UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE QUÍMICA

Pedro Henrique Schwanck Ramos de Morais

Comparação de Modelos Estatísticos para Calibração na Análise Espectrofotométrica UV-Vis de Corantes: Um Estudo de Precisão e Eficiência

Porto Alegre 2025

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL INSTITUTO DE QUÍMICA

CURSO DE BACHARELADO EM QUÍMICA

Pedro Henrique Schwanck Ramos de Morais

Comparação de Modelos Estatísticos para Calibração na análise espectrofotométrica UV-Vis de corantes: Um Estudo de Precisão e Eficiência

Trabalho de conclusão apresentado junto à atividade de ensino “Trabalho de Conclusão de Curso - QUI” do Curso de Bacharelado em Química, como requisito parcial para a obtenção do grau de Bacharel em Química

Profª. Drª. Elisa Coutinho
Orientadora

Prof. Dr. Marcio Schwaab
Coorientador

Porto Alegre 2025

CIP - Catalogação na Publicação

Morais, Pedro Henrique
Comparação de modelos estatísticos para calibração na análise espectrofotométrica uv-vis de corantes: um estudo de precisão e eficiência / Pedro Henrique Morais. -- 2025.
25 f.
Orientadora: Elisa Coutinho.

Coorientadora: Marcio Schwaab.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Instituto de Química, Bacharelado em Química, Porto Alegre, BR-RS, 2025.

1. Calibração. 2. Corantes. 3. PLS. 4. Redes Neurais. 5. Seleção de Variáveis. I. Coutinho, Elisa, orient. II. Schwaab, Marcio, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

FOLHA DE APROVAÇÃO

Pedro Henrique Schwanck Ramos de Morais

Comparação de Modelos Estatísticos para Calibração na análise espectrofotométrica UV-Vis de corantes: Um Estudo de Precisão e Eficiência

Trabalho de conclusão apresentado junto à atividade de ensino “Trabalho de Conclusão de Curso - QUI” do Curso de Bacharelado em Química, como requisito parcial para a obtenção do grau de Bacharel em Química.

Banca Examinadora

Prof.a Dr.a Elisa Coutinho

Prof. Dr. Márcio Schwab

Dr. Roger Kober

Profª. Drª. Morgana B. Dessuy

Aprovada em: Porto Alegre, 09 de janeiro de 2025.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão a todos que contribuíram para a realização deste trabalho. Agradeço imensamente aos meus professores, cuja dedicação, orientação e conhecimento foram fundamentais para o meu crescimento acadêmico e pessoal. Agradeço também à Universidade Federal do Rio Grande do Sul (UFRGS), por me proporcionar uma formação sólida e por todo o apoio institucional ao longo dessa jornada. Minha gratidão se estende a todos os colegas, amigos, familiares e colegas de laboratório que me incentivaram e estiveram ao meu lado, oferecendo suporte, compreensão e motivação. No entanto, quero fazer um agradecimento especial aos meus pais, que sempre acreditaram em meu potencial, muitas vezes enxergando algo em mim que eu mesmo não era capaz de perceber. Sem o apoio e o amor incondicional deles, este trabalho não teria sido possível. A todos, o meu mais sincero agradecimento.

"*A cor é uma sombra de luz, e a luz é a substância que torna a cor visível.*"

- Leonardo Da Vinci

RESUMO

Os corantes são amplamente empregados na indústria, estando presente durante os processos produtivos bem como nos resíduos industriais, exigindo uma rigorosa quantificação de suas concentrações tanto para garantir a segurança alimentar como para cumprir a legislação ambiental. Em misturas reais, muitas vezes são empregados mais de um corante simultaneamente, e sua quantificação deve ser realizada pela técnica de análise multivariada. Nesse contexto, empregou-se esse método quimiométrico para avaliação de dados de amostras analisadas por espectrofotometria UV-Vis com o objetivo de determinar com precisão e exatidão a concentração de misturas com múltiplos corantes, mesmo em sistemas complexos. Este trabalho propõe o emprego de técnicas de seleção de variáveis, como a seleção baseada em variáveis ortogonalizadas por matriz de correlação e a seleção por diminuição do resíduo espectral teórico. Além disso, para comparar diferentes métodos de calibração para a determinação da concentração em amostras com múltiplos corantes foram empregados: regressão por mínimos quadrados parciais (PLS) e redes neurais (RN). A seleção de variáveis usando o critério de maior ortogonalidade da matriz de correlação mostrou uma melhoria significativa na regressão por PLS, embora o impacto da seleção de variáveis nas metodologias de regressão por redes neurais tenha sido menos expressivo. Entretanto, as técnicas de regressão por PLS e por redes neurais demonstraram melhores resultados gerais para a calibração da espectrofotometria por UV-vis para misturas com múltiplos corantes.

Palavras-chave:

Calibração; Corantes; UV-Vis; PLS; Redes Neurais; Seleção de variáveis

ABSTRACT

Dyes are widely employed in the industry, being present both during production processes and in industrial waste, requiring rigorous quantification of their concentrations to ensure food safety and comply with environmental legislation. In real mixtures, more than one dye is often used simultaneously, and their quantification must be performed using multivariate analysis techniques. In this context, this chemometric method was applied to evaluate sample data analyzed by UV-Vis spectrophotometry to accurately and precisely determine the concentration of mixtures containing multiple dyes, even in complex systems. This work proposes the use of variable selection techniques, such as selection based on orthogonalized variables through a correlation matrix and selection by reducing the theoretical spectral residue. Additionally, to compare different calibration methods for determining concentrations in samples with multiple dyes, partial least squares regression (PLS) and neural networks (NN) were employed. Variable selection using the criterion of greater orthogonality of the correlation matrix showed significant improvement in PLS regression, although the impact of variable selection on neural network regression methodologies was less pronounced. Nevertheless, PLS and neural network regression techniques demonstrated overall better results for UV-Vis spectrophotometric calibration in mixtures with multiple dyes.

Keywords:

Calibration; Dyes; UV-Vis; PLS; Neural Networks; Variable Selection

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PLS	Regressão por mínimos quadrados parciais
UV-VIs	Análise espectrofotométrica na faixa de ultravioleta-vísivel
RN	Redes Neurais
SNV	Standart normal variate
MSC	Multi Scatter Correction
MSE	Erro quadrático médio
Abs	Absorção

SUMÁRIO

1. INTRODUÇÃO

A quantificação precisa de corantes é um desafio crucial na indústria, uma vez que esses compostos não apenas influenciam a aparência dos produtos, mas também impactam diretamente na percepção e na segurança dos consumidores, em especial na indústria de alimentos. Dada sua relevância, é fundamental o desenvolvimento de métodos analíticos robustos que permitam identificar e quantificar corantes em misturas multicomponentes de maneira eficaz tanto durante os processos quanto nos resíduos industriais.

A espectrofotometria UV-Vis é uma técnica amplamente utilizada nesse contexto, devido à sua versatilidade, baixo custo e rapidez. Contudo, a sobreposição e as interações de espectros de absorção dos diferentes componentes em misturas de múltiplos corantes apresentam desafios significativos para a análise e a quantificação tradicional desta técnica. Métodos de calibração convencionais, como o uso de comprimentos de onda de absorção máxima, frequentemente são insuficientes para permitir a precisa quantificação dos componentes, especialmente em misturas complexas.

Nesse cenário, a aplicação de técnicas quimiométricas, como a Regressão por Mínimos Quadrados Parciais (PLS) e redes neurais surge como uma alternativa poderosa na análise de misturas reais. Essas ferramentas não apenas melhoram a capacidade preditiva em cenários com forte colinearidade entre variáveis espectrais, mas também permitem a integração de estratégias de seleção de variáveis para otimizar a modelagem da calibração. O desenvolvimento de métodos analíticos portáteis, baseados nesses avanços, pode ainda ampliar o impacto prático dessas soluções, tornando-as acessíveis para controle de qualidade e pesquisa aplicada.

Este trabalho, portanto, busca explorar e comparar diferentes estratégias de calibração, integrando técnicas estatísticas e avanços em quimiometria, com o objetivo de propor soluções inovadoras para a quantificação de corantes em misturas multicomponentes. A partir desse enfoque, espera-se contribuir para o avanço na análise espectrofotométrica de corantes, oferecendo alternativas viáveis e precisas para aplicações práticas.

2. OBJETIVOS

Objetivo Geral

Desenvolver e avaliar métodos de calibração para a quantificação de corantes em misturas multicomponentes utilizando espectrofotometria UV-Vis, visando precisão e exatidão no contexto de análises químicas.

Objetivos Específicos

- Estudar os princípios e as limitações dos métodos de calibração aplicados à quantificação de corantes.
- Comparar diferentes modelos estatísticos de calibração, incluindo PLS, redes neurais e sistemas de soluções lineares, quanto à sua eficácia e aplicabilidade.
- Investigar técnicas de seleção de variáveis para melhorar a robustez dos modelos de calibração.

3. REVISÃO BIBLIOGRÁFICA

3.1. CORANTES

Corantes são substâncias que possuem a capacidade de conferir cor a materiais como tecidos, papéis, plásticos, entre outros, devido à presença de grupos cromóforos (estruturas químicas que absorvem luz em certos comprimentos de onda). Diferente dos pigmentos[1], que são partículas insolúveis, os corantes são solúveis em seu meio de aplicação, permitindo que se difundam e se fixem quimicamente na superfície dos materiais. Eles são amplamente utilizados não apenas na indústria têxtil, mas também em alimentos, cosméticos, medicamentos e em processos de coloração industrial, sendo essenciais para criar variações estéticas e funcionais nos produtos.

A história dos corantes é milenar, remontando a civilizações antigas, como as da Índia, Egito e Mesopotâmia, que utilizavam corantes naturais para tingir tecidos e produzir tintas. Entre os primeiros corantes conhecidos estão a púrpura de Tiro, extraída de moluscos marinhos, e o índigo, obtido de plantas do gênero *Indigofera*. No entanto, o grande marco na história dos corantes ocorreu em 1856, quando o químico britânico William Henry Perkin[2] acidentalmente descobriu o primeiro corante sintético, a malvaína (também conhecida como anilina púrpura), ao tentar sintetizar um medicamento antimalárico. Essa descoberta revolucionou a indústria têxtil, pois os corantes sintéticos eram mais vibrantes, estáveis e fáceis de produzir em larga escala, em comparação com os corantes naturais.

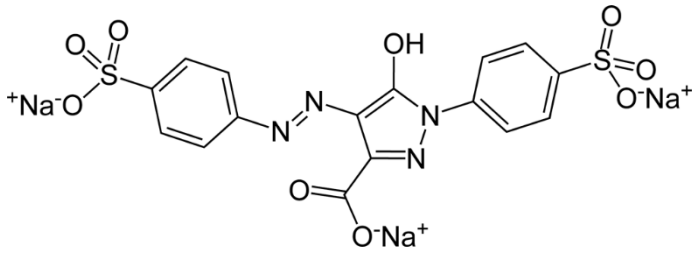
Os corantes podem ser classificados em diferentes categorias de acordo com sua origem e aplicação. Corantes naturais[3] são obtidos de fontes vegetais, animais ou minerais, como o açafrão, a clorofila, a carmim (extraída de insetos) e o índigo; embora sejam mais ecológicos, apresentam menor estabilidade à luz e têm um custo mais elevado devido à complexidade de sua extração. Por outro lado, os corantes sintéticos[4] dominam o mercado atual graças à sua maior estabilidade, ampla gama de cores e menor custo de produção, sendo subdivididos em tipos como ácidos, básicos, reativos, dispersos, diretos e de mordente, cada um destinado a aplicações específicas. Já os corantes alimentícios[5] utilizados para adicionar ou intensificar a cor de alimentos e bebidas, com regulamentações rigorosas para garantir sua segurança; podem ser tanto naturais quanto sintéticos, embora haja uma preferência crescente

por opções naturais devido às tendências de consumo mais saudáveis[6]. Por fim, há os corantes funcionais[7], que, além de colorir, desempenham papéis adicionais, como proteção contra radiação UV, atuação como sensores de pH ou uso como indicadores em processos laboratoriais, mostrando que a função dos corantes vai além da simples coloração, integrando-se a diversas aplicações tecnológicas e científicas.

Com o avanço da química orgânica no final do século XIX e início do século XX, a produção de corantes sintéticos tornou-se uma das bases da indústria química moderna.[8] A partir da malvaína, uma série de corantes derivados da anilina e outros compostos aromáticos foi desenvolvida, como o vermelho Congo e a fenolftaleína, além de outros que ganharam grande destaque, como a Tartrazina, o Amaranto e o Amarelo Crepúsculo, que serão os utilizados nesse estudo.

A tartrazina, cuja estrutura molecular é representada na Figura 1, é conhecida como E102, um corante sintético pertencente ao grupo funcional dos azo-compostos (compostos orgânicos que apresentam nitrogênio em sua estrutura química). Este tem propriedade de proporcionar a cor amarelo-limão se utilizada como corante alimentar, por exemplo. É derivada do creosoto mineral, e possui solubilidade na água e absorção máxima em solução aquosa em 427 ± 2 nm. Seu uso mais frequente se dá em condimentos, como também em cosméticos e medicamentos.[9]

Figura 1 – Estrutura Molecular do corante Tartrazina

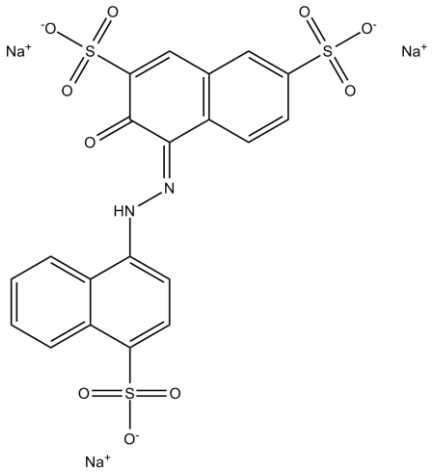


Fonte: Arquivo pessoal do autor. Figura produzida em Chemsketch.

Amaranto, também conhecido como Vermelho FD&C No. 2, E123, C.I. Vermelho para Alimentos 9 ou Vermelho Ácido 27, é também um corante azo de tonalidade que varia de vermelho escuro a púrpura, sendo popularmente chamado no Brasil de Vermelho Amaranto. Inicialmente utilizado como corante alimentício e em cosméticos, seu uso foi banido nos Estados Unidos em 1976 pela *Food and Drug Administration* (FDA) devido à suspeita de ser carcinogênico. No mercado, é normalmente encontrado na forma de um sal trissódico, apresentando-se como um pó

de coloração vermelho escuro a castanho-púrpura, solúvel em água, com decomposição a 120 °C, sem fundir-se. Sua solução aquosa exibe absorção máxima em torno de 520 nm, sendo classificado como um corante aniônico. Além de sua aplicação como aditivo alimentar (código E123), o Amaranto também é utilizado para tingir fibras naturais e sintéticas, couro, papel e resinas fenol-formaldeído.[10]

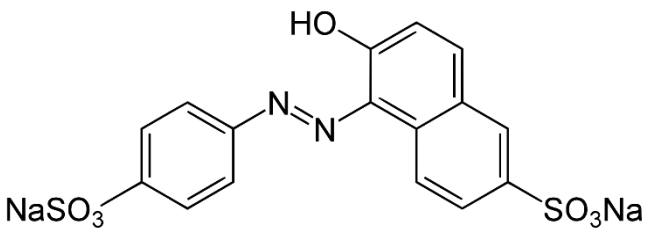
Figura 2 – Estrutura Molecular do corante Amaranto



Fonte: Arquivo pessoal do autor. Figura produzida em Chemsketch

O Amarelo Crepúsculo, frequentemente identificado pela sigla FCF (também conhecido como Amarelo Alaranjado S, Amarelo FD&C 6 ou C.I. 15985), é outro corante azo sintético de coloração amarela. Sua absorção máxima ocorre em comprimentos de onda entre 480 e 500 nm, variando conforme o pH. Esse corante é produzido a partir de hidrocarbonetos aromáticos derivados do petróleo e, quando utilizado em alimentos na Europa, é designado pelo número E110, ou INS 110 em outros sistemas de classificação. Apesar de haver relatos associando o Amarelo Crepúsculo a reações alérgicas em algumas pessoas, estudos científicos até o momento não confirmaram de forma conclusiva esses efeitos adversos.[11]

Figura 3 – Estrutura Molecular do corante Amarelo Crepúsculo



Fonte: Arquivo pessoal do autor. Figura produzida em Chemsketch.

3.2. QUIMIOMETRIA

Quimiometria é uma disciplina que combina métodos estatísticos e matemáticos para otimizar e interpretar dados de análises físico-químicas. Seu principal objetivo é extrair informações significativas a partir de conjuntos de dados complexos, auxiliando no controle de qualidade, monitoramento de processos e desenvolvimento de métodos analíticos.[12]

Além de sua ampla aplicação em diferentes técnicas analíticas, a quimiometria é particularmente útil em técnicas espectrofotométricas, como espectroscopia UV-Vis, infravermelho (IV) e espectrometria de massa. Nessas técnicas, grandes volumes de dados são gerados, como espectros com centenas de variáveis (comprimentos de onda), o que torna o uso de ferramentas quimiométricas essencial para a interpretação correta dos resultados. Métodos como a regressão por mínimos quadrados parciais (do inglês, *Partial Least Squares* – PLS) permitem correlacionar dados espectrais com concentrações de analitos, facilitando a quantificação precisa em misturas complexas. Além disso, a análise de componentes principais (do inglês, *Principal Component Analysis* – PCA) ajuda a identificar padrões e eliminar redundâncias, simplificando a interpretação de dados espectrais multivariados.[13]

3.2.1. Métodos de pré-tratamento de dados

O pré-tratamento de dados é uma etapa crucial em qualquer análise de dados, e sua importância reside no fato de que dados brutos muitas vezes contêm ruído, valores ausentes, outliers e variações não relevantes que podem prejudicar a qualidade dos modelos empregados. O objetivo do pré-tratamento é preparar os dados para que se tornem mais adequados para a análise subsequente, garantindo que as conclusões obtidas sejam mais confiáveis e precisas.[14]

3.2.2. Standard Normal Variate

A SNV (*Standard Normal Variate*) é uma técnica de pré-tratamento de dados muito usada em quimiometria e análise espectral para corrigir variações de espalhamento de luz em dados espectroscópicos.[15] A SNV é especialmente útil em espectros que possuem grandes variações de intensidade ou ruído, principalmente

quando esses problemas resultam de efeitos físicos, como o espalhamento de partículas em amostras heterogêneas.

Esta técnica consiste essencialmente de três passos: calcular a média do espectro, calcular o desvio padrão do espectro, e aplicar a normalização para cada ponto utilizando os dois valores citados anteriormente, conforme a Equação (1), onde x_i' é o valor normalizado da variável no ponto, x_i é o valor da variável medida no ponto, σ_x é a média da variável no espectro obtido e ρ é o desvio padrão neste espectro.[16]

$$x_i' = \frac{x_i - \sigma_x}{\rho} \quad (1)$$

3.2.3. Multiplicative Scatter Correction

O MSC (*Multiplicative Scatter Correction*) é uma técnica de pré-processamento de dados muito utilizada em espectroscopia, especialmente em espectros obtidos por infravermelho próximo (NIR) e de reflectância. Seu principal objetivo é corrigir o efeito de espalhamento de luz (*scattering*) que ocorre devido à interação desigual entre a luz e as partículas presentes na amostra.[17]

O ajuste do MSC consiste em definir um espectro médio de todas as amostras ou um espectro de uma amostra considerada idealmente representativa. A partir desse espectro de referência, é feita uma regressão linear entre este e o espectro da amostra desconhecida, conforme a Equação (2), onde x_i' é o valor medido no ponto no espectro original da amostra desconhecida, a é o coeficiente aditivo, b é o coeficiente multiplicativo, x_{ref} é o valor obtido no espectro de referência e $\in$: é o erro residual.[18]

$$x_i' = a + b * x_{ref} + \in \quad (2)$$

3.2.4. Modelos estatísticos de regressão multifatorial

Modelos estatísticos de regressão multifatorial são ferramentas utilizadas para analisar a relação entre uma variável dependente (ou resposta) e múltiplas variáveis independentes (ou preditoras). Ao contrário da regressão simples, que explora a influência de apenas uma variável preditora sobre a resposta, os modelos de

regressão multifatorial permitem avaliar como várias variáveis, simultaneamente, afetam a variável dependente. Esses modelos partem do princípio de que a variável dependente pode ser explicada por uma combinação linear de várias variáveis independentes.[19]

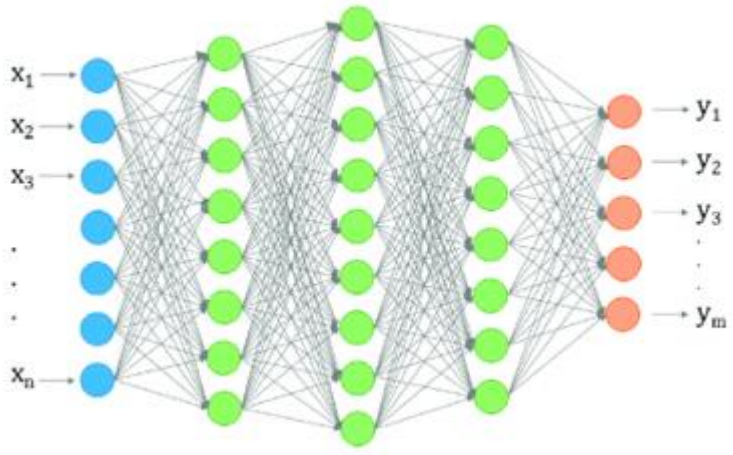
3.2.5. Método de mínimos quadrados parciais

O método de mínimos quadrados parciais (PLS) é uma técnica utilizada para criar modelos preditivos quando há um grande número de variáveis independentes. Em vez de focar diretamente nas variáveis originais, o PLS busca extrair componentes novos que melhor representam as informações relevantes para prever a variável dependente. Esses componentes são combinações lineares das variáveis iniciais, otimizados para capturar a máxima variabilidade explicativa. [20]

3.2.6. Regressão por rede neural

Uma rede neural funciona como uma estrutura matemática que visa prever um valor contínuo (da variável dependente) a partir de várias entradas (variáveis independentes).[21] Essas entradas representam as características ou variáveis do problema e são alimentadas na camada de entrada da rede neural. Cada uma dessas variáveis é atribuída a um neurônio nessa camada, e, em seguida, a rede processa essas informações por meio de conexões ponderadas.

Figura 4 – Estrutura de uma rede neural



Fonte: Arquivo pessoal do autor.

Nas redes neurais simples para regressão linear, a relação entre as variáveis de entrada e a saída pode ser direta, sem a necessidade de camadas ocultas. O modelo realiza operações lineares, onde as variáveis de entrada são multiplicadas por pesos, e os resultados dessas multiplicações são somados para gerar a saída. No entanto, caso existam camadas ocultas, estas introduzem um nível adicional de complexidade ao aprendizado, permitindo à rede capturar relações mais complexas entre as variáveis, obtendo-se assim uma análise mais robusta.[22]

Durante o treinamento da rede, os pesos associados a cada entrada são ajustados com base no erro observado entre a predição feita pela rede e o valor real. Esse processo utiliza algoritmos como a retropropagação e técnicas de otimização, como o gradiente descendente, para minimizar a função de custo (geralmente o erro quadrático médio, MSE). Dessa forma, a rede neural "aprende" a ajustar os pesos para melhor aproximar a relação entre as variáveis de entrada e a variável de saída, tornando-se uma versão mais flexível e ajustável de um modelo de regressão linear multivariada.

3.3. PARÂMETROS DE VALIDAÇÃO

A utilização de parâmetros de validação é fundamental em projetos de modelagem preditiva e de aprendizado de máquina, garantindo que ele não apenas se ajuste bem aos dados de treinamento, mas também generalize suas previsões para novos dados. Isso é crucial para evitar o *overfitting*, que ocorre quando o modelo aprende padrões específicos dos dados de treinamento que não se aplicam a novas informações. Além disso, os parâmetros de validação oferecem uma maneira objetiva de comparar diferentes modelos, permitindo que os desenvolvedores identifiquem qual abordagem apresenta o melhor desempenho e justifiquem suas escolhas. A validação também possibilita o aprimoramento contínuo dos modelos, já que as métricas observadas podem guiar ajustes, como a seleção de algoritmos, o ajuste de hiperparâmetros ou a modificação da engenharia de recursos. Assim, a análise das métricas de validação permite uma tomada de decisão informada, identificação de problemas nos dados e aumento da eficiência. Em suma, a utilização de parâmetros de validação não é apenas uma prática recomendada, mas uma necessidade,

garantindo que os modelos sejam eficazes, robustos e prontos para uso em situações reais, resultando em melhores resultados e maior confiança nas previsões feitas.

3.3.1. Coeficiente de determinação

O coeficiente de determinação, ou R^2, é uma métrica amplamente utilizada para avaliar a qualidade de ajuste de modelos de regressão, especialmente em análises estatísticas e de aprendizado de máquina. Ele mede a proporção da variabilidade na variável dependente que é explicada pelas variáveis independentes do modelo, e seus valores vão de zero a um. O valor de R^2 ser zero indica que o modelo não explica nenhuma variabilidade, enquanto se este coeficiente apresenta o valor de um significa que toda a variabilidade dos dados é explicada. Por exemplo, um R^2 de 0,80 implica que 80% da variabilidade nos dados pode ser atribuída ao modelo. O R^2 é matematicamente definido como demonstrado na Equação (7), onde SS_{res} é a soma dos quadrados dos resíduos e SS_{tot} é a soma total dos quadrados.[23]

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{7}$$

Embora o R^2 seja uma métrica útil para avaliar o ajuste do modelo, sua interpretação deve ser cautelosa, pois ele não considera a complexidade do modelo e pode ser enganoso em casos de *overfitting*. Além disso, o R^2 não é apropriado para modelos não lineares, o que torna essencial a utilização de outras métricas, como o MSE (Erro Quadrático Médio) e o MAE (Erro Absoluto Médio) para uma avaliação mais completa da qualidade do modelo.

3.3.2. Erro quadrático médio

O Erro Quadrático Médio, ou MSE (*Mean Squared Error*), é uma métrica amplamente utilizada para avaliar a performance de modelos de regressão, especialmente na medição da precisão de previsões. Como demonstrado na Equação (8), o MSE é calculado como a média das diferenças quadráticas entre os valores observados (y_i) e os valores previstos pelo modelo (y_i'), proporcionando uma medida direta da quantidade de erro em um modelo, dependendo do número de amostras, n .[24]

$$MSE = \frac{1}{n}\sum(y_i - y_i')^2 \quad (8)$$

O MSE é sempre um valor não negativo, sendo que um MSE igual a zero indica um modelo perfeito, que não apresenta erro em suas previsões. Apesar de sua utilidade, o MSE pode ser sensível a *outliers*, pois erros maiores são elevados ao quadrado, o que pode distorcer a interpretação do desempenho do modelo. Portanto, em casos onde a presença de *outliers* é uma preocupação, outras métricas, como o Erro Absoluto Médio (do inglês, *Mean Absolute Error* – MAE), podem ser consideradas para fornecer uma avaliação mais robusta da qualidade do modelo.

3.3.3. Erro absoluto médio

O Erro Absoluto Médio, ou MAE (*Mean Absolute Error*), é uma métrica utilizada para avaliar a precisão de modelos de regressão, medindo a média das diferenças absolutas entre os valores observados (y_i) e os valores preditos pelo modelo (y_i'), dependendo do número de amostras (n). O MAE, Equação (9), fornece uma representação direta do erro em unidades da variável de interesse, facilitando a interpretação dos resultados.[25]

$$MAE = \frac{1}{n}\sum | y_i - y_i' | \quad (9)$$

Diferentemente do MSE, o MAE não eleva os erros ao quadrado, o que o torna menos sensível a *outliers*, oferecendo uma visão mais robusta da precisão do modelo em situações onde dados extremos podem distorcer a avaliação. O MAE é sempre um valor não negativo, e um MAE igual a zero indica um modelo perfeito que não apresenta erro em suas previsões. Embora o MAE forneça informações valiosas sobre a performance do modelo, ele não penaliza erros maiores de forma tão intensa quanto o MSE, o que pode ser uma desvantagem em contextos onde é importante considerar a gravidade dos erros. Assim, o MAE é uma ferramenta útil na avaliação de modelos de regressão, especialmente quando se busca uma medida clara e interpretável do erro, mas deve ser utilizado em conjunto com outras métricas para uma avaliação completa do desempenho do modelo.

3.4. PROBLEMAS ENVOLVENDO MISTURAS DE CORANTES

3.4.1. Misturas com dois corantes

Quando dois corantes são misturados, suas interações podem alterar propriedades da mistura como cor, solubilidade e estabilidade. Essas interações podem ser complexas e influenciar a forma como os corantes se comportam em aplicações práticas, especialmente quando suas absorções ópticas se sobrepõem. A combinação de espectros pode resultar em novas cores que não são uma simples soma das cores individuais, além disso, o comprimento de onda de máxima absorção (λ_{max}) pode ser deslocado devido a interações moleculares entre os compostos, afetando a cor percebida. Corantes aromáticos com sistemas conjugados também podem interagir através de empilhamento π-π, formando agregados que alteram a absorção e a intensidade da cor. Em outros casos, corantes podem formar um complexo entre si via interações intermoleculares, como ligações de hidrogênio ou dipolo-dipolo, modificando suas propriedades eletrônicas. Essas interações podem levar à formação de estruturas estáveis que afetam a solubilidade e a cor da mistura. Por fim, a transferência de energia entre corantes pode ocorrer em misturas, onde um corante, após absorver luz, transfere essa energia a outro corante, que então emite luz em um comprimento de onda diferente. Este fenômeno, conhecido como transferência de Förster[26], é a base para sensores fluorescentes e dispositivos ópticos, onde é importante aproveitar a sobreposição de espectros de absorção e emissão para otimizar a emissão de luz.

Para a caracterização dessas misturas de dois corantes, normalmente é feita a calibração selecionando os comprimentos de onda de máxima absorção de cada corante e realizando uma calibração externa com apenas um corante por vez, ou seja, criando pontos de absorção em função da concentração do corante sozinho e utilizando a equação da reta destes pontos para a determinação da composição da solução da mistura original. Contudo, levando em contas os efeitos supracitados, podem existir problemas em determinar o comprimento de onda de máxima absorção e cada corante na mistura, e para esses casos as técnicas derivativas se mostram eficazes. Também pode ser utilizada técnicas como a Regressão por Mínimos Quadrados Parciais, *Partial Least Squares* - PLS,[27] ou a técnica derivativa, que envolve realizar a derivada do espectro, o que permite uma melhor separação dos

picos e até mesmo encurtar suas zonas de influência, melhorando assim o desempenho da calibração externa.[28]

A manipulação matemática do espectro se mostra uma ótima alternativa a lidar com algum grau de sobreposição, porém, se limita a desconsiderar outros efeitos que distorcem a linearidade das absorções, o que a torna eficaz somente em casos de dois corantes com baixo grau de sobreposição.

3.4.2. Misturas com três ou mais corantes

As interações químicas que causam a não-linearidade nos picos de absorção geralmente se tornam mais pronunciadas em misturas de três ou mais corantes em comparação com misturas binárias. Isso ocorre porque com mais corantes presentes, há um número maior de possíveis interações intermoleculares, como empilhamento π–π, formação de complexos e de interações dipolo-dipolo. Esses fenômenos podem alterar o estado eletrônico dos corantes, fazendo com que os espectros de absorção desviem da linearidade esperada, especialmente quando os corantes começam a se associar entre si de maneiras imprevisíveis.[29]

Além disso, a sobreposição dos espectros de absorção se torna mais significativa em misturas mais complexas, dificultando a separação clara dos picos individuais. À medida que mais corantes são adicionados, os espectros podem se sobrepor e interagir, resultando em mudanças no comprimento de onda máximo e em alterações nas intensidades de absorção. Essas mudanças não são apenas uma soma das contribuições individuais, mas refletem também o impacto das interações moleculares entre os corantes, levando a respostas não lineares. [30]

A influência do ambiente, como o pH e a polaridade do solvente, também se torna mais importante em misturas com múltiplos corantes. Cada corante pode reagir de forma diferente a mudanças no microambiente, alterando suas propriedades ópticas de maneiras que não são facilmente previsíveis. Esse comportamento é ainda mais complicado quando ocorrem fenômenos de transferência de energia entre moléculas, em que um corante pode excitar outro, ou quando há *quenching*, onde um corante diminui a fluorescência ou absorção do outro.[31]

Nessas situações, é necessário recorrer a técnicas de calibração multivariada e métodos mais avançados para lidar com a complexidade dos dados[32]. A Regressão por Mínimos Quadrados Parciais (PLS) é amplamente utilizada devido à sua

capacidade de lidar com sobreposição espectral e colinearidade, correlacionando diretamente os dados espectrais com a concentração. A Regressão por Componentes Principais (do inglês, *Principal Component Regression* – PCR) também é usada, combinando a redução de dimensionalidade com regressão, embora seja menos eficiente que o PLS em capturar interações. O MCR-ALS (*Multivariate Curve Resolution – Alternating Least Squares*) é uma abordagem iterativa que resolve perfis de concentração e espectros de componentes em misturas complexas sem a necessidade de padrões puros, sendo útil para resolver picos sobrepostos.

Por fim, técnicas de *Machine Learning*, como Redes Neurais Artificiais e *Random Forests*, podem ser aplicadas quando há interações não lineares mais complexas e em maior quantidade, aproveitando grandes volumes de dados para capturar padrões que métodos lineares não conseguem modelar.

4. METODOLOGIA

4.1. PREPARO DAS SOLUÇÕES

Foram selecionados três corantes para a análise. A partir deles, foram preparadas três soluções estoque, sendo a solução A de corante Tartrazina (72 mg/L), a solução B de corante Amaranto (72 mg L^{-1}) e a solução C de corante *Sunset Yellow* (72 mg L^{-1}). Após o preparo das soluções, foram adquiridos os espectros de absorção para cada solução-estoque.

O experimento foi realizado com um espectrofotômetro de absorção molecular na região do ultravioleta e visível (UV-Vis) modelo T80, equipado com sistema óptico de feixe duplo. O equipamento possui capacidade para realizar leituras na faixa de 190 a 1100 nm, embora, para este estudo, tenha sido utilizada a faixa de 390 a 700 nm. Ele acomoda até oito cubetas, sendo utilizadas duas quartzo para o experimento.

Após a obtenção dos espectros, foram preparadas outras soluções a partir das soluções estoque A, B e C. Essas novas soluções filhas (1 a 27) tiveram suas concentrações entre 0 mg L^{-1} e 6 mg L^{-1} para cada corante individualmente e de 0 a 18 mg L^{-1} para a soma dos 3 corantes nas misturas, conforme a **Tabela 1**.

Tabela 1: Concentrações dos corantes A (Tartazina), B (Amaranto) e C (*Sunset Yellow*) , em mg/L, nas soluções preparadas para o estudo.

Solução	[A]	[B]	[C]
1	0	0	0
2	0	0	3
3	0	0	6
4	0	3	0
5	0	3	3
6	0	3	6
7	0	6	0
8	0	6	3
9	0	6	6
10	3	0	0
11	3	0	3
12	3	0	6
13	3	3	0
14	3	3	3
15	3	3	6
16	3	6	0
17	3	6	3
18	3	6	6
19	6	0	0
20	6	0	3
21	6	0	6
22	6	3	0
23	6	3	3
24	6	3	6
25	6	6	0
26	6	6	3
27	6	6	6

4.2. SELEÇÃO DE VARIÁVEIS

A seleção de variáveis é uma etapa crucial em muitos métodos de análise multivariada, especialmente quando se lida com grandes conjuntos de dados, como aqueles encontrados em processos de espectrofotometria. A partir dos dados de perfis espectrais dos corantes foram desenvolvidos dois métodos de seleção de variáveis.

4.2.1. Seleção de variáveis por matriz linear

Inicialmente é utilizado o método *dataframe.corr()* em Python para criar a matriz de correlação entre os diferentes comprimentos de onda. Foram então selecionados os comprimentos de onda que apresentaram um grau de correlação menor que um valor limite, chamado de “limite colinear”. Esse valor inicialmente foi configurado como sendo 0,6.

Foi desenvolvido um algoritmo que seleciona o primeiro comprimento de onda, chamado de V_1, e então lê a correlação desse com os comprimentos de onda adjacentes até que um desses estivesse abaixo do limite colinear, sendo esse novo comprimento de onda chamado de V_2. O algoritmo então seleciona o V_2 como a nova variável de referência, ou seja, ele irá ler as variáveis adjacentes até achar uma que possua uma correlação abaixo do limite colinear em relação a variável de referência V_2. Essas varreduras serão feitas até a última variável de referência V_f não possuir mais outras variáveis que estejam abaixo do limite colinear dentro do conjunto de dados.

Após isso, as variáveis selecionadas foram organizadas em uma matriz, que apresenta em cada coluna um comprimento de onda previamente selecionado (V_n) e em cada linha um comprimento de onda diferente do espectro. Em seguida, foram realizados os métodos de regressão para essa matriz utilizando os valores de X sendo os comprimentos de onda selecionados e comparado com o resultado obtido ao utilizar os comprimentos de onda sem essa seleção inicial.

4.2.2. Seleção de variáveis por resíduos minimizados (SVRMi)

Outro método para seleção de variáveis empregado foi o algoritmo desenvolvido de seleção de variáveis por resíduos minimizados (SVRMi). Esse método apresenta duas etapas de seleção, onde a primeira etapa consiste no desenvolvimento de um espectro teórico, que seria criado pela soma linear de outros n-espectros diferentes de corantes que possuam a mesma concentração.

Uma vez tendo obtido os espectros experimentais de cada um dos corantes individualmente e da mistura dos corantes, o algoritmo então itera para cada comprimento de onda, subtraindo o valor da absorção real da mistura pelo valor da absorvância teórica, e elevando o resultado ao quadrado, Equação (10).

$$X\,nm = (abs_{real} - abs_{teor})^2 \qquad (10)$$

Após isso, os comprimentos de onda que apresentarem um valor de X_{nm} inferior a um valor crítico serão selecionados. Essa seleção inicial permite acessar os comprimentos de onda que satisfazem o teorema de independência da lei de Lambert-Beer, sendo assim selecionadas as variáveis cuja absorvância naquele determinado comprimento de onda seja uma combinação linear da absorvância de todas as entidades na solução.

A linearidade em um comprimento de onda pode não se manter em diferentes absorvâncias, ou seja, por mais que o método acima seleciona um comprimento de onda cuja absorvância é a soma linear exata das absorvância dos corantes, pode ser que em outras concentrações essa linearidade se perca. Para isso, é realizada uma segunda etapa que consiste em verificar se o coeficiente de absorção molar permanecerá o mesmo em diferentes concentrações desses mesmos corantes. Isso é feito através da determinação dos coeficientes de absortividade molar de cada espécie tanto na solução com 9 mg L^{-1} quanto para 18 mg L^{-1}, e selecionar as variáveis cujos resíduos sejam minimizados, ou seja, que haja pouca variação entre as absortividades molares daquele comprimento de onda naquela faixa de absorvância.

Estas seleções permitem assegurar um conjunto de variáveis (ou números de onda) em que o coeficiente de absortividade molar permaneça constante independentemente do meio reacional e da concentração do corante de interesse.

4.3. DESENVOLVIMENTO DO MODELO

4.3.1. Regressão linear por mínimos quadrados parciais (PLS)

O método de mínimos quadrados parciais é uma técnica estatística poderosa para modelar a relação entre as variáveis, especialmente quando há um grande número de variáveis. No caso em questão, as variáveis independentes (absorvâncias em cada comprimento de onda) foram utilizadas para prever as variáveis dependentes (concentração de cada corante).

Foi separado o conjunto de dados (com as variáveis dependentes e independentes) em treino e teste, sendo o conjunto de treino utilizado para treinar o modelo utilizando o método *PLSRegression()* da biblioteca *sklearn*, e as métricas foram obtidas ajustando o modelo obtido do conjunto de treino ao conjunto de dados de teste.

4.3.2. Regressão linear por redes neurais

O conjunto de dados com as variáveis independentes (absorvâncias em seus respectivos comprimentos de onda) e as variáveis dependentes (concentração de cada corante) foi separado em treino, para calibrar o modelo, e teste, para validar a robustez e exatidão do mesmo. Para a regressão foi utilizado o método *MLPRegressor()* da biblioteca *sklearn*, que é um modelo de rede neural artificial que usa retropropagação para ajustar os pesos entre os neurônios, a fim de melhorar a precisão da previsão. O primeiro parâmetro calibrado foi o número de camadas ocultas (*hidden layers*), que crescem linearmente em tamanho à medida que a complexidade do modelo aumenta. Esse ajuste é fundamental, pois um número inadequado de camadas pode comprometer o desempenho da rede, seja por não capturar a complexidade dos dados (*underfitting*) ou por modelar ruídos e detalhes irrelevantes, resultando em um modelo mais complexo do que o necessário.

O processo de calibração também envolveu a execução de diversos testes variando o número de camadas ocultas de 1, 2, 3, 4 e 5 enquanto foi monitorado o desempenho do modelo por meio de métricas como erro quadrático médio (MSE) e coeficiente de determinação (R^2), como pode ser observado na Tabela 2. Estas

métricas foram utilizadas para identificar o equilíbrio ideal entre a capacidade de generalização e o ajuste ao conjunto de treinamento.

Tabela 2 – Efeito de diferentes números de camadas ocultas (1 a 5) para rede neural no ajuste do modelo de calibração.

	5	**4**	**3**	**2**	**1**
R^2	0,93644	0,98201	0,99373	0,99114	0,56915
MSE	0,29102	0,072709	0,0282	0,039728	1,8849
MAE	0,4238	0,16993	0,11689	0,14167	1,1082

A partir da Tabela 2 foi determinado como sendo 3 o número ideal de camadas ocultas para o modelo. A partir desse resultado, foi analisado o número crescente de gerações, e foi observado um número constante de parâmetros de validação, ou seja, a rede neural obteve um grau de convergência a partir de 1000 gerações.

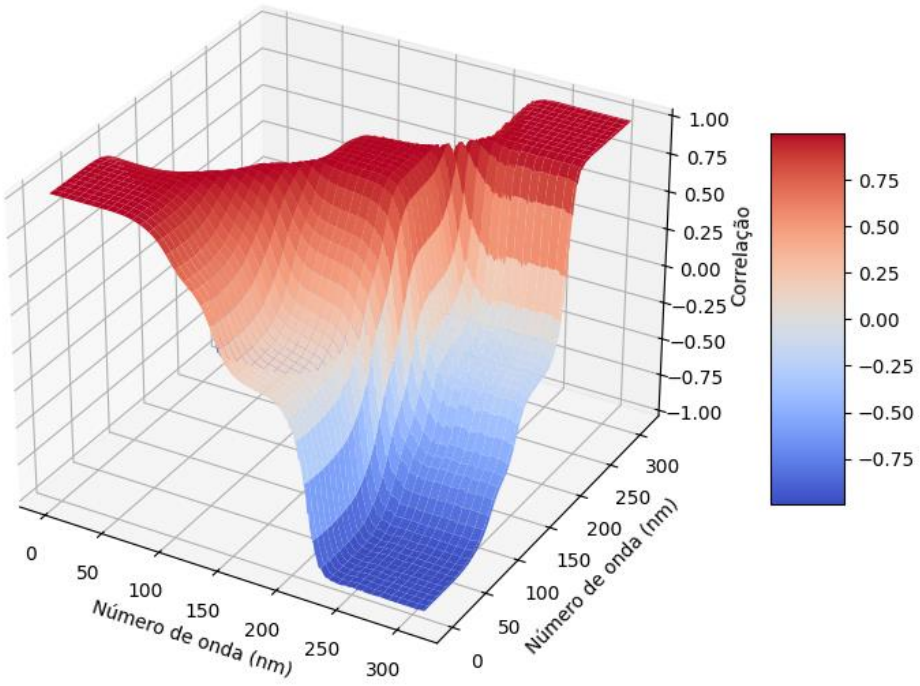
5. RESULTADOS E DISCUSSÃO

5.1. SELEÇÃO DE VARIAVEIS

5.1.1. Seleção de variáveis por matriz linear

A partir da matriz de correlação linear obtida pelo método *dataframe.corr()*, é possível observar graficamente na Figura 5 a relação entre correlação para os diferentes números de onda. Foi realizada uma codificação matemática através da subtração nos eixos dos gráficos, com os comprimentos de onda reais sendo os respectivos valores do eixo X ou Y somado a 300, ou seja, a correlação entre 250 nm e 300 nm seria a relação entre 550 nm e 600 nm para o conjundo de dados real.

Figura 5 – Correlação entre os diferentes números de onda para a mistura de concentrações iguais de 3 corantes.



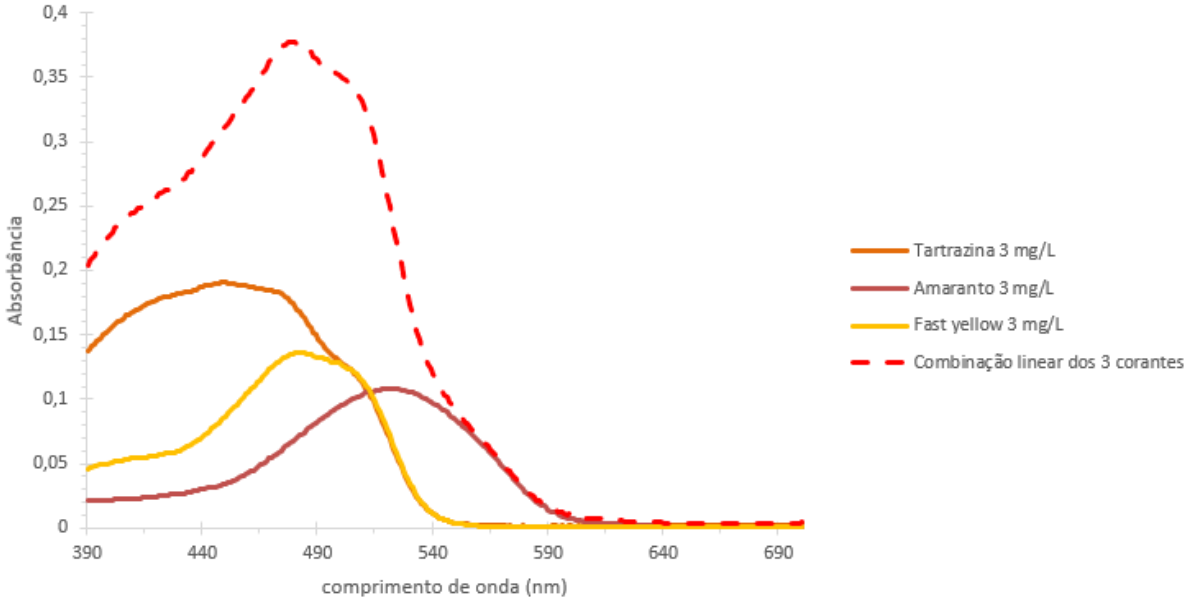
Como pode ser visto, existe uma baixíssima correlação entre os comprimentos de onda extremos, como 300 nm e 100 nm, mostrando que eles explicam variações muito diferentes do sistema. Utilizando o algoritmo desenvolvido de seleção de variáveis, é possível percorrer o mapa selecionando e agrupando os números de onda que melhor descrevem o sistema. Para a escolha do limite colinear é necessário um grau de parcimônia, já que um limite de seleção muito alto como 0,9 irá selecionar variáveis que tenham no máximo uma correlação de 0,9, ou seja, ainda serão muito colineares (como o caso de 300 e 290, que explicam a mesma variabilidade no conjunto de dados e só iriam inserir redundâncias no modelo). Contudo, um limite muito baixo como -0,75 irá selecionar variáveis que tenham no máximo -0,75 de correlação, ou seja, que sejam praticamente ortogonalizadas (o que na figura 5 poderia ser visto entre as variáveis 300 e 75 por exemplo) o que reduziria demasiadamente o número de variáveis a serem utilizadas no modelo, prejudicando a sua calibração e teste, e que comprometeria a robustez, já que desconsideraria a variabilidade das variáveis que possuem uma correlação de -0,25, por exemplo. Portanto, se faz necessário um balanço na hora de escolher o limite colinear, para que o mesmo inclua somente variáveis que não adicionem redundância no modelo (sendo

muito colineares) mas que também não acabe por perder informação de variáveis que tenham um grau moderado de colinearidade.

5.1.2. Seleção de variáveis por resíduos minimizados (SVRMi)

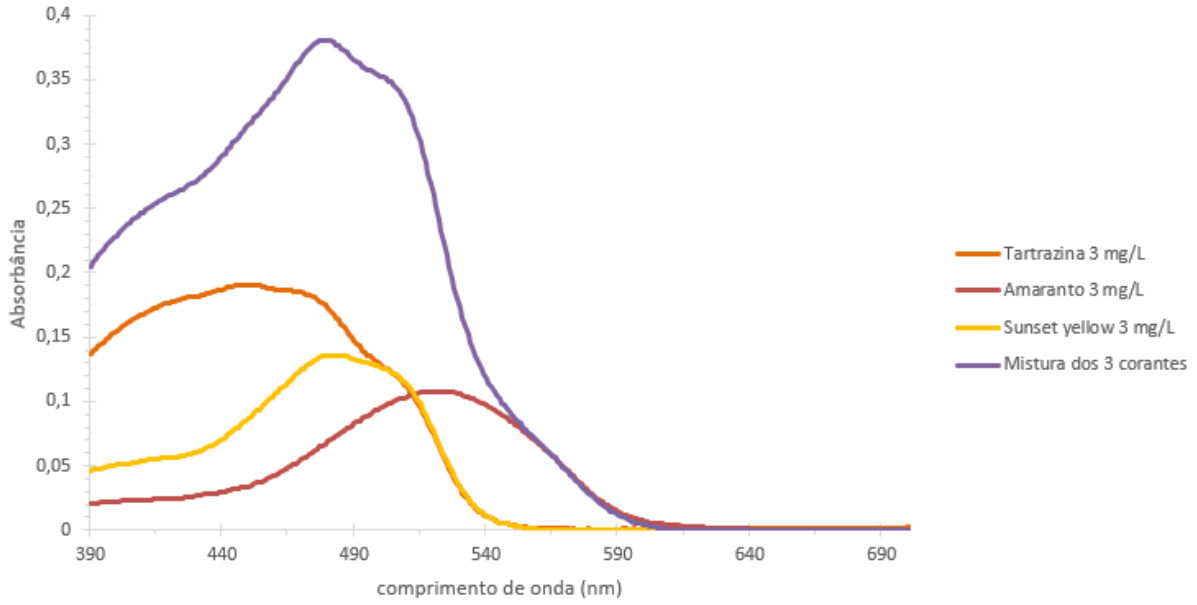
Foram obtidos os espectros de UV-vis individuais para cada um dos corantes (tartrazina, amarando e fast yellow) na concentração de 3 mg/L. A partir destes espectros, foi calculado o espectro teórico da mistura pela combinação linear dos três espectros individuais dos corantes, ou seja, foram somadas as absorções de cada um dos três corantes na concentração de 3mg/L em cada número de onda, assim obtendo o perfil espectral pontilhado na Figura 6.

Figura 6 – Espectros de UV-vis dos corantes tartrazina, amaranto e *fast yellow* na concentração de 3 mg/L, e sua combinação linear.



Foi também obtido o espectro da mistura dos três corantes, cada um na concentração de 3 mg/L, Figura 7.

Figura 7 – Espectros de tartrazina, amaranto, *fast yellow* na concentração de 3 mg/L, e o espectro real da mistura dos três na mesma concentração



Comparando ambas figuras, é possível notar uma alta relação no perfil entre o espectro teórico obtido a partir da combinação linear da absorção dos três corantes (soma das três absorções em cada número de onda) e do espectro real obtido a partir da leitura de uma amostra contendo concentrações iguais para os três corantes. Por mais que visualmente este seja o caso, apenas através da Equação 10 vista anteriormente poderemos verificar se as absorções teóricas (obtidas no espectro teóricos) e as absorções reais (obtida através da análise da solução de 3 mg/L de tartrazina, amaranto e fast yellow) se aproximam numericamente para cada número de onda. Foram verificadas diferentes linearidades para diferentes números de onda, assim mostrando a necessidade de uma seleção nesse conjunto de dados para a análise subsequente, já que a não linearidade (diferença entre absorção real e teórica) indica uma não linearidade na absortividade molar para os corantes em determinados números de onda.

Foi utilizada a equação 10 para um limite de linearidade de 0,001, ou seja, foram selecionados apenas os números de onda cuja a diferença entre a absorção do espectro teórico e a absorção do espectro real estejam na terceira casa decimal, assim, selecionando números de onda que possam ser utilizados para o modelo de resolução de sistemas de equação lineares como poderemos ver posteriormente.

5.2. DESENVOLVIMENTO DO MODELO

5.2.1. ANÁLISE POR SVRMi

A análise por seleção de variáveis de resíduos minimizados resultou em uma série de números de onda que poderiam ser utilizados para a resolução linear do sistema. Para essa técnica, foi empregada a biblioteca *lsq_linear* do *SciPy*, uma ferramenta poderosa do *Python* para resolver sistemas lineares. Nesse contexto, a matriz X foi composta pelas absortividades molares (coeficientes do sistema), enquanto o vetor Y foi formado pelas absorções observadas, com cada linha do sistema representando um comprimento de onda distinto. O objetivo principal era determinar as concentrações molares dos corantes presentes na amostra a partir de um conjunto extenso de sistema de equações lineares que seriam resolvidos pelas bibliotecas supracitadas:

$$Y_{518} = \varepsilon_{518}\,[Tartrazina] + \varepsilon_{518}\,[Amaranto] + \varepsilon_{518}\,[Fast\ Yellow]$$

$$Y_{529} = \varepsilon_{529}\,[Tartrazina] + \varepsilon_{529}\,[Amaranto] + \varepsilon_{529}\,[Fast\ Yellow]$$

$$Y_{534} = \varepsilon_{534}\,[Tartrazina] + \varepsilon_{534}\,[Amaranto] + \varepsilon_{534}\,[Fast\ Yellow]$$

$$Y_n = \varepsilon_n\,[Tartrazina] + \varepsilon_n\,[Amaranto] + \varepsilon_n\,[Fast\ Yellow] \quad (11)$$

Contudo, os resultados obtidos apresentaram grande variação, pois o valor das concentrações calculadas pelo modelo variava de 99% até valores até cinco vezes maiores. Essa discrepância pode ser atribuída à extrema sensibilidade do sistema linear a pequenas flutuações nos coeficientes, especificamente nas absortividades molares. Mesmo ao selecionar comprimentos de onda nos quais essas absortividades fossem minimizadas (ou seja, com valores praticamente constantes), qualquer pequena variação (como a quarta ou quinta casa decimal) gerava mudanças abruptas na solução do sistema de equações. Isso ocorre porque o sistema é altamente interconectado, o que amplifica a propagação de erros, especialmente em equações lineares mal condicionadas.

Essa análise destaca a complexidade de utilizar modelos lineares para resolver problemas em que as variáveis de entrada são extremamente sensíveis, exigindo

métodos mais robustos que não sejam tão influenciados por pequenas flutuações de dados de entrada, como modelos de regressão linear.

5.2.2. Seleção de Variáveis na Regressão por Mínimos Quadrados Parciais (PLS)

A seleção de variáveis por limite colinear melhora o desempenho de modelos PLS principalmente porque evita redundância de informações. A colinearidade ocorre quando duas ou mais variáveis independentes carregam informações muito semelhantes, o que faz com que o modelo desperdice componentes explicando a mesma informação. Ao eliminar variáveis redundantes, cada variável passa a contribuir com informações distintas, tornando o uso dos componentes latentes mais eficiente e significativos, como pode ser visto na Tabela 2.

Tabela 3 – Comparação de parâmetros de eficiência entre PLS com e sem seleção de variáveis por limite colinear.

	R_2	MSE	MAE
Sem Seleção	0,9970	0,0133	0,0976
Com Seleção	0,9988	0,0064	0,0582

Outro ponto importante é que a redução da colinearidade ajuda a evitar o superdimensionamento do modelo. Como o PLS busca maximizar a variância explicada da variável dependente, a presença de variáveis colineares pode fazer com que o modelo crie componentes que acabam "superajustando" o ruído presente nos dados. Selecionando apenas variáveis com baixa colinearidade, o modelo consegue focar nas relações mais relevantes entre os preditores e a variável resposta, garantindo melhor generalização.

5.2.3. Seleção de variáveis na regressão por redes neurais

Foi comparado, incialmente, a utilização de seleção de variáveis por limite colinear para a rede neural de três camadas ocultas, com cada camada possuindo 700, 300 e 30 neurônios, respectivamente. As figuras de mérito obtidas para esse modelo podem ser vistas na Tabela 4.

Tabela 4 – Comparação de parâmetros de eficiência entre rede neural com e sem seleção de variáveis por limite colinear

	R_2	MSE	MAE
Sem Seleção	0,9984	0,007	0,0677
Com Seleção	0,9993	0,0032	0,0417

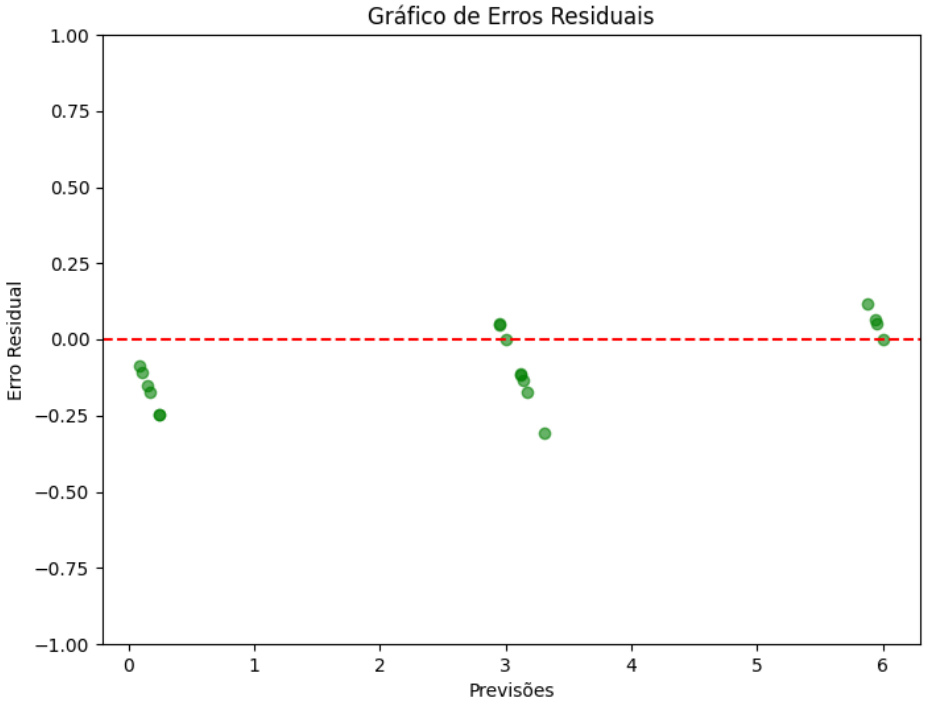
A melhora no desempenho observada na tabela 4 se deve ao fato de que variáveis colineares carregam essencialmente os mesmos padrões ou informações semelhantes, o que pode confundir o modelo durante o treinamento. Esse excesso de dados redundantes faz com que a rede neural desperdice recursos computacionais tentando aprender representações semelhantes, o que diminui a eficiência do processo de aprendizado.

Além disso, a colinearidade pode introduzir instabilidade no processo de ajuste dos pesos da rede. Durante o treinamento, pequenas variações nos dados de entrada podem resultar em grandes mudanças nos pesos associados às variáveis colineares, dificultando a convergência do modelo. Isso torna o treinamento menos robusto e aumenta o risco de *overfitting*, uma vez que o modelo pode acabar ajustando-se demais aos ruídos presentes nos dados em vez de aprender padrões generalizáveis. Em contrapartida, ao excluir variáveis redundantes, o modelo é treinado apenas com as informações essenciais, tornando-o mais robusto e capaz de identificar padrões de forma mais exata.

Utilizando um gráfico de erros residuais podemos observar as previsões feitas pelo modelo e os erros residuais obtidos quando não utilizamos a seleção de variáveis.

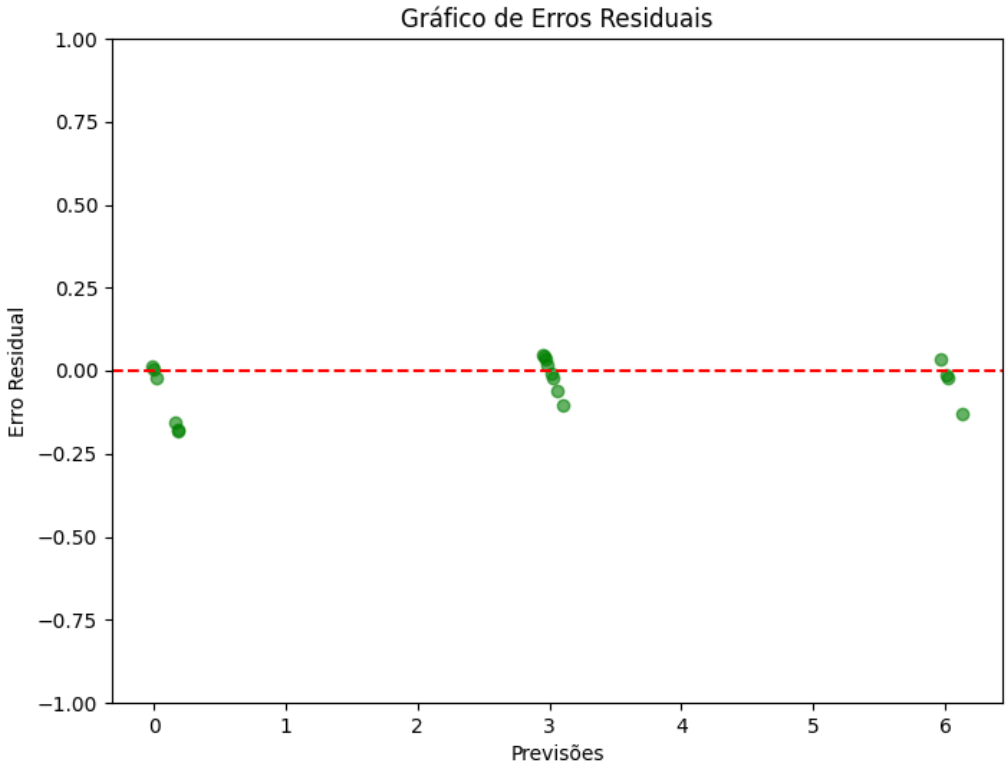
Como podemos verificar pela Figura 8, a distribuição do erro residual deveria seguir uma distribuição normal com os pontos flutuando ao redor do 0, mas verificamos uma variância heterocedástica, o que indica um viés ou um tendencia não captura pelo modelo.

Figura 8 – Previsões realizadas pelo modelo e erro residual obtido na análise sem seleção de variáveis por limite colinear.



Por mais que os pontos estejam distribuídos entre as concentrações utilizadas para calibrar e validar o modelo (0mg/L, 3mg/L e 6mg/L), a distribuição dos erros se assemelha a erros sistemáticos, o que pode comprometer a reprodutibilidade desse método de calibração. Contudo, utilizando a seleção de variáveis proposta (figura 10), podemos observar um perfil mais homocedástico, o que indica um maior ajuste do modelo aos dados e que a maior parte das tendências foi capturada pelo modelo.

Figura 9 – Previsões realizadas pelo modelo e erro residual obtido na análise com seleção de variáveis por limite colinear.



6. CONCLUSÕES

Dentre os principais modelos utilizados para o cálculo de calibração de espectros UV-vis para mistura de múltiplos corantes, foi observado que o método que apresentou melhor desempenho foi o de redes neurais com seleção de variáveis. Isso se deve pelas redes neurais terem a capacidade de modelar relações não lineares entre as variáveis de entrada e de saída, enquanto a regressão por PLS limita-se a capturar apenas relações lineares. Portanto, o método de regressão por redes neurais consegue ajustar comportamentos mais complexos por meio das camadas ocultas e funções de ativação não lineares. Isso torna os métodos com redes neurais mais adequadas para problemas onde os dados apresentam padrões não lineares. Outro ponto importante é a possibilidade de generalização dos modelos, pois os métodos com redes neurais, quando bem treinados e regularizados, permitem generalizar e inferir as concentrações das misturas melhor em relação a novos dados, capturando padrões relevantes sem sofrer com *overfitting*. A regressão por PLS, por depender de relações lineares e de uma redução inicial da dimensionalidade, pode perder informações importantes ao simplificar demais os dados para os cálculos. Dessa forma, para problemas complexos, com grandes volumes de dados ou relações não triviais, a regressão por redes neurais tende a ser mais eficientes e oferecer melhores resultados preditivos do que a regressão por PLS, como foi o caso observado, neste trabalho, para calibração de espectros UV-vis de misturas de corantes.

A seleção de variáveis por limite colinear se demonstrou eficaz em ambos os casos, mas proporcionou um ajuste melhor para a regressão por redes neurais. No caso da regressão por PLS, a redução da colinearidade é essencial para evitar que o modelo fique excessivamente complexo, já que a presença de variáveis colineares pode levar à criação de componentes que capturam o ruído dos dados em vez de padrões importantes. Ao selecionar variáveis com baixa colinearidade, o modelo foca nas relações mais relevantes entre os preditores e a variável resposta, melhorando sua capacidade de generalização.

Já para a regressão por redes neurais, a colinearidade afeta a estabilidade do treinamento ao dificultar o ajuste preciso dos pesos sinápticos. A sensibilidade excessiva a pequenas variações nos dados de entrada pode amplificar flutuações nos pesos associados às variáveis redundantes, resultando em um treinamento instável e menos eficiente. Esse fenômeno compromete a convergência do modelo e, muitas

vezes, favorece o *overfitting*, onde a rede memoriza o ruído dos dados em vez de generalizar padrões robustos e interpretáveis.

Ao eliminar variáveis redundantes, o modelo consegue explorar de forma mais eficiente a representatividade das informações presentes nos dados. A regressão por rede neural concentra seus esforços em aprender características mais relevantes e abstratas, potencializando o uso das camadas ocultas e garantindo uma representação mais robusta e generalizável dos padrões complexos. Isso resulta não apenas em um treinamento mais eficiente, mas também em um modelo mais estável, interpretável e capaz de lidar com cenários reais, onde a robustez e a generalização são essenciais.

REFERÊNCIAS BIBLIOGRÁFICAS

1. UNIVERSIDADE DE CAMPINAS. A química dos pigmentos. Campinas: UNICAMP, 2010.
2. HOLME, I. Sir William Henry Perkin: a review of his life, work and legacy. *Coloration Technology*, v. 122, p. 235–251, 2006.
3. SANTOS, N.; SILVA, F. da S.; MARQUES, I. Corantes naturais: importância e fontes de obtenção. 2675, 2022.
4. PRADO, M.; GOMES, H. A. Nutrição e corantes artificiais em alimentos. 2003.
5. CONSTANT, P.; STRINGHETA, P.; CEPPA, D. S. B. Corantes alimentícios. 2002.
6. KOBYLEWSKI, S.; JAKES, M. Toxicology of food dyes. *International Journal of Occupational and Environmental Health*, Taylor & Francis, 2012..
7. VALDIVIA, V. O. Estudio comparativo en el uso de colorantes naturales y sintéticos en alimentos, desde el punto de vista funcional y toxicológico. Chile, 2004.
8. BURROWS, A. Palette of our palates: a brief history of food coloring and its regulation. *Comprehensive Reviews in Food Science and Food Safety*, v. 8, p. 394–408, 2009.
9. FREITAS, M. S. Corante artificial amarelo tartrazina: uma revisão das propriedades e análises de quantificação. *Acta Tecnológica*, v. 7, p. 65–72, 2013.
10. SABNIS, R. W. *Handbook of biological dyes and stains: synthesis and industrial applications*. John Wiley & Sons, 2010.
11. GUIMARÃES, R. de C. A.; NASCIMENTO, V. A. do; BOGO, D.; HIANE, P. A. Corantes artificiais: uma revisão. *Multitemas*, p. 67–82, 2023.
12. FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L. O. Quimiometria I: calibração multivariada, um tutorial. *Química Nova*, v. 22, p. 724–731, 1999.

13. TENENHAUS, M.; VINZI, V. E.; CHATELIN, Y.-M.; LAURO, C. PLS path modeling. *Computational Statistics & Data Analysis*, v. 48, p. 159–205, 2005.

14. MAHARANA, K.; MONDAL, S.; NEMADE, B. A review: data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, v. 3, p. 91–99, 2022.

15. STANDARD normal variate – an overview. *ScienceDirect Topics*. Disponível em: https://www.sciencedirect.com/topics. Acesso em: 14 jan. 2025.

16. GRISANTI, E. et al. Dynamic Localized SNV, Peak SNV, and Partial Peak SNV: novel standardization methods for preprocessing of spectroscopic data used in predictive modeling. *Journal of Spectroscopy*, v. 2018, p. 1–14, 2018.

17. MALEKI, M. R.; MOUAZEN, A. M.; RAMON, H.; DE BAERDEMAEKER, J. Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. *Biosystems Engineering*, v. 96, p. 427–433, 2007.

18. DHANOA, M. S.; LISTER, S. J.; SANDERSON, R.; BARNES, R. J. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. *Journal of Near Infrared Spectroscopy*, v. 2, p. 43–47, 1994.

19. ANDERSON, M.; TER BRAAK, C. Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation*, v. 73, p. 85–113, 2003.

20. PARTIAL least squares regression – an overview. *ScienceDirect Topics*. Disponível em: https://www.sciencedirect.com/topics/medicine-and-dentistry/partial-least-squares-regression. Acesso em: 14 jan. 2025.

21. NEURAL network – an overview. *ScienceDirect Topics*. Disponível em: https://www.sciencedirect.com/topics/social-sciences/neural-network. Acesso em: 14 jan. 2025.

22. HAYKIN, S. *Redes neurais: princípios e prática*. 2001.

23. ELEMENTAR, E. M. R. C. Coeficiente de determinação. 2018.

24. SCHLUCHTER, M. D. Mean square error. *Encyclopedia of Biostatistics*, 2005. DOI: 10.1002/0470011815.B2A15087.

25. SURYANTO, A. A. Penerapan do método mean absolute error (MAE) no algoritmo de regressão linear para previsão de produção de arroz. *SAINTEKBU*, v. 11, p. 78–83, 2019.

26. SOUZA, E. R.; SIGOLI, F. A. Princípios fundamentais e modelos de transferência de energia inter e intramolecular. *Química Nova*, v. 35, p. 1841–1847, 2012.

27. SANTOS, M. E. dos; DEMIATE, I. M.; NAGATA, N. Determinação simultânea de amarelo tartrazina e amarelo crepúsculo em alimentos via espectrofotometria UV-VIS e métodos de calibração multivariada. *Ciência e Tecnologia de Alimentos*, v. 30, p. 903–909, 2010.

28. VIDOTTI, E. C.; ROLLEMBERG, M. do C. E. Espectrofotometria derivativa: uma estratégia simples para a determinação simultânea de corantes em alimentos. *Química Nova*, v. 29, p. 230–233, 2006.

29. CHEN, Z. et al. Photoluminescence and conductivity of self-assembled π-π stacks of perylene bisimide dyes. *Chemistry - A European Journal*, v. 13, p. 436–449, 2007.

30. NI, Y.; GONG, X. Simultaneous spectrophotometric determination of mixtures of food colorants. *Analytica Chimica Acta*, v. 354, p. 163–171, 1997.

31. DOOSE, S.; NEUWEILER, H.; SAUER, M. A close look at fluorescence quenching of organic dyes by tryptophan. *ChemPhysChem*, v. 6, p. 2277–2285, 2005.

32. NEVADO, J. Simultaneous spectrophotometric determination of tartrazine, patent blue V, and indigo carmine in commercial products by partial least squares and principal component regression methods. *Talanta*, v. 48, p. 895–903, 1999.