UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Veterinária
Programa de Pós-Graduação em Ciência Animal

Mariana de Assis Lopes Frankó

Uso de *Machine Learning* para previsão de contagem padrão em placas de leite cru refrigerado antes de seu processamento tecnológico

Belo Horizonte
2022

Mariana de Assis Lopes Frankó

Uso de *Machine Learning* para previsão de contagem padrão em placas de leite cru refrigerado antes de seu processamento tecnológico

Dissertação apresentada à Universidade Federal de Minas Gerais, Escola de Veterinária, como requisito parcial para obtenção do grau de Mestre em Ciência Animal

Orientador**:** Prof. Dr. Marcos Xavier Silva

Coorientadores: Profa Dra. Mônica M. O. Pinho Cerqueira; Prof Dr. Frederico Gualberto Ferreira Coelho

Belo Horizonte
2022

Ficha catalográfica

Frankó, Mariana de Assis Lopes,1988 -
F834u Uso de Machine Learning para previsão de contagem padrão em placas de leite cru refrigerado antes de seu processamento tecnológico /Mariana de Assis Lopes Frankó. – 2022.
86f: il

Orientador: Marcos Xavier Silva
Coorientadores: Mônica M. O. Pinho Cerqueira
Frederico Gualberto Ferreira Coelho

Dissertação apresentada à Universidade Federal de Minas Gerais, Escola de Veterinária, como requisito parcial para obtenção do grau de Mestre em Ciência Animal.
Área de concentração: Epidemiologia
Inclui bibliografia
Anexos: f. 86

1. Leite - Qualidade - Teses - 2. Leite - Análise - Teses – I. Silva, Marcos Xavier – II. Cerqueira, Mônica M. O. Pinho - III. Coelho, Frederico Gualberto Ferreira - IV. Universidade Federal de Minas Gerais, Escola de Veterinária – V. Título.

CDD – 637

Bibliotecária responsável Cristiane Patrícia Gomes – CRB2569
Biblioteca da Escola de Veterinária, Universidade Federal de Minas Gerais

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE VETERINÁRIA
COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA ANIMAL

FOLHA DE APROVAÇÃO

MARIANA DE ASSIS LOPES FRANKÓ

Dissertação submetida à banca examinadora designada pelo Colegiado do Programa de Pós-Graduação em CIÊNCIA ANIMAL, como requisito para obtenção do grau de MESTRE em CIÊNCIA ANIMAL, área de concentração Epidemiologia.

Aprovado(a) em 30 de novembro de 2022, pela banca constituída pelos membros:

Dr.(a). Marcos Xavier Silva - Presidente - Orientador(a)

Dr.(a). Elisa Helena Paz Andrade

Dr.(a). Soraia de Araujo Diniz

Documento assinado eletronicamente por **Marcos Xavier Silva**, **Professor do Magistério Superior**, em 30/11/2022, às 19:39, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Elisa Helena Paz Andrade**, **Professora do Magistério Superior**, em 30/11/2022, às 20:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **SORAIA DE ARAUJO DINIZ**, **Usuário Externo**, em 03/12/2022, às 14:53, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1929615** e o código CRC **71E02210**.

Dedico este trabalho ao meu esposo Bernard e meus filhos por serem meu suporte nessa caminhada.

Agradecimentos

Agradeço primeiramente a Deus por tantas bênçãos recebidas e pela oportunidade de usufruir de uma vida em Cristo.

Agradeço ao meu esposo Bernard por ser meu grande companheiro nessa jornada e por me apoiar sempre, com tanto amor e paciência. Ao meu filho João por ser uma luz na minha vida.

Agradeço aos meus pais André e Lucia, minhas irmãs Sarah e Marcela por todo amor e incentivo.

Agradeço ao meu orientador Marcos por acreditar no meu potencial mesmo quando eu duvidei.

Professora Mônica, minha co-orientadora por acreditar no projeto e não medir esforços para que conseguíssemos parceiros para viabilizá-lo. Minha eterna gratidão.

Professor Frederico, obrigada pela oportunidade de trabalhar com você, pela generosidade e disponibilidade, não seria possível sem seu apoio.

À Isabela, minha colega e amiga agradeço por todo ensinamento e por ter me acompanhado durante essa trajetória.

Meu agradecimento ao laticínio que me recebeu, aos colegas (em especial Daniela e Felipe) que me ajudaram neste projeto, e também todo o meu agradecimento e respeito aos transportadores de leite.

Agradeço ao laboratório LITC da escola de engenharia da UFMG, ao professor Braga e aos colegas por me receberem de braços abertos sempre dispostos a ensinar.

Não poderia deixar de agradecer a minha rede de apoio materna que muitas vezes cuidaram do meu filho pra que eu pudesse estudar: Mariana, Luísa, Rosinha, mamãe e Dalva.

Resumo

O leite é uma das *commodities* mais produzidas e relevantes por ser um produto com alto valor nutricional e baixo custo para o consumidor se comparado a outras fontes de nutrientes. Por este motivo pode ser um importante meio de cultura e de transmissão de doenças se não for cuidadosamente manipulador, estocado, transportado e processado. Objetivou-se avaliar e comparar a performance de três modelos de ML para previsão da CPP do leite cru que chega às plataformas do laticínio. Comparou-se três modelos de ML, o *Support Vector Machine* (SVM), XGBoosting e redes neurais *MultiLayer Perceptron* (MLP). Obtivemos um resultado com RMSE de 5,4887410; 4,7333138; 6,0639758 respectivamente e um MAPE de 0,63%; 0,06%; 0,92%, demonstrando que a rede XGBoosting foi a que apresentou menor erro, porém os três modelos são eficientes para prever a CPP do leite e são estatisticamente semelhantes.

Palavra-chave: leite cru, contagem bacteriana, *machine learning*, previsão.

Abstract

Milk is one of the most produced and relevant commodities because it is a product with high nutritional value and low cost for the consumer compared to other sources of nutrients. For this reason, it can be an important means of culture and transmission of diseases if it is not carefully handled, stored, transported and processed. The objective was to evaluate and compare the performance of three ML models for predicting the CPP of raw milk arriving at the dairy platforms. Three ML models were compared, the Support Vector Machine (SVM), XGBoosting and MultiLayer Perceptron (MLP) neural networks. We got a result with RMSE of 5.4887410; 4.7333138; 6.0639758 respectively and a MAPE of 0.63%; 0.06%; 0.92%, demonstrating that the XGBoosting network had the lowest error, but the three models are efficient in predicting milk CPP and are statistically similar.

Keywords: raw milk, bacterial count, machine learning, prediction.

Lista de tabela

Tabela 1: Padrões de CPP e CCS do leite cru refrigerado de tanques individuais e coletivos

Tabela 2: Resultado das métricas de avaliação da performance do algoritmo MICE para predição de dados de CPP do produtor.

Tabela 3: Métrica de RMSE para os modelos.

Tabela 4: Métrica de MAPE para os modelos.

Lista de figura

Figura 1: Arquiteturas de RNAs.

Figura 2: Modelo esquemático do neurônio MCP.

Figura 3: Mapa da região noroeste de São Paulo.

Figura 4: Fluxograma da formação dos bancos de dados utilizados para análise de desempenho do algoritmo de imputação múltipla MICE para dados faltantes de CPP de produtor.

Figura 5: Gráfico representando valor real e valor imputado para 5% de falha no banco de dados.

Figura 6: Gráfico representando valor real e valor imputado para 10% de falha no banco de dados.

Figura 7: Processo de validação cruzada k-fold, com k=5.

Figura 8: Métricas para a avaliação dos modelos RMSE, MAPE e suas respectivas fórmulas.

Figura 9: Matriz de correlação de Pearson (r) para seleção de variáveis, a serem utilizadas na previsão da CPP final do leite entre janeiro de 2021 e junho de 2021.

Figura 10: *Recursive Feature Elimination* (RFE) para seleção de variáveis, para previsão da CPP final do leite.

Figura 11: Algoritmo Boruta aplicado aos dados para seleção de variáveis, para previsão da CPP final do leite.

Figura 12: Representação gráfica dos resultados das previsões de CPP do leite do compartimento, no modelo MLP para cada *fold* e respectivos histogramas de erros RMSE (%).

Figura 13: Representação gráfica dos resultados das previsões de CPP do leite do compartimento, no modelo SVM para cada *fold* e respectivos histogramas de erros RMSE (%).

Figura 14: Representação gráfica dos resultados das previsões de CPP do leite do compartimento, no modelo *XGboosting* para cada *fold* e respectivos histogramas de erros RMSE (%).

Figura 15: *Boxplot teste de Wilcoxon comparartivo de erros dos modelos MLP, SVM e XGboosting.*

Figura 16: Resultado do pacote *shapley* para a rede neural MLP.

Lista de Siglas

CPP- Contagem em Placas Padrão

CCS- Contagem de Células Somáticas

MAPA- Ministério da Agricultura Pecuária e Abastecimento

MAPE- *Mean Absolute Percentage Error*

ML- *Machine Learning*

MLP- *MultiLayer Perceptron*

MICE- *Multivariate Imputation by Chained Equations*

RFE- *Recursive Feature Elimination*

RNAs- Redes Neurais Artificiais

SNG- Sólidos não gordurosos

RIISPOA- Regulamento da Inspeção Industrial e Sanitária de Produtos de Origem Animal

RMSE- *Root Mean Squared Error*

SVM- *Support Vector Machine*

Sumário

Resumo **6**
Sumário 10
Introdução 12
Objetivos 12
Objetivo geral **12**
Objetivos específicos **12**
Capítulo 1 13
Parâmetros técnicos importantes no serviço de inspeção do leite e em sua qualidade 13
1.1 Conceito de leite **13**
1.2 Panorama do leite **13**
1.3 Aspectos físico-químicos e microbiológicos **15**
1.4 Legislação vigente **17**
1.5 Qualidade do leite **19**
1.6 Parâmetros para avaliação da qualidade do leite **22**
1.7 Contagem de células somáticas (CCS) **23**
1.8 Contagem padrão em placas (CPP) **24**
1.9 Transporte do leite cru a granel **26**
1.10 Coleta de leite a granel **28**
1.11 Referências Bibliográficas **29**
Capítulo 2 35
***Machine Learning*: conceitos importantes** 35
2.1 Aprendizado Supervisionado **35**
2.2 Qualidade dos dados **37**
2.3 Pré-processamento dos dados **37**
2.4 Seleção de variáveis **38**
2.4.1 Algoritmo Boruta 39
2.4.2 *Recursive Feature Elimination* (RFE) 39
2.5 Modelos de predição **40**
2.5.1 Redes Neurais Artificiais 40
As Redes Neurais Artificiais (RNA) surgiram no final da década de 80, e são uma forma de computação não-algorítmica que simula a estrutura de funcionamento do cérebro humano (Braga *et al*., 2007). 40
2.5.2 Principais Arquiteturas das RNA 41
2.8 Neurônio Artificial MCP 42
2.9 Redes Neurais *MultiLayer Perceptron* (MLP) 43
2.10 XGboosting 43

2.11 *Support vector machine* (SVM) 44
2.12 Validação cruzada 45
2.13 Referências Bibliográficas 46
Capítulo 3 48
Imputação múltipla de dados faltantes nas análises de contagem bacteriana do leite cru do produtor, utilizando o algoritmo MICE 48
3.1 Resumo 48
3.2 Introdução 48
3.3 Material e Métodos 50
3.3.1 Delineamento do estudo 50
3.3.2 Dados do Laticínio 51
3.3.3 Dados Meteorológicos 52
3.3.4 Limpeza dos dados 52
3.3.5 Aspectos éticos 54
3.4 Resultados e discussão 54
3.5 Conclusão: 57
3.6 Referências bibliográficas 58
Capítulo 4 60
Comparação de modelos de *machine learning* para a previsão da contagem bacteriana do leite cru antes do seu processamento na indústria 60
4.1 Resumo 60
4.2 Introdução 60
4.2.1 Modelos de ML 61
4.3 Material e Métodos 63
4.3.1 Delineamento do estudo 63
4.3.2 Dados Do Laticínio 63
4.3.3 Dados Meteorológicos 64
4.3.4 Limpeza dos dados 64
4.3.5 Etapa de treinamento e teste 66
4.3.6 Aspectos éticos 67
4.4 Resultados e Discussão 67
4.4.1 Seleção de variáveis 67
4.4.2 Previsão da CPP do leite do compartimento dos caminhões 70
4.4.3 Métricas de avaliação dos modelos 73
4.5 Conclusão 76
4.6 Referências Bibliográficas 77
Anexo I 81

Introdução

O Brasil é um dos maiores produtores de leite do mundo e Minas Gerais se destaca pela maior produção, correspondendo a quase 24,7% da produção nacional em 2021 (Anuário Leite, 2022).

O leite fluido ainda é uma das principais fontes de proteína animal, de altíssimo valor biológico para a maioria da população brasileira e mundial. Neste sentido, garantir a oferta de produtos seguros, com elevado valor nutricional, qualidade físico-química e microbiológica e com baixo custo para o consumidor final tem grande relevância social.

O desenvolvimento de tecnologias que possam melhorar o produto, em qualquer ponto da cadeia de produção do leite, pode influenciar beneficamente o consumidor final. Tecnologias de *Machine Learning*, estão sendo amplamente utilizadas para beneficiar a sociedade e também o agronegócio.

Este trabalho teve como objetivo, desenvolver uma tecnologia baseada em inteligência artificial e redes neurais para monitoramento da qualidade do leite fluido em suas rotas de transporte, para garantir a recepção de leite cru refrigerado e processamento de derivados lácteos com melhor qualidade.

Os resultados foram promissores e demonstram uma área de desenvolvimento tecnológico a ser ainda mais explorada, com a popularização das tecnologias chamadas “internet das coisas”.

Objetivos

Objetivo geral

- Prever a contagem padrão em placa do leite cru refrigerado recebido na plataforma de recepção de uma indústria de laticínios antes de seu processamento tecnológico.

Objetivos específicos

- Selecionar variáveis com maior importância, para previsão da contagem padrão em placas do leite cru refrigerado, relacionadas às condições de sua estocagem no tanque de expansão nas propriedades leiteiras, de seu transporte e temperatura ambiente.
- Modelar e treinar modelos de redes neurais, para prever a contagem padrão em placas do leite cru refrigerado que chega a indústria de laticínio, utilizando variáveis com maior importância neste parâmetro.

Capítulo 1

Parâmetros técnicos importantes no serviço de inspeção do leite e em sua qualidade

1.1 Conceito de leite

A produção de leite para o consumo humano data da pré-história. Inicialmente o leite era obtido da vaca da família e consumido em poucas horas após sua obtenção. Com o passar dos anos, o ser humano foi criando formas e técnicas para conservar mais aquela fonte de nutrientes, tais como a produção de queijo, fervura e resfriamento do leite (Food…, 1989).

A definição de leite segundo o Decreto nº 9013/2017 (Brasil, 2017) alterado pelo Decreto no 10468/2020 (Brasil, 2020a) é,

> "*[...]Entende-se por leite, sem outra especificação, o produto oriundo da ordenha completa e ininterrupta, em condições de higiene, de vacas sadias, bem alimentadas e descansadas. O leite de outros animais deve denominar-se segundo a espécie de que proceda[...]*".

O leite possui cerca de 87% de água e 13% de sólidos, sendo que em média 3,9% referem-se à gordura e o restante (9,1% em média), aos sólidos não gordurosos (SNG) (Bylund, 2003).

1.2 Panorama do leite

O leite de vaca é o mais consumido em todo o mundo, representando 83% do consumo total. Por constituir uma rica fonte de proteínas, minerais, carboidratos de boa qualidade e gorduras e ser um alimento de fácil acesso e relativo baixo custo, o leite representa a base da alimentação mundial (Verduci *et al*., 2019).

As vacas primitivas produziam em média 1000 litros de leite por lactação que era o suficiente para suprir o crescimento da cria. No entanto, depois que o homem domesticou a vaca, isso mudou por meio de seleção genética, pode-se chegar a 6000 litros de leite por lactação, o que é seis vezes mais que a vaca primitiva (Bylund, 2003).

Com tantos nutrientes, o leite é um importante meio de cultura e de transmissão de doenças se não for cuidadosamente obtido, estocado, transportado e processado. Cada uma destas etapas

pode apresentar riscos de contaminação e perda de sua qualidade (Philpot; Nickerson, 2002; Bylund, 2003).

A análise dos perigos e pontos críticos de controle na produção de leite devem ser observados desde a sua obtenção na fazenda (sanidade animal, manejo de ordenha, estocagem, resfriamento), até a coleta, transporte, estocagem na indústria, processamento, envase, transporte e estocagem do produto final (Philpot; Nickerson, 2002; Bylund, 2003).

Atualmente cerca de 150 milhões de famílias no mundo estão envolvidas na cadeia leiteira. Nos últimos 30 anos a produção de leite aumentou 59%, passando de 530 milhões de toneladas em 1988, para 843 milhões de toneladas em 2018. A Índia é o maior produtor de leite do mundo detendo 22% da produção, seguida por, Estados Unidos, China, Paquistão e Brasil (Food…, 2020).

O leite é uma das *commodities* agrícolas mais produzidas e relevantes, por ser um dos produtos com alto valor nutricional e baixo custo para o consumidor se comparado a outras fontes de nutrientes. Além do contexto nutricional, sua importância do ponto de vista econômico e social é indiscutível, sendo caracterizado, em sua maioria, por uma produção familiar (Marangoni *et al*., 2019; Food…, 2020; International…, 2019).

Em decorrência da pandemia, o ano de 2021 foi desafiador para o setor de lácteos. Houve redução da renda *per capita*, aumento da inflação reduzindo o poder de compra do consumidor e aumento do custo de produção, causando uma queda histórica no setor. O reflexo no setor foi uma retração de 3,7% (Associação…, 2022).

De acordo com a Associação Brasileira de Leite Longa Vida (ABLV), o Brasil produziu cerca de 33.690 milhões de quilos de leite em 2021. O volume de matéria-prima captado pela indústria diminuiu 1,8% em comparação ao ano de 2020. Isso ocorreu pelo desestímulo no campo, em decorrência dos altos custos de produção e condições climáticas desfavoráveis (Associação…, 2022).

São Paulo apresentou uma produção de 2.566 milhões de litros em 2021, representando 10,2% da produção do país. O valor médio pago ao produtor no primeiro semestre foi R$2,19 (Associação…, 2022).

De acordo com relatório da Organização das Nações Unidas para a Alimentação e a Agricultura (FAO) de 2021, um dos grandes entraves na produção leiteira nos países subdesenvolvidos é a dificuldade no transporte do leite cru a granel, devido à quantidade de pequenos produtores,

rotas em estradas ruins, baixa qualidade do leite cru e dificuldade de refrigerar o leite. Corroborando com a realidade que temos observado no Brasil, apesar da existência de regiões mais desenvolvidas quanto à tecnificação da produção, ainda há dificuldades com rotas extensas, de difícil acesso e com muitos produtores na mesma rota, sendo eles em sua maioria com baixa tecnificação (Food…, 2020).

Toda esta diversidade de cenários representa um desafio para os órgãos governamentais estabelecerem padrões de qualidade para todo o Brasil, levando em consideração a sua extensão territorial e a diversidade de climas. É um desafio! Muitas vezes, fica a cargo dos laticínios, trabalhar políticas de bonificação por qualidade, para obter um bom rendimento, encontrar os parâmetros que melhor se enquadram aquela realidade e padrões exigidos pelo ministério da agricultura em normas específicas.

1.3 Aspectos físico-químicos e microbiológicos

O leite para ter qualidade e ser inócuo para a população, deve ser obtido de forma higiênica na propriedade e a partir de animais saudáveis. Dessa forma, é necessário atentar para os aspectos de sanidade do rebanho atestado por médico veterinário; imediata refrigeração do leite após ordenha a uma temperatura inferior a 4°C para retardar o crescimento bacteriano e transporte no máximo 48 após a ordenha (Brasil, 2020a; Cruz *et al.*, 2019).

O tempo de estocagem e o transporte do leite interferem na qualidade do produto final devido a rotas de difícil acesso, pulverização da produção e altas temperaturas ambientais (Brasil, 2020a; Cruz *et al.*, 2019; Ferrari, 2018).

O leite é uma mistura complexa composta de água, gorduras, proteínas (caseína e albumina), carboidratos, vitaminas e minerais. A proporção de cada componente pode variar em função da alimentação, espécie, raça, idade, individualidade do animal e o estado fisiológico do animal pode alterar a sua composição (Ordonez, *et al.*, 2005).

Do ponto de vista físico-químico o leite é uma mistura homogênea com grande número de substâncias (lactose, glicerídeos, sais, proteínas, vitaminas, enzimas e etc.), das quais algumas estão em emulsão (gordura e as substâncias associadas), algumas em suspensão (caseínas ligadas aos sais minerais) e outras em dissolução verdadeira (lactose, vitaminas hidrossolúveis, proteínas do soro, sais e etc) (Ordonez, *et al*., 2005).

A composição do leite das vacas varia muito, mas as médias percentuais dos componentes são 3,7% de gordura, 4,8% de carboidrato, 4,5% lactose, 3,5% proteínas totais, 0,7% proteínas do

soro do leite, 3,7% gordura e 0,7% de minerais. A água é o componente que entra em maior proporção na composição do leite e influencia sensivelmente na densidade do mesmo. De modo geral, o leite contém 87% de água e 13% de sólidos, podendo variar (Bylund, 2003, Silva *et al*., 2019).

As proteínas do soro do leite estão em solução coloidal e a caseína em suspensão coloidal. As características típicas de um colóide são: tamanho de partícula pequeno, carga elétrica e afinidade das partículas pelas moléculas de água. Alguns fatores podem alterar a afinidade das partículas por estas moléculas de água e uma delas é o pH da solução. O pH pode ser alterado por exemplo pela temperatura e carga bacteriana do leite. Desta forma fatores externos podem afetar na qualidade final do leite e seus derivados (Bylund, 2003).

A gordura do leite consiste principalmente em triglicerídeos, di- e monoglicerídeos, ácidos graxos, esteróis, carotenoides que são responsáveis pela coloração amarela da gordura, vitaminas A, D, E e K e outros oligoelementos que são componentes secundários (Bylund, 2003).

A membrana do glóbulo de gordura consiste em fosfolípides, lipoproteínas, cerebrosídeos, proteínas, ácidos nucleicos, enzimas, oligoelementos. A água ligada a composição e a espessura da membrana não são constantes porque os componentes estão sendo trocados constantemente com o soro de leite circundante. Como os glóbulos de gordura são as maiores partículas do leite e as mais leves, elas tendem a ficar em suspensão quando o leite é deixado em repouso (Bylund, 2003).

Ao ser secretado da glândula mamária a uma temperatura média de 35° Celsius, o leite em condições saudáveis apresenta uma contagem bacteriana baixa, constituída de bactérias Gram-positivas e em sua maioria, de bactérias lácticas. Contudo, as contaminações podem ocorrer durante ou após a ordenha devido a contato com fezes de animais, contaminações no ambiente, equipamentos mal higienizados e falta de higiene operacional e também em decorrência de infecções da glândula mamária (Brito *et al*., 2004; Cruz *et al*., 2019; Gonçalves; Vieira, 2002; Philpot; Nickerson, 2002).

Quando o leite é obtido em condições higiênico sanitárias ruins por falhas nas boas práticas agropecuárias e manejo, haverá maior contaminação bacteriana do leite. Depois de ordenhado ele será normalmente mantido sob refrigeração até a coleta pelo laticínio e um leite com qualidade higiênico-sanitária insatisfatória, mantido sob refrigeração, poderá ter maior

contagem de bactérias psicrotróficas com predominância de *Pseudomonas e Acinetobacter* (Cruz *et al.*, 2019; Lopes *et al*, 2022; Brito *et al*., 2004; Gonçalves; Vieira, 2002).

1.4 Legislação vigente

Em 1860, foi aprovado o primeiro Decreto no país por Dom Pedro II, que criava a Secretaria do Estado dos Negócios da Agricultura. Somente em 29 de dezembro de 1906, o então presidente Afonso Penna promulgou a Lei n° 1.606 que transformou a Secretaria em Ministério dos Negócios da Agricultura e com base nesta lei, criou a Diretoria da Indústria Animal.

No ano de 1910 foi regulamentado o Serviço Veterinário e em 27 de janeiro de 1915, o Serviço de Indústria Pastoril pelo Decreto n° 11.460 que reorganiza os Serviços de Veterinária e definia o conceito de polícia sanitária animal. No mesmo ano, foi aprovado o primeiro Regulamento de Inspeção de Fábrica e Produtos Animais, que era bastante simples e continha apenas 23 artigos (Costa *et al*., 2015).

Já na fase de industrialização do Brasil, que foi alcançado graças às melhorias obtidas nos processos produtivos e nas legislações, o mercado do país foi aberto para exportações. Em 18 de dezembro de 1950 foi promulgada a Lei n°1.283, chamada de “Lei mãe”, que estabeleceu a obrigatoriedade da inspeção industrial de produtos de origem animal e a fiscalização de acordo com o âmbito do comércio do estabelecimento.

Em 1952 foi aprovado o Decreto n° 30.691 de 29 de março referente ao novo Regulamento da Inspeção Industrial e Sanitária de Produtos de Origem Animal (RIISPOA) que continha 952 artigos e foi um marco na inspeção no Brasil. Este regulamento passou por algumas alterações, mas ficou vigente até 2017, quando houve a modernização do RIISPOA, em resposta a “Operação carne fraca”, por meio do Decreto n° 9.013 de 29 de março de 2017 (Brasil, 1997; Costa *et al*., 2015; Brasil, 2017).

No que se refere à indústria de laticínios, o novo Regulamento da Inspeção Industrial e Sanitária de Produtos de Origem Animal (RIISPOA) trouxe modificações. Dentre elas, conforme descrito no artigo 53, destacou-se a necessidade de os estabelecimentos assegurarem que todas as etapas do processo produtivo sejam realizadas de forma higiênica, atendendo aos padrões de qualidade, não representando riscos à saúde do consumidor (Brasil, 2017).

O artigo 75 do RIISPOA, discorre a respeito da necessidade de os estabelecimentos disporem de mecanismos para garantir a rastreabilidade das matérias-primas e produtos com disponibilidade de informação em toda a cadeia produtiva. Isso trouxe uma nova visão ao que

se refere à responsabilidade da indústria sobre o controle da qualidade dos processos produtivos, na propriedade rural e no transporte do leite cru, e em acordo com isso, as Instruções Normativas (IN 76 e77) (Brasil, 2017; Brasil, 2018a, Brasil, 2018b).

A IN 77 estabelece critérios para produção, acondicionamento, conservação, transporte, seleção e recepção do leite cru. Nesta IN, destacam-se os artigos 6° e 7° que definem que o estabelecimento deve incluir junto aos seus Programas de Autocontrole (PAC), o Programa de Qualificação de Fornecedores de Leite (PQFL) (Brasil, 2018b). Estabeleceu-se que assistência técnica, gerencial e capacitação de fornecedores, com foco na gestão da propriedade rural e Boas Práticas Agropecuárias (BPA) é de responsabilidade das indústrias de laticínios (Brasil, 2018b).

A implantação de BPA é de extrema importância para a obtenção de matéria-prima de boa qualidade, em condições higiênicas e estas práticas devem ser incluídas nos Programas de Autocontrole (PAC) elaborados pela indústria. Nestes programas devem ser mantidos registros auditáveis que evidenciem a execução, atingimento de metas pelos fornecedores por um período mínimo de 12 meses (Vallin *et al*., 2009; Brasil, 2018b).

A IN 76 fixa os padrões de identidade e qualidade do leite cru refrigerado, leite pasteurizado, e leite pasteurizado tipo A, estabelecendo outros limites legais. Dispõe que o leite cru refrigerado de tanque individual ou de uso comunitário deve apresentar médias geométricas trimestrais de Contagem Padrão em Placas de no máximo 300.000 UFC/mL (trezentas mil unidades formadoras de colônia por mililitro) e de Contagem de Células Somáticas de no máximo 500.000 CS/mL (quinhentas mil células por mililitro). O leite cru refrigerado deve apresentar limite máximo para Contagem Padrão em Placas de até 900.000 UFC/mL antes do seu processamento no estabelecimento beneficiador (Brasil, 2018b).

As médias geométricas devem considerar as análises realizadas no período de três meses consecutivos e ininterruptos com no mínimo uma amostra mensal de cada tanque (Brasil, 2018b).

Os limites de temperatura de recebimento do leite devem ser de no máximo 7°C, excepcionalmente de 9º C (Brasil, 2018b).

A IN 76 foi alterada pela IN 58 (Brasil, 2019a). As alterações incluíram: no caso de ausência de resultado mensal para a composição das médias geométricas trimestrais o resultado deve ser substituído pela média geométrica trimestral calculada, até o restabelecimento, e o leite cru

refrigerado deve apresentar limite máximo para Contagem Padrão em Placas (CPP) de até 900.000 UFC/mL (novecentas mil unidades formadoras de colônia por mililitro) antes do seu processamento na indústria (Brasil, 2019a).

De acordo com IN 77 a coleta de leite deve ser suspensa de produtores que apresentarem por três meses consecutivos resultados de médias geométricas de CPP fora do padrão, sendo que, de acordo com IN 59, para restabelecimento da coleta, é necessário identificar causa do desvio e adotar medidas e ações corretivas e na apresentação de um resultado de CPP abaixo de 300.000 UFC/mL a coleta pode ser restabelecida (Brasil, 2018b; Brasil, 2019a, Brasil, 2019b).

A partir das alterações realizadas na legislação nos últimos anos, a indústria é responsável por garantir controle e melhorias em todo o processo produtivo desde a sua obtenção na propriedade rural, transporte até o seu processamento na indústria, baseando-se no Plano de Qualificação de Fornecedores de Leite e nas alterações no novo RIISPOA. Uma grande alteração foi feita no Decreto n° 9.013 por meio da publicação do Decreto n° 10.468 de 18 de agosto de 2020 com intuito de ter uma racionalização, simplificação e a virtualização de processos e procedimentos, afim de obter qualidade no processo e rastreabilidade (Brasil, 2017, Brasil, 2018a, Brasil, 2018b, Brasil, 2020a).

O Programa de Autocontrole (PAC) que já estavam contemplados no Decreto nº 9013 (Brasil, 2017), deve ser desenvolvido, implantado, mantido, monitorado, verificado pelo próprio estabelecimento, com registros sistematizados e auditáveis que comprovem o atendimento dos requisitos higiênico-sanitário e tecnológico. O Plano de Qualificação de Fornecedores de Leite (PQFL) deve ser incluído ao PAC, conforme norma complementar, ou seja, segundo a IN 77 de 2018 (Brasil, 2017; Brasil, 2018a; Brasil, 2018b).

O leite é um alimento importante para a base alimentar humana por ser fonte de nutrientes e acessível à maior parte da população (Confederação…, 2020). Assim, ações para obtenção, transporte e processamento higiênico devem ser implementadas para garantir a obtenção de um alimento inócuo e seguro ao consumidor e a eliminação de riscos de veiculação de patógenos pelos alimentos (Brasil, 2017).

1.5 Qualidade do leite

A demanda do mercado no Brasil por leite de melhor qualidade tem aumentado e tem sido impulsionada por mudanças recorrentes nos padrões de qualidade exigidos pelos órgãos governamentais competentes. Isso tem acontecido com intuito de equiparar nosso produto com

o leite de países considerados referência como, por exemplo, os Estados Unidos e os que compõem a União Europeia (Dias; Ant

es, 2014).

Em 2002, foi aprovada a IN 51 de 2002 (Brasil, 2002) que padronizava os critérios de qualidade e identidade do leite. Posteriormente, em 2011, foram publicadas as IN 32 e 62 e em 2018, foram publicadas as IN 76 e 77 que além de estabelecer novos padrões de qualidade para o leite cru, estabeleceram também orientações sobre o PQFL, determinaram a indústria como responsável por desenvolver, implantar, manter e monitorar o Plano (Brasil, 2011a; Brasil, 2011b; Brasil, 2018a, Brasil, 2018b).

O rigor com a segurança alimentar tende a aumentar, a cada dia, devido aos problemas socioeconômicos que diversos países em todo o mundo têm enfrentado por doenças transmitidas por alimentos (DTA). De acordo com a Organização Mundial de Saúde (OMS), cerca de 600 milhões de pessoas adoecem por ano após ingerirem alimentos contaminados e 420 mil pessoas vêm a óbito. Dentre os patógenos mais comuns estão *Salmonella, Campylobacter, Escherichia coli enterohemorrágica*, *Listeria monocytogenes* e *Vibrio cholerae.* Além desses contaminantes também podem ser incluídos vírus, parasitas, "príons" e produtos químicos (Marder *et al*., 2018; Who, 2015).

Com o advento do novo coronavírus, as fábricas e indústrias tiveram que introduzir medidas de redução dos riscos de transmissão do COVID-19. Essas medidas além de diminuírem a transmissão do vírus, também serão muito observadas nas movimentações comerciais entre países (WHO, 2020).

No Brasil, em relação às doenças veiculadas pelo leite destacam-se a intoxicação por enterotoxinas estafilocócicas e por *Bacillus cereus* e a infecção alimentar causada por *Campylobacter jejuni*, *Listeria monocytogenes*, *Brucella abortus* e *Mycobacterium bovis* (Vasconcellos; Ito, 2011; Melo, *et al*., 2018).

Os problemas com a qualidade do leite no Brasil têm sido relacionados principalmente a falhas no processo de obtenção nas propriedades rurais e no transporte deste produto. Entre os principais problemas relatados estão as altas contagens bacterianas e de células somáticas, mastite clínica e subclínica e presença de substâncias inibidoras (Brito *et al*., 2004; Vallin *et al*., 2009; Sequetto *et al*., 2017).

Estes problemas estão relacionados a falhas no manejo, não implementação de boas práticas agropecuárias, carência de treinamento de mão-de-obra, falta de incentivo por parte da indústria e compreensão por parte do produtor da responsabilidade relacionada à produção de leite seguro para o consumo da população (Marcondes *et al*., 2017; Marioto *et al*., 2020).

Um dos grandes desafios referem-se à falta de equidade na produção quando se comparam produtores de grande volume de leite com os que produzem baixo volume de leite, ou seja, igualdade de treinamento e tecnificação. O produtor de pequeno volume de leite, em suma, é representado por produção familiar é responsável por uma menor parte da produção global, agrega o maior contingente de pessoas em ocupação, demonstrando a importância socioeconômica da atividade (Bressan; Martins, 2004; Marcondes *et al*., 2017).

Buscar meios para auxiliar produtores de pequeno volume de leite a se especializarem e tecnificarem, é importante não só do ponto de vista de saúde pública, mas também do ponto de vista econômico e social (Marcondes et *al*., 2017).

Problemas relacionados à falta de qualidade da matéria-prima também são relatados por Vallin *et al*. (2009) que observaram que a adoção de Boas Práticas Agropecuárias (BPA) reduziu a CPP do leite em 87,9% em propriedades com ordenha manual e em 86,99% com ordenha mecânica. Isto demonstra que a melhoria na qualidade microbiológica do leite pode ocorrer também em propriedades com baixa tecnificação.

Paixão *et al*. (2014) também demonstraram melhoria da qualidade do leite com a adoção de BPA. Quando praticada em seu nível máximo, levou a um rápido retorno do capital investido.

A implantação de BPA e assistência técnica ao produtor de leite vão proporcionar equidade nos processos produtivos de produtores de grande e pequeno volume de leite levando à obtenção de leite de melhor qualidade microbiológica e mais seguro para o consumidor (Bressan; Martins, 2004; Nero *et al*., 2005).

Em concordância com as IN 76 e 77, as indústrias devem realizar um diagnóstico de situação dos fornecedores de leite, por meio de um levantamento detalhado e cadastramento desses fornecedores: avaliação sanitária do rebanho, categorização dos fornecedores, realização de ações corretivas por categorias de produtores, monitoramentos e registros, verificação das ações implementadas, além de auditoria da continuidade das ações propostas para o produtor e registros auditáveis (Brasil, 2018a; Brasil, 2018b).

A partir do diagnóstico, pode-se caracterizar o perfil dos produtores de leite do laticínio e categorizá-los para tomada de decisão com relação às ações individuais e coletivas, de acordo com o tipo de não conformidade observada em cada item. Este diagnóstico é importante pois o tipo de ação pode ser diferente para cada região e as necessidades podem variar bastante.

Atualmente o Brasil encontra-se em um processo de adequação bastante rigoroso, em que as exigências pela captação de leite de melhor qualidade são cada vez maiores.

1.6 Parâmetros para avaliação da qualidade do leite

As normas que vigoram no Brasil para a qualidade do leite foram definidas com base na Instrução normativa 51 (Brasil, 2002) e Instruções normativas 76 e 77 (Brasil, 2018a; Brasil, 2018b). Desde então o setor vem passando por um processo de adequação e reorganização, tendo em vista atender os novos padrões exigidos pelo Ministério da Agricultura.

Identificar os fatores intrínsecos e extrínsecos relacionados à qualidade do leite é necessário para orientar a assistência técnica, a indústria, instituições de pesquisa e órgãos governamentais, para definição de estratégias relacionadas a melhoria da qualidade do leite (Dias *et al*., 2021).

De acordo com as recomendações do Ministério da Agricultura Pecuária e Abastecimento (MAPA), nos tanques de refrigeração de uso individual e coletivo devem ser realizados no mínimo uma vez no mês, coleta de uma amostra para realização das seguintes análises: teor de gordura, proteína total, lactose anidra, sólidos não gordurosos, sólidos totais, contagem de células somáticas (CCS), contagem padrão em placas (CPP), resíduos de produtos veterinários, e outros que venham a ser exigidos em normas complementares, pelo laboratório da Rede Brasileira de Laboratórios de Controle da Qualidade do Leite (RBQL) (Brasil, 2018b).

Além das análises mensais de CCS e CPP do leite cru refrigerado coletado dos tanques de expansões, na plataforma recepção das indústrias de laticínios são realizadas mais de vinte análises no leite dos caminhões de coleta a granel, com o intuito de detectar fraudes, resíduos de antibióticos e garantir o processamento de um alimento inócuo ao consumidor. Todas estas análises são realizadas para garantir que o leite atenda os requisitos físico-químicos e microbiológicos estabelecidos na legislação brasileira (Brasil, 2018b).

Na tabela 1 podemos relembrar os valores de referência para CPP e CCS para tanque individual e coletivo, e CPP dos caminhões após o transporte e silo (Tab. 1).

Tabela 1: Padrões de CPP e CCS do leite cru refrigerado de tanques individuais e coletivos

Parâmetro	Valor máximo
CPP	300.000 UFC/mL
CCS	500.000 CS/mL
Leite cru refrigerado antes do processamento	
CPP	900.000 UFC/mL

Fonte: Adaptado de (Brasil, 2018b)

1.7 Contagem de células somáticas (CCS)

A contagem de células somáticas (CCS) é indicativa do número de células leucocitárias do sangue e de células epiteliais presentes no leite. É usada para monitorar a inflamação no úbere, como parâmetro de infecção ou inflamação intramamária (Philpot; Nickerson, 2002).

Com a infecção, há aumento de leucócitos na corrente circulatória devido à resposta inflamatória do tecido mamário que leva a um aumento da CCS. A maior parte das células somáticas presentes na CCS corresponde aos leucócitos, principalmente neutrófilos (Philpot; Nickerson, 2002).

Diversos fatores podem afetar a contagem de células somáticas e entre eles, destacam-se: ocorrência de mastite, tipo de microrganismo envolvido, idade do animal, estágio de lactação, variações diurnas e sazonais, estresse e frequência de ordenha. Desses o principal é a mastite, que deve ser monitorada e controlada com orientação do médico veterinário, por meio de implantação de boas práticas agropecuárias, limpeza dos equipamentos, higiene do ordenhador, práticas de pré e pós-*dipping* e manejo ambiental (Arcuri, 2006).

Leite com altas contagens de células somáticas trazem grandes danos à indústria, como a coagulação e floculação que ocorrem no processamento térmico do leite pasteurizado e do leite em pó. Devido à alteração na qualidade do leite de vacas com mastitie, mudanças significativas podem ocorrer nos produtos lácteos como: alteração na viscosidade e sabor do iogurte, geleificação e coagulação das proteínas do leite UHT durante a estocagem, alterações na fabricação de queijos, redução no rendimento industrial, aumento no conteúdo de água no coágulo, alterações negativas nas propriedades sensoriais, aumento do tempo para formação do

coágulo; baixa taxa de firmeza do coágulo, defeitos de textura e elevada perda de sólidos no soro do queijo (Ordonez *et al*, 2005; Philpot; Nickerson, 2002).

1.8 Contagem padrão em placas (CPP)

A CPP é um importante parâmetro de qualidade do leite que indica deficiência na adoção de práticas higiênico-sanitárias durante a ordenha, refrigeração ou transporte (Cruz *et al*., 2019; Philpot; Nickerson, 2002).

A contagem bacteriana em animais sadios ordenhados higienicamente é inferior a 100 UFC/mL, mas ela pode aumentar em decorrência de manejo inadequado de ordenha, falta de limpeza e higiene do equipamento de ordenha, refrigeração inadequada do leite (Cruz *et al*., 2019; Philpot; Nickerson, 2002).

A IN 77/20218 estabelece que a interrupção da coleta do leite de produtores que apresentarem médias geométricas em três meses consecutivos acima de 300.000 UFC/ mL, para restabelecimento da coleta deve ser identificada a causa do desvio, adotadas as ações corretivas e apresentado um resultado de análise de CPP dentro do padrão, emitido por laboratório da RBQL, apresentação do resultado de análise de CPP dentro do padrão no mesmo mês referente à terceira média geométrica fora do padrão, a interrupção de que trata o caput não se aplicará (Brasil, 2018b, Brasil, 2019).

Tem sido um desafio para o produtor brasileiro atender os novos parâmetros de qualidade, porém já se sabe que modificações de manejo simples podem apresentar resultados significativos na redução da CPP e consequentemente, redução da inflamação por microrganismos oportunistas, o que pode refletir também em redução da CCS (Lopes *et al*., 2022).

Lopes *et al*. (2022) analisaram 11 artigos sobre implementação de BPA e observaram que a adoção de práticas simples de manejo como descarte dos três primeiros jatos, pré- *dipping* e pós- *dipping*, secagem dos tetos com toalha descartável, higienização correta dos utensílios de ordenha e eliminação da água residual, são práticas eficientes para redução da CPP. Fato que já foi relatado por diversos autores a quase duas décadas como Brito *et al*. (2004), Philpot e Nickerson (2002), Gonçalves; Vieira (2002).

O leite possui alto valor nutricional, sendo importante meio para multiplicação de microrganismos que podem ser classificados neste caso em dois grupos: deteriorantes e

patogênicos. Os contaminantes associados ao leite incluem bactérias, vírus, fungos filamentosos e leveduras (Cruz *et al*., 2019; Philpot; Nickerson, 2002).

Os grupos de microrganismos deteriorantes são classificados em sacarolíticos, proteolíticos e lipolíticos, que impactam na cadeia produtiva do leite, causando perdas na qualidade nutricional do produto, interferindo na lucratividade do laticínio por diminuir o rendimento dos derivados (Araújo *et al*., 2019; Cruz *et al.*, 2019).

O armazenamento nos tanques individuais sob refrigeração e a coleta de leite a granel representaram um grande avanço para a indústria no Brasil, trazendo melhoria da qualidade do leite recebido e reduzindo o problema da proliferação de microrganismos mesófilos, que aumentam a acidez do produto, prejudicando o beneficiamento. Entretanto, com a permanência do leite em temperaturas de refrigeração por períodos prolongados, observa-se a substituição da microbiota deteriorante mesófila, por uma microbiota de bactérias psicrotróficas que são capazes de se multiplicar e produzir enzimas deteriorantes do leite, mesmo em temperatura de refrigeração (Araújo *et al*., 2019; Cruz *et al*., 2019).

Reserva-se o termo psicrotróficos para bactérias que apresentam temperaturas ótimas de crescimento entre 20 e 40 ºC, mas que podem crescer em temperaturas abaixo de 7 ºC (International…, 1976). Este gênero é composto por bactérias Gram-positivas e Gram-negativas, bacilos, cocos, vibrios, formadores ou não de esporos, assim como microrganismos aeróbios e anaeróbios. Alguns gêneros de bolores e leveduras também apresentam características do grupo dos psicrotróficos e podem causar problemas de qualidade do leite (Santos; Fonseca, 2001).

O gênero *Pseudomonas* é considerado o mais importante entre as bactérias psicrotróficas, podendo ser encontrado em aproximadamente 10% da microbiota do leite recém-ordenhado, sendo que sob condições de refrigeração este gênero rapidamente predomina sobre a microbiota, tanto do leite cru como do leite pasteurizado (Muir, 1996).

Os microrganismos psicrotróficos podem produzir enzimas termorresistentes (lipases e proteases) que provocam alterações no leite e principalmente nos derivados (Santana; Beloti; Barros, 2001). Algumas bactérias psicrotróficas carreiam o gene *apr*, que é responsável pela informação de produção de metaloprotease alcalina, enzima termorresistente, capaz de degradar proteínas do leite. Após o processo térmico na indústria a enzima continuará deteriorando o leite, reduzindo o tempo de prateleira deste produto (Araujo *et al*., 2019).

Dentre os componentes do leite, as proteínas são as de maior valor para a industrialização. A lucratividade das indústrias depende do rendimento representado pelo extrato seco total e eficiência da transformação do leite em derivados e leite fluido, que por sua vez dependem da qualidade da matéria-prima, baseada nos parâmetros de sanidade do rebanho (Brasil *et al*., 2015).

As caseínas representam cerca de 80% das proteínas do leite e consistem em quatro proteínas principais: αs1-, αs2-, β- e κ-caseína. A hidrólise enzimática da k-caseína que ocorre por meio da temperatura, pH, excesso de Ca^{2+} e adição de etanol afetam a estabilidade dessas proteínas, que estão em grande parte presentes no leite, na forma de partículas coloidais, conhecidas como micelas (Brasil *et al*., 2015).

Os microrganismos patogênicos comprometem a inocuidade e qualidade do produto e estão relacionados a agentes etiológicos que podem ser transmitidos para os seres humanos pelo consumo do leite e de seus derivados. Outro fator é a ocorrência de surtos de intoxicação ou infecção de origem alimentar, daí a importância de se produzir alimentos seguros desde a sua obtenção na propriedade rural, processamento na indústria até o consumo (Cruz *et al*., 2019).

1.9 Transporte do leite cru a granel

Nas últimas décadas, o transporte do leite no Brasil passou por diversas mudanças. Até a década de 90 o leite era armazenado em latões e entregue na indústria de beneficiamento e para isso era necessário que o transporte fosse realizado no início da manhã e que fosse em pequenas distâncias. Essa realidade limitava a expansão do setor devido a baixa qualidade do leite produzido e dificuldade de escalar a produção com este sistema de transporte (Paixão, 2011).

Neste contexto a IN n° 51 foi um marco, pois regulamentou a coleta de leite cru refrigerado e seu transporte a granel:

> "[...]*O processo de coleta de Leite Cru Refrigerado a Granel consiste em recolher o produto em caminhões com tanques isotérmicos construídos internamente de aço inoxidável, através de mangote flexível e bomba sanitária, acionada pela energia elétrica da propriedade rural, pelo sistema de transmissão ou caixa de câmbio do próprio caminhão, diretamente do tanque de refrigeração por expansão direta ou dos latões contidos nos tanques de expansões de imersão* (Brasil, 2002). [...]"

Essa legislação foi posteriormente atualizada por meio de vários documentos normativos e entre eles, pela IN n° 77 que estabelece a coleta nos tanques de expansão individuais e coletivos, nos quais os latões são entregues a um titular treinado, responsável pelo tanque. Desta forma o caminhão não coleta leite de latões individualmente e sim dos tanques de expansão devidamente cadastrados conforme estabelecido pelo MAPA.

Como o leite é transportado em condições isotérmicas, a produção higiênica do leite, a sua manutenção na fazenda e seu transporte em temperatura próxima de 4º C irão reduzir os riscos de aumento da carga bacteriana, tornando menos suscetível a alterações durante principalmente o transporte (Teixeira; Ribeiro, 2000; Brasil, 2018b).

Quando o leite passa pelo equipamento de ordenha ou pelos utensílios utilizados, ele pode ser contaminado. Se a contaminação inicial do leite na fazenda for baixa, o leite consegue manter a contagem bacteriana baixa até a chegada na indústria, mesmo que o tempo de percurso seja longo (Ferrari, 2018).

O leite cru deve chegar ao laticínio a uma temperatura máxima de 7°C podendo excepcionalmente chegar a 9°C, devendo o laticínio se responsabilizar pela organização da logística para que não haja excepcionalidade (Brasil, 2018b).

Pelos desafios que o transporte representa na manutenção da qualidade do leite, o estudo e gerenciamento de logística industrial se torna cada dia mais importante, para oferecer um produto com alta qualidade e melhor custo-benefício para o consumidor. Reduzir perdas nesta etapa tem sido uma meta das empresas do setor, levando em consideração que o leite pode perder qualidade no transporte (Dutra *et al.*, 2014; Ferrari, 2018).

As condições de higiene e limpeza dos caminhões de transporte a granel estão diretamente relacionadas ao aumento da contagem bacteriana do leite que chega à indústria. Ferrari (2018) observou que a contagem padrão em placas alterou no decorrer do percurso, mostrando-se mais elevada na chegada à indústria.

Outro problema do setor está relacionado à oscilação da energia elétrica nas propriedades rurais, fazendo com que o transportador capte o leite fora da temperatura adequada. Além disto, falhas no monitoramento da temperatura do leite durante o transporte, dificuldade em manter uma carga bacteriana inicial baixa, oscilações de temperaturas regionais, rotas longas e más condições das estradas, comprometem a qualidade do produto, gerando aumento de custos e potenciais impactos ambientais, quando há necessidade de descarte de carga.

De acordo com Paixão *et al.* (2011), o reparo de estradas e a implantação de programas de educação continuada aos produtores de leite, aos responsáveis pelos tanques comunitários e motoristas dos caminhões tanques, são fundamentais para manutenção da qualidade do leite, eficiência do processamento e diminuição de custos na coleta e transporte do leite.

1.10 Coleta de leite a granel

O leite cru é coletado por caminhões isotérmicos, que mantêm a temperatura do leite coletado, por isso é necessário que ele esteja a uma temperatura inferior a 4°C, para que mesmo com o aumento da temperatura durante o percurso, o produto chegue a uma temperatura inferior a 7°C no laticínio (Brasil, 2018b; Bylund, 2003).

O caminhão tanque deve ter acesso até o local de refrigeração e armazenagem do leite na fazenda. A mangueira de carregamento do caminhão tanque é conectada à válvula de saída do tanque de resfriamento da fazenda, através de mangueira e bomba sanitária em circuito fechado (Brasil, 2018b).

Alguns caminhões tanque são equipados com um medidor de vazão e uma bomba para que o volume seja registrado automaticamente, mas geralmente o volume é medido utilizando-se uma régua medidora do próprio tanque de resfriamento da fazenda (Brasil, 2018b; Bylund, 2003).

O bombeamento é interrompido assim que o tanque de resfriamento é esvaziado. Isso evita que o ar seja misturado ao leite. O veículo de coleta é dividido em vários compartimentos para evitar que o leite se espalhe durante o transporte, e geralmente as análises realizadas na plataforma de recebimento do leite são feitas em cada compartimento. Cada compartimento é preenchido por sua vez e, quando o caminhão tanque completa sua rota programada, ele entrega o leite ao laticínio (Brasil, 2018b; Bylund, 2003).

1.11 Referências Bibliográficas

ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE LEITE LONGA VIDA. Relatório Anual 2021(ABVL). São Paulo, 2022.

ARAÚJO, Lorena Gonçalves *et al*. Identificação de atividade deteriorante e do gene *apr* na microbiota isolada de leite cru em Caxias, MA. *Revista do Instituto de Laticínios Cândido Tostes*, v. 74, n. 4, p. 219-230, 2019.

ARCURI, Ellen Fernandes et al. Qualidade microbiológica do leite refrigerado nas fazendas. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, v. 58, p. 440-446, 2006.

BRASIL, R. B.; NICOLAU, E. S.; CABRAL, J. F.; SILVA, M. A. P. da. Estrutura e estabilidade das micelas de caseína do leite bovino. *Ciência Animal*, Fortaleza, v. 25, n. 2, p. 71-80, jun./set. 2015.

BRASIL. Ministério Da Agricultura Pecuária e Abastecimento. Regulamento de inspeção industrial e sanitária de produtos de origem animal. Aprovado pelo Decreto 30. 691, de 29/03/52, alterado pelos decretos 1. 255, de 25/06/1962; 1236, de 02/09/1994; 1812, de 08/02/1996, e 2244, de 04/06/1997.174 p. Brasília: Ministério da Agricultura, 1997.

BRASIL. Ministério da Agricultura Pecuária e Abastecimento. Instrução Normativa nº 51, de 18 de setembro de 2002 Regulamentos Técnicos de Produção, Identidade e Qualidade do Leite tipo A, do Leite tipo B, do Leite tipo C, do Leite Pasteurizado e do Leite Cru Refrigerado e o Regulamento Técnico da Coleta de Leite Cru Refrigerado e seu Transporte a Granel. Diário Oficial da União, Brasília. 2002.

BRASIL. Ministério da Agricultura Pecuária e Abastecimento. Instrução Normativa nº 32, de 30 de junho de 2011. Prorroga a vigência dos prazos estabelecidos para adoção de novos limites microbiológicos e de células somáticas. Diário Oficial da União, Brasília. 2011.(a)

BRASIL. Ministério da Agricultura Pecuária e Abastecimento, Instrução normativa nº 62, de 29 de dezembro de 2011 Regulamento Técnico de Produção, Identidade e Qualidade do Leite tipo A, Leite Cru Refrigerado, Leite Pasteurizado, Leite Cru Refrigerado e seu Transporte a Granel. Diário Oficial da União, Brasília. 2011(b).

BRASIL. Ministério da Agricultura Pecuária e Abastecimento. Decreto nº 9.013, de 29 de março de 2017. Regulamenta a inspeção industrial e sanitária de produtos de origem animal,

que disciplina a fiscalização e a inspeção industrial e sanitária de produtos de origem animal, Brasília, DF, 2017.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n°. 76 de 26 de novembro 2018. Aprova o Regulamento Técnico de Produção, Identidade e Qualidade do Leite Cru Refrigerado, Leite Pasteurizado e o Leite pasteurizado tipo A. Diário Oficial da República Federativa do Brasil. Brasília. 2018a.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n°. 77 de 26 de novembro 2018. Estabelece os critérios e procedimentos para a produção, acondicionamento, conservação, transporte, seleção e recepção do leite cru em estabelecimentos registrados no serviço de inspeção oficial, na forma desta Instrução Normativa e do seu Anexo Diário Oficial da República Federativa do Brasil. Brasília. 2018b.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa Nº 59, de 6 de Novembro de 2019. Confere o art. 87, parágrafo único, inciso II, da Constituição, tendo em vista o disposto na Lei nº 1.283, de 18 de dezembro de 1950, na Lei nº 7.889, de 23 de novembro de 1989, no Decreto nº 9.013, de 29 de março de 2017, e o que consta do Processo nº 21000.013698/2018-31.Diário Oficial da República Federativa do Brasil. Brasília. 2019.

BRASIL. Ministério da Agricultura Pecuária e Abastecimento. Decreto nº 10.468, de 18 de agosto de 2020. Altera o Decreto nº 9.013, de 29 de março de 2017, que regulamenta a Lei nº 1.283, de 18 de dezembro de 1950, e a Lei nº 7.889, de 23 de novembro de 1989, que dispõem sobre o regulamento da inspeção industrial e sanitária de produtos de origem animal, Brasília, DF, 2020a.

BRESSAN, M., MARTINS, M.. Segurança alimentar na cadeia produtiva do leite e alguns de seus desafios. Revista de Política Agrícola, Local de publicação (editar no plugin de tradução o arquivo da citação ABNT), 13, Jun. 2015. Disponível em: <https://seer.sede.embrapa.br/index.php/RPA/article/view/577>. Acesso em: 13 Ago. 2020.

BRITO, José R. F. et al. Adoção de boas práticas agropecuárias em propriedades leiteiras da Região Sudeste do Brasil como um passo para a produção de leite seguro. *Acta Scientiae Veterinariae*, v. 32, n. 2, p. 125-131, 2004.

BYLUND, Gösta. *Dairy Processing Handbook*. Tetra Pak Processing Systems AB, 2003.

CRUZ, Adriano Gomes da et al. Microbiologia, higiene e controle de qualidade no processamento de leites e derivados. 1 edição, Rio de Janeiro: Elieser, 2019.

COSTA, B. S.; CIRÍACO, N. M.; SANTOS, W. L. M.; ORNELLAS, C. B. D.; SANTOS, T. M. História e evolução da inspeção industrial e sanitária de produtos de origem animal no Brasil. *Cadernos Técnicos de Veterinária e Zootecnia*, nº 77 - setembro de 2015. FEP MVZ Editora, ed. p 9- 31.

DIAS, J. A.; ANTES, F. G. Qualidade físico-química, higiênico-sanitária e composicional do leite cru: indicadores e aplicações práticas da Instrução Normativa 62 /. -- Porto Velho, RO: Embrapa Rondônia. 2014.

DIAS, J. A.; CARVALHO, B. P.; MESQUITA, A. Q. de; LAMBERTUCCI, D. M.; CAVALCANTE, F. A. Caracterização epidemiológica dos indicadores de qualidade higiênico-sanitária do leite de rebanhos de três microrregiões do estado do Acre. Embrapa Rondônia, Porto Velho, RO, 2021. Acesso em 08/07/2022: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1136166/caracterizacao-epidemiologica-dos-indicadores-de-qualidade-higienico-sanitaria-do-leite-de-rebanhos-de-tres-microrregioes-do-estado-do-acre

DUTRA, Aline et al. Sistema logístico do transporte de leite a granel: um estudo de caso. CEP, v. 95070, p. 560, 2014.

FERRARI, VERENA AYRES SILVA. Transporte de leite a granel e sua influência na qualidade do leite que chega à indústria. Dissertação (Mestrado) – Universidade Estadual de Santa Cruz (UESC). Programa de Pós-graduação em Ciência Animal. 2018.

FOOD AND AGRICULTURE ORGANIZATION OF UNITED NATIONS-FAO ANIMAL PRODUCTION AND HEALTH PAPER 78 - Milking, milk production hygiene and udder health, 1989.

FOOD AND AGRICULTURE ORGANIZATION OF UNITED NATIONS. Dairy Production and Products –Milk Production. 2019. Disponível em: http://www.fao.org/dairy-production-products/production/en/.

FOOD AND AGRICULTURE ORGANIZATION OF UNITED NATIONS. Dairy Market Review. Mar., 2020. Disponível em: http://www.fao.org/3/ca8341en/CA8341EN.pdf. Acesso em: 08 de abril de 2021.

GONÇALVES, C. A.; VIEIRA, L. C. Obtenção e higienização do leite in natura. Embrapa Amazônia Oriental-Documentos (INFOTECA-E), 2002.

INTERNATIONAL DAIRY FEDERATION. PSYCHROTROPHS IN MILK AND MILK PRODUCTS, IDF E Doc 8 – *International Dairy Federation*, Brussel, 1976.

INTERNATIONAL DAIRY FEDERATION. THE GLOBAL DAIRY SECTOR: Facts 2019. Disponível em: https://www.fil-idf.org/wp-content/uploads/2021/01/DDOR-Global-Dairy-Facts2019.pdf. Acesso em: 08 de abril de 2021.

LOPES, Carla Machado DE ARAUJO et al. Influência das boas práticas agropecuárias na contagem padrão em placas (CPP) e na contagem de células somáticas (CCS) no leite cru. *Brazilian Journal of Development*, v. 8, n. 3, p. 21519-21536, 2022.

MARANGONI, F. et al. Cow's milk consumption and health: A health professional's guide. *Journal of the American College of Nutrition*, v. 38 n. 3, p. 197-208, 2019.

MARDER, E. P.; GRIFFIN, P. M.; CIESLAK, P. R.; DUNN, J.; HURD, S.; JERVIS, R., MARDER, E. P. et al. Preliminary Incidence and Trends of Infections with Pathogens Transmitted Commonly Through Food - Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2006-2017. *Morbidity & Mortality Weekly Report*, [s. l.], v. 67, n. 11, p. 324–328, 2018. Disponível em: http://search-ebscohost-com.ez27.periodicos.capes.gov.br/login.aspx?direct=true&db=c8h&AN=128749342&lang=pt-br&site=ehost-live. Acesso em: 5 out. 2020.

MARCONDES, M I; BRANDÃO, V. L. N.; FERREIRA, G. A. T. ; SILVA, A. L.. Impact of farm size on milk quality in the Brazilian dairy industry according to the seasons of the year. Ciência Rural, Santa Maria, v.47: 11, e20161004, 2017.

MARIOTO, L. R. M.; DANIEL, G. C.; GONZAGA, N.; MAREZE, J.; TAMANINI, R.; BELOTI, V. Potencial deteriorante da microbiota mesófila, psicrotrófica, termodúrica e esporulada do leite cru. *Ciência Animal Brasileira*, Goiânia, v. 21, e-44034, 2020.

MELO, E. S.; AMORIM, W. R.; PINHEIRO, R. E. E.; NASCIMENTO CORRÊA, P. G.; CARVALHO, S. M. R.; SANTOS, A. R. S. S., et al. Doenças transmitidas por alimentos e principais agentes bacterianos envolvidos em surtos no Brasil. PUBVET, 12, 131. 2018.

MUIR, D. DONALD. The shelf-life of dairy products: 2. Raw milk and fresh products. *International Journal of Dairy Technology*, v. 49, n. 2, p. 44-48, 1996.

NERO, L. A.; MATTOS, M. R. D.; BELOTI, V.; BARROS, M. D. A.; PINTO, J. P. D. A.; ANDRADE, N. J. D. *et al.* Leite cru de quatro regiões leiteiras brasileiras: perspectivas de

atendimento dos requisitos microbiológicos estabelecidos pela Instrução Normativa 51. *Ciência e Tecnologia de Alimentos.* Campinas, v. 25, n. 1, p. 191-195, Mar. 2005.

Acesso:http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-20612005000100031&lng=en&nrm=iso 11de agosto de 2020

ORDONEZ, J. *Tecnologia de Alimentos*. Volume 2. Alimentos de origem animal–1 a. Ed. Artmed–SP, 2005.

PAIXÃO, M. G. *et al*. Carretagem de Leite a Granel: Um Estudo de Caso. *Revista do Instituto de Laticínios Cândido Tostes*, v. 66, n. 382, p. 42-47, 2011.

PAIXÃO, M. G.; LOPES, M. A.; PINTO, S. M.; ABREU, L. R. D. Impacto econômico da implantação das boas práticas agropecuárias relacionadas com a qualidade do leite. *Revista Ceres*, Viçosa, v. 61, n. 5, p. 612-621, Oct. 2014. Disponível em: https://doi.org/10.1590/0034-737X201461050003

PHILPOT, W.; NICKERSON, S. *Vencendo a Luta Contra a Mastite*. Publicado por Westfalia Surge Inc. e Westfalia Landtechnik do Brasil Ltda. Brasil. Milkbizz. Edição Brasileira, p. 6-9, 2002.

SANTANA, Elsa Helena Walter de; BELOTI, Vanerli; BARROS, Márcia de Aguiar Ferreira. Microrganismos psicrotróficos em leite. Revista Higiene Alimentar, p. 27-33, 2001.

SANTOS, Marcos Veiga dos; FONSECA, Luis Fernando Laranja da. Importância e efeito de bactérias psicrotróficas sobre a qualidade do leite. Revista Higiene Alimentar, v. 15, n. 82, p. 13-19, 2001.

SEQUETTO, P. L.; ANTUNES, A. S.; NUNES, A. S.; ALCANTARA, L. K. S.; REZENDE, M. A. R.; PINTO, M. A. O. et al. Avaliação da qualidade microbiológica de leite cru refrigerado obtido de propriedades rurais da Zona da Mata mineira. *Revista Brasileira de Agropecuária Sustentável*, v. 7, n. 1, 2017.

SILVA, N. N. et al. Micelas de caseína: dos monômeros à estrutura supramolecular. *Brazilian Journal of Food Technology*, v. 22, 2019.

TEIXEIRA, S. R.; RIBEIRO, M. T. *Transporte do leite a granel*. Embrapa Gado de Leite, 2000.

VALLIN V.M.; BELOTI V.; BATTAGLINI A.P.P.; TAMANINI R.; FAGNANI R.; ANGELA H.L.; SILVA L.C.C. Melhoria da qualidade do leite após implantação de boas práticas de

fabricação em ordenha em 19 municípios da região central do Paraná. *Semina*: Ciências Agrárias, v.30, p.181-188, 2009. Disponível em: http://www.uel.br/revistas/uel/index.php/semagrarias/article/view/2661/ 2313 >.

VASCONCELLOS, S.A.; ITO, F.H. Principais zoonoses transmitidas pelo leite – Atualização / Major milk transmitted zoonoses – Update / *Revista de Educação Continuada em Medicina Veterinária e Zootecnia do CRMV-SP*. São Paulo: Conselho Regional de Medicina Veterinária, v. 9, n. 1 (2011), p. 32–37, 2011.

VERDUCI, E. et al. Cow's milk substitutes for children: Nutritional aspects of milk from different mammalian species, special formula and plant-based beverages. *Nutrients*, v. 11, n. 8, p. 1739, 2019.

WHO. World Health Organization. estimates of the global burden of foodborne diseases: foodborne disease burden epidemiology reference group 2007-2015. 255 p., 2015. Disponível em: https://www.who.int/foodsafety/publications/foodborne_disease/fergreport/en/

WHO. World Health Organization. COVID-19 and Food Safety: Guidance for Food Businesses: interim guidance. p.5. 2020. https://www.who.int/publications-detail/covid-19-and-food-safetyguidance-for-food-businesses.

Capítulo 2

Machine Learning: conceitos importantes

O termo *Machine Learning* (ML) é um ramo da inteligência artificial que permite que sistemas de computador aprendam por meio de exemplos, dados e experiências, de forma a realizar tarefas complexas, visando automatizar a criação de modelos analíticos e/ou estatísticos. Este ramo da inteligência artificial baseia-se na ideia de que sistemas podem aprender, principalmente através de dados, podendo identificar padrões e auxiliando humanos a tomarem melhores decisões (Deo, 2015; Vopham *et al.*, 2018).

Machine Learning faz parte do estudo de Inteligência Artificial (IA). Esse campo possui um vínculo forte com a estatística, matemática e a ciência da computação, sendo capazes de criar algoritmos e soluções que conseguem lidar com pequenas ou grandes quantidades de dados, auxiliando a ciência no avanço desde tópicos simples até os mais complexos (Deo, 2015; Vopham *et al*., 2018).

Machine Learning é o processo pelo qual os parâmetros de uma rede neural artificial (RNA) são ajustados e a RNA é o modelo ou a ferramenta utilizada para alcançar essa finalidade. O aprendizado pode ser dividido em três grandes grupos: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

Aprendizado é o processo pelo qual os parâmetros de uma rede são ajustados através de uma forma contínua de estímulo, pelo ambiente no qual a rede está operando, dado por um algoritmo de aprendizado. O algoritmo de aprendizado é um conjunto de procedimentos bem definidos, para adaptar os parâmetros de uma RNA, de forma que a mesma possa aprender uma determinada função (Braga *et al*., 2007).

O tipo de aprendizado realizado é definido pela maneira como ocorrem os ajustes realizados nos parâmetros. Neste trabalho vamos nos ater ao aprendizado supervisionado que é o modelo que será utilizado.

2.1 Aprendizado Supervisionado

No aprendizado supervisionado, os dados de entrada e saída são fornecidos de forma externa por um supervisor, que direciona o processo de treinamento da rede em relação a um comportamento bom ou ruim, através da resposta que a rede dá àquele problema. O supervisor

consegue fazer isso ajustando os pesos das conexões das variáveis de entrada, até chegar a uma resposta desejada (Braga *et al*., 2007).

Quando se fala em Aprendizado Supervisionado, o objetivo dos modelos/algoritmos é de prever uma variável já conhecida pelos dados coletados. Como exemplo de tarefas que já são conhecidas, tem-se o reconhecimento de imagens, classificação de documentos, e principalmente, nas atividades que envolvam regressão e classificação de dados. (Deo, 2015).

A abordagem de classificação funciona da seguinte maneira: já conhecendo em que subgrupos cada registro do conjunto de dados pertence, tenta-se prever em qual subgrupo os atuais e novos registros (os que ainda possuem a variável alvo desconhecida), tendem a se encaixar. Ainda dentro deste universo, podem-se falar de duas ramificações: Classificação Binária, quando apenas há a possibilidade da resposta pertencer ou não a uma classe (0 ou 1) e Classificação Múltipla, caso existam várias classes a serem identificadas.

Para realizar essa identificação, os modelos precisam ser treinados com dados que tenham a variável alvo conhecida, para aprender os padrões e comportamentos comuns a cada uma das classes existentes. Para isso, usa-se um grupo de variáveis, preditora/regressoras e a partir delas os modelos são treinados.

O trabalho de Valente *et al.* (2014) é um exemplo de classificação binária para detecção de leite normal ou adulterado com soro do queijo, onde foram apresentadas uma parte dos dados e omitida outra parte. As variáveis de entrada foram: temperatura, pH, teor de gordura, extrato seco desengordurado, proteínas, ponto de congelamento, condutividade, lactose e densidade das amostras. O resultado dessa classificação pode ser leite normal ou adulterado. Depois do treinamento a rede é testada com os dados de entrada omitidos para avaliar a rede.

Da Silva e Da Silva (1995), em seu trabalho de previsão de volume em séries temporais com uso de RNA em uma indústria de laticínios, observou uma previsão satisfatória para a época que o trabalho foi apresentado, demonstrando 68% de acerto pela rede. Este trabalho teria o objetivo de colaborar com indústrias de pequeno e médio porte a se organizarem quanto à diferença de demandas, durante os períodos do ano.

Na pesquisa realizada por Ferrão *et al*. (2007), a rede neural foi utilizada para regressão multivariada e comparada com métodos tradicionais de regressão estatística, para detecção de adulterantes em leite em pó. O modelo de regressão por RNA apresentou vantagens pela sua capacidade de generalização e flexibilidade, sendo capaz de inferir mesmo na ausência de um

ou mais adulterantes, em comparação com métodos tradicionais que apresentaram um falso positivo.

2.2 Qualidade dos dados

Quando se trabalha em uma pesquisa com banco de dados, a qualidade dos mesmos é determinante para a segurança dos resultados obtidos na pesquisa. Tratando-se de banco de dados secundários, principalmente aqueles manuscritos, surgem diversos desafios como dificuldade na compreensão da escrita, rasuras, avarias no documento, perda de páginas, dentre outros.

Outros problemas encontrados em bancos planilhados são dados faltantes, falta de padronização, informações incorretas ou imprecisas. Dessa forma, a etapa de pré-processamento dos dados tem como função minimizar e tratar os problemas, antes do processamento (Batista, 2003).

2.3 Pré-processamento dos dados

O pré-processamento dos dados é uma etapa que depende muito do conhecimento do analista de dados sobre a natureza do problema, com isso a tomada de decisão para solucioná-lo, optando por padronizar e imputar os dados ou por excluí-lo do banco (Batista, 2003).

Os três principais passos envolvidos na etapa de pré-processamento de dados são: limpeza de dados, transformação de dados e redução de dados, sendo que cada um deles envolve diversas atividades.

2.3.1 Limpeza dos dados

De modo geral, uma boa parte dos dados apresenta erros ou anormalidades. Essas anormalidades criam problemas à sua utilização, influenciando negativamente a validade dos resultados e conclusões obtidas, resultando em um custo maior e num proveito menor para o utilizador. Dessa forma, antes da escolha ou aplicação de qualquer ferramenta para análise, os dados devem ser “limpos” com o intuito de remover e reparar quaisquer anomalias que possam existir (Oliveira; Rodrigues; Henriques, 2004).

Em grandes bancos de dados podem-se eliminar os dados faltantes para tentar eliminar o ruído, mas em bancos de dados pequenos, em que todos os exemplos e conjuntos de dados são importantes procuram-se recursos para substituir os ruídos por valores consistentes. A etapa de limpeza visa apresentar os dados de forma apropriada aos algoritmos de mineração de dados.

Dentre as principais operações de limpeza tem-se: eliminação de dados errôneos, padronização de dados, eliminação de dados em duplicidade e tratamento de valores ausentes (Hsu *et al*., 2000).

2.3.2 Transformação de dados

Geralmente os algoritmos utilizados no *Machine Learning* requerem que os dados estejam em um formato apropriado, fazendo necessário a etapa de transformação dos dados. Dentre as principais operações de transformação podemos citar a Padronização dos dados, que consiste em converter valores de atributos para faixas de -1 e 1 ou 0 e 1, sendo de grande utilidade para algoritmos de classificação. Outra operação de transformação de dados é a conversão de valores simbólicos para valores numéricos, discretização de atributos e composição de atributos (Han; Kamber, 2006).

2.3.3 Redução de dados

Após as etapas de limpeza e transformação dos dados deve-se avaliar a seleção das variáveis que entrarão no treinamento do modelo. O motivo para selecionar as variáveis é o ganho de velocidade no treinamento do modelo. Ao eliminar as variáveis não relevantes ocorre redução na dimensionalidade do problema, reduzindo o tempo de treinamento e melhora da performance do modelo (Han; Kamber, 2006).

As variáveis que não contribuem para a predição do modelo causam ruídos e aumentam as chances de ocorrer *overfitting*, que é um sobreajuste do modelo aos dados estudados, fazendo com que o modelo perca a sua capacidade de generalização (Han; Kamber, 2006).

Há três métodos de seleção de variáveis: *filter methods*, *wrapper methods* e *embedded methods*. Os *filter methods*, como é o caso dos métodos estatísticos, fornecem um *ranking* das variáveis com relação a alguma medida de importância e um exemplo é a correlação de Pearson. *Wrapper methods* realiza manipulações nas variáveis para avaliar a correlação entre elas e o problema que se procura solucionar e um exemplo é o algoritmo Boruta. *Embedded methods* são métodos que aprendem quais variáveis contribuem para a acurácia do modelo no momento em que está sendo criado e como exemplos, podem ser citados: Randon forest, XGBoost e LASSO (Lira; Chaves Neto, 2006; Han; Kamber, 2006).

2.4 Seleção de variáveis

De modo geral, a maioria dos agravos ocorre de maneira não linear e estocástica, ou seja, diversas variáveis podem estar relacionadas para que um determinado agravo ocorra. A maioria

dos fenômenos apresenta características dinâmicas e dependentes de um número de variáveis interagindo, para que ocorram (Velázquez, 2006).

Para que se tenha uma boa previsão de dados é necessária uma boa escolha das variáveis, para qualquer modelo de ML. É necessário selecionar as variáveis preditoras, ou seja, aquelas que melhor explicam a variável que eu quero prever. Retirar variáveis que não acrescentam ou aquelas que querem dizer a mesma coisa, impedem que haja ruído no meu modelo atrapalhando o treinamento. (Velázquez, 2006).

2.4.1 Algoritmo Boruta

O algoritmo Boruta é um método de seleção de variáveis, criado inicialmente para a linguagem R por Miron B. Kursa e Witold R. Rudnicki (2010), baseado na implementação da Random Forest de R. Trata-se de um algoritmo de seleção de variáveis do tipo *Wrapper method* (método empacotador), que considera a seleção de um conjunto de variáveis, como um problema de busca no qual diferentes combinações de variáveis são preparadas, calculadas e comparadas com outras combinações. Assim o algoritmo decide qual variável deve ser removida do conjunto (Pathak, 2018; Kursa; Rudnicki, 2010).

Boruta é baseado na ideia de variável sombra e distribuição binomial. Quando o mesmo é utilizado, as características não são avaliadas como elas mesmas, mas como uma versão aleatória delas. Primeiro duplica-se o conjunto de dados criando variáveis sombras e mistura-se aleatoriamente os valores de cada coluna, verificando para cada variável real se a mesma tem uma importância maior que a variável sombra. Se houver, isso é chamado de acerto e continuam as demais iterações. Este passo é chamado de características de sombra e tem a função de remover suas correlações com a variável resposta (Kursa; Rudnicki, 2010).

2.4.2 *Recursive Feature Elimination* (RFE)

Recursive Feature Elimination (RFE) é um método simples e comum de seleção de variáveis, que por meio de um estimador externo, infere valor às variáveis e remove as mais fracas por meio de inúmeras iterações, até que um número especificado de variáveis é alcançado. Este método elimina um número de atributos definido por *step* de atributos a cada iteração, acabando por eliminar dependências e colinearidades que podem vir a existir em um modelo. RFE, por meio de um parâmetro K, define quantos atributos devem ser mantidos (Mendes; Jesus, 2021).

Não estão determinados quantos atributos devem de fato, ser mantidos. Assim, para encontrar o melhor número de atributos ou os que resultam em um melhor valor de classificação, uma

validação cruzada pode ser realizada. Junto com o RFE, essa validação cruzada separa em diferentes subconjuntos de atributos, quantificando-os e selecionando os melhores (Mendes; Jesus, 2021).

2.5 Modelos de predição

2.5.1 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA) surgiram no final da década de 80, e são uma forma de computação não-algorítmica que simula a estrutura de funcionamento do cérebro humano (Braga *et al*., 2007).

As RNA são sistemas constituídos de unidades de processamentos simples que calculam determinadas funções matemáticas, normalmente não-lineares. As unidades de processamentos podem possuir uma ou mais camadas que estão interligadas entre si por conexões, que geralmente são unidirecionais (Braga *et al*., 2007).

As redes são capazes de generalizar dados apresentados e aprender com eles. Assim, quando um determinado conjunto de dados é apresentado para o programa e dados de saída são dados a ele, o modelo aprende a solucionar aquele problema. Sua capacidade de aprender e solucionar aquele problema torna-se muito atrativa, pois oferece propriedades e capacidades úteis, tais como a não-linearidade, adaptabilidade, tolerância a falhas e mapeamento de entrada/saída. As principais áreas de aplicação das RNA são para solucionar problemas de regressão, previsão e classificação (Braga *et al*., 2007).

Com a evolução e maior acesso ao uso dos computadores nos últimos anos, termos como *Machine Learning* e Ciência de Dados se tornaram usuais em diversos ambientes corporativos, na academia e na maioria das áreas do conhecimento. As máquinas estão se aperfeiçoando no aprendizado autônomo e na indústria de alimentos não é diferente: o termo microbiologia preditiva vem sendo utilizado desde os anos 90 pelo uso de modelos estatísticos para auxiliar na melhoria da segurança e qualidade dos alimentos (Cruz *et al*., 2019).

Acredita-se que cerca de 30% da população seja acometida por doenças transmitidas por alimentos. Desta forma, a melhoria dos processos, produtos, rastreabilidade e logística são de extrema importância para conseguirmos garantir a segurança alimentar. No Brasil os problemas com transporte do leite cru a granel são um entrave para a melhoria do setor. Dessa forma, a utilização de ferramentas que auxiliam na melhor compreensão das variáveis envolvidas na

qualidade de leite é de extrema importância para a cadeia produtiva de leite do Brasil (Cruz *et al*., 2019; Food…, 2019).

2.5.2 Principais Arquiteturas das RNA

A arquitetura da RNA informa o tipo de problema que a rede é capaz de resolver, como por exemplo, as redes com camada única MCP pois ela somente será capaz de resolver problemas linearmente separáveis (Braga *et al*., 2007).

A figura 1 apresenta as principais arquiteturas de redes neurais artificiais. Existem três tipos de classificação de arquitetura de rede. Quanto ao número de camadas, podem ser de camada única (Fig1.5 a, e) ou múltiplas camadas. (Fig. 1.5 b, c, d). Quanto ao tipo de conexão dos nodos podem ser acíclica (Fig. 1.5 a, b, c) ou cíclica (Fig. 1.5 d, e). Por último tem-se a classificação quanto ao tipo de conectividade podendo ser, fracamente /parcialmente conectada (Fig. 1.5 b, c, d) ou completamente conectada (Fig. 1.5 a, e) (Braga *et al*., 2007).

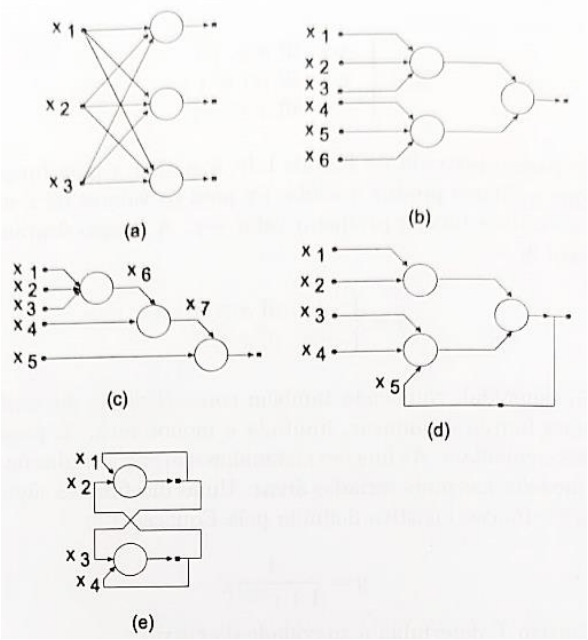


Figura 1: Arquiteturas de Redes Neurais Artificiais (RNA). RNA de camada única (Fig 1.5 a, e), RNA de múltiplas camadas (Fig. 1.5 b, c, d). RNA de conexão acíclica (Fig. 1.5 a, b, c), RNA cíclica (Fig. 1.5 d, e). RNA fracamente /parcialmente conectada (Fig. 1.5 b, c, d), RNA completamente conectada (Fig. 1.5 a, e).

Fonte: (Braga *et al*., 2007)

2.8 Neurônio Artificial MCP

A rede neural é caracterizada por uma estrutura de neurônios artificiais que simulam a função de um neurônio biológico. Pode-se dizer que o neurônio McCulloch e Pitts (MCP) é a unidade básica de uma rede neural, assim como o neurônio biológico é a unidade básica do sistema nervoso.

O neurônio MCP é uma representação simplificada do neurônio biológico. Na figura 2 observa-se o modelo simplificado do neurônio MCP, em que X representa os dados/variáveis de entrada; W, os pesos de cada variável de entrada; no centro acontece a soma das entradas Xi ponderadas pelos pesos Wi, por meio de uma função de ativação f(u); e Y são os dados de saída (Fig. 2) (Braga *et al.*, 2007).

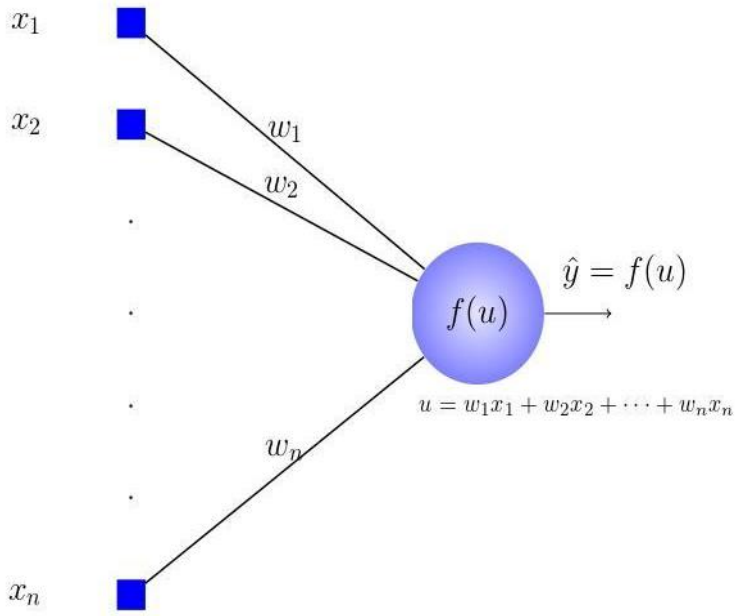


Figura 2: Modelo esquemático do neurônio McCulloch e Pitts (MCP), X representa os dados/variáveis de entrada; W, os pesos de cada variável de entrada; no centro acontece a soma das entradas Xi ponderadas pelos pesos Wi, por meio de uma função de ativação f(u); e Y são os dados de saída

Fonte: Braga *et al.* (2007)

2.9 Redes Neurais *MultiLayer Perceptron* (MLP)

As redes neurais *MultiLayer Perceptron* (MLP) apresentam pelo menos uma camada intermediária, sendo, desta forma, capazes de resolver problemas não linearmente separáveis e permitindo a aproximação de qualquer função matemática (Braga *et al*., 2000).

O treinamento da rede MLP supervisionado utilizando *backpropagation* compõe duas etapas. Na primeira, um padrão é apresentado à camada de entrada e, a partir desta camada as unidades calculam sua resposta que é produzida na camada de saída, o erro é calculado. No segundo passo, o erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo modificados utilizando a regra delta generalizada (Braga *et al*., 2000).

O desafio do trabalho com RNA é definir o número de camadas intermediárias e o número de nodos da camada que são definidos em função da complexidade do problema e dos dados disponíveis. Uma vez definida a geometria da rede inicial um processo de refinamento sucessivo é realizado até que chegue na estrutura final para o modelamento para que não ocorra *overfiting* ou *underfitting* (Braga *et al*., 2000).

Dentre os algoritmos de treinamento de Redes Neurais Artificiais, o algoritmo de *backpropagation* é um dos mais utilizados, sendo ele um algoritmo supervisionado que utiliza pares (entrada e saída desejadas) para, por meio de um mecanismo de correção de erros, ajustar os pesos da rede. O treinamento ocorre em duas fases que têm sentidos opostos. A fase *forward* (para frente) é usada para definir a saída da rede para um dado padrão de entrada. A fase *backward* (para trás) utiliza a saída desejada e a saída fornecida pela rede para atualizar os pesos de suas conexões (Braga *et al*., 2000).

2.10 XGboosting

Modelo XGBoost é um algoritmo de aprendizado de máquina, baseado em árvore de decisão e que utiliza uma estrutura de *Gradient boosting*. Constituem *ensembles* de árvores de regressão semelhantes que formulam hipóteses sobre os exemplos agregando as respostas de uma assembleia de preditores simples. O conjunto de árvores de regressão que compõem esta assembleia é elaborado em duas etapas: (1) uma árvore $T0$ é construída adicionando à sua estrutura a ramificação do atributo Fi que mais aperfeiçoa o preditor a cada etapa e, definida sua estrutura, são calculados os valores ótimos para as folhas l; (2) a inclusão de novas árvores

*T*1...*N* ao *ensemble* é orientada pelo erro residual do modelo, de maneira que novos preditores complementam as imperfeições dos anteriores (Friedman, 2002).

Modelos XGBoost diferem de outras técnicas de *Gradient boostin* pois empregam um algoritmo sensível à dispersão na busca por ramificações, que torna a complexidade computacional do modelo linear ao número de observações não ausentes. Assim, integram otimização do uso de recursos que permitem calcular paralelamente a aptidão dos atributos (Chen; Guestrin, 2016).

2.11 *Support vector machine* (SVM)

O *Support Vector Machine* (SVM) são ferramentas de classificação e regressão, são uma generalização do algoritmo *Generalized Portrait* que foi desenvolvido por Vapnik, Lemer, e Chervonenkis na década de 60, são modelos de aprendizagem supervisionados aplicados tanto em tarefas de classificação quanto de regressão e predição (Tapak *et al*., 2019, Vapnik, 1998).

O SVM é uma abordagem geral para problemas de estimativa de função, ou seja, para problemas de encontrar a função y = f(x) dada por suas medidas Yi com ruído em alguns vetores (geralmente aleatórios) x. Este método pode ser aplicado para reconhecimento de padrões (para estimar indicadores funções), para regressão (para estimar funções de valor real) e para resolver equações de operadores lineares.

O conceito de *Support Vector Machine* (SVM) foi introduzido nos anos 60, mas a implementação do algoritmo, como classificador não linear pela aplicação de funções Kernel, só surgiu no final do século XX (Boser *et al*., 1992).

SVM tenta encontrar o hiperplano ótimo que permita maximizar a separação de dados de classes distintas (Boser *et al*., 1992). A margem de separação dos dados define a fronteira de decisão (hiperplano ótimo) e os padrões que estão mais próximos dela são os vetores de suporte. Além do ótimo desempenho em problemas de classificação, o SVM também pode ser utilizado em problemas de regressão em que cada variável a ser predita é definida como *target*, enquanto as restantes são consideradas como dados de entrada. Este método apresenta uma fase de treino, em que são utilizados os dados completos com o intuito de determinar os parâmetros óptimos do modelo. Por fim, os dados são preditos seguindo o modelo com os parâmetros obtidos.

Em problemas de regressão, os SVMs já foram aplicados, por exemplo na previsão da qualidade do ar (Liu *et al*., 2017), demanda de água e previsão da qualidade da água (Ghalehkhondabi *et al*., 2017; Zhang *et al.*, 2017), e na previsão de surtos de Influenza (Tapak et al., 2019).

2.12 Validação cruzada

O método de validação cruzada é utilizado para avaliar a capacidade de generalização do modelo e evitar o *overfitting*. O *overfitting* é um termo estatístico usado quando ocorre um sobreajuste dos dados do modelo, ou seja, quando o modelo se ajusta muito bem ao conjunto de dados anteriormente apresentado, mas se mostra ineficaz para prever novos resultados (Berrar, 2019).

A validação cruzada pertence à classe de métodos de Monte Carlo e é utilizada por meio de uma série de divisões nos conjuntos de dados, apresentando em cada etapa um conjunto diferente de dados para teste, evitando assim um viés do modelo e melhor balanceamento deste em relação a realidade dos dados. Assim a validação cruzada pode ser utilizada para ajustar os parâmetros do modelo de forma mais fiel à realidade dos dados (Schaffer, 1993; Berrar, 2019).

2.13 Referências Bibliográficas

BERRAR, D. *Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology*, Volume 1, Elsevier, pp. 542-545, 2019.

BOSER., B. E.; GUYON, I. M.; VAPNIK V. N. "A Training Algorithm for Optimal Margin Classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152. 1992.

BRAGA, A. P., LURDEMIR, T. B., AND CARVALHO, A. C. P. L. F. *Redes Neurais Artificiais: Teoria e Aplicações*. LTC - Livros Técnicos e Científicos Editora S.A. Belo Horizonte, Brasil, 2000.

BRAGA, A. P. e LUDERMIR, T. B. e CARVALHO, A. C. P. de L. F. *Redes neurais artificiais: teoria e aplicações*. Rio de Janeiro: LTC. 2007.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: ACM. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. p. 785–794, 2016.

DEO, R. C. Machine learning in medicine. *Circulation*, v. 132, n. 20, p. 1920-1930, 2015.

FRIEDMAN, J. H. *Stochastic gradient boosting. Computational Statistics & Data Analysis*, Elsevier, v. 38, n. 4, p. 367–378, 2002.

GHALEHKHONDABI, I.; ARDJMAND, E.; YOUNG, W.N.; et al. Water demand forecasting: review of soft computing methods. *Environmental monitoring and assessment*. v. 189, n. 7, p. 1-13, 2017.

LIU, B. C.; BINAYKIA, A.; CHANG P. C. *et al*. Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. *Plos one*, 12 (7):e0179763, 2017.

SCHAFFER, C. Selecting a classification method by cross-validation. *Machine Learning*, v. 13, n. 1, p. 135-143, 1993.

TAPAK, L.; HAMIDI, O.; FATHIAN, M. *et al*. Comparative evaluation of time series models for predicting influenza outbreaks: application of influenza-like illness data from sentinel sites of healthcare centers in Iran. *BMC research notes*, v. 12, n. 1, p. 1-6, 2019.

VAPNIK, V. The support vector method of function estimation. In: *Nonlinear Modeling*. Springer, Boston, MA, 1998. p. 55-85.

VOPHAM, T; HART, J. E.; LADEN F. et al. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. *Eenvironmental Health*,17(1):40, 2018. doi: 10.1186/s12940-018-0386-x.

ZHANG, L.; ZOU, Z.; SHAN, W. Development of a method for comprehensive water quality forecasting and its application in Miyun reservoir of Beijing, China. *Journal of Environmental Sciences*, 56:240-246, 2017.

Capítulo 3

Imputação múltipla de dados faltantes nas análises de contagem bacteriana do leite cru do produtor, utilizando o algoritmo MICE

3.1 Resumo

O objetivo deste trabalho foi apresentar uma metodologia que pode ser utilizada na rotina do laticínio para imputar dados de leite em análises de CPP do produtor que por algum motivo foram perdidas. A perda de amostra de CPP do produtor pode gerar um grande impacto para o laticínio, justificando a busca por tecnologias que possam auxiliar o técnico a estimar estes dados, tendo em vista a importância em identificar não conformidades e corrigi-las o mais breve possível. Utilizou-se um modelo de imputação múltipla de dados, algoritmo MICE versão 3.14.0 com validação cruzada. Utilizou-se um método de avaliação do algoritmo MICE, realizando falhas aleatórias (5% e 10% de falhas) no banco e comparando ao valor real. Observou-se um MAE (Erro Médio Absoluto), para contagem bacteriana de 4,73 para 5% de falha e 16,93 para 10% de falha. O coeficiente de correlação de Pearson (r), para 5%, se falha foi 1, e para 10%, se falha foi 0,96 significando que há uma alta correlação entre o valor previsto e o valor realizado. O modelo se mostrou eficiente para imputação de dados faltantes para CPP do produtor.

3.2 Introdução

O leite possui aproximadamente 87% de água e 13% de sólidos, sendo que destes, 3,9% são de gordura e 9,1% de sólidos não gordurosos (SNG). Com tantos nutrientes, o leite pode ser um importante meio de cultura para a multiplicação de microrganismos, se não for obtido higienicamente, estocado sob refrigeração, transportado e processado em condições higiênico-sanitárias adequadas (Bylund, 2003).

Cada etapa da produção e processamento do leite pode apresentar riscos de contaminação que além de causar a sua decomposição, gerar danos à saúde do consumidor. Na análise dos perigos e pontos críticos de controle na produção de leite, devem ser observados desde a sua obtenção na fazenda (sanidade animal, manejo de ordenha, estocagem, resfriamento), até a coleta, transporte, estocagem na indústria, processamento, envase, transporte e estocagem do produto final (Philpot; Nickerson, 2002; Bylund, 2003).

A IN 77 estabelece critérios para produção, acondicionamento, conservação, transporte, seleção e recepção do leite cru. Dentre as recomendações, o estabelecimento deve incluir junto aos seus

Programas de Autocontrole (PAC), o Programa de Qualificação de Fornecedores de Leite (PQFL). A assistência técnica, gerencial e capacitação de fornecedores, com foco na gestão da propriedade rural e Boas Práticas Agropecuárias (BPA) são responsabilidades das indústrias de laticínios (Brasil, 2018a, Brasil, 2020b).

A implantação das BPA na fazenda é de extrema importância para a obtenção de matéria-prima de boa qualidade e deve ser incluída nos Programas de Autocontrole realizados pela indústria. Nestes programas devem ser mantidos registros auditáveis que evidenciem a execução, atingimento de metas pelos fornecedores, por um período mínimo de 12 meses (Vallin *et al.*, 2009; Brasil, 2018b).

A IN 76 fixa os padrões de identidade e qualidade do leite cru refrigerado, leite pasteurizado, e leite pasteurizado tipo A, estabelecendo outros limites legais. Dispõe que o leite cru refrigerado de tanque individual ou de uso comunitário deve apresentar médias geométricas trimestrais de Contagem Padrão em Placas de no máximo 300.000 UFC/mL (trezentas mil unidades formadoras de colônia por mililitro) e de Contagem de Células Somáticas de no máximo 500.000 CS/mL (quinhentas mil células por mililitro). O leite cru refrigerado deve apresentar limite máximo para Contagem Padrão em Placas de até 900.000 UFC/mL antes do seu processamento no estabelecimento beneficiador (Brasil, 2018b).

As médias geométricas devem considerar as análises realizadas no período de três meses consecutivos e ininterruptos com no mínimo uma amostra mensal de cada tanque (Brasil, 2018b).

Os limites de temperatura de recebimento do leite devem ser de no máximo 7°C, excepcionalmente de 9º C (Brasil, 2018b).

As análises de CPP do produtor de leite em laticínios de pequeno e médio porte são realizadas, geralmente, em uma frequência de uma a duas vezes no mês. Assim, para o técnico responsável pela qualidade da matéria-prima, a perda dessa amostra pode ter um grande impacto, já que a próxima análise somente será coletada 15 a 30 dias depois.

Diversos fatores podem contribuir para que ocorram erros na coleta de amostras de leite impedindo que a análise seja realizada: erro de coleta, perda da amostra, quantidade insuficiente, acondicionamento inadequado, coagulação do leite, falta de conservante, temperatura de chegada ao laboratório acima da prevista (maior do que 10 ºC), presença de sujidades, dentre outros. Pensando neste problema, foi proposta a utilização de um modelo de

imputação multivariada, nas amostras que foram perdidas por algum motivo para que ações preventivas possam ser adotadas pelas indústrias de laticínios.

A escolha do melhor método de imputação de dados neste estudo, baseia-se na técnica sugerida por Harrel (2001), que recomenda, entre 5 e 15 por cento de dados faltantes a imputação única pode ser usada. No entanto, o uso da imputação múltipla é mais indicado.

O método de imputação multivariada, utilizado neste estudo foi o MICE (*Multivariate Imputation by Chained Equations*), utilizado em problemas complexos.

O método MICE utiliza uma abordagem que chamamos de *Fully Conditional Specification* (FCS), este é um método Monte Carlo via Cadeias de Markov (MCMC). Para cada iteração e para cada variável na ordem especificada na lista de variáveis, o método FCS ajusta um modelo (variável dependente única) usando todas as outras variáveis disponíveis no modelo como preditores, em seguida, imputa os valores faltantes para a variável que está sendo ajustada. O método continua até que o número de iterações seja atingido e a média dos valores imputados nas interações sejam salvos no conjunto de dados imputados (Van Buuren; Groothuis-Oudshoorn, 2011).

O método de validação cruzada é utilizado para avaliar a capacidade de generalização do modelo e evitar o *overfitting*. O *overfitting* é um termo estatístico usado quando ocorre um sobreajuste dos dados do modelo, ou seja, é quando um modelo estatístico se ajusta muito bem ao conjunto de dados anteriormente apresentado, mas se mostra ineficaz para prever novos resultados (Berrar, 2019).

A validação cruzada é utilizada por meio de uma série de divisões nos conjuntos de dados, apresentando em cada etapa, um conjunto diferente de dados para teste, evitando assim um viés do modelo e melhor balanceamento deste em relação à realidade dos dados. Assim a validação cruzada pode ser utilizada para ajustar os parâmetros do modelo de forma mais fiel à realidade dos dados (Schaffer, 1993; Berrar, 2019).

3.3 Material e Métodos

3.3.1 Delineamento do estudo

O estudo foi realizado em um laticínio localizado na região noroeste do estado de São Paulo, que possui cerca de 50.000 km2 e é formada por 153 municípios, distribuídos em doze microrregiões (Figura 3). O maior município dessa região é São José do Rio Preto (Conceição; Tonietto, 2012).

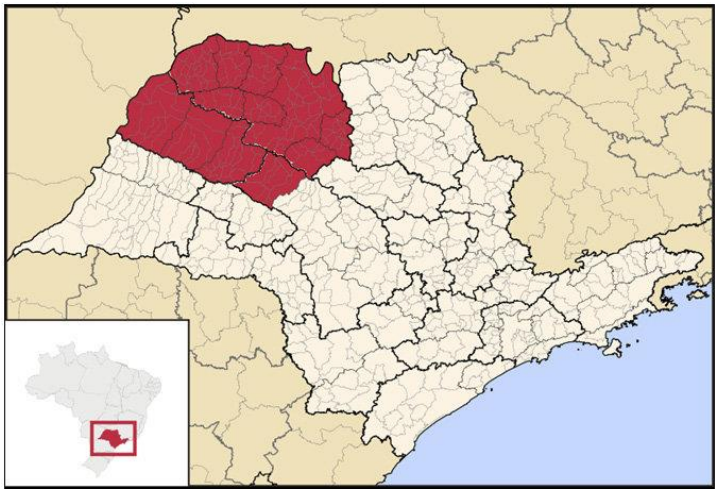


Figura 3: Mapa da região noroeste de São Paulo.

Fonte: Conceição; Tonietto (2012).

3.3.2 Dados do Laticínio

Os dados utilizados neste estudo foram coletados no laticínio no período de 01 de janeiro de 2021 a 31 de junho de 2021, sendo eles originados de planilha de controle de recebimento de leite a granel: planilhas que estavam impressas e preenchidas a mão pelos próprios transportadores durante a rota no mesmo momento da coleta; Planilha Boletim Diário da Plataforma; Planilha Análises de CPP dos Produtores; Planilha Análises de CPP dos Caminhões; Planilha de Localização das Propriedades e Planilha de Controle de Temperatura dos Caminhões.

Foram utilizadas, portanto, seis planilhas de controle diferentes, sendo que a planilha de controle de recebimento de leite a granel foi digitada em planilha de *Excel* versão 10. As outras cinco planilhas já estavam no formato Excel 10.

As variáveis de interesse de cada planilha foram agrupadas em uma única planilha. A partir da Planilha Controle de Recebimento a Granel foram extraídos os dados de data da coleta, código do produtor, volume coletados, hora da coleta, temperatura da coleta, compartimento (compartimento do caminhão transportador), responsável pela coleta. Na planilha CPP do produtor, foram obtidos os dados de CPP individual de cada produtor e a partir da planilha análises de CPP dos caminhões, extraiu-se o dado de CPP do compartimento. Na planilha de controle de temperatura dos caminhões foi obtido o dado de temperatura do leite no compartimento.

Uma nova variável chamada de tempo de rota foi criada considerando o horário de saída do transportador e o horário de chegada ao laticínio. Esta informação foi obtida na planilha boletim diário da plataforma.

A ordem de grandeza do tamanho da rota foi calculada com o uso de uma função de cálculo da distância usando uma métrica de distância geodésica de cada produtor até o destino final, que é a localização do laticínio, e foram somadas as distâncias dos produtores para cada rota. Foi necessário fazer isso para estimar o tamanho da rota, uma vez que não havia informação do tamanho de cada rota, mas apenas a localização dos produtores e do destino final.

3.3.3 Dados Meteorológicos

Os dados meteorológicos (temperatura, umidade e precipitação) foram coletados no banco de dados do INMET no site https://portal.inmet.gov.br/, aba dados meteorológicos, banco de dados meteorológicos, selecionado o período de 01 de janeiro de 2021 a 31 de dezembro de 2021.

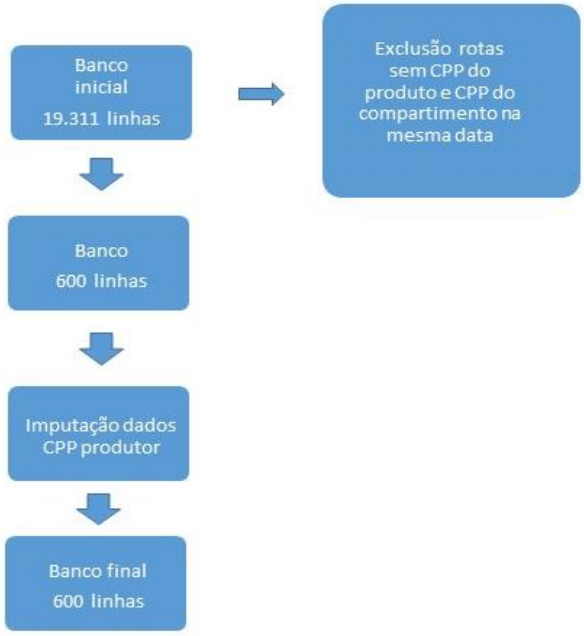
A estação meteorológica mais próxima da região de estudo é a de Votuporanga-SP e foram coletados dados de temperatura média compensada, precipitação total e umidade relativa do ar, utilizamos os dados das estações automáticas (A729) e convencionais (83623).

3.3.4 Limpeza dos dados

Após o agrupamento dos dados a serem utilizados na pesquisa selecionaram-se 18 variáveis (data, código do produtor, número de produtores do compartimento, volume médio por produtor/compartimento, volume individual, temperatura individual, temperatura média do compartimento, código do responsável pela coleta, CPP do produtor individual, CPP média do compartimento, umidade, temperatura, precipitação, nome da rota, CPP do compartimento, temperatura do compartimento, tempo de rota e tamanho da rota) que, em tese, seriam interessantes para o estudo.

Após agrupamento das variáveis, o total utilizado foi de 19.311 análises. Uma segunda seleção foi realizada neste banco e foram excluídas amostras que não continham análises de CPP ou que a coleta da CPP do produtor e CPP do compartimento foram realizadas em datas diferentes. A partir daí selecionou-se uma planilha com 600 amostras.

Das 600 amostras selecionadas, cerca de 36 análises não tinham o dado do CPP do leite do produtor, sendo necessário então utilizar o método de imputação. A variável CPP do leite do produtor foi utilizada para imputação e todas as outras variáveis foram utilizadas no modelo como preditoras. Foram excluídas as variáveis data e código do produtor.



Ao todo foram selecionadas 16 variáveis (número de produtores do compartimento, volume médio por produtor/compartimento, volume individual, temperatura individual, temperatura média do compartimento, código do responsável pela coleta, CPP do produtor individual, CPP média do compartimento, umidade, temperatura, precipitação, nome da rota, CPP do compartimento, Temperatura do compartimento, tempo de rota e tamanho da rota) e 600 amostras.

Para se fazer a imputação dos dados faltantes, como pode ser observado na figura 4, foi necessário testar se o método escolhido era realmente capaz de estimar dados faltantes de forma coerente. Foram excluídas as 36 análises contendo dados de CPP do leite do produtor faltante. No banco sem falhas (564 amostras), geraram-se 5% e 10% de falhas aleatórias (ou seja, excluímos de forma aleatória) na variável CPP do leite do produtor (Fig. 4).

Fonte: Elaborado pelos autores (2022)

Figura 4: Fluxograma da formação dos bancos de dados utilizados para análise de desempenho do algoritmo de imputação múltipla MICE para dados faltantes de CPP de produtor.

As métricas de avaliação foram: o coeficiente de correlação (r), MAE e R^2. Para análise de desempenho do algoritmo MICE, comparando-se os dados imputados com dados observados.

Utilizou-se a metodologia de imputação múltipla de dados com validação cruzada. Os dados foram divididos em cinco partes. Foram apresentados 80% dos dados e 20% foram omitidos.

Realizou-se a imputação cinco vezes, ou seja, cinco soluções para aquele problema e em cada vez, foi realizado o teste com uma parte dos dados, impedindo o vazamento e viés dos mesmos.

A técnicas de balanceamento possibilitaram a reamostragem de classes e atributos, assegurando que não haja um super ajuste dos dados também chamado de *overfitting*. Dados brutos raramente vem na forma necessária para o desempenho ideal do algoritmo de *Machine Learning*. Assim, realizou-se a padronização dos dados.

Optou-se por utilizar o método de imputação múltipla de dados utilizando o algoritmo MICE version 3.14.0. Este algoritmo realiza uma imputação múltipla usando especificação totalmente condicional (FCS) ou seja ele ajusta a variável faltante a partir de outras variáveis do modelo como preditoras (Van Buuren; Groothuis-Oudshoorn, 2011).

Os dados foram organizados em planilha de Excel e analisados em software R versão 4.2.1 de 2022 (R Core Team. R: https://www.R-project.org/).

Os códigos em R, elaborados para este estudo, estão em um repositório no Github e podem ser acessados pelo link: https://github.com/Marianafranko/scriptcppleite/tree/main.

3.3.5 Aspectos éticos

Os dados foram cedidos por um laticínio da região Noroeste do estado de São Paulo, e um acordo de confidencialidade foi firmado entre os envolvidos na pesquisa e a indústria, resguardando a todos os envolvidos (laticínio, técnicos, transportadores, proprietários rurais e funcionários) a confidencialidade dos dados ou qualquer tipo de variável que possa identificar qualquer pessoa ou empresa envolvida.

3.4 Resultados e discussão

Observou-se que quanto menor o percentual de falhas no banco, melhor o desempenho do algoritmo na imputação (Tab. 2).

Tabela 2: Resultado das métricas de avaliação da performance do algoritmo MICE para predição de dados de CPP do produtor

Métricas	5% de falha	10% de falha
R2	1.00	0.92
MAE	4.73	16.93
r	1.00	0.96

d	1.00	0.98

Fonte: Elaborado pelos autores (2022)

O R^2, no caso do resultado com 5% de falha, foi igual a 1, indicando que 100% da variância dos dados podem ser explicados pelo modelo. Com 10% de falha, observou-se R2 =0,98 indicando que 98% da variância dos dados podem ser explicados pelo modelo.

A métrica MAE (Erro Médio Absoluto), indica o quanto o modelo erra, ou seja, a regressão observada para contagem bacteriana erra em média 4,73 para 5% de falha e 16,93 para 10% de falha.

No coeficiente de correlação de Pearson (r), o valor ideal é r = 1 e o valor encontrado para 5% se falha foi 1 e para 10% se falha foi 0,96, o que significa que há uma alta correlação entre o valor previsto e o valor realizado.

O índice de acurácia de Willmott (d), desenvolvido na década de 80 por Cort J. Willmott (Willmott *et al.*,1985), demonstra a concordância entre o valor predito e o valor observado. Esse índice (d) varia de 0 a 1, sendo d = 0, uma total discordância e d =1 indica uma perfeita concordância entre os valores. No presente estudo foi observado um resultado de d = 1 para 5% de falha e de d = 0,98 para 10% de falha.

Vários gráficos de diagnóstico estão disponíveis para inspecionar a qualidade das imputações e neste estudo utilizou-se o pacote ggboost (https://ggplot2.tidyverse.org/reference/ggplot.html) para visualizar o resultado das imputações e os valores reais. Na figura 5 podem ser verificados os resultados para 5% de falhas.

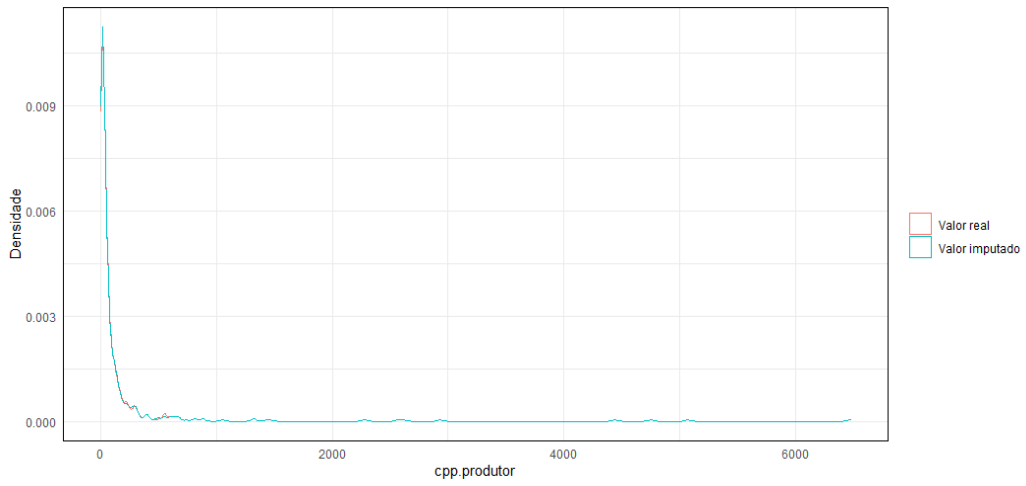


Figura 5: Gráfico representando valor real e valor imputado para 5% de falha no banco de dados.

Fonte: Elaborado pelos autores (2022)

Na figura 6 pode-se ver os resultados para 10% de falhas.

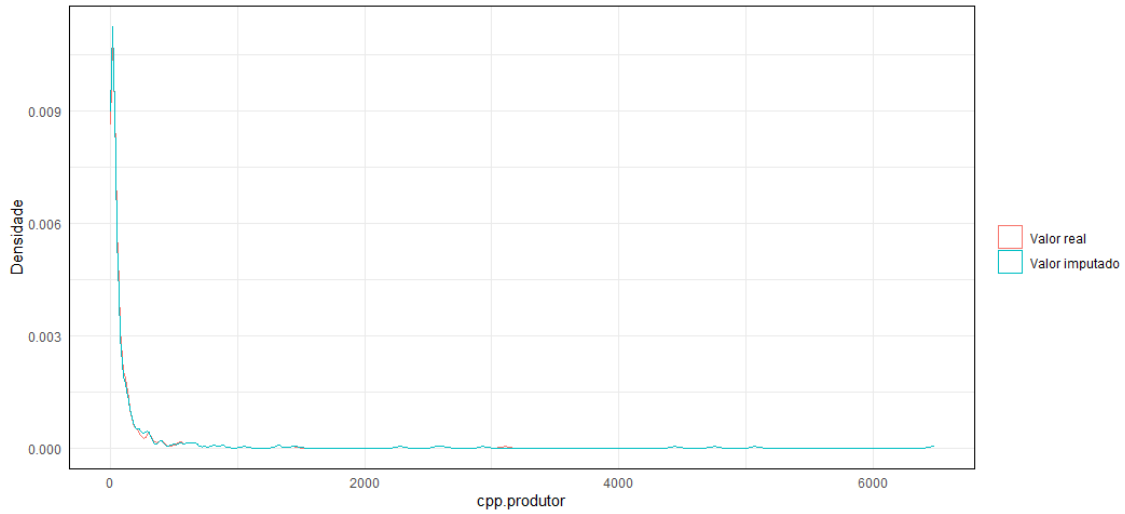


Figura 6: Gráfico representando valor real e valor imputado para 10% de falha no banco de dados.

Fonte: Elaborado pelos autores (2022).

Nunes, Klück e Fachel (2009) observaram que o uso da imputação múltipla de dados nos casos de dados faltantes ao invés da utilização apenas dos dados completos em um banco de dados pode ser uma importante ferramenta, pois excluir dados que contém alguma variável incompleta pode provocar coeficientes menos fidedignos e as estimativas podem ser “viesadas” se grupos homogêneos de dados forem excluídos da análise. Portanto, uma justificativa para o uso da imputação de dados é que quando se tem perda de dados o poder estatístico diminui, pois diminui o tamanho da amostra. Os autores concluíram que imputar dados faltantes pode aumentar consideravelmente a confiabilidade dos resultados obtidos e que a utilização do modelo MICE foi eficiente na imputação destes dados.

O resultado da validação das imputações realizadas neste trabalho corrobora com os encontrados por Alves e Gomes (2020) que utilizaram a metodologia e concluíram que esta era satisfatória para preenchimento de dados de chuva faltantes.

Mello e Nicolette (2021) concluíram que a utilização do método MICE, assim como neste estudo, foi eficiente para imputação de dados faltantes e para melhoria da eficácia da previsão de dados futuros.

Para resultados obtidos pelo coeficiente de determinação múltiplo (R2) e de MAE, os modelos podem ser utilizados por possuírem boa capacidade preditiva.

3.5 Conclusão:

O modelo se mostrou eficiente para imputação de dados faltantes para CPP do leite do produtor rural, podendo ser utilizado como um aliado na política interna do laticínio para avaliação e controle da contagem bacteriana do produtor quando há perda de amostra e riscos de não conformidade e até mesmo de suspensão de coleta por não atendimento do limite legal. Portanto, na falta do resultado de CPP, o modelo pode estimar o risco de não conformidade e estabelecimento de ações corretivas antes de nova coleta programada de amostras.

A ferramentas pode contribuir para que se possa estabelecer solução para um problema de não conformidade junto ao produtor e um monitoramente eficiente da qualidade da nossa matéria prima.

Como a responsabilidade de resguardar a qualidade do leite junto ao produtor rural é de responsabilidade do laticínio, a busca por ferramentas que auxiliem no processo de vigilância da qualidade do leite é muito importante.

Ainda que em nível legal esses dados não possam ser utilizados como resultados de análises junto ao Ministério da Agricultura Pecuária e Abastecimento (MAPA), acredita-se que para o laticínio é de suma importância saber a realidade da qualidade microbiológica daquele produtor para que decisões assertivas de orientação técnica possam ser estabelecidas antes que este tenha que ser desligado do programa de fornecimento de leite por não atendimento do limite de CPP estabelecido pela legislação brasileira.

3.6 Referências bibliográficas

ALVES, L. E. R.; GOMES, H. B. Validação da imputação múltipla via Predictive Mean Matching para preenchimento de falhas nos dados pluviométricos da Bacia do Médio São Francisco. *Anuário do Instituto de Geociências*, v. 43, n. 1, p. 199-206, 2020.

ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE LEITE LONGA VIDA (ABVL). Relatório Anual 2021. São Paulo, 2022.

BERRAR, Daniel. *Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology*, Volume 1, Elsevier, pp. 542-545, 2019.

BRASIL. Ministério da Agricultura Pecuária e Abastecimento. Decreto nº 10.468, de 18 de agosto de 2020. Altera o Decreto nº 9.013, de 29 de março de 2017, que regulamenta a Lei nº 1.283, de 18 de dezembro de 1950, e a Lei nº 7.889, de 23 de novembro de 1989, que dispõem sobre o regulamento da inspeção industrial e sanitária de produtos de origem animal, Brasília, DF, 2020a.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n° 76 de 26 de novembro 2018. Aprova o Regulamento Técnico de Produção, Identidade e Qualidade do Leite Cru Refrigerado, Leite Pasteurizado e o Leite pasteurizado tipo A. Diário Oficial da República Federativa do Brasil. Brasília. 2018.(a)

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n°. 77 de 26 de novembro 2018. Estabelece os critérios e procedimentos para a produção, acondicionamento, conservação, transporte, seleção e recepção do leite cru em estabelecimentos registrados no serviço de inspeção oficial, na forma desta Instrução Normativa e do seu Anexo Diário Oficial da República Federativa do Brasil. Brasília. 2018.(b)

BYLUND, G. *Dairy Processing Handbook*. Tetra Pak Processing Systems AB, 2003.

CONCEIÇÃO, M. A. F.; TONIETTO, J. Clima vitícola para a região de Jales (SP). Embrapa Uva e Vinho-Documentos (INFOTECA-E), 2012.

MELLO, K. Q.; DA FONTOURA, N., R. Imputação de dados faltantes em séries temporais de qualidade de água: o caso do Rio Doce. Anais do Salão Internacional de Ensino, Pesquisa e Extensão, v. 13, n. 3, 2021.

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G.. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. *Cadernos de Saúde Pública*, v. 25, p. 268-278, 2009.

PHILPOT, W.; NICKERSON, S. *Vencendo a Luta Contra a Mastite*. Publicado por Westfalia Surge Inc. e Westfalia Landtechnik do Brasil Ltda. Brasil. Milkbizz. Edição Brasileira, p. 6-9, 2002.

VALLIN V.M.; BELOTI V.; BATTAGLINI A.P.P.; TAMANINI R.; FAGNANI R.; ANGELA H.L.; SILVA L.C.C. Melhoria da qualidade do leite após implantação de boas práticas de fabricação em ordenha em 19 municípios da região central do Paraná. Semina: Ciências Agrárias, v.30, p.181-188, 2009. Disponível em: http://www.uel.br/revistas/uel/index.php/semagrarias/article/view/2661/ 2313 >.

VAN BUUREN, Stef; GROOTHUIS-OUDSHOORN, Karin. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, v. 45, p. 1-67, 2011.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

SCHAFFER, Cullen. Selecting a classification method by cross-validation. *Machine Learning*, v. 13, n. 1, p. 135-143, 1993.

WILLMOTT, C. J. *et al*. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research*, Hoboken, v. 90, n. C5, p. 8995–9005, set. 1985.

Capítulo 4

Comparação de modelos de *machine learning* para a previsão da contagem bacteriana do leite cru antes do seu processamento na indústria

4.1 Resumo

O leite é uma das *commodities* mais produzidas e relevantes por ser um produto com alto valor nutricional e baixo custo para o consumidor se comparado a outras fontes de nutrientes. Por este motivo pode ser um importante meio de cultura e de transmissão de doenças se não for cuidadosamente manipulador, estocado, transportado e processado. Objetivou-se avaliar e comparar a performance de três modelos de ML para previsão da CPP do leite cru que chega às plataformas do laticínio. Comparou-se três modelos de ML, o *Support Vector Machine* (SVM), XGBoosting e redes neurais *MultiLayer Perceptron* (MLP). Obtivemos um resultado com RMSE de 5,4887410; 4,7333138; 6,0639758 respectivamente e um MAPE de 0,63%; 0,06%; 0,92%, demonstrando que a rede XGBoosting foi a que apresentou menor erro, porém os três modelos são eficientes para prever a CPP do leite e são estatisticamente semelhantes.

4.2 Introdução

O leite é um alimento com alto valor nutricional e baixo custo para o consumidor quando comparado a outras fontes de nutrientes e uma das *commodities* mais produzidas e relevantes no âmbito global. Além do contexto nutricional, sua importância do ponto de vista econômico e social é indiscutível, sendo caracterizado, em suma, por uma produção familiar (Marangoni *et al*., 2018; Food…, 2020; International…, 2019).

O leite pode ser um importante meio de cultura e de transmissão de doenças se não for obtido higienicamente e estocado, transportado e processado em condições higiênico-sanitárias adequadas. Cada etapa pode apresentar riscos de contaminação por diferentes microrganismos que podem deteriorá-lo ou representar riscos à saúde do consumidor. Por sua importância também como fonte de nutrientes para o desenvolvimento de microrganismos e a dificuldade de estabelecer, em tempo real, a sua qualidade microbiológica, desenvolver e aplicar ferramentas capazes de prever a CPP do leite é de grande utilidade para o monitoramento e melhoria da qualidade do leite (Philpot; Nickerson, 2002; Bylund, 2003).

Condições de higiene, limpeza e desinfecção dos caminhões de transporte a granel e temperatura estão diretamente relacionadas a alterações e aumento da contagem padrão em placas do leite que recebido pela chegada indústria (Ferrari, 2018).

Estudos e gerenciamento de logística industrial se tornam cada dia mais importantes para a oferta de produtos com alta qualidade e melhor custo-benefício para o consumidor. Desta forma, reduzir as perdas decorrentes do transporte tem sido uma meta das empresas do setor para evitar que o leite piore a sua qualidade e apresente menor rendimento industrial (Dutra *et al.*, 2014; Ferrari, 2018).

Atualmente a IN 76 que estabelece os limites legais para o leite cru refrigerado, determina em tanque individual ou de uso comunitário deve apresentar médias geométricas trimestrais de Contagem Padrão em Placas de no máximo 300.000 UFC/mL (trezentas mil unidades formadoras de colônia por mililitro) e de Contagem de Células Somáticas de no máximo 500.000 CS/mL (quinhentas mil células por mililitro). O leite cru refrigerado deve apresentar limite máximo para Contagem Padrão em Placas de até 900.000 UFC/mL antes do seu processamento no estabelecimento beneficiador (Brasil, 2018b).

Ferramentas de *Machine learning* (ML) permitem que sistemas de computador aprendam padrões por meio dados de forma a realizar tarefas complexas, visando automatizar a criação de modelos analíticos e/ou estatísticos. (Deo, 2015; Vopham *et al.*, 2018).

Modelos como *Support Vector Machine* (SVM) (Vapnik, 1998); XGBoost (Chen e Guestrin, 2016); e redes neurais *MultiLayer Perceptron* (MLP) (Braga *et al*., 2000) são amplamente utilizados em diversas áreas da ciência para previsão de dados.

4.2.1 Modelos de ML

Três modelos de ML foram utilizados na previsão da CPP do leite do compartimento do caminhão de transporte que chega ao laticínio: *Support Vector Machine* (SVM), XGBoosting e redes neurais *MultiLayer Perceptron* (MLP) (Braga *et al*., 2000; Friedman, 2002; Vapnik, 1998).

Support Vector Machine (SVM) são ferramentas de classificação e regressão. São uma generalização do algoritmo *Generalaized Portrait* que foi desenvolvido por Vapnik, Lemer, e Chervonenkis na década de 60 (Vapnik, 1998).

SVM é um aproximador de função, ou seja, para problemas de encontrar a função y = f(x) dada por suas medidas Yi com ruído em alguns vetores (geralmente aleatórios) x, este método pode ser aplicado para reconhecimento de padrões (para estimar indicadores funções), para regressão (para estimar funções de valor real) e para resolver equações de operadores lineares.

O método SVM foi descoberto em 1964 para a construção de hiperplanos em problemas de reconhecimento de padrões, então, em 1992 -1995, foi generalizado para a construção de funções de separação não lineares (mas lineares em característica espaço). Em 1995 foi generalizado para estimação de funções reais. Por último, em 1996 foi aplicado para resolver equações de operadores lineares (Vapnik, 1998).

Modelos XGBoost constituem um algoritmo de aprendizagem em árvore de modelo linear. Ele suporta várias funções objetivas, incluindo regressão e classificação. O pacote é feito para ser extensível, para que os usuários também possam definir suas próprias funções objetivas facilmente (https://xgboost.readthedocs.io/en/stable/R-package/xgboostPresentation.html).

O modelo é baseado em árvores de regressão semelhantes que formulam hipóteses sobre os exemplos agregando as respostas de um comitê de preditores simples. O conjunto de árvores de regressão que compõem este comitê é elaborado de forma que um modelo aperfeiçoe o erro do modelo anterior (Friedman, 2002).

Modelo XGBoost diferem de outras técnicas de GB pois empregam um algoritmo sensível à dispersão na busca por ramificações, que torna a complexidade computacional do modelo linear ao número de observações não ausentes. Assim como integram otimizações do uso de recursos que permitem calcular paralelamente a aptidão dos atributos (Chen e Guestrin, 2016).

As redes neurais MultiLayer Perceptron (MLP) apresentam pelo menos uma camada intermediária, sendo assim capazes de resolver problemas não linearmente separáveis e permitindo a aproximação de qualquer função matemática (Braga *et al*., 2000)

O treinamento da rede MLP supervisionado utilizando backpropagation compõe duas etapas. Na primeira, um padrão é apresentado à camada de entrada e, a partir desta camada as unidades calculam sua resposta que é produzida na camada de saída, o erro é calculado e no segundo passo, o erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo modificados utilizando a regra delta generalizada (Braga *et al*., 2000).

O desafio do trabalho com RNA é definir o número de camadas intermediárias e o número de nodos da camada que são definidos em função da complexidade do problema e dos dados disponíveis. Uma vez definida a geometria da rede inicial um processo de refinamento sucessivo é realizado até que a estrutura final para o modelamento para que não ocorra overfitting e underfitting. (Braga *et al*., 2000).

O objetivo deste trabalho foi avaliar e comparar a performance de três modelos de ML para previsão da CPP do leite cru refrigerado, armazenado no compartimento do caminhão de transporte, antes do processamento no laticínio.

4.3 Material e Métodos

4.3.1 Delineamento do estudo

O estudo foi realizado em um laticínio localizado na região noroeste do estado de São Paulo, a região noroeste paulista possui cerca de 50.000 km^2 e é formada por cento e cinquenta e três municípios, distribuídos em doze microrregiões. O maior município dessa região é São José do Rio Preto (Conceição; Tonietto, 2012).

4.3.2 Dados Do Laticínio

Os dados utilizados neste estudo foram coletados no laticínio no período de 01 de janeiro de 2021 a 31 de junho de 2021, sendo eles planilha de controle de recebimento de leite a granel, as planilhas que estavam impressas e preenchidas a mão pelos próprios transportadores durante a rota no mesmo momento da coleta, “Planilha Boletim Diário da Plataforma”, “Planilha Análises de CPP dos Produtores”, “Planilha Análises de CPP dos Caminhões”, “Planilha de Localização das Propriedades” e “Planilha de Controle de Temperatura dos Caminhões”.

Foram obtidas, portanto, 6 planilhas de controle diferentes, sendo que a planilha de controle de recebimento de leite a granel foi digitada em planilha de excel versão 10, e as outras 5 planilhas já estavam no formato excel 10:

As variáveis de interesse de cada planilha foram agrupadas em uma única planilha A partir da Planilha Controle de Recebimento a Granel extraímos os dados: data da coleta, código do produtor, volume coletados, hora da coleta, temperatura da coleta, compartimento (compartimento do caminhão transportador), responsável pela coleta. Na planilha CPP do produtor extraímos os dados de CPP individual de cada produtor, a partir da planilha análises

de CPP dos caminhões extraímos o dado de CPP do compartimento. Na planilha de controle de temperatura dos caminhões obtivemos o dado de temperatura do compartimento.

Uma nova variável chamada de tempo de rota foi criada, a partir do horário de saída do transportador e o horário de chegada ao laticínio, esta informação foi obtida na "Planilha boletim diário da plataforma".

A ordem de grandeza do tamanho da rota foi calculada com o uso de uma função de cálculo da distância usando uma métrica de distância geodésica de cada produtor até o destino, que é a localização do laticínio, e foram somadas as distâncias dos produtores para cada rota. Precisou-se fazer isso para estimar o tamanho da rota, uma vez que não havia informação do tamanho de cada rota, mas apenas a localização dos produtores e do laticínio.

As análises estatísticas dos dados utilizados estão nas tabelas 5 e 6 no anexo I.

4.3.3 Dados Meteorológicos

Os dados meteorológicos (temperatura, umidade e precipitação) foram coletados no banco de dados do INMET no site https://portal.inmet.gov.br/, aba dados meteorológicos, banco de dados meteorológicos, selecionado o período de 01 de janeiro de 2021 a 31 de dezembro de 2021.

A estação meteorológica mais próxima da região de estudo é a de Votuporanga-SP e foram coletados dados de temperatura média compensada, precipitação total e umidade relativa do ar, utilizamos os dados das estações automáticas (A729) e convencionais (83623).

4.3.4 Limpeza dos dados

Após o agrupamento dos dados a serem utilizados na pesquisa foram selecionadas 18 variáveis (data, código do produtor, número de produtores do compartimento, volume médio por produtor/compartimento, volume individual, temperatura individual, temperatura média do compartimento, código do responsável pela coleta, CPP do produtor individual, CPP média do compartimento, umidade, temperatura, precipitação, nome da rota, CPP do compartimento, temperatura do compartimento, tempo de rota e tamanho da rota) que seriam interessantes para o estudo.

Após agrupamento das variáveis totalizaram 19.311 amostras. Uma segunda seleção foi realizada neste banco, foram excluídas amostras que não continham análises de CPP ou que a coleta da CPP do produtor e CPP do compartimento foram realizadas em datas diferentes. A partir daí selecionou-se uma planilha com 600 amostras.

Das 600 amostras selecionadas, cerca de 36 análises não tinham o dado do CPP do produtor, sendo necessário então utilizar o método de imputação. A variável CPP do produtor foi utilizada para imputação e todas as outras variáveis foram utilizadas no modelo como preditoras, foram excluídas as variáveis data e código do produtor.

Ao todo foram selecionadas 16 variáveis (número de produtores do compartimento, volume médio por produtor/compartimento, volume individual, temperatura individual, temperatura média do compartimento, código do responsável pela coleta, CPP do produtor individual, CPP média do compartimento, umidade, temperatura, precipitação, nome da rota, CPP do compartimento, temperatura do compartimento, tempo de rota e tamanho da rota) e 600 amostras.

Após a etapa de exclusão de amostras com dados faltantes nas quais não seria possível imputar dados, realizou-se a imputação dos dados faltantes da variável CPP do produtor em 6% (36/600) da amostra. Nesses casos foi realizada análise de CPP do produtor e do compartimento na mesma data, entretanto por motivo desconhecido algum resultado não foi registrado ou houve perda de amostra.

As 600 amostras foram distribuídas em subgrupos: treinamento (80%), teste (20%).

O método de imputação multivariada utilizado neste estudo foi o MICE (Multivariate Imputation by Chained Equations) que utiliza além da variável predita outras covariáveis para aumentar a performance do modelo (Alves; Gomes, 2020; Nunes; Klück; Fachel, 2009; Van Buuren; Groothuis-Oudshoorn, 2011).

Para exclusão das variáveis altamente correlacionadas e daquelas que não auxiliam no desempenho do modelo, utilizaram-se três métodos: matriz de correlação de Pearson (r), algoritmo Boruta e algoritmo *Recursive Feature Elimination* (RFE), ficando apenas as preditoras (Chen; Jeong, 2007; Snedecor; Cochran, 1971).

Os objetivos da seleção de variáveis são melhorar o desempenho de previsão dos preditores, fornecer preditores mais rápidos e econômicos e fornecer uma melhor compreensão do processo subjacente que gerou os dados (Guyon; Elisseeff, 2003; Chen; Jeorge, 2007).

Os dados foram organizados em planilha de Excel versão e analisados em software R versão 4.2.1 de 2022. (R CORE TEAM. R: https://www.R-project.org/).

Os códigos em R, elaborados para este estudo estão em um repositório no Github e podem ser acessados pelo link a seguir:

https://github.com/Marianafranko/scriptprevis-o/tree/main.

4.3.5 Etapa de treinamento e teste

Método de validação cruzada

Foi utilizado o método de validação cruzada, que em consiste dividir aleatoriamente o banco de dados original em k partes (k-*folds*) de tamanhos aproximadamente iguais, sendo separados em dados para treinamento e validação (Kuhn; Johnson, 2013).

A figura 7, representa o processo de validação cruzada k-*fold*, com k=5. Os dados de treinamento foram divididos em 5 partes, e em cada iteração, 4 (80%) partes foram utilizadas para o treinamento do modelo com diferentes hiperparâmetros, e uma (20%) para estimar sua performance preditiva. Os mesmos *folds* foram utilizados para todos os modelos de previsão para que eles possam ser comparados posteriormente (Fig. 7).

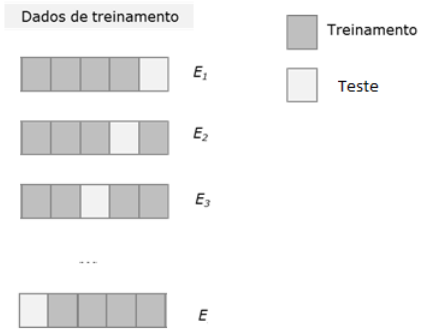


Figura 7: Processo de validação cruzada k-*fold*, com k=5.

Fonte: Adaptado de Raschka (2017)

Ao final do processo, as performances estimadas de cada modelo, Ek, são utilizadas para calcular sua performance média, as interações são realizadas até que todas as partes tenham participado tanto do treinamento como da validação do modelo (Kuhn; Johnson, 2013).

A métrica utilizada para a avaliação dos modelos foi a raiz quadrada do erro médio quadrático ou *root mean squared error* (RMSE) e *mean absolut percentage error* (MAPE) (Scikit-learn, 2021), que pode ser observado na figura 8, em que n é o número da posição total de amostra; N é o número total de elementos, da amostra; X0t é o valor observado na posição t; Xpt é o valor previsto na posição t (Scikit-learn, 2021) (Fig.8).

RMSE	Raiz do erro quadrado médio	$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(x_{o_t} - x_{p_t})^2}{N}}$
MAPE	Erro percentual absoluto médio	$MAPE = \frac{\sum_{t=1}^{n}\left\lvert\frac{(x_{o_t} - x_{p_t})}{x_{o_t}}\right\rvert}{N} \times 100$

Figura 8: Métricas para a avaliação dos modelos RMSE, MAPE e suas respectivas fórmulas.

Fonte: Adaptado Scikit-learn (2021)

4.3.6 Aspectos éticos

Os dados foram cedidos por um laticínio da região Noroeste do estado de São Paulo, e um acordo de confidencialidade foi firmado entre os envolvidos na pesquisa e a indústria, resguardando a todos os envolvidos (laticínio, técnicos, transportadores, proprietários rurais e funcionários) a confidencialidade dos dados ou qualquer tipo de variável que possa identificar qualquer pessoa ou empresa envolvida.

4.4 Resultados e Discussão

4.4.1 Seleção de variáveis

Na fase de pré-processamento dos dados foram realizados testes para a seleção de variáveis preditoras utilizadas nos modelos. Observou-se no teste de Correlação de Pearson que umidade e precipitação apresentaram alta correlação ($r \geq 0{,}80$), indicando possível redundância entre as mesmas (Fig.9).

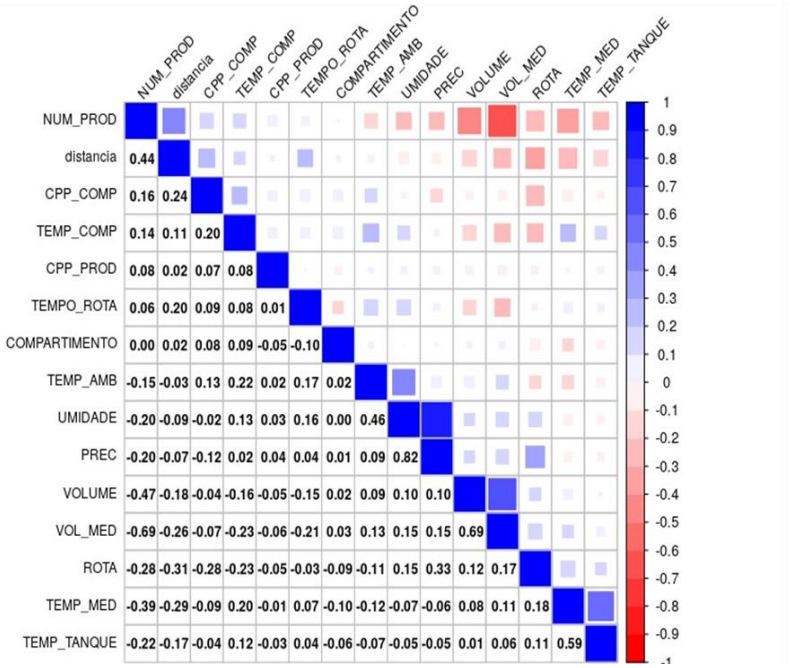


Figura 9: Matriz de correlação de Pearson (r) para seleção de variáveis, a serem utilizadas na previsão da CPP final do leite entre janeiro de 2021 e junho de 2021.

Fonte: Elaborado pelos autores (2022)

Diante da ocorrência de multicolinearidade, optou-se por retirar a variável umidade por se tratar de um problema que envolve logística de transporte de leite.

O algoritmo *Recursive Feature Elimination* (RFE) foi utilizado para selecionar variáveis de interesse para os modelos de previsão. O RFE é um *wrapper methods* que determina quais variáveis mais influenciam no seu modelo e que é importante para evitar problemas de *overfitting* e melhorar o desempenho da previsão (Chen; Jeong, 2007) (Fig. 10).

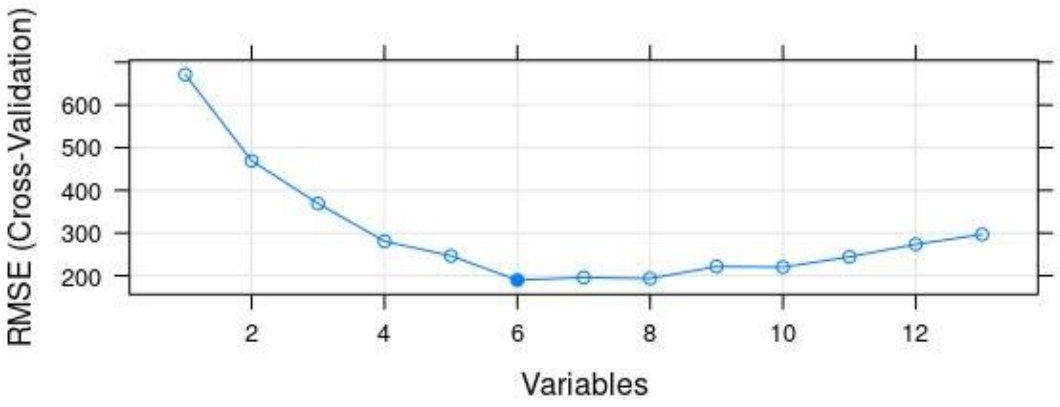


Figura 10: *Recursive Feature Elimination* (RFE) para seleção de variáveis, para previsão da CPP final do leite.

Fonte: Elaborado pelos autores (2022)

Observou-se pelo resultado do RFE que nenhuma das variáveis de entrada selecionadas poderia ser descartada, apresentando, portanto, relevância para previsão da CPP final do leite que chega ao laticínio. Dentre as variáveis selecionadas pelo RFE, seis variáveis que mais influenciaram na CPP do leite do compartimento dos caminhões foram: temperatura do leite no compartimento, número de produtores que forneceram leite por compartimento, rota, tempo de rota, volume médio e temperatura média de entrada do leite no compartimento do caminhão.

Utilizou-se o algoritmo Boruta para seleção das variáveis de entrada ou preditoras em que não houve exclusão, havendo, portanto, relevância para todas as variáveis apresentadas (Fig. 11).

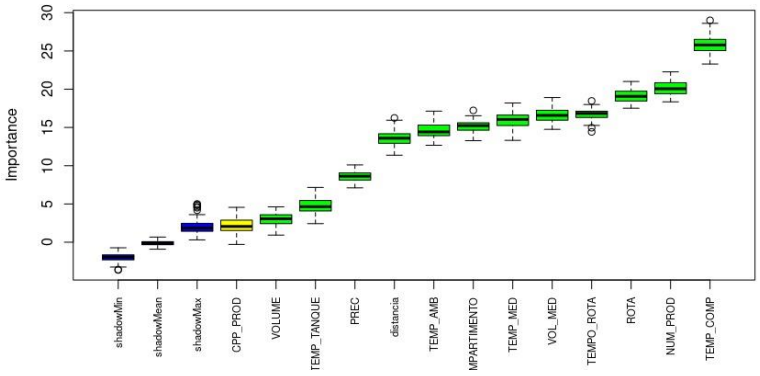


Figura 11: Algoritmo Boruta aplicado aos dados para seleção de variáveis, para previsão da CPP final do leite.

Fonte: Elaborado pelos autores (2022)

Os *boxplots* azuis correspondem à pontuação Z mínima, média e máxima de um atributo de sombra. Os *boxplots* amarelos e verdes representam as pontuações Z de atributos provisórios e confirmados, respectivamente. Não foi rejeitada nenhuma variável, mas caso isso tivesse acontecido o *Boxploxs* estaria em vermelho.

Determinou-se que todas as variáveis seriam utilizadas no modelo, exceto a variável “umidade” que havia sido removida anteriormente após o teste de Correlação de Pearson.

A variável CPP produtor ficou como provisória. No entanto, decidiu-se manter esta variável em função de trabalhos publicados na área que demonstram a importância da uma baixa contagem

bacteriana para manutenção da qualidade do leite. Neste sentido, quanto maior a contagem bacteriana inicial, maior é a taxa de crescimento bacteriano e a CPP final do leite (Philpot; Nickerson, 2002; Bylund, 2003; Cruz *et al*., 2019).

4.4.2 Previsão da CPP do leite do compartimento dos caminhões

Rede neural *multilayer perceptron*

Na figura 12 verificam-se os resultados das previsões da rede neural *multilayer perceptron* para cada um dos cinco *folds* e suas respectivas frequências de erros representadas na forma de histograma (Fig.12).

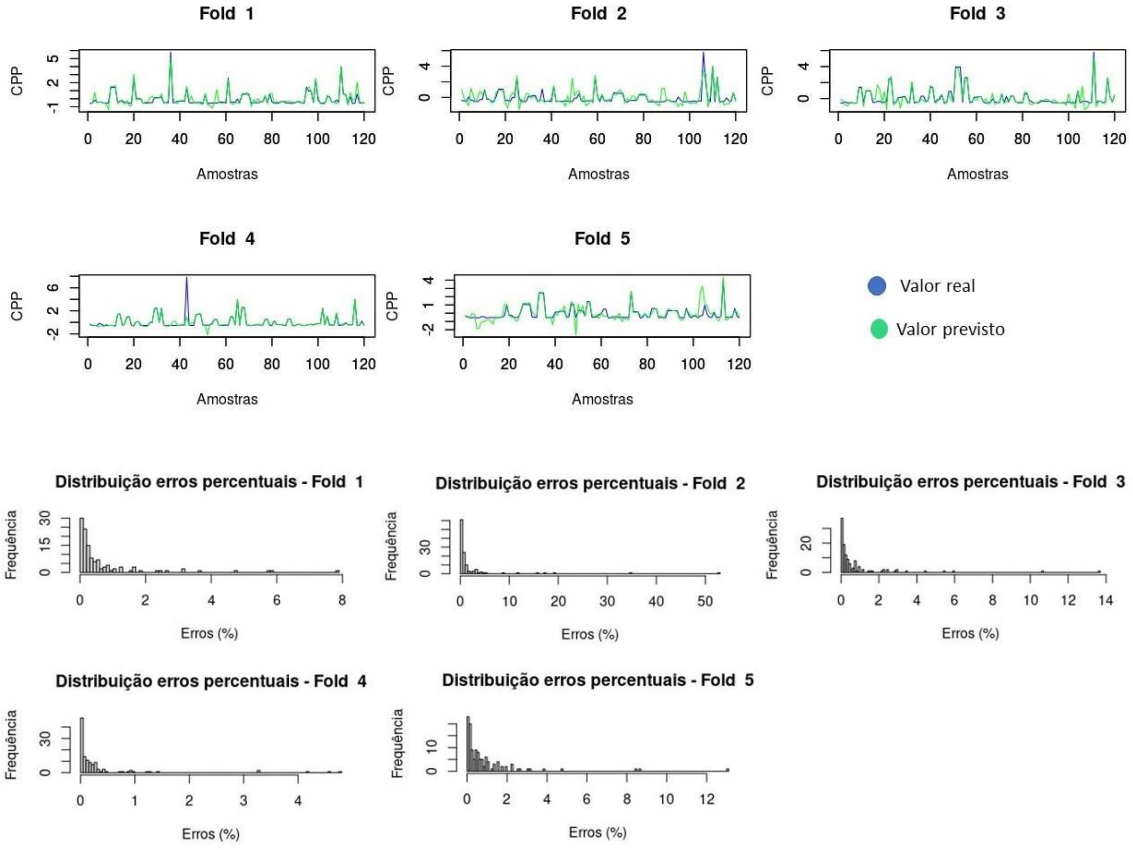


Figura 12: Representação gráfica dos resultados das previsões de CPP do leite dos compartimentos dos caminhões, no modelo MLP para cada *fold* e respectivos histogramas de erros RMSE (%).

Fonte: Elaborado pelos autores (2022)

Observou-se no *fold* 4 que a maior parte da frequência de erros ocorreu abaixo de 1% de erro. Nos *folds* 1, 3 e 5 a maior parte da frequência de erros ocorreu abaixo de 4% enquanto no *fold*

2 a maior parte da frequência de erros ocorreu abaixo de 10% de erro. Resumindo, em todos os *folds* os erros que aconteceram em sua maioria se concentraram abaixo de 4%, exceto no *fold* 2 que ocorreram erros abaixo de 10% do valor real.

Ventura *et al*. (2007) observaram grande versatilidade da rede MLP em relação ao desempenho operacional e ao tempo de processamento e vantagem em relação às outras metodologias na previsão.

Modelo SVM

Na figura 13 podemos observar os resultados das previsões do modelo SVM para cada um dos cinco *folds* e suas respectivas frequências de erros representadas na forma de histograma (Fig.13).

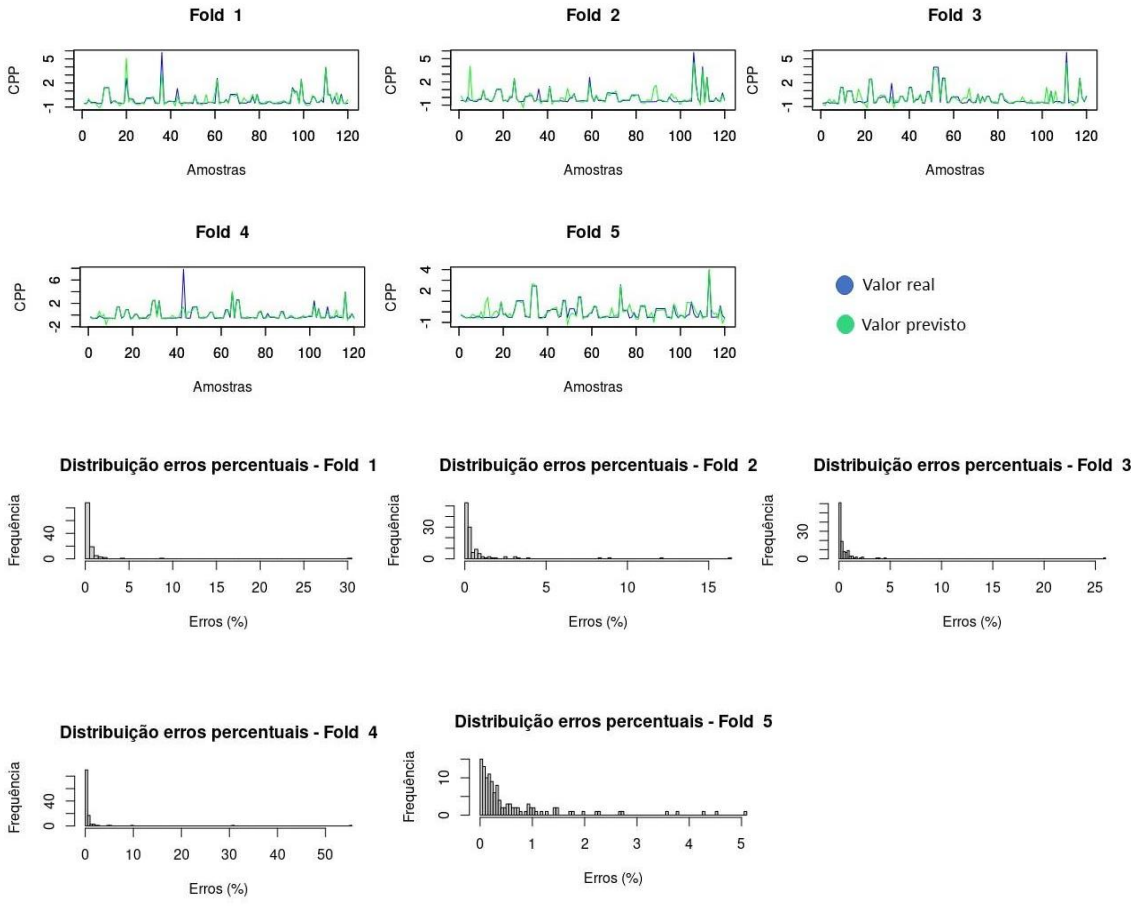


Figura 13: Representação gráfica dos resultados das previsões de CPP do leite dos compartimentos dos caminhões no modelo SVM para cada *fold* e respectivos histogramas de erros RMSE (%).

Fonte: Elaborado pelos autores (2022)

Na rede SVM a frequência de erro em todos os *folds* foi abaixo de 5%. Ferrão *et al*. (2007) observaram com este modelo, uma alta capacidade de previsão dos dados mesmo na ausência de outros dados, e que a rede SMV apresentou alta capacidade de generalização e flexibilidade.

XGboosting

Na figura 14 podem ser observados os resultados das previsões do modelo *XGboosting* para cada um dos cinco *folds* e suas respectivas frequências de erros representadas na forma de histograma (Fig.14)

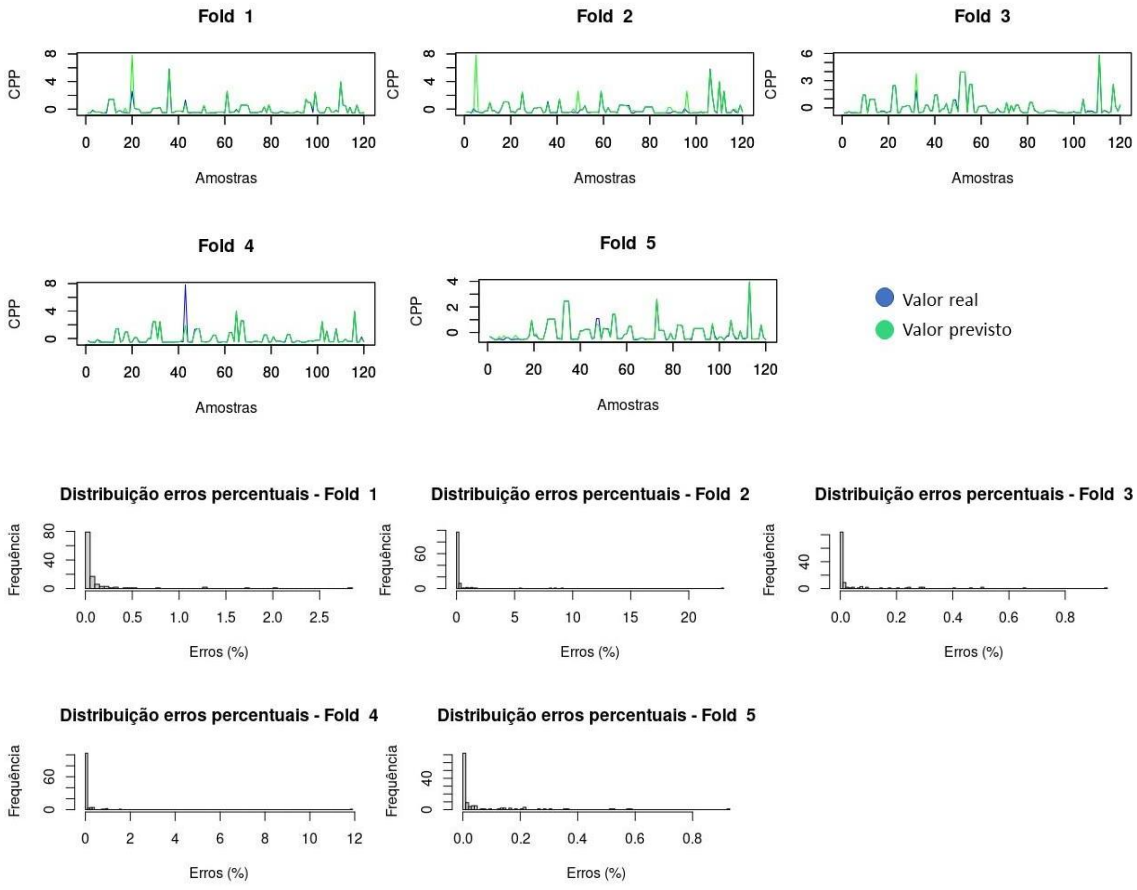


Figura 14: Representação gráfica dos resultados das previsões de CPP do leite do compartimento, no modelo *XGboosting* para cada *fold* e respectivos histogramas de erros RMSE (%).

Fonte: Elaborado pelos autores (2022).

No modelo *XGboosting*, a distribuição da frequência de erros no *fold* 1 se concentrou abaixo de 2,5%; no *fold* 2 abaixo de 5%; no *fold* 3 abaixo de 0,8%; no *fold* 4 abaixo de 2% e no *fold* 5 abaixo de 0,6%.

4.4.3 Métricas de avaliação dos modelos

Na tabela 1 pode-se observar um comparativo dos resultados do RMSE dos três modelos. O erro do modelo *XGboosting* foi menor comparado aos outros modelos (Tab. 3).

Tabela 3: Métrica de RMSE para os modelos utilizados

Fold	Resultados de RMSE dos Modelos		
	MLP	SVM	*XGboosting*
1	3,925773	4,738330	5,3348846
2	7,371892	6,290227	9,4170137
3	5,204014	4,759422	1,9780577
4	7,303085	7,407580	5,9996462
5	6,515115	4,248146	0,9369666
Média	6,0639758	5,4887410	4,7333138

Fonte: Elaborado pelos autores (2022)

Apesar do erro do modelo *XGboosting* ser menor comparado aos outros modelos, pelo teste de Wilcoxon, demonstrou-se como pode ser visto na figura 15 que os três modelos de previsão são estatisticamente equivalentes (Fig. 15).

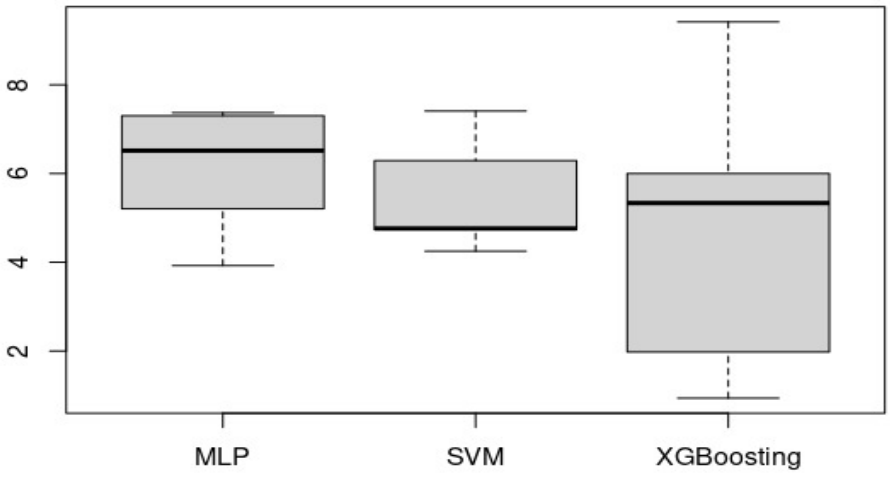


Figura 15: *Boxplot* teste de Wilcoxon comparartivo de erros dos modelos MLP, SVM e XGboosting.

Fonte: Elaborado pelos autores (2022)

Na tabela 4 podem-se avaliar a performance dos modelos utilizados. No modelo de rede neural MLP observa-se um percentual de erro na predição de 0,987% e 99,013 acerto. No modelo SVM, nota-se 0,831% de erro e 99,169% de acerto e no modelo *XGboosting*, 0,199% de erro e 99,801% de acerto. Estes resultados demonstram um alto percentual de acerto nos três modelos (Tab. 4).

Tabela 4: Métrica de MAPE para os modelos

Fold	Resultados de MAPE dos Modelos		
	MLP	SVM	*XGboosting*
1	0,681	0,740	0,133
2	2,176	0,843	0,561
3	0,816	0,711	0,056
4	0,344	1,228	0,179
5	0,920	0,631	0,064
Média MAPE	0,987	0,831	0,199
Média Acerto	99,013	99,169	99,801

Fonte: Elaborado pelos autores (2022)

Na busca pelas variáveis que mais influenciaram o modelo de rede neural MLP utilizou-se o método *Shapley* (*Shapley Additive exPlanations*) que tem como objetivo, explicar como foi realizada a previsão. O Método *Shapley* indica a importância aditiva de cada variável no modelo desenvolvido, seu artigo original data de 1953 (Algaba; Fragnelli; Sánchez-Soriano, 2019).

O *Shapley* foi realizado no *fold* 5, aquele com menor RMSE médio, o que indica melhor representatividade do problema. Na figura 16 pode-se observar o resultado do pacote *Shapley* para a rede neural MLP (Fig. 16).

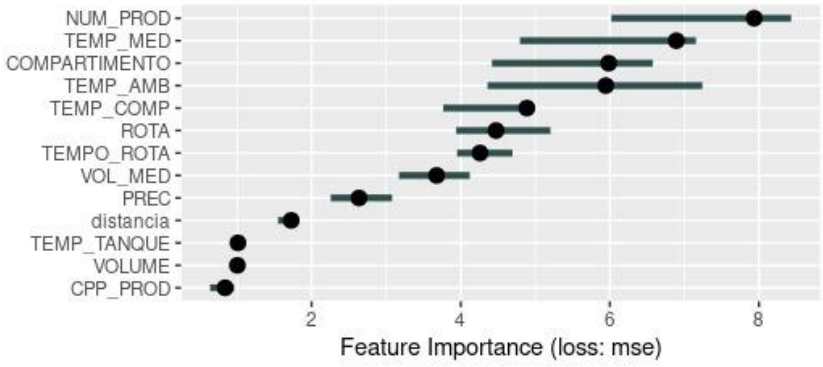


Figura 16: Resultado do pacote *shapley* para a rede neural MLP.

Fonte: Elaborado pelos autores (2022).

Neste resultado do Shapley pode-se observar que as variáveis mais importantes para o modelo são número de produtores no compartimento (NUM_PROD), temperatura média de entrada do leite no compartimento do caminhão (TEMP_MED), compartimento (COMPARTIMENTO) e temperatura ambiente (TEMP_AMB).

Verfica-se que a primeira variável que mais influenciou na previsão foi o número de produtores no compartimento. Isso acontece, pois, quanto maior o número de produtores maior será a variação da temperatura do compartimento e por se tratar de um caminhão isotérmico, ele chega ao primeiro produtor vazio e quente. Dessa forma, com maior número de produtores e menor volume de leite por propriedade fica mais difícil manter a temperatura, já que o caminhão não refrigera o leite.

A segunda variável foi exatamente a temperatura média de entrada do leite no compartimento, corroborando com Silva *et al.* (2009) que observou temperatura do leite e o tanque isotérmico os maiores responsáveis pelo aumento da contagem bacteriana total.

Outro resultado importante e inesperado foi que a CPP do produtor (CPP_PROD) assim como no resultado do RFE foi a variável considerada menos importante para a predição.

A temperatura ambiente é uma variável muito importante principalmente devido à sua influência na temperatura do leite. Pelo fato de o Brasil ser um país tropical em que as altas temperaturas são comuns, a qualidade do leite está relacionada com o grau de contaminação inicial e com o binômio tempo/ temperatura, em que o leite permanece desde a ordenha até o processamento (Silveira *et al*., 2000).

Recentemente diversas tecnologias de *Machine Learning* vêm sendo utilizadas nos mais diversos campos e inclusive na indústria de alimentos para auxiliar na obtenção de um produto de melhor qualidade. Lima (2021) e Valente *et al.* (2014) avaliaram a utilização de *Machine Learning* para detecção de fraude no leite e Ferrão *et al.* (2007) utilizaram ferramentas de ML para quantificação de adulterantes no leite em pó.

4.5 Conclusão

Os resultados deste estudo estão restritos a esta amostra e apresentam como principal limitação, o tamanho da amostragem. No entanto, os modelos utilizados apresentaram bons resultados de predição para CPP do leite antes do processamento.

Modelos preditivos têm sido amplamente utilizados na área da saúde e esta tem sido nossa expectativa com este estudo, ou seja, contribuir para que mais modelos possam ser aperfeiçoados para que a indústria de alimentos possa utilizá-los em prol da melhoria de sua qualidade .

A boa performance dos modelos abre muitas possibilidades de estudos adicionais para uma melhor aplicação e desempenho no setor produtivo de leite.

Por meio da previsão da qualidade do leite em rotas de coleta a granel pode-se identificar as rotas mais problemáticas do ponto de vista microbiológico e propor ações focadas em melhoria da qualidade do leite nas diferentes etapas (produção primária e transporte).

4.6 Referências Bibliográficas

ALGABA, E.; FRAGNELLI, V.; SÁNCHEZ-SORIANO, J. (Ed.). *Handbook of the Shapley value*. CRC Press, 2019.

ALVES, L. E. R.; GOMES, H. B. Validação da imputação múltipla via Predictive Mean Matching para preenchimento de falhas nos dados pluviométricos da Bacia do Médio São Francisco. Anuário do Instituto de Geociências, v. 43, n. 1, p. 199-206, 2020.

BREIMAN, L. et al. Cart. Classification and Regression Trees, 1984.

BERRAR, D.. Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology*, Volume 1, Elsevier, pp. 542-545, 2019.

BRAGA, A. P., LURDEMIR, T. B., AND CARVALHO, A. C. P. L. F. *Redes Neurais Artificiais: Teoria e Aplicações*. LTC - Livros Técnicos e Científicos Editora S.A. Belo Horizonte, Brasil, 2000.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n°. 76 de 26 de novembro 2018. Aprova o Regulamento Técnico de Produção, Identidade e Qualidade do Leite Cru Refrigerado, Leite Pasteurizado e o Leite pasteurizado tipo A. Diário Oficial da República Federativa do Brasil. Brasília. 2018ª.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n°. 77 de 26 de novembro 2018. Estabelece os critérios e procedimentos para a produção, acondicionamento, conservação, transporte, seleção e recepção do leite cru em estabelecimentos registrados no serviço de inspeção oficial, na forma desta Instrução Normativa e do seu Anexo Diário Oficial da República Federativa do Brasil. Brasília. 2018b.

BYLUND, G. *Dairy Processing Handbook*. Tetra Pak Processing Systems AB, 2003.

CHEN, X.; JEONG, J. C. Enhanced recursive feature elimination. In: *Sixth International Conference on Machine Learning and Applications* (ICMLA 2007). IEEE, 2007. p. 429-435.

CHEN, T.; GUESTRIN, C. X. A scalable tree boosting system. In: ACM. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. p. 785–794, 2016.

CONCEIÇÃO, M. A. F.; TONIETTO, J. Clima vitícola para a região de Jales (SP). Embrapa Uva e Vinho-Documentos (INFOTECA-E), 2012.

CRUZ, Adriano Gomes da et al. Microbiologia, higiene e controle de qualidade no processamento de leites e derivados. 1 edição, Rio de Janeiro: Elieser, 2019.

DA SILVA, Carlos Arthur B.; DA SILVA, Elaine Berges. Previsão da recepção de leite em usinas laticinistas: uma aplicação de redes neurais. *Revista de Economia e Sociologia Rural*, v. 33, n. 1, p. 89-97, 1995.

DEO, Rahul C. Machine learning in medicine. *Circulation*, v. 132, n. 20, p. 1920-1930, 2015.

DUTRA, A. et al. Sistema logístico do transporte de leite a granel: um estudo de caso. CEP, v. 95070, p. 560, 2014.

FERRÃO, M. F. et al. LS-SVM: uma nova ferramenta quimiométrica para regressão multivariada. Comparação de modelos de regressão LS-SVM e PLS na quantificação de adulterantes em leite em pó empregando NIR. *Química Nova*, v. 30, p. 852-859, 2007.

FERRARI, V. A. S. Transporte de leite a granel e sua influência na qualidade do leite que chega à indústria. Dissertação (Mestrado) – Universidade Estadual de Santa Cruz, UESC. Programa de Pós-graduação em Ciência Animal. Ilhéus-BA, 2017.

2018.

FRIEDMAN, J. H. *Stochastic gradient boosting. Computational Statistics & Data Analysis,* Elsevier, v. 38, n. 4, p. 367–378, 2002.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, v. 3, n. Mar, p. 1157-1182, 2003.

HAN, J.; KAMBER, M. *Data mining: concepts and techniques*, 2nd. University of Illinois at Urbana Champaign: Morgan Kaufmann, 2006.

INTERNATIONAL DAIRY FEDERATION. The Global Dairy Sector: Facts 2019. Disponível em: https://www.fil-idf.org/wp-content/uploads/2021/01/DDOR-Global-Dairy-Facts2019.pdf. Acesso em: 08 de abril de 2021.

KUHN, Max. Building predictive models in R using the caret package. Journal of statistical software, v. 28, p. 1-26, 2008.

KUHN, M.; JOHNSON, K. *Applied Predictive Modeling*. New York: Springer, 2013.

KURSA, M. B.; RUDNICKI, W. R. Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1–13. 2010. https://doi.org/10.18637/jss.v036.i11.

LIMA, J. S. et al. Espectrofotometria FTIR (Fourier Transform Infrared) e técnicas de aprendizado de máquina para a detecção de fraude por adição de soro de queijo ao leite cru. 2021. Tese (Doutorado Ciência Animal)- Escola de veterinária, Universidade Federal de Minas Gerais (UFMG). Minas Gerais, 2021.

MARANGONI, F. et al. Cow's milk consumption and health: A health professional's guide. *Journal of the American College of Nutrition*, v. 38 n. 3, p. 197-208, 2019.

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. *Cadernos de Saúde Pública*, v. 25, p. 268-278, 2009.

PHILPOT, W.; NICKERSON, S. *Vencendo a Luta Contra a Mastite*. Publicado por Westfalia Surge Inc. e Westfalia Landtechnik do Brasil Ltda. Brasil. Milkbizz. Edição Brasileira, p. 6-9, 2002.

RASCHKA, S. *Python Machine Learning*. 2. ed. Birmingham: Packt Publishing Ltd, 2017.

SCIKIT-LEARN DEVELOPERS. Regression metrics. Disponível em: <https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics>. Acesso em: 2 de setembro de 2021.

SILVA, M. A. P. da et al. Influência do transporte a granel na qualidade do leite cru refrigerado. Revista do Instituto Adolfo Lutz, São Paulo, v. 68, n. 3, p. 381-387, jul./set. 2009.

SILVEIRA, I. A.; CARVALHO, E. P.; TEIXEIRA, D. Influência de microrganismos psicrotrófcos sobre a qualidade do leite cru refrigerado. Uma revisão. *Higiene Alimentar*, 12(55): 21–7, 2000.

SNEDECOR, G. W.; COCHRAN, W. G. *Métodos Estadísticos*. Cía. Ed. Continental SA, Mexico, 1971.

VALENTE, G. F. S. et al. Aplicação de redes neurais artificiais como teste de detecção de fraude de leite por adição de soro de queijo. *Revista do Instituto de Laticínios Cândido Tostes*, v. 69, n. 6, p. 425-432, 2014.

VAN BUUREN, S.; GROOTHUIS-OUDSHOORN, K. Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, v. 45, p. 1-67, 2011.

VAPNIK, V. The support vector method of function estimation. In: *Nonlinear modeling*. Springer, Boston, MA, p. 55-85, 1988.

VENTURA, R. V. et al. Uso de redes neurais artificais na predição de valores genéticos para peso aos 205 dias em bovinos da raça Tabapuã. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, v. 64, p. 411-418, 2012.

VOPHAM, T.; HART, J. E.; LADEN, F. et al. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. *Environ Health*, 17(1):40, doi: 10.1186/s12940-018-0386-x, 2018.

Anexo I

Tabela 5: Estatística dos dados completos, contendo 19.331 linhas.

	VOLUME	TEMP. COLETA	CPP PROD.	CPP COMPART.	TEMP. AMB.	PRECIP.
MÍNIMO	2,0	0,5	1	6	10,32	0
1° QUADRANTE	193,0	3,1	15	60	22,86	0
MEDIANA	354,0	3,4	36	144	24,42	0
MÉDIA	664,5	3,5	167	526,6	24,17	2,08
3° QUADRANTE	690,0	3,9	91	674,5	26,08	0
MÁXIMO	7132,0	18	9999	7241,0	28,56	57

Tabela 6: Estatística dos após etapa de pré-processamento dos dados, contendo 600 linhas.

	VOLUME	TEMP. COLETA	CPP PROD.	CPP COMPART.	TEMP. AMB.	PRECIP.
MÍNIMO	32	1,9	2	6	22,6	0
1° QUADRANTE	185,5	3,1	14	54	22,92	0
MEDIANA	349,5	3,5	35	141,5	24,38	0
MÉDIA	554,8	3,5	156,5	505,8	24,23	4,86
3° QUADRANTE	647,2	3,9	104	648	24,49	14,0
MÁXIMO	2923	10	6474	7241	26,33	23,9