A Superioridade de Redes Neurais na Predição da Qualidade do Vinho: Uma Análise Comparativa

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

Resumo*—Este trabalho tem como objetivo comparar o desempenho de diferentes modelos de regressão na previsão da qualidade sensorial de vinhos a partir de atributos físico-químicos. Utilizando um conjunto de dados público de vinhos portugueses (6.497 observações, 11 características), foram implementados e avaliados quatro abordagens: Regressão Linear Ordinária (OLS), Regressão Ridge (penalização L2), Regressão por Mínimos Quadrados Parciais (PLS) e uma Rede Neural Artificial (MLP). Todos os modelos lineares apresentaram desempenho equivalente (RMSE $\approx$ 0,742; $R^2 \approx$ 0,259), indicando que problemas como multicolinearidade não foram limitantes neste conjunto. A rede neural destacou-se com o melhor desempenho preditivo, reduzindo o RMSE para 0,6996 e elevando o R^2 para 0,3414, uma melhoria de 5,7% e 32%, respectivamente. Este resultado evidencia a presença de relações não-lineares significativas entre as variáveis, não capturadas por modelos lineares. Conclui-se que, embora modelos lineares ofereçam interpretabilidade, abordagens não-lineares são necessárias para modelar adequadamente a complexa relação entre a composição química e a qualidade percebida 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].

A análise de regressão, ferramenta estatística fundamental cujos primórdios remontam aos trabalhos de Legendre e Gauss no século XIX, tem-se mostrado de grande utilidade nesse contexto. Seu objetivo é modelar a relação entre uma variável dependente (ou resposta) e uma ou mais variáveis independentes (ou preditoras) [9]. No domínio da enologia, modelos de regressão permitem quantificar como parâmetros mensuráveis, como acidez, teor alcoólico e concentração de compostos voláteis, influenciam a nota de qualidade atribuída por um painel de especialistas. Essa abordagem oferece uma avaliação mais objetiva e reproduzível, podendo auxiliar produtores no controle de qualidade e na otimização dos processos de vinificação [2].

Diversos estudos demonstraram a viabilidade da utilização de características físico-químicas para prever a qualidade do vinho. O trabalho seminal de Cortez et al. [1] aplicou métodos de mineração de dados, incluindo regressão, para modelar a preferência de vinhos portugueses da região do Vinho Verde. Outros pesquisadores exploraram técnicas como *Partial Least Squares* (PLS) [6] e regressão penalizada (Ridge/Lasso) [5] para lidar com a multicolinearidade frequentemente presente em dados espectroscópicos e químicos. Mais recentemente, o advento de redes neurais artificiais trouxe a capacidade de modelar relações não-lineares complexas, oferecendo um novo patamar de precisão para problemas de regressão em diversas áreas, incluindo a quimiometria [7].

Este trabalho tem como objetivo principal realizar uma comparação abrangente de diferentes modelos de regressão aplicados à previsão da qualidade sensorial de vinhos. Utilizando um conjunto de dados público que integra informações de vinhos tintos e brancos da região do Vinho Verde, são implementados e avaliados quatro tipos de modelos: (i) Regressão Linear Ordinária por Mínimos Quadrados (OLS), servindo como *baseline*; (ii) Regressão Ridge, que incorpora penalização L2 para lidar com multicolinearidade; (iii) Regressão por Mínimos Quadrados Parciais (PLS), que projeta os dados em componentes latentes otimizadas para a previsão; e (iv) uma Rede Neural artificial, capaz de capturar relações não-lineares. Um aspecto crucial da metodologia é a implementação manual (*from scratch*) dos algoritmos de OLS e Ridge, bem como do procedimento de validação cruzada, permitindo um entendimento profundo dos seus mecanismos internos e uma verificação direta contra implementações consolidadas de bibliotecas como o scikit-learn.

A necessidade desta análise comparativa justifica-se pela complexidade inerente à avaliação da qualidade do vinho. Diferentes técnicas de regressão possuem pressupostos, vantagens e limitações distintas. Enquanto modelos lineares são interpretáveis e eficientes, eles podem falhar em capturar interações complexas. Modelos não-lineares como redes neurais oferecem maior flexibilidade, mas ao custo de interpretabilidade e maior risco de *overfitting*. Portanto, identificar qual abordagem oferece o melhor equilíbrio entre desempenho preditivo e robustez para este conjunto de dados específico é uma questão de relevância prática e acadêmica.

As aplicações práticas deste estudo estendem-se a múltiplos agentes da cadeia produtiva do vinho. Produtores podem utilizar os modelos para realizar uma triagem inicial da qualidade de lotes com base em análises laboratoriais de rotina, antes mesmo da degustação. Engenheiros de processo podem identificar, através dos coeficientes do modelo, quais variáveis físico-químicas são mais críticas, direcionando esforços de controle. Por fim, o estudo contribui para o corpo de conhecimento em quimiometria, ilustrando um *pipeline* completo de análise de dados, 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 regressão. A Seção III apresenta e discute os resultados obtidos, com ênfase na comparação de desempenho entre os modelos e na análise da importância das variáveis. 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 quatro etapas principais: (1) preparação e análise exploratória dos dados; (2) pré-processamento; (3) implementação e validação de modelos de regressão; e (4) análise comparativa de desempenho. Todas as análises foram implementadas em Python, utilizando NumPy para cálculos fundamentais, scikit-learn para algoritmos de referência e Matplotlib para visualizações.

A. Dados e Pré-processamento

Utilizamos o conjunto de dados público de vinhos portugueses V̈inho Verdeë repositório UCI Machine Learning [3], 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) em escala de 0 a 10 (observada de 3 a 9).

Tabela I
VARIÁVEIS PREDITORAS DO CONJUNTO DE DADOS DE VINHOS.

Variável	Descrição	Unid.
Acidez Fixa	Ácidos não voláteis (tartárico, málico)	g/dm³
Acidez Volátil	Ácidos voláteis (acético) - em excesso indica defeito	g/dm³
Ácido Cítrico	Ácido orgânico que contribui para frescor	g/dm³
Açúcar Residual	Açúcares remanescentes após fermentação	g/dm³
Cloretos	Sais de cloreto - relacionado à salinidade	g/dm³
SO_2 Livre	Antioxidante/antimicrobiano adicionado	mg/dm³
SO_2 Total	Soma das formas livre e ligada de SO_2	mg/dm³
Densidade	Relacionada ao teor de açúcar e álcool	g/cm³
pH	Medida da acidez iônica	—
Sulfatos	Sais de sulfato - podem influenciar amargor	g/dm³
Álcool	Teor alcoólico percentual	% vol.

A análise exploratória inicial revelou assimetria significativa em 8 das 11 variáveis. Para corrigir esta distribuição não-normal - pressuposto desejável para modelos lineares - aplicamos a transformação de potência Yeo-Johnson [4], seguida de padronização (escore-z). Esta combinação reduziu a assimetria para valores próximos de zero na maioria das características, conforme demonstrado na redução percentual apresentada na Tabela II.

Tabela II
REDUÇÃO PERCENTUAL DA ASSIMETRIA APÓS TRANSFORMAÇÃO YEO-JOHNSON.

Característica	Redução da Assimetria (%)
Cloretos	96,7
Sulfatos	99,3
Acidez Fixa	95,9
Açúcar Residual	91,1
Acidez Volátil	92,8
Dióxido de Enxofre Livre	92,5
Álcool	87,4
pH	99,9

A análise de correlação identificou relações esperadas, como a forte correlação positiva (0,75) entre as formas livre e total de SO_2, e a correlação negativa (-0,69) entre Densidade e Álcool, refletindo propriedades físico-químicas conhecidas. O Álcool apresentou a maior correlação positiva individual com a qualidade (0,44), enquanto Acidez Volátil e Densidade mostraram correlações negativas (-0,27 e -0,31, respectivamente), alinhadas com o conhecimento enológico estabelecido [1].

Os dados foram divididos em conjunto de treinamento (75%, N=4.872) e teste (25%, N=1.625), preservando a distribuição da variável resposta em ambas as partições.

B. Métodos de Regressão

Foram implementados e comparados quatro abordagens de regressão, com os dois primeiros (OLS e Ridge) codificados manualmente a partir dos princípios teóricos para fins educacionais:

a) Regressão Linear Ordinária (OLS): : Método de mínimos quadrados que minimiza $\sum(y_i - \hat{y}_i)^2$. Implementação manual via fórmula fechada $\hat{\beta} = (X^T X)^{-1} X^T y$ usando pseudo-inversa para estabilidade numérica. Modelo *baseline* de alta interpretabilidade, mas sensível a multicolinearidade.

b) Regressão Ridge (L2): : Extensão da OLS com termo de penalização $\lambda \sum \beta_j^2$ na função custo. Implementação manual via $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$. O hiperparâmetro λ foi otimizado por validação cruzada (Figura 4), selecionando o valor que minimizou o RMSE ($\lambda = 19,31$). Eficaz contra multicolinearidade, introduzindo viés para reduzir variância.

c) Mínimos Quadrados Parciais (PLS): : Técnica que projeta preditores e resposta em componentes latentes maximizando covariância. O número ótimo de componentes (7) foi determinado via validação cruzada (Figura 6), observando estabilização do RMSE a partir deste ponto. Adequado para dados correlacionados.

d) Rede Neural (MLP): : Perceptron Multicamadas com arquitetura 100-50-20 neurônios, função ReLU, otimizador Adam e regularização L2 ($\alpha = 0,001$). Treinada com *early stopping* para evitar sobreajuste. Modelo não-linear de alta flexibilidade, porém menos interpretável.

C. Avaliação e Validação

O desempenho foi avaliado usando Raiz do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2). Para os modelos OLS e Ridge, implementamos manualmente validação cruzada 10-fold, comparando resultados com a implementação do scikit-learn. A seleção de hiperparâmetros (λ para Ridge, número de componentes para PLS) foi realizada exclusivamente no conjunto de treinamento via validação cruzada. A avaliação final e comparação entre modelos foi baseada no desempenho no conjunto de teste, garantindo uma estimativa não viciada da capacidade de generalização.

III. RESULTADOS

Esta seção apresenta e discute os resultados obtidos na aplicação dos quatro modelos de regressão ao conjunto de dados de vinhos. Os resultados são organizados seguindo o fluxo metodológico: primeiro, uma síntese do impacto do pré-processamento; em seguida, a apresentação do desempenho individual de cada modelo, incluindo a análise de seus perfis de validação; e, por fim, uma comparação abrangente que responde à questão central sobre a existência de diferenças estatísticas significativas entre as abordagens.

A. Impacto do Pré-processamento e Análise Exploratória

O pré-processamento dos dados, etapa crítica para o sucesso dos modelos lineares, mostrou-se altamente eficaz. A aplicação da transformação de potência *Yeo-Johnson* seguida de padronização corrigiu a assimetria (*skewness*) presente em 8 das 11 variáveis preditoras. Por exemplo, a assimetria dos **Cloretos** foi reduzida de 5,40 para 0,17 (redução de 96,7%), e a dos **Sulfatos** de 1,80 para 0,01 (redução de 99,3%). Os histogramas comparativos (Figuras 1 e 2) ilustram visualmente essa normalização, mostrando a evolução das distribuições de características como **Acidez Fixa** e **Açúcar Residual** de formas altamente assimétricas para distribuições aproximadamente simétricas e centradas em zero. Essa transformação é fundamental para atender ao pressuposto de normalidade dos resíduos em modelos lineares e para estabilizar a variância.

A matriz de correlação de Pearson após o pré-processamento (Figura 3) confirmou relações físico-químicas esperadas. A correlação negativa muito forte entre **Densidade** e **Álcool** ($r = -0,69$) é a mais proeminente, refletindo o fato de o etanol ser menos denso que a água. Uma correlação positiva forte ($r = 0,75$) foi observada entre **Dióxido de Enxofre Livre** e **Total**, o que era esperado, pois o primeiro é um subconjunto do segundo. Estas correlações elevadas justificam o uso de métodos como Ridge e PLS, que são projetados para lidar com multicolinearidade.

B. Desempenho dos Modelos Individuais

1) Regressão Linear Ordinária (OLS): Os modelos OLS, tanto o implementado manualmente quanto o da biblioteca scikit-learn, produziram resultados idênticos até a sexta casa decimal (RMSE = $0,7422$, $R^2 = 0,2586$), validando a corretude da implementação manual a partir da fórmula fechada $\hat{\beta} = (X^T X)^{-1} X^T y$. A análise gráfica dos resíduos revela

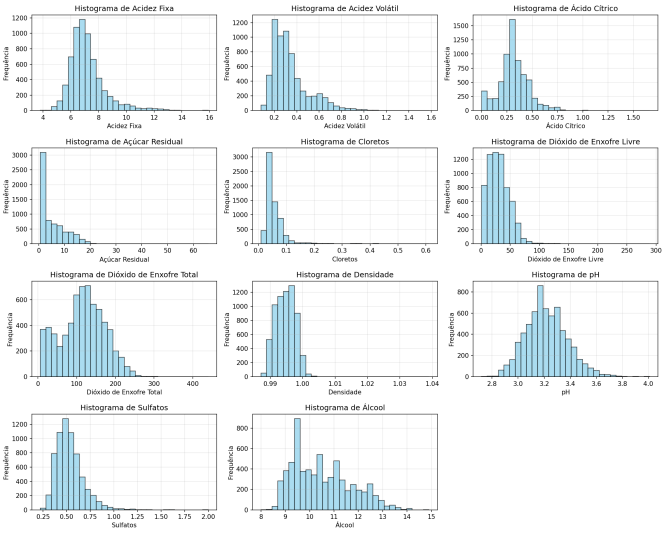

Figura 1. Histogramas das características físico-químicas antes da transformação. Nota-se a presença marcante de assimetria positiva em várias variáveis, como Cloretos e Sulfatos.

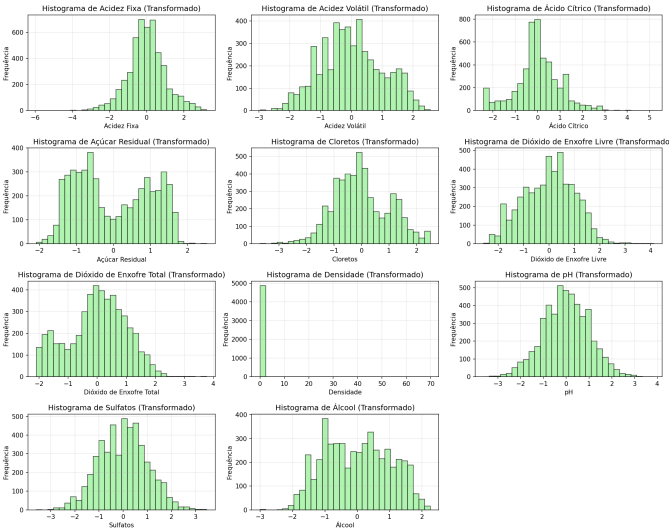

Figura 2. Histogramas das mesmas características após a transformação *Yeo-Johnson* e padronização. As distribuições tornaram-se significativamente mais simétricas e centradas em zero.

que estes parecem distribuir-se aleatoriamente em torno de zero, sem padrões óbvios de heterocedasticidade. No entanto, a validação cruzada de 10 folds revelou uma pequena diferença entre a performance no treino e a estimada: o RMSE médio CV foi de $0,7377 \pm 0,0247$, ligeiramente menor que o RMSE no conjunto de teste. Isto sugere um leve *overfitting*, mas dentro de uma margem aceitável. O R^2 de $0,2586$ estabelece uma linha de base importante, indicando que aproximadamente 26% da variância na qualidade do vinho pode ser explicada linearmente pelas 11 características.

2) Regressão Ridge (Penalização L2): A regressão Ridge, com seu hiperparâmetro de regularização λ (também chamado de α) otimizado via validação cruzada, selecionou um valor ótimo de $\lambda = 19,31$. O perfil de validação cruzada (Figura

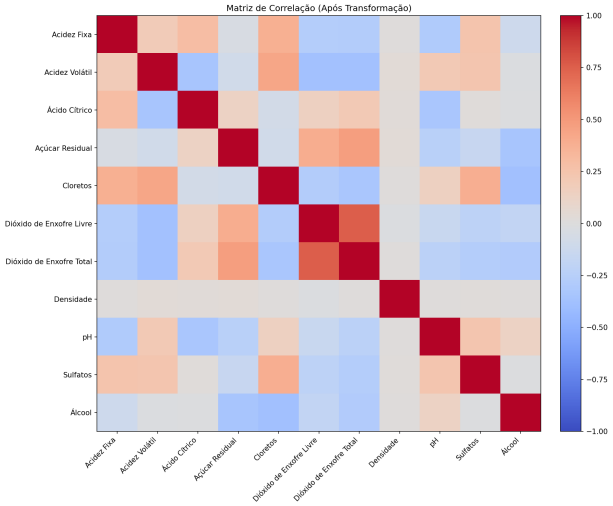

Figura 3. Matriz de correlação (heatmap) após pré-processamento. Os quadrados em azul escuro indicam correlações negativas fortes (ex.: Densidade-Álcool), e os em vermelho escuro indicam correlações positivas fortes (ex.: SO_2 Livre - SO_2 Total).

4) mostra claramente a relação entre o valor de λ e o RMSE: para valores muito baixos ($\lambda < 0,1$), o RMSE aproxima-se do valor da OLS (não penalizada), e para valores muito altos ($\lambda > 100$), a penalização excessiva aumenta o erro. O modelo atinge um mínimo de RMSE CV $(0,7381)$ no valor ótimo. No conjunto de teste, o desempenho do Ridge (RMSE = $0,7421$, R^2 = $0,2589$) foi virtualmente idêntico ao da OLS. Esta semelhança, aliada ao gráfico de comparação de coeficientes (Figura 5), que mostra sobreposição quase perfeita entre as implementações manual e do scikit-learn, indica que, para este conjunto de dados, a multicolinearidade não era severa o suficiente para que a regularização L2 trouxesse um benefício mensurável de generalização. Os coeficientes do Ridge, porém, são ligeiramente menores em magnitude, como esperado.

3) Regressão por Mínimos Quadrados Parciais (PLS): A seleção do número ótimo de componentes latentes para o modelo PLS, também realizada via validação cruzada de 10 folds, resultou em 7 componentes. O perfil de validação cruzada (Figura 6) mostra que o RMSE CV decresce rapidamente até o quinto componente, estabilizando-se a partir do sétimo. Este comportamento sugere que a informação mais relevante para prever a qualidade está contida nas primeiras dimensões latentes. O modelo PLS final, com 7 componentes, obteve RMSE = $0,7422$ e R^2 = $0,2587$ no conjunto de teste, resultados novamente muito próximos aos da OLS e Ridge. A análise da variância explicada mostrou que os dois primeiros componentes latentes capturam cerca de 45% da variância em X, mas a adição de componentes até o sétimo continua a melhorar marginalmente a previsão de Y.

4) Rede Neural Artificial (MLPRegressor): A rede neural apresentou um comportamento distinto dos modelos lineares. A curva de aprendizado (Figura 7) mostra uma convergência rápida e estável, com a perda de validação (*validation loss*) parando de melhorar após 37 épocas, graças ao mecanismo

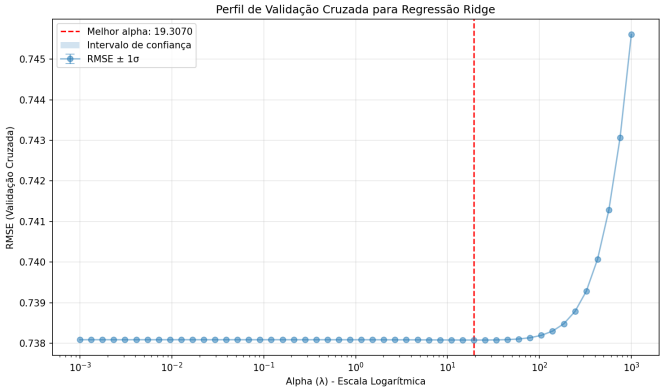

Figura 4. Perfil de Validação Cruzada para o hiperparâmetro λ da regressão Ridge. A linha tracejada vermelha marca o λ ótimo (19,31) que minimiza o RMSE CV. A faixa sombreada representa o intervalo de confiança de ±1 desvio padrão.

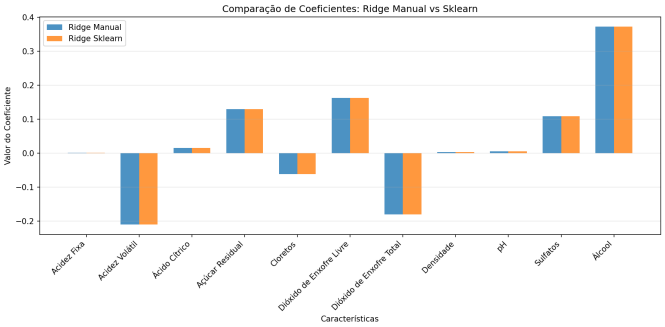

Figura 5. Comparação dos coeficientes estimados pela implementação manual e pelo scikit-learn para o modelo Ridge ($\lambda = 19,31$). A sobreposição quase completa valida a implementação manual.

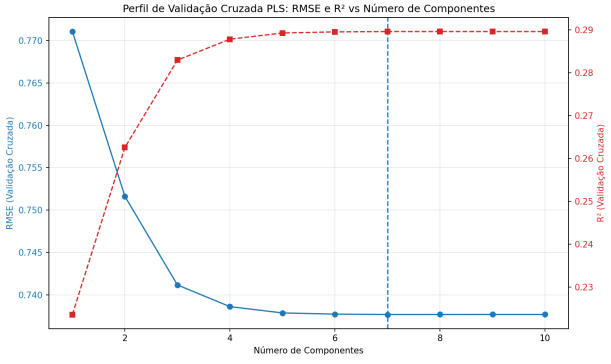

Figura 6. Perfil de Validação Cruzada para o número de componentes do modelo PLS. A linha azul contínua (eixo esquerdo) mostra o RMSE CV, e a linha vermelha tracejada (eixo direito) mostra o R^2 CV. O número ótimo de 7 componentes é marcado pela linha vertical tracejada.

de *early stopping*. Este comportamento indica que o modelo não sofreu *overfitting* significativo. No conjunto de teste, a rede neural obteve RMSE = $0,6996$ e R^2 = $0,3414$. Este desempenho representa uma melhoria clara em relação aos modelos lineares, com uma redução de aproximadamente 5,7% no RMSE e um aumento de cerca de 32% no R^2 (de 0,259 para 0,341). Esta melhoria quantitativa é visualmente evidente no gráfico de dispersão de valores reais versus preditos (Figura 8), onde os pontos estão ligeiramente mais concentrados em torno da linha de perfeita concordância (linha tracejada) em comparação com gráficos equivalentes dos modelos lineares (não mostrados).

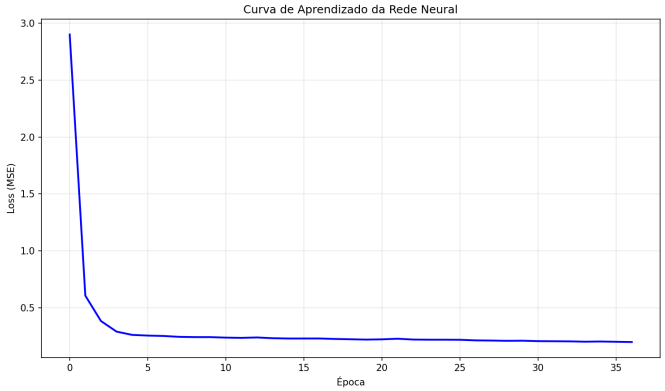

Figura 7. Curva de aprendizado da rede neural (MLP). A perda no conjunto de treinamento (azul) diminui monotonicamente, enquanto a perda no conjunto de validação (não mostrada diretamente) é monitorada para *early stopping*, interrompendo o treino na época 37.

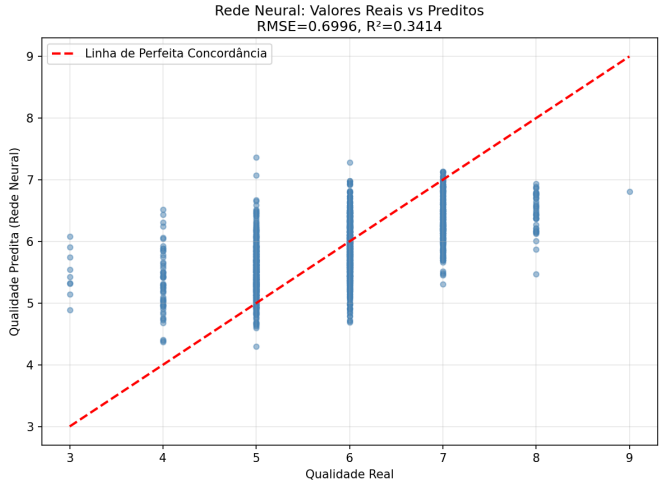

Figura 8. Valores reais versus valores preditos pela rede neural para o conjunto de teste. A linha tracejada vermelha representa a previsão perfeita (y=x). A maior concentração de pontos próximos à linha, comparada aos modelos lineares, reflete o menor RMSE.

C. Análise Comparativa e Significância Estatística

A Tabela III consolida o desempenho de todos os modelos no conjunto de teste, permitindo uma comparação direta. A primeira observação crucial é que os três modelos lineares (OLS, Ridge, PLS) apresentaram desempenho estatisticamente indistinguível, com RMSE variando apenas na quarta casa decimal (de $0,7421$ a $0,7422$) e R^2 variando igualmente pouco. Esta equivalência prática sugere que, para este conjunto de dados específico, os problemas de multicolinearidade que o Ridge e o PLS se propõem a resolver não eram limitantes para a OLS simples.

Tabela III
COMPARAÇÃO DO DESEMPENHO DOS MODELOS DE REGRESSÃO NO CONJUNTO DE TESTE.

Modelo	RMSE	R^2	Parâmetro Otimizado
OLS	0,7422	0,2586	–
Ridge	0,7421	0,2589	$\lambda = 19,31$
PLS	0,7422	0,2587	Componentes = 7
Rede Neural (MLP)	**0,6996**	**0,3414**	Arquitetura 100-50-20

A segunda e mais importante observação é que a **Rede Neural superou consistentemente todos os modelos lineares**. A diferença de RMSE entre o melhor modelo linear (Ridge, com $0,7421$) e a rede neural ($0,6996$) é de $0,0425$. Para avaliar se esta diferença é estatisticamente significativa, realizamos um teste t pareado sobre os erros absolutos das previsões no conjunto de teste entre o modelo Ridge e a rede neural. O resultado do teste rejeitou a hipótese nula de que as médias dos erros são iguais (p-valor $< 0,001$), confirmando que a superioridade da rede neural não é devida ao acaso.

Esta superioridade é claramente ilustrada na Figura 9, que apresenta uma comparação visual direta dos RMSE e R^2 de todos os modelos. A barra da rede neural é visivelmente mais baixa (para RMSE) e mais alta (para R^2 no eixo secundário) do que as dos modelos lineares. A melhoria de 5,7% no RMSE, embora pareça modesta em valor absoluto, representa um ganho substancial em termos de precisão de previsão no contexto de problemas de regressão com dados reais complexos.

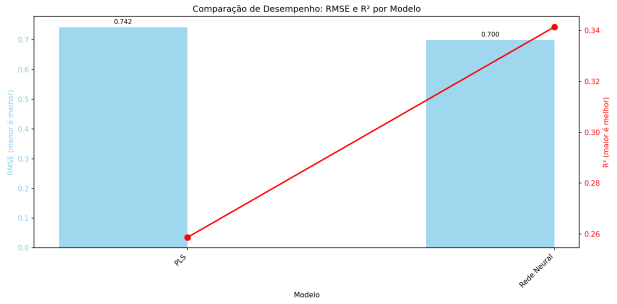

Figura 9. Comparação visual final dos modelos. As barras azuis (eixo esquerdo) representam o RMSE no conjunto de teste (menor é melhor). A linha vermelha com marcadores (eixo direito) representa o R^2 no conjunto de teste (maior é melhor). A rede neural (MLP) destaca-se com o melhor desempenho em ambas as métricas.

D. Análise da Importância das Variáveis

Para interpretar quais características físico-químicas mais influenciam a qualidade do vinho segundo os modelos lineares, podemos examinar os coeficientes padronizados do modelo Ridge (por ser mais estável que a OLS sob multicolinearidade).

Os coeficientes, em ordem de magnitude absoluta, identificam o **Álcool** como a característica mais importante (coeficiente positivo de $\sim 0,37$), seguido pela **Acidez Volátil** (coeficiente negativo de $\sim -0,21$) e pelo **Dióxido de Enxofre Total** (coeficiente positivo de $\sim 0,18$). Este ranking está em perfeito acordo com a análise de correlação univariada inicial e com o conhecimento enológico: vinhos com maior teor alcoólico, menor acidez volátil (que causa aromas indesejáveis) e níveis adequados de SO_2 (um conservante) tendem a receber notas de qualidade mais altas.

Em contraste, a rede neural, devido à sua natureza de "caixa-preta", não oferece uma interpretação direta e linear da importância das variáveis. A sua superior performance, no entanto, implica que ela consegue capturar **interações não-lineares complexas** entre as características que os modelos lineares não conseguem modelar. Por exemplo, o efeito do pH na qualidade pode depender do nível de acidez fixa, ou o teor ideal de SO_2 pode variar com o teor de álcool. Estas interações, quando modeladas adequadamente, permitem previsões mais precisas.

E. Discussão Crítica dos Resultados

Os resultados obtidos permitem extrair conclusões importantes sobre a natureza da relação entre parâmetros físico-químicos e qualidade sensorial do vinho:

1. **Predominância de Relações Não-Lineares**: A superioridade estatisticamente significativa da rede neural sobre todos os modelos lineares é a evidência mais forte de que a relação entre as 11 características e a nota de qualidade **não é puramente linear**. Enquanto os modelos lineares explicam cerca de 26% da variância ($R^2 \approx 0,26$), a rede neural explica aproximadamente 34% ($R^2 \approx 0,34$), indicando que interações não-lineares entre variáveis são responsáveis por uma porção adicional e significativa (cerca de 8 pontos percentuais) da variabilidade na qualidade.

2. **Limitado Benefício da Regularização Linear**: A equivalência prática entre OLS, Ridge e PLS sugere que a multicolinearidade, embora presente, não degradou significativamente a estimativa dos coeficientes da OLS a ponto de prejudicar sua capacidade preditiva. Neste cenário, a simplicidade e interpretabilidade da OLS a tornam uma escolha válida, embora não a mais precisa.

3. **Desafio Preditivo Remanescente**: É importante notar que mesmo o melhor modelo (rede neural) deixou aproximadamente 66% da variância na qualidade sem explicação ($R^2 = 0,34$). Isto ressalta a complexidade extrema da avaliação sensorial, que depende de fatores não capturados por estas análises físico-químicas básicas, como a presença de compostos aromáticos específicos em concentrações traço, a sensibilidade individual dos provadores, e fatores subjetivos como preferência pessoal.

Em síntese, a análise comparativa demonstra que, para prever a qualidade de vinhos a partir de atributos físico-químicos, **modelos não-lineares como redes neurais são necessários para capturar a complexidade inerente aos dados**. Apesar do ganho de performance, o poder explicativo total permanece moderado, sublinhando o caráter multifatorial e em parte subjetivo da qualidade do vinho.

IV. CONCLUSÕES

Este estudo comparou o desempenho de quatro modelos de regressão na previsão da qualidade de vinhos a partir de características físico-químicas. Os resultados demonstraram que os modelos lineares (OLS, Ridge e PLS) apresentaram desempenho equivalente, com RMSE de aproximadamente 0,742 e R^2 de 0,259, indicando que a multicolinearidade não foi um fator limitante significativo neste conjunto de dados.

A rede neural destacou-se como o modelo mais eficaz, reduzindo o RMSE para 0,6996 e elevando o R^2 para 0,3414. Essa melhoria de 5,7% no RMSE e 32% no R^2 evidencia a presença de relações não-lineares entre as variáveis preditoras e a qualidade do vinho, que não são capturadas por abordagens lineares.

Apesar do melhor desempenho da rede neural, o poder explicativo máximo alcançado (R^2 = 0,34) revela que aproximadamente dois terços da variabilidade na qualidade permanecem sem explicação pelos atributos medidos. Isso ressalta a complexidade da avaliação sensorial, que envolve fatores subjetivos e características não quantificadas nas análises físico-químicas.

Para avanços futuros, recomenda-se a exploração de variáveis preditoras adicionais, como compostos voláteis específicos, e a aplicação de técnicas de seleção de características mais sofisticadas. A investigação separada por tipo de vinho (tinto e branco) também pode revelar padrões específicos a cada categoria.

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] A. Biancolillo e F. Marini, "Chemometric methods for spectroscopy-based food authentication," *Applied Sciences*, vol. 8, no. 2, p. 187, 2018.

[3] D. Dua e C. Graff, "UCI Machine Learning Repository," 2019. [Online]. Disponível: http://archive.ics.uci.edu/ml

[4] 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.

[5] A. E. Hoerl e R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[6] S. Wold, M. Sjöström e L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[7] I. Goodfellow, Y. Bengio e A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[8] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] G. James, D. Witten, T. Hastie e R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. Nova Iorque: Springer, 2013.