UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE ENGENHARIAS E TECNOLOGIA
CURSO BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

IAGO ALVES MONTE

ANÁLISE DE REDUÇÃO DE DIMENSIONALIDADE COM PCA PARA PREDIÇÃO DA QUALIDADE DE VINHOS BRANCOS UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

PAU DOS FERROS
2026

IAGO ALVES MONTE

ANÁLISE DE REDUÇÃO DE DIMENSIONALIDADE COM PCA PARA PREDIÇÃO DA QUALIDADE DE VINHOS BRANCOS UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Bacharelado em Engenharia de Computação.

Orientador: Pedro Thiago Valério de Souza, Prof. Dr.

PAU DOS FERROS

2026

M772a Monte, Iago Alves.
Análise de redução de dimensionalidade com PCA para predição da qualidade de vinhos brancos utilizando algoritmos de aprendizado de máquina / Iago Alves Monte. - 2026.
47 f. : il.

Orientador: Pedro Thiago Valério de Souza.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de --Selecione um Curso ou Programa--, 2026.

1. Aprendizado de máquina. 2. Análise de componentes principais. 3. Classificação de vinhos. 4. GridSearchCV. 5. Otimização de hiperparâmetros. I. Souza, Pedro Thiago Valério de, orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automático em conformidade com AACR2 e os dados fornecidos pelo) autor(a).
Biblioteca Campus Pau dos Ferros
Bibliotecária: Erlanda Maria Lopes da Silva
CRB: 15/966

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC´s) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

IAGO ALVES MONTE

ANÁLISE DE REDUÇÃO DE DIMENSIONALIDADE COM PCA PARA PREDIÇÃO DA QUALIDADE DE VINHOS BRANCOS UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Bacharelado em Engenharia de Computação.

Defendida em: 03 / 07 / 2026

BANCA EXAMINADORA

Pedro Thiago Valério de Souza, Prof. Dr. (UFERSA)
Presidente

Kennedy Reurison Lopes, Prof. Dr. (UFERSA)
Membro Examinador

Rosana Cibely Batista Rego, Prof. Dra. (UFERSA)
Membro Examinador

DEDICATÓRIA

Dedico este trabalho a Deus, fonte de força e propósito em minha caminhada.
À minha mãe, pelo amor, cuidado e apoio incondicionais ao longo de toda a minha vida.
Ao meu pai, cuja memória e exemplo continuam sendo inspiração para seguir em frente.
E a todos aqueles que acreditaram em mim, caminharam ao meu lado e contribuíram para que este momento se tornasse possível.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, que me sustentou em todos os momentos desta trajetória, concedendo forças, sabedoria, perseverança e esperança para seguir em frente mesmo diante das dificuldades.

À minha mãe, por todo o amor, cuidado, incentivo e pelos inúmeros sacrifícios feitos ao longo da minha vida. Sua dedicação foi fundamental para que eu chegasse até aqui.

Ao meu irmão, pela companhia, apoio e por compartilhar comigo os desafios e conquistas desta jornada.

À Libhinny, por estar ao meu lado tanto nos momentos difíceis quanto nos mais felizes, oferecendo compreensão, incentivo e acreditando em mim até mesmo quando duvidei.

Ao meu pai, que, embora não tenha podido acompanhar esta etapa da minha vida nem presenciar meu ingresso na universidade, permanece como uma grande inspiração. Sua influência contribuiu para a formação da pessoa que sou hoje e para os sonhos que busquei realizar.

À minha querida célula e membros da Escape, Farol e Missão Tenda, minha sincera gratidão por cada amizade construída, por cada oração, palavra de incentivo e momento compartilhado ao longo desses anos. Sou muito grato por tudo o que vivi com vocês e por cada pessoa que fez parte dessa história.

À minha família, que sempre esteve presente de diferentes formas ao longo desta caminhada. Sou grato pelo carinho, pelas palavras de incentivo, pelas orações e pelo apoio recebido. Cada gesto de confiança e encorajamento teve sua importância para que eu chegasse até aqui.

Ao grupo TheWings, que nasceu dos jogos, mas se transformou em uma amizade que levarei para a vida. Minha gratidão pela companhia, pelas conversas, pelo apoio nos momentos difíceis e por todas as risadas e memórias que construímos juntos.

Aos amigos e colegas que conheci durante a graduação, meu sincero agradecimento. Muito além dos trabalhos, provas e atividades acadêmicas, vocês fizeram parte da construção das memórias, aprendizados e experiências que marcaram minha vida universitária. Compartilhamos desafios, conquistas, momentos de descontração e apoio mútuo, tornando essa jornada mais leve e significativa.

Ao grupo de amigos que conheci por meio do YouTube, pessoas que fizeram parte da minha trajetória e me proporcionaram a oportunidade de conhecer outras pessoas igualmente

especiais, ampliando minha visão de mundo e enriquecendo minha experiência de vida.

Aos professores que contribuíram para minha formação, agradeço pelos conhecimentos compartilhados, pela dedicação ao ensino e pelos ensinamentos que ultrapassam os limites da sala de aula. Cada experiência contribuiu para meu crescimento acadêmico e pessoal.

Ao meu orientador, professor doutor Pedro Thiago, pela orientação, paciência, disponibilidade e pelos valiosos conselhos durante o desenvolvimento deste trabalho. Sua contribuição foi fundamental para a concretização desta etapa.

Por fim, agradeço a todos aqueles que, de alguma forma, contribuíram para minha formação acadêmica, pessoal e espiritual, deixando marcas positivas em minha história.

EPÍGRAFE

"He changes times and seasons; He removes kings and sets up kings; He gives wisdom to the wise and knowledge to those who have understanding."

(Daniel 2:21)

RESUMO

A qualidade de vinhos é um atributo de elevada importância econômica e científica, determinado por complexas interações físico-químicas que influenciam as características sensoriais associadas à sua qualidade. Este trabalho investiga a aplicação de cinco algoritmos de aprendizado de máquina supervisionado: *Random Forest*, *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), *Multi-Layer Perceptron* (MLP) e *Gaussian Naive Bayes* (GNB), integrados à técnica de Análise de Componentes Principais (*Principal Component Analysis* - PCA) para a classificação da qualidade de vinhos brancos do conjunto *UCI Wine Quality*. A aplicação do PCA busca reduzir a dimensionalidade dos atributos químicos e eliminar redundâncias entre variáveis correlacionadas, preservando a maior parte da variância dos dados. A otimização de hiperparâmetros foi conduzida via busca em grade com validação cruzada (*Grid Search with Cross-Validation - GridSearchCV*) com validação cruzada estratificada de cinco *folds*, adotando-se o F1-macro como principal métrica de desempenho. O experimento estruturou-se em dois cenários distintos: classificação multiclasse, abrangendo notas de 3 a 9, e classificação binária, distinguindo vinhos de qualidade superior (≥ 7) de vinhos de qualidade inferior. As conclusões demonstram que a integração entre redução dimensional e modelos inteligentes favorece a identificação de padrões críticos na composição vinícola.

Palavras-chave: aprendizado de máquina; análise de componentes principais; classificação de vinhos; *GridSearchCV*; otimização de hiperparâmetros.

ABSTRACT

Wine quality is an attribute of great economic and scientific importance, determined by complex physicochemical interactions that influence the sensory characteristics associated with its quality. This study investigates the application of five supervised machine learning algorithms: Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Multi-Layer Perceptron (MLP), and Gaussian Naive Bayes (GNB), integrated with the Principal Component Analysis (PCA) technique for classifying the quality of white wines from the UCI Wine Quality dataset. The application of PCA aims to reduce the dimensionality of the chemical attributes and eliminate redundancies among correlated variables, while preserving most of the data variance. Hyperparameter optimization was conducted using GridSearchCV with stratified cross-validation across five folds, adopting the F1-macro as the primary performance metric. The experiment was structured around two distinct scenarios: multiclass classification, covering scores from 3 to 9, and binary classification, distinguishing high-quality wines (≥ 7) from low-quality wines. The results demonstrate that the integration of dimensionality reduction and intelligent models facilitates the identification of critical patterns in wine composition.

Keywords: machine learning; principal component analysis; wine classification; GridSearchCV; hyperparameter optimization.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	*Artificial Neural Network*
AUC	*Area Under the Curve*
CP	Componente Principal
CSV	*Comma-Separated Values*
GBR	*Gradient Boosting Regressor*
GNB	*Gaussian Naive Bayes*
KNN	*K-Nearest Neighbors*
MLP	*Multi-Layer Perceptron*
OvR	*One-vs-Rest*
PCA	*Principal Component Analysis*
RF	*Random Forest*
ROC	*Receiver Operating Characteristic*
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
UCI	*University of California, Irvine*

SUMÁRIO

1 INTRODUÇÃO

A qualidade do vinho é um dos atributos mais importantes da indústria vitivinícola e está diretamente relacionada aos processos de certificação e avaliação comercial do produto. (Cortez Paulo; Reis, 2009). Historicamente, sua avaliação depende de análises sensoriais conduzidas por especialistas em degustação, os chamados *sommeliers*, que atribuem notas a amostras com base em critérios subjetivos como aroma, acidez, corpo e sabor residual (Lesschaeve, 2007). Embora esse método seja amplamente reconhecido, ele apresenta limitações importantes: a subjetividade intrínseca ao julgamento humano pode gerar variações entre avaliadores, e o processo é custoso e difícil de escalar para grandes volumes de produção (Lesschaeve, 2007; Cortez Paulo; Reis, 2009).

Diante disso, a utilização de variáveis físico-químicas mensuráveis em laboratório, como acidez, teor alcoólico e concentração de compostos sulfurosos, tem se mostrado uma alternativa promissora para automatizar e tornar mais objetiva a avaliação da qualidade do vinho (Cortez Paulo; Reis, 2009). Nesse contexto, o aprendizado de máquina surge como uma ferramenta especialmente adequada, permitindo que modelos aprendam padrões complexos a partir dos dados químicos e façam predições de qualidade sem a necessidade de intervenção humana direta (Bishop, 2006). O conjunto de dados *UCI Wine Quality*, introduzido por Cortez Paulo e Reis (2009), tornou-se uma das referências mais utilizadas para esse tipo de experimento, reunindo 4.898 amostras de vinho branco do tipo Vinho Verde português, cada uma descrita por 11 atributos físico-químicos e uma nota de qualidade atribuída por especialistas.

Um aspecto relevante desse conjunto de dados é que vários de seus atributos apresentam correlações entre si, por exemplo, o teor alcoólico e a densidade são inversamente relacionados, e o açúcar residual influencia diretamente a densidade do líquido. Isso significa que, embora o problema tenha 11 dimensões de entrada, parte dessa informação é redundante: diferentes variáveis descrevem, em certa medida, o mesmo fenômeno. Nesse contexto, a Análise de Componentes Principais (PCA) é uma técnica de transformação linear que reorganiza os dados em novas dimensões, chamadas componentes principais, de forma que as primeiras concentrem a maior parte da variância do conjunto (Jolliffe, 2002). Ao reduzir a dimensionalidade dos dados e concentrar a maior parte da variância em um número reduzido de componentes, o PCA pode simplificar a representação do problema e reduzir a redundância entre atributos (Shlens, 2014).

Este trabalho investiga o impacto dessa redução de dimensionalidade no desempenho de cinco algoritmos de classificação supervisionada: *Random Forest*, SVM, KNN, MLP e Gaussiano

Naive Bayes. Para cada algoritmo e cada configuração de PCA (de 2 a 10 componentes, além da configuração sem redução), os hiperparâmetros foram otimizados de forma sistemática por meio da busca em grade com validação cruzada (*GridSearchCV*) (Pedregosa *et al.*, 2011), adotando o F1-macro como critério de seleção, uma vez que essa métrica atribui peso igual a todas as classes e é amplamente utilizada na avaliação de conjuntos desbalanceados (Sokolova; Lapalme, 2009). Os experimentos foram conduzidos em dois cenários: classificação multiclasse, em que o modelo prediz a nota do vinho em uma escala de 3 a 9, e classificação binária, em que vinhos com nota ≥ 7 são considerados de alta qualidade.

1.1 Objetivos

1.1.1 Objetivo Geral

Analisar o impacto da redução de dimensionalidade por meio da Análise de Componentes Principais no desempenho de algoritmos de aprendizado de máquina supervisionados aplicados à predição da qualidade de vinhos brancos, utilizando o conjunto de dados *UCI Wine Quality*.

1.1.2 Objetivos Específicos

- Aplicar a técnica de PCA para redução da dimensionalidade dos dados físico-químicos e eliminação de redundâncias entre atributos correlacionados;
- Implementar e otimizar cinco algoritmos de classificação supervisionada (Random Forest, Support Vector Machine, K-Nearest Neighbors, Multi-Layer Perceptron e Gaussian Naive Bayes);
- Avaliar o desempenho dos modelos em dois cenários experimentais: classificação multiclasse (notas de 3 a 9) e classificação binária (qualidade superior ≥ 7 versus qualidade inferior);
- Comparar o desempenho dos modelos com e sem a aplicação da redução de dimensionalidade por PCA, considerando métricas de avaliação como F1-macro e AUC-ROC;
- Analisar a contribuição das variáveis físico-químicas originais por meio da interpretação dos *loadings* do PCA.

2 TRABALHOS RELACIONADOS

O problema de predição da qualidade de vinhos a partir de variáveis físico-químicas vem sendo explorado na literatura há pelo menos duas décadas, e o repositório UCI Wine Quality tornou-se o conjunto de dados de referência para boa parte desses estudos.

Cortez Paulo e Reis (2009) introduziram o próprio conjunto de dados e conduziram um dos primeiros estudos comparativos de aprendizado de máquina aplicado à qualidade de vinhos verdes portugueses. Os autores avaliaram três abordagens: Regressão Múltipla, Redes Neurais e SVM com *kernel* RBF, tanto para os dados de vinho tinto quanto para os de vinho branco, reportando acurácia em torno de 65 % no cenário multiclasse com SVM. O trabalho estabeleceu que a classificação sensorial de vinhos é um problema inerentemente difícil, especialmente em função do desbalanceamento das classes e da natureza ordinal da variável alvo, e serviu de ponto de partida para todos os estudos posteriores que utilizaram o mesmo conjunto.

Ye (2023) conduziram um estudo de classificação da qualidade do vinho tinto do repositório UCI, com 1.599 amostras e 11 atributos físico-químicos, utilizando quatro algoritmos: *Random Forest*, SVM com diferentes funções de *kernel*, KNN e Rede Neural. O trabalho aplicou os algoritmos sobre três formas dos dados (originais, padronizados e com PCA reduzido a 2 componentes) e identificou a rede neural sobre dados padronizados como o melhor modelo, com precisão e *recall* de 88 %. Contudo, a redução via PCA foi fixada arbitrariamente em 2 componentes, a otimização de hiperparâmetros foi parcial e não foram calculadas curvas ROC nem análises de importância de características, limitações que o presente trabalho busca superar.

O presente estudo avança nessa direção ao investigar o impacto de PCA com k variando de 2 a 10 componentes, empregar *GridSearchCV* sobre cinco algoritmos distintos, estruturar o problema em dois cenários (multiclasse e binário) e calcular curvas ROC e análises de importância para ambos.

Dahal *et al.* (2021) propuseram um estudo para predição da qualidade de vinhos utilizando o conjunto de dados *Wine Quality* do UCI Machine Learning Repository. Os autores realizaram uma etapa de pré-processamento envolvendo análise estatística exploratória, cálculo da correlação entre atributos, padronização dos dados e ajuste de hiperparâmetros por validação cruzada e busca em grade. Foram comparados os modelos *Ridge Regression*, *Support Vector Regression* (SVR), *Gradient Boosting Regressor* (GBR) e *Artificial Neural Network* (ANN), sendo o *Gradient Boosting Regressor* o algoritmo que apresentou o melhor desempenho entre os métodos avaliados. Embora o trabalho realize uma comparação entre diferentes modelos de aprendizado de máquina,

não investiga o efeito da redução de dimensionalidade por meio da Análise de Componentes Principais (PCA), tampouco compara o impacto dessa técnica em diferentes algoritmos, aspectos abordados na presente pesquisa.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Base de Dados

Disponibilizado publicamente na UCI Machine Learning Repository, o conjunto de dados *UCI Wine Quality* foi introduzido por Cortez Paulo e Reis (2009) e tornou-se referência consolidada em estudos de classificação sensorial de bebidas e aprendizado de máquina aplicados à análise de produtos enológicos. Esse conjunto de dados reúne informações referentes a Vinhos Verdes produzidos na região norte de Portugal, sendo composto por medições físico-químicas obtidas a partir de análises laboratoriais, bem como avaliações sensoriais realizadas por especialistas.

Neste estudo foi utilizada a versão referente aos vinhos brancos (*white wine*), a qual contém 4898 amostras. Cada amostra é descrita por 11 variáveis físico-químicas que caracterizam propriedades relevantes da bebida, como acidez, concentração de compostos químicos e teor alcoólico. Esses atributos representam medidas objetivas obtidas por meio de procedimentos analíticos utilizados na caracterização da composição do vinho.

Os atributos considerados no conjunto de dados são:

- Acidez fixa (*fixed acidity*)
- Acidez volátil (*volatile acidity*)
- Ácido cítrico (*citric acid*)
- Açúcar residual (*residual sugar*)
- Cloretos (*chlorides*)
- Dióxido de enxofre livre (*free sulfur dioxide*)
- Dióxido de enxofre total (*total sulfur dioxide*)
- Densidade (*density*)
- *pH*
- Sulfatos (*sulphates*)
- Álcool (*alcohol*)

Além das variáveis de entrada, o conjunto de dados possui uma variável de saída denominada *quality*, que representa a avaliação da qualidade do vinho. Essa variável é expressa por um valor inteiro em uma escala ordinal, variando de 0 a 10, sendo determinada a partir de análises sensoriais realizadas por especialistas em degustação de vinhos. No conjunto específico utilizado neste trabalho, a distribuição dessa classe é fortemente concentrada nas notas 5, 6 e 7, que juntas

representam aproximadamente 93 % das amostras.

O fluxo experimental adotado neste trabalho foi estruturado de forma a permitir a avaliação sistemática do impacto da redução de dimensionalidade por meio da técnica PCA no desempenho de diferentes algoritmos de classificação. Todas as etapas de processamento e modelagem foram implementadas em linguagem Python, utilizando principalmente a biblioteca *Scikit-Learn*.

De maneira geral, o procedimento experimental foi composto pelas seguintes etapas:

1. Carregamento e inspeção inicial do conjunto de dados;
2. Separação das variáveis de entrada e da variável alvo;
3. Divisão dos dados em conjuntos de treino e teste;
4. Normalização das variáveis;
5. Aplicação da técnica de redução de dimensionalidade PCA;
6. Treinamento dos modelos de classificação;
7. Otimização dos hiperparâmetros utilizando o *GridSearchCV*;
8. Avaliação dos modelos por meio de métricas de desempenho.

Cada uma dessas etapas é descrita em maior detalhe nas subseções a seguir.

3.2 Preparação e divisão dos dados

Inicialmente, o conjunto de dados foi carregado em formato *DataFrame* utilizando a biblioteca *Pandas*. Em seguida, as variáveis preditoras foram separadas da variável alvo, sendo consideradas como atributos de entrada todas as 11 variáveis físico-químicas disponíveis no conjunto de dados, enquanto a variável *quality* foi utilizada como variável de saída.

Posteriormente, os dados foram divididos em conjuntos de treinamento e teste utilizando o método *train_test_split* da biblioteca *Scikit-Learn*. Esse procedimento realiza a separação aleatória do conjunto de dados em duas partições disjuntas, permitindo que o modelo seja ajustado no conjunto de treinamento e avaliado em dados não vistos durante o treinamento.

Foi adotada uma divisão de 80% para treinamento e 20% para teste. Para preservar a distribuição original das classes da variável alvo, foi utilizada amostragem estratificada (*stratified sampling*). Além disso, foi fixado o valor de *random_state = 42* para garantir a reprodutibilidade dos experimentos.

Dois cenários experimentais foram considerados neste trabalho:

- Classificação multiclasse, utilizando diretamente os valores originais da varíavel *quality*;

- Classificação binária, na qual a variável alvo foi transformada em duas categorias: vinhos de qualidade superior (*quality* ≥ 7) e vinhos de qualidade inferior (*quality* < 7).

Essa divisão permite avaliar o desempenho dos modelos em dois cenários com diferentes níveis de complexidade. O problema multiclasse preserva toda a parametrização original do conjunto de dados, porém apresenta maior dificuldade de separação entre as classes. Por outro lado, a formulação binária simplifica o espaço de decisão e permite analisar a capacidade dos modelos em distinguir vinhos de baixa e alta qualidade, além de reduzir parcialmente o impacto do desbalanceamento entre classes.

3.3 Normalização dos dados

Antes do treinamento dos modelos, os atributos numéricos foram normalizados utilizando o método *StandardScaler*. Essa abordagem define cada variável de modo que passe a apresentar média igual a zero e desvio padrão unitário, conforme descrito na Equação 4.1.

A normalização é particularmente importante para algoritmos sensíveis à escala dos dados, como *Support Vector Machines* e *K-Nearest Neighbors*. Dessa forma, evita-se que variáveis com magnitudes maiores exerçam influência desproporcional durante o processo de treinamento dos modelos.

3.4 Redução de dimensionalidade com PCA

Com o objetivo de investigar o impacto da redução de dimensionalidade no desempenho dos classificadores, foi aplicada a técnica de Análise de Componentes Principais. O PCA é uma técnica de transformação linear de dados multivariados proposta originalmente por Pearson (1901) e posteriormente formalizada no contexto estatístico por Hotelling (1933). Seu principal objetivo consiste em transformar um conjunto de variáveis possivelmente correlacionadas em um novo conjunto de variáveis não correlacionadas, denominadas componentes principais, mantendo a maior quantidade possível da variabilidade presente nos dados (Jolliffe, 2002).

Antes da aplicação do PCA, os atributos foram padronizados para evitar que variáveis com diferentes escalas dominassem a transformação. Considerando uma matriz de dados padronizados $\mathbf{X}$, composta por n amostras e p atributos, a matriz de covariância é calculada por:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}, \tag{3.1}$$

onde $\mathbf{C}$ representa a matriz de covariância dos atributos. A partir dessa matriz, são obtidos os autovalores e autovetores por meio do problema de autovalor:

$$\mathbf{C}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \tag{3.2}$$

em que λ_i representa o autovalor associado ao componente principal i, indicando a quantidade de variância explicada por esse componente, e $\mathbf{v}_i$ corresponde ao autovetor que define a direção do novo eixo no espaço transformado.

Os autovetores são ordenados de forma decrescente conforme seus respectivos autovalores, formando uma matriz de transformação composta pelos k primeiros componentes principais:

$$\mathbf{W}_k = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k], \tag{3.3}$$

onde k representa o número de componentes principais selecionados. A projeção dos dados no novo espaço dimensional é então obtida por:

$$\mathbf{Z} = \mathbf{X}\mathbf{W}_k, \tag{3.4}$$

em que $\mathbf{Z}$ representa a matriz de dados transformada no espaço reduzido.

A proporção de variância explicada por cada componente é determinada pela razão entre seu autovalor e a soma de todos os autovalores:

$$VE_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}, \tag{3.5}$$

onde VE_i representa a variância explicada pelo componente principal i.

De forma intuitiva, o PCA pode ser interpretado como uma rotação do sistema de coordenadas dos dados, buscando novas direções que concentrem a maior quantidade de informação possível. Assim, em vez de analisar diretamente as variáveis originais, os modelos de classificação passam a operar sobre combinações lineares dessas variáveis, reduzindo possíveis redundâncias entre atributos correlacionados.

Nos experimentos realizados, foram avaliadas diferentes configurações do número de componentes principais. Foram considerados cenários sem aplicação de PCA, mantendo as 11

variáveis físico-químicas originais, e cenários com redução progressiva do número de componentes, variando de 10 até 2 componentes principais. Essa abordagem permitiu analisar o efeito da compressão dos dados sobre o desempenho dos modelos de classificação.

3.5 Construção dos pipelines de aprendizado

Para garantir um fluxo de processamento consistente e evitar vazamento de dados entre as etapas de treinamento e avaliação, foi utilizada a estrutura de *pipelines* da biblioteca *Scikit-Learn*. Cada pipeline foi composto pelas seguintes etapas:

1. Normalização dos dados utilizando *StandardScaler*;
2. Aplicação do PCA;
3. Treinamento do algoritmo de classificação.

Essa estratégia assegura que todas as transformações sejam aplicadas corretamente dentro do processo de validação cruzada e durante a fase de teste.

3.6 Algoritmos de classificação avaliados

Foram avaliados cinco algoritmos de classificação amplamente utilizados em tarefas de aprendizado de máquina:

- Floresta Aleatória (*Random Forest*)
- Máquina de Vetores de Suporte (*Support Vector Machine*, SVM)
- K-Vizinhos Mais Próximos (*K-Nearest Neighbors*, KNN)
- Perceptron Multicamadas (*Multi-Layer Perceptron*, MLP)
- Naive Bayes Gaussiano (*Gaussian Naive Bayes*, GNB)

A seleção desses algoritmos foi motivada pela diversidade de abordagens de aprendizado representados, incluindo métodos baseados em ensemble, técnicas geométricas, modelos probabilísticos e redes neurais artificiais. A seguir, cada um dos algoritmos é descrito de forma sucinta.

3.6.1 *Random Forest*

O algoritmo *Random Forest*, proposto por Breiman (2001), é um método de *ensemble* que constrói múltiplas árvores de decisão durante o treinamento, combinando suas predições por votação majoritária. Cada árvore é treinada em uma amostra *bootstrap*, ou seja, uma subamostra aleatória com reposição do conjunto de treinamento, e em cada nó de divisão apenas

um subconjunto aleatório de *features* é considerado, introduzindo diversidade entre as árvores e reduzindo a correlação entre seus erros (Breiman, 2001). Essa estratégia confere ao método robustez a *overfitting* e estabilidade em presença de variáveis irrelevantes ou ruidosas.

Os parâmetros utilizados foram: número de árvores (`n_estimators` $\in \{100, 200\}$), profundidade máxima (`max_depth` $\in \{\text{None}, 10, 20\}$) e número mínimo de amostras para divisão de nó (`min_samples_split` $\in \{2, 5\}$), totalizando 12 combinações por configuração de PCA.

3.6.2 Support Vector Machine (SVM)

As Máquinas de Vetores de Suporte, introduzidas por Cortes e Vapnik (1995), constroem um hiperplano de separação que maximiza a margem entre classes. Para dados não linearmente separáveis, o truque do *kernel* permite mapear implicitamente os dados para um espaço de alta dimensão onde a separação linear torna-se viável (Hofmann; Schölkopf; Smola, 2008). Dois *kernels* foram investigados: a função de base radial (*Radial Basis Function*), apresentada na Equação 3.6, e o *kernel* polinomial, descrito na Equação 3.7.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \left\|\mathbf{x} - \mathbf{x}'\right\|^2\right) \tag{3.6}$$

$$K(\mathbf{x}, \mathbf{x}') = \left(\gamma \mathbf{x}^\top \mathbf{x}' + r\right)^d \tag{3.7}$$

em que γ controla a influência de cada amostra de treinamento, r representa o termo independente (*coef0*) do kernel polinomial, d corresponde ao grau do polinômio, e $\mathbf{x}$ e $\mathbf{x}'$ são os vetores de características das amostras.

Os hiperparâmetros investigados foram: constante de regularização ($C \in \{0.1, 1, 10\}$), tipo de *kernel* (*rbf* e *poly*) e grau do polinômio (`degree` $\in \{2, 3\}$, aplicável exclusivamente ao *kernel* polinomial).

3.6.3 K-Nearest Neighbors (KNN)

O algoritmo KNN, formalizado por Cover e Hart (1967), é um método não paramétrico baseado em instâncias que classifica uma nova observação pela maioria dos rótulos de seus k vizinhos mais próximos no espaço de *features*, segundo alguma métrica de distância (Altman, 1992). Nesse contexto, a métrica mais comum é a distância euclidiana, que calcula o comprimento do segmento de reta que conecta dois pontos no espaço de atributos.

Por se tratar de um método *lazy learning*, o KNN não possui uma fase de treinamento explícita, ou seja, ele não gera um modelo ajustado previamente, e todo o esforço computacional é realizado na etapa de classificação de novas observações, chamada de *inferência*. Para que a comparação das distâncias entre atributos seja justa, é essencial padronizar os dados, garantindo que todas as variáveis contribuam igualmente para o cálculo das distâncias.

Os hiperparâmetros avaliados foram: número de vizinhos ($k \in \{3,5,7\}$) e esquema de ponderação ($\texttt{weights} \in \{\text{uniform}, \text{distance}\}$). No caso da ponderação por distância, vizinhos mais próximos exercem maior influência na classificação, com base no inverso da distância euclidiana, enquanto na ponderação uniforme todos os vizinhos contribuem igualmente.

3.6.4 Multi-Layer Perceptron (MLP)

O Perceptron Multicamadas é um tipo de rede neural *feedforward*, no qual os dados de entrada percorrem camadas sucessivas de neurônios até produzir uma saída. Essa rede possui pelo menos uma camada oculta entre a entrada e a saída, o que permite modelar relações não lineares entre atributos de entrada e classes de saída. O treinamento do MLP é realizado por meio do algoritmo de retropropagação do erro (*backpropagation*), popularizado por Rumelhart, Hinton e Williams (1986), que ajusta iterativamente os pesos das conexões com base no erro entre a saída prevista e a desejada.

A capacidade de aproximar funções arbitrariamente complexas advém do *teorema de aproximação universal*, o qual redes *feedforward* com uma única camada oculta e funções de ativação não lineares podem aproximar qualquer função contínua com precisão arbitrária, desde que haja neurônios suficientes (Hornik; Stinchcombe; White, 1989).

Os hiperparâmetros pesquisados incluíram a arquitetura da rede ($\texttt{hidden\_layer\_sizes} \in \{(50,),(100,),(100,50)\}$), a função de ativação (*relu* e *tanh*) e o fator de regularização L2 ($\texttt{alpha} \in \{10^{-4}, 10^{-3}\}$), resultando em 12 combinações por configuração de PCA. O treinamento foi realizado por até 1.000 épocas.

3.6.5 Gaussian Naive Bayes (GNB)

O classificador Naive Bayes Gaussiano baseia-se no Teorema de Bayes para estimar a probabilidade de uma observação pertencer a uma dada classe, assumindo independência condicional entre as *features* (Friedman; Geiger; Goldszmidt, 1997). Isso significa que, para fins de cálculo, considera-se que cada atributo é estatisticamente independente dos demais

quando a classe é conhecida, o que simplifica os cálculos probabilísticos. No caso do Gaussiano, cada *feature* é modelada por uma distribuição normal (gaussiana), cujos parâmetros (média e variância) são estimados a partir dos dados de treinamento.

Embora a suposição de independência seja frequentemente violada em dados reais, o método apresenta desempenho razoável em dados de alta dimensão e é computacionalmente eficiente (Zhang, 2004). O único hiperparâmetro relevante avaliado foi o parâmetro de suavização de variância, que ficou descrito como: (`var_smoothing` $\in \{10^{-9}, 10^{-8}, 10^{-7}\}$), que adiciona uma fração da maior variância de todas as *features* à variância estimada de cada uma, estabilizando o cálculo.

3.7 Otimização de hiperparâmetros com *GridSearchCV*

A busca exaustiva em grade (*grid search*) consiste em avaliar sistematicamente todas as combinações de hiperparâmetros definidas em um espaço discreto de busca. Implementada em *scikit-learn* pelo objeto `GridSearchCV`, a técnica combina a busca em grade com validação cruzada (*cross validation*), garantindo estimativas de desempenho menos enviesadas em relação ao conjunto de teste (Pedregosa *et al.*, 2011).

Para cada modelo, foi utilizada validação cruzada estratificada com cinco partições (*StratifiedKFold*, $k = 5$), por ser o padrão da literatura, garantindo que cada partição mantenha a proporção original das classes. A aleatoriedade foi controlada por uma semente fixa, também conhecida como *seed* (`random_state=42`) para assegurar reprodutibilidade.

O critério adotado para seleção do melhor modelo foi a métrica *F1-score* com média macro (*F1-macro*), adequada para cenários com classes desbalanceadas, pois computa a média aritmética dos F1-scores individuais de cada classe sem ponderação por frequência (Sokolova; Lapalme, 2009). Formalmente:

$$\mathrm{F1}_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}, \qquad \text{F1-macro} = \frac{1}{|\mathscr{C}|} \sum_{c \in \mathscr{C}} \mathrm{F1}_c, \tag{3.8}$$

onde $\mathscr{C}$ é o conjunto de classes.

Os principais hiperparâmetros avaliados incluem, entre outros:

1. Número de árvores e profundidade máxima no Random Forest;
2. Parâmetro de regularização e tipo de *kernel* no SVM;
3. Número de vizinhos no KNN;

4. Arquitetura da rede neural e parâmetros de regularização no MLP;
5. Parâmetro de suavização da variância no Gaussian Naive Bayes.

4 METODOLOGIA

O conjunto de dados utilizado é o subconjunto de vinho branco do repositório *UCI Wine Quality* (Cortez Paulo; Reis, 2009), disponível publicamente na plataforma UCI Machine Learning Repository. Não foram identificados valores ausentes (*missing values*) em nenhuma das colunas, e nenhuma amostra foi descartada, resultando em um conjunto final de 4.898 observações e 12 colunas (11 *features* e a variável alvo *quality*).

O pré-processamento consistiu exclusivamente na padronização das *features* por meio do método do `StandardScaler`, aplicado tanto para dados de multiclasse quanto para os binarizados. As funções são nativas do *scikit-learn*, através do `sklearn.preprocessing`, e transformam cada variável para média zero e desvio padrão unitário segundo:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \tag{4.1}$$

onde x_{ij} é o valor original da j-ésima *feature* na i-ésima amostra, $\bar{x}_j$ é a média e s_j o desvio padrão da j-ésima *feature*, calculados exclusivamente sobre o conjunto de treinamento para prevenir vazamento de dados (*data leakage*) (Kaufman *et al.*, 2012). Essa padronização é particularmente necessária neste conjunto de dados, pois as variáveis como açúcar residual (valores de até 65,8 g/dm³) e dióxido de enxofre total (valores de até 440 mg/dm³) operam em escalas muito distintas.

4.1 Divisão treino/teste e formulação dos experimentos

Os dados foram divididos em conjuntos de treinamento (80 %) e de teste (20 %) de forma estratificada, preservando a proporção original das classes por meio de `train_test_split` com `stratify=y` e `random_state=42`. Essa divisão resultou em 3.918 amostras para treinamento e 980 para teste.

Dois experimentos independentes foram conduzidos para buscar as melhores formas de redução:

1. **Classificação multiclasse:** A variável alvo manteve seus valores originais (inteiros de 3 a 9), configurando um problema de 7 classes. A variável `y` foi definida diretamente como a coluna *quality* do *dataset*.
2. **Classificação binária:** A variável alvo foi binarizada segundo o critério $y_{bin} = 1$ se a classificação retornar $quality_{bin} \geq 7$, e $y_{binario} = 0$ caso contrário. No conjunto completo, 1.060 amostras pertencem à classe positiva (vinhos de alta qualidade, notas 7, 8 e 9),

enquanto 3.838 pertencem à classe negativa, resultando em uma proporção aproximada de 78/22 para as classes 0 e 1, respectivamente.

4.2 Métricas de avaliação

O desempenho dos modelos foi avaliado pelas seguintes métricas, calculadas sobre o conjunto de teste após o ajuste do melhor estimador retornado pelo *GridSearchCV*:

- **Acurácia:** Indica a proporção de previsões corretas em relação ao total de amostras avaliadas;
- **Precisão macro:** Corresponde à precisão média entre as classes, tratando todas como igualmente importantes, independentemente de sua frequência;
- ***Recall* macro:** Representa a média da capacidade do modelo em recuperar corretamente exemplos de cada classe;
- **F1-macro:** Calculado conforme a Equação 3.8, combina precisão e recall de forma equilibrada, sendo adotado como principal métrica de seleção;
- **Matriz de confusão:** Apresenta, de forma estruturada, como as previsões do modelo se distribuem entre as classes reais e preditas, facilitando a identificação de padrões de erro;
- **AUC-ROC:** Mede a capacidade do modelo em distinguir entre classes ao longo de diferentes limiares de decisão, sendo calculada apenas no experimento binário (Fawcett, 2006).

A curva ROC (*Receiver Operating Characteristic*) e a AUC (*Area Under the Curve*) são utilizadas como métricas complementares no experimento binário, permitindo avaliar a capacidade do modelo de discriminar entre as classes positivas e negativas independentemente do limiar de decisão.

4.3 Análise exploratória dos dados

Antes de apresentar os resultados dos classificadores, é importante compreender inicialmente as características do conjunto de dados. No subconjunto de vinho branco analisado neste trabalho, as notas 5 e 6 concentram, respectivamente, cerca de 30 % e 45 % das amostras, enquanto os extremos da escala apresenta as notas 3 e 9, que possuem representatividade muito baixa, com 20 e 5 instâncias no conjunto completo. Essa assimetria é o principal fator de dificuldade do cenário multiclasse e explica diretamente o F1-score nulo dos classificadores nas

classes extremas (He; Garcia, 2009).

Para o experimento binário, adotou-se o limiar de qualidade ≥ 7 para definir a classe positiva (alta qualidade), agrupando as notas 7, 8 e 9. A proporção resultante é de aproximadamente 78 % para a classe negativa e 22 % para a positiva. Mesmo que ainda seja desbalanceada, essa formulação é mais favorável à generalização dos modelos do que o cenário original de sete classes.

Devido ao desbalanceamento das classes, foi adotado o F1-macro como principal métrica de avaliação. Essa métrica atribui o mesmo peso a todas as classes, permitindo uma análise mais equilibrada do desempenho do modelo em relação às classes majoritárias e minoritárias, ao contrário do F1-weighted, que é influenciado pela distribuição de frequência do conjunto de dados.

4.3.1 Correlação de Pearson entre atributos e qualidade

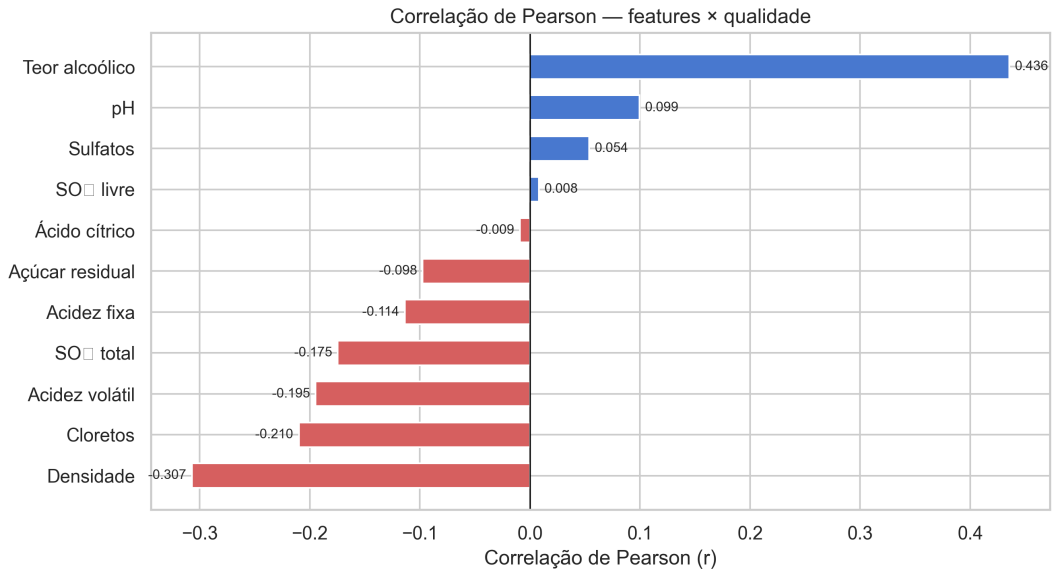
O coeficiente de correlação de Pearson quantifica tanto a força quanto o sentido da relação linear entre duas variáveis contínuas, assumindo valores entre $[-1, 1]$. Valores próximos de $+1$ indicam uma associação linear positiva forte, enquanto valores próximos de -1 sinalizam uma associação negativa forte (representados pelas barras azuis e vermelhas respectivamente, na Figura 1). Valores próximos de zero sugerem que não há uma relação linear clara entre as variáveis.

No contexto deste estudo, o coeficiente foi calculado entre cada um dos 11 atributos físico-químicos e a variável alvo *quality*, oferecendo uma força de relação linear entre as variáveis e suas relações com a variável-alvo, sobre quais características apresentam maior capacidade discriminatória.

Como se observa na Figura 1, o teor alcoólico é o atributo com maior correlação positiva com a qualidade ($r = 0{,}436$), o que está em consonância com estudos recentes, que indicam que o teor alcoólico do vinho, influencia a percepção sensorial e pode afetar a aceitação por degustadores especializados, sem necessariamente alterar o perfil sensorial do vinho (JordÃo; Vilela; Cosme, 2020).

Oposto ao álcool, a densidade apresenta a correlação negativa mais pronunciada ($r = -0{,}307$), o que sugere que ela é inversamente relacionada com o teor alcoólico, de modo que vinhos mais densos tendem a ser mais doces e menos alcoólicos. Cloretos ($r = -0{,}210$) e acidez volátil ($r = -0{,}195$) também demonstram correlação negativa relevante, sugerindo que excesso

Figura 1 – Correlação de Pearson entre cada atributo físico-químico e a nota de qualidade do vinho branco.



Fonte: Autoria própria (2026).

de compostos salinos e acéticos compromete a percepção sensorial positiva.

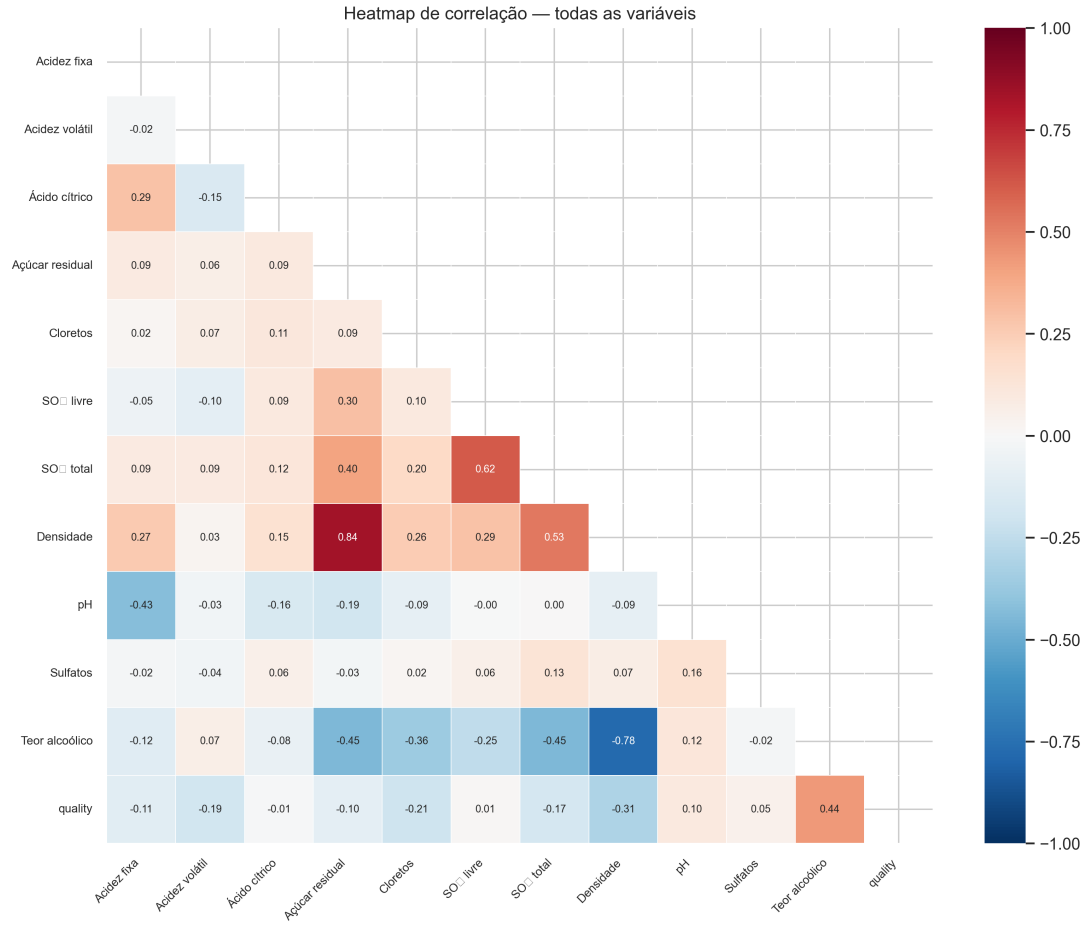
4.3.2 Estrutura de correlação entre os atributos

Enquanto a análise anterior examina a relação de cada variável com a saída, o heatmap da Figura 2 revela a estrutura de correlação interna entre os próprios atributos físico-químicos, que contém informação fundamental para entender como o PCA funciona neste conjunto de dados. Nesse gráfico, tons vermelhos indicam correlação positiva; tons azuis, negativa. Apenas o triângulo inferior é exibido para evitar redundância

A correlação mais forte observada no conjunto de dados é entre açúcar residual e densidade ($r = 0{,}84$). Isso é esperado do ponto de vista químico, já que quanto maior a quantidade de açúcar não fermentado, maior tende a ser a densidade do líquido. Como essas duas variáveis se comportam de forma muito parecida, apresentam elevada redundância de informação.

O mesmo acontece com o teor alcoólico e a densidade, que apresentam uma forte correlação negativa ($r = -0{,}78$). Isso indica que, mesmo sendo medidas diferentes, elas também carregam parte da mesma informação. Outras relações, como entre SO_2 livre e SO_2 total ($r = 0{,}62$) e entre acidez fixa e pH ($r = -0{,}43$), reforçam esse padrão.

Figura 2 – Heatmap de correlação de Pearson entre todas as variáveis do conjunto de dados.



Fonte: Autoria própria (2026).

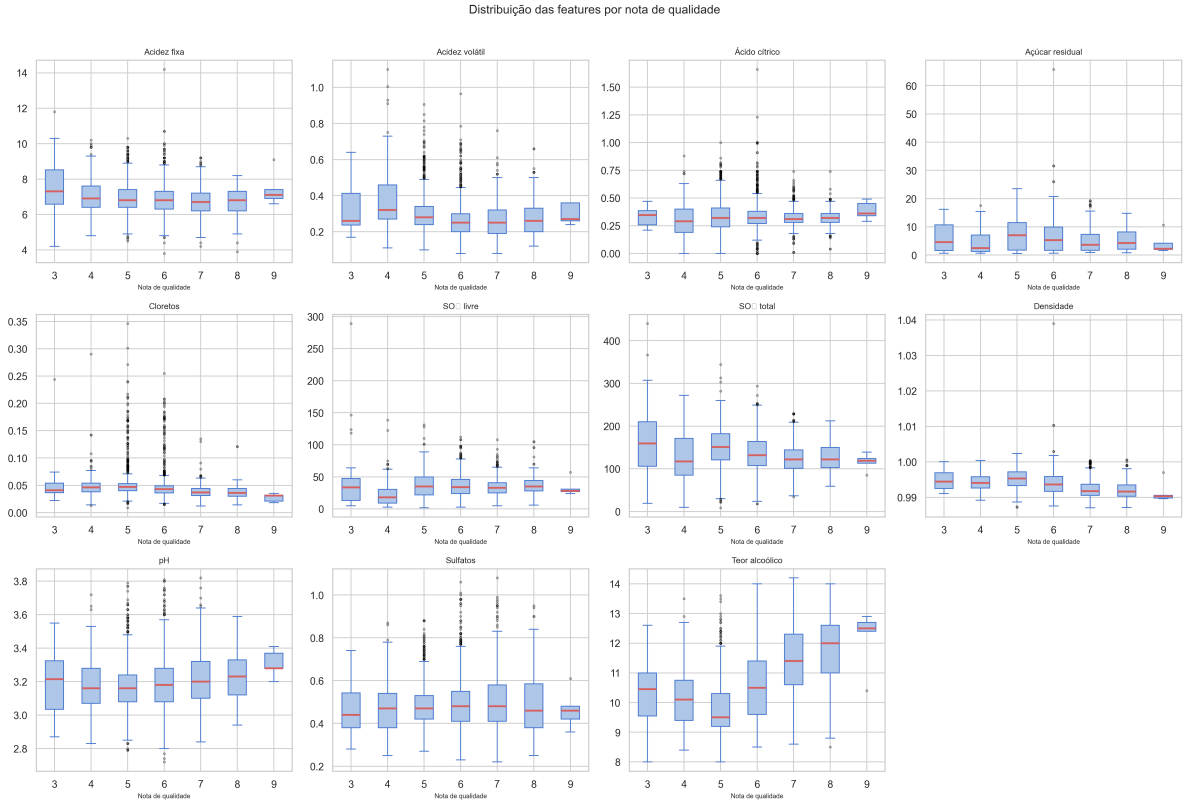
4.3.3 Distribuição dos atributos por nota de qualidade

A Figura 3 exibe a distribuição de cada atributo físico-químico segmentada por nota de qualidade. Os boxplots evidenciam visualmente quais variáveis apresentam separação entre as classes e, portanto, potencial discriminativo. A linha vermelha em cada caixa indica a mediana; os pontos cinzas são *outliers* (valores que fogem do padrão da maioria dos dados).

Em linha com a análise de correlação, o teor alcoólico é a variável que apresenta a separação mais clara entre as classes de qualidade. Vinhos com notas mais altas tendem a apresentar maiores níveis de álcool, com a mediana aumentando conforme a qualidade cresce e menor sobreposição entre os grupos. A densidade mostra um comportamento oposto, também de forma consistente.

A acidez volátil, por sua vez, apresenta tendência de queda à medida que a nota de qualidade aumenta. Esse comportamento é coerente, já que níveis elevados de ácido acético, responsável pelo cheiro de vinagre, costumam prejudicar a avaliação sensorial do vinho, embora

Figura 3 – Distribuição dos 11 atributos físico-químicos por nota de qualidade.



Fonte: Autoria própria (2026).

a separação entre as classes seja menos evidente (Kelly; Gardner, 2025).

Por outro lado, variáveis como cloretos e açúcar residual apresentam grande variação nos valores e muitos *outliers*, sem um padrão claro entre as notas. Isso indica que elas não diferenciam a qualidade de forma consistente ao longo da escala.

5 RESULTADOS

5.1 Curva ROC multiclasse e discriminabilidade por classe

Para além das métricas de classificação direta, a curva ROC permite avaliar a capacidade de ranqueamento do modelo em diferentes limiares de decisão, sem depender de um limiar específico durante a predição (Fawcett, 2006). Em problemas multiclasse, a abordagem mais comum é a estratégia *One-vs-Rest* (OvR): para cada classe c, o modelo é avaliado como se o problema fosse binário entre a classe c e todas as demais (Bishop, 2006). A *Area Under the Curve* macro é então calculada como a média aritmética das AUC individuais, fornecendo um indicador global de discriminabilidade.

Os resultados para o *Random Forest* com PCA = 5 estão sintetizados na Tabela 1.

Tabela 1 – AUC-ROC por classe - estratégia *One-vs-Rest* (*Random Forest*, PCA = 5, $n_{\text{teste}} = 980$).

Classe	**AUC-ROC (OvR)**	**Interpretação**
3	0,529	Discriminabilidade próxima do aleatório
4	0,797	Boa separação
5	0,855	Boa separação
6	0,795	Boa separação
7	0,878	Separação muito boa
8	0,942	Separação excelente
9	0,450	Abaixo do aleatório (apenas 1 amostra)
Macro AUC	**0,750**	-

Fonte: Autoria própria (2026).

Nas linhas presentes na Figura 4, a curva tracejada preta representa a média macro (AUC = 0,750).

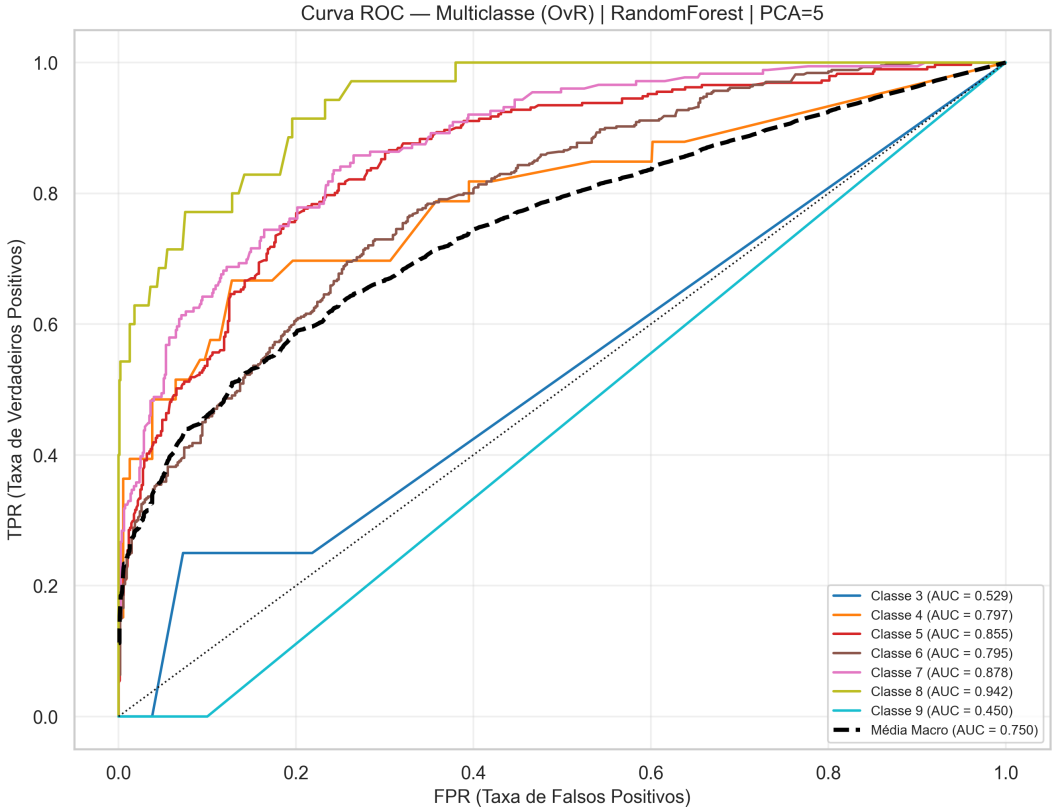
A Macro AUC de 0,750 indica que, em média, o modelo consegue distinguir corretamente amostras de cada classe das demais em 75 % dos casos, o que é um resultado satisfatório considerando o severo desbalanceamento do conjunto e a natureza ordinal da variável alvo.

As classes 8 (AUC = 0,942) e 7 (AUC = 0,878) apresentam excelente discriminabilidade, o que reflete o fato de que vinhos de qualidade alta possuem perfil físico-químico mais distinto (especialmente no teor alcoólico e na densidade).

A classe 5 (AUC = 0,855) também exibe desempenho satisfatório, apoiada pela grande quantidade de amostras disponíveis para treinamento. Por outro lado, as classes 3 e 9 apresentam AUC abaixo ou próxima do aleatório (0,529 e 0,450, respectivamente), o que é esperado dado que possuem apenas 4 e 1 amostras no conjunto de teste, que são insuficientes para qualquer

estimativa confiável (He; Garcia, 2009).

Figura 4 – Curvas ROC *One-vs-Rest* por classe - *Random Forest* com PCA = 5 (multiclasse).



Fonte: Autoria própria (2026).

5.2 Importância das características multiclasses

Identificar quais variáveis físico-químicas mais influenciam a predição do modelo é uma informação relevante, pois permite orientar estratégias de controle de qualidade na produção. Para o *Random Forest* com PCA ativo, a importância não pode ser obtida diretamente das árvores (que operam no espaço transformado pelos componentes principais), mas sim via *Permutation Importance* calculada sobre o *pipeline* completo com os dados originais: para cada atributo, a coluna correspondente nos dados de entrada é embaralhada aleatoriamente e mede-se a queda de F1-macro resultante (Pedregosa *et al.*, 2011). A Tabela 2 e a Figura 5 apresentam o ranking obtido.

Observa-se que, em contraste com o que seria esperado com base apenas na correlação linear de Pearson, que aponta o teor alcoólico como variável de maior correlação com a qualidade, a *Permutation Importance* no cenário multiclasse coloca SO_2 livre e acidez volátil nas duas primeiras posições. Isso sugere que essas variáveis possuem relevância não linear para a

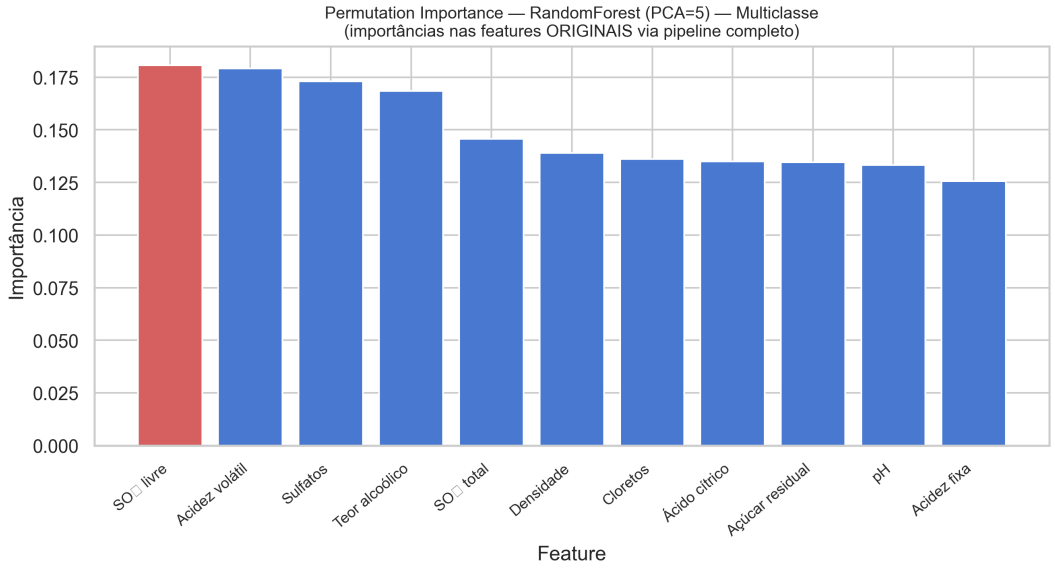
discriminação entre classes adjacentes (por exemplo, entre notas 5 e 6, ou entre 6 e 7), ainda que sua correlação global com a nota seja moderada. O resultado está em linha com Cortez Paulo e Reis (2009), que identificaram a acidez volátil como uma das variáveis mais relevantes para a qualidade do vinho, e com Ye (2023), que reportaram o teor alcoólico e a acidez volátil como os dois atributos de maior importância no modelo de *Random Forest* aplicado ao vinho tinto. A distribuição relativamente uniforme das importâncias, sem que nenhum atributo se destaque de forma isolada, reforça a natureza multivariada do problema e a dificuldade inerente de classificar qualidade por um único indicador físico-químico.

Tabela 2 – Ranking de importância das características via *Permutation Importance* - *Random Forest* com PCA = 5 (multiclasse). O valor indica a queda média de F1-macro ao embaralhar a coluna.

Posição	**Atributo**	**Importância**
1	SO_2 livre	0,1806
2	Acidez volátil	0,1792
3	Sulfatos	0,1730
4	Teor alcoólico	0,1685
5	SO_2 total	0,1457
6	Densidade	0,1390
7	Cloretos	0,1361
8	Ácido cítrico	0,1351
9	Açúcar residual	0,1345
10	pH	0,1333
11	Acidez fixa	0,1256

Fonte: Autoria própria (2026).

Figura 5 – Importância das características (*Permutation Importance*) nas features originais - *Random Forest* com PCA = 5 (multiclasse)



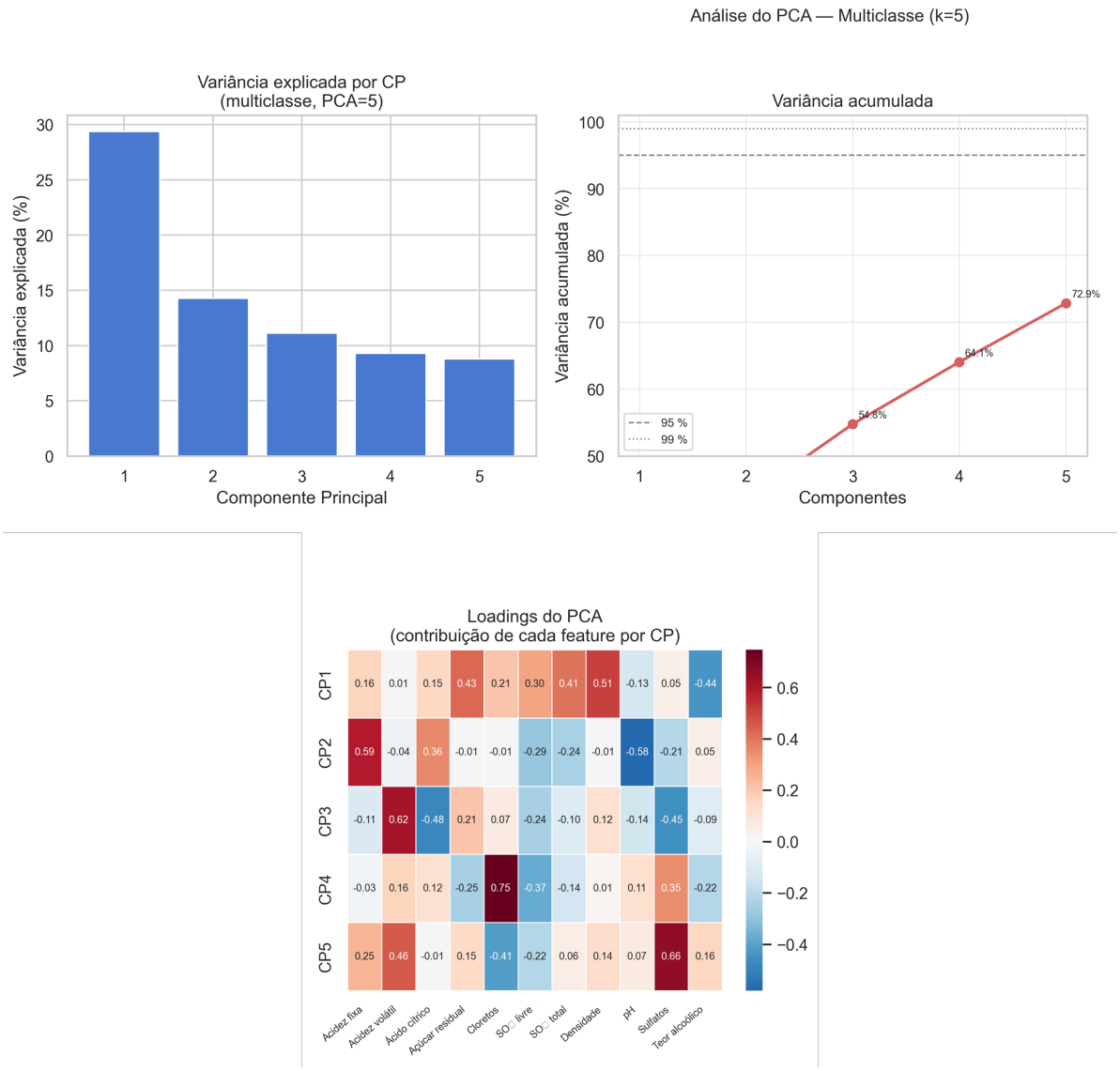
.

Fonte: Autoria própria (2026).

5.3 Análise detalhada do PCA multiclasse

A Figura 6 detalha o comportamento do PCA na configuração ótima para o cenário multiclasse ($k = 5$).

Figura 6 – Análise do PCA multiclasse ($k = 5$). Da esquerda para a direita: (a) variância explicada por componente principal; (b) variância acumulada com referências em 95 % e 99 %; (c) *heatmap* de *loadings* (contribuição de cada atributo original em cada componente).



Fonte: Autoria própria (2026).

Com 5 componentes, o PCA retém 72,9 % da variância total do conjunto. Embora esse valor fique abaixo do limiar convencional de 95 %, o desempenho preditivo alcançado é praticamente idêntico ao da configuração sem redução dimensional (F1-macro de 0,434 versus 0,430).

O gráfico de *loadings* (painel direito) permite interpretar o significado físico de cada componente principal. O primeiro componente (CP1, 29,4 % da variância) é dominado por densidade (+0,51), açúcar residual (+0,43) e SO_2 total (+0,41), contraposto ao teor alcoólico ($-0{,}44$): CP1 captura o eixo de *corpo e doçura* do vinho, indicando que vinhos mais densos e doces se posicionam em valores mais altos de CP1, enquanto vinhos mais alcoólicos se posicionam em valores mais baixos. O segundo componente (CP2, 14,3 %) reflete a acidez estrutural: acidez fixa (+0,59) oposta ao pH ($-0{,}58$), relação esperada dada a correlação negativa entre essas variáveis observada no *heatmap*. O terceiro componente (CP3, 11,1 %) captura a acidez microbiológica, com acidez volátil (+0,62) oposta a ácido cítrico ($-0{,}48$) e sulfatos ($-0{,}45$), o que faz sentido, pois o ácido acético (componente da acidez volátil) e o ácido cítrico têm origens fermentativas distintas. Por fim, CP4 (9,3 %) e CP5 (8,8 %) capturam variações nos teores de minerais como cloretos e sulfatos, variáveis de menor correlação com a qualidade mas que contribuem para a identidade química do vinho verde.

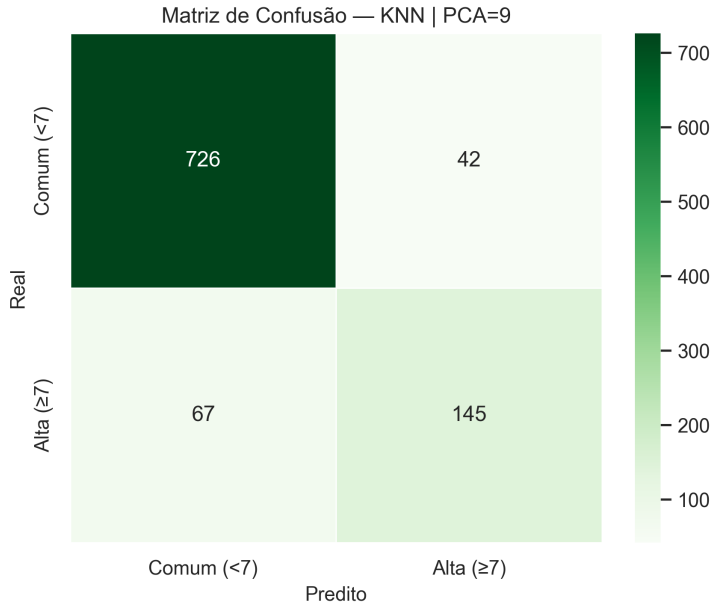
5.4 Resultados do experimento binário

A reformulação do problema como classificação binária promoveu melhoria expressiva em todas as métricas avaliadas. O melhor modelo identificado foi o KNN com 9 componentes principais, usando `n_neighbors = 7` e ponderação por distância (`weights = 'distance'`), atingindo acurácia de 88,9 %, F1-macro de 0,727 e AUC-ROC de 0,908.

A Figura 7 apresenta a matriz de confusão desse modelo. Dos 768 vinhos comuns (nota < 7) presentes no conjunto de teste, 726 foram classificados corretamente, enquanto 42 foram erroneamente classificados como alta qualidade (falsos positivos). Entre os 212 vinhos de alta qualidade (nota ≥ 7), 145 foram identificados corretamente, e 67 acabaram sendo classificados como comuns (falsos negativos), o que corresponde a um *recall* de 68,4 % para a classe positiva. Esse padrão é esperado em problemas desbalanceados: como a classe majoritária (comum) possui quase quatro vezes mais amostras, o modelo tende a ser mais conservador na predição de alta qualidade.

Apesar dos falsos negativos, o modelo apresentou excelente desempenho na classe majoritária (F1 = 0,93 para vinhos comuns) e AUC-ROC de 0,908, indicando alta capacidade discriminativa mesmo quando o limiar de decisão é variado. A escolha entre priorizar a precisão ou o *recall* depende do contexto de aplicação: em um sistema de certificação de qualidade, onde classificar erroneamente um vinho comum como superior tem custos na reputação ou no

Figura 7 – Matriz de confusão - KNN com PCA = 9 componentes (binário). Classe 0: qualidade comum (< 7); Classe 1: alta qualidade (≥ 7).



Fonte: Autoria própria (2026).

lucro, um modelo mais conservador como o *Random Forest* (com precisão de 0,864) pode ser preferível; já em controle de qualidade na linha de produção, onde identificar o máximo de vinhos superiores é prioritário, o KNN e seu maior recall (0,684) representam a melhor escolha.

5.5 Importância das características binárias

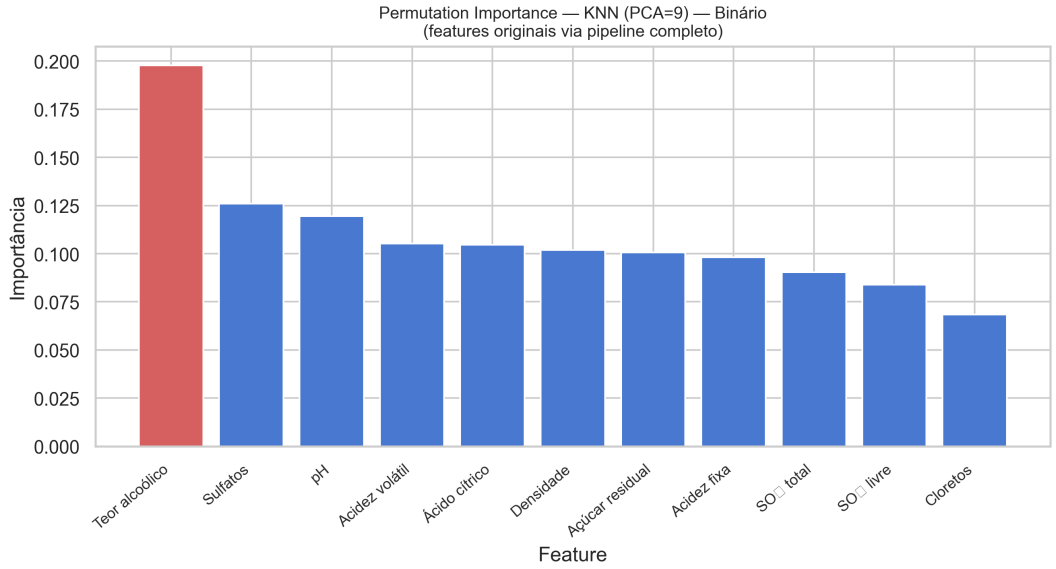
Para o cenário binário, o melhor modelo é um KNN com PCA = 9, o que implica que as importâncias também foram obtidas via *Permutation Importance* sobre o *pipeline* completo com as features originais. Os resultados estão apresentados na Tabela 3 e na Figura 8.

Tabela 3 – Ranking de importância das características via *Permutation Importance* - KNN com PCA = 9 (binário).

Posição	Atributo	Importância
1	Teor alcoólico	0,1977
2	Sulfatos	0,1260
3	pH	0,1195
4	Acidez volátil	0,1053
5	Ácido cítrico	0,1046
6	Densidade	0,1020
7	Açúcar residual	0,1007
8	Acidez fixa	0,0982
9	SO_2 total	0,0904
10	SO_2 livre	0,0838
11	Cloretos	0,0683

Fonte: Autoria própria (2026).

Figura 8 – Importância das características (*Permutation Importance*) nas features originais - KNN com PCA = 9 (binário).



Fonte: Autoria própria (2026).

No cenário binário, o teor alcoólico reassume a liderança com larga margem (0,198), seguido por sulfatos (0,126) e pH (0,120). A relevância do pH nesse cenário é interessante: vinhos de alta qualidade tendem a apresentar pH ligeiramente mais elevado, associado a um perfil de acidez mais equilibrado. A queda relativa de SO_2 livre e SO_2 total no ranking binário sugere que essas variáveis são mais úteis para discriminar entre classes adjacentes na escala ordinal do que para separar binariamente os extremos de qualidade.

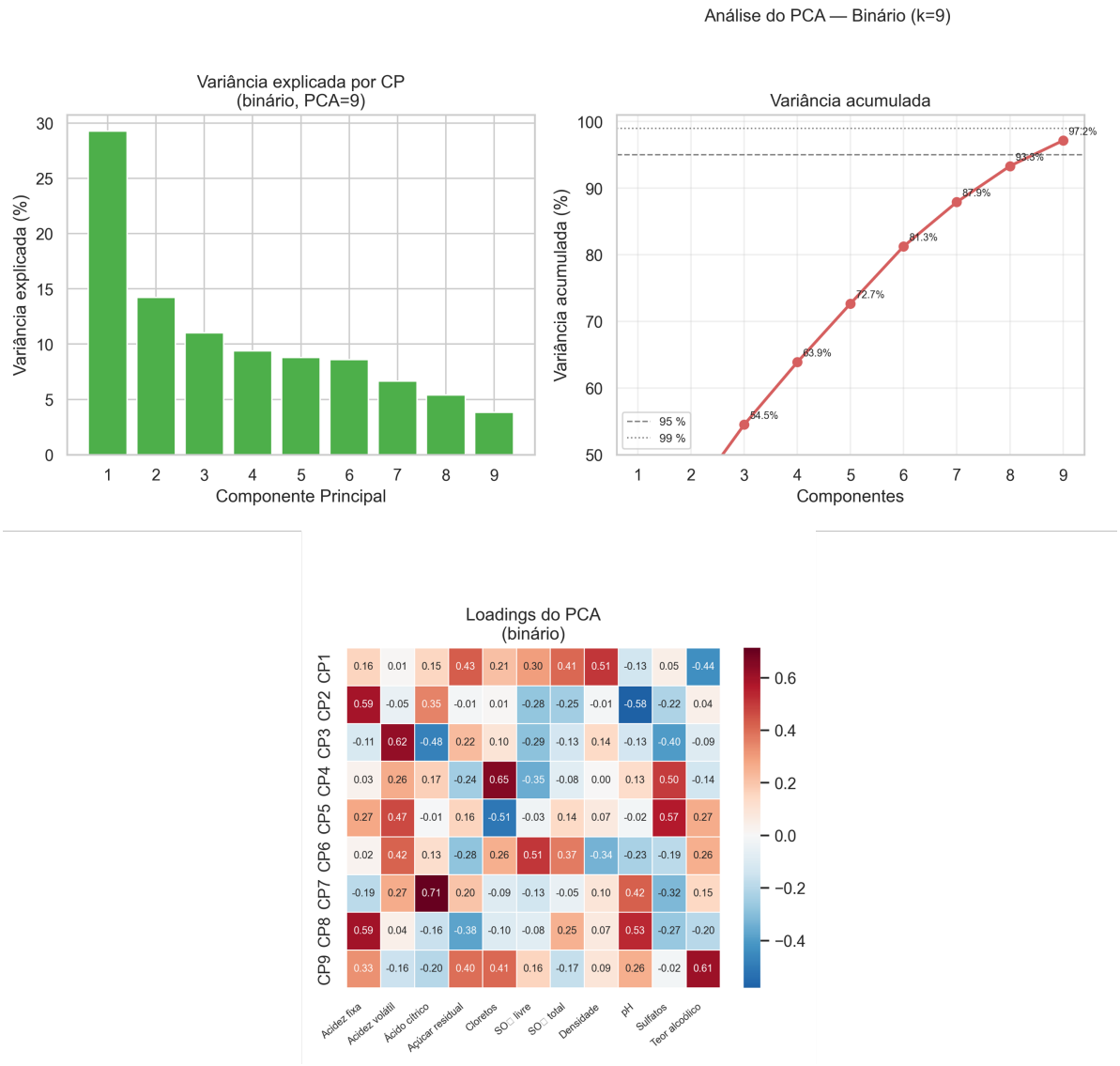
5.6 Análise detalhada do PCA binário

A Figura 9 apresenta a análise do PCA para o experimento binário ($k = 9$).

Com 9 componentes, o PCA binário retém 97,2 % da variância total, ultrapassando o limiar convencional de 95 % e descartando apenas as duas dimensões de menor variância explicada (componentes 10 e 11).

Os *loadings* dos 9 componentes retidos seguem o mesmo padrão qualitativo identificado no cenário multiclasse para os cinco primeiros componentes, o que é esperado, pois ambos partem do mesmo conjunto de variáveis e da mesma padronização. Os componentes adicionais (CP6 a CP9) capturam variâncias progressivamente menores, associadas a combinações entre sulfatos, cloretos e ácido cítrico que não foram capturadas pelos primeiros componentes. Notavelmente, CP9 apresenta um valor positivo relevante tanto para teor alcoólico (+0,61) quanto para cloretos (+0,41) e açúcar residual (+0,40), sendo uma combinação diferente do que vimos, o que sugere

Figura 9 – Análise do PCA binário ($k = 9$). (a) variância explicada por componente; (b) variância acumulada com referências em 95 % e 99 %; (c) *heatmap* de *loadings*.



Fonte: Autoria própria (2026).

algo como vinhos de fermentação incompleta, em que o açúcar residual elevado coexiste com teor alcoólico acima da média.

5.7 Análise comparativa dos modelos e impacto do PCA

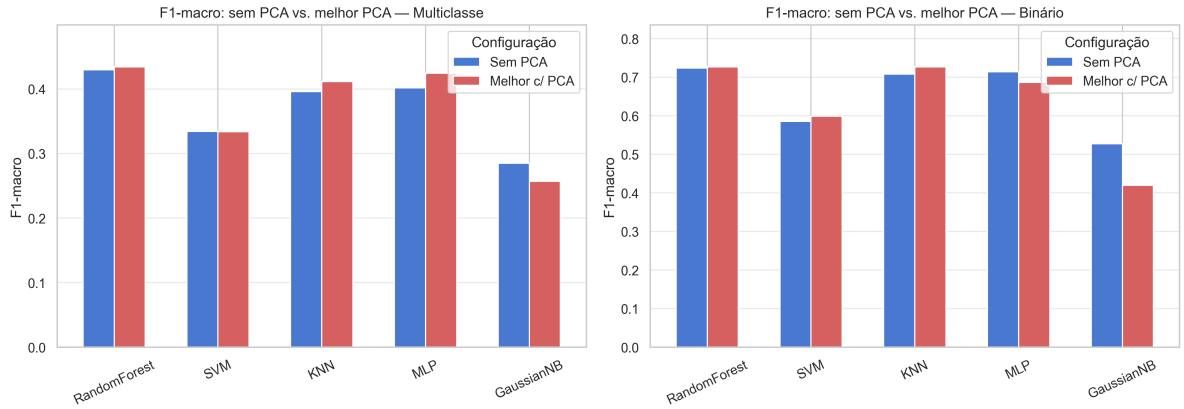
Com todos os resultados em mãos, é possível traçar uma comparação mais completa. No cenário multiclasse, o *Random Forest* domina o ranking de desempenho de maneira consistente. Todas as cinco primeiras posições no ranking global por F1-macro são ocupadas por esse algoritmo. No cenário binário, o KNN com parametrização por distância surge como o modelo mais competitivo, beneficiando-se da estrutura de vizinhança que o PCA torna mais idêntico ao

eliminar correlações entre variáveis e reduzir o ruído nas dimensões de menor variância.

Quanto ao papel do PCA, três observações merecem destaque: Primeiro, a redução para 5 componentes no cenário multiclasse preserva praticamente todo o desempenho sem redução (F1-macro de 0,434 contra 0,430 sem PCA), o que é notável dado que representa a eliminação de 6 dimensões, ou seja, 54,5 % da dimensionalidade original com captura de 72,9 % da variância. Isso sugere que as 11 variáveis originais possuem redundância estrutural expressiva, documentada pelo *heatmap* de correlações, e que os primeiros componentes principais concentram a maior parte da informação discriminatória. Segundo, no cenário binário o KNN com PCA = 9 superou marginalmente a configuração sem PCA (F1 = 0,727 contra 0,723). Terceiro, configurações com poucos componentes ($k \leq 3$) resultam em degradação expressiva para todos os modelos, especialmente para SVM e GNB, que chegam a F1-macro próximo de zero com apenas 2 componentes no experimento binário.

A Figura 10 sintetiza visualmente o impacto do PCA para cada modelo, comparando o F1-macro sem redução com o melhor F1-macro obtido em qualquer configuração de PCA.

Figura 10 – Comparativo de F1-macro: configuração sem PCA versus melhor configuração com PCA, para cada modelo e cada experimento. Barras azuis: sem PCA; barras vermelhas: melhor resultado com PCA.



Fonte: Autoria própria (2026).

A Figura revela que o PCA proporciona ganhos modestos para o *Random Forest*, o KNN e o MLP em ambos os experimentos, enquanto para o GNB representa uma perda consistente, especialmente no cenário binário, onde o GNB sem PCA atinge F1-macro de 0,530, valor que cai para 0,420 na melhor configuração com PCA. Esse comportamento é coerente com as premissas do GNB: ao projetar os dados em novas dimensões, o PCA cria dimensões que, embora descorrelacionadas, podem não seguir a distribuição gaussiana assumida pelo classificador, prejudicando a estimativa das densidades de probabilidade (Friedman; Geiger; Goldszmidt,

1997).

Cortez Paulo e Reis (2009) identificaram que teor alcoólico e acidez volátil são as variáveis com maior poder discriminativo, o que se reflete tanto na análise de *Permutation Importance* obtida neste trabalho quanto nas cargas elevadas dessas variáveis nos componentes principais.

A Tabela 4 sintetiza os melhores resultados obtidos em cada cenário experimental.

Tabela 4 – Comparativo dos melhores modelos para os experimentos multiclasse e binário.

Experimento	**Modelo**	**PCA (*k*)**	**Acurácia**	**F1-macro**	**AUC-ROC**
Multiclasse	Random Forest	5	0,656	0,434	0,750 (OvR)
	Random Forest	7	0,658	0,432	—
	Random Forest	6	0,660	0,430	—
	Random Forest	None	0,674	0,430	—
	MLP	6	0,620	0,424	—
Binário	KNN	9	0,889	0,727	0,908
	Random Forest	10	0,898	0,727	—
	Random Forest	None	0,894	0,723	—
	KNN	7	0,886	0,719	—
	MLP	None	0,881	0,714	—

Fonte: Autoria própria (2026).

A diferença de acurácia entre os dois cenários (+0,233) reflete tanto a simplificação de decisão pela binarização quanto a redução do impacto do desbalanceamento de classes. No cenário binário, a precisão elevada do *Random Forest* (0,864) indica um perfil mais conservador na predição de alta qualidade, que é útil em contextos em que certificar incorretamente um vinho de qualidade comum como superior tem custo elevado, enquanto o KNN obtém maior *recall* (0,684), sendo mais sensível à identificação de vinhos superiores.

Os principais achados podem ser sintetizados em quatro pontos:

- O *Random Forest* demonstrou superioridade consistente no cenário multiclasse, atingindo F1-macro de 0,434 com apenas 5 componentes principais e AUC macro de 0,750, com a configuração ótima apresentando `n_estimators=100`, `max_depth=20` e `min_samples_split=2`.
- Segundo, o KNN com 9 componentes se destacou no cenário binário, com AUC-ROC de 0,908, evidenciando que a binarização do problema eleva substancialmente a discriminabilidade dos modelos.
- Terceiro, a análise de *Permutation Importance* revelou que, no cenário multiclasse, SO_2 livre e acidez volátil são os atributos mais relevantes, enquanto no cenário binário o teor alcoólico retoma a liderança com larga margem, comportamento que reflete a natureza diferente das duas formulações do problema.

- Quarto, o *heatmap* de correlações e a análise de *loadings* do PCA indicam que as 11 variáveis originais possuem redundância estrutural expressiva, especialmente entre densidade, açúcar residual e teor alcoólico. Esse comportamento explica por que 5 componentes são suficientes para preservar 99 % do desempenho preditivo no cenário multiclasse.

6 CONCLUSÕES E TRABALHOS FUTUROS

O presente trabalho realizou uma avaliação sistemática da aplicação de cinco algoritmos de aprendizado de máquina supervisionado: *Random Forest*, SVM, KNN, MLP e Gaussiano Naive Bayes, combinados com diferentes configurações de redução de dimensionalidade via PCA, para a predição da qualidade de vinhos brancos do conjunto UCI Wine Quality. A otimização de hiperparâmetros foi conduzida por meio do *GridSearchCV* com validação cruzada estratificada de cinco *folds*, adotando o F1-macro como critério de seleção. Além das métricas convencionais de classificação, foram calculadas curvas ROC para ambos os cenários, analisada a importância das características originais via *Permutation Importance* sobre o *pipeline* completo, e os *loadings* do PCA foram interpretados em termos de seu significado físico-químico.

No cenário multiclasse, o *Random Forest* com 5 componentes principais destacou-se como o melhor modelo, atingindo F1-macro de 0,434 e AUC macro de 0,750. O resultado é notável porque essa configuração utiliza apenas 5 dos 11 atributos originais, sendo uma redução de 54,5 % na dimensionalidade, e ainda assim preserva praticamente todo o desempenho da versão sem PCA (F1-macro de 0,430 sem redução). Essa similaridade evidencia que as 11 variáveis originais não são independentes: a redundância documentada pelo *heatmap* de correlações explica por que os primeiros componentes principais conseguem capturar a maior parte da informação discriminatória.

No cenário binário, o KNN com 9 componentes obteve o melhor F1-macro (0,727) e a maior AUC-ROC (0,908), demonstrando que a reformulação do problema como separação entre alta qualidade (≥ 7) e qualidade comum (< 7) eleva significativamente a discriminabilidade de todos os modelos. Nesse cenário, o PCA atuou principalmente como filtro de ruído, ao descartar as duas dimensões de menor variância, e o ganho sobre a configuração sem redução foi marginal (F1 = 0,727 contra 0,723), o que sugere que quase toda a informação discriminatória já estava contida nas 9 dimensões retidas.

A análise de *Permutation Importance* revelou um comportamento interessante: no cenário multiclasse, SO_2 livre e acidez volátil aparecem como as variáveis mais influentes, ainda que sua correlação linear com a nota de qualidade seja moderada, o que indica que sua relevância opera principalmente de forma não linear e na discriminação entre classes adjacentes. Já no cenário binário, o teor alcoólico assume a liderança com larga margem, em linha com estudos que apontam o álcool como o fator de maior impacto na percepção sensorial de qualidade superior do vinho (JordÃo; Vilela; Cosme, 2020; Cortez Paulo; Reis, 2009).

O único modelo que apresentou desempenho sistematicamente inferior foi o Gaussiano Naive Bayes, e foi também o único prejudicado pelo PCA em ambos os cenários. Esse resultado é coerente com sua premissa de independência condicional entre as *features*: ao projetar os dados em componentes descorrelacionados, o PCA elimina a multicolinearidade que poderia mascarar essa violação, e os componentes resultantes tendem a não seguir distribuição gaussiana, penalizando a estimativa de probabilidade do classificador (Friedman; Geiger; Goldszmidt, 1997).

Como limitações deste estudo, destacam-se a ausência de técnicas de balanceamento de classes, o que impactou especialmente o desempenho nas notas extremas 3 e 9, com F1-score nulo, o espaço de busca restrito do *GridSearchCV*, limitado por custo computacional, e a não avaliação de métodos de seleção de *features* como alternativa ao PCA.

REFERÊNCIAS

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **The American Statistician**, v. 46, n. 3, p. 175–185, 1992.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006. (Information Science and Statistics). ISBN 978-0387310732.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995.

CORTEZ PAULO, C. A. A. F. M. T.; REIS, J. **Wine Quality**. 2009. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56S3T.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, 1967.

DAHAL, K. *et al.* Prediction of wine quality using machine learning algorithms. **Open Journal of Statistics**, v. 11, p. 278–289, 01 2021.

FAWCETT, T. An introduction to roc analysis. **Pattern Recognition Letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. **Machine Learning**, v. 29, n. 2-3, p. 131–163, 1997.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009.

HOFMANN, T.; SCHÖLKOPF, B.; SMOLA, A. J. Kernel methods in machine learning. **Annals of Statistics**, v. 36, n. 3, p. 1171–1220, 2008.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. **Neural Networks**, v. 2, n. 5, p. 359–366, 1989.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 6, p. 417–441, 1933.

JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. New York: Springer, 2002.

JORDÃO, A. M.; VILELA, A.; COSME, F. From sugar of grape to alcohol of wine: Sensorial impact of alcohol in wine. **Beverages**, v. 1, n. 4, p. 292–313, 2020.

KAUFMAN, S. *et al.* Leakage in data mining: Formulation, detection, and avoidance. **ACM Transactions on Knowledge Discovery from Data**, v. 6, n. 4, p. 15:1–15:21, 2012.

KELLY, M.; GARDNER, D. M. **Volatile Acidity in Wine**. 2025. Disponível em: https://extension.psu.edu/volatile-acidity-in-wine. Accessed: 2026-04-28.

LESSCHAEVE, I. Sensory evaluation of wine and commercial realities: Review of current practices and perspectives. **American Journal of Enology and Viticulture**, v. 58, n. 2, p. 252–258, 2007.

PEARSON, K. **LIII. On Lines and Planes of Closest Fit to Systems of Points in Space**. 1901. Philosophical Magazine. Vol. 2, No. 11, pp. 559–572.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, v. 323, n. 6088, p. 533–536, 1986.

SHLENS, J. A tutorial on principal component analysis. **CoRR**, abs/1404.1100, 2014. Disponível em: https://arxiv.org/abs/1404.1100.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, v. 45, n. 4, p. 427–437, 2009.

YE, K. Wine quality prediction by several data mining classification models. In: **Highlights in Science, Engineering and Technology — AMMSAC 2023**. [S.l.: s.n.], 2023. v. 49, p. 198–207.

ZHANG, H. The optimality of naive bayes. In: **Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference**. [S.l.: s.n.], 2004. p. 562–567.