

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA & INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM
COMPUTACIONAL

Douglas Lima Fonseca

Inteligência Artificial para Identificação de Mastite Bovina em Sistemas de
Produção *Compost Barn*

Juiz de Fora

2023

Douglas Lima Fonseca

Inteligência Artificial para Identificação de Mastite Bovina em Sistemas de
Produção *Compost Barn*

Dissertação apresentada ao PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional.

Orientador: Prof. Dr. Leonardo Goliatt da Fonseca

Juiz de Fora

2023

Ficha catalográfica elaborada através do programa de geração automática da Biblioteca Universitária da UFJF, com os dados fornecidos pelo(a) autor(a)

Fonseca, Douglas Lima.

Inteligência artificial para identificação de mastite bovina em sistemas de produção compost barn / Douglas Lima Fonseca. -- 2023.

87 p. : il.

Orientador: Leonardo Goliatt da Fonseca

Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Faculdade de Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2023.

1. Aprendizado de máquina automatizado. 2. Mastite. 3. Compost Barns. I. Fonseca, Leonardo Goliatt da, orient. II. Título.

Douglas Lima Fonseca

**Inteligência Artificial para Identificação de Mastite Bovina em Sistemas de Produção
Compost Barn**

Dissertação
apresentada ao
Programa de Pós-
graduação em
Modelagem
Computacional
da Universidade
Federal de Juiz de
Fora como requisito
parcial à obtenção do
título de Mestre em
Modelagem
Computacional. Área
de concentração:
Modelagem
Computacional.

Aprovada em 28 de setembro de 2023.

BANCA EXAMINADORA

Prof. Dr. Leonardo Goliatt da Fonseca - Orientador

Universidade Federal de Juiz de Fora

Prof. Dr. Wagner Antonio Arbex

Universidade Federal de Juiz de Fora

Prof. Dr. Jefferson Filgueira Alcindo

Universidade Federal Rural do Semi-Árido

Juiz de Fora, 01/11/2023.



Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca, Professor(a)**, em 06/11/2023, às 11:28, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Jefferson Filgueira Alcindo, Usuário Externo**, em 06/11/2023, às 15:25, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Wagner Antonio Arbex, Professor(a)**, em 08/11/2023, às 19:14, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **1559037** e o código CRC **B76412D2**.

AGRADECIMENTOS

Antes de mais nada, quero expressar minha gratidão a Deus por Sua orientação e bênçãos em todos os momentos da minha jornada. Sem Sua ajuda e proteção divina, eu não teria sido capaz de chegar até aqui e concluir este projeto com sucesso. Sou imensamente grato por Seu amor incondicional e por me proporcionar a força e a coragem necessárias para enfrentar os desafios e obstáculos que surgiram ao longo do caminho. Agradeço a Deus por me conceder a oportunidade de crescer e aprender com essa experiência e por Sua presença constante em minha vida.

Gostaria de expressar minha gratidão especial ao meu filho, Joaquim, e à minha esposa, Flávia, por todos os sacrifícios que fizeram durante o processo de elaboração desta dissertação. Joaquim foi uma fonte constante de inspiração e motivação para mim, mesmo quando eu precisava trabalhar horas extras para concluir minha pesquisa. Flávia sempre esteve ao meu lado, encorajando-me e me ajudando a manter o foco, mesmo quando isso significava adiar nossos próprios planos e compromissos. Sem o apoio incondicional e o amor deles, este projeto não teria sido possível.

Não posso expressar o suficiente minha gratidão à minha mãe, Vilma, e meu pai, Jânio, por terem me ajudado a conciliar as demandas da minha vida acadêmica com a paternidade. Eles sempre estiveram disponíveis para cuidar do meu filho quando eu precisava comparecer às aulas ou estudar para as disciplinas. Sem esse suporte, seria impossível para mim concluir essa jornada. Eles também me ensinaram valores importantes, como dedicação e perseverança, que me ajudaram a chegar onde estou hoje. Sou grato por tudo o que fizeram por mim e por serem exemplos tão valiosos em minha vida. Eu os amo muito!

Gostaria de expressar minha profunda gratidão ao meu irmão querido, Tales, e a sua esposa, Janaína, que foram uma fonte constante de incentivo e apoio durante todo o processo de elaboração desta dissertação. Tales não só é meu irmão, mas também meu melhor amigo e maior inspiração. Ele sempre esteve disponível para me ajudar com seu vasto conhecimento e experiência, e sua presença tem sido uma grande referência profissional e acadêmica para mim. Além disso, agradeço a Janaína, que também está enfrentando os desafios do mestrado e tem sido uma companheira valiosa nesta jornada. Sou muito grato por tê-los em minha vida.

Quero prestar uma homenagem especial ao meu avô Joaquim, que, apesar de já ter falecido, continua sendo uma referência paterna para mim. Suas memórias e ensinamentos continuam vivos em mim e me ajudaram a chegar até aqui. Sempre me lembrarei dos seus conselhos sábios e da sua presença amorosa, que fizeram uma grande diferença na minha vida. Sou grato por tê-lo tido como avô e por tudo o que ele me ensinou.

À minha vó Iris, gostaria de agradecer pelo carinho que sempre me ofereceu. Seu

amor incondicional e sua dedicação foram uma fonte constante de apoio e inspiração para mim. Sou muito grato por tê-la em minha vida.

Agradeço do fundo do meu coração ao meu avô Antônio, que já partiu deste mundo, e à minha avó Maria, que também já nos deixou. Mesmo não tendo tido a oportunidade de conhecê-la, eles são parte fundamental da minha essência e da minha história de vida. Sou grato por todo o amor e ensinamentos que eles transmitiram aos meus pais e à minha família como um todo. Seus valores e legado continuam vivos em mim e sempre serão lembrados com carinho e gratidão.

Gostaria de expressar minha sincera gratidão ao meu sogro, Guilherme, e a minha sogra, Alda, por todo o apoio e carinho que recebi deles durante o período de elaboração desta dissertação. Sou muito grato por tê-los como parte da minha família. Agradeço aos meus companheiros da Modelagem, que foram fundamentais nesta jornada de aprendizado e crescimento. Trocamos informações, experiências e dificuldades, o que nos ajudou a superar os desafios.

Quanto ao meu orientador, Leonardo, não tenho palavras para expressar minha gratidão. Ele foi um guia sábio e paciente, sempre disponível para me ajudar em qualquer momento, seja por mensagens ou reuniões. Mesmo quando eu o procurava em horários pouco convencionais, ele estava lá, pronto para me ajudar a tirar dúvidas ou debater temas importantes. Agradeço por todo o apoio e pela confiança que depositou em mim ao longo deste processo e por ser um exemplo inspirador de excelência acadêmica e profissional.

Gostaria de expressar minha gratidão à Universidade Federal de Juiz de Fora e à Embrapa pelo apoio acadêmico, financeiro e estrutural durante todo o período em que estive envolvido neste estudo. A universidade me proporcionou um ambiente rico em conhecimento, com professores dedicados e recursos essenciais para a realização da pesquisa. Além disso, a Embrapa me deu a oportunidade de fazer parte de um projeto inovador e desafiador, o que me permitiu aplicar meus conhecimentos em um contexto real e relevante. Sou grato por toda a ajuda e suporte que recebi dessas instituições e por ter tido a oportunidade de contribuir para o avanço do conhecimento em minha área de pesquisa.

Mais uma vez, agradeço a todos que de alguma forma contribuíram para a realização deste projeto. Foi uma jornada desafiadora, mas também muito gratificante, e não teria sido possível sem o apoio de cada um de vocês.

“Não se coloque dentro de uma forma, se adapte e construa sua própria, e deixa-a expandir, como a água. Se colocarmos a água num copo, ela se torna o copo; se você colocar água numa garrafa, ela se torna a garrafa. A água pode fluir ou pode colidir. Seja água, meu amigo“.

Bruce Lee.

RESUMO

Os inúmeros fatores que afetam a contaminação de mastite subclínica em vacas em ambientes controlados, como os *Compost Barns*, tornam o estudo dessa detecção um desafio complexo. Embora existam pesquisas na literatura que se concentrem na detecção de mastite subclínica por meio de aprendizado de máquina, com uma variedade de modelos e variáveis, a aplicação desse tipo de técnica para a detecção da mastite subclínica ainda é pouco explorada, especialmente em um contexto de aprendizado de máquina automatizado. Portanto, o uso de modelos de aprendizado de máquina automatizado para a detecção da mastite subclínica representa um tópico de estudo com grande potencial. Inspirado por essa lacuna, este estudo tem como objetivo aplicar técnicas avançadas de aprendizado de máquina automatizado para aprimorar a detecção precoce da mastite subclínica em bovinos leiteiros mantidos em *Compost Barns*, com ênfase no bem-estar animal e na identificação precoce da doença. Além disso, busca-se auxiliar os produtores a reduzir os impactos financeiros significativos causados por essa doença. O aprendizado de máquina desempenha um papel crucial na detecção da mastite, permitindo a análise eficiente de padrões sutis nos dados dos bovinos leiteiros. Essa abordagem viabiliza a identificação antecipada da doença, resultando em intervenções rápidas que preservam o bem-estar dos animais e reduzem os custos associados a tratamentos tardios. A capacidade de processar grandes volumes de informações contribui para uma pecuária mais saudável e produtiva. Para abordar essa questão, o estudo propôs a aplicação de quatro modelos de aprendizado de máquina automatizado: AutoGluon, H2O, TPOT e uma combinação dos três. Esses modelos foram testados em três conjuntos de dados distintos e demonstraram resultados promissores em todas as análises. Vale destacar a eficácia da abordagem H2O, que alcançou um desempenho de 0,80% na métrica de Área sob a Curva ROC usando apenas 12 variáveis. Por outro lado, o TPOT também revelou grande potencial devido à sua programação genética. O melhor resultado percentual, independentemente da quantidade de variáveis, foi obtido com a combinação dos três modelos, alcançando 0,81% na métrica de Área sob a Curva ROC, ao ser aplicado a um conjunto de dados composto por 24 variáveis.

Palavras-chave: Aprendizagem de máquina automatizada. Mastite. *Compost Barns*

ABSTRACT

The numerous factors that affect subclinical mastitis contamination in cows within controlled environments, such as *Compost Barns*, make the study of its detection a complex challenge. While there is research in the literature that focuses on subclinical mastitis detection through machine learning using various models and variables, the application of such a technique for subclinical mastitis detection remains relatively unexplored, especially in the context of automated machine learning. Therefore, using automated machine learning models for subclinical mastitis detection is a promising research topic. Inspired by this gap, this study aims to apply advanced automated machine-learning techniques to enhance early subclinical mastitis detection in dairy cattle housed in *Compost Barns*, focusing on animal welfare and early disease identification. Additionally, it aims to assist producers in reducing the significant financial impacts caused by this disease. Machine learning is crucial in mastitis detection, enabling efficient analysis of subtle patterns in dairy cattle data. This approach facilitates early disease identification, resulting in prompt interventions that preserve animal welfare and reduce costs associated with late treatments. The ability to process large volumes of information contributes to a healthier and more productive livestock industry. To address this issue, the study proposed applying four automated machine learning models: AutoGluon, H2O, TPOT, and a combination. These models were tested on three distinct datasets and demonstrated promising results in all analyses. It is worth highlighting the effectiveness of the H2O approach, which achieved a performance of 0.80% in the Area under the ROC Curve metric using only 12 variables. On the other hand, TPOT also showed great potential due to its genetic programming. The best percentage result, regardless of the number of variables, was achieved with the combination of the three models, reaching 0.81% in the Area under the ROC Curve metric when applied to a dataset composed of 24 variables.

Keywords: Automated Machine Learning. Mastitis. Compost Barns.

LISTA DE ILUSTRAÇÕES

Figura 1 - Evolução da Produção do Leite.	16
Figura 2 - Sistema de confinamento do tipo <i>Tie Stall</i>	22
Figura 3 - Sistema de confinamento do tipo <i>Loose Housing</i>	23
Figura 4 - Sistema de confinamento do tipo <i>Loose Housing</i> . - Área de descanso . .	24
Figura 5 - Sistema de confinamento do tipo <i>Free Stall</i>	25
Figura 6 - Sistema de confinamento do tipo <i>Compost Barn</i>	26
Figura 7 - Conjunto de exaustores do <i>Compost Barn</i>	32
Figura 8 - Lonas defletoras mais área de alimentação do <i>Compost Barn</i>	32
Figura 9 - Vaca com teto removido devido à contaminação por <i>Clostridium perfringens</i> . .	36
Figura 10 - Estratégia de empilhamento em várias camadas do AutoGluon, um caso de duas camadas de empilhamento e n tipos de modelos de aprendizado de máquina ⁽¹⁾	42
Figura 11 - Pipeline otimizado baseado em árvore.	45
Figura 12 - Tela de login.	49
Figura 13 - Menu do sistema.	50
Figura 14 - Cadastro de Usuário.	50
Figura 15 - Dashboard de Peso.	51
Figura 16 - Dashboard de Controle leiteiro.	52
Figura 17 - Conjunto de dashboards do sistema.	53
Figura 18 - Dashboard de Ocorrências dos Patógenos de Mastite.	55
Figura 19 - Dashboard de Mastite.	56
Figura 20 - Conjunto de dashboards de Mastite do sistema.	57
Figura 21 - Tela de Predição de Mastite.	57
Figura 22 - Dashboard de Mastite.	58
Figura 23 - Dashboard de Mastite.	59
Figura 24 - Estrutura do <i>Compost Barn</i> da Embrapa.	61
Figura 25 - Sonda TU utilizada no composto.	61
Figura 26 - TERMO-HIGRO-ANEMÔMETRO DIGITAL PORTÁTIL.	62
Figura 27 - Localização do <i>Compost Barn</i>	63
Figura 28 - Processo de união dos dados.	64
Figura 29 - Matriz de correlação das 24 variáveis.	68
Figura 30 - Matriz de Confusão para base de dados de seis variáveis.	71
Figura 31 - Métrica ROC Curve para seis variáveis.	74
Figura 32 - Matriz de Confusão para base de dados de 12 variáveis.	75
Figura 33 - Métrica ROC Curve para 12 variáveis.	77
Figura 34 - Matriz de Confusão para base de dados de 24 variáveis.	79
Figura 35 - Métrica ROC Curve para 24 variáveis.	80

LISTA DE TABELAS

Tabela 1 – Comparação de Custos entre o Sistema Tradicional e o Sistema de Compost Barn.	28
Tabela 2 – Comparação de índices técnicos/gerenciais e econômicos para sistemas de produção de <i>Compost Barn</i> e <i>free stall</i> localizados no estado de Minas Gerais.	30
Tabela 3 – Melhorias identificadas a serem feitas.	31
Tabela 4 – Benefícios do aumento do conforto das vacas	31
Tabela 5 – Comparativo da influência da mastite em leite com alta contagem de células somáticas (CCS)	35
Tabela 6 – Estimativa da representação da célula somática em relação à incidência da mastite subclínica	38
Tabela 7 – Importância das variáveis	66
Tabela 8 – Métricas para base de dados de seis variáveis	73
Tabela 9 – Métricas para base de dados de 12 variáveis	75
Tabela 10 – Métricas para base de dados de 24 variáveis	79

LISTA DE ABREVIATURAS E SIGLAS

ANOVA	Análise de Variância
API	Interface de Programação de Aplicação / Application Programming Interface
AutoML	Aprendizado de máquina automatizado
CB	<i>Compost Barn</i>
CMT	<i>California Mastitis Test</i>
CCS	Contagem de células somáticas
Embrapa	Empresa Brasileira de Pesquisa Agropecuária
FN	Falso Negativo
FP	Falso Positivo
FS	<i>Free Stall</i>
FSS	<i>Feature Set Selector</i>
GLDG	Grau de Liberdade Dentro dos Grupos
GLEG	Grau de Liberdade Entre Grupos
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
ILSVRC	<i>Large Scale Visual Recognition Challenge</i>
Inmetro	Instituto Nacional de Metrologia, Qualidade e Tecnologia
IoT	Internet das Coisas
JSON	Notação de Objetos JavaScript
KG	Quilograma
LED	Diodo Emissor de Luz
LH	<i>Loose Housing</i>
ML	<i>Machine Learning</i>
ODS	Objetivo de Desenvolvimento Sustentável
OOF	<i>Out-of-fold</i>
SQDG	Soma dos Quadrados Dentro dos Grupos
SQEG	Soma dos Quadrados Entre Grupos
TS	<i>Tie Stall</i>
UCP	Unidade Central de Processamento
UFJF	Universidade Federal de Juiz de Fora
UPG	Unidade de Processamento Gráfico
VDG	Variância Dentro Grupos
VEG	Variância Entre Grupos
VGG	<i>Visual Geometry Group</i>
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1	INTRODUÇÃO	14
1.1	APRESENTAÇÃO DO TEMA E CONTEXTUALIZAÇÃO DO PROBLEMA	14
1.2	JUSTIFICATIVA	17
1.3	OBJETIVOS	19
1.3.1	Objetivo Geral	19
1.3.2	Objetivos Específicos	19
1.4	ORGANIZAÇÃO DO TEXTO	19
2	SISTEMA DE CONFINAMENTO PARA BOVINOS LEITEIROS	21
2.1	TIE STALL	21
2.2	LOOSE HOUSING	22
2.3	FREE STALL	23
2.4	COMPOST BARN	25
2.4.1	<i>Instalação do Compost Barn na Embrapa</i>	29
3	MASTITE	33
3.1	MASTITE CLÍNICA	36
3.2	MASTITE SUBCLÍNICA	37
4	APRENDIZAGEM DE MÁQUINA	39
4.1	APRENDIZADO DE MÁQUINA AUTOMATIZADO	39
4.1.1	<i>AutoGluon</i>	40
4.1.2	<i>H2O AutoML</i>	43
4.1.3	<i>TPOT</i>	44
5	SISTEMA	47
5.1	API	47
5.2	SISTEMA WEB	48
6	EXPERIMENTO	60
6.1	AQUISIÇÃO DOS DADOS	60
6.2	BASE DE DADOS	63
6.3	RESULTADOS E DISCUSSÃO	67
6.3.1	Experimentos computacionais	67
6.3.2	Análises	69
<i>6.3.2.1</i>	<i>Análise na base de dados de seis variáveis</i>	<i>71</i>
<i>6.3.2.2</i>	<i>Análise na base de dados de 12 variáveis</i>	<i>73</i>
<i>6.3.2.3</i>	<i>Análise na base de dados de 24 variáveis</i>	<i>76</i>
6.3.3	Vantagens e Limitações	80
7	CONCLUSÃO	82
7.1	TRABALHOS FUTUROS	83
	REFERÊNCIAS	84

1 INTRODUÇÃO

1.1 APRESENTAÇÃO DO TEMA E CONTEXTUALIZAÇÃO DO PROBLEMA

Por milênios, as marcas do consumo de leite e seus derivados remontam a eras distantes, datando de aproximadamente 4 mil anos a.C. O valor nutricional notável do leite de mamíferos, incluindo cabras, búfalas e, notadamente, vacas, tem sido uma força impulsionadora por trás da adoção generalizada desse produto lácteo. Impregnando uma escala global, essa preferência impulsionou não apenas a expansão das opções disponíveis, mas também uma inclusão cada vez mais profunda nas dietas de variadas culturas em todo o mundo (2).

Nesse contexto, o leite surge como um reservatório nutricional abrangente, enriquecido com elementos fundamentais para o desenvolvimento e crescimento de todas as fases da vida, desde a infância até a idade adulta. Para além do estímulo ao crescimento, ele também desempenha um papel crucial na manutenção da saúde dos ossos, músculos e dentes. Adicionalmente, o leite reforça a capacidade do organismo de enfrentar doenças infecciosas, conferindo uma variedade de vantagens para a saúde (3).

De acordo com Siqueira⁽⁴⁾, a produção anual de leite atinge 816 milhões de toneladas globalmente. Em média, cada habitante consome cerca de 116,5 quilogramas (kg) de leite por ano. Esse consumo tem aumentado anualmente, com crescimento médio de 1,2% desde 1999. A diversidade de consumo é notável, com variações marcantes entre nações. Enquanto algumas, como Nova Zelândia, Austrália, Ucrânia e países da União Europeia, consomem mais de 300 kg de leite anualmente por habitante, outras, como Indonésia e diversas nações africanas, registram consumos inferiores a 30 kg por habitantes por ano.

Segundo a Associação Brasileira da Indústria de Alimentos, no ano de 2022, o setor de laticínios no Brasil registrou um faturamento de R\$ 100,9 bilhões, representando uma parcela significativa de 9,4% do faturamento global da indústria de alimentos, que atingiu a marca de R\$ 1,075 trilhão¹. De acordo com as informações da Embrapa (5), o cenário da pandemia desempenhou um papel notável na drástica diminuição do consumo médio de leite por habitante no Brasil. No ano de 2020, esse indicador registrava 172 litros anuais por pessoa, e, em 2021, reduziu-se para 170 litros, chegando a 162 litros em 2022.

O consumo aparente per capita no Brasil atingiu 166,4 litros por habitante em 2018, situando-se abaixo dos níveis de países desenvolvidos, porém consideravelmente acima da realidade de duas décadas atrás (4). No ano de 2017, a média de consumo de leite foi de 175 equivalentes quilogramas por habitante. Embora não haja uma diretriz global específica, o padrão de consumo brasileiro está alinhado a orientações de diversas nações,

¹ Informação obtida através da assessoria de comunicação da Associação Brasileira da Indústria de Alimentos no dia 31/10/2023 através do telefone de contato divulgado no site <https://www.abia.org.br/>.

destacando-se o crescimento de mais de 60% no consumo de produtos lácteos no Brasil desde 1990, com ênfase no consumo de leite longa vida e queijos, conforme evidenciado por Siqueira⁽⁴⁾.

A Figura 1a apresenta informações do Instituto Brasileiro de Geografia e Estatística (IBGE) entre os anos 1997 e 2022, baseadas em dados de estabelecimentos inspecionados por governo federal, estadual ou municipal para garantir padrões de segurança. Esses dados mostram a quantidade de leite cru, tanto resfriado quanto não resfriado. Por sua vez, a Figura 1b destaca a quantidade de leite cru, resfriado ou não, que passou por processos de transformação industrial. Essas ilustrações fornecem uma visão clara da produção e processamento de leite ao longo do tempo no Brasil.

A produção de leite no Brasil é fundamentada em diversos sistemas, entre os quais se destacam três sistemas de produção a pasto: extensivo, semi-intensivo e intensivo (6). O sistema extensivo caracteriza-se pela ausência de períodos de descanso da pastagem ao longo do ano. Já o sistema semi-intensivo adota a prática de pastejo rotacionado, com períodos de descanso que favorecem a produção de forragem, por meio da subdivisão do pasto em piquetes que são utilizados sequencialmente. Por sua vez, o sistema intensivo assemelha-se ao semi-intensivo, mas destaca-se pela melhor adesão às tecnologias e opções de manejo, tais como *Free Stall*, *Tie Stall*, *Loose Housing* e *Compost Barn* (7).

O sistema de criação conhecido como *Compost Barn* foi inicialmente desenvolvido nos Estados Unidos e posteriormente adotado em outros países, como Israel e Brasil. Consiste em celeiros com áreas de descanso compostas principalmente de material orgânico, onde ocorre o processo de compostagem. A área de alimentação é pavimentada com concreto e fornece aos animais acesso a alimentos e água. A estrutura oferece liberdade de movimento, permitindo que os animais se deitem em diferentes posições na cama. Esse sistema apresenta benefícios em termos de saúde animal, incluindo melhorias nos índices de higiene e claudicação, bem como um aumento na produtividade (8).

Além dos proprietários, os consumidores também estão cada vez mais preocupados com os métodos aplicados aos animais e a qualidade dos produtos gerados, buscando novos valores ao consumir. Dessa forma, os conceitos relacionados ao bem-estar de rebanhos leiteiros são essenciais para alcançar um produto com qualidade e segurança ainda melhores para ambos: consumidores e animais.

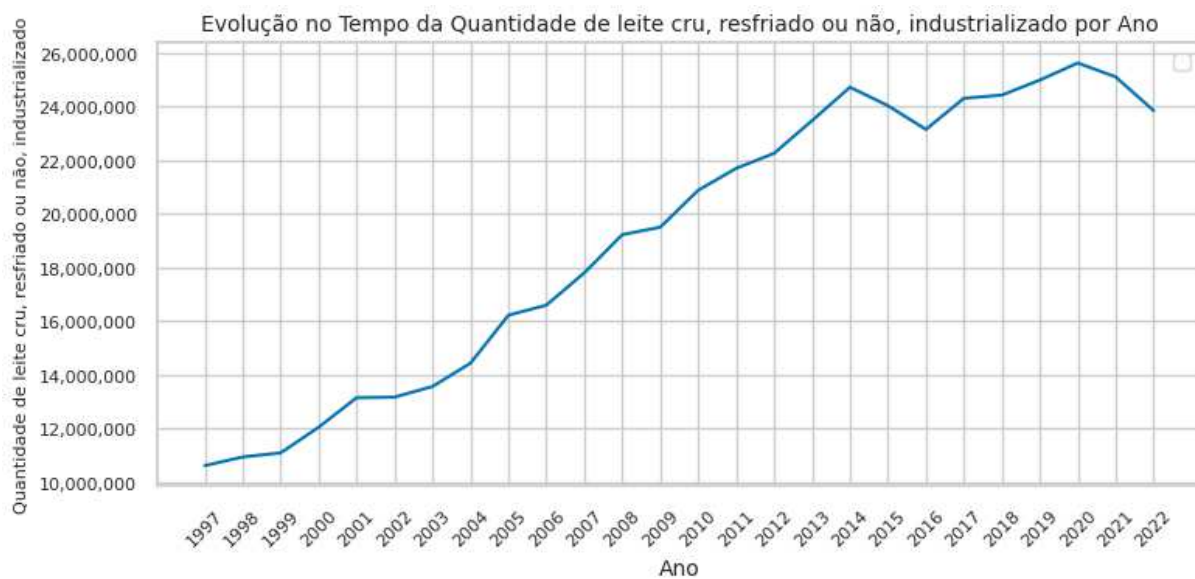
Para tornar isso possível, é necessário aplicar algumas medidas de biossegurança externa e interna na propriedade. As medidas de biossegurança são essenciais para fornecer o bem-estar dos animais. O controle deve ser mantido ao realizar diagnósticos para que seja possível monitorar as doenças e outras necessidades dos animais, permitindo a tomada de decisões mais eficientes.

Hoje em dia, empresas diversas utilizam o conceito de Internet das Coisas (IoT) para coletar dados por meio de uma rede de dispositivos conectados que geram informações

Figura 1 - Evolução da Produção do Leite.



(a) Evolução no Tempo da Quantidade de leite cru, resfriado ou não, adquirido por Ano.



(b) Evolução no Tempo da Quantidade de leite cru, resfriado ou não, industrializado por Ano.

Fonte: Elaborada pelo autor (2023).

constantes. A análise dessas informações geralmente produz conhecimento sobre o domínio de aplicação, levando as empresas a tomar decisões orientadas por dados. Uma medida de decisão baseada em dados foi desenvolvida em um estudo. Como resultado, estatisticamente, quanto mais orientada a dados for a empresa, mais produtiva ela será (9).

Algumas empresas têm realizado pesquisas que combinam a IoT, a Inteligência Artificial e o uso de campos de *Compost Barn* para tomar decisões baseadas em dados que garantam o bem-estar das vacas leiteiras. Dessa maneira, utilizando a ciência de dados, é possível tomar decisões que beneficiem a saúde dos animais e das pessoas que

os gerenciam, cuidam e consomem seus produtos, como a identificação de problemas de saúde, incluindo a mastite (10), a previsão de produtos gerados pelos animais (11) e a verificação de possíveis sazonalidades na produção (12).

Nesse contexto, segundo Nielsen (13), faz-se necessário entender que a mastite é uma inflamação da glândula mamária em vacas, causada por microorganismos patogênicos que entram no canal do teto e produzem toxinas danosas para a glândula. Isso pode resultar em alterações na composição do leite e diminuição da qualidade da gordura. A doença é multifatorial e sua incidência depende de exposição a patógenos, eficácia dos mecanismos de defesa da glândula mamária e presença de fatores de risco ambientais.

A mastite em vacas é um problema de saúde que pode causar impactos financeiros significativos para os produtores. O tratamento da mastite pode envolver o uso de antibióticos, e, durante o período de recuperação, ocorre uma redução na produção de leite, podendo chegar até 7%, além do descarte total do leite em alguns casos. Essas consequências têm um impacto negativo no bem-estar dos animais e no desempenho financeiro da produção leiteira (14).

Sendo assim, há a necessidade de aprimoramento do bem-estar dos animais, e a utilização de técnicas de aprendizagem de máquina para a classificação dos animais acometidos por mastite e os saudáveis é fundamental para que haja uma tomada de decisão rápida e precisa a fim de evitar quaisquer impactos para os animais e produtores. Nesse sentido, a parceria entre a Embrapa Gado de Leite e a Universidade Federal de Juiz de Fora (UFJF) é essencial para o sucesso do projeto PECUÁRIA DIGITAL DE PRECISÃO - DECISION e-LIVESTOCK, que tem como objetivo a detecção de mastite em animais por meio de uma *Application Programming Interface* - API que utilizará modelos matemáticos junto com inteligência artificial.

De forma geral, uma API desempenha o papel de uma conexão que viabiliza a interação harmonizada entre sistemas diversos. Seu valor é notório quando um desenvolvedor busca aproveitar os recursos ou serviços de um aplicativo ou plataforma, sem necessitar mergulhar nas minúcias intrincadas dessa plataforma. As APIs permitem requisitar dados, transmitir informações ou até mesmo orquestrar operações em diferentes programas ou sistemas.

A plataforma desenvolvida proporcionará aos produtores rurais uma ferramenta valiosa para o gerenciamento da saúde de seus rebanhos, de forma simplificada, rápida e de fácil acesso.

1.2 JUSTIFICATIVA

Esta dissertação tem como objetivo central investigar a detecção de mastite em bovinos leiteiros, enquadrando esta pesquisa no contexto do Objetivo de Desenvolvimento

Sustentável (ODS) número 2: Fome Zero e Agricultura Sustentável e número 3: Saúde e bem-estar. Inserido na agenda global de desenvolvimento sustentável da Organização das Nações Unidas, o ODS 2 não somente busca eliminar a fome, mas também promover práticas agrícolas de longo prazo que otimizem a produção de alimentos de maneira ecologicamente consciente e socialmente inclusiva. Esse objetivo almeja garantir o acesso universal a alimentos seguros, acessíveis e nutritivos, ao mesmo tempo que impulsiona transformações positivas nos sistemas alimentares e agrícolas.

A mastite, uma condição inflamatória que acomete as glândulas mamárias de bovinos, tem efeitos negativos tanto na produção de leite quanto na qualidade do produto final. Detectar essa enfermidade de forma precoce e eficaz é crucial para preservar a saúde do rebanho e garantir a qualidade do leite produzido. Essa abordagem se harmoniza de maneira coesa com diversos aspectos do ODS número 2:

- **Aprimoramento da Produção Agrícola Sustentável:** a mastite impacta negativamente a produtividade leiteira, acarretando prejuízos financeiros consideráveis aos agricultores. A identificação e o tratamento precoces da mastite representam contribuições significativas para manter a produção de leite em níveis sustentáveis.
- **Promoção da Saúde Animal:** o combate à mastite exige adoção de práticas de manejo e higiene criteriosas, cruciais para zelar pela saúde global do rebanho. Essa ênfase está diretamente ligada à promoção de práticas agrícolas sustentáveis que visam ao bem-estar e à saúde dos animais.
- **Garantia da Segurança Alimentar:** dado que o leite é uma fonte fundamental de nutrição, sua qualidade está intrinsecamente relacionada à saúde das vacas leiteiras. A detecção e o tratamento eficazes da mastite, sem dúvida, contribuem para uma produção de leite segura e rica em nutrientes, sustentando a segurança alimentar da população.
- **Incentivo a Práticas Sustentáveis:** a detecção precoce da mastite tem o potencial de reduzir a necessidade excessiva de antibióticos, promovendo práticas agrícolas sustentáveis. Essa abordagem está alinhada com os esforços globais de redução do uso indiscriminado de produtos químicos na agricultura.

Dessa maneira, esta dissertação fundamenta sua justificativa na conexão direta com o ODS número 2, Fome Zero e Agricultura Sustentável. Esse alinhamento é particularmente relevante no cenário brasileiro e em outras nações que dependem da atividade pecuária leiteira como componente essencial de suas práticas agrícolas.

Além disso, assume um papel crucial na realização do Objetivo de Desenvolvimento Sustentável número três, que visa assegurar uma vida saudável e promover o bem-estar para todas as pessoas. A mastite bovina, é uma preocupação crítica na produção de leite

e pode afetar diretamente a saúde dos rebanhos e a qualidade dos produtos lácteos. Este estudo emprega a inovação da inteligência artificial para a identificação precoce e precisa da mastite, possibilitando tratamentos mais eficazes e, por conseguinte, contribuindo para a saúde dos animais e a segurança alimentar.

Ao concentrar-se na detecção da mastite em sistemas de produção Compost Barn, essa pesquisa alinha-se também com a visão do ODS 3 de promover a saúde e o bem-estar tanto no âmbito humano quanto no animal. A aplicação de técnicas avançadas de inteligência artificial neste contexto não apenas aprimora o manejo dos rebanhos, mas também atua como uma medida preventiva importante para a saúde pública, garantindo que os produtos lácteos sejam seguros e de alta qualidade. Portanto, esse estudo desempenha um papel vital na realização dos objetivos de saúde e bem-estar sustentáveis, demonstrando o impacto positivo que a tecnologia pode ter na agricultura e na vida das pessoas.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo principal deste trabalho reside na proposição de quatro modelos de aprendizado de máquina automatizado, destinados à classificação da mastite subclínica em sistemas de confinamento do tipo *Compost Barns*, bem como à classificação do risco global associado ao rebanho.

1.3.2 Objetivos Específicos

- Construir uma base de dados através das diferentes fontes de informações.
- Desenvolver de forma integrada a API com o sistema web, incorporando os modelos de aprendizado de máquina, a fim de criar uma plataforma abrangente que disponibilize aos produtores rurais uma poderosa ferramenta para tomada de decisões estratégicas.
- Comparar o desempenho dos modelos de Aprendizado de Máquina Automatizado (AutoML): AutoGluon, H2O, TPOT e a fusão desses três modelos através das métricas de Acurácia, Recall, F1, Precisão e Área sob a Curva ROC.

1.4 ORGANIZAÇÃO DO TEXTO

A presente dissertação está dividida em sete capítulos. O primeiro capítulo apresenta a introdução do trabalho. O segundo capítulo aborda o cenário em que os animais estão confinados. No terceiro capítulo, aborda-se a doença da mastite. No quarto capítulo, serão descritos os modelos de aprendizado de máquina utilizados para o desenvolvimento do trabalho. O quinto capítulo apresenta informações sobre o sistema web e a API desenvolvida. Já o sexto consiste em uma análise sobre o experimento da dissertação, contendo a forma

como os dados foram adquiridos e a análise sobre os resultados encontrados. Por fim, o sétimo conclui este trabalho, abordando as principais contribuições deste estudo e sugerindo possibilidades para trabalhos futuros.

2 SISTEMA DE CONFINAMENTO PARA BOVINOS LEITEIROS

Este capítulo tem como objetivo fornecer uma visão geral dos diferentes tipos de confinamento para bovinos leiteiros, com um foco maior no sistema de *Compost Barn*. O composto é o ambiente onde o experimento em questão foi executado. Além do composto, também abordaremos outros sistemas comumente utilizados, como as baias individuais conhecidas como *tie stall* e os sistemas de estabulação livre, como *loose housing* e *free stall*.

A evolução do confinamento para bovinos leiteiros apresenta marcos transformadores. A partir de 1946, o confinamento ganhou adeptos entre produtores de leite, embora já houvesse estruturas para abrigar bovinos desde 1880, inicialmente usadas para armazenar colheitas e gado. Durante a II Guerra Mundial, muitos agricultores recorreram a galpões devido ao clima adverso, impulsionando a transição de pastoreio para sistemas mais tecnológicos (15).

Nos anos 50, os agricultores adotaram o *loose housing*, proporcionando áreas de descanso para as vacas. No entanto, problemas de manutenção higiênica limitaram sua adoção em algumas regiões. Surgiu então o sistema *tie stall*, bem aceito em Minnesota e Wisconsin, aproveitando a tradição local de silos verticais para armazenamento de alimentos. O sistema *free stall* emergiu na metade do século XX como uma alternativa para rebanhos diferentes, se tornando-se altamente recomendado e se expandindo globalmente (16, 15).

No Brasil, a Embrapa Gado de Leite introduziu sistemas de confinamento nos anos 80, destacando-se o *free stall*, amplamente adotado com apoio da Embrapa. Em Minnesota, o sistema *Compost Barn* surgiu em 2001, priorizando bem-estar animal e sustentabilidade. Baseado na aeração da cama, ganhou popularidade nos EUA e em países como o Brasil (15).

O *Compost Barn* chegou ao Brasil via fazenda Santa Andrea, em Itararé – SP, destacando-se por custo menor e conforto animal. O sucesso impulsionou mais de 30 instalações até 2014, incluindo Minas Gerais, valorizando a ecologia e a viabilidade econômica. Apesar do crescimento, a pesquisa científica é escassa, o que evidencia a necessidade de estudos para orientar produtores (17).

2.1 TIE STALL

O *tie stall* é um sistema de confinamento de rebanho que utiliza baias individuais, onde as vacas ficam lado a lado, dentro de suas respectivas baias. Na maior parte do tempo, são mantidas presas por uma corrente no pescoço e ficam soltas geralmente na hora da ordenha (15). Esse sistema era o principal tipo de alojamento para os bovinos nos Estados Unidos e, segundo seu Departamento de Agricultura, no ano de 2007, cerca de 62% dos produtores ainda utilizavam essas instalações(18).

De acordo com Morabito e Bewley⁽¹⁸⁾, os proprietários devem tomar precauções específicas ao construir ou reformar essas acomodações. Um projeto inadequado pode ter um impacto negativo na produção e no bem-estar desses animais. Por conseguinte, é de suma importância que o sistema seja dimensionado corretamente. Além disso, de acordo com os autores (18), uma das vantagens do *tie stall* é que ele requer um investimento inicial baixo e é altamente conveniente para serviços veterinários, reprodução e administração de tratamentos, graças às baias estacionárias disponíveis. Em contrapartida, Mota *et al.* ⁽¹⁵⁾ relatam que o *tie stall* é um sistema caracterizado por um alto investimento por animal e apresenta pouca eficiência no manejo do ambiente e do rebanho.

A desvantagem na ordenha do rebanho pode acarretar problemas graves, uma vez que exige muito esforço em seu manejo, podendo resultar em lesões e doenças nos animais(18). Outra preocupação nesse tipo de alojamento é a falta de interação e exercício do rebanho, já que passam a maior parte do tempo confinados, aumentando a possibilidade de estresses, como ilustrado na Figura 2.

Figura 2 - Sistema de confinamento do tipo *Tie Stall*.



Fonte: (15)

2.2 LOOSE HOUSING

Nos anos 50, houve um aumento no uso do sistema *loose housing* pelos pecuaristas como uma alternativa econômica e eficiente para a produção de leite. Geralmente, esse sistema inclui uma área separada para descanso do rebanho, que é totalmente distinta da área de ordenha. Além disso, há uma área de circulação onde as vacas podem se

movimentar livremente. A alimentação é fornecida em um local diferente da área de descanso (19).

Conforme mencionado por Lewis e Johnson⁽¹⁹⁾, é recomendado ter uma pavimentação adequada para facilitar a remoção dos dejetos dos animais e, para a área de livre movimentação e descanso do rebanho, sugere-se um piso de terra para melhor drenagem. Para diminuir o risco de lesões e assegurar o aumento no conforto dos animais na área de descanso, é recomendado que o piso seja coberto por palha, tapete de borracha ou outro material que seja macio e adequado para o conforto(20).

De acordo com Mota *et al.* ⁽¹⁵⁾, é de extrema importância que essas construções do *loose housing* tenham uma boa iluminação natural, uma boa ventilação e que a cama seja coberta o tempo todo com material seco para a proteção das vacas contra lesões e contra ventos frios e chuva. É necessário que diariamente seja feita a substituição da serragem que estiver molhada e suja. Não menos importante é necessário que a remoção do excesso das fezes seja realizada diariamente para manter o ambiente saudável para o rebanho. Na Figura 3, é possível observar um ambiente limpo e saudável para o rebanho.

Figura 3 - Sistema de confinamento do tipo *Loose Housing*.



Fonte: (20)

Pisos gradeados ou de cimento ou qualquer superfície dura, suja ou úmida, devem ser considerados inaceitáveis. Pode-se observar na Figura 4 uma área de descanso que está com nível de sujeira acima do ideal (20).

2.3 FREE STALL

Na década de 80, iniciou-se o processo de migração na utilização dos pecuaristas do sistema *tie stall* para o *free stall*. O *free stall* é um composto que tem como objetivo

Figura 4 - Sistema de confinamento do tipo *Loose Housing* - Área de descanso.



Fonte: (20)

principal tornar o confinamento das vacas mais confortável e fornecer um manejo mais fácil. Geralmente possuem três principais áreas para o rebanho: atividade física, descanso e alimentação. A qualidade de vida dos animais e seu desempenho são diretamente influenciados pelas particularidades dessas áreas (21).

De acordo com Mota *et al.* (15), a utilização desse sistema ocorreu devido à facilidade do manejo dos animais. Quando não estão no processo de ordenha, eles podem ficar livres, onde é disponibilizado alimentação de fácil acesso. Fernández; Mariño e Carreira (21) reforça que esses aspectos são de extrema importância para esse tipo de composto, a facilidade de manejo do rebanho e o bem-estar dele.

Para que esses aspectos sejam atendidos, é importante que as instalações do *free stall* sejam capazes de manter os animais com o mínimo de estresse possível. Para isso, é importante observar a radiação solar para evitar sobrecarga térmica dos animais. Levando isso em consideração, é necessário verificar os materiais utilizados na construção dessas estruturas, podendo ser o tipo de piso, as telhas, tamanho das camas e assim por diante, priorizando um espaço global que permita o máximo de conforto possível(15).

Em determinadas situações, é comum utilizar telhas de cimento de amianto como cobertura do composto. Dependendo da área, essas telhas são colocadas verticalmente para reduzir o impacto da radiação solar no ambiente. Quanto às camas, elas podem ser compostas por diferentes materiais. Nas imagens referentes às baias ilustradas na Figura 5, foram utilizados areia ou colchões de borracha (22).

Figura 5 - Sistema de confinamento do tipo *Free Stall*.



(a) Borracha picada



(b) Colchão de borracha



(c) Roletes recheados com borracha picada, entremeados com borracha



(d) Areia

Fonte: (22)

2.4 COMPOST BARN

O sistema *Compost Barn* é uma alternativa de confinamento para bovinos leiteiros, com o objetivo de fornecer um ambiente confortável e propício ao bem-estar animal, além de facilitar a gestão diária e melhorar a gestão dos dejetos. Consiste em um galpão com área de cama comum, que é preenchida geralmente com material orgânico, proporcionando um ambiente seguro e controlado para os animais, o que contribui para a redução de doenças e estresse térmico (23).

É válido mencionar que o sistema em questão surgiu no final da década de 1980, quando produtores de leite do estado de Virgínia, nos Estados Unidos, buscavam inovações para sua produção. Posteriormente se difundiu para outras partes do mundo, incluindo o Brasil. Embora implementado há relativamente pouco tempo, o sistema de *Compost Barn* tem se revelado altamente efetivo no que concerne ao bem-estar animal, tendo demonstrado resultados significativos nesse sentido. Ademais, é importante destacar que sua implementação tem gerado impactos positivos na produtividade dos animais, bem como em índices como higiene e claudicação(8).

Segundo Shane; Endres e Janni ⁽²⁴⁾, o sistema de *Compost Barn*, em linhas gerais, possui uma área destinada ao descanso dos animais, cujo leito é constituído predominantemente por material orgânico que passa por um processo de compostagem. Tal sistema proporciona às vacas maior liberdade de movimento e possibilidade de se deitarem em qualquer lugar que lhes seja conveniente. Ademais, a alimentação e a água são oferecidas em um corredor específico para esse fim, conforme é possível observar na Figura 6.

Figura 6 - Sistema de confinamento do tipo *Compost Barn*.



Fonte: Elaborada pelo autor (2023).

Em linhas gerais, os *Compost Barns*, compreendem uma área contendo uma cama revolvida duas vezes ao dia, circundada por um muro de 1,2 metros de altura. Essa cama é projetada para fornecer espaço suficiente para as vacas e é inicialmente preenchida com serragem fina ou serragem de madeira. Camadas adicionais de serragem fresca são incorporadas periodicamente. A cama é submetida a aeração duas vezes ao dia por meio de equipamentos adequados. Esse processo estimula a atividade biológica e mantém a superfície sempre fresca, proporcionando um ambiente higiênico e propício ao descanso das vacas ⁽²⁵⁾.

Segundo Janni *et al.* ⁽²⁵⁾, no contexto da compostagem, a área destinada à cama revolvida é dimensionada de forma a garantir o espaço adequado para as vacas, variando de 7,4 m² a 9,4 m² por vaca, dependendo do tamanho e das necessidades individuais. A área da cama é planejada de modo a acomodar simultaneamente todas as vacas em repouso, garantindo que elas possam se levantar para alimentar-se ou beber água. A quantidade de espaço requerido por vaca depende diretamente da quantidade de esterco e urina produzidos, sendo fundamental equilibrar a atividade microbiana com a taxa de secagem.

A largura do celeiro pode variar conforme as demandas do cocho e da área da cama, sendo comumente realizada a alimentação em corredores de concreto com 3,6 metros de largura. Os cochos seguem as recomendações típicas de celeiros do tipo *freestall*. Para garantir um acesso adequado à água, os bebedouros são posicionados ao longo do corredor de alimentação (25).

De acordo com Janni *et al.* (25), uma ventilação eficaz é um aspecto crítico para remover o calor e a umidade gerados tanto pelas vacas quanto pela área da cama revolvida. Os *Compost Barns* geralmente adotam uma ventilação natural e, portanto, requerem uma localização estratégica que permita a circulação de ar durante os meses de verão. Além disso, a altura das paredes laterais é uma consideração fundamental para promover a ventilação natural e fornecer espaço para a passagem de equipamentos. Recomenda-se que as paredes laterais dos celeiros de compostagem tenham uma altura de 4,9 metros para assegurar uma ventilação adequada e permitir o acesso de caminhões de cama fresca quando necessário. Em alguns *Compost Barns*, são utilizados ventiladores que direcionam o ar para baixo, auxiliando na secagem da superfície da cama. Para isso, esses ventiladores são instalados a uma altura que não interfira na operação dos equipamentos de revolvimento da cama.

Em um estudo de custos conduzido por Michels *et al.* (26), foi realizado um comparativo entre o sistema tradicional, em que os animais se alimentavam principalmente a pasto, e o sistema de *Compost Barn*. A propriedade adotou o sistema tradicional até o segundo semestre de 2017, e posteriormente, fez a transição para o *Compost Barn*. O resultado desse estudo pode ser encontrado na Tabela 1.

Mediante a análise dos dados apresentados na Tabela 1, torna-se evidente o incremento na produção de leite subsequente à transição do sistema tradicional para o *Compost Barn*. Concomitantemente, observa-se que os custos associados à alimentação experimentaram um aumento substancial, em virtude da liberdade de pastoreio do gado no sistema tradicional.

Conforme destacado no estudo conduzido por Silva *et al.* (27), a qual realizou uma comparação com o propósito de orientar produtores e técnicos na escolha da instalação mais adequada para suas propriedades, a análise minuciosa da rentabilidade dos sistemas de produção de leite, incluindo tanto o *Compost Barn* quanto o *free stall*, foi conduzida. Essa análise abarcou dados coletados ao longo do ano completo de 2016, abrangendo quatro fazendas, com uma distribuição equitativa entre elas.

A Tabela 2 apresenta uma comparação detalhada dos resultados deste estudo. Nela, podemos observar uma significativa redução na contagem de células somáticas e uma notável diminuição na contagem total de bactérias quando comparamos os dois sistemas. Além disso, houve uma melhoria notável na produtividade das vacas no sistema *Compost Barn* em comparação com o sistema de *free stall*. Vale ressaltar que a taxa de serviço e a

Tabela 1 – Comparação de Custos entre o Sistema Tradicional e o Sistema de Compost Barn.

	Tradicional		<i>Compost Barn</i>	
	1º Trimestre	2º Trimestre	3º Trimestre	4º Trimestre
Mão de obra	R\$10.500,00	R\$10.200,00	R\$10.200,00	R\$10.800,00
Encargos sociais	R\$5.636,36	R\$5.654,23	R\$6.246,88	R\$6.572,72
Ração/Alimentação	R\$90.662,13	R\$ 96.643,30	R\$158.847,00	R\$151.105,28
Sêmen/Material/Ultrassom	R\$1.380,00	R\$1.391,00	R\$2.303,07	R\$2.495,00
Medicamentos	R\$7.415,69	R\$8.535,90	R\$8522,01	R\$6.632,24
Material de ordenha	R\$3.857,32	R\$3.725,00	R\$2.705,25	R\$4.198,00
Combustível	R\$1.850,00	R\$3.581,24	R\$5.751,00	R\$5.325,13
Manutenção máquinas/Instalação	R\$555,00	R\$8.466,00	19.970,00	895,00
Energia elétrica/Telefone	R\$2.946,50	R\$2.777,78	R\$2.595,92	R\$4.595,11
Arrendamentos	R\$2.640,00	R\$2.640,00	R\$2.640,00	R\$2.640,00
Total	R\$127.443,00	R\$143.614,45	R\$219.781,13	R\$195.258,48
Leite produzido (Litros)	183.203	176.120	233.992	238.116
Custo/Litro	R\$0,70	R\$0,82	R\$0,94	R\$0,82

Fonte: Michels *et al.* ⁽²⁶⁾

taxa de concepção apresentaram equilíbrio relativo entre esses sistemas.

Em outro estudo realizado com 42 produtores de leite que utilizam *Compost Barn*, Damasceno ⁽²⁸⁾ observou a necessidade de melhoria no sistema de uma maneira geral. A Tabela 3 representa essas características que os produtores apontaram a importância de ser melhoradas. Além disso, o estudo de Damasceno ⁽²⁸⁾ constatou benefícios significativos na qualidade de vida dos animais, incluindo um aumento no conforto e uma melhora no escore de sujidade, conforme demonstrado na Tabela 4.

2.4.1 Instalação do Compost Barn na Embrapa

Na cidade de Coronel Pacheco em Minas Gerais, a Embrapa estabeleceu a primeira instalação de *Compost Barn* para pesquisa do tipo túnel de vento no Brasil, denominada Vacas e Pessoas Felizes, que iniciou suas operações em 12 de março de 2020. O referido complexo, cujo nome se deve aos resultados promissores que apresentou, como o conforto proporcionado tanto ao rebanho quanto aos trabalhadores diretamente envolvidos no sistema (29).

De acordo com a Embrapa ⁽²⁹⁾, em uma das extremidades, ficam 22 conjuntos de exaustores, conforme observado na Figura 7, os quais têm como objetivo puxar o ar de dentro para fora do ambiente, e essa corrente de vento gerada pode ser controlada de acordo com a necessidade do lugar. Na outra extremidade, há uma parede composta de placas evaporativas em forma de colmeia, na qual circula água, que ajuda na climatização do ambiente. Além disso, as laterais são fechadas por uma lona especial que reflete a luz do sol.

Devido ao impacto das condições tropicais na produção de leite, a temperatura interna do ambiente pode ser até 8 graus Celsius mais baixa em comparação com a temperatura externa, o que ajuda a reduzir o estresse térmico no rebanho. Além disso, o sistema de compostagem conta com um conjunto de lonas defletoras posicionadas em um ângulo de 45 graus em relação ao teto, conforme se observa na Figura 8. Essas lonas têm a função de direcionar o fluxo de vento de cima para baixo, mantendo uma temperatura mais fresca no ambiente(29).

Segundo a Embrapa ⁽²⁹⁾, o *Compost Barn* possui cinco metros de pé direito e tem capacidade de receber até cem vacas, esses animais são da raça Holandesa. A cama orgânica, composta de maravalha de eucalipto possui 1200 m² e 60 centímetros de profundidade. Possui ainda uma área em que o trator faz o revolvimento da cama pelo menos duas vezes ao dia e uma pista para a distribuição dos alimentos do rebanho.

O ambiente traz diversas tecnologias de ponta, como colares com chips para monitoramento dos animais em tempo real, controladores de luzes de Diodo Emissor de Luz (LED), câmeras, faz também recolhimento e armazenamento da água da chuva para reuso e possui captação de energia solar através das células fotovoltaicas (29).

Tabela 2 – Comparação de índices técnicos/gerenciais e econômicos para sistemas de produção de *Compost Barn* e *free stall* localizados no estado de Minas Gerais.

	<i>Compost Barn</i>		<i>Free stall</i>	
	Fazenda 1	Fazenda 2	Fazenda 3	Fazenda 4
Contagem de células somáticas (milhares de células por mL)	295	212	372	397
Contagem total de bactérias (milhares de unidades formadoras de colônias por mL)	20	28	98	272
Taxa de concepção (%)	35	43	26	33
Taxa de serviço (%)	72	63	59	70
Produtividade (kg/vaca/dia)	27,3	33,1	21,1	23,1
Produção de leite/área (kg/ha/ano)	29.487	23.245	27.288	33.087
Ponto de equilíbrio operacional diário (kg de leite)	1.253	1.050	764	1.334
Ponto de equilíbrio operacional anual (kg de leite)	458.493	384.219	279.618	488.143
Remuneração de capital sem animais e terra (% a.a.)	155.235,00	196.827,11	96.035	172.005
Remuneração de capital com terra (% a.a.)	407.235,00	535.227,11	291.035	282.405
Custo operacional efetivo/preço do leite (%)	73,07	62,93	66,05	72,66
Custo operacional total/preço do leite (%)	77,18	69,65	68,81	77,18
Despesas com concentrado/preço do leite (%)	26,53	30,80	32,97	27,62

Fonte: Extraída de Silva *et al.* ⁽²⁷⁾ e adaptada pelo Autor.

Tabela 3 – Melhorias identificadas a serem feitas.

Recomendação	Porcentagem (%)
Aumentar o tamanho ou capacidade do estábulo	25
Deixar as paredes laterais mais altas e melhor a ventilação	20
Adicionar muro de contenção	10
Adicionar cortinas	8.3
Inserir mais ventiladores	8.3
Aumentar a ventilação do telhado	8.3
Retirar postes da área de descanso	6.7
Alterar o número ou a localização dos bebedouros	6.7
Alterar a localização ou o tamanho do cocho de alimentação	6.7

Fonte: (28)

Tabela 4 – Benefícios do aumento do conforto das vacas

Benefício	Porcentagem (%)
Melhora no conforto das vacas	24.1
Melhora na limpeza das vacas	12.1
Baixa manutenção	9.5
Benéfico para novilhas, vacas com problemas de claudicação, vacas recém-paridas, vacas com problemas de saúde e vacas idosas	8.6
Posicionamento natural de descanso	7.8
Melhora nos cascos e nas pernas	6.9
Proximidade da sala de ordenha	6.9
Redução na contagem de células somáticas	5.2
Aumento na detecção de cio	5.2
Facilidade no manejo de esterco	2.6
Aumento no consumo de matéria seca	2.6
Aumento na produção	2.6
Aumento na longevidade	2.6
Diminuição de lesões nas pernas e tetas	1.7
Redução do tempo em pé sobre concreto	1.7

Fonte: (28)

Segundo a Embrapa ⁽³⁰⁾, após um ano da implementação do composto pela Embrapa, houve aumento na produção de leite, passando de 31,6 kg/vaca/dia para 37 kg/vaca/dia. O índice de mastite clínica ficou abaixo dos 2% desejáveis para propriedades leiteiras, com isso, houve uma redução na utilização de antibióticos.

Figura 7 - Conjunto de exaustores do *Compost Barn*.



Fonte: Elaborada pelo autor (2023).

Figura 8 - Lonas defletoras mais área de alimentação do *Compost Barn*.



Fonte: Elaborada pelo autor (2023).

3 MASTITE

Em 1787, com Von Willburg, começaram a ser realizados estudos científicos das doenças da glândula mamária. No entanto, esses estudos inicialmente se concentraram na classificação das doenças sem explorar os aspectos relacionados à sua causa, uma vez que, naquela época, ainda não se tinha conhecimento da existência de microrganismos (31).

Segundo (31), o interesse dos veterinários e criadores de gado leiteiro marcou o início de uma nova fase no estudo das doenças das glândulas mamárias, motivado pela necessidade de classificar as mastites de maneira prática e científica. Diversas classificações foram propostas ao longo do tempo para as mastites, visando identificar suas formas clínicas. Inicialmente, em 1835, Rychner categorizou as mastites em ativas, passivas e reumáticas. Já em 1867, Haubner dividiu-as em formas clínicas distintas, como superficial, profunda e parenquimatosa.

No ano de 1876, Von Franck realizou pesquisas pioneiras sobre a etiologia infecciosa das mastites, identificando microrganismos em amostras de leite de glândulas doentes. Entretanto, não determinou os agentes etiológicos específicos. A descoberta da microscopia óptica por Antoni van Leeuwenhoek em 1676 permitiu essas observações. No mesmo sentido, em 1885, Escherich identificou germes coliformes como possíveis causadores. Mais tarde, em 1885, Nocard e Mollereau evidenciaram germes do gênero *Streptococcus* em amostras de leite de glândulas afetadas (31).

Ao longo dos anos, surgiram outras classificações. Em 1910, Poels introduziu a primeira classificação etiológica com base nos agentes causadores das mastites. Já em 1908, Kitt e Glage avaliaram a patogênese, sintomas e alterações anatômicas, classificando-as em infecciosas e não infecciosas. Em 1931, Gótze propôs uma classificação baseada nas características da glândula e do leite. No ano de 1971, Schalm e demais colaboradores categorizaram mastites por formas clínicas (hiperaguda, aguda, subaguda e subclínica), mas os critérios foram mais relacionados à evolução da doença do que à manifestação clínica (31).

Segundo Vlieghe *et al.* (32), a mastite em novilhas, inicialmente descrita há décadas, só foi reconhecida como um problema significativo posteriormente. Sua prevalência global foi documentada em várias regiões, como América do Norte, Europa, Nova Zelândia, Japão e América do Sul. A mastite em novilhas difere daquela em vacas, tendo uma maior incidência nos primeiros dias após o parto. As bactérias causadoras variam com a idade, com novilhas apresentando mais *Streptococcus uberis* e *Staphylococcus coagulase-negativos*, e menos *Staphylococcus aureus* em comparação com animais mais velhos.

A mastite, de acordo com Massote *et al.* (33), é uma doença caracterizada pela inflamação da glândula mamária, podendo ser causada por diferentes fatores, como processos fisiológicos, traumas ou infecções. Por meio de uma pesquisa conduzida por

Watts ⁽³⁴⁾, foi possível identificar a presença de uma ampla variedade de 137 espécies microbianas, abrangendo subespécies e variações sorovares¹, no interior da glândula mamária bovina.

Conforme mencionado por Bradley ⁽³⁵⁾, a mastite é a doença de maior importância econômica no gado leiteiro, respondendo por 38% dos custos totais das doenças comuns de produção. Avaliar as perdas da mastite clínica é complexo devido aos custos de tratamento, descarte, óbito e redução na produção de leite. O custo médio de um caso é cerca de 175 libras esterlinas. Com uma média de 40 casos/100 vacas/ano e 2,4 milhões de vacas no rebanho nacional, a mastite custa mais de 168 milhões de libras esterlinas anualmente à indústria leiteira do Reino Unido.

Ainda segundo Bradley ⁽³⁵⁾, a taxa de mortalidade anual devido à mastite é de 0,6% das vacas em lactação. Também é a causa mais comum de morte em vacas leiteiras adultas. Avaliar as perdas da mastite subclínica é ainda mais difícil, englobando tratamento, redução da produção e risco de descarte. Philpot ⁽³⁶⁾ mostrou relação inversa entre contagem de células somáticas acima de 200.000 células/mL e produção (2,5% de queda na produção a cada 100.000 células/mL de aumento na contagem de células somáticas).

Afora os custos, tanto a mastite clínica quanto a subclínica prejudicam a fertilidade subsequente. Além das questões financeiras, a importância para a saúde pública não deve ser subestimada. O uso de antibióticos no tratamento e controle pode gerar cepas resistentes e risco na cadeia alimentar. Há potencial de disseminação de patógenos zoonóticos pelo leite, especialmente em mercados de produtos não pasteurizados e em falhas na pasteurização ⁽³⁵⁾.

O impacto no bem-estar dos animais também foi abordado, como ressaltado no Relatório de Bem-Estar de Vacas Leiteiras, conforme apresentado por Bradley ⁽³⁵⁾. Além disso, a mastite tóxica acarreta consequências evidentes, e estudos revelam a ocorrência de hiperalgesia após casos leves de episódios clínicos.

A infecção ocorre quando micro-organismos patogênicos entram na glândula mamária através da teta da vaca e conseguem superar o sistema imunológico, resultando na multiplicação desses micro-organismos, podendo causar danos ao tecido mamário ^(33, 37).

A mastite é uma doença complexa e multifatorial, influenciada pela exposição a diversos patógenos, bem como por fatores individuais do animal e do ambiente em que ele está confinado, que podem aumentar o risco de contaminação ^(33, 37). De acordo com Nielsen ⁽¹³⁾ e Harmon ⁽³⁸⁾, a lesão no tecido mamário causada pela mastite pode resultar em um aumento na permeabilidade vascular. Isso pode acarretar vários problemas para o leite, incluindo a diminuição da qualidade da gordura, alterações na composição, vazamento de sangue, enzimas e sais, além da redução na síntese de caseínas e lactose.

¹ Referentes a grupos de microrganismos, como bactérias, que compartilham características antigênicas específicas.

A Tabela 5 desenvolvida por Harmon ⁽³⁸⁾, descreve de forma numérica as alterações que ocorrem no leite quando um animal está com mastite.

Tabela 5 – Comparativo da influência da mastite em leite com alta contagem de células somáticas (CCS)

Constituinte	Leite Normal	Leite com alta CCS	Porcentagem do normal
			(%)
Sólido não gorduroso	8.9	8.8	99
Gordura	3.5	3.2	91
Lactose	4.9	4.4	90
Proteína total	3.61	3.56	99
Caseína total	2.8	2.3	82
Proteína de soro	0.8	1.3	162
Albumina sérica	0.02	0.07	350
Lactofenina	0.02	0.10	500
Imunoglobulinas	0.10	0.60	600
Sódio	0.057	0.105	184
Cloreto	0.091	0.147	161
Potássio	0.173	0.157	91
Cálcio	0.12	0.04	33

A mastite bovina gera consequências negativas para todo o ecossistema leiteiro. Os prejuízos acontecem com gastos com medicamentos para o tratamento do animal, serviços específicos de veterinários, baixa qualidade do leite gerado (o que se pode observar na Tabela 5 comentada anteriormente), diminuição na produção leiteira, descarte do leite contaminado, descarte do animal, quando o mesmo passa a ter uma incidência constante de mastite. Pode ocorrer também no pior cenário no qual o animal vem a óbito (39, 40, 13). Conforme mencionado por Da Fonseca *et al.* ⁽³⁹⁾, estudos apontam que a mastite bovina acarreta um significativo impacto econômico, estimado em cerca de 10% do faturamento das fazendas leiteiras no Brasil.

Segundo Brito e Sales ⁽⁴¹⁾, as bactérias *Streptococcus agalactiae*, *Streptococcus aureus*, *Streptococcus dysgalactiae*, *Streptococcus uberis* e *Escherichia coli* são os principais agentes causadores da mastite. A maioria dos casos de mastite é atribuída a essas bactérias, representando cerca de 95% das infecções. Podem-se classificar esses agentes causadores da mastite de contagiosos e ambientais.

A probabilidade de infecção da glândula mamária é maior durante o período o inicial e final da lactação e nas duas primeiras e duas últimas semanas do período seco ²(41). Ainda de acordo com Brito e Sales ⁽⁴¹⁾, foi observado em diversos lugares do mundo um acréscimo de infecções quando o clima se apresentava mais quente ou chuvoso. Além dessas

² O período seco da vaca compreende os dois últimos meses de gestação. - <https://is.gd/periodoseco> - Acessado em 29 abr. 2023

duas maneiras já citadas anteriormente, pode haver mais duas formas de novas infecções. Isso ocorre quando se introduz no rebanho uma novilha ou uma vaca já infectada.

Pode-se subdividir a mastite em duas categorias, mastite clínica e subclínica, sua classificação irá depender dos diferentes sintomas da doença(40).

3.1 MASTITE CLÍNICA

A mastite clínica pode ser classificada em diferentes graus de gravidade. De acordo com Ramos *et al.* (42), existem várias formas de classificação, incluindo aguda, subaguda crônica, superaguda e gangrenosa. Durante a fase aguda, os sinais da doença são sutis, o que dificulta sua detecção, já que a evolução é lenta. Na forma subaguda, é possível identificar a presença de grumos no teste da caneca, embora os sinais inflamatórios sejam menos evidentes. A fase crônica é caracterizada pela persistência da infecção no úbere do animal, podendo durar dias, meses ou até anos. Nesses casos, é comum o desenvolvimento de fibrose nos quartos afetados.

Situações superagudas, conforme observado por Ramos *et al.* (42), ocorrem quando o animal apresenta prostração intensa e febre. Esses casos são associados a uma inflamação intensa, geralmente provocada por agentes ambientais. Por fim, há a forma gangrenosa da mastite, na qual o quarto mamário pode apresentar gotejamento de soro e uma coloração anormal. Pode-se observar, na Figura 9, um caso de um animal contaminado pela bactéria *Clostridium perfringens* que precisou fazer a remoção de um quarto mamário devido à mastite clínica gangrenosa.

Figura 9 - Vaca com teto removido devido à contaminação por *Clostridium perfringens*.



Fonte: (43)

3.2 MASTITE SUBCLÍNICA

A mastite subclínica é uma doença intramamária mais importante em bovinos e tem uma incidência bem maior que a mastite clínica. Ela, em poucas ocasiões, representa uma ameaça imediata ao animal (37). Segundo Ramos *et al.* ⁽⁴²⁾, a mastite subclínica é uma doença silenciosa por não apresentar sinais visíveis e, por isso, acaba passando despercebida pelos produtores. Com isso, a doença vai se espalhando para o rebanho e infectando outras vacas.

Raça, idade, alta produção leiteira, paridade, estágio da lactação, defesa e conformação da glândula mamária, intervalo de ordenha, contagem de células somáticas do leite, período seco, lesões nos tetos, resistência genética são fatores de risco associados à mastite subclínica. Além disso, fatores como práticas de manejo inadequadas (higiene do galpão e da glândula mamária, condição ruim dos tetos, higiene ambiental inadequada, saneamento, tamanho do rebanho, uso de pano para lavagem das mãos, imersão inadequada dos tetos, técnica de ordenha, máquina de ordenha) também estão relacionados à prevalência da mastite subclínica (44, 45).

De acordo com Kumari; Bhakat e Choudhary ⁽⁴⁵⁾, os agentes causadores da mastite subclínica em vacas incluem *Streptococcus dysgalactiae*, *Staphylococcus aureus*, *Actinomyces pyogenes*, *Pseudomonas aeruginosa*, *Nocardia asteroides*, *Clostridium perfringens*, *Mycobacterium*, *Mycoplasma*, *Pastuerella*, *Prototheca* e *leveduras*. Os principais patógenos envolvidos são espécies de *Staphylococcus*, *Streptococcus*, coliformes e *Actinomyces pyogenes*.

A mastite subclínica pode ser avaliada através de exames microbiológicos para identificar o agente causador e determinar a resistência aos antibióticos. No entanto, esses exames são caros e nem sempre disponíveis. Uma alternativa mais acessível é a CCS no leite, que pode ser feita usando o método *California Mastitis Test* - CMT. Esse método é prático e fornece resultados imediatos, mas a interpretação dos resultados é subjetiva e requer treinamento. Outras formas de contar as células somáticas exigem o envio de amostras de leite para laboratórios especializados (41).

A Tabela 6 mostra uma estimativa da representação da célula somática em relação à incidência da mastite subclínica.

Portanto, a mastite subclínica é uma forma mais grave e causa perdas significativamente maiores para a indústria leiteira. As perdas devido à mastite subclínica são mais de três vezes superiores em comparação com a mastite clínica. Nessa forma de mastite, o leite parece normal e não há sinais visíveis de anormalidades, como inchaço da glândula mamária, endurecimento do quarto afetado, dor ou presença de leite aquoso. No entanto, ocorrem alterações físicas e químicas no leite, o que facilita a detecção da mastite subclínica por meio de vários testes diagnósticos (45).

Tabela 6 – Estimativa da representação da célula somática em relação à incidência da mastite subclínica

CCS no leite do tanque (x1.000/mL)	Estimativa da gravidade da mastite	Redução na produção (%)	% de animais infectados
< 250	Pouca ou nenhuma	Irrelevante	6
250 - 500	Média	4	26
500 - 750	Acima da Média	7	±42
750 - 1000	Ruim	15	±42
1000	Muito ruim	18	±54

Fonte: Retirado de Brito e Sales ⁽⁴¹⁾

4 APRENDIZAGEM DE MÁQUINA

O campo do aprendizado de máquina teve sua origem na década de 1960, emergindo como um subcampo da inteligência artificial voltado para a extração de padrões a partir de dados. Inicialmente, as aplicações do aprendizado de máquina tinham um foco predominantemente computacional. No entanto, a partir do final da década de 1990, essa área expandiu seus horizontes e adquiriu uma identidade própria como um campo de estudo independente (46).

Na indústria de bovinos leiteiros, o campo da aprendizagem de máquina tem desempenhado um papel fundamental em várias frentes de aplicação, trazendo consigo um impacto extremamente positivo. Essas aplicações abrangem a previsão de diversos aspectos vitais, como estresse térmico (47), o estado metabólico de vacas leiteiras no início da lactação (48), a detecção de claudicação (49), a previsão do parto com base no monitoramento do comportamento de atividade, repouso e ruminação dos bovinos (50), além da avaliação do estado de saúde da úbere por meio da análise da contagem de células somáticas (51).

A integração da aprendizagem de máquina nesses cenários tem proporcionado uma melhoria substancial na precisão e eficiência das previsões, resultando em práticas mais eficazes e na otimização da gestão do gado leiteiro, como demonstrado em trabalhos anteriores (47, 48, 49, 50, 51). Isso tem se traduzido em benefícios tanto para os produtores quanto para o bem-estar dos animais, além de contribuir para a sustentabilidade e produtividade da indústria de bovinos leiteiros. O uso dessas técnicas modernas promete continuar a impulsionar avanços significativos em diferentes aspectos da produção leiteira, como evidenciado pelos estudos mencionados.

4.1 APRENDIZADO DE MÁQUINA AUTOMATIZADO

Nos últimos tempos, o aprendizado profundo tem sido amplamente utilizado em diversas áreas de IA, como a categorização de imagens, detecção de objetos e modelagem de linguagem. Em 2012, no Desafio de Reconhecimento Visual em Grande Escala do ImageNet, mais conhecido como ILSVRC (52), conseguiu superar todos os modelos tradicionais existentes. Desde então, têm surgido redes neurais cada vez mais sofisticadas e profundas, como o VGG-16 (53), com uma grande quantidade de parâmetros e demanda computacional elevada. Contudo, esses modelos são criados manualmente, requerendo recursos e tempo consideráveis por parte dos especialistas para alcançar um desempenho satisfatório (54).

Por esse fato, surge uma nova ideia de automatizar o pipeline de AutoML para reduzir os custos de desenvolvimento e permitir que especialistas em domínio construam aplicações de aprendizagem de máquina sem exigir um amplo conhecimento de estatística

e ML (54).

Conforme destacado por He; Zhao e Chu ⁽⁵⁴⁾, o avanço veloz do poder computacional elevou o AutoML à posição de destaque tanto na indústria quanto na academia. Com a combinação de diversas técnicas, os sistemas completos de AutoML proporcionam a criação de pipelines de ML acessíveis e eficazes, incentivando empresas de IA a compartilharem soluções, como o Cloud AutoML 1 da Google, para auxiliar usuários sem conhecimento prévio em ML na criação de modelos personalizados de alta qualidade.

De acordo com Hutter; Kotthoff e Vanschoren ⁽⁵⁵⁾, uma das vantagens do AutoML é a otimização automatizada dos hiperparâmetros. Segundo os autores, essa abordagem oferece redução do esforço humano, melhora no desempenho dos algoritmos ao adaptá-los melhor aos problemas específicos e, além disso, proporciona um aumento na reprodutibilidade e imparcialidade de estudos científicos. Com a automatização dos hiperparâmetros, é possível realizar comparações entre diferentes métodos quando todos recebem o mesmo tipo de ajuste.

Atualmente, uma variedade de ferramentas de AutoML está disponível, incluindo Amazon Lex (56), AutoKeras (57), Google Cloud AutoML (58), AutoGluon (1), H2O (59) e TPOT (60), entre outras. O foco deste trabalho se direcionará à exploração dos modelos AutoGluon, H2O e TPOT. A seleção desses modelos ocorreu após uma análise cuidadosa da facilidade de implementação, a qual é respaldada por suas documentações bem organizadas e explicativas. Vale ressaltar que a utilização dessas ferramentas é simplificada, visto que requer apenas a instalação da biblioteca em um ambiente virtual para se ter acesso aos seus recursos.

4.1.1 *AutoGluon*

O AutoGluon é uma poderosa ferramenta de aprendizado de máquina automatizado, desenvolvida com base em princípios fundamentais (1). Sua principal característica é a robustez, o que significa que ele pode lidar efetivamente com diversos conjuntos de dados estruturados, assegurando o sucesso do treinamento mesmo quando alguns modelos individuais falham.

A simplicidade é outra característica importante, possibilita que qualquer usuário treine um modelo diretamente em seus dados brutos, sem precisar conhecer detalhes específicos sobre os dados ou modelos de aprendizado de máquina. Além disso, o AutoGluon demonstra tolerância a falhas, permitindo retomar o treinamento a partir do ponto em que foi interrompido. Por fim, ele oferece a capacidade de fornecer resultados dentro dos limites definidos pelos usuários (1).

Quando o usuário não determina para o AutoGluon com que tipo de problema irá lidar, ele realiza o processamento de dados de forma automática, inferindo o tipo de problema de previsão (classificação binária, multiclasse ou regressão) com base nos valores

da coluna de rótulo e identificando os tipos de variáveis em outras características(61).

O pré-processamento inclui categorização de características em numéricas, categóricas, texto ou data/hora. Colunas não categorizadas são descartadas, e datas/horas são transformadas em números ordenados. Valores de colunas de texto são convertidos em vetores numéricos de n-gramas (retendo os mais frequentes para economizar memória). Variáveis discretas ausentes são tratadas como *Desconhecidas*, permitindo ao AutoGluon lidar com categorias desconhecidas durante os testes (61).

Segundo Shirkov *et al.* (61), o AutoGluon oferece uma variedade de modelos disponíveis para abordar diferentes problemas, incluindo LightGBM, CatBoost, XGBoost, Random Forest, Extra Trees, KNearestNeighbors, modelo linear, TabularNeuralNetTorchModel, NNFastAiTabularModel, VowpalWabbitModel, MultiModalPredictorModel, TextPredictorModel, ImagePredictorModel e outros.

Usando uma estratégia de conjunto de modelos em várias camadas, o AutoGluon melhora a precisão das previsões ao combinar previsões de diversos modelos para reduzir a variância nas previsões finais(1).

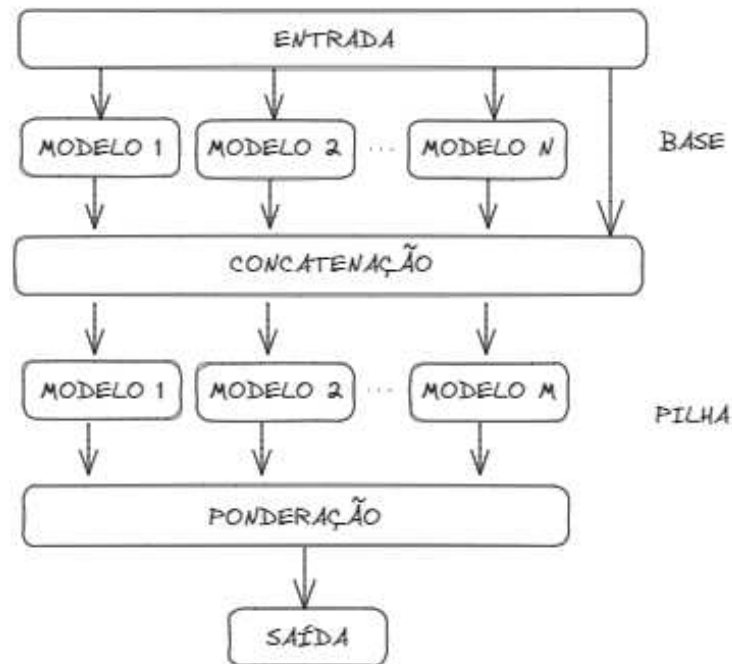
De acordo com Qi; Xu e Xu (62), o AutoGluon reutiliza todos os modelos da camada básica como empilhadores, com os mesmos valores de hiperparâmetros. As previsões do modelo de camada inferior, combinadas com as características originais dos dados, são usadas como entrada para os empilhadores de camadas superiores.

A última camada do modelo empilhador usa a seleção do conjunto para ponderar as previsões dos empilhadores, e esse processo é repetido para obter o modelo empilhador de várias camadas. A estratégia de conjunto em várias camadas do AutoGluon é ilustrada na Figura 10, com um exemplo de duas camadas de empilhamento (1).

Utilizando o *ensemble bagging* com k-fold de todos os modelos em todas as camadas, o AutoGluon melhora o desempenho de empilhamento ao utilizar todos os dados disponíveis para treinamento e validação. Os dados são divididos aleatoriamente em k partes (estratificadas por etiquetas), e, em seguida, é treinado k cópias de um modelo, sendo cada uma com uma parte diferente dos dados como conjunto de validação. Após isso, o AutoGluon empilha todos os modelos, e cada um gera previsões *out-of-fold* (OOF) no conjunto de dados não utilizado no treinamento, permitindo obter previsões OOF de todos os modelos para cada exemplo de treinamento (61).

Segundo Erickson *et al.* (1), é crucial que os modelos de camadas superiores sejam treinados com base nas previsões OOF das camadas inferiores. Ao usar previsões OOF dos *ensembles bagged*, os modelos de empilhamento de camadas superiores aproveitam a mesma quantidade de dados de treinamento que os da camada anterior. No entanto, o k-fold *bagging* introduz uma pequena incompatibilidade: os modelos de empilhamento recebem previsões OOF de modelos individuais durante o treinamento, mas, durante a inferência, eles operam com as previsões médias das camadas inferiores sobre todo o conjunto *bagged*.

Figura 10 - Estratégia de empilhamento em várias camadas do AutoGluon, um caso de duas camadas de empilhamento e n tipos de modelos de aprendizado de máquina (1).



Fonte: Modificada pelo autor, retirada de Erickson *et al.* (1)

Para mitigar esse problema, é proposto um processo de *bagging* repetido. Se houver tempo de treinamento restante após o *k-fold bagging*, repete-se o processo de *bagging* *k-fold* em n partições diferentes dos dados de treinamento, calculando a média de todas as previsões OOF nos bags repetidos(61).

A estratégia de treinamento envolve um controle de tempo para cada camada de empilhamento. É verificado o tempo necessário para treinar cada modelo e avançar somente se o tempo disponível for suficiente. Após o treinamento de cada novo modelo, ele é imediatamente salvo em disco para garantir a tolerância a falhas. Essa abordagem torna o *framework* altamente previsível em seu comportamento, tanto em relação ao tempo quanto em relação às possíveis falhas durante o treinamento (1).

Ao contrário de outros *frameworks*, o treinamento é realizado de forma sequencial e é confiada a cada implementação individual a tarefa de aproveitar eficientemente múltiplos núcleos de processamento. Essa abordagem permite superar problemas de falta de memória em conjuntos de dados maiores. Com o AutoGluon, é possível produzir previsões mesmo sob limites de tempo apertados, assegurando que, ao menos um modelo possa ser treinado em uma etapa dentro do tempo disponível, resultando em informações valiosas (61).

A utilização do AutoGluon tem se mostrado eficaz em diferentes estudos recentes, como os de Qi; Xu e Xu (62) na análise de riscos de deslizamentos sísmicos e Xiong; Pan e Chen

4.1.2 H2O AutoML

A empresa H2O.ai tem como missão possibilitar a incorporação da inteligência artificial nas operações de negócios através do uso de seu principal produto, o H2O. Essa plataforma de código aberto, de aprendizado de máquina distribuído e escalável viabiliza a implementação de técnicas de IA e aprendizado profundo em uma ampla gama de setores (63).

Com recursos de compressão em memória para enfrentar volumes significativos de dados e interfaces amigáveis para várias linguagens de programação, o H2O disponibiliza uma ampla variedade de algoritmos de aprendizado de máquina, permitindo a criação de inúmeros modelos para obter previsões precisas. A H2O.ai também cultiva uma cultura de colaboração com a comunidade acadêmica e científica, buscando avançar no campo da ciência de dados (63).

O H2O AutoML é um algoritmo de aprendizagem supervisionada completamente automatizado que foi desenvolvido no H2O. Além de oferecer suporte para Python, R, Java e Scala, também possui uma interface gráfica acessível através da web. Embora o algoritmo seja altamente automatizado, diversos ajustes e configurações estão disponíveis como parâmetros para os usuários, permitindo a personalização de certos aspectos nas etapas de modelagem (59).

Conforme mencionado por LeDell e Poirier ⁽⁵⁹⁾, o AutoML em questão realiza o pré-processamento automático de dados em algoritmos de aprendizado supervisionado, incluindo imputação automática e, quando necessário, normalização e codificação *one-hot* para modelos *XGBoost*. Além disso, os modelos baseados em árvore, como *Random Forest* e *Gradient Boosting Machines*, são suportados pelo *framework*, permitindo a divisão em grupos de variáveis categóricas e o tratamento nativo dos dados categóricos.

Oferecendo suporte para o *wrapper* do software *XGBoost*, o H2O AutoML permite a inclusão desse algoritmo de terceiros, possibilitando a aceleração do treinamento por meio de Unidade de Processamento Gráfico (UPG). Além disso, ele ainda é capaz de abranger diversos modelos de aprendizado de máquina, como *Gradient Boosting Machines* do *XGBoost* e H2O, *Random Forests* (variedade de árvores padrão e extremamente aleatórias), Redes Neurais Profundas e Modelos Lineares Generalizados do H2O (59).

A versão, H2O v3.30.0.3, do *framework* é capaz de treinar e validar cruzadamente vários modelos, incluindo pré-especificados *XGBoost GBMs* e *H2O Gradient Boosting Machines*, H2O Modelos Lineares Generalizados em uma grade fixa, *H2O Random Forest* padrão, *H2O Deep Neural Nets* e *H2O Extremely Randomized Trees* (63).

Segundo LeDell e Poirier ⁽⁵⁹⁾, a busca pelos hiperparâmetros é realizada de forma aleatória, com base em *benchmarks* internos e na experiência de cientistas de dados especializados. Essa abordagem é constantemente aprimorada para obter melhores resultados

de previsão.

Os modelos pré-especificados são fornecidos para oferecer opções rápidas e confiáveis para cada algoritmo, sendo a ordem dos algoritmos ajustável pelo usuário. Além disso, ele busca aumentar a diversidade do conjunto de modelos para a criação de *Stacked Ensembles* finais. Na versão H2O v3.30.0.1, um parâmetro experimental chamado taxa de exploração foi introduzido, permitindo o ajuste do modelo *XGBoost*, *Gradient Boosting Machines* e do modelo *H2O Gradient Boosting Machines* para otimização de resultados (59).

De acordo com LeDell e Poirier ⁽⁵⁹⁾, trata-se de um algoritmo altamente escalável para grandes conjuntos de dados, suportando treinamento paralelo em Unidade Central de Processamento (UCP) ou UPG. Ele inclui *XGBoost* e outros algoritmos com eficientes implementações. Graças a isso, pode treinar mais modelos em menos tempo do que bibliotecas AutoML com algoritmos mais lentos. O uso de técnicas simples, como busca aleatória e *stacking*, resulta em excelente desempenho comparável a métodos mais complexos, como otimização Bayesiana ou algoritmos genéticos.

Em um estudo conduzido para prever dados epidemiológicos da covid-19, Marques *et al.* ⁽⁶⁴⁾ observou que o AutoML H2O apresentou resultados superiores em algumas métricas e inferiores em outras em comparação com a rede neuronal *Long Short-Term Memory Networks*. Essa performance diferenciada sugere que a utilização de um seletor de algoritmo de aprendizado de máquina pode ser uma abordagem promissora para aprimorar a precisão na previsão de séries temporais epidêmicas.

4.1.3 TPOT

Desenvolvido no Laboratório de Genética Computacional da Universidade da Pensilvânia, a Ferramenta de Otimização de Pipeline Baseada em Árvores (TPOT) é um sistema AutoML baseado em Algoritmo Evolucionário que utiliza programação genética (GP) para otimizar uma série de seletores de características, pré-processadores e modelos de aprendizado de máquina com o objetivo de maximizar a precisão da classificação (60).

Utilizando Programação Genética com otimização Pareto NSGA-II, o TPOT otimiza a precisão do pipeline, considerando sua complexidade. Para a implementação do algoritmo GP, o TPOT utiliza o pacote *Distributed Evolutionary Algorithms* em *Python* de (65).

De acordo com Le; Fu e Moore ⁽⁶⁰⁾, enquanto a maioria dos sistemas AutoML concentra-se principalmente na escolha do modelo e na otimização de hiperparâmetros, o TPOT destaca-se ao também considerar a seleção e a engenharia de características. Ele avalia os pipelines completos com base em pontuações cruzadas, como erro médio quadrático ou precisão equilibrada, sem requerer conhecimento prévio sobre o problema.

De acordo com Olson e Moore ⁽⁶⁶⁾, o TPOP emprega diversos algoritmos de aprendizado de máquina, incluindo *DecisionTree*, *RandomForest*, *Extreme LogisticRegression*

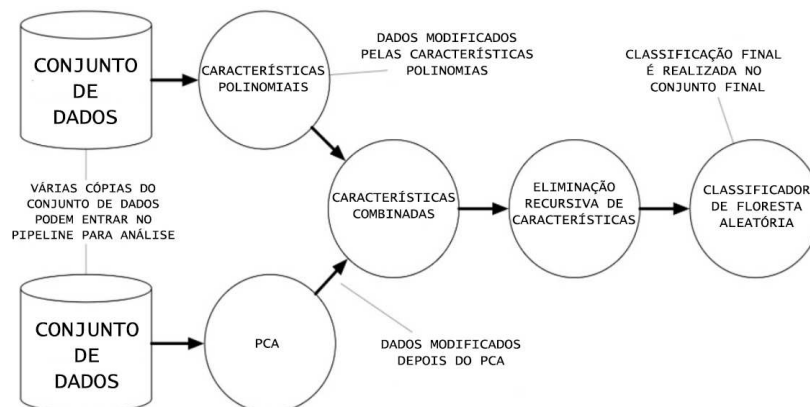
e *KNearestNeighborClassifier*, todos provenientes do pacote *scikit-learn*. Além disso, o classificador *Extreme Gradient Boosting XGBoost* é fundamentado no algoritmo proposto por Chen e Guestrin ⁽⁶⁷⁾. Esses classificadores armazenam as previsões como uma nova característica.

O TPOT, conforme mencionado por Le; Fu e Moore ⁽⁶⁰⁾, automatiza o processo de projetar pipelines de aprendizado de máquina, representando-os como árvores de expressão binárias com operadores de ML como primitivas. As implementações eficientes, como o *extreme gradient boosting*, são incluídas nos pipelines, juntamente com algoritmos da biblioteca *scikit-learn*.

Na etapa de pré-processamento de características, o objetivo é realizar transformações nos dados antes de alimentá-los aos algoritmos de aprendizado de máquina. No TPOT, existem diversas opções disponíveis para esse propósito, tais como: *StandardScaler*, *RobustScaler*, *MinMaxScaler*, *MaxAbsScaler*, *RandomizedPCA*, *Binarizer* e *PolynomialFeatures* ⁽⁶⁶⁾.

Por outro lado, na etapa de seleção de características, segundo Olson e Moore ⁽⁶⁶⁾, o objetivo é reduzir, caso necessário, o número de características no conjunto de dados, utilizando critérios específicos e, posteriormente, retornar o conjunto de dados modificado. O TPOT oferece diversas opções para realizar essa tarefa, incluindo: *VarianceThreshold*, *SelectKBest*, *SelectPercentile*, *SelectFwe* e *Recursive Feature Elimination*. Outra funcionalidade é a possibilidade de combinar conjuntos de dados distintos, transformando-os em um único conjunto. Na Figura 11, é possível observar um exemplo de um pipeline baseado em árvore.

Figura 11 - Pipeline otimizado baseado em árvore.



Fonte: Extraída e adaptada de Olson e Moore ⁽⁶⁶⁾

Com base em critérios biológicos, o TPOT incorpora em sua gama de funcionalidades, um novo operador chamado *Feature Set Selector (FSS)*. Ele permite a seleção de recursos em nível de grupo. O operador FSS capacita o TPOT a selecionar o melhor

conjunto que otimiza a precisão média em validação cruzada k-fold, utilizando subconjuntos predefinidos de recursos. Implementado na etapa inicial do pipeline, o FSS encaminha apenas um subconjunto específico de recursos, efetivamente dividindo o conjunto de dados original em partes menores. Essa funcionalidade permite que os usuários, na inicialização do pipeline, especifiquem subconjuntos relevantes de recursos, reduzindo a dimensão do espaço de recursos (60).

Os estudos conduzidos por Radzi *et al.* ⁽⁶⁸⁾, Fati *et al.* ⁽⁶⁹⁾ e Kiala; Odindi e Mutanga ⁽⁷⁰⁾ destacam a notável eficácia da ferramenta de otimização de pipelines baseada em árvores (TPOT) em diversos domínios científicos. Entre os exemplos notáveis, pode-se mencionar a aplicação do TPOT na predição de câncer de mama, em que obteve resultados promissores, bem como sua capacidade de realizar estimativa contínua de pressão arterial sem o uso de manguito. Além disso, o TPOT também demonstrou sua efetividade no mapeamento da Erva Parthenium e em outras aplicações relevantes. Essas descobertas ressaltam o potencial do TPOT como uma ferramenta versátil e valiosa na otimização de pipelines e análise de dados em diferentes áreas de estudo.

5 SISTEMA

5.1 API

As APIs constituem um meio essencial para que tanto pessoas quanto aplicativos interajam com software. Enquanto as pessoas usam interfaces de usuário, os aplicativos podem acessar software por meio de APIs, que oferecem uma maneira simplificada de conectar, integrar e ampliar sistemas de software. As APIs são versáteis e podem ser definidas de várias maneiras, como um serviço web que fornece recursos via tecnologias como HTTP. Elas desempenham um papel crucial na construção de sistemas de software distribuídos, viabilizando um acoplamento flexível (71).

Ainda segundo Biehl ⁽⁷¹⁾, uma API atua como uma ponte entre provedores internos e consumidores externos de dados empresariais, permitindo uma interação fluida e padronizada. Geralmente ocultas aos olhos do usuário final, as APIs operam nos bastidores, sendo invocadas por outros softwares. Usadas na comunicação entre máquinas e na integração de sistemas, elas são um elemento fundamental na construção de aplicativos e soluções. A criação de APIs deve priorizar a facilidade de uso para os desenvolvedores que as incorporarão em novas aplicações, implicando muitas vezes uma mudança de perspectiva no processo de desenvolvimento. Em vez de serem construídas a partir da visão do provedor de serviço, as APIs devem ser concebidas considerando a perspectiva do consumidor.

No cenário global, uma ampla variedade de APIs está disponível, com algumas delas ganhando maior notoriedade do que outras. Por exemplo, a API do Google Drive oferece funcionalidades que permitem aos usuários ler, gravar e sincronizar arquivos com a plataforma. Além disso, menciona-se a API do *OpenWeatherMap*, que fornece informações essenciais sobre temperatura e vento. Essa API é uma ferramenta valiosa para diversas plataformas que executam atividades influenciadas pelas condições climáticas. Setores como agricultura, construção, esportes e eventos podem se beneficiar ao utilizar essa API para obter insights relevantes.

A API desenvolvida no contexto deste projeto tem como principal propósito a integração entre os modelos de aprendizado de máquina e o sistema web. Entretanto, sua utilidade transcende essa aplicação específica, já que a API pode ser empregada e personalizada por qualquer entidade autorizada, permitindo que pequenos produtores se beneficiem desta ferramenta como um recurso auxiliar em suas atividades pecuárias.

Nesta implementação, utilizou-se o *framework Flask* na versão 2.3.2, criado por Armin Ronacher, empregando a linguagem de programação Python. A API elaborada abrange duas funcionalidades essenciais: a primeira oferece uma Notação de Objetos JavaScript (JSON), que encapsula os resultados de predição individualizada para cada animal, e a segunda função proporciona um JSON, que apresenta o índice de risco global de contaminação do rebanho.

A escolha do modelo a ser utilizado dependerá do consumidor da API. Para isso, é necessário selecionar um dos três modelos disponíveis no formato pkl, os quais foram escolhidos com base nos resultados do treinamento durante a fase de aprendizado de máquina, conforme detalhado na subseção de Modelos. O consumidor deverá especificar a opção de modelo desejada e fornecer os dados essenciais para a realização da predição. São disponibilizadas três alternativas de modelos, com configurações de 6, 12 e 24 variáveis, conforme previamente explicado. Portanto, a seleção do modelo apropriado ficará a cargo do usuário que estiver consumindo a API.

Dentro da estrutura da API, antes de executar a predição, um estágio crucial de pré-processamento é implementado. Esse procedimento, conhecido como MinMaxScaler, conforme previamente elucidado na subseção de Modelos, desencadeia uma transformação linear nas características dos dados, ajustando cada valor em direção a um intervalo predefinido, que tipicamente varia entre 0 e 1. Tal procedimento se faz necessário em virtude de o modelo ter sido treinado sob essas mesmas condições, em que os dados passaram por esse pré-processamento. Isso resulta em uma harmonização dos dados, permitindo uma interpretação uniforme pelo modelo treinado.

Uma vez que todos os procedimentos da API tenham sido concluídos, a plataforma fornecerá ao consumidor um JSON contendo informações intrínsecas à funcionalidade selecionada pelo usuário. Essa saída concederá ao consumidor a liberdade de manipular os resultados de acordo com suas preferências e necessidades.

5.2 SISTEMA WEB

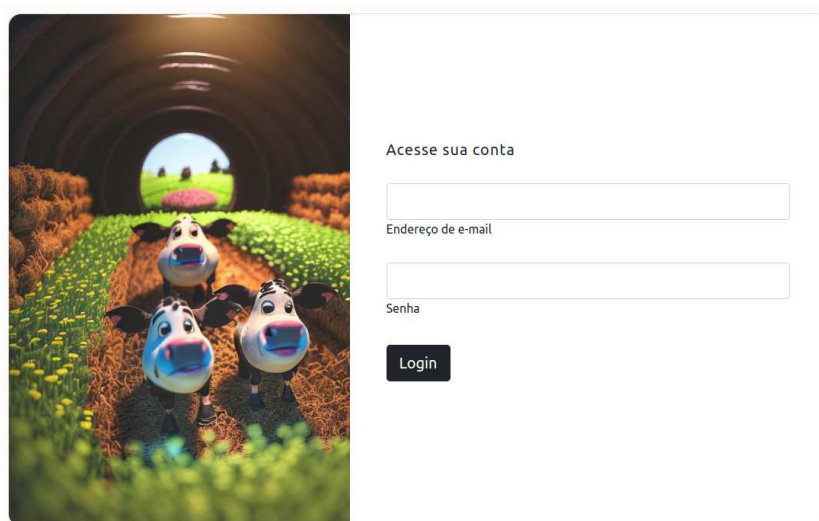
O sistema web foi concebido utilizando tecnologias de ponta. A implementação foi realizada com PHP na sua versão 8.1.2, que oferece recursos atualizados e aprimorados para o desenvolvimento. Além disso, a estilização das páginas foi alcançada com o estilizador *Bootstrap* na sua versão 5, o que permitiu um design moderno e responsivo.

O dinamismo e a interatividade foram alcançados por meio da linguagem *JavaScript*, permitindo a criação de funcionalidades avançadas e uma experiência de usuário envolvente. O *framework* Laravel, na sua versão 10.13.5, foi a base do desenvolvimento, proporcionando uma estrutura robusta e bem organizada para a construção do sistema. Isso possibilitou uma codificação mais eficiente e o aproveitamento de recursos poderosos disponibilizados pelo Laravel.

Por fim, o PostgreSQL, versão 14.8, foi a escolha para o sistema de gerenciamento de banco de dados, garantindo confiabilidade, segurança e desempenho na manipulação dos dados essenciais ao funcionamento da aplicação. Essa combinação de tecnologias contribuiu para a criação de um sistema web completo, eficiente e sofisticado, capaz de atender às demandas e expectativas dos usuários de forma exemplar.

A Tela de Login representada pela Figura 12 é o ponto de entrada para os usuários acessarem o sistema. Nessa tela, os usuários devem inserir suas credenciais (endereço de e-mail e senha) para autenticação. Ela oferece uma camada fundamental de segurança, garantindo que apenas usuários autorizados possam acessar as funcionalidades do sistema.

Figura 12 - Tela de login.



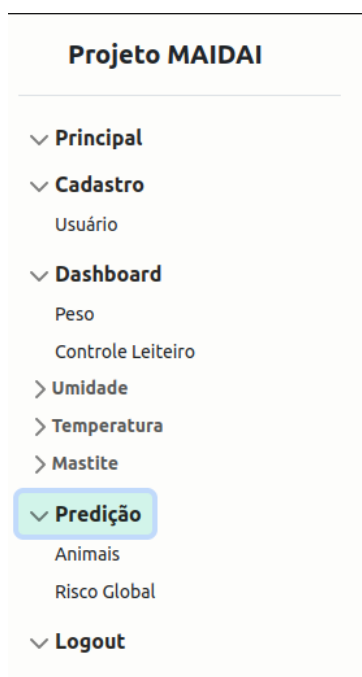
Fonte: Elaborada pelo autor (2023).

O centro de navegação do sistema é representado pela Tela do Menu, cuja imagem está disponível na Figura 13. Nessa interface, os usuários encontram uma experiência amigável e intuitiva, permitindo a exploração e acesso a todas as funcionalidades disponíveis. A organização hierárquica do menu em categorias e subcategorias facilita a localização das opções desejadas, garantindo uma navegação eficiente e satisfatória. Ao consultar a Figura 13, os leitores poderão visualizar como o sistema apresenta esse importante recurso, essencial para a interação do usuário e a utilização plena das funcionalidades disponíveis.

A Tela de Cadastro de Usuário é destinada à inclusão de novos usuários no sistema. Ela oferece um formulário abrangente para a inserção de informações pessoais e dados de acesso, além de fornecer opções para definir níveis de permissões e papéis no sistema. Essa tela é fundamental para a administração do sistema, permitindo que administradores criem e gerenciem contas de usuários com diferentes níveis de acesso e privilégios, conforme se pode observar na Figura 14.

O painel exibido na Figura 15 de forma confiável ilustra a média de peso dos animais no composto ao longo dos anos de 2021 e 2022, oferecendo uma visão do padrão de comportamento em relação a cada mês individual, fornecendo dados cruciais para avaliar a adequação da alimentação à política de nutrição. A cor azul destaca o peso médio dos animais, enquanto a cor cinza representa a quantidade total presente no período. Essas informações estratégicas permitem monitorar o desempenho nutricional do rebanho de

Figura 13 - Menu do sistema.



Fonte: Elaborada pelo autor (2023).

Figura 14 - Cadastro de Usuário.

Cadastro de Usuários				
ID	Nome	Email	Permissão	Ações
3	Douglas Lima Fonseca	teste@teste.com	Administrador	Adicionar Usuário Editar Excluir

Fonte: Elaborada pelo autor (2023).

forma clara e intuitiva.

Na Figura 16, apresenta-se a tela do controle leiteiro do rebanho, uma informação de extrema relevância capaz de gerar *insights* valiosos. Através dessa ferramenta, é possível identificar reduções na produção de leite, o que pode indicar problemas de saúde nos animais, tais como doenças ou questões relacionadas à alimentação, bem como o término da lactação.

Essa análise minuciosa dos dados permite uma abordagem proativa no manejo da pecuária, garantindo a detecção precoce de possíveis enfermidades e o ajuste adequado na nutrição das vacas leiteiras. Com a utilização desse recurso, é possível tomar decisões estratégicas para otimizar a produção de leite, maximizar a saúde do rebanho e,

Figura 15 - Dashboard de Peso.



Fonte: Elaborada pelo autor (2023).

consequentemente, aumentar a eficiência e lucratividade da atividade.

A representação gráfica da informação torna mais acessível e compreensível o acompanhamento do controle leiteiro, possibilitando que os produtores tomem decisões informadas e embasadas em dados precisos. Essa integração entre tecnologia e pecuária contribui significativamente para o bem-estar animal, a sustentabilidade e o sucesso geral da produção leiteira.

Os dashboards dedicados a umidade 17a, temperatura 17b, luminosidade 17c e velocidade do ar 17d são organizados por lotes, com cada lote sendo representado por uma

Figura 16 - Dashboard de Controle leiteiro.



pele autor (2023).

cor única. Essas informações são de suma importância para analisar e compreender os eventos que ocorrem no composto, como variações na produção de leite e peso, bem como o aumento da contaminação no rebanho.

Através dessas visualizações gráficas detalhadas, é possível identificar padrões e tendências que podem afetar diretamente a saúde e o desempenho dos animais. O monitoramento constante desses fatores ambientais e de seu impacto nas diferentes categorias de animais permite a tomada de decisões precisas e oportunas para otimizar a produção e garantir o bem-estar dos animais.

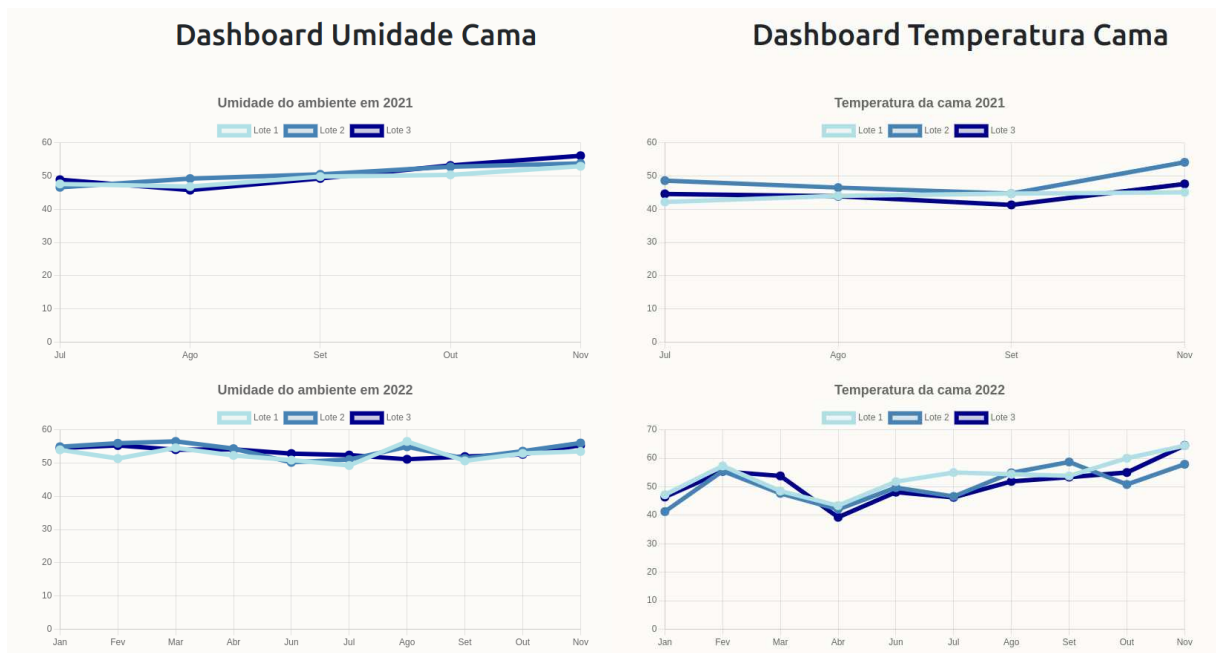
Essa abordagem tecnológica na gestão do composto agrega valor à pecuária, oferecendo aos produtores uma visão holística e abrangente das condições ambientais e seu efeito no rebanho. O uso inteligente desses dashboards contribui para uma produção mais eficiente, sustentável e responsável, resultando em melhores resultados econômicos e maior qualidade do produto final.

Os Dashboards de Mastite possuem telas estratégicas e informativas que oferecem uma visão geral abrangente dos principais aspectos relacionados à mastite no rebanho. Eles englobam diversas informações essenciais para o gerenciamento da saúde das vacas leiteiras e a manutenção da qualidade do leite produzido. Ao explorar as Figuras 18, 19, 20a, 20b e 20c, podem-se destacar os seguintes componentes:

Figura 17 - Conjunto de dashboards do sistema.

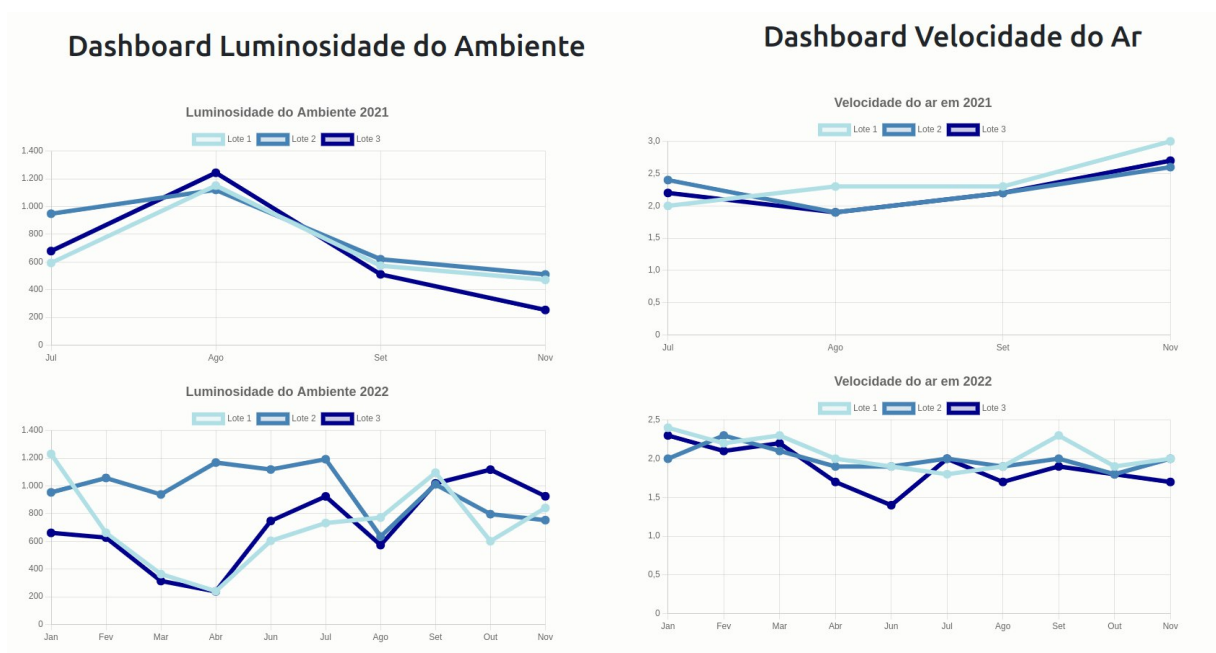
(a) Umidade da Cama.

(b) Temperatura da Cama.



(c) Luminosidade do ambiente.

(d) Velocidade do ambiente.



Fonte: Elaborada pelo autor (2023).

- Tela de Ocorrência de Casos: a seção dedicada à *Tela de Ocorrência* apresenta uma representação gráfica da quantidade de casos de mastites confirmados ao longo do mês. Essa visualização permite aos usuários acompanhar a evolução dos casos e identificar possíveis picos ou tendências sazonais. Isso é crucial para a tomada de

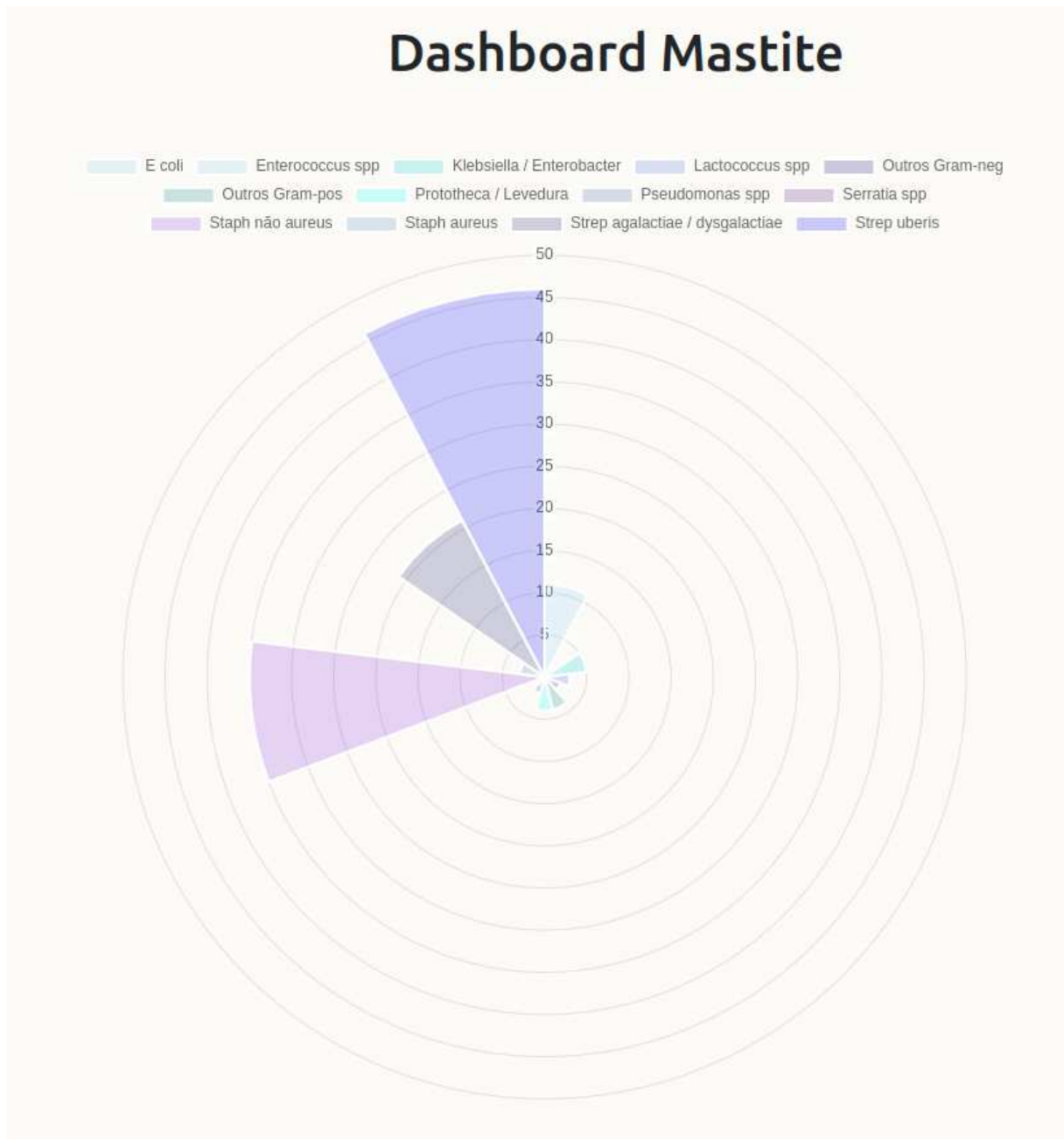
decisões proativas, visando à saúde das vacas e à qualidade do leite;

- Tela de Graus de Mastite: a aba referente aos *Graus de Mastite* fornece *insights* sobre a distribuição dos diferentes graus de mastite identificados no rebanho. Por meio de gráficos e estatísticas visuais, é possível compreender quais graus de infecção são mais prevalentes, auxiliando na alocação de recursos e estratégias de tratamento específicas para cada caso;
- Tela de Quartos Mamários: a seção dos *Quartos Mamários* apresenta uma análise detalhada sobre quais quartos mamários foram afetados pela mastite e a quantidade de ocorrências em cada um deles. Isso permite uma abordagem mais direcionada no tratamento e monitoramento da saúde individual das vacas, além de auxiliar na detecção precoce de possíveis focos de infecção;
- Tela de Resultados Laboratoriais: a tela final exhibe os *Resultados Laboratoriais* das amostras coletadas. Esses resultados não se limitam apenas a diagnósticos positivos ou negativos, mas também levam em consideração possíveis problemas de leitura, contaminação de lâminas e outras variáveis. Essa abordagem mais ampla e realista permite uma interpretação precisa das informações e evita conclusões precipitadas;
- Tela de Patógenos: o gráfico de radar é uma ferramenta poderosa para visualizar as bactérias mais recorrentes ao longo do tempo. Através desse gráfico, os profissionais podem identificar padrões sazonais, variações significativas e tomar decisões embasadas em dados concretos. Com o Dashboard de Mastite, os produtores de leite e equipes veterinárias têm à disposição uma plataforma informacional valiosa para melhorar a saúde e o bem-estar das vacas leiteiras, bem como garantir a qualidade do leite produzido.

A tela de *Predição de Mastite do Rebanho* é uma etapa crucial no processo de análise e predição de dados. Nessa tela, representada pela Figura 21, os usuários têm a oportunidade de optar entre diferentes níveis de predição oferecidos pelo sistema. Com o propósito de alcançar esse objetivo, o sistema utiliza a API desenvolvida e detalhada na seção anterior, viabilizando uma integração contínua e eficaz entre a interface do usuário e o tratamento dos dados. São disponibilizadas três alternativas:

- Opção simplificada: requer menos variáveis para gerar resultados, embora sua precisão seja menor;
- Opção intermediária: demanda um maior número de variáveis em comparação à anterior e oferece resultados um pouco mais precisos;
- Opção avançada: exige um conjunto mais significativo de variáveis em relação às outras opções, mas entrega resultados mais expressivos.

Figura 18 - Dashboard de Ocorrências dos Patógenos de Mastite.

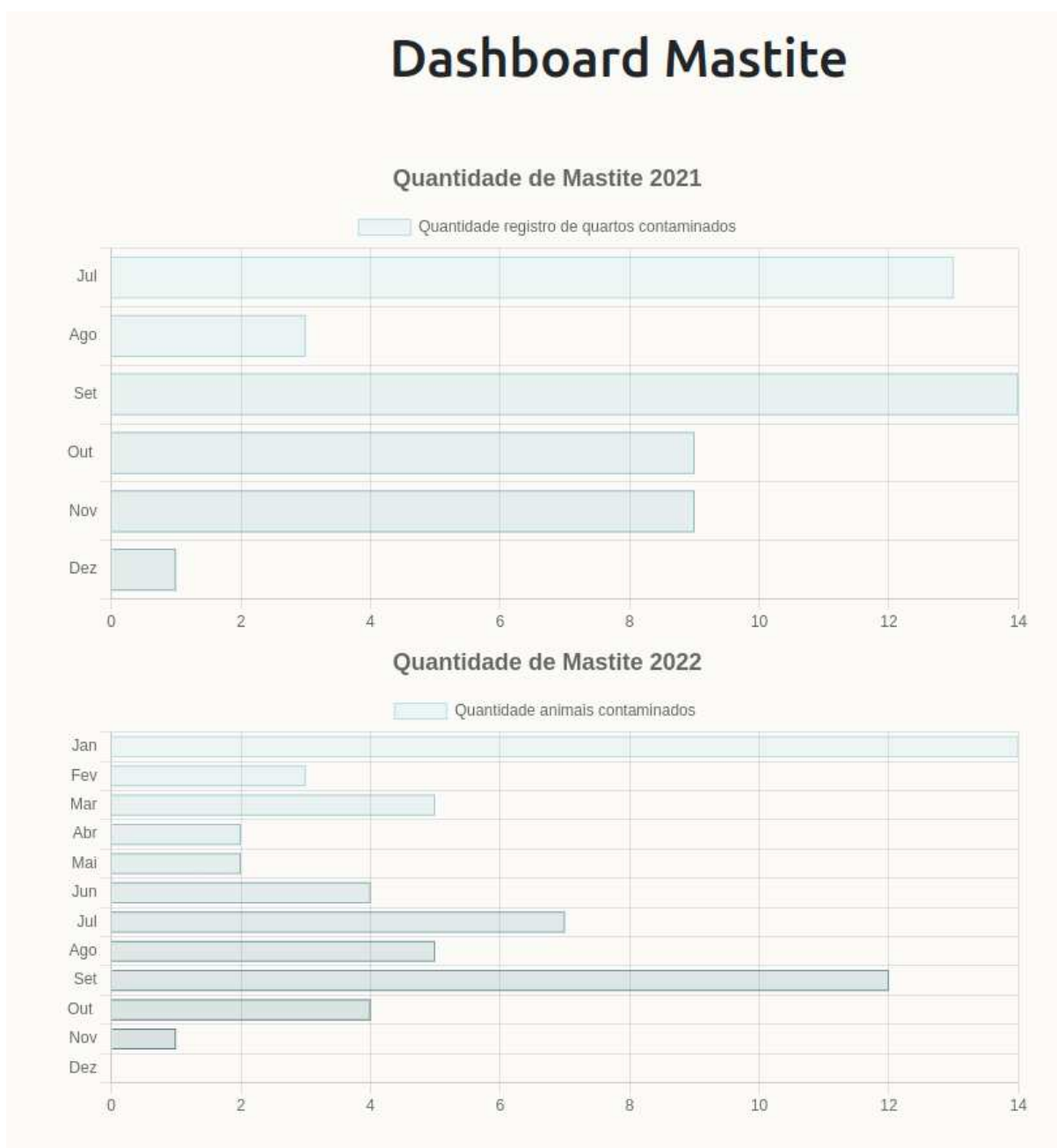


Fonte: Elaborada pelo autor (2023).

Além dessas escolhas, a tela também oferece a opção de *download* contendo o exemplo de como os dados devem ser organizados para que o sistema funcione adequadamente. Além disso, os usuários podem fazer o *upload* dos registros que desejam submeter à predição utilizando um botão específico. Essa abordagem interativa e orientada a dados permite que os usuários aproveitem ao máximo as capacidades do sistema.

A Tela de Resultado de predição, como ilustrado na Figura 22, exibe os resultados da análise de dados realizada pelo modelo de aprendizado de máquina selecionado. Nessa

Figura 19 - Dashboard de Mastite.



Fonte: Elaborada pelo autor (2023).

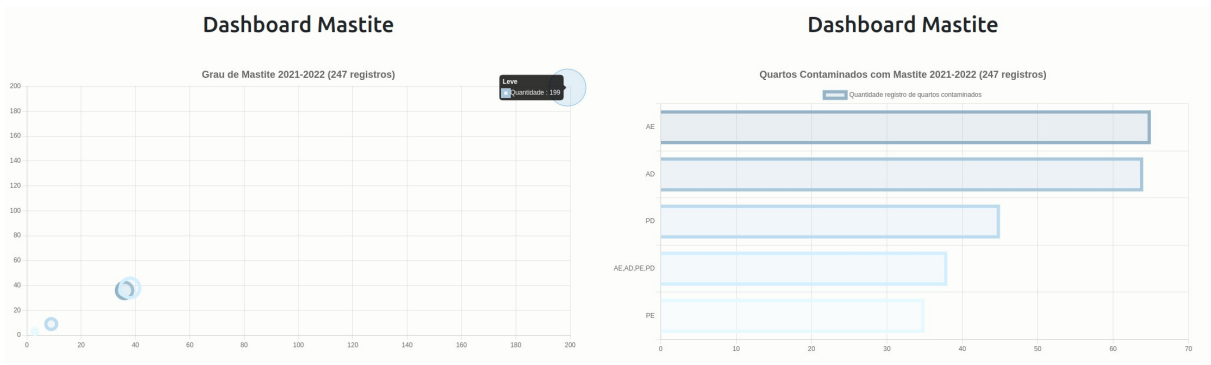
tela, os usuários têm acesso a uma tabela contendo os resultados positivos ou negativos relacionados aos animais em questão.

Além do comentado anteriormente, uma funcionalidade relevante é o *download* do arquivo previamente carregado, agora com uma coluna extra que representa o resultado da predição. Essa informação é de extrema importância para a tomada de decisões estratégicas e pode oferecer *insights* valiosos para a gestão do rebanho e a detecção precoce de problemas de saúde.

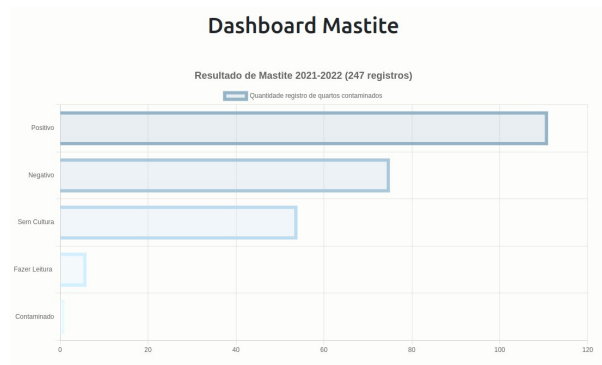
Figura 20 - Conjunto de dashboards de Mastite do sistema.

(a) Grau da Mastite.

(b) Quarto mamários contaminados.



(c) Resultado do exame.



Fonte: Elaborada pelo autor (2023).

Figura 21 - Tela de Predição de Mastite.

Predição de Mastite do Rebanho

Essa abordagem envolve o uso de um número maior de variáveis, o que a torna a opção mais precisa disponível.

Antes de prosseguir, baixe o arquivo de exemplo:

Escolher arquivo: Nenhum arquivo escolhido

Fonte: Elaborada pelo autor (2023).

Figura 22 - Dashboard de Mastite.

Resultado da Predição

[Download do Resultado em CSV](#)

Registro	Brinco	Resultado
1º	2499-1	Com Mastite
2º	2849-1	Sem Mastite
3º	7141-1	Sem Mastite
4º	4690-2	Com Mastite
5º	7552-2	Sem Mastite
6º	3781-2	Com Mastite
7º	0113-2	Com Mastite
8º	2874-1	Sem Mastite
9º	6719-2	Sem Mastite
10º	5405-1	Sem Mastite
11º	1160-2	Sem Mastite
12º	4504-1	Sem Mastite
13º	8461-1	Sem Mastite
14º	4272-1	Sem Mastite

Fonte: Elaborada pelo autor (2023).

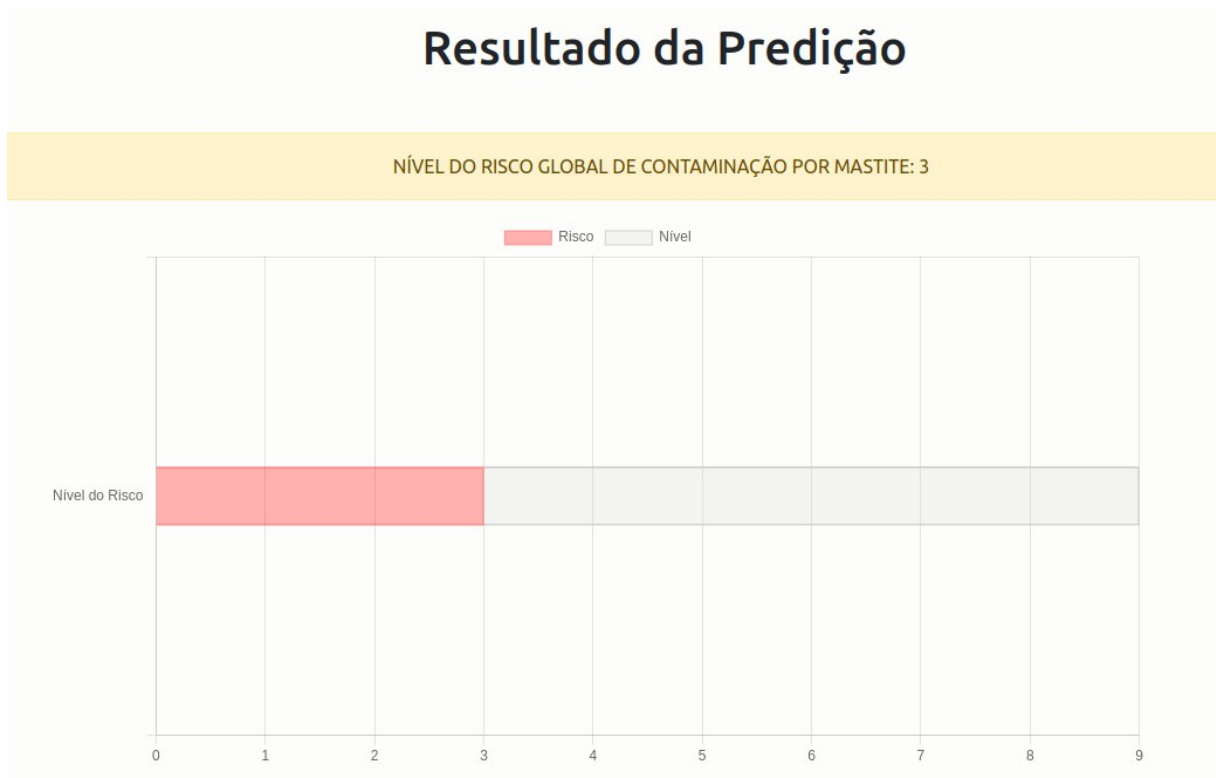
O sistema apresenta uma tela preliminar similar àquela previamente demonstrada, com uma nova finalidade de realizar a predição do Risco Global de contaminação do rebanho. Essa interface mantém as funcionalidades da versão anterior, incluindo a seleção de três níveis de opção de modelo, bem como a capacidade de efetuar o download de um arquivo de exemplo e também a opção de carregar dados para a previsão em análise. Isso estabelece uma integração harmoniosa entre a interface do usuário e o processamento de dados, impulsionada pela API detalhada na seção anterior, para uma experiência eficaz e contínua.

O cálculo do risco global é obtido somando as probabilidades individuais de cada animal estar contaminado, dividida pelo total de animais. Cada nível de risco global é representado por uma faixa de 10%, de modo que o nível 2 de risco global inclui porcentagens maiores que 20% e menores ou iguais a 30% de animais contaminados.

Mostrando a saída da previsão do risco global em uma escala de 0 a 9, a Figura 23 reflete que um valor mais alto representa uma situação mais desfavorável para o rebanho. A intensidade das cores varia gradualmente de um vermelho suave a um vermelho intenso, correspondendo à gravidade da situação. Na ilustração, é perceptível que, no cenário avaliado, o rebanho possui um nível de risco de 3 em 10, o que pode ser considerado relativamente baixo. No entanto, essa avaliação proporciona ao proprietário a oportunidade de examinar e avaliar as circunstâncias, bem como adotar as medidas necessárias para evitar um aumento nos casos, em vez de uma exacerbada incidência da doença.

Assim concluímos a exposição da plataforma web, uma ferramenta que, apesar de

Figura 23 - Dashboard de Mastite.



Fonte: Elaborada pelo autor (2023).

sua simplicidade, mostra-se abrangente e capaz de oferecer apoio aos produtores.

6 EXPERIMENTO

O experimento apresentado neste trabalho foi realizado no primeiro *Compost Barn* destinado a pesquisas do Brasil, através da Embrapa, localizado na cidade de Coronel Pacheco, Minas Gerais. O *Compost Barn* em questão foi projetado com uma capacidade média para abrigar até cem vacas, todas pertencentes à raça Holandesa.

O composto possui uma cama com dimensões de 77 x 16 x 0,9 metros, garantindo um amplo espaço para os animais. Conta com 22 exaustores, sendo oito manuais e o restante em funcionamento automático, permitindo um controle adequado da ventilação. Além disso, possui uma placa evaporativa que auxilia na regulação da umidade e temperatura internas, proporcionando um ambiente confortável e saudável para o bem-estar dos animais alojados.

A Figura 24 apresenta a estruturação do sistema de confinamento adotado. A ilustração detalha a disposição de dois corredores distintos, um designado para manejo e outro para trato, enfatizando a organização eficiente do ambiente. Adicionalmente, os pontos estratégicos onde o rebanho é alojado são identificados por meio de Ps, divididos em P1, P2 e P3, P4, P5 e P6, P7, P8 e P9, e, por fim, P10, P11 e P12. A flexibilidade do tamanho dos lotes é ressaltada, uma vez que são ajustados de acordo com as diferentes fases do rebanho, como lactação ou pré-parto, produção, dias em lactação e até mesmo o status reprodutivo dos animais, garantindo uma abordagem personalizada para melhor atender às necessidades específicas dos animais.

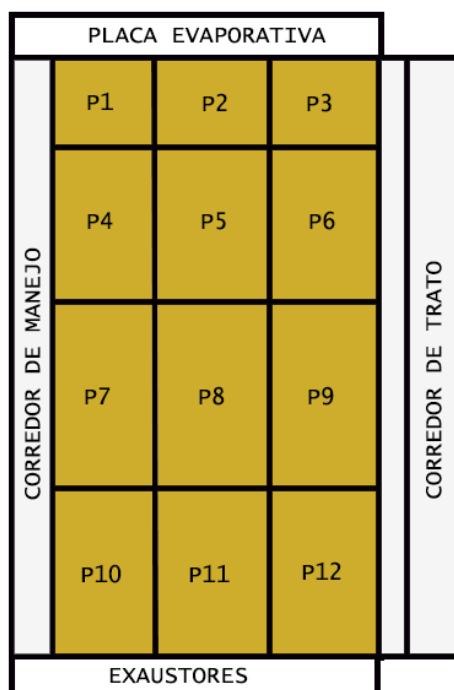
Para adquirirmos o *dataset* utilizado nesta pesquisa, foi necessário conduzir uma metódica consolidação de seis conjuntos de dados distintos, coletados de múltiplos equipamentos. Nas próximas seções, apresentaremos uma análise mais aprofundada de cada uma dessas fontes de dados.

6.1 AQUISIÇÃO DOS DADOS

Para os dados relativos à temperatura e à umidade, o composto é equipado com uma sonda TU, como exemplificado na imagem 25, altamente confiável e precisa para medir temperatura (0 a 60° C) e umidade (0% a 99%). Sua resolução de leitura é de 0,1° C e 1%, e o erro máximo é de apenas 0,5° C e 5%, respectivamente. Com velocidade de resposta de 1 segundo, fornece dados atualizados em tempo real.

Além da sonda, o composto também conta com o Termo-Higro-Anemômetro modelo THAR-300 26, um equipamento altamente versátil e preciso que fornece informações abrangentes sobre as condições ambientais. Com um Termopar tipo K, possui faixa de medição de temperatura de -100 a 1300°C e mede a umidade de 10% a 95%. Além disso, integra um anemômetro para medir velocidades de vento de 0,4 a 30 m/s e registrar leituras máximas e mínimas para referência futura.

Figura 24 - Estrutura do *Compost Barn* da Embrapa.



Fonte: Elaborada pelo autor (2023).

Figura 25 - Sonda TU utilizada no composto.



Fonte: Extraída de Inobram ⁽⁷²⁾ e adaptada pelo Autor

Além disso, integrando-se ao conjunto de dados, incorporam-se os registros obtidos pela sofisticada estação meteorológica automática A557, pertencente ao Instituto Nacional de Meteorologia (Inmet), órgão vinculado ao Ministério da Agricultura e Pecuária, localizada em Coronel Pacheco. Essa estação se mostrou fundamental ao disponibilizar

Figura 26 - TERMO-HIGRO-ANEMÔMETRO DIGITAL PORTÁTIL.



Fonte: Extraída de Inobram ⁽⁷²⁾ e adaptada pelo Autor

informações de suma relevância para o desfecho conclusivo deste estudo, abrangendo parâmetros como temperatura máxima, temperatura mínima, pressão máxima, dentre outros, que desempenharam papel crucial na obtenção dos resultados alcançados.

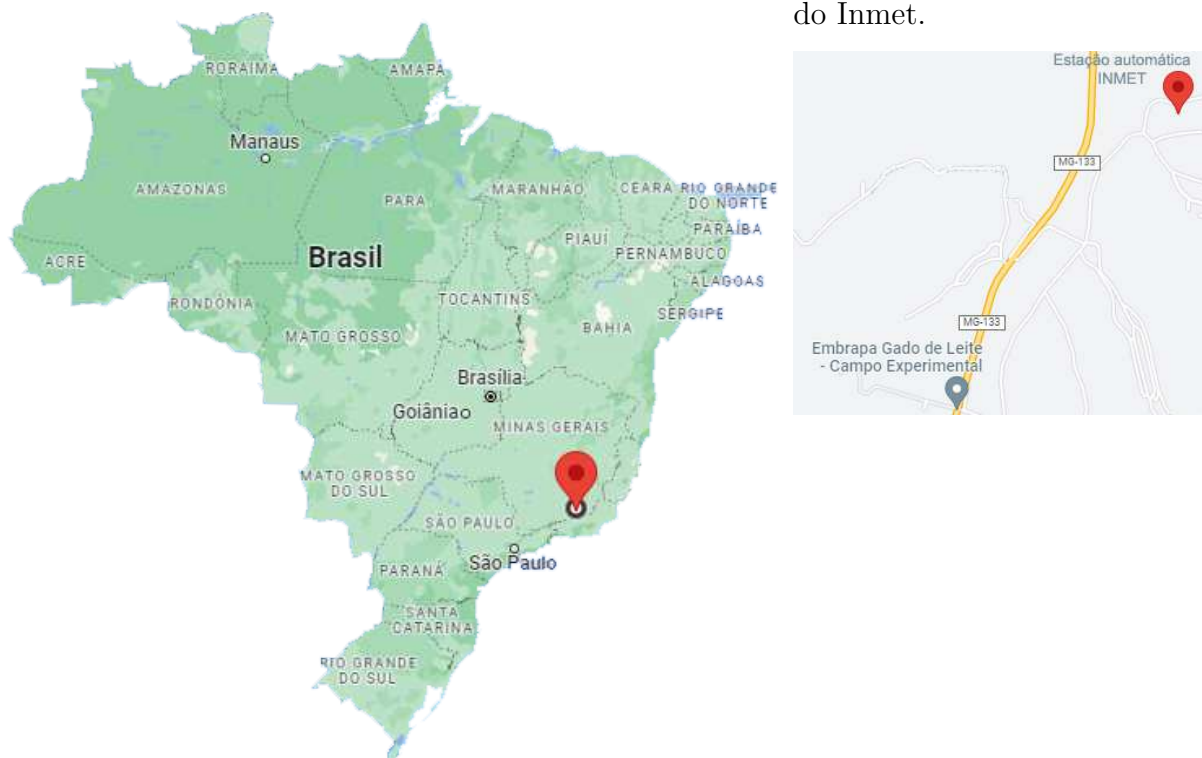
Conforme os dados fornecidos pelo Inmet, a estação meteorológica em questão está localizada precisamente nas coordenadas geográficas (-21.546667, -43.261111). Essa localização geográfica é de particular relevância, pois permite visualizar, na Figura 27a, a notável proximidade da estação com o campo experimental da Embrapa, onde o composto de interesse está localizado. A distância entre ambos é de menos de 2 km, o que reforça a significativa vinculação espacial entre a estação e o campo experimental, tornando os dados climáticos obtidos pela estação altamente representativos.

Além dos equipamentos mencionados anteriormente, há um controle realizado pelo composto para registrar a presença dos animais dentro da instalação em datas específicas. Adicionalmente, há registros provenientes da ordenhadeira que coletam informações sobre a quantidade de leite produzido por cada animal. Em alguns casos, também são registradas informações laboratoriais do leite, como a contagem de células somáticas, a contagem de bactérias totais e outros dados relevantes. Outro registro essencial é o peso do animal, que é registrado mensalmente.

Ademais, são registrados casos de mastite no rebanho, com informações detalhadas sobre qual teto foi afetado, o grau de infecção e quais bactérias foram identificadas como causadoras.

Figura 27 - Localização do *Compost Barn*

(a) Localização do *Compost Barn* em relação ao Mapa do Brasil
 (b) Distância do *Compost Barn* em relação à estação automática do Inmet.



Fonte: Elaborada pelo autor (2023).

6.2 BASE DE DADOS

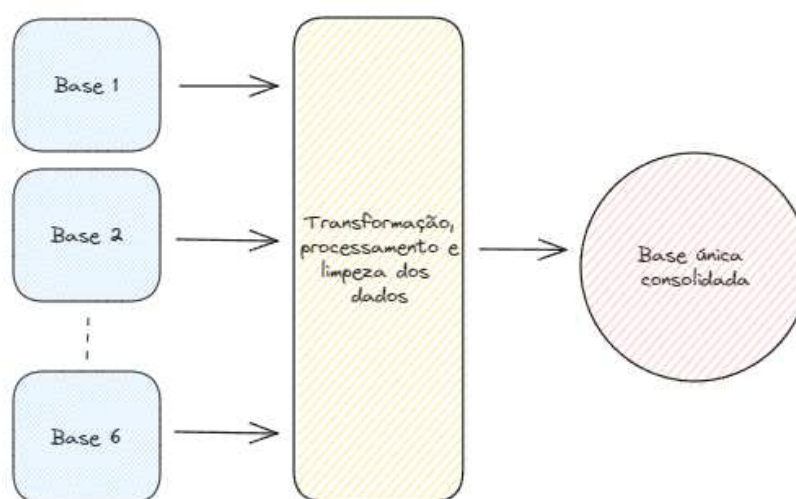
A criação da base de dados necessária para o treinamento do modelo de aprendizado de máquina envolveu a compilação não apenas dos registros provenientes dos sensores mencionados anteriormente, mas também a aquisição de dados provenientes de outras fontes relevantes mencionadas na seção 6.1. Antes de consolidar esses dados, foi essencial realizar um processo de organização e correção das informações, a fim de resolver problemas de inconsistência. Posteriormente, as informações de múltiplas origens foram agregadas em uma base de dados centralizada e coesa.

A etapa de preparação dos dados envolveu uma rigorosa abordagem para lidar com a ausência de alguns registros. Estabeleceu-se um limiar de 30% de dados faltantes como critério de corte, resultando na exclusão das colunas afetadas. Em determinadas instâncias, optou-se por preencher as lacunas utilizando a média das informações, enquanto em outras situações, empregou-se o método ffill, que é frequentemente utilizado para completar valores em branco em sequências de dados, baseando-se em observações anteriores. Adicionalmente, utilizou-se o método bfill, que preenche os valores ausentes com informações imediatamente subsequentes. Essas estratégias de tratamento de dados visaram assegurar a integridade e

a coesão dos conjuntos de dados, essenciais para análises sólidas e confiáveis.

Considerando que se está lidando com seis bases de dados, cada uma apresentando características distintas, foi preciso unificá-las. Nesse sentido, a temporalidade dos dados foi usada como critério para combinar as informações. Um período abrangendo julho de 2021 até dezembro de 2022 foi selecionado para a construção da base de dados. A Figura 28 mostra como esse processo de consolidação aconteceu.

Figura 28 - Processo de união dos dados.



Fonte: Elaborada pelo autor (2023).

Dentro das seis bases mencionadas, foi preciso criar informações artificiais para preencher lacunas de dados. Isso foi necessário porque certas datas no período mencionado anteriormente não tinham registros disponíveis. Esse preenchimento foi feito de maneira aleatória, seguindo uma faixa de valores mínimos e máximos específicos para cada variável afetada por essa falta de dados.

Após a consolidação das informações, criou-se uma base de dados contendo 49 colunas, sendo a última destinada à variável alvo que indica a presença ou ausência de contaminação por mastite no animal. Ademais, a referida base é composta de um total de 447 registros.

No entanto, esse processo não se encerra aqui. Dado que, nesta abordagem, os usuários têm a flexibilidade de escolher entre três opções, conforme identificado durante a avaliação do sistema web, houve a necessidade de estratificar os conjuntos de dados de acordo com a significância das variáveis. Esse procedimento permite uma análise mais aprofundada das diferentes variáveis e sua magnitude de influência.

Dado o grande número de variáveis envolvidas, optou-se por empregar a abordagem do *SelectKBest*. Trata-se de um componente integrante da biblioteca *scikit-learn* que seleciona as *k* características mais relevantes, considerando suas pontuações distintivas.

Essas pontuações são derivadas por meio de uma análise estatística univariada, na qual cada variável é avaliada isoladamente (73). No nosso cenário específico deste estudo, decidiu-se adotar o parâmetro $f_classif$, o qual aplica o cálculo do valor F da Análise de Variância (ANOVA) para a amostra disponibilizada.

A análise de variância avalia se as diferenças médias entre grupos de pontuações são estatisticamente significativas. Esses grupos podem ser definidos por diferentes níveis de uma variável independente ou combinações de níveis de várias variáveis independentes. A análise compara a diferença entre as médias com a variabilidade natural das medições para determinar se a diferença é estatisticamente relevante. Por exemplo, em um estudo de medicamentos, compara-se se as médias das avaliações de dor são diferentes entre um grupo de controle e um grupo de tratamento. A análise de variância é usada para discernir se as diferenças observadas são mais do que o esperado por acaso (74).

Para calcular o valor F, é essencial inicialmente determinar a soma dos quadrados dentro dos grupos (SQDG). Essa soma, abrangendo todos os valores em todos os grupos, envolve as diferenças ao quadrado entre cada valor individual e a média do respectivo grupo. Adicionalmente, é necessário derivar os grau de liberdade dentro dos grupos (GLDG). Essa métrica é obtida ao subtrair o número de grupos (k) do total de observações (N), no qual N representa o somatório completo de observações e k representa o número de grupos.

Com base nas informações fornecidas, é necessário efetuar o cálculo da variância entre grupos (VEG). Esse cálculo envolve a determinação da soma dos quadrados entre grupos (SQEG), que consiste em uma soma ponderada das diferenças ao quadrado entre as médias dos grupos individuais e a média geral. Essa ponderação leva em consideração os tamanhos das amostras presentes em cada um dos grupos. Adicionalmente, os grau de liberdade entre grupos (GLEG) são estabelecidos através da subtração de 1 do número total de grupos (k - 1, em que k representa o número de grupos). Os detalhes sobre como essa operação é executada estão apresentados na equação 6.1.

$$VEG = \frac{SQEG}{GLEG} \quad (6.1)$$

Além da variância entre grupos (VEG), é necessário encontrar a variância dentro dos grupos (VDG), conforme vemos a equação 6.2.

$$VDG = \frac{SQDG}{GLDG} \quad (6.2)$$

A essência da soma dos quadrados dentro dos grupos reside na agregação das diferenças elevadas ao quadrado entre cada valor individual e a média correspondente ao seu respectivo grupo. Esse processo é aplicado a todos os valores presentes em cada um dos grupos. Por sua vez, os graus de liberdade dentro dos grupos emergem da subtração

do número de grupos a partir do total de observações ($N - k$, em que N representa o total de observações e k simboliza o número de grupos).

A Tabela 7 apresenta as 24 variáveis que foram selecionadas como a base fundamental para o experimento devido à sua significativa contribuição na previsão da ocorrência de mastite. As restantes 24 colunas foram excluídas em virtude de sua importância percentual reduzida. Os testes efetuados demonstraram uma aprimorada performance na predição após a eliminação dessas 24 colunas. Por outro lado, quando todas as 48 variáveis foram mantidas, constatou-se uma diminuição na taxa percentual de precisão.

Tabela 7 – Importância das variáveis

Variável	Porcentagem (%)
Pressão mínima em hectopascal	17.27
Pressão máxima em hectopascal	16.10
Pressão instantânea em hectopascal	16.07
Temperatura ambiente média	15.14
Rajada de vento (em metros por segundo) externa	15.12
Contagem de células somáticas (x1000)	12.80
Peso	12.73
Temperatura máxima externa	9.37
Temperatura instantânea externa	8.83
Temperatura mínima externa	8.44
Ponto de orvalho mínimo externo	8.35
Ponto de orvalho instantâneo externo	8.16
Velocidade do vento externa	7.75
Ponto de orvalho máximo externo	7.66
Umidade mínima externa	7.39
Produção da segunda ordenha	7.25
Umidade da cama	7.03
Chuva (mm)	6.71
Umidade instantânea externa	6.50
Produção da primeira ordenha	6.40
Umidade máxima externa	6.25
Umidade externa	4.90
Produção Total de Leite	4.42
Temperatura média da cama	4.26

Fonte: Elaborada pelo autor (2023).

Após a seleção criteriosa dessas 24 variáveis, foi deliberado proceder à segmentação da base de dados em três conjuntos distintos, levando em consideração a relevância de cada variável, expressa em termos percentuais. Ficou estabelecido que seriam realizados testes utilizando as 24 variáveis restantes, bem como um conjunto contendo apenas as 12 variáveis principais, selecionadas com base nas suas maiores porcentagens de importância. Além disso, um terceiro conjunto foi composto de apenas seis variáveis, também escolhidas com base em suas elevadas porcentagens de importância. Essa abordagem metódica foi

adotada com o objetivo de simplificar e viabilizar a participação de agricultores com recursos limitados que possuam capacidade mínima para implantar um sistema de *Compost Barn*, os quais por vezes não dispõem dos recursos para conduzir análises de maior complexidade.

É fundamental salientar que a segmentação em três bases de dados implica custos subjacentes. Em certas circunstâncias, os pequenos produtores que eventualmente desejem empregar esses modelos para as três bases de dados podem se deparar com limitações financeiras que os impeçam de investir na infraestrutura tecnológica necessária para obter todas as informações contidas na base de dados deste estudo, composta de 24 variáveis. Isso, por sua vez, dificulta a extração das informações por meio de sensores, indispensável para alcançar o resultado desejado. Assim, é importante ressaltar que à medida que o número de variáveis aumenta, os custos de investimento também crescem, uma vez que a obtenção dessas informações requer uma variedade de sensores e estratégias distintas, a fim de serem incorporadas com eficácia ao modelo treinado.

Na instância de dados consolidada, é possível identificar a presença de 300 registros caracterizados pela ausência de mastite, enquanto 147 registros apresentam a presença dessa condição.

6.3 RESULTADOS E DISCUSSÃO

6.3.1 Experimentos computacionais

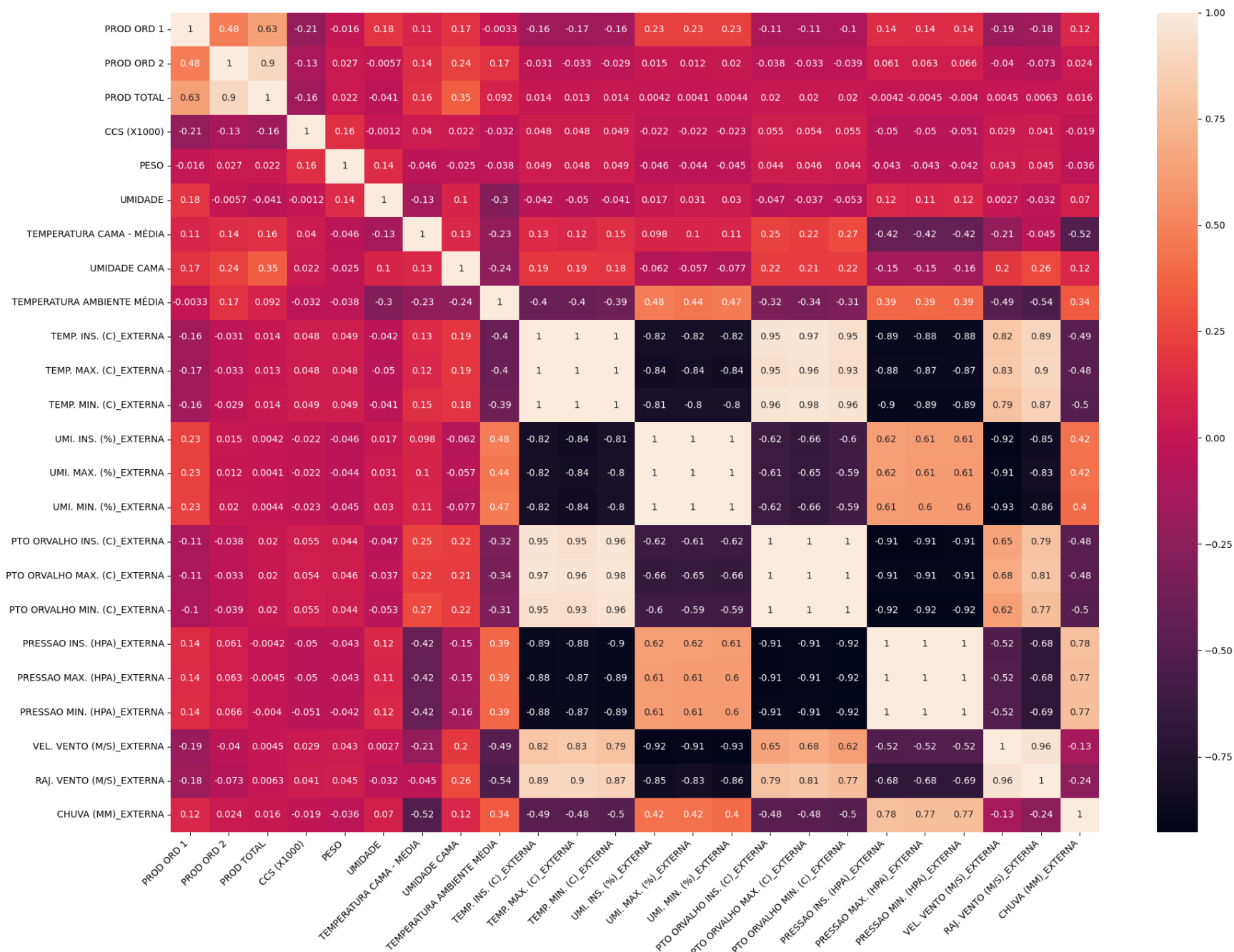
A pesquisa descrita neste estudo foi conduzida utilizando um sistema computacional portátil Dell G15 5520, equipado com um processador Core i7-12700H, uma placa de vídeo RTX 3060, uma capacidade de memória RAM de 16 GB e um armazenamento SSD de 512 GB.

Na Figura 29, é possível visualizar e compreender como a base de dados se estrutura, bem como perceber as relações de interdependência que se estabelecem entre as variáveis presentes. Essa representação gráfica permite identificar quais variáveis possuem padrões de variação similares ou opostos, contribuindo para uma compreensão mais profunda das associações e influências mútuas entre os diferentes elementos da base de dados.

O experimento foi implementado por meio da execução de três modelos distintos - AutoGluon, H2O e TPOT - sobre as três bases de dados que haviam sido construídas previamente. Cada um desses modelos foi aplicado individualmente a cada uma das três bases de dados distintas. Mantendo consistência, a mesma semente foi adotada em todas as iterações do experimento, assegurando resultados comparáveis. A divisão entre os conjuntos de treinamento e teste foi fixada em uma proporção de 80/20 para manter uma abordagem padronizada de avaliação.

Além dos experimentos individuais com os modelos AutoGluon, H2O e TPOT, uma análise adicional foi conduzida para avaliar a possibilidade de melhorar o desempenho

Figura 29 - Matriz de correlação das 24 variáveis.



Fonte: Elaborada pelo autor (2023).

geral por meio da fusão dos resultados desses modelos. Nessa abordagem de fusão, os resultados dos três modelos foram combinados de forma a tirar uma média ponderada das previsões geradas por cada um deles. Essa média ponderada foi então utilizada como a previsão final para as instâncias de dados em questão.

Adicionalmente, foi empregada uma abordagem de validação cruzada utilizando cinco conjuntos (folds), o que contribuiu para uma avaliação mais robusta e confiável do desempenho dos modelos em diferentes cenários. Essa estratégia permitiu a obtenção de métricas de avaliação mais consistentes e representativas do desempenho geral dos modelos ao longo de várias iterações do experimento.

A validação cruzada K-fold é uma técnica que avalia modelos de aprendizado de máquina de forma confiável, especialmente para evitar *overfitting*¹ e *underfitting*² em conjuntos de dados limitados. Ela envolve dividir os dados em K partes de tamanho similar. O modelo é treinado K vezes, usando cada parte como conjunto de teste e as demais como treinamento. Isso garante que todas as partes do conjunto sejam usadas tanto para treinamento quanto para teste, evitando depender de uma única divisão. Essa abordagem contribui para uma avaliação abrangente e imparcial do desempenho do modelo (75).

6.3.2 Análises

Prosseguindo, uma análise detalhada do desempenho dos modelos de AutoML foi conduzida, submetendo-os a uma avaliação rigorosa com base em métricas essenciais, incluindo acurácia, precisão, recall, F1 e a Área sob a Curva ROC (AUC). Essa análise foi conduzida separadamente para cada uma das três bases de dados testadas. Realizou-se uma comparação minuciosa entre os diferentes modelos de AutoML empregados em relação a cada base de dados. No entanto, antes de entrar em detalhes, foi fundamental examinar a metodologia subjacente ao cálculo de cada uma dessas métricas.

A acurácia, em geral, mede a proporção de previsões corretas em relação ao número total de instâncias avaliadas. Ela é calculada levando em conta a soma dos elementos Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN) no numerador, e a soma de todos os elementos da matriz de confusão, incluindo Falso Positivo (FP) e Falso Negativo (FN), no denominador, conforme a equação 6.3. Os elementos VP e VN representam as classificações corretas feitas pelo modelo e são encontrados na diagonal principal da matriz de confusão. Além disso, o denominador engloba todos os elementos fora da diagonal principal que

¹ Ocorre quando um modelo se ajusta excessivamente aos dados de treinamento, capturando até o ruído. Isso resulta em má generalização para novos dados, prejudicando o desempenho real do modelo. É um problema de complexidade excessiva.

² Acontece quando um modelo é muito simples para capturar os padrões dos dados de treinamento, levando a um desempenho insuficiente tanto nos dados de treinamento quanto em novos dados. O modelo não consegue capturar a complexidade dos dados, resultando em baixa precisão e generalização inadequada.

foram incorretamente classificados pelo modelo, incluindo tanto os falsos positivos quanto os falsos negativos (76).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (6.3)$$

A fórmula apresentada na equação 6.4 está relacionada à métrica de Precisão. Essa métrica estabelece uma relação entre os elementos VP e o total de unidades que foram previstas como pertencentes à classe positiva (a soma dos casos previstos como positivos). No contexto específico, o termo VP refere-se aos elementos que o modelo classificou como positivos e que de fato são positivos, enquanto o termo FP indica os elementos que o modelo classificou como positivos, mas que, na realidade, pertencem à classe negativa (76).

$$Precisão = \frac{VP}{VP + FP} \quad (6.4)$$

O *Recall* é a proporção de VP em relação às unidades classificadas como positivas. Ele é calculado pela equação 6.5. Essa métrica avalia a habilidade do modelo em identificar unidades Positivas no conjunto de dados (77).

$$Recall = \frac{VP}{VP + VN} \quad (6.5)$$

A métrica F1, conhecida como F1-Score, também avalia o desempenho de modelos de classificação a partir da matriz de confusão, combinando as métricas de Precisão e Recall através da média harmônica. A equação 6.6 pertencente ao F1-Score pode ser entendida como uma média ponderada entre Precisão e Recall, onde o F1-Score atinge sua melhor pontuação em 1 e a pior pontuação em 0. A contribuição relativa de precisão e recall é equilibrada no F1-Score, sendo a média harmônica útil para determinar a melhor harmonização entre essas duas medidas (76).

$$F1 = \left(\frac{2}{\text{Precisão}^{-1} + \text{Recall}^{-1}} \right) = 2 \times \left(\frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}} \right) \quad (6.6)$$

Por fim, mas não menos importante, tem-se a Área sob a Curva ROC, também conhecida como ROC AUC. É uma medida quantitativa que captura a essência da Curva ROC em um único valor conciso. Sua escala varia de 0 a 1, oferecendo uma avaliação da capacidade discriminatória de um modelo. Quanto maior o valor associado à métrica ROC AUC, mais eficaz é o desempenho do modelo em discernir entre as categorias positiva e negativa (78).

Uma ROC AUC de 0.5 indica que o modelo está realizando classificações de maneira equivalente a uma seleção aleatória, ao passo que um valor próximo de 1 indica que o modelo demonstra uma competência notável em separar eficazmente as duas classes em consideração (78). Ao examinar a equação 6.7, tem-se a oportunidade de investigar como

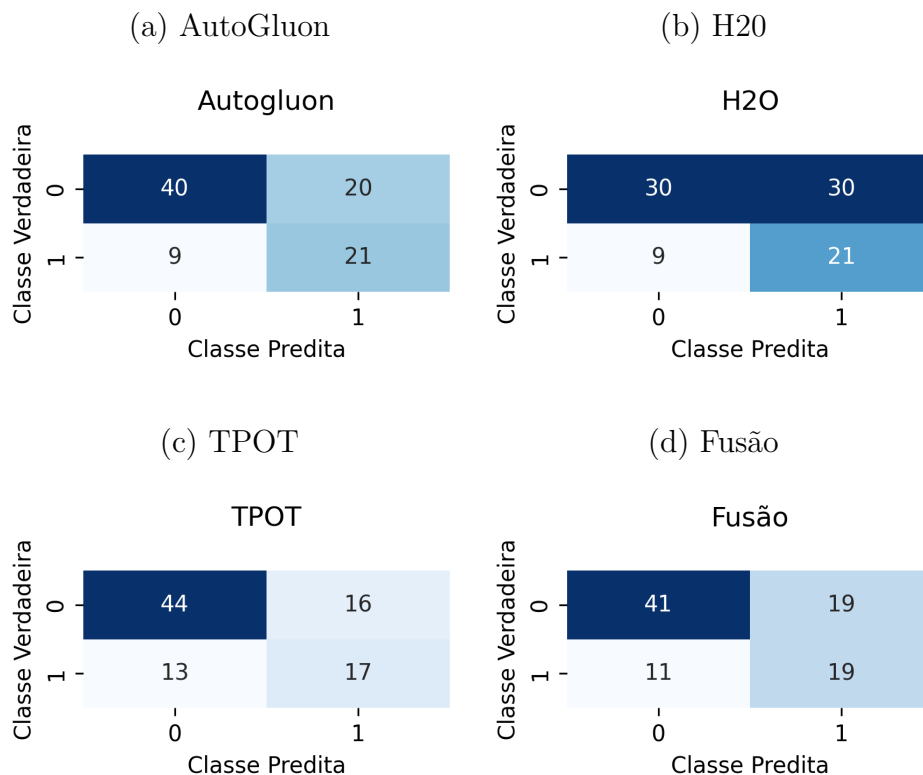
a sua expressão matemática se desdobra. Nesse contexto, destaca-se que TVP denota a taxa de verdadeiros positivos e TFP corresponde à taxa de falsos positivos. É relevante mencionar que $TFP^{-1}(t)$ se configura como a função inversa da TFP. Adicionalmente, vale ressaltar que a amplitude da integral, que se estende de 0 a 1, espelha a ampla gama de limiares de classificação considerados.

$$\text{ROC AUC} = \int_0^1 \text{TVP}(TFP^{-1}(t)) dt \quad (6.7)$$

6.3.2.1 Análise na base de dados de seis variáveis

Nesta seção, os resultados das métricas derivadas do experimento que foi conduzido na base de dados foram analisados. Esse experimento se concentrou nas seis variáveis mais relevantes, que foram escolhidas a partir de um conjunto total de 24 variáveis disponíveis na base de dados deste estudo. Para uma análise mais aprofundada, a matriz de confusão 30 foi explorada inicialmente.

Figura 30 - Matriz de Confusão para base de dados de seis variáveis.



Fonte: Elaborada pelo autor (2023).

Ao examinar a matriz de confusão do AutoGluon 30a, diferentes cenários de classificação foram identificados. Notou-se que havia 40 instâncias que foram corretamente classificadas como pertencentes à classe negativa, conhecidas como VN. Além disso, foram

encontradas 20 instâncias que foram erroneamente classificadas como pertencentes à classe positiva, apesar de serem da classe negativa, caracterizando os FPs.

Adicionalmente, observou-se que havia nove instâncias que foram classificadas erroneamente como pertencentes à classe negativa, embora fossem da classe positiva, configurando os FNs. Por último, identificou-se 21 instâncias que foram corretamente classificadas como pertencentes à classe positiva, representando os VPs.

Já a matriz de confusão 30b, proveniente do modelo H2O, foi capaz de identificar 30 casos como negativos e 21 como positivos, demonstrando VN e VP.

Contudo, em sua busca por precisão, o modelo também errou: nove exemplos foram classificados equivocadamente como negativos, quando deveriam ser positivos. Tais lapsos são os FNs, retratando momentos em que identificou padrões inexistentes.

O modelo H2O deixou passar 30 exemplos que deveriam ser negativos, porém foram considerados positivos, refletindo os FPs.

Ao examinar a matriz gerada pelo modelo TPOT 30c, pôde-se observar que ele teve um acerto notável em 44 situações, identificando corretamente casos sem mastite. Adicionalmente, em 17 ocasiões, o modelo se saiu bem novamente, dessa vez, identificando corretamente situações em que os animais estavam contaminados por mastite.

Contudo, é importante notar que o modelo também apresentou alguns equívocos. Ele erroneamente assinalou 16 situações como casos de mastite, quando, na verdade, eram situações de animais não contaminados por mastite. Da mesma forma, o modelo identificou incorretamente 13 situações como não contaminadas por mastite, quando, na realidade, eram casos de mastite.

Através da combinação dos três modelos, a matriz de confusão representada na Figura 30d acertou em suas previsões 41 vezes, diagnosticando corretamente a não presença de mastite nos animais. No entanto, a matriz também apresentou 19 situações em que houve erro: nesses casos, o modelo indicou erroneamente que os animais tinham mastite, porém, estes não estavam contaminados.

Além disso, o modelo demonstrou uma capacidade correta de identificar 19 casos em que os animais estavam contaminados com mastite. No entanto, houve 11 instâncias em que o modelo cometeu um equívoco, prevendo erroneamente a não presença de mastite nos animais que estavam contaminados.

Após examinar a matriz de confusão, é importante concentrar a atenção na avaliação do desempenho dos modelos com base nas métricas estabelecidas neste estudo. A referência à Tabela 8 foi fundamental para orientar as análises que se seguiram.

É evidente que os modelos AutoGluon e TPOT alcançaram a maior pontuação na métrica de acurácia, registrando um valor de 0,6778. Isso sugere um desempenho mais preciso quando comparado com as outras abordagens.

Quanto à métrica de precisão, o modelo TPOT também se destacou, obtendo uma pontuação mais elevada de 0,5152. Isso indica que suas previsões de falsos positivos são menos frequentes em relação aos demais modelos. No que diz respeito ao recall, tanto o AutoGluon quanto o H2O exibiram um desempenho notável, alcançando uma pontuação de 0,7000 em comparação com os demais modelos. Isso sugere uma capacidade consistente desses dois modelos em identificar e recuperar efetivamente os casos relevantes, superando seus concorrentes.

O modelo AutoGluon obteve o maior valor de F1, atingindo 0.5915. Essa pontuação realça a habilidade do AutoGluon em harmonizar o equilíbrio entre precisão e recall, resultando em um desempenho global competitivo.

Tabela 8 – Métricas para base de dados de seis variáveis

Métrica	AutoGluon	H2O	TPOT	Fusão
Acurácia	0.6778	0.5667	0.6778	0.6667
Precisão	0.5122	0.4118	0.5152	0.5000
Recall	0.7000	0.7000	0.5667	0.6333
F1	0.5915	0.5185	0.5397	0.5588

Fonte: Elaborada pelo autor (2023).

Já para métrica da Área sob a Curva ROC, todos os modelos têm um desempenho razoável na métrica ROC AUC, indicando que eles são capazes de distinguir entre as classes positiva e negativa em um grau considerável. Quanto mais próximo o valor estiver de 1, melhor o modelo é em separar as classes. Observando a Figura 31, podemos ver que os modelos estão muito próximos, com uma leve vantagem para o AutoGluon.

6.3.2.2 Análise na base de dados de 12 variáveis

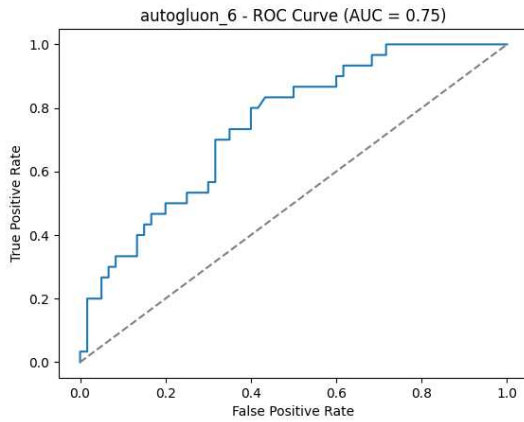
Nesta seção, foram investigados os resultados das métricas obtidas a partir do experimento que foi conduzido com a base de dados deste estudo, com as 12 variáveis consideradas mais relevantes, selecionadas dentre um conjunto total de 24 variáveis disponíveis no conjunto de dados da pesquisa. Para uma análise mais detalhada, inicialmente, examinou-se a matriz de confusão apresentada na Figura 32.

A matriz de confusão 32a revelou que o modelo AutoGluon teve sucesso em identificar corretamente as vacas sem mastite em 43 casos e as vacas com mastite em 17 casos. No entanto, o modelo errou ao prever incorretamente em 13 casos que vacas não tinham mastite quando na realidade tinham e em 17 casos, que vacas tinham mastite quando na verdade não tinham.

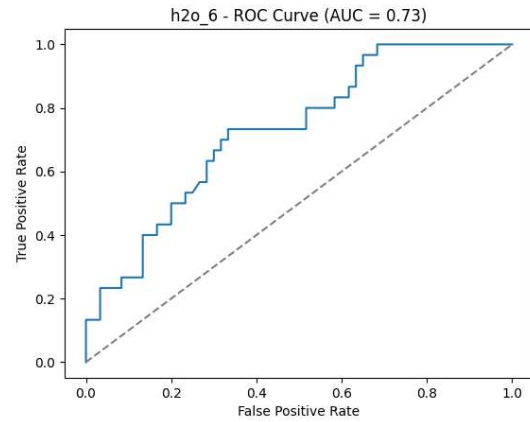
Ao analisar a matriz de confusão do modelo H2O 32b, foi possível concluir que o modelo teve sucesso em identificar corretamente 49 casos em que as vacas não tinham mastite e 16 casos em que as vacas tinham mastite. No entanto, o modelo errou ao prever

Figura 31 - Métrica ROC Curve para seis variáveis.

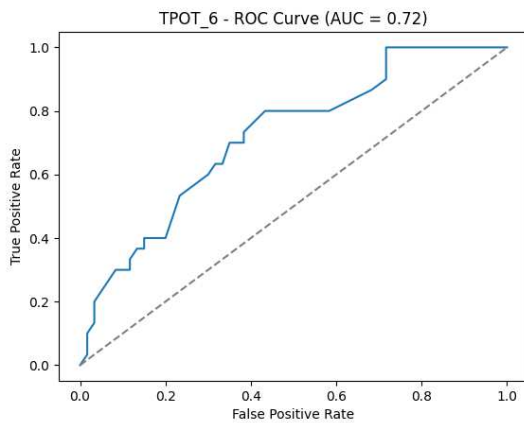
(a) AutoGluon



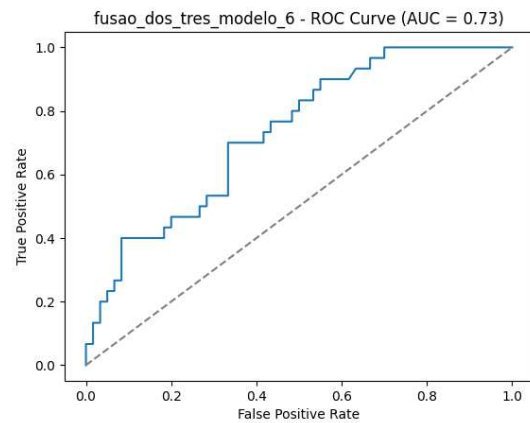
(b) H2O



(c) TPOT



(d) Fusão



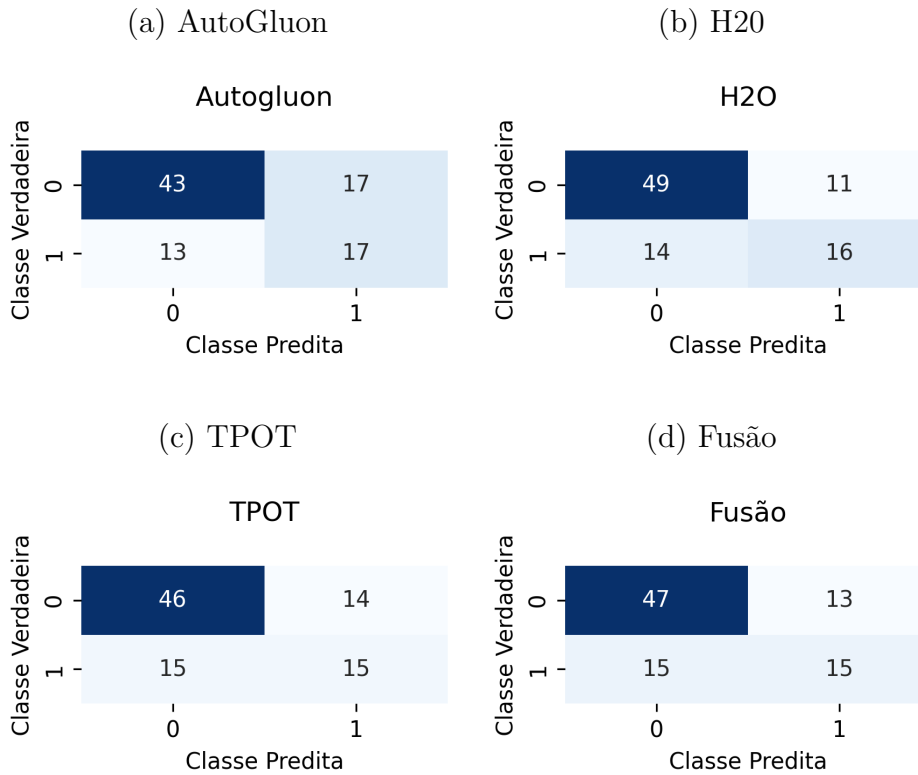
Fonte: Elaborada pelo autor (2023).

incorretamente em 14 casos que vacas não tinham mastite quando tinham e em 11 casos, que vacas tinham mastite quando na realidade não tinham.

No que diz respeito à matriz de confusão 32c, foi possível concluir que o modelo TPOT teve sucesso em identificar corretamente 46 casos em que as vacas não tinham mastite e 15 casos em que as vacas tinham mastite. No entanto, o modelo errou ao prever incorretamente em 15 casos que vacas não tinham mastite quando tinham e em 14 casos, que vacas tinham mastite quando na verdade não tinham.

A matriz de confusão da fusão dos modelos 32d mostrou que o modelo teve sucesso em identificar corretamente 47 casos em que as vacas não tinham mastite e 15 casos em que as vacas tinham mastite. No entanto, a matriz errou ao prever em 15 casos que vacas não tinham mastite quando tinham e em 13 casos, que vacas tinham mastite quando na realidade não tinham.

Figura 32 - Matriz de Confusão para base de dados de 12 variáveis.



Fonte: Elaborada pelo autor (2023).

Depois de analisar a matriz de confusão, é essencial direcionar o foco para a avaliação do desempenho dos modelos, usando as métricas definidas neste estudo como base. A utilização da Tabela 9 é crucial para guiar as análises subsequentes.

Tabela 9 – Métricas para base de dados de 12 variáveis

Métrica	AutoGluon	H2O	TPOT	Fusão
Acurácia	0.6667	0.7222	0.6778	0.6889
Precisão	0.5000	0.5926	0.5172	0.5357
Recall	0.5667	0.5333	0.5000	0.5000
F1	0.5313	0.5614	0.5085	0.5172

Fonte: Elaborada pelo autor (2023).

O modelo que apresentou o desempenho mais promissor com base em três métricas distintas - acurácia, precisão e F1-Score - foi o H2O. Ao analisar os resultados das métricas, observou-se um padrão consistente de destaque do modelo H2O em termos de performance.

Quando se examina a métrica de acurácia, fica claro que o modelo H2O alcançou uma taxa de 72,22%. Isso significava que suas previsões foram corretas para aproximadamente 72,22% das instâncias avaliadas, destacando sua habilidade em compreender padrões e generalizar a partir dos dados.

No quesito precisão, o H2O também demonstrou um desempenho notável, atingindo uma precisão de 0,5926. Em termos simples, isso indicava que cerca de 59,26% das vezes em que o modelo emitia uma previsão positiva, essa previsão estava correta. Essa alta precisão refletia a capacidade do modelo H2O em fazer distinções precisas entre as instâncias positivas e negativas do conjunto de dados.

Adicionalmente, o modelo H2O se destacou na métrica de F1-Score, obtendo um valor de 0,5614. O F1-Score é especialmente relevante pois considera tanto a precisão quanto o recall das previsões de um modelo, e esse resultado aponta para um equilíbrio sólido entre esses dois aspectos cruciais da avaliação de um modelo de machine learning.

O desempenho consistente do modelo H2O em todas essas métricas ressalta sua capacidade de oferecer previsões precisas e abrangentes. Já o modelo que se destacou como o líder em termos de métrica Recall foi o AutoGluon. Ao observar os resultados apresentados, pode-se perceber que o modelo AutoGluon atingiu um Recall de 0,5667. Isso significa que ele foi capaz de identificar aproximadamente 56,67% das instâncias positivas contidas no conjunto de dados.

Examinando a representação gráfica da Área sob a Curva ROC no Gráfico 33, torna-se aparente que, com base nos valores de ROC (AUC), o modelo H2O se destacou consideravelmente, apresentando o desempenho mais notável entre os modelos submetidos à avaliação. Seu valor superior sugeriu a eficácia particular do H2O em discriminar entre as classes positiva e negativa, um aspecto crucial em tarefas de classificação. Isso enfatizou sua posição como um modelo em destaque em relação aos demais, indicando uma habilidade sólida para tomar decisões precisas em diversos cenários.

Vale a pena ressaltar, no entanto, que os outros três modelos - AutoGluon, TPOT e a fusão dos modelos - também obtiveram desempenhos dignos de nota na métrica ROC (AUC). Embora o modelo H2O tenha apresentado a maior eficiência, os demais modelos demonstraram habilidades apreciáveis na diferenciação entre as classes, sugerindo que também eram opções viáveis para tarefas de classificação. Suas contribuições reforçaram a importância de considerar o contexto específico da aplicação e as métricas relevantes ao determinar a melhor abordagem para cada cenário.

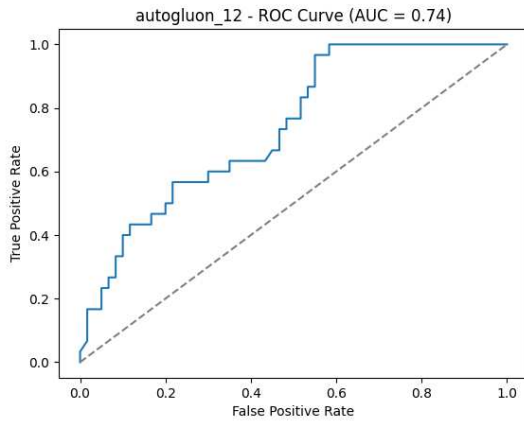
6.3.2.3 Análise na base de dados de 24 variáveis

Nesta seção, o enfoque direcionou-se à análise da base de dados, que engloba as 24 variáveis da base de dados. Iniciou-se essa incursão investigativa pelo exame da matriz de confusão 34, um instrumento essencial utilizado na avaliação dos modelos de classificação adotados.

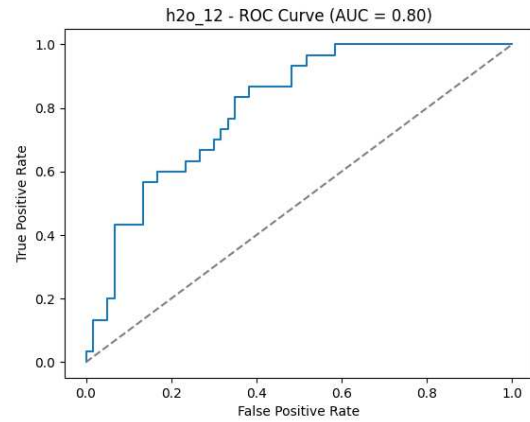
A matriz de confusão, ponto de partida da avaliação, proporciona uma representação concisa das previsões realizadas pelos modelos, em relação às classes reais das instâncias do conjunto de dados. Por meio dessa ferramenta, é possível observar a eficácia das previsões

Figura 33 - Métrica ROC Curve para 12 variáveis.

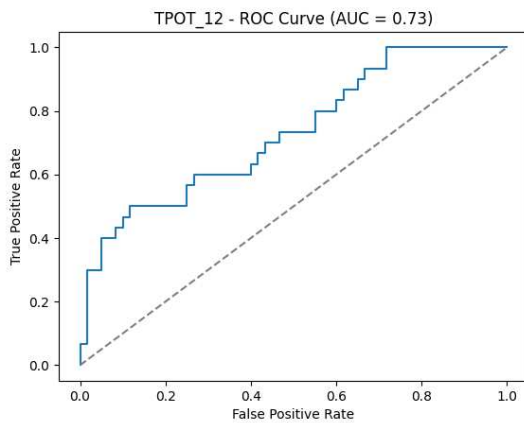
(a) AutoGluon



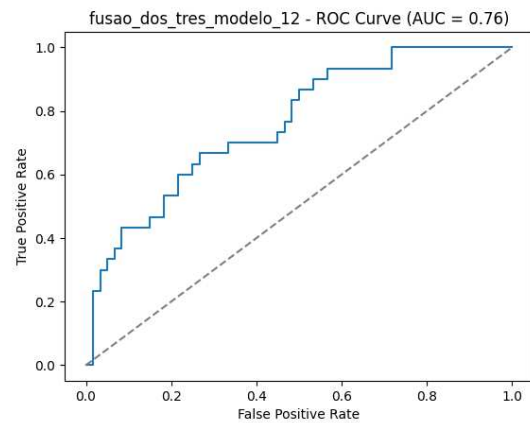
(b) H2O



(c) TPOT



(d) Fusão



Fonte: Elaborada pelo autor (2023).

feitas por cada modelo, assim como identificar áreas que demandam um aprimoramento.

Inicia-se pela abordagem da matriz de confusão do modelo AutoGluon 34a. Nesse contexto, os elementos situados na diagonal principal denotam as classificações precisas: 49 vacas sem mastite foram identificadas corretamente, ao passo que 19 vacas com mastite também receberam a classificação exata. Os valores localizados fora dessa diagonal espelham as classificações incorretas: 11 vacas sem mastite foram erradamente categorizadas como portadoras de mastite, enquanto outras 11 vacas com mastite foram equivocadamente classificadas como isentas da doença.

Prosseguindo com a análise dos resultados, a atenção se direciona agora para a matriz de confusão associada ao modelo H2O 34b. Nela, pode-se verificar que os elementos situados na diagonal principal, representados por 52 e 15, denotam as classificações corretas do modelo, refletindo que 52 instâncias de ausência de mastite foram adequadamente

identificadas e 15 casos de mastite foram corretamente classificados. No entanto, os valores fora dessa diagonal, indicando oito e 15, destacam erros de classificação. Notadamente, 8 instâncias que não apresentavam mastite foram incorretamente categorizadas como positivas, enquanto 15 instâncias com mastite foram equivocadamente classificadas como negativas.

Direcionando o foco da análise para a matriz de confusão do modelo TPOT 34c, ao examinar seus elementos, observamos que os valores 45 e 18, situados na diagonal principal, representam as classificações corretas. Isso indica que 45 instâncias sem mastite foram corretamente identificadas e 18 casos de mastite foram devidamente classificados. No entanto, os valores fora dessa diagonal, 15 e 12, indicam classificações incorretas. Especificamente, 15 casos que não apresentavam mastite foram categorizados erroneamente como positivos, enquanto 12 instâncias com mastite foram equivocadamente classificadas como negativas.

Por fim, a matriz de confusão 34d, representou a avaliação do modelo de fusão na tarefa de classificação de vacas com e sem mastite. Nessa matriz, os valores na diagonal principal (51 e 17) indicam as classificações corretas realizadas pelo modelo. Ou seja, 51 instâncias de vacas sem mastite foram corretamente identificadas e 17 instâncias de vacas com mastite foram classificadas de forma precisa. No entanto, os valores fora dessa diagonal (9 e 13) corresponderam a classificações incorretas. Nesse contexto, nove instâncias que não apresentavam mastite foram erroneamente categorizadas como positivas, enquanto 13 instâncias de vacas com mastite foram equivocadamente classificadas como negativas.

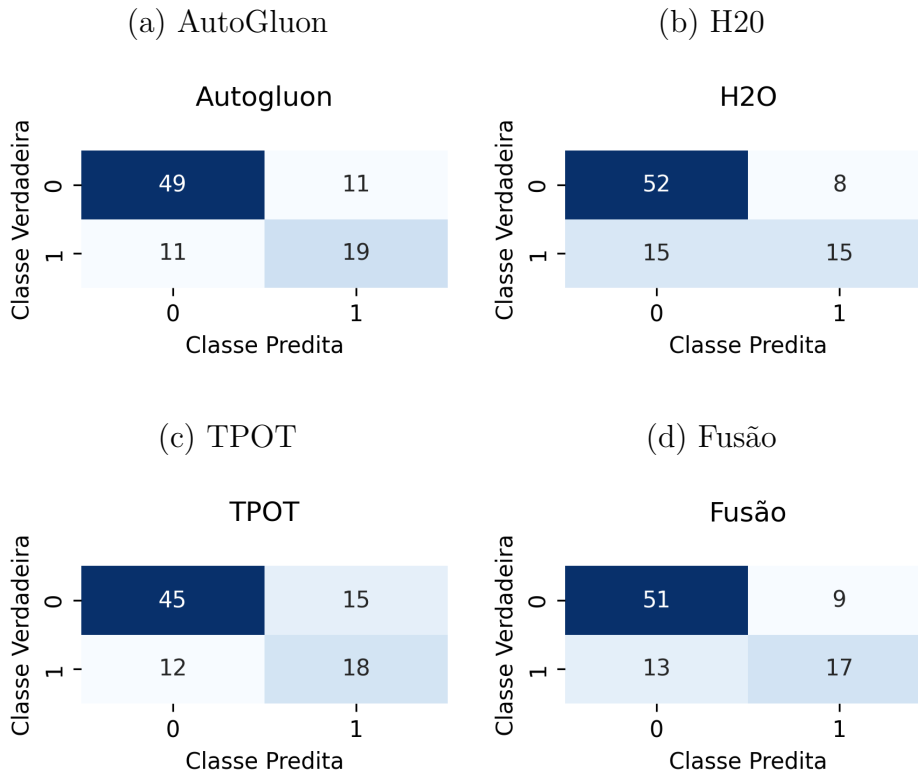
Após a análise da matriz de confusão referente às 24 variáveis, tornou-se fundamental direcionar a atenção para a avaliação do desempenho dos modelos, com base nas métricas estabelecidas neste estudo. Na Tabela 10 de métricas apresentada para a base de dados composta por 24 variáveis, é perceptível que diferentes modelos obtiveram desempenhos destacados em cada métrica específica.

Os modelos AutoGluon e a Fusão se destacaram ao alcançarem o melhor desempenho, registrando uma acurácia de 75.56%. Esse resultado indica que eles foram capazes de fazer previsões corretas para aproximadamente 75.56% das instâncias avaliadas.

Na métrica de precisão, que avalia a proporção de instâncias positivas corretamente classificadas em relação às instâncias que o modelo classificou como positivas, o modelo Fusão se destacou. Ele alcançou uma precisão de 65.38%, indicando que aproximadamente 65.38% das instâncias classificadas como positivas pelo modelo eram realmente positivas. Por outro lado, o modelo H2O obteve uma precisão ligeiramente inferior, com 65.22%.

O modelo que obteve o melhor desempenho no recall foi o AutoGluon, com um recall de 0.6333 (ou cerca de 63.33%). Isso significa que o modelo AutoGluon foi capaz de capturar e identificar corretamente aproximadamente 63.33% das instâncias positivas presentes no conjunto de dados

Figura 34 - Matriz de Confusão para base de dados de 24 variáveis.



Fonte: Elaborada pelo autor (2023).

Ao considerar a métrica F1, é importante destacar que o modelo que sobressaiu foi o AutoGluon, atingindo um valor de 0.6333. Esse resultado enfatiza a habilidade do AutoGluon em alcançar um equilíbrio eficaz entre precisão e recall, resultando em um desempenho sólido na tarefa de classificação.

Tabela 10 – Métricas para base de dados de 24 variáveis

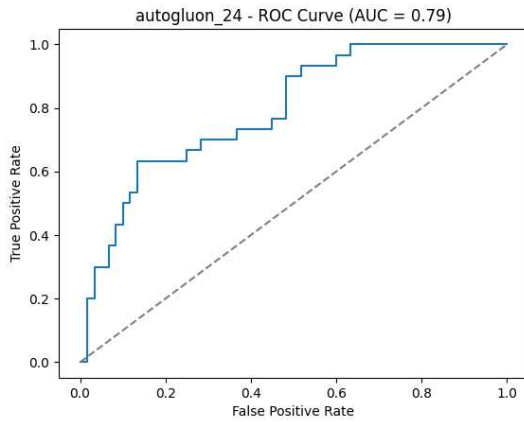
Métrica	AutoGluon	H2O	TPOT	Fusão
Acurácia	0.7556	0.7444	0.7000	0.7556
Precisão	0.6333	0.6522	0.5455	0.6538
Recall	0.6333	0.5000	0.6000	0.5667
F1	0.6333	0.5660	0.5714	0.6071

Fonte: Elaborada pelo autor (2023).

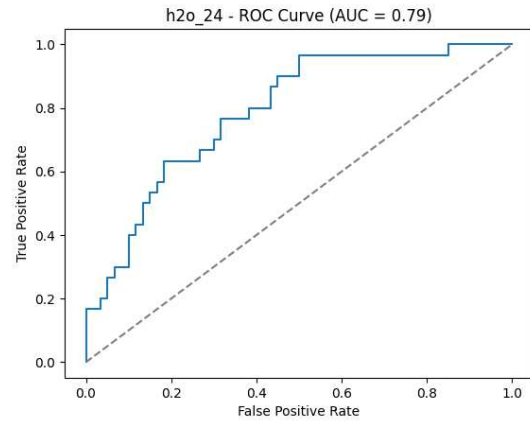
Apesar de a diferença entre os valores ser discretamente pequena, o modelo de Fusão sobressai ao apresentar o resultado mais elevado na métrica Área sob a Curva ROC, evidenciando sua habilidade superior em diferenciar entre as classes positiva e negativa. Além disso, ao analisar a representação gráfica na imagem 35, fica claro que a discrepância entre os desempenhos dos modelos é bastante sutil.

Figura 35 - Métrica ROC Curve para 24 variáveis.

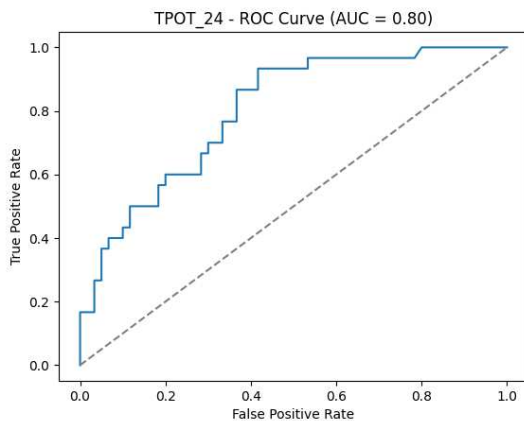
(a) AutoGluon



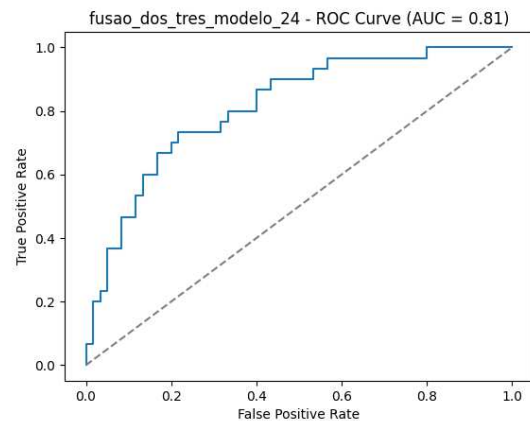
(b) H2O



(c) TPOT



(d) Fusão



Fonte: Elaborada pelo autor (2023).

6.3.3 Vantagens e Limitações

A presença de limitações, como a baixa qualidade dos dados decorrente de falhas nos sensores, apresenta-se como um desafio substancial. Além disso, a dispersão dos dados em múltiplas fontes também complica a situação. No entanto, é crucial ressaltar que uma análise de dados rigorosa e a adoção de técnicas eficientes de geração de dados sintéticos podem oferecer um apoio significativo em determinados cenários. É inegável que a condição ideal seria a utilização de dados totalmente oriundos dos sensores e das coletas realizadas pelos pecuaristas.

A busca incessante por aprimorar os modelos de classificação levou ao desenvolvimento de várias abordagens, cada uma delas com suas próprias vantagens e desvantagens. Nesse contexto, foram analisados três modelos - H2O, AutoGluon e TPOT - no sentido de alcançar um nível satisfatório de classificação. Além disso, foi considerada a fusão desses

modelos.

Os modelos H2O e Autoglun utilizam uma técnica conhecida como *random grid* para a seleção de hiperparâmetros. Esse método oferece a vantagem de uma execução mais rápida, o que é especialmente benéfico quando se trata de eficiência computacional. Ademais, essa abordagem pode levar a um bom desempenho inicial, sendo suficiente para atingir métricas de qualidade, mesmo quando os dados são limitados em quantidade. A simplicidade é outra característica positiva dessa abordagem, pois ela exige menos ajustes manuais e configurações detalhadas (59, 62).

No entanto, há algumas desvantagens a serem consideradas. A escolha aleatória de hiperparâmetros pode restringir a exploração do espaço de possibilidades, limitando o desempenho potencial do modelo. Com o tempo, é possível que esses modelos alcancem um patamar de desempenho subótimo, devido à falta de uma busca sistemática por aprimoramentos.

Por outro lado, o modelo TPOT utiliza a evolução genética para otimizar os parâmetros (66). Isso traz a vantagem de uma exploração mais completa do espaço de hiperparâmetros. Essa abordagem é especialmente útil para descobrir combinações que podem não ser alcançadas através de abordagens aleatórias. Também o modelo TPOT tem o potencial de melhorar continuamente, uma vez que a evolução genética permite iterações sucessivas em busca de um desempenho ainda melhor.

Entretanto, essa abordagem tem algumas desvantagens. Ela requer um tempo de execução maior em comparação com a busca aleatória, o que pode ser um obstáculo em cenários em que a eficiência é crucial. A implementação e a configuração do modelo TPOT com evolução genética podem ser mais complexas e exigir mais ajustes finos.

A fusão dos modelos também foi explorada como uma abordagem. Isso traz a vantagem de combinar diferentes pontos fortes dos modelos individuais e mitigar suas fraquezas. A diversidade de abordagens pode levar a um desempenho mais confiável e estável, especialmente quando um dos modelos não é adequado para determinadas situações.

Porém, a fusão dos modelos também possui desvantagens. Ela adiciona uma camada de complexidade ao processo de classificação, o que pode dificultar a interpretação e a manutenção do sistema como um todo, como também pode exigir mais recursos computacionais devido à coordenação e ao uso de múltiplos modelos.

Em resumo, a escolha entre essas abordagens depende das necessidades e limitações específicas do problema em questão. A busca aleatória é vantajosa em termos de velocidade, enquanto a evolução genética oferece um potencial de desempenho contínuo e adaptativo. A fusão dos modelos pode ser uma estratégia eficaz para equilibrar as vantagens e desvantagens individuais.

7 CONCLUSÃO

Uma série de estudos tem sido conduzida com foco no bem-estar de rebanhos leiteiros. Essas investigações englobam desde abordagens termográficas, análise de dados laboratoriais e sensores de movimento até técnicas de visão computacional. No caso específico deste estudo, a ênfase recai sobre a detecção de mastite, uma problemática essencial, e seu casamento com o aprendizado de máquina.

Uma diferença notável emerge ao considerar a diversidade dos dados em relação a vários estudos anteriores e à abordagem adotada nesta pesquisa. Um aspecto essencial que merece destaque é a composição abrangente dos dados, incluindo fatores como temperatura e umidade, os quais desempenham um papel crucial na influência sobre a detecção da mastite. O objetivo central deste estudo consistiu em elaborar um método capaz de antecipar a ocorrência de mastite, utilizando os dados acessíveis, e esse objetivo foi plenamente alcançado. Além disso, desenvolveu-se um sistema on-line concebido para assumir um papel fundamental na detecção precoce da mastite, oferecendo auxílio aos pecuaristas. Isso, por sua vez, evita possíveis consequências negativas para o bem-estar dos animais e perdas financeiras para a exploração pecuária.

O diferencial deste trabalho reside, ainda, na adoção da abordagem AutoML. Embora já existam pesquisas nesse domínio, a aplicação das tecnologias empregadas neste estudo é escassa. Uma faceta adicional consiste na apresentação de uma comparação e demonstração de três modelos distintos.

Para o desenvolvimento do sistema web, optou-se pelo modelo TPOT. Embora não tenha apresentado a melhor performance nas métricas adotadas, a capacidade inerente de evolução genética do TPOT endossa seu potencial para superar essa situação e emergir como uma solução ainda mais robusta, especialmente em um cenário de evolução do modelo e em produção em larga escala.

Assim sendo, é possível afirmar que as contribuições centrais delineadas por este estudo se condensam em:

- Fomentar a criação de ferramentas que apoiem a tomada de decisões dos produtores rurais;
- Impulsionar a adoção da aprendizagem de máquina automatizada como método para detectar mastite bovina;
- Promover o conhecimento sobre aprendizagem de máquina automatizada;
- Contribuir para o avanço da detecção de mastite em bovinos em ambientes controlados.

Com satisfação, é possível concluir que todos os objetivos específicos delineados foram alcançados com êxito ao longo deste trabalho. A realização de análises exploratórias minuciosas nos conjuntos de dados permitiu uma compreensão aprofundada das informações, fornecendo uma base sólida para o desenvolvimento subsequente do sistema. A concretização da integração perfeita entre a API e o sistema web, juntamente com a incorporação harmoniosa dos modelos de aprendizado de máquina, resultou na criação de uma plataforma ampla e eficaz. Essa plataforma agora oferece aos produtores rurais uma ferramenta poderosa para tomadas de decisões estratégicas, enriquecendo suas práticas agrícolas e visando ao bem-estar dos animais.

Ademais, a avaliação comparativa detalhada dos modelos de AutoML, como o AutoGluon, H2O, TPOT e a fusão dos modelos, foi realizada com amplitude, revelando *insights* valiosos sobre suas performances relativas. Em suma, os resultados obtidos em cada objetivo específico contribuíram de maneira significativa para a concretização bem-sucedida do objetivo geral deste trabalho, representando um avanço promissor no campo da detecção de mastite bovina subclínica e no suporte à tomada de decisões estratégicas no âmbito da produção pecuária.

7.1 TRABALHOS FUTUROS

Uma direção promissora para trabalhos futuros seria concentrar-se na melhoria significativa do sistema web e explorar a viabilidade de criar um aplicativo móvel adicional, o qual traria uma simplificação substancial para os produtores. Concomitantemente, uma sugestão é direcionar esforços para potencializar a capacidade do modelo TPOT, conferindo-lhe a capacidade de evoluir e se adaptar a cada novo conjunto de dados incorporado. A adoção dessa abordagem de treinamento contínuo permitiria ao modelo expandir seu domínio de conhecimento e aprimorar suas métricas, resultando em uma eficácia ainda mais acentuada na detecção precisa da presença ou ausência de mastite nos animais.

Outrossim, seria interessante considerar a possibilidade de incorporar uma opção que permita aos usuários escolher quais variáveis desejam utilizar. Isso ofereceria uma maior flexibilidade em comparação com o modelo atual, que possui variáveis fixas. Essa inclusão permitiria que os produtores personalizassem a análise de acordo com suas demandas específicas, o que poderia potencialmente enriquecer ainda mais a eficácia do sistema. Essa personalização também poderia se estender à escolha das variáveis mais relevantes para a detecção da mastite, promovendo um aprimoramento adicional nos resultados alcançados.

REFERÊNCIAS

- 1 Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.
- 2 R Patiño Pardo, L Altahona, and J Pérez Palencia. O leite como alimento funcional: revisão. *Livestock Research for Rural Development*, 25(139):0121–3784, 2013.
- 3 APAC Gracindo. *Produzindo leite de alta qualidade*. EMPARN, 2009.
- 4 Kennya Beatriz Siqueira. O mercado consumidor de leite e derivados. *Circular Técnica Embrapa*, 120:1–17, 2019.
- 5 Embrapa Gado de Leite (CNPGL). Anuário leite 2023: leite baixo carbono. 2023.
- 6 AG De Assis, Lorildo Aldo STOCK, OF de CAMPOS, AT Gomes, R Zoccal, and MR Silva. Sistemas de produção de leite no brasil. 2005.
- 7 Andre Rozemberg Peixoto Simões, Rosana Moreira da Silva, Marcus Vinicius Moraes de Oliveira, Roberta Oliveira Cristaldo, and Mírian Carla Becker Brito. Avaliação econômica de três diferentes sistemas de produção de leite na região do alto pantanal sul-mato-grossense. *Agrarian*, 2(5):153–167, 2009.
- 8 Willian Maurício Radavelli, Beatriz Danieli, Maria Luisa Appendino Nunes Zotti, Fábio José Gomes, Márcia I Endres, and Ana Luiza Bachmann Schogor. Compost barns in brazilian subtropical region (part 1): facility, barn management and herd characteristics. *Research, Society and Development*, 9(8):e445985198–e445985198, 2020.
- 9 Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59, 2013.
- 10 Robert M Hyde, Peter M Down, Andrew J Bradley, James E Breen, Chris Hudson, Katharine A Leach, and Martin J Green. Automated prediction of mastitis infection patterns in dairy herds using machine learning. *Scientific reports*, 10(1):1–8, 2020.
- 11 Maria Frizzarin, IC Gormley, DP Berry, TB Murphy, A Casa, A Lynch, and Sinead McParland. Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods. *Journal of Dairy Science*, 104(7):7438–7447, 2021.
- 12 Roberta Cristina Bruza Alves, Alessandra Laffranchi, and Cláudio Lima Aguiar. Impacto da variação sazonal na produção e composição de leite produzido por vacas da raça jersey na fazenda experimental, tamarana, paraná. *Journal of Health Sciences*, 8(1), 2006.
- 13 Christel Nielsen. *Economic impact of mastitis in dairy cows*, volume 2009. 2009.
- 14 Naeem Abdul Ghafoor and Beata Sitkowska. Maspas: A machine learning application to predict risk of mastitis in cattle from ams sensor data. *AgriEngineering*, 3(3):575–583, 2021.

- 15 Vania Corrêa Mota, Alessandro Torres Campos, Flávio Alves Damasceno, Everton Augusto de Melo Resende, Creuza Pedroso do Amaral Rezende, Luiz Ronaldo de Abreu, and Teodora Vareiro. Confinamento para bovinos leiteiros: Histórico e características. *Pubvet*, 11:424–537, 2017.
- 16 Ronaldo Braga REIS and D Combs. Atividade leiteira nos estados unidos da américa. *MADALENA, FE; MATOS, LL; HOLANDA JÚNIOR, EV Produção de leite e sociedade. Belo Horizonte: FEPMVZ*, 2001.
- 17 Juan Leandro Monge, Flávio Alves Damasceno, Carlos Eduardo Alves Oliveira, Jairo Alexander Osorio Saraz, and Cristian Chiavassa. Instalaciones para confinamiento de vacas lecheras del tipo compost barn en la argentina y brasil: revisión. In *Conference: II Simpósio Internacional de Ambiência e Engenharia na Produção Animal Sustentável*, 2019.
- 18 Emily Morabito and Jeffrey Bewley. Tie-stall facilities: Design, dimensions, and cow comfort, 2020.
- 19 RC Lewis and JD Johnson. Observations of dairy cow activities in loose-housing. *Journal of Dairy Science*, 37(3):269–275, 1954.
- 20 LUIGI Bertocchi and FRANCESCA Fusi. Guidelines for the assessment of welfare and biosecurity in dairy cattle in loose housing systems. *Centro di Referenza Nazionale per il Benessere Animale (CReNBA): Brescia, Italy*, 2014.
- 21 ME Fernández, RA Mariño, and XC Carreira. Relationship between layout and timber structures in freestall dairy cattle barns: influence of internal features. *Biosystems Engineering*, 100(2):266–280, 2008.
- 22 Daiane Cecchin, Alessandro T Campos, Maria de FA Pires, Renato R de Lima, Tadayuki Yanagi Junior, and Myriam Souza. Avaliação de diferentes materiais para recobrimento de camas em baias de galpão modelo free-stall. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 18:109–115, 2014.
- 23 Rafaella Resende Andrade. Ambiência e bem-estar animal na produção intensiva de leite em sistemas compost barn fechados para a tipologia construtiva e clima do brasil. 2021.
- 24 EM Shane, MI Endres, and KA Janni. Alternative bedding materials for compost bedded pack barns in minnesota: a descriptive study. *Applied engineering in agriculture*, 26(3):465–473, 2010.
- 25 Kevin A Janni, Marcia I Endres, Jeffrey K Reneau, and Wayne W Schoper. Compost dairy barn layout and management recommendations. *Applied engineering in agriculture*, 23(1):97–102, 2007.
- 26 Andressa MICHELS, Valmir Roque Sott, Luana Maria Pigosso, and Silvana Fátima Lunkes. Custo na produção leiteira: um comparativo entre o sistema tradicional e o sistema compost barn em uma propriedade do oeste de santa catarina. In *Anais do Congresso Brasileiro de Custos-ABC*, 2019.

- 27 Gustavo Rafael de Oliveira Silva, Marcos Aurélio Lopes, André Luís Ribeiro Lima, Geraldo Márcio Da Costa, Flávio Alves Damasceno, Vitor Pires Barros, Matteo Barbari, et al. Profitability analysis of compost barn and free stall milk-production systems: a comparison. *Semina. Ciências Agrárias*, 40:1165–1183, 2019.
- 28 Flávio Alves Damasceno. Compost bedded pack barns system and computational simulation of airflow through naturally ventilated reduced model. 2012.
- 29 Embrapa Gado de Leite (CNPGL). Anuário leite 2020: leite de vacas felizes. 2020.
- 30 Embrapa Gado de Leite (CNPGL). Anuário leite 2021: saúde única e total. 2021.
- 31 Lilian Gregory, Eduardo Harry Birgel, Martina Hoedemaker, and Eberhard Grunert. Mastite dos bovinos: histórico de suas formas clínicas. *Revista de Educação Continuada em Medicina Veterinária e Zootecnia do CRMV-SP*, 4(3):31–38, 2001.
- 32 Sarne De Vlieghe, LK Fox, Sofie Piepers, S McDougall, and HW Barkema. Invited review: Mastitis in dairy heifers: Nature of the disease, potential impact, prevention, and control. *Journal of dairy science*, 95(3):1025–1040, 2012.
- 33 Vitória Pereira Massote, Bruna Mariana Zanateli, Geovana Vilela Alves, Elaine Santana Gonçalves, and Elizângela Guedes. Diagnóstico e controle de mastite bovina: uma revisão de literatura. *Revista Agroveterinária do Sul de Minas-ISSN: 2674-9661*, 1(1):41–54, 2019.
- 34 Jeffrey L Watts. Etiological agents of bovine mastitis. *Veterinary microbiology*, 16(1):41–66, 1988.
- 35 Andrew J Bradley. Bovine mastitis: an evolving disease. *The veterinary journal*, 164(2):116–128, 2002.
- 36 W Nelson Philpot. Economics of mastitis control. *The Veterinary Clinics of North America. Large Animal Practice*, 6(2):233–245, 1984.
- 37 Henri Seegers, Christine Fourichon, and François Beaudeau. Production effects related to mastitis and mastitis economics in dairy cattle herds. *Veterinary research*, 34(5):475–491, 2003.
- 38 RJ Harmon. Physiology of mastitis and factors affecting somatic cell counts. *Journal of dairy science*, 77(7):2103–2112, 1994.
- 39 Maria Eduarda Barbosa da Fonseca, Anderson Moreira Mourão, Jônathan David Ribas Chagas, Letícia Meirelles Ávila, Thiago Luiz Pereira Marques, Bruna de Azevedo Baêta, Renata Fernandes Ferreira de Moraes, and Erica Cristina Rocha Roier. Mastite bovina: Revisão. *Pubvet*, 15:162, 2020.
- 40 T. Halasa, K. Huijps, O. Østerås, and H. Hogeveen. Economic effects of bovine mastitis and mastitis management: A review. *Veterinary Quarterly*, 29(1):18–31, 2007. PMID: 17471788.
- 41 José Renaldi Feitosa Brito and Ronaldo de Oliveira Sales. Saúde do úbere. uma revisão. *Revista Brasileira de Higiene e Sanidade Animal*, 1(1):67–90, 2007.

- 42 Fernando Silva Ramos, Lysandra Martineli Fonseca, Joerbertson Caixeta de Sousa, Sabrina Silva Almeida, and Leandra Aparecida Leite Borges. Importância do diagnóstico de mastite subclínica e seus impactos econômicos nas propriedades leiteiras-revisão da literatura. *Revista Coleta Científica*, 1(1):17–27, 2017.
- 43 Franciele Maboni, Marcia Regina Skorei, Ricardo Xavier da Rocha, Agueda Castagna de Vargas, and Marcelo Cecim. Mastite gangrenosa bovina por *Clostridium perfringens*. *Veterinária e Zootecnia*, 15(3):456–459, 2008.
- 44 Débora Brito Goulart and Melha Mellata. *Escherichia coli* mastitis in dairy cattle: etiology, diagnosis, and treatment challenges. *Frontiers in Microbiology*, 13:928346, 2022.
- 45 Tripti Kumari, Champak Bhakat, and Rajeev Kumar Choudhary. A review on subclinical mastitis in dairy cattle. *Int. J. Pure Appl. Biosci*, 6(2):1291–1299, 2018.
- 46 Rafael Izbicki and Tiago Mendonça dos Santos. *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki, 2020.
- 47 CA Becker, A Aghalari, M Marufuzzaman, and AE Stone. Predicting dairy cattle heat stress using machine learning techniques. *Journal of dairy science*, 104(1):501–524, 2021.
- 48 Wei Xu, Ariette TM van Knegsel, Jacques JM Vervoort, Rupert M Bruckmaier, Renny J van Hoeij, Bas Kemp, and Edoardo Saccenti. Prediction of metabolic status of dairy cows in early lactation with on-farm cow data and machine learning algorithms. *Journal of dairy science*, 102(11):10186–10201, 2019.
- 49 S Shahinfar, Majid Khansefid, Mekonnen Haile-Mariam, and JE Pryce. Machine learning approaches for the prediction of lameness in dairy cows. *Animal*, 15(11):100391, 2021.
- 50 MR Borchers, YM Chang, KL Proudfoot, BA Wadsworth, AE Stone, and JM Bewley. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *Journal of dairy science*, 100(7):5664–5674, 2017.
- 51 Tania Bobbo, Stefano Biffani, Cristian Taccioli, Mauro Penasa, and Martino Cassandro. Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. *Scientific Reports*, 11(1):13642, 2021.
- 52 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- 53 Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- 54 Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.
- 55 Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

- 56 Devendra Singh, Pawan Kumar Pant, Himanshu Pant, and Dinesh C Dobhal. Robust automated machine learning (automl) system for early stage hepatic disease detection. In *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020*, pages 65–76. Springer, 2021.
- 57 Haifeng Jin, Qingquan Song, and Xia Hu. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1946–1956, 2019.
- 58 Jason Liang, Elliot Meyerson, Babak Hodjat, Dan Fink, Karl Mutch, and Risto Miikkulainen. Evolutionary neural automl for deep learning. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 401–409, 2019.
- 59 Erin LeDell and Sebastien Poirier. H2O AutoML: Scalable automatic machine learning. *7th ICML Workshop on Automated Machine Learning (AutoML)*, July 2020.
- 60 Trang T Le, Weixuan Fu, and Jason H Moore. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, 36(1):250–256, 2020.
- 61 Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. 2020.
- 62 Wenwen Qi, Chong Xu, and Xiwei Xu. Autogluon: A revolutionary framework for landslide hazard analysis. *Natural Hazards Research*, 1(3):103–108, 2021.
- 63 Arno Candell, Viraj Parmar, Erin LeDell, and Anisha Arora. Deep learning with h2o. *H2O. ai Inc*, pages 1–21, 2016.
- 64 Joao Alexandre Lobo Marques, Francisco Nauber Bernardo Gois, José Xavier-Neto, and Simon James Fong. *Predictive models for decision support in the COVID-19 crisis*. Springer, 2021.
- 65 Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner Gardner, Marc Parizeau, and Christian Gagné. Deap: Evolutionary algorithms made easy. *The Journal of Machine Learning Research*, 13(1):2171–2175, 2012.
- 66 Randal S Olson and Jason H Moore. Tpot: A tree-based pipeline optimization tool for automating machine learning. In *Workshop on automatic machine learning*, pages 66–74. PMLR, 2016.
- 67 Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- 68 Siti Fairuz Mat Radzi, Muhammad Khalis Abdul Karim, M Iqbal Saripan, Mohd Amiruddin Abd Rahman, Iza Nurzawani Che Isa, and Mohammad Johari Ibahim. Hyperparameter tuning and pipeline optimization via grid search method and tree-based automl in breast cancer prediction. *Journal of personalized medicine*, 11(10):978, 2021.
- 69 Suliman Mohamed Fati, Amgad Muneer, Nur Arifin Akbar, and Shakirah Mohd Taib. A continuous cuffless blood pressure estimation using tree-based pipeline optimization tool. *Symmetry*, 13(4):686, 2021.

- 70 Zolo Kiala, John Odindi, and Onesimo Mutanga. Determining the capability of the tree-based pipeline optimization tool (tpot) in mapping parthenium weed using multi-date sentinel-2 image data. *Remote Sensing*, 14(7):1687, 2022.
- 71 Matthias Biehl. *API Architecture*, volume 2. API-University Press, 2015.
- 72 Inobram. <https://www.inobram.com.br/sondas-smaai>. Acesso em: 30 Julho 2023.
- 73 T Desyani, A Saifudin, and Y Yulianti. Feature selection based on naive bayes for caesarean section prediction. In *IOP Conference Series: Materials Science and Engineering*, volume 879, page 012091. IOP Publishing, 2020.
- 74 Barbara G Tabachnick and Linda S Fidell. *Experimental designs using ANOVA*, volume 724. Thomson/Brooks/Cole Belmont, CA, 2007.
- 75 Michael W Browne. Cross-validation methods. *Journal of mathematical psychology*, 44(1):108–132, 2000.
- 76 Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.
- 77 Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.
- 78 James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.