Análise exploratória da qualidade de vinhos

1st Cícero Oliveira da Silva Júnior
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
junioroliver@alu.ufc.br

2nd David Lima dos Santos
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
davidls10203040@gmail.com

3rd Gabriel Moreira de Andrade
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
gabrielm03@alu.ufc.br

4th Igor Pereira Gouveia
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
igorpereiragv@alu.ufc.br

***Resumo*—Este trabalho apresenta a análise dos dados provenientes do conjunto de dados "Wine Quality", o qual é composto por dois subconjuntos de dados com amostras de vinhos tintos e brancos originários de Portugal. Sobre este conjunto foi realizado a análise exploratória de dados, em que técnicas de visualização de dados e análise estatística foram utilizadas com o objetivo de interpretar e expor suas informações mais relevantes.**

***Palavras-chave*—Visualização de dados, Estatística, PCA.**

I. Introdução

A Análise Exploratória de Dados (EDA, em inglês) é uma etapa inicial fundamental no trabalho com dados, antecedendo a criação de modelos preditivos ou a aplicação de algoritmos de *Machine Learning*. Apesar de ser preliminar, essa fase desempenha um papel crucial, podendo influenciar diretamente o sucesso ou o fracasso do modelo ou da aplicação dos algoritmos. Por meio da análise explorátoria dos dados é possível compreender em detalhes a natureza do conjunto de dados. Esse processo é realizado usando os insumos da estatística descritiva que permite apresentar, organizar e sintetizar o conjunto de dados, o que torna possível a interpretação acerca das amostras análisadas e a redução as incertezas durante o processo de modelagem [1].

Este processo é essencial para identificar problemas de qualidade nos dados, como inconsistências, duplicações, anomalias e valores ausentes. Ignorar essas questões pode comprometer a eficácia dos modelos preditivos, tornando-os imprecisos ou enviesados. A correção desses problemas durante a EDA garante que os dados estejam prontos para uso em abordagens preditivas, aumentando a confiabilidade dos resultados [1].

Além disso, a EDA possibilita compreender a natureza das variáveis, ou seja, seus padrões, correlações e tendências. Ela permite entender melhor as distribuições e características de cada variável, possibilitando a exclusão ou transformação de variáveis que possam causar distorções nos resultados. Isso reduz o risco de gerar previsões enviesadas ou tomar decisões erradas com base em dados mal interpretados.

Neste trabalho, utilizaremos esta técnica para examinar o conjunto de dados "Wine Quality". Nosso objetivo é extrair insights sobre o comportamento das variáveis e associar os resultados observados à natureza e ao contexto das amostras, promovendo uma compreensão mais profunda dos padrões presentes nos dados.

II. Metodologia

O conjunto de dados em análise possui 13 variáveis, sendo uma variável qualitativa, que representa a cor do vinho, e uma variável quantitativa discreta, referente à qualidade do vinho. As demais variáveis são quantitativas contínuas e estão relacionadas às características físico-químicas do vinho, como acidez fixa, acidez volátil, ácido cítrico, açúcar residual, cloretos, dióxido de enxofre livre, dióxido de enxofre total, densidade, pH, sulfatos e teor alcoólico. A variável cor possui dois valores, vermelho/tinto e branco. Ao todo, são 6497 observações, sendo 1599 amostras de vinho vermelho e 4898 amostras de vinho branco.

A variável qualidade assume 7 valores distintos no conjunto de dados, pertencentes ao conjunto 3, 4, 5, 6, 7, 8, 9. Com o objetivo de realizar uma análise com maior abstração para compreensão dos dados, esta variável foi separada em três classes, conforme descrito abaixo:

$$\text{Classe qualidade(x)} = \begin{cases} \text{Ruim, se } x \leq 5 \\ \text{Médio, se } 6 \leq x \leq 7 \\ \text{Bom, se } x \geq 8 \end{cases}$$

Para a análise dos dados seccionamos a abordagem em três etapas:

1) Análise univariada
2) Análise bivariada
3) Análise multivariada

Essa separação foi utilizada para possibilitar a compreensão de forma individual e pareada do comportamento e da causalidade das variáveis presentes no conjunto de dados.

Durante a análise univariada é possível averiguar, de modo individual à uma variável, a distribuição dos dados por meio de histogramas, a variabilidade e *outliers* por meio de boxplots, assim como os valores referentes média (Equação 1), desvio

padrão (Equação 2) e assimetria (Equação 3). Segue abaixo as fórmulas utilizados para o cálculo das estatísticas.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{2}$$

$$\gamma = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3 \tag{3}$$

Já na etapa de análise bivariada é averiguada a relação entre os features de forma pareada. O objetivo é avaliar a correlação entre os pares de variáveis, além de verificar se essas relações apresentavam um comportamento linear ou não. A análise do comportamento linear pode ser realizada por meio de scatter plots, enquanto a correlação entre os features pode ser calculada utilizando o coeficiente de Pearson e seus valores podem ser dispostos em um heatmap, de modo a facilitar a interpretação visual das associações. Essa análise pode ser realizada tanto no âmbito condicional quanto incondicional. Segue abaixo a formula correspondente ao coeficiente de Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

Por fim, na análise multivariada é averiguada as relações entre as múltiplas variáveis e os padrões nos dados. Dentre as abordagens utilizadas, destaca-se a Análise de Componentes Principais (PCA, em inglês). Tal processo tem como objetivo encontrar projeções em um espaço vetorial de menor dimensão que sejam semelhantes aos pontos de dados originais, minimizando a perda de informação. Este método permite resumir o conteúdo informacional de grandes conjuntos de dados em um conjunto menor de variáveis não correlacionadas, denominadas componentes principais. Estes componentes são combinações lineares das variáveis originais que maximizam a variância, em comparação com outras combinações lineares [2].

Para realização da PCA os seguintes passos são seguidos:

- Padronização dos Dados
- Cálculo da Matriz de Covariância
- Cálculo dos Autovalores e Autovetores
- Projeção das Componentes Principais

A. Padronização dos Dados

Para utilizar o PCA, inicia-se padronizando os dados, de forma que estes reflitam uma distribuição com média zero e desvio-padrão igual a 1. A padronização é realizada em duas etapas. Primeiramente, calcula-se a média de cada variável e subtrai-se este valor de cada elemento da variável correspondente, centralizando os dados. Em seguida, cada elemento é dividido pelo desvio-padrão da variável correspondente, garantindo que todas as variáveis estejam na mesma escala.

B. Cálculo da Matriz de Covariância

A variância é um indicador do espalhamento dos dados, e o PCA pode ser descrito como um algoritmo de redução de dimensionalidade que maximiza a variância na nova representação de menor dimensão, retendo o máximo possível da informação dos dados originais. Para isso, calcula-se a matriz de covariância, que é uma medida estatística que captura a relação entre duas variáveis.

A covariância entre duas variáveis x_1 e x_2 é dada por:

$$\text{Cov}(x_1, x_2) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)(x_{i,2} - \bar{x}_2), \tag{5}$$

onde $\bar{x}_1$ e $\bar{x}_2$ são as médias das variáveis x_1 e x_2, respectivamente. Se a covariância for positiva, x_1 e x_2 aumentam simultaneamente; se for negativa, x_1 aumenta enquanto x_2 diminui; se for zero, as variáveis não apresentam correlação linear direta.

C. Cálculo dos Autovalores e Autovetores

Com a matriz de covariância calculada, calcula-se seus autovalores e autovetores. Cada autovalor indica a quantidade de variância explicada pelo autovetor correspondente. Os autovetores, por sua vez, representam as direções principais dos dados.

$$\text{Taxa de variância explicada (EVR)}_i = \frac{\lambda_i}{\sum_{j=1}^{m}\lambda_j} \tag{6}$$

- λ_i: Variância capturada pelo i-ésimo componente principal.
- $\sum_{j=1}^{m}\lambda_j$: Soma total das variâncias (autovalores), representando a variância total dos dados.
- m: Número total de variáveis numéricas consideradas no PCA.

A soma das taxas de variância explicada para todos os componentes principais será sempre igual a 1 (100%). A taxa de variância explicada de cada componente pode ser observada por meio de um scree plot.

D. Projeção das Componentes Principais

Para reduzir a dimensionalidade, ordenam-se os autovalores em ordem decrescente e selecionam-se os autovetores correspondentes aos maiores autovalores. O autovetor associado ao maior autovalor define o primeiro componente principal, que é a direção que captura a maior variância dos dados. O segundo maior autovalor corresponde ao segundo componente principal, e assim por diante, até o n-ésimo maior autovalor, que define o n-ésimo componente principal. A relação entre o espaço projetado e o espaço original pode ser obtida projetando os dados nas direções dos componentes principais.

$$X_{\text{projetado}} = X \cdot W \tag{7}$$

- X: Matriz de dados original com dimensão $n \times m$, onde n é o número de observações (amostras) e m é o número de atributos (variáveis).

- W: Matriz de projeção composta pelos autovetores correspondentes aos k maiores autovalores, com dimensão $m \times k$, onde k é o número de componentes principais selecionados.
- $X_{\text{projetado}}$: Matriz resultante no espaço reduzido, com dimensão $n \times k$, representando as observações projetadas nos k componentes principais.

III. Resultados

A. Análise monovariada não condicionada à classes

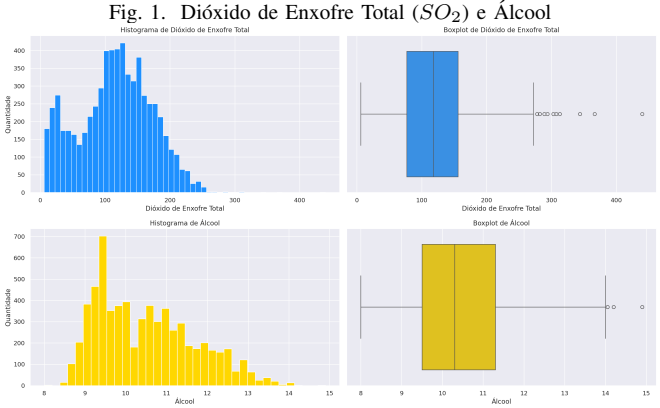
A abordagem em relação as dados teve o âmbito populacional e de classes. Primeiramente avaliamos os parâmetros estatísticos das amostras de vinho como um todo e os seguinte resultados referentes a média, desvio padrão e distorção foram encontrados:

TABLE I
RESUMO ESTATÍSTICO DAS COLUNAS DO DATASET COMPLETO.

Geral	***Skewness***	***Média***	***Desvio Padrão***
residual sugar	1.4354	5.4432	4.7578
total sulfur dioxide	-0.0012	115.7446	56.5219
pH	0.3868	3.2185	0.1608
alcohol	0.5657	10.4918	1.1927

A fim de visualizar gráficamente a distribuição do conjunto de amostras, os histogramas e *boxplot* da Densidade, do Álcool, do pH e do Açucar residual foram gerados:

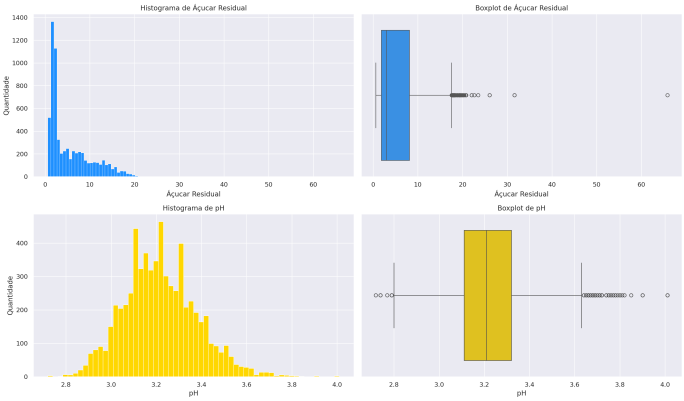
Fig. 1. Dióxido de Enxofre Total (SO_2) e Álcool



Como é possível analisar na Figura 1, os dados do histograma do álcool estão bem distribuídos, apresentando poucos *outliers* no *boxplot*. Observa-se que 75% dos dados (intervalo interquartil) estão concentrados entre 9,5% e 11,5%, indicando uma produção predominantemente homogênea.

Já ao analisar a variável dióxido de enxofre total também na Figura 1, percebe-se uma maior presença de outliers, o que indica maior variabilidade nos dados. A concentração de valores entre 0 e 50 e entre 100 e 150 sugere diferentes práticas de utilização desse composto em vinhos distintos. Isso pode estar relacionado a diferentes estilos de vinhos (como tintos, brancos ou espumantes) e seus requisitos de preservação. Como será visto em análises posteriores, o dióxido de enxofre total é uma variável importante para determinar o tipo de vinho.

Fig. 2. pH e Açucar residual.



Analisando o pH na Figura 2, percebe-se uma alta quantidade de *outliers*, embora a distribuição geral seja aproximadamente normal. O histograma indica que a maioria dos dados está concentrada em torno da média, refletindo uma distribuição tendencialmente simétrica.

Já o açúcar residual, além de apresentar uma quantidade relevante de *outliers*, exibe uma alta concentração de dados à esquerda dos gráficos. Essa assimetria indica que a maioria dos vinhos do dataset possui baixos níveis de açúcar residual.

B. Análise monovariada condicionada à classes

Como dito inicialmente, os nossos dados apresentam classes relativas a cor do vinho, assim como a sua qualidade. Para compreendermos melhor o conjunto de dados, é necessário visualizar como estes se comportam sob a ótica dessas classificações.

Abaixo temos as medidas estatísticas referentes às amostras de cada cor. Na Tabela II, estão os valores para as amostras de vinho vermelho, e, na Tabela III, para as amostras de vinho branco.

TABLE II
RESUMO ESTATÍSTICO DAS FEATURES PRINCIPAIS

Tinto	***Skewness***	***Média***	***Desvio Padrão***
total sulfur dioxide	1.5155	46.4678	32.8953
pH	0.1937	3.3111	0.1544

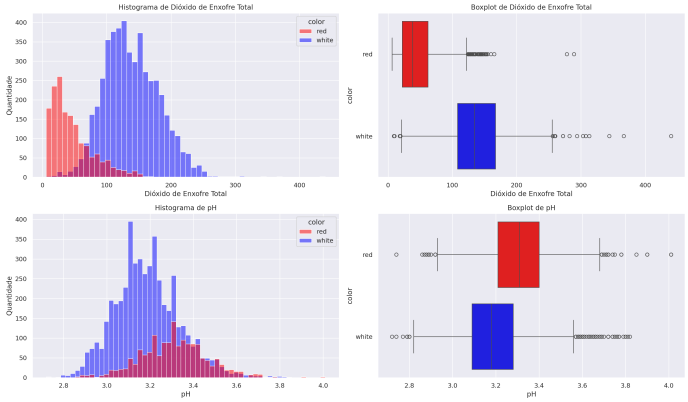
TABLE III
RESUMO ESTATÍSTICO DAS FEATURES PRINCIPAIS

Branco	***Skewness***	***Média***	***Desvio Padrão***
total sulfur dioxide	0.3907	138.3607	42.4981
pH	0.4578	3.1883	0.1510

Na sequência, temos os histogramas e *boxplot* para algumas variáveis do *dataset*, só que, dessa vez, contendo duas cores em cada plot. Azul para representar as amostras brancas e vermelho para representar as amostras de vinho tinto.

Na Figura 3, analisando a variável de dióxido de enxofre, notamos que a média da concentração é maior nos vinhos

Fig. 3. Dióxido de Enxofre (SO_2) Total e pH dos vinhos tintos e brancos



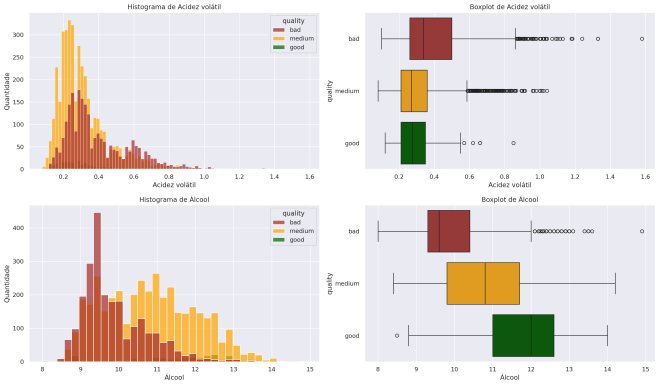
brancos do que nos tintos. Amplamente utilizado para controle de microorganismos durante a preparação do vinho, esse composto químico é comum de ser mais encontrado nos vinhos brancos. A curva dos vinhos tintos (1.51) tem uma assimetria positiva consideravelmente maior que ao valor da curva azul (0.39), cuja distribuição tende a ser normal. Enquanto que nos vinhos brancos se tem um nível médio de 138 mg/L de dióxido de enxofre molecular, os vinhos tintos estao com uma média de 46 mg/L. Isso se deve porque, além de poder ser aplicado em diferentes fases da preparação do vinho branco, como na antes da fermentação ou do engarrafamento, o vinho tinto possui essa substância mais moderadamente, uma vez que o SO_2 afeta nos aromas e no clareamento das cores dos insumos das uvas tintas, o que contribui para diminuir a presença do sabor forte dos vinhos com cor [3].

Quanto a acidez dos vinhos, medida pelo pH nos gráficos, podemos inferir que os vinhos brancos tem um valor menor que os vinhos tintos. A média destes é cerca de 0.12 pontos maior que a média daqueles. No geral, os vinhos brancos possuem um pH no intervalo entre 3.0 e 3.4, e os vinhos tintos possuem no intervalo de 3.3 a 3.6 [4]. Por outro lado, ambos possuem desvio padrão similar, com os *outliers* dos vinhos brancos alcançando valores mais baixos, ao passo que os dos vinhos tintos alcamçam valores mais altos. A acidez é essencial para um bom envelhecimento do vinho, porém, principalmente para os tintos, ele afeta no seu sabor frutoso e complexo, uma vez que o nível de acidez pode facilitar a digestão rápida do vinho [5]. Em vista disso, os vinhos brancos são mais ácidos e possuem um sabor mais equilibrado e refrescante.

Ademais, podemos perceber que essas duas variáveis se comportam de maneira inversa em cada um dos vinhos. Os vinhos tintos se apresentam com uma média de pH maior, mas tem um nível de concentração médio de SO_2 total menor, enquanto que com os vinhos branco ocorre o contrário. Um valor de pH influencia muito na vida microbiana durante a preparação dos vinhos [4]. Um pH menor inibe mais alguns tipos de micróbios, fazendo com que o dióxido de enxofre, principalmente na sua forma livre, tenha uma concentração maior. Vinhos com um pH maior exigem serem nutridos por mais conservantes como o SO_2 durante o processo de armazenamento, uma vez que o nível presente neles não é suficiente para atrasar a deterioração [4]. Em vista disso, vinhos brancos possuem uma longevidade melhor que os vinhos tintos.

Olhando no aspecto geral, sem considerar a distinção de cores dos vinhos, na Figura 4 temos histogramas e boxplots das variáveis de acidez volátil e álcool, sob à segmentação das classes de qualidade, explicada anteriormente. Torna-se necessário analisar como dois dos principais indicadores de qualidade de vinho se apresentam neste *dataset*.

Fig. 4. Qualidade dos vinhos a partir da Acidez Volátil e do Álcool



Se o pH é a medição da quantidade de ácidos dissociados presentes na solução, a acidez volátil mede o grau de concentração dos ácidos gasosos dos vinhos, mais especificamente, do ácido acético e compostos relacionados [6]. No histograma, vemos que as três classes tem curvas de assimetria positiva, sendo as classes ruim e média as que possuem maior número de elementos na cauda à direita, o que se confirmar ao visualizar a grande presença de *outliers* no *boxplot* de ambas. O nível aceitável de ácido volátil varia entre 0.3 e 0.5 g/L [6], o que corrobora com os gráficos das 3 classes, as quais possuem média entre 0.2 e 0.4 g/L. Entretanto, notemos que a classe ruim possui mais amostras com uma concentração maior que 0.5g/L, o que nos leva a interpretar que uma maior quantidade desse composto afeta negativamente o gosto e a sensação aromática do vinho para as pessoas. Isso se deve, pois, essa elevada presença desse composto leva a bebida a ter um cheiro mais semelhante ao de um vinagre. Quando equilibrado, o ácido volátil é responsável por garantir um sabor refrescante e que equilibra doçura e acidez.

O teor alcóolico de um vinho é um dos fatores mais levado em consideração na hora de se comprar um vinho. No *dataset* em questão, as distribuições das 3 classes nessa variável estão bem distintas uma das outras. Enquanto que a qualidade ruim tem uma distribuição com assimetria positiva, a média tem uma distribuição quase normal, e a boa, apesar da baixa quantidade de amostras, tem uma distribuição com assimetria negativa. O álcool do vinho deriva de processo de sintetização dos açúcares das uvas em dióxido de carbono e álcool, perdurando pela fase de armazenamento em barris [7]. Abaixo de 11.5% é considerado nível abaixo, em que o gosto do vinho é mais leve e sutil, enquanto que acima de 14% é nível alto. Mesmo que o gosto pessoal acerca da concentração de álcool desejável mude

entre as pessoas, é mais comum que a faixa aceitável esteja entre 11% e 13%, como demonstra a Tabela IV as médias das classes boas e médias estão entre 10.5% e 12%. Dessa forma, as pessoas procuram um nível suficiente para deixar o vinho encorpado, doce, ácido e equilibrado, mas sem exagerar, evitando o teor de queimação que níveis elevados de álcool provocam.

TABLE IV
RESUMO ESTATÍSTICO DAS COLUNAS DO DATASET POR CATEGORIA DE QUALIDADE

Categoria	Coluna	*Skewness*	*Média*	*Desvio Padrão*
Ruim	volatile acidity	1.2511	0.3974	0.1880
	alcohol	1.1685	9.8735	0.8417
Médio	volatile acidity	1.5123	0.3070	0.1398
	alcohol	0.2438	10.8076	1.2011
Bom	volatile acidity	1.2237	0.2912	0.1181
	alcohol	-0.8637	11.6914	1.2733

C. Análise Bivariada

Por meio da figura 5, podemos verificar que os dados de em um panorama geral apresentam um baixo nível de correlação e que nestes casos em questão tal relação aparenta ser linear.

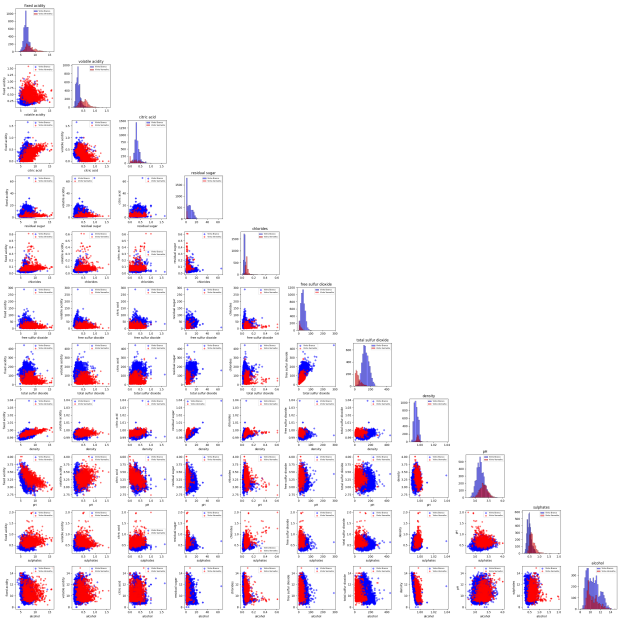


Fig. 5. Scatter plot condicional por cor do vinho

Por meio da Figura 6, podemos observar com maior clareza quais variáveis possuem um maior grau de correlação, seja ela positiva ou negativa. Podemos destacar os seguintes pares: (dióxido de enxofre total, dióxido de enxofre livre), (densidade e álcool) e (densidade, açúcar residual).

Para compreendermos esses resultados, é importante analisar o processo de fabricação do vinho, especialmente o momento da fermentação. Durante esse processo, o suco de uva se transforma em álcool etílico, uma substância orgânica gerada pela fermentação de açúcares como a glicose e a frutose. No

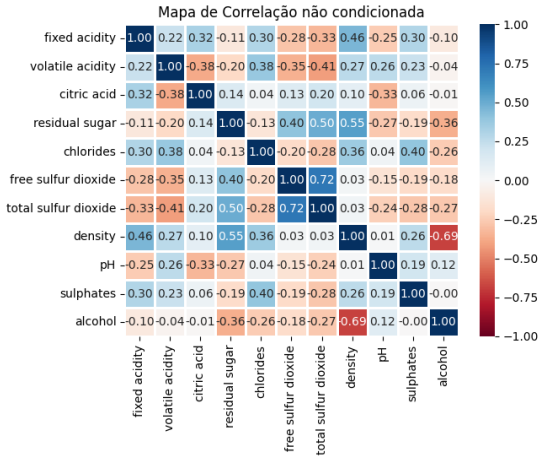


Fig. 6. Mapa de correlação das variáveis

entanto, em alguns casos, ocorre um fenômeno denominado ”parada prematura da fermentação”, no qual o processo é interrompido antes que todo o açúcar seja convertido em álcool, resultando em um nível de açúcar residual.

Outro aspecto relevante é que uma maior concentração de açúcares na mistura aumenta sua densidade, uma vez que a água compõe cerca de 80% da uva e tem densidade inferior à do açúcar, enquanto o álcool etílico é menos denso que a água. Dessa forma, podemos explicar por que os pares (densidade, álcool) e (densidade, açúcares residuais) apresentam as características de correlação observadas na Figura 6.

Já correlação entre o dióxido de enxofre total e o dióxido de enxofre livre pode ser explicada pela adição de SO_2 durante a elaboração do vinho, parte do SO_2 adicionado se combina com compostos como oxigênio, açúcares e aldeídos, enquanto apenas o SO_2 livre permanece ativo, o que justifica a alta correlação positiva observada.

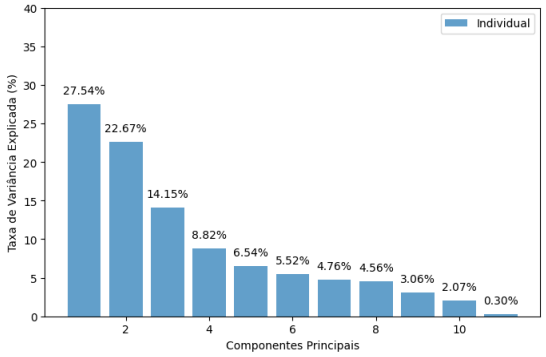
D. Análise Multivariada

Após a realização de todas as etapas da PCA, obtivemos as informações referentes à taxa de variância explicada por cada componente principal. Verificamos que a primeira e a segunda componentes, em conjunto, representam 50,21% da variação total dos dados. Essa informação pode ser visualizada na Figura 7, que ilustra a proporção de variância explicada por cada componente principal obtida no processo.

Em seguida, realizamos a projeção das duas primeiras componentes principais. Inicialmente, essa projeção foi feita de forma incondicional, considerando todas as amostras de maneira geral e depois aplicamos uma projeção condicional, na qual identificamos a classe de cada amostra utilizando cores e formas geométricas distintas.

Observando gráfico das duas primeiras componentes principais da PCA incondicional, Figura 8, verificamos que os dados não estão nem completamente bem distribuídos, nem mal distribuídos, o que era esperado, considerando que as duas primeiras componentes principais explicam aproximadamente 50% da variabilidade total dos dados. Isso indica que, embora

Fig. 7. Scree plot



essas componentes capturem uma boa parte da variação, ainda há informações importantes distribuídas em direções adicionais que não são totalmente representadas.

Fig. 8. PCA não condicional

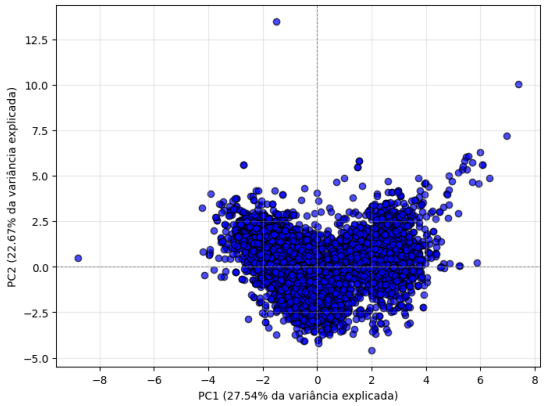
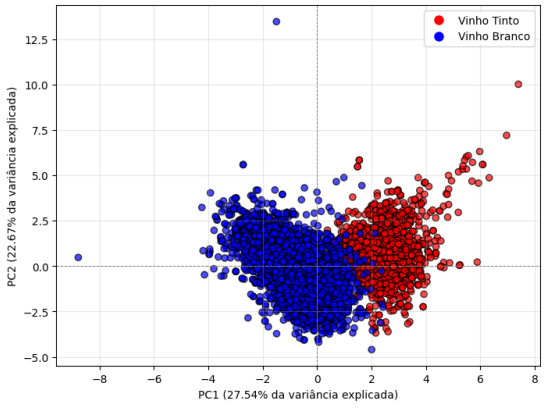
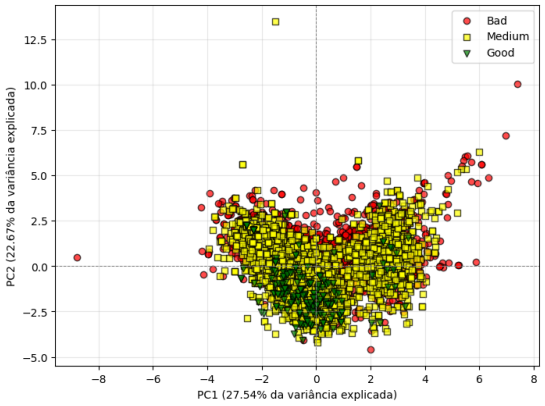


Fig. 9. PCA condicionada à variável cor



Ao analisarmos os dados da projeção condicional referente a cor, podemos observar, pela Figura 9, uma transição/fronteira entre as amostras de vinho branco e vinho tinto. É possível visualizar, de maneira razoavelmente clara, os agrupamentos correspondentes a cada tipo de vinho. Vale lembrar que nesse conjunto de dados a quantidade de amostras referentes ao vinho tinto é cerca de 3 vezes menor que a de vinho branco, caso tivéssemos mais amostras possivelmente poderíamos visualizar um agrupamento mais bem definido.

Fig. 10. PCA condicionada à variável qualidade



Por fim, ao analisarmos a projeção condicional referente à qualidade, não conseguimos observar pela Figura 10 um agrupamento claro entre as diferentes classes de qualidade do vinho. Isso indica que, com base apenas nas duas primeiras componentes principais, não é possível separar de maneira eficaz as amostras de diferentes qualidades.

References

[1] DSAcademy Blog, *A importância da Análise Exploratória de Dados (EDA) no sucesso de projetos de Data Science*, 29-set-2024. [Online]. Disponível em: https://blog.dsacademy.com.br/a-importancia-da-analise-exploratoria-de-dados-eda-no-sucesso-de-projetos-de-data-science/

[2] IBM, "Principal Component Analysis (PCA)". [Online]. Disponível em: https://www.ibm.com/br-pt/topics/principal-component-analysis. Acessado em: 06-dez-2024.

[3] P. Henderson, *Sulfur Dioxide and Wine*, Practical Winery and Vineyard Journal, jan-2009. [Online]. Disponível em: https://home.sandiego.edu/~josephprovost/SulfurDioxideandWine.pdf

[4] Atlas Scientific: Environmental Robotics, *The importance of pH in wine making*, 6-jul-2023. [Online]. Disponível em: https://atlas-scientific.com/blog/the-importance-of-ph-in-wine-making/?srsltid=AfmBOorPmRHs5NsZRjkTS5-kXDFa8xpyJPoWdUuAYGmeHCwuQposhQAc.

[5] B. Robson, *12 Noteworthy Differences between Red Wine and White Wine*. [Online]. Disponível em: https://wdd.my/blog/comprehensive-guide-white-red-wines/.

[6] WineFun, *Acidez volátil: conheça um dos aspectos mais controvertidos do mundo dos vinhos*, 24-nov-2021. [Online]. Disponível em: https://winefun.com.br/acidez-volatil-conheca-um-dos-defeitos-mais-controvertidos-do-mundo-dos-vinhos/

[7] Art des Caves, *A qualidade do vinho depende do teor alcoólico?*, 16-jun-2016. [Online]. Disponível em: https://blog.artdescaves.com.br/qualidade-vinho-depende-teor-alcoolico

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in Python*, 2023rd ed., Springer Texts in Statistics, 2023.

[9] Divvino, "O que é o açúcar residual no vinho?". [Online]. Disponível em: https://www.divvino.com.br/blog/acucar-residual-vinho/. Acessado em: 06-dez-2024.

[10] A. Bender, J. F. Martinez, P. S. Lúcio, V. BrasilCosta, R. S. E. Silva, and M. B. Malgarim, "Avaliação de dióxido de enxofre total e livre em vinhos artesanais e comparação com especificações legais", *SIEPE 2014*, [Online]. Disponível em: https://cti.ufpel.edu.br/siepe/arquivos/2014/CA_03403.pdf. Acessado em: 06-dez-2024.

[11] Google Colab: Códigos utilizados. Link: https://colab.research.google.com/drive/1G4On5N8m-r6FwGup8adWWbC-8eRMAFVP?usp=sharing