Uso de técnicas de aprendizado de máquina na avaliação da qualidade do leite

Eduardo Antonio Borges

Universidade Federal de Uberlândia
Faculdade de Computação
Bacharelado em Sistemas de Informação

Monte Carmelo - MG
2023

Eduardo Antonio Borges

Uso de técnicas de aprendizado de máquina na avaliação da qualidade do leite

Trabalho de Conclusão de Curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Área de concentração: Ciência da Computação

Orientador: Fernanda Maria da Cunha Santos

Monte Carmelo - MG
2023

Este trabalho é dedicado às crianças adultas que, quando pequenas, sonharam em se tornar cientistas.

Agradecimentos

O desenvolvimento deste trabalho de conclusão de curso contou com a ajuda de diversas pessoas, dentre as quais agradeço:

A professora orientadora Fernanda Maria, me acompanhou em todo processo, dando auxílio necessário para a elaboração do projeto.

Aos professores do curso de Sistemas de Informação que através de seus ensinamentos permitiram que eu pudesse estar hoje concluindo este trabalho.

Aos meus pais que me deram todo suporte possível para a realização deste sonho.

"A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo"
(Albert Einstein)

Resumo

A cadeia produtiva do leite destaca-se como uma das principais atividades econômicas do Brasil em termos de geração de emprego e renda. Logo, a qualidade em que o leite é produzido deve ser estritamente fiscalizada, pois a qualidade inadequada acarreta diversos prejuízos para as indústrias e produtores. Assim, busca-se técnicas preditivas que definirão um padrão para a qualidade do leite produzido, com o intuito de garantir níveis de qualidade em todas épocas do ano. Diante esse cenário, o objetivo deste trabalho foi criar um modelo computacional usando técnicas de aprendizado de máquina para classificar a qualidade do leite. Dentre as técnicas de Aprendizado de máquinas, optou-se por implementar Redes Neurais Perceptron Multicamdas (PMC) e Support Vector Machine (SVM) aplicadas à uma base de dados formada por propriedades do leite independentes (pH, temperatura, sabor, odor, gordura, turbidez, cor). As redes neurais PMC obtiveram resultados melhores gerando acima de 98% nas medidas de avaliação acurácia, recall, precisão e f1-score para todas as três classes de qualidade de leite. O modelo implementando o SVM também teve resultados semelhantes. Conclui, que o modelo proposto classificará a qualidade do leite e, consequentemente, contribuirá com os produtores que terão o entendimento de quando haverá perda de qualidade do leite diante das alterações dos atributos analisados.

Palavras-chave: Qualidade do leite, aprendizado de máquina, redes neurais artificiais, *support vector machine*.

Lista de ilustrações

Lista de tabelas

Lista de siglas

AM Aprendizado de Máquinas

BDMEP Banco de Dados Meteorológicos para Ensino e Pesquisa

ITU Índice de Temperatura e Umidade

INMET Instituto Nacional de Meteorologia

MLP *Multilayer Perceptron*

MOS *Metal Oxide Semiconductor*

PMC Perceptron Multicamdas

PANDAS Panel Data

RNA Redes Neurais Artificiais

SVM Support Vector Machine

SRM Minimização do Risco Estrutural

TDE transdutor duplo-elemento

UHT *Ultra High Temperature*

Sumário

Capítulo 1

Introdução

O leite é uma mistura complexa, nutritiva e estável de gordura, proteínas, minerais e vitaminas (CARVALHO; STOCK, 2006). O leite é um produto delicado e altamente perecível, tendo suas características físicas, químicas e biológicas facilmente alteradas pela ação de microrganismos (WINCK et al., 2011). Ademais, o seu valor nutricional também pode ser reduzido pela adição de água ou de alguns produtos químicos, o representando sério risco à saúde do consumidor. Desta forma, todas as etapas que compõem a produção do leite devem ser minuciosamente observada para preservá-lo.

A cadeia produtiva do leite destaca-se como uma das principais atividades econômicas do Brasil em termos de geração de emprego e renda. Com mais de 1 milhão de produtores distribuídos em praticamente todos os municípios brasileiros, estima-se que a cadeia gere 4 milhões de empregos nos seus diferentes segmentos, resultando em valor bruto da produção superior a RS 27,2 bilhões (6º maior dentre os produtos agropecuários nacionais) e faturamento da indústria de laticínios de RS 70,9 bilhões, atrás apenas do setor de derivados de carne e de beneficiados de café, chá e cereais (ALIMENTAÇÃO, 2019).

Existe uma enorme cadeia trabalhística na produção do leite, envolvendo os produtores que fazem todo o serviço de extração do leite dos animais, até a disponibilidade do leite e seus derivados nas prateleiras dos supermercados. As pessoas e serviços envolvidos na produção do leite compreendem os agricultores responsáveis pelo fornecimento da alimentação dos animais, a infraestrutura específica das fazendas produtoras (pastos demarcados, ordenhadeiras, tanques, resfriadores, etc), as transportadoras de leite e laticínios destinados ao processamento do produto e outros. Contudo, como é um produto altamente perecível, suas características devem ser rigorosamente analisadas.

O leite com qualidade inadequada acarreta diversos prejuízos para as indústrias. Calcula-se que haja uma perda diária em torno de 2% do leite entregue à usina e uma redução de 7,6% da produção do leite devido à mastite, pois esta doença reduz o volume produzido devido á problemas que causa no animal e que durante o tratamento o leite deve ser descartado. No comércio varejista ocorrem perdas no período de vida de prateleira. Devido à má qualidade da matéria-prima, os produtos lácteos brasileiros têm

um tempo de prateleira bastante curto, quando comparado com os similares de países desenvolvidos, quando a matéria-prima é de qualidade inferior ao mínimo recomendável (GARCIA SILVA CRISTIANE ANDRADE; ROCHA, 2006).

Sendo um produto que deriva inúmeros outros usados na alimentação da população brasileira, a qualidade em que o leite é produzido deve ser estritamente fiscalizada. Quanto maior a qualidade do leite, melhor será o produto final e, consequentemente, um retorno financeiro satisfatório será gerado para os produtores.

1.1 Motivação

Alguns critérios físico-químicos são normalmente utilizados pelas indústrias de laticínios para avaliar os cuidados que o produtor deve ter com o leite. Esses critérios são úteis para a indústria controlar o rendimento industrial e a qualidade do produto final. Entretanto, se o intervalo de valores e as interpretações destes critérios forem utilizados de forma inadequada, podem penalizar injustamente os fornecedores da matéria-prima, uma vez que tais parâmetros são influenciados por características individuais do rebanho, pelo tipo de alimentação dos animais, pela época do ano, pelas condições climáticas, pelas formas de transporte do leite e outros (WINCK et al., 2011).

Devido as inúmeras variáveis que podem afetar a qualidade do leite, a predição de algumas destas contribuiria para garantir a integridade do produto, o que, consequentemente, levaria à manufatura de produtos com qualidade para a população e evitaria prejuízos financeiros aos produtores.

A variação da qualidade do leite durante alguns períodos do ano prejudica financeiramente o produtor, pois o preço final do seu produto é reduzido, o que interfere direto na renda final do mês e prejudica nos possíveis investimentos nas etapas diversas da produção de leite. Assim, busca-se técnicas preditivas que definirão um padrão para a qualidade do leite produzido, com o intuito de garantir níveis de qualidade em determinadas épocas do ano.

A obtenção dos valores de certas características do leite trará uma ideia de quando haverá perda de qualidade do leite diante as alterações preditas, para que assim seja possível tratar uma solução para o problema ou prever quando ocorrerá um maior declínio da qualidade pelo produtor.

O uso de sistemas computacionais para auxiliar na classificação da qualidade do leite é uma ferramenta sugestiva, principalmente, se for constituída por técnicas de aprendizado de máquina, como as Redes Neurais Artificiais (RNA), Árvores de Decisão, SVM e outros algoritmos. As redes neurais artificiais são uma excelente alternativa para resolução de problemas de classificação, uma vez que o processamento é estruturalmente paralelo e apresenta diversas funcionalidades, como adaptabilidade, tolerância à falha e abstração (BRAGA; LUDERMIR, 2000).

1.2 Hipótese

A implementação de um modelo computacional estruturado sob os algoritmos de Aprendizado de Máquinas para classificar a qualidade do leite perante algumas características do leite.

1.3 Objetivos

O objetivo deste trabalho é criar um modelo computacional que avalia a qualidade do leite através de um conjunto de atributos independentes do leite. Os valores destes atributos servirão como dados de entrada para o treinamento e teste das RNA e do algoritmo de SVM.

O uso das técnicas de mineração de dados no processamento de dados massivos, permite fazer uma análise bem detalhada de várias informações, levando a ter conclusões que podem ser interessantes para a linha de produção. Analisando sob esse aspecto, pode ser feita uma analise das variáveis sob a qualidade do leite para que se obtenha alguns padrões para classificar a qualificação do leite em bom, ruim ou moderada, principalmente, em determinadas épocas do ano. Consequentemente, possibilitará definir um modelo que se possa desenvolver métodos de prevenção da perda da qualidade em períodos críticos.

1.4 Contribuições

Espera com o fim do estudo obter um resultado que possa beneficiar os produtores de leite e os laticínios de forma em que possam tomar medidas que evitem a perda da qualidade do leite em determinados momentos do ano produtivo.

1.5 Organização da Monografia

No Capítulo 2 é apresentado o conteúdo teórico utilizado neste trabalho, bem como o referencial teórico pesquisado. Na sequência, no Capítulo 3, são descritas as etapas de desenvolvimento do modelo computacional e os resultados obtidos após o treinamento e validação do mesmo. Por último, no Capítulo 4, são descritos os principais resultados e dificuldades encontradas durante a evolução do trabalho.

Capítulo 2

Fundamentação Teórica

Neste capítulo, serão apresentados os conceitos sobre redes neurais artificiais e sobre SVM que foram relevantes para o desenvolvimento do trabalho. Além disso, haverá uma descrição sobre as características da qualidade do leite e os trabalhos correlatos que auxiliaram na estruturação do tema proposto.

2.1 Redes Neurais Artificiais

Uma rede neural é um sistema projetado para simular a maneira como o cérebro humano realiza uma tarefa particular, sendo normalmente implementada via componentes eletrônicos ou por sistemas computacionais. Para alcançarem bons desempenhos, as redes neurais empregam uma interligação maciça de células computacionais simples, denominadas de "neurônios" ou unidades de processamento (HAYKIN, 2007). De acordo com Haykin (2007), a rede neural se assemelha ao cérebro humano em dois aspectos básicos:

- ❏ o conhecimento é adquirido pela rede a partir de seu ambiente, por intermédio do processo de aprendizagem;
- ❏ forças de conexão entre neurônios (pesos sinápticos) são utilizadas para armazenar o conhecimento adquirido.

As redes neurais facilitam os estudos e trazem resultados interessantes na análise de dados massivos, pois operam sob diferentes tipos de dados, como imagens e planilhas com dados diversos. Elas são organizadas em multicamadas, onde a primeira parte se posiciona a receber os padrões a serem seguidos para determinado tipo analisado. A segunda parte se desenvolve para apreender sobre as variáveis que são designadas para a caracterização do estudo e a terceira camada da rede vai gerar os resultados obtidos a partir do que foi apreendido nas fases anteriores. A Figura 1 mostra a organização em camadas de uma rede neural.

A arquitetura da rede neural PMC é capaz de realizar aproximações universais de funções a partir de um conjunto de dados de treino. Existem muitos campos de aplicação

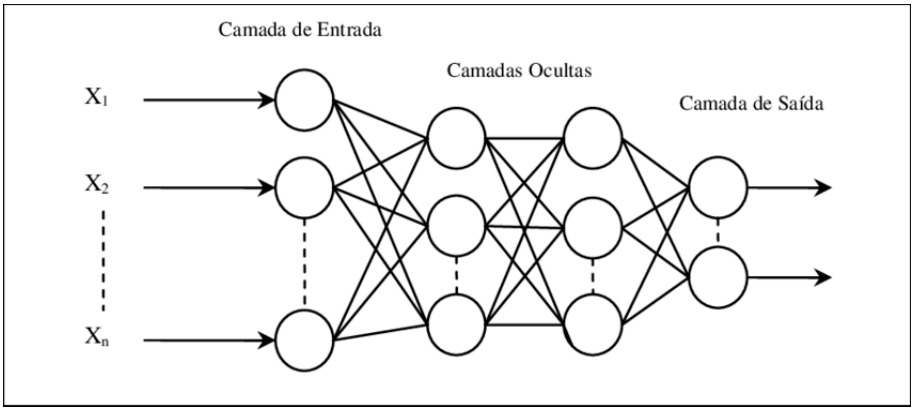

Figura 1 – Representação das camadas de uma rede neural artificial.

Fonte: https://www.researchgate.net/figure/Figura-2-Perceptron-Multicamadas-Adaptado-de-LNCC-2008$_fig1_2$8432919

que permitem mapear o problema para uma rede neural e adaptar os pesos pelos dados de treino especificados.

Uma desvantagem da PMC é o tempo de treinamento extenso. Em um problema complexo pode-se levar muito tempo até se obter um conjunto de pesos adequados, além da escolha da taxa de aprendizagem, que desempenha um papel fundamental.

No processo de aprendizagem, a rede neural aprende pelo método supervisionado ou não-supervisionado. No supervisionado, cada conjunto de atributos, representando uma amostra, é acompanhado por um valor de classificação, que é denominado alvo ou o valor desejado. Na aprendizagem não-supervisionada, o algoritmo irá aprende apenas com as amostras de treino disponíveis, sem a presença do alvo.

A base de dados usada na execução da rede neural é dividida em conjunto de treinamento e conjunto de testes. No primeiro, a rede neural aprende pela definição de todos os parâmetros para que o algoritmo identifique os padrões que precisa classificar. O conjunto de dados de teste é apresentado para a rede, e a eficiência da rede neural é verificada.

2.2 Support Vector Machine

SVM é um algoritmo de classificação supervisionado que, por meio de vetores de amostras de treinamento, estabelece um hiperplano ótimo de separação entre as classes a fim de maximizar a distância entre elas (HAYKIN, 2007), além de classificar adequadamente os dados de teste, os quais não foram vistos previamente.

Dado um conjunto de treinamento composto por n amostras e pertencentes a duas classes linearmente separáveis. O conjunto de treinamento é denominado por vetores de suporte e o objetivo é definir um hiperplano que separe os vetores. Entre os muitos hiperplanos possíveis, o hiperplano separador ótimo é o plano que maximiza a margem,

ou seja, a distância entre o hiperplano e o vetor mais próximo de cada classe. A Figura 2 ilustra um hiperplano separador ótimo.

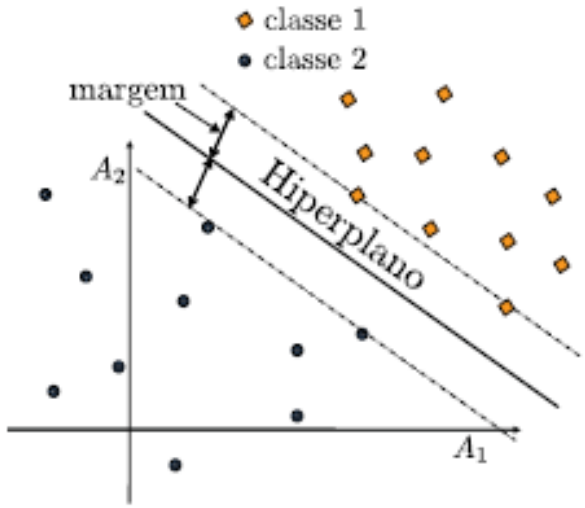

Figura 2 – Hiperplano de separação SVM de maior margem (adaptada de (HAYKIN, 2007)).

As margens grandes tendem a ter erros de generalização melhores, como pode ser explicado pelo princípio de Minimização do Risco Estrutural (SRM). Esse princípio define um limite mais alto para o erro de generalização de um classificador a partir de seus erros de treinamento, o número de exemplos de treinamento e sua capacidade (TAN; STEINBACH; KUMAR, 2009).

Um SVM é eficiente na classificação de dados linearmente separáveis, pela definição de um hiperplano com a maior margem. A este classificador da-se o nome de SVM linear. Também, aplica-se SVM em conjunto de dados que tenham limites de decisão não lineares, transformando os dados do seu espaço de coordenadas original x para um novo espaço $\Phi(x)$ de maior dimensão (LORENA; CARVALHO, 2007), como pode ser visto na Figura 3

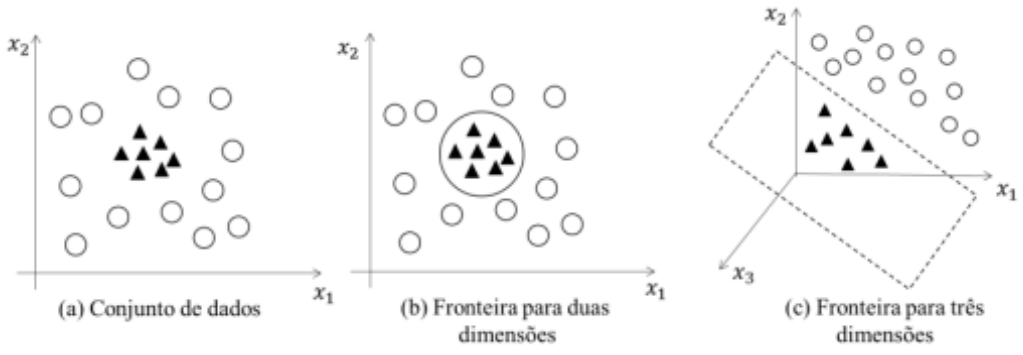

Figura 3 – Exemplo da transformação ocorrida num conjunto de dados não-linear para outro espaço de coordenadas de dimensão maior (GAMA et al., 2011).

Os principais parâmetros de ajuste do SVM para o modelo proposto neste trabalho foram:

- *kernel*: define a função que cria os hiperplanos de separação das classes.
- C: trata a penalidade do erro, além de controlar o trade off entre o limite de decisão suave e a classificação correta dos pontos de treinamento (SOUZA, 2019).
- *gamma*: pode ser considerado como erro permitido, ou seja, quanto maior o seu valor, tentará ajustar exatamente o conjunto de teste ao conjunto de treinamento (SOUZA, 2019). Esse coeficiente é usado como parâmetro auxiliar as funções de Kernel *rbf, poly e sigmoid*.

2.3 Qualidade do Leite

A qualidade do leite é de extrema importância para todos envolvidos na sua produção e processamento, pois ter uma boa qualidade do leite representa bônus financeiros. Ao se referir à qualidade do leite, deve-se ater principalmente à qualidade da matéria-prima, que é ponto de extrema importância no processo de inserção do Brasil no mercado mundial de lácteos. Essa questão envolve mudança radical nas normas de recepção do leite (contagem bacteriana, crioscopia, acidez, células somáticas, etc.) e introdução de normas de origem (refrigeração na propriedade, coleta a granel e ordenha mecânica), conforme preconizado no Programa de Melhoria da Qualidade do Leite (SANTOS; FARIA, 2005).

Sistemas atuais de estimação das exigências nutricionais e energéticas dos animais (NRC, 2001) consideram as interações entre alimentos, entre energia e proteína, de maneira que as referidas exigências, expressas na base diária, podem variar segundo as características dos alimentos utilizados. Isto tem contribuído para adequação de dietas para vacas leiteiras, permitindo, por exemplo, reduções consideráveis no fornecimento proteína e excreção de nitrogênio pelos animais (COUNCIL, 2001).

Essas características do leite podem sofrer alterações devido aos fatores climáticos que afetam o conforto dos animais, tipos de alimentação dos animais, que afetam diretamente na proteína do leite produzido. A alimentação é influenciada na maioria dos casos pela época do ano, onde o clima afeta diretamente sobre que tipo de alimentação que vai ser dada aos animais, que podem diferenciar o teor de proteína, devido suas composições nutritivas.

2.4 Trabalhos Correlatos

O estudo desenvolvido por Borges et al. (2019) teve a finalidade de predizer perdas na produção leiteira pelas condições climáticas, ou seja, com base no Índice de Temperatura e Umidade (ITU). O impacto de fatores ambientais na produção animal tem sido

avaliado por meio de diversos índices de conforto térmico e são estimados, em função da temperatura, a umidade relativa do ar, a velocidade do vento e a radiação solar. A partir desta problemática, objetivou-se estimar as perdas na produção leiteira, em função do Índice de Temperatura e Umidade, aplicando-se redes neurais artificiais.

A base de dados foi formada por informações meteorológicas provenientes do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET). Os dados referem-se ao período compreendido entre 01/01/1988 a 31/12/2017, totalizando 29 anos sobre a Estação Meteorológica Convencional do município de Xavantina(MT). Foram coletados valores de temperatura de bulbo seco e umidade relativa do ar nos horários correspondentes ás 00:00 horas, 12:00 horas, 18:00 horas. Os dados foram armazenados em planilhas do Excel e organizados pelo dia e hora, coletando cerca de 1095 leituras anuais para a formação da base de dados.

As Redes Neurais Perceptron Multicamadas foram escolhidas e para definir a arquitetura adequada para a base de dados em análise, realizou testes a em diferentes arquiteturas de redes contendo até duas camadas intermediárias ou ocultas e cada uma possuindo como máximo oito neurônios, sendo avaliadas um total de 26 combinações.

Conclui-se neste trabalho que a Rede Neural possibilita estimar o ITU para qualquer dia do ano, ela pode auxiliar no planejamento de medidas simples para amenizar o ambiente térmico das vacas leiteiras.

No trabalho desenvolvido por Silva Jomar S. Vasconcelos (2017) objetivou desenvolver um nariz eletrônico portátil capaz de reconhecer diferentes substâncias adulterantes no leite. Tal adulteração geralmente inclui produtos químicos como formaldeído, hidróxido de sódio e ureia. O dispositivo, mais conhecido como *e-nose* ou nariz eletrônico, é um sistema embarcado composto por uma matriz de sensores químicos não seletivos, por um sistema microcontrolador para o processamento de sinais, e uma rede neural artificial para a classificação e o reconhecimento de perfis aromáticos de cada contaminante.

Para a construção da matriz foram utilizados três tipos de sensores do tipo *Metal Oxide Semiconductor* (MOS). Para embarcação do algoritmo de controle foi escolhido a placa de desenvolvimento Arduino UNO baseada no microcontrolador ATMega 328P de arquitetura AVR. Para a rede neural optou-se pela implementação da arquitetura *Multilayer Perceptron* (MLP), e o software escolhido foi o Matlab® devido à sua diversidade de ferramentas, e por apresentar uma linguagem de programação simples e de alto desempenho.

Em relação a base de dados, as amostras foram recolhidas de cinco marcas comerciais de leite *Ultra High Temperature* (UHT) e submetidas à adição de diferentes concentrações de contaminantes.

Para o reconhecimento e classificação de cada adulterante a rede neural MLP apresentou desempenho satisfatório reconhecendo todos os contaminantes do conjunto de teste. Os resultados apresentaram que das amostras utilizadas para treinamento da rede 97,1%

foram corretamente classificadas bem como 95,7% das amostras de validação. Os parâmetros da rede foram ajustados pelo método de otimização simplex sequencial.

O trabalho desenvolvido por Nazário et al. (2009) mostra a aplicação de técnicas de ultra-som e de redes neurais para classificar amostras de leite líquido em função do teor de gordura e de água adicionada, com o objetivo de detectar adulterações.

As caracterizações do leite líquido são feitas em laboratórios existentes nos próprios laticínios, onde medem-se propriedades como o teor de gordura, proteína, sólidos totais e água adicionada, entre outros. Estas propriedades são medidas e balanceadas de acordo com padrões aceitos pela pela indústria. Uma alternativa a ser considerada na caracterização do leite é o uso de ultra-som, que é uma técnica relativamente simples e não-destrutiva.

A célula de medição por ultra-som, apresenta três partes: transdutor duplo-elemento (TDE), câmara de amostra e refletor. O TDE é, por sua vez, composto por um transdutor emissor piezoelétrico compósito, uma linha de retardo de acrílico, um receptor de grande abertura de P(VDF-TrFE) (Polivinilideno de Flúor Tri-Flúor Etileno) e uma linha de retardo de vidro. Foram utilizadas amostras de leite UHT para se fazer a base de dados utilizada no estudo. Os sinais dos sensores são digitalizados por uma placa de aquisição com elevada resolução (National Instruments NI-4351, 24 bits) e um programa escrito em MATLAB faz o controle da placa e a leitura dos dados.

A célula de medição de propriedades de líquidos no ultra-som obteve a densidade, a velocidade de propagação e o coeficiente de atenuação, que foram relacionados com as concentrações de gordura e água adicionada em amostras de leite bovino. Esses dados foram utilizados por uma rede neural do tipo MLP para classificar os teores de gordura do leite UHT. As redes neurais criadas resultaram em mais de 95% de amostras classificadas corretamente.

Capítulo 3

Experimentos e Análise dos Resultados

Neste capítulo, serão apresentadas as etapas do modelo computacional proposto para a classificação da qualidade do leite, bem como as características definidas em cada etapa. Também, são descritas a linguagem e bibliotecas das ferramentas computacionais utilizadas na implementação do modelo. Por fim, é descrito os resultados alcançados e uma análise detalhada dos mesmos.

3.1 Métodos do Modelo Proposto

A Figura 4 exibe as etapas do modelo proposto para classificação da qualidade do leite. Nas sessões subsequentes, haverá uma descrição detalhada de cada etapa.

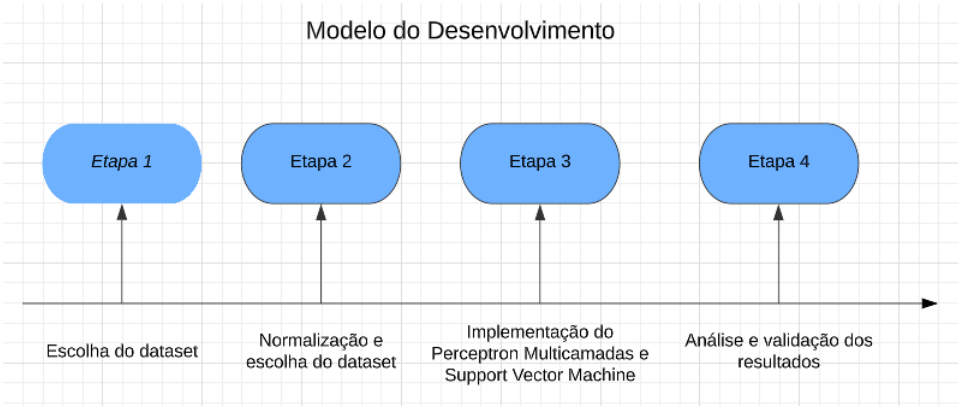

Figura 4 – Etapas da metodologia do sistema computacional proposto.

Fonte:Autoria própria.

3.1.1 Base de Dados

A base de dados aplicada neste trabalho foi retirada do site Kaggle (GNV, 2020). Ela é composta por 1059 amostras e 8 atributos, sendo os sete primeiros variáveis inde-

pendentes relacionadas às propriedades do leite: pH, temperatura, sabor, odor, gordura, turbidez, cor. Geralmente, o grau ou qualidade do leite depende desses parâmetros, que desempenham um papel vital na análise preditiva do leite.

O oitavo atributo traz a classificação do leite, definido pelos valores 0, 0,5 ou 1, ou seja, a qualidade do leite é classificada como ruim, moderada e boa, respectivamente. A Figura 5 mostra a distribuição das amostras e suas classes.

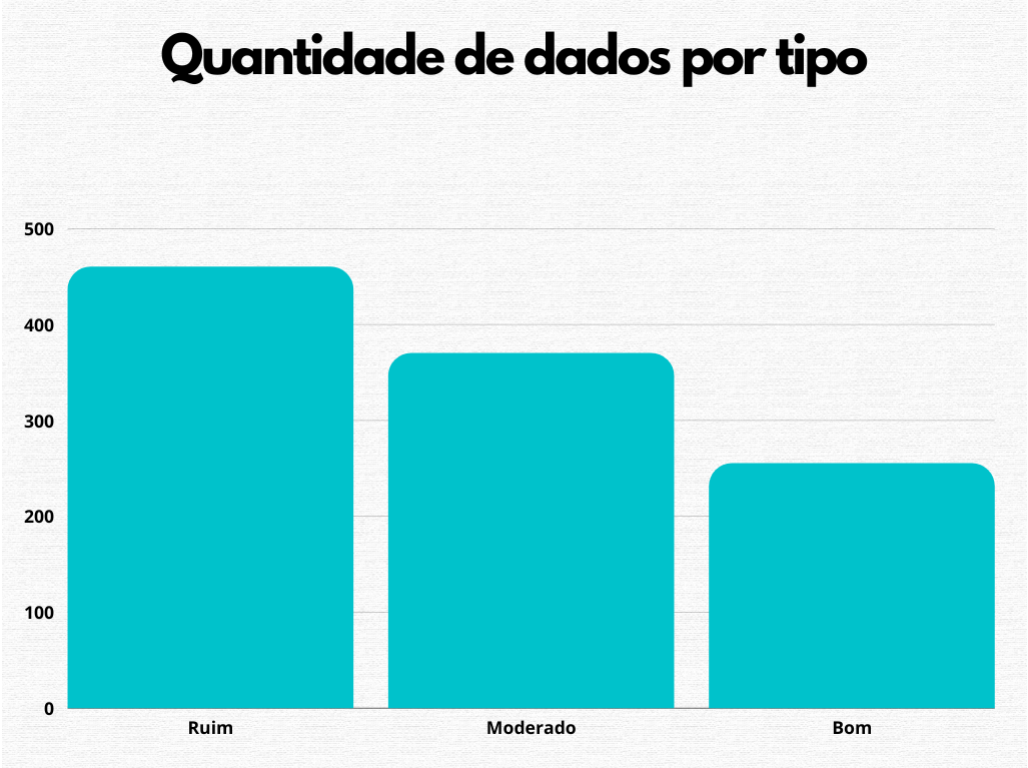

Figura 5 – Distribuição das amostras da base de dados e suas classificações.

Fonte: Autoria própria.

Ao analisar os valores presentes na base de dados, os intervalos de valores das características temperatura, pH e cor são 34 a 90, 3 a 9,5 e 240 a 255, respectivamente. A Tabela 1 mostra, detalhadamente, os intervalos de valores das características temperatura, pH e cor para as classes de qualidade do leite.

Tabela 1 – Intervalos de valores para classificação da qualidade do leite.

	Temperatura	**pH**	**Cor**
Bom	35 - 45	6.5 - 6.8	245 - 255
Moderado	34 - 45	6.4 - 6.8	240 - 255
Ruim	34 - 90	3.0 - 9.5	245 - 255

Fonte: GNV (2020)

Observando os valores da Tabela 1, os atributos temperatura e pH apresentam semelhanças entre si para as classificações moderado e bom. Isso indica que precisará critérios diferenciados para decidir qual classe ou se esse atributo não atrapalhará os sistemas inteligentes definir a classificação..

3.1.2 Linguagem de Programação do Modelo

A linguagem de programação Phyton foi a linguagem escolhida para implementar o modelo de classificação devido a grande quantidade de bibliotecas para algoritmos de aprendizado de máquina (PYTHON, 2022). Foram utilizadas as bibliotecas Tensorflow que permite o uso de funcionalidades que auxiliam no desenvolvimento de funções para calculo de pesos, função de ativação e otimização(INC, 2022). A biblioteca Panel Data (PANDAS) foi usada para a manipulação dos dados, normalizando a base de dados. A biblioteca Scikit-learn para fazer o uso do SVM(SCIKIT-LEARN, 2022).

3.1.3 Pré-Processamento

Todos os atributos da base de dados passaram por uma normalização para facilitar os cálculos do modelo. Foi usada a biblioteca Pandas para aplicar a escala máxima absoluta, onde redimensiona cada recurso entre -1 e 1, dividindo cada observação por seu valor absoluto máximo, usando os métodos $.max()$ e $.abs()$.

3.1.4 Parametrização das Técnicas de Aprendizado de Máquina

As arquiteturas das redes neurais MLP implementadas foram definidas depois de realizar alguns testes com a base de dados. Assim, segundo a descrição da Tabela 2, foram realizados testes e analisados os resultados das redes neurais com dois valores para a taxa de aprendizagem com três diferentes quantidades de neurônios na camada neural escondida.

A função de ativação definida nos neurônios da camada escondida foi o Relu. A função Relu retorna 0 para todos os valores negativos, e o próprio valor para valores positivos. O número de épocas foi 5000.

Tabela 2 – Configurações das arquiteturas das RNAs para o modelo

Taxa de aprendizagem	Quantidade de neurônios
0,1	10
	15
	20
0,01	10
	15
	20

Support Vector Machine é um algoritmo de aprendizado de máquina supervisionado que pode ser usado para desafios de classificação ou regressão. Ele utiliza o método de plotagem dos dados em um plano e define uma reta entre as classes diferentes, fazendo com que em cada interação distancie a posição dos dados da reta linear definida.

3.1.5 Medidas de Avaliação

Após desenvolver a etapa de aprendizagem de um classificador, é importante verificar se ele apresenta um bom desempenho aplicando o modelo no conjunto de testes. Para isso, obtém a matriz de confusão que é responsável por gerar variáveis usadas no cálculo das medidas quantitativas de desempenho. São elas:

- *Acurácia*: total de acertos que a rede neural fez de acordo com a base de dados de teste.
- *Precisão*: de todos os tipos de leite que a rede neural indicou como sendo da classe x, quais realmente são dessa classe?
- *Recall*: de todos os tipos de leite do tipo x presentes na minha base de dados, quantas a rede neural conseguiu identificar?
- *F1-score*: média harmônica entre precisão e o recall.

3.2 Experimentos e Avaliação dos Resultados

Foram realizados diversos experimentos durante a fase de treinamento do modelo, testando diferentes valores para os parâmetros dos algoritmos, até obter os melhores que definiram as redes neurais MLP e o SVM para atuar na classificação da qualidade do leite. Nesta seção, serão apresentados os resultados médios obtidos com o conjunto de dados de teste aplicado as redes MLP e ao algoritmo do SVM por 10 vezes. As variações das arquiteturas das redes MLP são caracterizadas pelas diferentes taxas de aprendizagem e pela quantidade de neurônios na camada oculta.

3.2.1 Experimento 1

Este experimento tem o propósito de identificar qual a melhor divisão do conjunto de treinamento e do conjunto de teste para a base de dados em estudo. Foram escolhidas as divisões de 80%,50%,e 20% para treino e 20%,50%,e 80% para teste, respectivamente.

Para avaliar o desempenho do algoritmo para cada divisão foi aplicado as configurações de 0,1 de taxa de aprendizagem e 20 neurônios na camada oculta.

Tabela 3 – Resultados do modelo formado por 80% treino e 20% teste.

	Medidas de avaliação			
Tipo de leite	Precisão	Recall	f1-score	Acurácia
Ruim	1,00	0,98	0,99	
Moderado	1,00	1,00	1,00	0, 99
Bom	0,97	1,00	0,99	

Tabela 4 – Resultados do modelo formado por 50% treino e 50% teste.

	Medidas de avaliação			
Tipo de leite	Precisão	Recall	f1-score	Acurácia
Ruim	0,97	0,99	0,98	
Moderado	1,00	0,99	0,99	0, 98
Bom	0,98	0,96	0,97	

Tabela 5 – Resultados do modelo formado por 20% treino e 80% teste.

	Medidas de avaliação			
Tipo de leite	Precisão	Recall	f1-score	Acurácia
Ruim	1,00	1,00	1,00	
Moderado	1,00	0,87	0,93	0, 95
Bom	0,83	0,99	0,90	

Foi colocado nas tabelas os resultados que apareceram mais em algumas execuções, portanto escolhi a divisão de 80% para treino e 20% para teste nos próximos experimentos, pois teve um melhor desempenho nestes experimento.

3.2.2 Experimento 2

No primeiro experimento foi feito testes com a taxa de aprendizagem fixa em 0,01 e com diferentes quantidades de neurônios na camada oculta. A Tabela 6 mostra os valores médios alcançados pelas medidas de desempenho com a arquitetura da rede MLP formada por 10 neurônios, a Tabela 7 com 15 neurônios e a Tabela 8 com 20 neurônios.

Tabela 6 – Resultados do modelo formado por 10 neurônios na camada oculta.

	Medidas de avaliação			
Tipo do Leite	Precisão	Recall	f1-score	Acurácia
Ruim	0,99	0,98	0,98	
Moderado	0,97	0,99	0,98	0, 98
Bom	0,97	0,96	0,96	

Diante os resultados apresentados nas Tabelas 6, 7 e 8, percebe-se que os valores são bem parelhos e com resultados significativos, acima de 96% de acertos. Entretanto, a Tabela 8 mostra que a rede neural MLP com 20 neurônios apresentou uma melhor classificação, dado os valores gerados pelas medidas de desempenho.

Tabela 7 – Resultados do modelo formado por 15 neurônios na camada oculta.

Tipo de leite	Medidas de avaliação			
	Precisão	Recall	f1-score	Acurácia
Ruim	0,99	0,98	0,98	
Moderado	0,97	0,99	0,98	0, 98
Bom	0,97	0,96	0,96	

Tabela 8 – Resultados do modelo formado por 20 neurônios na camada oculta.

	Medidas de avaliação			
Tipo de leite	Precisão	Recall	f1-score	Acurácia
Ruim	1,00	0,98	0,99	
Moderado	1,00	1,00	1,00	0, 99
Bom	0,97	1,00	0,99	

3.2.3 Experimento 3

Os resultados apresentados nesta subseção referem-se as arquiteturas das redes neurais MLP com taxa de aprendizagem igual a 0,1, e com 10, 15 e 20 neurônios na camada escondida. Os resultados das medidas de desempenho para as respectivas redes MLP descritas, podem ser vistas nas Tabelas 9, 10 e 11.

Tabela 9 – Resultados do modelo formado por 10 neurônios na camada oculta.

	Medidas de avaliação			
Tipo de leite	Precisão	Recall	f1-score	Acurácia
Ruim	0,99	0,98	0,98	
Moderado	1,00	0,99	0,99	0, 99
Bom	0,97	1,00	0,99	

Tabela 10 – Resultados do modelo formado por 15 neurônios na camada oculta.

	Medidas de avaliação			
Tipo de leite	Precisão	Recall	f1-score	Acurácia
Ruim	0,98	0,98	0,98	
Moderado	0,86	0,98	0,91	0, 93
Bom	0,96	0,79	0,87	

Diante os resultados apresentados nas Tabelas 9, 10 e 11, percebe-se uma maior variação da acurácia da taxa de aprendizado 0,1 em comparação com a taxa 0,01. Também, conclui que o modelo da RNA com 10 neurônios apresentou os valores das medidas de desempenho superior aos modelos de 15 e 20 neurônios.

3.2.4 Experimento 4

Os valores dos parâmetros do SVM foram obtidos após realizar vários testes, buscando sempre os melhores resultados para as medidas de desempenho. Por isso, o algoritmo do

Tabela 11 – Resultados do modelo formado por 20 neurônios na camada oculta.

	Medidas de avaliação			
Tipo de leite	Precisão	Recall	f1-score	Acurácia
Ruim	0,99	0,98	0,98	
Moderado	0,96	0,95	0,95	0, 96
Bom	0,92	0,94	0,93	

SVM foi definido pelos parâmetros *kernel* igual a "poly", C igual a 4.0 e *gamma* igual a 2.0.

Tabela 12 – Resultados do modelo SVM.

	Medidas de avaliação			
Qualidade do leite	Precisão	Recall	f1-score	Acurácia
Ruim	0,98	0,98	0,98	
Moderado	1,00	0,98	0,99	0, 98
Bom	0,97	1,00	0,99	

A Tabela 12 apresenta os valores obtidos pelas medidas de desempenho para o algoritmo SVM alcançando resultados acima de 97% de acertos para todos métodos de avaliação.

3.2.5 Experimento 5

Aplicação das redes neurais e do SVM apenas para os atributos temperatura, pH e cor, de acordo com as informações da Tabela 1, é afirmado que pode gerar uma confusão no algoritmo na classificação dos tipos bom e moderado. Foi feito experimentos somente com os atributos citados acima, com taxa de aprendizagem 0,01 e 20 neurônios na camada oculta.

As linhas da matriz confusão é representada pela quantidade de dados do tipo ruim, moderado e bom, e na coluna é representada a classificação do algoritmo para cada tipo de dado.

Matriz confusão do algoritmo SVM.

$$\begin{bmatrix} 94 & 2 & 0 \\ 0 & 83 & 10 \\ 0 & 41 & 29 \end{bmatrix}$$

Matriz confusão do algoritmo MLP.

$$\begin{bmatrix} 94 & 2 & 0 \\ 0 & 89 & 4 \\ 0 & 24 & 46 \end{bmatrix}$$

Em seguida foi feito testes com todas as características da base de dados e a mesma configuração do algoritmo.

Matriz confusão do algoritmo SVM.

$$\begin{bmatrix} 94 & 0 & 2 \\ 2 & 91 & 0 \\ 0 & 0 & 70 \end{bmatrix}$$

Matriz confusão do algoritmo MLP.

$$\begin{bmatrix} 94 & 0 & 2 \\ 0 & 93 & 0 \\ 0 & 3 & 67 \end{bmatrix}$$

3.3 Avaliação dos Resultados

Os resultados das tabelas descritas na seção 3.2 validam a hipótese de gerar um modelo computacional, implementado a partir de algoritmos de Aprendizado de Máquinas (AM), para classificar a qualidade do leite. As técnicas RNA e SVM alcançaram resultados significativos ao distinguir as três classes propostas.

Em relação as arquiteturas analisadas das redes neurais MLP destacam-se a com 20 neurônios na camada escondida e taxa de aprendizagem igual a 0,01 e aquela com 10 neurônios e taxa de aprendizagem igual a 0,1. Estas obtiveram taxas de acertos superior a 98% em todas as medidas de avaliação. De maneira semelhante, o SVM também atingiu acima de 97% os acertos das medidas de avaliação.

Entretanto, é importante destacar que a base de dados estava desbalanceada em relação a quantidade de amostras de cada classe, o que poderia ter prejudicado a classificação. Ademais, os atributos do leite presentes na base de dados em estudo devem ser consideradas poucas, dado a quantidade de exigências recomendadas pelos produtores e pelos distribuidores.

Capítulo 4

Conclusão

A classificação da qualidade do leite por meio de um modelo computacional foi o fator preponderante no estudo deste projeto, que com o auxílio das redes neurais MLP e do SVM obteve êxito ao atingir assertividade acima de 97% das medidas de avaliação.

Ressalta que as redes neurais MLP e o SVM são duas eficazes técnicas de classificação aplicadas em base de dados com rápido poder de generalização, o que resulta num bom aprendizado dado a escolha adequada da arquitetura. Desta forma, os resultados obtidos foram satisfatórios tanto com o uso das redes MLP quanto para o SVM.

No entanto, a base de dados utilizada neste estudo, é pequena dada a proporção de propriedades e estados que o leite recebe influências, atuando em sua qualidade. Assim, uma base de dados com um número maior de amostras e de atributos, poderia auxiliar melhor todos os profissionais envolvidos.

Com a finalização deste estudo, sugere como trabalho futuro a criação de uma base de dados mais complexa que possa agregar diferentes atributos, ou seja, as diferentes variáveis do leite que são exigidas tanto pelos produtores quanto pelas indústrias de laticínio. Como por exemplo uma base coma atributos específicos para um determinado setor de produção ou produto feito a partir do leite. Com uma base da dados mais completa, seria possível testar outras técnicas de AM ou outras arquiteturas de RNA.

Para uma evolução do trabalho realizado foi colocado o código usado para realizar este estudo, onde é encontrado no usuário "eduardo-1110"no github.

Referências

ALIMENTAÇÃO, A. B. das Indústrias de. **Números do Setor – Faturamento**. 2019. Disponível em: <http://www.abia.org.br/vsn/anexos/faturamento2016.pdf>. Citado na página 12.

BORGES, P. H. de M. et al. Predição do declínio na produção leiteira com auxílio de redes neurais artificiais. **Scientia Agraria Paranaensis**, v. 17, n. 4, p. 71–97, 2019. Citado na página 18.

BRAGA, A. P. d. L. F. d. C. Antonio de P.; LUDERMIR, T. B. **Redes neurais artificiais: teoria e aplicações**. [S.l.]: LTC, 2000. Citado na página 13.

CARVALHO, A. V. C. G. R.; STOCK, L. A. **Brasil no cenário mundial de lácteos.** [S.l.], 2006. Disponível em: <https://www.embrapa.br/busca-de-publicacoes/-/publicacao/595890/o-brasil-no-cenario-mundial-de-lacteos>. Citado na página 12.

COUNCIL, N. R. **Nutrient Requirements of Dairy Cattle**. [S.l.]: National Academy Press, 2001. Citado na página 18.

GAMA, J. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Grupo Gen - LTC, 2011. ISBN 9788521618805. Disponível em: <https://books.google.com.br/books?id=4DwelAEACAAJ>. Citado 2 vezes nas páginas 6 e 17.

GARCIA SILVA CRISTIANE ANDRADE, V. F. M. P. R.; ROCHA, R. M. E. Produção e qualidade do leite na bacia leiteira de pelotas-rs em diferentes meses do ano. **Ciência Rural**, v. 36, n. 1, p. 209–214, 2006. Citado na página 13.

GNV, P. **Milk Grading**. 2020. Disponível em: <https://www.kaggle.com/datasets/prudhvignv/milk-grading>. Citado 2 vezes nas páginas 21 e 22.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007. Citado 4 vezes nas páginas 6, 15, 16 e 17.

INC, G. **TensorFlow**. 2022. Disponível em: <https://www.tensorflow.org/learn?hl=pt-br>. Citado na página 23.

LORENA, A. C.; CARVALHO, A. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007. ISSN 21752745. Disponível em: <https://seer.ufrgs.br/rita/article/view/rita_v14_n2_p43-67>. Citado na página 17.

NAZáRIO, S. L. S. et al. Caracterização de leite bovino utilizando ultra-som e redes neurais artificiais. **Revista Controle & Automação**, v. 20, n. 4, p. 627–636, 2009. Citado na página 20.

PYTHON. 2022. Disponível em: <https://www.python.org/>. Citado na página 23.

SANTOS, J. C. d. M. F. A. P.; FARIA, V. P. de. **VISÃO técnica e econômica da produção leiteira**. [S.l.]: FEALQ, 2005. Citado na página 18.

SCIKIT-LEARN. 2022. Disponível em: <https://scikit-learn.org/stable/>. Citado na página 23.

SILVA JOMAR S. VASCONCELOS, A. C. S. V. F. L. Dispositivo sensorial olfativo associado À rede neural artificial para identificação de contaminantes no leite. In: **XIII Simpósio Brasileiro de Automação Inteligente**. [s.n.], 2017. p. 2177–2182. Disponível em: <https://www.ufrgs.br/sbai17/program.php>. Citado na página 19.

SOUZA, A. **Algoritmo SVM (Máquina de vetores de suporte) a partir de exemplos e código (Python e R)**. 2019. Disponível em: <https://blogdozouza.wordpress.com/2019/04/10/algoritmo-svm-maquina-de-vetores-de-suporte-a-partir-de-exemplos-e-codigo-python-e-r/>. Citado na página 18.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining: mineração de dados**. [S.l.]: Ciência Moderna, 2009. Citado na página 17.

WINCK, C. A. et al. Produção de leite no brasil: qualidade, mercado internacional e agricultura familiar. **PUBVET**, v. 5, n. 32, p. 1205–1211, 2011. Citado 2 vezes nas páginas 12 e 13.