Análise Exploratória da Qualidade dos Vinhos Tinto com Base em Propriedades Físico-Químicas e Sensoriais

1st Marina Vasques Rodrigues
Dept. de TeleInformática
Universidade Federal do Ceará
Fortaleza, Brazil
marinavasq18@alu.ufc.br

2nd Fábio Gabriel Esteves Ivo Gomes
Dept. de TeleInformática
Universidade Federal do Ceará
Fortaleza, Brazil
fabiogabriel@alu.ufc.br

3rd Caio Vinícius Pessoa Freires
Dept. de TeleInformática
Universidade Federal do Ceará
Fortaleza, Brazil
caiopessoa145@gmail.com

4th Fábio Agostinho da Silva Nascimento Filho
Dept. de TeleInformática
Universidade Federal do Ceará
Fortaleza, Brazil
fabinhosnf@gmail.com

***Resumo*—Este trabalho apresenta uma análise exploratória do conjunto de dados "Wine Quality", focada em vinhos tintos. Foram avaliadas propriedades físico-químicas e sensoriais, considerando seus principais valores estatísticos. A análise inclui quatro abordagens: univariada incondicional, univariada condicional por classe, bivariada incondicional e multivariada incondicional. O objetivo é compreender e identificar relações entre os preditores e a qualidade dos vinhos tintos, fornecendo insights para possíveis modelos preditivos.**

***Index Terms*—análise exploratória de dados, vinhos tintos, estatística descritiva, boxplot, histogramas**

I. Introdução

A análise da qualidade de vinhos é uma área de interesse acadêmico e industrial, pois envolve diversos fatores físico-químicos e sensoriais que determinam o valor do produto final e aceitação do produto pelo consumidor. Entender das relações entre esses fatores é fundamental para a melhoria dos processos de vinificação e controle de qualidade.

Trabalhos anteriores ([1], [3]–[8]) mostram que atributos como acidez volátil, teor alcoólico e teor de sulfatos estão fortemente relacionados à avaliação sensorial de vinhos tintos e variáveis como densidade e dióxido de enxofre exibem correlação inversa com a qualidade. Entretanto, essas relações são complexas e multivariadas, o que torna necessário aplicar técnicas de análise capazes de sintetizar padrões e relações entre múltiplos preditores.

Neste contexto, o presente trabalho realiza uma análise exploratória focada em vinhos tintos, com o objetivo de identificar padrões estatísticos e correlações nas propriedades físico-químicas do produto. São empregadas quatro abordagens complementares: análise univariada incondicional, análise univariada condicional por classe de qualidade, análise bivariada incondicional e análise multivariada incondicional.

Dessa forma, o estudo contribui com uma análise estatística voltada à enologia, fornecendo uma base para trabalhos futuros de modelagem preditiva e otimização da qualidade de vinhos tintos.

II. Métodos

A. Descrição do Dataset

O dataset "Wine Quality"[1] contém 6.497 amostras, sendo 1.599 vinhos **tintos** e 4.898 vinhos **brancos**. As variáveis de entrada são fatores físico-químicos (como pH e densidade), enquanto a saída representa a avaliação sensorial, obtida pela média de pelo menos três especialistas, em uma escala de 0 a 10. Para esta análise, serão considerados apenas os dados referentes aos vinhos **tintos**, uma vez que pesquisas indicam que eles são os mais consumidos e preferidos pelos brasileiros. [2]

As 11 variáveis de entrada são:

1) **Acidez fixa (g/L):** ácidos naturais predominantes, influenciam frescor e aroma [3].
2) **Acidez volátil (g/L):** ácidos que evaporam facilmente, impactam sabor e aroma [4].
3) **Ácido cítrico (g/L):** presente em menor quantidade, equilibra acidez [3],[6].
4) **Açúcar residual (g/L):** açúcar restante após fermentação, influencia doçura [4].
5) **Cloretos:** teor de cloretos, maior próximo ao mar [5].
6) **Dióxido de enxofre livre (mg/L):** influência na preservação e estabilidade do vinho.[8]
7) **Dióxido de enxofre total (mg/L):** soma do livre e do ligado à outras moléculas.[8]
8) **Densidade (g/L):** concentração de ácidos, açúcares e outros compostos.
9) **pH:** nível de acidez total, influencia sabor e estabilidade.
10) **Sulfatos (g/L):** compostos químicos que contêm enxofre. Sua presença pode influenciar o sabor, a acidez e a longevidade do vinho. [9]

11) **Álcool (% vol):** teor alcoólico.

B. Análise Monovariada Incondicional

A análise monovariada incondicional avalia cada preditor X_d individualmente, usando todas as N observações. Os passos são:

1) Plotagem de histogramas (incondicional)
2) Cálculo da média μ_d:

$$\mu_d = \frac{1}{N}\sum_{i=1}^{N} X_{i,d}$$

3) Cálculo do desvio padrão populacional σ_d:

$$\sigma_d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_{i,d} - \mu_d)^2}$$

4) Cálculo da assimetria (skewness) γ_d:

$$\gamma_d = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_{i,d} - \mu_d)^3}{\sigma_d^3}$$

C. Análise Monovariada Condicional por Classe de Qualidade

Nesta etapa, foi feita a análise monovariada condicional, cujo objetivo é compreender o comportamento de cada variável de acordo com as diferentes classes de qualidade do vinho. Essa abordagem permite identificar padrões nas distribuições dos preditores que podem estar relacionados ao nível de qualidade percebido.

As classes qualitativas foram definidas a partir da variável quality, seguindo os seguintes intervalos:

- 0 a 4 — Ruim
- 5 — Regular
- 6 — Médio
- 7 — Bom
- 8 a 10 — Excelente

Para cada variável, foram calculadas estatísticas descritivas condicionadas a cada classe, permitindo observar tendências e variações importantes entre os grupos.

Inicialmente, todas as onze variáveis físico-químicas foram analisadas, mas, para fins de apresentação neste artigo, foram selecionados quatro preditores que se mostraram mais representativos:

- **Álcool** — fortemente correlacionado com a qualidade e indicador de corpo e sabor.
- **Acidez Volátil** — associada a defeitos de fermentação e aroma indesejado.
- **Ácido Cítrico** — relacionado à sensação de frescor e equilíbrio do vinho.
- **Sulfatos** — contribuem para a conservação e potencializam o sabor.

A escolha desses preditores foi motivada por sua relevância em estudos sobre vinhos (literatura enológica) e pela variação observada entre as classes, que os tornam bons indicadores da qualidade percebida. As demais variáveis e suas estatísticas detalhadas estão disponíveis no material suplementar.

D. Análise Bivariada Incondicional

Nesta etapa, foi realizada uma **análise bivariada incondicional** com o objetivo de identificar a relação entre os preditores físico-químicos do vinho tinto. Essa análise permite investigar o grau de associação entre duas variáveis numéricas, fornecendo indícios de possíveis colinearidades e interdependências que podem influenciar a variável resposta (*quality*).

1) Correlação linear de Pearson: A medida utilizada para quantificar o grau de associação linear entre duas variáveis X e Y foi o **coeficiente de correlação de Pearson** (r_{xy}), definido como:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

onde:

- x_i e y_i são os valores observados das variáveis X e Y;
- $\bar{x}$ e $\bar{y}$ são as respectivas médias amostrais;
- n representa o número de observações.

O valor de r_{xy} varia entre -1 e 1, indicando:

- $r_{xy} \approx 1$: forte correlação linear positiva (as variáveis aumentam juntas);
- $r_{xy} \approx -1$: forte correlação linear negativa (uma aumenta enquanto a outra diminui);
- $r_{xy} \approx 0$: ausência de correlação linear significativa.

2) Procedimentos de análise: Para o conjunto de dados do vinho tinto (*red wine*), foram consideradas as 11 variáveis físico-químicas disponíveis, excluindo-se a variável resposta *quality*. Assim, foram avaliadas todas as possíveis combinações de pares de preditores, totalizando:

$$\frac{11 \times 10}{2} = 55 \text{ pares distintos.}$$

Além disso, foram gerados gráficos de dispersão (*scatter plots*) para cada par de variáveis, com o objetivo de visualizar o padrão de relacionamento entre elas. Essa visualização auxilia na identificação de possíveis relações lineares, outliers e agrupamentos.

3) Critério de interpretação: Para fins interpretativos, adotaram-se as faixas de intensidade de correlação linear indicadas por Dancey e Reidy (2006):

Valor de $\lvert r\rvert$	**Interpretação**
$0,00 \leq \lvert r\rvert < 0,10$	Correlação desprezível
$0,10 \leq \lvert r\rvert < 0,30$	Correlação fraca
$0,30 \leq \lvert r\rvert < 0,50$	Correlação moderada
$0,50 \leq \lvert r\rvert < 0,70$	Correlação forte
$\lvert r\rvert \geq 0,70$	Correlação muito forte

Esses critérios foram utilizados na seção de Resultados para destacar as relações mais relevantes entre os preditores físico-químicos.

E. Análise Multivariada Incondicional

Nesta etapa, foi realizada a análise multivariada incondicional utilizando a técnica de Principal Component Analysis (PCA), implementada manualmente, sem o uso de funções

pré-existentes. O objetivo é reduzir a dimensionalidade do conjunto de dados, preservando o máximo possível da variância original e permitindo a visualização das relações entre as amostras.

O algoritmo foi desenvolvido conforme os seguintes passos:

1) Padronização dos dados:
   Cada variável foi centralizada pela média e normalizada pelo desvio-padrão, assegurando que todas as características contribuíssem de forma equitativa.
2) Cálculo da matriz de covariância:
   Avaliou-se a variabilidade conjunta entre os preditores.
3) Decomposição:
   Foram obtidos autovalores e autovetores da matriz de covariância.
4) Ordenação dos componentes principais:
   Os autovetores foram ordenados conforme seus autovalores (variâncias explicadas).
5) Projeção dos dados:
   As observações foram projetadas sobre os dois primeiros componentes principais.

III. Resultados

A. Descrição do Dataset

A Tabela I apresenta um resumo estatístico das variáveis do dataset de vinhos tintos. Esses valores indicam diferenças importantes nas características físico-químicas dos vinhos, que podem influenciar diretamente a avaliação de sua qualidade. Por exemplo, a média da qualidade dos vinhos tintos é 5,64, com valor máximo de 8,0, demonstrando que há uma baixa variação entre as avaliações.

Tabela I: Resumo Estatístico dos Preditores do Vinho

Variáveis	count	mean	std	min	25%	50%	75%	max
fixed acidity	1599	8.32	1.74	4.60	7.10	7.90	9.20	15.90
volatile acidity	1599	0.53	0.18	0.12	0.39	0.52	0.64	1.58
citric acid	1599	0.27	0.19	0.00	0.09	0.26	0.42	1.00
residual sugar	1599	2.54	1.41	0.90	1.90	2.20	2.60	15.50
chlorides	1599	0.09	0.05	0.01	0.07	0.08	0.09	0.61
free sulfur dioxide	1599	15.87	10.46	1.00	7.00	14.00	21.00	72.00
total sulfur dioxide	1599	46.47	32.89	6.00	22.00	38.00	62.00	289.00
density	1599	2.21	0.97	0.99	0.996	0.997	0.998	1.004
pH	1599	3.31	0.15	2.74	3.21	3.31	3.40	4.01
sulphates	1599	0.66	0.17	0.33	0.55	0.62	0.73	2.00
alcohol	1599	10.40	1.16	1.00	9.50	10.20	11.10	14.90
quality	1599	5.64	0.81	3.00	5.00	6.00	6.00	8.00

B. Análise Monovariada Incondicional

A análise monovariada incondicional permite estudar estatisticamente cada preditor individualmente. Para os vinhos tintos, observamos os histogramas de frequência de alguns preditores, valores de média, desvio padrão e assimetria apresentados na Figura 1 e na Tabela II.

A partir desses resultados, podemos destacar:

1) **Média**: Os preditores com maiores valores médios nos vinhos tintos são o *Dióxido de Enxofre Livre* e o *Dióxido de Enxofre Total*, indicando que, em geral, esses atributos apresentam níveis altos nos vinhos tintos.
2) **Desvio Padrão**: O *Dióxido de Enxofre Livre* e o *Dióxido de Enxofre Total* apresentam desvio padrão elevado, o que indica alta variação entre os valores desses preditores.
3) **Assimetria**: Os preditores *Densidade*, *Cloretos* e *Açucar Residual* apresentam as maiores assimetrias, sugerindo a presença de alguns vinhos com os valores, para esses preditores, significativamente menores que a média.

Tabela II: Média, Desvio Padrão e Assimetria dos Preditores

Preditor	Média (μ)	Desvio Padrão (σ)	Assimetria (γ)
Ácido fixo	8.3196	1.7411	0.9828
Acidez volátil	0.5278	0.1791	0.6716
Ácido cítrico	0.2710	0.1948	0.3183
Açúcar residual	2.5388	1.4099	4.5407
Cloretos	0.0875	0.0471	5.6803
Dióxido de enxofre livre	15.8749	10.4602	1.2506
Dióxido de enxofre total	46.4678	32.8953	1.5155
Densidade	2.2087	9.6641	9.8039
pH	3.3111	0.1544	0.1937
Sulfatos	0.6581	0.1695	2.4287
Álcool	10.3978	1.1599	-0.5768

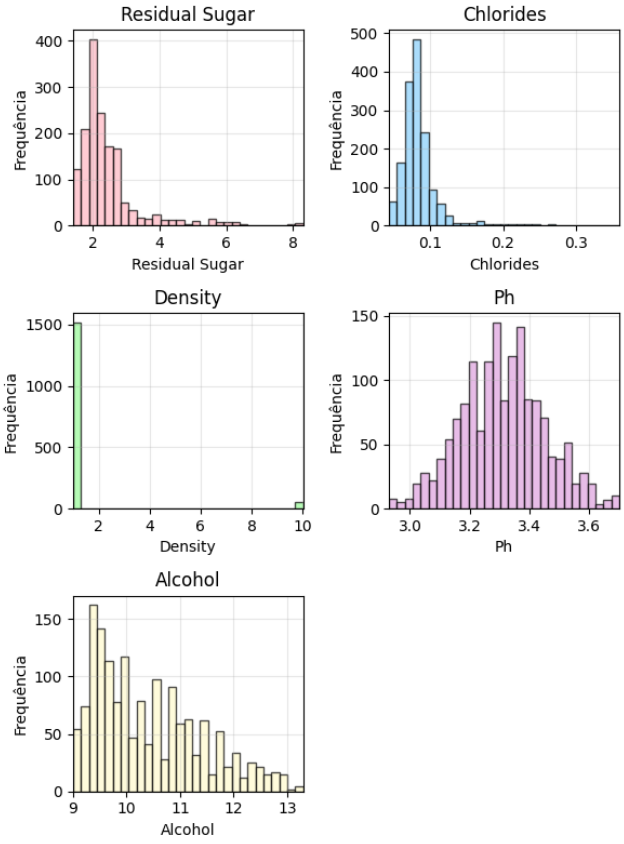

Figura 1: Distribuição do teor alcoólico por classe de qualidade.

Baseado nos histogramas e nos valores de assimetria, pode-se destacar as seguintes afirmações sobre os seguintes preditores:

Açúcar Residual: O histograma tem maior frequência em valores muito baixos, entre 0 e 1, caracterizando vinhos predominantemente secos (até 4 g de açúcar por litro, na legislação brasileira) [10]. A assimetria confirma que há poucos vinhos

tintos com açúcar residual mais elevado.

Cloretos: Distribuição concentrada e próxima a zero, com assimetria alta, indicando que a maior parte dos vinhos tem valores de cloreto menores que a média.

Densidade: Distribuição concentrada em valores próximos de zero, com assimetria alta, evidenciando que grande maioria dos vinhos tem densidade abaixo da média, mas há alguns poucos com densidade muito alta.

pH: Distribuição aproximadamente normal, centralizada entre 3 e 3,5, com assimetria baixa, ou seja, a maioria dos dados está concentrada próximo à média.

Álcool: Distribuição concentrada entre 9 e 10%, com assimetria negativa, indicando teor alcoólico maior que a média na maioria dos vinhos.

C. Análise Monovariada Condicional por Classe de Qualidade

Determinadas variáveis exibem padrões distintos entre as categorias de qualidade, sugerindo uma relação direta com a qualidade final do vinho.

Álcool: O teor alcoólico apresentou aumento contínuo com a melhora da qualidade (Figura 2). Essa tendência indica que vinhos mais alcoólicos tendem a ser avaliados como de melhor qualidade, possivelmente por apresentarem corpo mais robusto e melhor equilíbrio entre acidez e sabor.

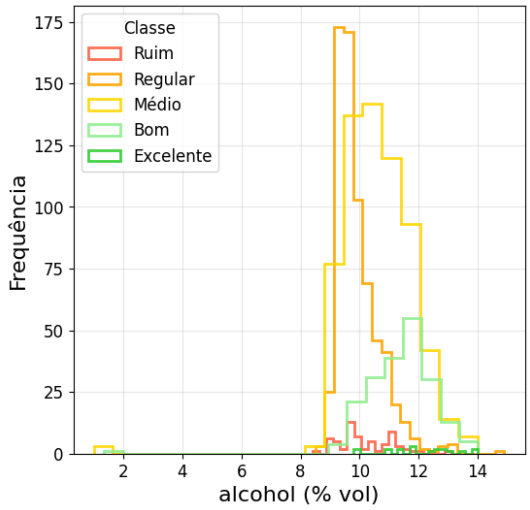

Figura 2: Distribuição do teor alcoólico por classe de qualidade.

Acidez Volátil: A acidez volátil mostrou uma tendência contratia, diminuindo conforme a qualidade aumenta (Figura 3). Vinhos com menor acidez volátil são percebidos como mais suaves, pois altos níveis desse componente estão relacionados a aromas desagradáveis provenientes de processos fermentativos inadequados.

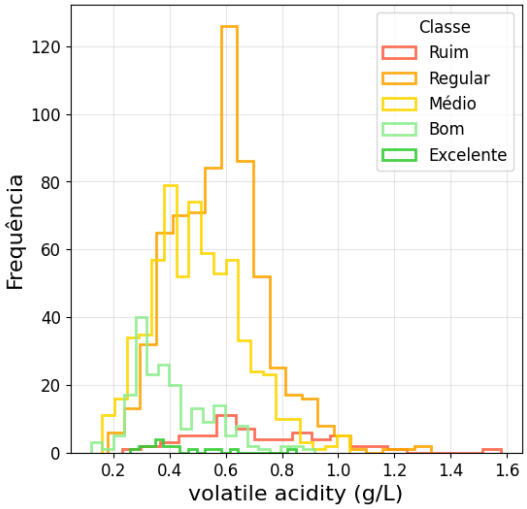

Figura 3: Distribuição da acidez volátil por classe de qualidade.

Ácido Cítrico: Aqui conseguimos observar o aumento gradual do teor de ácido cítrico nas classes de maior qualidade (Figura 4). Esse comportamento sugere que o equilíbrio entre acidez e frescor é um fator determinante para se ter um vinho de qualidade.

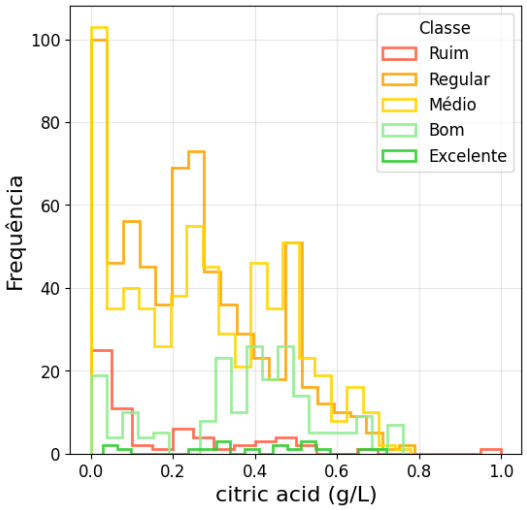

Figura 4: Distribuição do ácido cítrico por classe de qualidade.

Sulfatos: Os níveis de sulfatos também aumentaram gradualmente com a qualidade (Figura 5). Isso está de acordo com o fato de que o enxofre auxilia na preservação e pode estar associado a vinhos mais bem conservados e equilibrados.

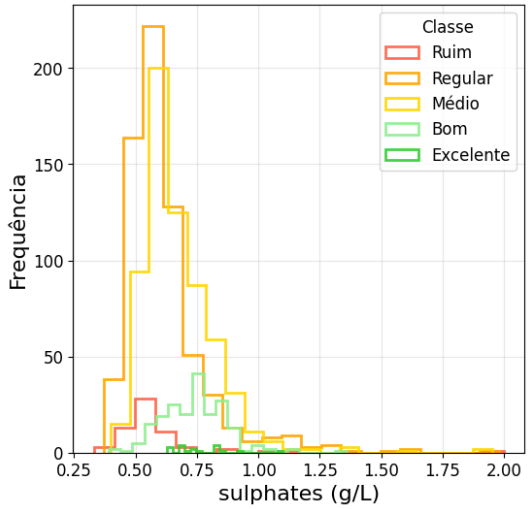

Figura 5: Distribuição dos teores de sulfatos por classe de qualidade.

No geral, esses quatro preditores mostraram os padrões mais consistentes e que contém mais informação para a diferenciação entre as classes. As tabelas com as estatísticas completas (média, desvio padrão e assimetria) de todos os preditores podem ser consultadas no *notebook* disponível no Colab, referenciado ao final deste trabalho, onde também constam os gráficos adicionais das demais variáveis.

D. Análise Bivariada Incondicional

1) Gráficos de Dispersão: A Figura 6 apresenta os gráficos de dispersão (*scatter plots*) gerados para alguns pares de variáveis físico-químicas do vinho tinto. Essa visualização permite identificar padrões lineares ou não lineares, agrupamentos e possíveis *outliers* nas variáveis.

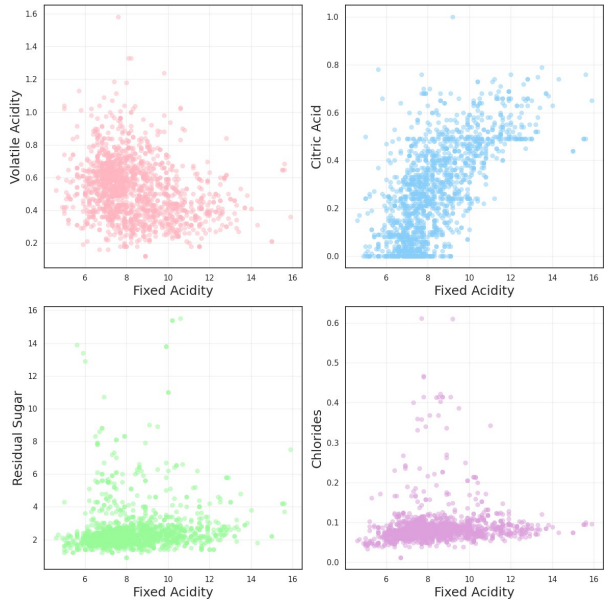

Figura 6: Gráficos de dispersão entre os pares de variáveis físico-químicas do vinho tinto.

2) Matriz de Correlação: Para quantificar as relações lineares entre os preditores, foi calculada a matriz de correlação de Pearson (Figura 7). Observa-se que algumas variáveis possuem correlação forte, como *free sulfur dioxide* e *total sulfur dioxide* ($r = 0.67$), enquanto outras apresentam correlação negativa, como *fixed acidity* e *pH* ($r = -0.68$).

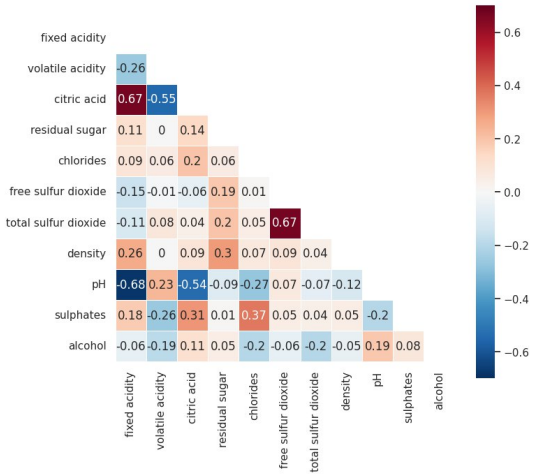

Figura 7: Gráficos de dispersão entre os pares de variáveis físico-químicas do vinho tinto.

Observa-se, de maneira geral, que:

- Variáveis relacionadas ao enxofre (*free sulfur dioxide* e *total sulfur dioxide*) apresentam forte correlação positiva.
- Algumas variáveis químicas possuem correlação negativa, como *fixed acidity* e *pH*, indicando que vinhos mais ácidos tendem a ter pH menor.
- A maioria das demais variáveis apresenta correlações fracas ou moderadas, sugerindo relações menos lineares ou mais complexas entre si.

E. Análise Multivariada Incondicional

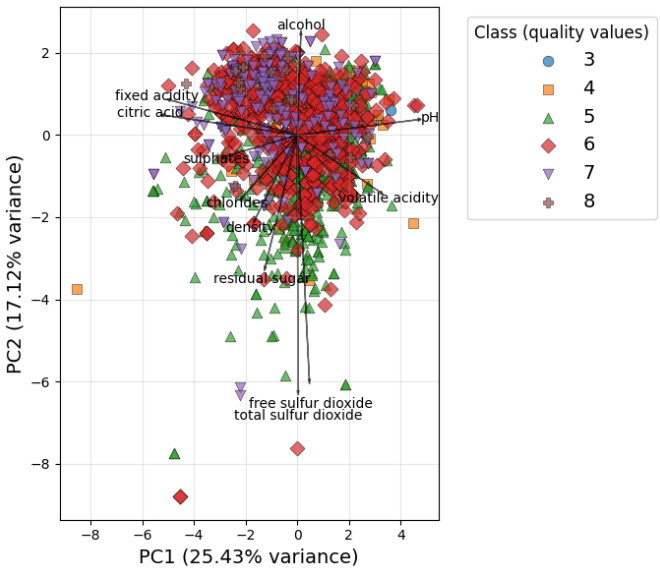

Figura 8: Projeção do PCA - Qualidade do vinho tinto

A Figura 8 apresenta o gráfico de dispersão das amostras projetadas sobre os dois primeiros componentes principais, as setas pretas (vetores de carga) indicam a contribuição de cada variável original na formação desses componentes.

Na figura, o primeiro Componente Principal (PC1) explica 25,43% da variância total e o segundo Componente Principal (PC2) explica 17,12%, resultando em aproximadamente 42,55% da variabilidade sendo preservada. Esse valor representa menos da metade da variância, mas permite uma visualização da estrutura geral dos dados.

No gráfico de dispersão, as amostras estão pintadas com base na classe de qualidade do vinho. As classes estão em grande parte agrupadas, indicando que as variáveis físico-químicas, sozinhas, não são suficientes para uma categorização clara entre os diferentes níveis de qualidade. Essa sobreposição de classes indica que a qualidade do vinho é de natureza complexa e envolve diversos fatores.

A distribuição dos vetores de carga mostra que variáveis relacionadas estão orientadas em direções semelhantes, mostrando que existe redundância em alguns atributos. Exemplos são a concentração de ácido cítrico e acidez fixa sendo proximamente relacionados, indicando que o ácido cítrico tem uma influência considerável na acidez total da bebida. Variáveis inversamente relacionadas apontam em direções diferentes, exemplo é o volume de álcool e concentração de dióxido de enxofre. O PCA representou essas correlações em um espaço bidimensional, destacando os preditores mais relevantes na variação do conjunto.

Dessarte, os resultados indicam que, apesar de o PCA ser útil por diminuir a dimensionalidade dos dados, mostrar padrões de relações entre variáveis e redundâncias, a predição da qualidade do vinho exige outros métodos capazes de capturar relações não lineares entre os preditores.

IV. Referências

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.

[2] PORTAL INSIGHTS. Qual o vinho preferido dos brasileiros? Disponível em: ¡https://www.portalinsights.com.br/perguntas-frequentes/qual-o-vinho-preferido-dos-brasileiros¿. Acesso em: 19 out. 2025.

[3] Caveroyale, "Ácido Cítrico: Importância e Aplicações em Vinhos Premium," [Online]. Available: https://www.caveroyale.com.br/glossario/acido-citrico-importancia-aplicacoes-vinhos-premium/, acesso em: 28 set. 2025.

[4] Caveroyale, "Acidez Volátil: Entenda seu Impacto nos Vinhos Premium," [Online]. Available: https://www.caveroyale.com.br/glossario/acidez-volatil-vinhos-premium/, acesso em: 28 set. 2025.

[5] Agrovin, "Técnicas para corrigir a acidez do vinho," [Online]. Available: https://agrovin.com/pt-pt/tecnicas-para-corrigir-a-acidez-do-vinho/, acesso em: 28 set. 2025.

[6] Embrapa, "Metodologia de Análise de Vinho Tinto," [Online]. Available: https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/887323/1/Metodologiaanalisevinhotintoed012010.pdf, acesso em: 28 set. 2025.

[7] Famiglia Valduga, "A importância da acidez no vinho," [Online]. Available: https://blog.famigliavalduga.com.br/qual-a-importancia-da-acidez-no-vinho/, acesso em: 28 set. 2025.

[8] Duarte, I., "Dióxido de enxofre na vinificação: Importância e Funções," [Online]. Available: https://www.caveroyale.com.br/glossario/dioxido-de-enxofre-na-vinificacao-importancia-e-funcoes/, acesso em: 29 set. 2024.

[9] Caveroyale, "Sulfatos no vinho"[Online].Available: https://www.caveroyale.com.br/glossario/sulfatos-no-vinho-importancia-impacto/

[10] Acionista, "O açúcar residual no vinho é vilão ou aliado?," [Online]. Available: https://acionista.com.br/o-acucar-residual-no-vinho-e-vilao-ou-aliado/, acesso em: 18 out. 2025.