Reconhecimento de Padrões em Vinhos

Guilherme Filipe Tomé Duarte- Nº 96571

Trabalho de Reconhecimento de Padrões

Docente: José Gonçalves Dias

Índice

Introdução

Existe evidência que o início da produção de vinho remonta o período Neolítico (8500 a.c. e 4000 a.c) [1]. Portugal é atualmente um dos maiores países exportadores de vinho a nível mundial [2]. Devido ao rápido crescimento da produção do vinho, novas técnicas e tecnologias tanto na sua produção como na sua venda têm vindo a surgir. Para que a sua produção cumpra um nível adequado de qualidade e segurança foram criadas certificados de qualidade para melhorar a produção do mesmo, bem como garantir que este não origine problemas de saúde. Estes certificados são atribuídos através de testes sensoriais, bem como testes físico-químicos [3]. Contudo estes testes sensoriais são feitos por humanos os quais são propensos a cometer erros, para não falar que a relação destes testes com os testes físico-químicos é complexa e não é completamente compreendida [4].

Este trabalho teve como objectivo encontrar padrões nestes testes a partir de duas bases de dados de vinhos portugueses (uma com vinhos brancos e outra com vinhos tinto) que foram agrupadas numa só, com intuito de facilitar os produtores na escolha das características do vinho e de garantir uma melhor qualidade na produção deste.

Dados

Relativamente a base de dados original esta era constituída por 11 variáveis ativas (acidez fixa (ácido tartárico em g/dm^3), acidez volátil (ácido acético em g/dm^3), ácido cítrico (em g/dm^3), açúcar residual (em g/dm^3), cloretos (cloreto de sódio em g/dm^3), dióxido de enxofre livre (em mg/dm^3), dióxido de enxofre total (em mg/dm^3), densidade (em g/cm^3), pH, sulfatos (sulfato de potássio em g/dm^3), álcool (%Vol)) que foram utilizadas na análise e duas passivas (qualidade (1-10) e tipo de vinho (branco ou tinto)) que foram utilizadas na caracterização. Na tabela em baixo (Tabela 1) é apresentado um sumário destas variáveis. Esta base de dados continha 6497 vinhos.

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	vinho
Min. : 3.800	Min. :0.0800	Min. :0.0000	Min. : 0.600	Min. :0.00900	Min. : 1.00	Length:6497
1st Qu.: 6.400	1st Qu.:0.2300	1st Qu.:0.2500	1st Qu.: 1.800	1st Qu.:0.03800	1st Qu.: 17.00	Class :character
Median : 7.000	Median :0.2900	Median :0.3100	Median : 3.000	Median :0.04700	Median : 29.00	Mode :character
Mean : 7.215	Mean :0.3397	Mean :0.3186	Mean : 5.443	Mean :0.05603	Mean : 30.53	
3rd Qu.: 7.700	3rd Qu.:0.4000	3rd Qu.:0.3900	3rd Qu.: 8.100	3rd Qu.:0.06500	3rd Qu.: 41.00	
Max. :15.900	Max. :1.5800	Max. :1.6600	Max. :65.800	Max. :0.61100	Max. :289.00	

total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
Min. : 6.0	Min. :0.9871	Min. :2.720	Min. :0.2200	Min. : 8.00	Min. :3.000
1st Qu.: 77.0	1st Qu.:0.9923	1st Qu.:3.110	1st Qu.:0.4300	1st Qu.: 9.50	1st Qu.:5.000
Median :118.0	Median :0.9949	Median :3.210	Median :0.5100	Median :10.30	Median :6.000
Mean :115.7	Mean :0.9947	Mean :3.219	Mean :0.5313	Mean :10.49	Mean :5.818
3rd Qu.:156.0	3rd Qu.:0.9970	3rd Qu.:3.320	3rd Qu.:0.6000	3rd Qu.:11.30	3rd Qu.:6.000
Max. :440.0	Max. :1.0390	Max. :4.010	Max. :2.0000	Max. :14.90	Max. :9.000

Tabela 1-Sumário das variáveis utilizadas

Identificação das dimensões da análise

Esta fase consistiu na redução da dimensionalidade da base de dados. Esta fase tem o propósito de tentar reduzir o nº de variáveis, através do método de análise de componentes principais, de forma a facilitar a interpretação dos dados recolhidos. Começou-se inicialmente por converter a matriz numa matriz de correlações e analisar a correlação entre as variáveis através da figura 1. A variável álcool não é aqui apresentada por esta ter sido removida devido ao teste de teste de Kaiser-Meyer-Olkin (KMO), sendo a razão pela sua remoção explicada mais adiante.

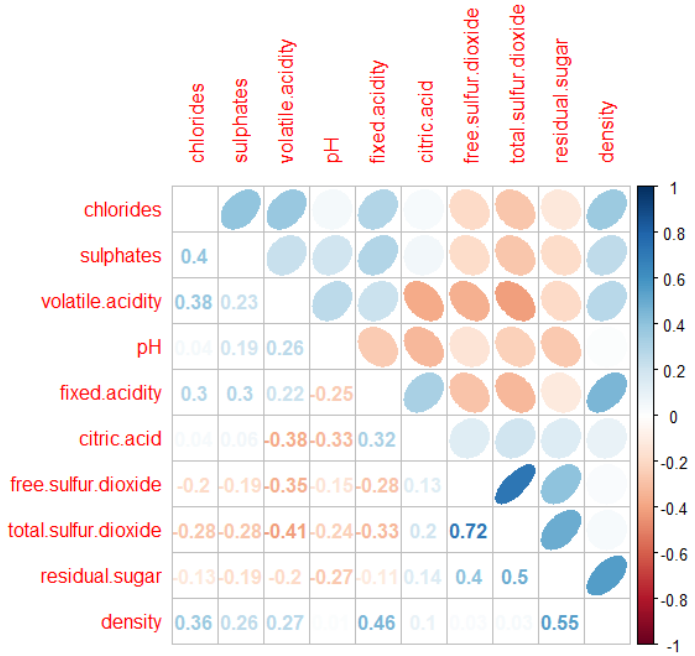

Figura 1- Gráfico de correlações entre as variáveis

Observando as correlações que em módulo têm uma correlação igual ou superior a 0,4, de forma a testar a validade destas correlações, podemos observar que o dióxido de enxofre livre e o dióxido de enxofre total tem uma elevadíssima correlação (0,72), devendo se ao facto de que o dióxido de enxofre total é a porção de dióxido de enxofre que está livre no vinho mais a porção que está ligada a outros produtos químicos no vinho, como aldeídos, pigmentos ou açúcares. O dióxido de enxofre é utilizado como conservante e como um estabilizador de pH, e por esse motivo existe uma correlação negativa com este, de forma a que outros componentes do vinho se formem em maior quantidade caso este não esteja presente e que contribuem para um pH baixo, tais como o ácido tartárico e o ácido acético, sendo por isso também que estes tenham também uma correlação negativa com o dióxido de enxofre [5].

Relativamente à densidade, esta tem uma relação positiva com o cloreto de sódio, sulfato de potássio, ácido tartárico, o ácido acético e com o açúcar residual devido ao facto de que estes têm uma massa molar maior que a água e por isso contribuem para uma densidade maior, ou seja, quanto maior for a massa molar de cada uma das componentes do vinho em relação à massa molar da água mais forte é a correlação destes com a densidade do vinho.

O cloreto de sódio e o sulfato de potássio têm uma correlação média fraca positiva (0,4). Esta correlação deve-se ao facto de que o cloreto de sódio é utilizado para melhorar a performance das leveduras no processo de fermentação, enquanto que o sulfato de potássio serve para impedir o crescimento bacteriano de bactérias que venham a danificar o vinho e reduzir o conteúdo de sal de algum produto alimentar e desta forma balancear o uso de cloreto de sódio [6].

Relativamente ao açúcar residual este tem uma correlação de 0,4 e 0,5 com o dióxido de enxofre livre e total, respetivamente. O açúcar residual é o açúcar natural das uvas que sobra após o processo de fermentação. Como o processo de fermentação consiste na conversão de açúcar em álcool pelas leveduras, o que leva a que quanto maior for a quantidade de dióxido de enxofre menor é quantidade de leveduras para que ocorra a fermentação e por isso maior é quantidade de açúcar que sobra após o processo de fermentação [7].

Apesar de as variáveis não apresentarem uma elevada correlação entre elas, é possível que se possa agrupar as variáveis de forma a reduzir a sua quantidade, visto que existe algumas correlações superiores a 0,3 (em módulo), e por isso foi decidido prosseguir com o método.

O passo seguinte foi fazer o teste de Bartlett. O teste de Bartlett é usado para testar a hipótese nula de que as variáveis não esta correlacionadas na população. Por outras palavras este teste verifica se existe pelo menos duas variáveis que estão correlacionadas assumindo um nível significância de 0,05. Se o valor de p-value for inferior a este valor, é pouco provável que a matriz de correlação populacional é uma matriz identidade (rejeitamos assim a hipótese nula). A partir da nossa base de dados foi obtido um p-value de 3.49e-57, e consequentemente, podemos rejeitar a hipótese nula e afirmar que existe alguma correlação entre as nossas variáveis.

O teste de KMO foi feito posteriormente, sendo que este teste indica a proporção da variância das variáveis iniciais que pode ser causada por fatores latentes/comuns. Por outras palavras este teste serve para averiguar se amostra que temos é adequada fazer uma principal component analysis (PCA), sendo um valor adequado, um valor superior a 0,5.

Inicialmente este teste foi feito inicialmente com as 11 variáveis iniciais, porém o valor de KMO não era adequado para a realização da PCA (ver anexo 1 pág.16). Visto que a variável álcool era a que menos contribuía para adequação do método, esta foi removida e feito novamente o teste verificando-se um valor de KMO de 0,53 (Tabela 2).

	Kaiser-Meyer-Olkin (KMO)	Overall MSA = 0.53			
MSA for each item	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides
	0.4	0.77	0.56	0.41	0.59
	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates
	0.73	0.73	0.34	0.36	0.82

Tabela 2- Teste de KMO

Após a realização de todos estes testes iniciou-se então a análise de componentes principais, tendo sido necessário estandardizar as variáveis (transformar as variáveis para que estas tenham uma média de 0 e uma variância de 1). Para a

realização deste foi utilizado 3 critérios: Critério de Kaiser, Scree Plot e o mínimo de variância explicada. Para avaliar estas métricas foram criadas 10 componentes (uma para cada variável).

O Critério de Kaiser retém componentes cujo eigenvalues são superiores à média dos eigenvalues, isto é, retém componentes cuja os eigenvalues são maiores que 1. Na tabela a baixo é apresentado os eigenvalues obtidos para a base de dados utilizada (Tabela 3).

	eigenvalues
1	3.015
2	2.100
3	1.432
4	0.962
5	0.710
6	0.546
7	0.521
8	0.368
9	0.250
10	0.097

Tabela 3- Eigenvalues obtidos pelo critério de Kaiser. Assinalado a laranja estão as componentes que cumprem este critério.

A partir da observação da tabela, este método sugere (assinalado a laranja) que se crie apenas 3 componentes.

O método de Scree Plot consiste num gráfico que representa os eigenvalues pelo número de componentes por ordem de extração. Este método consiste na procura do "cotovelo" que normalmente está associado a eigenvalues menores que 1. A figura abaixo representa este método (Figura 2).

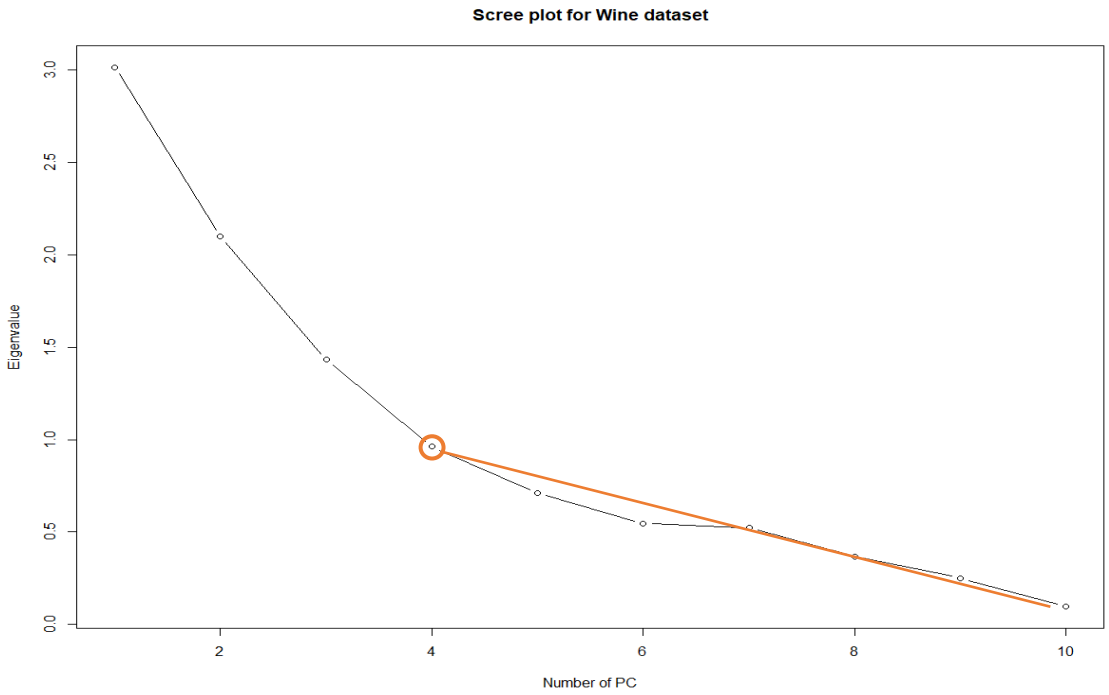

Figura 2- Gráfico de Scree Plot. Assinalado a laranja estão os números de componentes que o método "sugere" que sejam utilizadas (4 componentes).

A partir da observação do gráfico de Scree Plot, este parece sugerir 4 componentes apesar do "cotovelo" não ser significativo.

Relativamente ao 3º e último método, o método de variância mínima explicada. Este método consistiu em garantir que as componentes extraídas consigam explicar 75% da variância explicada. A tabela apresentada abaixo mostra os pesos e variância explicada para cada uma das componentes, bem como a variância explicada acumulada (Tabela 4).

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
SS Loadings	3.015	2.100	1.432	0.962	0.710	0.546	0.521	0.368	0.250	0.097
Proportation Var.	0.301	0.210	0.143	0.096	0.071	0.055	0.052	0.037	0.025	0.010
Cumulative Var.	0.301	0.511	0.655	0.751	0.822	0.876	0.928	0.965	0.990	1.000

Tabela 4- Resultados obtidos dos pesos e variância explicada para cada uma das componentes, bem como a variância explicada acumulada para 10 componentes. Assinalado estão o número de componentes necessárias para que haja uma variância mínima explicada acumulada de 75%.

Tendo em conta os resultados obtidos pode-se verificar que o número de componentes sugeridas são de 3 e 4 componentes. Deste modo foi feito uma tentativa de explicar a base de dados com 3 componentes e com 4 componentes. Sendo que apenas com 4 componentes foi possível definir as componentes, apenas estes são aqui apresentados. Os resultados obtidos para 3 componentes encontram-se em anexo (anexo 2 pág.16). Foi feito também uma rotação Varimax para ser possível interpretar melhor os resultados obtidos sendo que esta rotação o que faz é tentar maximizar a variância dos loadings de uma variável apenas numa componente.

Na tabela abaixo são apresentados os pesos que permitam explicar cada uma das variáveis (Tabela 5).

	RC1	RC2	RC3	RC4
fixed.acidity	-0.50	0.34	0.57	0.31
volatile.acidity	-0.56	0.38	-0.45	0.21
citric.acid	0.23		0.81	0.22
residual.sugar	0.51	0.71	0.10	-0.24
chlorides	-0.26	0.28		0.64
free.sulfur.dioxide	0.85	0.10		
total.sulfur.dioxide	0.85	0.14	0.11	-0.21
density		0.88		0.33
pH		-0.19	-0.72	0.38
sulphates	-0.11			0.86

Tabela 5- Variáveis e os loadings que estas têm em cada uma das componentes.

A partir dos resultados obtidos podemos ver que a 1º componente apresenta loadings elevados positivos tanto de dióxido de enxofre livre e total, bem como de açúcar residual, o que confirma o que foi explicado anteriormente nas correlações, e loadings negativos elevados para acidez fixa e volátil devido à presença do açúcar residual. Devido a estes resultados a esta componente foi atribuída o nome de “Nível de Doçura”. A 2º componente apresenta loadings elevados para a densidade, mas tendo também loadings positivos para as outras componentes, principalmente para o açúcar residual. A esta variável ficou atribuída o nome de “Densidade” pois os loadings são proporcionais à massa molar de cada um dos componentes e que contribuem para o aumento de densidade. A 3º componente mostra um loading elevado para o ácido

cítrico e para a acidez fixa, bem como um loading negativo para o pH (visto que quanto mais ácido for o vinho, menor é o pH), ficando atribuído a esta componente o nome de "Nível de Acidez". A 4º e última componente apresenta loadings elevados para os sulfatos e para os clorídricos visto que estes são usados em conjunto para controlar o crescimento microbiano e o nível de sal ficando a esta componente atribuída o nome de "nível de proteção microbiano".

Identificação da heterogeneidade na base de dados

Esta fase teve como objectivo tentar encontrar padrões nos vinhos ao tentar agrupar os vinhos em grupos de acordo com as semelhanças e as diferenças que estes têm entre si (através das componentes principais e das variáveis de profile). Para tal foi usado 3 métodos: Hierarchical Clustering, K-Means Clustering e PAM clustering.

Começou-se inicialmente com o Hierarchical Clustering. Este foi utilizado para dar uma ideia do número de clusters (grupos) que serão formados para que depois os outros dois métodos sejam mais fáceis de se aplicar, sendo estes dois usados para comparar um com os outros. Esta técnica usa uma matriz de distância para ver as semelhanças entre indivíduos e produz uma hierarquia de partições. O método que foi utilizado foi o método de Ward. Este método utiliza a distância euclidiana ao quadrado para calcular estas semelhanças. Sendo este método um método aglomerativo, este método começa por ter todas as observações (vinhos) individualizados (clusters individuais). A cada iteração dois clusters se juntam (de acordo com a sua aproximação em termos de distância), até que exista apenas um cluster no final. Para analisar estes resultados, recorreu-se a um dendograma (Figura 3). Tendo em conta os resultados considerou-se um agrupamento de 3 clusters.

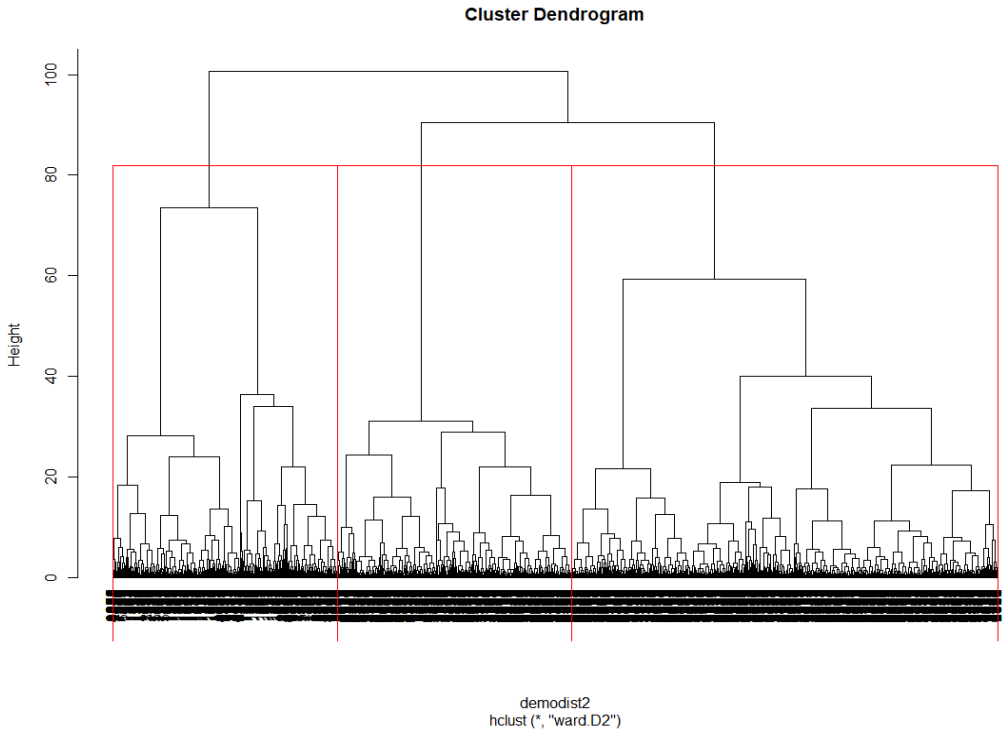

Figura 3- Dendograma com 3 clusters (assinalados a vermelho)

Foi ainda feito a representação gráfica da silhueta e determinado o seu valor (silhouette score), sendo que esta representação nos indica o quão semelhante as observações são ao cluster onde ficaram atribuídas em comparação com os outros clusters sendo que o seu valor pode variar entre -1 e 1, em que um valor de -1 indica que as observações estão atribuídas incorretamente a um cluster e um valor de 1 indica que as observações estão perfeitamente atribuídas a um cluster e os clusters são facilmente distinguíveis. No caso de ter um valor de próximo de 0 indica que os clusters estão sobrepostos. Foi obtido um score de 0,25 o que indica que a estrutura é fraca e os clusters estão próximos uns dos outros. A silhueta encontra-se representada na Figura 4.

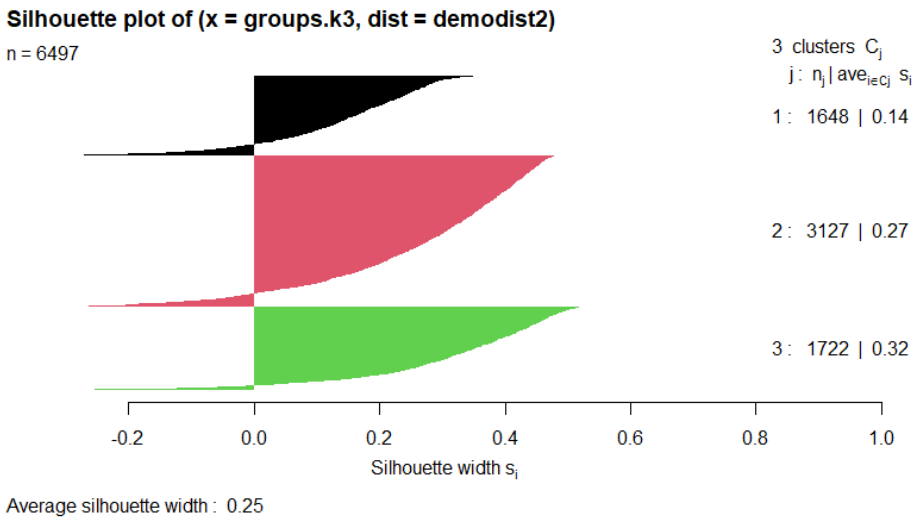

Figura 4- Silhueta do Hierarchical Clustering

Sabendo agora o número de clusters que devem ser utilizados, foi utilizado o método de K-Means para observar os agrupamentos que este gera. O método de K-Means é um método de partição, ou seja, este começa por considerar todas as observações (vinhos) como um grupo só e a cada iteração estas observações são agrupadas em grupos (clusters) tendo em conta a distância euclidiana e a posição dos centroides, até minimizar a variância dentro dos clusters (soma dos quadrados das distâncias euclidianas). Tendo em conta que o hierarchical clustering sugeria 3 clusters, a técnica de k-means foi feita com 3 clusters também. O primeiro cluster continha 1593 vinhos (24,51%), o segundo cluster continha 3016 vinhos (46,4%) e o terceiro cluster continha 1888 vinhos (29,06%). Em baixo é apresentado tabelas que correlacionam os clusters obtidos com as variáveis profile, tendo em conta a quantidade de vinhos (Tabela 6 e Tabela 7).

	1	2	3
branco	53	2968	1877
tinto	1540	48	11

Tabela 6- tipo de vinho em cada cluster (nº de vinhos)

	1	2	3
3	10	8	12
4	69	103	44
5	690	656	792
6	621	1382	833
7	187	728	164
8	16	135	42
9	0	4	1

Tabela 7- qualidade do vinho em cada cluster (nº de vinhos)

A partir da Tabela 6 é possível aferir que o cluster 1 é constituído principalmente por vinhos do tipo tinto e o cluster 2 e 3 por vinhos brancos e o cluster 2 ter aproximadamente o dobro das observações. Da Tabela 7 é possível aferir que os vinhos no cluster 2 têm uma qualidade em média (média=6,04) superior ao cluster 3 (média=5,65), apesar de a diferença não seja muito significativa.

Tal como para o Hierarchical Clustering, no K-Means também foi representada a silhueta obtendo-se um score de 0,27 o que indica que a estrutura é fraca e os clusters estão próximos uns dos outros (Figura 5).

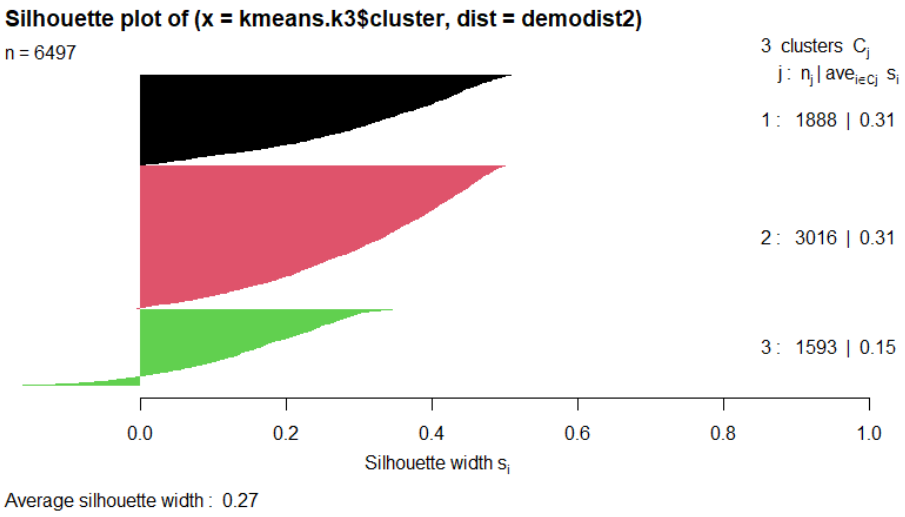

Figura 5 - Silhueta do K-means

Foi feito ainda um gráfico que relaciona a soma dos quadrados das distâncias dentro dos clusters e o número de clusters (Figura 6) com intuito de ver se este método sugeria um número diferente de clusters. Este gráfico foi feito com o propósito de encontrar o "cotovelo" (quando a soma dos quadrados das distâncias dentro dos clusters varia de forma reduzida com o incremento do nº de clusters). Este gráfico "sugere" utilizar 5 clusters, porém não se encontrou melhores padrões comparativamente com a utilização de 3 clusters e como tal apenas os resultados para 3 clusters são apresentados.

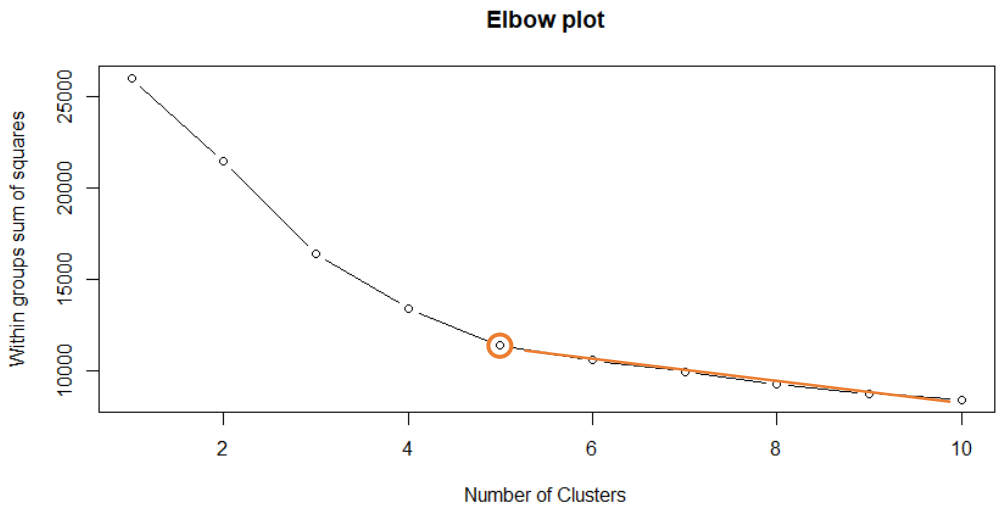

Figura 6- Gráfico que relaciona a soma dos quadrados das distâncias dentro dos clusters e o número de clusters. Assinalado a laranja estão os números de componentes que o método "sugere" que sejam utilizadas (4 componentes).

Outro gráfico foi feito com o intuito de tentar descobrir mais padrões nos dados (Figura 7).

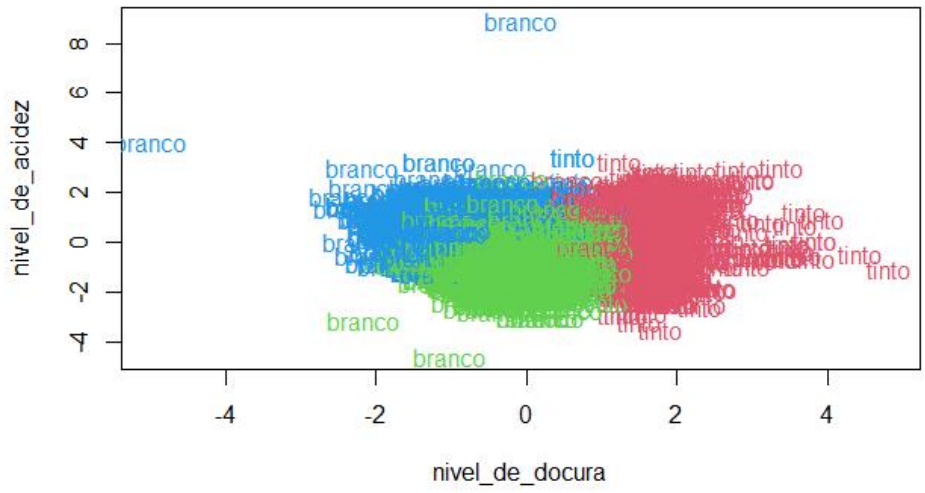

Figura 7-Relação entre o nível de acidez e o nível de doçura com o método de K-Means (cluster 1-vermelho, cluster 2-verde, cluster 3-azul)

A partir da figura acima é possível verificar que os vinhos do tipo tinto são relativamente mais doces do que os vinhos brancos e também que os vinhos no cluster 2 tendem a ser ligeiramente mais doces do que os vinhos no cluster 3.

Para atribuir os nomes aos clusters foi observado os valores médios de cada uma das componentes em relação aos clusters (Tabela 8) bem como os resultados obtidos anteriormente.

	nível_de_docura	densidade	nível_de_acidez	nível_de_protecção_microbiano
1	1.50	0.31	0.26	-0.03
2	-0.25	-0.64	-0.55	0.10
3	-0.87	0.76	0.65	-0.14

Tabela 8- valores médios de cada uma das componentes em relação aos clusters

A partir dos resultados obtidos, ao cluster 1 foi atribuído o nome de "vinhos doces do tipo tinto". Ao cluster 2 foi dado o nome de "vinhos brancos de baixa densidade e de qualidade média-alta" e ao cluster 3 foi dado o nome de "vinhos azedos do tipo branco e de qualidade média".

Foi feito por fim o clustering com o método PAM. O método PAM é um método parecido ao K-Means, porém mais robusto e que em vez de usar centroides, usa medoides.

Os resultados obtidos para este método não diferem em muito os resultados obtidos pelo método de K-Means e, consequentemente, as conclusões que se conseguiam alcançar e os nomes que seriam atribuídos aos clusters também não. Em baixo são apresentados os resultados obtidos (Tabela 9 e 10 e Figura 8).

	1	2	3
branco	107	2772	2019
tinto	1545	47	7

Tabela 9- tipo de vinho em cada cluster (nº de vinhos)

	1	2	3
3	10	8	12
4	83	88	45
5	719	606	813
6	633	1299	904
7	189	684	206
8	18	130	45
9	0	4	1

Tabela 10- qualidade do vinho em cada cluster (nº de vinhos)

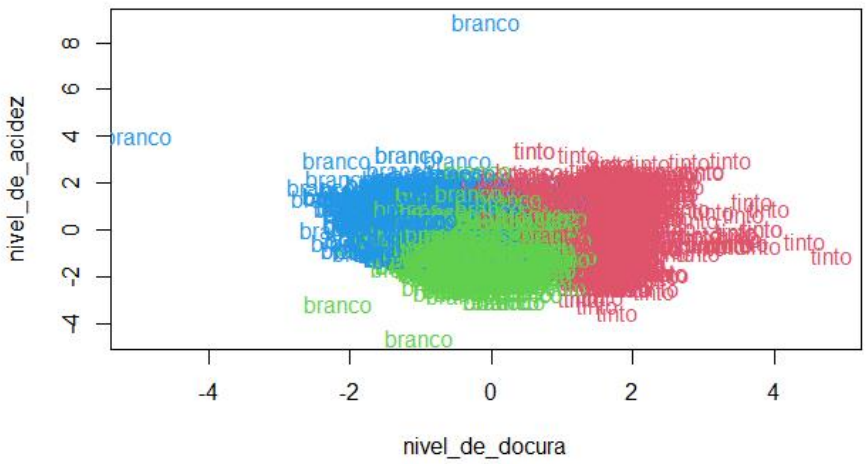

Figura 8-Relação entre o nível de acidez e o nível de doçura com o método de PAM. (cluster 1-vermelho, cluster 2-verde, cluster 3-azul)

Conclusão

A partir dos resultados obtidos foi possível chegar a algumas conclusões. Os vinhos do tipo tinto têm tendência a serem mais doces do que os vinhos brancos o que o sugere que o processo de fermentação dos vinhos do tipo tinto deve ser mais rápido comparativamente com os vinhos brancos de forma a que leveduras não consumam todo o açúcar ficando este presente no vinho no processo. Esta informação pode ser utilizada pelos produtores de vinhos para que não demorem tanto o tempo de fermentação do vinho do tipo tinto, visto não ser preciso, sendo que desta forma possam economizar dinheiro. A partir dos valores médios de cada uma das componentes em relação aos clusters e das figuras X e Y foi possível concluir que os vinhos brancos mais doces e por consequente menos azedos (cluster 2) aparentam ter em média uma qualidade ligeiramente acima da média comparativamente com os vinhos brancos azedos (cluster 3). Sendo esta informação tido em conta no processo de produção do vinho os produtores poderão obter um maior lucro neste mercado.

Referências Bibliográficas

[1] Hua Li, Hua Wang, Huanmei Li, Steve Goodman, Paul van der Lee, Zhimin Xu, Alessio Fortunato, Ping Yang,The worlds of wine: Old, new and ancient,Wine Economics and Policy,Volume 7, Issue 2,2018,Pages 178-182,ISSN2212-9774, https://doi.org/10.1016/j.wep.2018.10.002.

[2] FAO,FAOSTAT—Food and Agriculture Organization Agriculture Trade Domain Statistics,July2008,http://faostat.fao.org/site/535/DesktopDefault.aspx?PageID=535.

[3] S. Ebeler, Flavor Chemistry — Thirty Years of Progress, Kluwer Academic Publishers, 1999, pp. 409–422, chapter Linking flavour chemistry to sensory analysis of wine.

[4] A. Legin, A. Rudnitskaya, L. Luvova, Y. Vlasov, C. Natale, A. D'Amico, Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception, Analytica Chimica Acta 484 (1) (2003) 33–34.

[5] https://en.wikipedia.org/wiki/Sulfur_dioxide

[6] https://www.dovepress.com/influence-of-sodium-chloride-on-wine-yeast-fermentation-performance-peer-reviewed-article-IJWR

[7] https://winefolly.com/deep-dive/what-is-residual-sugar-in-wine/

Anexos

	Kaiser-Meyer-Olkin (KMO)	Overall MSA = 0.41				
MSA for each item	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	alcohol
	0.28	0.61	0.62	0.29	0.73	0.73
	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	
	0.75	0.72	0.30	0.21	0.56	

Anexo1-Teste de KMO para as 11 variáveis ativas

	RC1	RC2	RC3
fixed.acidity	-0.25	0.65	0.51
volatile.acidity	-0.23	0.57	-0.51
citric.acid			0.82
residual.sugar	0.86	0.14	0.11
chlorides	-0.14	0.69	
free.sulfur.dioxide	0.74	-0.29	
total.sulfur.dioxide	0.78	-0.36	0.16
density	0.48	0.80	
pH	-0.16		-0.72
sulphates	-0.22	0.58	

Anexo 2- Variáveis e os pesos que estas têm em cada uma das componentes (para 3 componentes).