UNIVERSIDADE CATÓLICA DE SANTOS

CENTRO DE CIÊNCIAS EXATAS, ARQUITETURA E ENGENHARIAS

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO/SISTEMAS DE INFORMAÇÃO

FABRIZIO JACINTO
GABRIEL AUGUSTO SIMÕES FAIA CONRADO
SAMUEL MENESES ABREU MAGALHÃES
THIAGO COUTINHO DE JESUS

USO DE TÉCNICAS DE ANÁLISE DE DADOS PARA AVALIAÇÃO QUALITATIVA DO VINHO TINTO

Santos

2025

FABRIZIO JACINTO
GABRIEL AUGUSTO SIMÕES FAIA CONRADO
SAMUEL MENESES ABREU MAGALHÃES
THIAGO COUTINHO DE JESUS

USO DE TÉCNICAS DE ANÁLISE DE DADOS PARA AVALIAÇÃO QUALITATIVA DO VINHO TINTO

Pesquisa Curricularizada de Graduação apresentada nos componentes: arquitetura de redes de computados, interação homem-computador e programação orientada à objetos; como exigência para aprovação nos componentes citados, durante o 4º semestre do curso de Ciência da Computação/Sistemas de Informação.

Santos

2025

Sumário

1. INTRODUÇÃO

O vinho é uma bebida milenar, citada, inclusive, nas Sagradas Escrituras e liturgia católicas, sendo o fruto do trabalho do homem que traz júbilo e gozo para os que ingerem. Há uma profissão chamada sommelier de vinho, cuja atribuição é determinar a qualidade de tal. Sendo assim, o grupo escolheu esse tema por sua relevância histórica, profissional e por ser uma bebida muito apreciada por alguns integrantes do grupo.

A qualidade do vinho é tradicionalmente avaliada por especialistas por meio de testes sensoriais, que podem ser subjetivos e inconsistentes. Este trabalho explora uma abordagem baseada em mineração de dados para prever a qualidade do vinho a partir de propriedades físico-químicas mensuráveis, utilizando um conjunto de dados de vinhos portugueses "vinho verde". O estudo se baseia no artigo "Modeling wine preferences by data mining from physicochemical properties" (Cortez, Paulo et al., 2009), que aplica técnicas como Support Vector Machines (SVM), redes neurais e regressão múltipla para modelar preferências sensoriais.

A relevância deste tema está na aplicação de técnicas computacionais para otimizar processos na indústria vinícola, desde a produção até a certificação, reduzindo a dependência de avaliações humanas e identificando fatores críticos para a qualidade.

Para além da análise de dados sobre a qualidade do vinho, que constitui o foco central desta pesquisa, o presente trabalho também se dedica a uma avaliação da plataforma digital de origem do dataset. Esta análise secundária foi conduzida com o objetivo de examinar a interface e a experiência do usuário do website. Para isso, foram aplicadas as 10 Heurísticas de Usabilidade de Jakob Nielsen, e, adicionalmente, foram realizados testes de acessibilidade com base nas diretrizes do eMAG (Modelo de Acessibilidade em Governo Eletrônico) e da WCAG (Web Content Accessibility Guidelines).

2. DEMONSTRATIVO ANALÍTICO

2.1. Tipo de dado

A primeira ação realizada foi uma verificação para assegurar que as colunas possuem tipos únicos de dados, para isso foi utilizado a função dtypes, da biblioteca pandas. O resultado é a tabela abaixo:

Tabela 1 - Tipos das colunas.

Variável	Tipo de Dado
fixed acidity	float64
volatile acidity	float64
citric acid	float64
residual sugar	float64
chlorides	float64
free sulfur dioxide	float64
total sulfur dioxide	float64
density	float64
pH	float64
sulphates	float64
alcohol	float64
quality	int64

Fonte: os autores.

2.2. Análise estatística

Posteriormente foi feita a análise estatística geral do dataset por colunas, ou seja, foram obtidos os seguintes valores de cada coluna:

- Count: número de linhas
- Mean: média
- Std: desvio padrão

- Min: menor valor da coluna
- 25%: primeiro quartil
- 50%: segundo quartil
- 75%: terceiro quartil
- Max: maior valor da coluna

Abaixo é possível visualizar a tabela com os valores calculados:

Tabela *2* - Descrição do dataset.

Estatística	**fixed acidity**	**volatile acidity**	**citric acid**	**residual sugar**	**chlorides**	**free sulfur dioxide**
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000
Max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000

Fonte: os autores.

Tabela 2 – Continuação.

Estatística	**total sulfur dioxide**	**density**	**pH**	**sulphates**	**alcohol**	**quality**
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000

Estatística	total sulfur dioxide	density	pH	sulphates	alcohol	quality
50%	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

Fonte: os autores.

2.2.1. Desvio padrão

O desvio padrão é uma medida estatística com a seguinte equação (1):

$$\sqrt{\frac{\sum(x_i + u)^2}{n}} \quad (1)$$

Ele indica o quão "espalhados" os valores estão: um desvio padrão baixo significa que os dados estão muito próximos da média, enquanto um desvio padrão alto indica que os dados estão distribuídos de uma forma mais ampla.

Segundo Montgomery e Runger (2018), o desvio padrão é a medida de dispersão mais importante e amplamente utilizada na estatística, sendo simplesmente a raiz quadrada da variância. Ele é expresso na mesma unidade dos dados originais, o que facilita enormemente sua interpretação. Um valor pequeno para o desvio padrão indica que as observações estão em sua maioria, próximos da média, o que mostra consistência e baixa variabilidade. Por outro lado, um valor grande diz que os dados estão mais dispersos.

2.2.2. Quartil

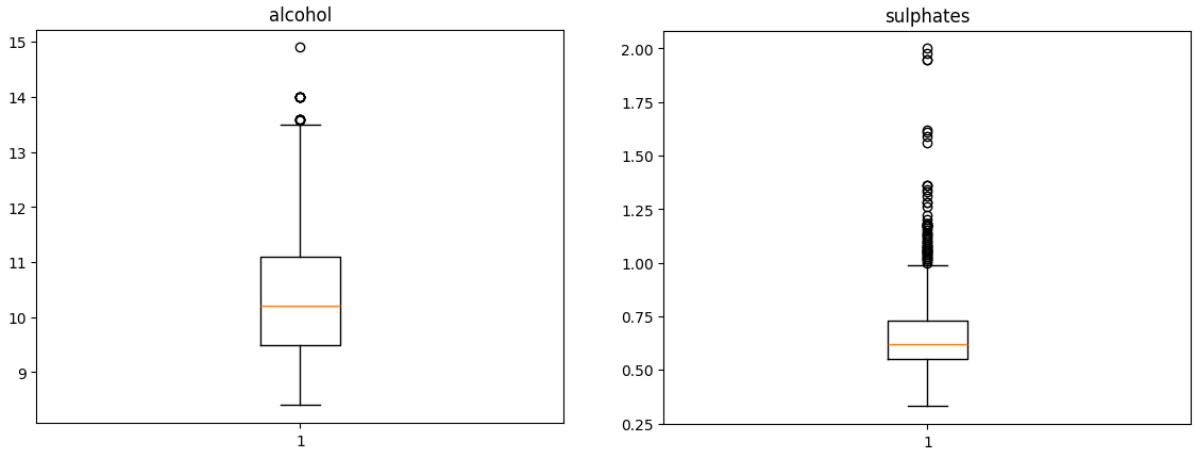
Os quartis são medidas estatísticas que dividem um conjunto de dados ordenado em quatro partes iguais, sendo essenciais para entender a distribuição e a dispersão dos dados. Segundo Triola (2018), são medidas de posição úteis para resumir as características principais de um conjunto de dados. O Primeiro Quartil (Q1) corresponde ao valor que deixa 25% das observações abaixo dele; o Segundo Quartil (Q2) é a mediana, que divide o conjunto ao meio; e o Terceiro Quartil (Q3) corresponde ao valor abaixo do qual se encontram 75% das observações.

A principal aplicação dos quartis neste trabalho é a identificação de outliers (valores discrepantes) através do cálculo do Intervalo Interquartil (IIQ = Q3 - Q1), conforme será demonstrado nos gráficos de boxplot.

2.3.Outliers

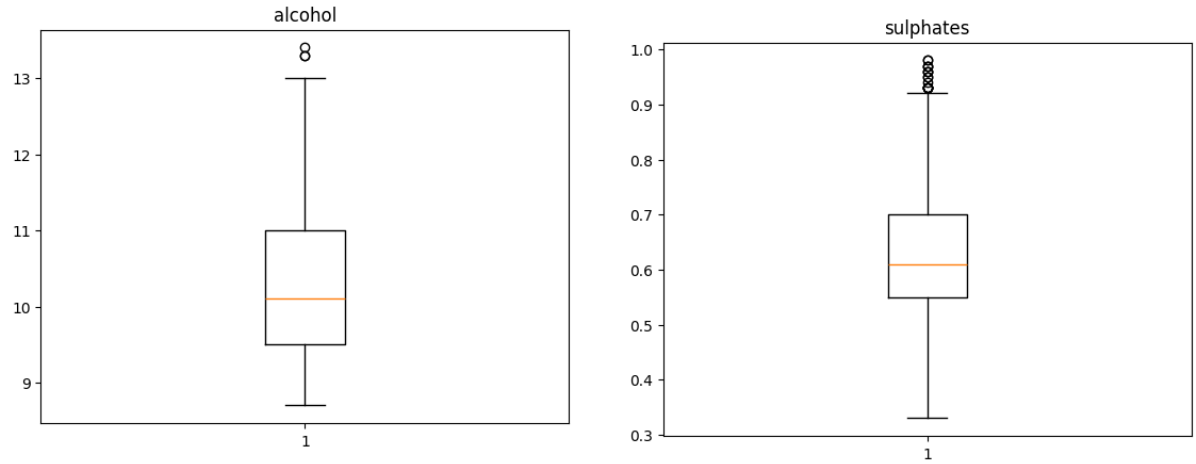
Para focar a análise, são apresentados a seguir os gráficos de boxplot que ilustram o impacto do tratamento de outliers. São exibidas as variáveis que tiveram melhora na correlação com a qualidade, como álcool e sulfatos (Figuras 1 e 2), e aquelas que demonstraram piora, como a acidez fixa e a acidez volátil (Figuras 3 e 4), comparando os dados antes e depois do tratamento.

Figura 1 - Dados que melhoraram depois do tratamento antes.



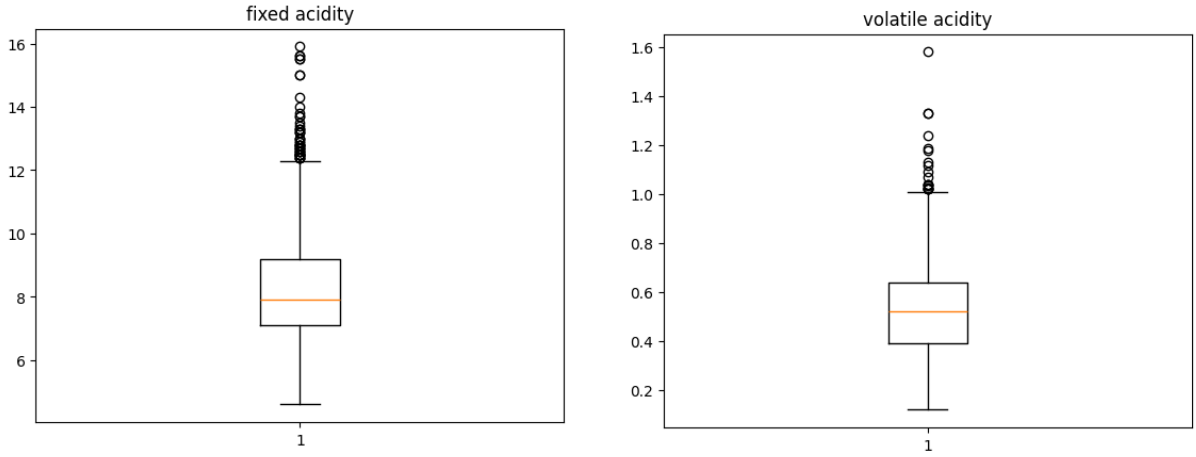
Fonte: os autores.

Figura 2 - Dados que melhoraram depois do tratamento depois.



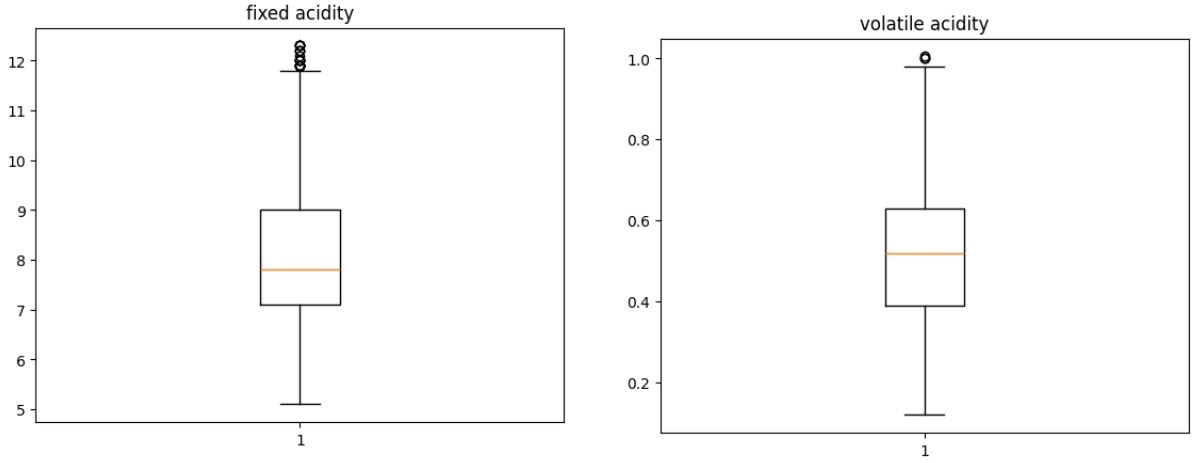
Fonte: os autores.

Figura 3 - Dados que pioraram depois do tratamento antes.



Fonte: os autores.

Figura 4 - Dados que pioraram depois do tratamento depois.



Fonte: os autores.

As tendências visuais observadas nos gráficos são confirmadas numericamente pelas tabelas de correlação a seguir. A Tabela 3 apresenta a correlação de cada variável com a qualidade do vinho antes do tratamento de outliers, enquanto a Tabela 4 mostra os mesmos dados após a remoção dos valores discrepantes, permitindo uma comparação direta do impacto dessa limpeza.

Tabela 3 - Correlação antes do tratamento.

Variável	**Correlação com Quality**
fixed acidity	0.124052
volatile acidity	-0.390558
citric acid	0.226373
residual sugar	0.013732
chlorides	-0.128907
free sulfur dioxide	-0.050656
total sulfur dioxide	-0.185100
density	-0.174919
pH	-0.057731
sulphates	0.251397
alcohol	0.476166
quality	1.000000

Fonte: os autores.

Tabela 4 - Tabela de correlação depois do tratamento.

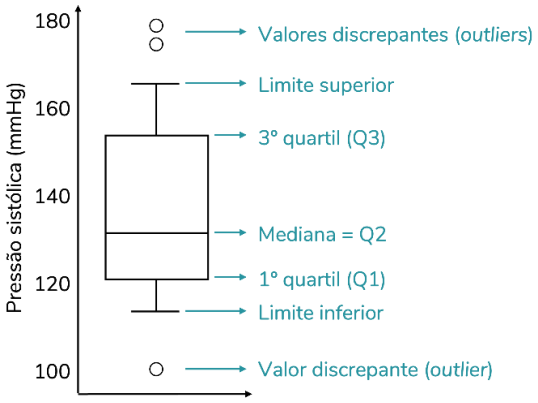
Variável	Correlação com Quality
fixed acidity	0.113422
volatile acidity	-0.346962

Variável	Correlação com Quality
citric acid	0.212133
residual sugar	0.007934
chlorides	-0.190869
free sulfur dioxide	-0.003609
total sulfur dioxide	-0.203374
density	-0.215375
pH	-0.060288
sulphates	0.413533
alcohol	0.492551
quality	1.000000

Fonte: os autores.

Esses gráficos são relevantes para a análise, sendo necessário, primeiramente, compreender como são formados. Anteriormente, foi abordado o conceito de quartil, e qual e utilizado na construção do gráfico de boxplot, conforme explicado a seguir pela figura 5:

Figura 5 - Construção do boxplot.



Fonte: Peres, Fernandes 2022.

A análise das correlações após o tratamento dos outliers revelou resultados mistos. Embora algumas variáveis tenham enfraquecido sua correlação com a qualidade, a concentração de sulfatos demonstrou uma melhora substancial. Ponderando esses efeitos, optou-se por manter os dados tratados, priorizando o fortalecimento de um indicador que se mostrou relevante.

Como consequência, as classes de qualidade com baixa representatividade no conjunto de dados foram removidas. Essa decisão alinha-se ao objetivo do estudo, que foca na análise dos vinhos de qualidade 5, 6 e 7, por constituírem a grande maioria das amostras. A nova análise estatística descritiva, referente a este conjunto de dados filtrado, é apresentada na Tabela 5.

Tabela 5 - Análise depois da limpeza.

Estatística	**fixed acidity**	**volatile acidity**	**citric acid**	**residual sugar**	**chlorides**	**free sulfur dioxide**
count	1146.000000	1146.000000	1146.000000	1146.000000	1146.000000	1146.000000
mean	8.165620	0.518783	0.249040	2.186998	0.078540	15.166667
std	1.461328	0.161485	0.179149	0.441114	0.014306	8.808946
min	5.100000	0.120000	0.000000	1.200000	0.041000	1.000000
25%	7.100000	0.390000	0.090000	1.900000	0.069000	8.000000
50%	7.800000	0.520000	0.240000	2.100000	0.078000	13.500000
75%	9.000000	0.630000	0.390000	2.500000	0.087000	20.000000
max	12.300000	1.005000	0.730000	3.600000	0.119000	42.000000

Fonte: os autores.

Tabela 5 – Continuação.

Estatística	**total sulfur dioxide**	**density**	**pH**	**sulphates**	**alcohol**	**quality**
count	1146.000000	1146.000000	1146.000000	1146.000000	1146.000000	1146.000000
mean	42.566318	0.996576	3.323412	0.633848	10.360849	6.670157

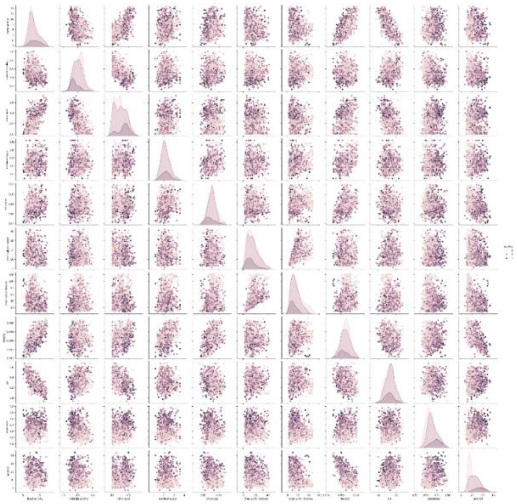
Estatística	total sulfur dioxide	density	pH	sulphates	alcohol	quality
std	26.112913	0.001596	0.131807	0.115844	0.968572	0.676653
min	6.000000	0.992360	2.940000	0.370000	8.700000	5.000000
25%	23.000000	0.995503	3.230000	0.550000	9.600000	5.000000
50%	37.000000	0.996600	3.320000	0.610000	10.100000	6.000000
75%	56.000000	0.997600	3.410000	0.700000	11.000000	6.000000
max	122.000000	1.001000	3.680000	0.980000	13.400000	7.000000

Fonte: os autores.

3. STORYTELLING

O gráfico utilizado para visualizar a relação entre cada variável é o pairplot (Figura 6), por ser extenso é incluído no trabalho. No entanto, para melhor visualização por parte do usuário recomenda-se o acesso à versão disponível do Colab, além disso partes fundamentais do trabalho encontram-se aqui.

Figura 6 - Pairplot em relação a qualidade.



Fonte: os autores.

3.1. Pairplot

O Pairplot é um gráfico que mostra como cada variável se relaciona com relação a qualidade, exemplo disso é o gráfico álcool x sulfato, é possível visualizar várias “bolinhas”, que advém do gráfico scatter, cada ponto indica um valor, como era feito para montar os gráficos de funções no papel, e dependendo da cor, ele indica o valor da qualidade naquele ponto. Isso é interessante, tendo em vista que é possível enxergar como as variáveis se comportam.

3.2. Conclusões dos gráficos

Para uma introdução à construção do pairplot, foi realizada uma agregação dos dados por meio da função groupby com qualidade, e então, calculando-se as médias das demais variáveis conforme demonstrado na tabela 6. Essa análise permitiu observar que, em média, o aumento do teor alcoólico está associado a uma elevação na qualidade do vinho. Da mesma forma, verificou-se que os níveis de sulfato, ácido cítrico e acidez fixa também aumentam. Já os ácidos voláteis, densidade, cloridratos e dióxido de enxofre total diminuem enquanto a qualidade aumenta.

Tabela 6 - Média dos valores.

Variável	Qualidade Baixa (3-4)	Qualidade Média (5-6)	Qualidade Alta (7-8)
fixed acidity	8.023782	8.189558	8.616296
volatile acidity	0.566930	0.500341	0.403852
citric acid	0.220799	0.251044	0.348963
residual sugar	2.192885	2.175000	2.208889
chlorides	0.081374	0.077187	0.072763
free sulfur dioxide	5.280702	15.548193	13.325926
total sulfur dioxide	49.276803	38.945783	30.422222
density	0.996905	0.996456	0.995772
pH	3.825614	3.327771	3.298963
sulphates	0.589220	0.653815	0.729780
alcohol	9.892008	10.555388	11.424815

Fonte: os autores.

O álcool em maior quantidade no vinho é o etílico que é doce e intensifica o gosto dos açucares, ou seja, é um fator muito importante uma vez que forma o equilíbrio gustativo. Já o gosto ácido causa um frescor, mas tem que ser bem dosado, uma vez que pode causar amargor. Ácidos fixos são uma "classe" de ácidos, sendo eles: tartárico, málico, cítrico, succínico e lático; de forma geral, trazem frescor, um gosto amanteigado, verde e uma leve adstringência.

Muitos desses fatores advêm da fermentação, o que demanda tempo, por isso existe a famosa frase: "envelheceu como vinho", pois a todo tempo o vinho fermenta, ou seja, sua qualidade "aumenta", o vinho fica melhor.

Observando o pairplot, é possível notar que as variáveis que realmente interferem na qualidade do vinho tinto são o sulfato e o álcool em algum dos eixos, as outras variáveis quando se relacionam, se dispersam, então é possível concluir que a harmonização dos açucares, causado pelo álcool, em conjunto com os sulfatos que trazem o gosto adstringente, causam uma maior qualidade do vinho.

4. AVALIAÇÃO DA INTERFACE DO REPOSITÓRIO UC IRVINE MACHINE LEARNING

4.1. Método de avaliação

Para realizar análise da interface quanto a usabilidade, foi utilizado o método de avaliação heurística que consiste em uma verificação sistemática da interface com foco na identificação de violações das diretrizes estabelecidas por Jakob Nielsen (1994) e para avaliar a acessibilidade o conjunto de diretrizes da WCAG (Web Content Accessibility Guidelines) foi utilizado como referência.

4.2. Resultado da avaliação heurística de usabilidade

Foram identificadas três violações das heurísticas de usabilidade, que serão explicadas no decorrer do trabalho.

4.2.1. Violação da Heurística de visibilidade do status do sistema

Utilizar a barra de busca, não há indicação visual de que a ação foi processada. A página simplesmente é congelada antes de exibir os resultados.

- **Local na interface**: Tela inicial, canto superior direito.

- **Justificativa**: A ausência de um *feedback* visual pode levar o usuário a interpretações equivocadas, como acreditar que o sistema travou ou que seu clique não foi registrado, potencialmente resultando em múltiplos cliques desnecessários.
- **Recomendação:** Implementar um indicador de carregamento, como uma barra de progresso ou um *spinner*.
- **Gravidade:** 3 - Problema Maior (Escala de 0 a 4).

4.2.2. Violação da heurística de consistência e padrões

O site emprega termos distintos para a mesma ação: "contribuir com um conjunto de dados" no topo da página e "doe um conjunto de dados" no rodapé.

- **Local na interface**: Topo e rodapé da tela inicial.
- **Justificativa:** A inconsistência terminológica pode induzir o usuário a crer que se trata de funcionalidades diferentes, causando confusão e perda de eficiência na navegação.
- **Recomendação**: Padronizar a nomenclatura em toda a interface, utilizando o mesmo termo para a mesma ação, por exemplo, "Contribuir com um conjunto de dados".
- **Gravidade:** 2 - Problema menor.

4.2.3. Violação da heurística de ajuda e documentação

A plataforma não dispõe de uma seção de "Ajuda" ou "Perguntas Frequentes" (FAQ) de fácil acesso.

- **Local na interface**: Tela inicial.
- **Justificativa:** Usuários inexperientes ou não familiarizados com repositórios de dados podem encontrar dificuldades para utilizar a plataforma e compreender sua terminologia específica.
- **Recomendação:** Incluir um link claramente visível para uma seção de ajuda abrangente no cabeçalho do site.
- **Gravidade:** 2 - Problema menor.

4.3. Resultado da avaliação de acessibilidade

Conforme utilizado o site da ACHECKER e alguns outros princípios, foram identificadas três violações de acessibilidade, conforme explicado a seguir:

4.3.1 Violação da WCAG 1.1.1

Conteúdo Não Textual (Nível A): O conteúdo do site é principalmente textual, a ausência de formatos alternativos (áudios, visualizações interativas) pode limitar o acesso a usuários com diferentes necessidades.

4.3.2 Violação da WCAG 2.1.1

Teclado (Nível A): Não há suporte para navegação por teclado, que é essencial para usuários com deficiência motora que não podem usar o mouse para interagir.

4.3.3 Violação da WCAG 1.3.1

Informação e Relações (Nível A): Falta de descrição de imagens, o que prejudica os usuários que precisam utilizar leitores de tela.

5. LINGUAGEM, PLATAFORMA E BIBLIOTECA USADAS

Foi usada a linguagem Python na plataforma colab. As bibliotecas utilizadas podem ser vistas na primeira parte do Google Colab, organizado todas as bibliotecas em um só bloco, mas trazendo para cá, terá:

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from google.colab import drive
```

Pandas é a principal biblioteca utilizada para lidar com o dataset, com ela é possível abrir e realizar as primeiras análises. O Matplolib, bem como o seaborn, servem para os gráficos e o Google Colab é para estabelecer conexão com o drive.

6. CONCLUSÃO

O trabalho tinha o objetivo de realizar uma análise dos parâmetros físico-químicos associados à qualidade do vinho tinto, a partir de técnicas de mineração de dados e análise estatística. Com base no conjunto de dados sobre vinho e na literatura especializada, foi possível

identificar relações significativas entre determinadas variáveis e a nota de qualidade atribuída aos vinhos.

A análise estatística descritiva, aliada à detecção e tratamento de outliers, permitiu uma compreensão mais precisa do comportamento das variáveis envolvidas. Através da visualização de dados com pairplots e do cálculo de correlações, constatou-se que fatores como teor alcoólico e concentração de sulfatos apresentam impacto relevante na percepção de qualidade do vinho.

Esses elementos, relacionados ao equilíbrio do paladar, revelaram-se determinantes na construção sensorial da bebida. A utilização da linguagem Python, associada a bibliotecas como Pandas, Matplotlib e Seaborn, demonstrou ser eficaz para a manipulação e visualização dos dados, evidenciando o potencial da computação aplicada à resolução de problemas interdisciplinares.

Dessa forma, conclui-se que a integração entre ciência de dados e enologia pode contribuir significativamente para a modernização dos processos de avaliação da qualidade do vinho, promovendo maior objetividade, reprodutibilidade e eficiência, com impactos positivos para a indústria vinícola e para o campo da pesquisa acadêmica.

REFERÊNCIAS

ACHECKER. **Free Web Accessibility Checker – WCAG Audit**. Disponível em: https://achecker.ca/. Acesso em: 21 set. 2025.

BARBOSA, S. D. J.; SILVA, B. S. Interação Humano-Computador. Rio de Janeiro, Editora Campus,2010. NETTO, A. A. O. IHC - Engenharia Pedagógica. São Paulo, Editora Visual Books, 2010. SOMMERVILLE, Ian. Engenharia de software. 9. ed. São Paulo: Pearson Education do Brasil, 2011.

CORTEZ, P. et al. Modeling wine preferences by data mining from physicochemical properties. **Decision Support Systems**, v. 47, n. 4, p. 547–553, nov. 2009.

LEONARDI, M. **Como e porque degustar vinhos**. [s.l: s.n.]. Disponível em: https://missaosommelier.com.br/wp-content/uploads/2017/03/missao-sommelier.pdf. Acesso em: 25 maio. 2025.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. 6. ed. Rio de Janeiro: LTC, 2018.

RESERVA85. **Como o vinho é feito**. Disponível em: https://reserva85.com.br/vinho/como-vinho-e-feito. Acesso em: 25 maio. 2025.

TRIOLA, M. F. **Introdução à estatística**. 12. ed. Rio de Janeiro: LTC, 2018.