Análise de qualidade de vinho português por estatística descritiva e visualização de dados

1st Catherine Bezerra Markert
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
catherinemarkert@alu.ufc.br

2nd Davi Queiroz Albuquerque
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
daviqueiroz2002@alu.ufc.br

3rd Marcos Augusto Pereira Lima
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
marcospl@alu.ufc.br

4th Yago Costa de Oliveira
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
yagocosta@alu.ufc.br

***Abstract*—O uso de dados emerge como uma prática cada vez mais popular na sociedade atual, dado que a analisá-los auxilia a realizar as mais diversas escolhas no cotidiano da humanidade. Percebendo esse importante contexto, esse artigo tem a finalidade de apresentar como a análise de dados de um certo conjunto pode ajudar a conhecê-lo, apontando características como assimetria de preditores e presença de outliers. Para isso, foram feitas as análises incondicional monovariada, classe-condicional monovariada, incondicional bi-variada e incondicional multivariada para conhecer as principais características desse conjunto de amostras. Com efeito, os resultados obtidos mostram correlações entre preditores e destacam como as suas classes possuem características diferentes e como as features influenciam na qualidade de uma amostra de vinho.**

***Index Terms*—visualização de dados, estatística, *principal component analysis*, analise multivariada, vinho**

I. Introdução

Dados são extremamente importantes para a tomada de decisões na atualidade, visto que auxiliam a deduzir resultados de relevante significado e a tomar decisões conscientes e informadas. Contudo, dados brutos, ou seja, sem tratamento, não conseguem oferecer resultados totalmente confiáveis e significativos, necessitando passar por um processamento para apresentarem tais características. Ou seja, para conseguir analisar os dados coletados de forma a obter resultados assertivos, é necessário tratá-los para que se tenha o melhor entendimento possível do que eles representam. Segundo [2], "Aprendizagem Estatística se refere a um vasto conjunto de ferramentas para entender dados."

Nesse sentido, o pré-processamento de dados, em conjunto com os conceitos da Aprendizagem Estatística, se firma como uma importante ferramenta para tornar dados não-tratados em elementos analisáveis e compreensíveis. De acordo com [3], as técnicas de processamento de dados geralmente fazem referências à adição, à deleção ou transformação do conjunto de dados separado para treino de um certo modelo, ou seja, realizam-se modificações em um dataset que será utilizado para desenvolver um certo modelo de aprendizado de máquina. Ademais, transformações dos dados reduzem o impacto da assimetria, a qual, segundo [3], "Uma distribuição assimétrica é uma que é raramente simétrica. Isso significa que a probabilidade de uma amostra estar em qualquer um dos lados da média da distribuição é praticamente igual. Uma distribuição assimétrica para a direita tem um número maior de pontos no lado esquerdo da distribuição do que no direito. Consideram-se assimétricos dados cuja razão entre maior e menor valor é maior que 20.", e dos outliers de algumas features, que são amostras cujo valor ou característica são discrepantes da maioria dos dados coletados, promovendo melhorias significativas no desempenho de modelos de regressão que utilizam esses dados, por exemplo.

Considerando o contexto de coleta automatizada de dados, no qual o número de preditores pode crescer de maneira extremamente rápida, muitos deles não são tão úteis ou são redundantes, como abordado por [4]. Dessa maneira, se faz necessário outro tipo de transformação nos preditores, a de redução de dimensionalidade, cujo principal objetivo é mapear o conjunto de dados original para um conjunto menor, preservando o máximo possível de informação com o intuito de melhorar o desempenho do modelo preditivo, seja na taxa de acerto ou no tempo de execução. Uma das técnicas mais utilizadas para esse tipo de transformação é a de análise dos componentes principais (*Principal Component Analysis*, PCA), cujo propósito é diminuir a dimensionalidade dos dados, enquanto se preserva sua variação.

Para exemplificar a importância da análise de dados, investiga-se, um conjunto de dados de amostras de vinho português [1], obtendo-se informações relevantes sobre o comportamento dos preditores desse conjunto de dados mediante análises incondicional monovariada, classe-condicional monovariada, incondicional bi-variada e incondicional multivariada, com o intuito de entender o comportamento dos dados colocados para futuramente prepará-los para a construção de

modelos de regressão.

II. Métodos

A. Descrição do Dataset

O Dataset "Wine Quality" [1] é composto por 4898 amostras da variante branca e por 1599 amostras da variante vermelha do Vinho Verde português, coletadas do norte de Portugal. Nesse sentido, dois subdatasets foram criados para representar as 2 classes do Dataset, um para as amostras da variante branca do vinho e outro para a variação vermelha da bebida, cujas entradas incluem testes realizados de forma objetiva e suas saídas têm como base dados sensoriais (mediana de pelo menos três avaliações realizadas por especialistas de vinhos). Cada conhecedor de vinhos avaliou a qualidade de cada amostra entre 0 e 10, sendo o 0 classificado como muito ruim e 10 como muito excelente.

Ademais, as doze features do dataset são:

1) Acidez Fixa
2) Acidez Volátil
3) Ácido Cítrico
4) Açúcar Residual
5) Cloretos
6) Dióxido Sulfúrico Livre
7) Dióxido Sulfúrico Total
8) Densidade
9) pH
10) Sulfatos
11) Álcool
12) Qualidade (nota entre 0 e 10)

Os onze primeiros preditores são baseados em testes psicoquímicos e o último é baseado em dados sensoriais.

B. Análise incondicional monovariada

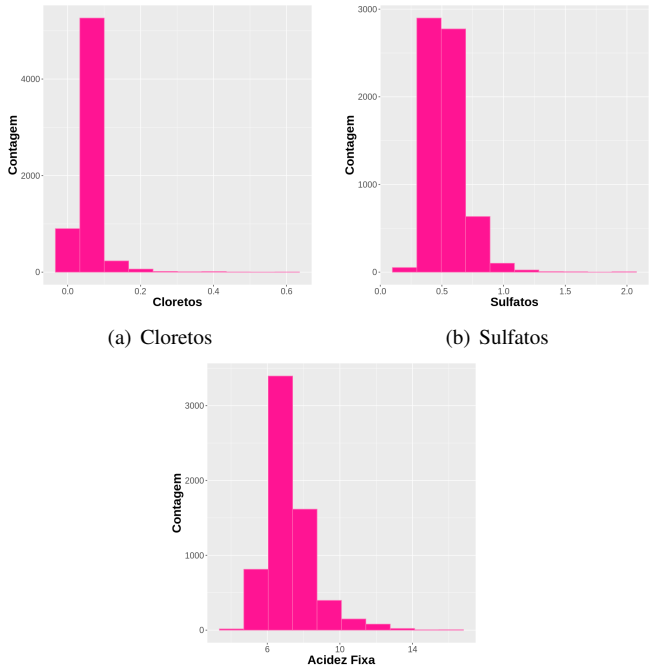
Considerada uma forma de análise de dados simples, a análise incondicional monovariada é uma observação de cada preditor de um certo dataset, sem se realizar uma divisão das classes dele, visando investigar como cada um dos preditores do conjunto de dados se distribui.

Para realizar tal análise, plota-se um histograma para cada um dos preditores, além de se calcular a média incondicional, o desvio padrão e a assimetria de cada uma das features. Desse modo, podemos observar o comportamento de cada preditor em relação a cada um desses cálculos.

Considerando o conjunto de dados utilizado nesse artigo, realizou-se o cálculo das métricas citadas acima e a plotagem de histograma para as 12 features contidas nesse dataset. Os gráficos mais relevantes estão plotados abaixo, precedidos por uma breve análise.

A maioria dos gráficos possui uma assimetria para a direita, destacando-se os preditores Cloretos, Sulfatos e Acidez fixa como os que possuem maior assimetria, a qual pode prejudicar a construção de modelos de predição para tais dados por uma grande adição de viés e deve ser tratada antes de se construir um modelo com o conjunto de dados. A figura 1 apresenta a distribuição de tais preditores.

Fig. 1. Observando distribuições assimétricas da análise incondicional monovariada

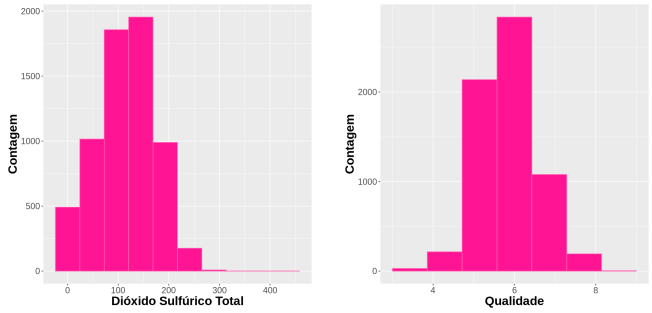


(a) Cloretos (b) Sulfatos

(c) Dióxido sulfúrico total

Outrossim, as distribuições mais simétricas são as de Dióxido Sulfúrico Total e Qualidade, possuindo os valores de assimetria calculados mais baixos.

Fig. 2. Observando distribuições simétricas da análise incondicional monovariada



(a) Dióxido Sulfúrico Total (b) Qualidade

C. Análise classe-condicional monovariada

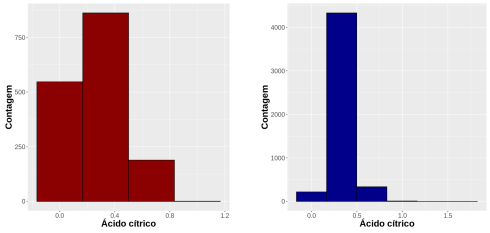
A análise classe-condicional monovariada consiste em aplicar os procedimentos da análise anterior, ou seja, plotar os histogramas, calcular as médias, o desvio padrão e a assimetria de cada uma das features, mas dessa vez utilizando somente as amostras de uma única classe por vez.

Esse estudo é de suma importância pois permite avaliar as tendências dos preditores tanto considerando só as amostras de vinho tinto, quanto somente as de vinho branco, procurando de forma inicial entender as particularidades de cada classe.

Ao plotar os histogramas, percebe-se já algumas características para cada classe em particular, pois enquanto as

unidades de vinho tinto apresentam uma distribuição pouco mais balanceada para o ácido cítrico, para os vinhos brancos há uma concentração perceptível no intervalo $0.25 \vdash 0.5$, podendo visualizar isso nas Figuras 3 a seguir.

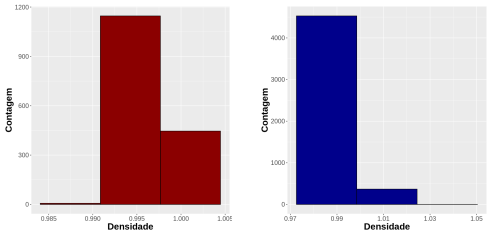
Fig. 3. Histogramas do preditor "ácido cítrico" quando condicionado às classes.



(a) Vinhos tintos. (b) Vinhos brancos.

Quanto à densidade, temos que, pelas Figuras 4, ambos os tipos de vinho apresentam valores próximos a 0.99, porém as amostras de vinho tinto estão mais concentradas neste valor central, enquanto as de vinho branco destoam com uma escala maior em relação a outra classe.

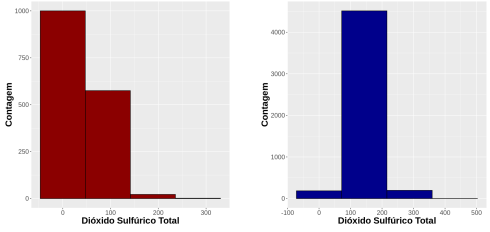
Fig. 4. Histogramas do preditor "densidade" quando condicionado às classes.



(a) Vinhos tintos. (b) Vinhos brancos.

Para a feature "dióxido sulfúrico total", como visto nas Figuras 5, os valores com maior frequência para cada classe são, relativamente, bem distintos, já que para o tinto a classe modal é $0 \vdash 50$, enquanto para o branco é $100 \vdash 200$. Nas Tabelas II e III, também é possível perceber que para o dióxido sulfúrico livre, as amostras de vinho branco têm em média maior quantidade desse componente do que as de vinho tinto.

Fig. 5. Histogramas do preditor "dióxido sulfúrico total" quando condicionado às classes.



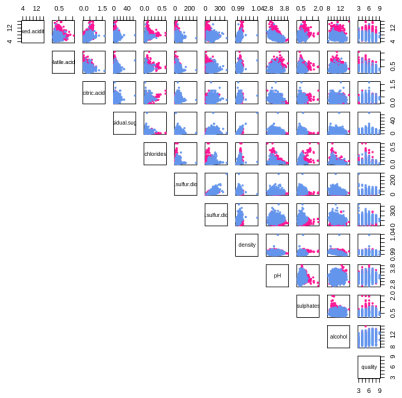
(a) Vinhos tintos. (b) Vinhos brancos.

D. Análise incondicional bi-variada

A análise incondicional bi-variada é a observação do comportamento de dois preditores na presença mútua de ambos, sem que haja uma separação das classes, objetivando, assim, perceber a influência de uma feature na outra.

Por meio disso, realizamos a criação do gráfico de todos pontos da relação de pares de preditores, representado na figura 6. No qual, definimos a cor rosa para representar a classe do vinho vermelho e o azul para o vinho branco.

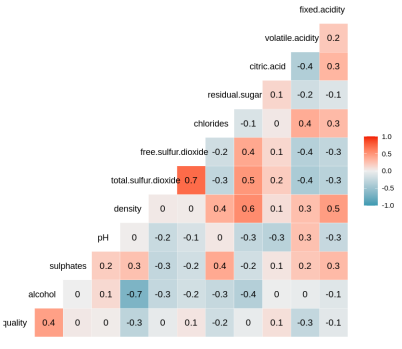
Fig. 6. Gráfico de dispersão de todos os preditores



Sendo verificado, a partir dessa interação, a correlação entre as variáveis, sendo estabelecido a possibilidade de existir uma relação positiva ou negativa, sendo perfeita na correlação em 1 e -1, ou seja, um preditor influência totalmente no outro, sendo inversamente ou não. Ademais, pode haver a ausência de relação, caso a correlação seja 0, indicando que uma feature não influência na outra.

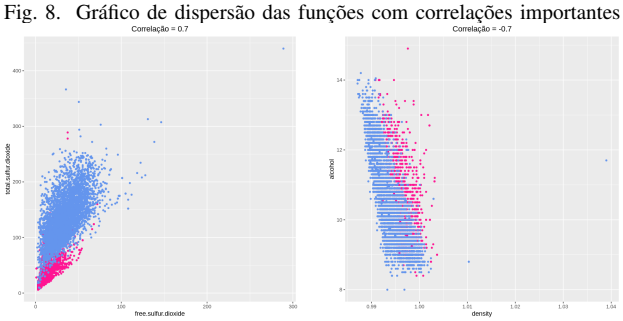
Em conhecimento disso, realizamos a plotagem de todas as correlações, conseguindo então visualizar algumas relações entre os pares de preditores.

Fig. 7. Correlação Linear de todos os preditores



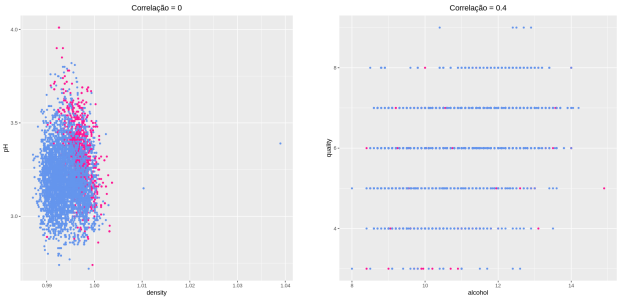
Analisando os valores da correlação obtida, ampliamos o gráfico de algumas relações mais interessantes, que demonstram o comportamento dito anteriormente. Separamos 4 exemplos, em que temos os casos que mais se aproximam dos extremos na correlação e alguns valores mediano.

Fig. 8. Gráfico de dispersão das funções com correlações importantes



(a) Dióxido Sulfúrico Total e Dióxido Sulfúrico Livre

(b) Álcool e Densidade

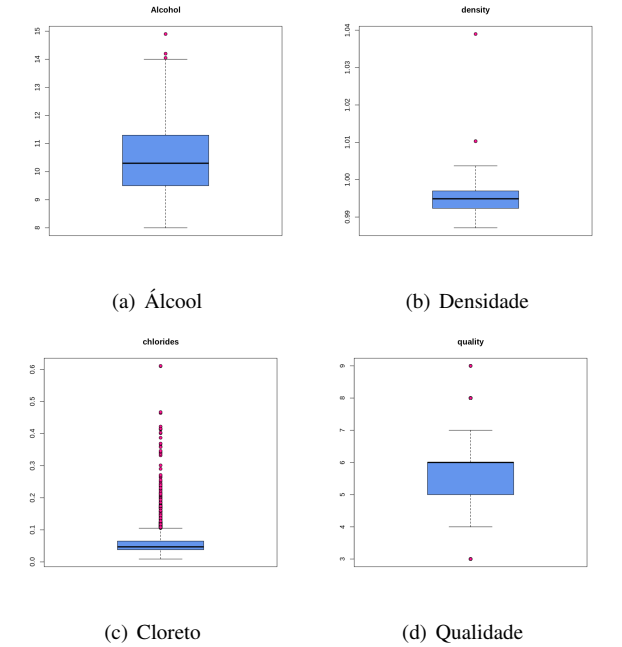


(c) pH e Densidade

(d) Álcool e Qualidade

Todavia, para avaliar corretamente os dados, é preciso verificar a quantidade de outliers presente nos preditores, pois essas exceções podem alterar os valores obtidos na análise, mas não alteram o resultado na prática. Assim, realizamos a plotagem de alguns outliers interessantes.

Fig. 9. Box plot de preditores com outliers importantes



(a) Álcool

(b) Densidade

(c) Cloreto

(d) Qualidade

Logo, com a produção dessas figuras, conseguimos visualizar os dados de maneira diferente, podendo perceber relações entre preditores e possíveis influências entre as variáveis, sendo essa análise aprofundada em resultados.

E. Análise incondicional multivariada

Conforme o número de preditores cresce, o número de análises que temos que fazer com cada par de preditor cresce na ordem quadrática, nessa perspectiva, se faz necessário utilizar técnicas que nos possibilitem analisar diversas variáveis de maneira simultânea, tais métodos são classificados na categoria de análise multivariada, elas se aproveitam do fato de que dados com grandes dimensionalidades muitas vezes têm dimensões redundantes e que podem ser explicadas por combinações de outras [6] .

Neste trabalho, utilizamos a Análise de Componentes principais ela consiste em encontrar os componentes principais que são combinações lineares dos preditores. Considerando o conjunto de dados $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}, \mathbf{x}_i \in \mathbb{R}^D$, com média igual a 0, nosso conjunto tem N observações e D preditores $\{p_1, p_2, ..., p_D\}$, a matriz de covariância, S, é definida por:

$$S = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T$$

Onde $S \in \mathbb{R}^{D\times D}$ e o termo s_{ij} é a covariância entre p_i e p_j. Conforme é mostrado em [6], os componentes principais podem ser encontrados como sendo o autovetores da matriz S e a suas variâncias são iguais aos autovalores associados a eles, portanto se queremos maximizar a variância total dos dados, devemos escolher os componentes principais associados aos maiores autovalores e então fazer uma projeção dos dados originais sobre o espaço vetorial cujo a base é os componentes principais escolhidos.

Sendo $\{\lambda_1, \lambda_2, ..., \lambda_k\}$ o espectro (conjunto de autovalores) de S, a contribuição do $i-$ésimo componente principal associado ao autovalor λ_i pode ser calculada como:

$$\frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j} = \frac{\lambda_i}{traço(S)}$$

Como o conjunto de dados utilizado nesse artigo não tem média 0, antes de calcular os componentes principais devemos escalonar os dados de maneira tal que eles fiquem com média 0, podemos fazer isso utilizando a seguinte transformação:

$$\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i - \mu}{\sigma} \tag{1}$$

Onde $\mu, \sigma \in \mathbb{R}^D$ são a média e o desvio padrão do dataset respectivamente. Desta forma o novo conjunto de observações, $\tilde{\mathcal{X}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_N\}$, terá média 0 e desvio padrão 1.

Aplicando esta transformação e calculando os autovalores da matriz de covariância dos dados, obtemos a contribuição de cada componente principal, figura 10,

Projetando os dados sobre os dois primeiros componentes principais, obtemos resultado da figura 11, onde os pontos rosas representam as observações da classe vinho tinto e os azuis da classe branco.

Fig. 10. Scree plot

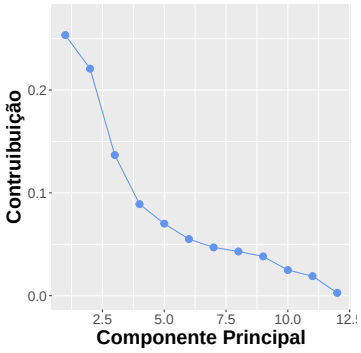
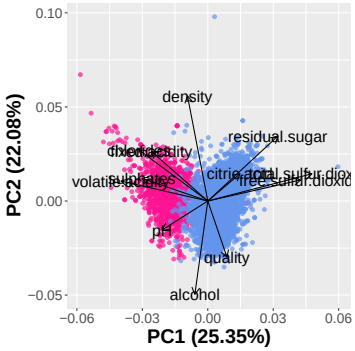


Fig. 11. Analise dos componentes principais



III. Resultados

A. Resultados da análise incondicional monovariada

A tabela I apresenta os dados obtidos na análise monovariada incondicional. Percebe-se que as medidas calculadas têm valores diferentes a depender da feature analisada pela relevante diferença dos valores das médias de cada preditor, comprovando a sua heterogeneidade. Além disso, tem-se que o maior desvio padrão é o do Dióxido Sulfúrico Total e o menor desvio padrão é o da Densidade.

Ademais, quando se analisam as assimetrias de cada preditor, nota-se que a maior, de maneira discrepante, é a de Cloretos, enquanto a menor é a de Dióxido Sulfúrico Total, a qual é a única negativa, indicando, portanto, que tal distribuição tem uma leve assimetria para a esquerda.

Dessa maneira, percebe-se que, embora a medida de assimetria de cada preditor possa ter um valor baixo, ainda é possível notar a assimetria em sua distribuição ao se realizar a observação de tal feature mediante plotagem de histograma e do cálculo de medidas como média, visto que, com auxílio da observação gráfica, consegue-se perceber como os dados se organizam em torno dela, e desvio padrão, o qual expressa como os dados coletados estão dispersos dentro do conjunto de amostras de cada preditor, destacando a importância da análise monovariada incondicional em um conjunto de dados.

Assim, conclui-se que a análise monovariada incondicional consegue auxiliar na análise dos preditores de um certo conjunto de dados, informando relevantes características deles.

TABLE I
Resultados obtidos com a análise monovariada incondicional

Preditor (d)	$\mu_{\mathbf{d}}$	$\sigma_{\mathbf{d}}$	$\gamma_{\mathbf{d}}$
Acidez fixa	7.22	1.30	1.72
Acidez volátil	0.34	0.16	1.49
Ácido cítrico	0.32	0.15	0.47
Açúcar residual	5.44	4.76	1.43
Cloretos	0.06	0.04	5.40
Dióxido sulfúrico livre	30.53	17.75	1.22
Dióxido sulfúrico total	115.74	56.52	−0.0012
Densidade	0.99	0.0030	0.50
pH	3.22	0.16	0.39
Sulfatos	0.53	0.15	1.80
Álcool	10.49	1.19	0.57
Qualidade	5.82	0.87	0.19

B. Resultados da análise classe-condicional monovariada

Os resultados das medidas estatísticas obtidas para as classes de vinho tinto e branco estão disponíveis nas Tabelas II e III, respectivamente.

Através da análise dos resultados encontrados, podemos verificar que, em média, quantidade de ácidos fixos e voláteis em vinhos tintos é maior, mas o grau de acidez revelado pelo pH indica que os vinhos brancos são levemente mais ácidos.

As amostras de vinho branco também apresentaram maior quantidade de dióxido sulfúrico, tanto o livre quanto total, substância essa utilizada para conservar o vinho [7]. A variante branca também apresentou média significativamente maior para o preditor açúcar residual, indicando que as amostras desse tipo possam ser mais doces que as do tinto.

Vale também ressaltar as médias semelhantes para a concentração de álcool e a qualidade de ambos os tipos, com a variante branca sobressaindo-se por pouco, além de apresentar menor assimetria classe-condicionada.

TABLE II
Vinho tinto (l = "tinto").

Preditor (d)	$\mu_{\mathbf{d}\|\mathbf{l}}$	$\sigma_{\mathbf{d}\|\mathbf{l}}$	$\gamma_{\mathbf{d}\|\mathbf{l}}$
Acidez fixa	8.32	1.74	0.98
Acidez volátil	0.53	0.18	0.67
Ácido cítrico	0.27	0.19	0.32
Açúcar residual	2.54	1.41	4.53
Cloretos	0.09	0.05	5.67
Dióxido sulfúrico livre	15.87	10.46	1.25
Dióxido sulfúrico total	46.47	32.90	1.51
Densidade	1.00	0.00	0.07
pH	3.31	0.15	0.19
Sulfatos	0.66	0.17	2.42
Álcool	10.42	1.07	0.86
Qualidade	5.64	1.23	0.49

C. Resultados da análise incondicional bi-variada

Conforme foi apresentado na figura 6 e 7, percebemos a relação gráfica e os valores de correlação de todos os pares de preditores.

Entre elas, existem algumas relações relevantes, como a relação do Dióxido sulfúrico total e Dióxido sulfúrico livre

TABLE III
VINHO BRANCO (l = "BRANCO").

Preditor (d)	$\mu_{\mathbf{d}\|\mathbf{l}}$	$\sigma_{\mathbf{d}\|\mathbf{l}}$	$\gamma_{\mathbf{d}\|\mathbf{l}}$
Acidez fixa	6.85	0.84	0.65
Acidez volátil	0.27	0.10	1.58
Ácido cítrico	0.33	0.12	1.28
Açúcar residual	6.39	5.07	1.08
Cloretos	0.05	0.02	5.02
Dióxido sulfúrico livre	35.31	17.00	1.41
Dióxido sulfúrico total	138.36	42.50	0.39
Densidade	0.99	0.00	0.98
pH	3.19	0.15	0.46
Sulfatos	0.49	0.11	0.98
Álcool	10.51	1.23	0.49
Qualidade	5.88	0.89	0.16

o qual apresenta a maior correlação positiva, sendo aproximadamente de 0.7. Percebendo essa relevância, realizamos a plotagem do gráfico separado, na imagem 8 (a). Conforme, já percebemos pela correlação positiva, os valores possuem uma tendencia a se tornarem linear e diretamente proporcional, o que foi confirmado no gráfico, que ambas variáveis crescem juntas, o que em alguns pontos, possuem valores muito próximos do linear.

Ademais, nos dados utilizados encontramos o valor oposto, que seria na relação do Álcool e Densidade, uma vez que a correlação é de -0.7, visualizamos que as variáveis decrescem juntas, mantendo uma característica linear bem semelhante com o do exemplo anterior, sendo isso analisado na figura 8(b).

Por utilizarmos dados com muitos preditores, é possível de se observar diversos tipos de relações nas variáveis, inclusive valores que se relacionam, mas não chegam a ser lineares como os exemplos anteriores. A relação do pH e Densidade, que possui uma correlação 0, ou seja, não existe influência entre esses preditores, não existindo linearidade entre os pontos dispersos no gráfico 8(c).

Com as análises sobre as relações entre os preditores, conseguimos, então compreender a composição dos vinhos e como a presença de cada variável afeta as outros, seja de maneira linear ou não, podendo também ser diretamente ou inversamente proporcional o crescimento dos valores. Além disso, observamos que o Álcool é o principal preditor que influência na qualidade. Por possuir uma correlação positiva de 0.4, percebemos que quanto mais Álcool, melhor será a qualidade do vinho, mas não possuirá um um crescimento linear.

Como já mencionado e visualizado nas figuras 9, é importante verificar a existência de outliers. Foram colocados alguns boxplot interessantes. Entre os preditores dos dados encontramos 2 tipos de outliers, alguns como o do Álcool e densidades que são os preditores com poucos outliers e que se encontram em um intervalo pequeno, representados na figura 9(a) e (b). Ademais, percebemos a ocorrência de alguns preditores com muitos outliers e em um intervalo grande, como o Cloreto e o Açúcar Residual, demonstrado na figura 9(c).

Logo, notamos a existência de outliers em algumas variáveis, que possivelmente alteraram os valores da correlação, sendo então passível de um tratamento, para tornar os dados melhores. Por fim, percebemos que existem poucos outliers nos valores de qualidade do vinho e de Álcool, confirmando a análise anterior de que realmente existe uma relação entre esses preditores, além de demonstrar que algumas features já são adequadas, não necessitando de um pré-processamento.

D. Resultados da análise incondicional multivariada

Podemos observar na figura 10 que os dois primeiro componentes principais não explicam tanto da variância dos dados, apenas 47.43%, apesar disso na figura 11, há uma excelente separação das classes de vinho branco e tinto, especialmente sobre o primeiro componente e, observando os *loadings*, notamos que as variáveis que mais influenciam nele são a volatilidade e a quantidade de dióxido de enxofre livre e total.

REFERENCES

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

[2] James, Gareth, et al. An Introduction to Statistical Learning. 1st ed., PDF, Springer, 2013.

[3] Kuhn, Max, and Kjell Johnson. Applied Predictive Modeling. Springer, 2018.

[4] Zheng Alice and Casari Amanda. 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Inc, 99

[5] Richardson, Mark. "Principal component analysis." URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si 6 (2009): 16.

[6] Deisenroth, M. P., Faisal, A. A.,, Ong, C. S. (2020). Mathematics for Machine Learning. Cambridge University Press.

[7] Sauceda, M. F. D. M. (2015). Polifenóis totais e dióxido de enxofre em variedades de vinho de mesa e sua relação com a ingestão diária estimada.