UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

DISSERTAÇÃO DE MESTRADO

Análise de Dados Físico-Químicos de Amostras de Leite Cru do Sul do Brasil Utilizando Métodos Multivariados Exploratórios e Classificatórios

LUCAS HANSEN

Porto Alegre, março de 2019

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

LUCAS HANSEN

Análise de Dados Físico-Químicos de Amostras de Leite Cru do Sul do Brasil Utilizando Métodos Multivariados Exploratórios e Classificatórios

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Química

Prof. Dr. Marco Flôres Ferrão

Orientador

Porto Alegre, março de 2019

A presente dissertação foi realizada inteiramente pelo autor, exceto as colaborações as quais serão devidamente citadas nos agradecimentos, no período entre março/2017 e março/2019 no Instituto de Química da Universidade Federal do Rio Grande do Sul sob Orientação do Professor Doutor Marco Flôres Ferrão. A dissertação foi julgada adequada para a obtenção do título de Mestre em Química pela seguinte banca examinadora:

Comissão Examinadora:

Prof. Dr. Adriano de Araújo Gomes

Prof. Dr. Jorge Otávio Trierweiler

Prof. Dr. Paulo Augusto Netz

Prof. Dr. Marco Flôres Ferrão
Orientador

Lucas Hansen

AGRADECIMENTOS

Agradeço aos meus pais e ao meu irmão, e também à minha companheira Ana, pelo apoio.

Agradeço também ao professor Marco F. Ferrão pela orientação e aos colegas do Laboratório de Quimiometria e Instrumentação Analítica pelo conhecimento trocado.

Por fim, dedico este trabalho ao meu filho Gabriel.

PRODUÇÃO CIENTÍFICA

Artigos publicados em periódicos:

1. Hansen, L.; Ferrão, M.F. Identification of Possible Milk Adulteration Using Physicochemical Data and Multivariate Analysis. *Food Anal. Methods,* **2018**, *11*, *7*, 1994–2003.

2. Hansen, L.; Ferrão, M.F. Classification of Milk Samples Using CART. *Food Anal. Methods,* **2019**, *XX*, 1-8.

Apresentações de trabalhos:

Identification of Possible Milk Adulteration using PCA. III Winter School of Chemometrics. Araraquara-SP, Brasil, julho de 2017.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

ABREVIATURAS

A – Taxa de Acerto

AG – Algoritmo Genético

CART – Árvore de Classficação e Regressão (do inglês *Classification and Regression Tree*)

E – Taxa de Erro

EP – Especificidade

FP – Falso Positivo

FN – Falso Negativo

NV – Negativo Verdadeiro

kNN – *k* Vizinhos Mais Próximos (do inglês k-*Nearest Neighbors*)

KSOM – Mapa (ou Rede) Auto-Organizável de Kohonen (do inglês *Kohonen Self-Organizing Maps*)

PC – Componente Principal (do inglês *Principal Component*)

PCA – Análise de Componentes Principais (do inglês *Principal Components Analysis*)

PLS-DA – Mínimos Quadrados Parciais com Análise Discriminante (do inglês Partial Least Squares with Discriminant Analysis)

PV – Positivo Verdadeiro

S – Sensibilidade

SIMCA – Modelagem Independente Flexível por Analogia de Classes (do inglês *Soft Independent Modelling by Class Analogy*)

SVM – Máquinas de Vetores de Suporte (do inglês *Support Vector Machines*)

RESUMO

O leite é um dos alimentos mais amplamente consumidos no mundo, tendo sua produção global sido de 723 milhões de toneladas na segunda metade da década passada. Porém, para aumentar a rentabilidade com a venda do produto, alguns atores na sua cadeia de produção podem adulterá-lo, alterando sua composição química e consequentemente diminuindo a qualidade nutricional do leite. Dessa maneira, a qualidade desse alimento é avaliada através dos resultados de análises físico-químicas como extrato seco total, nitrogênio total pelo método Kjeldahl (proteínas), cinzas, acidez, gordura, açúcares redutores totais em lactose, depressão do ponto de congelamento (crioscopia) e densidade relativa. No presente trabalho, foi utilizada a Análise de Componentes Principais (PCA) e métodos supervisionados (PLS-DA, SIMCA, kNN, SVM, KSOMs e CART) para a análise exploratória de dados físico-químicos dessas amostras e classificação das mesmas como conformes ou não conformes aos padrões estabelecidos no Regulamento brasileiro de Inspeção Industrial e Sanitária de Produtos de Origem Animal. Os resultados de classificação para o conjunto de teste relativos à taxa de acerto variaram de 72 a 98%, à especificidade para as amostras não-conformes variaram de 76 a 100% e aqueles relativos à sensibilidade para as amostras não-conformes variaram entre 67 e 97%. O método que demostrou melhor desempenho de classificação entre todos os métodos testados, levando em consideração as figuras de mérito avaliadas e o número de amostras não-classificadas, foi o PLS-DA. O PCA mostrou que todos os parâmetros legais utilizados para aferir a qualidade do leite cru são pertinentes de serem analisados, porém apontou possíveis deficiências na legislação brasileira no que tange ao estabelecimento de parâmetros para a qualidade do leite cru, o que foi corroborado pelo diagrama CART resultante.

ABSTRACT

Milk is one the most widely consumed foods in the world, with an average annual production of 723 million tons in the second half of the past decade. However, to increase milk's profitability, some actors in the dairy chain may adulterate it, altering its chemical composition and reducing the nutritional value of this food. The quality of milk is therefore assessed through chemical and physical analyses such as total solids, total Kjeldahl nitrogen, ash, acidity, fat, reducing sugars, depression of freezing point and relative density. In this work, we have used Principal Components Analysis (PCA) and supervised methods (PLS-DA, SIMCA, kNN, SVM, KSOMs and CART) to explore physicochemical data from milk and to classify and discriminate between samples that were compliant or not to the parameters set in the Brazilian Regulation for the Inspection of Animal Products. Classification results for the test set regarding non-error rate ranged between 72 and 98%, and those regarding specificity towards noncompliant samples ranged from 76 to 100%. Finally, results regarding sensitivity towards noncompliant samples ranged from 67 to 97%, with PLS-DA as the method that presented the best classification performance among all the tested methods, when the evaluated figures of merit and number of non-assigned samples are taken into account. PCA confirmed the pertinence of the parameters set in Brazilian regulation in assessing the quality of raw milk, however pointed towards possible deficiencies in Brazilian regulations regarding the establishment of parameters for the quality of raw milk, an observation corroborated by the resulting CART diagram.

1 INTRODUÇÃO

O leite é um produto de extrema importância econômica e social em praticamente todos os países do mundo, provendo sustento direto para cerca de 1 bilhão de pessoas no planeta no início da década, com uma produção de cerca de 750 milhões de toneladas nesse mesmo período; quase 621 milhões de toneladas dessa quantidade são correspondentes a leite de vaca[1].

Dada a sua importância econômica, o leite está sujeito a adulteração por diversos atores na sua cadeia de produção, como os produtores, transportadores ou beneficiadores. A adulteração é o ato de alterar intencionalmente a composição do leite destinado a venda, ou deliberadamente vender o mesmo em condições deterioradas, resultando em um produto de qualidade inferior do ponto de vista nutricional, ou mesmo nocivo à saúde. A adulteração pode resultar diretamente da ação humana, como no caso da adição de substâncias estranhas, substituição de constituintes por outros de qualidade inferior ou remoção de constituintes, como gordura[2]. Pode também ser resultante da ação de fatores alheios à vontade humana, como a ação de microrganismos que causa a acidificação do leite.

A adulteração do leite é um problema antigo, bem como a procura por métodos que pudessem auxiliar na sua detecção: como exemplo, desde o final do século 19 já se usava nos Países Baixos a depressão do ponto de congelamento do leite como meio de detectar fraudes na composição do mesmo[3]. Nos dias de hoje, porém, existem vários métodos disponíveis para esse fim. Entre eles encontram-se os métodos quimiométricos, que fazem uso das ferramentas da matemática, estatística e informática para a solução de problemas químicos.

Nesse contexto, é proposto no seguinte trabalho a utilização de ferramentas quimiométricas, mais especificamente métodos de análise exploratória (não-supervisionados) e métodos de classificação (supervisionados) para a detecção de possíveis adulterações na composição do leite com base em dados físico-químicos de amostras de leite cru, e a

classificação dessas amostras como conformes ou não-conformes aos padrões estabelecidos na legislação brasileira para esse produto.

Neste trabalho pretende-se comparar métodos quimiométricos na classificação de dados físico-químicos de leite cru, obtidos por métodos analíticos amplamente utilizados na avaliação da qualidade do leite e tidos como altamente confiáveis ou de referência. O que se busca é estabelecer um padrão de comparação para futuros trabalhos que venham a avaliar o desempenho dos classificadores aqui testados sobre dados oriundos, por exemplo, de métodos como a espectroscopia de infravermelho próximo ou de infravermelho médio, ou para trabalhos que venham a avaliar o desempenho de outros métodos de classificação sobre o mesmo tipo de dados utilizado neste trabalho.

2 OBJETIVOS

2.1 OBJETIVOS GERAIS

São os objetivos gerais desse trabalho realizar o estudo exploratório e a avaliação da conformidade de amostras de leite cru da região Sul do Brasil, mais especificamente dos estados do Rio Grande do Sul e de Santa Catarina, com base nos dados físico-químicos como extrato seco total, nitrogênio total pelo método Kjeldahl (proteínas), cinzas, acidez, gordura, açúcares redutores totais em lactose, depressão do ponto de congelamento (crioscopia) e densidade relativa dessas amostras.

2.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos do presente trabalho são:

1) realizar estudo exploratório de amostras de leite cru através do método de análise de componentes principais (PCA), com o objetivo de identificar possíveis padrões físico-químicos nessas amostras.

2) classificar essas mesmas amostras como conformes ou não conformes aos padrões físico-químicos brasileiros de conformidade para o leite cru, lançando mão de métodos de multivariados, quais sejam, o método PLS-DA, SIMCA, kNN, SVM, CART e o método dos mapas ou redes auto-organizáveis de Kohonen (KSOMs).

3) com base nos resultados encontrados para os métodos multivariados avaliar a pertinência dos parâmetros adotados pela legislação vigente no que tange a conformidade ou não do leite cru bovino.

3 REVISÃO BIBLIOGRÁFICA

3.1 LEITE

Leite é o produto oriundo da ordenha completa, ininterrupta, em condições de higiene, de vacas sadias, bem alimentadas e descansadas[4]. Embora seja composto principalmente por água, consiste em um alimento extremamente complexo: estima-se que o leite tenha em torno de 100.000 constituintes distintos, a maioria deles ainda não identificada[5]. Tem sido descrito como um dos alimentos naturais mais próximos da perfeição, devido ao seu elevado teor de nutrientes[6,7]. A tabela a seguir apresenta a composição centesimal média do leite de vaca e a sua variação:

Tabela 1: Composição centesimal média do leite de vaca[5]

Constituinte	Teor % (m/m)	Variação % (m/m)
Água	87,30	85,5 – 88,7
Extrato seco desengordurado	8,80	7,9 – 10,0
Gordura	3,90	2,4 – 5,5
Lactose	4,60	3,8 – 5,3
Proteínas	3,25	2,3 – 4,4
Substâncias Minerais	0,65	0,53 – 0,80

A água é o componente quantitativamente mais importante do leite, nela estando dissolvidos, dispersos ou emulsionados todos os demais. A maior parte encontra-se como água livre, no entanto existe também água ligada às proteínas e água de solvatação dos minerais e da lactose.

A gordura no leite se encontra emulsionada em glóbulos envoltos por uma membrana lipoproteica[8]. Consiste principalmente em triacilgliceróis com 437 variedades de ácidos graxos, sendo os principais o ácido palmítico, de cadeia C_{16}, e o ácido oleico, monoinsaturado de cadeia C_{18}. A gordura é o componente do leite que mais sofre variações em razão de alimentação, raça e período de lactação do animal e estação do ano[5].

Em relação aos compostos nitrogenados presentes no leite bovino, 95% ocorrem como proteínas e 5% como compostos não-proteicos. Do nitrogênio proteico, 80% é oriundo de caseínas e 20% de proteínas não-caseínicas.

A lactose é o glicídio característico do leite. É formada por glicose e galactose, sendo o constituinte sólido mais abundante e menos variável no alimento. Tratamentos térmicos podem causar escurecimento do leite devido especialmente à reação de Maillard sofrida pela lactose, com diminuição do valor nutricional.

Os principais minerais presentes no leite são cloro, fósforo, potássio, sódio, cálcio e magnésio. Estão também presentes, embora em menor teor, ferro, alumínio, bromo, zinco e manganês[9]. Todos esses minerais formam sais orgânicos e inorgânicos, sendo a associação entre tais sais e as proteínas do leite um fator determinante para a estabilidade das caseínas. O fosfato de cálcio, especificamente, faz parte da estrutura das micelas de caseína.

No leite encontram-se também todas as vitaminas conhecidas[10], estando as lipossolúveis A, D, E e K associadas aos glóbulos de gordura e as demais na fase aquosa. Enzimas estão da mesma maneira presentes em profusão, como lipases, proteinases, óxido-redutases, fosfatases, catalases e peroxidases. O desenvolvimento de microrganismos contribui para o complexo enzimático.

3.2 ANÁLISES FÍSICO-QUÍMICAS DO LEITE

As análises físico-químicas a seguir descritas são efetuadas para a determinação da composição do leite a fim de caracterizar e controlar fraudes no mesmo, que podem ser realizadas pelo produtor ou durante o transporte.

3.2.1 Acidez

A acidez no leite tem origem na atividade de bactérias dos gêneros *Lactococcus* e *Lactobacillus*, que produz principalmente ácido lático e ácido pirúvico a partir da fermentação (hidrólise) da lactose. Outros gêneros que produzem ácidos são *Micrococcus* e *Microbacterium*, além de várias espécies do grupo coliforme. A titulação da acidez tem amplo uso na inspeção industrial e sanitária do leite, permitindo avaliar o estado de conservação e eventuais anomalias no produto. A acidez natural do leite varia de 0,12 a 0,17% de ácido lático, podendo o valor máximo chegar a 0,23% no leite fresco pela ação microbiana (acidez desenvolvida)[5,11].

3.2.2 Densidade

A densidade é simplesmente a relação entre a massa e o volume de um corpo, ou seja, d = m/V. No caso da determinação de densidade em leite, o valor dessa é relativo, sendo o quociente da divisão da massa de um volume de leite pela massa de um igual volume de água (é determinada a densidade relativa do leite). Esse parâmetro serve para controlar, até certos limites, fraudes no leite por desnatação prévia ou adição de água. A densidade média do leite a 15℃ varia entre 1,027 e 1,034 $g.mL^{-1}$.

3.2.3 Gordura

A determinação do teor de gordura é de interesse para o sistema de pagamento do leite. Os métodos clássicos para determinar esse parâmetro se baseiam na destruição da estrutura globular das gorduras no leite e a posterior quantificação da quantidade liberada.

3.2.4 Extrato seco total

O extrato seco total (EST) consiste em todos os componentes do leite menos a água. Geralmente é determinado por gravimetria. A diferença entre o extrato seco total e o teor de gordura dá o extrato seco desengordurado (ESD) do leite.

3.2.5 Índice crioscópico

O índice crioscópico, também chamado depressão do ponto de congelamento ou simplesmente crioscopia, tem por objetivo a detecção de fraudes, principalmente por adição de água. Nesse ensaio é determinada a temperatura em que o leite passa do estado líquido para o estado sólido, sendo essa a característica mais constante do produto. É medida em graus Hortvath (ºH; 1ºH = 1,03562 *x* ºC). No Brasil a legislação determina um padrão de -0,550ºH ± 0,01ºC para os leites do tipo A e B e -0, 530 a -0,560ºH para o leite do tipo C[12].

3.2.6 Lactose

A lactose é um dissacarídeo formado por uma molécula de glicose e uma de galactose. Sua determinação é importante para aferição do valor nutricional do leite e para indicar a ocorrência de processos fermentativos. É comumente

determinada pelo método de Fehling. O método se baseia na redução de íons Cu (II) a Cu (I) pelo grupamento aldeídico livre presente em C_5 da glicose que compõe a lactose.

3.2.7 Proteína

O teor proteico do leite é de grande interesse principalmente para as indústrias queijeiras, e de produtos fermentados e concentrados. O método mais utilizado é o de Kjeldahl e a sua variante micro-Kjeldahl. O método é baseado na destruição da matéria orgânica através da digestão da amostra em ácido sulfúrico, ocorrendo a conversão do nitrogênio na amostra a íons amônio. Os íons amônio são por sua vez convertidos a amônia pela adição de hidróxido, e a amônia assim liberada é destilada em ácido bórico e titulada. O teor de N obtido nesse ensaio é multiplicado por um fator de 6,38 (no caso do leite) para a expressão do resultado em teor de proteínas.

3.2.8 Resíduo mineral fixo (Cinzas)

O resíduo mineral fixo é uma estimativa do teor de minerais na amostra de leite. O método consiste em dessecar a amostra e submetê-la a altas temperaturas (geralmente 550℃), de maneira que a matéria orgânica seja volatilizada na forma de CO_2 e H_2O, restando as cinzas, que uma vez pesadas dão o valor do resíduo mineral fixo.

3.3 MÉTODOS EXPLORATÓRIOS E SUPERVISIONADOS

Nesta seção serão descritos os métodos não-supervisionados ou exploratórios utilizados no presente trabalho, bem como os métodos supervisionados. Não é o objetivo deste texto prover uma descrição detalhada, em termos matemáticos e computacionais, dos métodos quimiométricos mais consagrados, como o PCA, kNN, SIMCA, PLS-DA e SVM. Métodos menos

conhecidos, como os mapas auto-organizáveis de Kohonen e as CART, serão descritos em maior detalhe nas próximas sub-seções.

3.3.1 Métodos não-supervisionados

3.3.1.1 PCA

A Análise por Componentes Principais (PCA, do inglês *Principal Component Analysis*) foi introduzido por Pearson em 1901[13], embora o primeiro relato de sua utilização tenha sido publicado por Hotelling mais de 30 anos depois, em 1933[14]. É um método exploratório, não-supervisionado, utilizado para projetar um conjunto de dados multivariados em um espaço de dimensão menor, sem afetar as relações entre as amostras. Assim, as informações relevantes são separadas e amplificadas, tornando possível descobrir, visualizar e interpretar as diferenças existentes entre as variáveis e examinar as relações que possam existir entre as amostras, bem como identificar amostras discrepantes do conjunto de dados[15]. Detalhes sobre o método e exemplos de aplicações na área da Química podem ser encontrados em artigos de revisão por Wold *et al.*[16] e por Bro e Smilde[17], para citar apenas alguns.

3.3.2 Métodos supervisionados

3.3.2.1 PLS-DA

O método dos mínimos quadrados parciais (PLS, do inglês *Partial Least Squares*), amplamente conhecido e utilizado em diversas ciências, não foi concebido como um método de classificação, mas sim regressão. Tem sido usado como ferramenta de regressão na Quimiometria há mais de 30 anos[18]. Como método de classificação, é denominado de método de mínimos quadrados parciais com análise discriminante (PLS-DA, do inglês *Partial Least Squares with Discriminant Analysis*), tendo seu primeiro uso sido relatado há mais de vinte anos atrás[19]. Foi formalmente introduzido há 15 anos por Barker e Rayens[20]. O

método consiste em construir um separador linear, ou uma fronteira de decisão linear, que divide um espaço de variáveis latentes em duas regiões, permitindo discriminar entre duas (ou mais) classes amostrais distintas[21]. O separador é definido pela sua posição e sua inclinação. A Figura 1 abaixo ilustra um separador linear criado pelo PLS-DA para discriminação entre duas classes baseada em duas variáveis latentes:

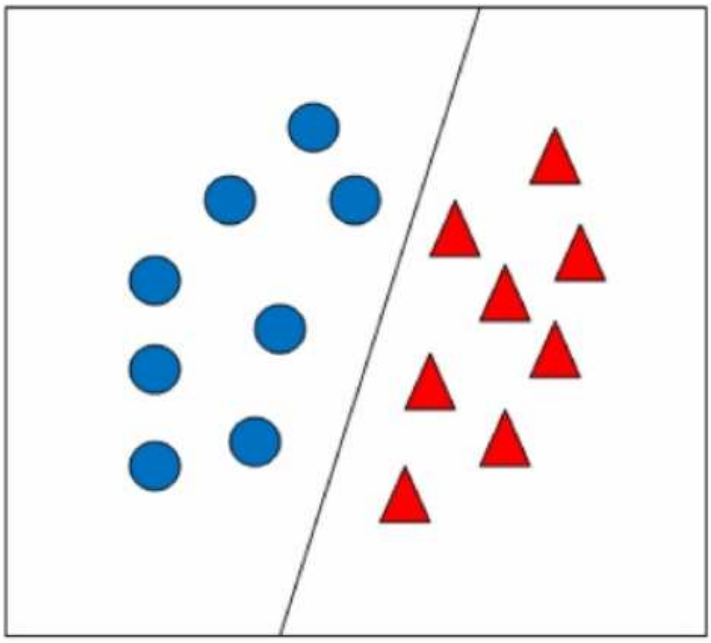

Figura 1: Separador de classes linear para classificação binária[21].

3.3.2.2 SVM

As máquinas de vetores de suporte, conhecidas pela sigla SVM (*Support Vector Machines*), são uma versão generalizada de classificadores lineares[22], como o PLS-DA. Foram introduzidos no final da década de 70 por Vapnik[23], embora tenham ganhado atenção apenas na metade da década de 90[24] Funcionam tanto para problemas de classificação nos quais apenas duas classes estão presentes quanto para problemas de classificação multiclasse, embora tenham sido concebidas inicialmente para classificação binária. As SVM traçam uma fronteira de decisão de dimensão superior àquela traçada por um separador linear, podendo ela ser também linear ou não. Assim, uma fronteira

de decisão construída por uma SVM é chamada de *hiperplano*, podendo esse ser representado por uma combinação de linear de funções parametrizadas por vetores de suporte, que consistem em amostras do conjunto de treinamento que se situam próximas à fronteira de decisão. A regra de classificação é derivada desses vetores de suporte ou amostras. Durante a otimização, o algoritmo das máquinas de vetores de suporte busca uma máxima margem de separação dentre todas as possíveis entre as classes de amostras presentes, onde a margem é a distância entre o hiperplano e a amostra mais próxima para cada classe[25]. Devido a isso, são incluídas na família de classificadores ditos *de máxima margem*.

No caso de as amostras não serem linearmente separáveis, as coordenadas das amostras no espaço de entrada (Figura 2(a)) são mapeadas através de uma função Φ em um espaço de dimensão superior chamado de espaço característico, onde elas podem ser linearmente separadas por um hiperplano H[26] (Figura 2(b)), fazendo com que a ideia geral do método se assemelhe, de fato, à dos separadores lineares.

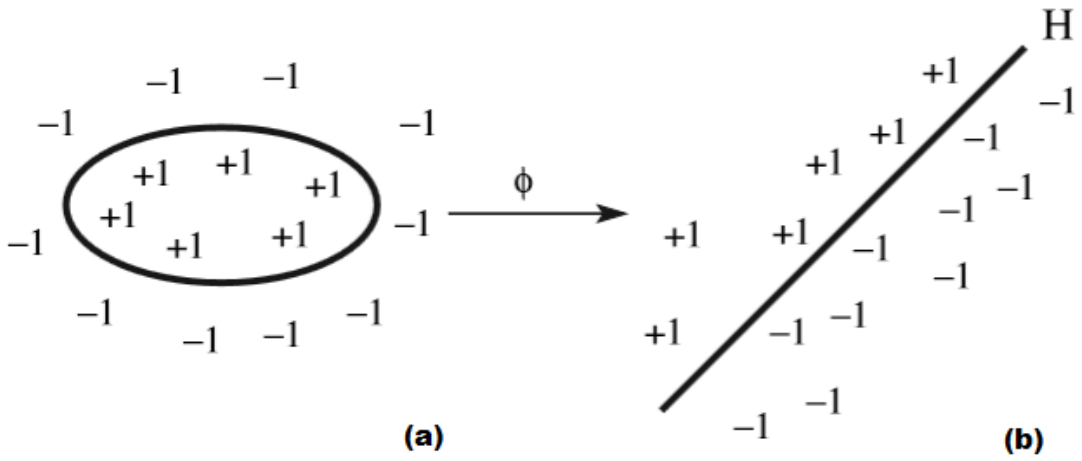

Figura 2: Separação linear de classes para classificação binária por SVM através de mapeamento das amostras no espaço característico. Adaptada de Ivanciuc, 2007.[26]

3.3.2.3 SIMCA

O método SIMCA (*Soft Independent Modelling by Class Analogy* ou modelagem independente flexível por analogia de classes) foi introduzido por Wold e Sjostrom em 1977[27]. Foi um método especialmente concebido para a solução de problemas químicos. É um método chamado de “flexível” pelo fato de haver a possiblidade de sobreposição de classes no espaço compreendido pela amostras[28], ou seja, não se supõe que as amostras pertençam a algum tipo de distribuição estatística[25]. É também dito “independente” pelo fato de ser construído um modelo para cada classe, de maneira independente, havendo a possiblidade de serem introduzidos novos modelos sem interferir nos modelos já construídos[29]. Basicamente, o modelo SIMCA consiste em um conjunto de modelos de PCA (análise de componentes principais, apresentada anteriormente, seção **3.3.1.1**), um modelo para cada classe presente nas amostras, conforme já explicado. Cada modelo pode ter um número ótimo de componentes principais próprio, a depender das particularidades das amostras pertencentes às diferentes classes. Para fins de designar uma amostra à uma determinada classe, a mesma é projetada sobre cada modelo criado e comparada a ele de maneira a avaliar a distância geométrica da amostra ao modelo. A amostra é designada a uma classe de acordo com uma regra de decisão, que pode ser, por exemplo, probabilística, como a posição em uma região no subespaço de cada modelo na qual se tem 95% de confiança que a amostra pertença à classe em questão. A Figura 3, adaptada de Vandengiste *et al.*[30] ilustra dois exemplos de modelos calculados com base em uma (Fig. 3 (a)) e duas (Fig. 3(b)) componentes principais, encerrando uma região cilíndrica e outra prismática retangular, respectivamente.

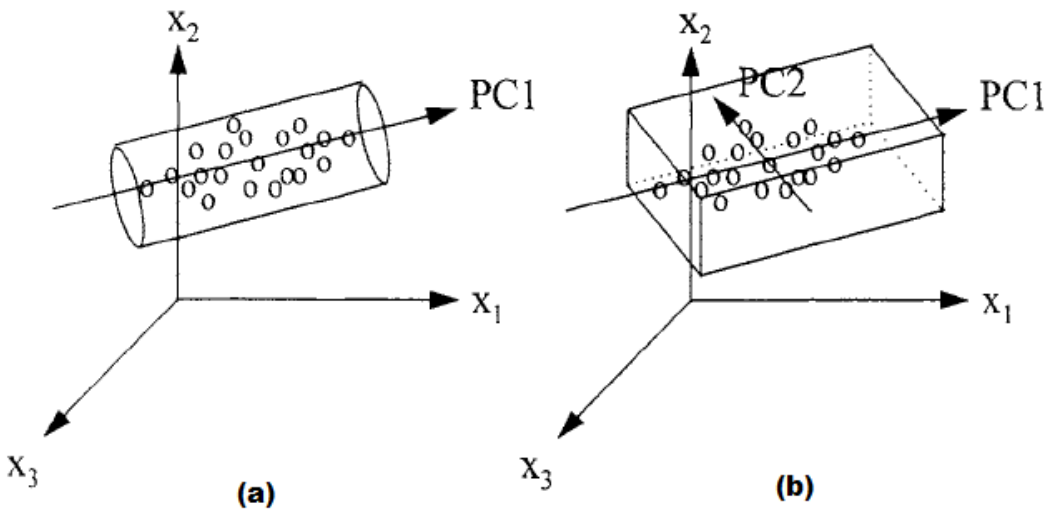

Figura 3: Modelos SIMCA calculados com base em uma (**a**) e duas (**b**) componentes principais[30]. Adaptada de Vandengiste *et al.*, 1998. [30]

3.3.2.4 kNN

O método dos *k* vizinhos mais próximos, mais conhecido pela sua sigla kNN (*k-Nearest Neighbors)* foi introduzido por Cover e Hart em 1966[31]. É um método conceitualmente e computacionalmente muito simples[25], que consiste basicamente em classificar uma amostra de acordo com as classes dos seus *k* vizinhos mais próximos. É computada a distância da amostra a cada um desses vizinhos para determinar quais são os mais próximos, e a classe à qual cada um pertence é equivalente a um "voto" para classificar a amostra desconhecida. Em caso de empate, são atribuídos pesos maiores ao vizinhos sucessivamente mais próximos, de modo que seja atribuído um valor de classe ao objeto ou amostra desconhecida. O procedimento do algoritmo do kNN pode ser sequencialmente descrito como[28]: 1) com base em um conjunto de treinamento cujas classes das amostras são conhecidas, calcular a distância (euclidiana, por exemplo) de uma amostra de classe desconhecida em relação a todas as amostras no conjunto de treinamento; 2) ordenar as distâncias de acordo com algum índice (por exemplo, atribuir o valor 1 à menor distância de todas); 3) selecionar as *k* menores distâncias e determinar a qual classe a amostra desconhecida pertence, com base no número de "votos" dados utilizando como critério de decisão essas distâncias. A Figura 4, onde x_1 e x_2 são variáveis, a mostra de classe

desconhecida é representada por um círculo e as amostras de treino pertencentes a diferentes classes são marcadas por quadrados e losangos, ilustra o processo de decisão por voto:

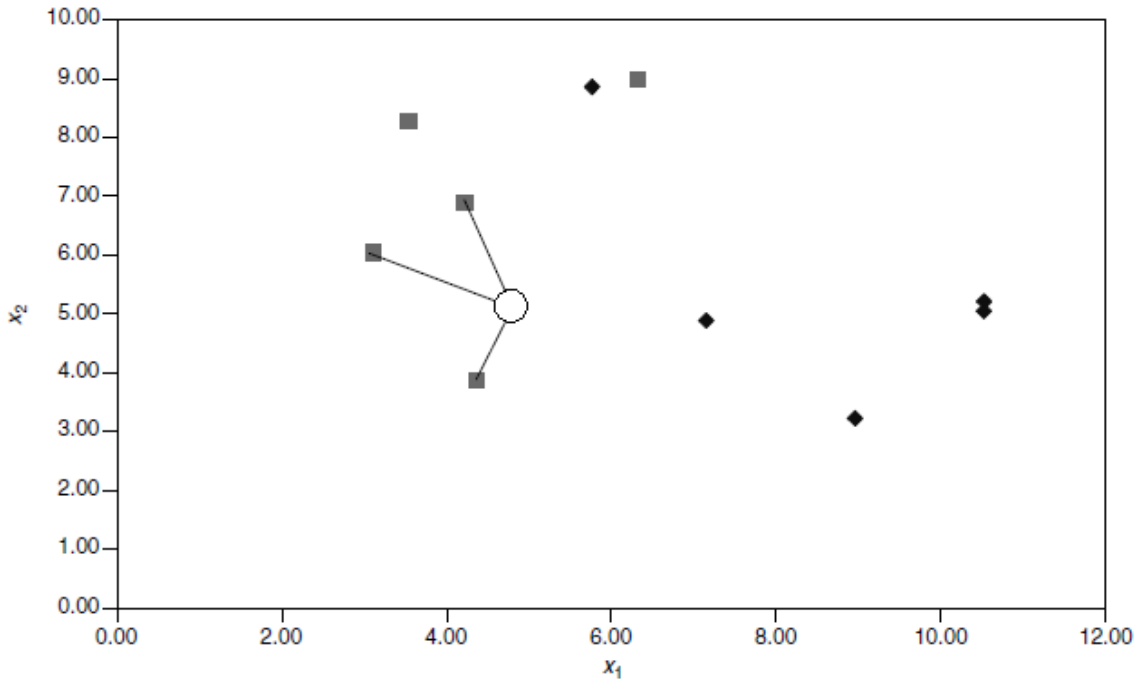

Figura 4. Ilustração do critério de votos dos *k* vizinhos mais próximos para atribuição de classe a amostra de classe desconhecida no método kNN. Adaptada de Brereton, 2003.[28]

3.3.2.5 Mapas auto-organizáveis de Kohonen

Mapas auto-organizáveis de Kohonen são redes neurais que imitam o aprendizado biológico. Essa ferramenta foi introduzida por Kohonen em 1981[32,33] e publicada na formada de artigo pouco depois[34]. Foi inicialmente proposta para modelar um fenômeno biológico chamado *retinotopia*, a auto-organização de ligações neurais entre o córtex visual e as células da retina quando as últimas são excitadas por sucessivos estímulos independentes[35]. Conceitualmente, os mapas de Kohonen são um método que agrupa dados e os relaciona por similaridade[36]. Em outras palavras, são um método que combina as características de métodos de aprendizado tanto supervisionados como não-

supervisionados, sendo capaz de executar o agrupamento de dados multivariados e de criar modelos de regressão e classificação[37].

Mais precisamente, os mapas ou redes auto-organizáveis de Kohonen (da sigla em inglês, KSOMs – *Kohonen Self-Organizing Maps*) são um tipo específico de redes neurais[37]. Redes neurais são constructos matemáticos e computacionais cuja estrutura consiste em "camadas". Uma arquitetura de três camadas é a mais comum para redes neurais, a primeira sendo a camada de entrada, cujo número de unidades ou neurônios é o mesmo que o número de variáveis (e possivelmente uma unidade adicional, chamada de "unidade de viés; isto é, a primeira camada é constituída basicamente pelas variáveis de um experimento e seus respectivos valores medidos, mais uma unidade de viés comumente constituída de neurônios com o valor 1 (a qual pode ser omitida, então a camada de entrada pode ser um espaço de dimensão *N* ou *N+1*). A segunda camada é chamada de camada "oculta" (a qual pode possuir unidades de viés também), onde os dados da camada de entrada são processados, e a terceira camada é a camada "de saída", a qual processa as entradas supridas pela camada oculta e retorna os resultados das computações dos dados iniciais, aqueles da camada de entrada.

As redes de Kohonen possuem uma arquitetura de duas camadas, ou seja, uma camada de entrada e outra de saída; não há camada oculta. Uma característica importante das redes de Kohonen é a redução de dimensionalidade: um conjunto de dados multi-dimensional (suprido pela camada de entrada) é projetado de maneira não-linear em um espaço de dimensionalidade reduzida (a camada de saída processa os dados *N*- ou *N+1*-dimensionais e os projeta em um mapa bidimensional), fazendo com que a ideia subjacente às redes de Kohonen seja a mesma de outros métodos exploratórios[38], como o PCA. Porém, como já mencionado, os mapas de Kohonen podem também ser utilizados como ferramentas classificatórias.

No que diz respeito aos aspectos funcionais dos KSOMs, Kohonen[39] ilustra seus processos matemáticos e de aprendizagem no que pode ser resumido como:

1. Cada amostra, que pode ser representada por um vetor, seleciona uma unidade de melhor ajuste (*Best Matching Unit*, ou BMU), que também pode ser representada como um vetor, na camada de saída, dentre um conjunto de vetores-modelo (ou unidades-modelo). Os vetores-modelo são criados na fase de treinamento do mapa de Kohonen, e representam nodos específicos na camada de saída. Aquele que apresentar a menor distância em relação ao vetor de dados da amostra corresponde à unidade de melhor ajuste, e é chamado de *neurônio vencedor*.
2. Neurônios (vetores-modelo) na vizinhança do neurônio vencedor são rearranjados de maneira que outros neurônios com distâncias sucessivamente maiores ao vetor de dados ocupem a vizinhança desse. O processo é repetido até que seja atingida uma configuração estável, definida em termos do erro de quantificação (o qual será apresentado a seguir).

O processo de quantificação e minimização de distância, acima descrito, pode ser matematicamente formulado de acordo com a Equação 1:

$$d = \text{argmin}_i \, || \mathbf{x} - \mathbf{m_i} || \qquad (1)$$

Onde:

d = distância euclidiana mínima

x = vetor de dados no espaço N-dimensional

$\mathbf{m_i}$ = *i*-ésimo vetor-modelo, também no espaço N-dimensional

O vetor-modelo ***m*** cuja distância euclidiana em relação ao vetor de dados ***x*** é a menor entre todos os vetores do conjunto de vetores-modelo corresponde à unidade de melhor ajuste, ou neurônio vencedor. A distância entre os neurônios vencedores e os neurônios de dados é utilizada para computar o *erro de quantificação*, de acordo com a Equação (2):

$$\int_V || \mathbf{x} - \mathbf{m_c} ||^2 \, . \, p(\mathbf{x}) . \, dV \qquad (2)$$

Onde:

x = vetor de dados

$\mathbf{m_c}$ = *c*-ésimo neurônio vencedor

p(**x**) = função densidade de probabilidade de **x**

dV = diferencial de volume para o espaço de dados (N-dimensional)

O mapa com o menor erro de quantificação é aquele que possui configuração ótima. O processo de rearranjo de neurônios descrito anteriormente é realizado de maneira a minimizar o erro de quantificação. Não existe solução linear para a Equação (2), porém foi demonstrado por Kohonen que ela converge para mínimos locais[40], portanto foi provado matematicamente que é possível encontrar uma configuração ótima para um certo mapa de Kohonen, embora ela possa ser ou não a melhor dentre o conjunto de todas as configurações possíveis.

Computacionalmente, o algoritmo das redes de Kohonen (como inicialmente proposto), lembra o procedimento de otimização chamado de gradiente descendente. Tal procedimento treina pesos em um modelo de acordo com o descrito na Equação (3):

$$\mathbf{w_i}\,(t+1) = \mathbf{w_i}\,(t) - [1/(2.N)].\ \alpha.\ \frac{d}{dw}\sum (\mathbf{w_i}(t)\ \mathbf{x}_i - \mathbf{y_i})^2 \qquad (3)$$

Onde:

$\mathbf{w_i}$ = vetor ou matriz de pesos

t = *t*-ésima iteração do algoritmo

N = número de exemplos (amostras) no conjunto de treino

α = taxa de aprendizado (tamanho do “passo” dado pelo algoritmo a cada iteração)

$\mathbf{x_i}$ = vetor ou matriz de dados correspondente a uma variável independente

$\mathbf{y_i}$ = vetor ou matriz de valores para a variável categórica y

Ou seja, o gradiente descendente tenta minimizar a distância entre os valores de saída calculados $\boldsymbol{w_i x_i}$ e os valores de saída observados $\boldsymbol{y_i}$, e a cada iteração ajusta o valor de cada peso até convergir, i.e., treina um modelo para os dados cujo erro de predição é o menor possível.

De maneira semelhante, os mapas de Kohonen tentam minimizar a distância entre cada vetor de dados e cada vetor-modelo, de acordo com a equação (4):

$$\mathbf{m_i}(t + 1) = \mathbf{m_i}(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m_i}(t)] \qquad (4)$$

Onde:

$\mathbf{m_i}$ = *i*-ésimo vetor-modelo

t = *t*-ésima iteração (chamada de *época*) do algoritmo

x = vetor de dados

$h_{ci}(t)$ = função de vizinhança

A função de vizinhança $h_{ci}(t)$ é uma função implícita da taxa de aprendizado α, que por sua vez também é função de t nesse caso. O índice *c* se deve ao fato de a função ser calculada para um vetor $\boldsymbol{m_c}$ (correspondente ao neurônio vencedor ou unidade de melhor ajuste), cuja distância ao vetor de dados é a mínima, de acordo com a Equação (4). A Equação (5) exemplifica uma função de vizinhança:

$$h_{ci}(t) = \alpha(t). \exp[-\text{sqdist}(c, i)/2\,\sigma^2(t))] \qquad (5)$$

Onde:

$h_{ci}(t)$ = função de vizinhança

t = *t*-ésima iteração (época) do algoritmo

α(t) = função monotonicamente decrescente de t (linear, exponencial, etc.) para a taxa de aprendizado α

sqdist(c, i) = distância quadrática entre os neurônios c (neurônio vencedor) e o neurônio i em sua vizinhança

$\sigma^2(t)$ = função monotonicamente decrescente de t

Dessa maneira, a taxa de aprendizado decresce com o número de iterações do algoritmo (o mesmo ajuste pode ser feito para o gradiente descendente). Os vetores-modelo são atualizados a cada iteração até que suas respectivas distâncias aos vetores de dados sejam minimizadas. O algoritmo não apenas atualiza os vetores-modelo (equivalentes às respostas calculadas no gradiente descendente), como também rearranja os neurônios com respostas semelhantes a uma variável de maneira a minimizar a distância entre eles, alterando dessa maneira a configuração inicial do mapa para a sua configuração (localmente) ótima, ou até sua convergência.

3.3.2.5.1 Algoritmo Genético

Para encontrar a arquitetura ideal e o número ótimo de iterações para os mapas de Kohonen, foi utilizado no presente trabalho o chamado Algoritmo Genético. Tal algoritmo consiste em um poderoso método de otimização que segue a ideia da seleção natural: soluções progressivamente melhores evoluem de soluções anteriores até que uma solução próxima da ótima seja alcançada[41]. Mais especificamente, o Algoritmo Genético é um algoritmo de busca probabilístico baseado na mecânica da seleção natural e da genética natural. O algoritmo é iniciado com um conjunto de soluções chamado de população, que permanece constante em número durante a execução do algoritmo. A cada geração (iteração), cromossomos (soluções) são aleatoriamente selecionados com base em seus respectivos valores de aptidão; cromossomos com altos valores de aptidão tem uma alta probabilidade de serem selecionados. Assim, cromossomos da próxima geração podem apresentar valores mais altos de aptidão do que aqueles da geração anterior, em um processo que imita a evolução biológica. Copulação e mutação aleatórias podem também ocorrer. O processo de evolução é repetido até que uma condição de terminação seja

satisfeita[42]. De acordo com Mukhopadhyay *et al.*[43], o algoritmo genético pode ser representado em pseudo-código da seguinte maneira:

INÍCIO

Computar população inicial B_0;

ENQUANTO condição de terminação não for cumprida FAÇA

Selecionar indivíduos para reprodução;

Criar filhos pelo cruzamento de indivíduos;

Eventualmente mutar indivíduos;

Computar nova geração

FIM

3.3.2.6 Árvores de Classificação e Regressão

Uma árvore de classificação e regressão (CART ou *Classification and Regression Trees*, na sigla em inglês) é um modelo de previsão que pode ser representado por uma árvore, como obviamente descrito pelo nome[44]. As árvores de classificação e regressão foram introduzidas por Breiman *et al.* em 1984[45], embora o algoritmo inicial para sua construção, que permitia apenas criar modelos de regressão, foi concebido por Morgan e Sondquist cerca de 20 anos antes[46]. Adaptações do método por Messenger e Mandel em 1973[47] permitiram que as árvores fossem utilizadas também para classificação.

Matematicamente, o método consiste em mapear um ponto de dados $\mathbf{x}_i$ a um ponto de dados de treinamento $\mathbf{y}_i$ via uma função d($\mathbf{x}$). O critério para a escolha de d($\mathbf{x}$) é normalmente o erro quadrático médio de previsão $E\{d(\mathbf{x}) - E(\mathbf{y}|\mathbf{x})\}^2$, onde, no caso da classificação, E($\mathbf{y}$|$\mathbf{x}$) é o custo esperado para a classificação equivocada de uma amostra. O algoritmo desse método consiste em criar uma grande árvore, e então “podá-la” a um tamanho que possua o menor erro estimado possível na validação cruzada[44]. Após crescer a árvore ao

seu tamanho máximo, o algoritmo remove todas as ramificações produzidas anteriormente, deixando apenas duas delas, as quais não aumentam a exatidão da árvore em classificar os dados do conjunto de treinamento. A poda começa desse ponto, gerando uma sequência de sub-árvores, e iterativamente removendo as ramificações que menos contribuem para o desempenho de classificação dos dados de treinamento. A Figura 5 mostra um diagrama de uma CART, adaptado de um artigo publicado por Loh[44] onde foi utilizado o conjunto de dados de flores do gênero iris, por sua vez utilizado por Fisher em 1936[48] para a introdução da Análise de Discriminantes Linear:

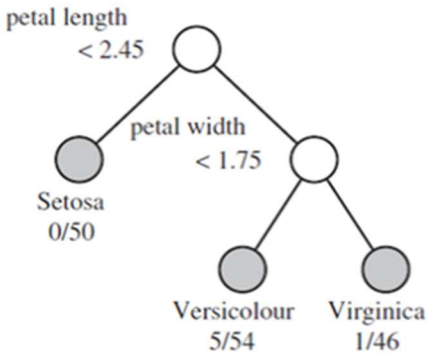

Figura 5: Diagrama CART para o conjunto de dados “Iris”[44]

A cada nodo intermediário, uma observação se dirige à esquerda se, e somente se, a condição apresentada for verdadeira (por exemplo, na Fig.1, um espécime de iris seria classificado como pertencente à espécie *Setosa* se e somente sua largura de pétala (petal width) for menor que 1,75 unidades de medida). Fisher originalmente mediu quarto variáveis: largura de pétala (petal width), comprimento de pétala (petal length), largura de sépala (sepal width) e comprimento de sépala (sepal length), porém apenas duas dessas variáveis estão presentes na árvore. Essa é uma característica do algoritmo: se duas ou mais variáveis estiverem fortemente correlacionadas, no máximo uma delas aparecerá no diagrama final[49]. Isso pode ser problemático e levar a conclusões espúrias, o que é chamado de *mascaramento* das variáveis.

3.3.3 Utilização de Métodos Multivariados na Análise de Dados Físico-Químicos

A seguir, serão descritos trabalhos que fizeram uso dos mesmos métodos multivariados aqui empregados para a análise de dados físico-químicos de leite fluido e outras matrizes alimentícias.

Liu[50] utilizou o método SVM-DA para detectar a adulteração de amostras de leite em pó por adição de gordura ou desnatamento, utilizando dados de espectroscopia de frequência na região dos terahertz. Todas as amostras no conjunto de treinamento (*n* = 40) foram corretamente classificadas pelo SVM-DA, com especificidade de até 100% e sensibilidade de até 98,62% para as amostras de leite em pó desnatado (desengordurado), porém não foram relatados os resultados de classificação para o conjunto de teste. Bougrini *et al.*[51] utilizou o mesmo método para classificar amostras de leite UHT de diferentes marcas com diferentes períodos de armazenamento. O método SVM classificou 100% das amostras do grupo de teste de uma marca de leite corretamente e 96,67% de outra marca de acordo com os dias de armazenamento (1-5 dias, 5 classes). Os dados foram adquiridos através de um nariz eletrônico (*e-nose*) multissensor e uma língua eletrônica (*e-tongue*) voltamétrica.

Scholl *et al.*[52] conseguiram diferenciar leite em pó com adição de melamina de leite em pó sem adulteração utilizando o método SIMCA. O perfil espectral de cada amostra foi obtido por espectroscopia de infravermelho próximo (NIR). No conjunto de teste, todas as amostras sem adição e 91% (10 de 11) das amostras com adição foram corretamente classificadas. Outros dois trabalhos por Gondim *et al.*[53,54] descreveram a detecção de vários adulterantes em leite com o uso de espectroscopia na região no infravermelho médio (MID) e o método SIMCA. Levando em consideração ambos os trabalhos, a sensibilidade foi maior que 62% e a especificidade maior que 70% para os conjuntos de treinamento. Até 24,1% das amostras foram inconclusivamente classificadas. O mesmo método supervisionado foi empregado por Santos *et al.*[55] em amostras de leite fluido analisadas por microespectroscopia no infravermelho com o objetivo de detectar e classificar amostras adulteradas com diversas

substâncias. A porcentagem de amostras corretamente classificadas no conjunto de treinamento variou de 90% (amostras não-adulteradas) a 98% (amostras adulteradas com ureia); apenas 3% das amostras não foram classificadas pelo método. Em outro trabalho, Santos *et al.*[56] empregaram tanto o SIMCA como o kNN para a detecção de adulterantes em amostras de leite fluido analisadas por TD-NMR, alcançando valores de seletividade e especificidade maiores que 66% para ambos os métodos.

Jaiswal *et al.*[57] relataram que o SIMCA teve uma taxa de sucesso de ao menos 93,33% na classificação de amostras de leite fluido não-adulteradas e amostras adulteradas com detergente aniônico. Em outro trabalho, Wu *et al.*[58] compararam o desempenho do PLS-DA ao do SIMCA na classificação de amostras de leite com adição de melamina analisadas via espectroscopia de infravermelho próximo (NIR). O método PLS-DA classificou corretamente todas as amostras no conjunto de teste, alcançando uma seletividade e especificidade ambas com um valor de 100%. O método SIMCA classificou corretamente 90% das amostras com uma seletividade de 81,8% e especificidade de 100% para o conjunto de teste.

Nenhum dos trabalhos acima aplicou os métodos SIMCA, PLS-DA, kNN e SVM em dados oriundos de leite cru, conforme descrito. Também não foram encontradas menções a outros trabalhos que tenham utilizado dados físico-químicos de leite cru em um artigo de revisão recente sobre métodos multivariados qualitativos em análise de alimentos por Callao e Ruisánchez (2018)[59].

Em relação às redes de Kohonen, um artigo de Marengo *et al.*[60] descreveu o emprego desse método como uma ferramenta exploratória para o agrupamento de 68 amostras de vinho de acordo com 4 diferentes variedades e diferentes períodos de envelhecimento. Compostos orgânicos voláteis de cada uma das amostras foram extraídos por micro-extração em fase sólida (SPME) e analisados por cromatografia gasosa acoplada a detector de massas (GC-MS). Foram testadas diversas arquiteturas de mapas de Kohonen, e aquela que apresentou melhor desempenho continha 7X7 neurônios, treinada com 500 iterações com taxa de aprendizado linearmente decrescente de 0,5 a 0,01 e

número de pesos treinados a cada iteração também decrescente, de 7 até 1. Os mapas de Kohonen mostraram ser uma ferramenta poderosa para agrupar as amostras de acordo com suas respectivas variedades e tempo de envelhecimento, e os resultados observados apontaram aspectos que a análise por PCA não foi capaz de apontar. Outro trabalho de Diaz *et al.*[61] também empregou esse mesmo método para diferenciar vinhos de 4 ilhas diferentes no arquipélago das Canárias. A concentração de 11 metais foi determinada em triplicata para cada amostra de vinho via espectroscopia de absorção de chama. Várias arquiteturas e parâmetros foram testados para os mapas de Kohonen e aquele que resultou em separação de classes perfeita (nenhuma amostra agrupada erroneamente de acordo com a ilha de origem) foi uma rede de 6X6 neurônios, treinada por 2000 ciclos com taxa de aprendizado decrescente de 0,5 até 0,01. Como no trabalho citado anteriormente, foi observado que os mapas de Kohonen podem melhorar o agrupamento das amostras em comparação à PCA.

Outros dois trabalhos sequenciais de Urruty *et al.*[62] e De Boishebert *et al.*[63] utilizaram dados de compostos orgânicos voláteis de morangos obtidos através de SPME e GC-MS. No primeiro artigo, os resultados foram processados por 5 diferentes redes de Kohonen com tamanhos de 4 a 24 unidades. Elas foram treinadas em duas fases: fase de ordenamento (2000 iterações) e fase de ajuste (500 iterações). O mapa de 24 unidades foi capaz de separar perfeitamente as 70 amostras analisadas para 23 compostos em 17 classes. No segundo artigo, o objetivo foi o de separar as amostras de acordo com a variedade e ano de colheita de cada uma. Combinações de 25 variedades cultivadas em 3 diferentes anos (164 amostras no total) foram analisadas para compostos voláteis da mesma maneira que no trabalho anterior. Novamente foram testadas várias arquiteturas de mapas de Kohonen, e a que apresentou melhor desempenho foi uma de 42 unidades, que conseguiu separar completamente as amostras de acordo com variedade e ano de colheita, sem erros no agrupamento delas.

Nadal *et al.*[64] empregaram mapas de Kohonen para encontrar a relação entre a concentração de dibenzo-*p*-dioxinas (PCDDs) e dibenzofuranos (PCDFs) no leite humano e os hábitos alimentares de diversos países. Um dos resultados encontrados foi o de que o leite humano em países com alto consumo de peixes

tende a ter maiores concentrações de PCDDs e PCDFs do que países com consumo relativamente mais alto de carnes e leite, por exemplo. Assim, foi demonstrado que os mapas de Kohonen podem ser ferramentas úteis para encontrar correlações entre diferentes dados.

Novamente, nenhum dos trabalhos citados aplicou o método dos mapas de Kohonen a dados físico-químicos de leite cru. Não foram também encontrados relatos do uso desse método nesse tipo de dados em um artigo de revisão de literatura de Callao e Ruizsánches[59] e em outros dois de Kamal e Karoui[61] e de Karoui e De Beardemaker[66] sobre a aplicação de ferramentas quimiométricas na análise de dados de leite e derivados. Outra revisão da literatura de Lohumi *et al.*[67] sobre métodos espectroscópicos acoplados a ferramentas quimiométricas na análise de alimentos tampouco menciona redes de Kohonen ou sua aplicação em dados físico-químicos de leite cru. No campo mais amplo da Química Analítica, poucos foram os trabalhos encontrados que fazem uso das redes de Kohonen[68-73], embora haja um número crescente de trabalhos sendo publicados nos últimos anos[38,74,75]. Assim, percebe-se que o método foi pouco explorado desde a sua concepção há quase 40 anos.

4 MATERIAIS E MÉTODOS

4.1 AMOSTRAS E PARÂMETROS

Foram compilados dados de 328 amostras de leite cru dos estados brasileiros do Rio Grande do Sul (RS) e de Santa Catarina (SC) analisadas entre 2013 e 2016 e amostradas por conveniência, sendo 260 pertencentes ao RS e 68 a SC. As amostras foram coletadas de 68 produtores diferentes e 142 delas apresentavam ao menos um parâmetro fora das especificações da legislação brasileira para o leite cru. As análises foram realizadas por diversos analistas, incluindo o autor (de 2014 a 2016), em um laboratório governamental de controle de qualidade de alimentos. A Tabela 2 relaciona o valor mínimo ou a faixa legal para cada parâmetro analisado para que amostra seja considerada conforme à legislação brasileira:

Tabela 2: Valores mínimos ou faixas de valor para os parâmetros analisados

Parâmetro	**Faixa legal/valor mínimo[76]**
Acidez	0,14 – 0,18% em ácido lático
Glicídios redutores em lactose	Mín. 4,3%
Densidade (15℃, relativa)	1,0280 – 1,0340
Extrato seco total	Min. 11,5%
Depressão do ponto de congelamento	-0,530 – -0,550 ℉
Gordura	Mín. 3%
Proteína	Mín. 2,9%

4.2 MÉTODOS FÍSICO-QUÍMICOS

A Tabela 3 relaciona os parâmetros analisados para cada amostra e a respectiva referência metodológica adotada para os ensaios.

Tabela 3: Ensaios físico-químicos e metodologia empregada

Ensaio	Método[77]
Acidez	IN MAPA n°68, 2006
Glicídios redutores em lactose	IN MAPA n°68, 2006
Densidade	IN MAPA n°68, 2006
Extrato seco total	IN MAPA n°68, 2006
Depressão do ponto de congelamento	IN MAPA n°68, 2006
Gordura	IN MAPA n°68, 2006
Proteína	IN MAPA n°68, 2006
Resíduo mineral fixo	IN MAPA n°68, 2006

4.2.1 Acidez

O método titulométrico foi utilizado para a determinação da acidez. Dez mililitros de amostra são titulados com NaOH 0,1 N com fenolftaleína como indicador. O resultado é expresso como percentagem de massa de por volume de ácido lático na amostra.

4.2.2 Lactose

O método Lane-Eynon[78] (Método A) foi utilizado para a determinação da percentagem de lactose em massa. Vinte e cinco gramas de amostra são pesadas e transferidas quantitativamente a um balão volumétrico de 250 mL.

Cinco mililitros de sulfato de zinco a 30% e 5 mL de ferrocianeto de potássio triidratado a 15% são adicionados para a precipitação de gorduras, proteínas e demais componentes emulsionados. A suspensão resultante é diluída com água deionizada e filtrada. O filtrado é titulado contra uma solução padronizada de sulfato de cobre (II) 0,05 N sob aquecimento, utilizando azul de metileno a 1% como indicador redox.

4.2.3 Densidade

Para as amostras anteriores a 2014, foi utilizado o método do termolactodensímetro, que consiste na transferência de 250 mL da amostra resfriada a 8℃ para uma proveta. Aguardar-se-á tem peratura atingir 15℃, conforme leitura no termolactodensímetro, e a densidade é lida no próprio instrumento

Para as demais amostras, a densidade relativa a 15°C foi determinada em densímetro digital. Três mililitros de amostra são injetados no equipamento e o valor para o parâmetro é lido diretamente no equipamento. O método do densímetro digital foi devidamente validado contra o método do termolactodensímetro antes de ser colocado em uso, produzindo valores estatisticamente idênticos de leituras de densidade. Os relatórios de validação não se encontram anexos ao presente trabalho por tratarem-se de documentos confidenciais.

4.2.4 Extrato seco total

A percentagem mássica de sólidos é determinada gravimetricamente. Cinco gramas de amostras são pesados em uma cápsula de alumínio previamente mantida em estufa a 100℃ e resfriada e m dessecador. A amostra é então pré-secada em chapa de aquecimento elétrica a 100℃ por 30 min. e depois levada a estufa a 100 ± 2 ℃ por 2h, resfria da em dessecador e então pesada. Após a primeira pesagem, repete-se ciclos de 1 h de secagem em estufa

seguida de resfriamento e pesagem até que a diferença entre pesagens seja igual ou menor a 0,5% da massa de prova (aprox. 0,0025g)

4.2.5 Depressão do ponto de congelamento

O método crioscópico foi o empregado para determinar esse parâmetro. Um crioscópio digital é calibrado em relação a soluções de NaCl a 0,6859 % (m/v) (- 0,422 ºH ou -0,408 ºC) e a 1,0155 % (m/v) (-0,621 ºH ou -0,600 ºC). Um volume de amostra de 2,5 mL é transferido para um tubo de crioscópio e o ponto de congelamento é lido diretamente no equipamento.

4.2.6 Gordura

O método butirométrico para leite fluido (Método C) foi utilizado para a quantificação da gordura. Em um butirômetro de Gerber, 11 mL de amostras são transferidos para o butirômetro contendo 10 mL de H_2SO_4 (d = 1,820 – 1,825 a 20℃) e 1 mL de álcool isoamílico. O butirômetro é então fechado com rolha de borracha, agitado e centrifugado a 1000 – 1200 rpm por 5 minutos em uma centrífuga termostatizada mantida a 65℃. O percentual mássico de gordura é lido diretamente pelo analista na escala do butirômetro.

4.2.7 *Proteína*

O clássico método de Kjeldahl (na sua variante micro-Kjeldahl) foi empregado. Um grama e meio de amostra é pesado para dentro de um tubo micro-Kjeldahl e 2,5 g de mistura catalítica (10% de sulfato cúprico em sulfato de sódio) são adicionados. A amostra é digerida em 7 mL de ácido sulfúrico concentrado em um ciclo de temperatura programado que inicia a 100℃ e termina a 400℃ por aproximadamente 3h, até que o conteúdo do tubo esteja límpido e com coloração azul-esverdeada. A amônia é liberada em destilador Kjeldahl pela adição de NaOH 50% à mistura e 50 mL são destilados para um

Erlenmeyer de 125 mL contendo 25mL de ácido bórico a 4% com indicador misto verde de bromocresol/vermelho de metila. O destilado é então titulado com ácido sulfúrico 0,1 N e é calculada a percentagem de proteína na amostra.

4.2.8 Cinzas

Embora não seja um parâmetro legal para o controle da qualidade do leite no Brasil, o teor de cinzas ou resíduo mineral fixo é determinado para fechamento das outras analises: a soma dos teores de proteína, lactose e cinzas deve ser igual ao extrato seco desengordurado, isto é, o valor do extrato seco subtraído do valor de gordura. O teor de cinzas é determinado gravimetricamente. Cinco gramas da amostra são pesados em um cadinho de porcelana, que é levado a uma chapa de aquecimento elétrico para queima lenta. Após a pré-queima, o cadinho é levado a forno mufla a 550℃ por 3h. A amostra é resfriada em dessecador e pesada apenas uma vez.

4.2.9 Importância dos métodos físico-químicos de bancada no controle de qualidade do leite

Medidas de parâmetros físico-químicos são utilizadas para avaliar a qualidade do leite nos países (e regiões) mais importantes e populosos do mundo, como Estados Unidos[79], China[80], União Europeia[81], Índia[82], Rússia[83], Japão[84], Austrália[85] e Brasil[76]. A maioria desses países não adota explicitamente métodos clássicos ou de bancada para a determinação dos parâmetros físico-químicos do leite, porém todos os métodos de referência ISO são métodos de bancada ou métodos instrumentais clássicos: método gravimétrico para sólidos totais ((ISO 6731:2010 (IDF 21:2010))[86]; método gravimétrico para gordura em leite (ISO 1211:2010 (IDF 1:2010))[87]; método Kjeldahl para nitrogênio (proteico e não-proteico) (ISO 8968-4:2016)[88]; método cromatográfico para lactose (ISO 22662:2007)[89]; método crioscópico (ISO 5764:2009 (IDF 108:2009))[90], e o método titulométrico para determinação da acidez (ISO 6091:2010)[91]. O método butirométrico ou de Gerber para gordura em leite fluido (ISO 488:2008 (IDF 105:2008))[92] e o método alternativo para determinação de acidez em leite (ISO

6092:1980)[93] também são métodos fisico-químicos reconhecidos como métodos de rotina pela ISO. O método Lane-Eynon[78] para a quantificação de lactose é o método oficial no Brasil[76], Índia[94] e China[95].Os países que explicitamente adotam métodos clássicos incluem Brasil[76], China[95-101] e Índia[94]. Para a validação de métodos mais rápidos para utilização em análises de rotina, é necessário compará-los direta ou indiretamente (através do uso de materiais de referência certificados, que por sua vez devem ser analisados por métodos de referência clássicos) a métodos de referência[102-105]. Assim, justifica-se a escolha do uso de dados oriundos de métodos físico-químicos clássicos ao invés de outros métodos, dada a confiabilidade com que se percebem os métodos de bancada e sua importância e amplo uso para o controle da qualidade do leite em nível mundial.

4.3 ANÁLISE MULTIVARIADA

A análise de componentes principais (PCA) foi feita utilizando o programa ChemoStat®[106]. A análise multivariada por PLS-DA, kNN, SIMCA, SVM, e CART foi realizada em MATLAB (versão R2017a) utilizando-se os módulos criados para ambiente MATLAB por Ballabio e Consonni[107]. A análise por mapas de Kohonen também foi realizada em ambiente MATLAB em módulos criados por Ballabio *et al*.[108]. As amostras para o conjunto de teste (1/3 do total das amostras, arredondado para baixo) e para o conjunto de treinamento (2/3 do total) foram escolhidas através do algoritmo Kennard-Stone[109]. A escolha dos parâmetros de treino para cada método foi feita utilizando-se uma rotina de otimização já embutida nos módulos, com exceção do CART, para o qual o processo de otimização não é aplicável. Para os métodos em que a otimização é aplicável, a escolha do modelo mais adequado foi baseada no menor erro médio de designação para as classes. O erro de designação *E*, como definido em ambos os pacotes de análise multivariada[107,108], é dado por (Equação 9):

$$E = 1 - A \qquad (9)$$

Onde:

E = taxa de erro

A = taxa de acerto (*non-error rate*)

A taxa de acerto *A*, por sua vez, é definida como a média aritmética da especificidade e sensibilidade do modelo. A especificidade *EP* descreve a habilidade do modelo em rejeitar amostras não-pertencentes a certa classe. Consiste na razão entre o número de amostras corretamente identificadas como não-pertencentes à classe em questão (negativos verdadeiros) e a soma dos negativos verdadeiros *NV* com falsos positivos *FP* (número de amostras incorretamente classificadas como pertencentes à classe. É definida de acordo com a Equação 10:

$$EP = NV / (FP + NV) \qquad (10)$$

Onde:

EP = especificidade

NV = número de negativos verdadeiros

FP = número de falsos positivos

Já a sensibilidade *S*, que descreve a habilidade do método em corretamente identificar amostras de fato pertencentes à classe em questão, consiste na razão entre o número de amostras corretamente identificadas como pertencentes à classe (positivos verdadeiros, *PV*) e a soma entre os positivos verdadeiros e falsos negativos *FN*, conforme a Equação 11:

$$S = PV / (FN + PV) \qquad (11)$$

Onde:

S = sensibilidade

PV = número de positivos verdadeiros

FN = número de negativos verdadeiros

Os dados foram autoescalados tanto para a PCA quanto para os métodos supervisionados. O autoescalamento consiste em subtrair de um determinado valor de uma variável em uma matriz de dados o valor médio do vetor que representa a variável, e dividir esse valor pelo desvio-padrão da variável, de acordo com a Equação 12:

$$v_a = (v_i - v_m)/s \qquad (12)$$

Onde:

v_a = valor autoescalado

v_i = *i*-ésimo valor para determinada variável em uma matriz de dados

v_m = valor médio para a variável

s = desvio-padrão para a variável

Para cada método supervisionado, foram adotados os critérios de designação de classe padrão apresentados pelos programas. Para o PLS-DA, o critério de designação de classe escolhido foi o Bayesiano, com uma amostra sendo designada como pertencente a uma certa classe se a probabilidade de a mesma ser daquela classe for cumulativamente maior que 50% e maior que a probabilidade de pertencimento à outra classe. O algoritmo do PLS-DA calcula uma função de densidade de probabilidade para cada classe supondo distribuição normal das amostras no espaço em que são projetadas. Para o kNN, o critério utilizado foi a menor distância Euclidiana ao k vizinhos escolhidos no processo de otimização. No caso do SIMCA, o critério adotado foi modelagem de classe, tendo como valor-limite 95% de probabilidade de pertencimento a uma classe, dado pela função densidade de probabilidade construída para cada classe pelo método. Para a escolha da margem de decisão por parte do SVM, foi determinado a utilização de uma função de núcleo (Kernel) linear, sem parâmetro de núcleo, com aplicação automática em componentes principais. No caso da CART, a regra de decisão é determinada pelo próprio algoritmo durante o processo de treinamento do modelo, sendo explicitada no diagrama criado após o treino.

Para os métodos PLS-DA, SVM, kNN, e SIMCA, foi utilizada na otimização a validação de blocos contíguos com 5 blocos (parâmetro que levou ao melhor desempenho dos modelos após treinamento). Já no treinamento dos modelos, foi utilizada a validação cruzada pelo método Monte Carlo (conhecida pela sigla MCCV, *Monte Carlo Cross-Validation*) com 1000 iterações e 20% das amostras selecionadas para o conjunto de treinamento de validação. Para o processo de otimização, não se encontra implementado no programa utilizado a validação Monte Carlo, por ser um método mais lento do que o de blocos contíguos. Porém, como já mencionado, o método Monte Carlo se encontra implementado para o treinamento dos modelos, sendo utilizado para esse fim no presente trabalho.

A validação cruzada Monte Carlo guarda alguma semelhança com o conhecido método *leave-one-out* (LOO). Porém, ao invés de retirar apenas uma amostra para fins de teste e deixar as restantes para treino, o método Monte Carlo retira um número maior (no caso do presente trabalho, 20%), de amostras aleatoriamente escolhidas, com reposição, utilizadas no entanto para treino. O restante das amostras (80%, nesse trabalho) são utilizadas para teste (validação). A escolha do MCCV mostra-se apropriada para o espaço amostral aqui investigado devido ao grande número de amostras nele presente: a probabilidade de uma amostra no conjunto de teste (n = 109) ser escolhida duas vezes em sequência para compor o conjunto de validação é de cerca de 0,008% (1/109 x 1/109), e menor ainda no conjunto de treino (n = 219). A probabilidade de uma amostra ser escolhida mais do que duas vezes em sequência é obviamente menor, de maneira que o problema da escolha de amostras idênticas ou conjuntos de validação idênticos serem escolhidos é virtualmente inexistente. Adicionalmente, o número de conjuntos de treino de validação cruzada possíveis no conjunto de teste é de cerca de $\binom{109}{21} = 1{,}5 \times 10^{22}$, sendo esse número igual ao de possíveis conjuntos de teste de validação cruzada. Logo, o problema de conjuntos de validação cruzada serem escolhidas não-sequencialmente nas 1000 iterações do algoritmo é também praticamente inexistente. A escolha do MCCV é também corroborada pelo fato de ele ser um método dito assintoticamente consistente: ele é capaz de escolher o modelo ideal (ou, como no presente caso, confirmar ou não como apropriada a escolha do modelo) conforme o número de amostras n tende ao infinito, i.e., $n \rightarrow \infty$, como

demonstrado por Haddad *et al.*[110], sendo um método apropriado para grandes espaços amostrais. Haddad *et al.*[110] também demonstraram que a validação cruzada pelo método de Monte Carlo tem melhor desempenho do que o método *leave-one-out* mesmo em conjuntos de dados menores (*n* = 96, 12 variáveis), assim como Xu e Liang[111], (*n* = 80, 4 variáveis, embora tenham encontrado que o número de amostras escolhidas para o conjunto de treino na validação cruzada deva ser na faixa de 40-60% do número total de amostras).

No caso específico dos mapas de Kohonen, foi utilizado o algoritmo genético (AG) para escolha do modelo mais adequado (otimização). Para todos os métodos supervisionados com exceção das redes de Kohonen, a escolha do modelo ótimo foi baseada no critério de menor erro de classificação no conjunto de treinamento, conforme já descrito. Para os mapas de Kohonen, porém, o critério utilizado foi a maior taxa de escolha do modelo pelo algoritmo genético, independentemente de o erro de classificação ser o menor ou não entre todos pertencentes à população inicial. Dentre os modelos com mesma taxa de escolha pelo AG, se aplicou o critério de maior taxa de acerto. Justifica-se o uso da taxa de escolha pelo fato de o modelo com maior taxa de escolha pelo algoritmo ser o menos propenso a sobre ajuste. A função de adequação (*fitness function*) F_i, descrita na Equação 13 conforme implementada no pacote, explica a seleção de um modelo com base no critério de sobre ajuste:

$$F_i = A_{teste} \cdot (1 - (A_{trein} - A_{teste})) \qquad (13)$$

Onde:

F_i = adequação do cromossomo i

A_{teste} = taxa de acerto na etapa de teste (validação) do AG

A_{trein} = taxa de acerto na etapa de treino do AG

A etapa de teste do algoritmo genético utiliza 20% das amostras do conjunto de treino e a etapa de treino usa 80% delas, todas aleatoriamente escolhidas, com reposição. Porém, não deve ser confundida a otimização por AG com a otimização pelo método Monte Carlo, onde, no caso do presente

trabalho, 20% das amostras são utilizadas para treino e 80% para teste ou validação.

Para a escolha do melhor modelo por meio do algoritmo genético, a única regra de decisão para a qual o algoritmo mostrou convergência foi aquela intitulada no pacote como “Método 1”. A regra de decisão para esse método é mostrada na Equação 14:

$$c_i = k \quad \text{se } y_{i1} > y_{i2} \qquad (14)$$

Onde:

c_i = *i*-ésimo neurônio na rede de Kohonen

k = número identificador de classe (1 para conforme e 2 para não-conforme)

y_{i1} = peso da classe 1 no neurônio c_i

y_{i2} = peso da classe 2 no neurônio c_i

Na otimização foi também utilizada a inicialização por autovalores, para garantir convergência ao mesmo mínimo local para cada solução testada pelo Algoritmo Genético, e evitar que a solução escolhida como mais adequada pelo mesmo fosse o produto de um processo aleatório, evitando também a escolha do modelo errado como mais adequado. O método de treinamento tanto na otimização quanto no treino do mapa mais adequado foi o por batelada (*batch*).

Após a otimização pelo AG, foram treinados 25 mapas com inicialização aleatória utilizando o Método 1 e mais dois dos outros três métodos disponíveis no pacote, Método 2 e Método 3 (sendo treinados 25 mapas para cada um desses métodos. O Método 2 atribui uma amostra a uma certa classe *k* se a diferença entre o maior peso de classe no neurônio e o segundo maior peso de classe for maior que um limite predefinido, nesse caso 0,3 (Equação 15):

$$c_i = k \quad \text{se } y_{ka} - y_{kb} > 0{,}3 \qquad (15)$$

onde:

c_i = *i*-ésimo neurônio na rede de Kohonen

k = número de classe

y_{ka} = maior peso de classe no neurônio c_i (classe 1 ou classe 2)

y_{kb} = segundo maior peso de classe no neurônio c_i (classe 2 se y_{ka} for um peso para a classe 1 e vice-versa)

O último método de designação de classe testado foi o Método 3, que atribui um valor de classe a uma amostra apenas se o peso da classe no neurônio for maior que um limite estabelecido, nesse caso 0,5. A Equação 16 expressa matematicamente a regra de decisão desse método:

$$c_i = k \quad \text{se} \quad y_k > 0{,}5 \qquad (16)$$

Onde:

c_i = *i*-ésimo neurônio

k = valor de classe (1 ou 2)

y_k = maior peso de classe no neurônio c_i

Quanto ao último método de decisão disponível no pacote de programas, chamado Método 4, não se testou esse devido ao fato de ele ser um método de alisamento, apropriado para dados espectrais, o que não é caso no presente trabalho.

Em relação à inicialização aleatória no treinamento das redes de Kohonen, adotada após a otimização, sua escolha se justifica pelo fato de haver a possibilidade de serem encontradas soluções melhores do que na inicialização por autovalores, uma vez que o algoritmo de treino inicia em pontos diferentes de uma hipersuperfície de erro a cada inicialização, podendo convergir para diferentes mínimos (locais ou não), com erro possivelmente menor do que aquele do mínimo atingido pela inicialização por autovalores.

O método de validação cruzada utilizado para a otimização e o treinamento dos mapas de Kohonen (utilizando as 219 do conjunto de treino) foi o de blocos contíguos, com 10 blocos (parâmetro que apresentou melhor desempenho). O método de Monte Carlo não se encontra implementado no pacote, por ser de convergência excessivamente lenta para o caso da validação cruzada de redes de Kohonen. A taxa de aprendizado foi ajustada para decrescer de 0,5 até 0,01 com o crescimento do número de iterações de treino, para garantir convergência. A condição de contorno escolhida para o mapa foi toroidal (lados opostos do mapa se conectam), porque ela melhora a distribuição de erro no mapa[112]. Finalmente, o parâmetro de classe SKN, que determina o efeito relativo da informação da classe de uma amostra sobre o treino da rede de Kohonen[73], foi ajustado para o valor de 1 (sem efeito no treinamento), e 5 repetições de cada modelo foram executadas no treinamento (5 iterações de validação cruzada). Arbitrou-se a escolha de uma arquitetura quadricular, e não hexagonal, para melhor visualização dos resultados.

5 RESULTADOS E DISCUSSÃO

5.1 RESULTADOS DA ANÁLISE EXPLORATÓRIA

Os resultados da análise de componentes principais (PCA) podem ser observados nas Figuras 6 e 7, que consistem no gráfico de escores das componentes principais com a contribuição das variáveis para essas componentes. As amostras não-conformes à legislação brasileira que procedem do estado do Rio Grande do Sul estão coloridas em vermelho e aquelas procedentes do estado de Santa Catarina estão coloridas em laranja. As amostras conformes de ambos os estados estão coloridas em azul. As variáveis estão abreviadas como *Ac%* (acidez percentual em ácido lático), d (densidade relativa), *DPCH* (depressão do ponto de congelamento em graus Hortvath), *EST%* (extrato seco total percentual, em massa), *gord%* (gordura percentual, em massa), *L%* (lactose percentual, em massa), *prot%* (proteína percentual, em massa) e *RMF%* (cinzas, ou resíduo mineral fixo percentual, em massa). As linhas verdes representam a contribuição de cada variável para as componentes principais. A PC1 e a PC3 (componentes principais 1 e 3) explicam 47,33% da variabilidade total dos dados. Embora essa não tenha sido a combinação de PC's que explicou a maior quantidade de variabilidade dos dados, foi a que se mostrou mais informativa de um ponto de vista exploratório do que o gráfico de PC1 *versus* PC2, uma vez que praticamente separou as amostras conformes das não-conformes. Além do mais, o gráfico de PC1 *versus* PC3 separou melhor as amostras não conformes do estado de Santa Catarina (laranja) das amostras conformes (azul), ao contrário do gráfico PC1 *versus* PC2.

A Figura 6 mostra que a PC3 é direta e fortemente relacionada a acidez (*Ac%*) e inversamente relacionada a todos as outras varáveis. Mais da metade das variáveis inversamente relacionadas à PC3 tiveram contribuições notáveis, exceto por *L%*, *d*, e *RMF%*, que tiveram menores contribuições. A PC1, embora explique maior parte da variabilidade dos dados, não foi capaz de separar as amostras conformes das não-conformes tão bem quanto a PC3.

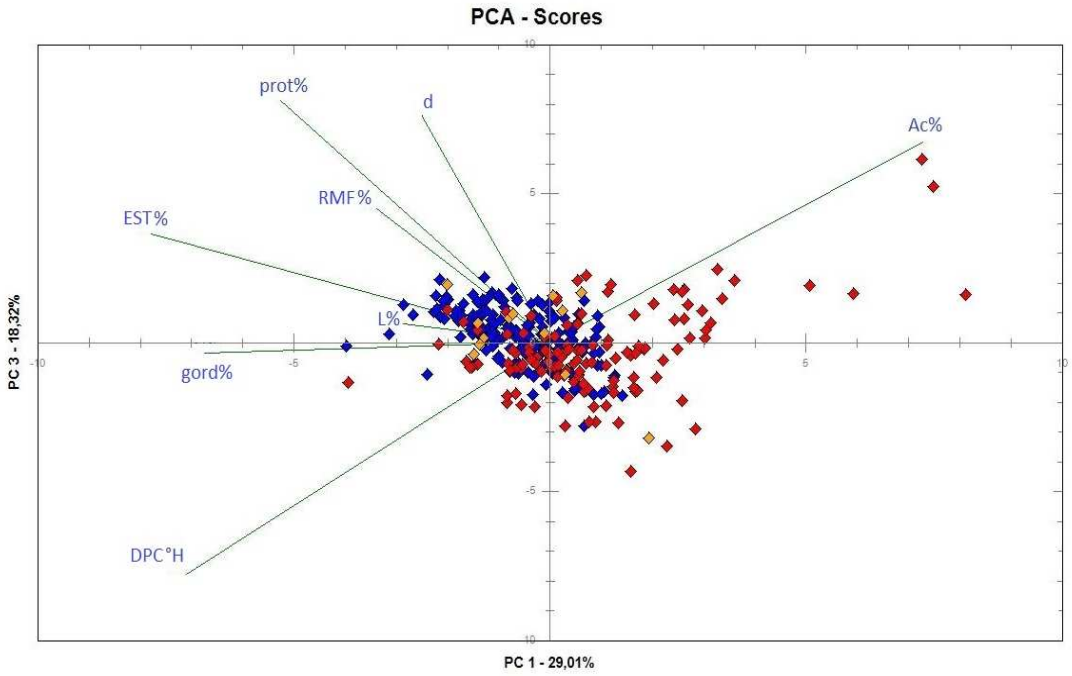

Figura 6: Gráfico biplot (escores e pesos) de PC1 *versus* PC3 para as amostras de leite cru, com base nos dados físico-químicos das mesmas. Amostras em vermelho são amostras não-conformes do estado do Rio Grande do Sul e amostras em laranja são amostras não-conformes do estado de Santa Catarina. Amostras em azul são amostras conformes à legislação brasileira.

Observando o gráfico de escores da Figura 6, nota-se que a maior parte das amostras se localiza na direção de acidez crescente, o que leva a conclusão que a não-conformidade mais frequente observada no conjunto de dados analisado é a acidez elevada ou azedamento do leite. Como a PC1 teve menor contribuição da *Ac%* do que a PC3, entende-se o porquê de a PC3 ter separado melhor as amostras não-conformes das conformes do que a PC1. A Figura 7 corrobora e esclarece a tendência observada, uma vez que as amostras de acidez elevada são colocadas em evidência na direção de *Ac%* crescente na PC2.

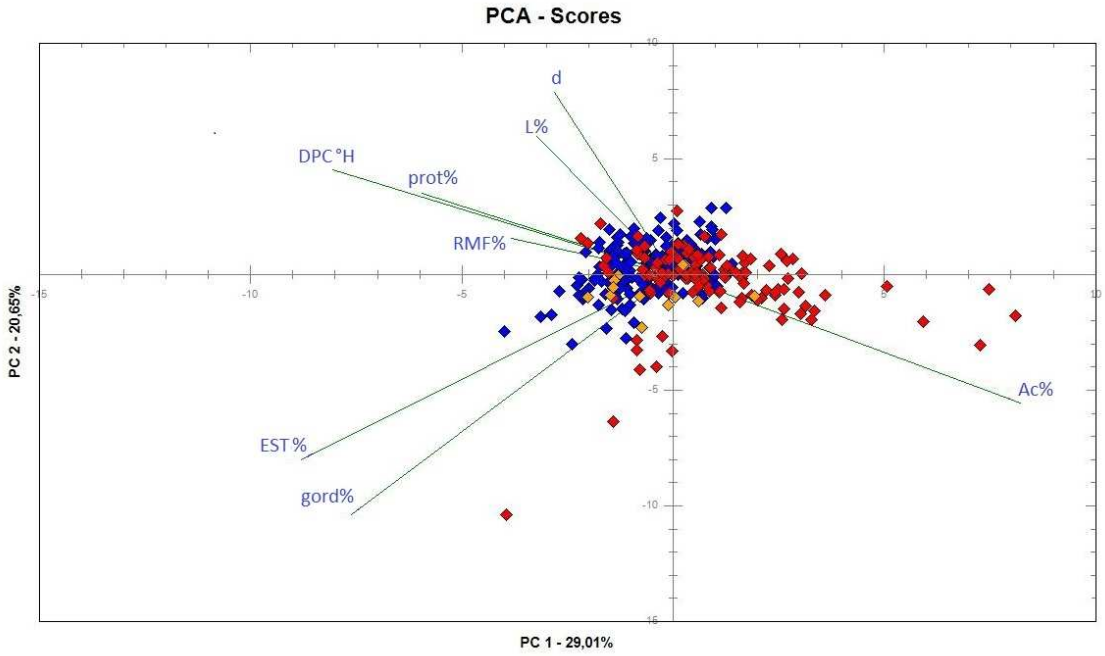

Figura 7: Gráfico biplot (escores e pesos) de PC1 *versus* PC2 para as amostras de leite cru, com base nos dados físico-químicos das mesmas. Amostras em vermelho são amostras não-conformes do estado do Rio Grande do Sul e amostras em laranja são amostras não-conformes do estado de Santa Catarina. Amostras em azul são amostras conformes à legislação brasileira

De todas essas amostras, apenas um pequeno número provinha de Santa Catarina (coloridas em laranja); a maior parte era proveniente do Rio Grande do Sul (vermelho). Isso mostra que a alta acidez pode ser uma não-conformidade mais frequente no estado do Rio Grande do Sul do que no estado de Santa Catarina no intervalo de tempo em que as amostras foram analisadas (2013-2016). Uma conclusão definitiva não pode ser tirada porque o número de amostras não-conformes do Rio Grande do Sul (129), é muito maior do que o número de amostras não-conformes do estado de Santa Catarina (13), logo mais amostras com essa última origem geográfica deveriam ser incluídas no conjunto para certificar-se de que esse padrão de não-conformidade seria reproduzido em um maior número de amostras. Porém, pela informação visual dada nas Figuras 6 e 7, é possível concluir que as amostras do Rio Grande do Sul tiveram acidez como não-conformidade mais comum. De fato, 39 das 129 amostras com aquela

origem geográfica (em torno de 30%), apresentaram acidez demasiadamente elevada para serem consideradas conformes à legislação brasileira.

Ainda em relação às Figuras 6 e 7, nota-se também que existem amostras conformes muito próximas a amostras não conformes no espaço de variáveis reduzidas dos gráficos de escores. Isso indica que essas amostras conformes têm forte caráter de não-conformidade à legislação, o que, por sua vez, indica que a legislação brasileira vigente pode não ser suficientemente precisa em determinar parâmetros que definam se uma amostra de leite cru está apta a ser comercializada, mantendo seu valor nutricional e/ou inocuidade à saúde humana, ou não.

5.2 RESULTADOS DOS MÉTODOS SUPERVISIONADOS

5.2.1 Resultados das otimizações

A otimização dos modelos de cada método supervisionado, conforme descrita no capítulo anterior, resultou nos seguintes parâmetros para o treinamento dos mesmos: 4 componentes para o PLS–DA, k = 3 para o kNN, 7 componentes para a classe 1 (conforme) e 3 componentes para a classe 2 (não-conforme) para o SIMCA e um custo de 1 para o SVM, resultando em 138 vetores de suporte e 4 componentes principais no conjunto de treinamento.

No caso dos mapas de Kohonen, a melhor arquitetura dada pelo algoritmo genético foi a de 6 *x* 6 neurônios (36 no total) com 350 iterações de treinamento, tendo uma taxa de escolha pelo AG de 0,5 e taxa de acerto de 72%. A Figura 8 mostra o “diagrama de bolhas”, que auxiliou na escolha do melhor modelo. Bolhas maiores significam maior número de neurônios na rede, e bolhas mais escuras significam maior número de iterações para treinamento.

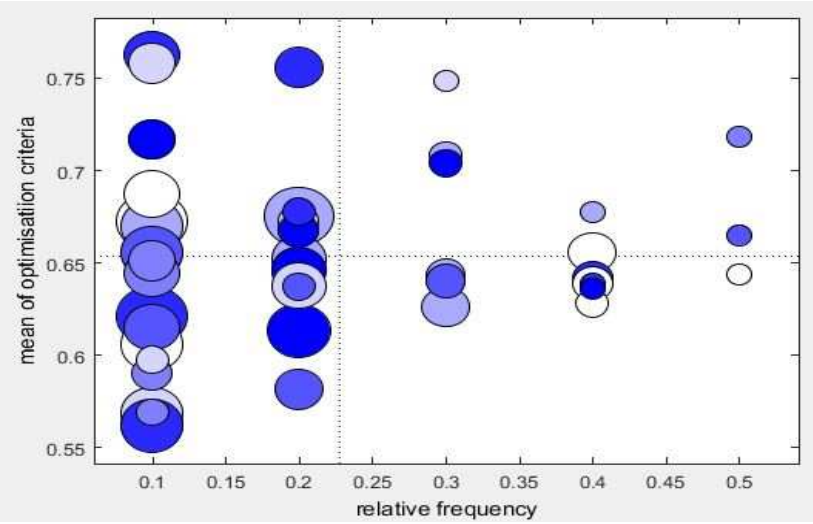

Figura 8: Diagrama de bolhas para escolha da arquitetura e parâmetros ótimos de treino da rede de Kohonen. Bolhas maiores representam maior número de neurônios, e bolhas mais azuladas maior número de iterações de treinamento. Os valores no eixo das abscissas representam a frequência de escolha de determinada arquitetura pelo algoritmo genético, e os valores no eixo das ordenadas representam a média do critério de otimização, a taxa de acerto de determinado modelo na classificação das amostras de treino na etapa de teste do algoritmo.

5.2.2 Resultados de classificação

A Tabela 4 mostra os resultados de classificação para a classe 2 (amostras não-conformes aos regulamentos brasileiros de qualidade do leite), todos para o conjunto de treinamento. No caso da rede de Kohonen, o método de decisão que mostrou melhores resultados foi o Método 2, descrito no capítulo anterior; o resultado mostrado se refere ao melhor modelo entre os 2 treinados para o Método 2. Já a Tabela 5 mostra os resultados de classificação no conjunto de teste para os métodos PLS-DA, SIMCA, kNN e SVM, para fins de comparação com a literatura, uma vez que alguns dos artigos que empregaram esses métodos continham dados referentes também ao desempenho no treinamento. Conforme relatado na Revisão Bibliográfica, não foram encontrados trabalhos que relatassem resultados de classificação de amostras de leite ou derivados usando os métodos de Redes de Kohonen (KSOM) e CART. Foram considerados resultados satisfatórios aqueles maiores que 70% na classificação

do conjunto de teste, para cada figura de mérito. Em outras palavras, todas as figuras de mérito na classificação das amostras do conjunto de teste deveriam ser maiores que 70% para o método ser considerado satisfatório.

Tabela 4. Percentagem de amostras corretamente designadas (taxa de acerto), especificidade, seletividade percentual em relação à classe 2 (não conforme), especificidade percentual em relação à classe 2 (não conforme) e percentual de amostras não classificadas (quando aplicável) para cada método supervisionado no conjunto de treino.

Método	Taxa de acerto %	Especificidade%, classe 2 (não conforme)	Sensibilidade%, classe 2 (não conforme)	Amostras não classificadas%
PLS-DA	89	86	83	0
CART	87	87	86	-
KSOM	84,4	90	79	5,9
SVM	69	90	50	-
kNN	67	86	50	-
SIMCA	90	100	79	58

Tabela 5. Percentagem de amostras corretamente designadas (taxa de acerto), especificidade, seletividade percentual em relação à classe 2 (não conforme), especificidade percentual em relação à classe 2 (não conforme) e percentual de amostras não classificadas (quando aplicável) para cada método supervisionado no conjunto de teste.

Método	Taxa de acerto %	Especificidade%, classe 2 (não conforme)	Sensibilidade%, classes 2 (não conforme)	Amostras não classificadas%
PLS-DA	89	88	83	0
KSOM	81,6	87	76	0,9
CART	78	76	81	-
SVM	75	78	71	-
kNN	72	78	67	-
SIMCA	98	100	97	32

5.2.3 Comparação inter-métodos dos resultados do SIMCA, PLS-DA, kNN e SVM e comparação dos resultados desses métodos com resultados da literatura

Observando-se os resultados contidos as Tabelas 4 e 5, percebe-se que os métodos supervisionados tiveram diferentes desempenhos quanto à taxa de amostras corretamente designadas, especificidade e seletividade em relação às amostras não-conformes e quanto ao número de amostras não-classificadas

Para a classificação das amostras, o método SIMCA aparentou ser aquele de maior sucesso, tendo corretamente classificado 98% delas. Porém, o número de amostras não-classificadas no conjunto de treino foi de 56%, e no de teste, 32%. Isso mostra que, apesar de ter apresentado os melhores resultados de classificação em termos de figuras de mérito, o método pode não ser o mais apropriado para a classificação deste conjunto de dados. O motivo pelo qual o SIMCA deixou de classificar um maior número de amostras no conjunto de treino do que no de teste pode ser atribuído à seleção das amostras pelo algoritmo Kennard-Stone. Este algoritmo seleciona as amostras mais discrepantes (ou

representativas) de um conjunto de dados inicial para o conjunto de treinamento, e escolhe aquelas de menor variabilidade para o conjunto de teste. Assim, um maior número de amostras no conjunto de treinamento foram consideradas anômalas (ou consideradas como *outliers*). Porém, os resultados encontrados foram melhores ou similares do que aqueles descritos na literatura[52-58], onde se almejava a detecção de adulterantes no leite, mesmo sendo os métodos de análise química muito diferentes daqueles aqui empregados. Ainda assim, o SIMCA deixou de classificar até 32% das amostras quando o modelo foi testado sobre as amostras de teste, enquanto o pior desempenho para essa figura de mérito foi de 24,1% na classificação de amostras (teste), relatado por Gondim *et al.* (2017a)[53]. O fato de que o método analítico empregado naquele trabalho foi o da espectroscopia no infravermelho, que disponibiliza uma quantidade de informação muito maior do que o conjunto de dados de oito variáveis aqui empregado, pode explicar o menor número de amostras classificadas de maneira inconclusiva. Além do mais, a adição de quantidades distintas de adulterantes às amostras naquele trabalho pode ter criado perfis químicos amostrais distintos, facilmente diferenciáveis por SIMCA, e por consequência uma boa separação entre classes, reduzindo a classificação de amostras a múltiplas classes ou a nenhuma. Os valores de especificidade foram também melhores ou similares àqueles relatados por Gondim *et al.* [53,54], que variaram de 56,7% (em um modelo de 5 classes) a 100% (tanto em um modelo de 2 como em um de 5 classes). O SIMCA gerou bons resultados no que tange à sensibilidade também, novamente melhores ou similares ao de Gondim *et al.*[54] em um modelo de 2 classes: 65% a 90%, naquele trabalho.

O PLS-DA também teve bons resultados quando comparado aos outros métodos, sendo o método com segunda maior taxa de acerto entre eles quando aplicado ao conjunto de teste. Porém, a taxa de acerto nos conjuntos de treinamento e de teste não foram tão altas como as relatadas por Wu *et al.*[58], as quais foram 100% para ambos os conjuntos. Para o conjunto de treinamento os valores foram próximos aos relatados no trabalho de Wu *et al.* (100% e 82,6%, respectivamente). Diferenças no tipo de amostra (leite em pó vs. leite fluido) e métodos analíticos (espectroscopia no infravermelho vs. métodos físico-químicos clássicos) podem explicar os resultados diferentes.

Embora o PLS-DA tenha apresentado resultados inferiores aos do SIMCA em relação às figuras de mérito avaliadas, o primeiro não deixou de classificar amostras tanto no conjunto de treino quanto no conjunto de teste. Isso mostra que o PLS-DA é, provavelmente, o método mais confiável para a classificação dos dados aqui analisados, pois exibiu desempenho superior ao SIMCA em relação ao número de amostras não-classificadas, e superior aos demais métodos em relação às figuras de mérito avaliadas.

Quando aplicado ao conjunto de teste, o SVM resultou em 55 vetores de suporte e 3 componentes principais. Os resultados de classificação do conjunto de teste pelo SVM foram razoavelmente bons quando comparados aos outros métodos (taxas de acerto de 70% e 75% para os conjuntos de treinamento e de teste, respectivamente, Tabelas 4 e 5); porém inferiores aos relatados por Liu[50] e Bougrini *et al.*[51], os quais foram de 100% para o conjunto de teste no trabalho de Liu e 100% no trabalho de Bougrini *et al.*, também no conjunto de teste. Os valores de especificidade também foram satisfatórios em ambos os conjuntos de treinamento e de teste (90% e 78%, respectivamente), conforme mostrado nas Tabelas 4 e 5. Porém, o conjunto de treino teve uma seletividade de 50%, considerada aqui insatisfatória, enquanto essa mesma figura de mérito apresentou um valor maior para o conjunto de teste (71%). Esses resultados também foram inferiores aos descritos por Liu[50], de 100% de especificidade e 96,67% de seletividade. Novamente, diferenças nos tipos de amostra e nos métodos analíticos podem explicar esses resultados não-similares entre si.

Por fim, o método kNN foi o que apresentou o desempenho mais baixo entre os métodos testados (Tabelas 4 e 5), levando em conta a taxa de acerto, especificidade e sensibilidade tanto para o conjunto de treinamento quanto para o conjunto de teste. Os resultados de sensibilidade e especificidade foram em geral inferiores aos relatados por Santos *et al.*[55], que variaram de 66% a 100%. Mais uma vez, diferenças entre as amostras e métodos de análise podem ter papel determinante para explicar os diferentes resultados observados.

5.2.4 Comparação entre os resultados de teste do KSOM e CART com os resultados de teste do PLS-DA, SIMCA, kNN e SVM

Compara-se os resultados apenas de teste dos métodos KSOM e CART com aqueles dos métodos SIMCA, PLS-DA, kNN e SVM tendo em vista que na aplicação desses métodos em uma situação de rotina para a classificação de amostras de leite, os resultados da classificação sobre o conjunto de teste são os relevantes para avaliar a possiblidade de implementação desses métodos em um laboratório analítico. Um modelo pode apresentar resultados de classificação insatisfatórios para um conjunto de treinamento e ainda assim apresentar bons resultados para a classificação de amostras de um conjunto de teste: é o que se espera de um modelo livre de sobre ajuste, conforme comentado na seção de Materiais e métodos. Assim, dada a ausência de literatura que relata resultados de classificação no treinamento de modelos dos métodos KSOM e CART sobre amostras de leite ou derivados, e dada a irrelevância desses dados para a avaliação da aplicabilidade desses métodos em laboratório, conforme anteriormente citado, serão comparados nessa subseção apenas os resultados de teste dos métodos aqui empregados sobre o conjunto de dados utilizado no presente trabalho.

Conforme os resultados apresentados na Tabela 4, os métodos de mapas de Kohonen (KSOM) e CART apresentaram resultados satisfatórios de acordo o critério aqui estabelecido: todas as figuras de mérito para cada um desses métodos foi maior do que 70%. No caso do KSOM, a taxa de acerto foi de 81,6%, e a especificidade e sensibilidade em relação às amostras não-conformes à legislação brasileira foram, respectivamente, 87 e 76%, com uma taxa de amostras não-classificadas de 0,9%. Essas mesmas figuras de mérito foram, na mesma ordem anterior, 78, 76 e 81%. A Figura 9 mostra o diagrama CART e a regra de decisão criada pelo algoritmo para classificar as amostras do conjunto de teste. As variáveis denominadas “var 1” até “var 7” são, respectivamente, acidez, lactose, densidade, extrato seco total, depressão do ponto de congelamento, gordura e proteína.

Comparando-se os resultados de classificação do conjunto de teste do método KSOM com os outros métodos utilizados (Tabela 4), percebe-se que

esses foram superiores aos dos métodos SVM (75% de taxa de acerto, 78% de especificidade e 71% de sensibilidade), kNN (72%, 78% e 67%, respectivamente) e CART (78%, 76% e 81%, respectivamente). Os resultados do KSOM foram inferiores, porém, aos do PLS-DA (89, 88 e 83%) e aos do SIMCA (98, 100 e 97%). O SIMCA, por outro lado, deixou de classificar quase um terço (32%) das amostras no conjunto de teste, enquanto o KSOM classificou de maneira inconclusiva apenas 0,9% das amostras desse mesmo conjunto.

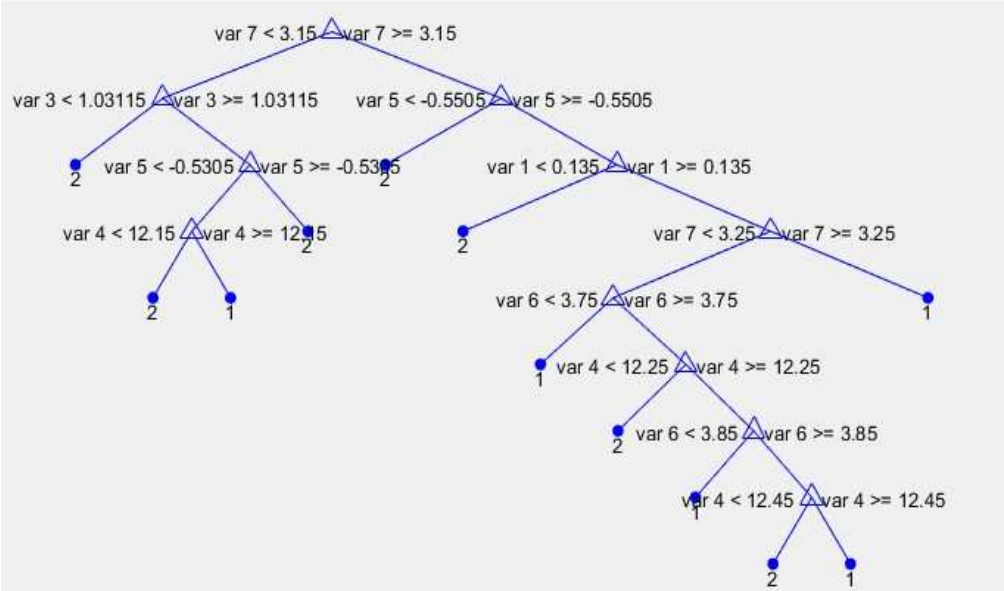

Figura 9: Diagrama CART resultante do treino com as amostras do conjunto de treinamento As variáveis denominadas “var 1” até “var 7” são, respectivamente, acidez, lactose, densidade, extrato seco total, depressão do ponto de congelamento, gordura e proteína. Não são explicitadas as unidades de medida.

O método CART (78% de taxa de acerto, 76% de especificidade e 81% de seletividade) teve desempenho superior apenas aos métodos kNN e e SVM, e próximo ao do KSOM (81,6% de taxa de acerto, 87% de especificidade e 76% de seletividade), tendo apresentado maior seletividade em relação à classe 2 (amostras não-conformes) do que esse último método.

Tendo em vista esses números, é possível afirmar que o método que com mais sucesso classificou as amostras de leite cru como conformes ou não-

conformes à legislação brasileira, baseado nos seus dados físico-químicos, foi o PLS-DA, com menor número de amostras não-classificadas do que o SIMCA e com maiores valores de taxa de acerto, especificidade e seletividade em relação às amostras não-conformes do que todos os demais métodos, com exceção do SIMCA. Ambos os métodos, conforme aplicados no presente trabalho, tem a característica comum de serem métodos probabilísticos, ou seja, métodos que empregam uma regra de decisão probabilística, supondo distribuição normal das amostras em um hiperespaço de variáveis (SIMCA) ou espaço de dimensionalidade reduzida (PLS-DA) para decidir se uma amostra pertence à uma classe ou à outra. Assim, é possível afirmar que métodos probabilísticos foram os mais apropriados para a classificação das amostras do conjunto de dados aqui utilizados, provavelmente pelas mesmas obedecerem à distribuição normal nos espaços em questão nas quais foram projetadas.

Dos métodos não-probabilísticos, o KSOM teve o maior sucesso em classificar as amostras, embora tenha apresentado sensibilidade menor do que o CART (76% contra 81%, respectivamente). O KSOM é um método computacionalmente mais complexo que o CART, e provavelmente tenha tido um melhor desempenho em relação à taxa de sucesso e especificidade pela própria complexidade das amostras utilizadas para o treinamento dos modelos, resultando em um modelo mais ou menos apropriado para a classificação das amostras do conjunto de teste de acordo com o algoritmo empregado. Essa mesma complexidade na distribuição das amostras em um hiperespaço pode explicar o desempenho relativamente inferior dos métodos kNN e SVM, bem como a relativa simplicidade das regras de decisão desses dois métodos quando comparados aos demais.

De volta à Figura 7, nota-se a ausência da variável cinzas. Isso ocorre devido ao mascaramento da variável por outra. De fato, uma regressão linear da variável cinzas em função da variável extrato seco total tem R^2 = 0,1075, considerado como forte correlação por Falk e Miller[113]. Ainda em relação ao CART, percebe-se que os resultados de treinamento foram melhores do que os de teste, o que não foi observado para nenhum dos outros métodos, com exceção das redes de Kohonen (Tabelas 4 e 5). Isso pode ser explicado também pela maneira com que o algoritmo Kennard-Stone escolhe as amostras para

treino e teste, conforme já discutido. O CART, ao contrário dos demais métodos, consegue classificar melhor amostras quando essas apresentam valores pouco similares para as suas variáveis. Esse é justamente o caso das amostras do conjunto de treino quando escolhidas pelo algoritmo Kennard-Stone, que separa as amostras mais discrepantes do conjunto de dados original para comporem o conjunto de treinamento. A mesma explicação pode ser dada para o método de redes de Kohonen, que também exibe desempenho superior quando as amostras têm valores de variáveis discrepantes, sendo capaz de segregar melhor neurônios com forte resposta a uma variável se ela apresentar resultados bem distintos.

Por fim, percebe-se também, ao se observar o diagrama CART resultante, que os valores determinados pelo algoritmo para a construção da regra de decisão não coincidem com os limites ou faixas legais presentes na legislação brasileira. Isso indica uma possível deficiência nos limites estabelecidos pela legislação, que não permitem classificar de maneira não-ambígua uma amostra como conforme ou não, confirmando que existem amostras conformes com forte caráter não-conforme, de acordo com o observado no PCA.

5.2.5 Pertinência dos parâmetros adotados para o leite cru na legislação brasileira

Com base nos biplots, considerando os pesos das variáveis para cada componente principal nas Figuras 6 e 7, nota-se que todas as variáveis contribuem para a separação das amostras em torno das componentes principais, uma vez que todas exibem contribuições (quando projetadas sobre cada eixo coordenado) para ambas as PCs conforme gráficos de escores. Isso mostra que todas elas trazem informação relevante para a classificação do leite cru como conforme ou não à legislação brasileira, e confirma a necessidade da análise de todos os parâmetros elencados nela. Porém, o critério univariado utilizado para determinar a conformidade de uma amostra (basta um parâmetro ser não-conforme para a amostra ser classificada como não conforme, independente dos valores dos outros parâmetros) não se mostrou suficientemente adequado para a classificação das amostras como conformes

ou não. Isso é demonstrado pelo fato de amostras conformes se localizarem muito próximas a amostras não conformes nos gráficos de escores de PCA, tendo forte caráter não conforme, assim como algumas amostras conformes que se localizam próximas às não conformes. O desempenho do método SIMCA pode também ser explicado por esse fato, no que diz respeito ao número de amostras não-classificadas. Amostras conformes com forte caráter não-conforme, e vice versa, podem não ter sido classificadas pelo método. Amostras classificadas erroneamente pelo SIMCA e pelos outros métodos supervisionados provavelmente apresentavam forte caráter (peso de classe, probabilidade de pertencimento, ou distância) da classe oposta.

5.2.6 Aplicação da metodologia em regulamentos de países que não o Brasil

Conforme mostrado na Tabela 6, os parâmetros de conformidade do leite cru em diferentes países e regiões do mundo guardam ao menos alguma similaridade com os padrões brasileiros. Os regulamentos cujos valores se apresentam mais próximos aos do Brasil são aqueles de China e Japão, o que sugere que os métodos supervisionados utilizados poderiam ser empregados com sucesso para classificar amostras daqueles países como conformes ou não aos seus respectivos regulamentos. Porém, a faixa legal para a depressão do ponto de congelamento no regulamento chinês (-0,518 – -0,580 °H) é bastante diferente daquela do brasileiro (-0,530 – -0,550°H). Esse problema pode ser contornado, uma vez que a faixa legal para a China é provavelmente baseada nos valores de ponto de congelamento observados naquele país, logo podem ser treinados modelos com amostras que apresentem aquela procedência geográfica, e assim essas amostras podem ser corretamente classificadas com os métodos utilizados. De fato, a composição do leite depende de uma série de fatores que podem ser específicos de cada país[117]. O parâmetro de depressão do ponto de congelamento para a União Europeia (UE) é simplesmente a média de valores para essa variável observada na região de origem do leite, o que sugere que amostras da EU que tenham valores não-conformes ou discrepantes de ponto de congelamento poderiam ser identificadas por modelos treinados com amostras provenientes de regiões específicas da Europa.

A densidade relativa é o parâmetro que apresenta os valores mais similares dentre as legislações dos diferentes países da Tabela 6, que regulam essa propriedade do leite. Os regulamentos de Brasil e Japão apresentam a mesma faixa legal (1,028 – 1,034), e aqueles da Rússia, China, e UE apresentam não apenas valores mínimos ou específicos que estão (ou podem estar) em uma mesma faixa (1,029 e 1,030, para China e Rússia, respectivamente), mas que são também muito semelhantes entre si. Isso sugere que amostras procedentes desses países podem também ser classificadas de acordo com a estratégia aqui empregada. O mesmo é válido para proteína e extrato seco desengordurado, com valores similares através das legislações. Além disso, a construção de modelos com amostras nativas de cada país pode fazer com que os modelos sejam adaptáveis a realidade de cada país.

Tabela 6: Parâmetros legais para o leite cru em diferentes países[77, 79-85]

Parâmetro	Faixa legal/valor máximo ou mínimo			
	Brasil	**China**	**Japão [c]**	**União Europeia**
Acidez	0,14 – 0,18% ácido lático	0.14 – 0,18% ácido lático	Máx. 0,18% ácido lático	-
Densidade (relativa)	1,0280 – 1,0340 (15℃)	1,0290 (20℃) [a]	1,0280 – 1,0340 (15℃)	1,030 (20℃) [d]
Extrato seco total	Mín. 11,5%	-	-	-
Extrato seco desengordurado	Mín. 8,4%	Mín. 8,1%	Mín. 8,0%	Mín. 8,3%
Lactose	Mín. 4,3%	-	-	-
Gordura	Mín. 3,0%	Mín. 3,1%	Mín. 3,0%	Mín. 3,5%
Dep. ponto de congelamento	-0,530 – -0,550 ℃H	-0,518– -0,580 ℃H [b]	-	Nenhum valor especificado [e]
Proteína	Mín. 2,9%	Mín. 2,8%	-	Min. 2,9% [f]
	Rússia	**Austrália [h]**	**Estados Unidos [i,j]**	**Índia**
Acidez	-	-	-	-
Densidade (relativa)	1,029 (20℃) [g]	-	-	-
Extrato seco total	-	-	-	-
Extrato seco desengordurado	Mín. 8,2%	-	Mín. 8,25%	Mín. 8,3%
Lactose	-	-	-	-
Gordura	Mín. 3,5%	Mín. 3,2%	Mín. 3,25%	Mín. 3,2%
Dep. ponto de congelamento	-	-	-	-
Proteína	-	Mín. 3,0%	-	-

Tabela 6 (continuação): Parâmetros legais para o leite cru em diferentes países[77, 80-86]

[a] Conversão aproximada do valor legal utilizando 998,2 kg/m^{-3} como o valor de densidade da água a 20℃ e 1,01325 bar[114].

[b] Conversão aproximada da faixa legal (-0,500 – -0,560 ℃) usando a relação 1℉ = 1,03562 x ℃ [115].

[c] Parâmetros legais para leite de vacas que não sejam da raça Jersey.

[d] Conversão aproximada do valor legal (1,029) utilizando 998,2 kg/m^{-3} como o valor de densidade da água a 20℃ e 1,01325 bar[114].

[e] O ponto de congelamento deve ser “próximo ao ponto de congelamento médio do leite cru registrado na região de origem” (no original: “close to the average freezing point for raw milk recorded in the area of origin”)[82].

[f] Conteúdo proteico para leite com 3,5% de gordura. Outros teores de gordura devem ter teores proporcionais de proteína.

[g] Conversão aproximada do valor legal utilizando 998,2 kg/m^{-3} como o valor de densidade da água a 20℃ e 1,01325 bar[114].

[h] Padrão para o leite “a ser vendido como leite de vaca” (no original: “to be sold as cow’s milk”)[85]. Infere-se que pode ser aplicado ao leite cru de vaca.

[i] Valores para o leite na forma final de envase para o consumo humano. Alguns estados permitem a venda de leite cru nessa forma[116], então infere-se que os parâmetros sejam válidos também para o leite cru.

[j] Padrões para o leite cru válidos em nível federal.

6 CONCLUSÃO

Com base nos resultados observados, conclui-se que a análise exploratória utilizando o método de Análise de Componentes Principais foi capaz de identificar um padrão físico-químico, o de acidez elevada, para as amostras de leite dos estados do Rio Grande do Sul e possivelmente de Santa Catarina que não se apresentavam conformes ao regulamento brasileiro para a qualidade do leite cru. Dado o número relativamente pequeno de amostras não-conformes de Santa Catarina em relação às amostras não-conformes do Rio Grande do Sul, seria necessário incluir mais amostras que tenham a primeira origem geográfica para confirmar esse padrão para essas amostras. Mesmo assim, a análise exploratória por PCA demonstrou que a não-conformidade mais frequente no conjunto de dados estudado é a acidez elevada. Dessa maneira, a ferramenta se mostrou útil para a identificação de padrões físico-químicos em amostras de leite cru. A análise por PCA também mostrou que os parâmetros legais brasileiros utilizados para avaliar a qualidade do leite são pertinentes de serem analisados e avaliados, uma vez que cada um deles traz, individualmente, informação relevante para a discriminação de amostras conformes daquelas não-conformes. Porém, a PCA indicou que a legislação brasileira vigente pode ser imprecisa em relação ao estabelecimento de critérios de não-conformidade para o leite cru, uma vez que diversas amostras conformes apresentavam forte caráter de não-conformidade quando projetadas nos gráficos de escores. Essa observação foi corroborada pelo diagrama de decisão resultante do método supervisionado CART, cujos valores utilizados para a construção da regra de decisão não coincidiram com os valores preconizados na legislação brasileira para a conformidade do leite.

Em relação aos resultados dos métodos de classificação, é possível concluir que o método mais confiável entre todos os métodos testados para a classificação das amostras de leite foi o PLS-DA. Embora o método SIMCA tenha apresentado melhor desempenho com base na avaliação das figuras de mérito resultantes da classificação das amostras do conjunto de teste e de treino, esse método foi inconclusivo para um número relativamente alto de amostras. O PLS-

DA teve desempenho inferior ao do SIMCA, porém nenhuma das amostras foi classificada de maneira inconclusiva. O método das redes de Kohonen, por sua vez, apresentou desempenho inferior ao PLS-DA, porém ainda satisfatório de acordo com o critério aqui adotado (todas as figuras de mérito maiores que 70%), uma vez que o pior resultado apresentado pelo método no conjunto de teste foi 76% para a sensibilidade em relação às amostras não-conformes. Após as redes de Kohonen, o próximo método que apresentou desempenho inferior foi o CART, porém também satisfatório, apresentando 76% de especificidade para a classe 2 (amostras não-conformes) como menor figura de mérito. Os resultados do SVM foram da mesma maneira satisfatórios, com valores superiores a 71% para todas as figuras de mérito. Por último, o método kNN foi o que apresentou o pior desempenho de todos em relação às figuras de método avaliadas, considerado aqui insatisfatório por apresentar sensibilidade de 67% em relação às amostras não conformes no conjunto de treino. Ainda assim, esses resultados foram similares aos apresentados na literatura para o método kNN.

REFERÊNCIAS

[1]IDF (2017). The economic impact of milk. IDF factsheet. Disponível em https://www.fil-idf.org.

[2]Poonia, A; Jha, A., Sharma, R., Singh, H.B., Rai, A.K., Sharma, N. Detection of adulteration in milk: A review. *Int. J. Dairy Technol.,* **2017**, *70 (1),* 1–19.

[3]Mackay, H. The detection of milk adulteration. *Can. Med. Assoc. J.,* **1929**, *21 (3)*, 309.

[4]Brasil. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa de 18 de setembro de 2002. Disponível em: http://extranet.agricultura.gov.br/sislegis-consulta/consultarLegislacao.do?operacao=visualizar&id=8932. Acesso em 24/01/2017.

[5]Da Silva, P.HF.; Pereira, D. B. C.; De Oliveira, L. L.; Costa Júnior, L. C. G. Físico-Química do leite e derivados: métodos analíticos. Oficina de Impressão Gráfica, Juiz de Fora, 1997, 190 p.

[6]Nunes, G. F. M.; De Paula, A.V.; De Castro, H. F.; Dos Santos, J. C. Modificação bioquímica da gordura do leite. *Quím. Nova,* **2010**, *33*, 431–437.

[7]Soares, K. M. P.; Bezerra, N.M. Características de identidade e qualidade do leite bovino brasileiro. PUBVET, 2010, 6, art. 750. Disponível em: http://www.pubvet.com.br/uploads/b452f6d75728a0204f9f4b01929fcac2.pdf. Acesso em 27/01/2017.

[8]Gantner, V.; Mijic, P.; Baban, M.; Skrtic, Z.; Turalija, A. The overall and fat composition of milk of various species. *Mljekarstvo*, **2015**, *65,* 223–231.

[9]Muehlhoff, E.; Bennett, A.; MacMahon, D. Milk and dairy products in human nutrition. Food and agriculture organization of the United Nations, Roma, 2013, 376 p.

[10]Milk Composition and Synthesis Resource Library. Disponível em: http://ansci.illinois.edu/static/ansc438/Milkcompsynth/milkcomp_vitamins.html. Acesso em 25/01/2017.

[11]Tronco, V. M. Manual para inspeção da qualidade do leite. 2ª ed. Editora UFSM, Santa Maria, 2003, 192 p.

[12]Brasil. Ministério da Agricultura, Pecuária e Abastecimento. RIISPOA. MAPA/SDA/DIPOA, 2007, 252 p.

[13]Pearson K. On lines and planes of closest fit to systems of points in space. *Phil. Mag.,* **1901,** *2,* 559–572.

[14]Hotelling H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.,* **1933**, *24,* 417–441.

[15]Ferreira, M. M. C. Quimiometria: conceitos, métodos e aplicações. Editora Unicamp, Campinas, 2015, 493 p.

[16]Wold, S.; Esbensen, K.; Geladi, S. Principal Component Analysis. *Chemom. Intell. Lab. Syst.,* **1987**, *2*, 237–252.

[17]Bro, R. & Smilde, A.K. Principal Component Analysis. *Anal. Methods,* **2014**, *6*, 2812–2831.

[18]Geladi, P. Notes on the history and nature of partial least squares (PLS) modelling. *J. Chemom,.* **1988**, *2 (4)*, 273–288.

[19]Gottfries J, Blennow K, Wallin A, Gottfries CG. Diagnosis of dementia using partial least squares discriminant analysis. *Dementia,* **1995**, *6*, 83–88.

[20]Barker, M., Rayens, W. Partial least squares for discrimination. *J. Chemom.,* **2003**, *17*, 166–173.

[21]Brereton, R.G., Lloyd, G.R. Partial least squares discriminant analysis: taking the magic away. *J. Chemom.,* **2014**, *28*, 213–225.

[22]Bhavsar, H., Panchal, M. H. A . *Int. J. Adv. Res. Comp. Eng. Tech.,* **2012,** *1 (10)*, 185–189.

[23]Vapnik, V. Estimation of Dependences Based on Empirical Data [em Russo]. Nauka, Moscow, 1979. 105p. (Tradução para o inglês: Springer Verlag, New York, 1982).

[24]Burges, C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.,* **1998**, *2*, 121–167.

[25]Ballabio, D., Todeschini, R. Multivariate Classification for Qualitative Analysis. In: Infrared Spectroscopy for Food Quality Analysis and Control, Part I: Fundamentals and Instruments. Ed.: Sun, W.D. Academic Press, 2009, 424 p.

[26]Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In: Reviews in Computational Chemistry, Volume 23, Eds.: K. B. Lipkowitz and T. R. Cundari. Wiley-VCH, Weinheim, 2007, pp. 291–400.

[27]Wold, S., Sjostrom, M. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity. *Am. Chem. Soc. Symp. Ser.* **1977**, *XX*, *52*.

[28]Brereton, R. G. Chemometrics. Data Analysis for the Laboratory and Chemical Plant. Wiley, Chichester, 2003 pp 241-243; 249–250.

[29]Rácz, A., Gere, A., Bajusz, D., Héberger, K. Is soft independent modeling of class analogies a reasonable choice for supervised pattern recognition? *RSC Adv.,* **2018**, *8*, 10–21.

[30]Vandeginste G.M., Massart D.L., Buydens M.C.-Handbook Of Chemometrics and Qualimetrics part B. Elsevier: Amsterdam, 1998. p.224.

[31]Cover, P.M., Hart, P.E. Nearest neighbor pattern classification*. IEEE Trans. Inf. Theory IT-13,* **1966**, *1*, 21–27.

[32]Kohonen, T. Automatic formation of topological maps of patterns in a self-organizing system. *In*: Oja,E., Simula, O. (Eds.). Proc. 2SCIA, *Scand. Conf. on Image Analysis,* **1981**, 214-220.

[33]Kohonen, T. Construction of similarity diagram for phonemes by a self-organizing algorithm. Report TKK – F A463, Helsinki Uni. Technol., Finland, 1981.

[34]Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.,* **1982**, *43*, 4359–4360.

[35]Flanagan, J. Self-organisation in Kohonen's SOM. *Neural Netw.,* **1996**, *7 (9)*, 1185–1197.

[36]Milone, D.H., Stegmayer, G., Kamenetzky, L., López, M., Carrari, F. Clustering biological data with SOMs: On topology preservation in non-linear dimensional reduction. *Expert Syst. Appl.,* **2013**, *40 (9)*, 3841–3845.

[37]Ballabio, D., Vasighi, M., Consonni, V*.* Effects of supervised Self Organising Maps parameters on classification performance*. Anal. Chim. Acta,* **2013**, *46*, 45–53.

[38]Ghobadi, M.Z., Kompany-Zareh, A. Application of supervised Kohonen map and counter propagation neuralnetwork for classification of nucleic acid structures based on their circular dichroism spectra. *Spectrochim. Acta A,* **2014**, *132*, 345–354.

[39]Kohonen, T. Essentials of the self-organizing map. *Neural Netw.,* **2013**, *37*, 52–65.

[40]Kohonen, T. Self-organizing maps: optimization approaches. In T. Kohonen,K. Mäkisara, O. Simula, & J. Kangas (Eds.), Artificial neural networks, II (pp. 981–990). Amsterdam, Netherlands: North-Holland, 1991.

[41]Abuiziah,I., Shakarneh, N. A Review of Genetic Algorithm Optimization:Operations and Applications to Water Pipeline Systems. *Int. J. Math. Comp. Sci.,* **2013**, 7 (12), 1782–1788.

[42]Kumar, M., Huslan M., Upreti, N., Gupta, D. Genetic Algorithm: Review and Application. *Int. J. Inf. Tec. Know. Manag.,* **2010**, *2 (2)*, 2451– 2454.

[43]Mukhopadhyay, D.M., Balitanas, M.O., Farkhod, A., Jeon,S.-H., Bhattacharyya, D. Genetic Algorithm: A Tutorial Review. *Int. J. Grid. Dist. Comp.,* **2009**, *2 (3)*, 25–32.

[44]Loh, W.-Y. Classification and Regression Tree Methods. In Encyclopedia of Statistics in Quality and Reliability. Ruggeri, Kenett and Faltin (eds.) pp. 315–323: Wiley, 2008.

[45]Breiman, L., Friedman, J., Olshen, R. and Stone, C. Classification and Regression Trees. Belmont, California: Wadsworth, 1984, 368 p.

[46]Morgan, J.N., Sonquist, J.A. Problems in the analysis of survey data, and a proposal. *J. Am. Stat. Assoc.,* **1963**, *58*, 415–434.

[47]Messenger, R.C., Mandell, M.L. A model search technique for predictive nominal scale multivariate analysis. *J. Am. Stat. Assoc.,* **1972**, *67*, 768–772.

[48]Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.,* **1936**, *7 (2)*, 179–188.

[49]Doyle, P. The use of automatic interaction detector and similar search procedures. *Oper. Res. Quart.,* **1973**, *24*, 465–467.

[50]Liu, J. Terahertz spectroscopy and chemometric tools for rapid identification of adulterated dairy product. *Opt. Quant. Electron.,* **2017**, *49*, 1.

[51]Bougrini, M., Thari, K., Haddi, Z., El Bari, N., Llobet, E., Jaffrezic-Renault, N., Bouchikhi, B. Aging time and brand determination of pasteurized milk using a multisensor e-nose combined with a voltammetric e-tongue. *Mater. Sci. Eng. C,* **2014**, *45*, 348–358.

[52]Scholl, P.F., Bergana, M.M., Yakes, B.J., Xie, Z., Zbylut S., Downey, G., Mossoba, M., Jablonski, J., Magaletta, R., Holroyd, S.E., Buehler, M., Qin, J., Hurst, W., LaPointe, J.H., Roberts, D., Zrybko, C., Mackey, A., Holton, J.D., Israelson, G.A., Payne, A., Kim, M.S., Chao, K., Moore, J.C. Effects of the Adulteration Technique on the Near-Infrared Detection of Melamine in Milk Powder. *J. Agric. Food Chem.* **2017** *65*, 5799–5809.

[53]Gondim, C. dos S., Junqueira, R.G., Souza, S.V.C., Ruisánchez I., Callao, M.P. Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies. *Food Chem.,* **2017**, *230*, 68–75. (a)

[54]Gondim, C. dos S., Junqueira, R.G., de Souza, S.V.C., Callao, M.P., Ruisánchez, I.Determining performance parameters in qualitative multivariate methods using probability of detection (POD) curves. Case study: Two common milk adulterants. *Talanta,* **2017**, *168*, 23–30. (b)

[55]Santos, P.M., Pereira-Filho, E.R., Rodriguez-Saona. Rapid detection and quantification of milk adulteration using infrared microspectroscopy and chemometrics analysis. *Food Chem.,* **2013**, *138*, 19–24.

[56]Jaiswal, P., Jha, S.N., Kaur, J., Borah, A. Detection and quantification of anionic detergent (lissapol) in milk using attenuated total reflectance-Fourier Transform Infrared spectroscopy. *Food Chem.,* **2017**, *221*, 815-821.

[57]Santos, P.M., Pereira-Filho, E.R., Colnago, L.A. Detection and quantification of milk adulteration using time domain nuclear magnetic resonance (TD-NMR). *Microchem. J.,* **2013**, *124*, 15–19.

[58]Wu, T., Chen, H., Lin, Z., Tan, C. Identification and Quantitation of Melamine in Milk by Near-Infrared Spectroscopy and Chemometrics. *J Spectrosc.,* **2016**, *XX*, 1–8.

[59]Callao, M.P; Ruisánchez, I. An overview of multivariate qualitative methods for food fraud detection (review). *Food Control,* **2018**, *86*, 283–293.

[60]Marengo, M., Aceto, A., Maurino, V. *J.* [60]Marengo, M., Aceto, A., Maurino, V. Classification of Nebbiolo-based wines from Piedmont (Italy) by means of solid-phase microextraction–gas chromatography–mass spectrometry of volatile compounds. *J. Chromatogr. A,* **2001**, *943*, 123–137.

[61]Díaz, C., Conde, J.E., Estévez, D., Olivero, S.J.P., Trujillo, J.P.P. Application of Multivariate Analysis and Artificial Neural Networks for the Differentiation of Red Wines from the Canary Islands According to the Island of Origin. *J. Agric. Food. Chem.,* **2003**, *51*, 4303–4307.

[62]Urruty, L., Giraudel, J.-L., Lek, S., Roudeillac, P., Montury, M. Assessment of Strawberry Aroma through SPME/GC and ANN Methods. Classification and Discrimination of Varieties. *J. Agric. Food. Chem.*, **2002**, *50*, 3129–3136.

[63]De Boishebert, V., Urruty, L., Giraudel, J.-C., Montury, M. Assessment of Strawberry Aroma through SPME/GC and ANN Methods. Classification and Discrimination of Varieties. *J. Agric. Food. Chem.,* **2002**, *52,* 2472–2478.

[64]Nadal M., Espinosa G., Schuhmacher M., Domingo J.L. Patterns of PCDDs and PCDFs in human milk and food and their characterization by artificial neural networks. *Chemosphere,* **2004**, *54*, 1375–1382.

[65]Kamal, M., Karoui, R. Analytical methods coupled with chemometric tools for determining the authenticity and detecting the adulteration of dairy products: A review (review). *Trends Food Sci. Technol.,* **2015**, *46*, 27–48.

[66]De Baerdemaker, J., Karoui, R. *Food Chem.* A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products (review). *Food Chem.,* **2007**, *102*, 621–640.

[67]Lohumi, S., Lee, S., Lee, H., Cho, B.-K. review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration (review). *Trends Food Sci. Technol.,* **2015**, *46*, 85–98.

[68]Zupan, J., Novič, M., Ruizsánchez, I. Kohonen and counterpropagation artificial neural networks in analytical chemistry (tutorial). *Chemom. Intell. Lab. Syst.,* **1997**, *38*, 1–23.

[69]Borges, C., Gómez-Carracedo, M.P., Andrade, J.M., Duarte, M.F., Biscaya, J.L., Aires-de-Sousa, J. Geographical classification of weathered crude oil samples with unsupervised self-organizing maps and a consensus criterion. *Chemom. Intell. Lab. Syst.*, **2010**, *101*, 43–55.

[70]Gómez-Carracedo, M.P., Andrade, J.M., Carrera, G.V.S.M., Aires-de-Sousa, J., Carlosena, A., Prada, D. Combining Kohonen neural networks and variable selection by classification trees to cluster road soil samples. *Chemom. Intell. Lab. Syst.,* **2010**, *102*, 20–34.

[71]Kittiwachana, S., Ferreira, D. L. S., Fido, L.A., Thompson, D. R, Escott, R. E. A., Brereton, R. G. Self-Organizing Map Quality Control Index. *Anal. Chem*., **2010**, *8*, 5972–5982.

[72]Affonso, G.A. Mapas Auto-organizáveis de Kohonen (SOM) aplicados na avaliação dos parâmetros da qualidade da água (Dissertação de mestrado). Insituto de Pesquisas Energéticas e Nucleares, 89 p., 2011.

[73]Shahbazy, M. Vasighi M., Kompany-Zareh M., Ballabio D. Oblique rotation of factors: a novel pattern recognition strategy to classify fluorescence excitation-emission matrices of human blood plasma for early diagnosis of colorectal cancer. *Mol. Biosyst.,* **2016**, *12 (6)*, 1963–1975.

[74]Burikov, S. A., Dolenko, T. A., Gushchin, K. A., Dolenko, S. A. Kohonen Self-Organizing Maps as a New Method for Determination of Salt Composition of Multi-Component Solutions. *Spectrochim. Acta A,* **2014**, *8 (10)*, 1130–1134.

[75]Deljanin, I., Antanasijević, D., Urošević, M.A., Tomašević, M., Perić-Grujić, A., Ristić, M. The novel approach to the biomonitor survey using one- and two-dimensional Kohonen networks. *Environ. Monit. Assess.,* **2015**, *187* (10), 618–628.

[76]Brasil. Ministério da Agricultura, Pecuária e Abastecimento. RIISPOA. MAPA/SDA/DIPOA, 2007, 252 p.

[77]Brasil. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n° 68 de 12 de dezembro de 2006. Disponível em: http://extranet.agricultura.gov.br/sislegis-consulta/consultarLegislacao.do?operacao=visualizar&id=17472. Acesso em 25/01/2017.

[78]Lane, J. H., & Eynon, L. (1934). Determination of reducing sugars by Fehling's solution with methylene blue indicator. London: Norman Rodger, 8p.

[79]United States of America (2006). Food and Drug Administration, HHS. 21 CFR Ch. I (4–1–06 Edition), § 131.110. Milk.

[80]People's Republic of China (2010). Ministry of Health of the People's Republic of China.China GB 19301-2010 National food safety standard. Raw milk. (a)

[81]European Union (2013). The European Parliament and the Council of the European Union. Regulation (EU) No 1308/2013 of the European Parliament and of the Council of 17 December 2013 establishing a common organisation of the markets in agricultural products and repealing Council Regulations (EEC) No 922/72, (EEC) No 234/79, (EC) No 1037/2001 and (EC) No 1234/2007.

[82]India (2017). Food Safety and Standards Authority of India. Directions under Section 16(5) of the Food Safety and Standards Act, 2006 regarding operationalization of amendment regulations regarding revised standards for milk, dated August 2nd, 2017.

[83]Russian Federation (2008). Federal Law dated 22.07.2010 No. 163- FZ. Technical regulations for milk and milk products.

[84]Japan (1951). Ministry of Health and Welfare Ordinance No. 52, December 27, 1951. Ministerial Ordinance on Milk and Milk products Concerning Compositional Standards, etc.

[85]Australia (2015). Australia New Zealand Food Standards Code – Standard 2.5.1 – Milk.

[86]ISO (2010). ISO 6731:2010 (IDF 21:2010). Milk, cream and evaporated milk – Determination of total solids content (Reference method). (a)

[87]ISO (2010). ISO 1211:2010 (IDF 1:2010). Milk – Determination of fat content – Gravimetric method (Reference method). (b)

[88]ISO (2016). ISO 8968-4:2016 (IDF 20-4). Milk and milk products – Determination of nitrogen content — Part 4: Determination of protein and non-protein nitrogen content and true protein content calculation (Reference method).

[89]ISO (2007) ISO 22662:2007. Milk and milk products — Determination of lactose content by high-performance liquid chromatography (Reference method).

[90]ISO (2009). ISO 5764:2009 (IDF 108:2009). Milk — Determination of freezing point — Thermistor cryoscope method (Reference method).

[91]ISO (2010). ISO 6091:2010. Dried milk — Determination of titratable acidity (Reference method). (c)

[92]ISO (2008). ISO 488:2008 (IDF 105:2008). Milk — Determination of fat content — Gerber butyrometers.

[93]ISO (1980) ISO 6092:1980. Dried milk -- Determination of titratable acidity (Routine method).

[94]India (2015). Ministry of Health and Family Welfare. Manual of Methods of Analysis of Foods. Milk and Milk products.

[95]People’s Republic of China (2010). Ministry of Health of the People’s Republic of China. GB 5413.5-2010. National food safety standard. Determination of

lactose and sucrose in foods for infants and young children, milk and milk products. (b)

[96]People's Republic of China (2010). Ministry of Health of the People's Republic of China. GB 5009.4-2010. National food safety standard. Determination of Ash in Foods. (c)

[97]People's Republic of China (2010). Ministry of Health of the People's Republic of China. China GB 5413.33-2010. National food safety standard. Determination of specific gravity in raw milk. (d)

[99]People's Republic of China (2010). Ministry of Health of the People's Republic of China. China GB 5009.5-2010 National Food Safety Standard Determination of protein in foods. (e)

[100]People's Republic of China (2010). Ministry of Health of the People's Republic of China.China GB 5413.34-2010 National food safety standard. Determination of acidity in milk and milk products. (f)

[101]People's Republic of China (2010). Ministry of Health of the People's Republic of China.China GB 5413.39— 2010 National food safety standard. Determination of Nonfat Total Milk Solids in Milk and Milk Products. (g)

[102]Mironiuk, M., Barańska, M., Chojnacka, K., Górecki, H. Determination of the reference value of nitrogen mass fraction in the reference material of Polish soil. *Accred. Qual. Assur.,* **2016**, *21*, 409–415.

[103]Wojciechowski, K.L., Melilli C., Barbano D.M. A proficiency test system to improve performance of milk analysis methods and produce reference values for component calibration samples for infrared milk analysis. *J. Dairy Sci.,* **2016**, *99*, 6808–6827.

[104]Rezende, P.S., Carmo G.P.D., Esteves E.G. Optimization and validation of a method for the determination of the refractive index of milk serum based on the reaction between milk and copper(II) sulfate to detect milk dilutions. *Talanta,* **2015**, *138*, 196–202.

[105]Vacchina, V., Séby, F., Chekri, R., Verdeil, J., Dumont, J., Hulin, M., Sirot, V., Volatier, J.L., Serreau, R., Rousseau, A., Simon, T., Guérin, T. Optimization and validation of the methods for the total mercury and methylmercury determination in breast milk. *Talanta,* **2017**, *167*, 404–410.

[106]Helfer, G. A., Bock, F., Marder, L., Furtado, J.C., da Costa, A.B., Ferrão, M.F. CHEMOSTAT, um software gratuito para análise exploratória de dados multivariados. *Quím. Nova,* **2015**, *38*, 575–579.

[107]Ballabio, D., Consonni, V. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Anal. Methods,* **2013**, *5*, 3790-3798.

[108]Ballabio, D., Consonni, V., Todeschini R. The Kohonen and CP-ANN toolbox: A collection of MATLAB modules for Self Organizing Maps and Counterpropagation Artificial Neural Networks. *Chemom. Intell. Lab. Syst.,* **2009**, *98*, 115–122

[109]Kennard, R.W., Stone, L.A. Computer Aided Design of Experiments. *Technometrics,* **1969**, *1*, 137–148.

[110]Haddad, K., Rahman, A., Zaman, M., Shrestha, S. Applicability of Monte Carlo cross validation technique for model development and validation using generalised least squares regression. *J. Hydrol.* (Amst.), **2013**, *482*, 119–128.

[111]Xu, Q.S.; Liang, Y-Z. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst*., **2001**, *56*, 1–11.

[112]Mount, NJ, Weaver, D. Self-organizing maps and boundary effects: Quantifying the benefits of torus wrapping for mapping SOM trajectories. *Pattern Anal. Appl.,* **2011**, *142*, 139–148.

[113]Falk, F.R., Miller, N.B. A Primer for Soft Modelling. First Edition. University of Akron: Akron, Ohio, 1992. p.80.

[114]SRD 69 NIST Chemistry WebBook. Acesso em 12/08/2017.

[115]BSI (1988). BS 3095-1.2:1988. Methods for determination of the freezing-point depression of milk. Methods. Hortvet method.

[116]United States National Conference of State Legislatures (NSCL) (2016). State Milk Laws. http://www.ncsl.org/research/agriculture-and-rural-development/raw-milk-2012.aspx. Acesso on 12/08/2017.

[117]Schönfeldt, H.C., Hall, N.G, Smit, L.E. The need for country specific composition data on milk. *Food Res. Int.,* **2012**, *47*, 207-209.