

**UNIVERSIDADE ESTADUAL PAULISTA – UNESP
CÂMPUS DE JABOTICABAL**

**ESTRATÉGIAS PARA SELEÇÃO DE VARIÁVEIS EM
DIFERENTES MODELOS PREDITIVOS PARA
TUBERCULOSE ANIMAL**

Leyde Emanuelle Costa Pereira

Médica Veterinária

2022

**UNIVERSIDADE ESTADUAL PAULISTA – UNESP
CÂMPUS DE JABOTICABAL**

**Estratégias para seleção de variáveis em diferentes
modelos preditivos para tuberculose animal**

Leyde Emanuelle Costa Pereira

**Orientador: Prof. Dr. Luís Antônio Mathias
Coorientador: Prof. Dr. Ricardo Alexandre Arcêncio**

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutora em Medicina Veterinária (Medicina Veterinária Preventiva)

2022

P436e Pereira, Leyde Emanuelle Costa
Estratégias para seleção de variáveis em diferentes modelos
preditivos para tuberculose animal / Leyde Emanuelle Costa
Pereira. -- Jaboticabal, 2022
64 p. : tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp),
Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Luís Antônio Mathias
Coorientador: Ricardo Alexandre Arcêncio

1. Bovino. 2. Teoria da previsão. 3. Banco de dados. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da
Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo
autor(a).

Essa ficha não pode ser modificada.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: ESTRATÉGIAS PARA SELEÇÃO DE VARIÁVEIS EM DIFERENTES MODELOS
PREDITIVOS PARA TUBERCULOSE ANIMAL

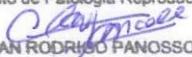
AUTORA: LEYDE EMANUELLE COSTA PEREIRA

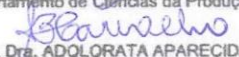
ORIENTADOR: LUÍS ANTONIO MATHIAS

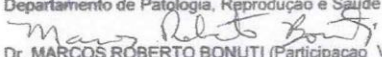
COORDENADOR: RICARDO ALEXANDRE ARCÊNCIO

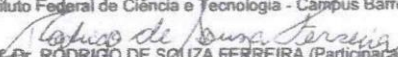
Aprovada como parte das exigências para obtenção do Título de Doutora em MEDICINA
VETERINÁRIA, área: Medicina Veterinária Preventiva pela Comissão Examinadora:


Prof. Dr. LUÍS ANTONIO MATHIAS (Participação Virtual)
Departamento de Patologia Reprodutiva e Saúde Única / FCAV UNESP - Jaboticabal


Prof. Dr. ALAN RODRIGO PANOSSO (Participação Virtual)
Departamento de Ciências da Produção Agrícola / FCAV / UNESP - Jaboticabal


Profa. Dra. ADOLORATA APARECIDA BIANCO CARVALHO (Participação Virtual)
Departamento de Patologia, Reprodução e Saúde Única / FCAV / UNESP - Jaboticabal


Dr. MARCOS ROBERTO BONUTI (Participação Virtual)
Instituto Federal de Ciência e Tecnologia - Câmpus Barretos / Barretos/SP


Prof. Dr. RODRIGO DE SOUZA FERREIRA (Participação Virtual)
Coordenadoria de Defesa Agropecuária-Escritório de Defesa Agropecuária / São João da Boa Vista/SP

Jaboticabal, 06 de maio de 2022

DADOS CURRICULARES DA AUTORA

LEYDE EMANUELLE COSTA PEREIRA – nasceu em Fernandópolis, São Paulo, no dia 06 de julho de 1990. Em dezembro de 2008 concluiu o ensino médio no Centro Educacional Profissionalizante (CEPRO), Ouroeste, SP. Em 2011 ingressou no curso de Medicina Veterinária da Universidade Camilo Castelo Branco, Câmpus Fernandópolis, São Paulo. Em dezembro de 2015 defendeu o assunto de Interesse, o trabalho de conclusão de curso intitulado: “Utilização de óleo de girassol ozonizado no tratamento de dermatofitose em cães”, sob orientação da Profa. Dra. Dora Inês Kozusny-Andreani. Em janeiro de 2016 recebeu o grau de médica veterinária. Em março de 2016 iniciou o mestrado pelo Programa de pós-graduação em Medicina Veterinária (Medicina Veterinária Preventiva) da Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista “Júlio de Mesquita Filho” – Câmpus de Jaboticabal, SP, sob a orientação do Prof. Dr. Antonio Sergio Ferraudó, e defendeu sua dissertação em junho de 2018. Em agosto de 2018 iniciou o curso de Doutorado em Medicina Veterinária pela mesma instituição sob a orientação do Prof. Dr. Luís Antônio Mathias.

“Que a Vossa bondade permita aos Espíritos consoladores derramarem por
toda parte a paz, a esperança e a fé.”

(Prece de Cáritas)

DEDICO:

A Deus,

Que me deu o dom da vida e seu infinito amor. Ele está presente em cada dia, protegendo-me e tornando-me capaz de alcançar meus objetivos. O que seria de mim sem a fé que tenho Nele?

Aos meus pais, Clair e Katia,

Com o apoio e o carinho de vocês consegui cumprir mais esta etapa de minha vida. Um amor sem medidas. Obrigada por vocês me fazerem uma pessoa melhor a cada dia e por estarem sempre ao meu lado. Amo vocês!

À minha irmã, Ana Claudia,

Que sempre me apoiou, orientou e ajudou a superar dificuldades ao longo deste caminho, tornando-se exemplo.

Aos meus avós paternos e maternos

Pelo exemplo, incentivo e pelas orações em meu favor. Obrigada pelo carinho. Também amo muito vocês.

Aos meus grandes amigos que conquistei ao longo desta jornada.

AGRADECIMENTOS

Ao querido Prof. Dr. Luís Antônio Mathias pela oportunidade da orientação em tempos difíceis do doutorado, aprendizado, amizade, confiança depositada, ensinamentos essenciais para o meu crescimento profissional.

Ao Prof. Dr. Alan Rodrigo Panosso e ao Prof. Dr. Ricardo Alexandre Arcêncio, por todo ensinamento, paciência, orientação e correções.

Ao Prof. Dr. Antonio Sergio Ferraudo, pela amizade, pela atenção, compreensão, preocupação, paciência e dedicação para concretização deste trabalho. Agradeço de coração por tudo o que fez por mim, pela oportunidade oferecida, por ter me encorajado cada vez mais a seguir em frente e pelos ensinamentos valiosos durante boa parte do doutorado.

Aos professores: Dr. Alan Rodrigo Panosso e Dr. Marcos Roberto Bonuti, pela atenção, e disponibilidade em participar da banca de qualificação, que brilhantemente fizeram parte deste trabalho.

Aos amigos que fiz durante minha estadia em Jaboticabal, Mariana, Giovanni, Elka, Eliéder, Fernando, Priscila, Leticía, Rafael... Muito obrigada pela amizade, companheirismo, apoio, carinho.

Aos amigos, colegas, professores e funcionários dos Departamentos de Medicina Veterinária Preventiva e Ciências Exatas da Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista “Júlio de Mesquita Filho”, UNESP, Jaboticabal, SP, pela agradável convivência durante o curso.

À Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista “Júlio de Mesquita Filho”, UNESP, Jaboticabal, SP, pela oportunidade da realização desta pesquisa, bem como o título de Doutora em Medicina Veterinária.

Ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), pela bolsa concedida nos quatro anos de curso, sem a qual não teria sido possível a minha dedicação total ao presente trabalho de doutorado.

A todos que de forma direta ou indireta contribuíram para a realização deste trabalho.

A Deus, pela permissão de cada experiência desta jornada, pela presença até mesmo quando eu não sou capaz de reconhecer, por ser meu guia, meu sentido e o meu porquê para me tornar cada dia uma pessoa melhor.

Que Deus os abençoe sempre!

SUMÁRIO

Página

CAPÍTULO 1 – Considerações gerais	1
Introdução.....	1
1. REVISÃO DE LITERATURA	3
1.1 Tuberculose bovina e suas implicações	3
1.2 Variáveis relacionadas à distribuição da tuberculose bovina.....	5
1.3 Inteligência Artificial e Machine Learning.....	6
1.4 <i>Panorama do uso de Machine Learning na área da saúde animal.....</i>	<i>7</i>
REFERÊNCIAS	9
CAPÍTULO 2 - Estratégias para seleção de variáveis em um modelo logístico	15
INTRODUÇÃO	16
MATERIAL E MÉTODOS.....	17
Descrição do banco de dados	17
Descrição das variáveis	18
Análise dos dados	20
Estratégias de seleção de variáveis explanatórias de modelos logísticos múltiplos.....	20
RESULTADOS	21
Estratégia A	21
Estratégia B	22
Estratégia C	22
DISCUSSÃO	25
CONCLUSÃO	29
REFERÊNCIAS	29
CAPÍTULO 3 – Machine Learning para análises preditivas em tuberculose bovina	34
INTRODUÇÃO	35
MATERIAL E MÉTODOS.....	36
Pré-processamento	38
Aprendizado: técnicas de reamostragem para otimização de hiperparâmetros.....	39
Aprendizado: medidas para avaliação de performance	40
Mineração de modelos para o risco de animais positivos utilizando Boosting.....	41
Coleta de dados	41

	ii
Questionário epidemiológico.....	42
RESULTADOS.....	41
DISCUSSÃO	45
CONCLUSÃO	47
REFERÊNCIAS.....	47

Estratégias para seleção de variáveis em diferentes modelos preditivos para tuberculose animal

RESUMO – O objetivo deste estudo foi usar diferentes metodologias estatísticas para seleção de variáveis para prever a presença da tuberculose bovina em um rebanho, e desta forma encontrar as estratégias que identificassem variáveis associadas com a presença de pelo menos um animal positivo dentro de um rebanho. Este procedimento foi aplicado a um banco de dados resultante de um questionário e de testes de tuberculinização aplicados por médicos veterinários da Coordenadoria de Defesa Agropecuária do Estado de São Paulo (CDA). Com base na combinação de métodos clássicos de regressão, foram utilizadas, no capítulo 2, três estratégias de seleção para que se obtivesse uma redução gradativa do número de variáveis com base no conhecimento da epidemiologia da doença. A partir dos dados originais, algumas variáveis foram mantidas e outras criadas, sendo o rebanho a unidade de observação. O capítulo 3 apresenta as etapas relacionadas à utilização de algoritmos de “machine learning” para análises preditivas em saúde animal. A variável resposta foi o rebanho com pelo menos um animal positivo no teste diagnóstico de tuberculose.

Palavras-chave: bovino, predição, critérios

STRATEGIES FOR SELECTING VARIABLES IN DIFFERENT PREDICTIVE MODELS FOR ANIMAL TUBERCULOSIS

ABSTRACT – The aim of this study was to use different statistical methodologies for selection of variables to predict bovine tuberculosis within a herd, and in this way find the strategies to select variables associated with the presence of a positive animal within a herd. This procedure was applied to a database that resulted from a questionnaire and tuberculin tests applied by veterinarians from the Agricultural Defense Coordination of the State of São Paulo (CDA). Based on the combination of classical regression methods, in Chapter 2, three selection strategies were used in order to obtain a gradual reduction in the number of variables based on the epidemiology of the disease. From the original data, some variables were maintained and others created, with the herd being the unit of observation. Chapter 3 presents the steps related to the use of “machine learning” algorithms for predictive analysis in animal health. The response variable was the positive herd for tuberculosis.

Key-words: bovine, prediction, criteria

CAPÍTULO 1 – Considerações gerais

Introdução

A tuberculose bovina (bTB) é uma doença crônica debilitante causada por *Mycobacterium bovis* do complexo de bactérias *Mycobacterium tuberculosis*. A bactéria é ubíqua e tem distribuição mundial, porém os países em desenvolvimento são os mais afetados (Covisi et al., 1998; Shitaye et al., 2007; O'reilly et al., 1995). Está bem documentado que o gado atua como principal reservatório de *M. bovis*, enquanto humanos e outras espécies como cabras, porcos, búfalos, cães, primatas, texugos, veados, gambás e bisões são hospedeiros suscetíveis. Em bovinos a tuberculose é responsável por uma redução da produtividade (O'Reilly et al., 1995; Thoen et al., 2006; Barlow et al., 1999; Bolske et al., 1995).

Infecções por bTB em humanos são muito difíceis de diferenciar clinicamente de infecções por *M. tuberculosis* (Granja, 2001; Collins, 1997). Pessoas com sistemas imunológicos fracos ou subdesenvolvidos, como pacientes com HIV, jovens e idosos, são particularmente vulneráveis às infecções por micobactérias, incluindo infecções por *M. bovis* (Chaisson e Hopewell, 1980; Grange et al., 1994). Isso, sem dúvida, significa que o tratamento é muitas vezes prolongado, levando ao aumento da morbidade e mortalidade como resultado da resistência aos medicamentos, com enormes implicações financeiras.

Independentemente da ameaça zoonótica e dos custos econômicos significativos associados à doença, estimativas precisas da prevalência de bTB são deficientes em muitos países, incluindo o Brasil, onde os programas nacionais de controle vêm sendo aos poucos implementados. A verdadeira carga da doença permanece desconhecida devido à falta de dados de vigilância de rotina da maioria dos países em desenvolvimento.

O principal teste diagnóstico para tuberculose em bovinos é o teste tuberculínico intradérmico - ou teste cutâneo - que provoca uma reação de hipersensibilidade tardia localizada em animais expostos ao *M. bovis* (De La rua-Domenech et al., 2006).

A reinterpretação em testes de nível severo é uma medida que pode ser tomada pelos veterinários do governo, a seu critério, em cenários de alto risco. Ainda assim, é necessário um método sistemático para identificar os testes direcionados para reinterpretação. O aprendizado de máquina, que tem sido usado para desenvolver modelos sensíveis e específicos para avaliar a probabilidade de um incidente de bTB

(Romero et al., 2021, 2020), oferece um mecanismo que possibilita a seleção de testes mais apropriados.

A análise epidemiológica dos dados deve ser incorporada aos programas de controle e erradicação (Good e Duignan, 2011; White et al., 2013). O objetivo deste estudo foi usar dados em nível de rebanho, segundo o inquérito realizado em 2011 para conhecer a situação epidemiológica da brucelose e tuberculose bovina do Estado de São Paulo, e assim construir modelos estatísticos, classificando as variáveis significativas que pudessem explicar a ocorrência de rebanhos bovinos com animais reagentes no teste de diagnóstico de tuberculose. Isso pode permitir a introdução direcionada de medidas adicionais de controle de doenças para mitigar o risco representado.

1. REVISÃO DE LITERATURA

1.1 Tuberculose bovina e suas implicações

A tuberculose (TB), doença infecciosa causada por membros do complexo *M. tuberculosis* (OMS, 2020), é uma das principais causas de mortalidade por doenças infecciosas em todo o mundo. Como uma doença zoonótica, a tuberculose afeta humanos e várias espécies animais e tem sido reconhecida como um grande risco para a saúde de humanos e animais há mais de um século (Olea-Popelka et al., 2017). Em 2019, 1,4 milhão de pessoas morreram de tuberculose (OMS, 2020). A tuberculose bovina é uma doença crônica de bovinos causada por membros do complexo *M. tuberculosis*, principalmente por *M. bovis*, mas também por *M. caprae* e, em menor grau, por *M. tuberculosis* (OIE, 2019). Além de infectar bovinos e humanos, a mesma micobactéria também infecta outros animais domésticos e populações selvagens (Kanipe e Palmer, 2020), causando doenças gerais, pneumonia, perda de peso e mortes. A tuberculose humana é causada pela espécie *M. tuberculosis*. (Smith et al., 2006). O *M. bovis* pode infectar, além dos bovinos, outros mamíferos domésticos e silvestres, e os seres humanos, causando, então, a tuberculose zoonótica (Ragan, 2002).

Essa enfermidade possui grande importância social e econômica, pois acarreta perdas e restrições à atividade pecuária, além dos prejuízos aos sistemas públicos de saúde (Valente et al., 2011). Presente em diversas regiões do mundo (Ruggiero, 2004), a tuberculose bovina apresenta maior prevalência em rebanhos leiteiros (Figueiredo, 2010), principalmente quando estabulados, e menor prevalência no gado de corte, pois são abatidos precocemente e criados, frequentemente, em manejo extensivo (Abrahão, 1998).

A transmissão do *M. bovis* para os bovinos ocorre com maior frequência por via aerógena (Abrahão, 1998). Em seres humanos, a transmissão ocorre principalmente pela via digestiva, por meio da ingestão de leite *in natura* e derivados não pasteurizados contaminados (Menzies e Neill, 2000).

O *M. tuberculosis* e o *M. bovis* são altamente patogênicos (Une e Mori, 2007) e causam enfermidade clínica e patologicamente indistinguível (Cosivi et al., 1998).

Concomitante a isso, formas pulmonares e extrapulmonares de tuberculose humana de origem animal continuam ser um problema em áreas onde a prevalência da infecção nos bovinos é alta (Acha e Szyfres, 2003). Porém, de acordo com Humblet et al. (2009), a incidência do *M. bovis* permanecerá provavelmente subestimada enquanto a diferenciação entre as espécies de micobactérias não for sistematicamente realizada, contribuindo dessa forma, para que continue sendo amplamente distribuída e negligenciada na maior parte dos países em desenvolvimento.

Nesse contexto, o Programa Nacional de Controle e Erradicação da Brucelose e da Tuberculose Animal (PNCEBT), instituído em 2001, objetivou a redução da prevalência e incidência de tuberculose e de brucelose (Brasil, 2016). Apesar da existência de um programa oficial de controle e erradicação da doença animal e da doença em humanos, ambos os programas não trabalham em ações conjuntas e a tuberculose representa uma importante enfermidade em várias regiões do Brasil (Lilenbaum et al., 2007).

Segundo Cosivi et al. (1998), devido aos impactos do *M. bovis* na saúde animal e saúde pública, programas de vigilância da doença em humanos devem ser considerados uma prioridade, especialmente em áreas onde fatores de risco estão presentes. Pesquisas evidenciaram que a conscientização sobre as possíveis implicações para a saúde pública, devido ao gado infectado, não pode ser banalizada (Ameni e Erkihun, 2007; Munyeme et al., 2010), pois a consciência da população quanto à tuberculose bovina, principalmente pessoas do grupo de risco, é de extrema importância para as políticas públicas de saúde quando se considera o controle e a erradicação.

Alguns aspectos interferem na execução das referidas medidas, dentre eles fazem parte a insuficiência de recursos, escassez de profissionais treinados para atuar na área, falta de decisão política, além da subestimação da importância da tuberculose bovina nos setores de saúde pública e animal pelos órgãos estaduais e nacionais de controle sanitário (Cosivi et al., 1998).

Embora existam limitações ao medir a verdadeira dimensão do problema, é necessária a realização de estudos para compreender os casos de tuberculose em

animais, e assim, conhecer a epidemiologia da tuberculose bovina no Brasil e promover melhorias na qualidade da saúde animal e consequentemente humana.

1.2 Variáveis relacionadas à distribuição da tuberculose bovina

Fatores associados à ocorrência da tuberculose bovina (TB) foram avaliados em muitos estudos no Brasil, por meio de levantamentos epidemiológicos realizados em diferentes estados, que demonstram diferentes variáveis de risco associados à doença (Bahense et al., 2016; Dias et al., 2016; Barbieri et al., 2016; Galvis et al., 2016; Guedes et al., 2016; Lima et al., 2016; Néspoli et al., 2016; Queiroz et al., 2016; Ribeiro et al., 2016; Rocha et al., 2016; Silva et al., 2016; Veloso et al., 2016; Vendrame et al., 2016).

Em nível de rebanho, a infecção por *M. bovis* é mais provável quanto maior a proximidade entre os animais; por isso, a introdução e a manutenção da tuberculose bovina em um rebanho são fortemente influenciadas pelas características da unidade de criação (Goodchild e Clifton-hadley, 2001; Barbieri et al., 2016; Vendrame et al., 2016). Quando se menciona tamanho de rebanho, está se referindo à probabilidade de contato efetivo entre o animal infectado e um susceptível (Cook et al., 1996; Cleaveland et al., 2007). A principal via de transmissão da tuberculose bovina nestes casos é a aerógena, pois o contato entre animais facilita a transmissão via aerossóis (Menzies e Neill, 2000). Com relação ao tipo de exploração como fator de risco, grande parte das pesquisas mostra que os rebanhos de produção leiteira apresentam prevalências mais elevadas de tuberculose comparadas aos rebanhos de corte (Cosivi et al., 1998; Humblet et al., 2009; Belchior et al., 2016; Silva et al., 2016; Bahense et al., 2016; Vendrame et al., 2016; Néspoli et al., 2016; Veloso et al., 2016; Rocha et al., 2016). De acordo com Salazar (2005), o risco de infecção em bovinos de corte é mínimo quando mantidos em baixa densidade populacional e em pastagens, mas quando mantidos em confinamento, o risco de transmissão é semelhante ao dos rebanhos leiteiros. Outro fator de risco considerado para a tuberculose está relacionado a reservatórios na fauna silvestre e os possíveis contatos que podem manter com as explorações pecuárias, pois favorecem um ciclo de transmissão do agente entre bovinos-espécies silvestres-bovinos (Michel et al., 2010).

No Brasil, estudos de prevalência e fatores de risco conduzidos em Minas Gerais (Belchior et al., 2016), Paraná (Silva et al., 2016), Bahia (Bahense et al., 2016), Mato Grosso (Néspoli et al., 2016), Rondônia (Vendrame et al., 2016), Goiás (Rocha et al., 2016) e Santa Catarina (Veloso et al., 2016) identificaram como principal fator de risco a exploração leiteira, rebanhos maiores e com maior grau de tecnificação. Segundo Belchior et al. (2016), a probabilidade de infecção aumenta nas propriedades de produção mais intensiva e tecnificada, o que pode estar relacionado a sistemas de criação animal em confinamento parcial ou total.

1.3 Inteligência Artificial e *Machine Learning*

Inteligência artificial (IA) é a capacidade de um algoritmo de assimilar informações para realizar tarefas características da inteligência humana, como reconhecer objetos e sons, contextualizar a linguagem, aprender com o ambiente e resolver problemas. Pode ser entendida como o potencial da máquina de tomar a melhor decisão possível dada a quantidade de informação disponível, com habilidade de adaptação a diferentes situações (Kuhn e Johnson, 2013; Chollet e Allaire, 2017). Atualmente a IA tem sido utilizada em vários campos da ciência, principalmente nas áreas da saúde, da indústria e ambiental (Ge et al., 2017; Chiavegatto Filho et al., 2018; Di Minin et al., 2018; Santos et al., 2019b). A aplicação dessas técnicas aos dados de mídia social, por exemplo, permite que o comportamento humano seja investigado em uma escala sem precedentes, tornando a IA uma constante no cotidiano da sociedade contemporânea (Di Minin et al., 2018).

Estudos a respeito da IA iniciaram-se na década de 1950 dentro da área da ciência da computação com o intuito de automatizar tarefas intelectuais, normalmente executadas por seres humanos. Até o final dos anos 80 acreditava-se que a IA, em nível humano, poderia ser alcançada mediante a programação de um grande conjunto de regras explícitas para manipular o conhecimento (Chollet e Allaire, 2017). Esse paradigma ficou conhecido como IA simbólica e, embora capaz de solucionar uma série de problemas lógicos bem definidos, não foi eficiente na solução de problemas mais complexos e confusos como a classificação de imagens, o reconhecimento de linguagem natural e tradução. Isso porque a IA simbólica não consegue descobrir regras explícitas para a solução de um problema, uma vez que as regras devem ser previamente programadas no sistema, e os dados são processados de acordo com essas regras (Kuhn e Johnson, 2013; Bruce e Bruce, 2017; Chollet e Allaire, 2017).

Em meados dos anos 80 e início dos anos 90 surgiu uma nova abordagem, o aprendizado de máquina estatístico, ou simplesmente, aprendizado de máquina (*Machine Learning*). Nesse novo paradigma os dados a respeito de um problema são fornecidos à máquina, bem como as respostas esperadas, e o aprendizado consiste na geração das regras de decisão, que é a saída do sistema computacional. Em outras palavras, a máquina aprende por conta própria as regras para a tomada de decisão (Bruce e Bruce, 2017). Assim, o aprendizado de máquina se concentra no desenvolvimento de algoritmos eficientes que agrupam grandes quantidades de dados visando otimizar modelos preditivos para, a partir da compreensão da estrutura dos dados, compor as regras de predição. Os métodos de aprendizado de máquina podem ser divididos em quatro categorias diferentes: aprendizado não supervisionado, aprendizado supervisionado, aprendizado semi-supervisionado e aprendizado por reforço, sendo que o aprendizado supervisionado e o aprendizado não supervisionado são os métodos mais amplamente adotados (Ge et al., 2017). No cenário não supervisionado apenas covariáveis preditoras estão disponíveis no conjunto de dados, em problemas supervisionados, além das covariáveis, está disponível também uma resposta de interesse, responsável por guiar a análise (Bruce e Bruce, 2017; Santos et al., 2019a). Em resumo, aprendizado de máquina é um campo de pesquisa estatística para o treinamento de algoritmos computacionais que dividem, classificam e transformam uma base de dados para maximizar a capacidade de classificar, prever, agrupar ou descobrir padrões em um conjunto de dados de destino (Reichstein et al., 2019).

1.4 Panorama do uso de *Machine Learning* na área da saúde animal.

Diversos estudos têm sido desenvolvidos no âmbito de *Machine Learning* aplicado à saúde animal. O crescimento das pesquisas nessa área ocorre conjuntamente ao aumento da demanda por métodos que possam facilitar diagnósticos e otimizar o tempo de profissionais da área da saúde.

Romero et al. (2021) objetivaram comparar modelos resultantes de três algoritmos de *Machine Learning* para conceber estratégias de controle e/ou prevenção de doenças baseadas no risco em nível de rebanho na Inglaterra. O modelo de *Machine Learning* apresentou o melhor desempenho dentre os modelos analisados, com modelos de saída de alto desempenho.

Romero et al. (2020) demonstraram uma nova aplicação da análise de árvore de classificação para melhorar a compreensão da tuberculose (TB) como uma doença

complexa e multifatorial e como uma ferramenta de seleção para modelos paramétricos. Os modelos de árvore criados ajudaram a explicar como os fatores de risco de TB estavam inter-relacionados e caracterizavam grupos de rebanhos de alto risco em relação à probabilidade de um incidente. A natureza da análise permitiu que grandes conjuntos de dados com vários fatores de risco fossem analisados.

No contexto brasileiro, Larios et al. (2021) investigaram o uso de espectroscopia no infravermelho médio e análise multivariada para realizar um diagnóstico rápido, preciso e fácil da leishmaniose visceral canina. Soros sanguíneos de 20 caninos não infectados, 20 por *Leishmania infantum*, e oito cães infectados por *Trypanosoma evansi* foram estudados. Os dados demonstram que a análise de componentes principais com algoritmos de aprendizado de máquina alcançou uma precisão geral acima de 85% no diagnóstico.

REFERÊNCIAS

- Abrahão, RMCM (1998) **Tuberculose humana causada pelo Mycobacterium bovis: considerações gerais e a importância dos reservatórios animais**. 318 f. Dissertação (Mestrado) - Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo.
- Acha PN, Szyfres B (2003) Zoonosis y enfermedades transmissibles comunes al hombre y los animales. clamidiosis, rickettsiosis y virosis. **Organización Panamericana de la Salud. Washington: OPAS** 580:28-56.
- Ameni G, Erkihun A (2007) Bovine tuberculosis on small-scale dairy farms in Adama Town, central Ethiopia, and farmer awareness of the disease. **Revue Scientifique Et Technique** 3:711–719.
- Bahiense L, Ávila LN de, Bavia ME, Amaku M, Dias RA, Grisi-Filho JHH, Ferreira F, Telles EO, Gonçalves VSP, Heinemann MB, Ferreira Neto JS (2016) Prevalence and risk factors for bovine tuberculosis in the State of Bahia, Brazil. **Semina: Ciências Agrárias, Londrina** 5:3549-3560.
- Barbieri JM, Oliveira LF, Dorneles EMS, Mota ALAA, Gonçalves VSP, Maluf PP, Ferreira Neto JS, Ferreira F, Dias RA, Telles EO, Grisi-Filho JHH, Heinemann MB, Amaku M, Lage AP (2016) Epidemiological status of bovine tuberculosis in the state of Minas Gerais, Brazil, 2013. **Semina: Ciências Agrárias, Londrina** 5:3531-3548.
- Barlow D, Mitchell K, Visram K (1999) Tuberculose bovina em lllama (Lama glama) no Reino Unido. **Veterinário Rec** 22:639–640.
- Belchior APC, Lopes LB, Gonçalves VSP, Leite RC (2016) Prevalence and risk factors for bovine tuberculosis in Minas Gerais State, Brazil. **Tropical Animal Health and Production, Edinburgh** 2:373-378.
- Bolske G, Englund L, Wahlstrom H, Lisle GW, Collins DM, Croston P.S (1995) Tuberculose bovina em fazendas de veados suecos: investigações epidemiológicas e rastreamento usando análise de fragmentos de restrição. **Veterinário Rec** 12:414–417.
- Brasil (2016) **Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa nº 19, de 10 de outubro de 2016. Regulamento técnico do programa nacional de controle e erradicação da brucelose e tuberculose animal**. Diário Oficial da União, Brasília, 3 de nov. Seção 1, 8 p.
- Bruce P, Bruce A (2017) **Practical Statistics for Data Scientists**. United States of America: O'Reilly Media, Inc 562 p.
- Chaisson R, Hopewell P (1980) Mycobacteria e mortalidade por AIDS. **Am Revista respira Dis** 139 :1–3.

Chiavegatto Filho ADP, Dos Santos HG, Do Nascimento CF, Massa K, Kawachi I (2018) Overachieving Municipalities in Public Health: A Machine-learning Approach. **Epidemiology** 6: 836-840.

Chollet F, Allaire JJ (2017) **Deep Learning with R** 341 p.

Cleaveland S, Shaw DJ, Mfinanga SG, Shirima G, Kazwala RR, Eblate E, Sharp M (2007) Mycobacterium bovis in rural Tanzania: Risk factors for infection in human and cattle populations. **Tuberculosis, Edinburgh** 87:30-43.

Collins CH, Grange JM, Yates MC (1997) Bacteriologia, Organização e Prática da Tuberculose. Oxford: Butterworth-Heinemann.

Cook AJ, Tuchili LM, Buve A, Foster SD, Godfrey-Fausett P, Pandey GS, Mcadam KP (1996) Human and bovine tuberculosis in the monze district of zambia-a cross-sectional study. **British Veterinary Journal, London** 1: 37-46.

Cosivi JM., et al (1998) Zoonotic Tuberculosis due to Mycobacterium bovis in Developing Countries. *Emerging Infectious Diseases* 1.

De La Rua-Domenech R, AT, Goodchild HM, Vordermeier RG, Hewinson KH, Christiansen RS, Clifton-Hadley (2006) Ante mortem diagnosis of tuberculosis in cattle: A review of the tuberculin tests, γ -interferon assay and other ancillary diagnostic techniques. **Res. Vet. Sci** 81:190–210. <https://doi.org/10.1016/j.rvsc.2005.11.005>.

Di Minin E, Fink C, Tenkanen H, Hiippala T (2018) Machine learning for tracking illegal wildlife trade on social media. **Nature Ecology & Evolution** 3:406-407.

Dias RA, Gonçalves VSP, Figueiredo VCF, Lôbo JR, Lima ZMB, Paulin LMS, Gunnewiek MFK, Amaku M, Ferreira Neto JS, Ferreira F (2009) Situação epidemiológica da brucelose bovina no Estado de São Paulo. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia, Belo Horizonte** 61 :118-125.

Dias RA, Stanojlovic FMU, Belchior APC, Ferreira RS, Gonçalves RC, Aguiar RSCB, Sousa PR, Santos AMA, Amaku M, Ferreira F, Telles EO, Grisi-Filho JHH, Gonçalves VSP, Heinemann MB, Ferreira Neto JS (2016) Prevalence and risk factors for bovine tuberculosis in the State of São Paulo, Brazil. **Semina: Ciências Agrárias, Londrina** 5:3673-3684.

Figueiredo SM (2010) **Brucelose e Tuberculose Bovina no Estado da Paraíba: Estudo retrospectivo**. 53 f. Dissertação (Mestrado em Medicina Veterinária – Centro de Saúde e Tecnologia Rural) - Universidade Federal de Campina Grande, Paraíba.

Galvis JOA, Grisi-Filho JHH, Costa D, Said ALPR, Amaku M, Dias RA, Ferreira F, Gonçalves VSP, Heinemann MB, Telles EO, Ferreira Neto JS (2016) Epidemiologic characterization of bovine tuberculosis in the state of Espírito Santo, Brazil. **Semina: Ciências Agrárias, Londrina**, 5:3567-3578.

Ge Z, Song Z, Ding SX, Huang B (2017) Data Mining and Analytics in the Process Industry: The Role of Machine Learning. **IEEE Access** 20590-20616.

Good M, Duignan A (2011) An evaluation of the Irish single reactor breakdown protocol for 2005 to 2008 and its use as a monitor of tuberculin test performance. **Veterinary Microbiology** 151:85-90. doi:10.1016/j.vetmic.2011.02.029.

Goodchild AVE, Clifton-hadley RS (2001) Cattle-to-cattle transmission of *Mycobacterium bovis*. **Tuberculosis** 1/2: 23-41

Grange J, Daborn C, Cosivi O (1994) Tuberculose relacionada ao HIV devido a *Mycobacterium bovis*: Tubercle and Lung Disease. **Eur Respir** 7:1564–6.

Granja J.M (2001) Infecção por *Mycobacterium bovis* em seres humanos. **Tuberc Edinb Scotl** 1-2: 71-77.

Guedes IB, Bottene IFN, Monteiro LARC, Leal Filho JM, Heinemann MB, Amaku M, Grisi-Filho JHH, Dias RA, Ferreira F, Telles EO, Gonçalves VSP, Ferreira Neto JS (2016) Prevalence and risk factors for bovine tuberculosis in the state of Mato Grosso do Sul, Brazil. **Semina: Ciências Agrárias**, Londrina 5: 3579-3588.

Humblet MF, Boschioli ML, Saegerman C (2009) Classification of worldwide bovine tuberculosis risk factors in cattle: a stratified approach. **Veterinary Research, London** 5:50.

Kanipe C, Palmer MV (2020) *Mycobacterium bovis* e você: uma visão abrangente da bactéria, suas semelhanças com o *Mycobacterium tuberculosis* e sua relação com a doença humana. **Tuberculose** 125 :102006. 10.1016/j.tube.2020.102006

Kuhn M, Johnson K (2013) Applied predictive modeling.

Larios G, Ribeiro M, Arruda C, Samuel L.O, Canassa T, Matthew JB, Marangoni B, Ramos C, Cena C (2021) Uma nova estratégia para o diagnóstico da leishmaniose visceral canina com base na espectroscopia FTIR e aprendizado de máquina. **Journal of Biophotonics**. <https://doi.org/10.1002/jbio.202100141>

Lilenbaum W, Souza GN, Fonseca LS (2007) Management factors associated with bovine tuberculosis on dairy herds in Rio de Janeiro, Brazil. **Revista Brasileira Ciências Veterinárias** 2:98-100.

Lima PB, Nascimento DL, Almeida EC, Pontual KAQ, Amaku M, Dias RA, Ferreira F, Gonçalves VSP, Telles EO, Grisi-Filho JHH, Heinemann MB, SILVA, J. C. R.; Ferreira Neto, J. S (2016) Epidemiological situation of bovine tuberculosis in the state of Pernambuco, Brazil. **Semina: Ciências Agrárias**, Londrina 5:3601-3610.

Menzies FD, Neill SD (2000) Cattle-to-Cattle Transmission of Bovine Tuberculosis. **The Veterinary Journal** 160: 92–106.

Michel AL, Müller B, Van HPD (2010) Mycobacterium bovis at the animal-human interface: a problem, or not? **Veterinary Microbiology**, Barcelona 3-4:371-381.

Munyeme M, Muma JB, Munang'andu HM, Kankya C, Skjerve E, Tryland M (2010) Cattle owners' awareness of bovine tuberculosis in high and low prevalence settings of the wildlife-livestock interface areas in Zambia. **BMC Vet Res** 6: 21.

Néspoli JMB, Negreiros RL, Amaku M, Dias RA, Ferreira F, Telles EO, Heinemann MB, Grisi-Filho JHH, Gonçalves VSP, Ferreira Neto JS (2016) Epidemiological situation of bovine tuberculosis in the state of Mato Grosso, Brazil. **Semina: Ciências Agrárias**, Londrina 5:3589-3600.

O'Reilly LM, Daborn CJ (1995) The epidemiology of mycobacterium bovis infections in animals and man: a review. **Tubercle and Lung Disease**, Edinburgh 76:1-46.

OIE (2019) *Código Sanitário dos Animais Terrestres Capítulo 8.11*. P: Office International des Epizooties.

Olea-Popelka F, Muwonge A, Perera A, Dean AS, Mumford E, Erlacher-Vindel E, Forcella S, Silk BJ, Ditiu L, Ildrissi AE; Raviglione M, Covisi O, Lobue P, Fujiwara IP (2017) Tuberculose zoonótica em seres humanos causada por Mycobacterium bovis - um apelo à ação. **Lancet Infect Dis**. 17 :21-5. 10.1016/S1473-3099(16)30139-6

OMS (2020) *Relatório Global de Tuberculose*. Genebra: Organização Mundial da Saúde.

Queiroz MR, Groff ACM, Silva NS, Grisi-Filho JHH, AMAKU M, Dias RA, Telles EO, Heinemann MB, Ferreira Neto JS, Gonçalves VSP, Ferreira, F (2016) Epidemiological status of bovine tuberculosis in the state of Rio Grande do Sul, Brazil. **Semina: Ciências Agrárias**, Londrina 5:3647-3658.

Ragan V (2002) The Animal and Plant Health Inspection Service: Brucellosis eradication program in the United States. **Veterinary Microbiology** 90:11-18.

Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, PRABHAT (2019) Deep learning and process understanding for data-driven Earth system science. **Nature** 7743:195-204.

Ribeiro LA, Gonçalves VSP, Francisco PFC, Mota ALA, Nascimento GT, Licurgo JB, Ferreira F, Grisi-Filho JHH, Ferreira Neto JS, Amaku M, Dias RA, Telles EO, Heinemann MB, Borges JRJ (2016) Epidemiological status of bovine tuberculosis in the Federal District of Brazil. **Semina: Ciências Agrárias**, Londrina 5:3561-3566.

Rocha WV, Jayme VS, Mota ALAA, Brito WMED, Pires GRC, Grisi-Filho JHH, Dias RA, Amaku M, Telles EO, Heinemann MB, Ferreira F, Ferreira Neto JS, Gonçalves VSP (2016) Prevalence and herd-level risk factors of bovine tuberculosis in the State of Goiás, Brazil. **Semina: Ciências Agrárias**, Londrina 5:3625-3628.

Romero MP, Chang YM, Brunton LA, Parry J, Prosser A, Upton P, Rees E, Tearne O, Arnold M, Stevens K, Drewe J.A (2019) Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making. **Preventive Veterinary Medicine** 175:104860. <https://doi.org/10.1016/j.prevetmed.2019.104860>.

Romero MP, Chang YM, Brunton LA, Prosser A, Upton P, Rees E, Tearne O, Arnold M, Stevens K, Drewe JA (2021) A comparison of the value of two machine learning predictive models to support bovine tuberculosis disease control in England. **Preventive Veterinary Medicine** 188:105264. doi: 10.1016/j.prevetmed.2021.105264.

Ruggiero APM (2004) **Métodos moleculares aplicados ao diagnóstico da tuberculose bovina**. 68 f. Dissertação (Mestrado em Epidemiologia Experimental e Aplicada às Zoonoses) - Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, São Paulo.

Salazar FHP (2005) **Ocorrência de tuberculose causada por Mycobacterium bovis em bovinos abatidos em frigoríficos no estado de Mato Grosso, Brasil**. Dissertação (Mestrado)– Faculdade de Medicina Veterinária e Zootecnia, Universidade Federal de Mato Grosso do Sul, Campo Grande.

Santos GAD, Moitinho MR, Silva BD, Xavier CV, Teixeira DD, Cora JE, La Scala N (2019a) Effects of long-term no-tillage systems with different succession cropping strategies on the variation of soil CO₂ emission. **Science of the Total Environment**, 413-424.

Santos HGD, Nascimento CFD, Izbicki R, Duarte Y.P, Chiavegatto Filho AD (2019b) Machine learning para análises preditivas em saúde: exemplo de aplicação para prever óbito em idosos de São Paulo, Brasil. **Cadernos de Saúde Pública**.

Shitaye JE, Tsegaye W, Pavlik I (2007) Infecção por tuberculose bovina em populações animais e humanas na Etiópia: uma revisão. **Medicina Veterinária** 8:317–332.

Silva MCP, Gonçalves VSP, Mota ALAA, Koloda M, Ferreira Neto JS, Grisi-Filho JHH, Dias RA, Amaku M, Telles EO, Ferreira F, Heinemann MB, Alfieri AA, Muller EE (2016) Prevalence and herd-level risk factors for bovine tuberculosis in the state of Paraná, Brazil. **Semina: Ciências Agrárias**, Londrina 5:3611-3624.

Smith NH, Gordon SV, De La Rua-Domenech R, Clifton-Hadley RS, Hewinson RG (2006) Bottlenecks and broomsticks: the molecular evolution of *Mycobacterium bovis*. **Nature**, 4:670-681.

Thoen C, Lobue P, De Kantor I (2006) A importância do *Mycobacterium bovis* como zoonose. *Microbiologia Veterinária*. (2-4): 339-345.

Une Y, Mori T (2007) Tuberculosis as a zoonosis from a veterinary perspective. **Comparative Immunology, Microbiology & Infectious Diseases**. 30:415-425.

Valente LCM, Vale SMLR, Braga MJ (2011) Determinantes do Uso de Medidas Sanitárias de Controle da Brucelose e Tuberculose Bovinas. **Revista de Economia e Sociologia Rural** 1:215-232.

Veloso FP, Baumgarten KD, Mota ALAA, Ferreira F, Ferreira Neto, JS, Grisi-Filho JHH, Dias RA, Amaku M, Telles EO, Heinemann MB, GONÇALVES VSP (2016) Prevalence and herd-level risk factors of bovine tuberculosis in the State of Santa Catarina, Brazil. **Semina: Ciências Agrárias**, Londrina 5:3659-3672.

Vendrame FB, Amaku M, Ferreira F, Telles EO, Grisi-Filho JHH, Gonçalves VSP, Heinemann MB, Ferreira Neto JS, Dias RA (2016) Epidemiologic characterization of bovine tuberculosis in the State of Rondônia, Brazil. **Semina: Ciências Agrárias**, Londrina 5:3639-3646.

White P, Martin SW, De Jong MCM, O'keeffe JJ, More SJ, Frankena K (2013) The importance of “neighbourhood” in the persistence of bovine tuberculosis in Irish cattle herds. **Preventive Veterinary Medicine** 110:346–355.

CAPÍTULO 2 - Estratégias para seleção de variáveis em um modelo logístico

RESUMO

A regressão logística vem sendo amplamente utilizada em procedimentos de seleção de variáveis. Em grandes bancos de dados a seleção de variáveis se torna imprescindível, pois elege as variáveis com maior peso explicativo dentro do que foi estudado. Isso é facilmente contornado pelo experimentador, a partir de um processo criterioso de seleção de variáveis, concomitante ao conhecimento específico do problema. Em pesquisas de saúde animal a utilização desse método pode acarretar multicolinearidade e podem ser selecionadas variáveis que não condizem com a biologia da doença, além da possibilidade de excluir variáveis condizentes com a causa analisada, comprometendo o resultado da pesquisa. Com base na combinação de métodos clássicos de regressão, foram utilizadas, neste estudo, três estratégias de seleção que possibilitassem uma redução gradativa do número de variáveis com base na epidemiologia da doença. Este procedimento foi aplicado a um banco de dados de uma pesquisa de saúde animal, no caso, rebanhos bovinos com a presença de tuberculose. Os resultados mostraram que notadamente nos três métodos as variáveis que estão correlacionadas a produção leiteira foram as mais significativas em cada estratégia, sendo elas: rebanho maior OR =1,72, 95% IC=1,20-2,45; número de ordenha OR=1,71, 95% IC=1,23-2,38, entrega leite OR =1,71, 95% IC=1,11-2,61, raça bovina OR=1,18, 95% IC=1,03-1,37, em que as três estratégias destacaram a aquisição de animais como uma variável associada à presença da doença. Conclui-se que a estratégia de seleção influi acentuadamente nas variáveis selecionadas, sugerindo a necessidade de muita cautela ao fazer inferências causais a partir de banco de dados com grande número de variáveis explanatórias, a fim de que se possa chegar a modelos com variáveis que de fato influam na ocorrência do desfecho estudado.

Palavras-chave: saúde animal, data mining, tuberculose

INTRODUÇÃO

Uma compreensão abrangente de como a transmissão ocorre é necessária para controlar a doença e projetar e implantar medidas de controle eficazes. No Brasil, a erradicação da tuberculose bovina (bTB) está se mostrando difícil, contrastando com a situação em países europeus (Allen et al., 2018). A doença tem uma epidemiologia notoriamente complexa em todo mundo, envolvendo não apenas bovinos, mas também outros hospedeiros de vida selvagem. Diante da possibilidade de diferenças epidemiológicas entre regiões e ao longo do tempo seria imprudente generalizar ou extrapolar sem dados de vigilância abrangentes.

Supõe-se que a transmissão indireta por meio de ambiente contaminado envolve a geração de núcleos de gotículas com bacilos, a partir de fômites, como solo, pastagens, excretas e ambiente construído, seja por inalação ou ingestão de animais hospedeiros (Palmer et al., 2004; Phillips et al., 2003; Vicente et al., 2007). Uma componente chave do potencial de um ambiente contaminado contribuir para a dinâmica de transmissão da bTB é a sobrevivência dos bacilos da tuberculose nesses fômites ambientais.

Diante de tantas variáveis para a ocorrência da doença, se faz necessário avaliar cada uma para que se tenha um diagnóstico preciso para doença. Atualmente o uso de bancos de dados para avaliar a situação epidemiológica de doenças se tornou uma chave de pesquisa e prevenção. No entanto, estes bancos de dados são compostos por muitas variáveis, e, portanto, é preciso analisar todas com cuidado e selecionar as que mais fazem sentido em relação à epidemiologia da doença.

O objetivo principal da seleção de variáveis é incorporar variáveis epidemiológicas relevantes e estatisticamente significativas no modelo, enquanto exclui ruído e variáveis redundantes (Tolles e Meurer, 2016; Kiezun et al., 2009). Existem muitos procedimentos para esse propósito, como seleção proposital e melhor subconjunto. Esses procedimentos têm se mostrado poderosos na construção de modelos e são amplamente utilizados. No entanto, nenhum desses procedimentos é uma panaceia, especialmente na era dos big data, quando muitas variáveis estão disponíveis.

A seleção proposital de variáveis geralmente envolve testes univariados para rastrear variáveis que estão significativamente associadas ao resultado de interesse. Em seguida, um modelo de regressão é construído com base nessas variáveis. No

entanto, tal procedimento pode ignorar algumas variáveis importantes que funcionam em conjunto para fazerem efeito, mas que não atingem o nível de significância estatística quando testadas de forma independente (Zhang, 2016). Porém esse procedimento pode ser útil em um conjunto de dados com pequeno número de variáveis, mas pode não ser a melhor escolha diante de muitas variáveis.

Frequentemente, os conjuntos de variáveis devem ser reduzidos para obter os modelos estatísticos mais simples com a perda mínima de informação. O problema reside no peso estatístico diferente de cada variável, dependendo de sua associação direta mais forte ou mais fraca com a variável resposta avaliada, mascarando-se mutuamente. Isso faz com que muitas variáveis relevantes sejam eliminadas durante o processo de seleção de variáveis, bem como possíveis interações entre as variáveis que não são endereçadas corretamente (Lozano et al., 2018).

Todas as variáveis preditivas podem afetar, em maior ou menor grau, e de maneiras diferentes, a variável resposta (Peña, 2002). A fim de reduzir tais interferências, é indicado o uso de procedimentos reconhecidos para seleção de variáveis, bem como a seleção adequada das informações mais relevantes, negligenciando aquelas com menos poder explicativo ou tentando alcançar um ou mais modelos equilibrados, sacrificando a capacidade seletiva (Lozano et al., 2018). Atualmente o modelo de regressão logística é o método mais utilizado em pesquisas epidemiológicas, porque na maioria dos casos a variável dependente é categórica.

Para lidar com esse fato, este estudo comparou três estratégias de seleção de variáveis com o objetivo de obter modelos logísticos múltiplos com todas as variáveis significativas que pudessem explicar a ocorrência do desfecho analisado, ou seja, rebanhos bovinos com animais reagentes no teste de diagnóstico de tuberculose.

MATERIAL E MÉTODOS

Descrição do banco de dados

Utilizou-se um banco de dados obtidos por ocasião do segundo inquérito, realizado em 2011, para conhecer a situação epidemiológica da brucelose e da tuberculose bovina no Estado de São Paulo. Os dados resultaram de um questionário e de testes de tuberculina aplicados por médicos veterinários da Coordenadoria de Defesa Agropecuária do Estado de São Paulo (CDA) em rebanhos bovinos ou

bubalinos, selecionados por procedimento probabilístico, elaborado por pesquisadores que assessoram o Ministério da Agricultura, Pecuária e Abastecimento (MAPA) no Programa Nacional de Controle e Erradicação da Brucelose e Tuberculose (PNCEBT). Detalhes da amostragem podem ser obtidos em Dias et al. (2016). Esses dados foram disponibilizados pela CDA. A partir dos dados originais, algumas variáveis foram mantidas e outras foram criadas, ficando o banco de dados com 37 variáveis e 1.743 observações, sendo o rebanho a unidade de observação.

Descrição das variáveis

A variável analisada como desfecho, identificada como tuberculose bovina (TB), representa rebanho com pelo menos um animal da amostra reagente no teste de tuberculina (0=não; 1=sim). As variáveis investigadas como explanatórias estão descritas na Tabela 1.

Reb_ma_19	Rebanho com mais de 19 animais	Não=0 Sim=1
Tipo_exp	Tipo de exploração	1=corte 2=leite 3=confinado
Tipo_cria	Tipo de criação	1=extensivo 2=semiconfinado 3=confinado
N_ordenha	Número de ordenhas por dia	1=não ordenha 2=uma ordenha 3=duas ou três ordenhas
Tipo_ordenha	Tipo de ordenha	1=não ordenha 2=ordenha mecânica em sala de ordenha 3=mecânica ao pé 4>manual
N_vacas	Número de vacas em lactação	Variável quantitativa
Prod_diaria	Produção diária de leite	Variável quantitativa
inseminação	Usa inseminação artificial	1=não 2=usa inseminação e touro 3=usa apenas inseminação
Raca_bov	Raça bovina predominante	1=zebu 2=europeia leite 3=europeia de corte 4=mestiça 5=outras raças
Ovi_cap	Presença de ovinos e caprinos	0=não 1=sim
equi	Presença de equinos	0=não 1=sim
sui	Presença de suínos	0=não 1=sim
aves	Presença de aves comerciais	0=não 1=sim
cao	Presença de cães	0=não

		1=sim
gato	Presença de gatos	0=não 1=sim
nsilvestre	Presença de animais silvestres de vida livre	0=não 1=sim
cervideo	Presença de cervídeo	0=não 1=sim
capivara	Presença de capivaras	0=não 1=sim
felideos	Presença de felídeos de vida livre	0=não 1=sim
marsu	Presença de marsupiais	0=não 1=sim
macaco	Presença de macacos	0=não 1=sim
Outra_sil	Presença de outras espécies	0=não 1=sim
Testa_tb	Faz teste para diagnóstico de tuberculose	0=não 1=sim
Regul_testa_tb	Regularidade dos testes de tuberculose	1=uma vez no ano 2=duas vezes no ano 3= quando compra animais 4=quando há aborto na fazenda 5=quando exigido para trânsito/eventos/crédito
Aquis_bovideos	Aquisição de bovídeos	0=não 1=sim
Aluga_pasto	Aluga pastos em alguma época do ano	0=não 1=sim
Pasto_comum	Tem pasto em comum com outras propriedades	0=não 1=sim
Comp_itens	Compartilha outros itens com outras propriedades	0=não 1=sim
17reas_alagadas	Presença de áreas alagadiças às quais o gado tem acesso	0=não 1=sim
Gado_conc	Área onde o gado permanece concentrado durante o dia ou à noite	0=não 1=sim
Piquete_parto	Tem piquete separado para fêmeas na fase de parto e/ou pós-parto	0=não 1=sim
Entr_leite	Entrega leite	0=não 1=sim
Lact_venda	Produz leite para venda	0=não 1=sim
Assis_vet	Tem assistência veterinária	0=não 1=sim
Alim_soro	Alimenta bovinos com soro de leite bovino	0=não 1=sim
Comp_aguada	Compartilha aguadas/bebedouros com animais de outra(s) propriedade(s)	0=não 1=sim
Classif_prop	Classificação da propriedade	1=rural clássica 2=aldeia indígena 3=assentamento 4=periferia urbana

Análise dos dados

Inicialmente foram obtidas estatísticas descritivas, por meio do software Epi Info 7.2, de modo a conhecer as frequências da variável resposta nas categorias das variáveis explanatórias. Na análise de regressão logística empregou-se a linguagem R (Rcore Team, 2021), e o ajuste do modelo foi avaliado por meio do teste de Hosmer e Lemeshow, o qual foi efetuado utilizando o pacote Resource Selection do R. Os modelos obtidos também foram avaliados com base na área sob a curva ROC, e para isso utilizou-se o pacote pROC do R. Para a investigação de colinearidade entre as variáveis explanatórias utilizou-se, também empregando o R, correlação de Spearman, para que um modelo logístico não mantivesse variáveis com coeficiente de correlação maior do que 0,70.

Estratégias de seleção de variáveis explanatórias de modelos logísticos múltiplos

Para a seleção das variáveis explanatórias em modelos logísticos múltiplos com todas as variáveis significativas, foram utilizadas três estratégias, aqui denominadas A, B e C.

A estratégia A consistiu em adotar inicialmente seleção automática, em ambas as direções, forward e backward, com base no menor AIC (Akaike Information Criterion), a partir de todas as variáveis; o próximo passo foi retirar variáveis altamente correlacionadas; em seguida, foram retiradas do modelo, uma por vez, por ordem inversa do valor de P, as variáveis não significativas, até obter um modelo múltiplo com todas as variáveis significativas ($P < 0,05$).

Na estratégia B, foi inicialmente feita triagem por meio de regressão logística simples, sendo selecionadas variáveis com valor de $P < 0,10$ e cujo sentido da associação apresentou nexos causais; essas variáveis foram submetidas então à seleção automática, como descrito no parágrafo anterior; do modelo obtido foram retiradas as variáveis altamente correlacionadas, sendo em seguida retiradas, uma a uma, as variáveis não significativas.

Na estratégia C, foram selecionadas variáveis com valor de $P < 0,10$ na regressão logística simples, descartando aquelas cuja direção da associação não tinha nexos do ponto de vista biológico. A partir dessas variáveis, foram escolhidas variáveis apontadas na literatura como tendo potencial associação causal com o

desfecho e descartando as variáveis que possam estar associadas a essas outras, evitando colocar no modelo variáveis intervenientes (Dohoo et al., 2010). Após checagem da correlação entre as variáveis até então selecionadas, foram retiradas, uma por vez, por ordem inversa do valor de P, as variáveis não significativas, até obter modelo múltiplo com todas as variáveis significativas ($P < 0,05$).

RESULTADOS

Estratégia A

Seguindo os passos descritos na metodologia da estratégia A, que consistiu em seleção automática, foram obtidas as variáveis rebanho maior que 19, número de ordenhas, tipo de ordenha, número de vacas em lactação, raça bovina, presença de suínos, de animais silvestres, de marsupiais, aquisição de bovídeos, pasto comum, compartilha itens, entrega leite, assistência veterinária e compartilha aguada. Após a checagem da correlação entre as variáveis, por meio de correlação de Spearman, procurando coeficientes acima de 0,70, observou-se correlação elevada entre tipo de ordenha e número de ordenhas, sendo então mantida apenas “número de ordenhas”, por apresentar o menor valor de P (Tabela 1). Na sequência, foram retiradas, uma a uma, assis_vet, pasto_comum e suínos. A Tabela 1 mostra o modelo obtido após a eliminação gradativa das variáveis.

Tabela 1. Variáveis selecionadas pelo modelo A com todas as variáveis significativas ($P < 0,05$).

Variáveis	OR	IC 95%	Coeficiente	Erro Padrão	Estat Z	Valor P
reb_ma_19	1,72	1,20–2,45	0,54196	0,18204	2,977	0,00291
n_ordenha	1,71	1,23 – 2,38	0,53723	0,16782	3,201	0,00137
raca_bov	1,18	1,03 – 1,37	0,16782	0,07435	2,268	0,02332
nsilvestre	1,82	1,17 – 2,82	0,59798	0,22373	2,673	0,00752
marsu	0,66	0,43–1,00	-0,42219	0,21252	-1,987	0,04697
aquis_bovideos	1,50	1,07–2,11	0,40733	0,17256	2,361	0,01825
comp_itens	1,83	1,19 – 2,75	0,60211	0,21300	2,827	0,0470
entre_leite	1,71	1,11–2,61	0,53365	0,21688	2,461	0,01387
comp_agua	0,45	0,25-0,75	-0,80777	0,27525	-2,935	0,00334

Estratégia B

Nesta estratégia, foi inicialmente feita a triagem por meio de regressão logística simples, e o descarte das variáveis explanatórias com valor de $P > 0,10$ (ovi_cap, cao, cervideo, capivara, felideos, marsu, macaco, outra_silv, pasto_comum, areas_alagadas, lact_venda, assis_vet, alim_soro, classif_prop). Também foram desconsideradas as variáveis cujas associações não apresentam nexos do ponto de vista biológico (testa_tb, regul_test_tb, comp_aguada). Após seleção automática, com as variáveis reb_ma_19, tipo_exp, tipo_cria, n_orderha, tipo_orderha, n_vacas, prod_diaria, inseminação, raca_bov, equi, sui, aves, gato, nsilvestre, aquis_bovideos, aluga_pasto, comp_itens, gado_conc, piquete_parto, entr_leite, obteve-se o seguinte modelo: reb_ma_19, n_orderha, tipo_orderha, n_vacas, raca_bov, cui, nsilvestre, aquis_bovideos, comp_itens, entr_leite. Após a checagem da correlação entre as variáveis, por meio do método de Spearman, procurando coeficientes acima de 0,70, observou-se correlação elevada entre as variáveis n_orderha, tipo_orderha, n_vacas, e optou-se por manter apenas “número de ordenhas”, por apresentar o menor valor P (Tabela 2). Na sequência foram retiradas, uma a uma, as variáveis nsilvestre e suíno. A Tabela 2 mostra o modelo obtido após a eliminação gradativa das variáveis.

Tabela 2. Variáveis selecionadas pelo modelo B com todas as variáveis significativas ($P < 0,05$).

Variáveis	OR	IC 95%	Coeficiente	Erro padrão	Estat Z	Valor P
reb_ma-19	1,75	1,23-2,49	0,56012	0,18082	3,098	0,00195
n_orderha	1,71	1,23-2,36	0,53459	0,16565	3,227	0,00125
raca_bovina	1,19	1,04-1,38	0,17771	0,07329	2,425	0,01531
aquis_bovideos	1,49	1,07-2,09	0,40095	0,17026	2,355	0,01853
comp_itens	1,51	1,00-2,24	0,41371	0,20471	2,021	0,04328
entre_leite	1,67	1,10-2,55	0,51572	0,21487	2,400	0,01639

Estratégia C

Foi realizada uma análise preliminar das variáveis importantes segundo a literatura; (tipo de exploração, pois a exploração leiteira está mais associada à ocorrência de tuberculose), excluindo variáveis que possam depender dela e mantendo variáveis com valor

de $P < 0,10$ na triagem e que não estejam associadas ao tipo de exploração. As variáveis selecionadas inicialmente foram reb_ma_19, tipo_exp, inseminacao, equi, sui, aves, gato, nsilvestre, aquis_bovideos, aluga_pasto, comp_itens, gado_conc, piquete_parto. Posteriormente foi retirada, uma por vez, das variáveis não significativas; inseminação, aves, equinos, presença de animais silvestres, aluga pasto, gado concentrado, gato, piquete de parto, compartilha itens. Observa-se na tabela 3 as variáveis significativas ($P < 0,05$) selecionadas no modelo.

Tabela 3. Variáveis selecionadas pelo modelo C com todas as variáveis significativas ($P < 0,05$).

Variáveis	OR	IC 95%	Coeficiente	Erro padrão	Estat Z	Valor P
reb_ma_19	1,99	1,42-2,77	0,6884	0,1699	4,053	5,06 e-05
tipo_exp	1,35	1,10–1,67	0,3028	0,1056	2,866	0,00415
suinos	1,57	1,13–2,17	0,4487	0,1667	2,692	0,00710
aquis_bovideos	1,69	1,22–2,34	0,5234	0,1660	3,153	0,00162

Na Figura 1 podemos observar com mais facilidade as variáveis comuns entre os modelos.

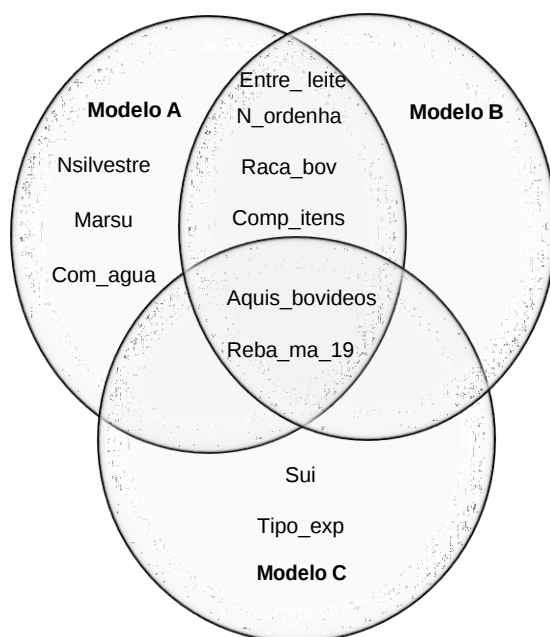


Figure 1. Diagrama de Venn, comparando as variáveis de cada estratégia.

A técnica adotada para resumir o desempenho do modelo em diferentes situações, como no caso as três estratégias, foi a construção da curva Receiver Operating Characteristic(ROC), como no demonstrado na Figura 2 e na tabela 4 em números para melhor especificação dos valores de cada estratégia, em que o modelo obtido na estratégia A apresentou maior área sob a curva.

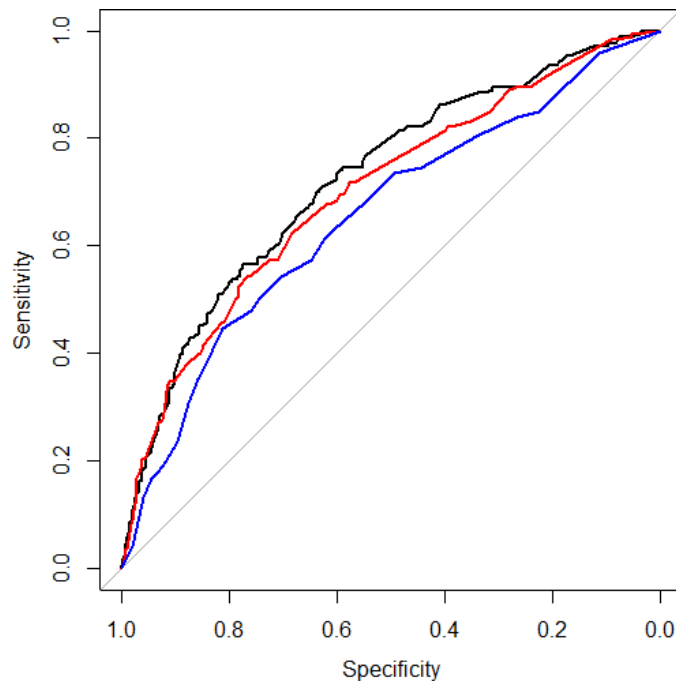


Figura 2. Curva ROC. Preto: modelo A; Vermelho: modelo B; Azul: modelo C

Tabela 4. Área sob a curva ROC (AUC)		
Modelo	AUC	IC 95% AUC
A	0,7246	0,6837—0,7655
B	0,7011	0,6582—0,744
C	0,6552	0,6102—0,7003

DISCUSSÃO

Algumas áreas de pesquisa utilizam conjunto de dados com os desafios da alta dimensionalidade e da multicolinearidade a serem superados, mas de forma prática a seleção de variáveis pode ser uma ferramenta útil. Embora os métodos multivariados existentes sejam, na maioria das vezes, eficientes para construção do modelo, frequentemente se faz necessário selecionar as variáveis mais importantes em explicar o modelo (Ferreira, 2015).

Segundo Zimmer e Anzanello (2014), engenheiros e pesquisadores têm buscado métodos para selecionar o melhor subconjunto de variáveis, visando a diminuição de custos e o aumento da precisão dos resultados. Alguns estudos desenvolveram novas métricas, mas a maior parte é em aprimoramento dos métodos estatísticos já existentes (Lin et al., 2015).

Este estudo avaliou as variáveis e selecionou as mais condizentes com a doença estudada e assim compara os resultados com os de estudos semelhantes. Dessa forma, pode-se observar nas três estratégias que elas apresentam variáveis semelhantes, entre elas rebanho maior que 19 animais e aquisição de bovídeos, que se mantiveram nos três modelos obtidos. Estudos com base no mesmo modelo e a respeito do mesmo assunto em diferentes estados do Brasil evidenciaram as mesmas variáveis, o que está de acordo com o conhecimento que se tem que rebanhos maiores são mais propensos a terem animais infectados. Segundo Dias et al. (2016), agricultores que possuem grandes rebanhos tendem a comprar mais gado, sendo esse fator um dos associados a um risco de introdução da tuberculose bovina no rebanho. Em rebanhos maiores também é mais difícil erradicar uma infecção crônica, porque, quanto maior o número de animais, maior a probabilidade de pelo menos um animal infectado não ser detectado nos testes de diagnóstico e continuar a disseminar a infecção no rebanho. Os três modelos deixaram evidente que a variável relacionada ao tamanho do rebanho está fortemente associada à presença de animais reagentes.

Estratégias como estas que foram utilizadas como ferramenta de seleção de variáveis necessitam de uma análise antes da seleção, para que a variável resposta seja condizente com a relação causal. Em vista disso, revisamos alguns estudos para verificar se as variáveis selecionadas pelos modelos eram coerentes e permitem conclusões mais seguras (Galvis et al., 2016; Bahiense et al., 2016; Silva et al., 2016; Veloso et al., 2016; Dias et al., 2016).

A estratégia C, quando comparada com as estratégias A e B, resultou em modelo com menor número de variáveis, ou seja, apenas quatro variáveis, sendo duas delas também observadas nos dois outros modelos (rebanho maior que 19 animais e aquisição de bovídeos) e duas que não foram mantidas nas outras estratégias (suínos e tipo exploração). A associação entre a tuberculose bovina e o tipo de exploração está amplamente descrita na literatura. Essa associação deve-se ao fato de que a transmissão do agente etiológico da tuberculose bovina exige proximidade e contato constante, o que ocorre mais frequentemente em explorações leiteiras. No entanto, nos modelos das estratégias B e C essa variável acaba sendo mascarada pela presença de outras variáveis relacionadas a ela.

Estudos em outros estados do Brasil e em outros países, nos quais a análise de múltiplas variáveis foi empregada, dão suporte aos achados deste estudo quanto à variável tipo de exploração, no caso leite, corte e misto (Bahiense et al., 2016;

Queiroz et al., 2016). Embora nem todos os estudos realizados em outros estados do Brasil tenham apontado significância dessa variável (Bahense et al., 2016), sua contribuição para a explicação do desfecho tem sido apontada como importante, conforme observado por Marangon et al. (1998), que verificaram que o tipo de exploração em rebanhos da região de Veneto na Itália foi uma variável de risco de ocorrência de tuberculose. Também o estudo de Dias et al. (2016) mostra que rebanhos leiteiros correram 2,70 vezes o risco que correram os rebanhos de corte, usados como referência.

Muitos estudos evidenciam a tuberculose em outras espécies animais. No caso dos suínos, podem não ocorrer tantas perdas econômicas relacionadas a doença clínica, porém vísceras com lesões tuberculoideas são descartadas e carcaças inteiras podem ser condenadas ou requerer tratamento térmico antes de liberadas para o consumo humano (Morés et al., 2007; Hibiya et al. 2008). Visto que a criação de suínos em condições precárias de higiene tende a aumentar a disseminação da doença no plantel, as instalações e os equipamentos podem facilmente ser contaminadas por fezes de animais infectados (Radostits et al., 2007). Lara et al. (2009) afirmam que para o controle e a profilaxia de tuberculose deve-se evitar a coabitação de bovinos com suínos, particularmente em áreas endêmicas de tuberculose bovina.

As três estratégias indicaram que a variável aquisição bovídeos está significativamente associada com a variável resposta. Vários autores também relataram que a aquisição de animais está fortemente associada à persistência da doença em rebanhos (Vendrame et al., 2016; Veloso et al., 2016; De la cruz et al., 2014; Dommergues et al., 2012; Gilbert et al., 2005; Grisi-Filho et al., 2011). As mesmas variáveis foram observadas em outros estados brasileiros, como Paraná, Bahia, Santa Catarina e Espírito Santo (Galvis et al., 2016; Bahense et al., 2016; Silva et al., 2016; Veloso et al., 2016; Dias et al., 2016). Esses estudos constataram que, quanto maior a população bovina, maior é a probabilidade de a infecção persistir no rebanho, no caso, em produção leiteira.

As estratégias A e B frisam as variáveis relacionadas ao gado de leite. Nota-se a significância das variáveis raça bovina, número de ordenhas e entrega de leite, todas relacionadas aos rebanhos leiteiros. Os resultados deste estudo indicam que a tuberculose bovina é mais frequente em rebanhos leiteiros, notadamente naqueles com produção mais intensiva, o que pode estar relacionado aos sistemas de produção sob confinamento parcial ou total. Salientamos que as variáveis encontradas

espelham semelhanças com as descobertas de estudos anteriores sobre as mesmas variáveis relatadas no Paraná (Silva et al., 2016), Bahia (Bahense et al., 2016), Rondônia (Vendrame et al., 2016), Mato Grosso (Néspoli et al., 2016), Santa Catarina (Veloso et al., 2016), Minas Gerais (Belchior et al., 2016), Pernambuco (Lima et al., 2016), Goiás (Rocha et al., 2016) e São Paulo (Dias et al., 2016).

Nos estados de Pernambuco e São Paulo foi observado que a variável compartilhamento de pastagens é significativa como variável de risco associada à infecção (Lima et al., 2016; Dias et al., 2016), porém o mesmo não foi observado para as variáveis compartilhamento de água e pasto. Acredita-se que a interface áreas de exploração de bovinos e contato com animais de vida selvagem possa contribuir para a disseminação de tuberculose bovina. No entanto, a direção predominante de transmissão da interface vida selvagem/pecuária permanece pouco clara e provavelmente bidirecional e determinada por vários fatores, como densidade populacional, prevalência de infecção, distribuição e movimento de infectados e espécies suscetíveis (Bengis et al., 2002; Renwick et al., 2007; Musoke et al., 2015). No caso deste estudo, a estratégia A foi a única a ressaltar essa problemática, com as variáveis marsupiais e presença de outras espécies de animais silvestres relativamente significativas como variáveis potenciais. No entanto, é preciso considerar que essa estratégia baseou-se na seleção automática de variáveis, o que não é recomendado pelos estudiosos do assunto (Dohoo et al., 2010).

Não foi encontrada associação estatisticamente significativa em variáveis reconhecidas por representarem mecanismos de propagação da doença, destacando-se: aluguel de pasto e presença de pastagem em comum com outras propriedades. Tal resultado, entretanto, pode ser em consequência de: (a) baixo poder discriminante dessas variáveis na população estudada; (b) baixo quantitativo de focos detectados. De acordo com Dohoo et al.(2010), o reduzido número de casos positivos pode afetar o poder dos testes estatísticos.

No caso de algumas variáveis do banco de dados, seria útil a disponibilidade de informações adicionais para que se pudesse estabelecer mais seguramente relação causal com o desfecho investigado. Um exemplo dessa situação é a variável relacionada à presença de suínos na propriedade, porém sem informar se esses suínos eram criados soltos ou emconfinamento, ou seja, tendo ou não contato com o rebanho bovino estudado.

Estudos que se assemelham ao que propomos são aqueles que avaliam a

seleção de diferentes variáveis. No entanto, alguns estudos mencionados acima reforçam que todas as variáveis encontradas pelas três estratégias foram compatíveis em termos de associação causal com o desfecho, uma vez que implementaram métodos para avaliar o risco das variáveis em uma única etapa e tendem a usar uma abordagem de regressão logística multivariada para atingir seus objetivos.

CONCLUSÃO

Os resultados mostraram que a estratégia de seleção das variáveis explanatórias influibastante no modelo final obtido, o que sugere cautela ao fazer inferência causal a partir de bancos de dados com muitas variáveis e a necessidade de associar outros critérios de pré-seleção e também a consideração da opinião de especialistas na enfermidade, a fim de obter conclusões mais seguras quanto à relação causal.

REFERÊNCIAS

Allen AR, Skuce RA, Byrne AW (2018) Tuberculose bovina na Grã-Bretanha e Irlanda – uma tempestade perfeita? A confluência de potenciais impedimentos ecológicos e epidemiológicos para o controle de uma doença infecciosa crônica. **Fronteiras na Ciência Veterinária**. 5 :109. doi: 10.3389/fvets.2018.00109.

Bahiense L, Ávila LN De, Bavia ME, Amaku M, Dias RA, Grisi-filho JHH, Ferreira F, Telles EO, Gonçalves VSP, Heinemann MB, Ferreira Neto JS (2016) Prevalence and risk factors for bovine tuberculosis in the State of Bahia, Brazil. **Semina: Ciências Agrárias**, Londrina, v. 37, 5: 3549-3560. Suplemento 2.

Belchior APC, Lopes LB, Gonçalves VSP, Leite RC (2016) Prevalence and risk factors for bovine tuberculosis in Minas Gerais State, Brazil. **Tropical Animal Health and Production**, Edinburg, v. 48, 2:373-378.

Bengis RG, Kock RA, Fischer J, (2002) Infectious animal diseases: the wildlife/livestock interface. **Revue scientifique et technique (International Office of Epizootics)** 21, 53–65. <https://doi.org/10.20506/rst.21.1.1322>.

De La Cruz ML, Perez A, Bezos J, Pages E, Casal C, Carpintero J, Romero B, Dominguez L, Barker CM, Diaz R, Alvarez J (2014) Spatial dynamics of bovine tuberculosis in the Autonomous Community of Madrid, Spain (2010- 2012). **PLoS One**, San Francisco, v. 9,12:115632.

Dias RA, Stanojlovic FMU, Belchior APC, Ferreira RS, Gonçalves SRC, Barão RSC, Sousa PR, Santos AMA, Marcos A, Ferreira F, Telles EO, Grisi Filho JHH, Gonçalves VSP, Heinemann MB, Neto JSF (2016) Prevalência e fatores de risco paratuberculose bovina no estado de São Paulo, Brasil. *Semina: Ciências Agrárias*, Londrina, v. 37,5: 3673-3684.

Dohoo I, Martin W, Stryhn H (2010) *Veterinary Epidemiologic Research*. 2 ed., 2 impr. Charlottetown: VER Inc. 865.

Dommergues L, Rautureau S, Petit E, Dufour B (2012) Network of contacts between cattle herds in a French area affected by bovine tuberculosis in 2010. *Transboundary and Emerging Diseases*, Malden, v. 59,4: 292-302.

Ferreira MMC (2015) *Quimiometria – Conceitos, Métodos e Aplicações*. Campinas, SP: Editora Unicamp. 493 f.

Galvis JOA, Grisi-Filho JHH, Costa D, Said ALPR, Amaku M, Dias RA, Ferreira F, Gonçalves VSP, Heinemann MB, Telles EO, Ferreira Neto JS (2016) Epidemiologic characterization of bovine tuberculosis in the state of Espírito Santo, Brazil. *Semina: Ciências Agrárias*, Londrina, v. 37, 5:3567-357. Suplemento 2.

Gilbert M, Mitchell A, Bourn D, Mawdsley J, Clifton-Hadley R, Wint W (2005) Cattle movements and bovine tuberculosis in Great Britain. *Nature*, London, v. 435, 7041: 491-496.

Grisi-Filho JHH, Rosales CAR, Ferreira F, Amaku M, Dias RA, Ferreira Neto JS (2011) Análise epidemiológica das condenações de bovinos por tuberculose em abatedouros do Estado de São Paulo. *Arquivos do Instituto Biológico*, São Paulo, v. 78, 2:175-181.

Hibiya K, Kasumi Y, Sugawara I, Fujita J (2008) Histopathological classification of systemic *Mycobacterium avium* complex infections in slaughtered domestic pigs. *Comp. Immunol. Microbiol. Infect. Dis.* 31:347-366.

Kiezun A, Lee IT, Shomron N (2009) Avaliação de técnicas de otimização para seleção de variáveis em regressão logística aplicada ao diagnóstico de infarto do miocárdio. *Bioinformation*; 3: 311-3. (<http://dx.doi.org/10.6026/97320630003311>) PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/19293999>)

Lara GHB, Ribeiro MG, Guazzelli A, Fernandes MC (2009) LINFADENITEINFECCIOSA EM SUÍNOS: ETIOLOGIA, EPIDEMIOLOGIA E ASPECTOS EM SAÚDE PÚBLICA. *Arquivos do Instituto Biológico* [online]. v. 76, 2:317-325. Disponível em: <<https://doi.org/10.1590/1808-1657v76p3172009>>. Epub 02 Jun 2021. ISSN1808-1657. <https://doi.org/10.1590/1808-1657v76p3172009>.

Lima PB, Nascimento DL, Almeida EC, Pontual KAQ, Amaku M, Dias RA, Ferreira F, Gonçalves VSP, Telles EO, Grisi Filho JHH, Heinemann MB, Silva JCR, Ferreira Neto JS (2016) Epidemiological situation of bovine tuberculosis in the state of Pernambuco, Brazil. *Semina: Ciências Agrárias*, Londrina, v. 37, 5:3601-3610, Suplemento 2.

Lin L, Wang Q, Sadek AW (2015) A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. *Transp Res Part C Emerg Technol*; 55: 444-59.

Lozano M, Manyes L, Peiró J, Iftimi A, Ramada JM (2018) Strategic procedure in three

stages for the selection of variables to obtain balanced results in public health research. *Cadernos de Saúde Pública* [online]. v. 34, n. 7, e00174017. Available from: <<https://doi.org/10.1590/0102-311X00174017>>. Epub 23 July 2018. ISSN 1678-4464. <https://doi.org/10.1590/0102-311X00174017>.

Marangon S, Matini M, Pozza MD, Neto F (1998) A case-control study on bovine tuberculosis in the Veneto Region (Italy). *Prev. Vet. Med.*, v.34, p.87-95.

Morés N, Ventura L, Dutra V, Silva VS, Junior WB, Oliveira SO, Kramer B & Neto JSF (2007) Linfadenite granulomatosa em suínos: linfonodos afetados e diagnóstico patológico da infecção causada por agentes do Complexo *Mycobacterium avium*. *Pesq. Vet. Bras.* 27:13-17.

Musoke J, Hlokwe T, Marcotty T, Du Plessis BJA, Michel AL (2015) Spillover of mycobacterium bovis from wildlife to livestock, south africa. *Emerging Infectious Diseases* 21, 448–451. <https://doi.org/10.3201/eid2103.131690>.

Néspoli JMB, Negreiros RL, Amaku M, Dias RA, Ferreira F, Telles EO, Heinemann MB, Grisi-Filho JHH, Gonçalves VSP, Neto JSF (2016) Epidemiological situation of bovine tuberculosis in the state of Mato Grosso, Brazil. *Semina: Ciências Agrárias, Londrina*, v. 37, n. 5, p. 3589-3600. Suplemento 2.

Oliveira VM, Fonseca AH, Pereira AMJS, Carneiro AV, Jesus VLT, Alves PAM (2008) Análise retrospectiva dos fatores associados à distribuição da tuberculose bovina no estado do Rio de Janeiro. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia* [online]. v.603:574-579. Disponível em: <<https://doi.org/10.1590/S0102-9352008000300008>>. Epub 10 Jul 2008. ISSN 1678-4162. <https://doi.org/10.1590/S0102-9352008000300008>.

Palmer MV, Waters WR, Whipple DL (2004) Investigação da transmissão de *Mycobacterium bovis* de veados para bovinos por contato indireto. *American Journal of Veterinary Research* . 65 (11):1483-1489. doi: 10.2460/ajvr.2004.65.1483.

Peña D (2002) Análisis de datos multivariantes. Madrid: McGraw-Hill Interamerica de España.

Phillips CJC, Foster CRW, Morris PA, Teverson R (2003) A transmissão da infecção por *Mycobacterium bovis* ao gado. *Pesquisa em Ciências Veterinárias*. 74 (1):1– 15. doi: 10.1016/s0034-5288(02)00145-5.

Queiroz MR, Groff ACM, Silva NS, Grisi-Filho JHH, Amaku M, Dias RA, Telles EO, Heinemann MB, Ferreira Neto JS, Gonçalves VSP, Ferreira F (2016) Epidemiological status of bovine tuberculosis in the state of Rio Grande do Sul, Brazil. *Semina: Ciências Agrárias, Londrina*, 5:3647-3658.

Radostits OM, Gay CC, Hinchcliff KW, Constable PD (2007) *Veterinary Medicine*. 10.ed., Madrid:Elsevier, .2156.

Renwick AR, White PCL, Bengis RG (2007) Bovine tuberculosis in southern African wildlife: a multi-species host-pathogen system. *Epidemiology and Infection* 135, 529– 40. <https://doi.org/10.1017/S0950268806007205>.

Rocha WV, Jayme VS, Gonçalves VSP, Brito WMED, Pires GRC, Mota ALAA, Grisi Filho O JHH, Dias RA, Amaku M, Telles EO, Heinemann MB, Ferreira F, Ferreira Neto JS (2016) Epidemiological situation of bovine tuberculosis in the State of Goiás, Brazil. *Semina:*

Ciências Agrárias, Londrina, v. 37, n. 5, suppl 2, 3625- 3638.

Silva MCP, Gonçalves VSP, Mota ALAA, Koloda M, Ferreira Neto JS, Grisi Filho JHH, Dias R A, Amaku M, Telles EO, Ferreira F, Heinemann MB, Alfieri AA, Muller EE (2016) Prevalence and herd-level risk factors for bovine tuberculosis in the state of Paraná, Brazil. *Semina: Ciências Agrárias*, Londrina, v37, 5, 3611-3624. Suplemento 2.

Tolles J, Meurer WJ (2016) Regressão logística: relacionando as características do paciente com os resultados. *JAMA*; 316: 533-4.
(<http://dx.doi.org/10.1001/jama.2016.7653>)]PubMed

Veloso FP, Baumgarten KD, Mota ALAA, Ferreira F, Ferreira Neto JS, Grisi- Filho JHH, Dias RA, Amaku M, Telles EO, Heinemann MB, Gonçalves VSP (2016) Prevalence and herd-level risk factors of bovine tuberculosis in the State of Santa Catarina, Brazil. **Semina: Ciências Agrárias**, Londrina, v. 37, n. 5, p. 3659- 3672. Suplemento 2.

Vendrame FB, Amaku M, Ferreira F, Telles EO, Grisi-Filho, JHH, Gonçalves VSP, Heinemann, MB, Ferreira Neto JS, Dias RA (2016) Epidemiologic characterization of bovine tuberculosis in the State of Rondônia, Brazil. **Semina: Ciências Agrárias**, Londrina, 5:3639-3646.

Vicente J, Höfle U, Garrido JM, Fernández-de-Mera IG, Acevedo P, Juste R, Barral M, Gortazar C (2007) Risk factors associated with the prevalence of tuberculosis-like lesions in fenced wild boar and red deer in south central Spain. **Vet Res.** (3):451-64. doi: 10.1051/vetres:2007002. Epub 2007 Apr 11. PMID: 17425933.

Zhang Z (2016) Variable selection with stepwise and best subset approaches. *Ann Transl Med.*4(7):136. doi: 10.21037/atm.2016.03.35. PMID: 27162786; PMCID: PMC4842399

Zimmer J, Anzanello MJ (2014) Um novo método para seleção de variáveis preditivas com base em índices de importância. **Production**. Porto Alegre, RS. v. 24, 1, 84- 93.

CAPÍTULO 3 – Aprendizado de máquina estatístico para análises preditivas em tuberculose bovina

Resumo

Identificar e compreender os fatores de risco de tuberculose bovina endêmica em rebanhos bovinos é fundamental para o controle desta doença. As técnicas exploratórias de aprendizado de máquina podem descobrir relações e interações não lineares complexas dentro das redes de causação de doenças e aprimorar nosso conhecimento sobre os fatores de risco da tuberculose bovina e como eles estão inter-relacionados. Modelos preditivos foram testados e comparados por meio de um banco de dados de uma pesquisa de saúde animal, no caso, para um rebanho infectado de tuberculose bovina, e suas saídas usadas para classificar as variáveis relacionadas a um rebanho positivo. Os resultados mostraram que notadamente nos cinco modelos as variáveis importantes estão correlacionadas a produção leiteira, número de animais, tipo de criação e ao manejo dos animais. Quanto ao desempenho dos algoritmos tivemos o seguinte valor AUC ROC de cada modelo; regressão logística (0,71), decision tree (0,70), random forest (0,70) e o XGBoost (0,74), enquanto todos os modelos forneceram excelente discriminação, o modelo de XGBoost alcançou maior sensibilidade e o modelo random forest alcançou maior especificidade. Conclui-se que os resultados dos nossos modelos preditivos podem ajudar a identificar as prováveis variáveis associadas a tuberculose bovina, porém se faz necessário cautela ao fazer inferências causais a partir de um banco de dados com grande número de variáveis explanatórias.

Palavras-chave: algoritmos, saúde animal, predição

Introdução

Tuberculose bovina é uma zoonose bacteriana global, relatada em 44% dos 188 territórios OIE (Organização Mundial de Saúde Animal) de janeiro de 2017 a junho de 2018 e sua prevalência varia de acordo com a realidade de cada país (Michel et al., 2010; OIE, 2016; Murai et al., 2019). Pode afetar quase todos os mamíferos, embora o gado seja o mais suscetível (Hamzi, 2014).

No Brasil, estudos epidemiológicos conduzidos em 13 Unidades Federativas (Distrito Federal, Espírito Santo, Bahia, São Paulo, Mato Grosso do Sul, Pernambuco, Mato Grosso, Rio Grande do Sul, Goiás, Santa Catarina, São Paulo, Rondônia e Paraná), realizados entre os anos de 1999 e 2015, demonstraram que a tuberculose bovina se encontra distribuída de forma homogênea no país, apresentando baixas prevalências em rebanhos e animais (Galviset al., 2016; Bahiense et al., 2016; Silva et al., 2016; Veloso et al., 2016; Vendrame et al., 2016; Néspoli et al., 2016; Dias et al., 2016; Rocha et al., 2016; Queiroz et al., 2016; Lima et al., 2016; Guedes et al., 2016; Ribeiro et al., 2016; Barbieri et al., 2016). Informações quanto a saúde do rebanho relacionadas ao risco de doença de prevalência elevada, é uma ferramenta importante para a implementação de um sistema de vigilância (Dias et al., 2016). Prever o risco da tuberculose bovina é de grande importância para produtores, laticínios e frigoríficos.

A análise de dados em epidemiologia humana e animal tomou dimensões relevantes nas últimas décadas com a maior disponibilidade computacional e o desenvolvimento de softwares livres. Inúmeros modelos estatísticos de regressão têm sido aplicados para a identificação de fatores de risco, porém o machine learning (ML) é subutilizada na identificação de variáveis preditoras em estudos epidemiológicos na medicina veterinária (Larios et al., 2021). Por meio das técnicas de aprendizado de máquina, computadores são programados para aprender com a experiência passada. Para isso, empregam um princípio de inferência denominado indução. Desta forma, algoritmos do aprendizado de máquina aprendem a induzir a função ou hipótese capaz de resolver um problema a partir de dados que representam observações do problema a ser resolvido (Faceliet al., 2011).

Tipicamente, aprendizado de máquina usado para treinar um modelo que permiteseparar amostras de diferentes classes (ex: saudável ou doente), baseado em

um conjunto de preditores (ex. hábitos alimentares, diabetes), para estimação de variáveis relevantes/importantes para o desfecho estudado (Bispo, 2006).

No Brasil, o uso de algoritmos em saúde animal até então era incipiente. Alguns estudos dentro da veterinária desenvolveram modelos preditivos para identificar a importância de cada variável preditora, pois é fundamental para interpretação e seleção de variáveis para aplicação do modelo (Fusaro et al., 2009; Abrahantes et al., 2009, Holtkamp et al., 2012, Larison et al., 2014).

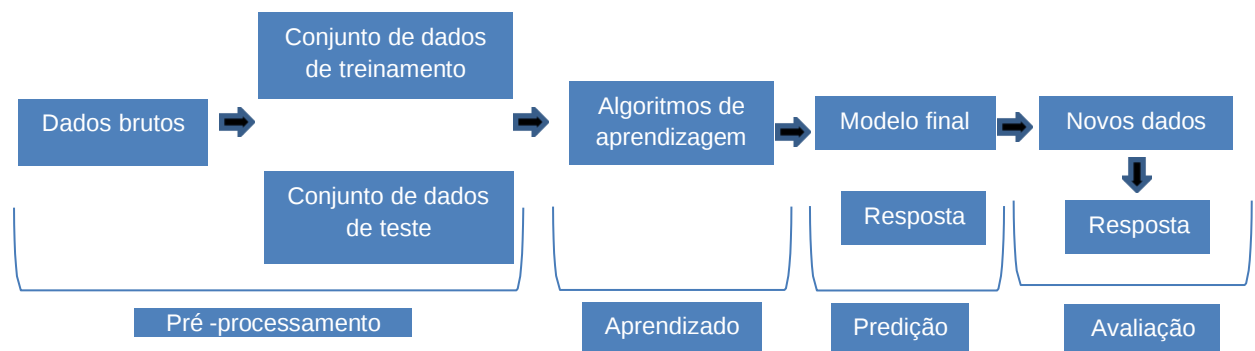
Nesse contexto, é necessário identificar de forma mais precisa as variáveis significativas que pudessem explicar a tuberculose no rebanho e possam se tornar um potencial de risco ambiental, com isso possibilitar a elaboração de estratégias de prevenção e controle da doença. Diante isso, objetivou-se testar os principais algoritmos de aprendizado de máquina e discutir as etapas que compõem uma análise preditiva, de tais algoritmos para prever o risco de tuberculose em bovinos nos municípios do Estado de São Paulo.

Material e métodos

Para a realização do estudo, algoritmos do aprendizado de máquina foram utilizados em um cenário supervisionado, no qual, cada observação do conjunto de dados dispõe de um vetor de mensurações para os preditores x_i , $i = 1, 2, \dots, n$, bem como para a resposta de interesse, y_i . O objetivo principal consiste em ajustar um modelo que relacione a resposta, Y , aos preditores, X , a fim de prever esse evento em observações futuras. O tipo de variável resposta a ser predita define dois subgrupos de aprendizado supervisionado: o de regressão, para variáveis quantitativas, e o de classificação, para as do tipo categórica (qualitativa). O ajuste de modelos preditivos, em ambos os casos, foi representado pelas seguintes etapas: divisão do conjunto de dados em treinamento e teste, pré-processamento, aprendizado e avaliação de modelos (Figura 1).

A divisão da amostra em dados de treinamento e de teste foi realizada para verificar se o modelo apresenta boa performance não apenas em dados utilizados para seu ajuste (treinamento), mas também capacidade de generalização para novas observações (teste). No presente estudo a divisão foi de 75:25, foram utilizados 75% dos dados para treinamento dos algoritmos ($n = 1.743$) e 25% para teste da performance preditiva dos modelos ajustados ($n = 51$). O seguinte passo é realizar uma predição acurada da resposta de interesse (no caso, Y: positivo tuberculose), apartir dos valores de um vetor de preditores, X, contendo informações como tipo de exploração, número de animais, raça bovina e as demais. O aprendizado de um problema de classificação consiste no particionamento do espaço dos preditores em regiões correspondentes às categorias da resposta de interesse.

Figura 1. Roteiro para aplicação de algoritmos do aprendizado de máquina em análise preditiva.



Fonte: Adaptado Raschka e Mirjalili.

A fronteira de decisão entre essas regiões, denominada classificador, representa o modelo preditivo estimado por determinado algoritmo. Portanto, nesse contexto, o objetivo do aprendizado é construir um classificador, $f(X)$, que faça boas predições da resposta de interesse em observações futuras (Izbicki e Santos, 2018).

Desse modo, os algoritmos do aprendizado de máquina são utilizados com o objetivo de estimar f e, portanto, de minimizar o erro redutível. Entre os diversos algoritmos disponíveis, alguns são pouco flexíveis (menos complexos), porém proporcionam uma grande facilidade de

interpretação, como é o caso da árvore de decisão.

É importante observar que modelos menos complexos podem apresentar melhor performance que aqueles que são mais flexíveis por estarem menos sujeitos ao sobreajuste (*overfitting*), quando o modelo gerado se torna muito especializado no conjunto de treinamento, obtendo baixo desempenho quando confrontado com novos dados, e, conseqüentemente, resultarem em predições mais acuradas para Y em novas observações, especialmente para o caso de bancos de dados pequenos, portanto o sobreajuste nada mais é que quando temos um baixo erro no treinamento e alto erro na validação, que é um problema de alta variância, máquina decora e não aprende. Entretanto, não há um único algoritmo capaz de apresentar boa performance em todas as aplicações, sendo importante comparar alguns algoritmos com características distintas para selecionar aquele que resulte em um modelo com performance preditiva satisfatória para o problema em questão. O gradiente boosting baseia-se no conceito de obtenção de algoritmos de classificação/regressão fortes, a partir da transformação de algoritmos de baixo desempenho. Assim como o algoritmo de *Random Forest*, ocorre a distribuição dos pesos às árvores individuais para, posteriormente gerar um consenso a partir da recombinação delas.

Nessa abordagem, o peso de um classificador será maior quanto pior for o desempenho de seu antecessor, previamente ajustado ao longo do processo de aprendizado. Esse método funciona empregando-se a partir da aplicação sequencial dos classificadores a versões reponderadas do conjunto de treinamento, dando maior peso aos registros/observações classificados erroneamente no passo anterior e menos peso aos classificados corretamente. Assim, gera-se uma sequência de árvores que usam a informação contida no treinamento de suas antecessoras, ou seja, o algoritmo vai "aprendendo com o erro".

Para a presente aplicação, foram selecionados para comparação os seguintes algoritmos: regressão logística, árvore de decisão, random forest e gradient boosted trees com implementação XGBoost (Extreme Gradiente Boosting).

Pré-processamento

Os dados originais em sua forma bruta normalmente precisam ser adequados aos modelos preditivos, de modo a otimizar o desempenho de tais modelos. A tais tarefas dá-se o nome de pré-processamento dos dados. O pré-processamento está relacionado às seguintes etapas: transformação de variáveis quantitativas (via padronização); redução de dimensionalidade do conjunto de dados (exclusão

de preditores altamente correlacionados ou utilização de análise de componentes principais); exclusão de variáveis/observações com dados inexistentes e utilização de técnicas de imputação (média, mediana ou valor mais frequente para variáveis categóricas); organização de variáveis qualitativas, (decomposição das variáveis categóricas em um conjunto de variáveis indicadoras que foram utilizadas como preditores (Kuhn e Johnson, 2013; Sakr et al., 2017; Géron, 2017).

Aprendizado: técnicas de reamostragem para otimização de hiperparâmetros

Os métodos do aprendizado de máquina consistem em algoritmos computacionais para relacionar todos ou alguns elementos de um conjunto de variáveis preditoras a um resultado. Para estimar o modelo, eles buscam de forma estocástica ou determinística o melhor ajuste. Esse processo de busca por um melhor ajuste de modelo aos dados difere entre os algoritmos existentes. No entanto, ao longo desse processo, cada algoritmo tenta equilibrar dois interesses: viés (bias) e variância. O viés é o ponto até o qual as previsões ajustadas pelo modelo correspondem aos valores verdadeiros, pode-se dizer que um modelo com alto viés não possui a complexidade necessária para classificar/estimar as observações corretamente. A variância é a sensibilidade das previsões na presença de alguma alteração nos dados de entrada. Rotineiramente busca-se reduzir tanto o viés quanto a variância, e esses dois objetivos geralmente estão em conflito, ou seja, o viés diminuindo pode ocasionar o aumento da variância e o contrário é verdadeiro.

Nesse cenário, o aprendizado de modelos preditivos é composto por dois objetivos principais: seleção e avaliação de modelos. No primeiro caso, para um dado algoritmo que possua hiperparâmetros, a performance de diferentes modelos, baseados em variações dos valores para os hiperparâmetros, é avaliada para selecionar aquele que resulte em melhor desempenho (equilíbrio entre viés- variância). Já no segundo, após a definição do modelo, busca-se estimar seu erro de predição (erro de generalização) em novas observações, geralmente separadas no próprio banco de dados e não utilizadas no processo de aprendizado.

Aprendizado: medidas para avaliação de performance

Uma vez estabelecido o valor de k , é preciso definir uma medida para estimar a performance dos modelos ajustados. Tais medidas são importantes tanto na etapa de seleção quanto na de avaliação dos modelos preditivos. Seu cálculo objetiva mensurar o quanto o valor predito para uma observação se aproxima de seu valor observado. Quando a resposta de interesse é uma variável categórica, dois tipos de predição podem ser obtidos: uma contínua (p_k^*), que é uma estimativa da probabilidade de a nova observação pertencer a cada uma das classes, $k, k = 1, 2, \dots, k$, da resposta de interesse, e outra categórica (por exemplo, 0: desfecho ausente e 1: desfecho presente), que é uma predição para o valor da resposta de uma nova observação.

As predições contínuas são especialmente interessantes por possibilitarem a utilização do classificador (modelo ajustado) em diferentes cenários, a partir do estabelecimento de pontos de corte de acordo com o interesse do pesquisador, em termos de sensibilidade (S) e especificidade (E). A matriz de confusão é a matriz quadrada em que se comparam os verdadeiros valores de uma classificação com os valores preditos através de algum modelo. Sua diagonal é composta pelos acertos do modelo e os demais valores são os erros cometidos. O caso binário, o mais comum, é representado pela seguinte matriz Figura 2, em que a e d denotam casos com resposta corretamente predita, e b e c representam erros de classificação. A sensibilidade ($a/(a+c)$) é a proporção de verdadeiros positivos (VP) entre todos os indivíduos cuja resposta de interesse foi observada, e a especificidade ($d/(b+d)$) refere-se à proporção de verdadeiros negativos (VN) entre aqueles com resposta de interesse ausente.

		Resposta Observada		
Resposta Predita	VP (a)	FP (b)	Total	
			a+b	
	FN (c)	VN (d)		
			b+d	
	a+c	b+d		

Figura 2. Exemplo de matriz de confusão para algoritmos de classificação

Mineração de modelos para o risco de animais positivos utilizando Boosting

XGBoosting (Extreme Gradiente Boosting) é um método geral para melhorar a precisão de qualquer algoritmo de aprendizagem proposto por Friedman (2001), cujas modificações estão descritas em Chen e Guestrin (2016). Foi implementado o método boosting para melhorar a precisão de cada algoritmo dessa aplicação, exceto a regressão logística, uma lista de valores candidatos para os hiperparâmetros foi estabelecida, e, após, utilizando validação cruzada *k-fold*, realizou-se a análise da performance preditiva de cada um desses modelos, por meio da *AUC ROC*, para selecionar aquele com melhor desempenho. Posteriormente, o modelo selecionado foi aplicado aos dados de teste para avaliar seu erro de predição em observações futuras, novamente utilizando a *AUC ROC*. Como a regressão logística não apresenta hiperparâmetros, essa foi ajustada uma única vez aos dados de treinamento, e, na sequência, o modelo com seus parâmetros ajustados foi avaliado nos dados de teste.

Coleta de dados

Utilizou-se um banco de dados obtidos por ocasião do segundo inquérito, realizado em 2011, para conhecer a situação epidemiológica da brucelose e da tuberculose bovina no Estado de São Paulo. Os dados resultaram de um questionário e de testes de tuberculina aplicados por médicos veterinários da Coordenadoria de Defesa Agropecuária do Estado de São Paulo (CDA) em rebanhos bovinos ou bubalinos selecionados por procedimento probabilístico elaborado por pesquisadores que assessoram o Ministério da Agricultura, Pecuária e Abastecimento (MAPA) no Programa Nacional de Controle e Erradicação da Brucelose e Tuberculose (PNCEBT). Detalhes da amostragem podem ser obtidos em Dias et al. (2016). Esses dados foram disponibilizados pela CDA.

Questionário epidemiológico

Foi aplicado um questionário epidemiológico, elaborado para obter informações sobre o tipo de exploração e as práticas de manejo empregadas. Foram selecionadas 46 variáveis como potenciais preditores de tuberculose bovina. Os preditores foram organizados da seguinte forma:

- Tipo de exploração (tipo_exp): animais de corte (1), animais de leite (2), animais leite e corte (3);
- Tipo de criação (tipo_cria): animais mantidos em sistema extensivo (1), animais mantidos semi-confinados (2), animais confinados (3);
- Número de ordenhas por dia (n_ordenha): não ordenha (1), uma ordenha por dia (2), duas ou mais ordenhas por dia (3)
- Número de vacas (n_vacas): quantidade de vacas em lactação (numérica contínua);
- Produção de leite (prod_diaria): produção diária de leite na fazenda (numérica contínua);
- Inseminação artificial (inseminação): não utiliza (1), usa inseminação artificial e touro (2), usa apenas inseminação artificial (3);
- Raça predominante (raca_bov): zebu (1), europeu de leite (2), europeu de corte (3), mestiço (4), outras raças (5);
- Outras espécies domésticas na propriedade: ovinos/caprinos(1), equídeos (1), suídeos (1), aves comerciais (1), cão (1), gato;
- Animais silvestres em vida livre na propriedade: cervídeos (1), capivaras (1), felídeos (1), marsupiais (gambá) (1), macacos (1), outras espécies silvestres (1);
- Aquisição de animais (aquis_bovideos): nos últimos dois anos houve introdução de bovinos ou bubalinos (binário);
- Arrendamento de pastagens (aluga_pasto): aluga pastos em alguma época do ano (binário);
- Compartilha pastagens (pasto_comum): tem pastos em comum com outras propriedades (binário);
- Compartilha utensílios (comp_itens): outros itens que compartilham com

outras propriedades (binário);

- Áreas alagadas (area_alagadas): existem na propriedade áreas alagadiças às quais o gado tem acesso (binário);
- Gado concentrado (gado_conc): existe na propriedade área(s) onde o gado permanece concentrado durante o dia ou à noite (binário);
- Piquete parto (piquete_parto): existe piquete separado para fêmeas na fase de parto e/ou pós-parto (binário);
- Entrega leite (entr_leite): a propriedade entrega leite (binário);
- Finalidade dos produtos lácteos: venda de lácteos (binário);
- Assistência veterinária: tem assistência veterinária (binário);
- Alimentação com soro: alimenta bovinos com soro de leite bovino (binário);
- Classificação da propriedade: rural clássica (1), aldeia indígena (2), assentamento (3) e periferia urbana (4)

As variáveis do questionário passaram por algumas seleções ao longo do estudo, sempre que necessário, foram excluídas ou inclusas no banco para tornar as informações compatíveis. O estudo teve aprovação do Comitê de Ética no Uso de Animais da Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal (nº do protocolo 06726/19). Todas as etapas foram realizadas com o auxílio da linguagem R.

Resultados

A Tabela 1 apresenta os resultados das etapas de aprendizado (treinamento) dos modelos preditivos. A métrica utilizada para otimização e seleção de modelos durante o aprendizado foi o AUC ROC, portanto, o modelo com maior AUC ROC, para cada algoritmo, foi escolhido para avaliação de performance nos dados de teste. Em relação ao *ranking* de importância dos preditores (cinco mais importantes) para os modelos selecionados ao fim da etapa de aprendizado, observa-se que diferentes abordagens produziram *rankings* com preditores similares, porém em sequência diferente.

Para o ponto de corte que maximiza a sensibilidade e a especificidade,

pode-se afirmar que, com uma frequência de falso positivos de 25%, é possível prever, com sucesso, 75% dos rebanhos com animais que irão testar positivo para doença.

Os modelos foram avaliados com base na área sob a curva característica de operação do receptor (AUC), que é uma curva de dependência Se-Sp. Em geral, modelos com AUC de 0.5 podem ser interpretado como tendo desempenho semelhante a uma estimativa aleatória, modelos com AUC maior que 0,7 obtêm bom desempenho, enquanto aqueles com AUC acima de 0,8 são excelentes preditores. A AUC também foi usada para selecionar o método de aprendizado de máquina de melhor desempenho.

Tabela 1. Aprendizado (treinamento) e avaliação (teste) de modelos preditivos

Algoritmos	Regressão logística	Decision tree	Random forest	XGBoost
Treinamento (resultados para o melhor modelo) Hiperparâmetros otimizados	Alpha =1 Lambda =0,001	Custo de complexidade 1,8 x10 ⁻⁷ Profundidade =5 N mínimo= 26	Mtry=1 Nyn = 26 Tree = 324	Learning rate=0,05 Simple size= 1 Loss reduction= -4 Mtry=0,1 Tree= 250 Mini(n)=5 Tree deep =3
AUC ROC				
Média (dp) da validaçãocruzada	0,71	0,70	0,71	0,74
Ranking de importância das variáveis				
1	Comp_aguada	N_vacas	Prod_diaria	N_vacas
2	N_ord enha	Prod_diaria	N_vacas	Prod_diaria
3	Tipo_ordenha	Entr_leite	Raca_bov	Macaco
4	N_silvestre	N_ordenha	N_ordenha	Comp_itens
5	Comp_itens	N_silvestre	Tipo_cria	Raca_bov
Matriz confusão	(VP)288	(VP)317	(VP)370	(VP)288
	(FN)18	(FN)22	(FN)35	(FN)12
	(FP)103	(FP)74	(FP)21	(FP)103
	(VN)27	(VN)23	(VN)10	(VN)33

Dp: desvio padrão; VP: verdadeiro positivo; FN: falso negativo; FP: falso positivo; VN: verdadeiro negativo.

A avaliação de performance dos modelos selecionados evidenciou desempenho satisfatório, com *AUC ROC* superior a 0,70 para todos os modelos. O melhor resultado foi observado para o modelo de XGBoost, seguindo-se random forest e regressão logística, sem grandes diferenças entre eles, e por fim o decision tree.

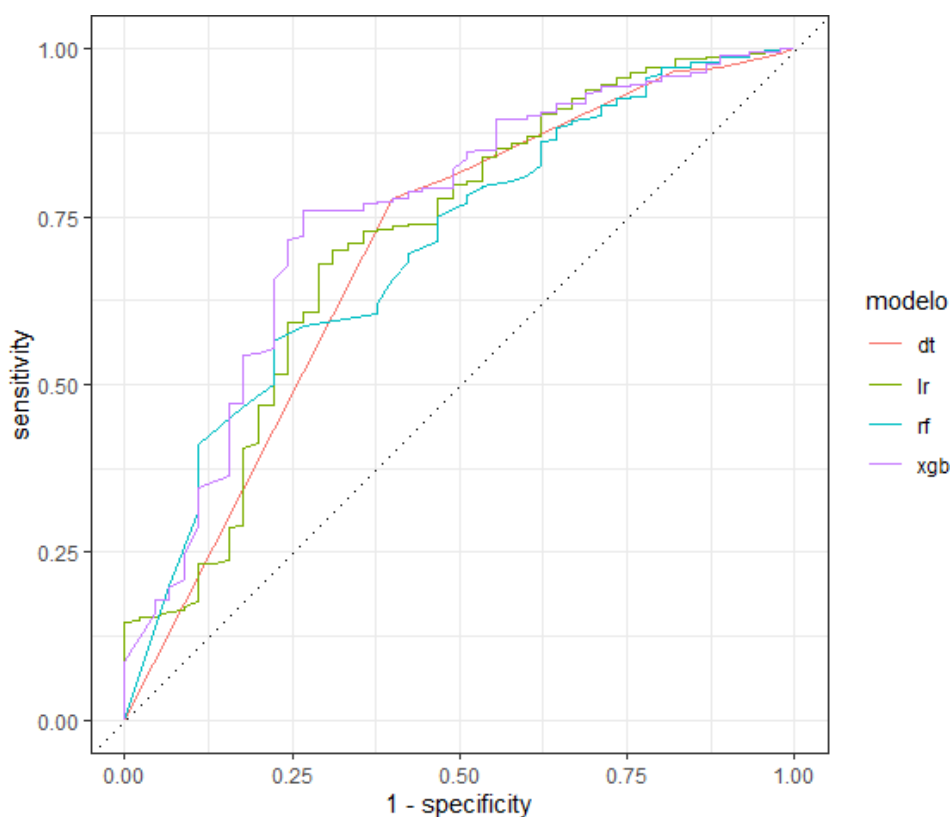


Figura 1. Curva ROC. Dt: decision tree, Lr: regressão logística, rf: random forest, xgb:XGBoost.

Após a avaliação de performance dos modelos selecionados, a xgboost representa uma escolha interessante como modelo final para utilização em dados futuros, por sua menor complexidade quando comparada aos demais algoritmos testados. Caso seja esse o modelo escolhido, as estimativas de seus parâmetros serão utilizadas na predição da resposta de interesse em novas observações podem ser definidas por meio do ajuste de um modelo xgboost ao conjunto de dados completo, utilizando os hiperparâmetros otimizados na etapa de treinamento.

DISCUSSÃO

O presente estudo fez uso de um banco de dados sobre um inquérito epidemiológico na área de tuberculose bovina, para testar e exemplificar as etapas envolvidas na utilização de algoritmos de aprendizado de máquina para análises preditivas em saúde animal. Sabe-se que a performance de um modelo preditivo é dependente do conhecimento relacionado ao problema que se deseja estudar e da disponibilidade de variáveis informativas e com poder discriminatório em relação à resposta de interesse. Portanto, de acordo com a literatura relacionada à tuberculose bovina, foram selecionados dentro do banco de dados preditores relativos à raça, exploração, manejo e outros.

De modo geral, todos os algoritmos avaliados alcançaram AUC ROC superior 0,70 nos dados de testes, e os melhores resultados foram XGBoost, regressão logística e random forest. Resultados semelhantes foram observados por Romero et al. (2021), no ajuste de modelos preditivos para o risco da tuberculose bovina, onde todos os modelos forneceram excelente discriminação, destacando o modelo random forest.

Quanto à classificação de preditores, foram destacadas as cinco principais variáveis de cada algoritmo, havendo semelhanças entre algumas repostas, o que forneceu saídas importantes que também pode ser usado para apoiar as decisões de controle de doenças (Verikas et al., 2011).

O hiperparâmetro XGBoost apresentou como variáveis importantes números de vacas, produção diária e raça bovina, variáveis essas correlacionadas à produção leiteira, apoiando a ideia de que animais de aptidão leiteira apresentam mais chance de positividade que animais de corte (Todeschini et al., 2018). Quanto às variáveis compartilhadas, sabemos que a doença pode se espalhar entre animais de explorações diferentes por via direta (gado arrombamentos ou contato nariz a nariz sobre cercas) ou contato indireto (fômites ou reservatórios de animais selvagens infectados acessando ambos os rebanhos) (Phillips et al., 2003). Vale ressaltar que a variável animais silvestres foi apresentada como variável importante na regressão logística e decision tree e o XGBoost destacou a variável macaco, considerando que a transmissão da interface vida selvagem/pecuária permanece pouco clara e provavelmente bidirecional e determinada por vários fatores, como densidade populacional, prevalência de infecção, distribuição e movimento de infectados e espécies suscetíveis (Bengis et al., 2002; Renwick et al., 2007; Musoke et al., 2015); o macaco é uma espécie suscetível às micobactérias, ocasionando a transmissão para o homem e animais.

Random forest destacou cinco principais variáveis fortemente relacionadas a produção leiteira, dentre elas número de ordenha e produção diária, que são semelhantes a resposta do algoritmo decision tree. No estudo de Romero et al. (2021) metodologias de regressão e aprendizado de máquina de random forest foram implementadas usando dados em nível de rebanho de 2016 para gerar os melhores modelos preditivos possíveis para um incidente de tuberculose bovina na Inglaterra e suas três áreas de risco de vigilância; enquanto todos os modelos forneceram excelente discriminação, os modelos de random forest alcançaram o maior equilíbrio precisão (ou seja, média de sensibilidade e especificidade).

No estado de Mato Grosso do Sul, Guedes et al. (2016), em estudo com 938 propriedades rurais, detectaram prevalências de 1,3% e 0,035% para propriedades rurais e animais positivos para tuberculose, respectivamente. Referiram maior concentração dos focos na região Planalto Sul, onde predominam as propriedades com atividade leiteira, ordenha mecanizada e sofisticação do manejo de produção pecuária, o mesmo observamos em variáveis encontradas pela regressão logística.

Quanto ao algoritmo decision tree, observamos uma variável que se destacou dentre as demais, a variável entrega leite. Franco et al. (2017) investigaram a ocorrência de micobactérias em amostras de leite cru bovino de tanques de expansão individuais, coletivos e 7 provenientes do comércio informal, no sudoeste do estado de São Paulo e detectaram 15 espécies distintas do gênero *Mycobacterium*, incluindo o *M. bovis*, cujas frequências nesses estabelecimentos foram 9%, 7% e 8%, respectivamente, além de espécies não tuberculosas.

A probabilidade prevista de um incidente foi calculada usando as variáveis que contribuíram para o modelo. Existem algumas limitações para este estudo, como não elucidar as vias de transmissão que levam a um incidente de tuberculose bovina em diferentes rebanhos. Seria útil a disponibilidade de informações adicionais para que se pudesse estabelecer mais seguramente relação causal com o desfecho investigado.

Demonstramos uma metodologia para informar uma abordagem para melhorar o controle da tuberculose bovina. Com essas informações, estratégias para implantar intervenções adequadas de prevenção e/ou controle de doenças também podem ser projetadas tanto no primário (ou seja, antes de um incidente, com base nos fatores de risco dos rebanhos) quanto no secundário (ou seja, uma vez que um incidente é declarado em um rebanho-alvo).

CONCLUSÃO

Esta metodologia pode ser adaptada a uma ampla variedade de cenários de controle de doenças em humanos, animais ou sistemas multi-hospedeiros como tuberculose bovina, desde que haja dados suficientes sobre a doença estudada. Os algoritmos utilizados mostraram que a exploração leiteira tem forte relação com a doença, destacando-se o algoritmo XGBoost que resultou em um modelo de saída de alto desempenho, em termos de precisão equilibrada. Finalmente, as saídas de nossos modelos preditivos podem ajudar a identificar a provável redução da tuberculose bovina após a implantação da prevenção de tuberculose bovina direcionada e medidas de controle.

REFERÊNCIAS

- Abrahantes JC, Bollaerts K, Aerts M, Ogunanya V, Stede Y (2009) Salmonella sorosurveillance: Different statistical methods to categorize pig herds based on sorological data. Preventive Veterinary Medicine, 89: 59-66.
- Bahiense L, Ávila LNDe, Bavia ME, Amaku M, Dias RA, Grisi-filho JHH, Ferreira F, Telles EO, Gonçalves VSP, Heinemann MB, Ferreira Neto JS (2016) Prevalence and risk factors for bovine tuberculosis in the State of Bahia, Brazil. Semina: Ciências Agrárias, Londrina, v. 37, 5: 3549-3560. Suplemento 2.
- Barbieri JM, Oliveira, LF, Dorneles EMS, Mota ALAA, Gonçalves VSP, Maluf PP, Ferreira Neto JS, Ferreira F, Dias RA, Telles EO, Grisi-Filho JHH, Heinemann MB, Amaku M, Lage AP (2016) Epidemiological status of bovine tuberculosis in the state of Minas Gerais, Brazil, 2013. **Semina: Ciências Agrárias, Londrina** 5:3531-3548.
- Bengis RG, Kock RA, Fischer J (2002) Infectious animal diseases: the wildlife/livestock interface. Revue scientifique et technique (International Office of Epizootics) 21, 53– 65. <https://doi.org/10.20506/rst.21.1.1322>.

Bispo CM (2006) *Reconhecimento de Padrões e Aprendizado de Máquina*. Nova York: Springer.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In ACM. Proceedings of the 22nd acm sigkdd international conference on knowledge Discovery and data mining, 785-794.

Dias RA, Stanojlovic FMU, Belchior APC, Ferreira RS, Gonçalves SRC, Barão RSC, Sousa PR, Santos AMA, Marcos A, Ferreira F, Telles EO, Grisi Filho JHH, Gonçalves VSP, Heinemann MB, Neto JSF (2016) Prevalência e fatores de risco para tuberculose bovina no estado de São Paulo, Brasil. *Semina: Ciências Agrárias, Londrina*, v. 37,5: 3673-3684.

Faceli K, Lorena AC, Gama J, Carvalho ACPdLFC (2011) *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Grupo Gen-LTC.

Franco MMJ, Ribeiro MG, Pavan FR, Miyata M, Heinemann MB, Souza Filho AF, Cardoso R, Almeida A, Sakate R, Paes AC (2017) Genotyping and rifampicin and isoniazid resistance in *Mycobacterium bovis* strains isolated from the lymph nodes of slaughtered cattle. *Tuberculosis (Edinburgh)*, 04:30-37.

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, 1189-1232.

Fusaro VA, Mani DR, Mesirov JP, Carr SA (2009) Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nature biotechnology*, 27:190-198.

Galvis JOA, Grisi-Filho JHH, Costa D, Said ALPR, Amaku M, Dias RA, Ferreira F, Gonçalves VSP, Heinemann MB, Telles EO, Ferreira Neto JS (2016) Epidemiologic characterization of bovine tuberculosis in the state of Espírito Santo, Brazil. *Semina: Ciências Agrárias, Londrina*, v. 37, 5:3567-357. Suplemento 2.

Guedes IB, Bottene IFN, Monteiro LARC, Leal Filho JM, Heinemann MB, Amaku M, Grisi-Filho JHH, Dias RA, Ferreira F, Telles EO, Gonçalves VSP, Ferreira Neto JS (2016) Prevalence and risk factors for bovine tuberculosis in the state of Mato Grosso do Sul, Brazil. **Semina: Ciências Agrárias**, Londrina 5: 3579-3588.

Hamzi A (2014) *Bovine TB. The Art & Science of Tuberculosis Vaccine Development*. FAO.

Holtkamp DJ, Lin H, Wang C, Oçannor AM (2012) Identifying questions in the American Association of Swine Veterinarians' PRRS risk assessment survey that are important for retrospectively classifying swine herds according to whether they reported clinical PRRS outbreaks in previous 3 years. **Preventive Veterinary Medicine**, 106:42-52.

Izbicki R, Santos TM (2018) Machine Learning sob a ótica estatística: uma abordagem preditiva para a estatística com exemplos em R, [versão em desenvolvimento]. <http://www.rizbicki.ufscar.br/sml.pdf>.

Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, 26.

Larios G, Ribeiro M, Arruda C, Samuel LO, Canassa T, Matthew JB, Marangoni B, Ramos C, Cena C (2021) Uma nova estratégia para o diagnóstico da leishmaniose visceral canina com base na espectroscopia FTIR e aprendizado de máquina. *Journal of Biophotonics*. <https://doi.org/10.1002/jbio.202100141>

Larison B, Njabo KY, Chasar A, Fuller T, Harrigan RJ, Smith TB (2014) Spillover of pH1N1 in Cameroon: an investigation of risk factors. **BMC Veterinary Medicine**, 10:1-8.

Lima PB, Nascimento DL, Almeida EC, Pontual KAQ, Amaku M, Dias RA, Ferreira F, Gonçalves VSP, Telles EO, Grisi Filho JHH, Heinemann MB, Silva JCR., Ferreira Neto, JS (2016) Epidemiological situation of bovine tuberculosis in the state of Pernambuco, Brazil. *Semina: Ciências Agrárias, Londrina*, v. 37, 5:3601-3610, Suplemento 2.

McLaren L, McIntyre L, Kirkpatrick S (2010) Rose's population strategy of prevention need not increase social inequalities in health. *Int. J. Epidemiol.* 39, 372–377. <https://doi.org/10.1093/ije/dyp315>.

Michel AL, Müller BE, Van Helden PD (2010) *Mycobacterium bovis* at the animal human interface: a problem, or not? *Veterinary Microbiology*, 140 (3-4): 371–381.

Murai K, Tizzani P, Awada L, Mur L, Mapitse NJ, Caceres P (2019) Controlling Bovine Tuberculosis: a One Health Challenge. *OIE Panorama Bulletin*.

Musoke J, Hlokw T, Marcotty T, Du Plessis BJA, Michel AL (2015) Spillover of *mycobacterium bovis* from wildlife to livestock, south africa. *Emerging Infectious Diseases* 21, 448–451. <https://doi.org/10.3201/eid2103.131690>.

Néspoli JMB, Negreiros RL, Amaku M, Dias RA, Ferreira F, Telles EO, Heinemann MB, Grisi-Filho JHH, Gonçalves VSP, Neto JSF (2016) Epidemiological situation of bovine tuberculosis in the state of Mato Grosso, Brazil. *Semina: Ciências Agrárias, Londrina*, v. 37, n. 5, p. 3589-3600. Suplemento 2.

OIE – Organización Mundial de Sanidad Animal – Portal WAHIS – Datos de Salud Animal. www.oie.int. Acessado em novembro de 2016.

Phillips CJC, Foster CRW, Morris PA, Teverson R (2003) The transmission of *Mycobacterium bovis* infection to cattle. *Res. Vet. Sci.* 74, 1–15. [https://doi.org/10.1016/S0034-5288\(02\)00145-5](https://doi.org/10.1016/S0034-5288(02)00145-5)

Queiroz MR, Groff ACM, Silva NS, Grisi-Filho JHH, Amaku M, Dias R A, Telles EO, Heinemann MB, Ferreira Neto JS, Gonçalves VSP, Ferreira F (2016) Epidemiological status of bovine tuberculosis in the state of Rio Grande do Sul, Brazil. *Semina: Ciências*

Agrárias, Londrina, 5:3647-3658.

Renwick AR, White PCL, Bengis RG (2007) Bovine tuberculosis in southern African wildlife: a multi-species host-pathogen system. *Epidemiology and infection* 135, 529– 40. <https://doi.org/10.1017/S0950268806007205>.

Ribeiro LA, Gonçalves VSP, Francisco PFC, Mota ALA, Nascimento GT, Licurgo JB, Ferreira F, Grisi-Filho JHH, Ferreira Neto JS, Amaku M, Dias RA, Telles EO, Heinemann MB, Borges JRJ (2016) Epidemiological status of bovine tuberculosis in the Federal District of Brazil. **Semina: Ciências Agrárias**, Londrina 5:3561-3566.

Rocha WV, Jayme VS, Gonçalves VSP, Brito WMED, Pires GRC, Mota ALAA, Grisi Filho JHH, Dias RA, Amaku M, Telles EO, Heinemann MB, Ferreira F, Ferreira Neto JS (2016) Epidemiological situation of bovine tuberculosis in the State of Goiás, Brazil. *Semina: Ciências Agrárias, Londrina*, v. 37, n. 5, suppl 2, 3625-3638.

Romero MP, Chang YM, Brunton LA, Parry J, Prosser A, Upton P, Rees E, Tearne O, Arnold M, Stevens K, Drewe JA (2020) Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making. *Prev. Vet. Med.* 175 <https://doi.org/10.1016/j.prevetmed.2019.104860>.

Romero MP, Chang YM, Brunton LA, Prosser A, Upton P, Rees E, Drewe JA (2021) Uma comparação do valor de dois modelos preditivos de aprendizado de máquina para apoiar o controle da tuberculose bovina na Inglaterra. *Medicina Veterinária Preventiva*, 188, 105264.

Sakr S, Elshaw R, Ahmed AM, Qureshi WT, Brawner CA, Keteyian SJ, Blaha MJ, Al-Mallah MH (2017) Comparison of machine learning techniques to predict all-cause mortality using fitness data: The Henry Ford exercise testing (FIT) Project. *BMC Medical Informatics and Decision Making*, BMC Medical Informatics and Decision Making, 1:1-15.

Silva MCP, Gonçalves VSP, Mota ALAA, Koloda M, Ferreira Neto JS, Grisi Filho JHH, Dias RA, Amaku M, Telles EO, Ferreira F, Heinemann MB, Alfieri AA, Muller EE (2016) Prevalence and herd-level risk factors for bovine tuberculosis in the state of Paraná, Brazil. *Semina: Ciências Agrárias, Londrina*, v37, 5, 3611-3624. Suplemento 2.

Todechini B, Costa EF, Santiago-Neto W, Santos DV, Groff ACM, Borba MR, Corbellini LG (2018) Ocorrência de brucelose e tuberculose bovinas no Rio Grande do Sul com base em dados secundários. *Pesquisa Veterinária Brasileira*, v. 38, 1:15- 22.

Veloso FP, Baumgarten KD, Mota ALAA, Ferreira F, Ferreira Neto JS, Grisi-Filho JHH, Dias RA, Amaku M, Telles EO, Heinemann MB, Gonçalves VSP (2016) Prevalence and herd-level risk factors of bovine tuberculosis in the State of Santa Catarina, Brazil. *Semina: Ciências Agrárias, Londrina*, v. 37, n. 5, p. 3659- 3672. Suplemento 2.

Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: a survey and results of new tests. *Pattern Recognit.* 44, 330–349.

Vendrame FB, Amaku M, Ferreira F, Telles EO, Grisi-Filho JHH, Gonçalves VS P, Heinemann MB, Ferreira Neto JS, Dias RA (2016) Epidemiologic characterization of bovine tuberculosis in the State of Rondônia, Brazil. *Semina: Ciências Agrárias, Londrina*, 5:3639-3646.