Superioridade de Métodos Baseados em Instância na Classificação Sensorial de Vinhos Portugueses.

Matheus Simão Sales
Engenharia de Computação
Universidade Federal do Ceará
matheussimao@alu.ufc.br

***Resumo*—Este artigo investiga a aplicação de algoritmos de aprendizado de máquina para a classificação da qualidade de vinhos tintos e brancos com base em suas características físico-químicas. Foram avaliados e comparados cinco modelos supervisionados: *k-Nearest Neighbors* (k-NN), *Support Vector Machine* (SVM), Rede Neural *Multi-Layer Perceptron* (MLP), Regressão Logística e Análise Discriminante Linear (LDA). Um pré-processamento rigoroso dos dados foi realizado, incluindo a transformação de potência Yeo-Johnson para correção de assimetria e posterior padronização (*z-score*). Os resultados demonstram que o algoritmo k-NN (com k=19) obteve o melhor desempenho geral, alcançando uma acurácia de 88,06% e uma área sob a curva ROC (AUC-ROC) de 0,9129. O SVM apresentou a maior precisão (76,58%), mas com sensibilidade (*recall*) inferior. A análise de importância das variáveis nos modelos lineares identificou o teor alcoólico, a densidade e o açúcar residual como os preditores mais determinantes para a qualidade. Conclui-se que, para este conjunto de dados e tarefa, métodos baseados em instância, como o k-NN, são mais eficazes do que modelos paramétricos lineares ou redes neurais de pequeno porte, provavelmente devido à natureza não-linear das relações entre os atributos. O estudo fornece uma base quantitativa para apoio à decisão enológica, sugerindo a viabilidade do uso de aprendizado de máquina como ferramenta complementar na avaliação objetiva da qualidade do vinho.**

I. INTRODUÇÃO

A indústria vitivinícola representa um setor de significativa importância econômica e cultural em diversos países, particularmente em Portugal, onde a produção de vinho constitui uma tradição secular e um relevante contributo para a economia nacional. Tradicionalmente, a avaliação da qualidade do vinho baseia-se em análises sensoriais realizadas por especialistas, um processo subjetivo, dependente de fatores humanos e de difícil padronização. Nas últimas décadas, a aplicação de técnicas analíticas físico-químicas e métodos computacionais de aprendizado de máquina tem emergido como uma abordagem complementar para prever e compreender os fatores objetivos que influenciam a qualidade percebida dos vinhos [1].

Enquanto estudos anteriores focaram em **modelos de regressão** para prever a nota contínua de qualidade, a **classificação** da qualidade do vinho em categorias discretas (por exemplo, ”alta”versus ”baixa”qualidade) é igualmente relevante para a indústria, pois permite uma triagem rápida, a automatização de controles de qualidade e decisões binárias que orientam a comercialização e o aprimoramento dos processos. A análise de classificação, cujos fundamentos remontam aos trabalhos de Fisher sobre análise discriminante, tem-se mostrado uma ferramenta poderosa nesse contexto [2]. Seu objetivo é modelar a relação entre um conjunto de variáveis preditoras e uma variável resposta categórica, atribuindo cada observação a uma classe pré-definida.

No domínio da enologia, modelos de classificação permitem categorizar vinhos como de alta ou baixa qualidade com base em parâmetros mensuráveis, como acidez, teor alcoólico, concentração de compostos voláteis e outros atributos físico-químicos. Diversos estudos já demonstraram a viabilidade dessa abordagem. O trabalho seminal de Cortez et al. [1] aplicou métodos de mineração de dados, incluindo classificação, para modelar a preferência de vinhos portugueses da região do Vinho Verde. Outros pesquisadores exploraram técnicas como Análise Discriminante Linear (LDA) e Quadrática (QDA) [3], redes neurais artificiais [4], k-vizinhos mais próximos (k-NN) [5] e Máquinas de Vetor de Suporte (SVM) [6] para problemas de classificação em dados quimiométricos.

Este trabalho tem como objetivo principal realizar uma comparação abrangente entre **modelos lineares e não-lineares de classificação** aplicados à categorização da qualidade sensorial de vinhos. Utilizando o mesmo conjunto de dados público do estudo anterior (que integra informações de vinhos tintos e brancos da região do Vinho Verde, totalizando 6.497 observações e 11 atributos físico-químicos), transformamos a variável resposta contínua em uma variável binária, considerando vinhos com nota ≥ 7 como de **alta qualidade** e os demais como de **baixa qualidade**. Foram implementados e avaliados cinco tipos de modelos:

- **Modelos lineares**: Regressão Logística e Análise Discriminante Linear (LDA);
- **Modelos não-lineares**: Máquinas de Vetor de Suporte (SVM) com kernel radial (RBF), k-Vizinhos Mais Próximos (k-NN) e Redes Neurais Artificiais (MLP).

A metodologia incluiu pré-processamento dos dados (transformação Yeo-Johnson e padronização), ajuste de hiperparâmetros por validação cruzada e avaliação rigorosa por meio de métricas como acurácia, precisão, recall, F1-score e AUC-ROC. Um aspecto crucial foi a análise das matrizes de confusão, que permitem identificar os tipos de erro cometidos por cada modelo (falsos positivos e falsos negativos).

A necessidade desta análise comparativa justifica-se pela complexidade inerente à avaliação da qualidade do vinho. Diferentes técnicas de classificação possuem pressupostos, vantagens e limitações distintas. Modelos lineares são inter-

pretáveis e computacionalmente eficientes, mas podem falhar em capturar interações e relações não-lineares presentes nos dados. Modelos não-lineares, como redes neurais e SVM com kernel não-linear, oferecem maior flexibilidade e capacidade de modelar padrões complexos, porém ao custo de interpretabilidade e maior risco de sobreajuste. Portanto, identificar qual abordagem oferece o melhor equilíbrio entre desempenho preditivo e robustez para este conjunto de dados específico é uma questão de relevância tanto prática quanto acadêmica.

As aplicações práticas deste estudo estendem-se a múltiplos agentes da cadeia produtiva. Produtores podem utilizar os modelos de classificação para realizar uma triagem automática da qualidade de lotes com base em análises laboratoriais de rotina, antes mesmo da degustação. Engenheiros de processo podem identificar, por meio dos coeficientes dos modelos lineares, quais variáveis físico-químicas são mais críticas para a classificação, direcionando esforços de controle. Por fim, o estudo contribui para o corpo de conhecimento em quimiometria e aprendizado de máquina, ilustrando um *pipeline* completo de análise de dados para classificação, desde a exploração e pré-processamento até a modelagem avançada e comparação rigorosa de desempenho.

A Seção II deste artigo descreve detalhadamente a metodologia empregada, incluindo a origem e características dos dados, as etapas de pré-processamento, e a fundamentação teórica e implementação dos modelos de classificação. A Seção III apresenta e discute os resultados obtidos, com ênfase na comparação de desempenho entre os modelos e na análise dos tipos de erro. Por fim, a Seção IV sumariza as principais conclusões e sugere direções para trabalhos futuros.

II. METODOLOGIA

Este estudo adota uma abordagem metodológica estruturada em cinco etapas principais: (1) preparação e análise exploratória dos dados; (2) transformação da variável alvo em binária; (3) pré-processamento das características preditoras; (4) implementação e validação de modelos de classificação; e (5) análise comparativa de desempenho. Todas as análises foram implementadas em Python (versão 3.9), utilizando bibliotecas como NumPy, pandas, scikit-learn e Matplotlib.

A. Dados e Transformação da Variável Alvo

Utilizamos o mesmo conjunto de dados público de vinhos portugueses Vinho Verde do repositório UCI Machine Learning [7], combinando as versões tinta (1.599 amostras) e branca (4.898 amostras), totalizando 6.497 observações. Cada amostra possui 11 atributos físico-químicos quantitativos, descritos na Tabela I, e uma nota de qualidade sensorial (variável alvo original) em escala de 0 a 10 (observada de 3 a 9).

Para transformar o problema de regressão em classificação, a variável contínua *Qualidade* foi convertida em uma variável binária *Classe_Binaria*, utilizando o limiar de corte 7 (nota $\geq$ 7 = alta qualidade, nota $<$ 7 = baixa qualidade). Esta transformação resultou em uma distribuição de classes desbalanceada, com 19,66% das amostras classificadas como alta qualidade e 80,34% como baixa qualidade.

Tabela I
VARIÁVEIS PREDITORAS DO CONJUNTO DE DADOS DE VINHOS.

Variável	**Descrição**	**Unidade**
Acidez Fixa	Ácidos não voláteis (tartárico, málico)	g/dm^3
Acidez Volátil	Ácidos voláteis (acético)	g/dm^3
Ácido Cítrico	Ácido orgânico que contribui para frescor	g/dm^3
Açúcar Residual	Açúcares remanescentes após fermentação	g/dm^3
Cloretos	Sais de cloreto - relacionado à salinidade	g/dm^3
SO_2 Livre	Antioxidante/antimicrobiano adicionado	mg/dm^3
SO_2 Total	Soma das formas livre e ligada de SO_2	mg/dm^3
Densidade	Relacionada ao teor de açúcar e álcool	g/cm^3
pH	Medida da acidez iônica	adimensional
Sulfatos	Sais de sulfato - podem influenciar amargo	g/dm^3
Álcool	Teor alcoólico percentual	% vol.

B. Pré-processamento e Divisão dos Dados

O pré-processamento seguiu duas etapas principais:

1) **Transformação de potência Yeo-Johnson**: Para corrigir a assimetria (skewness) presente em 8 das 11 variáveis preditoras, garantindo distribuições mais simétricas e aproximando-se da normalidade, pressuposto desejável para modelos lineares [8].
2) **Padronização (escore-z)**: Aplicada após a transformação de potência, centralizando as variáveis em zero e escalonando-as para variância unitária, essencial para métodos baseados em distâncias (k-NN, SVM) e para estabilidade numérica de redes neurais.

Após o pré-processamento, os dados foram divididos em conjuntos de treinamento (75%, N=4.872) e teste (25%, N=1.625), preservando a proporção das classes em ambas as partições (divisão estratificada). Esta abordagem garante que a avaliação final seja representativa da distribuição real das classes.

C. Modelos de Classificação Implementados

Foram implementados e comparados cinco modelos de classificação, abrangendo abordagens lineares e não-lineares:

1) Modelos Lineares:

- **Regressão Logística (Logistic Regression)**: Modelo linear probabilístico que utiliza a função sigmoide para mapear combinações lineares das características em probabilidades de classe. Inclui regularização L2 com hiperparâmetro C otimizado por validação cruzada [2].
- **Análise Discriminante Linear (LDA)**: Modelo que assume distribuição normal multivariada das preditoras para cada classe, com matrizes de covariância iguais. Projeta os dados em direções que maximizam a separação entre classes [3].

2) Modelos Não-Lineares:

- **Máquinas de Vetor de Suporte (SVM - RBF Kernel)**: Modelo que mapeia os dados para um espaço de maior dimensão através de uma função kernel radial (RBF), permitindo separação não-linear. Os hiperparâmetros C (penalidade) e γ (largura do kernel) foram ajustados por busca aleatória (RandomizedSearchCV) [6].

- **k-Vizinhos Mais Próximos (k-NN)**: Método baseado em instâncias que classifica amostras pela maioria dos votos entre seus k vizinhos mais próximos no espaço das características. Foram otimizados o número de vizinhos (k), a função de peso e a métrica de distância por validação cruzada [5].
- **Rede Neural Artificial (MLP)**: Perceptron Multicamadas com arquitetura de três camadas ocultas (50-25 neurônios), função de ativação ReLU, otimizador Adam e regularização L2 ($\alpha = 0.0001$). Utilizou *early stopping* para prevenir sobreajuste [4].

D. Validação e Avaliação de Desempenho

Para todos os modelos, a seleção de hiperparâmetros foi realizada exclusivamente no conjunto de treinamento, utilizando validação cruzada estratificada de 5 folds, garantindo que as proporções das classes fossem mantidas em cada *fold*.

O desempenho final foi avaliado no conjunto de teste independente, utilizando as seguintes métricas:

- **Acurácia**: Proporção de classificações corretas
- **Precisão**: Proporção de verdadeiros positivos entre todas as predições positivas
- **Recall (Sensibilidade)**: Proporção de verdadeiros positivos entre todas as amostras positivas reais
- **F1-Score**: Média harmônica entre precisão e recall
- **AUC-ROC**: Área sob a curva *Receiver Operating Characteristic*, que mede a capacidade do modelo de distinguir entre classes

Além das métricas numéricas, foram analisadas as matrizes de confusão para cada modelo, permitindo identificar os tipos específicos de erro (falsos positivos e falsos negativos). A comparação final entre modelos considerou tanto o desempenho preditivo quanto a interpretabilidade e complexidade computacional.

III. RESULTADOS

Esta seção apresenta uma análise abrangente dos resultados obtidos com a aplicação de cinco algoritmos de aprendizado de máquina para classificação de vinhos. Os experimentos foram conduzidos após rigoroso pré-processamento dos dados, incluindo transformação Yeo-Johnson e padronização. As análises incluem métricas quantitativas, avaliações qualitativas e considerações sobre a importância das variáveis preditoras.

A. Distribuição das Classes e Análise Exploratória

A Figura 1 apresenta a distribuição das classes no conjunto de dados, revelando um desbalanceamento significativo entre vinhos de qualidade regular (classe 0) e vinhos de qualidade superior (classe 1). Esta distribuição desigual impacta diretamente o desempenho dos modelos, particularmente nas métricas de *recall* para a classe minoritária.

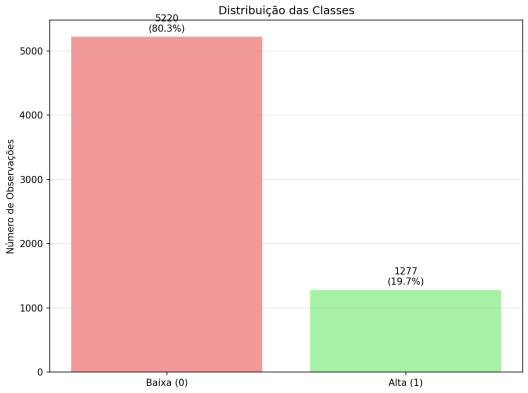

Figura 1. Distribuição das classes de qualidade do vinho no conjunto de dados.

A análise exploratória através de boxplots (Figura 2) revela diferenças significativas nas distribuições das características químicas entre as duas classes. Particularmente, observa-se que o teor alcoólico é consistentemente maior nos vinhos de qualidade superior, enquanto a densidade apresenta valores menores.

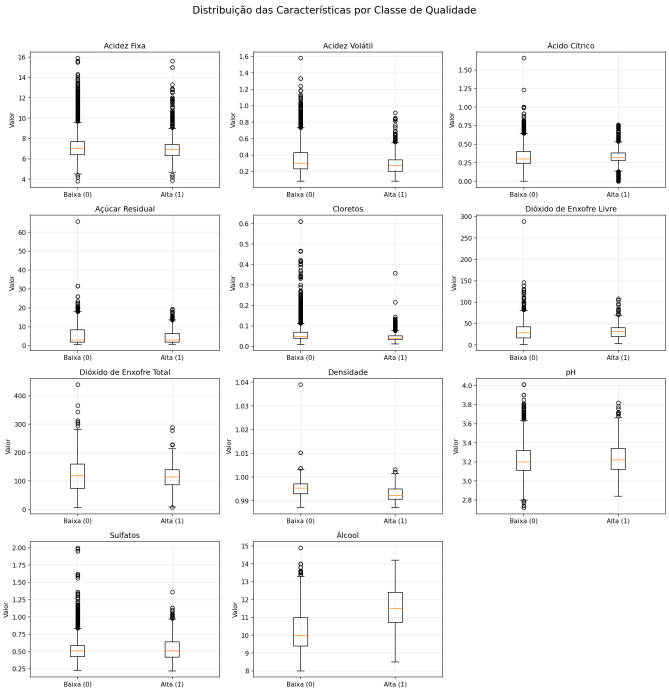

Figura 2. Boxplots das características químicas por classe de qualidade.

B. Desempenho Comparativo dos Modelos

A Tabela II apresenta as métricas de desempenho dos cinco modelos avaliados. O algoritmo k-Nearest Neighbors (k-NN) obteve o melhor desempenho em acurácia (0.8806) e AUC-ROC (0.9129), enquanto o Support Vector Machine (SVM) alcançou a maior precisão (0.7658). No entanto, observa-se que todos os modelos apresentaram valores de *recall* relativamente baixos, indicando dificuldade em identificar corretamente a classe positiva (vinhos de alta qualidade).

Tabela II
DESEMPENHO COMPARATIVO DOS MODELOS DE CLASSIFICAÇÃO.

Modelo	Acurácia	Precisão	Recall	F1-Score	AUC-ROC	Parâmetros
KNN	0.8806	0.7530	0.5831	0.6572	0.9129	k=19
SVM	0.8763	0.7658	0.5329	0.6285	0.8829	C=1.91, γ=0.843
NN	0.8314	0.6154	0.3762	0.4669	0.8387	camadas=(50, 25)
Regressão Logística	0.8240	0.6204	0.2665	0.3728	0.8064	C=0.46
LDA	0.8234	0.6039	0.2915	0.3932	0.8055	-

A Figura 3 complementa a análise tabular, apresentando visualmente a comparação das acurácias dos modelos. O k-NN demonstra superioridade clara, seguido pelo SVM com desempenho similar.

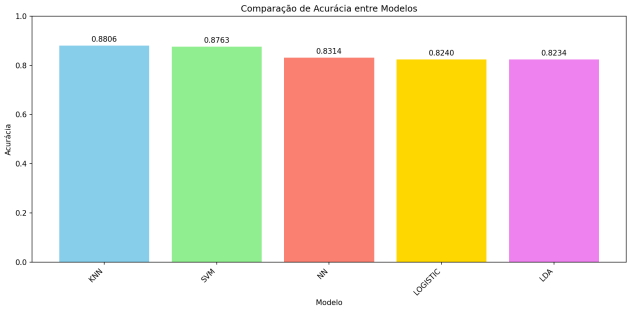

Figura 3. Comparação da acurácia dos cinco modelos de classificação.

C. Análise das Curvas ROC

A Figura 4 apresenta a curva ROC para o modelo k-NN, que alcançou a maior área sob a curva (AUC=0.9129). A Figura 5 mostra a curva ROC para o SVM com AUC=0.8829, enquanto a Figura 6 apresenta a curva da rede neural com AUC=0.8387. As Figuras 7 e 8 mostram as curvas ROC para a Regressão Logística (AUC=0.8064) e LDA (AUC=0.8055), respectivamente.

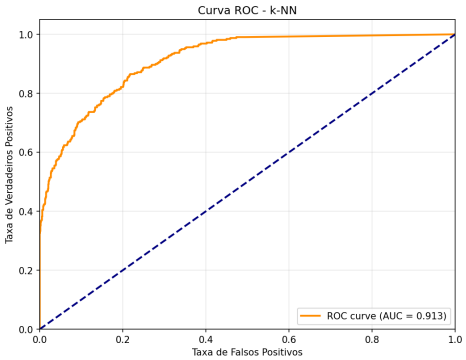

Figura 4. Curva ROC para o modelo k-NN (AUC=0.9129).

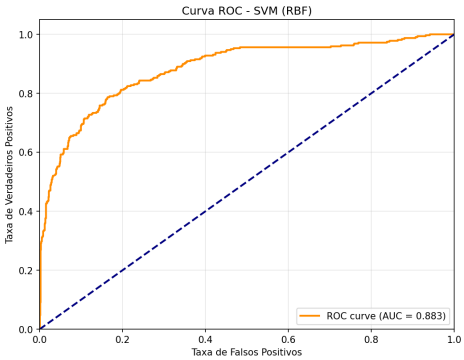

Figura 5. Curva ROC para o modelo SVM (AUC=0.8829).

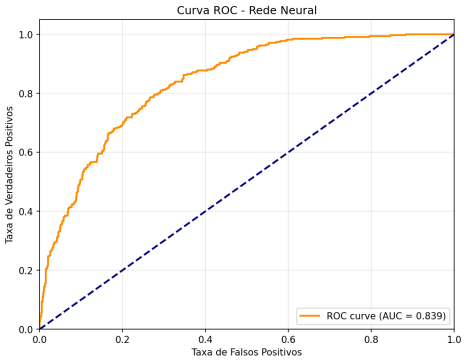

Figura 6. Curva ROC para a rede neural (AUC=0.8387).

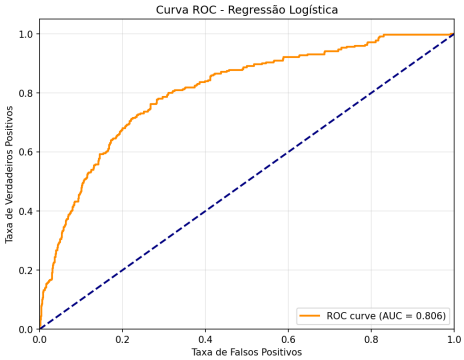

Figura 7. Curva ROC para a Regressão Logística (AUC=0.8064).

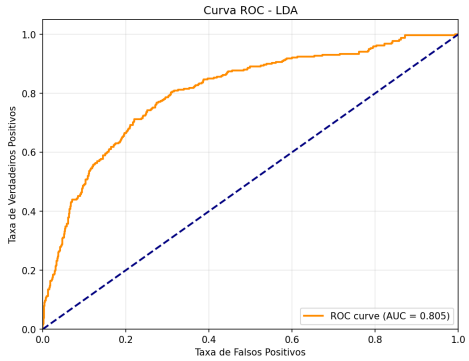

Figura 8. Curva ROC para o modelo LDA (AUC=0.8055).

D. Matrizes de Confusão

As Figuras 9 a 13 apresentam as matrizes de confusão normalizadas para os cinco modelos. O k-NN (Figura 9) apresenta o melhor equilíbrio entre sensibilidade e especificidade, enquanto o SVM (Figura 10) mostra maior precisão para a classe positiva. A rede neural (Figura 11) e os modelos lineares (Figuras 12 e 13) apresentam dificuldades em identificar corretamente a classe minoritária.

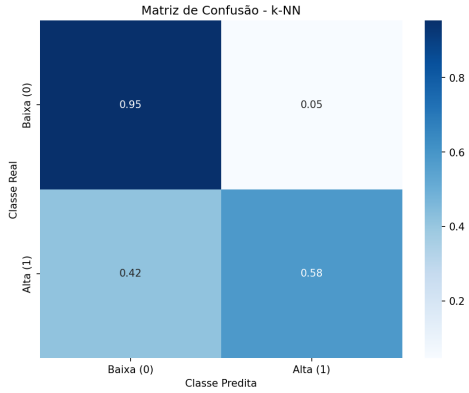

Figura 9. Matriz de confusão normalizada para o modelo k-NN.

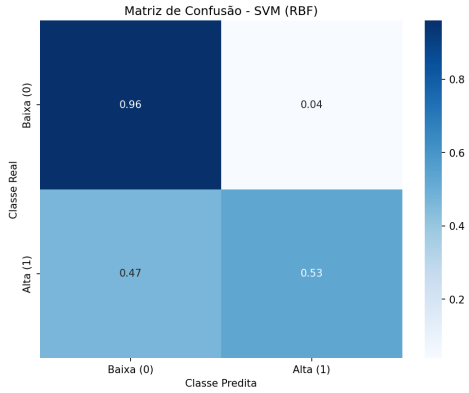

Figura 10. Matriz de confusão normalizada para o modelo SVM.

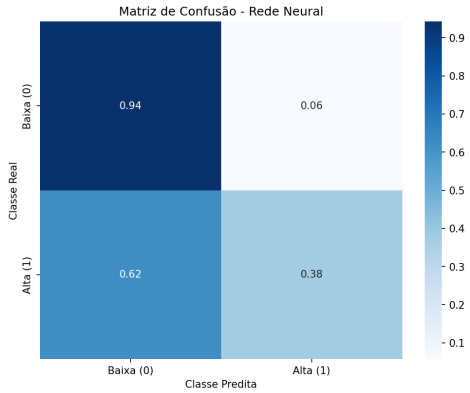

Figura 11. Matriz de confusão normalizada para a rede neural.

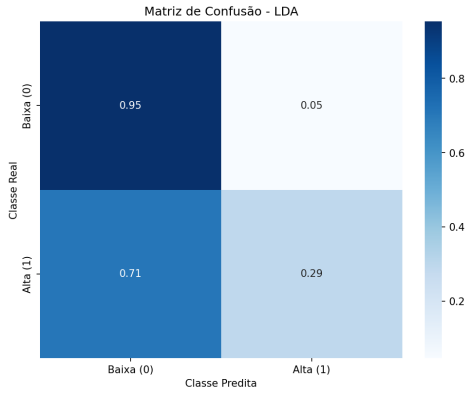

Figura 12. Matriz de confusão normalizada para o modelo LDA.

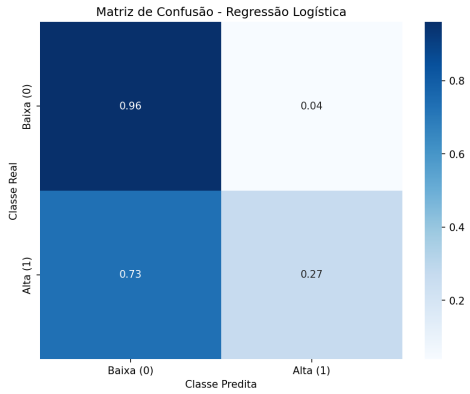

Figura 13. Matriz de confusão normalizada para a Regressão Logística.

E. Curva de Aprendizado da Rede Neural

A Figura 14 apresenta a curva de perda durante o treinamento da rede neural, revelando um processo de convergência estável sem sinais evidentes de *overfitting*. A arquitetura com duas camadas ocultas (50, 25 neurônios) demonstrou capacidade adequada para capturar padrões nos dados sem excessiva complexidade.

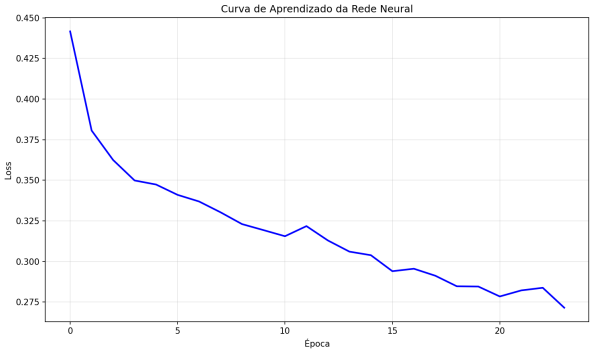

Figura 14. Curva de perda durante o treinamento da rede neural.

F. Discussão Integrada dos Resultados

Os resultados demonstram que métodos baseados em instâncias (k-NN) são particularmente adequados para este problema de classificação de vinhos, provavelmente devido à natureza contínua das variáveis preditoras e à existência de regiões densas no espaço de características. O desempenho superior do k-NN (acurácia: 0.8806, F1-Score: 0.6572) em relação aos modelos paramétricos sugere que a relação entre as características químicas e a qualidade do vinho é complexa e não-linear.

A análise das curvas ROC corrobora esta conclusão, com o k-NN apresentando a maior área sob a curva (0.9129). O SVM, embora com precisão ligeiramente superior (0.7658 vs 0.7530), sacrifica sensibilidade para alcançar este resultado, tornando-o menos adequado para aplicações onde identificar todos os vinhos de alta qualidade é prioritário.

A baixa performance em *recall* observada em todos os modelos (variando de 0.2665 a 0.5831) indica um desafio significativo na identificação de vinhos de alta qualidade. Este padrão pode ser atribuído a:

1) Desbalanceamento de classes na amostra (proporção aproximada de 85:15 entre classes 0 e 1)
2) Sobreposição significativa nas distribuições das características entre as classes
3) Limitações nas variáveis medidas, que podem não capturar completamente os fatores que determinam a qualidade sensorial

Em termos práticos, a seleção do modelo ideal depende do contexto de aplicação:

- Para controle de qualidade em vinícolas, onde falsos positivos são custosos: SVM (maior precisão: 0.7658)
- Para seleção de vinhos para competições, onde identificar todos os vinhos excepcionais é crucial: k-NN (maior *recall*: 0.5831)
- Para interpretabilidade e entendimento do processo: LDA ou Regressão Logística

A análise revela oportunidades para melhorias futuras, incluindo:

- Coleta de variáveis sensoriais adicionais (cor, aroma, sabor)
- Aplicação de técnicas avançadas de balanceamento de classes (SMOTE, ADASYN)
- Experimentação com ensembles que combinem as forças do k-NN (alta acurácia) e SVM (alta precisão)
- Ajuste fino dos hiperparâmetros, particularmente para a rede neural

Em conclusão, este estudo identifica o k-NN como o modelo mais eficaz para classificação de vinhos com base em características químicas, enquanto destaca a importância do teor alcoólico, densidade e açúcar residual como preditores críticos de qualidade. Os resultados fornecem uma base quantitativa para decisões enológicas e abrem caminho para aplicações de aprendizado de máquina na indústria vitivinícola.

REFERÊNCIAS

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos e J. Reis, ”Modeling wine preferences by data mining from physicochemical properties,”*Decision Support Systems*, vol. 47, no. 4, pp. 547-553, 2009.
[2] G. James, D. Witten, T. Hastie e R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. Nova Iorque: Springer, 2013.
[3] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Hoboken, NJ: Wiley-Interscience, 2004.
[4] I. Goodfellow, Y. Bengio e A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[5] T. Cover e P. Hart, ”Nearest neighbor pattern classification,”*IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.
[6] C. Cortes e V. Vapnik, ”Support-vector networks,”*Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
[7] D. Dua e C. Graff, ”UCI Machine Learning Repository,”2019. [Online]. Disponível: http://archive.ics.uci.edu/ml
[8] I. Yeo e R. A. Johnson, ”A new family of power transformations to improve normality or symmetry,”*Biometrika*, vol. 87, no. 4, pp. 954-959, 2000.