Análise da explicabilidade qualitativa do vinho português

Nelson Estevão[A76434] Rui Reis[A84930] Susana Marques[A84167]

Introdução & Contextualização

O vinho é de importância extrema, a vários níveis, para a sociedade portuguesa. Com isto em mente, e tendo em conta a importância económica, conseguimos ver utilidade na utilização de técnicas de aproveitamento de dados de forma a responder questões fulcrais para a comunidade de vinicultores, de forma a que esta consiga produzir vinhos que melhor se adequam à sociedade e competição de mercado existente.

Para alcançar esse objetivo, utilizamos dois conjuntos de dados, que representam dois tipos de vinhos, branco e tinto. A nossa análise é feita de forma abstrata na maioria dos pontos, focando-nos, quando necessário, na diferença entre os dois tipos. Os *data sets* utilizados possuem ambos um conjunto de 12 atributos, com 4898 amostras para o vinho verde e 1599 para o vinho tinto.

Neste documento, e com o intuito de aderir à metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), o nosso objetivo é o de esclarecer a importância de cada atributo em termos de conhecimento de domínio e inferir informação importante que possamos vir a retirar destes. De seguida, é conduzida uma análise exploratória com o intuito de encontrar conhecimento relevante nos dados, ou confirmar informação proveniente da fase conhecimento de domínio. Por fim, apresentamos a aplicação de diferentes modelos bem como uma análise ao tipo de inferências que esses modelos nos fornecem.

Dados e Estatística Descritiva

Compreensão de Negócio

Como referido anteriormente, os *data sets* encapsulam um total de 12 atributos, sendo que maioria deles são de uma especificidade técnica fora do nosso conhecimento. Pelo que, uma análise mais aprofundada é devida. De seguida, no formato de `Nome, Tipo : Descrição` apresentamos a nossa análise sobre cada atributo, quando conveniente apresentamos abaixo informação específica adicional.

- `fixed acidity`, numérico: Representa a acidez fixa que é a soma dos ácidos fixos. Tartárico e Málico são os mais importantes. Por princípio, quanto mais elevada for a acidez fixa, mais baixa é a volátil. As bactérias acéticas têm dificuldade em desenvolver-se em meios mais ácidos (Afonso 2017).
- `volatile acidity`, numérico: Representa a acidez volátil que é a soma dos ácidos voláteis, que se libertam pela ebulição ou destilação do vinho e traduz o nível de ataque aceto bacteriano ao vinho. Por lei não pode ultrapassar o valor de 1,2 g de ácido acético por litro (Afonso 2017).
- `citric acid`, numérico: O ácido cítrico é um ácido orgânico fraco, que se pode encontrar nos citrinos. É usado como conservante natural (antioxidante), dando um sabor ácido e refrescante na preparação de alimentos e de bebidas (Wikipédia 2020).
- `residual sugar`, numérico: O açúcar residual refere-se aos açúcares que não foram fermentados no vinho acabado. A quantidade de açúcar residual afeta a doçura do vinho. Na união europeia existe uma correspondência entre o açúcar residual e termos de rotulagem (Wu 2020).

- `chlorides`, numérico: Representa quantidade de sal no vinho, no entanto, concentrações moderadas a grandes de cloretos podem dar ao vinho um sabor salgado que pode desmotivar o consumidor. Quando os níveis desses elementos excedem certos limites, a comercialização e venda do vinho pode não ser permitida em alguns países (Coli et al. 2015).
- `free sulfur dioxide`, numérico: Representa a quantidade de dióxido de enxofre não reagido no vinho, por essa razão é capaz de garantir uma ação antioxidante (BRI, sem data).
- `total sulfur dioxide`, numérico: Representa a quantidade de dióxido de enxofre livre mais a que está associada a outros químicos no vinho. Esta quantidade é regulada, por exemplo, pelos Estados Unidos que limita a concentração total de dióxido de enxofre a 350 mg/L. Na EU, o limite legal é de 150mg/L para vinho tinto e 200mg/L para vinho branco (Moroney 2018; BRI, sem data).
- `density`, numérico: Quociente entre a massa e volume do vinho.
- `pH`, numérico: A importância da acidez para o vinho não pode ser subestimada, pois contribui com frescor, atua como um agente conservante e ajuda, notavelmente, com a estabilidade microbiana. O pH do vinho tende a estar entre 2 e 4 (Charest 2019).
- `sulphates`, numérico: Aditivo do vinho, age como um anti-microbiano e anti-oxidante. No entanto, não foi detetada informação relevante sobre este atributo.
- `alcohol`, numérico: Percentagem de total de álcool no vinho.
- `quality`, nominal: Classe objetivo, varia de 0 a 10 e representa a qualidade do vinho.

Percepções Valiosas

A partir das especificações acima, conseguimos encontrar as seguintes percepções valiosas sobre o significado dos atributos na lógica de negócio subjacente.

- Quanto mais elevada for a `fixed acidity`, menor é a `volatile acidity`.
- Bactérias acéticas têm mais dificuldade em desenvolver-se em meios ácidos, `pH < 7`.
- Em Portugal, a `volatile acidity`, por lei, não pode ultrapassar o valor de 1,2 g de ácido acético por litro.
- `citric acid` representa um conservante natural, dando um sabor ácido e refrescante.
- A quantidade de `residual sugar` afeta a doçura do vinho de forma diretamente proporcional.
- `chlorides` elevados podem dar um sabor salgado ao vinho, o que desmotiva o consumidor.
- `pH` do vinho tende a estar entre 2 e 4.

Análise Explorativa

A compreensão dos dados é tanto ou mais importante que compreender o negócio em questão. A partir dos dados podemos extrair relações, questões de interesse, bem como desenvolver modelos de inferência coerentes. Esta secção destina-se a agregar a análise dos nossos dados em duas componentes. Passo a passo, abordamos os atributos mais importantes e de que forma, para ambos os data sets considerados, funciona a dinâmica dos mesmos.

Unidades de Medida

Com base no estudo original de (Cortez et al. 2009), e com intuito de permitir observar os atributos de forma mais intuitiva, somos capazes de estabelecer as seguintes unidades de medida e domínios para cada um dos atributos.

Tabela 1: Unidades de medida e domínios considerados

	Unidade de Medida	Domínio
Fixed Acidity	g(ácido tartárico)/dm^3	$\Re$
Volatile Acidity	g(ácido acético)/dm^3	$[0, 1.2]$
Citric Acid	g/dm^3	$\Re$
Residual Sugar	g/dm^3	$\Re$
Chlorides	g(cloreto de sódio)/dm^3	$\Re$
Free Sulfur Dioxide	mg/dm^3	$\Re$
Total Sulfur Dioxide	mg/dm^3	$[0, 200]$
Density	g/cm^3	$\Re$
pH	sem unidade	[0,14]
Sulphates	g(sulfato de potássio)/dm^3	$\Re$
Alcohol	vol.%	$[0, 100]$
Quality	sem unidade	$\{0, 1, 2, \cdots, 10\}$

Acidez fixa e volátil

Como aferido pelo conhecimento de domínio, somos endereçados para o facto de que tende a existir uma forte correlação negativa, ou inversamente proporcional, entre a acidez fixa e volátil. Numa primeira tentativa, podemos obter as seguintes informações sobre os dados.

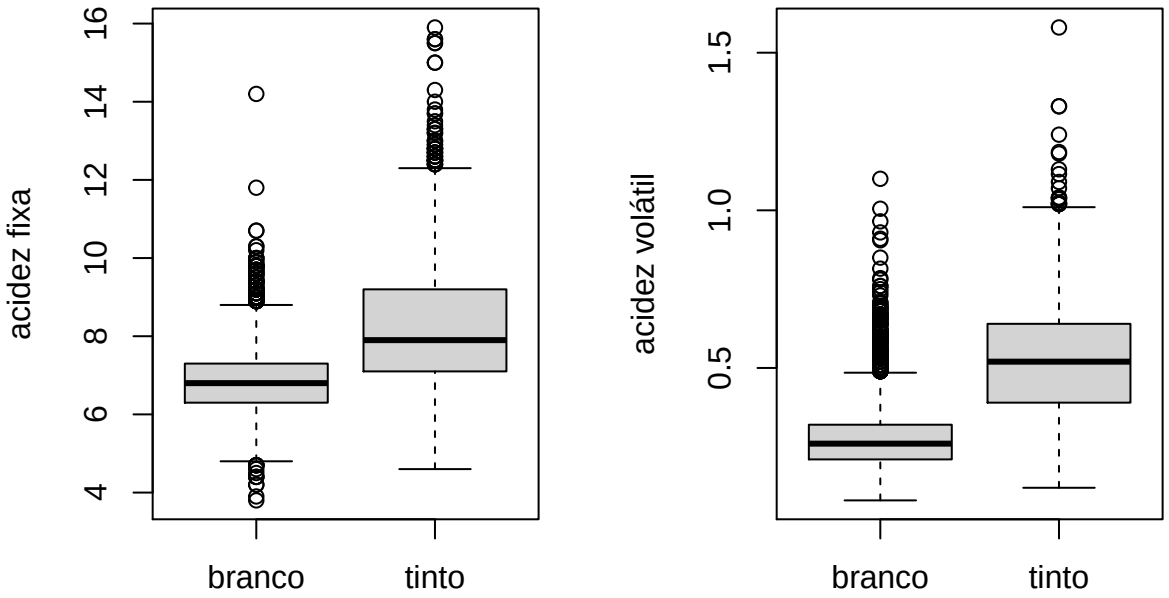


Figura 1: Comparação entre tipos de acidez

A partir dos dados acima somos capazes de perceber que parece existir uma relação, como evidenciado pelo negócio, entre os diferentes tipos de acidez. Para ambos os tipos de vinhos, a caixa de bigodes, dos diferentes atributos, parecem ser muito semelhantes. Numa análise mais especifica, experimentamos calcular a correlação entre os atributos em questão.

Tabela 2: Correlação entre atributos acídicos, para o vinho branco

fixed.acidity	1.000	-0.023
volatile.acidity	-0.023	1.000

Pela correlação não existe indícios que exista uma relação negativa entre aqueles atributos, no caso do vinho branco. Ao tentarmos visualizar a relação dos dois atributos na forma de um gráfico, obtemos o seguinte resultado.

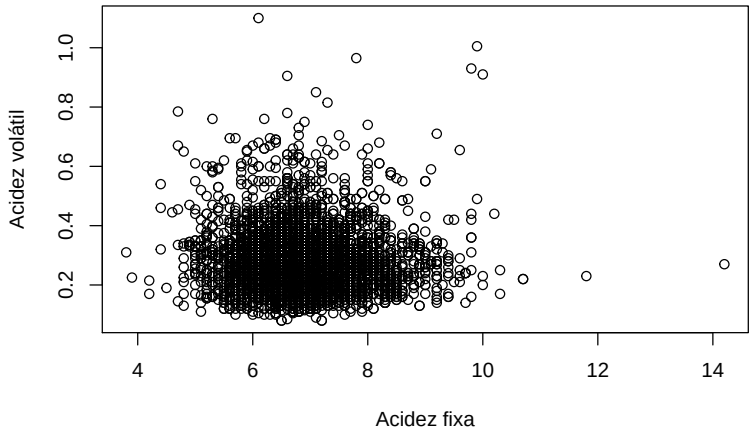


Figura 2: Visualização da relação entre os tipos de acidez

Facilmente verificamos que, apesar de haver uma grande concentração de pontos no gráfico, estes não definem nenhuma forma especifica. O que só nos confirma que o conhecimento que foi obtido pelo conhecimento de domínio não se verifica de forma significativa nos dados. Entre muitos fatores, isto pode-se dever ao tipo de *sampling* do vinho que foi utilizado. Se forem considerados só vinhos de uma determinada região, é evidente que podem existir caraterísticas mais especificas que depois não se verificam na população global.

Cloretos

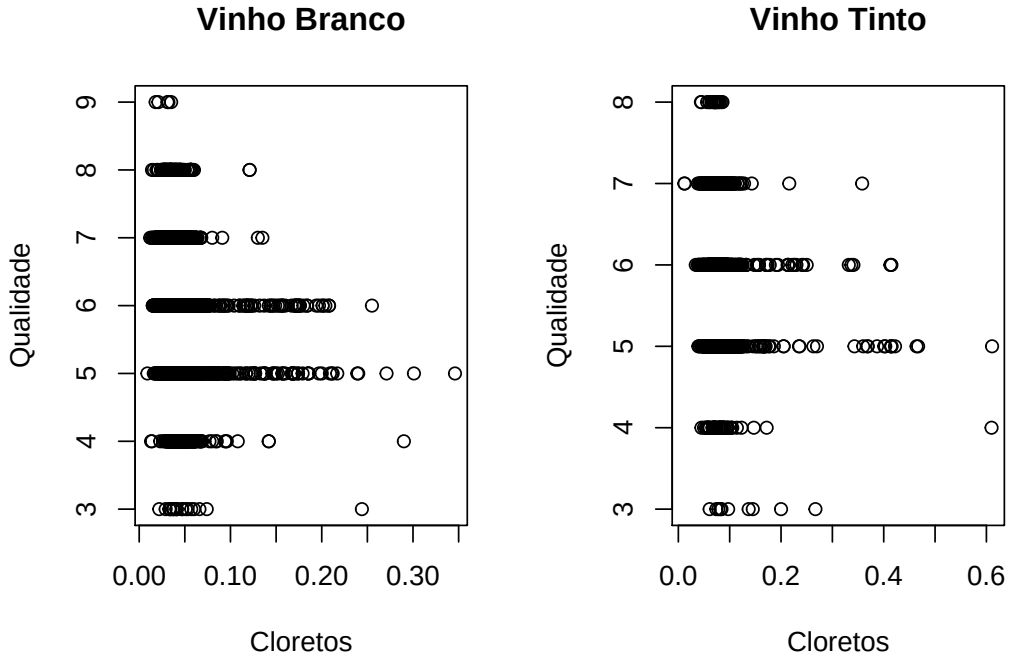


Figura 3: Representatividade dos Cloretos por tipo de Qualidade

Apesar de não ser tão evidente, conseguimos perceber, pela figura abaixo, que à medida que o sal no vinho aumenta, então a avaliação da qualidade desse vinho tende a ser menor, para ambos os tipos.

Dióxido de Enxofre Livre e Total

A relação entre os diferentes tipos de dióxido de enxofre parece ser intuitiva e há indícios para a existência de uma relação de linearlidade entre os dois atributos.

Vinho Tinto Começemos por analisar o caso do vinho tinto, como podemos verificar na figura seguinte. Podemos verificar a existência de uma relação parecida a linear, no entanto, com um formato afunilado. Este formato afunilado pode ser um sinal de heteroscedasticidade, caso onde se verifica variâncias não constantes nos erros. Este problema poder vir a ser uma fonte medidas de erro com elevado valor no futuro, especialmente em modelos que partem do pressuposto de um variância constante, como acontece na regressão linear.

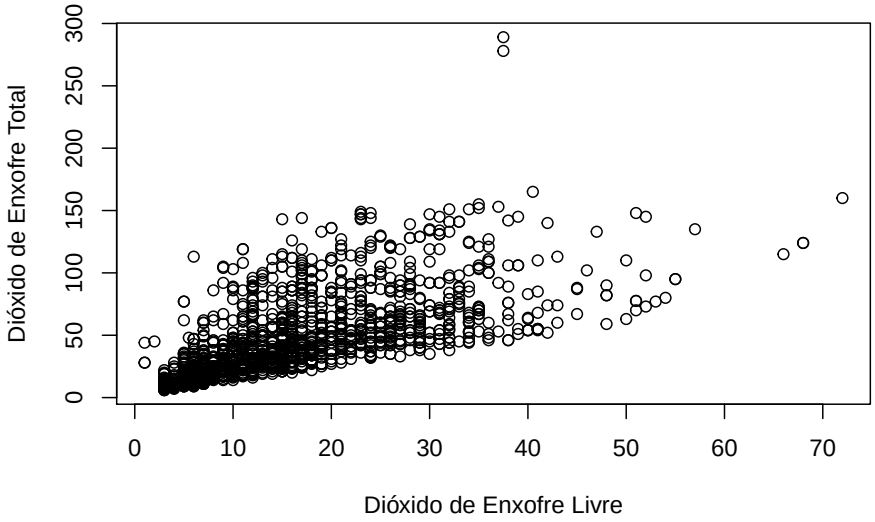


Figura 4: Visualização da relação entre os tipos de dióxido de enxofre

Se tentarmos fazer uma simples regressão linear, rapidamente percebemos que o modelo possui alguma dificuldade em se adaptar a dados com variâncias não constantes nos erros. Como tal, a reta resultante não aparenta espelhar bem a relação com os dados.

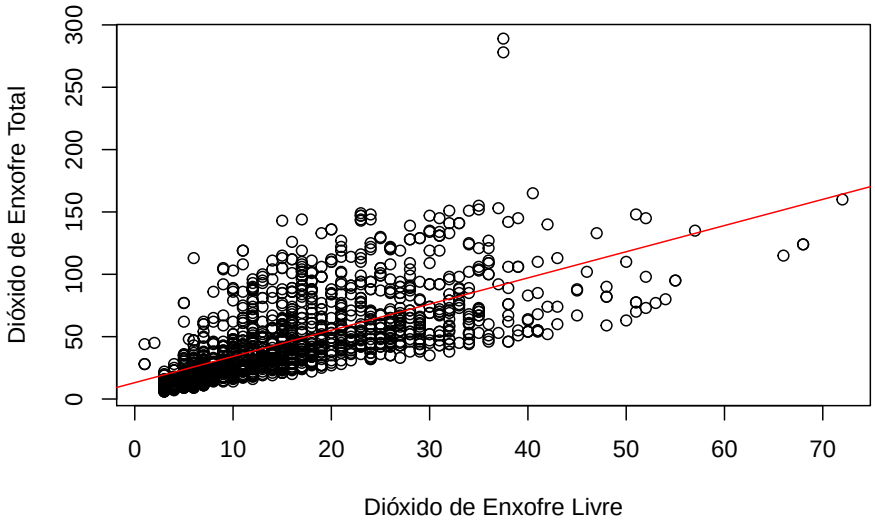


Figura 5: Visualização da relação entre os tipos de dióxido de enxofre com regressão linear

O modelo acima permite-nos calcular um RSE de 24.4816125.

Uma possível solução corresponde a mudar a escala utilizada, normalmente para logarítmica ou raiz quadrada. No nosso caso, escolhemos raiz quadrada por produzir os melhores resultados. Como podemos ver abaixo, a reta resultante da regressão linear parece acompanhar de forma mais adequada os dados. Deixou de se verificar a existência de uma forma afunilada.

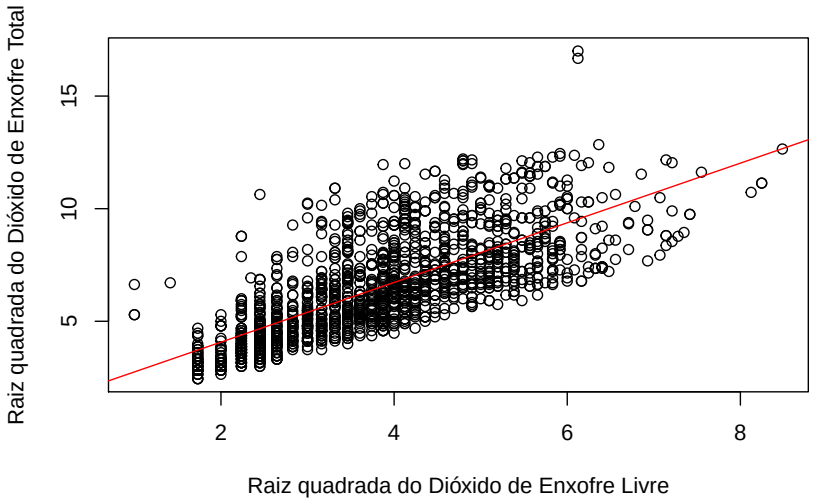


Figura 6: Visualização da relação entre a raiz quadrada dos tipos de dióxido de enxofre com regressão linear

O modelo acima permite-nos calcular um RSE de 1.5177552.

O que representa uma melhoria significativa aos resultados anteriormente obtidos, sem ter a consideração a existência de heteroscedasticidade.

Vinho Branco Para o conjunto do vinho branco conseguimos denotar uma figura também afunilada, evidenciando o mesmo tipo de problemas. Podemos também detetar a existência de um *high leverage point* à direita. Idealmente, deveriamos ter em consideração o impacto deste tipo de pontos. Porém, como estamos a fazer uma análise apenas à heteroscedasticidade, achamos desnecessário ter em consideração todas as componentes.

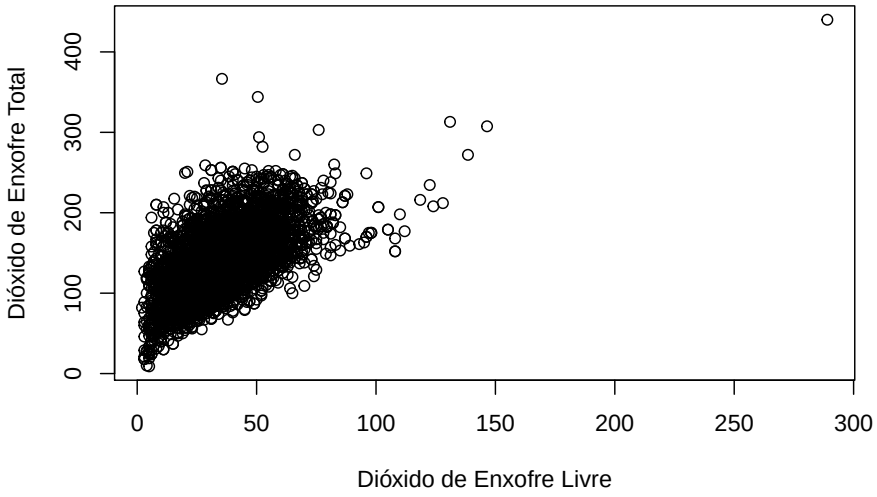


Figura 7: Visualização da relação entre os tipos de dióxido de enxofre

Ao tentarmos utilizar regressão linear, conseguimos perceber que o linha gerada acompanha muito mal os dados.

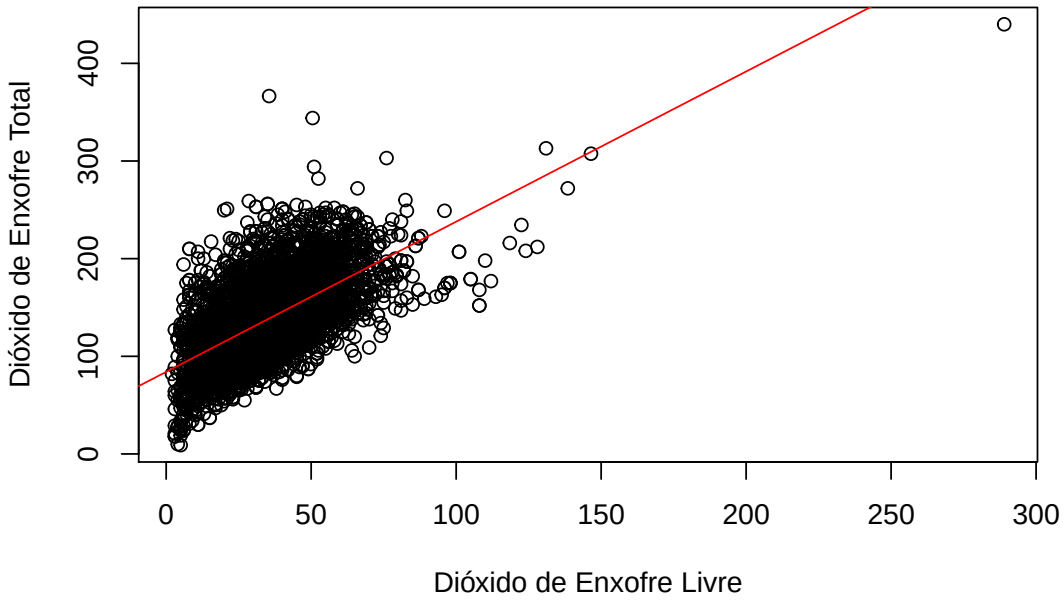


Figura 8: Visualização da relação entre os tipos de dióxido de enxofre com regressão linear

O modelo acima permite-nos calcular um RSE de 33.4908409.

De igual forma ao caso do vinho tinto, podemos alterar a escala dos nosso atributos para evitar o problema das variâncias não-constantes.

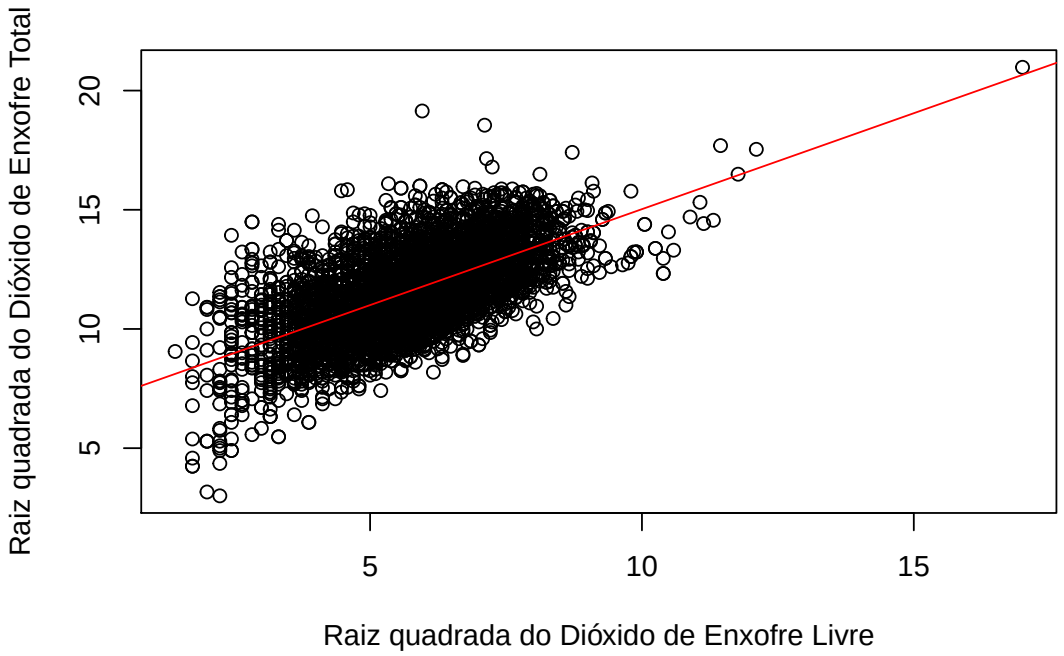


Figura 9: Visualização da relação entre a raiz quadrada dos tipos de dióxido de enxofre com regressão linear

O modelo acima permite-nos calcular um RSE de 1.430711.

A melhoria é extremamente significativa, e facilmente conseguimos perceber que, partindo de uma situação com um problema de heteroscedasticidade tão grave, conseguimos obter uma melhoria de performance cerca

de 20 vezes melhor. Com esta informação, conseguimos confirmar a existência e a solução para o problema de heteroscedasticidade neste par de atributos. O que poderá vir a ser relevante em modelos específicos.

Qualidade

Perceber a distribuição da classe objetivo é igualmente importante a compreender os outros atributos. Quando enfrentados com a variável de qualidade somos atingidos pelos seguinte gráficos.

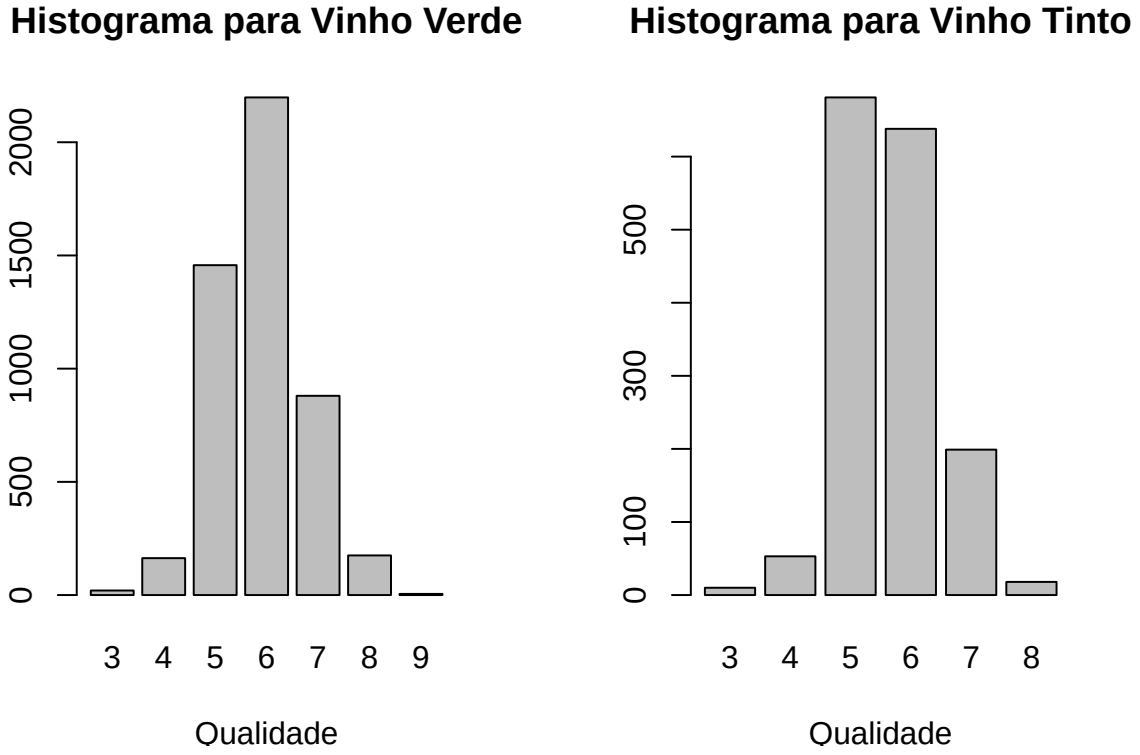


Figura 10: Distribuição da Qualidade entre vinhos

Que nos permitem perceber que existe uma péssima representatividade da classe objetivo, tendo em conta que a cardinalidade desta é de 10. Existem até qualidade que nunca se verificam. Assume-se que, por os dados terem sido recolhidos de forma sensorial, recorrendo a júris, existe um bias implícito associado. É aceitável considerar que os júris tendem a preferir não dar classificações extrema, nem muito boas, nem muito más, o que se verifica na prática, em ambos os *data sets*, pois os valores mais frequentes são o 5 e 6. Isto implica outro problema, nomeadamente, a má separação inter-classe devido a se tratar de um problema que depende da subjetividade do humano. É razoável assumir que 2 júris podem dar classificações completamente apostas a um mesmo vinho que, do ponto de vista laboratorial, seria considerado ótimo.

Posto isto, conseguimos prever *a priori* que qualquer modelo desenvolvimento deverá possuir uma *accuracy* pequena, devido, principalmente, ao problema da má separação entre as classes.

Análise Genérica

Com o intuito de tentar abordar atributos que achamos que não foram devidamente abordados, tentamos ilustrá-los utilizando a seguinte *scatter matrix*. Da qual denotamos os seguintes pontos.

- Parece existir uma forte relação positiva entre o açúcar residual e a densidade do vinho.
- Parece existir uma relação entre o sal e a densidade.
- Apesar de não tão óbvia, parece também existir uma relação entre o ácido cítrico e o sal.

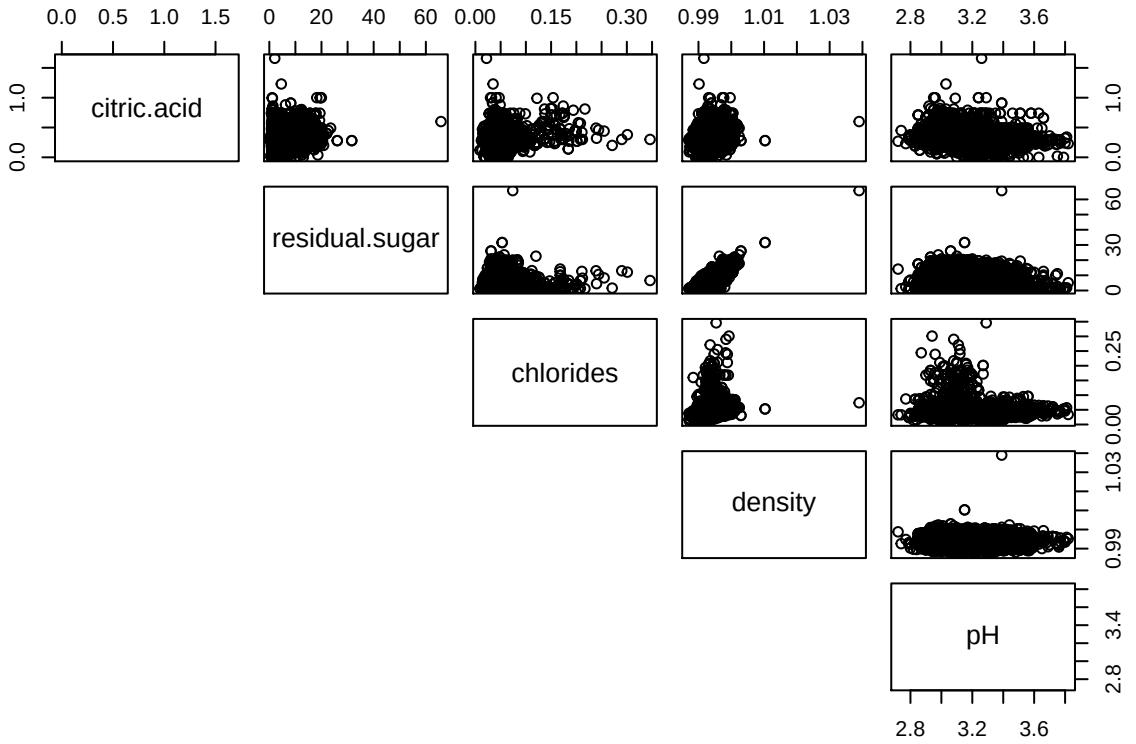


Figura 11: Visualização da relação entre os vários atributos

De forma a introduzir um nível mais técnico, utilizamos uma correlação para tentar determinar se estas relações existem de facto.

Tabela 3: Correlação entre atributos de interesse, para o vinho branco

citric.acid	1.000	0.094	0.114	0.150	-0.164
residual.sugar	0.094	1.000	0.089	0.839	-0.194
chlorides	0.114	0.089	1.000	0.257	-0.090
density	0.150	0.839	0.257	1.000	-0.094
pH	-0.164	-0.194	-0.090	-0.094	1.000

Partindo da tabela acima, conseguimos evidenciar que a única relação que verdadeiramente se evidencia como forte correlação positiva é entre o açúcar residual e a densidade. Poderá vir a ser interessante remover um destes atributos na tentiva de perceber como os modelos respondem.

Questões Relevantes

Do conjunto de análises efetuadas acima, surgem um conjunto de questões de interesse, que ilustram o foco do nosso estudo em secções adiante.

- **Questão 1**: Quais atributos devem ser considerados para melhorar a qualidade de um vinho?
- **Questão 2**: Como lidar com a má separação de classes inerente a avaliações sensoriais?
- **Questão 3**: Qual o impacto de atributos não significativos no mercado do vinho?

Modelação e Resultados

Regressão Logística

Descrição e Funcionamento

Quando a variável dependente é uma variável *dummy* (também conhecida como variável binária), ou seja, assume o valor 0 ou o valor 1, o modelo de probabilidade linear tem duas falhas principais que são as probabilidades ajustadas podem ser menores que zero ou maiores que um e o efeito parcial de qualquer variável explicativa é constante (Wooldridge 2015).

Essas falhas podem ser superadas usando modelos mais sofisticados, como por exemplo, um modelo logit (criado a partir de uma regressão logística). Em vez de modelar a variável dependente diretamente, a regressão logística modela a probabilidade de pertencer a uma categoria específica (James et al. 2013).

O modelo calcula $p(X) = P(Y = 1|X)$, variando entre 0 e 1. Para cumprir este requisito podemos partir da equação 1.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

Podemos chegar a uma equação equivalente sem a fração, ver equação 2. A quantidade $\frac{p(X)}{1-p(X)}$ são as *odds* de um evento acontecer. Estas podem tomar qualquer valor entre 0 e ∞. Valores próximos de 0 indicam baixa probabilidade e valores próximos de ∞ indicam muito elevada probabilidade.

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X} \tag{2}$$

Na equação anterior os paramentos não são lineares com as *odds*. Se aplicarmos a função do logaritmo natural em ambos os lados, obtemos a equação 3. Continua a não ser linear uma vez que o modelo não corresponde a uma reta afim. No entanto, já foi possível remover o expoente.

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \tag{3}$$

O lado esquerdo da equação 3 chama-se *log-odds* ou *logit*. O que quer dizer que o acréscimo de uma unidade em X, multiplica e^{β} as *odds* tendo em conta o valor atual de X, isto porque a equação não é linear (James et al. 2013).

Casos de Estudo

No caso dos dados em análise, geramos a partir do campo `quality` uma variável binária que indica se o vinho é de boa ou má qualidade. A escolha da classificação vai depender do limiar definido. Pela análise da variabilidade dos dados na secção da estatística descritiva iremos definir como limiar de qualidade do vinho o valor 6. Assim, vinhos com esta classificação ou mais serão considerados de boa qualidade.

A classificação foi guardada numa nova coluna chamada `is_good`. O valor 1 representa que estamos perante um vinho de boa qualidade e o valor representa o contrário. A distribuição dos dados neste caso pode ser observado na figura 12.

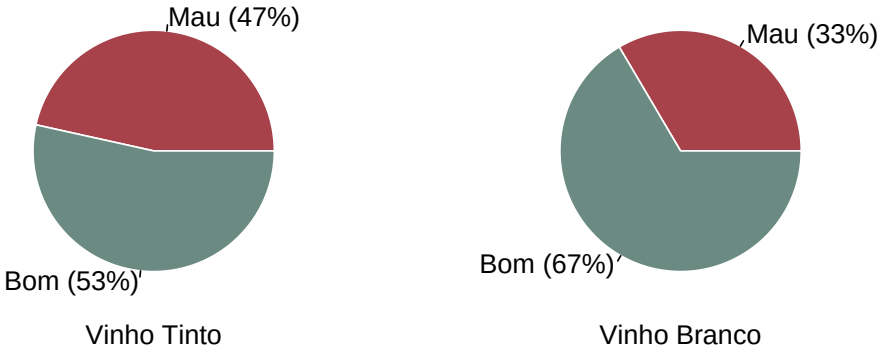


Figura 12: Gráfico com distribuição binária da qualidade

Vinho Branco O modelo será da forma apresentado na equação 3. A modelação para o vinho branco exigiu a remoção de atributos com p-valores não significativos. A lista é a seguinte:

- `citric.acid`
- `total.sulfur.dioxide`
- `alcohol`
- `chlorides`

Tabela 4: Modelo Logistico para o Vinho Branco

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	767.727	29.248	26.249	0.000
fixed.acidity	0.437	0.051	8.618	0.000
volatile.acidity	-6.058	0.387	-15.638	0.000
residual.sugar	0.341	0.016	21.796	0.000
free.sulfur.dioxide	0.007	0.002	3.213	0.001
density	-785.568	29.956	-26.224	0.000
pH	2.844	0.277	10.281	0.000
sulphates	2.521	0.342	7.372	0.000

A equação final é apresentada em 4 e o valores dos parametros podem ser consultados na tabela anterior.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \cdot fixed.acidity + \beta_2 \cdot volatile.acidity + \beta_3 \cdot residual.sugar + \beta_4 \cdot free.sulfur.dioxide + \beta_5 \cdot density \tag{4}$$

Por forma a testar o modelo apresentado na equação 4 foi feito *k-fold cross validation*, com $k = 10$. A precisão verificada foi de 74.7%, com uma taxa de falsos positivos de 2% e de falsos negativos de 0.8%.

Vinho Tinto No vinho tinto foram removidos diferentes atributos dos que foram removidos para o modelo logístico do vinho branco. Os atributos removidos foram os seguintes:

- `citric.acid`
- `fixed.acidity`
- `density`
- `residual.sugar`
- `pH`

Tabela 5: Modelo Logistico para o Vinho Tinto

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-8.150	0.809	-10.074	0.000
volatile.acidity	-2.896	0.371	-7.813	0.000
chlorides	-4.421	1.432	-3.087	0.002
free.sulfur.dioxide	0.024	0.008	2.988	0.003
total.sulfur.dioxide	-0.018	0.003	-6.468	0.000
sulphates	2.706	0.428	6.329	0.000
alcohol	0.859	0.071	12.151	0.000

Assim sendo, os resultados apresentados na tabela anterior correspondem à equação 5.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \cdot volatile.acidity + \beta_2 \cdot chlorides + \beta_3 \cdot total.sulfur.dioxide + \beta_4 \cdot free.sulfur.dioxide \tag{5}$$

Também para este modelo foi usado *k-fold cross validation*, obtendo a precisão de 74.7%, com uma taxa de falsos positivos de 1.3% e de falsos negativos de 1.5%.

Conclusões

Através da tabela 6 é possível concluir que para o vinho branco o pH e os sulfatos têm o maior impacto positivo no vinho, ou seja, são os que aumentam mais as chances de ser um bom vinho. No sentido oposto, a densidade e a acidez volátil são os os atributos que têm o maior impacto negativo.

Tabela 6: Modelo Logit em formato de Odds

	Vinho Branco			Vinho Tinto		
Characteristic	OR	95% CI	p-value	OR	95% CI	p-value
fixed.acidity	1.55	1.40, 1.71	<0.001			
volatile.acidity	0.00	0.00, 0.00	<0.001	0.06	0.03, 0.11	<0.001
residual.sugar	1.41	1.36, 1.45	<0.001			
free.sulfur.dioxide	1.01	1.00, 1.01	0.001	1.02	1.01, 1.04	0.003
density	0.00	0.00, 0.00	<0.001			
pH	17.2	10.0, 29.6	<0.001			
sulphates	12.4	6.40, 24.5	<0.001	15.0	6.60, 35.3	<0.001
chlorides				0.01	0.00, 0.19	0.002
total.sulfur.dioxide				0.98	0.98, 0.99	<0.001
alcohol				2.36	2.06, 2.72	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval, OR = Odds Ratio, CI = Confidence Interval

Já no vinho tinto, o álcool e os sulfatos são os atributos com maior impacto positivo e a acidez volátil e os cloretos são os atributos com maior impacto negativo nas chances de ser um vinho de boa qualidade.

Os restantes atributos apresentados são significativos no modelo, no entanto pequenas variações não são muito significativas nas chances.

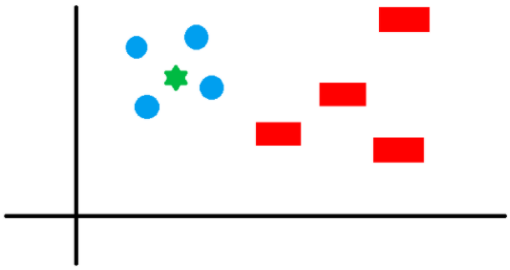
KNN

Descrição e Funcionamento

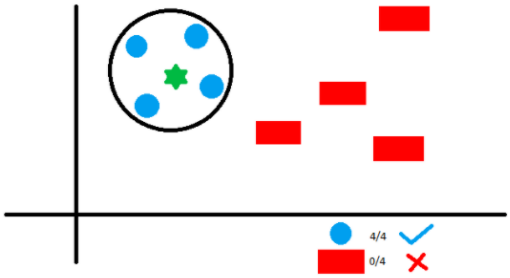
Assumindo-se que se tem vários grupos de amostras classificadas e que os itens presentes nos grupos são de natureza homogênea. Tem-se um exemplo de uma amostra que precisa de ser classificada num desses vários grupos. O algoritmo dos K vizinhos mais próximos (KNN) pode ser usado para resolver tal problema já que consiste em armazenar todos os casos disponíveis e classificar os novos casos pela maioria dos votos dos seus K vizinhos. Assim, este algoritmo consegue separar os dados sem classificação em grupos bem definidos.

É um algoritmo de classificação e regressão muito simples. No caso de classificação, novos pontos de dados são classificados em uma classe particular com base na votação dos vizinhos mais próximos. No caso da regressão, os novos dados são caraterizados com base nas médias do valor mais próximo. É um algoritmo de aprendizagem supervisionada mas não tem grande aprendizagem com os dados de treino. O método padrão é a distância euclidiana (distância mais curta entre 2 pontos) usada para variáveis contínuas, enquanto que para variáveis discretas, como para classificação de texto, a métrica de sobreposição (Distância de Hamming) é a melhor escolha.

A seguir está uma distribuição de círculos azuis (CA) e retângulos vermelhos (RV).



Pretende-se descobrir a classe da estrela verde que pode ser apenas CA ou RV. O "K" neste algoritmo é o número de vizinhos mais próximos nos quais desejamos votar. Assim, por exemplo, digamos K=4 e faremos um círculo com a estrela verda para incluir 4 pontos de dados no plano, conforme a figura seguinte:



Os 4 pontos mais próximos da estrela verde são todos CA, logo com um bom nível de confiança, pode-se afirmar que a estrela verde deve pertencer à classe CA.

Neste exemplo, a escolha tornou-se bastante óbvia porque todos os 4 votos do vizinho mais próximo foram para a classe CA, concluindo-se que a escolha do parâmetro K é muito importante neste algoritmo.

Este algoritmo tem certos requisitos específicos: É frequente decidir K tendo em conta a raiz quadrada do número de dados. Mas um K de valor elevado tem benefícios que incluem a redução da variância devido a dados ruidosos; o efeito colateral é o desenvolvimento de uma tendência para ignorar os baixos padrões que podem ter perceções úteis. Para colocar todos os dados na mesma escala (por exemplo: 0 a 1) deverá-se

escalar os dados ou ocorrer a normalização, caso contrário, irá dar-se mais peso aos dados que são mais elevados em valor, independentemente da escala/unidade.

Algo bastante positivo acerca deste algoritmo que nos levou a testa-lo foi ser altamente imparcial por natureza e não fazer nenhuma suposição prévia dos dados subjacentes, sendo um dos algoritmos mais populares por ser simples, eficaz e fácil de implementar.

Por outro lado, por ser um algoritmo extremamente simples, a construção deste algoritmo requer tempo investido na preparação de dados (especialmente tratando dados nulos e categóricos) para obter um modelo robusto que é sensível a *outliers*.

Caso de estudo

Vinho Tinto No nosso projeto, começamos por usar o caso do vinho tinto para experimentar a abordagem do algoritmo KNN.

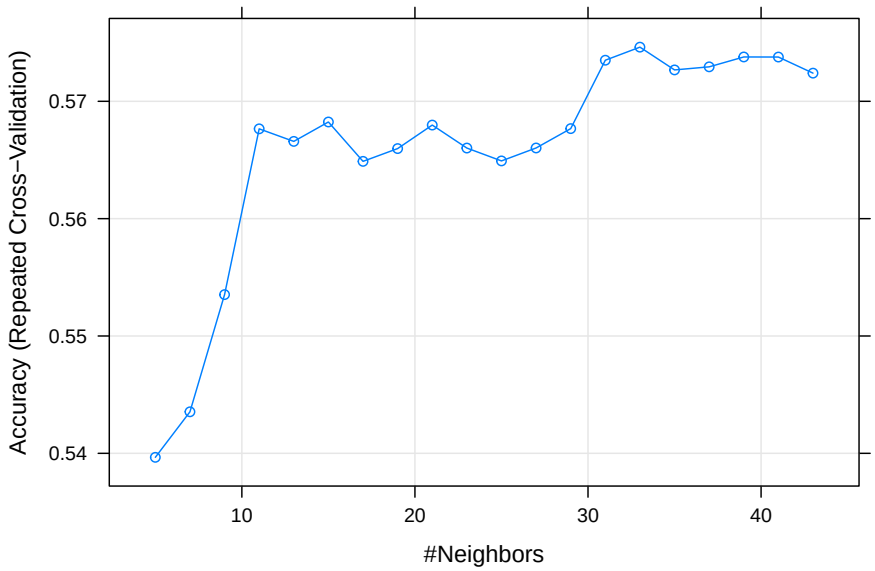
Em primeiro lugar repartimos os dados em dados de treino e em dados de teste usando a função `createDataPartition()` da biblioteca `caret`.

A seguir verificamos a distribuição dos dados originais e dos dados repartidos, concluindo que todos os valores para cada classe são idênticos.

Como já referido o algoritmo KNN requer que as variáveis sejam normalizadas ou escaladas. Optamos por as centralizar e escalar com métodos da biblioteca `caret`.

A seguir optamos por treinar o modelo, não nos esquecendo de antes fazer *cross-validation* que se repetiu 10 vezes para 10 *folds*.

Podemos observar no gráfico seguinte que o melhor valor de K para este dataset é K=33 com uma acurácia de 0.5746151.



Através da matriz de confusão podemos observar várias predições e estatísticas importantes:

Usando como exemplo o número que se encontra na linha 5 e na coluna 5, este número refere-se a 122 casos verdadeiros positivos, ou seja a 122 casos que previram um vinho de qualidade 5 e realmente aconteceu. Já o número da linha 5 e coluna 6 refere-se a 57 casos que previram um vinho de qualidade 5 mas o resultado é um vinho de qualidade 6, tornando-se um falso positivo (relativo a 5). Já na linha 6, coluna 5 o número 47 é um falso negativo porque não se preveu 5 (preveu-se 6) mas o resultado foi 5.

Também é confirmado que a precisão do modelo é 0.5894, podendo-se verificar a *sensitivity* e a *specifity* de cada classe e observar que para a qualidade 5 e 6 do vinho tinto esses valores são maiores indo de encontro

com a ideia de que o dataset estava desbalanceado.

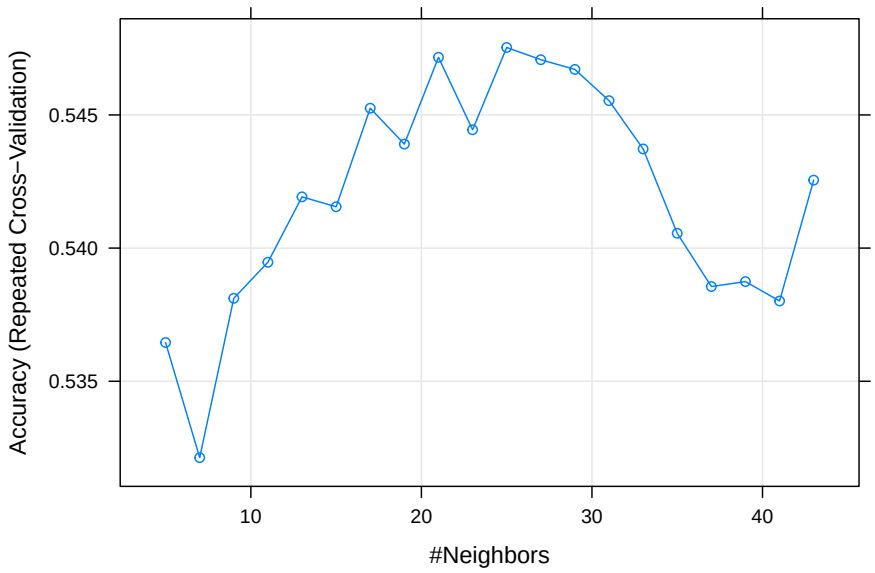
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   3   4   5   6   7   8
##          3   0   0   0   0   0   0
##          4   0   0   0   0   0   0
##          5   2   9 121  58   7   0
##          6   0   3  48  95  26   3
##          7   0   1   1   6  16   1
##          8   0   0   0   0   0   0
##
## Overall Statistics
##
##                Accuracy : 0.5844
##                  95% CI : (0.5342, 0.6333)
##     No Information Rate : 0.4282
##     P-Value [Acc > NIR] : 2.896e-10
##
##                   Kappa : 0.311
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
## Sensitivity          0.000000  0.00000   0.7118   0.5975  0.32653  0.00000
## Specificity          1.000000  1.00000   0.6652   0.6639  0.97414  1.00000
## Pos Pred Value            NaN      NaN   0.6142   0.5429  0.64000      NaN
## Neg Pred Value       0.994962  0.96725   0.7550   0.7117  0.91129  0.98992
## Prevalence           0.005038  0.03275   0.4282   0.4005  0.12343  0.01008
## Detection Rate       0.000000  0.00000   0.3048   0.2393  0.04030  0.00000
## Detection Prevalence 0.000000  0.00000   0.4962   0.4408  0.06297  0.00000
## Balanced Accuracy    0.500000  0.50000   0.6885   0.6307  0.65033  0.50000
```

Vinho Branco Experimentemos a abordagem do algoritmo KNN com o vinho branco agora. Todos os procedimentos foram os mesmos que para o vinho tinto com resultados obtidos ligeiramente diferentes.

Fazemos a distribuição dos dados originais e dos dados repartidos, mas desta vez para o vinho branco, concluindo que todos os valores para cada classe são idênticos.

Após treinarmos o modelo de igual forma a como se sucedeu com o vinho tinto:

Pelo gráfico seguinte podemos concluir que o melhor valor de K para este *dataset* é K=25 com uma precisão de 0.5470733.



Tal e qual como ocorria no vinho tinto, através da matriz de confusão podemos observar várias predições e estatísticas importantes:

Usando como exemplo o número que se encontra na linha 5 e na coluna 5, este número refere-se a 208 casos verdadeiros positivos, ou seja a 122 casos que previram um vinho de qualidade 5 e realmente aconteceu. Já o número da linha 5 e coluna 6 refere-se a 120 casos que previram um vinho de qualidade 5 mas o resultado é um vinho de qualidade 6, tornando-se um falso positivo (relativo a 5). Já na linha 6, coluna 5 o número 148 é um falso negativo porque não se preveu 5 (preveu-se 6) mas o resultado foi 5.

Também é confirmado que a precisão do modelo é 0.5556.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   3   4   5   6   7   8   9
##          3   0   0   0   0   0   0   0
##          4   0   1   0   0   0   0   0
##          5   0  21 205 119   4   2   0
##          6   5  16 152 385 133  19   1
##          7   0   2   7  45  83  21   0
##          8   0   0   0   0   0   1   0
##          9   0   0   0   0   0   0   0
##
## Overall Statistics
##
##                Accuracy : 0.5524
##                  95% CI : (0.524, 0.5805)
##     No Information Rate : 0.4493
##     P-Value [Acc > NIR] : 3.167e-13
##
##                   Kappa : 0.2892
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 3  Class: 4 Class: 5 Class: 6 Class: 7  Class: 8
```

```
## Sensitivity          0.000000 0.0250000   0.5632   0.7013  0.37727 0.0232558
## Specificity          1.000000 1.0000000   0.8298   0.5156  0.92515 1.0000000
## Pos Pred Value            NaN 1.0000000   0.5840   0.5415  0.52532 1.0000000
## Neg Pred Value       0.995908 0.9680590   0.8175   0.6791  0.87124 0.9656020
## Prevalence           0.004092 0.0327332   0.2979   0.4493  0.18003 0.0351882
## Detection Rate       0.000000 0.0008183   0.1678   0.3151  0.06792 0.0008183
## Detection Prevalence 0.000000 0.0008183   0.2872   0.5818  0.12930 0.0008183
## Balanced Accuracy    0.500000 0.5125000   0.6965   0.6084  0.65121 0.5116279
##                       Class: 9
## Sensitivity          0.0000000
## Specificity          1.0000000
## Pos Pred Value             NaN
## Neg Pred Value       0.9991817
## Prevalence           0.0008183
## Detection Rate       0.0000000
## Detection Prevalence 0.0000000
## Balanced Accuracy    0.5000000
```

Conclusões

Com este algoritmo, comprovou-se mais uma vez que os dados do *dataset* tinham uma distribuição fraca, sendo a precisão no vinho tinto de 0.5894 e no vinho branco de 0.5556. Estes valores vão de encontro com os resultados observados nos outros algoritmos, mostrando que apesar de haver bastantes verdadeiros positivos e falsos negativos, existe um grande número de falsos positivos e verdadeiros negativos ao treinar e testar este modelo.

Linear Discriminant Analysis

Descrição & Funcionamento

Apesar das suas vantagens, a regressão logística apresenta maus resultado para classes mal separadas, pos este torna-se bastante instável. O LDA tende a ser mais estável que a regressão logística e é normalmente utilizado para problemas de classificação com mais de duas classes de resposta.

No algoritmo de LDA, é assumido que os atributos são retirados de uma distribuição normal, com médias específicas por classe e um matriz de covariância comum entre todos os atributos.

Casos de Estudo

Predetendemos efetuar *backwards feature selection* de forma a aferir os atributos que mais influenciam o modelo produzido, minimizando a média de erros e variância. Assim sendo, para ambos os conjuntos começamos com um modelo que utiliza todos os atributos e incrementalmente reduzimos o número de elementos considerados, desde que estes forneçam melhorias ao nosso modelo. No entanto, podem existir casos em que a diferença de média de erros original e da versão reduzida é residual, nesse caso optamos pelo modelo mais simples. Para estes casos de estudo foi utilizado *cross validation* com $k = 10$ *folds*.

A metodologia de *backwards feature selection* consiste em realizar iterações sendo que uma iteração corresponde aos passos necessários para descobrir o primeiro elemento que leva ao modelo com menor erro. Na iteração seguinte, partimos desse modelo e voltamos a remover todos os atributos até encontrarmos o novo candidato. O caso de paragem corresponde à iteração em que nenhum atributo candidato seja identificado, na qual chegamos ao modelo simplista com menor erro.

Vinho Tinto Considerando um modelo inicial que contempla todos os atributos, somos capazes de obter a seguinte média de erro. Logo, qualquer modelo mais simples que produz uma média menor do erro ou muito semelhante, desde que mais simples, será preferido.

```
## [1] 0.4046502
```

Abaixo apresentado surge a tabela relativa aos resultados obtidos em cada iteração, indicando a respectiva iteração, atributo removido e média do modelo resultante.

Tabela 7: Resumo da aplicação de backwards feature selection

	Atributo Removido	Erro Obtido
Ponto de Partido	-	0.404650157232704
1ª Iteração	fixed acidity	0.401493710691824
2ª Iteração	residual sugar	0.399626572327044
3ª Iteração	pH	0.409005503144654
4ª Iteração	density	0.403368710691824
5ª Iteração	citric acid	0.406517295597484
6ª Iteração	free sulfur dioxide	0.410294811320755
Última Iteração	chlorides	0.412083333333333

Como se pode observar pela tabela acima, a aplicação do método de *backwards feature selection* permitiu reduzir o erro do nosso modelo na percentagem de -0.7433176.

O que representa uma diferença insignificante na precisão do modelo. No entanto, obtemos um modelo muito mais simples que produz resultos muito similares, o que nos indica que este é o conjunto de variáveis explicativas da variação da qualidade. Podemos observar os resultados da variação dos erros da seguinte forma, que nos permitem perceber que, mesmo alterando atributos, os erros continuam iguais.

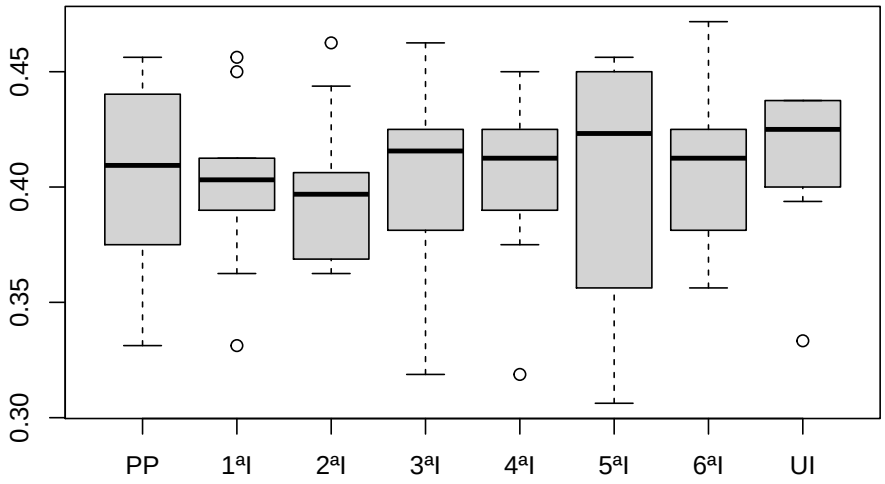


Figura 13: Distribuição dos erros utilizando backwards features selection

Em suma, para o conjunto de dados do vinho tinto, o modelo que mostrou ter mais sucesso foi aquele que apenas considerava os seguintes atributos como significativos:

- `volatile.acidity`
- `total.sulfur.dioxide`
- `sulphates`
- `alcohol`

Vinho Branco De igual forma, partindo de um modelo que considera todos os atributos existentes, aplicamos a mesma metodologia de forma incremental. Inicialmente, com o modelo completo, obtemos um resultado do erro de 0.4691753.

Com esta metodologia, fomos capazes de obter a seguinte tabela que resume os resultados obtidos.

Tabela 8: Resumo da aplicação de backwards feature selection

	Atributo Removido	Erro Obtido
Ponto de Partido	-	0.46917530946805
1ª Iteração	density	0.467134493141519
2ª Iteração	fixed acidity	0.467947474071596
3ª Iteração	citric acid	0.468568082970893
4ª Iteração	pH	0.469976580796253
5ª Iteração	total sulfur dioxide	0.470589662094346
6ª Iteração	free sulfur dioxide	0.469987453997993
Última Iteração	chlorides	0.468350618936099

Como se pode observar pela tabela acima, a aplicação do método de *backwards feature selection* permitiu reduzir o erro do nosso modelo na percentagem de 0.0824691.

Como sucede no caso do LDA no data set do vinho tinto, a diferença é insignificante, pelo que preferimos o modelo mais simples. Podemos compreender esta variação nos erros pelo seguinte gráfico.

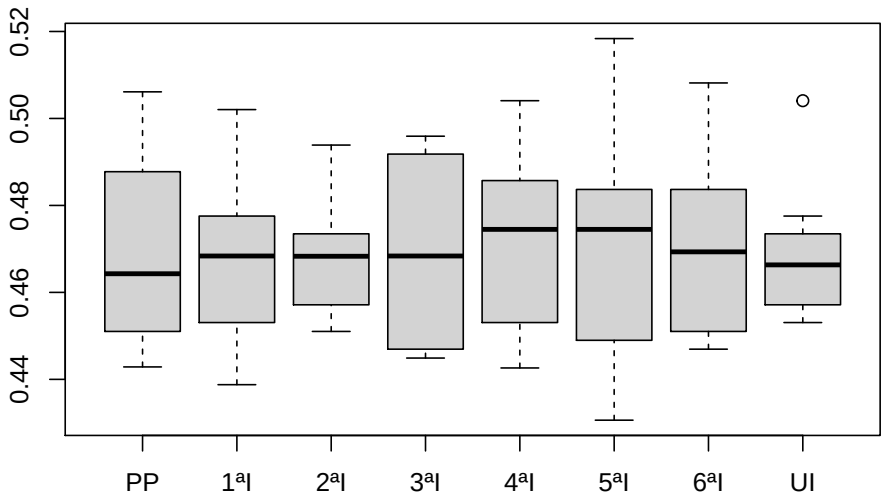


Figura 14: Distribuição dos erros utilizando backwards feature selection

Em suma, para o conjunto de dados do vinho branco, o modelo que mostrou ter mais sucesso foi aquele que apenas considerava os seguintes atributos como significativos:

- `volatile.acidity`
- `residual.sugar`
- `sulphates`
- `alcohol`

Conclusões

Através do método de LDA, conseguimos chegar a modelos distintos que utilizam diferentes atributos para, dentro do seu data set, prever os resultados de forma ótima. Com este modelo, conseguimos modelar os comportamentos e gostos dos provadores de vinho e conseguimos denotar que:

- Em *vinhos tintos* o nível de álcool é o atributo com mais impacto na previsão da qualidade, sendo o atributo de densidade ignorado. O dióxido de enxofre e sulfatos tem um impacto direto na qualidade

sensorial do vinho. Denota-se também que a quantidade de sulfato, que se considerava ser apenas para efeitos anti-oxidantes, tem impacto na qualidade, o que pode indiciar uma alteração do sabor.
- Em *vinhos brancos* o nível de álcool é o atributo com mais impacto na previsão da qualidade, a única diferença de atributos significativos responde ao açúcar residual. Nos vinhos brancos, este atributo é significativo, enquanto que nos vinhos tintos é preferido o dióxido de enxofre total.

Os resultados acima permitem chegar à conclusão que em vinhos brancos o mais importante é a doçura e acidez. Enquanto que em vinhos tintos, os fatores mais importantes são o nível do álcool e quantidade de dióxido de enxofre.

Quadratic Discriminant Analysis

Descrição & Funcionamento

À semelhança do LDA, o QDA também assume que os atributos são retidados de uma distribuição gaussiana. No entanto, e em contrário ao LDA, assume que existe uma matriz de covariância para cada um das classes considerar, sendo que o LDA considera uma matriz de covariância comum a todas as classes. Por esta razão, o QDA apresenta-se como um método alternativo ao LDA e mais flexível. O LDA, no entanto, é preferido para conjuntos com um número de amostras reduzidas, onde reduzir a variância total é crucial.

Casos de Estudo

Para garantir a consistência, foi aplicada a mesma metodologia presente na secção de LDA, com a vertente de *backwards feature selection* e *cross validation*.

Vinho Tinto Considerando um modelo inicial que contempla todos os atributos, somos capazes de obter a seguinte média de erro. Logo, qualquer modelo mais simples que produz uma média menor do erro ou muito semelhante, desde que mais simples, será preferido.

```
## [1] 0.441195
```

Abaixo apresentado surge a tabela relativa aos resultados obtidos em cada iteração, indicando a respectiva iteração, atributo removido e média do modelo resultante.

Tabela 9: Resumo da aplicação de backwards feature selection

	Atributo Removido	Erro Obtido
Ponto de Partido	-	0.441194968553459
1ª Iteração	pH	0.427314704243293
2ª Iteração	chlorides	0.420412387548762
3ª Iteração	free sulfur dioxide	0.429189554971738
4ª Iteração	density	0.422307141151182
5ª Iteração	fixed acidity	0.421634424010827
6ª Iteração	citric acid	0.417888703128732
Última Iteração	residual sugar	0.407797946023406

Como se pode observar pela tabela acima, a aplicação do método de *backwards feature selection* permitiu reduzir o erro do nosso modelo na percentagem de 3.3397023.

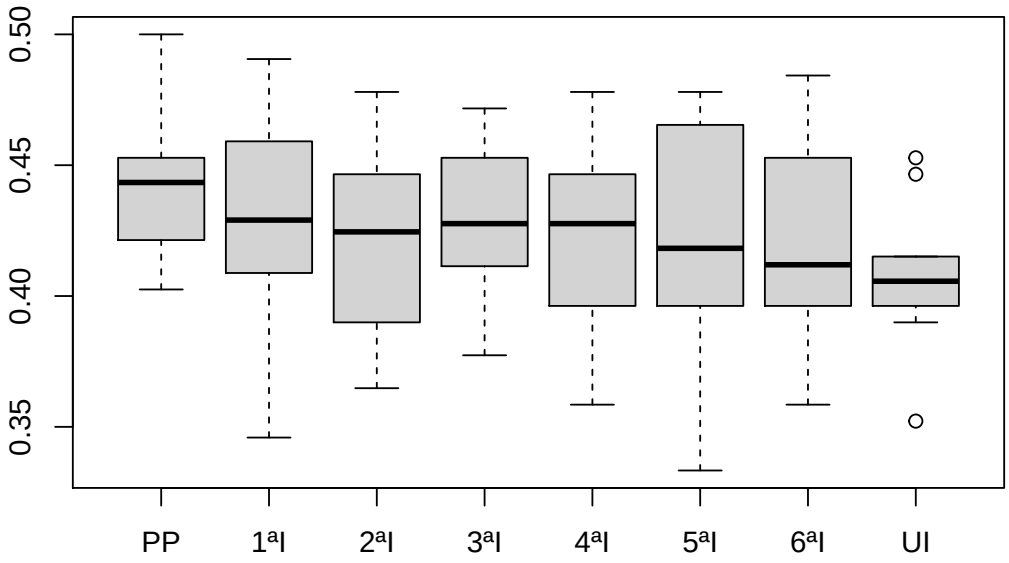


Figura 15: Distribuição dos erros utilizando backwards feature selection

O que representa uma melhoria significativa, garantindo também um modelo simplista que melhor será capaz de evitar casos de *overfitting*. Podemos também observar os erros ao longo das várias iterações na figura 15.

Em suma, para o conjunto de dados do vinho tinto, o modelo que mostrou ter mais sucesso foi aquele que apenas considerava os seguintes atributos como significativos:

- `volatile.acidity`
- `total.sulfur.dioxide`
- `sulphates`
- `alcohol`

Vinho Branco De igual forma, partindo de um modelo que considera todos os atributos existentes, aplicamos a mesma metodologia de forma incremental. Inicialmente, com o modelo completo, obtemos o seguinte resultado do erro.

```
## [1] 0.5213203
```

Com esta metodologia, fomos capazes de obter na tabela 10 que resume os resultados obtidos. Denota-se que o atributo mais significativo mostrou-se ser a densidade, enquanto que nos vinhos vermelhos tende a ser a percentagem de álcool.

Como se pode observar pela tabela 10, a aplicação do método de *backwards feature selection* permitiu reduzir o erro do nosso modelo na percentagem de 4.1442217.

A melhoria é notável e conseguimos compreender estes dados da seguinte forma:

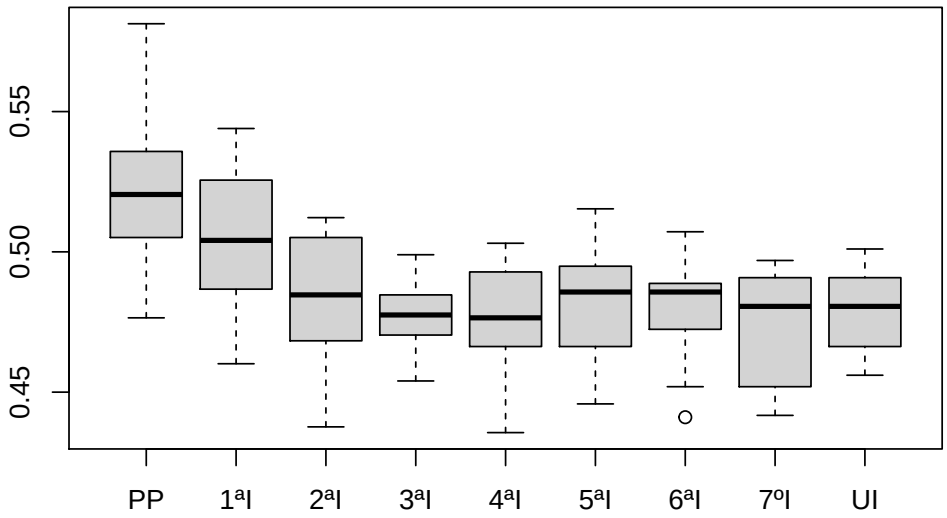


Figura 16: Distribuição dos erros utilizando backwards feature selection

Em suma, para o conjunto de dados do vinho branco, o modelo que mostrou ter mais sucesso foi aquele que apenas considerava os seguintes atributos como significativos:

- `volatile.acidity`
- `residual.sugar`
- `density`

Tabela 10: Resumo da aplicação de backwards feature selection

	Atributo Removido	Erro Obtido
Ponto de Partido	-	0.521320265349893
1ª Iteração	chlorides	0.504399221906329
2ª Iteração	alcohol	0.482712354730909
3ª Iteração	citric acid	0.478232081400569
4ª Iteração	free sulfur dioxide	0.47802134769814
5ª Iteração	total sulfur dioxide	0.480298269240361
6ª Iteração	sulphates	0.47927951518779
7ª Iteração	fixed acidity	0.474367798892713
Última Iteração	pH	0.479878048780488

Conclusões

Através do método de QDA, conseguimos chegar a modelos distintos que utilizam diferentes atributos para, dentro do seu data set, prever os resultados de forma ótima. Com este modelo, conseguimos modelar os comportamentos e gostos dos provadores de vinho e conseguimos denotar que:

- Em *vinhos tintos* verificou-se que os atributos mais significativos coincidem na totalidade com aqueles identificados pelo método de LDA, o que permitem colocar mais confiança na nossa inferência anterior.
- Em *vinhos brancos* a densidade é o atributo com mais impacto na previsão da qualidade, sendo o atributo de álcool ignorado. A acidez e açúcar residual, em conjunto com as densidade, permitem diretamente modelar a qualidade. Estes resultados diferem substancialmente dos resultados utilizando o LDA, especialmente sobre o atributo de densidade, este resultado pé indicativo de uma má separação de classes dentro do *data set* de vinhos brancos.

Conclusão

Para além do modelo de regressão logística, que separava o universo de classes em 2, é possível apresentar a seguinte comparação entre os melhores modelos obtidos para o problema de classificação com todas as classes, utilizando os modelos de: kNN, LDA e QDA.

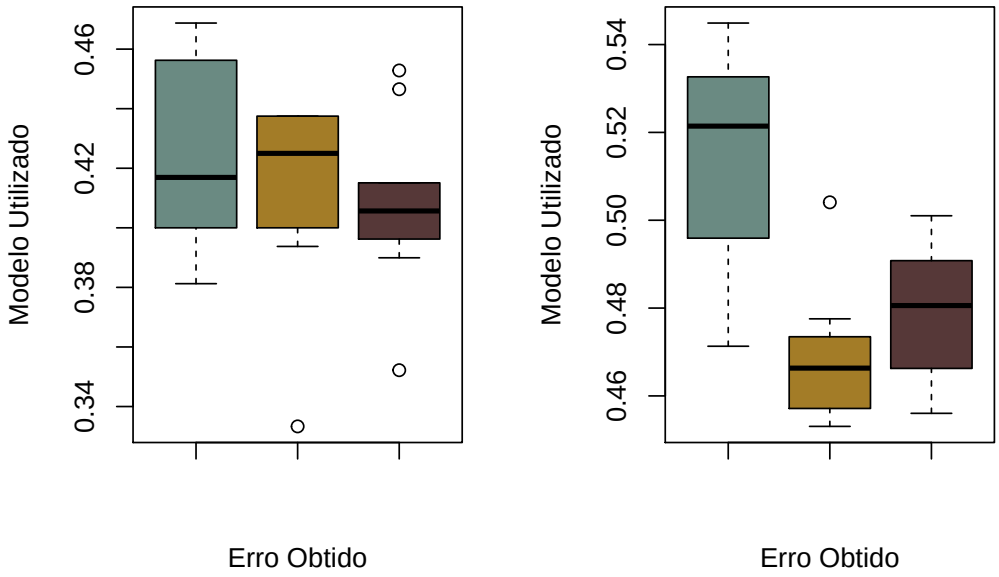


Figura 17: Distribuição dos erros nos vários modelos

Do gráfico acima, conseguimos perceber que os resultados são semelhantes em todos os modelos, pelo que não há nenhum que se distinga de forma significativa em termos de capacidade de previsão. No entanto, todos estes tornaram possível idealizar um conjunto de inferências pertinentes que não são contabilizadas na métricas de previsão.

Através da análise exploratória inicialmente conduzida, foi possível chegar a 3 perguntas basilares sobre o contexto no qual estamos inseridos. Estas questões visam transmitir conhecimento que futuramente pode ser utilizado para que uma empresa possa obter vantagem no mercado dos vinhos. A resposta a estas questões surge diretamente das inferências derivadas a partir dos vários modelos abrangidos. Por meio destas, conseguimos, com um elevado grau de certeza, inferir as características principais de um vinho na previsão da sua qualidade sensorial. As nossas conclusões relativas a essas questões principais são as descritas de seguida:

- **Quais atributos devem ser considerados para melhorar a qualidade de um vinho?**: Relativamente ao vinho branco existe associada uma maior incerteza sobre esta questão. No entanto, chegamos à conclusão de que a qualidade sensorial nos vinhos brancos tende a estar mais centrada à volta da relação entre açúcar residual e acidez volátil do vinho, o que indicia que a tipicidade do vinho, se é verde, maduro ou do douro, possui um fator determinante na classificação do mesmo. Porém, nos vinhos tintos observou-se que, apesar da acidez ser um atributo significativo, o açúcar residual, ao contrário do que sucedeu nos vinhos brancos, não o é, sendo este atributo substituído pelo total de dióxido de enxofre, o que indicia que o processo de fermentação do vinho, utilidade principal desta componente, tem uma maior importância para o consumidor final.
- **Como lidar com a má separação de classes inerente a avaliações sensoriais?**: Para lidarmos com este problema, e como apresentado anteriormente na secção de regressão logística, achamos adequado separar o universo de classes de qualidade em apenas 2. Como é impossível saber a classificação média de cada avaliador, é impossível calcular uma classificação centralizada para cada vinho. O que implica que as classificações mais representadas podem não corresponder à verdade, isto porque cada jurí possui uma tendência de classificação que não é modelada. Ao reduzir o conjunto em 2 somos capazes de reduzir o fator de variabilidade proveniente da tendência de cada júri, permitindo um modelo primitivo

da modelação da tendência das classificações. A aplicação desta metodologia provou-se útil no contexto da regressão logística e permitiu obter uma precisão notável. Razão pela qual achamos que, num contexto semelhante, mostra-se vantajoso uma modelação das classificações por meio de discretização.

- **Qual o impacto de atributos não significativos no mercado do vinho?**: Com a informação sobre os atributos mais irrelevantes na previsão da qualidade sensorial, os elementos deste mercado ficam equipados com conhecimento importante sobre conjuntos de pontos fortes em que se devem focar para aumentar a apreciação do mercado do vinho de interesse. No entanto, existem características que, apesar de não significativas para prever a qualidade, são de importância extrema na criação e manutenção da qualidade do vinho, mecanismos estes que não são contemplados por este estudo.

Em suma, achamos que conseguimos obter resultados favoráveis tendo em conta a informação fornecida. No entanto, não podemos deixar de mencionar que o erro da classificação sensorial com base em jurís, utilizada para obtenção do *data set*, possui um erro irredutível com um peso considerável. Tendo em conta que se trata da subjetividade humana, há fatores, para além daqueles inerestes ao vinho, que podem afetar a classificação de um jurí. Fatores como: a disposição, classificação média, experiência profissional de cada jurí, não são fornecidos, apesar de poderem ter um impacto significativo na análise dos vinhos.

Como tal, achamos que seria interessante, em trabalhos futuros, aplicar técnicas semelhantes em *data sets* com um erro irredutível menor, e tentar inferir, para além da subjetividade humana, os atributos de maior importância para prever a verdadeira qualidade de um vinho. Também seria muito interessante tentar uma modelação das classificações mais avançadas, que tivesse em conta, por exemplo, a classificação média de cada avaliador, bem como outras métricas pertinentes.

Referências

Afonso, João. 2017. «Acidez, Álcool e Tanino». https://grandesescolhas.com/acidez/.

BRI, Campden. sem data. «Wine analysis». https://www.campdenbri.co.uk/services/wine-analysis.php#sulphur-dioxide.

Charest, Rémy. 2019. «How Winemakers Analyze pH and Its Impact on Wine». https://daily.sevenfifty.com/how-winemakers-analyze-ph-and-its-impact-on-wine/.

Coli, Marina Sonegheti, Angelo Gil Pezzini Rangel, Elizangela Silva Souza, Margareth Ferraro Oliveira, e Ana Cristina Nascimento Chiaradia. 2015. «Chloride concentration in red wines: influence of terroir and grape type». *Food Science and Technology* 35 (Março): 95–99. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-20612015000100095&nrm=iso.

Cortez, P., A. Cerdeira, Fernando Almeida, T. Matos, e J. Reis. 2009. «Modeling wine preferences by data mining from physicochemical properties». *Decis. Support Syst.* 47: 547–53.

James, Gareth, Daniela Witten, Trevor Hastie, e Robert Tibshirani. 2013. *An introduction to statistical learning: with applications in R*. Spinger.

Moroney, Maureen. 2018. «Total Sulfur Dioxide - Why it Matters, Too!». https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too.

Wikipédia. 2020. «Ácido cítrico». https://pt.wikipedia.org/wiki/Ácido_cítrico.

Wooldridge, Jeffrey M. 2015. *Introductory econometrics: A modern approach*. Nelson Education.

Wu, Sylvia. 2020. «What is residual sugar in wine? – Ask Decanter». https://www.decanter.com/learn/residual-sugar-46007/.