IV SICITAL
SIMPÓSIO DE CIÊNCIA E TECNOLOGIA DE ALIMENTOS

PREDIÇÃO DE ANTOCIANINAS EM DIFERENTES MATÉRIAS-PRIMAS USANDO *MACHINE LEARNING*

Renires dos Santos Teixeira[1]*, Arliene do Socorro Batista dos Santos[2], Flavia Tayna Serra Silva[3], Gabriel Laquete de Barros[4], José Matheus Santos Oliveira[5], Andreza de Brito Leal[6], Natássia Rafaelle Medeiros Sirqueira[7], Márcia Vizzotto[8], Patrícia A. Jaques[9], Leonardo Nora[10]

[1*] Universidade Federal de Pelotas, Departamento de Ciência e Tecnologia Agroindustrial, Pelotas, Rio Grande do Sul, Brasil, ORCid: 0000-0002-6953-7063, E-mail: reniresantos@gmail.com
[2] Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico, Pelotas, Rio Grande do Sul, Brasil, ORCid: 0009-0009-9820-6296, E-mail: arlienebatista@gmail.com
[3] Universidade Federal de Pelotas, Departamento de Ciência e Tecnologia Agroindustrial, Pelotas, Rio Grande do Sul, Brasil, ORCid: 0009-0002-2383-8423, E-mail: flavia.belavista2@gmail.com
[4] Universidade Federal de Pelotas, Departamento de Ciência e Tecnologia Agroindustrial, ORCid: 0000-0001-5411-3906, E-mail: gabrielbarros95@yahoo.com.br
[5] Universidade Federal de Pelotas, Departamento de Ciência e Tecnologia Agroindustrial, Pelotas, Rio Grande do Sul, Brasil, ORCid: 0000-0003-2193-4788, E-mail: josematheussantos98@gmail.com
[6] Universidade Federal de Pelotas, Departamento de Ciência e Tecnologia Agroindustrial, Pelotas, Rio Grande do Sul, Brasil, ORCid: 0000-0001-5605-0507, E-mail: andrezaleal.tecno@gmail.com
[7] Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico, Pelotas, Rio Grande do Sul, Brasil, ORCid: 0000-0003-4601-7396, E-mail: natassiamsads@gmail.com
[8] Embrapa Clima Temperado, Laboratório de Ciência e Tecnologia de Alimentos, Pelotas, Rio Grande do Sul, Brasil, ORCid: 0000-0002-8071-4980, E-mail: marcia.vizzotto@embrapa.br
[9] Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico, Pelotas, Rio Grande do Sul, Brasil, ORCid: 0000-0002-2933-1052, E-mail: patricia.jaques@gmail.com
[10] Universidade Federal de Pelotas, Departamento de Ciência e Tecnologia Agroindustrial, Pelotas, Rio Grande do Sul, Brasil, ORCid: 0000-0002-4675-1403, E-mail: l.nora@me.com

RESUMO

As antocianinas, pigmentos naturais com reconhecidas propriedades bioativas e aplicações como corantes alimentícios, apresentam desafios significativos devido à sua instabilidade e degradação. A predição de antocianinas em fontes vegetais é essencial para controle de qualidade e estudos sobre estabilidade. O *Machine Learning* (ML), uma subárea da Inteligência Artificial que desenvolve algoritmos capazes de aprender padrões a partir de dados, surge como ferramenta poderosa na ciência de alimentos e química analítica, oferecendo novas abordagens para predição de antocianinas. Este trabalho explora as aplicações do ML na predição de antocianinas, incluindo a predição do seu comportamento sob diversas condições, análise de dados de degradação, a

validação de métodos de quantificação, a otimização de processos de extração e a implementação de análises não destrutivas. Os estudos analisados demonstram que técnicas de ML de análise de dados, combinadas com métodos não destrutivos, representam um avanço significativo na predição e monitoramento de compostos antociânicos. O sistema *Neuro Fuzzy Inference System* (ANFIS) mostrou-se eficiente na previsão da degradação de antocianinas em casca de uva, cenoura preta e repolho roxo, oferecendo uma alternativa robusta para a otimização de processos industriais, como pasteurização. Assim como, abordagens baseadas em imagens hiperespectrais (HSI) e algoritmos de aprendizado de máquina, como *Stacked Auto-Encoder- Genetic Algorithm - Extreme Learning Machine* (SAE-GA-ELM), *Random Forest* (RF) e *CatBoost*, comprovaram sua eficácia na predição não destrutiva de antocianinas em amora-preta, pétalas de *Rosa chinensis*, alface roxa e folhas de macieira. Os métodos superam as limitações das técnicas tradicionais (como espectrofotometria) ao reduzir custos, tempo de análise e danos às amostras, além de possibilitar análises mais rápidas, automatizadas e com maior precisão, mesmo em grandes volumes de dados. A integração entre espectroscopia, processamento de imagens e ML surge como uma ferramenta promissora para aplicações na indústria alimentícia, possibilitando maior eficiência na preservação de compostos antociânicos e na qualidade dos produtos.

Palavras-chave: Pigmentos Naturais; Aprendizado de Máquina; Algoritmos, Inteligência Artificial.

INTRODUÇÃO

As antocianinas constituem uma vasta classe de compostos fenólicos solúveis em água e são pigmentos naturais amplamente distribuídos no reino vegetal responsáveis pelas cores vibrantes que variam do vermelho, ao azul e roxo observadas em diversas frutas, vegetais, flores, entre outros (Chaves *et al.*, 2018; Sinopoli, Calogero e Bartolotta, 2019; Kaur *et al.,* 2021). Para além da sua função como corantes, as antocianinas são reconhecidas pelas suas significativas propriedades bioativas, incluindo atividade antioxidante e potenciais propriedades anticancerígenas (Anjos *et al*., 2020; Diaconeasa *et al*., 2020; Fakhri *et al*., 2020).

A concentração de antocianinas pode variar devido a diversos fatores ambientais e intrínsecos, como pH, temperatura, luz, oxigênio, presença de enzimas e íons metálicos (Sunarya *et al*., 2024; Dai *et al*., 2022). Essa variação no conteúdo pode comprometer a padronização da cor, a bioatividade e, consequentemente, limitar suas aplicações industriais e benefícios à saúde (Akther *et al*., 2020; Lin *et al*., 2023). Assim, a quantificação precisa da concentração de antocianinas representa um desafio importante para o desenvolvimento de métodos voltados à sua estabilidade, aplicação industrial e garantia da qualidade. A validação de métodos analíticos é fundamental para garantir a precisão na determinação da concentração de antocianinas em diferentes matérias-primas. Este processo envolve a avaliação rigorosa de parâmetros como seletividade, linearidade,

limites de detecção e quantificação, precisão, exatidão e robustez (Garcia-Oliveira *et al*., 2021).

Os métodos tradicionais para quantificação de antocianinas consistem em espectrofotométricos e cromatográficos. Os métodos espectrofotométricos baseiam-se na medição da absorbância em comprimento de onda específico para realizar a quantificação total (Dai *et al.,* 2022). Enquanto, o método cromatográfico consiste na separação, identificação e quantificação de antocianinas individuais (Gao *et al*., 2024; Mesquita et al., 2023).

Embora métodos tradicionais como a cromatografia ofereçam resultados confiáveis, apresentam limitações em termos de custo e tempo de análise, particularmente para grandes volumes de amostras ou monitoramento contínuo (Liu *et al*., 2024).

Neste contexto, o *Machine Learning*[1] (ML), subárea da Inteligência Artificial que se dedica ao desenvolvimento de algoritmos capazes de aprender a partir de dados, surge como uma abordagem promissora para auxiliar na predição de antocianinas, principalmente por sua capacidade de analisar grandes conjuntos de dados complexos e identificar padrões que seriam difíceis ou impossíveis de discernir por métodos tradicionais, o que ajuda a superar algumas das limitações existentes.

A capacidade preditiva do ML permite antecipar o comportamento das antocianinas sob diversas condições, auxiliando no desenvolvimento de produtos mais estáveis. O ML compreende um conjunto de técnicas computacionais que desenvolvem modelos matemáticos capazes de aprender padrões a partir de dados sem programação explícita. Suas principais vantagens incluem a análise eficiente de grandes volumes de dados, a identificação de relações complexas e não lineares, e a capacidade de melhorar continuamente a precisão das previsões com a incorporação de novos dados, otimizando assim processos de pesquisa e desenvolvimento (Revelou *et al*., 2025).

Diante do contexto, este trabalho de revisão tem como finalidade explorar a aplicação de técnicas de ML na predição da concentração de antocianinas totais. A abordagem adotada fundamenta-se na análise de estudos publicados em reconhecidos repositórios acadêmicos, tais como: *Open Access Journals, ScienceOpen, PubMed Central, ScienceDirect, IEEE, ResearchGate e Google Scholar*. A partir dessa base, se identificou os algoritmos utilizados para predição desse composto e como podem contribuir para o avanço de soluções analíticas voltadas ao controle de qualidade e à aplicação industrial das antocianinas.

[1] *Machine Learning* é um termo em inglês que significa "Aprendizado de Máquina" em português.

MÉTODOS TRADICIONAIS PARA PREDIÇÃO DE ANTOCIANINAS

Os métodos tradicionais para predição da concentração das antocianinas englobam uma variedade de técnicas, sendo os métodos espectrofotométricos e cromatográficos.

A Espectroscopia Ultravioleta-Visível (UV-Vis) é frequentemente utilizada para quantificar a concentração total de antocianinas e para monitorizar a sua degradação ao longo do tempo, através da medição da absorbância em comprimentos de onda específicos (Dai *et al.,* 2022). As antocianinas são capazes de absorver fortemente luz na região do visível. Essa característica particular permite a quantificação das antocianinas por métodos espectrofotométricos como método do pH único e pH diferencial (Fuleki e Francis, 1968; Teixeira; Stringheta; Oliveira, 2008).

Outros métodos que permitem a quantificação são os métodos cromatográficos, em particular a Cromatografia Líquida de Alta Eficiência (HPLC – ou mesmo suas variações - UPLC, UHPLC), pois oferecem uma maior resolução e especificidade na análise (Gao *et al*., 2024). Estas técnicas permitem a separação, identificação e quantificação de antocianinas individuais, bem como dos seus produtos de degradação, fornecendo informações detalhadas sobre a composição das amostras ao longo do tempo. Cabe salientar que, nestas técnicas há o acoplamento de detectores como o DAD (Arranjo de Diodos) e/ou MS (Espectrometria de Massa) que permitem, ainda mais, a especificidade da análise (Gao *et al*., 2024; Mesquita et al., 2023). Os métodos cromatográficos são cruciais para estudos de estabilidade, pois permitem o acompanhamento da degradação de cada componente individualmente.

Apesar da sua importância, os métodos tradicionais apresentam algumas limitações, pois, esses métodos, geralmente, são dispendiosos em termos de tempo e de custo, sobretudo para a análise de muitas amostras e, geralmente, envolvem etapas de pré-processamento (Gao *et al*., 2024). Nesse contexto, a aplicação de ML pode ser considerada uma aliada.

MACHINE LEARNING EM CIÊNCIA DE ALIMENTOS E QUÍMICA ANALÍTICA

O ML é uma área da inteligência artificial amplamente aplicada na construção de modelos preditivos e de estimativa. Por meio de algoritmos matemáticos ou computacionais, um modelo computacional é treinado para solucionar problemas ou realizar tarefas complexas com base em parâmetros de entrada fornecidos (Russell e Norvig, 2021). Um exemplo de abordagem

em ML seria treinar uma máquina para prever os resultados de reações químicas conhecidas, usando um grande conjunto de dados como base. Depois disso, o algoritmo treinado poderia ser usado para descobrir formas de criar moléculas mais complexas (Simeone, 2018).

Existem duas abordagens principais para o treinamento de modelos de Machine Learning. Na aprendizagem supervisionada, o algoritmo é treinado com exemplos de pares entrada-saída rotulados, aprendendo a mapear entradas para saídas conhecidas e permitindo, posteriormente, a predição de categorias ou valores para novos dados. Por outro lado, a aprendizagem não supervisionada é aplicada quando os dados não possuem rótulos pré-definidos, tendo como objetivo identificar padrões, estruturas ou relações intrínsecas nos dados, como agrupamentos ou redução de dimensionalidade, sem orientação externa explícita (Simeone, 2018).

O desenvolvimento de modelos de ML envolve etapas fundamentais, começando pela preparação de dados, que inclui a aquisição de dados relevantes de diversas fontes (Petrelli, 2022), o pré-processamento para tratar valores ausentes, discrepantes e normalizar recursos (More e Kumar, 2024; Munde, 2024), e a engenharia de *features* para criar ou modificar variáveis que aumentem a capacidade preditiva do modelo (More e Kumar, 2024). Em seguida, no treinamento de modelos, seleciona-se o algoritmo adequado ao tipo de problema (supervisionado ou não supervisionado) (Munde, 2024), divide-se o conjunto de dados em treinamento e teste, e aplicam-se técnicas de otimização para minimizar erros (More e Kumar, 2024; Petrelli, 2022). Por fim, a avaliação e implantação abrangem a validação do modelo com métricas como acurácia, precisão e *recall* (More e Kumar, 2024; Munde, 2024). A seleção de características relevantes (*feature selection*) e a criação de novas características a partir dos dados existentes (*feature engineering*) são etapas fundamentais para o sucesso dos modelos de ML. Como discutido por Russell e Norvig (2021), a qualidade e representatividade das características têm impacto direto na capacidade de generalização dos modelos.

O ML está revolucionando a análise de qualidade e segurança de alimentos ao substituir métodos tradicionais demorados por abordagens proativas e precisas, como monitoramento em tempo real e análise preditiva, que elevam os padrões de segurança alimentar. Ele contribui por meio da análise preditiva e automação, prevendo problemas de qualidade e reduzindo *recalls* e desperdícios (Abass *et al*., 2024), permite avaliações dinâmicas, precisas e automatizadas de diversos atributos, como sabor, textura, frescor e conteúdo nutricional, garantindo avaliações padronizadas e objetivas, além de técnicas avançadas, como narizes eletrônicos e cromatografia

gasosa, que permitem a detecção não destrutiva de contaminações microbianas e compostos perigosos (Pratap *et al*., 2024; Altaf e Ksouri, 2024). O ML também otimiza a cadeia de suprimentos, melhorando rastreabilidade, transparência, condições de armazenamento e sustentabilidade (Altaf e Ksouri, 2024). Apesar desses avanços, desafios como a integração com sistemas existentes ainda precisam ser superados para maximizar os benefícios do ML no setor alimentício.

No contexto da química analítica, o ML demonstra ser uma ferramenta valiosa na validação de métodos. Pode ser utilizado para avaliar a precisão, eficiência e automação, revolucionando práticas tradicionais. Ele melhora a análise de dados complexos, como em espectroscopia e cromatografia, e facilita a determinação de componentes em produtos naturais (Rial, 2024).

ALGORITMOS DE *MACHINE LEARNING* UTILIADOS NA PREDIÇÃO DE ANTOCIANINAS

O ML tem demonstrado grande potencial na avaliação de métodos de quantificação de antocianinas. Algoritmos de ML podem ser utilizados para avaliar o efeito do tratamento térmico, diferentes níveis de acidez e tempo na concentração e degradação de antocianinas (Ekici *et al*., 2014). Técnicas baseadas em análise de imagem digital são utilizadas para avaliação não destrutiva da concentração e da estabilidade de antocianinas permitindo o monitoramento em tempo real da sua degradação em plantas sem a necessidade de destruir as amostras (Askey *et al*., 2019; Zhang *et al*., 2025).

Como demonstrado por Modesto Junior *et al*. (2023), a degradação das antocianinas é significativamente influenciada pela temperatura e exposição à luz, com a faixa de 60 a 80 °C sendo crítica para sua estabilidade. Tal como observado em Polat Kaya *et al*. (2024), modelos de ML podem prever, com alta acurácia o rendimento, a estabilidade de antocianinas em função de parâmetros de extração (pH, temperatura e tempo). Esses dados experimentais podem ser utilizados em modelos de ML para prever a cinética de degradação e otimizar condições de processamento e armazenamento (Modesto Junior *et al*., 2023).

Diversos algoritmos de ML têm sido empregados na predição da concentração de antocianinas. As Redes Neurais Artificiais (RNAs), que são modelos computacionais inspirados no funcionamento do cérebro humano, são frequentemente utilizadas para modelar relações não lineares entre as variáveis e prever o conteúdo de antocianinas (Liu *et al*., 2024). As Redes Neurais

Convolucionais (RNCs), um tipo especializado de RNA particularmente eficaz no processamento e análise de dados visuais, têm sido aplicadas na análise de dados espectrais e de imagem para a extração de características relevantes nas amostras contendo antocianinas (Wang *et al*., 2024). As RNAs apresentam a vantagem de poderem modelar relações complexas e não lineares e de se adaptarem a diferentes tipos de dados. No entanto, elas requerem grandes conjuntos de dados para treinamento, apresentam risco de *overfitting* (sobreajuste, fenômeno em que o modelo se adapta excessivamente aos dados de treinamento, perdendo capacidade de generalização) e podem ser difíceis de interpretar (Russell e Norvig, 2021).

O algoritmo *Random Forest* (RF), um método de *ensemble* que combina múltiplas árvores de decisão, é utilizado tanto para tarefas de regressão como de classificação na predição do conteúdo de antocianinas e na avaliação da importância das variáveis (Liu *et al*., 2024). O RF é robusto ao sobreajuste, apresenta bom desempenho com dados de alta dimensionalidade e fornece informações sobre a importância das variáveis, embora possa ser menos eficaz para problemas altamente não lineares em comparação com as RNAs (Russell e Norvig, 2021).

As *Support Vector Machines* (SVMs) têm sido aplicadas na predição de antocianinas (Li *et al*., 2023; Zhang *et al*., 2025), tanto para tarefas de classificação (categorização de amostras em classes distintas, como alta ou baixa concentração de antocianinas) quanto para regressão (estimativa de valores contínuos, como a concentração exata de antocianinas em mg/L). As SVMs são eficazes em espaços de alta dimensão e apresentam bom desempenho com conjuntos de dados menores, mas a escolha do *kernel* pode ser crítica e podem ser computacionalmente intensivas para grandes conjuntos de dados (Russell e Norvig, 2021).

A Regressão Linear (RL), incluindo Mínimos quadrados ordinários (MQO) e Mínimos quadrados parciais (MQP), é um método clássico e utilizado para modelar relações entre variáveis contínuas, caracterizando-se por sua simplicidade e interpretabilidade, podendo ser utilizada como base de comparação com métodos mais complexos e para identificar relações lineares entre as variáveis e, assim, prever a concentração de antocianinas (Russell e Norvig, 2021; García-Curiel *et al*., 2023). Outros algoritmos como o *Extreme Learning Machine* (ELM), o *CatBoost* e o *Neuro Fuzzy Inference System* (ANFIS) também têm sido utilizados com sucesso na predição de antocianinas (Liu *et al*., 2024b; Zhang *et al*., 2025; Ekici *et al*., 2014).

A escolha do algoritmo de ML depende das características do problema e da natureza dos dados disponíveis. É comum comparar diferentes modelos utilizando técnicas de avaliação

robustas, como a validação cruzada, um procedimento que divide repetidamente os dados em subconjuntos para treinamento e teste, permitindo avaliar a consistência do desempenho do modelo em diferentes porções dos dados. Esta técnica visa selecionar o algoritmo com melhor capacidade de generalização, ou seja, aquele com melhor desempenho em dados não utilizados durante o treinamento.

A avaliação do desempenho envolve o uso de diversas métricas quantitativas. O erro quadrático médio (RMSE) mede a média dos quadrados dos erros entre valores previstos e observados, onde valores mais baixos indicam melhor precisão. O coeficiente de determinação (R^2) quantifica a proporção da variância na variável dependente que é previsível a partir das variáveis independentes, com valores próximos a 1 indicando excelente ajuste do modelo. A razão de desvio de performance (RPD), calculada como a razão entre o desvio padrão dos valores de referência e o erro padrão de predição, classifica a capacidade preditiva do modelo, onde valores superiores a 3 indicam excelente capacidade preditiva. Estas métricas, aplicadas em conjuntos de validação e teste, permitem uma estimativa mais confiável da capacidade de generalização do modelo (Russell e Norvig, 2021).

Em síntese, o ML pode fornecer melhores percepções sobre os mecanismos de degradação das antocianinas, permitindo o desenvolvimento de estratégias de proteção mais eficazes. Além disso, a análise comparativa de métricas (erro quadrático médio, correlação) validam a eficácia desses modelos na seleção de métodos de quantificação, assegurando confiabilidade nos resultados.

APLICAÇÃO DE *MACHINE LEARNING* NA PREDIÇÃO DA CONCENTRAÇÃO DE ANTOCIANINAS

A Tabela 1 apresenta estudos recentes sobre a aplicação de técnicas de ML na predição de antocianinas. Estes trabalhos demonstram a versatilidade e eficácia do ML para abordar diversos aspectos relacionados à quantificação destes compostos, evidenciando como diferentes algoritmos e métodos podem ser empregados para resolver questões específicas em análises de antocianinas. A tabela detalha os algoritmos aplicados em diferentes matrizes vegetais, desde uvas e amoras-pretas até alface roxa e folhas de macieira, incluindo métricas de desempenho como coeficientes de determinação (R^2) e erro quadrático médio (RMSE), permitindo uma comparação clara entre as diversas abordagens e seus respectivos resultados na predição de antocianinas.

Tabela 1 – Síntese dos resultados da aplicação de algoritmos de *machine learning* para predição de antocianinas.

Algoritmos	Aplicação	Principais Resultados
Neuro Fuzzy Inference System (ANFIS), Redes Neurais Artificiais (RNAs)	Casca de uva, cenoura preta e repolho roxo (Ekici *et al*., 2014)	O modelo ANFIS superou o RNAs na predição da degradação de antocianinas, apresentando menor erro (RMSE de até 0,0457) e maior precisão (R^2 de 0,9942).
Least Squares-Support Vector Machine (LS-SVM), *Extreme Learning Machine* (ELM) otimizado por Algoritmo Genético (GA) (SAE-GA-ELM)	Amora-preta (Li *et al*., 2023).	O modelo SAE-GA-ELM apresentou o melhor desempenho na detecção e visualização do conteúdo de antocianina, alcançando R^2 de 0.97 e RMSE de 0.22 mg/g na predição não destrutiva de antocianinas.
Back-Propagation Neural Network (BPNN), *Random Forest* (RF)	Pétalas de *Rosa chinensis* (Liu *et al*., 2024a)	O RF obteve maior precisão (R^2 = 0.958, RPD = 4.832) que o BPNN.
Extreme Learning Machine (ELM)	Alface roxa (Liu *et al*., 2024b)	O melhor desempenho foi alcançado pelo modelo UVE-CARS-SNV-DBO-ELM, com R^2 de 0,8623 (treino) e 0,8617 (validação), RMSE de 0,0095 mg/g e RPD de 2,7192.
Random Forest Regression (RFR), *Support Vector*	Folhas de macieira (Zhang *et al*., 2025)	O modelo *CatBoost* utilizando espectros originais + segunda

Regression (SVR), *CatBoost*	derivada e índices espectrais para todo o período de crescimento alcançou a maior precisão, com R^2=0.934 e RPD = 3.888.

Fonte: Autoria própria (2025).

Ekici *et al*., (2014) utilizaram o *Neuro Fuzzy Inference System* (ANFIS) para modelar a degradação e prever o conteúdo total de antocianinas em extratos de fontes como casca de uva, cenoura preta e repolho roxo. O estudo comparou o desempenho do ANFIS com o de RNAs para avaliar os efeitos de temperatura, tempo e pH (variáveis de entrada) na degradação de antocianinas. As variáveis de entrada foram normalizadas (0,2–0,8) para evitar viés, e os dados foram divididos em conjuntos de treinamento (50%), teste (25%) e validação (25%). O ANFIS foi superior a RNA em precisão, especialmente para repolho roxo, devido à sua capacidade de lidar com relações não lineares e incertezas nos dados. O estudo de Ekici *et al*., (2014) demonstrou que o ANFIS é ideal para modelar processos alimentares com múltiplas variáveis interdependentes, além de oferecer uma ferramenta eficiente e de baixo custo para a indústria alimentícia prever a degradação de corantes naturais, otimizando processos como pasteurização e armazenamento.

Outro estudo de Li *et al*., (2023) empregou algoritmos avançados para análise hiperespectral de amoras, utilizando o *Successive Projections Algorithm* (SPA) para selecionar 7 variáveis-chave, o *Competitive Adaptive Reweighted Sampling* (CARS) para identificar 15 bandas espectrais relevantes e o *Stacked Auto-Encoder* (SAE) para extrair 13 características não lineares profundas por meio de redes neurais *autoencoder*, capturando padrões complexos nos pixels das imagens (Li *et al*., 2023). Foram comparados dois modelos de ML. Para a modelagem preditiva, comparou-se o desempenho do GA-LS-SVM (que combina SVMs com otimização por algoritmo genético) com o GA-ELM (baseado em ELM e algoritmos genéticos). A combinação de imagens hiperespectrais com o modelo SAE-GA-ELM mostrou-se superior para a predição e visualização do teor de antocianinas, oferecendo alta precisão e eficiência.

Em pétalas de *Rosa chinensis*, a predição do conteúdo de antocianinas foi realizada utilizando imagens digitais e algoritmos como a *Back-Propagation Neural Network* (BPNN) e o *Random Forest* (RF) (Liu *et al*., 2024a). Foram capturadas imagens de 168 amostras de pétalas (3 pétalas por flor) com uma câmera digital (Canon EOS 500D). Utilizou-se o *software* ImageJ para

extrair valores médios dos canais R (*Red*), G (*Green*), B (*Blue*) de 9 regiões de interesse por amostra. 31 índices RGB foram calculados para correlacionar com o conteúdo de antocianinas. Os índices RGB das imagens foram utilizados como variáveis de entrada. O modelo RF demonstrou um desempenho superior ao modelo BPNN na predição do conteúdo de antocianinas, devido à sua capacidade de lidar com multicolinearidade entre índices RGB e menor sensibilidade a *overfitting* (Liu *et al*., 2024a).

O estudo de Liu *et al*., (2024b) propõe um método não destrutivo para monitorar os níveis de antocianina em alface roxa usando hiperspectralidade e algoritmos de ML otimizados. Os dados espectrais foram pré-processados usando processamento espectral de *Standard Normal Variate* (SNV) e *First-Derivative* (FD). Os comprimentos de onda das características foram selecionados usando *Uninformative Variable Elimination* (UVE) e UVE combinadas com *Competitive Adaptive Reweighted Sampling* (UVE + CARS), resultando na redução de 601 bandas espectrais para apenas 12 (2,8% do total), mantendo as mais correlacionadas com antocianinas. O índice de vegetação ideal de duas bandas (VI2) e o índice de vegetação de três bandas (VI3) foram então calculados. Finalmente, a otimização do *Dung Beetle Optimizer* (DBO), a *Subtraction-Average-Based Optimizer* (SABO) e o *Whale Optimization Algorithm* (WOA) otimizaram o *Extreme Learning Machine* (ELM) para modelagem. O melhor desempenho foi alcançado pelo modelo UVE-CARS-SNV-DBO-ELM.

Na estimativa de antocianinas em folhas da macieira, foram coletadas imagens hiperespectrais e medidas de antocianinas em folhas de macieira em três estágios de crescimento: final da floração, frutificação, e fruto em crescimento. Os espectros originais foram transformados usando segunda derivada para realçar características espectrais. Foram construídos índices espectrais de duas e três bandas para melhorar a correlação com as antocianinas. Três algoritmos de ML foram utilizados: *Support Vector Regression* (SVR), *Random Forest Regression* (RFR) e *CatBoost* (Zhang *et al*., 2025). O modelo *CatBoost* permitiu a visualização eficaz da distribuição de antocianinas em diferentes estágios de crescimento das folhas da macieira, mostrando maior concentração nas veias e extremidades.

BENEFÍCIOS E LIMITAÇÕES DA APLICAÇÃO DE *MACHINE LEARNING*

A aplicação de ML para predição da concentração de antocianinas oferece diversos benefícios. Uma das principais vantagens é a capacidade de analisar grandes volumes de dados de

forma eficiente, permitindo a identificação de padrões complexos e relações não lineares que podem não ser detectados com métodos estatísticos tradicionais (Russell e Norvig, 2021). O ML também melhora a precisão e a capacidade preditiva da concentração das antocianinas sob diferentes fatores, como temperatura, tempo e pH (Ekici *et al*., 2014). Além disso, o ML facilita o desenvolvimento de métodos não destrutivos e de baixo custo para avaliação da concentração de antocianinas. Particularmente quando combinado com técnicas de análise de imagem, o ML mostra-se promissor para a predição não invasiva destes compostos em diversos materiais vegetais, permitindo análises rápidas sem comprometer a integridade das amostras (Askey *et al*., 2019; Liu *et al*., 2024a)

No entanto, existem também algumas limitações associadas ao uso de ML neste contexto. Uma das principais limitações é a necessidade de grandes conjuntos de dados de treinamento de alta qualidade para construir modelos precisos e generalizáveis. A aplicação de ML em sistemas de predição de antocianinas pode enfrentar limitações devido à falta de bases de dados maduras e confiáveis específicas para esse domínio, já que, como visto por Russell e Norvig (2021) a escassez de dados estruturados e anotados impede o treinamento eficaz de modelos.

A qualidade dos dados de entrada é de grande importância para o desempenho dos modelos de ML. Dados ruidosos ou incompletos podem introduzir vieses e levar a previsões imprecisas, pois os modelos dependem fundamentalmente dos dados para generalização, ou seja, utilizar um modelo pronto e utilizar em novas aplicações (Russell e Norvig, 2021). Além da qualidade do pré-processamento ser de extrema importância para o processo de transformação dos dados, este segmento se aprofunda nos principais elementos que aumentam a eficácia do pré-processamento de dados, abrangendo metodologias de seleção de recursos, abordagens de normalização de dados, estratégias de gerenciamento de dados ausentes e técnicas de aumento de dados (Parashar *et al.*, 2023). Outro desafio significativo é o *overfitting* (sobreajuste), na qual modelos complexos se ajustam demais aos dados de treinamento, perdendo a capacidade de generalização para novos dados. Técnicas como validação cruzada e regularização são essenciais para mitigar esse problema (Russell e Norvig, 2021).

Um desafio importante na aplicação de modelos de ML é o fenômeno de *overfitting* (sobreajuste), que pode se manifestar de maneiras sutis durante o desenvolvimento dos modelos. Por exemplo, mesmo quando os modelos parecem ter bom desempenho segundo algumas métricas, podem estar apenas "memorizando" os dados de treinamento em vez de aprender padrões

generalizáveis (Gori, Betti, e Melacci, 2024). Para evitar este problema, é fundamental implementar estratégias de validação adequadas, como utilizar conjuntos de dados independentes para validação durante o treinamento e aplicar técnicas como a "parada antecipada" (*early stopping*), que interrompe o treinamento quando o desempenho no conjunto de validação começa a deteriorar, evitando assim que o modelo se ajuste excessivamente aos dados de treinamento.

Outro desafio importante relaciona-se à forma como os dados são processados durante o treinamento dos modelos. Quando os dados são analisados em pequenos grupos (conhecidos como "lotes" ou "mini-batches") durante o treinamento, podem ocorrer variações que dificultam a identificação do ponto ideal para interromper o treinamento (Gori, Betti, e Melacci, 2024). Além disso, problemas comuns em dados experimentais, como valores ausentes ou distribuição desigual de amostras, podem fazer com que os modelos desenvolvam tendências ou "vícios" que não representam adequadamente o fenômeno real. Isso é particularmente relevante na análise de antocianinas, onde fatores como variações sazonais e diferenças entre cultivares podem influenciar significativamente os resultados, comprometendo a aplicabilidade prática dos modelos desenvolvidos.

TENDÊNCIAS E PERSPECTIVAS DE USO DE *MACHINE LEARNING*

O futuro da aplicação de ML na validação de métodos para a estabilidade e predição de antocianinas apresenta diversas tendências promissoras. A integração do ML com outras técnicas computacionais, como as simulações de dinâmica molecular, poderá proporcionar uma compreensão mais profunda dos mecanismos de estabilidade das antocianinas a nível molecular e a criação de bases de dados específicas para sistemas de antocianinas é crucial para melhorar a precisão e a confiabilidade dos modelos de ML nesta área.

O desenvolvimento de modelos de ML mais interpretáveis é uma área de foco, pois permitirá obter *insights* mais claros sobre os fatores que afetam a estabilidade das antocianinas. A aplicação de ML na otimização de estratégias de estabilização, como o encapsulamento, a copigmentação e a acilação, é outra tendência futura importante. O ML poderá ser utilizado para prever a estabilidade de antocianinas em diferentes matrizes alimentícias e sob diversas condições de processamento, auxiliando no desenvolvimento de produtos mais estáveis (Ekici *et al*., 2014). A exploração de algoritmos de *redes neurais artificias*, capazes de aprender representações complexas a partir de grandes volumes de dados (Gori, Betti, e Melacci, 2024), poderá levar a

avanços significativos na análise da estabilidade de antocianinas a partir de dados espectrais e de imagem.

Na otimização de processos, estudos demonstraram a eficácia do uso do ML em métodos não destrutivos para prever concentração de antocianina (Liu *et al*., 2024b). O desenvolvimento de métodos de monitoração da estabilidade de antocianinas, baseados em ML e técnicas não destrutivas, poderá permitir um controle da concentração mais eficiente e em tempo real.

Finalmente, a utilização de ML para prever a biodisponibilidade e os efeitos na saúde das antocianinas com base na sua estabilidade é uma área de investigação promissora. Aplicações em biociência têm se beneficiado de modelos preditivos que correlacionam estabilidade molecular com biodisponibilidade (Schneider, 2013). Essa linha de pesquisa poderá abrir novas perspectivas para o desenvolvimento, por exemplo, de corantes naturais de antocianinas com propriedades bioativas. Sistemas baseados em *Internet of Things* (IoT) e ML podem possibilitar novas abordagens não destrutivas para análise de compostos bioativos em cadeias produtivas.

CONSIDERAÇÕES FINAIS

As antocianinas são compostos bioativos valiosos com prospecção significativas nas indústrias alimentícia e farmacêutica. No entanto, a sua instabilidade representa um desafio para prever a quantidade de antocianinas para a sua utilização plena. O uso de novas abordagens para auxiliar na otimização de processos como na predição da concentração das antocianinas é, portanto, crucial para garantir a qualidade e a eficácia dos produtos que as contêm. O ML emerge como uma ferramenta poderosa neste contexto, oferecendo abordagens inovadoras para a predição das antocianinas.

Com base nos estudos analisados, é notável que as abordagens não destrutivas baseadas em tecnologias espectroscópicas e algoritmos de inteligência artificial têm se mostrado altamente eficazes para a predição de antocianinas em diferentes matrizes vegetais. A utilização de modelos como ANFIS, SAE-GA-ELM, Random Forest e CatBoost, aliados a técnicas como imagens hiperespectrais (HSI) e digitais (RGB), possibilitou estimativas rápidas, precisas e econômicas, representando um avanço significativo em relação aos métodos tradicionais, que frequentemente etapas laboratoriais demoradas. Além disso, tais abordagens oferecem potencial para aplicações em tempo real e em larga escala, favorecendo o monitoramento da qualidade de produtos agrícolas

e processados ao longo de sua cadeia produtiva.

Apesar das limitações existentes, como a necessidade de grandes conjuntos de dados e a interpretabilidade dos modelos, as tendências futuras apontam para uma integração ainda maior do ML com outras técnicas, desenvolvimento de modelos híbridos, desenvolvimento de modelos mais interpretáveis, engenharia de *features* e a criação de bases de dados específicas. Em suma, o ML tem o potencial de revolucionar a forma como a predição das antocianinas é avaliada e validada, contribuindo para processos mais rápidos, modernos e o desenvolvimento de produtos mais estáveis e eficazes.

AGRADECIMENTOS

Agradecimentos à CAPES e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (409933/2021-0) pelo apoio financeiro, e à Embrapa Clima Temperado pelo apoio financeiro (projeto SEG 10.19.03.040.00.00).

REFERÊNCIAS BIBLIOGRÁFICAS

ABASS, T. *et al*. Concept paper: Innovative approaches to food quality control: AI and machine learning for predictive analysis. **World Journal of Advanced Research and Reviews**, v. 21, n. 3, p. 823-828, 2024. DOI: 10.30574/wjarr.2024.21.3.0719.

AKTHER, S. *et al*. Anthocyanin Stability Profile of Mango Powder: Temperature, pH, Light, Solvent and Sugar Content Effects. **Turkish Journal of Agriculture - Food Science and Technology**, *[S. l.]*, v. 8, n. 9, p. 1871–1877, 2020. DOI: 10.24925/turjaf.v8i9.1871-1877.3487.

ALTAF, Q. S.; KSOURI, R. Application of Machine Learning in the Food Industry. **Artificial Intelligence in the Food Industry: Enhancing Quality and Safety**, p. 23, 2025. DOI: 10.1201/9781032633602-2.

ANJOS, R. *et al*. Effect of agricultural practices, conventional vs organic, on the phytochemical composition of 'Kweli' and 'Tulameen' raspberries (*Rubus idaeus L.*). **Food Chemistry**, v. 328, p. 126833, 30 out. 2020.

ASKEY, B. C. *et al*. A noninvasive, machine learning–based method for monitoring anthocyanin accumulation in plants using digital color imaging. **Applications in Plant Sciences**, v. 7, n. 11, p. e11301, 10 nov. 2019. DOI: 10.1002/aps3.11301.

CHAVES, V. C. *et al*. Berries grown in Brazil: anthocyanin profiles and biological properties. **Journal of the Science of Food and Agriculture**, v. 98, n. 11, p. 4331–4338, 2018.

DAI, H. *et al*. Prediction of Anthocyanin Color Stability against Iron Co-Pigmentation by Surface-Enhanced Raman Spectroscopy. **Foods**, v. 11, n. 21, p. 3436, jan. 2022. DOI: 10.3390/foods11213436.

DIACONEASA, Z., *et al*. Anthocyanins, Vibrant Color Pigments, and Their Role in Skin Cancer Prevention. **Biomedicines**. 2020;8(9):336. Published 2020 Sep 9. DOI:10.3390/biomedicines8090336.

EKICI, L. *et al.* Effects of Temperature, Time, and pH on the Stability of Anthocyanin Extracts: Prediction of Total Anthocyanin Content Using Nonlinear Models. **Food Analytical Methods**, v. 7, n. 6, p. 1328–1336, 1 jul. 2014. DOI: 10.1007/s12161-013-9753-y.

FAKHRI, S. *et al*. The ameliorating effects of anthocyanins on the cross-linked signaling pathways of cancer dysregulated metabolism. **Pharmacological Research**, v. 159, p. 104895, 1 set. 2020. DOI: 10.1016/j.phrs.2020.104895.

FULEKI, T.; FRANCIS. F. J. Extraction and determination of total anthocyanin in Cranberries. **Journal of Food Science, Chicago**, v. 33, n. 1, p. 72–77, 1968. DOI: 10.1111/j.1365-2621.1968.tb00887.x.

GAO, Z. *et al.* Rapid measurement of anthocyanin content in grape and grape juice: Raman spectroscopy provides non-destructive, rapid methods. **Computers and Electronics in Agriculture**, [S.l.], v. 222, p. 109048, 2024. DOI: 10.1016/j.compag.2024.109048.

GARCÍA-CURIEL, L. *et al*. Anthocyanin content prediction in frozen strawberry puree. **Italian Journal of Food Science**, v. 35, n. 2, p. 88–97, 25 maio 2023. DOI: 10.15586/ijfs.v35i2.2315.
GARCIA-OLIVEIRA, P. *et al.* Identification, Quantification, and Method Validation of Anthocyanins. **Chemistry Proceedings**, v. 5, n. 1, p. 43, 2021. DOI: 10.3390/CSAC2021-10680.

GORI, M.; BETTI, A.; MELACCI, S. **Machine Learning: A Constraint-Based Approach**. 2. ed. Cambridge, MA: Morgan Kaufmann, 2024. ISBN 978-0-323-89859-1.

KAUR, S. *et al.* Spotlight on the overlapping routes and partners for anthocyanin transport in plants. **Physiologia Plantarum**. 2021. 171(4), 868-881. DOI:10.1111/ppl.13378.

LI, X. *et al*. Non-destructive prediction and visualization of anthocyanin content in mulberry fruits using hyperspectral imaging. **Frontiers in Plant Science**, v. 14, 27 mar. 2023. DOI: 10.3389/fpls.2023.1137198.

LIN, Y. *et al.* Anthocyanins: Modified New Technologies and Challenges. **Foods**, v. 12, n. 7, p. 1368, jan. 2023. DOI: 10.3390/foods12071368.

LIU, X.-Y.; YU, J.-R.; DENG, H.-N. Non-Destructive Prediction of Anthocyanin Content of Rosa chinensis Petals Using Digital Images and Machine Learning Algorithms. **Horticulturae**, v. 10, n. 5, p. 503, maio 2024a.

LIU, C. *et al.* Prediction of Anthocyanin Content in Purple-Leaf Lettuce Based on Spectral Features and Optimized Extreme Learning Machine Algorithm. **Agronomy**, v. 14, n. 12, p. 2915,

dez. 2024b. DOI: 10.3390/agronomy14122915.

MESQUITA, L. M. de S. *et al*. Fast and green universal method to analyze and quantify anthocyanins in natural products by UPLC-PDA. **Food Chemistry**, v. 428, p. 136814, 2023.

MODESTO JUNIOR, E. N. *et al*. Stability Kinetics of Anthocyanins of Grumixama Berries (Eugenia brasiliensis Lam.) during Thermal and Light Treatments. **Foods**, v. 12, p. 565, 2023. DOI: 10.3390/foods12030565.
MORE, L. S.; KUMAR, B. Understanding and Applying Machine Learning Models. In: **The Pioneering Applications of Generative AI**. IGI Global, 2024. p. 274-309. DOI: 10.4018/979-8-3693-3278-8.ch013.

MUNDE, A. The Machine Learning Pipeline: Algorithms, Applications, and Managerial Implications. In: **Deep Learning Concepts in Operations Research**. Auerbach Publications, 2024. p. 226-243. DOI: 10.1201/9781003433309-18.

PARASHAR, A. *et al*. Data preprocessing and feature selection techniques in gait recognition: A comparative study of machine learning and deep learning approaches. **Pattern Recognition Letters**, v. 172, p. 65-73, 2023. DOI: 10.1016/j.patrec.2023.05.021

PETRELLI, M. Machine Learning Workflow. In: **Machine Learning for Earth Sciences: Using Python to Solve Geological Problems**. Cham: Springer International Publishing, 2023. p. 29-58. DOI: 10.1007/978-3-031-35114-3_3.

POLAT KAYA, H. *et al*. Anthocyanin and bioactivity properties of berberis crategina DC. In buffer system and apple juice: impact of temperature, time, and pH; Prediction using artificial neural network. **Sigma Journal of Engineering and Natural Sciences – Sigma Mühendislik ve Fen Bilimleri Dergisi**, p. 438–449, 2024. DOI: 10.14744/sigma.2024.00028.

PRATAP, C. B. *et al.* Revolutionizing Culinary Quality: An Intelligent Food Grading System Employing Advanced Machine Learning Algorithms. In: **2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)**. IEEE, 2024. p. 1422-1425. DOI: 10.1109/icaccs60874.2024.10717263.

REVELOU, P.-K. *et al.* Applications of Machine Learning in Food Safety and HACCP Monitoring of Animal-Source Foods. **Foods**, v. 14, n. 6, p. 922, jan. 2025. DOI: 10.3390/foods14060922.

RIAL, R. C. IA em química analítica: avanços, desafios e direções futuras. **Talanta**, pág. 125949, 2024. DOI: 10.1016/j.talanta.2024.125949.

RUSSELL, S. J.; NORVIG, Peter. **Artificial intelligence: a modern approach**. 4. ed. Hoboken: Pearson, 2021. ISBN 978-01-34-61099-3.

SCHNEIDER, G. **Prediction of Drug-Like Properties**. *In*: Madame Curie Bioscience Database [Internet]. [s.l.] Landes Bioscience, 2013. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK6404/ Acesso: 10 abr 2025.

SIMEONE, O. A Brief Introduction to Machine Learning for Engineers. **arXiv**, pp 1-231, 17 maio 2018. DOI: 10.48550/arXiv.1709.02840.

SINOPOLI, A.; CALOGERO, G.; BARTOLOTTA, A. Computational aspects of anthocyanidins and anthocyanins: A review. **Food Chemistry**, v. 297, 2019.

SUNARYA, R. R. *et al*. The Effect of pH and Temperature on The Stability of Anthocyanins from Black Soybean Skin Extracts. **Al Kimiya: Jurnal Ilmu Kimia dan Terapan**, v. 11, n. 1, p. 77–83, 1 ago. 2024. DOI: 10.15575/ak.v11i1.35861.

TEIXEIRA, L. N.; STRINGHETA, P. C.; OLIVEIRA, F. A. Comparação de métodos para quantificação de antocianinas. **Revista Ceres**, Viçosa, v. 55, n. 4, p. 297–304. 2008.

WANG, F. *et al*. Detection of Anthocyanins in Potatoes Using Micro-Hyperspectral Images Based on Convolutional Neural Networks. **Foods**, v. 13, n. 13, p. 2096, jan. 2024. DOI: 10.3390/foods13132096.

ZHANG, Y. *et al*. Estimation of Anthocyanins in Apple Leaves Based on Ground Hyperspectral Imaging and Machine Learning Models. **Agronomy**, v. 15, n. 1, p. 140, jan. 2025. DOI: 10.3390/agronomy15010140.