DISCIPLINARUM SCIENTIA *Disciplinarum Scientia.* **Série: Naturais e Tecnológicas**
Santa Maria, v. 27, n. 2, p. 27-44, 2026
Recebido em: 05.12.2025. Aprovado em: 13.04.2026.
ISSN 2176-462X | DOI: doi.org/10.37779/nt.v27i2.5632

ANÁLISE EXPERIMENTAL DA RELAÇÃO ENTRE PH E COR DO EXTRATO DE REPOLHO ROXO SOB LED RGB UTILIZANDO APRENDIZAGEM DE MÁQUINA

EXPERIMENTAL ANALYSIS OF THE RELATIONSHIP BETWEEN PH AND THE COLOR OF RED CABBAGE EXTRACT UNDER RGB LED ILLUMINATION USING MACHINE LEARNING

Milton Baptista Filho[1], Bruna Lemos Marques Figueiredo[2], Cassiana Barreto Hygino Machado[3], Vantelfo Nunes Garcia[4], Messias Freire Cardoso Baptista[5] e Tiago Desteffani Admiral[6]

RESUMO

Os recursos de aprendizagem de máquina têm se consolidado como ferramentas relevantes na análise de conjuntos de dados multidimensionais, especialmente em estudos qualitativos e quantitativos de substâncias sob diferentes condições experimentais. Neste trabalho, investigamos a relação entre pH e cor do extrato de repolho roxo por meio de uma montagem experimental acessível, composta por uma estrutura plástica que alinha um LED RGB, uma cubeta de vidro e a câmera de um smartphone. Cada cor primária do LED foi utilizada separadamente para produzir condições específicas de transmitância, permitindo registrar as variações colorimétricas do extrato em níveis de pH entre 2,0 e 9,0. A partir de repetições independentes para cada combinação de pH e cor de iluminação, foi obtido um conjunto de dados com mais de cem amostras e cinco características, analisado por métodos de aprendizagem de máquina nos domínios não supervisionado (clusterização) e supervisionado (regressão). Os resultados indicaram que as amostras iluminadas pelo LED verde apresentaram maior separação no espaço HSV, possibilitando o desenvolvimento de um modelo preditivo com desempenho mais consistente para estimar o valor de pH a partir dos parâmetros de cor. Os achados reforçam a utilidade de abordagens experimentais de baixo custo associadas à análise computacional moderna para a investigação quantitativa de indicadores naturais e para a construção de modelos preditivos baseados em aprendizado de máquina.

Palavras-chave: Química; indicadores; pH; aprendizagem de máquina; colorimetria digital.

1 Professor de física no Núcleo de Pesquisa em Física e Ensino de Ciências (NPFEC) do MNPEF/IFFluminense. E-mail: mfilho@iff.edu.br. ORCID: https://orcid.org/0000-0002-4035-436X

2 Licencianda em Ciências da Natureza no IF Fluminense, campus Campos Centro. E-mail: figueiredo.m@gsuite.iff.edu.br. ORCID: https://orcid.org/0009-0006-4820-5525

3 Professora de física no Núcleo de Pesquisa em Física e Ensino de Ciências (NPFEC) do MNPEF/IFFluminense. E-mail: cassiana.h.machado@iff.edu.br. ORCID: https://orcid.org/0000-0002-0126-4169

4 Professor de física no Núcleo de Pesquisa em Física e Ensino de Ciências (NPFEC) do MNPEF/IFFluminense. E-mail: vantelfo.garcia@iff.edu.br. ORCID: https://orcid.org/0000-0001-8569-1012

5 Cosmetica Labs Inc., Toronto, Canada. E-mail: messiasbap@outlook.com. ORCID: https://orcid.org/0000-0002-3101-630X

6 Professor de física no Núcleo de Pesquisa em Física e Ensino de Ciências (NPFEC) do MNPEF/IFFluminense. E-mail: tiago.admiral@iff.edu.br. ORCID: https://orcid.org/0000-0002-3829-5778

ABSTRACT

Machine learning resources have become consolidated as relevant tools for analyzing multidimensional datasets, particularly in qualitative and quantitative studies of substances under different experimental conditions. In this work, we investigate the relationship between pH and the color of red cabbage extract using an accessible experimental setup composed of a plastic structure that aligns an RGB LED, a standard glass cuvette, and a smartphone camera. Each primary color of the LED was used separately to produce specific transmittance conditions, allowing the colorimetric variations of the extract to be recorded across pH levels ranging from 2.0 to 9.0. Based on independent repetitions for each combination of pH and illumination color, a dataset with more than one hundred samples and five features was obtained and analyzed using machine learning methods in both unsupervised (clustering) and supervised (regression) domains. The results indicated that samples illuminated by the green LED exhibited greater separation in HSV space, enabling the development of a more consistent predictive model for estimating pH from the color parameters. These findings reinforce the utility of low-cost experimental approaches combined with modern computational analysis for the quantitative investigation of natural indicators and for the construction of machine-learning-based predictive models.

Keywords: *Chemistry; indicators; pH; machine learning; digital colorimetry.*

1 INTRODUÇÃO

Indicadores ácido-base naturais têm recebido atenção crescente em pesquisas voltadas a soluções de baixo custo e menor impacto ambiental. Entre esses materiais, o extrato de repolho roxo destaca-se pela marcada variação de cor em função do pH, pela ampla disponibilidade e pela aplicabilidade em diferentes cenários experimentais (Amaral, 2023). Entretanto, o registro digital dessas variações depende também das condições de iluminação, uma vez que a luz refletida ou transmitida pela amostra influencia diretamente a resposta colorimétrica obtida.

A integração entre materiais acessíveis, técnicas instrumentais simples e métodos modernos de análise numérica permite explorar esse sistema de maneira mais rigorosa. A relação quantitativa entre pH e cor de extratos naturais (Wulan Hastuti, 2020), especialmente sob fontes de luz controladas, constitui um problema experimental relevante que pode se beneficiar do uso de espaços de cor apropriados e de algoritmos de análise de dados.

Neste trabalho, apresentamos uma montagem experimental destinada a investigar como o extrato de repolho roxo responde a dez níveis distintos de pH quando iluminado separadamente pelas três cores de um LED RGB (*Red, Green, Blue*). As amostras foram registradas pela câmera de um smartphone, e os dados foram estruturados no espaço HSV (*Hue, Saturation e Value*). A partir desse conjunto, aplicamos diferentes métodos de clusterização e classificação para comparar a consistência das imagens obtidas sob cada iluminação (Chioson, 2018). Também avaliamos um método de aprendizagem supervisionada com o objetivo de estimar o valor de pH a partir dos registros obtidos com o LED verde.

A proposta articula um experimento de baixo custo, baseado em materiais de fácil obtenção, como extrato de repolho roxo, LED RGB e soluções preparadas a partir de água sanitária, a procedimentos de análise quantitativa via aprendizagem de máquina. O uso de reagentes simples também permitiu discutir limitações práticas relevantes, como a distinção entre mudanças de cor motivadas por variação de pH e aquelas decorrentes de processos oxidativos. A partir destes elementos o trabalho se desenvolveu apoiado na seguinte questão-problema: Como a cor do extrato de repolho roxo, registrada sob diferentes iluminações (LEDs RGB), se relaciona com o pH da solução, e qual condição de iluminação permite estimar o pH com maior consistência?

Os aspectos conceituais que embasam esta investigação, incluindo o comportamento químico das antocianinas, os espaços de cor utilizados e a aplicação de técnicas de aprendizagem de máquina, são apresentados na Seção 2.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 ANTOCIANINAS, EXTRATO DE REPOLHO ROXO E COMPORTAMENTO ÁCIDO-BASE

As antocianinas são corantes naturais presentes em flores, frutos e vegetais, expressando tonalidades que variam do vermelho ao violeta e azul, ocorrendo em diversas espécies e partes de vegetais (Amaral, 2023). O extrato de repolho roxo (*Brassica oleracea var. capitata* f. *rubra*) tem sido amplamente apresentado em artigos científicos, assim como proposto em atividades educacionais, devido à sua eficiência como indicador ácido-base natural (Santos, 2013).

O repolho roxo possui elevado teor de antocianinas, cujo principal composto é a cianidina-3-glucosídeo (Amaral, 2023). Existem distintas formas de extração, incluindo métodos mais sofisticados; no entanto, a maceração em solvente de baixo ponto de ebulição, especialmente em água acidificada, apresenta melhor eficiência em comparação com etanol (Santos, 2013).

O comportamento colorimétrico do extrato apresenta zonas de inversão de tendência e regiões de saturação de cor em pH inferior a 2,0 e superior a 9,0, como discutido por Amaral *et al.* (2023), o que implica em variações reduzidas de tonalidade nesses extremos, aspecto relevante para análises quantitativas de cor.

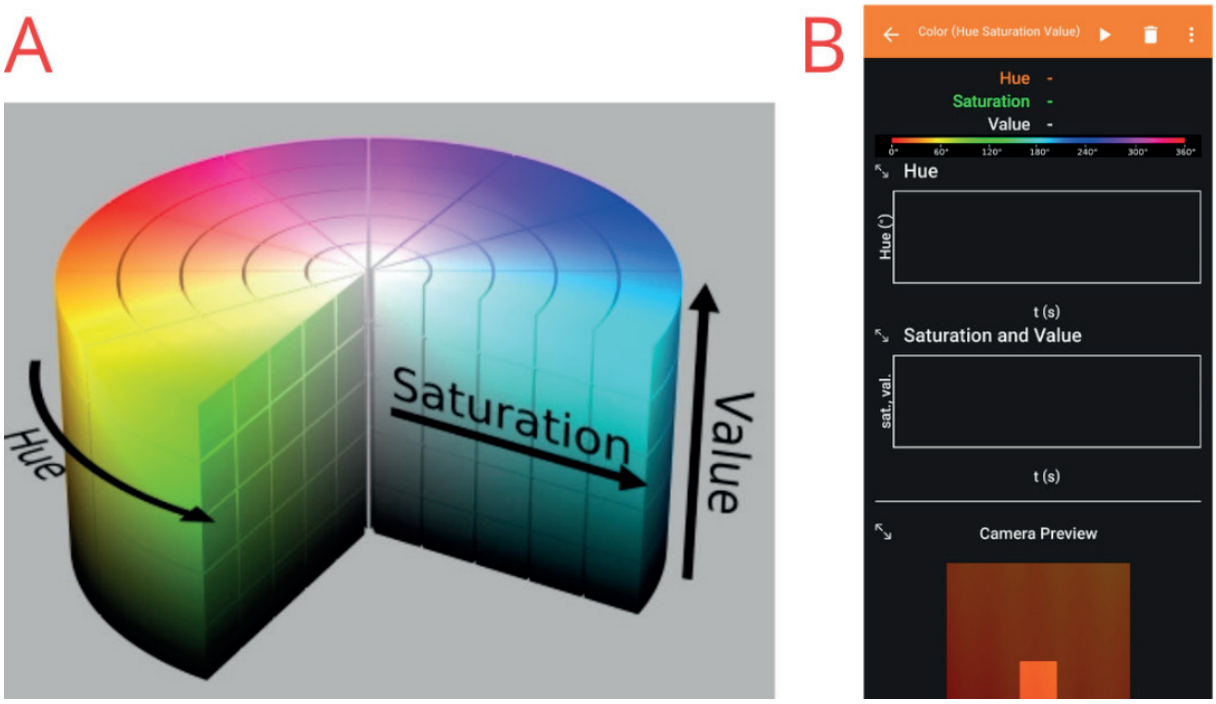
2.2. MEDIDAS DIGITAIS DE COR: RGB, HSV E SUAS APLICAÇÕES ANALÍTICAS

O registro digital de imagem constitui um recurso valioso para análise quantitativa de cor, sendo amplamente utilizado como método rápido e de baixo custo. O modelo RGB é o padrão mais comum, baseado em um espaço cúbico tridimensional que combina os três parâmetros. Entretanto,

sua capacidade de representação pode ser limitada em algumas aplicações, dada sua dependência da composição direta dos canais (Effendhy, 2024).

Em contraste, o espaço HSV (*Hue, Saturation, Value*) descreve cor a partir de três componentes, mais intuitivas no âmbito da percepção humana: tom (0° a 360°), saturação (0 a 1) e valor (0 a 1). O parâmetro H expressa a informação cognitiva primária de cor e é amplamente empregado na correlação entre mudança de tonalidade e transformação química de substâncias para fins de análise quantitativa (Erenas, 2012). A estrutura cilíndrica do espaço HSV, com a distribuição de H, S e V, pode ser visualizada na Figura 1.

Figura 1 - (A) Modelo cilíndrico destacando os eixos de matiz (Hue), saturação (Saturation) e valor (Value). (B) Interface do aplicativo utilizado para exibição dos parâmetros H, S e V a partir da captura de imagem.



Fonte: Autores.

Além disso, a determinação do módulo de cor pode ser obtida pelo comprimento do segmento geométrico euclidiano no espaço HSV, possibilitando relações métricas entre os três parâmetros (Park, 2012). O uso do HSV tem apresentado resultados expressivos em análises químicas portáteis, como no trabalho de Mathaweesansurn *et al.* (2017), que empregaram processamento de imagens capturadas por smartphone para distinguir tons de verde de albumina urinária em reação com Triton X-100.

Vários estudos vêm explorando RGB, HSV e outros espaços vetoriais para determinação quantitativa de substâncias, incluindo correlação entre tonalidade e pH (Park, 2021) e análise de águas subterrâneas (Peeters, 2014). Effendhy *et al* (2024) compararam HSV, YCbCr e YUV na determinação do pH utilizando extrato de repolho roxo, concluindo que o HSV apresentou melhor desempenho no método empregado.

No que diz respeito ao controle de iluminação, sabe-se que o registro digital de imagem é fortemente dependente da fonte de luz, o que exige padronização rigorosa para garantir coerência dos parâmetros H, S e V. A adoção de fontes específicas, como LEDs, constitui estratégia eficaz para assegurar uniformidade e repetibilidade dos dados. Em função dessa dependência espectral, fontes monocromáticas podem acentuar determinados contrastes de cor, motivo pelo qual um LED verde foi empregado nos experimentos discutidos neste trabalho.

2.3. APRENDIZAGEM DE MÁQUINA APLICADA À ANÁLISE COLORIMÉTRICA

O Aprendizado de Máquina (*Machine Learning*) é um conjunto de métodos matemáticos desenvolvidos ao longo das últimas décadas e amplamente difundido desde os anos 1990, especialmente devido ao acesso facilitado a linguagens e ambientes computacionais como Python (Zhai, 2020). Embora qualquer método possa ser implementado a partir de primeiros princípios, APIs modernas disponibilizam grande diversidade de algoritmos, permitindo aplicação direta a diferentes tipos de dados.

A aprendizagem de máquina já vem sendo aplicada à análise de contrastes entre pH e cor (Kim, 2021; Fairclough, 2020; Liu, 2019; Chioson, 2018; Kim, 2017; Moreno, 2012). Nos métodos supervisionados, os dados são divididos em conjuntos de treino e teste, com uma variável-alvo responsável pela convergência matemática do modelo. Em métodos não supervisionados, como os de clusterização, o algoritmo opera sem rótulos, utilizando métricas geométricas e aritméticas para determinar grupos naturais nos dados (Wulan Hastuti, 2020).

Métodos de classificação também podem ser empregados para separar setores distintos nos dados, enquanto técnicas de redução dimensional exploram relações matemáticas entre variáveis para analisar conjuntos multidimensionais, comuns em dados experimentais de ciências (Zhai, 2020). Ventura-Grandez (2025) utilizaram parâmetros colorimétricos derivados do extrato de repolho roxo para identificação de substâncias, reforçando o potencial da análise digital nesse tipo de aplicação.

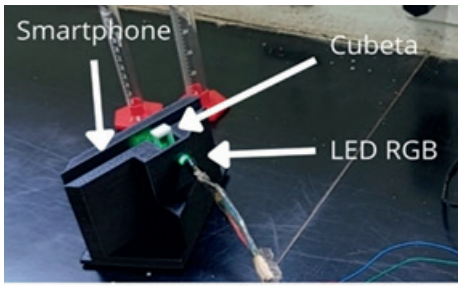
De forma paralela Wulan Hastuti *et al.* (2020) investigaram a correlação entre pH e cor do extrato de repolho roxo utilizando câmera digital profissional, fonte fluorescente e câmara escura; os dados foram adquiridos em RGB e convertidos para diferentes espaços vetoriais, sendo aplicados a métodos de classificação e a modelos de regressão, como *K-Nearest Neighbor* (KNNR) e Artificial *Neural Network* (ANNR). Os autores observaram melhor desempenho do RGB no método K-NN.

3 METODOLOGIA

O estudo é de natureza experimental e instrumental, com objetivo quantitativo (Gil, 2002). A montagem utilizada consistiu em uma peça plástica projetada para alojar uma cubeta de vidro, um

LED RGB, operado individualmente nas cores vermelha, verde e azul, e um smartphone com o modo *Color* disponível na versão beta do aplicativo *Phyphox* (Figura 2).

Figura 2 - Figura com o esquema experimental: fotos do kit.



Fonte: Autores (2025).

O extrato de repolho roxo foi obtido pela fervura de 200,0 g de repolho em 500,0 g de água. A partir desse extrato base, alíquotas foram ajustadas em dez níveis de pH, utilizando vinagre de álcool diluído para acidificação e água sanitária (hipoclorito de sódio 2,5%) para alcalinização. O pH foi medido com um medidor tipo caneta, precisão 0,01. A Tabela 1 apresenta o planejamento volumétrico empregado.

Tabela 1 - Planejamento de dosagem das amostras.

AMOSTRAS	EXTRATO (ML)	VOLUME ADITIVO (ML)	ADITIVO
1	16,0	16,0	Vinagre
2	16,0	12,0	
3	16,0	8,0	
4	16,0	4,0	
5	16,0	4,0	Água
6	20,0	0,4	Hipoclorito de Sódio 2,5%
7	20,0	0,8	
8	20,0	1,2	
9	20,0	1,6	
10	20,0	2,0	

Fonte: Autores (2025).

Cada amostra foi iluminada separadamente por cada cor do LED, sendo realizadas três repetições independentes para cada valor de pH. Em cada repetição, a solução original era descartada e uma nova alíquota era preparada, garantindo independência entre as coletas. Os parâmetros de cor Hue (H), Saturation (S) e Value (V) foram obtidos pelo *Phyphox* diretamente da câmera do smartphone.

O procedimento resultou em um conjunto de 189 amostras, estruturado em cinco características (pH, H, S, V e cor do LED). A hipótese central do estudo é que alguma das cores do LED produziria melhor contraste nos parâmetros de cor, permitindo construir um modelo supervisionado capaz de estimar o pH.

3.1 ANÁLISE DE AGRUPAMENTO DAS CARACTERÍSTICAS DE COR

A clusterização foi aplicada separadamente para os conjuntos obtidos com cada cor do LED. Utilizaram-se três métodos, o primeiro, *K-means*, consiste num método particional amplamente empregado pela boa precisão e baixo custo computacional (Rodriguez, 2019). O número de clusters foi ajustado pelo método do "cotovelo".

O segundo método é o *Agglomerative Clustering, que* utiliza modelo hierárquico baseado na fusão sucessiva de clusters conforme matrizes de distância (Oti, 2024). Inicia com cada ponto como um cluster individual e realiza combinações iterativas até formar um único agrupamento, permitindo avaliar a estrutura hierárquica dos dados.

E o terceiro método foi o *Gaussian Mixture Models* (GMM). Nesse método ocorre uma modelagem probabilística, em que se assume que os dados provêm de uma mistura de distribuições normais multivariadas (Patel, 2020). Mais flexível que o *K-means*, permite estimar verossimilhança e selecionar número de clusters via critérios como AIC/BIC.

Esses métodos foram aplicados para verificar se as amostras apresentavam padrões distintos ou sobreposição entre si para cada cor do LED, avaliando a consistência dos agrupamentos.

3.2 ANÁLISE SUPERVISIONADA A PARTIR DE MÉTODOS DE REGRESSÃO VARIADOS

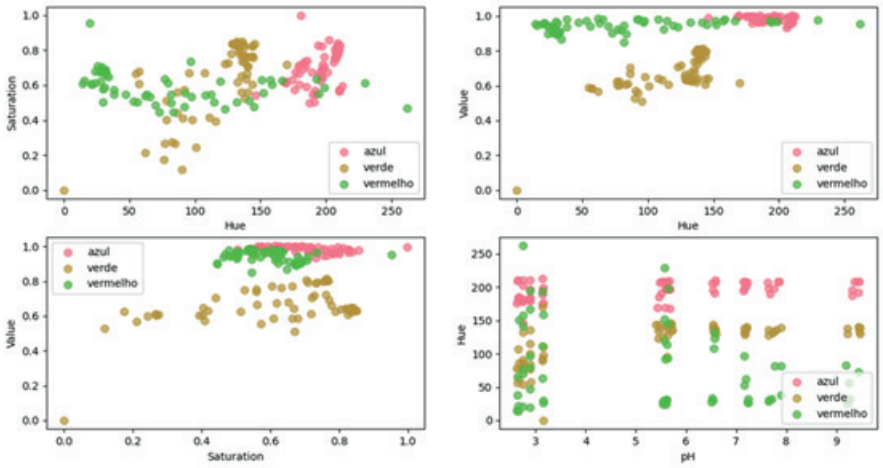
Para estimar o pH a partir dos parâmetros de cor (H, S e V), foram aplicados cinco métodos de regressão, analisados separadamente para cada cor do LED. Os indicadores de desempenho utilizados foram R^2, MAE (*Mean Absolute Error*) e MSE (*Mean Squared Error*), coletados em seus valores ótimos de ajuste. Os métodos empregados foram: *Ridge Regression*, que consiste num modelo de regularização L2 que reduz sobreajuste ao penalizar a magnitude dos coeficientes. *Linear Regression*, que é um modelo paramétrico baseado em mínimos quadrados ordinários, empregado como referência comparativa (Filzmoser, 2021). S*upport Vector Regression* (SVR), que aplica o princípio da margem máxima à regressão, com uso de kernels para capturar relações não lineares e robustez a outliers. *K-Nearest Neighbors Regression* (KNNR), que é baseado em instâncias, que prevê valores pela média dos k vizinhos mais próximos no espaço de características (Cunningham, 2021). E, por fim, *Random Forest Regression*, um método de ensemble que agrega múltiplas árvores de decisão, reduzindo variância e aumentando generalização (Grinsztajn, 2022).

Os modelos foram treinados com os dados obtidos individualmente para cada cor do LED, e posteriormente comparados quanto ao desempenho preditivo.

4 RESULTADOS E DISCUSSÕES

O conjunto de dados foi inicialmente explorado quanto ao cruzamento bidimensional dos quatro parâmetros numéricos. A figura 3 apresenta os quatro gráficos correlacionando os parâmetros de cor (H, S e V) entre si e o parâmetro de tom (*Hue*) com o valor de pH. Podemos observar uma segregação das amostras relativas ao LED verde e um deslocamento destas amostras no gráfico *Saturation x Hue*.

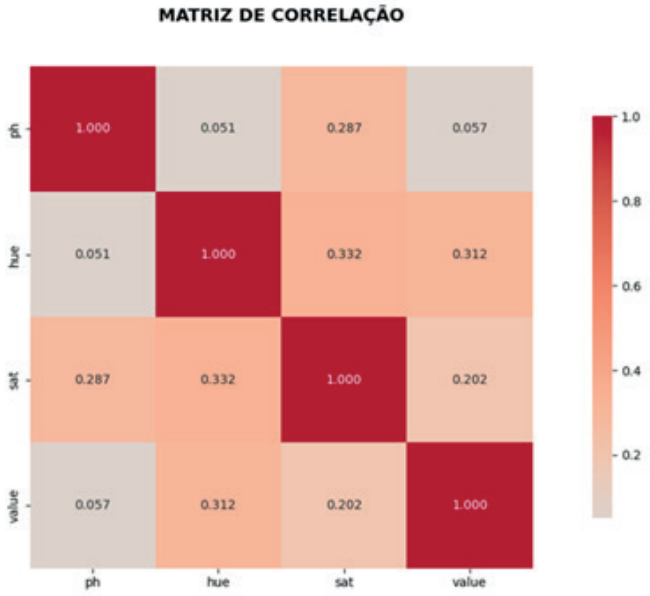
Figura 3 - Cruzamento aos pares entre as variáveis de cor e valor de pH de cada amostra.



Fonte: Autores (2025).

Outra forma de analisar este contexto de amostras deslocadas é a matriz de correlação da figura 4. Na matriz de correlação o valor mais elevado significa um peso de independência dos parâmetros entre estas características. Em especial, os valores de pH apresentam um valor baixo, o que pode indicar tendência relacionada aos demais valores, como tom e saturação, o que limita a independência destes parâmetros.

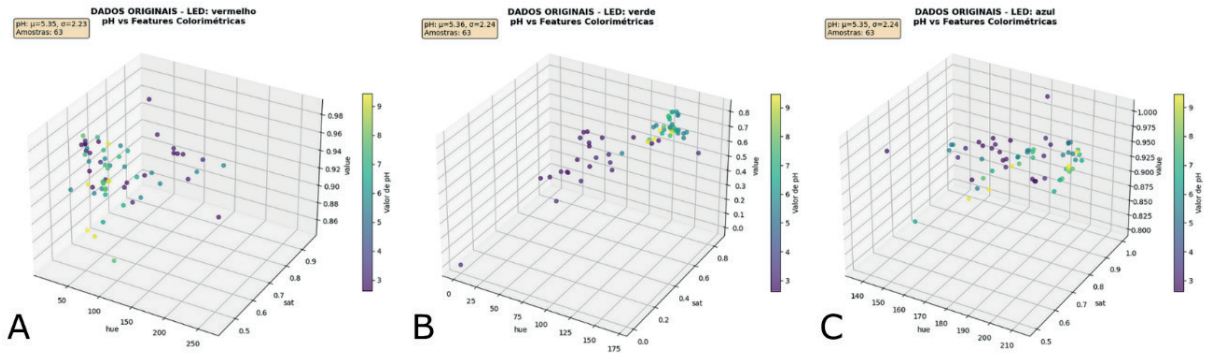
Figura 4 - Correlação de força entre as variáveis.



Fonte: Autores (2025).

A figura 5 apresenta em gráfico 3D dos parâmetros de cor (A. Vermelho, B. Verde e C. Azul) para cada LED com a distribuição dos valores de pH em escala de cor. As amostras relativas aos LEDs vermelho e azul apresentam um formato disperso em todas as direções e com escala de cor de tendência aleatória, o que difere dos valores do LED verde (Fig.5.B), que apresentam uma distribuição gradual, o que sugere tendência objetiva para ajuste preciso dos dados.

Figura 5 - Dados originais dos parâmetros de cor e escala colorida de valor de pH para cada LED, A) Vermelho, B) Verde e C) Azul.

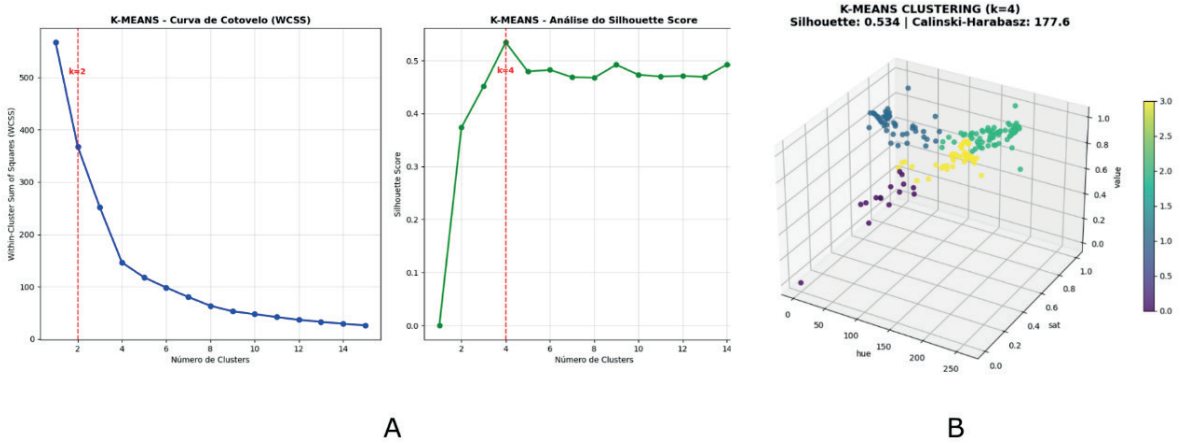


Fonte: Autores (2025).

4.1. ANÁLISE POR CLUSTERIZAÇÃO

Para avaliar a estrutura de agrupamento presente no conjunto de dados, iniciou-se a análise pelos métodos particionais, em especial o *K-Means*. Como primeira etapa, foi necessário determinar o número adequado de clusters. A Figura 6 apresenta a curva de cotovelo e a variação do *Silhouette* Score em função do número de grupos testados, permitindo identificar o ponto de melhor equilíbrio entre compactação interna e separação entre clusters.

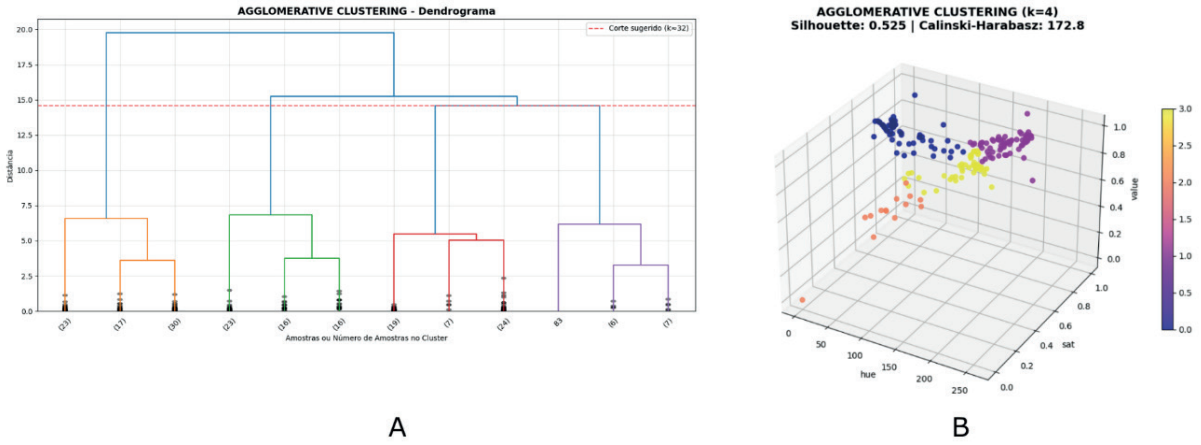
Figura 6 - Curva de cotovelo e otimização do parâmetro *Silhoutte Score.*



Fonte: Autores (2025).

A curva de cotovelo deve ser analisada através da identificação do ponto de inflexão onde a redução marginal na soma quadrática *intra-clusters* (WCSS) torna-se insignificante. Complementarmente, o *Silhouette Score* (Rousseeuw, 1987) fornece validação adicional, sendo que valores próximos de 1 indicam clusters bem definidos. Complementarmente, o dendrograma, que pode ser visto na figura 7, permite visualizar a estrutura hierárquica completa dos dados. A altura das fusões representa a dissimilaridade entre clusters, enquanto o padrão de ramificação revela relações de proximidade.

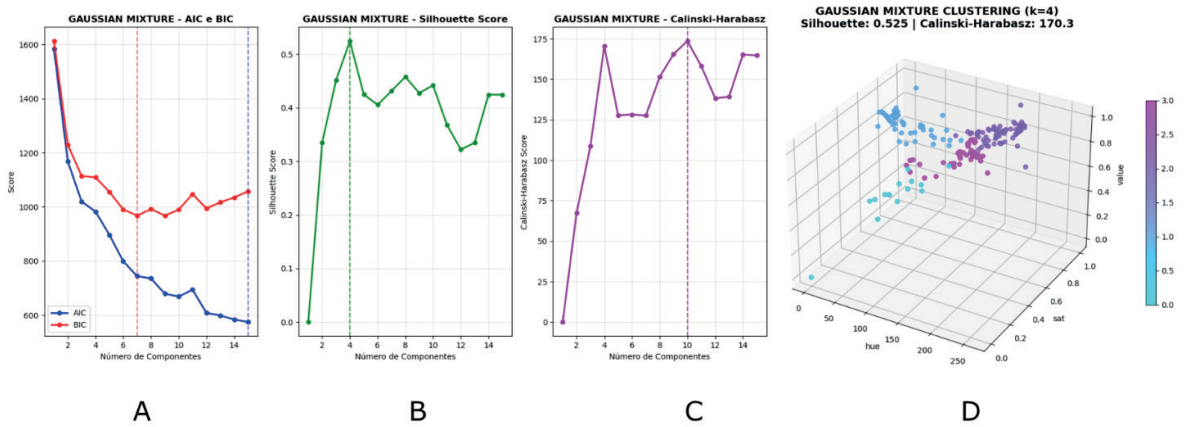
Figura 7 - Dendograma produzido a partir do método *Agglomerative Clustering.*



Fonte: Autores (2025).

Tendo como critérios de Avaliação a altura relativa das fusões: Fusões em baixa altura indicam alta similaridade, enquanto fusões em alta altura sugerem clusters distintivos, a estrutura do dendrograma: Dendrogramas balanceados sugerem clusters de tamanhos similares, enquanto estruturas assimétricas podem indicar presença de *outliers* ou *clusters* de densidade variável e o critério de corte: O ponto de corte ideal localiza-se onde se observa o maior aumento na distância de fusão (Müllner, 2011).

Figura 8 - Resultado de agrupamento, A, B e C) otimização dos parâmetros do método Gaussian Mixture e D) Imagem em 3D dos pontos agrupados.



Fonte: Autores (2025).

Os parâmetros de avaliação de clusterização são variados, destacamos: *Silhouette Coefficient*: Entre -1 e 1, quanto maior melhor, *Calinski-Harabasz Index*: Quanto maior, melhor, *Fowlkes-Mallows*: Similaridade com *K-means*, 1 = perfeita concordância, *Homogeneity*: Cada cluster contém apenas membros de uma única classe, *Completeness*: Todos os membros de uma classe estão no mesmo cluster, V-measure: Média harmônica entre *Homogeneity* e *Completeness* e *Rand Index*: Similaridade entre duas clusterizações.

Silhouette Coefficient (SC) é o coeficiente que mede quão similar um objeto é ao seu próprio cluster comparado aos outros *clusters*: SC > 0.7: Estrutura de *clustering* forte e bem definida, 0.5 < SC ≤ 0.7: Estrutura razoável, 0.25 < SC ≤ 0.5: Estrutura fraca, possibilidade de clusters artificiais e SC ≤ 0.25: Sem estrutura significativa de *clustering.* Já o parâmetro *Calinski-Harabasz Index* (CH), este índice calcula a razão entre dispersão *inter-cluster* e *intra-cluster*, onde valores mais altos indicam clusters mais bem definidos e particularmente eficaz para dados com distribuição esférica e o parâmetro *Fowlkes-Mallows* Score (FM) mede a similaridade entre duas partições, sendo FM = 1: Concordância perfeita entre os agrupamentos, FM = 0: Ausência completa de concordância, Valores intermediários indicam grau de sobreposição. Por fim, o parâmetro *Homogeneity*, *Completeness* e *V-measure*, estas métricas avaliam diferentes aspectos da qualidade do clustering. O parâmetro onde Homogeneidade considera que cada cluster contém apenas membros de uma única classe, completude: Todos os membros de uma dada classe são atribuídos ao mesmo cluster, *V-measure*: Média harmônica entre homogeneidade e completude, *Adjusted Rand Index* (ARI), corrige o Rand Index para o acaso, fornecendo medida mais confiável de similaridade entre partições.

Tabela 2 - Melhores valores de parâmetros de avaliação de cada método.

Métrica	K-Means	Agglomerative	Gaussian Mixture
Número de clusters (k)	4	4	4
Calinski–Harabasz Index	177,5953	172,7886	170,34
Silhouette Coefficient	0,5345	0,5250	0,5246
Fowlkes–Mallows Score	-	0,9478	0,9531
Homogeneity	-	0,9200	0,9162
Completeness	-	0,9200	0,9125
V-measure	-	0,9200	0,9144
Adjusted Rand Index	-	0,9259	0,9336

Fonte: Autores (2025).

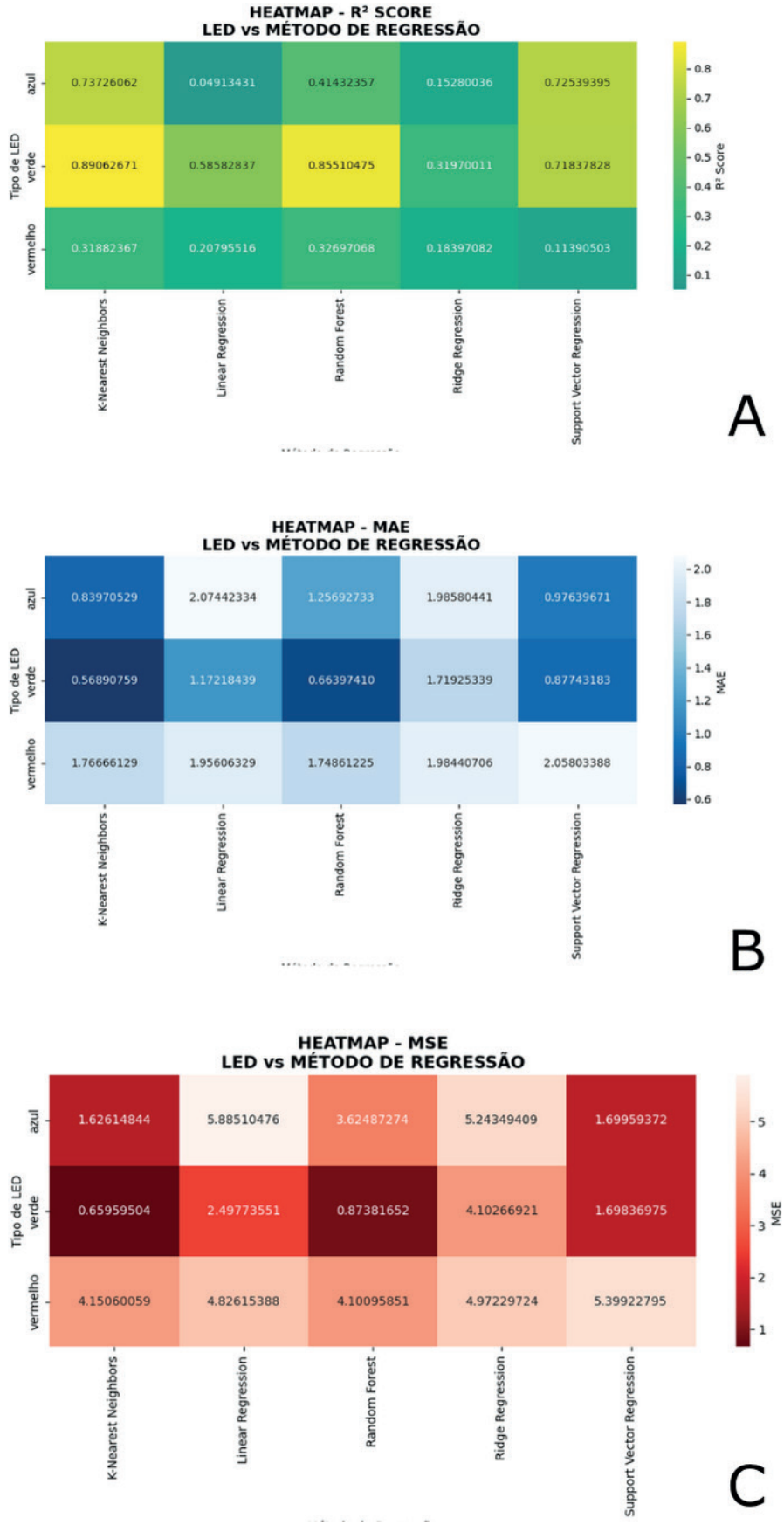
A partir da enorme diversidade de instrumentos matemáticos existentes em termos de metodologias de aprendizado de máquina tanto não supervisionado quanto supervisionado, foram empregados alguns destes instrumentos à título de comparação de performance, já que não existe um método com maior eficiência, cabendo sempre a comparação em cenários, de forma que o interesse principal reside nos índices e parâmetros que apresentem convergência entre os métodos, fortificando as conclusões obtidas.

4.2 ANÁLISE POR REGRESSÃO

Diferentes parâmetros de avaliação de performance foram empregados. A figura 9 sumariza os valores entregues pela avaliação de cada um destes parâmetros. Em termos da análise de conjuntos de dados, existe um potencial computacional muito diversificado onde o método mais adequado precisa levar em consideração características do conjunto de dados e a avaliação comparada. Especialmente no processo de regressão, comportamentos como *overfitting* e *underfitting* precisam ser considerados e analisados, mas dada a complexidade do próprio conjunto de dados, mesmo um modelo com limitantes próprios do processo de hiper-parametrização demandado por cada método, já entregou na forma de resultado uma visão numericamente objetiva sobre a análise do conjunto de dados.

As matrizes da figura 9.A, 9.B e 9.C sumarizam os parâmetros de performance para cada método aplicado a cada LED. A qualidade do valor R^2 score será melhor quanto mais se aproximar de 1, enquanto os valores de MAE e MSE indicarão melhor performance quanto menor for o valor. No caso do conjunto de dados analisado, os cinco métodos empregados reforçaram o destaque apresentado para as amostras produzidas usando a luz do LED verde, tendo resultado consistente para todos os métodos e para os três parâmetros neste trabalho utilizado.

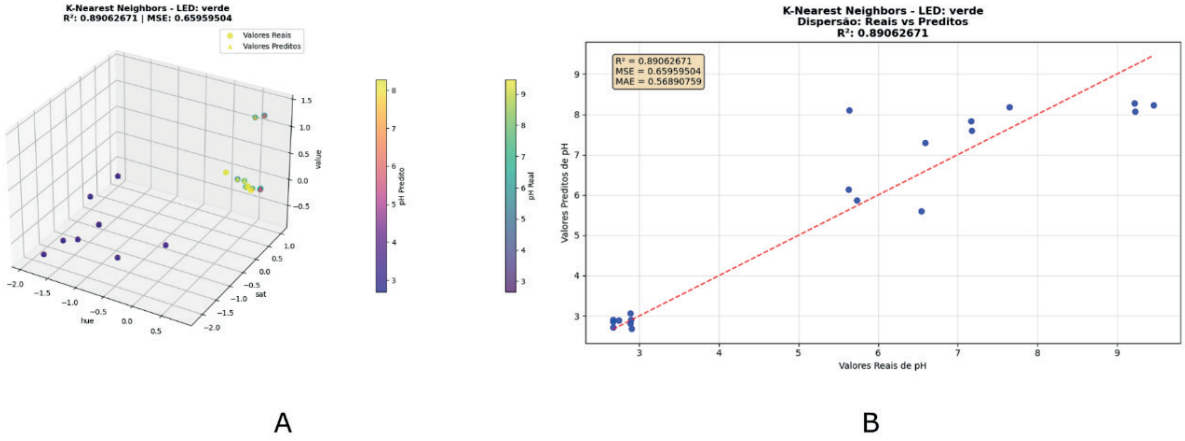
Figura 9 - Mapa de calor dos parâmetros de performance para cada método e para cada categoria LED, A) R^2 score, B) MAE e C) MSE.



Fonte: Autores (2025).

Trazendo um olhar mais aprofundado sobre os resultados obtidos para as amostras produzidas usando o LED verde, é apresentado na figura 10 a convergência principal de resultados para o método que apresentou indicadores de melhor performance. A figura apresenta os valores de pH predito e real em gradação de cor. Na parte A da figura 10, o gráfico 3D destaca em escala de cor a comparação entre os valores de pH reais e os valores preditos, onde pode se observar um erro mínimo. O método que apresentou a melhor performance de ajuste foi K-NN. A figura 10.B apresenta o ajuste entre valores de pH predito e real. Apesar de pontos de escape (*outlier*), um ajuste com fator R^2 Score de 0,89.

Figura 10 - A) distribuição de cada amostra pelos parâmetros de cor com duas escalas de cor: pH predito e pH real; B) Curva de ajuste pH predito versus pH real.



Fonte: Autores (2025).

O emprego de métodos de aprendizagem de máquina oferece elevado poder computacional para a análise de conjuntos de dados multidimensionais; contudo, a qualidade dos resultados permanece dependente da construção adequada do conjunto de dados e das condições experimentais que o originam. Os algoritmos utilizados neste estudo representam apenas uma fração das abordagens disponíveis e pertencem ao grupo de métodos de menor custo computacional. Ainda assim, os resultados obtidos indicam que fontes LED de baixa potência e baixo custo podem ser exploradas de forma mais refinada na caracterização de substâncias, especialmente em regiões de cor em que emissores individuais apresentam desempenho diferenciado.

A adoção de métodos de aprendizado de máquina é uma realidade do presente, que cada dia ocupa mais espaço e técnicas espectroscópicas, essencialmente, são técnicas que apresentam padrão regular e de características repetitivas, que embora tenham um olhar altamente qualificado que é aquele provido pela mente humana, tem potencial para apresentar visualização de micro detalhes que não passariam despercebidos matematicamente no âmbito estatístico. Embora o objetivo central deste trabalho tenha sido a caracterização quantitativa do extrato de repolho roxo sob iluminação controlada e sua análise por técnicas de aprendizagem de máquina, a montagem experimental proposta apresenta potencial de adaptação para ambientes educacionais. A disponibilidade de materiais

simples, como LED RGB, estrutura plástica e registro por smartphone, favorece sua reprodução em atividades de ensino de Química, permitindo discutir limitações de indicadores naturais, influência das condições de iluminação e interpretação de dados experimentais. Essa dimensão didática, no entanto, permanece como uma possibilidade complementar, não como finalidade principal da investigação.

5 CONSIDERAÇÕES FINAIS

O estudo apresentou uma estratégia experimental acessível para a análise quantitativa da resposta colorimétrica do extrato de repolho roxo sob diferentes níveis de pH e condições de iluminação controlada. A combinação entre uma montagem de baixo custo, composta por LED RGB, cubeta e registro por smartphone, e a aplicação de métodos de clusterização e regressão permitiu avaliar, de forma sistemática, como a escolha da fonte de luz influencia a distribuição dos parâmetros de cor no espaço HSV. Os resultados mostraram que a iluminação verde produz a separação mais consistente entre as amostras e possibilita o treinamento de um modelo supervisionado com desempenho robusto, mesmo diante das limitações químicas associadas ao uso de água sanitária como agente alcalinizante.

Esses achados reforçam a utilidade da metodologia proposta para estudos quantitativos envolvendo indicadores naturais, destacando a importância da padronização da iluminação e dos cuidados na escolha de reagentes que possam introduzir interferências não relacionadas ao pH. A abordagem baseada em machine learning demonstrou ser eficaz para explorar padrões nos dados e obter modelos preditivos com boa generalização.

Em resumo, o que manualmente pode se observar certa tendência quando usado o led verde para as mudanças de cor medida, a partir da aplicação de diferentes metodologias de aprendizado de máquina não supervisionada indicou uma estratificação matemática clara quando relacionado cor em função do nível de acidez e alcalinidade da amostra. Em analogia, os resultados apresentados contribuem para que segmentação deste tipo de espectroscopia avance nas metodologias a partir de processamento computacional que requer pouco investimento.

Embora não constitua o foco central do trabalho, a simplicidade do aparato experimental e o uso de instrumentos amplamente disponíveis sugerem que a metodologia pode ser adaptada a contextos educacionais, contribuindo para atividades que integrem experimentação, análise de dados e discussão crítica de limitações metodológicas. Desdobramentos futuros incluem a avaliação de outras fontes de luz, a comparação com diferentes indicadores naturais e a ampliação do conjunto de dados para aprimorar os modelos preditivos.

AGRADECIMENTOS

Os autores gostariam de agradecer o apoio concedido pelo Programação de Bolsas de Iniciação Tecnológica do CNPq no IF Fluminense.

REFERÊNCIAS

DO AMARAL, Fabio Augusto *et al.* Proposta e Aplicação de um Experimento Investigativo para a Construção de Curvas de Titulação com Extrato de Repolho Roxo. **Revista Virtual de Química**, v. 15, n. 1, 2023.

CHIOSON, Francheska B. *et al.* Classification and determination of pH value: a decision tree learning approach. In: **2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)**. IEEE, 2018. p. 1-4.

CUNNINGHAM, Padraig; DELANY, Sarah Jane. K-nearest neighbour classifiers: (with Python examples). **arXiv preprint arXiv:2004.04523**, 2020.

EFFENDHY, Nida Desri; ROTO, Roto; SISWANTA, Dwi. The Application of the HSV Color Model for Accurate Digital Colorimetric Analysis of Fluoride Detection Using a Thiourea Receptor. **Engineering Chemistry**, v. 8, p. 75-80, 2024.

ERENAS, M. M. *et al.* Use of digital reflection devices for measurement using hue-based optical sensors. **Sensors and Actuators B: Chemical**, v. 174, p. 10-17, 2012.

FAIRCLOUGH, Simon M. *et al.* Colorimetric sensor for pH monitoring of liquid samples using bubble wrap and mobile phone camera. In: **2020 IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS)**. IEEE, 2020. p. 1-4.

FILZMOSER, P.; NORDHAUSEN, K. Robust linear regression for high-dimensional data: an overview. WIREs Computational Statistics, v. 13, e1524, 2021. DOI: 10.1002/wics.1524.

GIL, Antonio Carlos *et al.* **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2002.

GRINSZTAJN, Léo; OYALLON, Edouard; VAROQUAUX, Gaël. Why do tree-based models still outperform deep learning on typical tabular data?. Advances in neural information processing systems, v. 35, p. 507-520, 2022.

WULAN HASTUTI, Dian *et al.* Prediction system for pH measurement on Brassica oleraceae (Red Cabbage) using machine learning regression. In: **Journal of Physics: Conference Series**. IOP Publishing, 2020. p. 012050.

KIM, Hyungi *et al.* Fluorescent sensor array for high-precision pH classification with machine learning-supported mobile devices. **Dyes And Pigments**, v. 193, p. 109492, 2021.

KIM, Sung Deuk; KOO, Youngmi; YUN, Yeoheung. A smartphone-based automatic measurement method for colorimetric pH detection using a color adaptation algorithm. **Sensors**, v. 17, n. 7, p. 1604, 2017.

LIU, Ting *et al.* Smartphone-based hand-held optical fiber fluorescence sensor for on-site pH detection. **IEEE Sensors Journal**, v. 19, n. 20, p. 9441-9446, 2019.

MATHAWEESANSURN, Arjnarong; MANEERAT, Noppadol; CHOENGCHAN, Nathawut. A mobile phone-based analyzer for quantitative determination of urinary albumin using self-calibration approach. **Sensors and Actuators B: Chemical**, v. 242, p. 476-483, 2017.

MORENO, Ivan. Image-like illumination with LED arrays: design. **Optics Letters**, v. 37, n. 5, p. 839-841, 2012.

MÜLLNER, Daniel. Modern hierarchical, agglomerative clustering algorithms. **arXiv preprint arXiv:1109.2378**, 2011.

OTI, Eric; OLUSOLA, Michael. Overview of agglomerative hierarchical clustering methods. **British Journal of Computer, Networking and Information Technology**, v. 7, n. 2, p. 14-23, 2024.

PATEL, Eva; KUSHWAHA, Dharmender Singh. Clustering cloud workloads: K-means vs gaussian mixture model. **Procedia Computer Science**, v. 171, p. 158-167, 2020. DOI: 10.1016/j.procs.2020.04.017.

PARK, Habeen; KOH, Young Gook; LEE, Wonmok. Smartphone-based colorimetric analysis of structural colors from pH-responsive photonic gel. **Sensors and Actuators B: Chemical**, v. 345, p. 130359, 2021.

PEETERS, Luk. A background color scheme for piper plots to spatially visualize hydrochemical patterns. **Groundwater**, v. 52, n. 1, p. 2-6, 2014.

RODRIGUEZ, Mayra Z. *et al.* Clustering algorithms: A comparative approach. **PloS one**, v. 14, n. 1, p. e0210236, 2019.

ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53-65, 1987. DOI: 10.1016/0377-0427(87)90125-7.

SANTOS, Gilcenir *et al.* Caracterização físico-química do repolho roxo (Brassica oleracea). **REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS**, v. 3, n. 5, p. 001-012, 2013.

VENTURA-GRANDEZ, Henry E. *et al.* Deep Neural Network Assisted Microfluidic pH Sensor. **IEEE Sensors Journal**, 2025.

ZHAI, Xiaoming *et al.* Applying machine learning in science assessment: a systematic review. **Studies in science education**, v. 56, n. 1, p. 111-151, 2020.