

**Universidade de Brasília**

**Instituto de Química**

## **Projeto Final de Quimiometria**

### **Análise de qualidade dos vinhos na região do Minho em Portugal**

**Aluno:** Lucas Fernandes Aguiar

**Matrícula FUB:** 242115910

**Professor:** Dr. Jez William Batista Braga

**Disciplina:** Quimiometria (PPGCF3744)

**Data:** 05 de dezembro de 2025

## **1. Introdução**

A indústria vinícola enfrenta o desafio constante de avaliar e garantir a qualidade de seus produtos. Tradicionalmente, a avaliação da qualidade do vinho depende de análises sensoriais realizadas por especialistas (sommeliers), um processo subjetivo, demorado e custoso. Alternativamente, análises físico-químicas podem fornecer informações objetivas sobre a composição do vinho, mas a interpretação dessas múltiplas variáveis simultaneamente requer ferramentas estatísticas avançadas.

A quimiometria oferece um conjunto de técnicas matemáticas e estatísticas que permitem extrair informações relevantes de dados químicos multivariados. Entre essas técnicas, destacam-se a Análise de Componentes Principais (PCA) para exploração de dados, a Análise Hierárquica de Agrupamentos (HCA) para identificação de padrões naturais e a classificação supervisionada por meio da regressão por quadradros parciais (PLS).<sup>1</sup>

A PCA é uma técnica de redução de dimensionalidade que transforma variáveis originais correlacionadas em componentes principais ortogonais, retendo a máxima variância dos dados e facilitando a visualização de amostras em gráficos de escores e a identificação de variáveis importantes por meio dos loadings (pesos) (Cortez et al., 2009a). A HCA agrupa amostras com base em suas similaridades, construindo dendrogramas que revelam a estrutura hierárquica dos dados sem necessidade de conhecimento prévio das classes, sendo útil para detectar outliers e padrões naturais de agrupamento (Everitt et al., 2011). Já a PLS é um método de regressão multivariada que estabelece relações entre blocos de variáveis preditoras e respostas, maximizando a covariância entre eles, sendo amplamente utilizada em calibração multivariada e modelos de classificação quando há alta colinearidade entre as variáveis independentes (Wold; Sjöström; Eriksson, 2001).

O estudo de qualidade realizado por Cortez et al. coletou amostras de vinhos na região do Minho em Portugal no período entre maio de 2004 a fevereiro de 2007 (Cortez et al., 2009a). Os dados foram disponibilizados sob o título “Wine Quality” no repositório público UCI Machine Learning, contendo dados físico-químicos de vinhos portugueses da região de “Vinho Verde”. Este conjunto de dados inclui 11 variáveis físico-químicas e uma variável de qualidade determinada sensorialmente (Cortez et al., 2009b).

---

<sup>1</sup>PCA = Principal Component Analysis; HCA = Hierarchical Clustering Analysis; PLS = Partial Least Squares.

## 2. Objetivo

O objetivo deste projeto é aplicar técnicas quimiométricas de reconhecimento de padrões e classificação para:

1. Explorar a estrutura dos dados através de PCA e HCA
2. Identificar as variáveis físico-químicas mais relevantes para a qualidade do vinho
3. Desenvolver modelos de classificação (PLS-DA) para discriminar vinhos de diferentes qualidades.<sup>2</sup>
4. Avaliar o desempenho dos modelos através de validação adequada (matriz de confusão)

## 3. Metodologia

### 3.1. Coleta de Dados

O estudo considerou amostras de vinho verde (tinto e branco), um produto exclusivo da região do Minho, em Portugal. Os dados foram coletados de maio de 2004 a fevereiro de 2007. A coleta se restringiu a amostras com Denominação de Origem Protegida (DOP) que foram testadas na entidade oficial de certificação (CVRVV).

Foi considerado um grande conjunto de dados em comparação com estudos anteriores neste domínio, totalizando 4.898 amostras de vinho branco e 1.599 amostras de vinho tinto.

Apenas os testes físico-químicos mais comuns, disponíveis na fase de certificação, foram selecionados. Esses testes incluíam 11 atributos mensuráveis, como densidade, álcool, pH, acidez fixa, acidez volátil, ácido cítrico, açúcar residual, cloretos, dióxido de enxofre livre, dióxido de enxofre total e sulfatos (ver Tabela 1).

As preferências sensoriais foram modeladas em uma abordagem de regressão que preserva a ordem das notas. Cada amostra foi avaliada por um mínimo de três avaliadores sensoriais (em degustações cegas). O vinho foi classificado numa escala de 0 (muito ruim) a 10 (excelente). A pontuação sensorial final utilizada no estudo foi dada pela mediana dessas avaliações.

### 3.2. Instalação das bibliotecas necessárias

A análise dos dados coletados foi realizada utilizando a linguagem python e suas bibliotecas. Para reproduzir o ambiente é necessário a instalação das seguintes bibliotecas:

bash

```
pip install pandas numpy matplotlib seaborn scikit-learn scipy ucimlrepo
```

### 3.3. Carregamento dos dados

Os dados disponibilizados no repositório da ucimlrepo já possui as tabelas necessárias.

python

```
from ucimlrepo import fetch_ucirepo
wine_quality = fetch_ucirepo(id=186)
X = wine_quality.data.features
y = wine_quality.data.targets

# Adicionar coluna de tipo de vinho
df_red['wine_type'] = 'red'
```

<sup>2</sup>DA = Discriminant Analysis.

```

df_white['wine_type'] = 'white'

# Combinação datasets e um dataframe
df = pd.concat([df_red, df_white], ignore_index=True)

# Estatística descritiva para os dados
df.describe().round(3)

```

### 3.4. Descrição das variáveis

Tabela 1: Descrição das variáveis encontradas nos conjuntos de dados e respectiva unidade.

Variável	Descrição	Unidade
fixed acidity	Ácidos não voláteis (tartárico)	g/mL
volatile acidity	Ácido acético	g/mL
citric acid	Ácido cítrico	g/mL
residual sugar	Açúcar residual após fermentação	g/mL
chlorides	Cloreto de sódio	g/mL
free sulfur dioxide	SO <sub>2</sub> livre	mg/mL
total sulfur dioxide	SO <sub>2</sub> total	mg/mL
density	Densidade	g/cm <sup>3</sup>
pH	pH	-
sulphates	Sulfato de potássio	g/mL
alcohol	Teor alcoólico	% vol
quality	Qualidade (avaliação sensorial)	0-10
wine_type	Tipo de vinho	tinto/branco

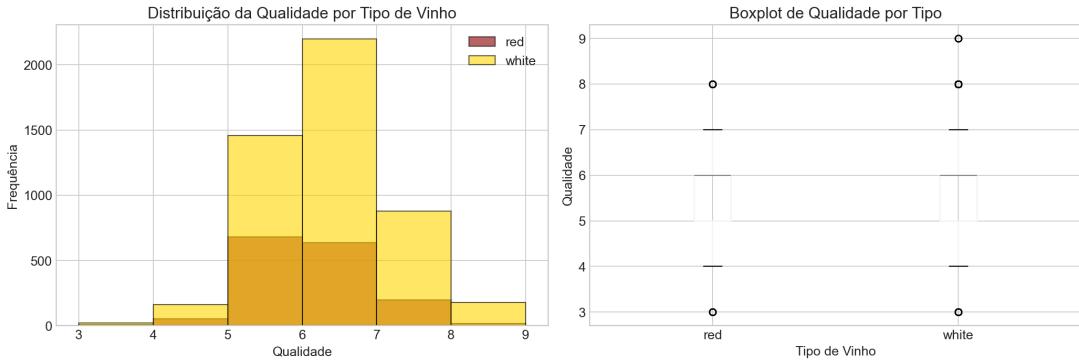
### 3.5. Pré-processamento dos dados

Essa etapa consiste em verificar se existem valores ausentes, duplicatas, fazer a separação de variáveis, criação de classes simplicadas para realizar classificação, a normalização dos dados. Na fase de pré-processamento, a base de dados foi transformada para que cada linha incluisse uma amostra distinta de vinho com todos os testes realizados. Devido às diferenças significativas no sabor, a análise foi realizada separadamente para os vinhos tinto e branco. Antes de ajustar os modelos, os dados foram padronizados para média zero e desvio padrão um.

## 4. Resultados e Discussão

### 4.1. Análise Exploratória dos dados

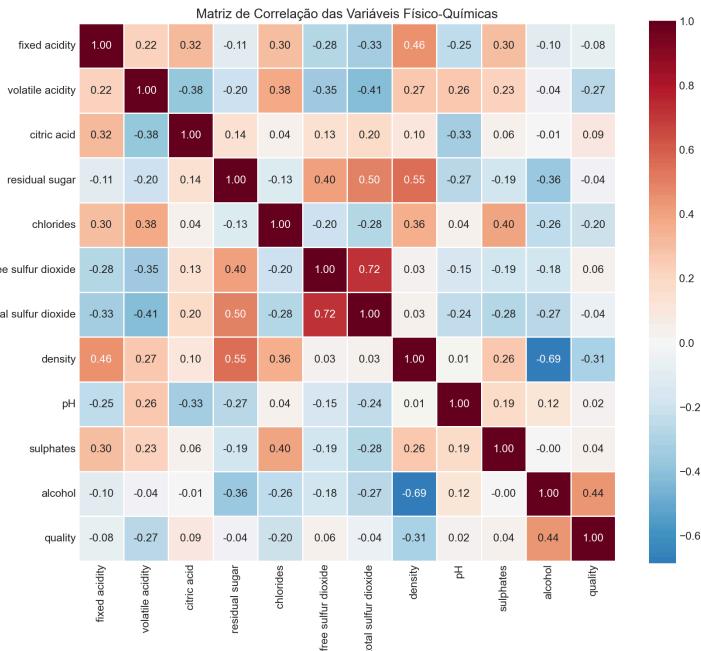
Com os dados pré-processados, podemos iniciar a exploração dos dados a fim de buscar identificar padrões.



**Figura 1:** **Esquerda:** Histograma com frequência de vinhos tintos e brancos de acordo com a avaliação de qualidade. **Direita:** Gráficos de boxplot com a distribuição das avaliações por tipo de vinho.

Os gráficos utilizados na Figura 1 não são capazes de demonstrar padrões claros que diferenciem a qualidade de acordo com o tipo de vinho.

Expandindo essa primeira análise, é possível através de uma matriz de correlação verificar as respostas de todos os fatores em uma única figura.



**Figura 2:** Matriz de correlação.

A matriz na Figura 2 é capaz de exibir um fator que sobressaiu sobre os demais na avaliação de qualidade, o teor alcoólico com uma correlação de 0,44. As demais variáveis não apresentaram um valores de correlação relevantes.

## 4.2. Análise de Componentes Principais (PCA)

Matematicamente, dada uma matriz de dados  $X$  com dimensões  $n \times p$  (onde  $n$  é o número de amostras e  $p$  é o número de variáveis), o PCA busca encontrar novas direções ortogonais que maximizam a variância projetada. A primeira componente principal (PC1) é dada por:

$$t_1 = Xp_1$$

onde  $p_1$  é o vetor de **loadings** (pesos) que maximiza a variância de  $t_1$  sujeito à restrição  $\|p_1\| = 1$ . Para mais componentes principais, haverá uma matriz de pesos  $P$  que informará a contribuição de cada variável em cada PC, de forma que a matriz  $X$  pode ser expressa como:

$$X = TP^T + E$$

#### 4.2.1. Algoritmo NIPALS

O algoritmo NIPALS é um método iterativo para calcular componentes principais, especialmente útil para matrizes de grande dimensão ou com valores ausentes.<sup>3</sup> Diferentemente da decomposição SVD tradicional, o NIPALS calcula um componente por vez através de um processo iterativo.<sup>4</sup> O NIPALS é recomendado nos casos em que matrizes são grandes e quando não é necessário o cálculo de todos os componentes.

#### 4.2.2. Seleção do Número de Componentes

Diversos critérios podem ser utilizados para determinar o número ótimo de componentes:

1. **Variância Acumulada:** Selecionar componentes até atingir um limiar de variância explicada (tipicamente 80-95%)
2. **Scree Plot:** Identificar o “cotovelo” no gráfico de autovalores, onde a taxa de decrescimento diminui significativamente;
3. **Validação Cruzada:** Minimizar o erro de predição (RMSECV) em um modelo de calibração.<sup>5</sup>

Tabela 2: Tabela com os autovalores e autovetores dos componentes principais.

PC	Autova-lor ( $\lambda$ )	$\log(\lambda)$	% Variânc- cia	% Var. Acumu- lada	RMSEC	RMSECV	Selecio-nado
1	2.988937	1.094918	27.167050	27.167050	0.874140	0.874686	
2	2.475732	0.906536	22.502425	49.669476	0.827811	0.828479	
3	1.587581	0.462211	14.429840	64.099315	0.803316	0.804072	✓
4	0.953581	-0.047531	8.667287	72.766602	0.787194	0.788547	
5	0.742375	-0.297901	6.747594	79.514196	0.763832	0.764970	
6	0.627156	-0.466560	5.700345	85.214541	0.752947	0.754181	
7	0.521138	-0.651741	4.736727	89.951268	0.749696	0.751001	
8	0.507694	-0.677876	4.614533	94.565801	0.748865	0.750367	
9	0.336784	-1.088313	3.061100	97.626900	0.732684	0.734560	
10	0.226001	-1.487217	2.054165	99.681066	0.732466	0.734651	

A escolha do número de componentes principais é baseado nos autovalores. Para isso, a seleção de componentes é feita baseada entre o critério de Kaiser (quando  $\lambda > 1$ ), de variância acumulada (95%), o scree plot (queda relativa de autovalores) e queda no log dos autovalores ( $\log(\lambda) > 0,1$ ). Os gráficos da Figura 3 evidenciam os diferentes métodos para seleção de componentes principais e é escolhido o critério de Kaiser, onde pois o PC4 possui autovalor que 1 ( $\lambda_{PC4} = 0.953581$ ).

<sup>3</sup>NIPALS = Nonlinear Iterative Partial Least Square.

<sup>4</sup>SVD = Single Value Decomposition, calcula todos os componentes principais de uma vez através da decomposição completa da matriz.

<sup>5</sup>RMSECV = Root mean square of error of cross-validation.

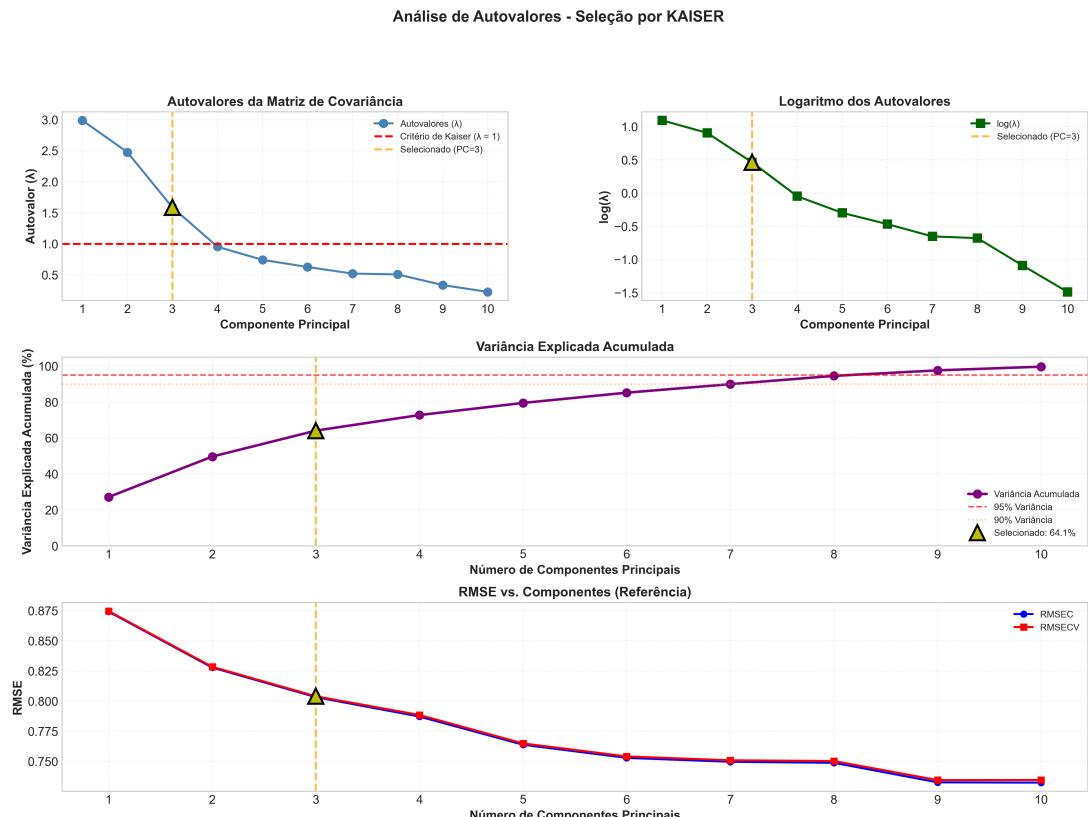


Figura 3: Gráficos com diferentes métodos de seleção de componentes principais. Método escolhido foi o critério de Kaiser (cima esquerda), com 3 componentes principais.

#### 4.2.3. Diagnóstico de Outliers

Para a identificação de outliers foi feito por meio do  $T^2$  de Hotteling (distância da amostra ao centro no espaço de escores), resíduo  $Q$  (variação não explicada) e leverage (influência individual de amostras no modelo).

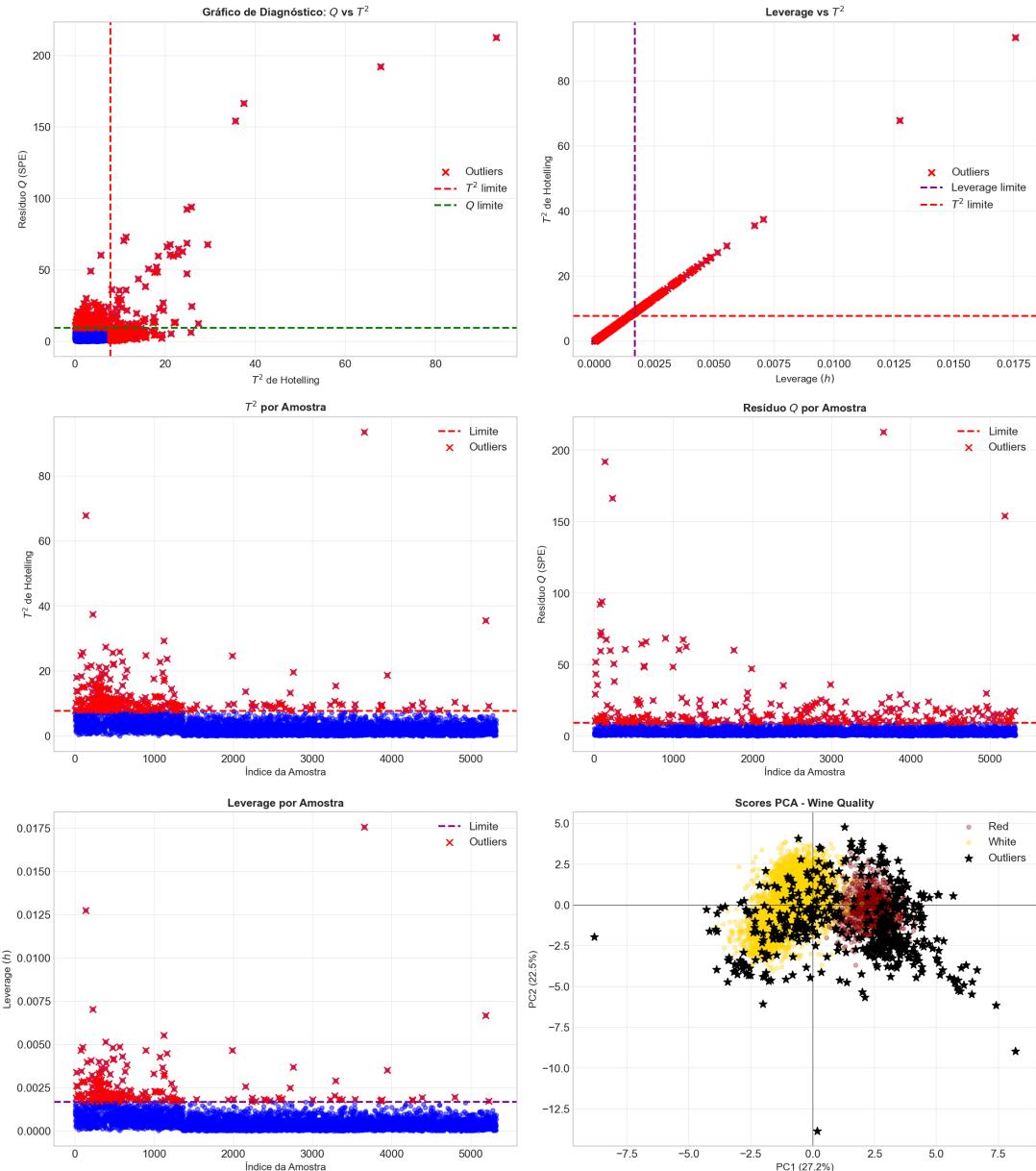


Figura 4: Diagnóstico de outliers para o modelo de PCA.

No entanto, essa escolha de apenas 3 componentes principais levou a situação de altos valores de resíduos  $Q$  e leverage, indicando que os outliers são influentes e problemáticos para o presente modelo.

### 4.3. Análise hierárquica de agrupamentos (HCA)

Para realizar a análise hierárquica de agrupamentos foram aplicadas diferentes métricas de distância em combinação com diferentes métodos de ligação.

#### 4.3.1. Métricas de distância

A partir da métrica de distância euclidiana foram gerados os dendogramas utilizando os seguintes métodos de ligação: simples (k vizinhos mais próximos, KNN), ligação completa, ligação por média, ligação por média ponderada, ligação centróide e método de Ward (ver Figura 5). Para cada uma dessas foi determinado a correlação cofenética. A correlação cofenética mede a correlação entre as distâncias originais entre pares de amostras e as distâncias apresentadas no dendrogramas, ou seja, indica o quanto bem o dendrograma representa os dados ( $\geq 0,85$  é uma excelente representação).

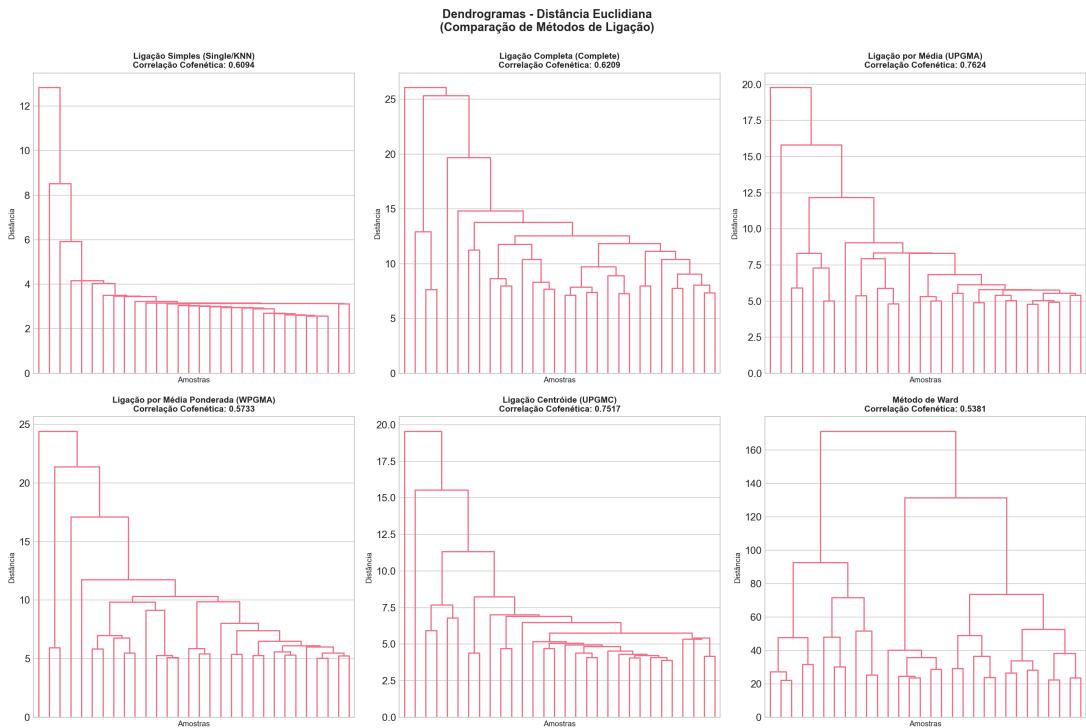


Figura 5: Dendogramas gerados a partir da métrica de distância euclidiana utilizando diferentes métodos de ligação.

A Tabela 3 a análise de todas as combinações realizadas, destacando o seu status a partir dos valores de correlação cofenética. O melhor resultado encontrado foi usando a distância de Mahalonobis com a ligação por média, exibindo uma correlação cofenética de 0,78 (classificado como “bom”).

Tabela 3: Análise de todas as combinações (Distância × Ligação)

Métrica	Método Linkage	Correlação Cofenética	Status
euclidean	single	0.6094	Fraco
euclidean	complete	0.6209	Fraco
euclidean	average	0.7624	Bom
euclidean	weighted	0.5733	Fraco
euclidean	centroid	0.7517	Bom
euclidean	ward	0.5381	Fraco
manhattan	single	0.6782	Moderado
manhattan	complete	0.4840	Fraco
manhattan	average	0.7612	Bom
manhattan	weighted	0.5876	Fraco
manhattan	centroid	N/A	Incompatível
manhattan	ward	N/A	Incompatível
mahalanobis	single	0.7410	Moderado
mahalanobis	complete	0.6076	Fraco
mahalanobis	average	0.7759	Bom
mahalanobis	weighted	0.5928	Fraco
mahalanobis	centroid	N/A	Incompatível
mahalanobis	ward	N/A	Incompatível

A determinação do número de agrupamentos (clusters) foi realizado por meio dos métodos Silhouette, Calinski-H, Davies-B, Inércia e Gap aparente no dendograma (ver Figura 6). Para definir o número de agrupamentos a partir dos diferentes métodos foi utilizado de consenso a partir dos resultados de cada método (conforme Tabela 4).

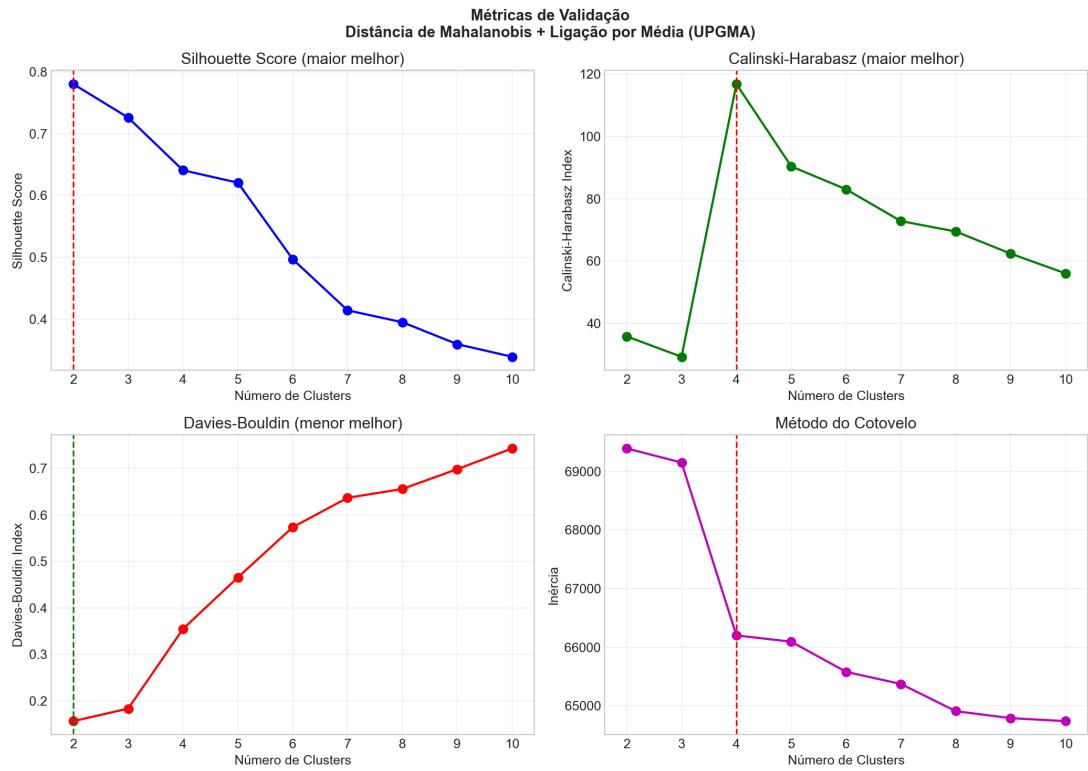


Figura 6: Gráficos com diferentes métodos de determinação do número adequado de agrupamentos presente nas amostras.

Tabela 4: Tabela com os resultados dos diferentes métodos de determinação do número adequado de agrupamentos presente nas amostras; Consenso (mediana):  $k = 2$ .

Métrica	$k$ ótimo
Silhouette	2
Calinski-Harabasz	4
Davies-Bouldin	2
Cotovelo	4
Gap Dendrograma	2

#### 4.4. Modelo PLS-DA

O PLS-DA é uma técnica de classificação supervisionada que combina a regressão por Mínimos Quadrados Parciais (PLS) com Análise Discriminante. Diferentemente do PCA (não supervisionado), o PLS incorpora a informação da variável resposta  $y$  durante a decomposição de  $X$ .

O algoritmo PLS é iterativo com as etapas:

1. Cálculo de pesos;
2. Escores;
3. Pesos de  $y$ ;
4. Pesos de  $X$ ;
5. Deflação (atualização das matrizes);

As etapas de 1 a 5 se repetem até encontrar o número de variáveis latentes desejado. As variáveis latentes (LVs) no PLS resultam da leve rotação dos componentes principais do PCA, otimizando simultaneamente a explicação da variância em  $X$  e a correlação/covariância com  $y$ .

A predição de dados por PLS inicia pela verificadação da distribuição das amostras entre as três classes, seguida pela divisão dos conjuntos de treino e de teste. O treino do modelo PLS utilizou os dados coletados para predizer a qualidade do vinho dentre as categorias baixa, média e alta. A Figura 7 demonstra a seleção do número de componentes com base na acurácia. É com base nesse número de componentes que é realizado o treino do modelo final.

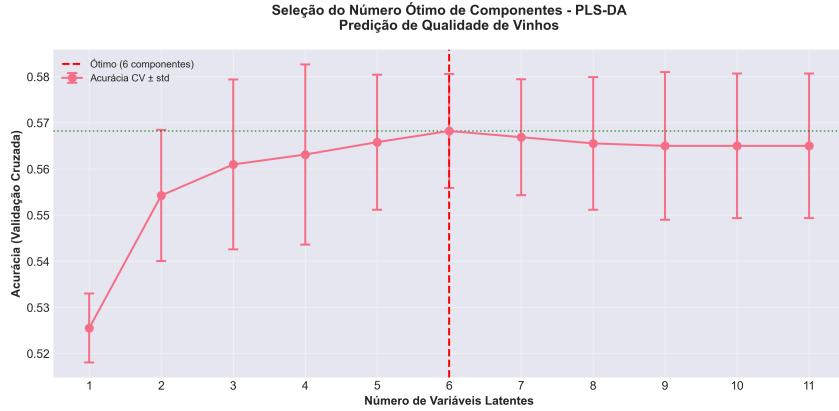


Figura 7: Gráfico de seleção do número ótimo de variáveis latentes.

Seguida a etapa de treino, tem-se a avaliação do modelo propriamente dito. Os valores encontrados para acurácia não foram elevados, por volta de 60%, conforme Tabela 5. A matriz de confusão demonstrou que o modelo construído teve bom desempenho para avaliar vinhos de qualidade baixa e média (acima de 70%), porém teve uma péssima performance na avaliação de vinho classificados como possuindo alta qualidade (14,85% de acurácia) (ver Figura 8). A escolha de 6 variáveis latentes foi capaz de explicar apenas 64,09% da variância.

Tabela 5: Avaliação de acurácia.

Conjunto	Acurácia
Treino	0.5720 (57.20%)
Teste	0.6040 (60.40%)

Tabela 6: Avaliação de desempenho do modelo.

Classe	Precisão	Reconhecimento	F1-Score	Supor te
Alta	0.6250	0.1485	0.2400	303
Baixa	0.6842	0.7198	0.7016	596
Média	0.5463	0.7030	0.6148	697
Acurácia			0.6040	1596
Média	0.6185	0.5238	0.5188	1596
Média ponderada	0.6127	0.6040	0.5760	1596

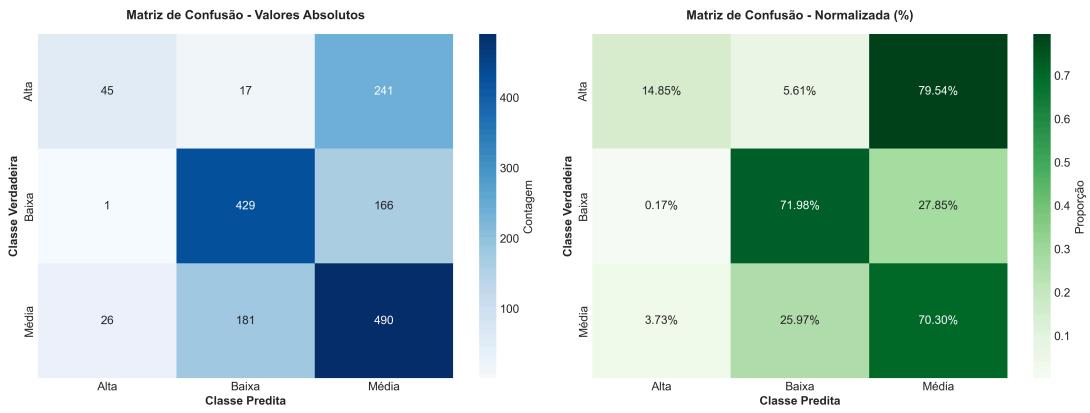


Figura 8: Matriz de confusão do modelo PLS.

Para demonstrar a dificuldade de se fazer a distinção das classes por meio deste método, um gráfico de escores com as variáveis latentes 1 e 2 demonstra como não existe regiões bem definidas para as três classes avaliadas (ver Figura 9).

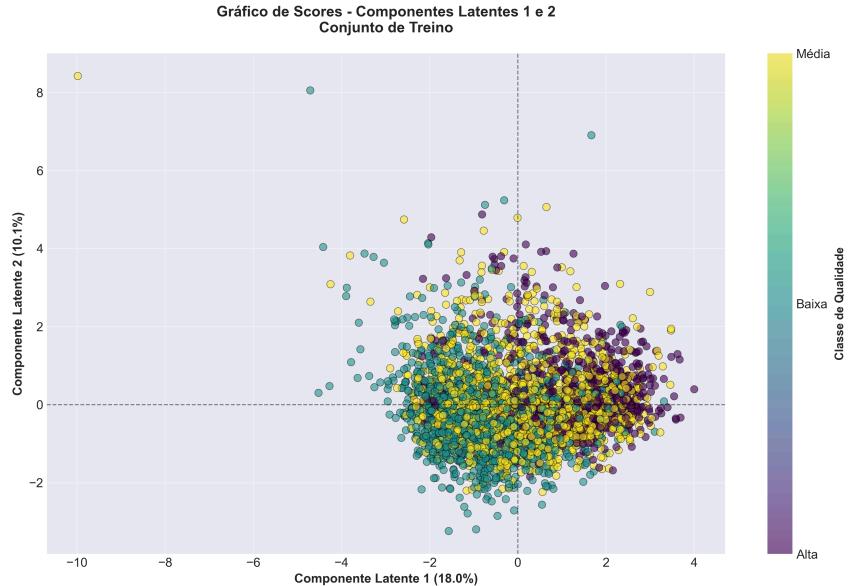


Figura 9: Gráfico de escores com as variáveis latentes 1 e 2.

O modelo foi capaz de destacar o grau de importância de variáveis para a predição da qualidade dos vinhos. Dos parâmetros, o pH e o teor de sulfatos foram os principais dada a soma dos pesos (ver Figura 10). A Tabela 7 apresenta os loadings (cargas) de cada variável nas 6 variáveis latentes (LV1-LV6) do modelo PLS. Os valores indicam a importância e a direção da contribuição de cada variável para cada componente latente.

Tabela 7: Loadings das variáveis por variável latente.

Variável	LV1	LV2	LV3	LV4	LV5	LV6
fixed acidity	-0.2258	0.5766	-0.4108	-0.0140	-0.4219	-0.2079
volatile acidity	-0.2645	0.0222	-0.1773	-0.5071	0.3348	-0.5676
citric acid	0.0163	0.4618	-0.1892	0.5848	0.0091	-0.0103
residual sugar	-0.3124	0.1618	0.5767	0.1832	-0.2547	-0.4283
chlorides	-0.3397	0.2586	-0.3050	-0.1890	-0.0203	0.6494
free sulfur dioxide	-0.0418	-0.0247	0.5916	0.3362	-0.1291	0.4663
total sulfur dioxide	-0.0844	-0.1713	0.4888	0.5266	0.1801	-0.0368
density	-0.6285	0.3258	0.0719	-0.0589	0.0426	-0.1653
pH	0.0674	-0.0949	0.1542	-0.4198	0.7443	0.5507
sulphates	-0.0892	0.6225	-0.1576	-0.1657	0.6178	0.3784
alcohol	0.6142	0.2614	0.0435	-0.1933	-0.0844	-0.3273

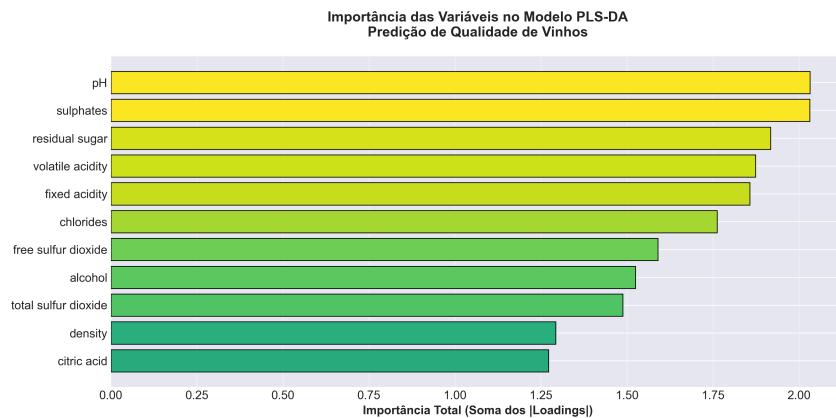


Figura 10: Soma dos pesos das variáveis latentes.

## 5. Conclusão

Este trabalho demonstrou a aplicabilidade de métodos quimiométricos para análise de dados químicos complexos. Os resultados obtidos indicam que as técnicas multivariadas empregadas foram parcialmente eficazes para predição da qualidade de vinhos a partir dos parâmetros medidos, tendo em vista os valores baixos de acurácia no modelo PLS-DA. O modelo PCA também demonstrou que o uso de apenas 3 componentes não foi o suficiente para reduzir significativamente o resíduo Q e o leverage, indicando que os outliers tem uma influência grande para o modelo.

## Bibliografia

CORTEZ, Paulo *et al.* Wine Quality. UCI Machine Learning Repository, , 2009.

CORTEZ, Paulo *et al.* Modeling Wine Preferences by Data Mining from Physicochemical Properties. **Decision Support Systems**, Smart Business Networks: Concepts and Empirical Evidence. v. 47, n. 4, p. 547–553, nov. 2009.

EVERITT, Brian S. *et al.* Cluster analysis. 2011.

WOLD, Svante; SJÖSTRÖM, Michael; ERIKSSON, Lennart. PLS-regression: a basic tool of chemometrics. **Chemometrics and intelligent laboratory systems**, v. 58, n. 2, p. 109–130, 2001.