UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

DISSERTAÇÃO DE MESTRADO

CLASSIFICAÇÃO DE IOGURTE QUANTO À PRESENÇA DE LACTOSE POR MEIO DA ESPECTROSCOPIA NIR E QUIMIOMETRIA

José Manuel Amancio da Silva

João Pessoa – PB - Brasil

Agosto/2024

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

DISSERTAÇÃO DE MESTRADO

CLASSIFICAÇÃO DE IOGURTE QUANTO À PRESENÇA DE LACTOSE POR MEIO DA ESPECTOSCOPIA NIR E QUIMIOMETRIA

José Manuel Amancio da Silva*

Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Federal da Paraíba como parte dos requisitos para obtenção do título de Mestre em Química, área de concentração Química Analítica.

Orientador: Prof. Dr. Edvan Cirino da Silva

***Bolsista CAPES**

João Pessoa – PB - Brasil

Catalogação na publicação
Seção de Catalogação e Classificação

S586c Silva, José Manuel Amancio da.
Classificação de iogurte quanto à presença de lactose por meio da espectoscopia NIR e quimiometria / José Manuel Amancio da Silva. - João Pessoa, 2024.
76 f. : il.

Orientação: Edvan Cirino da Silva.
Dissertação (Mestrado) - UFPB/CCEN.

1. Iogurte. 2. Lactose. 3. Espectroscopia NIR. I. Silva, Edvan Cirino da. II. Título.

UFPB/BC CDU 664(043)

Elaborado por CHRISTIANE CASTRO LIMA DA SILVA - CRB-15/865

Classificação de iogurte quanto à presença de lactose por meio da espectroscopia NIR e quimiometria.

Trabalho de Dissertação de Mestrado apresentado pelo aluno **JOSÉ MANUEL AMANCIO DA SILVA** e aprovado pela Comissão Examinadora composta abaixo, realizada no dia 12 de agosto de 2024.

gov.br
Documento assinado digitalmente
EDVAN CIRINO DA SILVA
Data: 12/08/2024 21:30:30-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Edvan Cirino da Silva
DQ/UFPB
Orientador/Presidente

gov.br
Documento assinado digitalmente
CLARIMAR JOSE COELHO
Data: 12/08/2024 17:20:54-0300
Verifique em https://validar.iti.gov.br

Dr. Clarimar José Coelho
PUC/Goiânia-GO
Examinador externo

gov.br
Documento assinado digitalmente
MARCIO JOSE COELHO DE PONTES
Data: 12/08/2024 18:02:45-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Márcio José Coelho de Pontes
DQ/UFPB
Examinador interno

Aos meus pais que sempre me incentivaram a nunca parar de estudar, e à minha saudosa avó Severina Maria. Para eles dedico com todo o meu amor.

AGRADECIMENTOS

Primeiramente à Deus, que é pai e filho ao lado do Cristo, por me permitir trilhar esse caminho e pelas pessoas que Ele colocou em minha vida que auxiliaram nessa jornada. Sou grato pela força e todas as graças e lições que Ele me proporcionou. À virgem Maria que, pela graça de Deus, me abençoou e me acalentou diversas vezes.

À minha família, que sempre me apoiou e me incentivou, mesmo que as vezes em pequenos detalhes nem tenham reparado. Em especial a minha mãe, Geralda Eliza de Jesus Silva e ao meu pai, José Amancio da Silva, por serem exemplos de pais e demonstrarem seu amor por mim. Às minhas irmãs, Micaelle Amancio, Michelle Amancio e Mirelle Amancio, que mesmo nos desentendimentos super costumeiros de irmãos nunca deixaram de ser queridas. Aos meus tios e tias que sempre demonstraram seu carinho por mim, em especial Maria de Fátima Amancio, que é uma segunda mãe para seus sobrinhos, à minha tia e madrinha Teodora Amancio e ao seu esposo e meu padrinho Francisco Geraldo.

Aos meus saudosos avós, Eliza Maria de Jesus (*in memoriam*), João Severino da Silva (*in memoriam*), Joaquim Amancio da Silva (*in memoriam*) e Severina Maria da Silva (*in memoriam*) minha querida dona Severina, como eu costuma chamá-la, que se foi no dia que eu fui aprovado para esse mestrado, sei que a senhora olhou por mim. E as minhas duas tias-avós que foram uma figura de avó para mim, Zumira Amancio (*in memoriam*) e Cândida Alves (*in memoriam*). Sou grato à Deus por Ele ter me permitido conhecer todos vocês.

Também agradeço à minha namorada, Patrícia Paloma de Sousa, sem seu apoio, sei que não teria conseguido concluir esses dois anos. Obrigado por segurar a barra e me acalmar quando eu estava prestes a desabar. Sou grato por tê-la ao meu lado e pelo seu amor por mim.

Aos meus amigos, que tanto me ajudaram e me escutaram, em especial à minha grande amiga Maire Gomes de Meneses, agradeço por ter paciência comigo [risos]. À Joyce Gomes, mesmo a gente se falando pouco [você demora a me responder, risos]. A Ruth Bezerra, que me apoiou e auxiliou a executar esse trabalho, sou grato por tudo que aprendi com você. A Francisco Antonio, meu colega de graduação, de mestrado e de apartamento. Aos meus companheiros de laboratório, os laquianos, e à Girlene pelas conversas e risos.

E, por fim, ao meu orientador, professor Dr. Edvan Cirino da Silva, agradeço a orientação e todos os conhecimentos compartilhados.

“Às vezes a vida é como um túnel escuro, nem sempre se pode ver a luz no fim do túnel, mas se você continuar em frente, você chegará a um lugar melhor”.

Tio Iroh, Avatar.

SUMÁRIO

RESUMO

Título: "Classificação de iogurte quanto à presença de lactose por meio da espectroscopia NIR e quimiometria".

Autor: José Manuel Amancio da Silva

O iogurte é uma bebida rica em diversos nutrientes e uma excelente fonte de cálcio, proteínas e vitaminas, obtida a partir da fermentação láctica mediante a ação de microrganismos. O iogurte pode ser classificado como um alimento funcional, possuindo propriedades nutritivas que auxiliam no bom funcionamento do sistema digestivo. Devido à condição de intolerância à lactose, que afeta mais de 75% da população mundial, a enzima β-galactosidase é empregada industrialmente para reduzir o teor de lactose no leite e seus derivados. O controle de qualidade de alimentos recorre geralmente a técnicas analíticas convencionais que são demorosas e de elevado custo operacional. Dessa forma, é necessário desenvolver técnicas alternativas que sejam rápidas, com pouco consumo de reagentes e que permitam fazer medições *in situ.* Assim, surge o interesse em métodos alternativos aos convencionais, como o uso da Espectroscopia no Infravermelho Próximo (NIR, Near Infrared) em combinação com a quimiometria. Portanto, o objetivo deste estudo foi propor uma estratégia analítica não destrutiva, rápida e de baixo custo – usando técnica de espectrometria NIR, combinada com ferramentas quimiométricas de análise multivariada – para a classificação de iogurtes quanto à presença de lactose. Para isso, um total de 102 amostras de iogurte isento e convencional foram adquiridas. Os espectros de reflectância difusa, obtidos na faixa de 900 a 1700 nm, foram adquiridos com equipamento NIR portátil. Diferentes técnicas de pré-processamento foram avaliadas, bem como o uso da validação cruzada e os modelos de classificação de Análise Discriminante Linear (LDA, Linear Discriminant Analysis), com seleção de variáveis por algoritmo dos morcegos (BA, Bat Algorithm), algoritmo genético (GA, Genétic Algorithm) e algoritmo das projeções sucessivas (SPA, Successive Projections Algorithm), e a Modelagem Independente e Flexível por Analogia de Classe, (SIMCA, Soft Independent Modeling of Class Analogy). Os modelos de classificação com seleção de variáveis BA-LDA e GA-LDA classificaram corretamente todas as amostras, atingindo valores máximos de desempenho. Apenas os modelos SPA-LDA com dados brutos e com pré-processamento SNV não alcançaram 100% de desempenho para o conjunto de treinamento. A modelagem de classes SIMCA apresentou sobreposição entre os modelos PCA das classes, o que resultou na classificação incorreta das amostras e em baixos valores de desempenho. Os resultados apresentados neste estudo apontam que a estratégia proposta, que combina espectroscopia NIR e análise multivariada, se mostrou uma ferramenta alternativa viável para realizar a classificação de iogurte com e sem lactose, apenas para modelagem LDA.

Palavras-chave: Iogurte. Lactose. Espectroscopia NIR. BA-LDA. GA-LDA. SPA-LDA. SIMCA.

ABSTRACT

Title: "Classification of yogurt for the presence of lactose using NIR spectroscopy and chemometrics".

Author: José Manuel Amancio da Silva

Yogurt is a nutrient-rich beverage and an excellent source of calcium, proteins, and vitamins, produced through lactic fermentation by microorganisms. Yogurt can be classified as a functional food due to its nutritional properties that support the proper functioning of the digestive system. Given that lactose intolerance affects over 75% of the global population, the enzyme β-galactosidase is used industrially to reduce lactose content in milk and its derivatives. Quality control of food typically relies on conventional analytical techniques that are time-consuming and have high operational costs. Therefore, it is necessary to develop alternative techniques that are rapid, reagent-efficient, and allow for in situ measurements. Hence, there is growing interest in alternatives to conventional methods, such as using Near-Infrared Spectroscopy (NIR) in combination with chemometrics. The purpose of this study was to propose a non-destructive, rapid, and cost-effective analytical strategy— using NIR spectrometry technique, combined with chemometric multivariate analysis tools — to classify yogurt according to lactose presence. For this, a total of 102 samples of zero and regular yogurt were acquired. Diffuse reflectance spectra, obtained in the range of 900 to 1700 nm, were collected using portable NIR equipment. Various pre-processing techniques were evaluated, as well as the use of cross-validation and the classification models Linear Discriminant Analysis (LDA), with selection of variables by Bat Algorithm (BA), Genétic Algorithm (GA) and Successive Projections Algorithm (SPA), and the Soft Independent Modeling of Class Analogy (SIMCA). The classification models with BA-LDA and GA-LDA variable selection correctly classified all samples, achieving maximum performance values. Only the SPA-LDA models with raw data and SNV pre-processing did not reach 100% performance for the training set. SIMCA class modeling showed overlap between PCA models of the classes, resulting in misclassification of samples and lower performance values. The results presented in this study indicate that the proposed strategy, combining NIR spectroscopy and multivariate analysis, proved to be a viable alternative tool for the classification of lactose-containing and lactose-free yogurts, particularly using LDA modeling.

Palavras-chave: Yogurt. Lactose. Spectroscopy NIR. BA-LDA. GA-LDA. SPA-LDA. SIMCA.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

AF - Alimentos funcionais;

ANVISA - Agência Nacional de Vigilância Sanitária;

BA - Algoritmo dos morcegos, do inglês: Bat Algorithm;

BO - Linha de base offset, do inglês: Baseline Offset;

ELSD - Detector de espalhamento de luz evaporativo, do inglês: Evaporative Light Scattering Detector;

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária

FAO - Organização das Nações Unidas para Alimentação e Agricultura;

FIESP - Federação das Indústrias do Estado de São Paulo;

FN – Falso negativo;

FP – Falso positivo;

GA - Algoritmos genéticos, do inglês: Genétic Algorithm;

GDP - Global Dairy Platform;

HCA - Análise por agrupamento hierárquico, do inglês: Hierarchical Cluster Analysis;

HPLC - Cromatografia líquida de alta eficiência, do inglês: High-Performance Liquid Chromatography;

IAL - Instituto Adolfo Lutz;

IBGE - Instituto Brasileiro de Geografia e Estatística;

IDF - International Dairy Federation;

IL- Intolerância à lactose;

KS - Algoritmo Kennard-Stone;

LDA - Análise Discriminante Linear, do inglês: Linear Discriminant Analysis;

MAPA - Ministério da Agricultura, Pecuária e Abastecimento;

MIR - Infravermelho médio. do inglês: Middle Infrared;

NIR - Infravermelho próximo, do inglês: Next Infrared;

OMS - Organização Mundial da Saúde;

PCA - Análise de componentes principais, do inglês: Principal Component Analysis;

PLS-DA - Regressão por Mínimos Quadrados Parciais para Análise Discriminante, do inglês: Partial Least Squares for Discriminant Analysis;

POF - Pesquisa de Orçamentos Familiares;

RP - Reconhecimento de padrões;

SG - Savitzky-Golay;

SIMCA - Modelagem Independente e Flexível por Analogia de Classe, do inglês: Soft Independent Modeling of Class Analogy;

SNV - Variação normal padrão, do inglês: Standard Normal Variation;

SPA - Algoritmo das projeções sucessivas, do inglês: Successive Projections Algorithm;

TCC – Taxa de classificação correta;

UHT – Temperatura ultra-alta, do inglês: Ultra-high Temperature;

VN – Verdadeiro negativo;

VP – Verdadeiro positivo.

Capítulo 1

INTRODUÇÃO

1 INTRODUÇÃO

1.1 CARACTERIZAÇÃO DA PROBLEMÁTICA E PROPOSTA DO TRABALHO

O leite está entre os alimentos mais consumidos do planeta, em 2023 sua produção mundial foi de 965,7 milhões de toneladas, de acordo com dados da Organização das Nações Unidas para Alimentação e Agricultura (FAO, 2024). Em 2022, a produção de leite no Brasil, conforme pesquisadores da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), foi de 23,81 bilhões de litros, já para o ano de 2023 houve um aumento 2,53%, fechando a produção anual em 24,5 bilhões de litros (CARVALHO, 2023; CARVALHO; OLIVEIRA; ARANTES, 2024; RENTERO, 2023).

Segundo a EMBRAPA, o mercado de leite não está restrito apenas à sua forma original, pois a comercialização de seus derivados é um ramo muito forte na indústria. Nesse contexto, o leite é transformado em diversos produtos, entre os quais se destacam os queijos e a manteiga. Ademais, é transformado em alimentos destinados à sobremesa, tais como iogurte, bebida láctea, leite condensado, fermentado e doce de leite (SIQUEIRA, 2019).

O cálcio é um dos principais nutrientes para a dieta humana, pois é o mineral mais abundante na composição da estrutura óssea. Juntamente com aminoácidos, vitaminas, enzimas e fósforo, o cálcio é responsável pela manutenção, crescimento e regeneração da estrutura óssea. A falta desses nutrientes pode levar a doenças que afetam todo o sistema esquelético (ZYLBERGELD, 2019). Portanto, a ingestão de alimentos ricos em cálcio é de extrema importância para assegurar a saúde e o bem-estar do ser humano.

O leite e seus derivados destacam-se como os principais alimentos que contribuem para o desenvolvimento e a nutrição humana. Por ser um alimento complexo, sua vasta e completa composição é rica em fontes nutritivas, devido à presença de macro e micronutrientes muito importantes para a nutrição humana, tais como: proteínas de alta qualidade, cálcio, magnésio, selênio, riboflavina, vitamina B12, gorduras, lactose, caseína (BOLAND; SINGH, 2020; SIQUEIRA, 2019).

Tendo em vista a riqueza nutricional do leite, é notória a necessidade da presença deste alimento como fonte de nutrientes na dieta humana. Entretanto, conforme a literatura, observa-se uma redução no consumo de leite, principalmente em países desenvolvidos. Isso se deve ao fato de muitos indivíduos adultos apresentarem diagnóstico de intolerância à lactose (IL) (DANTAS; VERRUCK; PRUDENCIO, 2019).

A intolerância à lactose consiste em um distúrbio gastrointestinal que pode acometer um indivíduo, impossibilitando-o de consumir leite e seus derivados. Este problema é causado pela

deficiência ou ausência da produção da lactase-florizina hidrolase, também conhecida como lactase ou β-galactosidase, que é a enzima responsável por hidrolisar a lactose, o açúcar presente no leite de origem animal (CASTELLANO, *et al.,* 2022; FORSGARD, 2019; VARELLA, 2018). Consequentemente, dificulta ou impossibilita o indivíduo de ingerir produtos que contenham o dissacarídeo, tendo em vista que este não poderá ser digerido pelo sistema digestivo.

A intolerância à lactose possui sintomas que podem gerar desconforto, tais como: perda de peso, diarreia, dor abdominal, flatulência e, em casos incomuns, constipação (SHARIF *et al.,* 2017). Nesse contexto, faz-se necessário o desenvolvimento de artigos isentos de lactose - comercialmente chamados de zero lactose - para suprir a necessidade de pessoas com essa enfermidade.

Conforme dados expressos na literatura, cerca de 75% da população mundial sofre de déficit de produção da enzima β-galactosidase (SURI, *et al,* 2019; PERATI; BORBA; ROHRER*,* 2018; PEREIRA, *et al.,* 2020). Esse fato contribui diariamente para a necessidade do mercado em disponibilizar alimentos isentos de lactose, mesmo essa não seja uma condição perigosa para a saúde. A fabricação de produtos sem lactose é realizada por meio de um processo de hidrólise enzimática com lactase, na qual a lactose é decomposta em glicose e galactose – que são dois monossacarídeos absorvíveis pelo sistema digestor, os quais servem como fonte de energia (PERATI; BORBA; ROHRER*,* 2018; SURI *et al.,* 2019).

Por se tratar de uma patologia que acomete grande parte da população mundial, a intolerância à lactose não afeta apenas o sistema digestório, mas também o psicológico. Isso pode provocar uma sensação de desânimo e de insegurança para se alimentar, manter um convívio social, bem como um receio da ocorrência de um mal súbito provocado pela ingestão de alimentos contendo a lactose. Para Suri e colaboradores (2019), evitar o consumo de produtos com lactose permite resolver os problemas nutricionais. No entanto, seguir essa dieta é uma tarefa difícil, pois necessita de um monitoramento constante e rigoroso dos alimentos consumidos.

Assim, é necessário garantir alimentos destinados a pessoas com intolerância à lactose e assegurar que estes sejam isentos do carboidrato, ou contenham a quantidade mínima recomendada. O Ministério da Agricultura, Pecuária e Abastecimento (MAPA) estipula os métodos de cromatografia iônica e enzimático com medição diferencial de pH para a determinação de lactose (BRASIL, 2022). Todavia, esses métodos são demorados e requerem um alto consumo de reagentes. Devido ao cenário atual de crescente procura por produtos sem

lactose, é necessário desenvolver métodos rápidos, precisos e com menor custo operacional, de modo que assegurem a qualidade destes produtos.

A espectroscopia no infravermelho próximo (NIR, do inglês: Near Infrared) tem sido utilizada largamente como uma alternativa para determinação qualitativa e/ou quantitativa de compostos, cujas moléculas contenham geralmente ligações C-H, N-H e O-H. De acordo com Nóbrega (2021), a espectrometria NIR surgiu como uma alternativa promissora e confiável em virtude dessa técnica permitir realização de análises não destrutivas e não invasivas, bem como não precisar do uso de reagentes contemplando os princípios da Química Verde.

Em face da necessidade alimentícia das pessoas com intolerância à lactose e de métodos que permitam sua determinação com rapidez, acurácia e baixo custo de análise, o presente trabalho propõe o uso da técnica NIR, em combinação com a quimiometria, visando à classificação de iogurtes quanto à presença de lactose.

1.2 OBJETIVOS

1.2.1 Geral

Propor uma estratégia analítica não destrutiva, rápida e de baixo custo – usando a técnica de espectrometria NIR, combinada com ferramentas quimiométricas de análise multivariadas – para a classificação de iogurtes quanto à presença de lactose.

1.2.2 Objetivos Específicos

Na consecução do objetivo geral, foram estabelecidos e implementados os objetivos específicos que são apresentados a seguir:

- Explorar a região NIR do espectro eletromagnético para obtenção dos espectros de absorção, com equipamento portátil, de iogurtes a fim de estudar a viabilidade da espectrometria NIR, como uma técnica instrumental adequada, combinada às técnicas quimiométricas multivariadas para a classificação de iogurtes com e sem lactose;

- Realizar uma análise exploratória dos dados por meio da Análise de Componentes Principais para investigar a existência de agrupamentos naturais das amostras de iogurtes com e sem lactose.

- Construir e validar modelos de classificação, usando técnicas de reconhecimento de padrão supervisionado, nomeadamente, o SIMCA (Soft Independent Modeling of Class Analogy) e LDA (Linear Discriminant Analysis) baseada em seleção de variáveis usando os algoritmos BA (bat algorithm), GA (genetic algorithm) e SPA (successive projections algorithm) acoplados à LDA;

- Avaliar o desempenho dos modelos de classificação de classe única, SIMCA, e de discriminação, a saber: BA-LDA, GA-LDA e SPA-LDA, por meio das métricas de desempenho sensibilidade, especificidade, precisão e taxa de classificação;

- Demonstrar a viabilidade do uso de um instrumento NIR portátil para a implementação de uma estratégia não destrutiva, rápida e de baixo custo para classificação de iogurtes quanto à presença de lactose – o que permite, ainda, um controle de qualidade *in situ*.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

2 FUNDAMENTAÇÃO TEÓRICA

2.1 CONSUMO E CARACTERÍSTICAS DO LEITE

O leite movimenta uma grande parte do mercado de alimentos, visto que ele e seus derivados estão presentes em diversos produtos ou receitas, como tortas, bolos, pães, dentre outros. Segundo a literatura, aproximadamente 1 bilhão de pessoas ao redor do mundo dependem do leite para garantir sua sobrevivência, sendo que 600 milhões vivem em fazendas produtoras de leite (GDP, 2017; SIQUEIRA, 2019).

No Brasil, não existe uma recomendação da quantidade ideal a ser consumida de leite, entretanto, o consumo médio do alimento por um adulto é de 455 ml/dia, sendo superior a países como África do Sul, Argentina e Austrália. Quando considerado apenas o leite fluido (leite processado em temperatura ultra-alta, UHT, do inglês: Ultra-high Temperature, e pasteurizado), o brasileiro consome 73 ml por dia, superior à média diária dos Estados Unidos que é de 64 ml (SIQUEIRA 2021).

Boland e Singh (2020) afirmam que as proteínas possuem o maior valor agregado entre os principais componentes do leite, característica associada à maior aceitação da proteína do leite bovino como uma fonte nutricional superior em comparação com outras fontes proteicas. Ainda segundo Boland e Singh, a preferência pela proteína do leite bovino alavancou a produção de diversos alimentos proteicos a partir desse alimento. Além disso, essa variedade apresenta diferentes composições e propriedades funcionais, possibilitando uma gama de aplicações industriais.

O leite é um fluido corporal secretado pelas fêmeas dos mamíferos, sua função primordial é atender as necessidades nutricionais e fisiológica do neonato[1], as proteínas presentes em sua composição fornecem aminoácidos essenciais e possibilitam a biossínteses de aminoácidos não essenciais (BOLAND; SINGH, 2020). Considerado um alimento nutricionalmente rico e consumido mundialmente, o leite possui um papel fundamental na dieta humana como fornecedor de energia e de nutrientes, seus principais componentes são: água, lipídios, proteínas, lactose, minerais e vitaminas (SILVA, 2023; ZEBIB; ABATE; WOLDEGIORGIS 2023).

A composição do leite é afetada por variações sazonais, tais como: raça do animal, idade, estado de saúde, estágio de lactação, dieta e padrões de alimentação, intervalo de ordenha e localizações geográficas (CHEN; LEWIS; GRANDISON, 2014; ZEBIB; ABATE;

[1] Recém-nascido, está entre o período inicial da vida, desde o nascimento até os primeiros 28 dias.

WOLDEGIORGIS, 2023). Nesse contexto, Boland e Singh (2020) discorrem acerca da composição do leite e seus constituintes mudarem de forma acentuada com cada período da lactação, sendo observada uma notável alteração durante os primeiros dias após o parto, em especial na fração imunoglobulina das proteínas. Após a fase inicial, nenhuma alteração é observada em sua composição durante o transcorrer do período de amamentação, permanecendo relativamente constante até se aproximar da fase final da lactação. Essa alteração é consequência da involução dos tecidos das glândulas mamárias e um aumento da concentração de sangue.

2.1.1 Composição e benefícios do Iogurte

O iogurte está cada vez mais presente na mesa do consumidor brasileiro, pois contém nutrientes essenciais e é um veículo de fortificação, sendo composto de probióticos, fibras, vitaminas e minerais (FISBERG; MACHADO, 2015). Nesse contexto, é um alimento bem aceito no mercado, podendo ser consumido por pessoas de todas as idades, inclusive aquelas que buscam por uma dieta menos calórica (SILVA; PANDOLFI, 2020).

O Ministério da Agricultura, Pecuária e Abastecimento (MAPA) define iogurte como um produto obtido por coagulação e diminuição do pH do leite através da fermentação láctica mediante a ação de microrganismos. Para realizar esse processo são utilizados os seguintes cultivos: *Streptococcus salivarius subsp. thermophilus* e *Lactobacillus delbrueckii subsp. Bulgaricus* (BRASIL, 2007).

Em 2017 e 2018, o Instituto Brasileiro de Geografia e Estatística (IBGE) realizou a Pesquisa de Orçamentos Familiares (POF), cujo objetivo era demonstrar o consumo e os gastos das famílias brasileiras. De acordo com esse estudo, o consumo médio *per capita* diária de leite (integral e desnatado) no Brasil é de 8,8 g, outros produtos lácteos, como o iogurte e o queijo, possuem um consumo de 8,1 e 5,8 g, respectivamente (IBGE, 2020). Uma pesquisa realizada pela EMBRAPA em 2020, para monitorar o consumo de leite e derivados no Brasil, constatou que 89% dos brasileiros possuem o hábito de comprar iogurte, como consequência, o alimento é um dos quatros principais laticínios mais consumidos no país (SIQUEIRA *et al*, 2021).

A composição do iogurte é rica em diversos nutrientes, sendo uma excelente fonte de proteína e cálcio, além de apresentar uma vasta variedade de vitaminas (BELADELI; GOLDINHO, 2022). A **Tabela 2.1** apresenta os principais constituintes encontrados o iogurte.

Ao longo dos anos, as preocupações com a saúde provocaram um aumento na busca por alimentos mais nutritivos e equilibrados. Segundo a Federação das Indústrias do Estado de São Paulo (FIESP), 81% dos brasileiros procuram manter uma alimentação mais saudável e 71%

afirmam estar dispostos a gastar mais para adquiri-la (FIESP, 2017). De acordo com Salgado (2017), um dos principais segmentos alimentícios desta classe são os alimentos funcionais (AF), um setor com um alto poder de crescimento e diversidade, que vem ganhando espaço no mercado principalmente por auxiliar no combate de doenças.

Tabela 2.1: Composição do iogurte em gramas.

	100 g de iogurte
Proteínas	5,0
Lipídeos	1,0
Lactose	4,5
Cálcio	0,2
Fósforo	0,1
Ácido láctico	1,0
Bactérias	0,1

Fonte: Adaptado de Beladeli e Goldinho (2022).

Alimentos funcionais podem ser definidos como alimentos ou ingredientes que oferecem benefícios à saúde, além das suas funções nutricionais básicas inerentes às suas composições químicas, desta forma desempenhando um papel na redução de doenças crônicas degenerativas, como câncer e diabetes (BRASIL, 2009; SIQUEIRA, 2021). Assim, almejando uma maior longevidade e qualidade de vida da população, os AF's desempenham um papel importante para construção de novos hábitos alimentares que visam benefícios para a saúde em longo prazo (SAFRAID *et al.,* 2022).

Os laticínios são um dos principais AF's, devido às propriedades de alguns produtos serem naturalmente funcionais e outros possuírem facilidade de manipulação. Graças a essas características, os laticínios ocuparam o primeiro lugar nas vendas de alimentos funcionais no ano de 2018, à frente de produtos como grãos e carnes (SIQUEIRA, 2021).

Alguns AF's visam a saúde intestinal, como os itens que contêm probióticos em sua composição, como leites fermentados e iogurte. A Organização das Nações Unidas para Alimentação e Agricultura (FAO) e a Organização Mundial da Saúde (OMS) definem probióticos como micro-organismos vivos que conferem benefícios à saúde do hospedeiro, influenciando o organismo com o intuito de melhorar o balanço microbiano (FAO, 2016; VALÉRIO; COSTA; CARDINES, 2022).

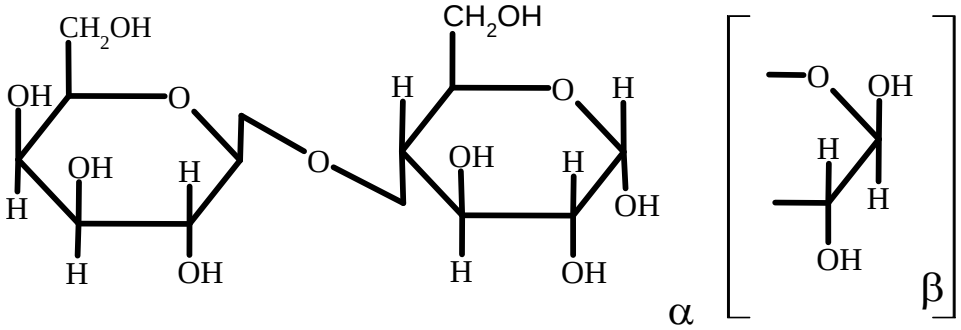
A presença de probióticos na composição de algumas linhas de iogurtes é um dos principais benefícios do uso da bebida. Cruz e Zacarchenco (2015) salientam que o iogurte e demais leites fermentados são os principais representantes da inclusão de culturas probióticas na dieta humana em todo o mundo, sendo a maior parte financeira do mercado na área de produtos dessa classe.

Assim, o iogurte possui propriedades nutritivas que auxiliam no bom funcionamento do intestino e é de fácil digestão, o que contribui para que ele seja considerado um bom produto para pessoas com problemas gastrointestinais. Por conter cálcio, o iogurte auxilia no fortalecimento de dentes, além de ser indicado em casos de osteoporose e para mulheres no período de menopausa que necessitam repor o mineral (SILVA *et al.,* 2017). Por ser uma excelente fonte de múltiplas vitaminas, minerais e proteínas, o iogurte contribui para a produção de anticorpos, hormônios e enzimas, reforçando o sistema imunológico e retardando o envelhecimento (PELEGRINE; AGUIAR; IODELES, 2015).

2.1.2 Lactose: características e intolerância

A lactose é encontrada exclusivamente no leite e tem como principal função ser uma fonte de energia. É um dissacarídeo redutor e o principal carboidrato do leite bovino, composta por dois monossacarídeos (a galactose e a glicose) ligados por intermédio de uma ligação glicosídica β1-4 (BOLAND; SINGH, 2020). Com forma molecular $C_{12}H_{22}O_{11}$ e nome sistemático β-D-galactopiranosil-D-glucopiranose, a lactose pode ocorrer nas formas α e β (**Figura 2.1**), sendo as duas formas esteriorisômeros que diferem pelo arranjo espacial da hidroxila do grupo hemiacetal da porção glicose (CARPIN *et al*., 2016).

Figura 2.1: Molécula da lactose.



Fonte: Ilustração própria (2024).

A concentração de lactose difere entre as espécies de mamíferos. No leite bovino, o teor do carboidrato pode variar em função da raça do animal, entre os indivíduos, infecção do úbere

(a mastite) e, principalmente, com o estágio de lactação, diminuindo gradativamente durante este período (FOX *et al.,* 2015).

A digestão da lactose ocorre por meio da hidrólise enzimática, portanto, a produção da enzima lactase é crucial, uma vez que, em sua ausência, a absorção intestinal da lactose não é possível. A atividade enzimática da lactase pode ser percebida desde a 8ª semana de gestação, aumentando até 34ª semana, contudo seu auge é atingido com o nascimento (DENG *et al*., 2015).

A hidrólise da lactose é crucial para a saúde dos recém-nascidos, uma vez que eles se alimentam do leite materno. Dessa forma, na maioria das vezes, a atividade enzimática da lactase é alta em recém-nascidos. Em contrapartida, a presença da enzima diminui gradualmente após o desmame, sendo observada uma variação que difere entre indivíduos (BRANCO *et* al., 2017; RAMALHO; GANECO, 2016; STORHAUG; FOSSE; FADNES, 2017). Além disso, o déficit de produção dessa enzima resulta em uma condição denominada hipolactasia, conhecida popularmente por intolerância à lactose, que é a inabilidade de consumir alimentos com lactose. Essa condição ocorre quando a lactose não hidrolisada chega ao intestino, contudo, não é absorvida, atuando como um substrato bacteriano podendo distender o cólon, provocando diarreia osmótica[2] (DENG, *et al.,* 2015; HARTWIG, 2014; STORHAUG; FOSSE; FADNES, 2017).

Em condições normais, após a lactose ser hidrolisada pela enzima lactase, a glicose e a galactose são transportadas pelas membranas celulares epiteliais para a corrente sanguínea. No entanto, em situações de déficit de lactase, a lactose ingerida não é degradada e passa para o cólon, onde algumas bactérias como bifidobactérias, lactobacilos e *Escherichia coli*, têm a capacidade de metabolizar a lactose. Ao serem liberadas através desse processo, a glicose e a galactose são convertidas pelas bactérias intestinais em diversos produtos, como ácidos graxos de cadeia curta e gás hidrogênio, ocasionando dor abdominal e inchaço (CORGNEAU *et al.,* 2017; DANTAS; VERRUCK; PRUDENCIO, 2019; SILVA, 2017).

Existem três formas distintas de intolerância à lactose: intolerância primária; intolerância secundária; intolerância congénita. A primeira forma de intolerância, a intolerância primária, é a condição mais comum caracterizada por um declínio gradual da atividade enzimática após o desmame, podendo ser observada em crianças, adolescentes ou adultos. Sua variação está intrinsicamente ligada com a quantidade de alimentos ingeridos que contenham lactose (BRANCO *et al.,* 2017; DENG, *et al.,* 2015; GUERRA *et al.,* 2018; SILVA; 2017).

[2] Situação na qual uma substância não pode ser absorvida pelo cólon e permanece no intestino, fazendo com que uma quantidade excessiva de água se converte nas fezes, ocasionando a diarreia.

A intolerância secundária é ocasionada por patologias, tais como lesão tecidual, que podem ser ocasionadas por quimioterapia, radioterapia, diarreia crônica, infecções virais agudas, parasitoses, doença de Crohn, dentre outras, que resulta na perda das células epiteliais, as quais são responsáveis pela produção da enzima lactase. Ao contrário da intolerância primária, a secundária pode ocorrer em qualquer idade e, em alguns casos é transitória, podendo ser revertida após o tratamento da doença precursora da intolerância (BRANCO *et al.,* 2017; GUERRA *et al.,* 2018; SILVA; 2017).

Por fim, a intolerância congênita é definida na literatura como uma condição rara em que um recém-nascidos é acometido de um problema genético que o impossibilita de produzir a enzima. É considerada extremamente grave, podendo ser fatal se não for detectada no momento inicial. Essa deficiência é caracterizada como uma herança genética e uma doença autossômica recessiva, ocasionada por uma mutação no código genético da lactase (BRANCO *et al.,* 2017; GUERRA *et al.,* 2018; SILVA; 2017).

Uma quarta condição pode ser encontrada na literatura, onde uma IL temporária pode ser observada em recém-nascidos prematuros. Esse estado é consequência da falta de maturidade do intestino, dessa forma sendo incapaz de produzir a enzima em quantidades adequadas, no entanto esse quadro é naturalmente revertido à medida que a criança cresce (GUERRA *et al.,* 2018).

Portanto, seja qual for a causa, a má absorção da lactose é uma condição que pode se tornar perigosa, assim, torna-se necessário normas que regulamentem a produção de alimentos destinados a pessoas portadoras dessa patologia. A resolução de nº 135, de 8 de fevereiro de 2017, da Agência Nacional de Vigilância Sanitária (ANVISA) delimita que alimentos isentos de lactose devem conter uma quantidade de lactose igual ou inferior a 0,1g/100 gramas ou mililitros do alimento pronto para consumo. Para ser considerado baixo teor de lactose, a concentração do carboidrato deve ser superior a 0,1g/100 gramas ou mililitros e igual ou inferior a 1 grama por 100 gramas ou miligramas do alimento (BRASIL, 2017).

Em seu trabalho, Pereira *et al.* (2012) afirmam que, devido a presença de micro-organismos em sua composição, os iogurtes e os leites fermentados possuem boa tolerabilidade pelos consumidores com intolerância a lactose. Entretanto, Mesquita Junior e colaboradores (2021), destacam que, apesar da etapa de fermentação presente na fabricação dos iogurtes, seu consumo ainda pode ocasionar desconfortos. Segundo Rosa e Alves (2019), o teor de lactose presente no iogurte pode variar de 2,4 a 4 g em 100 ml do produto. A **Tabela 2.2** apresenta alguns teores de lactose para diferentes tipos de iogurte e leite.

Tabela 2.2: Teores de lactose para diferentes tipos de leite e iogurte.

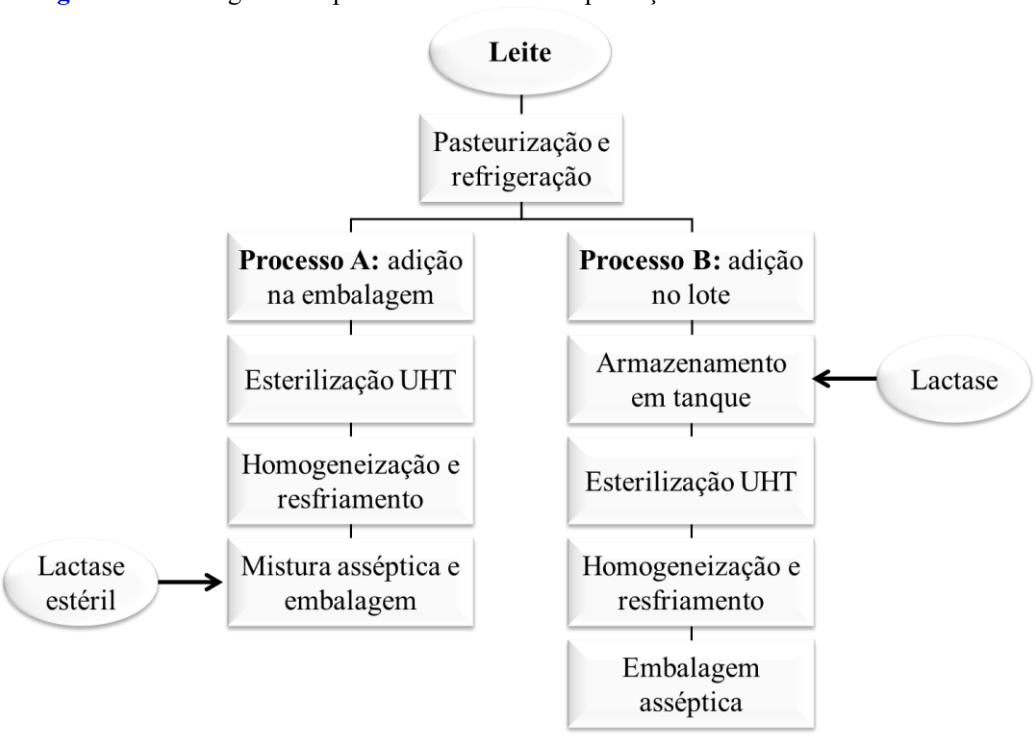
Leite	g/100	Iogurte	g/100
Desnatado	4,8	Natural	4,0
Semidesnatado	4,7	Fruta	4,0
Integral	4,6	Líquido	4,0
Pó desnatado	52,9	Fromage frais	3,0
Condensado	12,3	Tzatziki	4,7

Fonte: Adaptado de Mattar e Mazo (2010).

De acordo com os dados da literatura, o teor de lactose presente no iogurte é superior ao estipulado pela resolução nº 135 da ANVISA para baixo teor de lactose. Sendo assim, podendo ocasionar desconforto, portanto devem ser evitados por pessoas com intolerância a lactose (MATTAR; MAZO, 2010; ROSA; ALVES, 2019).

A enzima β-galactosidase é industrialmente utilizada para reduzir o teor de lactose em leite e seus derivados, a fim de produzir alimentos destinados a pessoas com restrição de consumo de lactose. A aplicação da enzima com essa finalidade é bastante difundida na literatura, sendo possível encontrar diversos trabalhos que abordem sua utilização (KURIBAYASHI *et al.,* 2021; RAMOS, 2022; RIBEIRO; CUBO; SALEM, 201; SILVA; SOUSA; FALLEIROS, 2019).

Figura 2.2: Fluxograma de processos utilizados na produção de leite baixo teor de lactose.



Fonte: Adaptado de Troise *et al* (2016).

Troise *et al* (2016) e Kuribayashi1 *et al* (2021) relatam dois processos de produção de leite baixo teor de lactose que se diferem na etapa de adição da enzima lactase (**Figura 2.2**). Os produtos obtidos via processo A apresentam atividade enzimática durante seu período de validade, visto que a adição da enzima se dá após a esterilização, enquanto os produtos obtidos pelo processo B tem sua atividade enzimática invalidados durante a etapa de esterilização. Portanto, leites obtidos pelo processo A, a hidrólise perdura durante todo o prazo de validade.

2.2 MÉTODOS PARA DETERMINAÇÃO DE LACTOSE

O crescimento desenfreado da população mundial e o aumento do consumo de alimentos ocasionam uma preocupação aos consumidores e ao setor industrial com relação à qualidade dos produtos alimentícios. A análise de qualidade de alimentos é relatada por muitos autores como sendo demorosa e de elevado custo, pois são realizadas convencionalmente, recorrendo ao emprego da analítica clássica, como processos titulométricos e gravimétricos ou com técnicas mais refinadas. Esses métodos envolvem um alto consumo de reagentes nocivos, diversas etapas de reparo de amostras, além de alguns necessitarem de equipamentos caros que exigem operadores treinados (CARAMÊS; LIMA-PALLONE, 2021; LUCAS; SCHÚ; NORA, 2021).

No entanto, diante do cenário atual, houve um aumento na procura por tecnologias alternativas que não causem ou diminuam o impacto negativo ao meio ambiente. Essas novas tecnologias, em geral, fazem o uso de pouco ou nenhum solvente. Segundo a literatura, existe uma necessidade de desenvolver técnicas que sigam os princípios da química verde, como redução do uso de reagentes tóxicos e medidas *in situ*. Assim, surge o apreço por técnicas como a espectroscopia no infravermelho ou imagens digitais, as quais demonstram grande potencial como métodos alternativos aos tradicionais, além da praticidade de poder empregar dispositivos portáteis (HUSSAIN; SUN; PU, 2019; LIMA-PALLONE, 2021; LUCAS; SCHÚ; NORA, 2021).

Sobre à lactose, o MAPA estabelece, por meio do manual "Métodos Oficiais para Análise de Produtos de Origem Animal", alguns métodos oficiais para determinação de lactose em alimentos específicos, como é o caso do leite fluido, para o qual deve seguir a norma IDF[3] 214, que utiliza o método enzimático por meio da diferença de pH. A enzima β-Galactosidase é utilizada para realizar a hidrólise da lactose, em pH 7,8, em seguida a glicose é fosforilada pela glucoquinase, liberando prótons que induzem a mudança no pH (BRASIL, 2022).

[3] Federação Internacional de Lacticínios, do inglês: *International Dairy Federation*.

A cromatografia iônica é recomendada pelo MAPA como método para a determinação da lactose em laticínios em geral, com a exceção dos caseinatos. São empregadas duas colunas de estireno/divinilbenzeno montadas em série, um Cromatógrafo de íons, com detector amperométrico de célula de ouro, e uma fase móvel composta por uma solução de hidróxido de sódio (NaOH) a 300 mmol L^{-1}e acetato de sódio a 1 mmol L^{-1} (BRASIL, 2022).

Outro método amplamente utilizado é a determinação de lactose pelo método de Lane-Eyon, uma titulação a quente que se baseia na determinação do volume de amostras necessário para reduzir uma medida da solução de Fehling (Silva *et al.*, 2020). Essa metodologia é recomendada pelo Instituto Adolfo Lutz em seu manual de métodos fisíco-químicos para análise de alimentos (IAL, 2008).

Diante dos obstáculos encontrados nos métodos convencionais empregados para o controle de qualidade, torna-se necessário o desenvolvimento de metodologias alternativas que atendam às necessidades de técnicas com alta velocidade analítica, que diminuam o consumo de reagentes, bem como o uso de equipamentos de elevado custo e de difícil operação.

Simião e colaboradores (2018) avaliaram o desempenho da Cromatografia líquida de alta eficiência (HPLC, do inglês: High-Performance Liquid Chromatography) com detector de espalhamento de luz evaporativo (ELSD, do inglês: Evaporative Light Scattering Detector), como metodologia alternativa para avaliar o teor de lactose em oito produtos zero lactose, utilizando acetonitrila e água deionizada como fases móveis. A identificação dos picos de lactose das amostras foi realizada por meio da comparação com o tempo de retenção do padrão analítico. De acordo com os autores, o método mostrou-se ser simples, com ótimos parâmetros de precisão, exatidão, repetibilidade e limites de quantificação e detecção que corroboram a viabilidade para análise de rotina.

Paula *et al* (2023) validaram um método de quantificação de lactose em leite e queijos Brie com ácido 3,5-dinitrosalicílico (ADNS), um reagente que pode ser reduzido por açucares redutores, passando de um composto amarelo para um colorido avermelhado, com absorção máxima na região do visível em comprimento de onda na região de 546 nm. Os resultados mostraram ótimos limites de quantificação e detecção (0,001 e 0,0004 mg/mL, respectivamente). O método apresentou valores de recuperação de 103% para o leite e 68% para os queijos. Por fim, os autores chegaram à conclusão de que o método proposto é satisfatório e adequado para quantificar lactose, permitindo uma análise precisa e confiável dos teores de lactose em produtos lácteos.

Outras formas alternativas podem ser encontradas na literatura, como o uso de um glicosímetro[4] para determinar a concentração de lactose soro no leite e em leite fluido (CAMPOS *et al,* 2014). Neste método, o aparelho foi utilizado para construir curvas de calibração por meio de simulação de matriz, adicionando diferentes volumes de uma solução estoque de lactose de 20 g/L em padrões de soro de leite de mesmo volume. Após a leitura, realizou-se correção do teor original de lactose no soro mediante valores de referência obtidos pelo método de cloramina-T. Por fim, os resultados apresentados mostraram-se satisfatórios para determinação de açucares redutores nas matrizes estudadas.

Podem ser encontrados na literatura alguns trabalhos que fizeram o uso da espectroscopia NIR para análise de alimentos, como o estudo de Caramês *et al* (2021), que avaliou o potencial bioativo de polpas de açaí liofilizadas comparando o desempenho da NIR e de imagens de Smartphones acoplados a ferramentas quimiométricas. Silva (2023) investigou a possibilidade da utilização de métodos quimiométricos de PLS para determinar o percentual de leite de vaca utilizado como adulterante em leite de cabra, bem como a concentração de lipídeos por meio da NIR, com o uso de equipamento portátil.

Lima *et al* (2018) propuseram uma metodologia analítica para classificação de leites processados em temperatura ultra-alta regulares e insetos de lactose combinando espectroscopia no infravermelho próximo e métodos de classificação multivariada. Para tanto foram utilizados equipamentos portáteis e de bancada, análise discriminante linear combinada com seleção de variáveis pelos algoritmos de genético e das projeções sucessivas, bem como a PLS-DA.

2.3 ESPECTROSCOPIA NIR APLICADA À ANÁLISE DE IOGURTE

A espectroscopia NIR, aplicada com ferramentas quimiométricas adequadas, é uma poderosa técnica para realizar análise de qualidade e adulteração de alimentos líquidos, como bebidas alcóolicas, laticínios e óleos, além de possibilitar a detecção simultânea de diferentes atributos (WANG *et al*., 2017). Diante disso, o uso da NIR para análise de qualidade de alimentos foi alavancado nos últimos anos, como, por exemplo, para a classificação de café instantâneo com base na cafeína e no grau de torra (NÓBREGA *et al*, 2023), a classificação de resíduo de algodão de diferentes países (ZHOU *et al*, 2023) e detecção de qualidade de carne (ZHENG *et al*, 2023).

A espectroscopia NIR é uma técnica rápida e não destrutiva que tem sido implementada com sucesso para a autentificação de produtos lácteos. Contudo, existe uma baixa demanda de

[4] Aparelho portátil destinado a medir a concentração de glicose no sangue baseada na conversão da glicose em gliconolactona por meio da ação enzimática da glicose desidrogenase (CAMPOS *et al.,* 2014).

pesquisas voltadas para o uso da NIR como técnica de identificação de adulteração de iogurte quando comparada a outros produtos lácteos (TAVARES; MEDEIROS; BARBIN, 2022). Ainda assim, é possível encontrar trabalhos que utilizaram a espectroscopia NIR para análise de iogurte.

Texeira *et al* (2021) avaliaram o uso da NIR, combinada com as ferramentas quimiométricas PCA e PLS-DA, para detecção de adulteração de iogurtes e queijos de cabras com leite de vaca. Para os espectros de iogurte, foram observadas duas bandas predominantes entorno de 1450 e 1930 nm, que podem ser atribuídas a bandas de absorção de O-H da água, o que pode mascarar as variações de outros componentes químicos. Contudo, os modelos resultantes demonstraram eficiência em discernir as amostras em autênticas das adulteradas com 10, 15 e 20% de leite de vaca.

Muncan, Tei e Tsenkova (2021) propuseram o uso da espectroscopia NIR par controle de produção e estudar a fermentação do iogurte, a fim de descobrir quais características espectrais portam informação sobre o processo e obter uma melhor compreensão da coagulação do iogurte. Os espectros obtidos apresentaram bandas características de absorção de OH da água (975, 1450 e 1900 nm). A análise de componentes principais de janela móvel foi utilizada elucidar mudanças físicas e químicas no processo de fermentação.

Por sua vez, He *et al* (2005), investigaram o uso da Vis/NIR em combinação com mínimos quadrados parciais (PLS, do inglês: Partial Least Squares), como um método não destrutivo para determinar o teor de açúcar no iogurte. Para tanto, foram adquiridas 160 amostras de iogurtes de cinco marcas diferentes, as quais foram analisadas em espectro de campo portátil. Os resultados obtidos no estudo indicaram que o método era eficaz para determinar o teor de açúcar em amostras de iogurte de forma rápida e não destrutiva.

Embora existam trabalhos voltados à análise de iogurte com espectroscopia NIR, a literatura científica apresenta-se limitada quanto ao uso de análises multivariadas e NIR para a classificação de iogurtes em função da presença de lactose.

2.4 ESPECTROMETRIA DO INFRAVERMELHO PRÓXIMO

Em 1800, ao decompor a radiação solar com um prisma e utilizando termômetros de mercúrios, Herschel mostrou a existência de uma radiação invisível além da região do vermelho que transporta calor (SCHRADER *et al*., 1995). Denominada Espectroscopia de Infravermelho, essa radiação descoberta por Herschel trata-se de uma espectroscopia vibracional, que se divide em três regiões em relação a faixa espectral do visível, sendo elas: infravermelho próximo, médio e distante.

A Espectroscopia do Infravermelho Próximo (NIR, do inglês: Near InfraRed), usa a energia de fótons na faixa de 2,65 * 10^{-19} e 7,96 * 10^{-20} J, correspondendo ao intervalo de comprimento de onda de 750 a 2500 nm, baseando-se no princípio de que diferentes ligações químicas da matéria orgânica absorvem ou emitem luz em diferentes comprimentos de onda. Contudo, a energia transportada pelas ondas eletromagnéticas da radiação NIR é relativamente baixa e sua interação com a matéria raramente promove excitação de eletrônicas. Porém, é suficiente para promover moléculas para seus estados vibracionais excitados mais baixos (PASQUINI, 2003, 2018; ZAREEF *et al*., 2020).

Os métodos analíticos que utilizam a região da espetroscopia NIR apresentam os seus aspectos mais relevantes, como a rapidez não ser destrutiva e invasiva, com alta penetração de feixe de sondagem e preparação mínima da amostra. A aplicação pode ser realizada em qualquer molécula que contenha as ligações C-H, N-H, S-H ou OH (PASQUINI, 2003). A técnica da espectroscopia NIR envolve irradiação de luz sobre a amostra, resultando na vibração das ligações entre os átomos. Esses fenômenos provocam o estiramento e a flexão das ligações, ocasionado mudanças no comprimento e no ângulo da ligação (WALSH *et al.,* 2020; ZAHIR *et al.,* 2022).

Os átomos participantes das ligações químicas estão constantemente se deslocando em comparação com os outros, com uma frequência estabelecida pela força da ligação e pela massa dos átomos envolvidos. Dessa forma, de acordo com a distribuição de Boltzmann, em temperatura ambiente, a maioria das moléculas estão em seu estado vibracional menos energético, n=0. Sendo assim, as transições ocasionadas pela absorção da radiação ocorrem entre o n = 0 e o estado adjacente (n = 1), chamadas de transições fundamentais (BURNS; CIURCZAK, 2008; PASQUINI, 2003; SILVA, 2017).

O comportamento vibracional do sistema é geralmente representado pela mecânica quântica na forma de um oscilador harmônico. Esse modelo impõe a restrição de que as transições vibracionais só podem ocorrer entre níveis adjacentes, (0 para o 1, do 1 para o 2 e assim sucessivamente). Além disso, outra restrição imposta é que a diferença de energia entre os estados adjacentes seja sempre a mesma (PASQUINI, 2003).

Contudo, considerando essas restrições impostas pelo modelo harmônico quântico, os fenômenos observados das interações das radiações NIR com o sistema vibracional não deveriam existir, pois a diferença de energia necessária para excitação vibracional de um sistema harmônico molecular situa-se na faixa de frequência que corresponde a região do infravermelho médio (MIR, do inglês: Middle Infrared) (PASQUINI, 2018).

Os sistemas moleculares não funcionam de forma perfeitamente harmônica, o modelo oscilador harmônico falha em não levar em consideração o comportamento não ideal do sistema vibracional. A repulsão dos átomos ao se aproximarem em movimentos de compressão das ligações químicas e o afastamento dos átomos em movimentos de distensão das ligações, que tente ao enfraquecimento e ruptura das ligações químicas, induzem o sistema vibracional a um comportamento anarmônico (PASQUINI, 2018).

O modelo de oscilador anarmônico diferencia do harmônico por considerar as forças de repulsão e dissociação dos átomos para vibrações de grande amplitude. Em decorrência da possibilidade de ruptura das ligações químicas proveniente do grande afastamento dos átomos ligados, há uma diminuição gradativa da energia potencial. Dessa forma, os níveis energéticos vibracionais adjacentes não são equidistantes, sendo a diferença entre níveis adjacentes é menor à medida que são considerados níveis vibracionais mais altos (PASQUINI, 2018).

Outra divergência entre os dois modelos de osciladores é o fato de o anarmônico permitir transições que diferem por mais de um nível de energia vibracional. Assim, a energia vibracional da molécula pode transitar diretamente do nível fundamental para os níveis 2, 3 ou 4, denominadas de transições de sobretom. A transição entre nível 0 e 2 é chamada de primeiro sobreton, para o nível 3 de segundo sobreton e assim sucessivamente (PASQUINI, 2018; SILVA, 2017).

A restrição quântica determina que a energia da radiação eletromagnética deve ser igual à diferença entre os níveis mais e menos energéticos, assim a faixa de energia necessária para realizar transição direta do nível 0 para o 2 é aproximadamente o dobro das transições fundamentais. Portanto, as transições que necessitam de energias relativamente altas da radiação MIR para serem realizadas terão suas energias de sobretons associadas a uma radiação mais energética, localizada em uma faixa característica da radiação NIR. Assim, as transições de sobreton podem ser induzidas pela radiação NIR, podendo ser observadas bandas de absorção na região de comprimento de onda referende a essa radiação (PASQUINI, 2018).

A anarmonicidade do sistema possibilita outro fenômeno relevante para a espectroscopia NIR, a combinação de modos vibracionais. Ao contrário de moléculas diatômicas, onde apenas o modo vibracional de estiramento e contração é possível, moléculas compostas por mais de dois átomos apresentam diferentes modos vibracionais. As combinações podem existir mediante absorção de um fóton com frequência eletromagnética igual à soma das frequências necessárias para excitar cada um dos modos. No entanto, uma vez que as combinações são permitidas apenas pela anarmonicidade, as intensidades das interações são

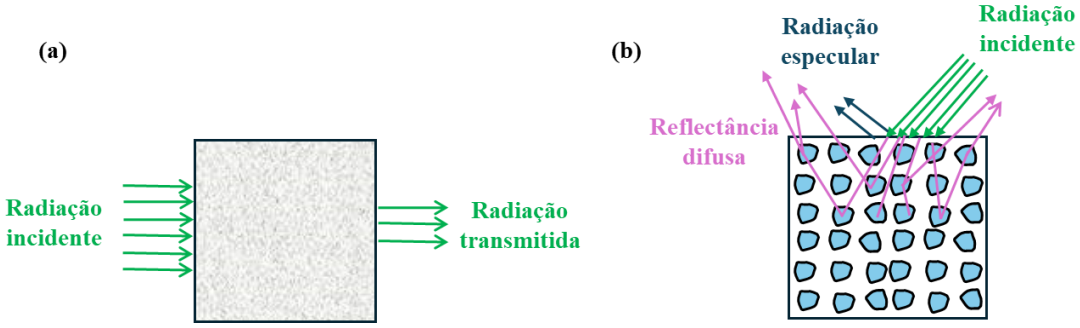
menores que as transições fundamentais registradas na região do MIR (PASQUINI, 2018; SILVA, 2017).

A região NIR contempla as bandas de absorção de sobretons e as bandas de combinação de modos vibracionais fundamentais de C-H, O-H e N-H. Sendo assim, a faixa do espectro eletromagnético referente à radiação NIR é subdividida em regiões de bandas de combinações e de sobretons. O conjunto de bandas de combinação de C-H manifestam-se entre 2000 e 2400 nm, enquanto o primeiro e segundo sobreton ocorrem, respectivamente, na faixa de 1600-1800 nm e 1100-1200 nm. No caso do grupo O-H, para álcoois e fenóis, as bandas de primeiro e segundo sobreton aparecem, respectivamente, em valores próximos a 1400 e 1000 nm. Por outro lado, a banda de combinação surge, aproximadamente, em 2000 nm. Para moléculas de água, a banda de combinação aparece em valores de "λ" em cerca 1940 nm, enquanto o primeiro sobreton surge em torno de 1440 nm (BOKOBZA, 1998).

Existem diversos modos de medição que podem ser aplicados à espectroscopia NIR, como a transmitância (**Figura 2.3a**), obtida da mesma forma como na espectroscopia ultravioleta-visível. Contudo, Karl Norris propôs que, ao invés de a medição espectral ser obtida de um sinal fraco de transmitância, seria possível obtê-la ao analisar o conteúdo de informação da porção de radiação refletida de forma difusa (**Figura 2.3b**) por amostras sólidas (PASQUINI, 2003).

As características físicas das amostras determinam se a radiação será transmitida, absorvida ou refletida. Em amostras com superfícies rugosas, coloridas ou em pó, a luz incidente é parcialmente espalhada e absorvida, provocando que a luz da radiação seja refletida em várias direções e ângulos diferentes do feixe incidente, gerando a reflectância difusa (BOLS *et al.,* 2024; ZAHIR *et al.,* 2022).

Figura 2.3: Medida de transmitância e reflectância difusa.



Fonte: Autoria própria (2024).

De acordo com os princípios da espectroscopia NIR e os espectros experimentais, observa-se a existência de interferentes (oriundos, por exemplo, de bandas de transições de modos vibracionais diferentes ou fontes de variação associadas ao tamanho e distribuição das partículas) que afetam a aquisição e seletividade do sinal analítico. Assim, os espectros NIR tornam-se susceptíveis à ocorrência de sobreposição ou alagamento de bandas e a variações (mudanças de linhas de base, por exemplo) não relacionadas às informações relevantes para análises qualitativas e quantitativas. Nesse contexto, a etapa de pré-processamento é crucial para a modelagem baseadas em dados NIR, pois minimiza as interferências indesejadas nos espectros obtidos. Isso possibilita a construção de modelos apropriados tanto em modelagem para calibração quanto para classificação multivariada (BI *et al*., 2016; ZAHIR *et al*., 2022).

2.5 ANÁLISE MULTIVARIADA DOS DADOS

Devido a grande quantidade de características registradas, as respostas espectrais obtidas por meio da radiação NIR apresentam uma grande quantidade de informações redundantes e ruidosas, além da ocorrência de sobreposição de bandas, dificultando a interpretação de informações úteis e a modelagem (CHEN *et al*., 2019; XIAOBO; 2010). Portanto, para a aplicação da NIR é necessário utilizar a análise multivariada de dados para desenvolver métodos analíticos capazes de extrair informações úteis dos espectros com alta complexidade oriunda de variações inerentes a natureza da amostra, sinais fracos ruídos instrumentais e espalhamentos (LIMA *et al*., 2018; SANTOS; PÁSCOA; LOPES; 2017).

O termo quimiometria foi usado pela primeira vez em 1971, pelo sueco e químico orgânico Svante Wold. Toda via, apenas em 1972 que o próprio Wold publicou o primeiro artigo na revista Kmisk Tidskrift que mencionava o termo (BRERETON *et al.,* 2017; FERREIRA, 2015).

A quimiometria surgiu a partir da carência de ferramentas matemáticas e estatísticas que pudessem suprir a necessidade de converter grandes quantidades de dados no máximo de informação útil possível, tarefa que a análise univariada não conseguia realizar. Os trabalhos pioneiros na quimiometria de Jurs, Kowalski, Isenhour e Reillye, em 1969, foram os responsáveis por ocasionar uma grande e inovadora mudança no campo de tratamento de dados químicos (FERREIRA, 2015).

As principais vertentes da quimiometria se concentram na utilização de ferramentas matemáticas e estatísticas com o intuito de planejar ou otimizar procedimentos experimentais, extrair o máximo de informação química relevante e obter conhecimento sobre sistemas químicos (FERREIRA, 2015). Para esses fins, a quimiometria usufrui de diversos métodos, tais

como: análise exploratória dos dados; calibração multivariada; planejamento e otimização de experimentos; e, reconhecimento de padrões.

A quimiometria tem sido bastante utilizada como ferramenta matemática nas diversas áreas da química, desde a medicinal até a química ambiental. Grande parte desse avanço pode ser percebida na sua utilização para o desenvolvimento de métodos de classificação multivariada usados para triagem na área de química analítica, com foco na autentificação de alimentos, evidenciada pelo crescimento de publicações científicas (JIMÉNEZ-CARVELO; CUADROS-RODRÍGUEZ, 2020). É importante salientar que a escolha do método empregado deve ser condizente com as informações que se desejadas e o objetivo da pesquisa.

2.4.1 Pré-processamento dos Dados

Os sinais registrados pelos instrumentos são compostos por duas contribuições distintas: um sinal verdadeiro, que contém informações sistemáticas correspondentes às contribuições determinísticas do sinal, e uma contribuição estocástica atribuída a variações aleatórias indesejáveis, denominada ruído. Outro interferente que pode estar presente no sinal são fontes de variações de baixa frequência e informações sistemáticas que não estão relacionadas ao processo de interesse (FERREIRA, 2015).

Em alguns casos, para evitar influências nos resultados, os dados devem ser pré-processados visando reduzir variações indesejadas que componham o sinal obtido. Portanto, o pré-processamento ideal deve ser estabelecido conforme a naturalidade dos dados analíticos (FERREIRA, 2015; NÓBREGA, 2021).

O processo de pré-processamento de dados compreende a limpeza, normalização, transformação, extração de característica e seleção, de modo que o produto final é utilizado para a criação de modelos preditivos, podendo ter um impacto significativo no desempenho dos algoritmos supervisionados (STOLL *et al.,* 2020). O pré-processamento de dados é dividido em dois tipos, quando aplicado às amostras e às variáveis, representadas pelas linhas e colunas da matriz, respectivamente (FERREIRA, 2015).

Existem diversos tipos de pré-processamento que podem ser aplicadas às amostras, incluindo a suavização por Savitzky-Golay (SG), correção de linha de base offset (BO, do inglês: Baseline Offset), variação normal padrão (SNV, do inglês: Standard Normal Variation) e derivada com filtro de Savitzky-Golay. Em relação aos pré-processamentos destinados às variáveis, a centralização dos dados na média é o mais utilizado.

As variações aleatórias podem ser reduzidas por meio de técnicas de alisamento, com o objetivo de aumentar a relação sinal ruído. Em geral, essas ferramentas usam as intensidades

de um segmento (ou uma janela) do espectro para determinar uma única resposta, a qual será atribuída ao centro da janela. Em seguida, a janela é deslocada e o processo é repetido por todo o percurso do espectro. Quanto maior o tamanho da janela, maior será a redução do ruído, entretanto, janelas grandes podem modificar o espectro, removendo ou distorcendo picos (FERREIRA, 2015).

Uma das principais técnicas de suavização utilizadas é por filtro Savitzky-Golay. O método consiste na remoção do ponto central da janela, posteriormente, ajusta-se um polinômio por mínimos quadrados para os demais pontos. Dessa forma, estima-se o valor do ponto central removido por meio do polinômio. Por fim, o intervalo é deslocado para o ponto seguinte e o processo é repetido (CERQUEIRA, 2000).

Por sua vez, as variações sistemáticas, intrínsecas de problemas instrumentais ou de amostragem, podem ser removidas ou minimizadas por meio da utilização dos métodos de derivadas de Savitzky-Golay, SNV e offset.

Um obstáculo que afeta significativamente espectros NIR é o desvio da linha de base, seja ele positivo ou negativo. Esses desvios podem ser lineares, quando provocados pela presença de um sinal de fundo (offset) devido variações instrumentais ou compensação inadequada do branco, ou não-lineares quando surgem por efeitos multiplicativos. Em geral, os desvios de linha de base podem ser corrigidos pela compensação do offset, derivadas ou métodos de correção de espalhamento multiplicativo (SENA; ALMEIDA, 2018).

As derivadas são excelentes ferramentas matemáticas para aprimorar a resolução dos sinais analíticos, mas aumentam o ruído, portanto, devem ser empregadas em combinação com técnicas de suavização. Desta forma, a primeira derivada elimina desvios aditivos da linha de base, corrigindo o deslocamento constante (offset), enquanto a segunda derivada elimina efeitos multiplicativos, os quais são variações sistemáticas que promovem a inclinação da linha de base (FERREIRA, 2015; SENA; ALMEIDA, 2018).

O método SNV é uma ferramenta de correção de efeitos de espalhamento aditivos e multiplicativos, resultantes de fenômenos físicos, tais como: mudanças no caminho óptico; sensibilidade do detector; variações na temperatura e na pressão; e diferenças no tamanho e nas formas das partículas. O SNV é uma espécie de normalização na qual cada espectro é centralizado e escalonado pelo desvio padrão correspondente, com o objetivo de deixar todos os espectros na mesma escala, além de ser robusto em relação à presença de amostras anômalas (FERREIRA, 2015; SENA; ALMEIDA, 2018; BI *et al.,* 2016).

Centrar na média é o pré-processamento mais comumente utilizado e pode ser aplicado a qualquer tipo de dados, inclusive aos dados espectrais. Neste método, a média de cada variável

é calculada para todas as amostras, ou seja, uma média para cada vetor coluna é calculada. Em seguida, é subtraído o valor da média de cada um dos valores da coluna correspondente. Dessarte, cada variável terá uma média zero, deslocando as coordenadas para o centro dos dados, salientando diferenças nas intensidades das variáveis (SENA; ALMEIDA, 2018).

2.4.2 Técnicas de Reconhecimento de Padrões

De acordo com Brereton (2015), a maioria das definições modernas de reconhecimento de padrões (RP) envolvem principalmente a classificação, ou seja, o processo de examinar as relações entre as amostras e variáveis, para atribuir objetos em grupos (classes). Destarte, os métodos de classificação podem ser divididos em dois tipos: supervisionados e não supervisionados.

A abordagem de RP não supervisionado tenta dividir o espaço de dados em grupo sem qualquer conjunto de treinamento, com o objetivo de verificar a existência de agrupamentos naturais das amostras. Enquanto o RP supervisionados usa um conjunto de treinamento para tentar dividir os objetos em grupos de acordo com suas características, dessa forma prevendo a que classe pertence uma amostra desconhecida (BRERETON, 2015; SENA; ALMEIDA, 2018).

A utilização de técnicas RP está bastante difundida na literatura, sendo fundamentais na maioria das áreas de aplicação da química analítica, como: estudos ambientais, ciências alimentares, análises biomédicas, clínicas, aplicações forenses e controle industrial (OLIVERI *et al*., 2021). Alguns trabalhos trazem a aplicação das técnicas de RP, tais como: autentificação de gengibre (TABBASSUM; ZEESHAN; LOW, 2022), discriminação de pele de animal (XU *et al*., 2022), e adulteração de polpa de goiaba (ALAMAR *et al.,* 2020).

Destarte, as técnicas de RP não supervisionadas mais utilizadas na literatura são: análise de componentes principais (PCA do inglês: Principal Component Analysis) e análise por agrupamento hierárquico (HCA, do inglês: Hierarchical Cluster Analysis). Já para os supervisionados, são: Modelagem Independente e Flexível por Analogia de Classe (SIMCA, do inglês: Soft Independent Modeling of Class Analogy), Análise Discriminante Linear (LDA, do inglês: Linear Discriminant Analysis), Regressão por Mínimos Quadrados Parciais para Análise Discriminante (PLS-DA, do inglês: Partial Least Squares for Discriminant Analysis).

2.4.3 Análises por Componentes Principais

PCA é o método multivariado mais conhecido e utilizado, sua finalidade é projetar e comprimir os dados em um espaço de dimensão menor, dessa forma, reduzindo a dimensionalidade do conjunto de dados, sem prejudicar a relação entre as amostras. É uma

ferramenta de análise exploratória, visto que auxilia na elaboração de hipóteses gerias a partir dos dados coletados. Essa metodologia distingue as informações relevantes das redundantes e aleatórias, permitindo identificar, visualizar e interpretar as diferenças existentes entre as variáveis e examinar as relações que podem existir entre as amostras (FERREIRA, 2015, 2022; SOUSA *et al.,* 2019).

A PCA reduz a dimensionalidade ao transformar um conjunto de dados com variáveis correlacionadas em um novo sistema de eixos não correlacionados, denominados componentes principais, obtidos por meio da combinação linear das variáveis originais. O modelo PCA é calculado de modo que a maior porção da variância seja explicada na primeira PC e, consequentemente, as porções menores de variância são explicadas de forma decrescente pelos componentes seguintes (BRERETON, 2017; SOUZA; POPPI, 2012).

No modelo PCA, a matriz **X** é decomposta em um produto de duas matrizes, denominadas de scores (**T**), que corresponde às coordenadas das amostras no sistema de eixos PC, e de loading ou pesos (**P**), que representa os coeficientes de combinação linear que determina quanto cada variável contribui para gerar os eixos de PC. Além disso, há um acréscimo de uma matriz de erro (**E**), também chamada de matriz de resíduo, que contém a variância não explicada pela PCA, sendo considerada nula quando não há compreensão dos dados nela contidos (**Equação 2.1**) (FERREIRA, 2022; SOUSA; POPPI, 2012).

$$\boldsymbol{X = TP^T + E} \qquad \textbf{(2. 1)}$$

A primeira PC é traçada no sentido da maior variância no conjunto de dados, enquanto PC2 é traçada ortogonalmente à primeira, com o propósito de abranger a maior parte da variância não explicada por PC1. O processo é repetido até que não haja mais compreensão dos dados do resíduo. Os resultados de um modelo PCA, ou seja, da decomposição da matriz X, podem ser interpretados de forma gráfica, pois possíveis agrupamentos podem ser revelados nos gráficos dos scores e a análise conjunta dos gráficos dos loading e scores permite determinar quais variáveis são responsáveis pelas diferenças entre as amostras (SOUSA; POPPI, 2012; VALDERRAMA *et al.,* 2016).

2.4.4 Modelos de Classificação

Problemas de classificação são recorrentes em Química Analítica, sendo utilizadas as diferentes abordagens, a exemplo, da linguagem de máquina e estatística para distinguir objetos com base nos dados obtidos em diferentes tipos de sinais instrumentais. No campo da quimiometria, as soluções de problemas de classificação multivariada são, frequentemente, uma abordagem que se baseiam em métodos que se enquadram em duas categorias: os baseados em

análise discriminante e os que se baseiam no princípio da modelagem de classes (VITALE *et al.,* 2023).

A análise discriminante analisa as diferenças entre as classes, delimitando uma fronteira global, dividindo o espaço multivariado das variáveis em sub-regiões correspondente à quantidade de classes atribuídas ao conjunto de treinamento. Por fim, sempre atribui um objeto (amostra) a apenas uma classe a qual ele melhor se adeque (CHEN; HARRINGTON, 2019; VITALE *et al.,* 2023). Os métodos discriminantes necessitam de mais de uma classe para construção de modelos e determinação dos limites de classificação de um novo objeto (MANUEL *et al.,* 2022).

Por sua vez, a modelagem de classes, também conhecida como classificação de classe única, estima independentemente um contorno individual para cada classe em estudo, de acordo com a similaridade dos seus objetos. Dessa forma, definindo uma área específica em torno de cada classe para determinar se um objeto pertence a categoria de destino. No entanto, se mais de uma classe for modelada, ao contrário da análise discriminante, as amostras podem ser classificadas como membros de nenhuma, uma ou múltiplas classes (CHEN; HARRINGTON, 2019; VITALE *et al.,* 2023).

2.4.4.1 Modelagem Independente e Flexível por Analogia de Classe (SIMCA)

A Modelagem Independente e Flexível por Analogia de Classe foi a primeira das técnicas de modelagem de classe a ser introduzida na quimiometria, por Svante Wold EM 1977 (OLIVERI *et al*., 2021). Sua utilização é bastante difundida para resolver problemas de classificação de uma classe, tais como classificação de sais comestíveis (LEE; HAN; NAM*,* 2017) e autentificação de cigarros quanto a fração de tabaco (ZELINKOVA; WENZL, 2020).

Conforme a literatura, SIMCA é um método totalmente baseado em dados, sem qualquer suposição a priori sobre a distribuição dos dados coletados, na qual cada classe é tratada separadamente. Trata-se de um método concentrado nas semelhanças das amostras pertencentes à classe individualmente investigada. Dessa forma, o SIMCA assume a premissa de poder captar a informação sistemática associada a essa semelhança por meio de uma representação em componentes principais dos dados para cada classe modelada individualmente, de maneira que novas observações, que por ventura lhe pertencem, podem ser avaliadas por meio de medidas estatísticas estimadas a partir da representação reduzida pelas PC's (VITALE *et al.,* 2023).

Portanto, SIMCA constrói modelos de classe com base nos scores da componente principal. O modelo PCA é calculado para a classe de interesse, caso haja mais de uma classe

de interesse, o procedimento é repetido, individualmente, para as demais classes. Sendo assim, cada modelo de classe compreende um número significativo de PC's. Dessa forma, determinar o número adequado de PC é uma etapa crucial para a construção dos modelos SIMCA, pois uma quantidade insuficiente pode excluir informações úteis para o modelo e prejudicar a seletividade. Por outro lado, a adição de muitas PC's agregará ruído ao modelo, o que resultará em um ajuste excessivo (superajuste) (CHEN; HARRINGTON, 2019; OLIVERI *et al*., 2021).

Sendo um método supervisionado, SIMCA define uma região de aceitação em torno do modelo PCA de cada classe. Esta região é definida por um limite de decisão, sendo utilizada para aceitar ou rejeitar novas medições. Um limite de decisão adequado é crucial para a capacidade do modelo de aceitar novos objetos pertencentes à classe de interesse e de rejeitar objetos que não sejam da classe de interesse. Dessa forma, a eficiência dos modelos SIMCA está diretamente relacionada à seleção adequada do número de PC's e ao limite de decisão (CHEN; HARRINGTON, 2019).

A distância de todos os objetos que compõe o conjunto de treinamento do modelo é calculada e seu desvio padrão é utilizado para definir a região de aceitação da classe em estudo, em determinado limite de confiança, conforme a estatística F de Fisher. Quando uma nova observação desconhecida é projetada, sua classificação final é dada por meio de um teste F, comparando a variação residual da observação com a variação residual de cada classe (OLIVERI *et al.,* 2021; ZELINKOVA; WENZL, 2020).

2.4.4.2 Análise Discriminante Linear (LDA)

Análise Discriminante Linear é um método de redução de dimensionalidade supervisionada que busca realizar uma combinação linear de características que proporcionem a maximização da separação entre as classes, ao mesmo tempo que minimiza a dispersão intraclasse. A premissa do LDA é projetar as amostras em um espaço vetorial discriminante, para obter o melhor efeito de extração de informações de classificação. Dessa forma, as amostras pertencentes a mesma classe irão se localizar o mais próxima possível no espaço projetando, enquanto as amostras de classes diferentes estarão o mais afastado possível (REZENDE; LOPES FILHO; VIEIRA, 2019; ZHU *et al.,* 2022).

Proposta por Fisher, o LDA é comumente utilizado para distinguir as observações entre classes e construir uma função discriminante capaz de realizar a classificação de observações futuras com classes desconhecidas (GARDNER-LUBBE, 2021). Em um problema de classificação binária, os dados multivariados são projetados em uma única função discriminante

que fornecerá os scores discriminantes para as classes 1 e 2, de modo a separar os grupos transformados o máximo possível (GONÇALVES, 2022; SOUZA, 2023).

Tendo obtido os vetores preditores, a classificação de novos objetos é dada pela proximidade da classe mais próxima com base em um limiar de classificação. Esse limiar pode ser estabelecido como sendo a metade da distância entre o centro das médias dos grupos. Dessa forma, uma amostra desconhecida é classificada por meio da comparação do seu score discriminante com o limiar de classificação (SILVA *et* al., 2021; SOUZA, 2023).

Dados espectrométricos apresentam, tipicamente, uma alta dimensionalidade e multicolinearidade, motivando o uso da LDA, tornando-se interessante para lidar com problemas de classificação de dados de natureza multivariada. No entanto, o princípio da LDA requer um condicionamento na modelagem, segundo o qual o número de objetos do conjunto de treinamento deve ser maior que o número de variáveis. Portanto, variáveis não informativas e/ou redundantes devem ser descartadas (SOUZA *et al.,* 2023).

A capacidade de generalização dos modelos LDA são frequentemente prejudicadas pela presença de colinearidade entre as variáveis, portanto o emprego de um algoritmo de seleção de variáveis é importante, com o objetivo reduzir a dimensionalidade e o ruído dos conjuntos de dados de classificação (PONTES *et al.,* 2020). Alguns estudos utilizaram técnicas de seleção de variáveis acopladas ao LDA para classificação multivariada, como o algoritmo genético (MORAIS; LIMA; MARTINS, 2019), o algoritmo das projeções sucessivas (FERNANDES, *et al.,* 2019) e o algoritmo dos morcegos (SOUZA *et al*., 2023).

2.4.5 Seleção de variáveis

2.4.5.1 Algoritmo das Projeções Sucessivas

Originalmente proposto para calibração multivariada (ARAÚJO *et al.*, 2001), o algoritmo das projeções sucessivas (SPA, do inglês: successive projections algorithm) foi adaptado para problemas de classificação e combinado com LDA por Pontes *et al* (2005). De acordo com a literatura, o SPA é uma técnica de seleção de variáveis que permite a redução de colinearidades na modelagem, baseando-se na escolha de grupos de variáveis que tenham máxima projeção entre si e que tenha o menor risco de classificação incorreta no conjunto de validação (CHEN; TAN; LIN, 2020; PONTES *et al.,* 2005; SILVA *et al.,* 2021).

No SPA, os vetores colunas, que corresponde as variáveis, são submetidos a sucessivas operações de projeção, resultando na criação de cadeias de variáveis. O procedimento do SPA consiste em iniciar uma cadeira com uma variável e acrescentar progressivamente variáveis que

apresentem menor colinearidade com as anteriores. O tamanho de cada cadeia em problemas de classificação é delimitado por N – C, onde N é o número de amostras de treinamento e C o número de classes envolvidas (CHEN; TAN; LIN, 2020; SOARES *et al.,* 2013).

A segunda fase do SPA consiste em avaliar os subconjuntos de variáveis por meio de uma função de custo relacionada ao risco médio de classificação incorreta do conjunto de validação – que é obtida com base na distância das amostras de validação, isto é, entre as classes verdadeiras e erradas. De tal modo que, um pequeno valor da função de custo indica que os objetos de validação estão próximos do centro de sua classe verdadeira e distantes dos centros das outras classes (CHEN; TAN; LIN, 2020; SOARES *et al.,* 2013).

2.4.5.2 Algoritmo Genético

Os algoritmos genéticos (GA, do inglês: genétic algorithm) são heurísticos e foram introduzidos no início da década de 1970 como uma abordagem de otimização baseados na seleção natural, segundo as regras clássicas do processo evolutivo de Charles Darwin. Tem sido amplamente utilizado para solução de problemas matemáticos, no qual se buscam informações que possibilitem encontrar o melhor desempenho dentro do espaço de busca (ANZANELLO *et al.,* 2017; NIAZI; LEARDI, 2012; PIRES, 2023).

O GA constrói cromossomos artificiais de maneira análoga ao cromossomo biológico, que é composto por genes. O algoritmo codifica os parâmetros do problema em genes que compõe os cromossomos artificiais (também chamados de indivíduos) (WEI *et al.,* 2020). Por meio da reprodução, cruzamento e mutação, o GA evolui as variáveis codificadas e produz descendentes (novas gerações) com maior nível de otimização (LIU *et al.,* 2018).

O procedimento de execução do GA consiste em codificar as informações, usando linguagem binária, em cromossomos, com cada variável correspondendo a um gene. De tal maneira que 1 significa que a amostra foi selecionada e 0 que não foi selecionada. Assim, é criada de forma randômica uma população inicial com um número determinado de indivíduos, de tal forma que todas as variáveis estejam presentes nos cromossomos. Cada indivíduo é submetido a uma avaliação e atribuído um valor associado à sua performance ao sistema (COSTA FILHO; POPPI, 1999; NIAZI; LEARDI, 2012).

Em seguida, uma nova população é criada por meio do cruzamento aleatório dos cromossomos, produzindo descendentes que recebem as informações dos seus progenitores, existindo uma tendência de transmissão das características dominantes. Uma mutação também pode ser aplicada, sendo a alteração do código genético de uma pequena parte da população (invertendo a codificação binária), direcionando a pesquisa para novas regiões do domínio

experimental, solucionando o problema de mínimos locais da otimização (COSTA FILHO; POPPI, 1999; NIAZI; LEARDI, 2012).

As etapas de avaliação, cruzamento e mutação se repetem em um ciclo até ser atingido um critério de parada. Para o GA esse critério é normalmente estabelecido como atingir o número máximo de gerações ou obter um erro mínimo desejado (COSTA FILHO; POPPI, 1999).

2.4.5.3 Algoritmo dos Morcegos

Proposto por Yang (2010), para solucionar problemas de otimização, e adaptado por Souza *et al.* (2023), para seleção de variáveis em modelagem de classificação usando LDA, o algoritmo dos morcegos (BA, do inglês: Bat Algorithm) foi concebido para simular o mecanismo de ecolocalização dos morcegos, que basicamente é a emissão de um pulso sonoro seguida de uma recepção de um eco de retorno dos objetos ao seu redor. Os pulsos ultrassônicos possuem uma frequência constante. Contundo, podem variar o comprimento de onda, diminuir a amplitude do som e aumentar a taxa de emissão de pulso de 10 a $20s^{-1}$ para 200 pulsos por segundo quando mudam de procurando presas para aproximando-se, a fim de aumentar a precisão (YANG, 2010).

O BA utiliza a frequência dos pulsos sonoros emitidos durante a movimentação dos morcegos para procurar as melhores soluções. No âmbito do algoritmo, essas soluções são as posições dos morcegos virtuais (x_i), que para problemas de seleção de variáveis, essas posições correspondem as variáveis selecionadas. Para tal, é criada uma população inicial de morcegos virtuais através de uma matriz de posições gerada aleatoriamente com valores de 0 a 1. Dessa forma, para problema binário de seleção de variáveis, valores acima de 0,5 são codificados como 1, representando uma variável selecionada, enquanto valores abaixo de 0,5 são codificados como 0, representando uma variável não selecionada. Assim, delimitando um subconjunto de variáveis (posições) para cada morcegos virtuais. Também é determinado de forma aleatória a taxa de emissão de pulsos (r_i), frequência (f_i) e velocidade (v_i) para cada indivíduo (SOUZA, 2023; SOUZA *et al.,* 2023).

Posteriormente, a aptidão de cada indivíduo é avaliada e armazenada, por meio de uma função de custo (G_{cost}) de risco médio de classificação incorreta, de tal modo que é atribuída a posição $x*$ para o melhor morcego, determinado pelo menor valor de G_{cost}. Em seguida, as posições dos morcegos são atualizadas por meio de alterações em suas velocidades e frequências, conforme as **Equações 2.2, 2.3** e **2.4**. As atualizações só acontecem se houver um aumento na taxa de emissão de pulsos, pois é indicativo do aumento na precisão de buscas dos

morcegos. Em caso da taxa de emissão de pulsos for menor que o ruído aleatório, significa que o morcego está distante da melhor posição (x_*). Dessa forma, realiza-se uma busca local para atualizar a posição do morcego através de um pequeno deslocamento aleatório da posição atual, de acordo com a **Equação 2.5** (SOUZA, 2023; SOUZA *et al.,* 2023).

$$fi = fmin + (fmax - fmin)\beta \quad \textbf{(2.2)}$$

$$\boldsymbol{v_i^t = v_i^{t-1} + (x^t - x_*)f_i} \quad \textbf{(2.3)}$$

$$\boldsymbol{x_i^t = x_i^{t-1} + x_i^t} \quad \textbf{(2.4)}$$

$$x_{novo} = x_{atual} + \varepsilon A^t \quad \textbf{(2.5)}$$

onde f_{mim} e f_{max} corresponde aos limites inferiores e superiores do espaço de busca, no qual o β corresponde a um valor escalar de uma distribuição normal gerado para cara iteração, *t* equivale ao número interações. Por sua vez, o parâmetro "ε" é um número aleatório, que pode assumir valores entre 0 e 1, e A^t corresponde a amplitude média dos pulsos emitidos pelos morcegos virtual para a atual iteração (SOUZA, 2023; SOUZA *et al.,* 2023; YANG, 2010).

Assim, as novas soluções são avaliadas, se a função de custo da nova iteração for menor que a anterior e a amplitude do pulso (A_i) for maior que uma função randômica, as novas posições são aceitas e são atualizadas a taxa de emissão e a amplitude do pulso. As atualizações das posições dos morcegos são repetidas para todos os morcegos até atingir o número de iterações estabelecidas. Por fim, determina o morcego que alcançou a melhor solução ($\boldsymbol{x_*}$), ou seja, ao menor valor de G_{cost} e utiliza as variáveis selecionadas para construção dos modelos LDA (SOUZA, 2023; SOUZA *et al.,* 2023).

2.4.6 Métricas de Desempenho dos Modelos

Conforme Diego e colaboradores (2022), as métricas de desempenho são usadas para avaliar se o modelo atende aos requisitos da classificação, sendo encontrado na literatura diversas métricas de desempenho de modelos em problemas de classificação, majoritariamente destinadas à classificação binária, podendo ser estendidas para multiclasses. Contudo, segundo os autores, não há uma métrica de desempenho que possa ser usada de forma geral, logo, deve-se estabelecer uma métrica que seja adequada ao domínio do problema e aos requisitos.

Em geral, as informações relativas à performance de um modelo de classificação são organizadas em uma matriz de confusão, construída a partir da comparação das classes observadas e preditas, codificando o número de predições corretas e incorretas de cada classe e

contendo todos os dados necessários para calcular as métricas de desempenho. Também conhecida como tabela de contingência, a matriz de confusão é uma matriz quadrada, cuja linhas e colunas representam classe real e preditas, respectivamente (BALLABIO; GRISONI; TODESCHINI, 2018; DIEGO *et al.,* 2022). A **Figura 2.4** apresenta uma forma genérica da matriz de confusão.

Partindo de um pressuposto conjunto de dados com n amostras e G classes, ng representa o número de amostras da g-ésima classe, enquanto n'_g expressa o número de amostras predita como pertencentes a g-ésima classe. O elemento C representa o número de objetos da classe g classificados como pertencente a classe k (g,k = 1, 2 ... G). Portanto, os elementos presentes na diagonal C_{gg} representam o número de objetos corretamente classificados, enquanto os demais elementos são os objetos erroneamente classificados (BALLABIO; GRISONI; TODESCHINI, 2018).

Figura 2.4: Matriz de confusão.

		Classe predita				
		1	2	...	G	
Classe real	1	C_{11}	C_{12}	...	C_{1G}	n_1
	2	C_{21}	C_{22}	...	C_{2G}	n_2
	...	...	...	...	...	...
	G	C_{G1}	C_{G2}	...	C_{GG}	n_G
		n'_1	n'_2	...	n'_G	

Fonte: Adaptada de Ballabio, Grisoni e Todeschini (2018).

A matriz de confusão contém as informações das distribuições das amostras dentro das classes, como por exemplo o número de amostras de cada classe equivale a somatória dos elementos da respectiva linha, bem como, o número de amostras preditas em uma data classe totaliza a soma dos elementos da coluna da classe predita (BALLABIO; GRISONI; TODESCHINI, 2018).

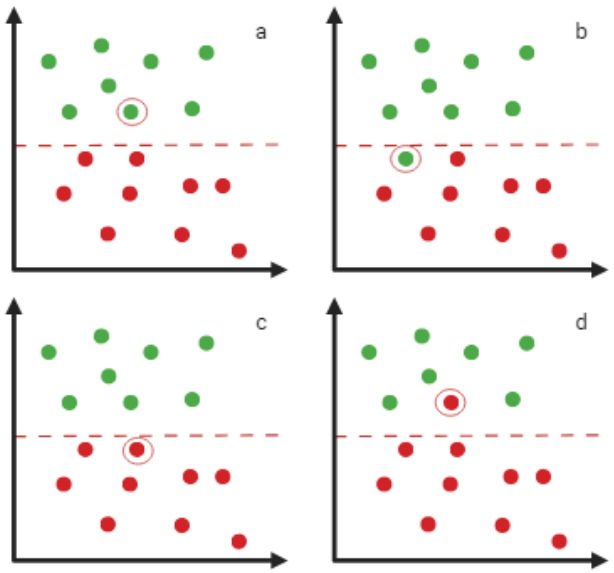
Por intermédio dessa matriz, é possível obter diferentes índices de desempenho, tais como: sensibilidade, especificidade e precisão. Essas métricas descrevem os resultados de classificação alcançados por cada classe e não devem ser usadas de forma individual, pois não consideram todas as informações do classificador (BALLABIO; GRISONI; TODESCHINI, 2018; DIEGO *et al.,* 2022).

As métricas de desempenho global, como a acurácia, são estimativas otimistas da capacidade do classificador, expressando a avaliação dos modelos em um único valor numérico, mas se comportam de forma diferente quando aplicadas a dados desequilibrados, que são

aqueles com número de objetos significativamente diferente entre as classes. Modelos de classificação podem ser categorizados como ótimos em medidas de classificação global, porém diferentes avaliações podem ser obtidas para o mesmo modelo quando empregadas outras medidas de desempenho (BALLABIO; GRISONI; TODESCHINI, 2018; DIEGO *et al.,* 2022).

É estabelecida uma codificação para rotular a forma como as amostras são preditas. Amostras da classe alvo podem ser classificadas como verdadeiro positivo (VP), quando são devidamente classificadas, e falso negativo (FN), quando são atribuídas a outra classe. As amostras pertencentes a classes não alvo são rotuladas como verdadeiro negativo (VN), classificação correta das amostras não alvo, e falso positivo (FP), que se refere às amostras classificadas incorretamente como sendo da classe alvo (SANTANA *et al*., 2020). A **Figura 2.5** ilustra a funcionalidade destes parâmetros.

Figura 2.5: Parâmetros de classificação das amostras para classe alvo (verde), tomando como base as amostras destacadas.



Fonte: Autoria própria (2024).
(a) Verdadeiro positivo, amostra classificada corretamente como sendo da classe alvo. (b) Falso negativo, amostra da classe alvo classificada erroneamente. (c) Verdadeiro negativo, amostra da classe estranha classificada como classe alvo. (d) Falso positivo, amostra da classe falsa classificada como sendo da classe alvo.

A sensibilidade é definida como a capacidade de um modelo de identificar corretamente as amostras de uma classe específica, determinada pela razão entre as amostras da classe corretamente classificadas (VP) e o número total de amostras pertencentes a classe. Por sua vez, a especificidade expressa a aptidão do classificador de rejeitar amostras de outra classe. É obtida pela razão entre amostras previstas corretamente como não pertencentes a uma determinada classe (VN) sobre o número total de amostras não pertencentes a essa classe. A sensibilidade e

especificidade são calculadas conforme as **Equações 2.6 e 2.7**, respectivamente (BALLABIO; GRISONI; TODESCHINI, 2018).

$$\boldsymbol{Sensibiidade} = \frac{\boldsymbol{VP}}{(\boldsymbol{VP+FN})} \quad \textbf{(2.6)}$$

$$\boldsymbol{Especificidade} = \frac{\boldsymbol{VN}}{(\boldsymbol{VN+FP})} \quad \textbf{(2.7)}$$

A precisão é estabelecida como a capacidade do modelo de evitar predições erradas em uma determinada classe, estabelecida como uma razão entre as amostras VP sobre o número total de amostras preditas na classe (BALLABIO; GRISONI; TODESCHINI, 2018). Acurácia, também conhecida como taxa e classificações corretas (TCC), é um parâmetro global que fornece um valor de desempenho para todo o modelo de forma a não considerar a informação da performance individual. Seu cálculo é realizado pela divisão do número total de amostras classificadas corretamente pelo número total de amostras do conjunto. A precisão e acurácia são calculadas conforme as **Equações 2.8 e 2.9**, respectivamente (SANTANA *et al*., 2020).

$$\boldsymbol{Precisão} = \frac{\boldsymbol{VP}}{(\boldsymbol{VP+FP})} \quad \textbf{(2. 8)}$$

$$\boldsymbol{TCC} = \frac{\boldsymbol{VP+VN}}{(\boldsymbol{VP+FN+VN+FP})} \quad \textbf{(2. 9)}$$

Capítulo 3

METODOLOGIA

3 METODOLOGIA

3.1 AQUISIÇÃO DAS AMOSTRAS

Foram adquiridas 102 amostras de diferentes sabores em mercados e supermercados da cidade de João Pessoa, sendo um total de 47 amostras com lactose (convencional) e 55 amostras isentas de lactose, de três marcas e lotes diferentes, a fim de assegurar a variabilidade da matriz. As marcas selecionadas possuíam as linhas de iogurte isenta de lactose e convencional. As duas classes de iogurte foram categorizadas utilizando IL (isento de lactose) e CL (com lactose). As amostras foram mantidas refrigeradas e lacradas até o ato da análise, a qual foi realizada sem nenhum pré-tratamento prévio.

3.2 AQUISIÇÃO DOS ESPECTROS NIR

As medidas espectrométricas foram realizadas na faixa de 900 a 1700 nm, utilizando-se de um equipamento portátil (NIRScasn InnoSpectra, modelo é NIR-S-G1) operando no modo de reflectância difusa. As leituras foram obtidas com resolução digital de 160 pontos, delimitada usando uma largura padrão de 10 nm, oversampling de 2 e 32 varreduras, e transformada de Hadamard. Os parâmetros foram empregados conforme indicação do fabricante. O padrão de politetrafluoretileno foi utilizado como branco.

Figura 3.1: Foto de leitura da amostra de iogurte com nanoNIR

Fonte: Autoria própria (2024).

As amostras de iogurte com e sem lactose foram homogeneizadas e dispostas em placas de Petri, com volume aproximado de 3 ml (**Figura 3.1**), sendo realizadas leituras em triplicatas. A aquisição dos espectros foi realizada em sala climatizada, com temperatura média de 23° C,

as amostras foram mantidas refrigeradas até o momento da análise. A água destilada foi usada para realizar a limpeza das placas entre as leituras.

3.3 PROCEDIMENTOS QUIMIOMÉTRICO

Diferentes técnicas de pré-processamento foram avaliadas com o objetivo de eliminar informações irrelevantes e melhorar o desempenho dos modelos de classificação. Dessarte, as técnicas empregadas neste estudo incluíram: Suavização com filtro Savitzky-Golay, SNV, 1° derivada de Savitzky-Golay e correção de linha de base por offset.

A análise exploratória para os dados brutos e pré-processados foi realizada empregando Análise de Componentes Principais. Os modelos de classificação foram construídos utilizando os métodos SIMCA e LDA com seleção de variáveis pelas técnicas GA, SPA e BA. O algoritmo de Kennard-Stone (KS) foi usado na seleção de amostras para compor os conjuntos de treinamento e teste.

Os pré-processamentos, bem como a modelagem SIMCA, foram realizados no The Unscrambler® X (CAMO S.A.), por sua vez foi implementado Matlab® 2010a (Mathworks, USA) para a construção dos modelos de classificação LDA. Os cálculos de LDA foram realizados utilizando o Toolbox LDA_VS_GUI. O pacote de LDA permitia realiza os métodos de seleção de variáveis SPA e GA, enquanto o BA foi realizado por meio do comando criado Souza *et al.* (2023), disponível em: https://www.ccen.ufpb.br/laqa/referencias/.

As amostras foram divididas em conjunto de treinamento e teste compostos por 70% e 30% das amostras de cada classe, consistindo, respectivamente, de 72 amostras (39 IL e 33 CL) e 30 amostras (16 IL e 14 CL). O procedimento de validação cruzada *leave-one-out* foi utilizada.

Para a modelagem SIMCA, foram construídos de forma individual modelos PCA para cada classes e foi utilizada a variância residual das amostras do conjunto de validação para determinar o número ideal de PC's. A distribuição de Fisher, com 95% de confiança (5% de significância), foi usada como limite de decisão para definir uma região de aceitação.

A análise discriminante linear foi empregada com seleção de variáveis pelo algoritmo dos morcegos com o objetivo de determinar as variáveis mais importante e reduzir a colinearidades existente entre elas. Os algoritmos SPA-LDA e GA-LDA foram utilizados como método de comparação com o BA-LDA. O processo de seleção de variáveis pelo BA-LDA foi realizado utilizando 30 morcegos virtuais com 500 iterações, os quais são estabelecidos com padrão no código do BA-LDA. As amostras de validação foram empregadas a uma função de custo, para selecionar o melhor conjunto de variáveis. O algoritmo foi executado em triplicata, de modo que foi selecionado o modelo com menor valor da função de custo.

Para seleção de variáveis por SPA, as amostras da validação foram utilizadas para selecionar o melhor subconjunto de comprimento de onda com base na função de custo do risco médio de classificação incorreta pelo LDA. Por fim, o conjunto de amostras de teste foi utilizado para verificar a capacidade de generalização do classificador (PONTES *et al.,* 2011). A seleção de variáveis pelo GA foi realizada com uma população de 100 indivíduos, em um total de 100 gerações. Os parâmetros de taxa de mutação e cruzamento foram estabelecidos em 5% e 60%, respectivamente, esses valores são indicados pelo pacote LDA. A rotina foi executada 3 vezes, sendo selecionados os modelos que obtiveram menor valor da função de custo.

Os desempenhos dos modelos foram avaliados conforme os parâmetros de sensibilidade, especificidade, precisão e TCC, seguindo as condições expressas na seção 3.6.6. A classe isenta foi tomada como classe alvo para os cálculos das métricas de desempenho dos modelos. Para os modelos LDA, os cálculos foram obtidos para os conjuntos de treinamento e de teste, contudo, para o SIMCA, apenas o conjunto de teste foi utilizado para o cálculo das métricas.

Capítulo 4

RESULTADOS E DISCUSSÃO

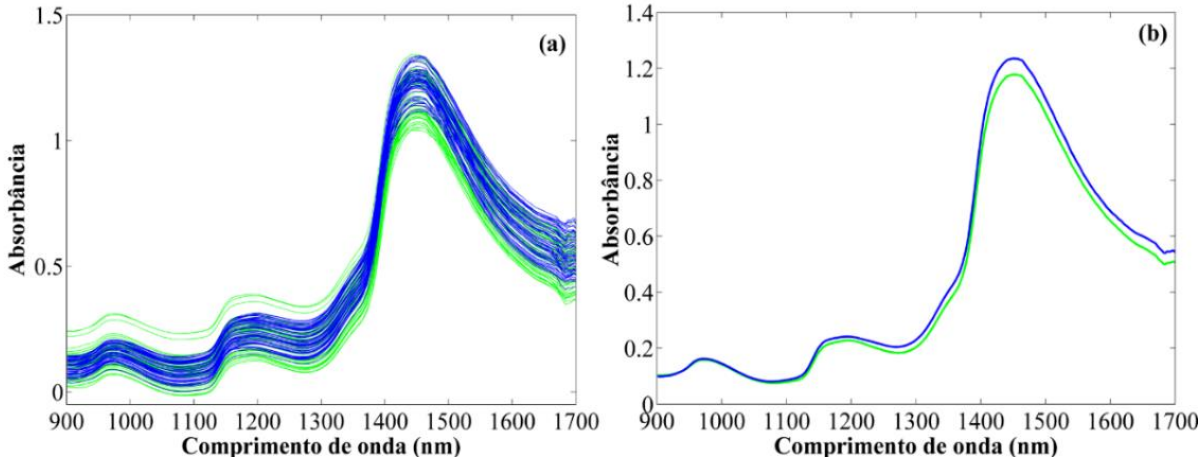
4 RESULTADOS E DISCUSSÃO

4.1 ESPECTROS NIR

Os espectros NIR das 102 amostras empregadas no estudo, obtidos na faixa de 900 a 1700 nm, são apresentados na **Figura 4.1a**. É possível observar uma banda predominante em torno de 1450 nm, correspondente ao primeiro sobreton de O-H. Outra banda, aproximadamente em 970, pode ser atribuída ao segundo sobreton de O-H. Apesar de estarem, predominantemente, associadas a bandas de água, é possível encontrar informações sobre a lactose e aos produtos da sua hidrólise. Isto, pois, de acordo com a literatura, essas regiões também podem ser atribuídas aos estiramentos de O-H de açúcares (DIAZ-OLIVARES *et al.,* 2024; GOLIC; WALSH; LAWSON, 2003; LIMA *et al.,* 2018; ROBERT; CADET, 1998).

Embora as bandas de O-H sejam dominadas pelo teor de água, ainda portam informações úteis para descriminar compostos lácteos quanto à presença ou ausência de lactose. Ademais, a presença de uma terceira banda, entre 1100 e 1200, pode ser crucial para a classificação, pois esta pode ser atribuída ao segundo sobreton de estiramento de C-H, assim associada a lactose e aos monossacarídeos resultantes da sua hidrólise (GOLIC; WALSH; LAWSON, 2003; LIMA *et al.,* 2018; MELFSEN; HARTUNG; HAEUSSERMANN, 2012). A **Figura 4.1b** apresenta o espectro médio das classes convencional (verde) e isenta de lactose (azul), sendo possível observar a semelhança no perfil das duas classes.

Figura 4.1: Espectros NIR da classe convencional (verde) e isenta (azul) (a) originais das 102 amostras e (b) espectros médios



Fonte: Autoria própria (2024).

4.2 PRÉ-PROCESSAMENTOS

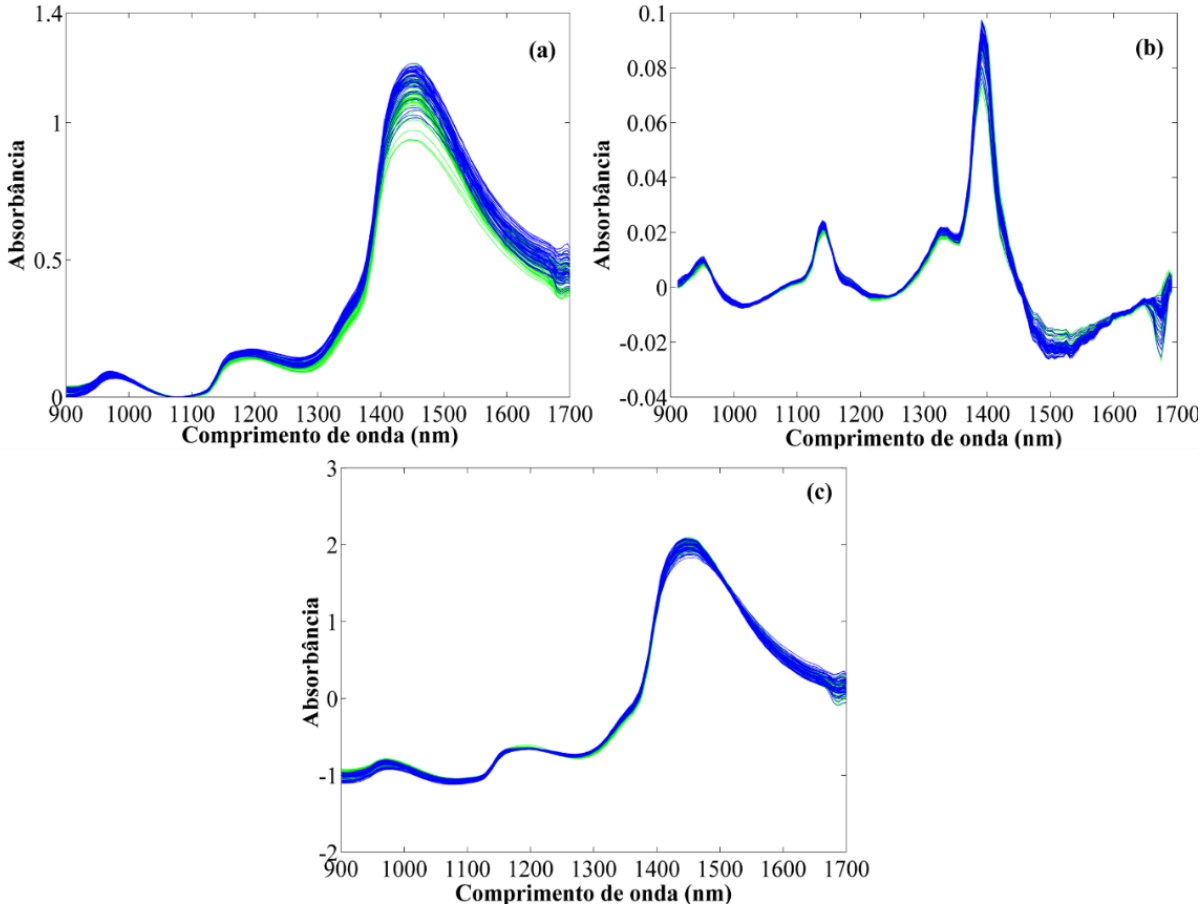
Ao analisar os espectros adquiridos (**Figura 4.1**), é possível notar um ruído instrumental, particularmente evidenciado após 1400 nm. Ademais, os espectros apresentam

uma variação sistemáticas na linha de base e efeito de espalhamento da radiação NIR. Assim, diferentes técnicas de pré-processamento foram aplicadas a fim de amenizar esses efeitos.

Para corrigir a variação aleatória, adotou-se um procedimento de suavização baseado no método Savitzky-Golay, no qual foram testadas janelas de tamanhos diferentes. No entanto, optou-se por utilizar uma janela de 5 pontos, uma vez que, com janelas de tamanho maior, houve a perda de informação. Após a aplicação da técnica foi observada uma redução do ruído. Os resultados obtidos foram submetidos aos pré-processamentos offset e SNV. Também foi aplicado o filtro derivativo Savitzky-Golay com primeira derivada, polinômio de segunda ordem e janela de 5 pontos foi realizada antes da suavização.

Para corrigir o efeito de deslocamento da linha de base, foram aplicadas técnicas de correção de linha de base por Offset e 1º derivada com filtro de SG com janela de 5 pontos. O método de SNV foi utilizado a fim de corrigir o espalhamento das amostras. Os espectros obtidos por meio desses procedimentos podem ser visualizados na **Figura 4.2**.

Figura 4.2: Espectros pré-processados das classes isenta (azul) e convencional (verde): (a) Offset, (b) derivação Savitzky-Golay com janela de 5 pontos, (c) SNV.



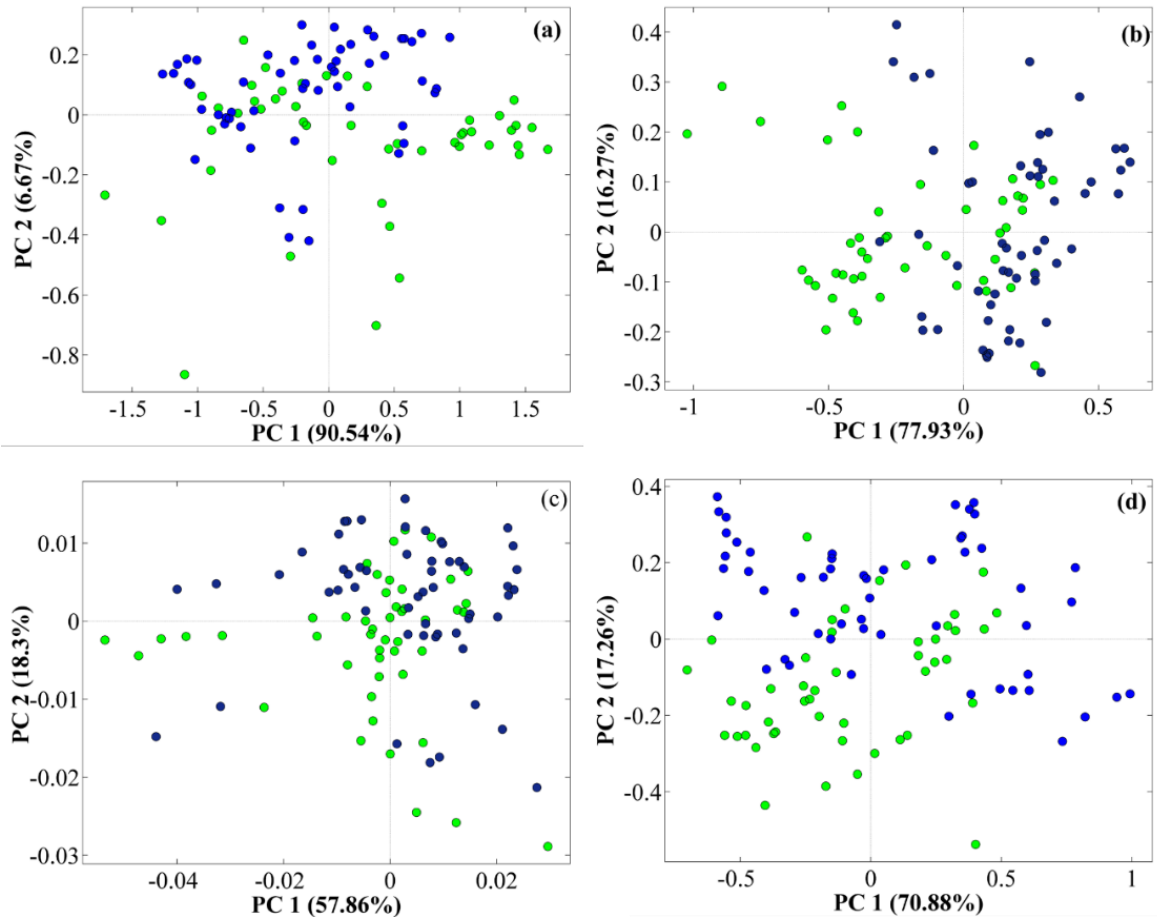
Fonte: Autoria própria (2024).

A primeira derivada evidenciou as bandas de absorção dos grupos CH e OH, as quais contêm informações sobre a lactose e os produtos da sua hidrólise (LIMA *et al.,* 2018). Além disso, é possível notar que, após a aplicação dos pré-processamentos, houve uma redução do ruído e do espalhamento dos espectros, bem como a correção do deslocamento da linha de base.

4.3 ANÁLISE POR COMPONENTES PRINCIPAIS

A análise de componentes principais foi aplicada às 102 amostras, com o objetivo de realizar um estudo exploratório para investigar a existência de tendências de agrupamento das classes. Os conjuntos de dados dos espectros brutos e com diferentes pré-processamentos foram submetidos a esta análise. Os resultados obtidos foram expressos na **Figura 4.3** por meio dos gráficos dos scores das 2 primeiras PC's para cada uma das situações.

Figura 4.3: Gráficos dos scores de PC1 versus PC2 para as 102 amostras de iogurte das classes isenta (azul) e convencional (verde): (a) Brutos, (b) Offset, (c) derivação Savitzky-Golay com janela de 5 pontos e (d) SNV.



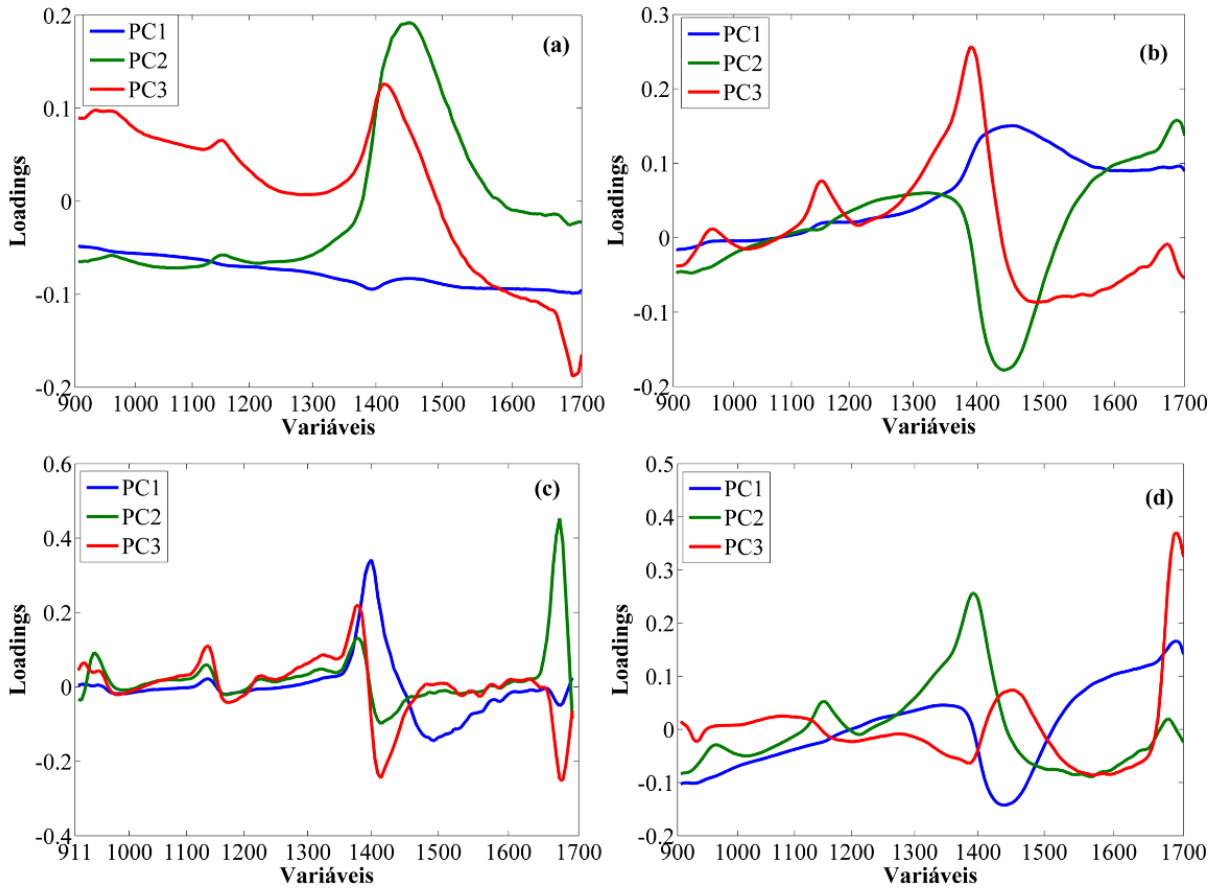
Fonte: Autoria própria (2024).

Ao examinar os gráficos dos escores de PC's (**Figura 4.3**), é possível verificar uma considerável sobreposição entre as amostras de iogurte isentas e convencionais. Dessa forma, não é possível visualizar a formação de agrupamentos bem definidos nos gráficos

bidimensionais de PC1 versus PC2. Assim, indicando que as duas classes não são facilmente distinguíveis pelos modelos PCA. Essa situação pode ser decorrente da semelhança existente entre os espectros das 102 amostras de iogurte, indicando que estes possam ter variáveis com alta colinearidade.

A **Figura 4.4** apresenta os gráficos de loadings para as três primeiras PC's dos espectros brutos e pré-processados. Ao analisar esses gráficos, é possível notar a influência majoritária das variáveis que estão relacionadas à banda de absorção de água, aproximadamente em 1450 nm, bem como das regiões em torno de 950 e 1200 nm, que correspondem, respectivamente, ao segundo sobreton de OH e ao segundo sobreton de CH. Essas informações sugerem que, embora os espectros sejam dominados pela banda de água, há potencial identificar marcadores de lactose e os produtos de sua hidrólise.

Figura 4.4: Gráfico dos loadings de PCA versus variáveis para amostras de iogurte isenta e com lactose para os dados (a) brutos e pré-processados com (b) offset, (c) derivada e (c) SNV.



Fonte: Autoria própria (2024).

Contudo, ainda pode-se destacar uma tendência de agrupamento das amostras isentas, no qual a maioria das amostras desta classe se encontram nos escores positivos de PC2, para os

dados brutos, e positivos de PC1, para dados processados com offset. Ao examinar os loadings dessas PC's, na **Figura 4.4a** e **4.4b,** é possível observar que a região de 1400 e 1500 nm apresenta loadings positivos, portanto, portam informação que favoreça esse agrupamento.

Embora PCA seja uma poderosa ferramenta de redução de dimensionalidade, amplamente utilizada, que auxilia a visualização dos dados, sua incapacidade em discriminar amostras de iogurte isentas de lactose e convencionais, sugere que informações sutis sobre as amostras, necessária para realizar a separação, podem não ser capturadas. Assim, técnicas de classificação mais avançadas como SIMCA, baseada em PCA, e LDA, com seleção de variáveis, foram empregadas.

4.4 CLASSIFICAÇÃO DE IOGURTES ISENTO DE LACTOSE UTILIZANDO ESPECTROSCOPIA NIR E SIMCA

Os modelos SIMCA propostos para classificação das amostras de iogurte foram realizados com os dados brutos e pré-processados apenas com a validação interna, todos os dados foram centrados na média antes da realização da classificação. Foi utilizado um conjunto de treinamento de cada classe, composto por 70% das amostras, para a construção dos modelos PCA individuais. Na etapa de teste, foi usado um conjunto que representava 30% das amostras de cada classe, assim como expressos na seção 3.3. Os modelos SIMCA foram construídos a partir dos modelos PCA individuais de cada classe, utilizando 5% de significância como um limite de classificação.

Para a modelagem SIMCA, foi possível observar amostras que foram incorretamente classificadas. Todas as amostras dos quatro conjuntos de dados (Bruto e pré-processados) foram classificadas em sua respectiva classe, entretanto, a maioria das amostras foi classificada em ambas as classes, o que resultou em um alto erro de classificação. A **Tabela 4.1** apresenta o número de amostras das classes IL e CL que foram classificadas como pertencentes a ambas as classes. Os modelos criados a partir dos dados pré-processados com derivada e SNV apresentaram o menor número de amostras classificadas para as duas classes.

Tabela 4.1: Número de amostras classificadas como ambas as classes.

	IL	CL
Bruto	12	6
Offset	12	7
Derivada	10	4
SNV	8	6

Fonte: Autoria própria (2024).

A **Tabela 4.2** resume os resultados da classificação para os modelos SIMCA.

Tabela 4.2: Matriz de confusão e resultados dos parâmetros de desempenhos dos modelos SIMCA.

Tratamento		Bruto		Offset		Derivada		SNV	
Método	Real/pred.	Teste		Teste		Teste		Teste	
	Métricas (%)	IL	CL	IL	CL	IL	CL	IL	CL
SIMCA	IL	4	12	4	12	6	10	8	8
	CL	6	8	7	7	4	10	6	8
	Sens.	25		25		37.5		50	
	Espe.	57.1		50		71.4		57.1	
	Pres.	40		36.4		60		57.1	
	TCC	40		36.7		53.3		53.3	

Fonte: Autoria própria (2024).

Devido à sobreposição dos modelos PCA, acarretando amostras classificadas em ambas as classes, a modelagem SIMCA apresentou baixo desempenho. A classificação incorreta das amostras pode ser atribuída à semelhança dos espectros de iogurte, assim, portando variáveis com alta colinearidade, sendo necessário a utilização de modelos de classificação que façam a seleção de variáveis, como os algoritmos BA-LDA, GA-LDA e SPA-LDA.

4.5 CLASSIFICAÇÃO DE IOGURTES ISENTO DE LACTOSE UTILIZANDO ESPECTROSCOPIA NIR E MODELAGEM DISCRIMINANTE LINEAR

O estudo de classificação com modelagem discriminante em combinação com os algoritmos de seleção de variáveis (BA, GA e SPA) foram construídos a partir dos espectros brutos e pré-processados. As amostras foram divididas em conjuntos de treinamento e teste conforme a expressos na seção 3.3

A **Tabela 4.3** apresenta a quantidade de vaiáveis (comprimento de onda) selecionadas pelos três algoritmos, nos conjuntos de dados originais e pré-processados com offset, derivada e SNV.

Tabela 4.3: Número de variáveis selecionadas pelos diferentes algoritmos.

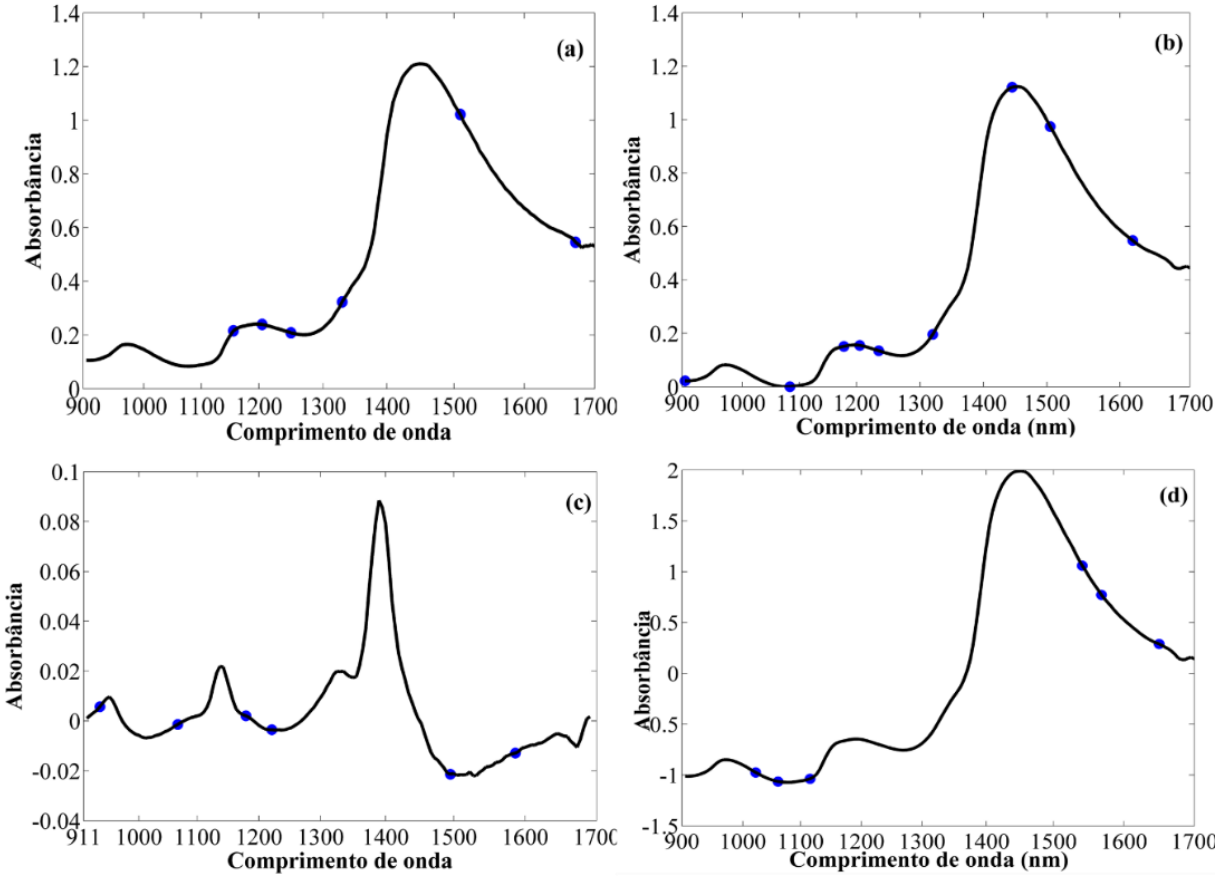
Pré-processamento	Modelos		
	SPA-LDA	GA-LDA	BA-LDA
Bruto	11	4	6
Offset	13	8	9
Derivada	7	2	6
SG-SNV	10	5	6

Fonte: Autoria própria (2024).

O GA foi empregado devido a sua capacidade de explorar um amplo espaço de buscar. Por sua vez, o BA foi utilizado por possuir a capacidade de convergir para regiões especificas, e por fim, a utilização do SPA é justificada devido ser um algoritmo que busca um conjunto de variáveis que possuam a maior variância entre si.

O BA selecionou um quantitativo de 6 variáveis para os dados brutos, SNV e derivada, e 9 variáveis para o pré-processamento offset, essas variáveis foram obtidas por todo espectro, como pode ser visualizado na **Figura 4.5**. Contudo, é possível notar a seleção de algumas variáveis em faixas que podem portar informação para distinção das amostras de iogurte, como a região de segundo sobreton de C-H (aproximadamente 1200 nm) e primeiro e segundo sobreton de O-H (1450 e 950). Os modelos BA-LDA resultantes classificaram corretamente todas as amostras de iogurte de treinamento e de predição, atingindo 100% de desempenho nas quatro métricas.

Figura 4.5: Variáveis selecionadas pelo BA para os (a) espectros brutos e pré-processados com (b) Offset, (c) derivada, (d) SNV.

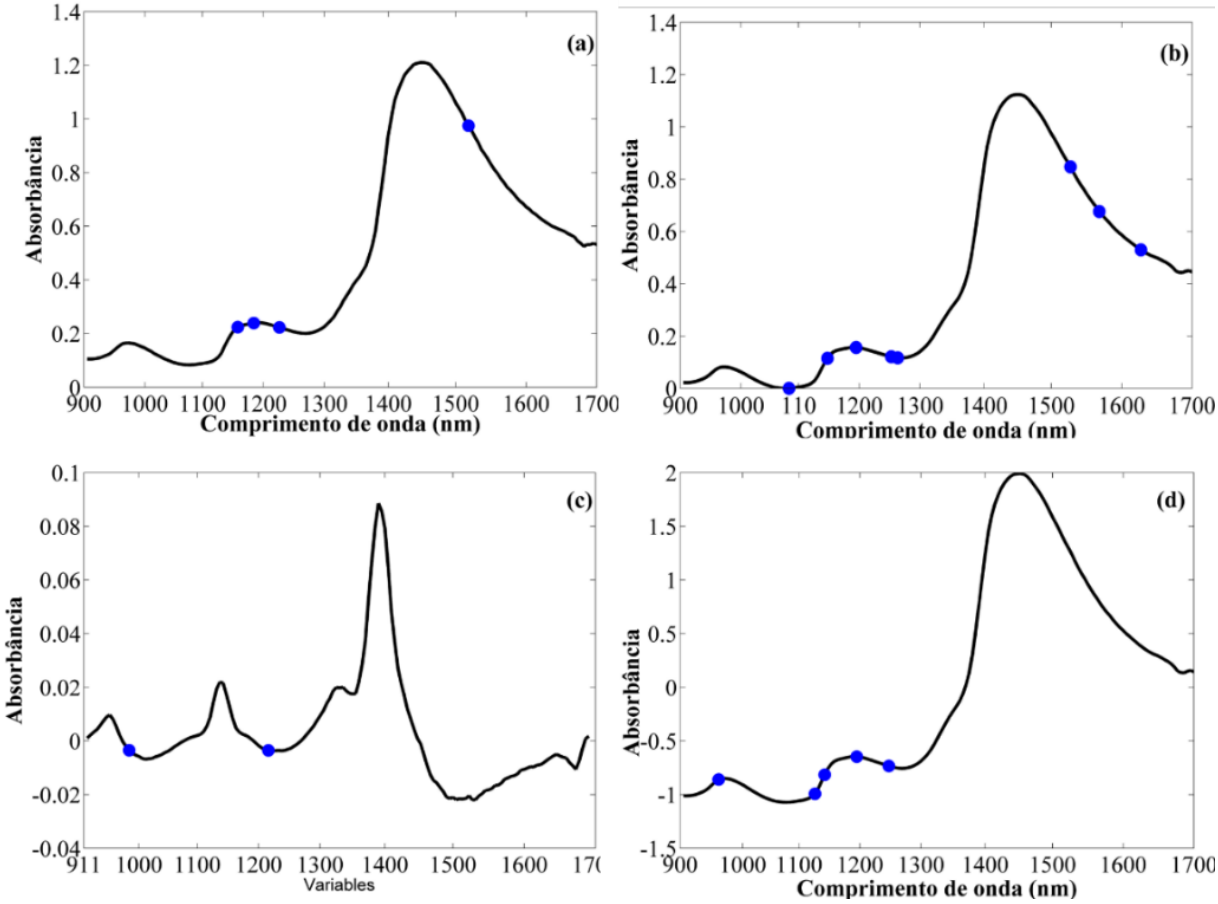


Fonte: Autoria própria (2024).

Os modelos LDA empregados com algoritmo genético selecionaram o menor quantitativo de variáveis, sendo 4 variáveis para dados brutos, 8 para Offset, 2 para derivada e

5 para SNV. Como pode ser visto na **Figura 4.6**, os quatro modelos selecionaram variáveis em regiões referentes à absorção de C-H, as quais podem ser atribuídas à lactose. Apenas a modelagem com dados brutos e offset selecionaram variáveis acima de 1500 nm. Os modelos GA-LDA resultantes foram aplicados aos conjuntos de teste, os quais conseguiram predizer adequadamente todas as amostras, obtendo um desempenho máximo nos conjuntos de treinamento e teste em todos os parâmetros utilizados.

Figura 4.6: Variáveis selecionadas pelo GA para os (a) espectros brutos e pré-processados com (b) Offset, (c) derivada, (d) SNV.



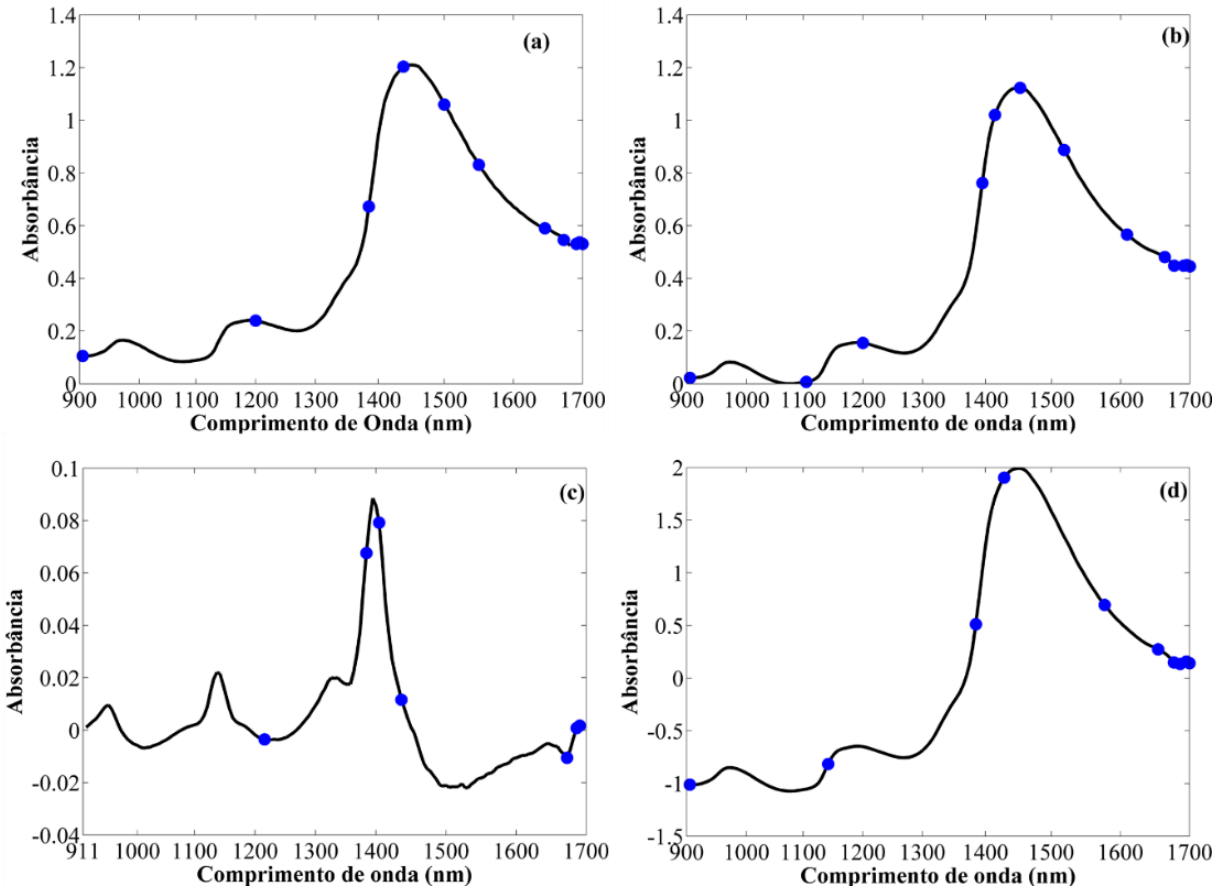
Fonte: Autoria própria (2024).

O SPA selecionou 11 variáveis para os dados brutos, 13 variáveis para o pré-processamento offset, 7 com derivada e 10 com SNV (**Figura 4.7**), sendo obtidas ao longo do espectro, com uma concentração em comprimentos de onda mais altos, entre 1600 e 1700 nm. Essa situação pode ser decorrente da variação aleatória que ocorre na região final dos espectros. A primeira e a última variáveis foram selecionadas para todas as condições, com exceção do modelo com derivada, que não foi selecionada a primeira variável. Apesar desse empecilho,

outras variáveis foram selecionadas em regiões importantes para lactose, como nas bandas de primeiro sobreton de O-H e segundo sobreton C-H.

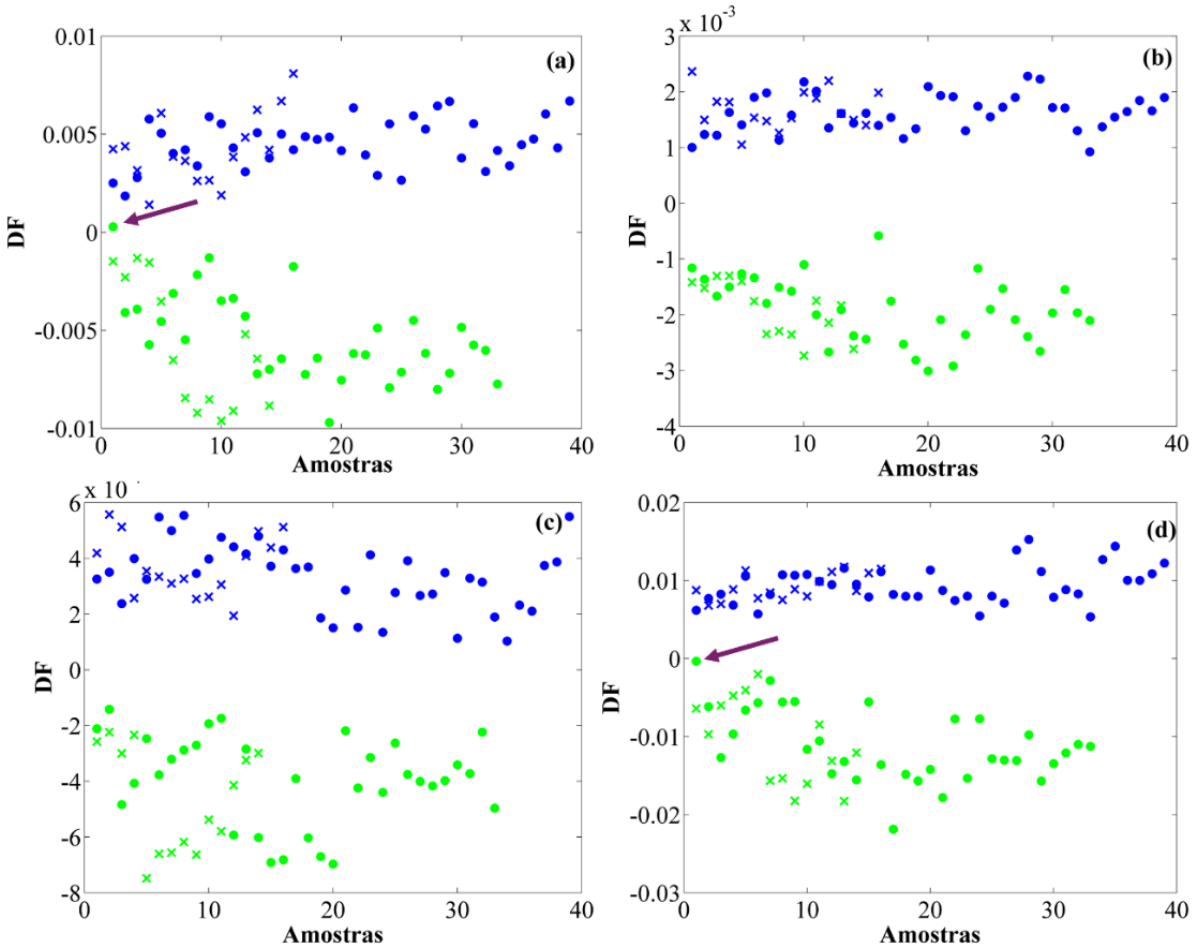
Os modelos SPA-LDA, construídos com base nos dados brutos e com SNV, obtiveram valores de sensibilidade máximos para os conjuntos de treinamento. Em contrapartida, os dois modelos atingiram valores de 97%, 97.5% e 98.6%, para especificidade, precisão e TCC, respectivamente. Esses resultados são decorrentes de uma única amostra convencional que foi incorretamente classificada como sendo isento de lactose, como pode ser observado no gráfico dos scores da função discriminante (**Figura 4.8**). Em compensação, os modelos com derivada e offset alcançaram valores máximos de eficiência para os conjuntos de treinamento. Para a fase de teste, todas as amostras das classes foram preditas corretamente pelos modelos SPA-LDA com dados brutos e pré-processados.

Figura 4.7: Variáveis selecionadas pelo SPA para os (a) espectros brutos e pré-processados com (b) Offset, (c) derivada e (d) SNV.



Fonte: Autoria própria (2024).

Figura 4.8: Gráfico dos scores da Função Discriminante 1 versus amostras do SPA-LDA para a classe isento (azul) e convencional (verde) dos (a) espectros brutos e pré-processados com (b) offset, (c) derivada e (d) SNV. (●: treinamento e x: teste).



Fonte: Autoria própria (2024).

De maneira geral, todos os modelos LDA criados em combinação com os três algoritmos de seleção de variáveis obtiveram ótimas taxas de desempenho, demonstrando assim eficiência para resolver o problema de classificação de iogurtes quanto à presença de lactose. Contudo os modelos criados sem a realização de pré-processamentos apresentaram bons resultados, indicando assim que não há necessidade de realizar os tratamentos para a modelagem LDA, portanto sendo uma metodologia mais simplificada e de maior facilidade de aplicação na indústria.

Os três algoritmos demonstraram ser igualmente eficientes na seleção de variáveis discriminantes, evidenciado pelos altos valores de eficiência. Os modelos BA-LDA e GA-LDA apresentaram ligeira diferença nas regiões das variáveis selecionadas, em contrapartida essa diferença foi mais atenuada nos conjuntos de variáveis determinados pelo SPA.

Como já mencionado, para os dados brutos, foi observado apenas um único erro, no qual uma amostra da classe convencional foi classificada como sendo isenta. Essa má classificação

pode ser oriunda das variáveis selecionadas pelo SPA não portarem informação discriminante das classes para esse espectro.

A **Tabela 4.4**, que apresenta a matriz de confusão e as métricas de desempenho, resume os resultados de classificação dos modelos LDA obtidos com os três algoritmos de seleção de variáveis.

Tabela 4.4: Resultado obtidos pelos diferentes métodos de classificação com LDA.

Tratamento		Brutos				Offset				Derivada				SNV			
Método	Real/pred.	Treina.		Teste		Treina.		Teste		Treina.		Teste		Treina.		Teste	
	Métricas (%)	IL	CL	IL	CL	IL	CL	IL	CL	IL	CL	IL	CL	IL	CL	IL	CL
SPA-LDA	IL	39	0	16	0	39	0	16	0	39	0	16	0	39	0	16	0
	CL	1	32	0	14	0	33	0	14	0	33	0	14	1	32	0	14
	Sens.	100		100		100		100		100		100		100		100	
	Espe.	97		100		100		100		100		100		97		100	
	Pres.	97.5		100		100		100		100		100		97.5		100	
	TCC	98.6		100		100		100		100		100		98.6		100	
GA-LDA	IL	39	0	16	0	39	0	16	0	39	0	16	0	39	0	16	0
	CL	0	33	0	14	0	33	0	14	0	33	0	14	0	33	0	14
	Sens.	100		100		100		100		100		100		100		100	
	Espe.	100		100		100		100		100		100		100		100	
	Pres.	100		100		100		100		100		100		100		100	
	TCC	100		100		100		100		100		100		100		100	
BA-LDA	IL	39	0	16	0	39	0	16	0	39	0	16	0	39	0	16	0
	CL	0	33	0	14	0	33	0	14	0	33	0	14	0	33	0	14
	Sens.	100		100		100		100		100		100		100		100	
	Espe.	100		100		100		100		100		100		100		100	
	Pres.	100		100		100		100		100		100		100		100	
	TCC	100		100		100		100		100		100		100		100	

IL – Isento de lactose; CL – com lactose. **Fonte:** Autoria própria (2024).

Com base nas regiões que as variáveis foram selecionadas (~1000, ~1200, ~1450), é possível destacar que a discriminação das duas classes de iogurte foi influenciada pela presença do dissacarídeo lactose, presentes nos iogurtes normais, e dos monossacarídeos glicose e galactose, provenientes da hidrólise da lactose, presentes nos iogurtes isento de lactose. Pois, essas regiões podem ser relacionadas às bandas vibracionais de primeiro e segundo sobreton de OH e CH, características de monossacarídeos e oligossacarídeos (LIMA, *et al.,* 2018; ROBERT; CADET, 1998).

Capítulo 5

CONCLUSÕES

5 CONCLUSÕES

O presente estudo apresentou uma estratégia analítica, baseada na espectroscopia NIR combinada a técnicas quimiométricas multivariadas, para classificação rápida e não destrutiva de iogurtes quanto à presença de lactose. Para essa finalidade, foram utilizadas técnicas de reconhecimento de padrão SIMCA e LDA combinadas com o algoritmo de seleção de variáveis (GA, BA e SPA) para seleção de variáveis (comprimentos de onda) em espectros NIR.

A análise exploratória dos dados por PCA apontou uma alta sobreposição das amostras de iogurte isentos e convencionais, com os dados brutos e pré-processados, o que refletiu no desempenho de classificação dos modelos SIMCA, apresentando um alto número de amostras classificadas como sendo de ambas as classes e baixos valores de desempenho. Sendo necessário recorrer a técnicas de classificação combinada com seleção de variáveis.

Os modelos LDA demonstraram eficácia para a classificação de amostras de iogurte das duas classes, evidenciado pelos valores máximos de desempenho. Os três algoritmos de seleção de variáveis mostraram ser igualmente eficientes, selecionando uma pequena quantidade de variáveis e em regiões que contêm informação referente à lactose e os derivados de sua hidrólise.

Os pré-processamentos foram utilizados com o objetivo de minimizar os interferentes existentes na aquisição dos espectros e para investigar uma possível melhora dos modelos SIMCA, no entanto, mesmo após a aplicação dos tratamentos, os modelos não foram capazes de realizar a classificação das amostras de iogurte. Por sua vez, os modelos LDA, combinados com os algoritmos de seleção de variáveis, apontaram ser eficientes para discriminar as duas classes com os dados originais sem tratamento.

A estratégia proposta, que se baseia na espectroscopia NIR e na análise multivariada, se mostrou uma ferramenta alternativa viável para realizar a classificação de iogurte com e sem lactose, apenas para modelagem LDA. A utilização de um equipamento portátil comercial mostrou-se adequada, além de apresentar a vantagem de utilização em *in situ*.

Embora os métodos propostos com LDA tenha apresentado ótimos resultados, utilização de técnicas de referências para certificar a autenticidade dos iogurtes isentos de lactose podem ser aplicadas como estratégia futuras. Bem como buscar possíveis planejamentos de aplicação industrial da metodologia.

REFERÊNCIAS

ALAMAR, Priscila D. *et al*. Detection of fruit pulp adulteration using multivariate analysis: Comparison of NIR, MIR and data fusion performance**. Food Analytical Methods**, v. 13, p. 1357-1365, 2020.

ANZANELLO, M. J. *et al*. A genetic algorithm-based framework for wavelength selection on sample categorization. **Drug testing and analysis**, v. 9, n. 8, p. 1172-1181, 2017.

ARAÚJO, M. C. U., *et al*. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. **Chemometrics and intelligent laboratory systems**, v. 57, n. 2, p. 65-73, 2001.

BALLABIO, D.; GRISONI. F.; TODESCHINI, R. Multivariate comparison of classification performace measures. **Chemometrics and Intelligent Laboratory Systems.** v. 174, p. 33-44, 2018.

BELADELI, M. N.; GODINHO, E. Z. Características físico-químicas na elaboração de iogurte com pedaços de frutas. **Revista Brasileira de Tecnologia Agroindustrial**, v. 16, n. 1, p. 3749-3766, 2022.

BI, Y. *et al.* A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation**. Analytica Chimica acta**, v. 909, p. 30-40, 2016.

BOKOBZA, L. Near infrared spectroscopy. **Journal of Near Infrared Spectroscopy**, v. 6, n. 1, p. 3-17, 1998.

BOLAND, M.; SINGH, H. **Milk proteins: from expression to food.** 3. ed. Londres: Academic Press. 2019.

BOLS, M. L. *et al*. In Situ UV–Vis–NIR Absorption Spectroscopy and Catalysis. **Chemical Reviews**, n. 124, p. 2352−2418, 2024.

BRANCO, M. de S. C. *et al*. Classificação da intolerância à lactose: uma visão geral sobre causas e tratamentos. **Revista de Ciências Médicas**, v. 26, n. 3, p. 117-125, 2017.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Métodos Oficiais para Análise de Produtos de Origem Animal.** 1. ed. Brasília, DF: MAPA, 2022.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução normativa nº 46, de 23 de outubro de 2007. Dispõe sobre a Inspeção Industrial e Sanitária dos Produtos de Origem Animal, considerando a Resolução MERCOSUL/GMC/RES. **Diário Oficial da União:** Seção 1, Brasília, n. 205, p. 4, 24 out. 2007. Disponível em: https://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=24/10/2007&jornal=1&pagina=4&totalArquivos=96. Acesso em: 19 maio 2024.

BRASIL. Ministério da Saúde. Biblioteca Virtual em Saúde. **Alimentos Funcionais.** Brasília: Ministério da Saúde, 2009. Disponível em < https://bvsms.saude.gov.br/alimento-funcionais/>. Acesso em: 19 maio 2024.

BRASIL. Ministério da Saúde. Agência Nacional de Vigilância Sanitária. RDC nº 135, de 08 de fevereiro de 2017. Aprova o regulamento técnico referente a alimentos para fins especiais, para dispor sobre os alimentos para dietas com restrição de lactose. **Diário Oficial da União:** Seção 1, Brasília, n. 29, p. 44, 9 fevereiro 2017. Disponível em <https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/20794561/do1-2017-02-09-resolucao-rdc-n-135-de-8-de-fevereiro-de-2017-20794490>. Acesso em: 19 maio 2024.

BRERETON, R. G. *et al*. Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools. **Analytical and bioanalytical chemistry**, v. 409, p. 5891-5899, 2017.

BRERETON, R. G. Pattern recognition in chemometrics**. Chemometrics and Intelligent Laboratory Systems,** v. 149, p. 90-96, 2015.

BURNS, D. A.; CIURCZAK, E. W. **Handbook of near-infrared analysis**. 3. ed. CRC press, 2008.

CAMPOS, N. S. *et al*. Utilização do Glicosímetro Accu-Chek® Para a Determinação de Lactose em Produtos Lácteos. **Revista Virtual de Química**, v. 6, n. 6, p. 1677-1686, 2014.

CASTELLANO, B. F. *et al*. Intolerância à lactose: diagnóstico clínico laboratorial e genético. **BioSCIENCE**, v. 80, n. 2, p. 12-12, 2022.

CARAMÊS, E. T. dos S. *et al*. Near infrared spectroscopy and smartphone-based imaging as fast alternatives for the evaluation of the bioactive potential of freeze-dried açai. **Food Research International**, v. 140, p. 109792, 2021.

CARAMÊS, E. T. dos S.; LIMA-PALLONE. Métodos analíticos verdes para a análise de alimentos. In: Verruck, S. **Avanços em Ciência e Tecnologia de Alimentos**, 1 ed. v. 3. Guarujá: Científica, 2021. p. 278-288, 2021.

CARPIN, M. *et al*. Caking of lactose: A critical review. **Trends in Food Science & Technology**, v. 53, p. 1-12, 2016.

CARVALHO, G. R. Oferta e demanda de leite no Brasil em 2022. **Anuário do leite.** v. 2023. p. 26-27, 2023.

CARVALHO, G. R.; OLIVEIRA, L. A. A. de; ARANTES, M. S. L. Oferta e demanda de leite no Brasil em 2023. **Anuário do leite.** v. 2024. p. 14-15, 2024.

CERQUEIRA, E. O. *et al*. Utilização de filtro de transformada de Fourier para a minimização de ruídos em sinais analíticos. **Química Nova**, v. 23, p. 690-698, 2000.

CHEN, B.; LEWIS, M. J.; GRANDISON, A. S. Effect of seasonal variation on the composition and properties of raw milk destined for processing in the UK**. Food chemistry**, v. 158, p. 216-223, 2014.

CHEN, H. *et al*. Quantifying several adulterants of notoginseng powder by near-infrared spectroscopy and multivariate calibration. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,** v. 211, p. 280-286, 2019.

CHEN, H.; TAN, C.; LIN, Z. Identification of ginseng according to geographical origin by near-infrared spectroscopy and pattern recognition. **Vibrational Spectroscopy**, v. 110, p. 103149, 2020.

CHEN, Z.; HARRINGTON, P. de B. Automatic soft independent modeling for class analogies. **Analytica Chimica Acta**, v. 1090, p. 47-56, 2019.

COSTA FILHO, P. A. da; POPPI, R. J. Algoritmo genético em química. **Química Nova**, v. 22, p. 405-411, 1999.

CORGNEAU, M. *et al*. Recent advances on lactose intolerance: Tolerance thresholds and currently available answers. **Critical reviews in food science and nutrition**, v. 57, n. 15, p. 3344-3356, 2017.

CRUZ, A. G. da; ZACARCHENCO, P. B. Desafio tecnológicos da aplicação de probióticos em produtos lácteos. *In:* SIMPÓSIO LÁCTEOS E SAÚDE, 1, 2016, Campinas. **Anais.** Campinas: ITAL, 2015. p. 67-71.

DANTAS, A.; VERRUCK, S.; PRUDENCIO, E. S. **Ciência e tecnologia de leite e produtos lácteos sem lactose**. Ponta Grossa-PR: Atena Editora, 2019.

DENG, Y. *et al*. Lactose intolerance in adults: biological mechanism and dietary management. **Nutrients**, v. 7, n. 9, p. 8020-8035, 2015.

DIAZ-OLIVARES, J. A. *et al*. Near-infrared spatially-resolved spectroscopy for milk quality analysis. **Computers and Electronics in Agriculture**, v. 219, p. 108783, 2024.

DIEGO, I. M. de. *et al*. General Performance Score for classification problems. **Applied Intelligence**, v. 52, n. 10, p. 12049-12063, 2022.

FAO - Food and Agriculture Organization of the United Nations. **Dairy Market Review – Overview of global dairy market and policy developments in 2023.** Rome, 2024.

FAO - Food and Agriculture Organization of the United Nations. **Probiotics in animal nutrition – Production, impact and regulation.** Rome, 2016.

FERNANDES, D. D. de S. *et al*. Simultaneous identification of the wood types in aged cachaças and their adulterations with wood extracts using digital images and SPA-LDA**. Food Chemistry,** v. 273, p. 77-84, 2019.

FERREIRA, M. M. C. **Quimiometria – Conceitos, Métodos e Aplicações.** 1. ed. Campinas: UNICAMP, 2015.

FERREIRA, M. M. C. Quimiometria III-Revisitando a análise exploratória dos dados multivariados. **Química Nova**, v. 10, pág. 1251-1264, 2022.

FISBERG, M.; MACHADO, R. History of yogurt and current patterns of consumption. **Nutrition reviews**, v. 73, n. suppl_1, p. 4-7, 2015.

FIESP - Federação das Indústrias do Estado de São Paulo. **A mesa dos brasileiros: Transformações, confirmações e contradições.** São Paulo: FIESP, 2017.

FORSGARD, R. A. Lactose digestion in humans: intestinal lactase appears to be constitutive whereas the colonic microbiome is adaptable. **The American journal of clinical nutrition, v. 110, n. 2, p. 273-279, 2019.**

FOX, P. F. *et al*. **Dairy Chemistry and Biochemistry.** 2. ed. Springer. 2015.

GARDNER-LUBBE, Sugnet. Linear discriminant analysis for multiple functional data analysis. **Journal of Applied Statistics**, v. 48, n. 11, p. 1917-1933, 2021.

GDP -GLOBAL DAIRY PLATFORM. **Dairy Everyday.Around the world.** 2017.

GOLIC, M.; WALSH, K.; LAWSON, P. Short-wavelength near-infrared spectra of sucrose, glucose, and fructose with respect to sugar concentration and temperature. **Applied spectroscopy**, v. 57, n. 2, p. 139-145, 2003.

GONÇALVES, V. P. **Monitoramento da estabilidade oxidativa de biodiesel empregando espectroscopia vibracional associada a ferramentas quimiométricas.** 2022. Dissertação (Mestrado em Química) – Instituto de Ciências Exatas – Universidade Federal de Minas Gerais, Belo Horizonte, 2022

GUERRA, P. V. P. *et al*. **Intolerância à lactose.** 1. ed. Belo Horizonte: GastroPED, 2018.

HARTWIG, F. P. **Intolerância à lactose: prevalência, determinantes e associação com consumo de lacticínios e osteoporose.** 2014. Dissertação (Mestrado em epidemiologia), Universidade Federal de Pelotas, Pelotas, 2014.

HE, Y. *et al*. Fast measurement of sugar content of yogurt using Vis/NIR-spectroscopy. **International Journal of Food Properties**, v. 10, n. 1, p. 1-7, 2007.

HUSSAIN, N.; SUN, D. W.; PU, H. Classical and emerging non-destructive technologies for safety and quality evaluation of cereals: **A review of recent applications**. Trends in Food Science & Technology, v. 91, p. 598-608, 2019.

IBGE -INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa de orçamentos familiares 2017-2018: análise do consumo alimentar pessoal no Brasil.** Rio de Janeiro: IBGE, 2020.

I. A. L - INSTITUTO ADOLFO LUTZ. **Métodos físico-químicos para análise de alimentos. 2008.**

JIMÉNEZ-CARVELO, A. M.; CUADROS-RODRÍGUEZ, L. The occurrence: A meaningful parameter to be considered in the validation of multivariate classification-based screening methods–Application for authenticating virgin olive oil. **Talanta**, v. 208, p. 120467, 2020.

KURIBAYASHI, L. M. *et al*. Immobilization of β-galactosidase from Bacillus licheniformis for application in the dairy industry. **Applied Microbiology and Biotechnology**, v. 105, p. 3601-3610, 2021.

LEE, Y.; HAN, S. H.; NAM, S. H. Soft independent modeling of class analogy (SIMCA) modeling of laser-induced plasma emission spectra of edible salts for accurate classification. **Applied Spectroscopy**, v. 71, n. 9, p. 2199-2210, 2017.

LIMA, G. F. de. *et al*. Multivariate classification of UHT milk as to the presence of lactose using benchtop and portable NIR spectrometers. **Food analytical methods**, v. 11, p. 2699-2706, 2018.

LIU, W. *et al*. Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics. **Food chemistry**, v. 251, p. 86-92, 2018.

LUCAS, B. N.; SCHÚ, A. I.; NORA, F. M.D. Uso de smartphone como alternativa inovadora no controle de qualidade de alimentos: uma breve revisão. *In:* Verruck, S. **Avanços em Ciência e Tecnologia de Alimentos**, 1 ed. v. 3. Guarujá: Científica, 2021. p. 278-288, 2021.

MANUEL, M. N. B. One-class classification of special agroforestry Brazilian coffee using NIR spectrometry and chemometric tools. **Food Chemistry,** v. 366, p. 130480, 2022.

MATTAR, R.; MAZO, D. F. de. Intolerância à lactose: mudança de paradigmas com a biologia molecular**. Revista da Associação Médica Brasileira**, v. 56, p. 230-236, 2010.

MELFSEN, A.; HARTUNG, E.; HAEUSSERMANN, A. Accuracy of milk composition analysis with near infrared spectroscopy in diffuse reflection mode. **Biosystems engineering**, v. 112, n. 3, p. 210-217, 2012.

MESQUITA JÚNIOR, G. A. de. *et al*. Disponibilidade de iogurtes para consumidores intolerantes à lactose. **Brazilian Journal of Development**, v. 7, n. 5, p. 44722-44736, 2021.

MORAIS, C. L. M.; LIMA, K. M. G.; MARTIN, F. L. Variable selection towards classification of digital images: identification of altered glucose levels in serum. **Analytical Letters**, v. 52, n. 14, p. 2239-2250, 2019.

MUNCAN, J.; TEI, K.; TSENKOVA, R. Real-time monitoring of yogurt fermentation process by aquaphotomics near-infrared spectroscopy. **Sensors**, v. 21, n. 1, p. 177, 2020.

NIAZI, A.; LEARDI, R. Genetic algorithms in chemometrics. **Journal of Chemometrics**, v. 26, n. 6, p. 345-351, 2012.

NÓBREGA, R. O. da. **Classificação de cafés solúveis usando espectroscopia NIR e quimiometria.** 2021. Dissertação (Mestrado em Química) – Universidade Federal da Paraíba, João Pessoa, 2021.

NÓBREGA, R. O. *et al*. Classification of instant coffees based on caffeine content and roasting degree using NIR spectrometry and multivariate analysis. **Microchemical Journal**, v. 190, p. 108624, 2023.

OLIVERI, P. *et al*. Qualitative pattern recognition in chemistry: Theoretical background and practical guidelines. **Microchemical Journal**, v. 162, p. 105725, 2021.

PASQUINI, C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. **Journal of the Brazilian chemical society**, v. 14, p. 198-219, 2003.

PASQUINI, C. Princípios da Espectroscopia no Infravermelho Próximo. *In:* TIBOLA, C. S. et al. **Espectroscopia no Infravermelho Próximo para Avaliar Indicadores de Qualidade Tecnológica e Contaminantes em Grãos.** 1. ed. Brasília, EMBRAPA, 2018. p. 13-30.

PAULA, L. N. de. *et al*. Teor de lactose em queijo fino tipo Brie com método alternativo. **Biblioteca Digital de TCC-UniAm? rica**, v. 6, n. 13, p. 570-579, 2023.

PELEGRINE, D. H. G.; AGUIAR, LF de S.; LODELIS, A. Iogurte de goiaba enriquecido com cereais: correlação da textura com os parâmetros sensoriais. **Revista de Ciência & Tecnologia**, v. 18, n. 36, p. 25-40, 2015.

PERATI, P.; BORBA, B. de; ROHRER, J. Determination of lactose in lactose-free milk products by high-performance anion-exchange chromatography with pulsed amperometric detection. **Thermo Fisher Scientific, Application Note**, p. 1-8, 2016.

PEREIRA, J. A. *et al*. Potentially symbiotic fermented milk: A preliminary approach using lactose-free milk. **LWT**, v. 118, p. 108847, 2020.

PEREIRA, M. C. S. *et al*. Lácteos com baixo teor de lactose: uma necessidade para portadores de má digestão da lactose e um nicho de mercado**. Revista do Instituto de Laticínios Cândido Tostes**, v. 67, n. 389, p. 57-65, 2012.

PONTES, M. J. C. *et al*. Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification. **Talanta**, v. 85, n. 4, p. 2159-2165, 2011.

PONTES, M. J. C. *et al*. The successive projections algorithm for spectral variable selection in classification problems. **Chemometrics and Intelligent Laboratory Systems,** v. 78, n. 1-2, p. 11-18, 2005.

PONTES, A. S. *et al*. Ant colony optimization for variable selection in discriminant linear analysis**. Journal of Chemometrics**, v. 34, n. 12, p. e3292, 2020.

PIRES, A. H. M. **Otimização de controladores fuzzy por algoritmos genéticos multiobjetivos no domínio wavelet.** 2023. Tese (Doutorado em Ciências), Universidade Federal do Rio Grande do Norte, Natal, 2023.

RAMALHO, M. E. O.; GANECO, A. G. Intolerância a lactose e o processamento dos produtos zero lactose. **Revista Interface Tecnológica**, v. 13, n. 1, p. 119-133, 2016.

RAMOS, R. de O. **Avaliação de estratégias de imobilização e estabilização de βgalactosidase de bacillus licheniformis.** 2022. Dissertação (Mestrado em Engenharia de Alimentos), Universidade Federal de Uberlândia, Patos de Minas, 2022.

RENTERO, N. Projeção de tendências para o leite aqui e lá fora. **Anuário de leite.** v. 2023, p. 22-25, 2023.

REZENDE, A. C. B.; LOPES FILHO, G.; VIEIRA, F. H. T. Aplicação da Análise Discriminante Linear (LDA) para Classificação de Sinais Eletromiográficos (EMG) de Movimentos da Mão. *In*: ESCOLA REGIONAL DE INFORMÁTICA DE GOIÁS, 7, 2019. **Anais.** Goiânia: SBC, 2019. p. 351-360.

RIBEIRO, E.; CUBO, M. F.; SALEM, R. D. S. Desenvolvimento e caracterização físico-química de iogurte sem lactose adicionado de chia (Salvia hispanica L.). **Uningá Review**, v. 34, n. 1, p. 26-39, 2019.

ROBERT, C.; CADET, F. Analysis of near-infrared spectra of some carbohydrates. **Applied Spectroscopy Reviews**, v. 33, n. 3, p. 253-266, 1998.

ROSA, F. dos S. da; ALVES, M. K. Teor de lactose em iogurtes naturais e leites fermentados. **UNICIÊNCIAS**, v. 23, n. 2, p. 66-69, 2019.

SAFRAID, G. F. *et al*. Perfil do consumidor de alimentos funcionais: identidade e hábitos de vida. **Brazilian Journal of Food Technology**, v. 25, p. e2021072, 2022.

SALGADO, J. **Alimentos funcionais**. 1. ed. São Paulo: Oficina de Textos, 2017.

SANTANA, F. B. de. *et al*. Experimento didático de quimiometria para classificação de óleos vegetais comestíveis por espectroscopia no infravermelho médio combinado com análise discriminante por mínimos quadrados parciais: Um tutorial, PARTE V. **Química Nova**, v. 43, n. 3, p. 371-381, 2020.

SANTOS, C. A. T. dos; PÁSCOA, R. N. M. J; LOPES, J. A. A review on the application of vibrational spectroscopy in the wine industry: From soil to bottle. TrAC **Trends in Analytical Chemistry**, v. 88, p. 100-118, 2017.

SCHRADER, B. (Ed.). **Infrared and Raman spectroscopy: methods and applications**. John Wiley & Sons, 1995.

SENA, M. M. de; ALMEIDA, M. R. de. Quimiometria Aplicada aos Dados Espectrais no Infravermelho Próximo. *In:* TIBOLA, C. S. et al. **Espectroscopia no Infravermelho Próximo para Avaliar Indicadores de Qualidade Tecnológica e Contaminantes em Grãos.** 1. ed. Brasília, EMBRAPA, 2018. p. 31-50.

SHARIF, A. *et al*. Lactose intolerance and inheritance of lactase Persistence: A review. **RADS Journal of Pharmacy and Pharmaceutical Sciences**, v. 5, n. 3, p. 70-74, 2017.

SILVA, A. C. da. *et al*. A fast and low-cost approach to quality control of alcohol-based hand sanitizer using a portable near infrared spectrometer and chemometrics**. Journal of Near Infrared Spectroscopy**, v. 29, n. 3, p. 119-127, 2021.

SILVA, B. L. **Sistema de medição não invasiva de glicose sanguínea baseado em princípios de espectroscopia de infravermelho próximo.** 2017 Dissertação (Mestrado em Engenharia Elétrica) -Universidade Federal de Santa Catarina, Florianópolis, 2017.

SILVA, C. C. de. *et al*. Produção de iogurte sabor tomate enriquecido de fibras. **Revista Online JCTOB**, v. 1, n. 1, p. 56-62, 2017.

SILVA, C. M. E. da. A Intolerância à Lactose e as Consequências na Absorção do Cálcio. **Rev. Eletrôn. Atualiza Saúde**, v. 6, p. 29-35, 2017.

SILVA, I. S. C. da; PANDOLFI, M. A. C. Análise das principais tendências no mercado brasileiro de iogurtes**. Revista Interface Tecnológica**, v. 17, n. 2, p. 523-534, 2020.

SILVA, K. C. M. da. *et al*. Determinação da lactose ante às metodologias contemporâneas. **Revista do Instituto de Laticínios Cândido Tostes**, v. 75, n. 1, p. 59-71, 2020.

SILVA, W. M. da. **Determinação simultânea de adulteração em leite de cabra por adição de leite de vaca e teor de lipídios usando NIR portátil e algoritmos PLS**. 2023. Dissertação (Mestrado em Ciências Farmacêuticas) -Universidade Estatual da Paraíba, Campina Grande, 2023.

SIMIÃO, S. C. G. *et al*. Desenvolvimento e validação de metodologia para determinação de lactose em produtos lácteos processados qualificados como "zero lactose". *In*: CONGRESSO INTERINSTITUCIONAL DE INICIAÇÃO CIENTÍFICA, 12., 2018. **Anais.** Campinas: EMBRAPA Meio Ambiente, 2018.

SIQUEIRA, K. B. **O mercado consumidor de leite e derivados**. 1. ed. Juiz de Fora: Embrapa Gado e Leite, 2019.

SIQUEIRA, K. B. *et al*. **Consumo de lácteos na pandemia: Principais mudanças no comportamento do consumidor brasileiro de leite e derivados durante a pandemia de Covid-19**. 1. ed. Juiz de Fora: Embrapa Gado e Leite, 2021.

SIQUEIRA, K. B. Reflexões sobre o nível de consumo de leite do brasileiro. *In:* SIQUEIRA, K. B. **Na era do consumidor: uma visão do mercado lácteo brasileiro.** 1. ed. Juiz de Fora: Embrapa Gado de Leite, 2021. p. 17-19.

SOARES, S. F. C. *et al*. The successive projections algorithm. **TrAC Trends in Analytical Chemistry**, v. 42, p. 84-98, 2013.

SOUSA, D. G. *et al*. Aplicação de técnicas de análise exploratória no monitoramento da qualidade da água do rio Cuiá, João Pessoa-PB**. Ambiência**, v. 15, n. 1, p. 131-145, 2019.

SOUZA, A. M. de; POPPI, R. J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte I. **Química nova**, v. 35, n. 1, p. 223-229, 2012.

SOUZA, J. da C. **Algoritmo inspirado nos morcegos para seleção de variáveis em problemas de classificação.** 2023. Tese (Doutorado em Ciências) -Universidade Federal da Paraíba, João Pessoa, 2023.

SOUZA, J. da C. *et al.* Bat algorithm for variable selection in multivariate classification modelingusing linear discriminant analysis. **Microchemical Journal,** v. 187, p. 108382, 2023.

STOLL, B. B. *et al*. Análise preditiva em bases desbalanceadas e comparação de técnicas de pré-processamento–Estudo de caso MOOC. **Revista de Sistemas e Computação-RSC**, v. 10, n. 1, 2020.

STORHAUG, C. L.; FOSSE, S. K.; FADNES, L. T. Country, regional, and global estimates for lactose malabsorption in adults: a systematic review and meta-analysis. **The Lancet Gastroenterology & Hepatology**, v. 2, n. 10, p. 738-746, 2017

SURI, S. *et al*. Considerations for development of lactose-free food. **Journal of Nutrition & Intermediary Metabolism,** v. 15, p. 27-34, 2019.

TABBASSUM, M.; ZEESHAN, F.; LOW, K. H. Discrimination and recognition of Bentong ginger based on multi-elemental fingerprints and chemometrics. **Food Analytical Methods**, v. 15, p. 1-10, 2022.

TAVARES, J. P. H. e; MEDEIROS, M. L. da S.; BARBIN, D. F. Near-infrared techniques for fraud detection in dairy products: A review. **Journal of Food Science,** v. 87, n 5, p. 1943-1960, 2022.

TEIXEIRA, J. L. da P. *et al*. Rapid adulteration detection of yogurt and cheese made from goat milk by vibrational spectroscopy and chemometric tools**. Journal of Food Composition and Analysis,** v. 96, p. 103712, 2021.

TROISE, A. D. *et al*. The quality of low lactose milk is affected by the side proteolytic activity of the lactase used in the production process. **Food Research International**, v. 89, p. 514-525, 2016.

VALDERRAMA, L. *et al*. Proposta experimental didática para o ensino de análise de componentes principais. **Química Nova,** v. 39, n. 2, p. 245-449, 2016.

VALÉRIO, G. S.; COSTA, I. F.; CARDINES, P. H. F. Desenvolvimento de iogurte enriquecido com batata Yacon: uma proposta de alimento funcional**. Revista Terra & Cultura: Cadernos de Ensino e Pesquisa,** v. 38, n. especial, p. 172-182, 2022.

VARELLA, D. Intolerância à Lactose. **Biblioteca virtual em saúde,** 2018. Disponível em: < https://bvsms.saude.gov.br/intolerancia-a-lactose/>. Acesso em: 26 jun. 2024.

VITALE, R. *et al*. Class modelling by Soft Independent Modelling of Class Analogy: why, when, how? A tutorial. **Analytica Chimica Acta,** v. 1270, 2023.

WANG, L. *et al*. Quality analysis, classification, and authentication of liquid foods by near-infrared spectroscopy: A review of recent research developments. **Critical reviews in food science and nutrition**, v. 57, n. 7, p. 1524-1538, 2017.

WEI, X. *et al*. Identification of soybean origin by terahertz spectroscopy and chemometrics. **IEEE Access**, v. 8, p. 184988-184996, 2020.

XIAOBO, Z. *et al*. Variables selection methods in near-infrared spectroscopy. **Analytica chimica acta**, v. 667, n. 1-2, p. 14-32, 2010.

XU, W. *et al*. Fourier transform infrared spectroscopy and chemometrics for the discrimination of animal fur types**. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 274, p. 121034, 2022.

YANG, X. S. A new metaheuristic bat-inspired algorithm. In: **Nature inspired cooperative strategies for optimization (NICSO 2010)**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 65-74.

ZAHIR, S. A. D. M. *et al*. A review of visible and near-infrared (Vis-NIR) spectroscopy application in plant stress detection. **Sensors and Actuators A: Physical**, v. 338, p. 113468, 2022.

ZAREEF, M. *et al*. An overview on the applications of typical non-linear algorithms coupled with NIR spectroscopy in food analysis. **Food Engineering Reviews**, v. 12, p. 173-190, 2020.

ZEBIB, H.; ABATE, D.; WOLDEGIORGIS, A. Z. Nutritional quality and adulterants of cow raw milk, pasteurized and cottage cheese collected along value chain from three regions of Ethiopia. **Heliyon**, v. 9, n. 5, 2023.

ZELINKOVA, Zuzana; WENZL, Thomas. Identification of cigarette brands by soft independent modeling of class analogy of volatile substances. **Nicotine and Tobacco Research**, v. 22, n. 6, p. 997-1003, 2020.

ZHENG, X. *et al*. Non-destructive detection of meat quality based on multiple spectral dimension reduction methods by near-infrared spectroscopy. **Foods**, v. 12, n. 2, p. 300, 2023.

ZHOU, C. *et al*. Classification of waste cotton from different countries using the near-infrared technique. **Textile Research Journal**, v. 93, n. 1-2, p. 133-139, 2023.

ZHU, F. *et al*. Neighborhood linear discriminant analysis. **Pattern Recognition**, v. 123, p. 108422, 2022.

ZYLBERGELD, B. Saúde óssea. **Revista de nutrição: BIOTEC.** v. 3, n. 5, p. 30-34, 2019.