629	5,5166
5,7326	5,2313388	5,47092
5,6282	4,7239662	5,42452
5,54128	4,6937306	5,78503
5,16998	4,4376188	5,28567

Figure 16: Calcul du BIC sur EXCEL