Abordagem Integrada para Análise de Qualidade do Leite Utilizando Parâmetros Físicos e Técnicas Computacionais

Victor Rodrigues Botelho[a,b,*], Jorge Otávio Trierweiler[a], Marcelo Farenzena[a], Renato Dutra Pereira Filho[b], Walter Augusto Ruiz[b]

[a]*Grupo de Intensificação, Modelagem, Simulação, Controle e Otimização de Processos (GIMSCOP) Departamento de Engenharia Química, Universidade Federal do Rio Grande do Sul (UFRGS)*
[b]*Universidade Federal do Rio Grande, EQA/LaCoPQ, Rio Grande/RS, Brasil*
**victor.botelho@ufrgs.br*

RESUMO

A qualidade dos alimentos deve atender às normas estabelecidas pelas autoridades sanitárias, e a fraude é um crime que pode reduzir o valor nutricional e causar riscos à saúde. Este trabalho teve como objetivo desenvolver um procedimento computacional para a detecção de fraudes em amostras de leite com base na espectroscopia elétrica de impedância. Foram utilizados conjuntos de dados quantitativos obtidos da literatura especializada, contendo propriedades do leite. A modelagem foi realizada por meio de métodos de aprendizado não supervisionados, como K-means, e supervisionados, como Random Forest, Support Vector Machine e K-Nearest Neighbor. A validação dos modelos foi conduzida por meio de métricas como acurácia, precisão, revocação e medida F. Os resultados indicaram que o K-Nearest Neighbors apresentou melhor desempenho na identificação de fraudes por adição de água, água deionizada e formaldeído, enquanto K-Nearest Neighbors e K-means se destacaram na detecção de fraudes relacionadas ao teor de gordura.

Palavras-chave: Agricultura 4.0; Aprendizado de máquina; Quimiometria.

1. Motivação

A garantia da qualidade e da identidade dos produtos lácteos é essencial para atender às normas sanitárias e evitar fraudes, que podem comprometer tanto a saúde dos consumidores quanto a cadeia produtiva. Nesse contexto, a química analítica tem avançado no desenvolvimento de técnicas instrumentais, como a espectroscopia elétrica de impedância (EEI), capaz de correlacionar propriedades físicas e elétricas para fins analíticos. Aliada a conceitos da Agricultura 4.0, como inteligência artificial e análise de dados, essa abordagem possibilitou métodos mais rápidos, confiáveis e eficientes para o monitoramento da qualidade do leite, especialmente diante de adulterações comuns, como a adição de água e de outros agentes químicos, ou a adulteração na quantidade de gordura (Caicedo-Eraso; Díaz-Arango; Osorio-Alturo, 2020).

Paralelamente, o uso de procedimentos computacionais e de modelagem empírica permitiu extrair relações entre variáveis complexas, mesmo na ausência de modelos teóricos bem definidos. A aplicação de técnicas de aprendizado de máquina, integrando métodos estatísticos e inteligência artificial, viabilizou a construção de modelos capazes de analisar padrões em dados experimentais. Dessa forma, a combinação entre EEI, processamento computacional e machine learning constituiu uma abordagem promissora para a avaliação da qualidade e da integridade de amostras de leite (Subasi, 2020).

2. Metodologia

2.1. Obtenção e avaliação de datasets

Foram utilizadas publicações que contivessem dados sobre as propriedades do leite e que pudessem ser correlacionados aos conceitos de EEI (Ahmad; Komolavanij; Chanvarasuth, 2010; Vique; Marichal; Steinfeld, 2020).

As informações foram categorizadas em sensoriamento, design experimental, eletroquímica, espectroscopia, modelos e simulação; assim, foram selecionados datasets de propriedades e análise do tipo de fraude do leite (Durante et al., 2016; Muyambo, 2018). Ambos constituíram os conjuntos de dados dataset Q e dataset G, respectivamente.

2.2. Construção dos cenários experimentais

A avaliação foi realizada em cenários sintéticos controlados para analisar o comportamento da abordagem em diferentes condições. Dados sintéticos permitem ajustar a separabilidade, o ruído e a correlação, sendo adequados para estudos metodológicos. Foram considerados dois cenários: um com alta separabilidade e outro com sobreposição entre as classes. Isso permitiu avaliar o desempenho e a sensibilidade à incerteza, trazendo múltiplos cenários controlados, essenciais para validar a robustez e a confiabilidade (Rajotte et al., 2022).

Para cada cenário, os dados foram organizados em diferentes níveis de dificuldade e divididos em conjuntos de treinamento e de teste. O primeiro cenário apresentou fronteiras bem definidas, favorecendo métodos tradicionais; já o segundo introduziu ambiguidade, simulando problemas reais, permitindo avaliar o desempenho e o comportamento da rejeição por confiança em diferentes níveis de incerteza (Amgoud; Doder; Vesic, 2022).

2.3. Análise Estatística Multivariada e escolha dos algoritmos

Foram realizadas análises de componentes principais e FreeViz no programa PAST (PHONE, 2021). As análises no PAST foram complementadas pelo Orange Data Mining, uma plataforma que oferece um front-end de programação visual, permitindo análises exploratórias de dados qualitativos, com visualização rápida e interativa. O algoritmo de aprendizado não supervisionado selecionado foi o k-means e os algoritmos de aprendizado supervisionado foram: random forest, support vector machine e k-nearest neighbors.

2.4. Análise de Projeção e Visualização de Dados Multivariados

A Análise de Componentes Principais (PCA) é uma técnica estatística utilizada para reduzir a dimensionalidade de conjuntos de dados, preservando a maior quantidade possível de informação relevante. A partir da centralização dos dados e da análise de sua variabilidade, o método identifica novas direções (componentes principais) que representam combinações das variáveis originais. Essas direções são determinadas de modo a capturar, em ordem decrescente, a maior variância presente nos dados, permitindo reorganizar a informação de forma mais compacta e interpretável (Jolliffe; Cadima, 2016).

O FreeViz é um método de visualização de dados que projeta conjuntos multidimensionais em um espaço de menor dimensão, geralmente bidimensional, com o objetivo de facilitar a interpretação e separação entre classes. Ele modela os dados como partículas em um sistema físico, no qual instâncias exercem forças de atração ou repulsão entre si, de acordo com sua similaridade ou diferença. A posição final dessas instâncias no espaço projetado é ajustada por meio de um processo iterativo que busca uma configuração de equilíbrio (Vique; Marichal; Steinfeld, 2020).

2.5. Análise de Algoritmos de Aprendizado de Máquina para Classificação e Agrupamento

O K-means é um algoritmo de aprendizado não supervisionado utilizado para agrupar dados em um número pré-definido de grupos, chamados clusters. Seu objetivo é organizar as amostras de modo que os elementos de um mesmo grupo sejam mais semelhantes entre si do que aos de outros grupos. Para isso, o método associa cada ponto de dado ao centroide mais próximo, que representa o centro de cada cluster (Ling, 2020).

O Random Forest é um algoritmo de aprendizado supervisionado baseado na combinação de múltiplas árvores de decisão. Cada árvore é treinada com subconjuntos aleatórios de dados e variáveis, o que aumenta a diversidade do modelo. A predição final é obtida pela agregação das respostas das árvores, geralmente por votação.

Essa abordagem reduz o sobreajuste e melhora a capacidade de generalização, sendo um método robusto e eficaz para tarefas de classificação e regressão (Breiman, 2001).

O Support Vector Machine (SVM) é um algoritmo de aprendizado supervisionado utilizado para classificação, que busca definir um limite capaz de separar diferentes classes de forma eficiente. Ele constrói um hiperplano que maximiza a distância entre os grupos de dados, melhorando a separação entre eles. Além disso, pode lidar com dados não linearmente separáveis por meio de transformações em espaços de maior dimensão, método amplamente difundido e reconhecidamente robusto (Yang; Shao; Zhang, 2013).

O K-Nearest Neighbors (KNN) é um algoritmo de aprendizado não supervisionado utilizado para classificação com base na proximidade entre os dados. Para uma nova amostra, identifica os vizinhos mais próximos no conjunto de treinamento e atribui a classe mais frequente entre eles. Seu desempenho depende da escolha do número de vizinhos e da métrica de distância, sendo um método simples e eficaz para reconhecimento de padrões (Bloice; Holzinger, 2016).

2.6. Procedimentos e validação

Foi utilizado código livre, com programação em Python, aliado a ambientes pré-estruturados, como as bibliotecas auxiliares. Por meio da plataforma de distribuição Anaconda, foi acessado o Jupyter Notebook, versão 6.3.0, no qual foi realizada a programação com o suporte de pacotes de ciência de dados, o que facilita o acesso às bibliotecas. Dentre as bibliotecas utilizadas, destaca-se a Scikit-Learn, que auxilia na implementação dos algoritmos citados anteriormente (Martinelli-Orlando; Shi; Angst, 2020).

Para validar e ajustar os procedimentos computacionais, foram utilizadas métricas de desempenho amplamente empregadas em modelos de classificação: acurácia, precisão, revocação e medida F. A acurácia indica a proporção de predições corretas realizadas pelo modelo, enquanto a precisão avalia a confiabilidade das predições positivas. Já a revocação mede a capacidade do modelo de identificar corretamente os casos positivos, também conhecida como sensibilidade. A medida-F combina precisão e revocação em um único indicador, permitindo avaliar o equilíbrio entre essas métricas (Parisi et al., 2021).

Além disso, foi considerada a performance média, obtida a partir da combinação dessas métricas, com o objetivo de fornecer uma visão global do desempenho dos modelos. Esse conjunto de indicadores permitiu uma análise mais completa e confiável, auxiliando na comparação entre diferentes abordagens e na escolha do modelo mais adequado. A Tabela 1 mostra a relação entre a performance média e a confiabilidade do algoritmo. A performance foi calculada através da média entre acurácia, precisão, revocação e medida-F (Khan; Lee, 2019).

Tabela 1. Relação entre performance média e confiabilidade do algoritmo.

Performance Média	Confiabilidade
0	Nula
$0 < PM < 0,4$	Baixa
$0,4 \leq PM < 0,7$	Moderada
$0,7 \leq PM < 1$	Alta
1	Exata

3. Resultados

3.1. Análises PCA e FreeViz

Para o dataset Q, o PCA indicou uma variância cumulativa de 0,664 e uma variância por componente de 0,161. Tais valores mostram que o somatório das razões de variância está abaixo do mínimo indicado (0,7), e que a variância entre componentes apresenta valores significativos, dada a magnitude das variáveis. Já para o dataset G foi perceptível uma variância cumulativa de 0,989 e uma variância por componente de 0,111. Tais valores mostram que o somatório das razões de variância está dentro da faixa indicada, entre 0,7 e 1, o que indica que os componentes analisados são importantes para o prosseguimento das análises, e que a variância entre componentes não apresenta valores significativos, dada a magnitude das variáveis.

Para o dataset Q, há uma correlação entre a adulteração com peróxido de hidrogênio e o leite sem adulteração. A análise FreeViz, possibilitou obter os primeiros indícios dessa possível sobreposição entre os dados

do leite sem adulteração e do leite adulterado com peróxido de hidrogênio, o que inicialmente não demandou nenhuma ação, mas é informação relevante no decorrer da apresentação dos resultados. Para o dataset G apenas algumas das variáveis de fato influenciavam a mineração a seguir, o que foi de suma importância no intuito de simplificar as variáveis de alimentação, possibilitando o uso das mais facilmente mensuráveis para prosseguir com a modelagem empírica.

3.2. Procedimentos computacionais utilizando dataset Q

A informação do dataset Q consistiu em impedância primária (Z'), impedância secundária (Z'') e frequência (F). Análise de correlação identificou que a frequência apresenta uma correlação entre moderada tendendo a alta, o que significa piores resultados utilizando esta variável enquanto os procedimentos computacionais utilizando apenas Z' e Z'' apresentaram melhores resultados como mostrado na matriz de correlação da Tabela 2.

Tabela 2. Matriz de correlação para o conjunto de dados de químicos.

	Z	**Z''**	**F**
Z'	1	0,214	0,194
Z''	0,214	1	0,611
F	0,194	0,611	1

Com esses resultados, foram testadas várias hipóteses para encontrar o melhor modelo através do aprendizado não supervisionado, testes K-means e de aprendizado supervisionado com Random Forest, Support Vector Machine e K-Nearest Neighbors. A Tabela 3 mostra quais componentes cada modelo empírico é capaz de estimar.

Tabela 3. Matriz de correlação para o conjunto de dados de químicos.

	Leite e água	**Leite e água deionizada**	**Leite e formaldeído 37%**	**Leite e peróxido de hidrogênio**
K-means	✓	A	X	B
Support Vector Machine	✓	X	✓	M
Random Forest	✓	A	✓	B
K-Nearest Neighbor	✓	✓	✓	X

Legenda: X – Incapaz de predizer; B – Baixa confiabilidade de predição; M – Moderada confiabilidade de predição; A – Alta confiabilidade de predição ✓ – Predição exata

Ao comparar os modelos empíricos apresentados com base nas métricas estabelecidas, percebeu-se que o melhor resultado foi obtido com o modelo K-Nearest Neighbors, pois foi capaz de predizer com exatidão as fraudes por água, água deionizada e formaldeído a 37%. De um modo geral, se tem que o modelo K-means apresentou os piores resultados, sendo capaz de predizer com exatidão apenas leite e água e com alta confiabilidade leite e água deionizada, a seguir, apresentando melhora, se tem o modelo Support Vector Machine, capaz de predizer com exatidão leite e água e leite e formaldeído 37% e com moderada confiabilidade leite e peróxido de hidrogênio, e, por fim, o modelo Random Forest, que conseguiu predizer com exatidão leite e água e leite e formaldeído 37% e com alta confiabilidade leite e água deionizada.

3.3. Procedimentos computacionais utilizando dataset G

Para o dataset G, as informações disponíveis eram percentual de gordura (%g), volume (V), condutividade elétrica (σ), temperatura (T), fração de volume de leite integral (VI/V) e fração de volume de leite desnatado (VD/V). Inicialmente, foi feita uma avaliação utilizando o Orange Data Mining, tanto através de PCA quanto de FreeViz, na qual se pôde verificar que seria possível prosseguir a análise utilizando apenas as variáveis mais facilmente mensuráveis, tais como temperatura, condutividade elétrica e volume.

Após tal definição, uma série de hipóteses foram testadas até se encontrar o melhor modelo. Os testes foram feitos utilizando aprendizado não supervisionado, através de K-means, e aprendizado supervisionado com

Random Forest, Support Vector Machine e K-Nearest Neighbors. A Tabela 4 apresenta qual tipo de leite e qual intervalo de confiabilidade cada modelo é capaz de predizer.

Tabela 4. Intervalo de confiabilidade das predições de acordo com os modelos utilizados.

	Leite integral	**Leite semidesnatado**
K-means	96,52% a 98,06%	61,99% a 98,78%
Support Vector Machine	87% a 100%	69% a 96%
Random Forest	84,48% a 100%	100%
K-Nearest Neighbor	100%	83,33% a 100%

Comparando os modelos empíricos apresentados com base nas métricas de validação estabelecidas, é possível perceber que todos foram capazes de realizar predições assertivas e com boa confiabilidade. Apesar desse bom desempenho de modo geral, dois modelos se destacaram: K-Nearest Neighbors e K-means. Isso se deu devido ao fato do modelo K-Nearest Neighbors afirmar com 100% de confiabilidade todas as predições inerentes ao leite integral, enquanto o modelo K-means afirmou com tal confiabilidade as predições inerentes ao leite semidesnatado.

O presente trabalho apresenta contribuições ao aplicar tanto algoritmos de aprendizado supervisionado quanto não supervisionado, aliados a uma técnica oriunda da química analítica de processos, a EEI, no intuito de predizer fraudes em leite, tanto por adição de substâncias químicas quanto por gordura.

4. Conclusão

Foram obtidos conjuntos de dados que utilizaram EEI: o dataset Q, com dados de adulteração do leite com água, água deionizada, formaldeído a 37% e peróxido de hidrogênio; e o dataset G, com dados referentes ao percentual de gordura do leite.

Através de análises estatísticas multivariadas e modelagens empíricas, utilizando técnicas de mineração de dados, aplicando algoritmos tanto de aprendizado não supervisionado quanto de supervisionado, foi possível construir um procedimento para detecção de fraudes no leite

Foram elaborados diversos procedimentos computacionais em Python, estruturados com 70% dos dados de cada conjunto e 30% para validação, e os resultados foram comparados por meio de métricas qualitativas, como performance média, e quantitativas, como percentual de confiabilidade das predições, a fim de decidir pelos melhores modelos para a detecção das fraudes estudadas.

O procedimento construído utilizando um modelo baseado em K-Nearest Neighbors foi capaz de identificar, com exatidão, adulterações com água, água deionizada e formaldeído e também foi capaz de identificar se o leite é classificado como integral, por meio de um modelo baseado em K-Nearest Neighbors ou se é semidesnatado, através de um modelo K-means, e em ambos os casos com exatidão.

Outros modelos empíricos, também apresentaram bons resultados, principalmente em se tratando do estudo do tipo de leite quanto à quantidade de gordura. Entretanto, nenhum deles foi capaz de atingir a exatidão que os modelos baseados em K-Nearest Neighbors e K-means alcançaram.

Houve limitações quanto à detecção do peróxido de hidrogênio, pois, perante a impedância primária, a impedância secundária e a frequência, que foram as variáveis disponíveis no dataset, o leite sem adulteração era praticamente indistinguível do adulterado com peróxido de hidrogênio, o que impossibilitou o treinamento do algoritmo exclusivamente para fraude utilizando essa substância.

Com os resultados obtidos, aplicando o procedimento elaborado e testado, embasado nos conceitos da Agricultura 4.0, foi possível concluir que há uma potencial redução na necessidade de análises e testes físico-químicos, tornando mais fácil, rápido e econômico o processo de detecção de fraudes nas três frentes possíveis: diretamente no produtor, no transporte e no processo de industrialização.

Referências

AHMAD, Imran; KOMOLAVANIJ, Somrote; CHANVARASUTH, Pisit. Prediction of Raw Milk Microbial Quality Using Data Mining Techniques. **Agricultural Information Research**, v. 19, n. 3, p. 64–70, 2010.

AMGOUD, Leila; DODER, Dragan; VESIC, Srdjan. Evaluation of argument strength in attack graphs: Foundations and semantics. **Artificial Intelligence**, v. 302, p. 103607, 1 jan. 2022.

BLOICE, Marcus D.; HOLZINGER, Andreas. A Tutorial on Machine Learning and Data Science Tools with Python. *In*: HOLZINGER, Andreas (Org.). **Machine Learning for Health Informatics**. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016. v. 9605 p. 435–480.

BREIMAN, Leo. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CAICEDO-ERASO, Julio César; DÍAZ-ARANGO, Félix Octavio; OSORIO-ALTURO, Andrea. Espectroscopia de impedancia eléctrica aplicada al control de la calidad en la industria alimentaria. **Ciencia & Tecnología Agropecuaria**, v. 21, n. 1, p. 1–20, 2020.

DURANTE, Gabriel *et al.* Electrical Impedance Sensor for Real-Time Detection of Bovine Milk Adulteration. **IEEE Sensors Journal**, v. 16, n. 4, p. 861–865, fev. 2016.

GORRIZ, J. M. *et al.* Statistical Agnostic Mapping: A framework in neuroimaging based on concentration inequalities. **Information Fusion**, v. 66, p. 198–212, 1 fev. 2021.

HENDRICKX, Kilian *et al.* Machine learning with a reject option: a survey. **Machine Learning**, v. 113, n. 5, p. 3073–3110, 1 maio 2024.

JOLLIFFE, Ian T.; CADIMA, Jorge. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 374, n. 2065, p. 20150202, 13 abr. 2016.

KHAN, Jebran; LEE, Sungchang. Implicit User Trust Modeling Based on User Attributes and Behavior in Online Social Networks. **IEEE Access**, v. 7, p. 142826–142842, 2019.

LING, Shuyang. k-means Clustering. 4 mar. 2020.

MARTINELLI-ORLANDO, Federico; SHI, Wei; ANGST, Ueli. Corrosion Behavior of Carbon Steel in Alkaline, Deaerated Solutions: Influence of Carbonate Ions. **Journal of The Electrochemical Society**, v. 167, n. 6, p. 061503, 20 mar. 2020.

MUYAMBO, Shadreck. Mathematical model for predicting butterfat concetration in bovine milk using electric conductivity properties, temperature and volume fraction of whole milk. v. 1, 15 dez. 2018.

PARISI, Luca *et al.* hyper-sinh: An accurate and reliable function from shallow to deep learning in TensorFlow and Keras. **Machine Learning with Applications**, v. 6, p. 100112, dez. 2021.

PHONE, VISITING ADDRESS Natural History MuseumSars' gate 1. 0562 OSLO Norway MAIL ADDRESS P. O. Box 1172 Blindern 0318 Oslo Norway. **Past 4 - the Past of the Future - Natural History Museum**. Disponível em: <https://www.nhm.uio.no/english/research/infrastructure/past/index.html>. Acesso em: 22 dez. 2021.

RAJOTTE, Jean-Francois *et al.* Synthetic data as an enabler for machine learning applications in medicine. **iScience**, v. 25, n. 11, p. 105331, 18 nov. 2022.

SUBASI, Abdulhamit. Machine learning techniques. *In*: **Practical Machine Learning for Data Analysis Using Python**. *[S.l.]*: Elsevier, 2020. p. 91–202.

VIQUE, Fabian; MARICHAL, Henry; STEINFELD, Leonardo. Inline mastitis detection system measuring the electrical conductivity of quarter milk. *In*: 2020 IEEE INTERNATIONAL CONFERENCE ON INDUSTRIAL TECHNOLOGY (ICIT). **2020 IEEE International Conference on Industrial Technology (ICIT)**. Buenos Aires, Argentina: IEEE, fev. 2020. Disponível em: <https://ieeexplore.ieee.org/document/9067172/>. Acesso em: 22 abr. 2021

YANG, Zhi-Xia; SHAO, Yuan-Hai; ZHANG, Xiang-Sun. Multiple birth support vector machine for multi-class classification. **Neural Computing and Applications**, v. 22, n. S1, p. 153–161, maio 2013.