CAIO DONIZETTI QUEIROZ ALVES

APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE BRUCELOSE BOVINA A PARTIR DE DADOS DESBALANCEADOS

LAVRAS – MG

2024

CAIO DONIZETTI QUEIROZ ALVES

APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE BRUCELOSE BOVINA A PARTIR DE DADOS DESBALANCEADOS

Dissertação apresentada à Universidade Federal de Lavras como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação para a obtenção do título de Mestre.

Prof. Dr. Danton Diego Ferreira
Orientador

Prof. Dra. Christiane Maria Barcellos Magalhães da Rocha
Coorientadora

LAVRAS – MG

2024

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Alves, Caio Donizetti Queiroz.
Aprendizado de Máquina para predição de brucelose bovina a partir de dados desbalanceado / Caio Donizetti Queiroz Alves. - 2024.
89 p.

Orientador(a): Danton Diego Ferreira.
Coorientador(a): Christiane Maria Barcellos Magalhães da Rocha.
Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2024.
Bibliografia.

1. Aprendizado de máquina. 2. Brucelose. 3. Balanceamento de classes. I. Ferreira, Danton Diego. II. da Rocha, Christiane Maria Barcellos Magalhães. III. Título.

CAIO DONIZETTI QUEIROZ ALVES

APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE BRUCELOSE BOVINA A PARTIR DE DADOS DESBALANCEADOS
MACHINE LEARNING TO PREDICT BOVINE BRUCELLOSIS FROM UNBALANCED DATA

Dissertação apresentada à Universidade Federal de Lavras como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação para a obtenção do título de Mestre.

APROVADA em 30 de janeiro de 2024.

Dr. Danton Diego Ferreira	(UFLA)
Dra. Christiane Maria Barcellos Magalhães da Rocha	(UFLA)
Dr. Bruno Henrique Groenner Barbosa	(UFLA)
Dra. Elaine Maria Seles Dorneles	(UFLA)
Dr. Adriano Olímpio Tonelli	(IFMG)

gov.br
Documento assinado digitalmente
DANTON DIEGO FERREIRA
Data: 25/11/2024 10:16:53-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Danton Diego Ferreira
Orientador

Prof. Dra. Christiane Maria Barcellos Magalhães da Rocha
Co-Orientadora

LAVRAS – MG
2024

Dedico esta monografia a Deus, minha família, minha namorada e a meus amigos.

AGRADECIMENTOS

Agradeço primeiramente a Deus por me possibilitar chegar aqui. Agradeço também à minha família por sempre apoiar minhas decisões, a minha namorada por sempre me incentivar durante a caminhada, e aos professores, profissionais e amigos por despertarem minha vontade de sempre aprimorar meus conhecimentos e tentar ir além. O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG).

Se eu vi mais longe, foi por estar sobre ombros de gigantes.
(Isaac Newton)

RESUMO

A expressividade da pecuária brasileira é inquestionável. Segundo dados da United States Department of Agriculture (USDA), em 2021 o Brasil foi o maior exportador mundial de carne bovina. A brucelose bovina é uma das doenças mais preocupantes para o setor. No Brasil, a brucelose bovina acarreta perdas anuais por volta de 448 milhões de dólares. A preocupação dos órgãos governamentais para controle e erradicação dessa zoonose é notável. Todavia, diversos fatores ameaçam o estabelecimento de ações dos programas de defesa animal vigentes no Brasil, sendo os principais: animais infectados permanecem assintomáticos quando infectados, extensa área territorial brasileira e grande efetivo de rebanhos. Modelos de inteligência computacional como Aprendizado de Máquina (AM) podem ser grandes aliados dos serviços de vigilância sanitária e epidemiológica. Tanto do ponto de vista agropecuária quanto do ponto de vista de saúde única, considerando que a brucelose é uma zoonose, o desenvolvimento de abordagens AM para predição de brucelose possuem elevadíssimo potencial de benefício para a sociedade. A predição da brucelose bovina por meio de questionários aplicados a pecuaristas, em conjunto com outras ferramentas de diagnóstico, podem auxiliar na triagem de propriedades com riscos diferenciados para a doença. Esses benefícios são capazes de possibilitar que programas de defesa animal desempenham ações de forma muita mais célere, eficaz e econômica. O desempenho dos modelos de AM está diretamente ligado com a qualidade e características intrínsecas da base dados utilizada durante a fase de projeto ou treinamento do modelo. O desequilíbrio dos dados disponíveis para as classes, envolvidas em determinado problema, é um exemplo de característica da base de dados que exige tratamento diferenciado. Dessa forma, a depender das características da base de dados, diversas técnicas podem ser adotadas visando tirar melhor proveito das informações ali disponíveis. Nesse trabalho são comparados e avaliados o desempenho de diversas abordagens de AM, combinadas com diferentes técnicas de balanceamento de classe, na predição de brucelose em rebanhos bovinos. Para criação das abordagens foram utilizados os dados do inquérito do serviço oficial de defesa sanitária animal do Estado de Minas Gerais (MAPA/IMA), de setembro de 2010 a dezembro de 2012. O banco de dados contou com registros de 2185 rebanhos, dentre eles, 2103 negativos e 82 positivos para brucelose. Os desempenhos das abordagens foram comparados utilizando diversas métricas recomendadas para problemas envolvendo base de dados com forte desequilíbrio entre classes, sendo elas: Sensibilidade (ou *Recall*), Especificidade, Precisão (ou Valor Preditivo Positivo), *F-measure*, G-mean, e *Index of Balanced Accuracy* (IBA). As abordagens que melhor desempenharam foram as *One-Class Classification* (OCC), as quais atingiram valores de G-mean e IBA próximos a 0,60 e 0,36, respectivamente. A princípio, o grande desafio do problema em questão foi o desequilíbrio entre classes da base de dados utilizada, contudo, os resultados poucos satisfatórios obtidos pelas abordagens de AM, instigaram a execução de uma análise exploratória dos dados. Durante a análise exploratória da base de dados, diversas características da base de dados evidenciaram um problema de alta complexidade a nível de reconhecimento de padrões. Os resultados aqui obtidos mostram que a aplicação de classificadores OCC são opções interessantes para lidar com base de dados que possuem desequilíbrio significativo entre classes e alta complexidade a nível de reconhecimento de padrões.

Palavras-chave: Aprendizado de máquina. Brucelose. Balanceamento de classes. Análise de dados.

ABSTRACT

The expressiveness of Brazilian livestock farming is unquestionable. According to data from the United States Department of Agriculture (USDA), in 2021 Brazil was the world's largest exporter of beef. Bovine brucellosis is one of the most worrying diseases for the sector. In Brazil, bovine brucellosis causes annual losses of around 448 million dollars. The concern of government agencies to control and eradicate this zoonosis is notable. However, several factors threaten the establishment of actions of the animal defense programs in force in Brazil, the main ones being: infected animals remain asymptomatic when infected, extensive Brazilian territorial area and large herds. Computational intelligence models such as Machine Learning (ML) can be great allies of health surveillance and epidemiological services. From both an agricultural and a One Health perspective, considering that brucellosis is a zoonosis, the development of ML approaches for predicting brucellosis has a very high potential benefit for society. Predicting bovine brucellosis through questionnaires administered to livestock farmers, together with other diagnostic tools, can help in screening properties with different risks for the disease. These benefits are capable of enabling animal defense programs to carry out actions much more quickly, effectively and economically. The performance of ML models is directly linked to the quality and intrinsic characteristics of the database used during the model design or training phase. The imbalance of data available for the classes involved in a given problem is an example of a database characteristic that requires different treatment. Therefore, depending on the characteristics of the database, different techniques can be adopted to make better use of the information available there. In this work, the performance of different ML approaches, combined with different class balancing techniques, in predicting brucellosis in cattle herds is compared and evaluated. To create the approaches, data from the survey of the official animal health defense service of the State of Minas Gerais (MAPA/IMA), from September 2010 to December 2012, were used. The database included records of 2185 herds, including , 2103 negative and 82 positive for brucellosis. The performances of the approaches were compared using several metrics recommended for problems involving databases with strong imbalance between classes, namely: Recall, Specificity, Precision (or Positive Predictive Value), *F-measure*, G-mean, and Index of Balanced Accuracy (IBA). The approaches that performed best were the One-Class Classification (OCC), which achieved G-mean and IBA values close to 0.60 and 0.36, respectively. Initially, the great challenge of the problem in question was the imbalance between classes in the database used, however, the unsatisfactory results obtained by ML approaches instigated the execution of an exploratory analysis of the data. During the exploratory analysis of the database, several characteristics of the database highlighted a highly complex problem in terms of pattern recognition. The results obtained here show that the application of OCC classifiers are interesting options for dealing with databases that have significant imbalance between classes and high complexity in terms of pattern recognition.

Keywords: Machine Learning. Brucellosis. Class balancing. Data analysis.

INDICADORES DE IMPACTO

O trabalho gera impactos significativos nos âmbitos social, tecnológico, econômico e ambiental, com potencial para transformar práticas relacionadas à pecuária brasileira. A brucelose bovina, considerada uma das principais zoonoses globais, acarreta prejuízos anuais estimados em 448 milhões de dólares ao setor agropecuário brasileiro, sendo responsável por perdas de produtividade, aumento de custos com sanidade animal e restrições comerciais. A aplicação de técnicas de aprendizado de máquina (AM) no diagnóstico da brucelose possibilita a identificação precoce de rebanhos infectados, mesmo em cenários de dados desbalanceados, onde a maioria das amostras corresponde a rebanhos saudáveis. Esse avanço tecnológico oferece uma alternativa eficiente e econômica aos métodos tradicionais de diagnóstico, permitindo a triagem de propriedades com maior risco de infecção e otimizando a alocação de recursos humanos e financeiros em programas de defesa sanitária. O impacto social desse trabalho é ampliado ao considerar a saúde única, que integra a saúde humana, animal e ambiental. A brucelose, sendo uma zoonose, compromete não apenas a produtividade animal, mas também representa um risco à saúde pública, com transmissão direta ou indireta para humanos por meio do contato com animais infectados ou consumo de seus produtos. No âmbito extensionista, o trabalho colabora com órgãos como o Ministério da Agricultura, Pecuária e Abastecimento (MAPA) e o Instituto Mineiro de Agropecuária (IMA), promovendo a integração entre academia, setor público e comunidades locais. Além disso, os resultados podem ser replicados para outras regiões do Brasil, ampliando o alcance dos benefícios. No campo tecnológico, os resultados obtidos mostram que a aplicação de classificadores One-Class Classification (OCC), associadas a métodos de balanceamento de classes, são opções interessantes para lidar com base de dados que possuem desequilíbrio significativo entre classes e alta complexidade a nível de reconhecimento de padrões. Os impactos econômicos diretos e indiretos incluem a redução de custos operacionais e o aumento da eficiência nos processos de controle e erradicação da brucelose. A mitigação das perdas econômicas associadas à doença favorece a competitividade do Brasil no mercado internacional de carne bovina, consolidando sua posição de liderança global no setor. Por fim, os alinhamentos do projeto com os Objetivos de Desenvolvimento Sustentável (ODS) da ONU, especialmente os ODS 2 (Fome Zero e Agricultura Sustentável), 3 (Saúde e Bem-Estar) e 12 (Consumo e Produção Responsáveis), reforçam a relevância da pesquisa para a agenda global de sustentabilidade e desenvolvimento.

IMPACT INDICATORS

The work generates significant impacts in social, technological, economic, and environmental spheres, with the potential to transform practices related to Brazilian livestock farming. Bovine brucellosis, considered one of the world's major zoonoses, causes annual losses estimated at $448 million in the Brazilian agricultural sector, leading to productivity losses, increased animal health costs, and trade restrictions. The application of machine learning (ML) techniques in brucellosis diagnosis enables the early identification of infected herds, even in scenarios with imbalanced datasets, where most samples correspond to healthy herds. This technological advancement offers an efficient and cost-effective alternative to traditional diagnostic methods, allowing for the targeted screening of properties at higher risk of infection and optimizing the allocation of human and financial resources in animal health programs. The social impact of this work is amplified by considering the One Health approach, which integrates human, animal, and environmental health. Brucellosis, as a zoonosis, not only compromises animal productivity but also poses a public health risk, with direct or indirect transmission to humans through contact with infected animals or consumption of their products. From an extensionist perspective, the work collaborates with organizations such as the Ministry of Agriculture, Livestock, and Supply (MAPA) and the Minas Gerais Agricultural Institute (IMA), fostering integration between academia, the public sector, and local communities. Additionally, the results can be replicated in other regions of Brazil, expanding the scope of the benefits. In the technological field, the results demonstrate that the application of One-Class Classification (OCC) classifiers, combined with class balancing methods, are effective options for dealing with datasets with significant class imbalances and high complexity in pattern recognition. The direct and indirect economic impacts include reducing operational costs and increasing efficiency in brucellosis control and eradication processes. Mitigating the economic losses associated with the disease enhances Brazil's competitiveness in the international beef market, solidifying its global leadership position in the sector. Finally, the project's alignment with the United Nations Sustainable Development Goals (SDGs), particularly SDG 2 (Zero Hunger and Sustainable Agriculture), SDG 3 (Good Health and Well-Being), and SDG 12 (Responsible Consumption and Production), underscores the relevance of the research to the global sustainability and development agenda.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O setor agropecuário é de fundamental importância econômica para o Brasil. Segundo dados da United States Department of Agriculture (USDA), em 2021 o Brasil foi o segundo maior produtor de carne bovina, sexto maior produtor de leite de vaca e o maior exportador de carne bovina, do *ranking* mundial. Ainda de acordo com dados da USDA, o Brasil detém o segundo maior rebanho bovino do planeta, ficando atrás apenas da Índia (USDA FOREIGN AGRICULTURAL SERVICE, 2022).

Diante da expressividade da pecuária Brasileira, a brucelose bovina é uma das doenças mais preocupantes para o setor agropecuário. A Food and Agriculture Organization (FAO) e a World Health Organization (WHO) consideram a brucelose como uma das zoonoses mais prevalentes, principalmente em países em desenvolvimento e de baixa renda (KHURANA et al., 2021). Causada por Cocobacilos Gram-negativos, geralmente *Brucella abortus* e raramente por *Brucella melitensis* e *Brucella suis*, a brucelose bovina é uma das zoonoses economicamente mais importantes do mundo (KOTHALAWALA et al., 2017). Abortos, perdas de bezerros recém-nascidos resultantes de abortos e natimortos, redução na produção de leite, impedimento na exportação e comércio de animais, são alguns exemplos de encargos econômicos que podem decorrer da brucelose bovina. Além do mais, há casos que culminam no abate sanitário de animais infectados (KHURANA et al., 2021). No Brasil foram estimadas percas anuais, atribuíveis à brucelose, em 448 milhões de dólares (CáRDENAS et al., 2019).

Para que programas de prevenção e erradicação da brucelose sejam eficazes, devem ser realizados diagnósticos acurados e precisos da doença. Apesar da Organização Mundial da Saúde (OMS) relatar em sua ficha informativa que cerca de milhões de casos de brucelose são contabilizados todos os anos, estima-se que a taxa real de incidência ainda é 10 a 25 vezes maior do que o número declarado de casos. O fato da maioria dos animais infectados permanecerem assintomáticos quando infectados e de não ser esperado lesões inflamatórias intensas, nem lesões patognomônicas causadas por Brucella spp., corrobora com a estatística (SOUZA et al., 2022). Estas características somadas a extensa área territorial Brasileira, grande efetivo de rebanhos, e outros fatores ameaçam o estabelecimento de ações dos programas de defesa animal vigentes no Brasil. À vista disso, a eficiência e a estratégia dos programas de sanidade animal podem ser otimizadas através de métodos que auxiliem na triagem de propriedades com riscos diferenciados para a doença. Neste sentido, a utilização de inteligência computacional para classificação tem contribuído no diagnóstico de doenças infecciosas (WANG et al., 2022).

Métodos preditivos, como algoritmos de Aprendizado de Máquina (AM), são ferramentas de inteligência computacional poderosas que podem mapear relações complexas e não lineares entre entradas e saídas (LIMA et al., 2022). Modelos de inteligência computacional são utilizadas amplamente nas áreas da psicologia, robótica, biologia, ciência da computação, engenharia e em diversas outras áreas (VARRECCHIA et al., 2021). Por possuírem grande capacidade de previsão de processos complexos e não lineares aliados à facilidade e flexibilidade de implementação, nos últimos anos houve crescente uso de diversos algoritmos de AM para fornecer modelos matemáticos de previsão e classificação de processos biológicos (BENFODIL et al., 2022; BIOURGE et al., 2020; BAUER; JAGUSIAK, 2022; VARRECCHIA et al., 2021; ACHARYA et al., 2018).

Os resultados da classificação feita por AM correm risco de serem comprometidas caso as categorias de classe não estejam representadas suficientemente. É importante que o banco de dados seja representativo o suficiente da classe a que se deseja classificar. Além disso, é importante que essa representação aconteça de forma equilibrada para as classes envolvidas no problema. Em outras palavras, supondo um problema de classificação de duas classes, A e B, sendo estas de complexidade similar, é importante que o número de eventos (amostras) da classe A seja equivalente ao da classe B. Quando este equilíbrio não está disponível na base de dados, podem ser adotadas técnicas de balanceamento das classes para o conjunto de dados de treinamento (SREEJITH; NEHEMIAH; KANNAN, 2020). Essas técnicas são discutidas com maior profundidade na Seção 2.3.

Outro ponto crítico que pode comprometer os resultados preditivos de um algoritmo de AM diz respeito ao número de variáveis do conjunto de dados. Grande número de variáveis não implica na melhoria de desempenho, e provavelmente a redundância de informações reduzirá o desempenho preditivo dos algoritmos (PAUL et al., 2016). Neste ponto, diversas abordagens podem ser empregadas visando reduzir a dimensionalidade dos dados. O método estatístico conhecido por Razão de Discriminação de Fisher (FDR, do inglês *Fisher Discriminant Ratio*) é bastante usado como ferramenta para pré-seleção de variáveis. Através da estratégia adotada pelo FDR é possível ranquear as variáveis que maximizam a separação entre as classes (DUDA; HART; STORK, 2012). A Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) é outro exemplo de método estatístico muito utilizado para este fim. Segundo Evsukoff (2020), o propósito do PCA é encontrar uma transformação linear que elimina a correlação entre as variáveis de entrada. Então, através da eliminação da redundância provocada

pela correlação entre as variáveis, o PCA permite a redução da dimensionalidade do conjunto de dados. Os algoritmos evolutivos também podem ser aplicados como ferramenta para determinar dependências de informações e diminuir o número de variáveis em um conjunto de dados. Os algoritmos evolutivos escolherão um subconjunto de variáveis com a mesma discernibilidade do conjunto de variáveis iniciais, resultando em melhores desempenhos de classificação. Os Algoritmos Genéticos são exemplos de algoritmos evolutivos e solucionam problemas complexos evoluindo de maneira semelhante à seleção natural (ALIčKOVIć; SUBASI, 2015; BRAND et al., 2021).

A classificação de rebanhos potencialmente positivos ou negativos para brucelose permite aperfeiçoamento das estratégias de diagnóstico e controle da doença. Deste modo, o uso de tecnologias que proporcionem esta classificação de forma automática é capaz de dar suporte aos órgãos de Defesa Sanitária, facilitando a triagem de propriedades e proporcionando melhor alocação de recursos humanos e financeiros.

1.1 Objetivo

O objetivo deste trabalho consiste na avaliação de diferentes abordagens de AM, combinadas com diferentes técnicas de balanceamento de classes, para a classificação de rebanhos bovinos quanto à soroprevalência para brucelose.

Os objetivos específicos incluem:

1. confrontar diferentes abordagens de AM frente ao desafio imposto pelo desbalanceamento da base de dados;

2. avaliar adoção técnicas de *undersampling* e *oversampling* para equilibrar classes;

1.2 Estrutura do trabalho

O presente trabalho está estruturado da seguinte forma: a fundamentação teórica, no Capítulo 2, aborda temas relacionados à AM incluindo as principais métricas de desempenhos em problemas de classificação, balanceamento de classes, além da apresentação de alguns trabalhos relacionados. A metodologia utilizada é apresentada no Capítulo 3, com detalhes sobre o banco de dados, ferramentas que foram utilizadas, as abordagens que foram criadas e avaliadas, e a forma como foram definidos os principais hiperparâmetros de cada abordagem. Os melhores hiperparâmetros para cada uma abordagem, bem como os resultados alcançados pelas

mesmas e as análises desses resultados são expostas no Capítulo 4. Um apanhado do trabalho com os principais resultados, conclusões e propostas para futuras pesquisas sobre os problemas abordados nesse trabalho são apresentados no Capítulo 5.

2 FUNDAMENTAÇÃO TEÓRICA

Nesse capítulo é contextualizado o cenário no qual o presente trabalho está inserido, abordando aspectos acerca da brucelose bovina, sua importância e suas implicações no contexto brasileiro. Ademais, elenca conceitos básicos de AM e balanceamento de classe necessários para o entendimento do trabalho. Por fim, são expostos alguns trabalhos relacionados com os temas abordados nessa pesquisa.

2.1 Brucelose

A brucelose bovina se apresenta como uma das mais importantes doenças na produção de bovinos no Brasil. Trata-se de uma doença infecciosa crônica causada principalmente pela bactéria *Brucella abortus*, caracterizada por sua capacidade de sobreviver como parasita intracelular facultativo (RODRIGUES et al., 2021). Essa propriedade permite que o microrganismo se adapte ao ambiente intracelular e supere os mecanismos imunológicos normais do hospedeiro. A bactéria invade e se multiplica em diferentes tipos celulares, incluindo macrófagos, células epiteliais, trofoblastos placentários e células dendríticas. Nos macrófagos, a principal estratégia de sobrevivência é a inibição da fusão entre o fagossoma e o lisossoma. Por ser uma bactéria gram-negativa, *Brucella abortus* possui lipopolissacarídeos na superfície celular que desencadeiam intensas respostas imunológicas no hospedeiro. Os sinais clínicos variam conforme o sexo dos animais. Nas fêmeas, a infecção está associada a abortos, nascimento de crias debilitadas e retenção de placenta. Nos machos, manifesta-se como orquite, epididimite e lesões em glândulas acessórias, resultando em subfertilidade ou infertilidade (DADAR RUCHI TIWARI; DHAMA, 2021; HULL; SCHUMAKER, 2018).

Sob a ótica estritamente econômica, a brucelose bovina acarreta alto prejuízo para o setor pecuário, seja na cadeia do corte ou do leite. As perdas econômicas podem decorrer de custos diretos (redução da produção de carne e leite, abortamento, aumento da infertilidade, entre outros) ou de custos indiretos (vacinação, descarte, abate sanitário, entre outros). De acordo com Deka et al. (2018), estudo realizado no Brasil em 2013 indicaram que a brucelose bovina acarretava uma perda de US$ 2,10 por animal. Observou-se também que cada aumento ou diminuição de 1% na prevalência da doença, aumenta ou decresce, respectivamente, o ônus econômico em US$ 0,37 por animal.

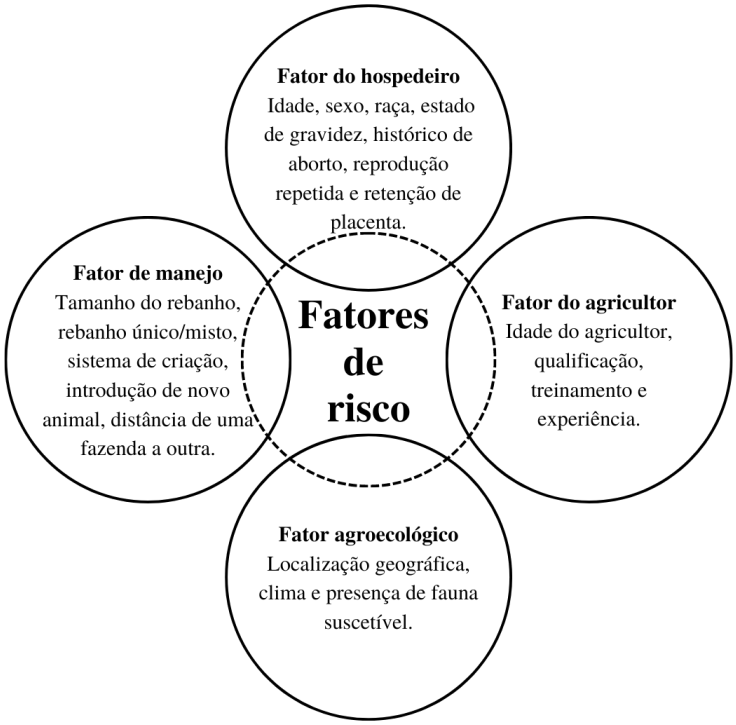
Sob a perspectiva da Saúde Única, que enfatiza a interdependência entre a saúde humana, animal e ambiental, a brucelose também é considerada uma doença de grande relevância.

Do ponto de vista de saúde humana, trata-se de uma zoonose que provoca uma doença grave, crônica e debilitante em humanos. Mesmo diante de sua gravidade, esta zoonose ainda é muito negligenciada, estima-se que aproximadamente 500.000 novos casos de brucelose humana ocorram globalmente a cada ano (ZHANG et al., 2018). A transmissão da brucelose dos animais para o ser humano pode ocorrer por meio do contato direto com secreções de animais infectados, ingestão de derivados lácteos ou leite cru originados de animais infectados, inoculação acidental durante programas de vacinação, ou por transmissão aérea para o vacinador, veterinários, laboratoristas, trabalhadores de matadouros e trabalhadores de campo (RAMIREZ et al., 2020). No contexto de saúde ambiental, a bactéria *Brucella* pode viver por um longo tempo no solo, água, pastagem e esterco. Pode se espalhar facilmente através de uma excreção profusa de organismos na placenta, fluidos fetais e secreções vaginais dos animais. A excreção de *Brucella* no meio ambiente é um fator extremamente preocupante para a saúde pública das famílias rurais, uma vez que não existem vacinas que possam prevenir a brucelose em humanos atualmente (RAMIREZ et al., 2020).

Apesarem de serem semelhantes, os fatores de risco da brucelose podem variar entre países e entre regiões. Na brucelose humana os fatores de risco estão geralmente ligados a ocupações que envolvem contato com gado, ou seus produtos, e o consumo humano de produtos lácteos crus (FRANC et al., 2018). Para a brucelose bovina os fatores de risco comumente estão relacionados com sistemas de produção, zonas agroecológicas, práticas e fatores de manejo. Deka et al. (2018) realizaram uma revisão de literatura e classificaram os principais fatores de risco para o brucelose bovina em quatro grupos. Uma adaptação da classificação elaborada por Deka et al. (2018) pode ser vista na Figura 2.1.

Dada a importância econômica e de Saúde Única da brucelose bovina, a implementação de programas de controle e erradicação da brucelose tem sido amplamente adotada, especialmente em países em desenvolvimento. As estratégias eficazes de controle erradicação da brucelose bovina incluem vigilância, prevenção da transmissão e controle do reservatório da infecção por diferentes métodos, incluindo o descarte. Porém, as dificuldades em realizar o diagnóstico clínico, devido a sintomas inespecíficos, em localizar os animais infectados com brucelose, em conter ou regular o movimento e compra de animais sem teste para brucelose, e falta de educação e interesse dos agricultores em relação à doença, atrapalham o controle e erradicação (CáRDENAS et al., 2019; KHURANA et al., 2021).

Figura 2.1 – Classificação em quatro grupos do principais fatores de risco para infecção por Brucella em animais.



Fonte: Adaptado de Deka et al. (2018).

Zhang et al. (2018) fizeram o levantamento de 23 programas de controle e erradicação de brucelose e resumiram os principais problemas que comprometem o processo de erradicação da brucelose em sete pontos. Dentre o sete pontos apontados por Zhang et al. (2018) estão: a recorrência da doença ou os animais positivos não são detectados a tempo devido à falta de vigilância, e ainda, e o fato dos animais infectados em cada rebanho serem difíceis de detectar porque a prevalência era extremamente baixa. Frente a todos os desafios, Zhang et al. (2018) ressaltam que, mesmo que todas as ferramentas necessárias para a erradicação da brucelose estejam disponíveis, um sistema de vigilância animal e de doenças que seja sistemático, integrado e eficiente é crucial para o sucesso. O registro e a identificação dos animais constituem o núcleo do programa de vigilância.

2.1.1 Programa Nacional de Controle e Erradicação da Brucelose e da Tuberculose Animal (PNCEBT)

Em 2001, o Brasil implementou o Programa Nacional de Controle e Erradicação da Brucelose e da Tuberculose Animal (PNCEBT) por meio da Instrução Normativa 02 de 10 de janeiro de 2001. Este programa inclui medidas de adesão compulsória e voluntária, com o objetivo de padronizar o controle da doença conforme as normas do Código Zoosanitário Internacional. Entre as medidas voluntárias, propriedades leiteiras podem obter a certificação de propriedades livres, enquanto os rebanhos de corte podem alcançar o status de rebanhos controlados. A erradicação da brucelose nessas propriedades é realizada por meio do diagnóstico e abate sanitário de animais reagentes, até que se obtenham três testes consecutivos negativos do rebanho. As medidas compulsórias incluem a vacinação obrigatória de bezerras contra brucelose, com idade entre três e oito meses, o controle de trânsito de animais destinados à reprodução e o abate sanitário de animais infectados (RODRIGUES et al., 2021; ROCHA et al., 2024).

Estudos recentes indicam avanços significativos, proveniente da implementação do PNCEBT, em algumas regiões. Barros et al. (2023) demonstraram que as medidas de controle da brucelose bovina no estado de Mato Grosso, com foco na vacinação de bezerras, geraram benefícios econômicos substanciais para o setor pecuário e a economia do estado. Vale ressaltar que o Mato Grosso é o estado Brasileiro com maior efetivo de rebanho bovino Barros et al. (2023). De forma similar a Barros et al. (2023), Ferreira et al. (2023) investigaram os resultados econômicos, para o setor pecuário e economia do estado Rondônia, produzidos pela implementação do programa de controle da brucelose bovina. Os autores também relataram que o controle da brucelose bovina no estado de Rondônia produziu resultados econômicos altamente vantajosos.

A notificação compulsória da brucelose no Brasil é regulamentada pelo Programa Nacional de Controle e Erradicação da Brucelose e Tuberculose Animal (PNCEBT). Desde 1999, os casos de Brucella abortus têm sido registrados no Sistema de Informação de Saúde Animal do Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Contudo, apenas a partir de 2012, passou-se a detalhar a distribuição dos casos por estado. Os dados oficiais revelam que a brucelose bovina está presente em todos os estados brasileiros, embora Alagoas não tenha registrado notificações nos últimos três anos. A ausência de informações epidemiológicas completas compromete a análise de tendências, o planejamento de estratégias de controle e a tomada de decisões por parte das autoridades sanitárias (ROCHA et al., 2024; FERREIRA et al., 2023).

Segundo Rocha et al. (2024), a subnotificação da brucelose bovina é um dos principais desafios enfrentados pelo Serviço Veterinário Oficial (SVO). Esse problema pode ser atribuído à falta de informações precisas, preocupações econômicas, como a necessidade de descarte de animais positivos, e outros fatores que influenciam a decisão de produtores e veterinários em relação à notificação. Segundo as diretrizes, resultados positivos ou inconclusivos devem ser reportados por veterinários credenciados. No entanto, a presença de animais assintomáticos ou soronegativos infectados por B. abortus também contribui para a subnotificação, dificultando a identificação e o controle eficaz da doença (ROCHA et al., 2024; FERREIRA et al., 2023; BARROS et al., 2023).

Outro aspecto relevante é a insuficiência de veterinários da iniciativa privada para atuar no programa sob supervisão do SVO, especialmente em algumas regiões. De acordo com o Diagnóstico Situacional do PNCEBT/MAPA, até 2018, os estados do Amazonas, Acre, Amapá e Roraima apresentavam escassez desses profissionais. Essa limitação pode ter influenciado os baixos índices de prevalência registrados em certas áreas do país, dificultando a implementação de medidas abrangentes de controle e erradicação da brucelose bovina (ROCHA et al., 2024).

2.2 Aprendizado de Máquina (AM)

Os estudos primordiais no campo da AM datam no início da década de 40. Próximo à essa época, em 1959, o engenheiro do *Massachusetts Institute of Technology* (MIT) chamado Arthur Samuel definiu AM como o campo de estudo que dá aos computadores a habilidade de aprender sem necessariamente terem sido programados para tal (SAMUEL, 1959). Diante disso, de forma sucinta e superficial, pode-se concluir que AM é a ciência da programação de computadores para que eles possam aprender com os dados.

A AM é fruto da combinação interdisciplinar de estatística e ciência da computação (XU et al., 2022). De acordo com Valletta et al. (2017) a AM envolve um conjunto de metodologias que aprendem padrões nos dados. Tais dados devem ser passíveis de previsão. O propósito é que o modelo preditivo generalize bem, ou seja, faça previsões precisas sobre dados inéditos que não foram usados durante a fase de aprendizagem. O modelo preditivo pode aprender o mapeamento entre um conjunto de características e um resultado contínuo (modelo de regressão) ou uma variável categórica (modelo de classificação). De maneira geral, um algoritmo (uma abordagem/modelo de AM) melhora seu desempenho (resultados preditivos) na realização de uma tarefa (regressão ou classificação) a partir da experiência (dados fornecidos na fase

de aprendizagem). A fase de aprendizagem de padrão é comumente denominada como fase de treinamento, e cada exemplo da base de dados usado para treinamento treinamento também é chamado de instância de treinamento (ou ainda observação/amostra) (PECHT; KANG, 2018).

Dentre as possíveis divisões das abordagens de AM, uma delas diz respeito a necessidade de supervisão humana durante a fase de treinamento. Nesse sentido, os modelos de AM podem ser divididos em:

1. **Aprendizado supervisionado:** na aprendizagem supervisionada cada instâncias de treinamento são rotuladas, ou seja, possuem suas respectivas soluções desejadas. As soluções desejadas são comumente denominadas de rótulos, alvo ou classe (PECHT; KANG, 2018). A classificação é uma tarefa característica de aprendizagem supervisionada. Os algoritmos comumente empregados em abordagens de aprendizado supervisionados são: Redes Neurais Artificiais, Árvores de Decisão, Florestas Aleatórias, Regressão linear, Regressão logística e *Support Vector Machine* (SVM) (GéRON, 2019).

2. **Aprendizado não Supervisionado:** o processo de aprendizagem não é supervisionado, para os caso em que os exemplos de entrada não são rotulados (HAN; PEI; KAMBER, 2011; MOHANTY et al., 2021).Métodos de aprendizagem não supervisionados revelam padrões em dados não rotulados. Esses padrões devem ser suficientemente diferentes do ruído puro não estruturado (VALLETTA et al., 2017). Os padrões podem ser descobertos visualizando os dados após redução do número de atributos das amostras (redução de dimensionalidade), ou identificando grupos de amostras que compartilham atributos semelhantes (agrupamento) e/ou determinando a distribuição dos dados (estimativa de densidade). Para abordagens de agrupamento os algoritmos comumente utilizados são: k-Means, *Hierarchical Cluster Analysis* (HCA) e Maximização da Expectativa. Algoritmos como *Principal Component Analysis* (PCA), *Locally-Linear Embedding* (LLE) e *t-distributed Stochastic Neighbor Embedding* (t-SNE) são comumente empregados para redução de dimensionalidade (GéRON, 2019).

3. **Aprendizado semi-supervisionado:** fazem uso de exemplos rotulados e não rotulados durante a fase de treinamento. Em geral, amostras rotulados são usadas para aprender modelos de classes e amostras não rotuladas são usadas para refinar os limites entre as classes. Esse modelos são muito usados para problemas que envolvem a detecção de anomalias (HAN; PEI; KAMBER, 2011; MOHANTY et al., 2021). A maior parte

dos algoritmos de aprendizado semi-supervisionado são resultantes de combinações de algoritmos supervisionados e não supervisionados. A combinação de redes de crenças profundas com componentes não supervisionados denominados de máquinas Boltzmann restritas é um exemplo de modelo de aprendizagem semi-supervisionada.

4. **Aprendizado por reforço:** os resultados preditivos são recompensados ou penalizados de acordo com critérios do usuário. A meta é otimizar a qualidade do modelo adquirindo ativamente conhecimento de usuários humanos (HAN; PEI; KAMBER, 2011).

Outra forma de categorizar os modelos de AM diz respeito ao formato de generalização. Os modelos de AM precisam ser capaz de generalizar (predizer um resultado) em amostras que nunca viu antes. Ter uma bons resultados preditivos nas instâncias de treinamento é bom, mas insuficiente; o principal objetivo é ter um bom desempenho em novos exemplos. Sendo assim, existem duas principais abordagens para a generalização: aprendizado baseado em instâncias e aprendizado baseado em modelo. No aprendizado baseado em instâncias o sistema aprende os exemplos por meio da memorização durante o treinamento e então generaliza para novas amostras utilizando uma medida de similaridade. O algoritmo *k-Nearest Neighbors* k-NN (será elucidado em seções futuras) pode ser um bom exemplo de aprendizagem baseada em instâncias. Por outro lado, aprendizado baseado em modelo a generalização é feita através da aprendizagem de padrões do conjunto de exemplos de treinamento. Assim que o modelo é então treinado utiliza-se os padrões aprendidos para fazer previsões para novas amostras (GéRON, 2019; PECHT; KANG, 2018).

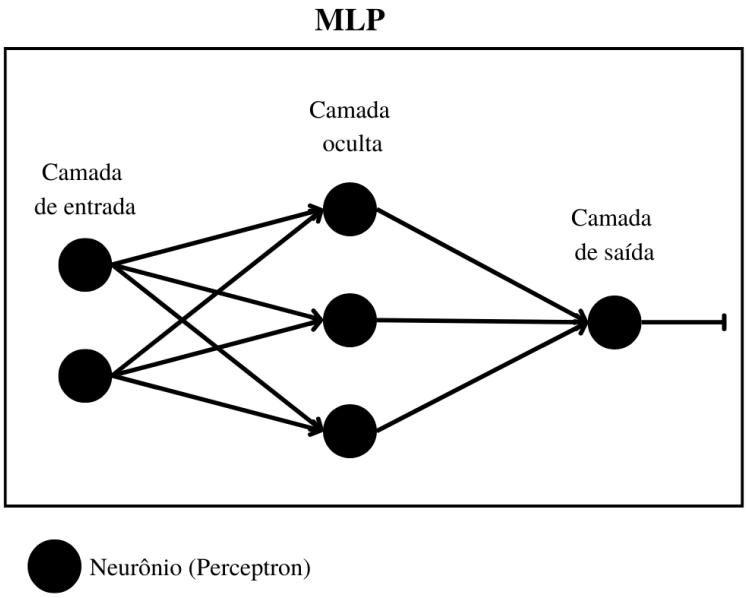
2.2.1 Redes Neurais Artificiais (RNA)

Uma Rede Neural Artificial é uma rede de neurônios artificiais desenvolvida para modelar como o cérebro humano reconhece novos objetos a partir de um aprendizado supervisionado anterior. Cada neurônio equivale a uma regressão logística. O conceito de neurônio artificial foi introduzido pela primeira vez por McCulloch e Pitts em 1943 (WANG et al., 2017; NARTOWT et al., 2019).

Atualmente existem diversas estruturas de RNAs, a mais utilizada em aplicações de classificação é a perceptrons de múltiplas camadas (MLP - *Multilayer Perceptron*). Geralmente uma RNA MLP contém uma camada de entrada, camadas ocultas, e uma camada de saída. As camadas de entrada e saída contêm um número fixo de neurônios, que são habitualmente definidos

com base na estrutura do problema e no conjunto de dados. Na RNA MLP, cada neurônio é uma unidade de algoritmo computacional simples que recebe múltiplas entradas. Após processar as entradas, o neurônio produz uma saída que será usada como entrada dos neurônios na camada seguinte, caso a sua camada não seja a camada de saída. Posteriormente ao calculo da saída, uma função de ativação pode ser usada para calcular a saída final do neurônio (AHMADIAN et al., 2022). Com base na saída dos neurônios da camada de saída é criado a lógica de classificação. A estrutura de uma RNA MLP com 2 neurônios na camada de entrada, uma camada oculta com 3 neurônios, e um neurônio na camada de saída é ilustrada na Figura 2.2.

Figura 2.2 – Estrutura básica de uma Rede Neural Artficial *Multilayer Perceptron*.



Fonte: Do autor (2024).

2.2.2 *k-Nearest Neighbors* (k-NN)

O algoritmo *K-Nearest Neighbors* (k-NN) ou K-Vizinhos Mais Próximos, é um algoritmo de aprendizagem supervisionada baseado em instância proposto inicialmente por Fix e Hodges (1951). O algoritmo k-NN não ganhou popularidade até a década de 1960, quando um maior poder de computação se tornou disponível. Desde então, tem sido amplamente utilizado na área de reconhecimento de padrões em problemas de classificação e regressão (HAN; PEI; KAMBER, 2011; GOODFELLOW; BENGIO; COURVILLE, 2016; PECHT; KANG, 2018).

O k-NN baseia-se no aprendizado por analogia, onde, compara uma determinada amostra com rótulo desconhecido com amostras que possuam rótulos conhecidos previamente semelhantes a ela. Desse forma, o algoritmo k-NN baseia no fato de que objetos semelhantes tendem a estarem próximos uns dos outros. Não existe necessariamente uma etapa de treinamento ou processo de aprendizagem. No momento do teste, encontra-se as k amostras vizinhas (do conjunto de amostras com rótulos conhecidos) mais próximas da nova entrada de teste X. Em seguida, para problemas de regressão, a saída y é produzida com a média dos rótulos das k amostras vizinhas mais próximas. Para problemas classificação, a saída y corresponde a classe (ou rótulo) mais comum entre as k amostras vizinhas mais próximas (HAN; PEI; KAMBER, 2011; GOODFELLOW; BENGIO; COURVILLE, 2016).

A proximidade entre duas amostras é definida em termos de uma métrica de distância entre elas. Supondo que cada amostra X representa um ponto em um espaço d-dimensional, em que d corresponde ao número de variáveis de X. Dessa forma, dados dois pontos $X1$ e $X2$, tem-se que, $X1 = (x_{11}, x_{12}, \ldots, x_{1n})$ e $X2 = (x_{21}, x_{22}, \ldots, x_{2d})$. A distância entre os pontos $X1$ e $X2$ pode ser obtida com a Equação (2.1):

$$\text{dist}(X1, X2) = (x_{11} - x_{21}) + (x_{12} - x_{22}) + \ldots + (x_{1d} - x_{2d}) \tag{2.1}$$

As medidas de distâncias comumente adotadas por algoritmos baseados no k-NN correspondem a distância Euclidiana ou distância Manhattan, definidas respectivamente pela Equação (2.2) e Equação (2.3).

$$\text{dist}(X1, X2) = \sqrt{\sum_{i=1}^{d} (x_{1i} - x_{2i})^2} \tag{2.2}$$

$$\text{dist}(X1, X2) = \sum_{i=1}^{d} |x_{1i} - x_{2i}| \tag{2.3}$$

2.2.3 *One-Class Classification* (OCC)

A maioria dos algoritmos de AM empregados em problemas de classificação atualmente utilizam, durante a fase de treinamento, amostras de treinamento de duas ou mais classes. A partir das amostras de treinamento os modelos criam limites de decisão para cada classe, o quais permitem que o modelo classifique novas amostras de rótulos desconhecido. Todavia, a maioria dos classificadores convencionais assumem que as classes envolvidas no problema são mais

ou menos igualmente equilibrados. Assim sendo, o desempenho preditivo dessas abordagens quando as classes estão desequilibradas normalmente é prejudicado. Nesse âmbito, os modelos de OCC são mais adequados para lidarem com ambientes onde há desequilíbrios entre as classes (WANG et al., 2018; TSAI; LIN, 2021).

Abordagens de OCC utilizam para treinamento amostras de uma única classe (chamada de "classe alvo") e às adotam como um modelo de normalidade e visa detectar dados de qualquer outra classe (chamada de "classe *outlier*") como *outliers*. Dadas as inúmeras aplicações de OCC, esses algoritmos são especialmente usados quando dados de treinamento de classes discrepantes são difíceis ou mesmo impossíveis de serem obtidos (WANG et al., 2018). Tsai e Lin (2021) concluíram em seus estudos que OCC são especialmente adequados para conjuntos de dados com altas taxas de desequilíbrio entre classes. Dentre os diversos modelos de OCC já criados, os métodos baseados em limites, como *One-class Support Vector Machine* (OC-SVM) proposto por Schölkopf et al. (1999), vem sendo amplamente usados em problemas de diversas áreas do conhecimento. .

2.2.3.1 *One-Class Support Vector Machines* (OC-SVM)

O OC-SVM é uma aplicação de algoritmos de *Support Vector Machines* (SVM) para problemas de classe única. O SVM foi proposto por Cortes e Vapnik (1995) e calcula um hiperplano de separação em um espaço $\mathbb{H}$ de alta dimensão maximizando a margem de separação entre as classes envolvidas no problema. O espaço de alta dimensão é induzido por *kernel* (função não linear $\phi(\cdot)$) que realizam produtos escalares entre pontos do espaço de entrada em um espaço de alta dimensão (DOMINGUES et al., 2018). Dado um conjunto de pares de amostras-saída $D = (X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$, onde $X_i \in \mathbb{R}^d$ é a i-ésima amostra de entrada, $y_i \in \{-1, 1\}$ é a saída (classe) correspondente de X_i, n é o número de amostras de treinamento e $\mathbb{R}^d$ indica o espaço de variáveis d-dimensional, a hiperplano de separação pode ser representado de acordo com a Equação (2.4):

$$g(X_i) = w^T X_i + b = 0 \tag{2.4}$$

onde $w \in \mathbb{H}$ e $b \in \mathbb{R}$, e correspondem, respectivamente, a direção e posição do hiperplano no espaço. O hiperplano determina a margem entre as classes, de modo que, as amostras de uma classe ($y_i = -1$) estão de um lado, enquanto as amostras de outra classe ($y_i = 1$) estão do outro:

$$\begin{cases} g(X_i) \leq -1, & \text{se } y_i = -1 \\ g(X_i) \geq 1, & \text{se } y_i = 1 \end{cases} \tag{2.5}$$

A distância z de qualquer amostra Xi ao hiperplano é calculada de acordo com a Equação (2.6):

$$z(X_i) = \frac{|g(X_i)|}{\|w\|} \tag{2.6}$$

Para as amostras mais próximas, denominadas de vetores de suporte, o valor de z corresponde a $\frac{margem}{2}$, dessa forma, o hiperplano construído procura a margem máxima entre as classes. É possível ainda projetar w e b de tal forma que, para os vetores de suportes, $g(X_i) = y_i$ resultando $z(X_i) = \frac{1}{\|w\|}$. Assim sendo, o valor da *margem* de separação será correspondente a Equação (2.7). Observa-se que uma das possíveis formas de maximizar a margem de separação entre as classes é através da redução do valor de $\|w\|$.

$$\frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|} \tag{2.7}$$

Em problemas nos quais as classes envolvidas são linearmente não separáveis, ou seja, não é possível separar as amostras das classes através de uma reta ou hiperplano de separação, é necessário aceitar que aconteça violações da margem. O ato de aceitar violação de margem corresponde ao fato de permitir que algumas amostras fiquem dentro da margem resultando em valores de $g(X)$ que violam a Equação (2.5). É necessário então preocupar-se em manter um bom equilíbrio entre tamanho da margem e violações da margem (PECHT; KANG, 2018).

As variáveis de folga ξ_i e a constante $C > 0$ são introduzidas para modelar o problema de otimização entre tamanho da margem e violações da margem. O valor de ξ_i determina o número de amostras dentro da margem, enquanto que a constante $C > 0$ controla o compromisso entre maximizar a margem e o número de amostras treinamento dentro da margem. Dessa forma, a função objetivo do SVM pode ser definida como um problema de minimização de acordo com a Equação (2.8):

$$\min_{w,b,\xi_i} \frac{\|w\|^2}{2} + C \sum_{i=1}^{n} \xi_i \tag{2.8}$$

sujeito a $y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \quad \text{e} \quad \xi_i \geq 0 \quad \text{para todo } i = 1, 2, \ldots, n.$

Valor grande de C resulta em hiperplano de margem menor, em contraste, valor muito pequeno de C origina um hiperplano com margem maior. Solucionando a Equação (2.8) através dos multiplicadores de Lagrange (CLARKE, 1976), a função de decisão para qualquer amostra X será dada pela Equação (2.9):

$$f(X) = \text{sgn}\left(\sum_{i=1}^{n} \lambda_i y_i K(X, X_i) + b\right) \tag{2.9}$$

sendo $\text{sgn}(\cdot)$ uma função de sinal, λ_i são os multiplicadores de Lagrange, onde, todo $\lambda_i > 0$ é ponderado na função de decisão $f(x)$. $K(X, X_i) = \phi(X)^T \phi(X_i)$ é conhecido como função *kernel*. A Equação (2.12), Equação (2.13), Equação (2.14) e Equação (2.15) são exemplo de alguns dos *kernels* mais utilizados sendo eles, respectivamente, Linear, Polinomial, Gaussiano (ou RBF, do inglês *Radial Basis Function*) e Sigmoide.

Para definirem OC-SVM Schölkopf et al. (1999) fizeram ligeiras modificações no problema de minimização, definido na Equação (2.8), de acordo com a Equação (2.10):

$$\min_{w, \xi_i, \rho} \frac{\|w\|^2}{2} + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i - \rho \tag{2.10}$$

sujeito a $w\phi(x_i) \geq \rho - \xi_i$ e $\xi_i \geq 0$ para todo $i = 1, 2, \ldots, n$. $\rho \in \mathbb{R}$ corresponde ao deslocamento do hiperplano. O parâmetro $\nu \in (0, 1)$ controla o limite superior das amostras de treinamento que são excluídas (consideradas *outliers*) pelo limite de decisão do OC-SVM e um limite inferior para o número de amostras de treinamento usadas como vetores de suporte. O ajuste do parâmetro ν evita *overfitting* do modelo. Provocar *overfitting* no modelo acarretará em desempenhos muito bons no conjunto de treinamento e muito ruins no conjunto de teste, ou seja, o modelo irá decorar as amostras de treinamento ao invés de generalizar o problema.

De forma similar à Equação (2.9), ao utilizar técnicas de Lagrange com função *kernel*, a função de decisão dos OC-SVM é definida de acordo com a Equação (2.11):

$$f(x) = \text{sgn}\left(\sum_{i=1}^{n} \lambda_i K(X, X_i) - \rho\right) \tag{2.11}$$

O OC-SVM cria um hiperplano de separação, caracterizado por w e ρ, em um espaço de alta dimensão induzido pelo *kernel* que realiza produtos escalares entre pontos do espaço de entrada (amostras) em um espaço de alta dimensão. O limite é ajustado às amostras de entrada de forma a maximizar a margem entre os dados e a origem no espaço de alta dimensão.

$$K(X,X_i) = X^T \cdot X_i \tag{2.12}$$

$$K(X,X_i) = (\gamma X^T \cdot X_i + r)^d \tag{2.13}$$

$$K(X,X_i) = \exp(-\gamma \|X^T - X_i\|^2) \tag{2.14}$$

$$K(X,X_i) = \tanh(\gamma X^T \cdot X_i + r) \tag{2.15}$$

Na Equação (2.15) o termo tanh corresponde a função Tangente Hiperbólica.

2.2.3.2 *One-Class Principal Curve* (OC-PC)

Também no campo de OCC, Borges et al. (2023) propuseram um modelo de classificação baseado em Curvas Principais (PCs) (HASTIE; STUETZLE, 1989) denominado de *One-Class Principal Curve* (OC-PC). De acordo com Borges et al. (2023) o OC-PC apresenta como destaque a independência de qualquer otimização de função de erro, baixo custo computacional durante a classificação (fase operacional) e baixo consumo de memória.

O OC-PC possui apenas dois parâmetros de entrada, sendo eles, o número de segmentos (*kmax*) do PC e a taxa de *outlier* (*OR*). Em primeiro momento o OC-PC constrói a curva principal para as amostras disponíveis para treinamento. O próximo passo consiste em calcular as distâncias euclidianas de cada evento do conjunto de dados ao PC. Para então calcular um limite para a distância dos dados até o PC (*dmax*), considera-se apenas a porcentagem de dados mais próxima da curva ($1 - OR$). Uma variável de folga (ξ_d) é usada para aplicar uma margem de tolerância ao classificador. ξ_d corresponde ao desvio padrão das distâncias do conjunto de instâncias de treinamento até a curva principal correspondente. Dessa forma, um limite (th) é definido de acordo com a Equação (2.16):

$$th = d_{max} + \xi_d \tag{2.16}$$

Uma região fechada ao redor do PC é construída com base no limite th. A regra de classificação então é dada por: amostras dentro desta região são atribuídos à classe de dados co-

nhecida e as amostras fora desta região são atribuídos como não pertencentes à classe conhecida (*outliers*).

Os valores atribuídos a *kmax* e *OR* influenciam diretamente no comportamento do classificador. Quanto maior *OR*, mais estreito será a região ao redor do PC. Por outro lado *kmax* controla a flexibilidade do limite em torno dos dados. Limites de decisão suaves e brutos são produzidos quando adota-se valores pequenos de *kmax* devido a alocação de menos segmentos. Por outro lado, limites de decisão curvos e flexíveis são produzidos quando valores grandes de *kmax* são adotados devido a alocação de mais segmentos. Para evitar *overfitting* o valor de *kmax* deve ser pequeno o suficiente, em compensação, deve ser grande o suficiente para capturar a complexidade dos dados e permitir que o modelo generalize o problema proposto.

2.2.3.3 Métricas de desempenho

Comumente é utilizado para avaliar o desempenho de algoritmos de AM para classificação as métricas de acurácia e/ou taxa de erro. Essas métricas podem ser derivadas facilmente de uma matriz de confusão de duas linhas e duas colunas como a representada na Tabela 2.1. Os valores da tabela são definidos como:

- **Verdadeiro Positivo (TP, do inglês *True Positive*):** no qual a classe de uma amostra pertencente a classe positiva foi predita pelo modelo de forma correta;
- **Falso Positivo (FP, do inglês *False Positive*):** diz respeito ao fato da classe de uma amostra pertencente a classe negativa ser predita pelo modelo como positiva;
- **Verdadeiro Negativo (TN, do inglês *True Negative*):** é relativo aos casos em que o modelo prevê a classe de uma amostra da classe negativa corretamente; e
- **Falso Negativo (FN, do inglês *False Negative*):** em que a classe de uma amostra pertencente a classe positiva ser predita pelo modelo como negativa.

Tabela 2.1 – Matriz de confusão para um problema de classificação com duas classes envolvidas (Classes dos Casos Positivos e Classes dos Casos Negativos).

	Predito Positivo	**Predito Negativo**
Real Positivo	TP	FN
Real Negativo	FP	TN

Fonte: Do autor (2024).

A acurácia e taxa de erro podem ser calculadas então, respectivamente, pelas Equação (2.17) e Equação (2.18).

$$\text{Acurácia} = \frac{TP+TN}{TP+FN+TN+FP} \tag{2.17}$$

$$\text{Erro} = 1 - \text{Acurácia} \tag{2.18}$$

Apesar da acurácia e taxa de erro serem comumente usadas, evidências teóricas e empíricas mostram que estas medidas são tendenciosas em ambientes com desiquilíbrio entre as classes envolvidas no problema. Se pouquíssimos exemplos pertencerem à classe positiva, é possível obter valor de acurácia muito alto apenas classificando todas as instâncias como negativas. No entanto, isto é muito prejudicial na maioria dos domínios reais considerando que a classe de interesse geralmente é a das amostras positivas. Como alternativa para contornar essa problemática tem-se adotado as métricas sensibilidade (ou *recall*) e especificidade. Essas métricas permitem monitorar o desempenho da classificação em cada classe separadamente. A sensibilidade Equação (2.19) corresponde a porcentagem de amostras da classe positiva que são classificadas corretamente, enquanto a especificidade Equação (2.20) é definida como a proporção de amostras da classe negativa que são classificados corretamente.

$$\text{Sensibilidade} = \frac{TP}{TP+FN} \tag{2.19}$$

$$\text{Especificidade} = \frac{TN}{TN+FP} \tag{2.20}$$

Em vários problemas pode ser especialmente interessante obter alto desempenho em apenas uma classe, como no caso de diagnóstico de doenças rara. Para este fim, outras métricas podem ser mais interessantes do que apenas sensibilidade e especificidade. A métrica de precisão (ou Valor Preditivo Positivo) tem sido muito adotada nesses casos. A precisão pode ser calculada com a Equação (2.21), e corresponde a porcentagem de exemplos que são corretamente rotulados como positivos.

$$\text{Precisão (ou Valor Preditivo Positivo)} = \frac{TP}{TP+FP} \tag{2.21}$$

Em adição a essas métricas simples, diversas métricas de avaliação mais complexas têm sido utilizadas em diferentes domínios práticos. A medida *F-measure* proposta por Lewis e Gale (1994) é usada para integrar precisão e sensibilidade em uma única medida, e representa uma média harmônica ponderada entre essas duas métricas. O valor de *F-measure* pode ser obtido com a Equação (2.22).

$$F\text{-}measure == \frac{(1+\beta^2)\cdot \text{Precisão}\cdot \text{Sensibilidade}}{\beta^2\cdot \text{Precisão} + \text{Sensibilidade}} \tag{2.22}$$

em que a constante β (não negativa) é um parâmetro para controlar a influência da sensibilidade e da precisão separadamente. A *F-measure* aproxima-se da sensibilidade quando $\beta \to \infty$ e, inversamente, reduz-se à precisão nos casos em que admite-se $\beta = 0$. Na maior parte dos casos, define-se $\beta = 1$ transformando a métrica *F-measure* na média harmônica de precisão e sensibilidade (métrica comumente denominada de $F1$-*score*).

Kubat, Holte e Matwin (1997) propuseram a métrica Média Geométrica (G-mean, do inglês *Geometric Mean*) como possível medida de desempenho para ambientes com desequilíbrio de classes. A ideia central dessa métrica é maximizar a sensibilidade e especificidade mantendo-as equilibradas. Valores elevados para G-mean são obtidos quando a sensibilidade e especificidade também são altas, enquanto valores baixos estão relacionados com pelo menos uma delas baixa. A métrica G-mean é obtida com a Equação (2.23).

$$\text{G-mean} = \sqrt{\text{Sensibilidade}\cdot \text{Especificidade}} \tag{2.23}$$

Apesar da métrica G-mean minimizar a influência negativa de desequilíbrio entre classes, não consegue distinguir entre a contribuição de cada classe para o desempenho global, nem qual é a classe dominante. Isto significa que um mesmo resultado de G-mean pode ser produzido com diferentes combinações de sensibilidade e especificidade. Frente a esse desafio, García, Mollineda e Sánchez (2009) criaram uma nova métrica denominada de *Index of Balanced Accuracy* (IBA), que pode ser definida para qualquer métrica de desempenho *M* de acordo com a Equação (2.24).

$$\text{IBA}_\alpha(M) = (1+\alpha\cdot \text{Dom})\cdot M \tag{2.24}$$

onde o parâmetro $Dom \in [-1,+1]$ corresponde a dominância sendo calculado como $Dom = Sensibilidade - Especificidade$, o qual é ponderado por $\alpha \geq 0$ para reduzir sua influência no

resultado da métrica M. A dominância é usada para estimar a relação entre sensibilidade e especificidade. Quanto mais próxima a dominância estiver de 0, mais próximos serão os valores de sensibilidade e especificidade. É possível observar que, se $\alpha = 0$ ou $Sensibilidade = Especificidade$, a métrica IBA se iguala a métrica M. Na prática, o valor ideal de α a ser escolhido depende da métrica utilizada. A métrica IBA quantifica um certo compromisso entre uma métrica M e um índice de quão equilibradas são os valores de sensibilidade e especificidade (dominância). A função IBA não apenas cuida da acurácia geral do classificador, mas também favorece modelos com melhores resultados na classe positiva (geralmente, a classe mais importante).

2.3 Balanceamento de classes

Domínios desequilibrados levantam desafios significativos ao construir modelos para realizar predição. Comumente a classe com menor quantidade de dados costuma ser a de maior interesse do ponto de vista do aprendizado. A escassa representação dos casos mais importantes leva a modelos que tendem a ser mais focados nos exemplos da classe com maior número de registros, negligenciando os eventos raros (BRANCO; TORGO; RIBEIRO, 2016). Em seu livro He e Ma (2013) dividem as questões e problemas específicos associados ao aprendizado a partir de dados desbalanceados em três categorias/níveis principais: questões de definição de problema, questões de dados e questões de algoritmo.

Diversas abordagens foram desenvolvidas na última década, dentro dos diferentes níveis de problemas propostos por He e Ma (2013), para lidar com a classificação em situações em que as classes estão desbalanceadas. De formar didática, de acordo com Loyola-González et al. (2016), as soluções propostas ao longo dos últimos anos, para mitigar as dificuldades que aparecem na classificação supervisionada usando bancos de dados desbalanceados, podem ser estratificadas em três abordagens básicas:

1. **Nível de dados:** permite aos classificadores atuar de maneira semelhante à classificação padrão, visto que, as instâncias de treinamento são modificadas de forma a produzir uma distribuição de classes mais balanceada. O equilíbrio da distribuição de classes é promovido por meio de reamostragem dos dados. A principal vantagem dos métodos desta abordagem é a possibilidade de serem aplicados, na maioria dos casos, com qualquer ferramenta de AM existente. Como principal inconveniente dessa estratégia é que pode ser difícil mapear a distribuição de dados fornecida em uma nova distribuição ideal,

prejudicando a fase de treinamento do classificador (BRANCO; TORGO; RIBEIRO, 2016).

2. **Nível de algoritmo:** envolve a adaptação dos métodos básicos de classificação supervisionada para estarem mais alinhados com problemas de desequilíbrio de classe. Essa abordagem está relacionada com a criação de novos classificadores, ou modificação dos já existentes, para resolver o problema de desequilíbrio de classes, dependendo fortemente da natureza do classificador. A maioria dos trabalhos que se encaixam nessa abordagem são focados na resolução de um problema específico. É esperado que os modelos dessa abordagem sejam mais compreensíveis para o usuário pelo fato dos objetivos do usuário serem incorporados diretamente nos modelos, sendo caracterizada como principal vantagem desta abordagem. Em compensação, o usuário fica restrito aos algoritmos que foram modificados para poder otimizar seus objetivos ou tem que desenvolver novos algoritmos para a tarefa. Para implementar estes métodos é necessário certo nível de conhecimento das implementações dos algoritmos de aprendizagem de máquina (BRANCO; TORGO; RIBEIRO, 2016).

3. **Sensível ao custo:** incorpora abordagens ao nível dos dados, ao nível de algoritmo ou a ambos os níveis. A ideia principal é atribuir um custo de classificação incorreta mais alto para objetos pertencentes à classe minoritária em relação ao custo de classificação incorreta para objetos pertencentes à classe majoritária, e minimizar erros de custo mais alto.

Métodos de aprendizado por agrupamento *ensemble* também são frequentemente adaptados a domínios desequilibrados. Estes métodos são conhecidos por melhorar o desempenho de um único classificador combinando vários classificadores básicos que superam todos os independentes. Os *ensembles* tornaram-se populares para solução de problemas provocados por desequilíbrio de classes. Podem ser trabalhados modificando o algoritmo de aprendizado por agrupamento na abordagem de nível de dados para pré-processar os dados antes do estágio de aprendizado de cada classificador ou incorporando uma estrutura sensível ao custo no processo de aprendizado conjunto (LóPEZ et al., 2012; HAIXIANG et al., 2017).

2.3.1 Reamostragem

As técnicas de reamostragem são as mais utilizadas para balanceamento de classe. Estas técnicas reequilibram o espaço amostral para um conjunto de dados desequilibrado com a finalidade de aliviar o efeito da distribuição de classes distorcida durante o treinamento da RNA. Os métodos de reamostragem independem do classificador selecionado o que os tornam mais versáteis. As técnicas de reamostragem se dividem basicamente em três grupos segundo Haixiang et al. (2017), sendo eles:

- ***Oversampling:*** elimina os danos da distribuição distorcida criando novas amostras para as classes minoritárias. A duplicação aleatória das amostras minoritárias e o SMOTE *(Synthetic Minority Over-sampling Technique)* proposto por Chawla et al. (2002) são os dois métodos amplamente utilizados para criar as amostras sintéticas. Apesar desses dois serem amplamente utilizados, existem atualmente inúmeros algoritmos propostos, desde métodos mais simples a abordagens mais robustas. Deve atentar-se ao utilizar técnicas de *oversampling*, pois, a depender da quantidade de objetos artificiais incluídos, podem provocar *overfitting* (treinamento em excesso) durante a fase de treinamento do classificador;

- ***Undersampling:*** elimina os danos da distribuição assimétrica, descartando as amostras da classe majoritária. O *Random Under Sampling* (RUS) proposto por Tahir et al. (2009) é uns dos algoritmos mais simples de *undersampling* e envolve a eliminação aleatória de exemplos das classes majoritárias. Existem diversos algoritmos propostos, com variado grau de complexidade. O ponto crítico desta abordagem é a possibilidade de excluir alguns objetos representativos do conjunto de treinamento, afetando assim o desempenho do classificador; e

- **Métodos híbridos:** combinam métodos de *oversampling* e métodos de *undersampling*. Essa abordagem tem os mesmos pontos críticos das técnicas de *oversampling* e *undersampling* citados anteriormente.

As estratégias usadas com maior frequência para criar ferramentas de *oversampling*, *undersampling* ou híbridos são métodos baseados em *cluster* (por exemplo: K-means), métodos baseados em distância (por exemplo: KNN (COVER; HART, 1967)) e métodos evolutivos (por exemplo: Algoritmo Genético) (HAIXIANG et al., 2017). Loyola-González et al. (2016)

analisaram o desempenho de métodos de reamostragem na estrutura de conjuntos de dados desbalanceados em relação a abordagens baseadas em aprendizado sensível ao custo. No estudo os autores consideraram dois métodos de *oversampling*: SMOTE e SMOTE-ENN (BATISTA; PRATI; MONARD, 2004), uma versão sensível ao custo e uma abordagem híbrida que integra ambas as abordagens. Loyola-González et al. (2016) observaram que ambas abordagens melhoram o desempenho geral, dado o problema de desequilíbrio, em todos os critérios utilizados no estudo. O estudo estatístico realizado permitiu dizer que, tanto as técnicas de reamostragem quanto o aprendizado sensível ao custo são abordagens boas e equivalentes para resolver o problema de desequilíbrio. Os autores enfatizaram ainda que razão de desequilíbrio - definida como a razão entre o número de instâncias da classe majoritária e da classe minoritária - é importante, contudo, questões como a sobreposição de classes e problemas relacionados com deslocamento do conjunto de dados podem surgir para alguns casos e prejudicar o desempenho do classificador.

Apesar da melhora apontada por Loyola-González et al. (2016), como dito anteriormente, a abordagem de *oversampling* adotada pode provocar *overfitting* durante a fase de treinamento do classificador, uma vez que, o classificador é exposto às mesmas informações. O SMOTE é uma abordagem alternativa de *oversampling* que gera novos dados de uma maneira que visa minimizar esse problema. Os dados sintéticos gerados com a técnica SMOTE são criados ao longo do segmento de linha que liga exemplos de classe minoritária. Apesar de mitigar a ocorrência *overfitting*, o SMOTE tem a desvantagem de criar exemplos sintéticos da classe minoritária, desconsiderando os exemplos da classe majoritária, podendo então criar dados ruidosos e na fronteira de decisão. Diante desta problemática várias modificações do SMOTE foram propostas. O Borderline-SMOTE (HAN; WANG; MAO, 2005), SMOTE-ENN (BATISTA; PRATI; MONARD, 2004), SMOTE-RSB (RAMENTOL et al., 2012), MSMOTE (HU et al., 2009), ADASYN (HE et al., 2008), Safe-Level-SMOTE (BUNKHUMPORNPAT; SINAPIROMSARAN; LURSINSAP, 2009) e diversos outros, são exemplos dessas variações do SMOTE.

Stefanowski (2016) estudou o impacto de várias características de dados no desempenho de algoritmos de AM e estratégias de reamostragem de dados. Dentre as características de dados, às quais foram dadas o nome de fatores de dificuldade de dados, Stefanowski (2016) também incluiu, assim como Loyola-González et al. (2016), o problema de sobreposição de classes. López et al. (2013) também estudaram várias características intrínsecas dos dados que

têm forte influência na classificação desequilibrada, e elencaram como as mais importantes às seguintes: presença de conjuntos disjuntos, tamanho do conjunto de dados e a falta de densidade no conjunto de treinamento, sobreposição de classes, presença de dados ruidosos, presença de instâncias na fronteira de decisão e sua relação com exemplos de ruído, e deslocamento do conjunto de dados. Superar esses problemas passa ser ponto chave para a melhoria do desempenho dos algoritmos de classificação.

É importante distinguir instâncias na fronteira de decisão de ruídos. Instância na fronteira de decisão correspondem a dados que estão situados na zona ao redor dos limites de classe, locais onde as classes minoritária e majoritária se sobrepõem. Estas instâncias são preocupantes, visto que, uma pequena quantidade de ruído pode movê-las para o lado errado do limite de decisão do classificador. Por outro lado, ruídos são indivíduos de uma classe ocorrendo nas áreas seguras da outra classe (BASHIR et al., 2020). Ruídos podem afetar negativamente qualquer sistema de classificação. Para o cenário com classes desbalanceadas a presença de ruído tem maior impacto nas classes minoritárias do que nas classes majoritária, em virtude de que a classe minoritária possui menos dados, sendo necessários menos exemplos ruidosos para impactar o aprendizado do classificador.

Dados ruidosos e o gerenciamento de exemplos na fronteira de decisão estão intimamente relacionados. Quanto melhor for a definição das fronteiras de decisão, melhor será o desempenho do classificar. Visando melhorar essa definição, existem diversas técnicas de limpeza que se fundamentam na detecção de instâncias na fronteiras de decisão e distinção de instâncias ruidosas que podem prejudicar a classificação. Muitas dessas técnicas tendem a melhorar a sensibilidade em detrimento da especificidade. Métodos como Borderline-SMOTE, ADASYN e Safe-Level-SMOTE implementam mecanismos de *oversampling* que têm potencial de amenizar esse problema.

2.3.1.1 ADASYN

A questão fundamental do algoritmo ADASYN *(Adaptive Synthetic Sampling)* proposto por He et al. (2008) é compensar as distribuições desbalanceadas alterando, de forma adaptativa, os pesos de diferentes exemplos minoritários. Para isso, o ADASYN utiliza uma distribuição de densidade como critério para decidir automaticamente o número de amostras sintéticas que precisam ser geradas para cada exemplo minoritário. Como resultado, os dados sintéticos adi-

cionais são gerados para exemplos da classes minoritária que são mais difíceis de aprender (DATTA et al., 2022).

O ADASYN é uma versão adaptativa do método SMOTE, e tem sido utilizado nas mais diversas áreas do conhecimento para problemas que envolvem classificação em ambientes de dados desbalanceados. Na área de sensoriamento remoto Datta et al. (2022) utilizaram o ADASYN visando equilibrar o conjunto de dados para treinamento de um classificador de imagem hiperespectral. Na medicina humana Alhudhaif (2021) aplicou o ADASYN para equilibrar classes de sinais cerebrais (EEG-Eletroencefalografia) que alimentaram um modelo de predição de epilepsia. Dentre as abordagens estudadas por Alhudhaif (2021) o método ADASYN aumentou o sucesso da classificação. Hasmita et al. (2019), em seus estudos para criarem um modelo que fosse capaz de prever preços do pimentão no mercado agrícola, empregaram o ADASYN para superar o problema de classes desbalanceadas dos dados obtidos. Segundo os autores o uso do ADASYN desempenhou um papel essencial no processo de previsão, e permitiu o classificador alcançar 100% de precisão. Como será visto na Seção 2.4, na medicina veterinária também existem diversos estudos de AM que fazem uso do ADASYN como algoritmo de *oversampling*.

De acordo com He et al. (2008), o ADASYN não apenas reduz o viés de aprendizado introduzido pela distribuição de dados desequilibrada, como também pode mudar o limite de decisão, de forma adaptativa, para se concentrar nos exemplos da classe minoritária mais difíceis de aprender. Como resultado, o ADASYN melhora o aprendizado da distribuição de dados de duas formas: (1) reduzindo o viés causado pelo desequilíbrio de classe e (2) alterando o limite de decisão de classificação na direção de exemplos mais difíceis de aprender.

O ADASYN atinge seus objetivos primeiramente calculando-se a quantidade total de exemplos da classe minoritária a serem gerados de acordo com a Equação (2.25).

$$G = (c_n - c_p) \cdot \beta, \tag{2.25}$$

na qual, c_n corresponde a quantidade de exemplos da classe majoritária, c_p a quantidade de exemplos da classe minoritária e $\beta \in [0,1]$ especifica o nível de balanceamento após a criação dos exemplos sintéticos. $\beta = 1$ implica em um conjunto de dados totalmente balanceado. Então, para cada exemplo $p_i \in \mathbb{P}$, em que $\mathbb{P}$ é o conjunto minoritário, encontra-se os K vizinhos mais próximos, com base na distância euclidiana no espaço v dimensional, e calcula-se a proporção

r_i de exemplos da classe majoritária vizinhos de p_i conforme a Equação (2.26).

$$r_i = \frac{\Delta_i}{K}, \tag{2.26}$$

sendo, Δ_i o número de exemplos nos K vizinhos mais próximos de p_i que pertencem à classe majoritária, portanto $r_i \in [0,1]$. Calculado r_i, deve-se normalizá-lo de acordo com a Equação (2.27), de modo que $\hat{r}_i$ é uma distribuição de densidade ($\sum_i \hat{r}_i = 1$).

$$\hat{r}_i = \frac{r_i}{\sum_{n=1}^{c_p} r_i}, \tag{2.27}$$

Por fim, é calculado o número de exemplos de dados sintéticos que precisam ser gerados (q_i) para cada exemplo minoritário p_i, para isso utiliza-se a Equação (2.28).

$$q_i = \hat{r}_i \cdot G, \tag{2.28}$$

Para gerar os exemplos sintéticos para cada exemplo p_i deve-se repetir os seguintes passo para cada p_i:

1. Escolha aleatoriamente um exemplo $p_{zi} \in \mathbb{P}$ dentre os K vizinhos mais próximos de p_i;
2. Gere o exemplo sintético s_i utilizando a Equação (2.29).

$$s_i = p_i + (p_{zi} - p_i) \cdot \lambda, \tag{2.29}$$

   onde, $(p_z i - p_i)$ é o vetor diferença no espaços de v dimensões, e λ corresponde a um número aleatório na qual $\lambda \in [0,1]$.

3. Repita os passos 1 e 2 até que a quantidade de exemplos sintéticos gerados para p_i seja igual a q_i.

2.4 Trabalhos relacionados

As RNAs são ferramentas de inteligência artificial que atualmente tem sido amplamente usadas para auxiliar pecuarista, médicos veterinários e demais integrantes do setor agropecuário a tomarem melhores decisões. RNAs são capazes de processar grandes quantidades de dados

para identificar padrões e prever resultados. Elas podem ser usadas para ajudar os veterinários a diagnosticar doenças, prever o comportamento de animais, prever respostas a tratamentos, prever prognóstico de um animal, entre outras diversas aplicações.

Gouda et al. (2022) avaliaram quatro abordagens de AM para predição de Febre Catarral Maligna (FCM) em animais. Os modelos analisados foram: *Random Forest* (RF), Árvore de Decisão (AD), RNA e Regressão Logística (RL). Utilizou-se os dados do estudo de soroprevalência de 233 animais aparentemente saudáveis (125 ovinos e 108 caprinos) de cinco províncias diferentes do Egito para avaliar o desempenho dos algoritmos na previsão da FCM. Dentre os quatro modelos aferidos os que obtiveram melhores resultados foram RF e RNA, com respectivos valores de acurácia obtidos no conjunto de dados de teste: 83% e 81%. Os autores descobriram ainda que as variáveis idade e estação do ano são as mais importantes para predição. Diante dos resultados obtidos, Gouda et al. (2022) concluíram que algoritmos de AM, principalmente RF e RNA, fornecem poder preditivo adequado e competência significativa na identificação e classificação de fatores de risco potenciais para FCM.

De modo similar, Bobbo et al. (2021) compararam oito métodos diferentes para prever o estado de saúde do úbere de vacas com base na contagem de células somáticas, sendo eles: Análise Discriminante Linear, Modelo Linear Generalizado com função de ligação logit, Naïve Bayes, Árvores de Classificação e Regressão, *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), *Random Forest* (RF) e RNA. Os melhores desempenhos, de acordo com diferentes métricas, na previsão de classes de saúde do úbere em um determinado dia de teste (saudável ou mastítico) com base nas características do leite de vaca registradas no dia teste anterior foram alcançados com os modelos RF e RNA. A classificação em saudável ou mastítico foi feita de acordo com a contagem de células somáticas abaixo ou acima de um limite predefinido de 200.000 células/ml. Os resultados encontrados mostraram que a RNA é uma promissora abordagem que ajudaria os pecuaristas identificarem com antecedência as vacas que possivelmente teriam alta contagem de células somáticas no dia de teste subsequente.

Também em rebanho leiteiro, Bobbo et al. (2022) compararam quatro abordagens de AM (Modelo Linear Generalizado, SVM, RF e RNA) para prever a presença ou ausência de mastite subclínica em búfalos mediterrâneos italianos. O conjunto de dados contou com 3891 registros de 1038 búfalos de 6 rebanhos situados na região de Basilicata (sul da Itália). A RNA foi o método que obteve melhor desempenho na base de dados de teste, com valores de sensibilidade e especificidade iguais a 67,7% e 83,6%, respectivamente. Indo de encontro a Bobbo

et al. (2021), Bobbo et al. (2022) concluíram que os métodos de AM, devido ao poder de exploração de grande quantidade de dados disponíveis atualmente, são ferramentas promissoras na prevenção e vigilância da mastite subclínica. Ainda na pecuária leiteira, Lima et al. (2022) utilizaram RNA aliada a técnica de espectroscopia de infravermelho por transformada de Fourier (FTIR) para detectar a adição de soro de queijo ao leite. Foram adicionadas 520 amostras de leite cru ao soro de queijo em diferentes concentrações e 65 amostras foram usadas como controle. Foram utilizados resultados complementares de 520 amostras de leite cru autêntico. O ponto de congelamento (°C) foram previstos usando FTIR e, juntamente com componentes selecionados (gordura, proteína, caseína, lactose, sólidos totais e sólidos não gordurosos), foram usados como recursos de entrada para a RNA. Os valores de sensibilidade e especificidade obtidos pela RNA ficaram acima de 95%. Através dos resultados atingidos os autores concluíram que RNA aliada a técnica de FTIR é relevante na prática, pois podem ser implantadas em uma rotina laboratorial de análise da qualidade do leite para auxiliarem na triagem de amostras suspeitas.

Na área de pequenos animais, Biourge et al. (2020) fizeram uso de RNA para predizer de forma precoce a doença renal crônica (DRC) em gatos utilizando dados de triagem de saúde de rotina. O modelo foi construído com dados de 218 gatos saudáveis com idade acima de 7 anos selecionados no *Royal Veterinary College* (RVC). O desempenho do modelo foi aferido em dados de 3.546 gatos nos registros do *Banfield Pet Hospital* e mais 60 do RVC - todos inicialmente sem diagnóstico de DRC. Os resultados de sensibilidade e especificidade alcançados pelo modelo, para predição de DRC em 12 meses, foram de 87% e 70%, respectivamente. Biourge et al. (2020) constataram então que RNA pode auxiliar na identificação de indivíduos na população geral de gatos (com 7 anos ou mais de idade) que estão em risco de desenvolver DRC em 12 meses, e portanto, se beneficiarem de um acompanhamento clínico mais frequente para maximizar as oportunidades de diagnóstico e intervenção precoce da DRC.

No âmbito da Brucelose, ZareBidaki et al. (2022) realizaram um estudo em Birjand, leste do Irã, de maio a agosto de 2018, com objetivo de determinar a ocorrência de espécies de Brucella. Foram coletadas 200 amostras pareadas, incluindo amostras de sangue (100) e leite (100), de maio a agosto de 2018, de 100 animais (57 cabras; 13 ovelhas; 30 vacas) durante os serviços de saúde de rotina. ZareBidaki et al. (2022) avaliaram também o papel de certos fatores de risco da brucelose na previsão da ocorrência do patógeno utilizando RNA. A sensibilidade e especificidade alcançada pela RNA na base de dados de teste foram de 81% e 62%, respectiva-

mente. Foi possível identificar que os fatores de risco tipo de rebanho, última data de vacinação (dias) e idade do animal (mês) foram as variáveis mais importantes no desenho do modelo, e as variáveis o contato com a fauna e aborto tiveram importância secundária.

Lopes (2016) desenvolveu um modelo matemático utilizando RNA para predizer rebanhos positivos e negativos para a brucelose bovina. Para criação dos modelos foram utilizados os dados do inquérito do serviço oficial de defesa sanitária animal do Estado de Minas Gerais (MAPA/ IMA), de setembro de 2010 a dezembro de 2012. O banco de dados contou com registros de 2185 rebanhos, dentre eles, 2103 negativos e 82 positivos para brucelose. Devido ao notável desequilíbrio entre as classes de rebanho (positivo para brucelose e negativo para brucelose), que poderia prejudicar o treinamento da RNA, Lopes (2016) utilizou técnicas de *oversampling* e *undersampling* para promover equilíbrio entre as classes no conjunto de treinamento. A técnica de *undersampling* contou com a utilização do algoritmo K-means para divisão dos dados da classe majoritária em 4 agrupamentos. Posteriormente foram selecionadas aleatoriamente (seguindo uma distribuição uniforme) alguns registros dos 4 grupos com o objetivo de formar um banco de dados de treinamento com 130 rebanhos negativos, garantindo a estatística equilibrada do grupo de rebanhos negativos e diminuiu o desvio padrão dos resultados. Por outro lado, a técnica de *oversampling* utilizada foi replicação. Foram selecionados 70 registros da classe minoritária e replicados, totalizando 140 rebanhos positivos para treinamento. O restante dos dados foi usado para conjunto de validação, o qual contou com 12 rebanhos da classe positivo para brucelose e 1973 da classe negativo para brucelose. Visando obter um modelo mais simplificado e acurado a RNA foi treinada com diferentes conjuntos de variáveis da base de dados. Lopes (2016) treinou o modelo com todas as variáveis do banco de dados (sem pré-seleção - 49 variáveis) e com pré-seleções das variáveis de maior relevância utilizando duas metodologias (Teste do Qui-quadrado de Pearson e FDR). Para sensibilidade todos os modelos demonstraram os melhores resultados igual a 83,33%. Para especificidade variou entre 53,89 e 66,60% e para eficiência de 68,61 a 74,96%, não havendo diferença significativa entre modelos. O modelo com método de seleção de variáveis pelo qui-quadrado atingiu o maior valor de especificidade, porém, a média foi significativamente semelhante a outros. Lopes (2016) concluiu que RNA para predição de rebanhos positivos e negativos para brucelose em veterinária é promissora e a pré-seleção de variáveis para buscar modelos mais simples deve ser utilizada, tendo em vista que a eficiência foi mantida.

De forma análoga, Keshavarzi et al. (2020) aplicou 2 técnicas de *undersampling* e *oversampling* no conjunto de dados de treinamento para criar 2 conjuntos de dados balanceados. O objetivo de Keshavarzi et al. (2020) foi prever a perda de gestação em rebanhos leiteiros iranianos utilizando algoritmos de AM. Os autores avaliaram o desempenho de 12 algoritmos. Para este propósito, foram coletados os dados de 6 grandes fazendas leiteiras comerciais com vacas paridas entre 2005 e 2014 no Irã. Os registros do histórico da vaca e informações genéticas do touro disponíveis foram extraídos de um software de manutenção de registros na fazenda. A base de dados contou então com 70130 registros de parto de 26289 vacas. Devido a taxa média de aborto de 15,4%, o conjunto de dados contava com desequilíbrio significante. Para solucionar o desequilíbrio Keshavarzi et al. (2020) fez uso do SpreedSunSample e do SMOTE. O SpreedSunSample foi adotado como técnica de *undersampling* para redução na classe majoritária, enquanto o SMOTE foi aplicado como técnica de *oversampling* para aumento da classe minoritária. Os algoritmos foram treinados usando os 4 conjuntos de dados: 1) o conjunto de dados incluindo informações de rebanho e vaca; 2) um conjunto de dados completo de rebanho e vaca, e informações genéticas de touro; 3) um conjunto de dados de *undersampling* com base em informações de rebanho e vaca; e 4) um conjunto de dados *oversampling* baseado em variáveis de rebanho e vaca. Keshavarzi et al. (2020) observou aumentos substanciais nas medidas de desempenho adotadas para os modelos nos conjuntos de dados balanceados. Apesar dos modelos de predição não terem alcançado o desempenho esperado por Keshavarzi et al. (2020), os modelos balanceados com método de *undersampling* superaram os demais, exibindo o melhor desempenho. Os autores concluíram que, apesar de sua precisão moderada, os modelos podem ser usados de forma benéfica para prever e reduzir as perdas de prenhez.

Martinez-Rau et al. (2022) empregaram o método de *oversampling* ADASYN, proposto por He et al. (2008), para criar eventos sintéticos adicionais para a classe minoritária visando controlar os desequilíbrios das classes de eventos durante a fase de treinamento do algoritmo de classificação. No estudo Martinez-Rau et al. (2022) aplicaram RNA MLP, aliado ao método acústico *Chew-Bite Energy Based Algorithm* (CBEBA), com intuito de detectar e classificar automaticamente os eventos mastigatórios de bovinos em pastejo. O modelo incorpora cálculos de sinal de potência instantânea para classificação de eventos de movimentos mandibulares associados a mastigações, mordidas e mastigações compostas. De forma adicional, o modelo classifica os eventos entre mastigações de baixa energia associadas à ruminação e mastigações de alta energia que estão associadas ao pastoreio. Os resultados demonstram que o modelo

atinge alta precisão de classificação, alcançando uma taxa de reconhecimento de 91,9% e 91,6% em condições silenciosas e ruidosas, respectivamente.

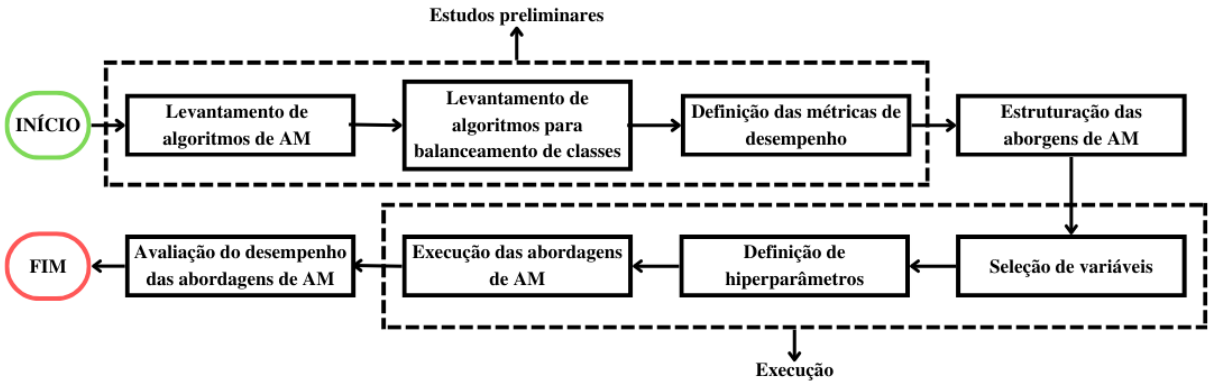
Semelhante a Martinez-Rau et al. (2022), Açıcı et al. (2019) superaram o problema do conjunto de dados desequilibrado gerando registros sintéticos através do algoritmo ADASYN. No trabalho Açıcı et al. (2019) empregaram uma arquitetura de Rede Neural Convolucional (CNN - do inglês *Convolutional Neural Ntwork*), para prever efetores IVA e IVB secretados pelo sistema de secreção tipo IV (T4SS). Presente em bactérias gram-negativas, o sistema de secreção tipo IV (T4SS) é um dos sistemas de secreção capaz de transportar proteínas e núcleoproteínas (proteína-proteína e proteína-DNA). As proteínas efetoras podem ser transportadas, através do T4SS, para o citoplasma de uma célula hospedeira diretamente. Esse transporte ocorre ao longo da membrana bacteriana e da membrana plasmática da célula hospedeira. CNN é uma classe de RNA utilizada especialmente em problemas de reconhecimento de imagem. As proteínas efetoras secretadas pelo T4SS são excretadas por vários patógenos nas células hospedeiras e geralmente são críticas para a virulência bacteriana. Através do T4SS as proteínas efetoras podem ser transportadas diretamente para o citoplasma de uma célula hospedeira, resultando em várias doenças. Para representar cada proteína efetora como uma imagem RGB, que posteriormente alimentaram a CNN, os autores aplicaram três métodos de extração de características, sendo eles: composições de aminoácidos, composições de dipeptídeos e matriz de pontuação específica de posição. A base de dados gerada contava com desequilíbrio muito grande entre a classe de proteínas efetoras e não-efetores, devido ao grande número de proteínas não efetoras em relação a proteínas efetoras. Portanto, para tornar o conjunto de dados mais equilibrado e fornecer uma fase de treinamento mais eficaz, Açıcı et al. (2019) criaram pseudo-proteínas para classe de proteínas efetoras com algoritmo ADASYN. O resultados mostraram que a utilização do algoritmo ADASYN para todos os métodos e cenários de representação de proteínas foi superior a todos os métodos sem ADASYN. Açıcı et al. (2019) demonstraram em seus resultados que a ferramenta criada previu todos os efetores IVA corretamente e, além disso, obteve desempenho de sensibilidade de 94,2% para predição do efetor IVB.

3 METODOLOGIA

Esse estudo aplica abordagens de AM combinando diferentes algoritmos de AM com técnicas de balanceamento de classe para a classificação de rebanhos como positivos e negativos para brucelose bovina. Para isso, utilizou-se os dados do inquérito de MG realizado de 2010 a 2012 pelo IMA/MAPA.

Esse capítulo descreve os materiais utilizados, a metodologia e as técnicas que foram empregadas na pesquisa e desenvolvimento do trabalho. O diagrama em blocos, ilustrado na Figura 3.1, resume as etapas desse estudo. Os três primeiros blocos compreendem aos estudos preliminares realizados com finalidade de identificar os algoritmos de AM, métricas de desempenho e outras técnicas mais recomendadas para problemas com desbalanceamento de classe. O próximo bloco corresponde a etapa de estruturação das abordagens, no qual foi codificado em linguagem de programação as abordagens de AM com base nos estudos preliminares. Os blocos de execução envolvem as etapas onde fez-se a seleção de variáveis, definição de hiperparâmetros e a execução em si das abordagens para coleta dos resultados das métricas de desempenho. Por fim, ilustrado no último bloco, foi feita a avaliação do desempenho de forma comparativa entre cada uma das abordagens estruturadas.

Figura 3.1 – Organograma das etapas percorridas na pesquisa e desenvolvimento desse estudo.



Fonte: Do autor (2024).

3.1 Ferramentas e *softwares*

O equipamento utilizado para a pesquisa foi um notebook de valor aproximado de R$5.400,00 com especificações capaz de suportar a execução dos *softwares* necessários para obtenção dos resultados. As especificações do notebook utilizado são: marca *Dell*, com processador *12th Gen Intel(R) Core(TM) i7-1255U 1.70 GHz*, gráficos *Intel(R) Iris(R) Xe Graphics*,

16GB de memória RAM, SSD de 512GB, GPU (*Graphics Processing Unit*, ou Unidade de Processamento Gráfico) *NVIDIA(R) GeForce(R) MX550 2GB*, sistema operacional *Windows 11 Home Single Language*.

Para implementação dos códigos foi utilizada a linguagem de programação *Python*[1], a qual é uma linguagem de programação gratuita e *open source*. As bibliotecas utilizadas neste trabalho são todas gratuitas e de código aberto, sendo elas:

- **Numpy**[2]**:** projeto que permite computação numérica com *Python*.
- **Pandas**[3]**:** ferramenta flexível de análise/manipulação de dados.
- **Scikit-Learn**[4]**:** biblioteca com conjunto de ferramentas simples e eficientes para análise preditiva de dados.
- **MatplotLib**[5]**:** biblioteca para criação de gráficos e visualizações de dados.
- **Seaborn**[6]**:** biblioteca que permite visualização de dados e desenho de gráficos estatísticos.
- **SciPy**[7]**:** possui diversos algoritmos fundamentais para computação científica em *Python*.
- **Imbalanced-learn**[8]**:** fornece ferramentas para problemas que envolvem classificação com classes desequilibradas.
- **OCPC Classifier**[9]**:** biblioteca com a implementação do classificador OC-PC.

3.2 Dados

A base de dados utilizada nesse trabalho é fruto do trabalho de Lopes (2016) e corresponde aos dados do inquérito de MG realizado de 2010 a 2012. O inquérito foi realizado pela equipe do serviço oficial de defesa sanitária animal do Estado de Minas Gerais (IMA), com o

[1] https://www.python.org
[2] https://numpy.org
[3] https://pandas.pydata.org
[4] https://scikit-learn.org/stable/index.html
[5] https://matplotlib.org
[6] https://seaborn.pydata.org
[7] https://scipy.org
[8] https://imbalanced-learn.org/stable
[9] https://pypi.org/project/ocpc-py

apoio da Coordenação Nacional do Programa Nacional de Controle e Erradicação da Brucelose e da Tuberculose Animal (PNCEBT) e com o suporte do Centro Colaborador do MAPA em saúde animal da FMVZ-USP e FAMV-UnB. As propriedades sorteadas foram visitadas no período de setembro de 2010 a dezembro de 2012 para preenchimento do formulário (entrevista) e coleta do sangue para o diagnóstico da Brucelose, em todo Estado de Minas Gerais. As informações sobre a coleta dos dados e diagnóstico ocorreram de acordo com o Manual de Procedimentos do estudo epidemiológico sobre Brucelose e Tuberculose em Bovinos e Bubalinos, como descrito por Oliveira et al. (2016).

O banco contendo os dados do inquérito de brucelose do Estado de Minas Gerais foi fornecido em forma de planilha do software Excel e foi composto pelas variáveis coletadas por meio dos questionários e pelos resultados dos testes de diagnóstico para a doença. Os casos e controles de Brucelose (propriedades com rebanho positivo e propriedades com rebanho negativo) foram consideradas como variáveis dependentes, e as coletadas por meio das entrevistas, relacionadas às características e ao manejo geral e sanitário da propriedade, como variáveis independentes. Realizou-se uma série de pré-processamentos na base de dados, como organização das variáveis e códigos, execução de transformações/agregações/recodificações das variáveis, verificação de consistência, análise descritiva preliminar, avaliação de *outliers*, e análise de associação, descritos em detalhes por Lopes (2016). Após pré-processamento o banco de dados passou a contar com dados de 2.185 propriedades (ou amostra) e 49 variáveis para cada propriedade. As 49 variáveis são: região, tipo de exploração, tipo de criação, número ordenha por dia, tipo de ordenha, realiza inseminação, raça predominante de bovinos, tamanho do rebanho (n), presença de ovinos e caprinos, presença de equinos, presença de suínos, presença de aves, presença de cão, presença de gato, presença de espécies silvestres, presença de cervídeos, presença de capivaras, presença de outras espécies silvestres, presença de aborto nos últimos 12 meses, destino do feto abortado e a placenta, realiza teste de brucelose, regularidade do teste brucelose, compra reprodutor, compra reprodutor em exposição, compra reprodutor em leilão, compra reprodutor de comerciante, compra reprodutor de fazendas, vende reprodutor, vende reprodutor em exposição, vende reprodutor em leilão, vende reprodutor a comerciante, vende reprodutor a fazendas, vacina contra brucelose, local de abate de reprodutor, aluga pasto, pasto em comum com outras propriedades, presença de áreas alagadas, presença de piquete de parição, destino do leite, resfria leite, como resfria leite, entrega leite a granel, produz queijo ou manteiga, finalidade do queijo ou manteiga produzido, consome leite cru, presença de assistên-

cia veterinária, tipo de assistência veterinária, compartilha aguadas/bebedouro com animais de outra propriedade e classificação da propriedade.

Lopes (2016) investigou diferentes formas de seleção de variáveis e comparou o desempenho de RNA com cada conjunto de variáveis selecionadas. As seleções de variáveis foram embasadas em Análise Univariada e FDR. A maioria das formas de seleção de variáveis avaliadas proporcionaram resultados preditivos sem diferença significativa entre as médias. Porém, através do FDR, foi possível selecionar apenas 10 variáveis sem comprometer os resultados preditivos da RNA utilizada.

O problema de identificação de risco de Brucelose Bovina a partir dos dados descritos é de alta complexidade, conforme relatado por Lopes (2016), e também é um problema de reconhecimento de padrões de alta dimensão. Dessa forma, decidiu-se, no contexto dessa Dissertação de Mestrado, considerar apenas as 10 variáveis elencadas por Lopes (2016) como base de dados para os estudos aqui relatados, e dar um foco maior nas estratégias de *machine learning* para lidar com dados desbalanceados. As 10 variáveis selecionadas, listadas em ordem decrescente de relevância de acordo com o FDR, são: tamanho do rebanho (n), presença de piquete de parição, realiza teste de brucelose, tipo de assistência veterinária, vacina contra brucelose, presença de assistência veterinária, vende reprodutor em leilão, compra reprodutor, vende reprodutor e realiza inseminação.

A base de dados possui desequilíbrio entre as classes de propriedades com diagnósticos positivos e propriedades com diagnósticos negativos muito marcante, conforme ilustrado na Figura 3.2. Das 2.185 propriedades da base de dados, 2.103 possuem diagnósticos negativos, enquanto que, apenas 82 possuem diagnósticos positivos para Brucelose. De acordo com Puri e Kumar Gupta (2021), o desequilíbrio de classe é representado pela razão de desequilíbrio de classe denotado por IR (do inglês *Class Imbalance Ratio*). O IR é dado pela razão entre a classe majoritária e a classe minoritária conforme Equação (3.1). Assim, o IR para base de dados em questão é 25,6. Segundo Leevy et al. (2018), do ponto de vista popular defendido por pesquisadores, define-se dados com um desequilíbrio de classe alto quando o IR varia de 100 a 10.000. Porém, ainda de acordo com Leevy et al. (2018), embora esta gama de desequilíbrio de classe possa ser observada em grandes volumes de dados, não é uma definição estrita de desequilíbrio de classe alta. Do ponto de vista de resolução dos problemas, qualquer nível de desequilíbrio de classe que torne a tarefa de classificação complexa e desafiadora, pode ser considerado desequilíbrio de classe alto.

$$IR = \frac{\text{nº de amostras da classe majoritária}}{\text{nº de amostras da classe minoritária}} \tag{3.1}$$

Figura 3.2 – Esquema de seleção de propriedades com técnica de balanceamento ADASYN.

Fonte: Do autor (2024).

3.2.1 Codificação e Normalização

As variáveis qualitativas foram codificadas atribuindo a elas valores numéricos padronizados de acordo com as respostas dos questionários. As variáveis quantitativas foram normalizadas (padronizadas em uma dada escala numérica) de acordo com a Equação (3.2).

$$X_i = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)} \tag{3.2}$$

em que X_i representa o valor da i-ésima variável da amostra X do conjunto de dados. $\min(X_i)$ e $\max(X_i)$ representam, respectivamente, os valores mínimo e máximos da i-ésima variável dentre todas as amostras da base de dados.

A normalização dessas variáveis é importante para evitar retardos no aprendizado da RNA causados pela saturação de neurônios, em decorrência de valores elevados de variáveis.

3.3 Aprendizado de Máquina (AM)

Neste trabalho foram abordados quatro algoritmos de AM frente a predição de Brucelose em rebanhos bovinos, sendo eles: RNA, k-NN, OC-SVM e OC-PC. A RNA foi empregada por

ser um dos algoritmos com maior capacidade de generalização e poder de classificação em dados complexos. O k-NN foi utilizado por não requerer treinamento, o que pode minimizar os efeitos causados pelo desequilíbrio entre as classes. As abordagens OCC aqui propostas (OC-SVM e OC-PC) foram no sentido de contornar os efeitos do desequilíbrio entre as classes.

Para os algoritmos RNA e k-NN empregou-se ainda duas abordagens de *Oversampling* a fim de minimizar o desequilíbrio dos dados de treinamento. Os algoritmos de *Oversampling* avaliados foram: replicação (CASTRO; BRAGA, 2011) e o ADASYN.

3.3.1 *Oversampling*

Das 82 propriedades positivas foram selecionadas aleatoriamente 60 para que, em conjunto com as 2103 propriedades negativas, fossem expostas ao ADASYN. O intuito foi gerar aproximadamente 180 propriedades positivas sintéticas. Para que isto fosse possível configurou-se o ADASYN com valor de $\beta = 0,12$. β é um parâmetro usado para especificar o nível de equilíbrio após a geração dos dados sintéticos, onde, $\beta \in [0,1]$, e $\beta = 1$ denota que um conjunto de dados totalmente balanceado será criado após o processo de generalização. Optou-se por utilizar o valor de β consideravelmente baixo visto que o aumento desse valor provocou redução acentuada no desempenho dos algoritmos de AM. Definiu-se também o valor de $K = 7$. K corresponde a quantidade de vizinhos mais próximos, com base na distância euclidiana no espaço de variáveis d-dimensional, que o algorítimo considerou para aplicar os cálculos necessários. Os valores de β e K foram definidos experimentalmente, através de variações em seus valores e observação dos resultados dos algoritmos de AM para a base de dados de treinamento.

A técnica de replicação de dados constitui-se da replicação dos dados de 60 propriedades positivas selecionadas aleatoriamente.

As 22 propriedades positivas que não participaram das técnicas de *oversampling* foram destinadas para compor o conjunto de dados da validação da abordagem de AM.

3.3.2 *Undersampling*

Mesmo aplicando a técnica de *oversampling* como ADASYN e de replicação de dados para amenizar o desequilíbrio das classes, a base de dados ainda contava com um desequilíbrio considerável entre as classes. Dessa forma, adotou-se técnica de *undersampling* para a classe de rebanhos negativos. A técnica de *undersampling* consistiu de divisão das propriedades negativas em 4 agrupamentos, de forma que, para compor o conjunto de amostras negativas usadas para

projetar (ou treinar) os modelos de AM eram selecionadas (aleatoriamente) amostras em cada um dos 4 grupos. Optou-se a divisão por quatro grupos, em razão da melhor distribuição dos dados nos mesmos, com equilíbrio do número de propriedades em cada grupo. Essa divisão foi alcançada usando o algoritmo k-means.

O algoritmo k-means foi proposto há mais de 50 anos, todavia, é um dos algoritmos mais usados para agrupamento. É um algoritmo simples e eficiente que na maioria dos problemas converge num número reduzido de iterações (JAIN, 2010; EVSUKOFF, 2020). A forma como é mais utilizado atualmente foi proposta por MacQueen (1967).

O k-means encontra a melhor divisão de um conjunto de dados em Q grupos (para este trabalho $Q = 4$), de maneira que a distância total entre os dados de um grupo e o seu respectivo centro, somada por todos os grupos, seja minimizada. O critério de agrupamento do k-means é descrito pela Equação (3.3).

$$E = \sum_{q=1}^{Q} \sum_{X_j \in G_q} d(X_i, X_{0q}), \tag{3.3}$$

sendo que, Q corresponde ao número de agrupamentos, X_{0q} é o centroide do agrupamento G_q e $d(X_i, X_{0q})$ é a distância entre os pontos X_i e X_{0q}.

A opção de dividir os dados de rebanhos negativos em 4 grupos e escolher rebanhos em cada um dos grupos foi adotada para minimizar possíveis vieses inseridos ao se escolher aleatoriamente uma minoria de rebanhos negativos.

3.3.3 RNA

Foram projetadas três diferentes abordagens com RNA, sendo elas: RNA com algoritmo de *oversampling* ADASYN (RNA+ADASYN), RNA com replicação de dados como técnica de *oversampling* (RNA+REPLICAÇÃO) e RNA sem utilização de técnica *oversampling* (RNA BASE).

Para as abordagens que empregaram o algoritmo ADASYN, o conjunto de treinamento contou com 240 propriedades positivas (180 sintéticas e 60 selecionadas aleatoriamente da base de dados) e 240 propriedades negativas. As 240 propriedades negativas foram concebidas através da seleção aleatória de 80 propriedades de cada um dos 4 grupos gerados pelo algoritmo k-means.

Nos casos em que a técnica de replicação dos dados foi empregada, o conjunto de treinamento contou com 120 propriedades positivas (60 propriedades que foram replicadas) e 120 propriedades negativas. As 120 propriedades negativas foram obtidas através da seleção aleatória de 30 propriedades de cada um dos 4 grupos gerados pelo algoritmo k-means.

Para fins de comparação, foi desenvolvida abordagem em que não foi aplicado nenhum dos algoritmos de *oversamplig* para criação do conjunto de treinamento. Sendo assim, o conjunto de treinamento contou com 64 propriedades positivas selecionadas aleatoriamente e 64 propriedades negativas. As 64 propriedades negativas foram obtidas através da seleção aleatória de 16 propriedades de cada um dos 4 grupos gerados pelo algoritmo k-means.

Para todas as abordagens, os dados que não participaram do treinamento da RNA foram usados para validação. A Figura 3.3 esquematiza as estruturas das três abordagens de AM criadas com RNA.

3.3.4 k-NN

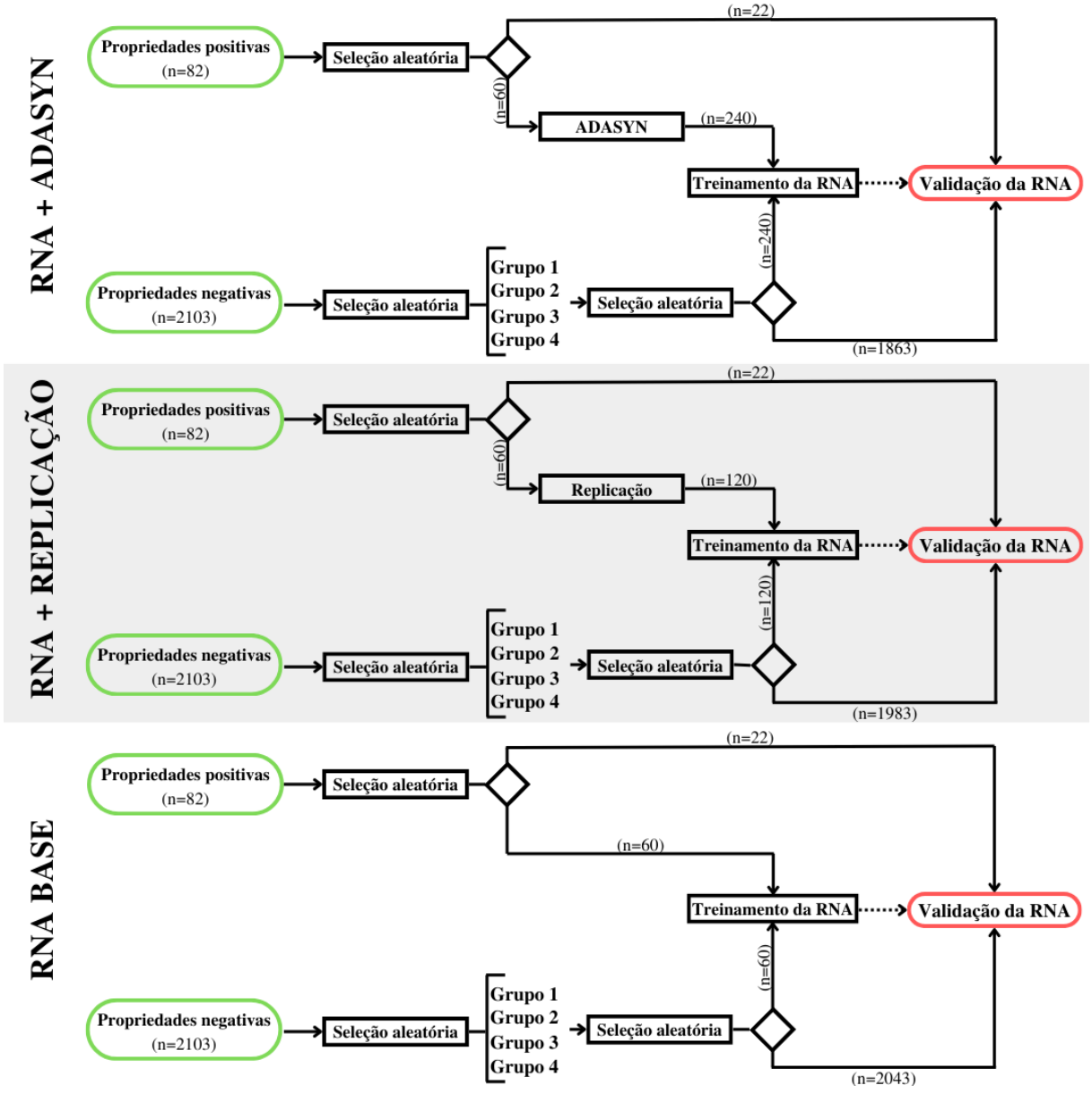
Seguindo a ideia das abordagens com RNA, foram estruturadas três diferentes abordagens com k-NN, sendo elas: k-NN com algoritmo de *oversampling* ADASYN (k-NN+ADASYN), k-NN com replicação dados como técnica de *oversampling* (k-NN+REPLICAÇÃO) e RNA sem utilização de técnica *oversampling* (k-NN BASE).

Nas abordagens que empregaram o algoritmo ADASYN, o conjunto de projeto (base de dados com rótulos) do k-NN contou com 240 propriedades positivas (180 sintéticas e 60 selecionadas aleatoriamente da base de dados) e 240 propriedades negativas. As 240 propriedades negativas foram concebidas através da seleção aleatória de 80 propriedades de cada um dos 4 grupos gerados pelo algoritmo k-means.

Para os casos que empregaram a técnica de replicação dos dados, o conjunto de projeto contou com 120 propriedades positivas (60 propriedades que foram replicadas) e 120 propriedades negativas. As 120 propriedades negativas foram obtidas através da seleção aleatória de 30 propriedades de cada um dos 4 grupos gerados pelo algoritmo k-means.

Para os modelos em que não foi aplicado nenhum dos algoritmos de *oversamplig* para criação, o conjunto de projeto contou com 64 propriedades positivas selecionadas aleatoriamente e 64 propriedades negativas. As 64 propriedades negativas foram obtidas através da seleção aleatória de 16 propriedades de cada um dos 4 grupos gerados pelo algoritmo k-means.

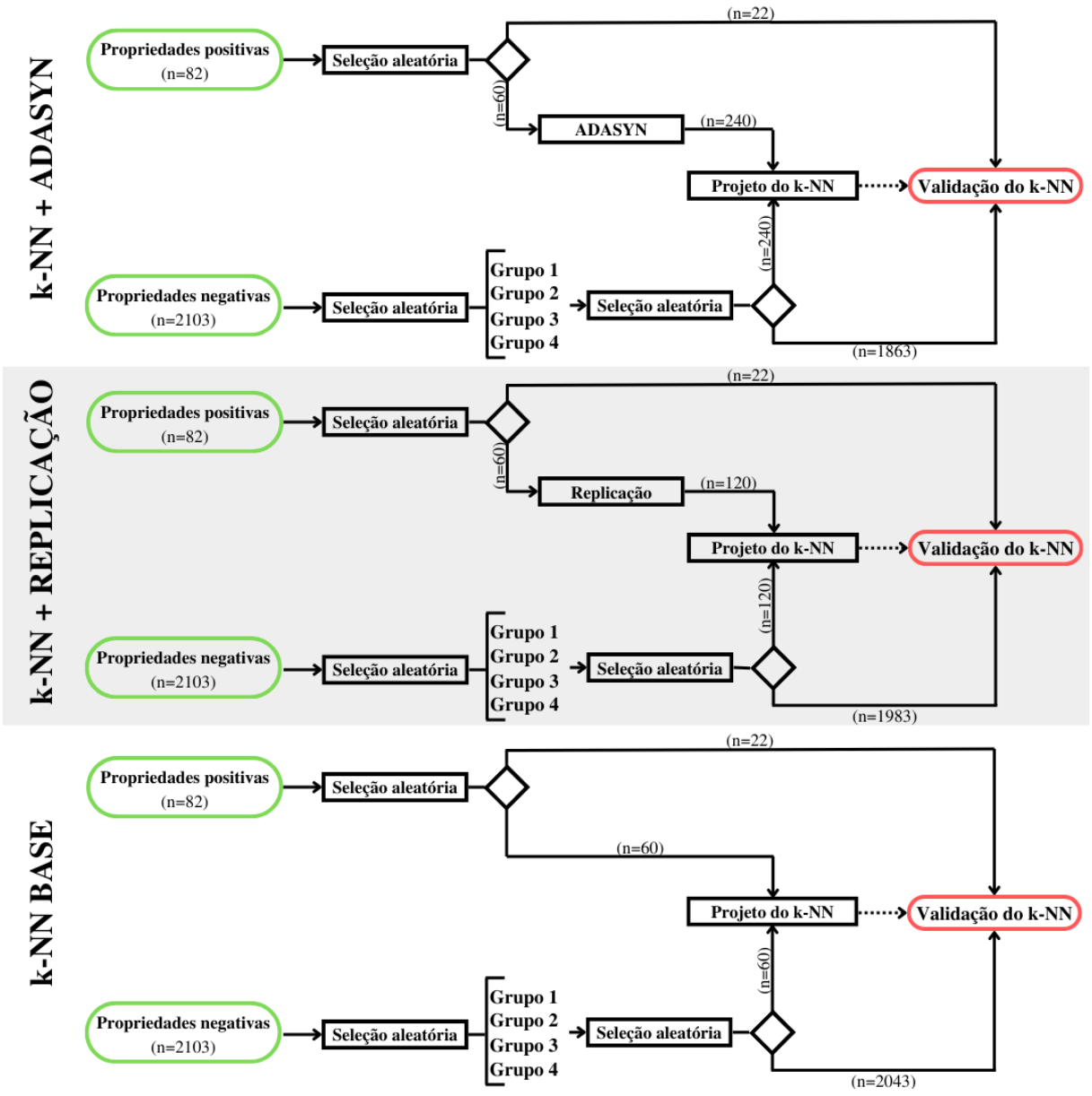
Figura 3.3 – Esquemas das estruturas das três abordagens de AM criadas com RNA.



Fonte: Do autor (2024).

Para todas as abordagens, as amostras que não fizeram parte do projeto do k-NN foram usadas para validação. A Figura 3.4 esquematiza as estruturas das três abordagens de AM criadas com k-NN.

Figura 3.4 – Esquemas das estruturas das três abordagens de AM criadas com k-NN.



Fonte: Do autor (2024).

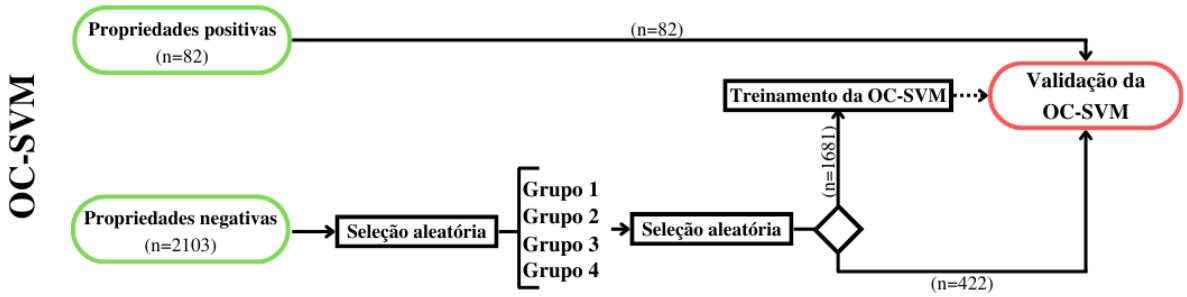
3.3.5 OC-SVM

O treinamento do classificador OC-SVM foi realizado adotando-se como classe alvo a classe de propriedades negativas e a classe de propriedades positivas foi definida como a classe *outlier*. Dessa forma, não foi necessário aplicar técnicas de *oversampling* para projetar o

classificador OC-SVM, visto que, todas as amostras de propriedades positivas foram utilizadas apenas na fase de validação. Foi criado então apenas uma abordagem utilizando OC-SVM.

Foram selecionadas 80% das amostras de cada um dos 4 grupos gerados pelo algoritmo k-means para compor o conjunto de treinamento do OC-SVM, totalizando 1861 propriedades. O restante das amostras de cada grupo (20%), juntamente com todas as amostras da classe positiva, foram utilizadas para validação do OC-SVM. A Figura 3.5 esquematiza a estrutura da abordagem de AM criada com OC-SVM.

Figura 3.5 – Esquema da estrutura da abordagem de AM criada com OC-SVM.



Fonte: Do autor (2024).

3.3.6 OC-PC

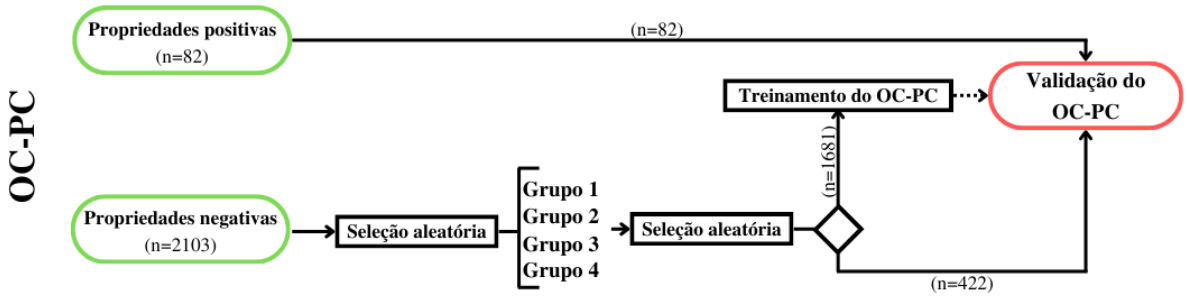
A abordagem criada utilizando o algoritmo OC-PC seguiu a ideia aplicada na abordagem criada com OC-SVM, resultando também na criação de apenas uma abordagem com OC-PC. Adotou-se como classe alvo a classe de propriedades negativas e a classe de propriedades positivas foi definida como a classe *outlier*.

Da mesma forma como foi feito para a abordagem com OC-SCM, foram selecionadas 80% das amostras de cada um dos 4 grupos gerados pelo algoritmo k-means para compor a conjunto de treinamento da OC-SVM, totalizando 1861 propriedades. Os 20% restante das amostras de cada grupo, juntamente com todas as amostras da classe positiva, foram utilizadas para validação do OC-PC. A Figura 3.6 esquematiza a estrutura da abordagem de AM criada com OC-PC.

3.4 Abordagens

Ao total foram estruturadas oito abordagens a fim de se comparar o desempenho, sendo elas:

Figura 3.6 – Esquema da estrutura da abordagem de AM criada com OC-PC.



Fonte: Do autor (2024).

1. **RNA+ADASYN** - RNA com técnica ADASYN;
2. **RNA+REP** - RNA com técnica de replicação de dados;
3. **RNA-BASE** - RNA sem aplicação de técnica para *oversampling*;
4. **k-NN+ADASYN** - k-NN com técnica ADASYN;
5. **k-NN+REP** - k-NN com técnica de replicação de dados;
6. **k-NN-BASE** - k-NN sem aplicação de técnica para *oversampling*;
7. **OC-SVM**;
8. **OC-PC**.

Vale ressaltar que, aferiu-se também o desempenho dos classificadores OC-SVM e OC-PC quando adotou-se a classe de propriedades negativas classe alvo. Porém, essas abordagens atingiram desempenhos inferiores quando comparados com a adoção da classe de propriedades negativas como classe alvo. Dessa forma, apenas as abordagens que adotaram como classe alvo a classe de propriedades negativas e a classe de propriedades positivas como a classe *outlier* foram comparadas com as demais abordagens dos outros classificadores.

3.5 Definição de hiperparâmetros

O ajuste de hiperparâmetros (parâmetros que especificam detalhes do processo de aprendizado dos algoritmos de AM) é o processo de identificação do conjunto de valores de hiperparâmetros que se espera produzir o melhores resultados com o algoritmo de AM em questão. Nesse trabalho os hiperparâmetros de cada uma das abordagens foram definidos via técnica de

otimização de hiperparâmetros exaustiva *Grid Search* (HSU; CHANG; LIN, 2003). A técnica de *Grid Search* itera o modelo de AM em todas as combinações possíveis de hiperparâmetros em produtos cartesianos a partir de um conjunto finito de valores definidos pelo usuário.

O conjunto de valores para cada hiperparâmetros (avaliados com a técnica de *Grid Search*) para abordagens que utilizaram RNA, k-NN, OC-SVM e OC-PC, estão dispostos, respectivamente, nas Tabelas 3.1, 3.2, 3.3 e 3.4.

Tabela 3.1 – Conjunto de valores para cada hiperparâmetros (avaliados com a técnica de *Grid Search*) para abordagens que utilizaram RNA.

Hiperparâmetros	**Valores**
Nº de neurônios na camada intermediária	[5, 10, 15, 20, 30]
Função de ativação	[Sigmoide, Tangente Hiperbólica, Relu]

Fonte: Do autor (2024).

Tabela 3.2 – Conjunto de valores para cada hiperparâmetros (avaliados com a técnica de *Grid Search*) para abordagens que utilizaram algoritmo k-NN.

Hiperparâmetros	**Valores**
k	[3, 4, 5, 6, 7, 8, 9, 10]
Medidas de distâncias	[Euclidiana, Manhattan]

Fonte: Do autor (2024).

Tabela 3.3 – Conjunto de valores para cada hiperparâmetros (avaliados com a técnica de *Grid Search*) para abordagens que utilizaram algoritmo OC-SVM.

Hiperparâmetros	**Valores**
kernel	['Linear', 'Polinomial', 'RBF', 'Sigmoide']
ν	[0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7]

Fonte: Do autor (2024).

A técnica de *Grid Search* foi executada 50 vezes para cada abordagem. Dessa forma, cada uma das abordagens foram executadas 50 vezes para cada uma das combinações possíveis de hiperparâmetros. Por fim, a combinação que obteve maior média para a métrica IBA foi adotada como hiperparâmetros da respectiva abordagem.

Os hiperparâmetros que não participaram da técnica de *Grid Search* foram fixos para todas as abordagens de cada classe de algoritmo. Em todas as abordagens que empregaram RNA adotou-se uma camada intermediária e um único neurônio na camada de saída. O algoritmo *Stochastic Gradient Descent* (SGD) com taxa de aprendizado constante (0,001) foi empregado

Tabela 3.4 – Conjunto de valores para cada hiperparâmetros (avaliados com a técnica de *Grid Search*) para abordagens que utilizaram algoritmo OC-PC.

Hiperparâmetros	**Valores**
kmax	[4, 5, 6, 7, 8, 9]
OR	[0.05, 0.10, 0.30, 0.50, 0.70]

Fonte: Do autor (2024).

para o treinamento das RNAs. O SGD apresentou bons resultados para a base de dados utilizada e por esse motivo optou-se por usá-lo. O treinamento foi encerrado quando o número máximo de 10000 épocas foi atingido ou quando o erro não decaiu em pelo menos 0,0001 por 10 épocas consecutivas.

Para a abordagem via OC-SVM adotou-se $\gamma = \frac{1}{\text{Nº de variáveis}}$ para os *Kernels* RBF, Polinomial e Sigmoide, e $r = 0$ para os *Kernels* Polinomial e Sigmoide. O treinamento foi encerrado quando o número máximo de 10000 iterações foi atingido ou quando o erro atingiu valor menor ou 0,001.

3.6 Comparação entre as abordagens de AM

A metodologia de seleção aleatória de amostras, treinamento e validação das oito abordagens, conforme ilustrado nas Figuras 3.3, 3.4, 3.5 e 3.6, foi realizada 100 vezes. Portanto, para cada abordagem foi possível efetuar a coleta de resultados para diferentes métricas em cada uma das 100 execuções. Com o objetivo de comparar os resultados das diferentes abordagens adotou-se o Teste Tukey (com 95% de intervalo de confiança) como teste de comparação de médias.

As métricas utilizadas para avaliação das abordagens foram: Sensibilidade (ou *Recall*), Especificidade, Precisão (ou Valor Preditivo Positivo), *F-measure*, G-mean, e IBA. Para a métrica *F-measure* adotou-se $\beta = 2$. Adotou-se $M = \text{G-mean}^2$ e $\alpha = 0.1$ como valores dos parâmetros da métrica IBA. Com esses respectivos valores de parâmetros, as métricas *F-measure* e IBA foram definidas de acordo com as Equação (3.4) e Equação (3.5), respectivamente.

$$F\text{-}measure = \frac{5 \cdot \text{Precisão} \cdot \text{Sensibilidade}}{4 \cdot \text{Precisão} + \text{Sensibilidade}} \tag{3.4}$$

$$\text{IBA}_{\alpha}(\text{G-mean}^2) = (1 + \alpha \cdot (Sensibilidade - Especificidade)) \cdot Sensibilidade \cdot Especificidade \tag{3.5}$$

A métrica IBA busca um compromisso entre uma determinada métrica e um índice de quão equilibrados são os valores de sensibilidade e especificidade. A métrica IBA tende favorecer modelos com melhores resultados na classe positiva. Dessa forma, penaliza abordagens que com valores de especificidade maiores que sensibilidade, e valoriza abordagens com valores de sensibilidade maiores que especificidade.

4 RESULTADOS E DISCUSSÃO

4.1 Hiperparâmetros

Para definição dos hiperparâmetros de cada abordagem de AM desenvolvida foi utilizada a técnica de otimização de hiperparâmetros *Grid Search*. A combinação de hiperparâmetros de cada abordagem que possibilitou obtenção da melhor média de IBA nas 50 execuções está disposta na Tabela 4.1. Essas combinações de hiperparâmetros, somadas aos hiperparâmetros fixos para todas as abordagens de cada modelo, foram então adotadas para comparação das abordagens.

Tabela 4.1 – Combinação de hiperparâmetros (para cada abordagem), obtidos durante o *Grid Search*, que possibilitou maior média (μ) de IBA e os respectivos desvios padrão (σ).

Abordagens	**Hiperparâmetros**	**IBA**	
		μ	σ
RNA+ADASYN	Func. Ativ. = Tan. Hip., Nº Neur. = 15	0,224	0,085
RNA+REP	Func. Ativ. = Tan. Hip., Nº Neur. = 10	0,231	0,065
RNA-BASE	Func. Ativ. = Tan. Hip., Nº Neur. = 30	0,226	0,073
k-NN+ADASYN	k = 7, Dist. = Euc.	0,281	0,061
k-NN+REP	k = 7, Dist. = Euc.	0,264	0,044
k-NN-BASE	k = 9, Dist. = Man.	0,289	0,064
OC-SVM	kernel = RBF, ν = 0,5	0,352	0,015
OC-PC	kmax = 5, OR = 0.5	0,372	0,014

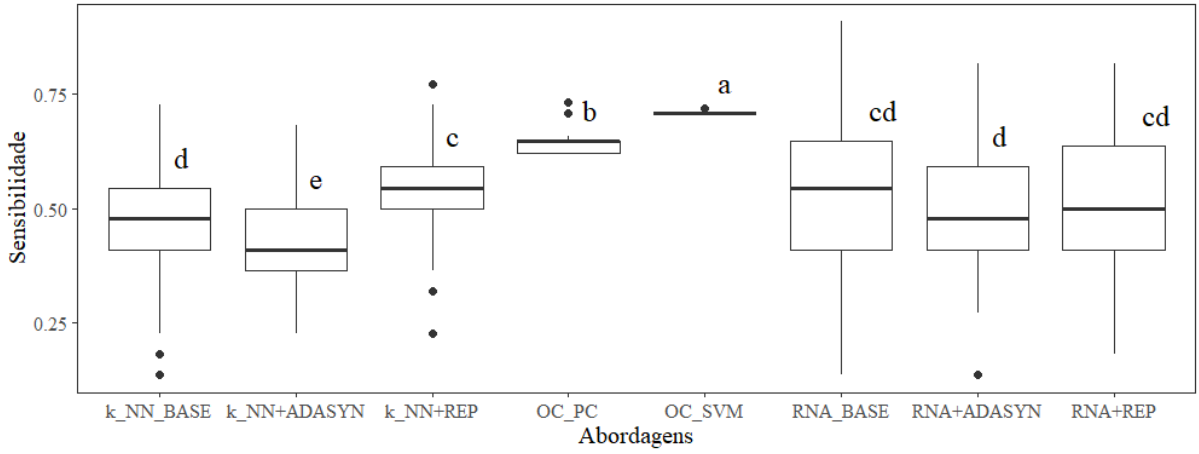
Fonte: Do autor (2024).

4.2 Análise das abordagens

Nos últimos anos presenciou-se um crescimento exponencial nas pesquisas envolvendo problemas de classificação em ambientes com classes desbalanceadas. Esse crescimento é fruto principalmente da democratização de algoritmos de inteligencia computacional. A aplicação desses algoritmos em problemas envolvendo questões de saúde geralmente esbarra em dados com desbalanceamento de classe. Isso deve-se ao fato que, em problemas envolvendo saúde, principalmente quando envolve doenças raras, os dados da classe de indivíduos negativos são consideravelmente mais numerosos. As pesquisas desenvolvidas nesse âmbito propõem diferentes condutas frente a cada desafio. Essas diferentes condutas são geralmente categorizadas em: nível de dados, nível de algoritmo e híbridas. Nesse trabalho foram empregadas diferentes condutas de diferentes categorias. O objetivo foi comparar e identificar as abordagens que obtiveram melhores desempenhos na base de dados utilizada.

A fim de se comparar o desempenho de cada abordagem frente ao desafio imposto pelo desequilíbrio entre as classe, os resultados obtidos para o conjunto de validação durante as 100 execuções de cada abordagem foram salvos. Os valores obtidos para Sensibilidade, Especificidade, Precisão, *F-measure*, G-mean, e IBA podem ser visualizados, no formato de Box Plot, respectivamente, nas Figuras 4.1, 4.2, 4.3, 4.4, 4.5 e 4.6. Nas Tabelas 4.2, 4.3 e 4.4 são apresentadas as média e desvios padrão do desempenho de cada abordagem de AM obtido durante as 100 execuções. Exceto para especificidade, as abordagens via OC-PC e OC-SVM foram significantemente (Intervalo de Confiança (IC) de 95%) superiores em todas as métricas.

Figura 4.1 – Resultados de sensibilidade obtidos no conjunto de validação pelas respectivas abordagens durante as 100 execuções.



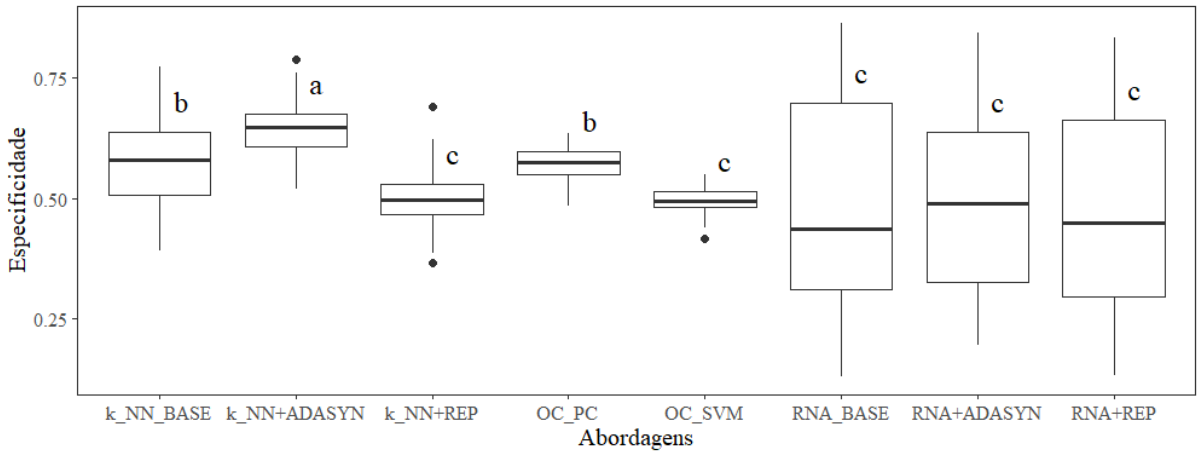
Legenda: Diferentes letras representam médias com diferenças significativas no Teste Tukey com nível de confiança de 95%.

Fonte: Do autor (2024).

Para sensibilidade, as abordagens OC-PC e OC-SVM obtiveram melhores resultados (Figura 4.1), sendo a OC-SVM a abordagem que melhor desempenhou nessa métrica, obtendo valor médio de 0,709. Apesar da OC-PC ter atingido média superior as demais abordagens, seu desempenho foi inferior à OC-SVM, resultando em uma média para sensibilidade de 0,653. Das demais abordagens, a maior média de desempenho para sensibilidade foi alcançada pela abordagem k-NN-REP, com um valor de 0,545 (Tabela 4.2). O restante das abordagens desempenharam para sensibilidade sem diferença significativa, no qual o pior resultado foi obtido pela abordagem k-NN-ADASYN (Figura 4.1).

A maioria das abordagens atingiram valores altos de desvios padrão (Tabela 4.2) para sensibilidade. Desvios padrão altos podem indicar maior dispersão da classe no espaço de características, ou seja, são geralmente encontrados para classes mais heterogêneas. Por outro

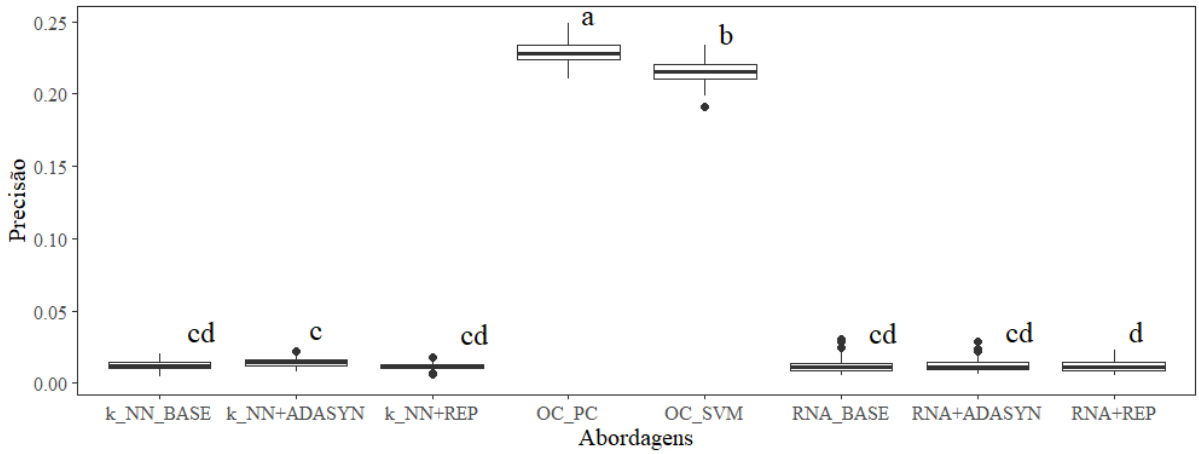
Figura 4.2 – Resultados de especificidade obtidos no conjunto de validação pelas respectivas abordagens durante as 100 execuções.



Legenda: Diferentes letras representam médias com diferenças significativas no Teste Tukey com nível de confiança de 95%.

Fonte: Do autor (2024).

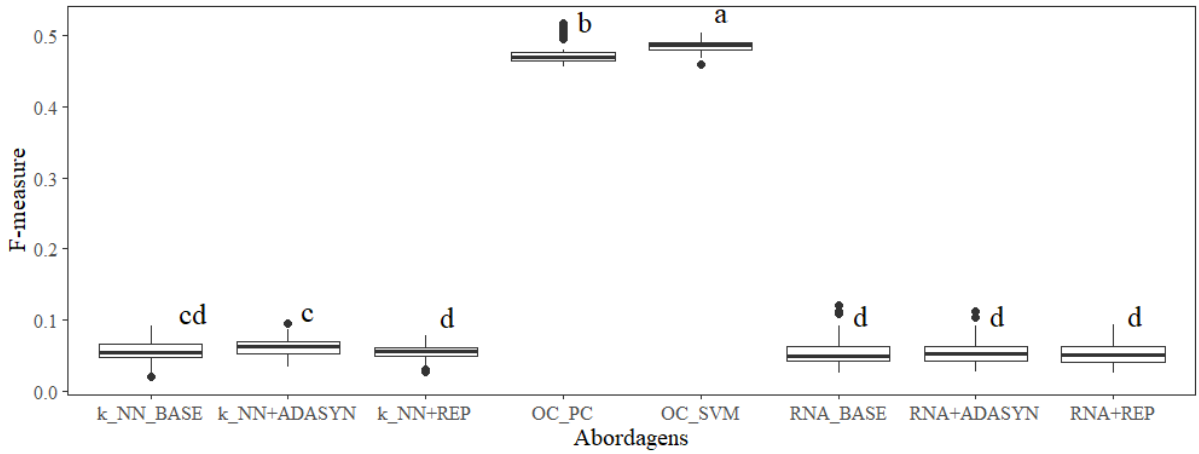
Figura 4.3 – Resultados de precisão obtidos no conjunto de validação pelas respectivas abordagens durante as 100 execuções.



Legenda: Diferentes letras representam médias com diferenças significativas no Teste Tukey com nível de confiança de 95%.

Fonte: Do autor (2024).

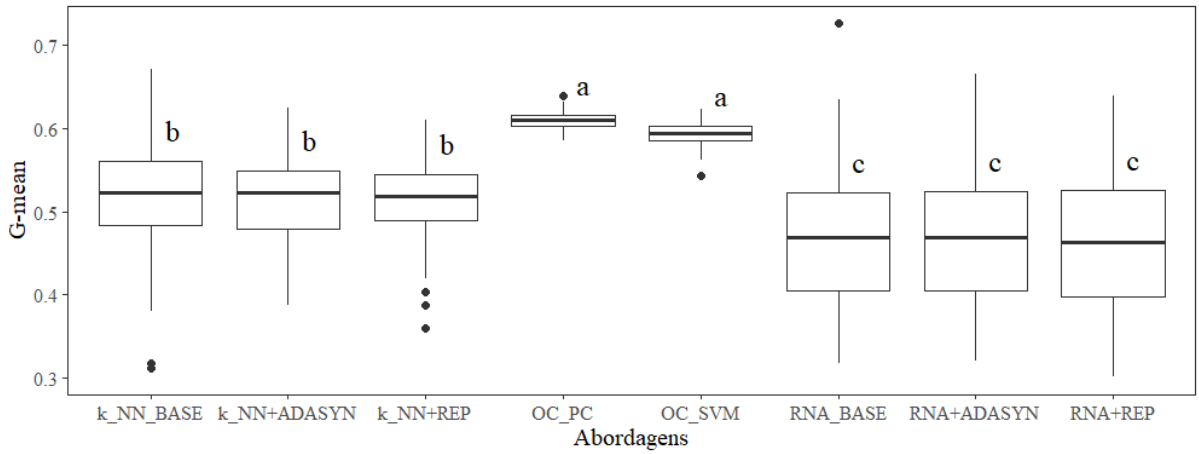
Figura 4.4 – Resultados de *F-measure* obtidos no conjunto de validação pelas respectivas abordagens durante as 100 execuções.



Legenda: Diferentes letras representam médias com diferenças significativas no Teste Tukey com nível de confiança de 95%.

Fonte: Do autor (2024).

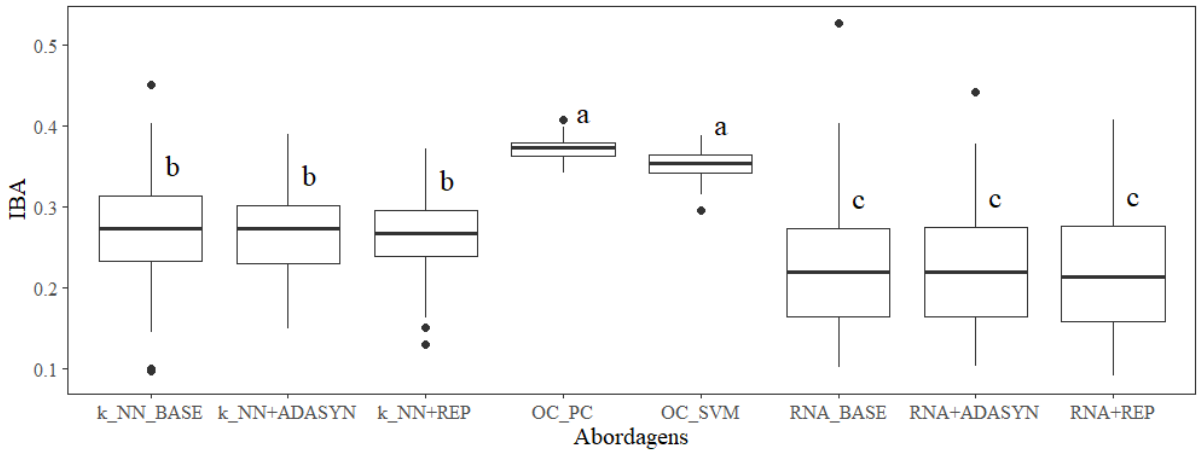
Figura 4.5 – Resultados de G-mean obtidos no conjunto de validação pelas respectivas abordagens durante as 100 execuções.



Legenda: Diferentes letras representam médias com diferenças significativas no Teste Tukey com nível de confiança de 95%.

Fonte: Do autor (2024).

Figura 4.6 – Resultados de IBA obtidos no conjunto de validação pelas respectivas abordagens durante as 100 execuções.



Legenda: Diferentes letras representam médias com diferenças significativas no Teste Tukey com nível de confiança de 95%.

Fonte: Do autor (2024).

Tabela 4.2 – Média (μ) e Desvio Padrão (σ) das métricas de sensibilidade e especificidade obtidos por cada abordagem durante as 100 execuções.

Abordagens	**Sensibilidade** $\mu \pm \sigma$	**Especificidade** $\mu \pm \sigma$
RNA+ADASYN	0,490 ± 0,126 **d**	0,482 ± 0,181 **c**
RNA+REP	0,507 ± 0,149 **cd**	0,479 ± 0,197 **c**
RNA-BASE	0,525 ± 0,174 **cd**	0,486 ± 0,216 **c**
k-NN+ADASYN	0,419 ± 0,096 **e**	0,648 ± 0,055 **a**
k-NN+REP	0,545 ± 0,100 **c**	0,497 ± 0,054 **c**
k-NN-BASE	0,480 ± 0,118 **d**	0,579 ± 0,088 **b**
OC-SVM	0,709 ± 0,004 **a**	0,497 ± 0,024 **c**
OC-PC	0,653 ± 0,038 **b**	0,571 ± 0,034 **b**

Legenda: Diferentes letras representam médias com diferenças significativas no Teste Tukey com nível de confiança de 95%.

Fonte: Do autor (2024).

Tabela 4.3 – Média (μ) e Desvio Padrão (σ) das métricas de precisão e *F-measure* obtidos por cada abordagem durante as 100 execuções.

Abordagens	**Precisão** $\mu \pm \sigma$	***F-measure*** $\mu \pm \sigma$
RNA+ADASYN	0,012 ± 0,004 **cd**	$0,054 \pm 0,016$ **d**
RNA+REP	0,011 ± 0,003 **d**	$0,052 \pm 0,015$ **d**
RNA-BASE	0,012 ± 0,004 **cd**	$0,053 \pm 0,018$ **d**
k-NN+ADASYN	0,014 ± 0,002 **c**	$0,061 \pm 0,012$ **c**
k-NN+REP	0,011 ± 0,002 **cd**	$0,054 \pm 0,009$ **d**
k-NN-BASE	0,012 ± 0,003 **cd**	$0,056 \pm 0,014$ **cd**
OC-SVM	0,215 ± 0,008 **b**	$0,486 \pm 0,008$ **a**
OC-PC	0,229 ± 0,007 **a**	$0,476 \pm 0,016$ **b**

Legenda: Diferentes letras representam médias com diferenças significativas no Teste Tukey com nível de confiança de 95%.

Fonte: Do autor (2024).

Tabela 4.4 – Média (μ) e Desvio Padrão (σ) das métricas de G-mean e IBA obtidos por cada abordagem durante as 100 execuções.

Abordagens	**G-mean** $\mu \pm \sigma$	**IBA** $\mu \pm \sigma$
RNA+ADASYN	$0,466 \pm 0,078$ **c**	$0,223 \pm 0,073$ **c**
RNA+REP	$0,465 \pm 0,078$ **c**	$0,222 \pm 0,074$ **c**
RNA-BASE	$0,469 \pm 0,084$ **c**	$0,227 \pm 0,082$ **c**
k-NN+ADASYN	$0,516 \pm 0,053$ **b**	$0,269 \pm 0,054$ **b**
k-NN+REP	$0,516 \pm 0,045$ **b**	$0,268 \pm 0,046$ **b**
k-NN-BASE	$0,519 \pm 0,064$ **b**	$0,274 \pm 0,065$ **b**
OC-SVM	$0,593 \pm 0,014$ **a**	$0,352 \pm 0,017$ **a**
OC-PC	$0,610 \pm 0,009$ **a**	$0,372 \pm 0,011$ **a**

Legenda: Diferentes letras representam médias com diferenças significativas no Teste Tukey com nível de confiança de 95%.

Fonte: Do autor (2024).

lado, os desvios padrão dos resultados de sensibilidade foram menores para as abordagens OC-PC e OC-SVM (vide Tabela 4.2 e Figura 4.1). Valores menores de desvio padrão indicam maior estabilidade nos resultados produzidos pelo algoritmo de AM. Dessa forma, mesmo diante de uma base de dados aparentemente heterogênea, as abordagens a nível de algoritmo produziram resultados concisos.

Deferentemente da sensibilidade, a abordagem que melhor desempenhou para especificidade foi a k-NN+ADASYN, atingindo média de 0,648, seguida pelas abordagens k-NN-BASE e OC-PC que obtiveram média de 0,579 e 0,571, respectivamente (vide Tabela 4.2). As demais abordagens desempenharam para especificidade sem diferença significativa (Figura 4.2). Assim como para a sensibilidade, os desvios padrão para especificidade foram menores nas abordagens OC-SVM e OC-PC (Tabela 4.2). A aplicação do algoritmo de *oversampling*, apesar de criar amostras sintéticas da classe positiva, possibilitou que um maior número de amostras negativas fossem destinadas para o projeto do classificador. Dessa forma, o classificar k-NN+ADASYN foi exposto a mais informações da classe negativa resultando em melhores resultados para especificidade quando comparado com demais abordagens que utilizaram o algoritmo k-NN.

Os desvios padrão para especificidade e sensibilidade das abordagens RNA+ADASYN, RNA+REP, k-NN+ADASYN e k-NN+REP foram menores quando comparados com as abordagens que utilizaram os mesmos algoritmos de AM e não utilizaram técnicas de *oversampling* (RNA-BASE e k-NN-BASE) (Tabela 4.2). Com aplicação de técnicas de *oversampling* foi possível explorar mais amostras da classe negativa durante a fase de treinamento ou projeto do classificador em cada uma das 100 execuções. Esse aumento da exploração das amostras negativas permitiu reduzir os desvios padrão nos resultados dos classificadores. Quando comparada as duas estratégias de *oversampling* (Replicação e ADASYN), o ADASYN possibilitou redução ainda mais significativa dos desvios padrão de sensibilidade e especificidade (Tabela 4.2). Isto submete ao fundamento do ADSYN de concentrar os dados sintéticos criados em lugar de difícil aprendizado. Todavia, esses fatos não colaboraram para melhoras substanciais na sensibilidade. Por outro lado, a aplicação do ADASYN em conjunto com o algoritmo de AM k-NN permitiu que a abordagem k-NN+ADASYN alcançasse a melhor média para especificidade dentre todas as abordagens avaliadas.

As abordagens OC-PC e OC-SVM atingiram níveis de precisão iguais a 0,229 e 0,215, e de *F-measure* 0,476 e 0,486, respectivamente (Tabela 4.3). Enquanto que para precisão a abordagem OC-PC foi significativamente superior (IC de 95%) a OC-SVM, para ***F-measure*** a

abordagem OC-SVM foi superior a OC-PC. A métrica *F-measure* com valores de β maiores aproxima-se da sensibilidade. Nesse trabalho, considerou-se $\beta = 2$, logo, como a sensibilidade foi maior para a abordagem OC-SVM, é esperado que essa superasse a abordagem OC-PC na métrica *F-measure* (Figura 4.4). As demais abordagens atingiram níveis de precisão próximos de 0,012 e de *F-measure* próximos a 0,015 (Tabela 4.3).

A superioridade das médias obtidas para as métricas de precisão e *F-measure* pelas abordagens OC-PC e OC-SVM em relação às demais abordagens foi marcante. Os principais fatores que promoveram tamanha discrepância foram o aumento na quantidade de TP e a redução substancial de FP. O aumento na quantidade de TP deve-se principalmente pelo fato de que todas as amostras da classe positiva foram destinadas para validação. No treinamento de classificadores OCC apenas a classe definida como classe alvo participa. Nesse trabalho definiu-se como classe alvo a classe de propriedades negativas. Dessa forma, participaram do treinamento do classificador OCC apenas as amostras negativas, e então, as amostras positivas foram todas destinadas para etapa de validação. Enquanto que as demais abordagens foram validadas apenas com 22 amostras da classe de propriedades positiva, as abordagens OC-PC e OC-SVM foram validadas com 82 amostras. A maior quantidade de amostras positivas dirigidas para a validação permitiu que a informação utilizada na etapa de validação representasse com mair fidelidade a classe de propriedades positivas. Assim sendo, a etapa de validação avaliou de maneira mais contundente a capacidade de generalização do classificador. Esse fato pode ser considerado uma grande vantagem dos algoritmos de OCC frente ao desafio de desbalanceamento de classe abordado neste trabalho.

Nas abordagens OCC o treinamento envolveu maior número de amostra da classe negativa. Isso foi possível, pois, ao não considerar a classe positiva no treinamento, o número de amostras da classe negativa destinada ao treinamento não se limitou ao da classe positiva como aconteceu nas outras abordagens. Dessa forma, a classe negativa foi dividida em treino e validação seguindo o padrão adotado pela academia, onde, 80% das amostras destinadas ao treinamento e 20% destinadas para validação. Com menor número de amostras negativas destinadas para a etapa de validação o número de FP reduziu consideravelmente.

A técnica de *undersampling* foi empregada para que as amostras de propriedades negativas, destinadas para treinamento ou projeto dos classificadores, apresentassem informações das diversidades da classe, ou seja, que represente bem a classe. O resultado de especificidade mostra que essa técnica foi eficaz. Mesmo as abordagens OCC sendo treinadas com maior

número de amostra da classe negativa, a abordagem que obteve maior especificidade foi a k-NN+ADASYN. Ou seja, as amostras negativas que participaram do projeto da abordagem k-NN+ADASYN foram representativas o suficiente para que a abordagem atingisse altos valores de TN, mesmo sendo projetada com número reduzido de propriedades negativas.

A ideia por trás da métrica IBA é garantir certo compromisso entre uma outra determinada métrica *M* qualquer e um índice de quão equilibradas são os valores de sensibilidade e especificidade. A métrica IBA tende favorecer modelos com melhores resultados na classe positiva. Era esperado então que a métrica IBA penalizasse as abordagens que obtivessem valores de especificidade maiores que sensibilidade, e valorizasse as abordagens com valores de sensibilidade maiores que especificidade. Porém, não houve valores discrepantes suficientes para promover mudança significativa (IC de 95%) na classificação das abordagens entre G-mean e IBA. Ou seja, as abordagens que melhor desempenharam na métrica IBA foram as mesmas da métrica G-mean (Figura 4.5 e Figura 4.6). As abordagens OCC (OC-PC e OC-SVM) foram as que melhor desempenharam, seguidas pelas abordagens que utilizaram o algoritmo k-NN (k-NN+ADASYN, k-NN+REP, e k-NN-BASE). As abordagens que utilizaram RNA (RNA+ADASYN, RNA+REP e RNA-BASE) foram as que pior desempenharam para G-mean e IBA (Tabela 4.4). Os algoritmos que utilizaram o classificador k-NN se beneficiam do fato do k-NN dispensar treinamento, o que é interessante para situações com poucos dados, como o caso da classe positiva. No caso das abordagens com RNA, o uso de poucos dados ou dados pouco representativos prejudica bastante a capacidade de aprendizado do modelo, fase crucial para garantir a generalização na fase operacional.

4.3 Análise de dados

Os resultados alcançados pelas abordagens de AM nas diferentes métricas avaliadas nesse trabalho não atingiram níveis satisfatórios. A abordagem que obteve maior média na métrica IBA foi a OC-PC, a qual atingiu o valor de 0,372. Esse valor de IBA foi resultante dos respectivos valores de sensibilidade, especificidade e G-mean: 0,653, 0,571 e 0,610 (Tabelas 4.2 e 4.2). Esses resultados ilustram o baixo desempenho atingido pelas abordagens de AM na base dados estudada nesse trabalho. Sabe-se que as características da base de dados como presença de conjuntos disjuntos, sobreposição de classes, presença de dados ruidosos, presença de instâncias na fronteira de decisão e sua relação com exemplos de ruído, influenciam diretamente o comportamento e desempenho dos algoritmos de AM frente ao desafio de desequilíbrio

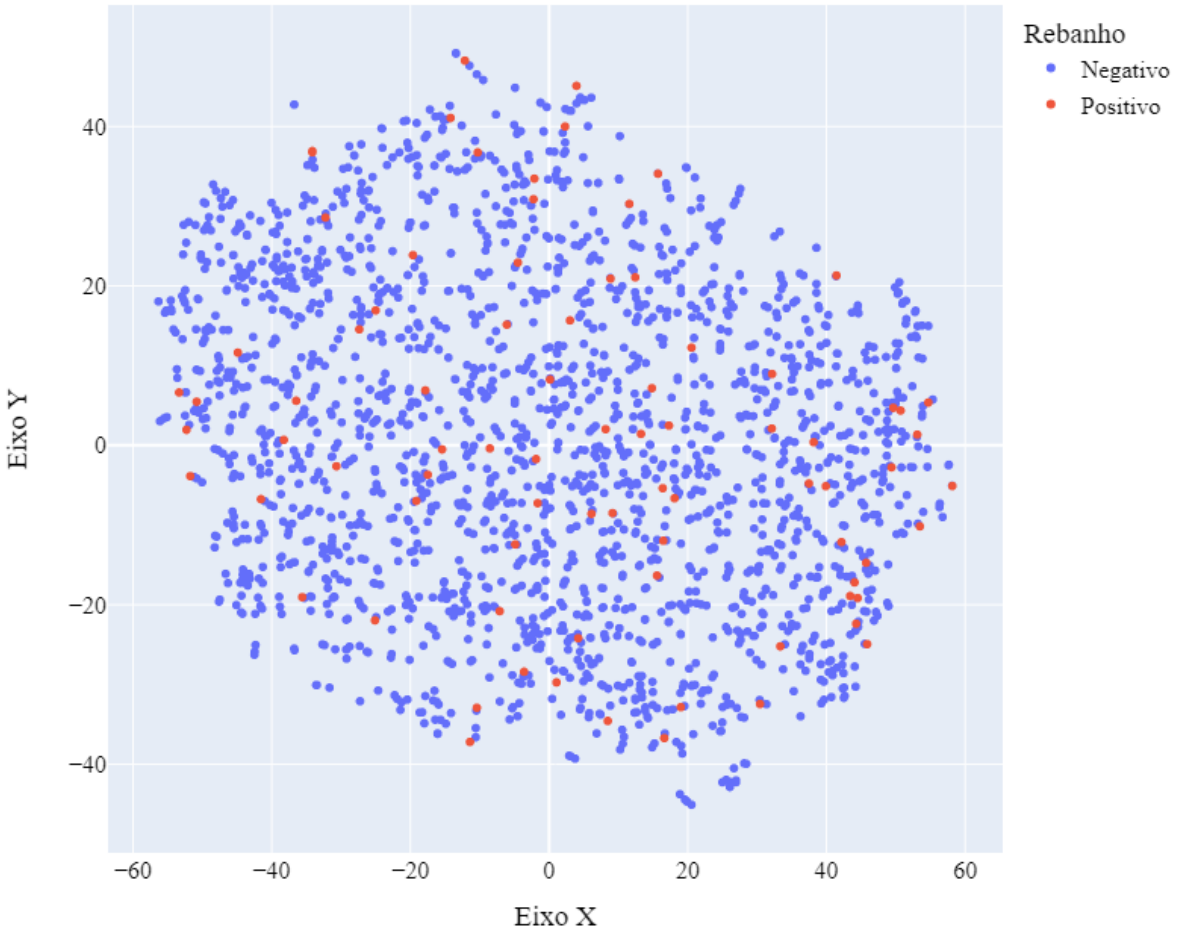
de classes (LóPEZ et al., 2013). Dessa forma, com o propósito de avaliar possíveis entraves que podem ter afetado o desempenho das abordagens, foi realizada uma análise exploratória da base de dados utilizada nesse trabalho.

Para visualizar a estrutura da base de dados bem como a distribuição das amostras, as variáveis foram transformadas usando o algoritmo *t-Distributed Stochastic Neighbour Embedding* (t-SNE) (MAATEN; HINTON, 2008). Através do algoritmo t-SNE é possível visualizar dados de alta dimensão, dando a cada amostra uma localização em um mapa bidimensional ou tridimensional. O algoritmo t-SNE converte um conjunto de dados de alta dimensão em uma matriz de similaridades de pares, após isso, aplica técnica para visualizar os dados de similaridade resultantes. De acordo com Maaten e Hinton (2008), o t-SNE é capaz de detectar muito bem grande parte da estrutura local dos dados de alta dimensão, ao mesmo tempo que revela a estrutura global, como a presença de *clusters* em diversas escalas.

A Figura 4.7 ilustra a distribuição da base de dados em sua totalidade, ou seja, todas as amostras com todas as 49 variáveis, em um espaço bidimensional alcançado pelo algoritmo t-SNE. Nota-se tamanha complexidade, do ponto de vista de reconhecimento de padrões, bem como elevado desequilíbrio da base. Tanto a classe de propriedades com rebanho negativo e a classe de propriedades com rebanho positivo são heterogêneas e esparsas. Não é possível identificar algum tipo de separação entre as classes nem presença de *clusters* bem definidos; as classes estão, praticamente, em sua totalidade sobrepostas. Essas características são prejudiciais para algoritmos de AM, o que explica um pouco dos resultados obtidos nessa Dissertação, apresentados na seção anterior.

A seleção de variáveis pode ser uma excelente aliada para melhorar o espaço amostral, permitindo então melhorias nos resultados dos algoritmos de AM. Em seu trabalho, Lopes (2016) explorou de forma aprofundada diferentes formas de seleção de variáveis e comparou o desempenho de RNA com cada conjunto de variáveis selecionadas para a mesma base de dados utilizada nesta Dissertação de Mestrado. Um dos seus objetivos foi determinar as variáveis mais discriminativas para a predição de brucelose em rebanhos bovinos, a fim de se obter um modelo de RNA mais simplificado, acurado e de fácil aplicação prática. As seleções de variáveis foram embasadas em Análise Univariada e FDR. Na Análise Univariada foram feitos estudos de associação das variáveis dependentes com as variáveis independentes ou explicativas. Adotou-se o Teste do Qui-quadrado de Pearson na análise dos dados qualitativos para medida das associações estatísticas. As variáveis quantitativas foram analisadas com teste U de Mann-Whitney, por

Figura 4.7 – Distribuição das amostras com todas as variáveis no espaço bidimensional



Fonte: Do autor (2024).

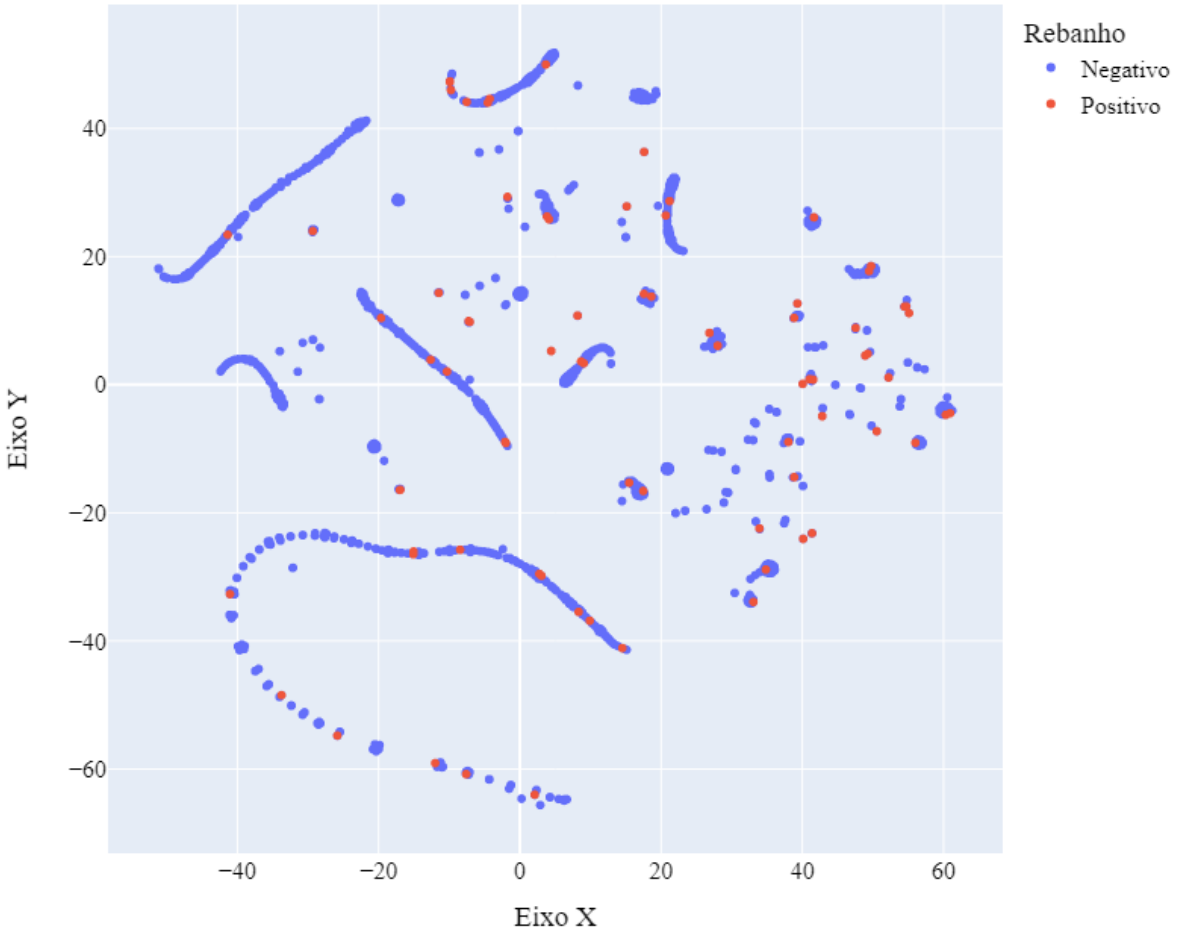
não apresentaram normalidade no teste de Shapiro-Wilk. Considerou-se os níveis de significância estatística de 5% ($p< 0,05$) e 10% ($p<0,10$) para seleção das variáveis. Quando admitiu-se $p<0,10$, 21 variáveis foram selecionadas para compor a RNA, enquanto que, para $p<0,05$, 13 variáveis foram selecionadas. O FDR indicou as variáveis com maior e menor relevância para o modelo, e foram feitas três formas de seleção de variáveis com FDR, sendo elas, as 30, 20 e 10 variáveis mais relevantes no FDR. Em seu estudo, Lopes (2016) avaliou também o desempenho da RNA sem seleção de variáveis, ou seja, utilizando todas as 49 variáveis presentes na base de dados.

No total, Lopes (2016) comparou 6 modelos de RNA combinados com diferentes formas de seleção de variáveis, sendo eles: um sem seleção de variáveis, dois com seleção de variáveis através de Análise Univariada e três com seleção de variáveis através do FDR. Não houve diferença significativa entre os resultados dos modelos, com relação à sensibilidade e eficiência. Porém, o modelo com as 10 variáveis mais relevantes no FDR apresentou maior especificidade em relação a outros, já que excluiu variáveis menos relevantes. Vale lembrar que a base de dados utilizada nesta Dissertação de Mestrado foi a base de dados com estas 10 variáveis mais relevantes descobertas no trabalho de Lopes (2016).

A Figura 4.8 mostra a distribuição das amostras, no espaço bidimensional, considerando as 10 variáveis mais relevantes no FDR selecionadas por Lopes (2016). Observa-se que com a seleção de variáveis foi possível agrupar significativamente as amostras da base de dados. Além do mais, a seleção de variáveis permitiu simplificar o espaço de características. Porém, mesmo com as melhorias provocadas pela seleção de variáveis, a base de dados continua sendo um problema de alta complexidade a nível de reconhecimento de padrões, haja vista que, as classes continuam muito sobrepostas e heterogêneas, apesar do aparecimento de alguns *clusters* alongados. Estes *clusters* com estrutura alongada podem justificar o bom desempenho do classificador OC-PC, uma vez que algoritmos baseados em Curvas Principais tem facilidade de representar dados com estruturas alongadas em alta dimensão, conforme mostrado em Moraes et al. (2020).

A fim de visualizar a separação entre classes proporcionada por cada variável gerou-se gráficos da Estimativa de Densidade de Kernel (KDE, do inglês *Kernel Density Estimation*) para cada variável. Os gráficos das KDE de cada variável estão ilustrados na Figura 4.10. Um exemplo de KDE para uma dada variável hipotética com alta separação entre classes pode ser observado na Figura 4.9. Pode-se observar que a KDE da Classe 1 não intersepta a KDE

Figura 4.8 – Distribuição das amostras com apenas 10 variáveis (selecionadas por Lopes (2016)) de maior relevância indicadas pelo FDR no espaço bidimensional



Fonte: Do autor (2024).

da Classe 2, essa não interseção indica que os dados da variável em questão possuem alta separabilidade entre as classes. Neste sentido, é fácil notar que nenhuma das variáveis exibidas na Figura 4.10 possuem boa separação entre classes, mesmo essas sendo as 10 variáveis de maior relevância indicadas pelo FDR. Esse fato reforça a ideia de que a base de dados utilizada nesse trabalho possui alta complexidade a nível de reconhecimento de padrões, que envolve: alta dimensão, alto desequilíbrio entre classes, mínima separabilidade e classes muito heterogêneas.

Figura 4.9 – Gráfico de KDE para variável hipotética com alta separação entre classes.

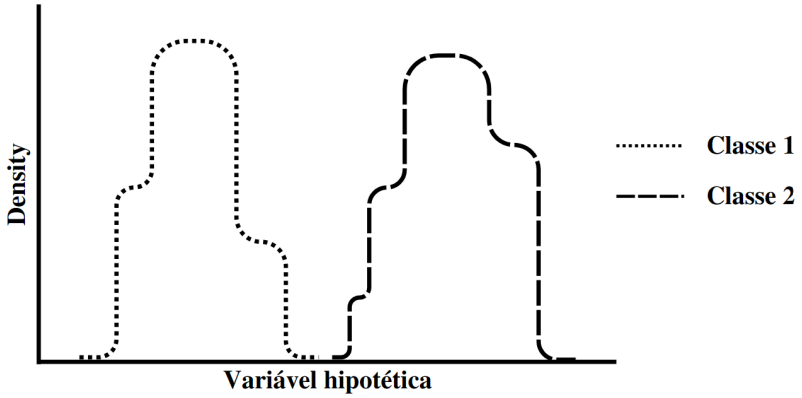
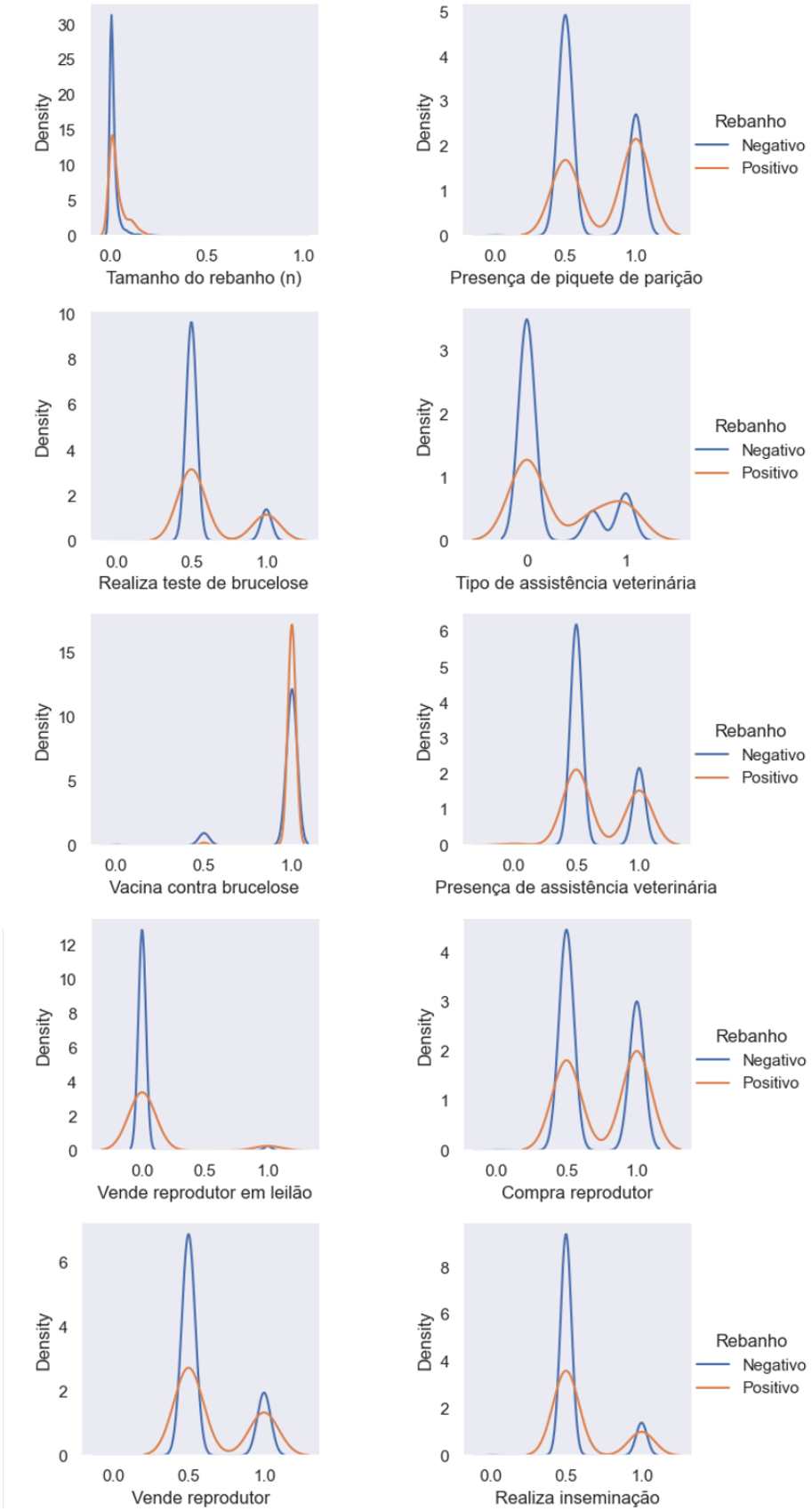


Figura 4.10 – Gráficos da KDE de cada uma das 10 variáveis (selecionadas por Lopes (2016)) de maior relevância indicadas pelo FDR no espaço bidimensional



Fonte: Do autor (2024).

5 CONSIDERAÇÕES FINAIS

Devido a diversas perdas para o setor pecuário provocadas pela brucelose bovina, essa tem sido considerada como uma das doenças mais preocupantes para o setor no Brasil. A preocupação dos órgãos governamentais para controle e erradicação dessa zoonose é notável, porém, diversos fatores ameaçam o estabelecimento de ações dos programas de defesa animal. Modelos de inteligência computacional como AM podem ser grandes aliados dos serviços de vigilância sanitária e epidemiológica. Os modelos de AM possuem grande capacidade de predição de informações, baseada em extração de padrões de base de dados, podendo fornecer subsídios para melhoria de processos e ações aos serviços de vigilância sanitária. Tanto do ponto de vista agropecuária quanto do ponto de vista de saúde única, considerando que a brucelose é uma zoonose, o desenvolvimento de abordagens AM para predição de brucelose possuem elevadíssimo potencial de benefício para a sociedade. A predição de brucelose, a partir de técnicas não invasivas como questionários, permitirão que os programas de defesa animal desempenham ações de forma muita mais célere, eficaz e econômica. Por exemplo, a disponibilização do questionário de forma on-line, aliado a predição feita por algoritmo de AM, possibilitaria a identificação de locais que necessitam de maior atenção dos fiscais de defesa sanitária de maneira ágil.

O desempenho dos modelos de AM está intimamente ligado com a qualidade e características intrínsecas da base dados utilizada durante a fase de projeto do modelo. O desequilíbrio dos dados disponíveis para as classes, envolvidas em determinado problema, é um exemplo de característica da base de dados que exige tratamento diferenciado. Dessa forma, a depender das características da base de dados, diversas técnicas podem ser adotadas visando tirar melhor proveito das informações ali disponíveis.

Esse trabalho comparou e avaliou o desempenho de diversas abordagens de AM, combinadas com diferentes técnicas de balanceamento de classe, na predição de brucelose em rebanhos bovinos. A princípio, o grande desafio do problema em questão foi o desequilíbrio entre classes da base de dados utilizada. Porém, os resultados pouco satisfatório obtidos pelas abordagens de AM, instigaram a execução de uma análise exploratória dos dados. Durante a análise exploratória da base de dados diversas características da base de dados evidenciaram um problema de alta complexidade a nível de reconhecimento de padrões. Os resultados aqui obtidos mostram que a aplicação de classificadores OCC são opções interessantes para lidar com base de dados que possuem desequilíbrio significativo entre classes e alta complexidade a nível de reconhecimento de padrões.

Para trabalhos futuros, podem ser adotadas outras técnicas de inteligência computacional, visando não apenas mitigar os efeitos do desequilíbrio entre classe, mas, também adequar de melhor forma à complexidade da base de dados. Implementações de algoritmos sensíveis ao custo ou *ensemble* são modelos com grande potencial para desempenharem em ambientes com alta complexidade de reconhecimento de padrões. Os métodos *ensemble* melhoram o desempenho de um único classificador combinando vários classificadores básicos que superam todos os independentes. A ideia principal dos métodos sensíveis ao custo é atribuir um custo de classificação incorreta mais alto para objetos pertencentes à classe de maior importância em relação ao custo de classificação incorreta para objetos pertencentes de menor importância, e minimizar erros de custo mais alto.

Dado tamanho potencial de benefício para a sociedade, para viabilizar a implementação de abordagens de AM que alcancem resultados satisfatórios na predição de brucelose, vale considerar como proposta futura a confecção de uma nova base de dados. Novas perguntas poderão ser incluídas no questionário, como as de pouca relevância poderiam ser excluídas. Ademais, as de maior relevância podem ser modificadas a fim de coletar respostas que maximizem a separação entre as classe de rebanhos positivos e negativos para brucelose.

6 CONCLUSÃO

O desempenho dos modelos de AM está intimamente ligado com a qualidade e características intrínsecas da base dados, como por exemplo o desequilíbrio entre os dados das classes envolvidas no problema. Esse trabalho comparou e avaliou o desempenho de diversas abordagens de AM, combinadas com diferentes técnicas de balanceamento de classe, na predição de brucelose em rebanhos bovinos. Os resultados aqui obtidos mostram que a aplicação de classificadores OCC são opções interessantes para lidar com base de dados que possuem desequilíbrio significativo entre classes e alta complexidade a nível de reconhecimento de padrões.

REFERÊNCIAS

ACHARYA, U. R. et al. Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals. **Computers in biology and medicine**, Elsevier Ltd, United States, v. 100, p. 270–278, 2018. ISSN 0010-4825.

AHMADIAN, S. et al. An efficient cardiovascular disease detection model based on multilayer perceptron and moth-flame optimization. **Expert systems**, Wiley Subscription Services, Inc, Oxford, v. 39, n. 4, p. n/a, 2022. ISSN 0266-4720.

ALHUDHAIF, A. A novel multi-class imbalanced eeg signals classification based on the adaptive synthetic sampling (adasyn) approach. **PeerJ. Computer science**, PeerJ. Ltd, United States, v. 7, p. e523–e523, 2021. ISSN 2376-5992.

ALIčKOVIć, E.; SUBASI, A. Breast cancer diagnosis using ga feature selection and rotation forest. **Neural computing applications**, Springer London, London, v. 28, n. 4, p. 753–763, 2015. ISSN 0941-0643.

AçıCı, K. et al. T4ss effector protein prediction with deep learning. **Data (Basel)**, MDPI AG, Basel, v. 4, n. 1, p. 45, 2019. ISSN 2306-5729.

BARROS, M. L. et al. Retrospective benefit-cost analysis of bovine brucellosis control in the state of mato grosso, brazil. **Preventive Veterinary Medicine**, v. 218, p. 105992, 2023. ISSN 0167-5877.

BASHIR, K. et al. Smotefris-inffc: Handling the challenge of borderline and noisy examples in imbalanced learning for software defect prediction. **Journal of intelligent fuzzy systems**, IOS Press BV, Amsterdam, v. 38, n. 1, p. 917–933, 2020. ISSN 1064-1246.

BATISTA, G.; PRATI, R.; MONARD, M. A study of the behavior of several methods for balancing machine learning training data. **SIGKDD explorations**, ACM, v. 6, n. 1, p. 20–29, 2004. ISSN 1931-0145.

BAUER, E. A.; JAGUSIAK, W. The use of multilayer perceptron artificial neural networks to detect dairy cows at risk of ketosis. **Animals (Basel)**, MDPI AG, Switzerland, v. 12, n. 3, p. 332, 2022. ISSN 2076-2615.

BENFODIL, K. et al. Prediction of trypanosoma evansi infection in dromedaries using artificial neural network (ann). **Veterinary parasitology**, Elsevier B.V, v. 306, p. 109716–109716, 2022. ISSN 0304-4017.

BIOURGE, V. et al. An artificial neural network-based model to predict chronic kidney disease in aged cats. **Journal of veterinary internal medicine**, John Wiley Sons, Inc, Hoboken, USA, v. 34, n. 5, p. 1920–1931, 2020. ISSN 0891-6640.

BOBBO, T. et al. Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. **Scientific reports**, Nature Publishing Group, London, v. 11, n. 1, p. 13642–13642, 2021. ISSN 2045-2322.

BOBBO, T. et al. Exploiting machine learning methods with monthly routine milk recording data and climatic information to predict subclinical mastitis in italian mediterranean buffaloes. **Journal of dairy science**, United States, 2022. ISSN 0022-0302.

BORGES, F. E. de M. et al. One-class classifier based on principal curves. **Neural Computing and Applications**, v. 35, n. 26, p. 19015–19024, 2023. ISSN 1433-3058.

BRANCO, P.; TORGO, L.; RIBEIRO, R. A survey of predictive modeling on imbalanced domains. **ACM computing surveys**, ACM, Baltimore, v. 49, n. 2, p. 1–50, 2016. ISSN 0360-0300.

BRAND, W. et al. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. **Journal of dairy science**, Elsevier Inc, United States, v. 104, n. 4, p. 4980–4990, 2021. ISSN 0022-0302.

BUNKHUMPORNPAT, C.; SINAPIROMSARAN, K.; LURSINSAP, C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: **Advances in Knowledge Discovery and Data Mining**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 475–482.

CASTRO, C. L. d.; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. **Sba: Controle Automação Sociedade Brasileira de Automatica**, Sociedade Brasileira de Automática, v. 22, n. Sba Controle Automação, 2011 22(5), 2011. ISSN 0103-1759.

CHAWLA, N. et al. Smote: Synthetic minority over-sampling technique. **The Journal of artificial intelligence research**, Ai Access Foundation, MARINA DEL REY, v. 16, p. 321–357, 2002. ISSN 1076-9757.

CLARKE, F. H. A new approach to lagrange multipliers. **Mathematics of Operations Research**, v. 1, n. 2, p. 165–174, 1976. ISSN 0364765X, 15265471.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, Sep 1995. ISSN 1573-0565.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE transactions on information theory**, IEEE, v. 13, n. 1, p. 21–27, 1967. ISSN 0018-9448.

CáRDENAS, L. et al. Characterization and evolution of countries affected by bovine brucellosis (1996–2014). **Transboundary and emerging diseases**, Wiley Subscription Services, Inc, Germany, v. 66, n. 3, p. 1280–1290, 2019. ISSN 1865-1674.

DADAR RUCHI TIWARI, K. S. M.; DHAMA, K. Importance of brucellosis control programs of livestock on the improvement of one health. **Veterinary Quarterly**, Taylor & Francis, v. 41, n. 1, p. 137–151, 2021.

DATTA, D. et al. A hybrid classification of imbalanced hyperspectral images using adasyn and enhanced deep subsampled multi-grained cascaded forest. **Remote sensing (Basel, Switzerland)**, MDPI AG, Basel, v. 14, n. 19, p. 4853, 2022. ISSN 2072-4292.

DEKA, R. P. et al. Bovine brucellosis: prevalence, risk factors, economic cost and control options with particular reference to india- a review. **Infection ecology epidemiology**, Taylor Francis, Abingdon, v. 8, n. 1, 2018. ISSN 2000-8686.

DOMINGUES, R. et al. A comparative evaluation of outlier detection algorithms: Experiments and analyses. **Pattern Recognition**, v. 74, p. 406–421, 2018. ISSN 0031-3203.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. New Jersey: Wiley, 2012.

EVSUKOFF, A. G. **Inteligência computacional: Fundamentos e aplicações [recurso eletrônico]**. Rio de Janeiro: E-papers, 2020.

FERREIRA, B. F. S. et al. Economic analysis of bovine brucellosis control in the rondônia state, brazil. **Tropical Animal Health and Production**, v. 55, n. 3, p. 225, May 2023.

FIX, E.; HODGES, J. **Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties**. Texas: USAF School of Aviation Medicine, 1951.

FRANC, K. A. et al. Brucellosis remains a neglected disease in the developing world: a call for interdisciplinary action. **BMC public health**, BioMed Central Ltd, England, v. 18, n. 1, p. 125–125, 2018. ISSN 1471-2458.

GARCÍA, V.; MOLLINEDA, R. A.; SÁNCHEZ, J. S. Index of balanced accuracy: A performance measure for skewed class distributions. In: **Pattern Recognition and Image Analysis**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 441–448. ISBN 978-3-642-02172-5.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

GOUDA, H. F. et al. Comparison of machine learning models for bluetongue risk prediction: a seroprevalence study on small ruminants. **BMC veterinary research**, BioMed Central Ltd, London, v. 18, n. 1, p. 1–394, 2022. ISSN 1746-6148.

GéRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**. [S.l.]: O'Reilly Media, 2019. ISBN 9781492032649.

HAIXIANG, G. et al. Learning from class-imbalanced data: Review of methods and applications. **Expert systems with applications**, Elsevier Ltd, v. 73, p. 220–239, 2017. ISSN 0957-4174.

HAN, H.; WANG, W.-Y.; MAO, B.-H. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In: **Advances in Intelligent Computing**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. p. 878–887.

HAN, J.; PEI, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. [S.l.]: Elsevier Science, 2011. ISBN 0123814804.

HASMITA, S. et al. Chili commodity price forecasting in bandung regency using the adaptive synthetic sampling (adasyn) and k-nearest neighbor (knn) algorithms. In: **2019 International Conference on Information and Communications Technology (ICOIACT)**. [S.l.: s.n.], 2019. p. 434–438.

HASTIE, T.; STUETZLE, W. Principal curves. **Journal of the American Statistical Association**, v. 84, n. 406, p. 502–516, 1989.

HE, H. et al. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: **2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)**. [S.l.: s.n.], 2008. p. 1322–1328. ISSN 2161-4407.

HE, H.; MA, Y. **Imbalanced Learning: Foundations, Algorithms, and Applications**. 1. ed. Somerset: Wiley, 2013.

HSU, C.-W.; CHANG, C.-C.; LIN, C.-J. **A Practical Guide to Support Vector Classification**. [S.l.], 2003.

HU, S. et al. Msmote: Improving classification performance when training data is imbalanced. In: **2009 Second International Workshop on Computer Science and Engineering**. [S.l.]: IEEE, 2009. v. 2, p. 13–17.

HULL, N. C.; SCHUMAKER, B. A. Comparisons of brucellosis between human and veterinary medicine. **Infection Ecology & Epidemiology**, Taylor & Francis, v. 8, n. 1, p. 1500846, 2018.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern recognition letters**, Elsevier B.V, AMSTERDAM, v. 31, n. 8, p. 651–666, 2010. ISSN 0167-8655.

KESHAVARZI, H. et al. Machine learning algorithms, bull genetic information, and imbalanced datasets used in abortion incidence prediction models for iranian holstein dairy cattle. **Preventive veterinary medicine**, Elsevier B.V, Netherlands, v. 175, p. 104869–104869, 2020. ISSN 0167-5877.

KHURANA, S. et al. Bovine brucellosis – a comprehensive review. **Veterinary Quarterly**, v. 41, p. 61–88, 2021.

KOTHALAWALA, K. A. C. et al. Association of farmers' socio-economics with bovine brucellosis epidemiology in the dry zone of sri lanka. **Preventive veterinary medicine**, Elsevier B.V, Netherlands, v. 147, p. 117–123, 2017. ISSN 0167-5877.

KUBAT, M.; HOLTE, R.; MATWIN, S. Learning when negative examples abound. In: **Machine Learning: ECML-97**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. p. 146–153. ISBN 978-3-540-68708-5.

LEEVY, J. L. et al. A survey on addressing high-class imbalance in big data. **Journal of Big Data**, v. 5, n. 1, p. 42, 2018. ISSN 2196-1115.

LEWIS, D. D.; GALE, W. A. A sequential algorithm for training text classifiers. In: **SIGIR '94**. London: Springer London, 1994. p. 3–12.

LIMA, J. S. et al. A machine learning proposal method to detect milk tainted with cheese whey. **Journal of dairy science**, Elsevier Inc, v. 105, n. 12, p. 9496–9508, 2022. ISSN 0022-0302.

LOPES, E. **Sistema neural para predição de brucelose em rebanhos bovinos**. 85 p. Tese (Doutorado em Ciências Veterinárias) — Universidade Federal de Lavras, Lavras, 2016.

LOYOLA-GONZáLEZ, O. et al. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. **Neurocomputing (Amsterdam)**, Elsevier B.V, v. 175, p. 935–947, 2016. ISSN 0925-2312.

LóPEZ, V. et al. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. **Information sciences**, Elsevier Inc, NEW YORK, v. 250, p. 113–141, 2013. ISSN 0020-0255.

LóPEZ, V. et al. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. **Expert systems with applications**, Elsevier Ltd, OXFORD, v. 39, n. 7, p. 6585–6608, 2012. ISSN 0957-4174.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008.

MACQUEEN, J. Classification and analysis of multivariate observations. In: UNIVERSITY OF CALIFORNIA LOS ANGELES LA USA. **5th Berkeley Symp. Math. Statist. Probability**. [S.l.], 1967. p. 281–297.

MARTINEZ-RAU, L. S. et al. A robust computational approach for jaw movement detection and classification in grazing cattle using acoustic signals. **Computers and electronics in agriculture**, Elsevier B.V, Amsterdam, v. 192, p. 106569, 2022. ISSN 0168-1699.

MOHANTY, S. N. et al. **Machine Learning for Healthcare Applications**. 1. ed. Newark: John Wiley Sons, Incorporated, 2021. ISBN 9781119791812.

MORAES, E. C. C. et al. Data clustering based on principal curves. **Advances in Data Analysis and Classification**, v. 14, n. 1, p. 77–96, 2020. ISSN 1862-5355.

NARTOWT, B. J. et al. Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. **PloS one**, Public Library of Science, San Francisco, CA USA, v. 14, n. 8, p. e0221421–e0221421, 2019. ISSN 1932-6203.

OLIVEIRA, L. F. d. et al. Seroprevalence and risk factors for bovine brucellosis in minas gerais state, brazil. **Semina. Ciências agrárias : revista cultural e científica da Universidade Estadual de Londrina**, Universidade Estadual de Londrina, v. 37, n. 5Supl2, p. 3449–3466, 2016. ISSN 1679-0359.

PAUL, D. et al. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. **Computerized medical imaging and graphics**, Elsevier Ltd, United States, v. 60, p. 42–49, 2016. ISSN 0895-6111.

PECHT, M. G.; KANG, M. **Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things**. Second edition. Newark: Wiley, 2018. ISBN 1119515335.

PURI, A.; Kumar Gupta, M. Knowledge discovery from noisy imbalanced and incomplete binary class data. **Expert Systems with Applications**, v. 181, p. 115179, 2021. ISSN 0957-4174.

RAMENTOL, E. et al. Smote-rsb: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory. **Knowledge and information systems**, Springer-Verlag, London, v. 33, n. 2, p. 245–265, 2012. ISSN 0219-1377.

RAMIREZ, O. L. H. et al. Seroepidemiology of bovine brucellosis in colombia's preeminent dairy region, and its potential public health impact. **Brazilian journal of microbiology**, Springer International Publishing, Cham, v. 51, n. 4, p. 2133–2143, 2020. ISSN 1517-8382.

ROCHA, I. D. dos S. et al. Distribution, seroprevalence and risk factors for bovine brucellosis in brazil: Official data, systematic review and meta-analysis. **Revista Argentina de Microbiología**, v. 56, n. 2, p. 153–164, 2024. ISSN 0325-7541.

RODRIGUES, D. L. et al. Seroprevalence and risk factors for bovine brucellosis in the state of paraná, brazil: an analysis after 18 years of ongoing control measures. **Tropical Animal Health and Production**, v. 53, n. 5, p. 503, Oct 2021.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210–229, 1959.

SCHöLKOPF, B. et al. Support vector method for novelty detection. In: **Proceedings of the 12th International Conference on Neural Information Processing Systems**. [S.l.]: MIT Press, 1999. (NIPS'99), p. 582–588.

SOUZA, M. R. S. et al. Evaluation of diagnostic tests' sensitivity, specificity and predictive values in bovine carcasses showing brucellosis suggestive lesions, condemned by brazilian federal meat inspection service in the amazon region of brazil. **Preventive veterinary medicine**, Elsevier B.V, Netherlands, v. 200, p. 105567–105567, 2022. ISSN 0167-5877.

SREEJITH, S.; NEHEMIAH, H. K.; KANNAN, A. Clinical data classification using an enhanced smote and chaotic evolutionary feature selection. **Computers in biology and medicine**, Elsevier Ltd, United States, v. 126, p. 103991–103991, 2020. ISSN 0010-4825.

STEFANOWSKI, J. Dealing with data difficulty factors while learning from imbalanced data. In: ____. **Challenges in Computational Statistics and Data Mining**. Cham: Springer International Publishing, 2016. p. 333–363.

TAHIR, M. A. et al. A multiple expert approach to the class imbalance problem using inverse random under sampling. In: **Multiple Classifier Systems**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 82–91.

TSAI, C.-F.; LIN, W.-C. Feature selection and ensemble learning techniques in one-class classifiers: An empirical study of two-class imbalanced datasets. **IEEE Access**, v. 9, p. 13717–13726, 2021.

USDA FOREIGN AGRICULTURAL SERVICE. **Production, Supply, and Distribution Database**. 2022.

VALLETTA, J. J. et al. Applications of machine learning in animal behaviour studies. **Animal Behaviour**, v. 124, p. 203–220, 2017. ISSN 0003-3472.

VARRECCHIA, T. et al. An artificial neural network approach to detect presence and severity of parkinson's disease via gait parameters. **PloS one**, Public Library of Science, United States, v. 16, n. 2, 2021. ISSN 1932-6203.

WANG, J. et al. A remote sensing data based artificial neural network approach for predicting climate-sensitive infectious disease outbreaks: A case study of human brucellosis. **Remote sensing (Basel, Switzerland)**, MDPI AG, Basel, v. 9, n. 10, p. 1018, 2017. ISSN 2072-4292.

WANG, M. et al. Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records. **BMC medical informatics and decision making**, BioMed Central Ltd, England, v. 22, n. 1, p. 41–41, 2022. ISSN 1472-6947.

WANG, S. et al. Hyperparameter selection of one-class support vector machine by self-adaptive data shifting. **Pattern Recognition**, v. 74, p. 198–211, 2018. ISSN 0031-3203.

XU, X. et al. Application of machine learning to predict the inhibitory activity of organic chemicals on thyroid stimulating hormone receptor. **Environmental Research**, v. 212, p. 113175, 2022. ISSN 0013-9351.

ZAREBIDAKI, M. et al. Occurrence and risk factors of brucellosis among domestic animals: an artificial neural network approach. **Tropical animal health and production**, Springer Netherlands, Dordrecht, v. 54, n. 1, p. 62–62, 2022. ISSN 0049-4747.

ZHANG, N. et al. Animal brucellosis control or eradication programs worldwide: A systematic review of experiences and lessons learned. **Preventive Veterinary Medicine**, v. 160, p. 105–115, 2018. ISSN 0167-5877.

APÊNDICE A – Lista de Publicações

Nesta seção, os artigos e resumos relacionados ao desenvolvimento desta, publicados em congressos e periódicos nacionais e internacionais, são apresentados. Estes estão organizados em ordem cronológica e acompanham breve descrição.

.1 Artigos Publicados em Anais de Congressos

1. ALVES, C. D. Q., FERREIRA, D. D., FORTUNATO, D. A., ROCHA, C. M. B. M., "Rede Neural Artificial para predição de brucelose bovina a partir de dados desbalanceados", XVI Brazilian Congress on Computational Intelligence - CBIC2023, Bahia, Brasil, outubro 2023

2. ALVES, C. D. Q., FERREIRA, D. D., ROCHA, C. M. B. M., "Rede neural artificial para predição de brucelose bovina a partir de dados desbalanceados", XXXII Congresso de Pós-Graduação da UFLA - CPG2023, Minas Gerais, Brasil, novembro 2023

   Estes trabalhos apresentam o desenvolvimento de RNA com técnicas de balanceamento de classes e seleção de variáveis via algoritmo genético, para classificação e segregação de rebanhos bovinos, quanto à soroprevalência para brucelose. Cinco RNAs foram projetadas combinando diferentes abordagens de técnica de balanceamento de classes e seleção de variáveis, a fim de comparar qual abordagem apresentaria melhores resultados. Os resultados mostraram que a RNA combinada com a técnica de seleção de variáveis, via Algoritmos Genéticos, e balanceamento de classes, é uma abordagem promissora.

.2 Resumos Publicados em Anais de Congressos

1. ALVES, C. D. Q., FERREIRA, D. D., ROCHA, C. M. B. M., "Rede Neural Artificial para predição de brucelose bovina a partir de dados desbalanceados", I ENCONTRO NACIONAL DA REDE EM PESQUISA E INOVAÇÃO EM SANIDADE E PECUÁRIA LEITEIRA - InovaLeite, Minas Gerais, Brasil, setembro 2023

   Este trabalho apresenta o desenvolvimento de RNA com técnicas de balanceamento de classes e seleção de variáveis via algoritmo genético, para classificação e segregação de rebanhos bovinos, quanto à soroprevalência para brucelose. Cinco RNAs foram projetadas combinando diferentes abordagens de técnica de balanceamento de classes e seleção

de variáveis, a fim de comparar qual abordagem apresentaria melhores resultados. Os resultados mostraram que a RNA combinada com a técnica de seleção de variáveis, via Algoritmos Genéticos, e balanceamento de classes, é uma abordagem promissora.