UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Carolina Deina

APRIMORANDO A TOMADA DE DECISÃO EM SAÚDE COM APRENDIZADO DE MÁQUINA EM PROBLEMAS DE CLASSIFICAÇÃO EM DADOS DESBALANCEADOS

Porto Alegre

2024

Carolina Deina

Aprimorando a Tomada de Decisão em Saúde com Aprendizado de Máquina em problemas de Classificação em Dados Desbalanceados

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Doutora em Engenharia, na área de concentração em Sistemas de Produção.

Orientador: Flávio Sanson Fogliatto, PhD.

Porto Alegre

2024

Carolina Deina

Aprimorando a Tomada de Decisão em Saúde com Aprendizado de Máquina em problemas de Classificação em Dados Desbalanceados

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

Prof. Flávio Sanson Fogliatto, PhD.
Orientador PPGEP/UFRGS

Prof. Michel José Anzanello, PhD.
Coordenador PPGEP/UFRGS

Banca Examinadora:

Professor Michel José Anzanello, PhD. (PPGEP/UFRGS)

Professor Rodolfo Benedito Zattar da Silva, Dr. (UFMT)

Professor Milad Yousefi, Dr.

À minha família, por seu amor, e a
todos que se inquietam e buscam.

Agradecimentos

Ao meu orientador, Prof. Flávio S. Fogliatto, pela constante disponibilidade em orientar, esclarecer dúvidas e compartilhar seus conhecimentos. Sua expertise foi fundamental para a condução deste trabalho e para o meu aprimoramento acadêmico e profissional. Foi uma honra contar com sua orientação ao longo desta jornada.

Aos meus orientadores de iniciação científica e mestrado — Ailton M. Liberato, Carla A. P. Schmidt, Neron A. C. Berghauser, Flávio Trojan e Hugo V. Siqueira — agradeço por compartilharem seus conhecimentos e por me abrirem portas durante minha trajetória acadêmica. Estendo esse agradecimento aos professores da UNIR, UTFPR e UFRGS, que contribuíram significativamente para minha formação.

Agradeço também aos meus pais, Fátima e Pedro, pelo amor e apoio de sempre. Às minhas irmãs, Jéssica, Gabriela e Miquelina, pela parceria e incentivo em todos os momentos. Aos meus sobrinhos, Théo e Saulo, por encherem minha vida de alegria. Ao meu companheiro, Alexandre, por estar ao meu lado em cada passo desta jornada, com afeto e paciência.

Aos amigos e colegas do PPGEP, obrigado pela parceria, conversas e pelos momentos que enriqueceram minha experiência acadêmica e pessoal. Aos amigos da vida, que, mesmo à distância, sempre ofereceram suporte e incentivo, minha sincera gratidão.

Ao PPGEP e à CAPES, agradeço pela oportunidade, pelos recursos e pelo apoio que tornaram possível esta trajetória.

A todos que, direta ou indiretamente, contribuíram para esta conquista, deixo meu mais sincero agradecimento.

RESUMO

Essa tese aborda a classificação de dados desbalanceados na área da saúde, propondo um *workflow* adequado para lidar com essa complexidade por meio de *Machine Learning* (ML) supervisionado. Composta por três artigos, a pesquisa apresenta estratégias para lidar com o desbalanceamento, além de explorar a generalização e a interpretabilidade dos algoritmos de ML. Os objetivos são: (i) definir um *framework* para classificação em bancos de dados desbalanceados; (ii) avaliar a capacidade de generalização dos modelos em situações do mundo real; (iii) propor direções futuras para a interpretabilidade dos resultados visando apoiar decisões médicas. Do ponto de vista prático, a aplicação da tese é validada através de dois estudos de caso: o primeiro identificando pacientes com diabetes com maior risco de hospitalização, otimizando o direcionamento de recursos hospitalares e o segundo sobre estimativas de *no-show* (não comparecimento) em consultas médicas, com o objetivo de melhorar o serviço prestado aos pacientes.

Palavras-chave: Dados desbalanceados; classificação desbalanceada; previsão de *no-show*; previsão de hospitalização; radiologia.

ABSTRACT

This dissertation addresses the classification of imbalanced datasets in the healthcare domain, proposing a workflow to deal with such complexity through supervised machine learning. Comprising three articles, the research presents strategies for handling imbalance while exploring the generalization and interpretability of machine learning models. The objectives are: (i) to define a framework for classification in imbalanced databases; (ii) to evaluate the models' generalization ability in real-world situations; (iii) to propose future directions for result interpretability to support medical decision-making. From a practical standpoint, the dissertation is validated through two case studies: the first identifying diabetic patients at higher risk of hospitalization, optimizing the allocation of hospital resources, and the second on predictions of no-show in medical appointments, aiming to enhance the service provided to patients.

Key words: Imbalanced datasets; imbalanced classification; no-show prediction; hospitalization prediction; radiology.

LISTA DE FIGURAS

Figura 1. Fluxograma PRISMA 27

Figura 2. Workflow para classificação desbalanceada 30

Figura 3. Outline of the proposed method 57

Figura 4. Features used as inputs and frequency with which they were retained in the 100 validations of the best predictive model 59

Figura 5. SHAP values (features impacts on model outputs) 62

Figura 6. Relationship between the features' real values (displayed along the horizontal axes) and their respective SHapley Additive exPlanations (SHAP) values (displayed along the vertical axes) 63

Figura 7. Outline of the proposed framework 88

Figura 8. Boxplot of sensitivity results in the validation set for all prediction models. 96

Figura 9. Features selected by top models, occurrence frequency in 100 test set replicates 97

Figura 10. Boxplot of sensitivity results in the validation set for all prediction models99

Figura 11. Features selected by top models, occurrence frequency in 100 test set replicates 99

LISTA DE TABELAS

Tabela 1. Estrutura das etapas da pesquisa desenvolvida 17

Tabela 2. Revisões de literatura sobre o tema 24

Tabela 3. Dimensões de pesquisa e palavras-chaves 28

Tabela 4. Average predictive performance and standard deviations (within parentheses) obtained from 100 replicates of the dataset validation portion for different combinations of resampling technique and 61

Tabela 5. Average predictive performance and standard deviations obtained from 100 replicates of Dataset 1's test portion 95

Tabela 6. Average predictive performance and standard deviations obtained from 100 replicates of Dataset 1's validation portion 95

Tabela 7. Average predictive performance and standard deviations obtained from 100 replicates of Dataset 2's test portion 98

Tabela 8. Average predictive performance and standard deviations obtained from 100 replicates of Dataset 2's validation portion 98

LISTA DE SIGLAS E ABREVIATURAS

AUC - Area Under the Receiver Operating Characteristic Curve

AUC - Urinary Albumin Concentration

DKD - Diabetic Kidney Disease

eGRF - Estimated Glomerular Filtration Rate Calculation

EMRs - Electronic medical records

FN - False Negative

FP - False Positive

GP - Genetic Programming

HbA1c - Glycated Hemoglobin

ICD - International Classification of Diseases

IHT - Instance Hardness Threshold

KNN - K-Nearest Neighbors

LR - Logistic Regression

ML - Machine Learning

NM - NearMiss

NPV - Negative Predictive Values

PPV - Positive Predictive Values

ROC - Receiver Operating Characteristic

RUS - Random under Sampling

SHAP - Shapley Additive Explanations

SMOTE - Synthetic Minority Oversampling Technique

SR - Symbolic Regression

SVM - Support Vector Machine

TN - True Negative

TP - True Positive

SUMÁRIO

1 INTRODUÇÃO 13

1.1 TEMA E OBJETIVOS 14

1.2 JUSTIFICATIVA DO TEMA E OBJETIVOS 15

1.3 DELINEAMENTO DO ESTUDO 16

1.3.1 Método de Pesquisa 16

1.3.2 Método de Trabalho 17

1.4 DELIMITAÇÕES DO ESTUDO 18

1.5 ESTRUTURA DA TESE 18

1.6 REFERÊNCIAS 19

2 ARTIGO 1: Desafios e soluções no uso de machine learning para dados de saúde desbalanceados: um workflow para robustez preditiva e interpretabilidade 21

2.1 INTRODUÇÃO 21

2.2 PESQUISAS ANTERIORES E CONTRIBUIÇÕES 23

2.3 MÉTODO 26

2.4 RESULTADOS 29

2.5 CONCLUSÃO 41

2.6 REFERENCES 42

2.7 APÊNDICE A1 50

3 ARTIGO 2: Machine learning for high-risk hospitalization prediction in outpatient patients with diabetes at a tertiary hospital 53

3.1 INTRODUCTION 54

3.2 METHODS 55

3.2.1 Study Database 55

3.2.2 Intelligent System Protocol 56

3.2.3 Statistical Analysis 58

3.3 RESULTS 58

3.4 DISCUSSION 64

3.5 CONCLUSION 66

3.6 REFERENCES 67

3.7 APPENDIX A2 75

3.8 APPENDIX B2 76

3.9 APPENDIX C2 77

4 ARTIGO 3: Decision analysis framework for predicting no-shows to appointments using machine learning algorithms 78

4.1 INTRODUCTION 79

4.2 BACKGROUND 82

4.2.1 Resampling techniques 85

4.2.2 Prediction models 86

4.3 METHOD 87

4.3.1 Performance metrics 92

4.4 DATASETS 93

4.5 RESULTS 94

4.6 DISCUSSION 100

4.6.1 Cross-Validation Strategies and Stratification 100

4.6.2 Resampling Techniques 102

4.6.3 Significant Predictors 104

4.6.4 Practical Implications 106

4.7 CONCLUSION 107

4.8 REFERENCES 107

4.9 APPENDIX A3 120

5 CONSIDERAÇÕES FINAIS 122

1 INTRODUÇÃO

As diversas abordagens de *Machine Learning* (ML), como supervisionado, não supervisionado e de reforço, oferecem uma ampla gama de algoritmos que são flexíveis e podem lidar com tarefas como classificação, regressão, clusterização e outras, adaptando-se a diferentes formatos de dados, como imagens, áudios, séries temporais, textos e tabelas. Isso expande consideravelmente o escopo e o potencial da ciência de dados para aplicações no mundo real (Gandhi & Narmawala, 2024).

Na área saúde, algoritmos de ML são ferramentas valiosas: orientam decisões, detectam precocemente condições médicas, possibilitam cuidados mais personalizados e eficazes, e fornecem *insights* para uma gestão otimizada dos recursos hospitalares. Desta forma, apresentam o potencial de trazer benefícios substanciais para pacientes, médicos e hospitais (Bai et al., 2023; Cui et al., 2018; Jia et al., 2022).

Tal diversidade de aplicações é evidenciada através de exemplos reportados na literatura. Algoritmos de aprendizado profundo foram utilizados por Shen et al. (2019) para detecção de câncer de mama em mamografia de rastreamento. Registros de eletroencefalograma (EEG), analisados como séries temporais de dados, foram utilizados por Gramacki & Gramacki (2022) em modelos de aprendizado supervisionado para detectar crises epilépticas. Algoritmos de ML supervisionadas também foram utilizadas por Valero-Bover et al. (2022) para prever *no-show* de pacientes a fim de otimizar a logística hospitalar e melhorar a eficiência dos serviços de saúde.

A implementação eficaz de algoritmos de ML na área da saúde demanda compreensão aprofundada dos problemas e dos dados envolvidos. A complexidade dos dados médicos, frequentemente desbalanceados quanto à incidência em categorias de interesse, requer um *workflow* de análise meticulosamente estruturado (Rekha et al., 2019). O desequilíbrio (i.e., desbalanceamento) entre classes pode distorcer os modelos de classificação e previsão, gerando resultados enviesados e pouco confiáveis. Por exemplo, ao treinar um modelo para prever a probabilidade de internação de pacientes com diabetes, o algoritmo pode favorecer a classe majoritária (pacientes que não requerem internação), negligenciando a identificação da classe minoritária,

comprometendo a precisão na identificação correta de casos de internação (Maheshwari et al., 2018).

Mesmo que a aplicação de técnicas de balanceamento de dados nas classes de interesse seja realizada, se essa etapa acontecer antes de uma divisão adequada entre conjuntos de treinamento e validação, os resultados podem ser tendenciosos. Por exemplo, ao usar técnicas de *oversampling* (que envolvem o aumento artificial da representação de instâncias da classe minoritária) antes da separação dos conjuntos de treino e teste, o modelo será exposto a dados de teste artificiais que podem resultar em superestimação de sua capacidade de generalização (Cabitza & Campagner, 2021). Ajustar o *workflow* de análise de acordo com as necessidades do problema é essencial para garantir resultados generalizáveis e, se interpretáveis, fornecer informações para auxiliar na tomada de decisões médicas ou gerenciais.

Nesse contexto, essa pesquisa se concentra na investigação de um *workflow* de análise robusto projetado para abordar o problema do desbalanceamento de dados em problemas de classificação na saúde.

1.1 TEMA E OBJETIVOS

O tema central desta tese é a melhoria do suporte à decisão na área da saúde, com uma ênfase especial na resolução de problemas de classificação em que as classes estão desbalanceadas quanto ao número de instâncias. O objetivo principal desta pesquisa é apresentar um *framework* abrangente que aborde o problema do desbalanceamento de classes, com o intuito de aprimorar a precisão das classificações.

Os objetivos específicos desta pesquisa compreendem:

(i) Definir as etapas fundamentais de um *framework* para análise de problemas da classificação em cenários de classes desequilibradas, proporcionando uma abordagem estruturada para lidar com essa complexidade;

(ii) Avaliar a capacidade de generalização de algoritmos de ML ao trabalhar com conjuntos de dados desbalanceados, fornecendo *insights* sobre a eficácia desses modelos em situações do mundo real;

(iii) Avaliar a interpretabilidade dos resultados de classificação ao utilizar algoritmos de ML tradicionais, a fim de oferecer informações para embasar decisões médicas informadas.

1.2 JUSTIFICATIVA DO TEMA E OBJETIVOS

Algoritmos de ML em problemas de classificação têm sido ferramentas valiosas no apoio à tomada de decisões em saúde (Habehh & Gohel, 2021; Obermeyer & Emanuel, 2016). Esses algoritmos têm a capacidade de analisar grandes volumes de dados de pacientes, identificar padrões e prever resultados com base em informações históricas. Isso é especialmente útil na identificação precoce de condições médicas, otimização de recursos hospitalares e personalização de tratamentos (Bai et al., 2023; Cui et al., 2018). Contudo, deve-se reconhecer e mensurar os impactos negativos de previsões imprecisas. Decisões clínicas ou administrativas baseadas em informações inadequadas podem resultar em tratamentos ineficazes, gastos excessivos e redução da qualidade do atendimento, afetando diretamente a condição dos pacientes (Ben-Israel et al., 2020; Chen et al., 2021; Haynes & Wilczynski, 2010).

Em relação às suas contribuições teóricas, esta tese se justifica ao propor um arcabouço analítico para lidar com problemas de classificação em dados médicos desbalanceados. Embora a literatura já ofereça estratégias para tratar o problema, uma abordagem eficaz vai além da compreensão de técnicas já existentes e do desenvolvimento de novos algoritmos. Nos artigos 1 e 3, enfatiza-se a importância de compreender plenamente o *workflow* de previsão para garantir resultados robustos e generalizáveis, minimizando viés nos conjuntos de calibração e validação. Isso fortalece a interpretabilidade dos resultados, a qual é essencial para decisões orientadas, questão discutida no artigo 2.

Quanto aos aspectos práticos, esta tese se justificativa pela validação das propostas apresentadas através de dois estudos de caso. No artigo 2, o modelo proposto permitiu identificar pacientes com diabetes com maior risco de hospitalização. Isso permite dedicar atenção especial a esses pacientes durante o acompanhamento ambulatorial e identificá-los assim que chegam ao serviço de emergência, otimizando a alocação de recursos. No artigo 3 foi relatado que estimar *no-show* em consultas médicas

pode melhorar significativamente a qualidade da assistência médica. Isso possibilita acesso transparente às informações dos pacientes, permitindo que gestores elaborem estratégias de intervenção direcionadas para pacientes com maior probabilidade de não comparecer às consultas. Previsões precisas de *no-show* também podem aprimorar o desempenho de estratégias conhecidas para reduzir os impactos das faltas no sistema, como o *overbooking* de agendamentos.

1.3 DELINEAMENTO DO ESTUDO

Estabelecidos os objetivos da tese e justificada a relevância da pesquisa nela apresentada, a presente seção delineia o estudo, detalhando os métodos de pesquisa utilizados para atingir esses objetivos.

1.3.1 Método de Pesquisa

As definições de Gil (2002) são adotadas no enquadramento metodológico deste estudo. Sob a ótica dos objetivos, este estudo tem natureza explicativa, inicialmente embasando-se na literatura técnica para construir a base teórica. Posteriormente, amplia essa base ao abordar questões de generalização e interpretabilidade no contexto da classificação de dados desbalanceados na área da saúde.

Quanto à natureza da pesquisa, classifica-se como aplicada, direcionando seus esforços para a aplicação prática de algoritmos de ML. A abordagem do problema adota duas perspectivas distintas: uma qualitativa, evidenciada no artigo 1, que revisita a literatura existente sobre classificação de dados desbalanceados na área médica, e outra quantitativa, evidenciada nos demais artigos, que aplicam algoritmos de ML.

No aspecto dos procedimentos técnicos, a pesquisa se delineia como bibliográfica, fundamentada na exploração de estudos científicos relacionados à classificação de dados desbalanceados na área médica. Também assume um caráter documental ao embasar suas previsões em materiais científicos e bases de dados já existentes. Adicionalmente, apresenta um viés operacional ao disponibilizar simulações e ferramentas quantitativas que contribuem significativamente para o processo decisório.

1.3.2 Método de Trabalho

Esta pesquisa consiste em três artigos, cada um direcionado a um objetivo específico em conformidade com o propósito geral do estudo, conforme demonstrado na Tabela 1.

Tabela 1. Estrutura das etapas da pesquisa desenvolvida

Problemas desbalanceados de classificação na área da saúde			
	Artigo 1	**Artigo 2**	**Artigo 3**
Questões de pesquisa	Quais etapas de *workflows* de previsão utilizadas na literatura? Como os autores abordam a interpretabilidade e generalização dos resultados?	Como identificar o impacto das variáveis ao utilizar técnicas tradicionais de *machine learning*?	Quais são as etapas necessárias em um modelo de previsão para reduzir o viés e garantir resultados generalizáveis?
Objetivos	Identificar, a partir de uma revisão sistemática da literatura, as etapas para a classificação de dados desbalanceados e como é abordada a generalização e de interpretabilidade dos modelos. Propor um *framework* para lidar com o desbalanceamento dos dados.	Desenvolver um modelo preditivo de internação de pacientes com diabetes, com enfoque na classificação de dados desbalanceados e interpretabilidade do modelo.	Desenvolver um modelo preditivo de *no-show* de pacientes ambulatoriais, com enfoque na classificação de dados desbalanceados e generalização do modelo.
Método de Pesquisa	Pesquisa qualitativa: 1. Revisão sistemática da literatura	Pesquisa quantitativa: 1. Validação do modelo proposto através de estudo de caso	Pesquisa quantitativa: 1. Validação do modelo proposto através de estudo de caso

O artigo 1 – "*Desafios e Soluções no uso de Machine Learning para Dados de Saúde Desbalanceados: um Workflow para Robustez Preditiva e Interpretabilidade*" – aborda o desbalanceamento nas classes de dados em problemas de classificação em saúde. Este estudo investiga como a literatura abordou essa questão, com um foco especial nas direções futuras de forma a propor um *framework*, além de abordar questões de generalização e de interpretabilidade dos modelos, questões discutidas também nos artigos 2 e 3.

O artigo 2 – "*Machine learning for high-risk hospitalization prediction in outpatient diabetic cases at a tertiary hospital*" – concentra-se na identificação de pacientes com diabetes que possuem maior probabilidade de hospitalização, um desafio crítico para otimizar tratamentos e alocar recursos hospitalares de forma eficiente. O estudo desenvolve um modelo preditivo que destaca os fatores mais relevantes na identificação de pacientes com risco de hospitalização, explorando a questão da interpretabilidade dos algoritmos convencionais de ML.

O Artigo 3 – "*Decision analysis framework for predicting no-shows to appointments using machine learning algorithms*" – aborda a previsão de não comparecimentos a consultas médicas, um problema que afeta a eficiência dos serviços de saúde e a qualidade da assistência prestada. Este artigo destaca a importância de evitar a dependência de um número limitado de validações cruzadas em conjuntos de dados desbalanceados. Além disso, explora o uso de algoritmos avançados de ML (e.g., a *Symbolic Regression (SR)* como modelo de previsão e a *Instance Hardness Threshold (IHT) technique* como estratégia de balanceamento de dados) para melhorar a precisão na previsão de não comparecimentos, explorando questões relacionadas à generalização dos resultados.

1.4 DELIMITAÇÕES DO ESTUDO

Esta pesquisa concentra-se na descoberta de conhecimento em problemas de classificação de dados desbalanceados e estruturados na área médica, visando oferecer soluções por meio de algoritmos tradicionais de ML. É fundamental ressaltar que esta pesquisa não explora modificações nos algoritmos para lidar com o desbalanceamento de dados.

Todas as metodologias e análises apresentadas são exclusivamente baseadas em dados desbalanceados e estruturados, sem incorporar análises de outros tipos de dados (e.g., baseados em opiniões subjetivas) para lidar com o desbalanceamento de classes. Além disso, o escopo desta pesquisa se limita apenas às técnicas supervisionadas dentro deste contexto específico de aplicação. O estudo se focaliza na análise da área médica, com ênfase especial na identificação de padrões de *no-show* de pacientes e na previsão de internações de pacientes com diabetes.

1.5 ESTRUTURA DA TESE

A tese está estruturada de cinco capítulos. No primeiro, aborda-se a contextualização do estudo, seus objetivos e as razões que fundamentam sua relevância tanto no âmbito acadêmico quanto prático. Além disso, são detalhados o método de trabalho, a estrutura e as limitações do estudo. Os capítulos 2 a 4 abrangem os artigos que desenvolvem as propostas delineadas, seguindo a estrutura anteriormente exposta.

Finalmente, o capítulo 5 traz as conclusões da tese e sugestões de direções para futuras pesquisas.

1.6 REFERÊNCIAS

Bai, B., Wu, Z., Weng, S., & Yang, Q. (2023). Application of interpretable machine learning algorithms to predict distant metastasis in osteosarcoma. *Cancer Medicine*, *12*(4), 5025–5034. https://doi.org/10.1002/cam4.5225

Ben-Israel, D., Jacobs, W. B., Casha, S., Lang, S., Ryu, W. H. A., de Lotbiniere-Bassett, M., & Cadotte, D. W. (2020). The impact of machine learning on patient care: A systematic review. *Artificial Intelligence in Medicine*, *103*, 101785. https://doi.org/10.1016/j.artmed.2019.101785

Cabitza, F., & Campagner, A. (2021). The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *International Journal of Medical Informatics*, *153*, 104510. https://doi.org/10.1016/j.ijmedinf.2021.104510

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, *2021*, 123–144. https://doi.org/10.1146/annurev-biodatasci-092820

Cui, S., Wang, D., Wang, Y., Yu, P. W., & Jin, Y. (2018). An improved support vector machine-based diabetic readmission prediction. *Computer Methods and Programs in Biomedicine*, *166*, 123–135. https://doi.org/10.1016/j.cmpb.2018.10.012

Gandhi, J., Narmawala, Z. (2024). A comprehensive survey on machine learning techniques in opportunistic networks: Advances challenges and future directions. *Pervasive and Mobile Computing*, 100, 101917. https://doi.org/10.1016/j.pmcj.2024.101917

Gramacki, A., & Gramacki, J. (2022). A deep learning framework for epileptic seizure detection based on neonatal EEG signals. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-15830-2

Habehh, H., & Gohel, S. (2021). Machine Learning in Healthcare. *Current Genomics*, *22*(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

Haynes, R. B., & Wilczynski, N. L. (2010). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: Methods of a decision-maker-researcher partnership systematic review. *Implementation Science*, 5, 12. https://doi.org/10.1186/1748-5908-5-12

Jia, L., Wang, Z., Lv, S., & Xu, Z. (2022). PE_DIM: An Efficient Probabilistic Ensemble Classification Algorithm for Diabetes Handling Class Imbalance Missing Values. *IEEE Access*, *10*, 107459–107476. https://doi.org/10.1109/ACCESS.2022.3212067

Maheshwari, S., Jain, R. C., & Jadon, R. S. (2018). An insight into rare class problem: Analysis and potential solutions. *Journal of Computer Science, 14(6)*, 777–792. https://doi.org/10.3844/jcssp.2018.777.792

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/nejmp1606181

Rekha, G., Tyagi, A. K., & Reddy, V. K. (2019). A wide scale classification of class imbalance problem and its solutions: A systematic literature review. *Journal of Computer Science, 15(7)*, 886–929. https://doi.org/10.3844/jcssp.2019.886.929

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-48995-4

Valero-Bover, D., González, P., Carot-Sans, G., Cano, I., Saura, P., Otermin, P., Garcia, C., Gálvez, M., Lupiáñez-Villanueva, F., & Piera-Jiménez, J. (2022). Reducing non-attendance in outpatient appointments: predictive model development, validation, and clinical assessment. *BMC Health Services Research*, *22*(1). https://doi.org/10.1186/s12913-022-07865-y

2 ARTIGO 1: DESAFIOS E SOLUÇÕES NO USO DE MACHINE LEARNING PARA DADOS DE SAÚDE DESBALANCEADOS: UM WORKFLOW PARA ROBUSTEZ PREDITIVA E INTERPRETABILIDADE

Resumo

O uso de algoritmos de *Machine Learning* (ML) na área da saúde pode oferecer suporte a diagnósticos e decisões clínicas. No entanto, desafios como o desbalanceamento das classes e a necessidade de modelos que proporcionem boa generalização e interpretabilidade persistem. Este artigo propõe um *workflow* para análise de dados de saúde desbalanceados, focando em etapas que promovem a robustez preditiva, interpretabilidade e a capacidade de generalização dos modelos, importantes para a confiabilidade das previsões em novos contextos. A partir de uma revisão da literatura, são mapeadas e analisadas etapas de pré-processamento, seleção de variáveis, reamostragem, aplicação de algoritmos de ML e validação, destacando os efeitos dessas etapas na mitigação de vieses e na sensibilidade na detecção de eventos de interesse. Abordagens para tornar algoritmos complexos mais interpretáveis e como o problema de generalização dos algoritmos de ML é abordado nas pesquisas analisadas também são discutidas. Os resultados indicam que o uso adequado de reamostragem e validação cruzada estratificada melhora a generalização e que técnicas de interpretabilidade como o *Shapley Additive Explanations* (SHAP) oferecem *insights* úteis para uso clínico. No entanto, foi identificado que a aplicação inconsistente das técnicas de pré-processamento e reamostragem, assim como o uso intensivo de algoritmos "caixa-preta", limitam a confiabilidade e a adoção desses modelos em contextos reais. Adicionalmente, a falta de métodos robustos para avaliar a generalização em classes minoritárias apresenta uma barreira para garantir previsões consistentes.

2.1 INTRODUÇÃO

O uso disseminado de algoritmos de *Machine Learning* (ML) na área da saúde desempenha um papel importante em diversas aplicações, como previsão de diagnósticos (Jia et al., 2022), prognósticos (Bai et al., 2023) e otimização de sistemas de consulta (Cui

et al., 2018). No entanto, essas aplicações frequentemente enfrentam o desafio do desequilíbrio nas classes de dados. Em contextos médicos, os desfechos de interesse são, frequentemente, eventos raros, tornando a classe minoritária, que representa tais eventos, fundamental para a detecção precoce e medidas proativas (Jia et al., 2022).

Os algoritmos tradicionais de ML operam sob a suposição de que as classes de dados estão equilibradas, o que leva à atribuição de custos iguais para erros de classificação, visando à minimização de erros gerais durante o treinamento (Rekha et al., 2019). No entanto, quando ocorre desequilíbrio nos dados, os modelos de classificação tendem a favorecer a classe majoritária, resultando em desempenho insatisfatório e detecção inadequada da classe minoritária, o que é especialmente problemático quando o evento de interesse corresponde a esta classe (Maheshwari et al., 2018). A baixa sensibilidade do algoritmo compromete a qualidade das decisões, algo crítico em aplicações do mundo real, especialmente na área médica. Desconsiderar o desequilíbrio entre classes podem resultar em diagnósticos tardios, tratamentos inadequados e impactos negativos na qualidade de vida dos pacientes. Nesse contexto, o desenvolvimento de sistemas inteligentes capazes de lidar com distribuições desequilibradas de classes e superar vieses é fundamental, constituindo o aprendizado a partir de dados desbalanceados (Ben-Israel et al., 2020; Haynes & Wilczynski, 2010).

A literatura frequentemente sugere estratégias para lidar com o desbalanceamento tanto a nível de dados, (e.g., através do uso de técnicas de reamostragem) quanto a nível de algoritmo, propondo modificações (Maheshwari et al., 2018; Spelmen & Porkodi, 2018; Susan & Kumar, 2021). Contudo, uma abordagem apropriada vai além da compreensão de técnicas e desenvolvimento de novos algoritmos, exigindo uma compreensão completa do *workflow* de previsão, que garante a generalização eficaz e a interpretabilidade dos classificadores, dificultadas em função do desbalanceamento das classes.

Em um editorial para o *International Journal of Medical Informatics*, Cabitza & Campagner (2021) enfatizam que praticantes e acadêmicos da área de ML médico frequentemente focam apenas no desempenho dos algoritmos. Cabitza & Campagner (2021) alertam que o acesso aos dados não é motivo suficiente para aplicar algoritmos de

ML no campo da medicina e que fatores críticos, como desequilíbrios nas classes de dados, e etapas essenciais de classificação, como a divisão de dados em treinamento e teste, não devem ter sua importância subestimada.

Esse artigo investiga como a literatura no campo médico tem abordado problemas de classificação considerando a presença de classes desbalanceadas, e propõe um *workflow* para classificação desbalanceada, enquanto avalia a importância da generalização e da interpretabilidade nesses contextos. O aprimoramento da prática médica e dos resultados de saúde é o motivador desta pesquisa, que busca promover uma avaliação mais precisa e criteriosa das ferramentas de informática médica, essencial para garantir que os avanços tecnológicos resultem em benefícios tangíveis para os pacientes e prática clínica.

2.2 PESQUISAS ANTERIORES E CONTRIBUIÇÕES

Revisões de literatura disponíveis sobre o tema, apresentadas na Tabela 2, exploram predominantemente abordagens destinadas a lidar com o problema do desbalanceamento de classes a nível de dados, algoritmos ou abordagem híbrida.

A nível de dados, as técnicas de reamostragem equilibram a distribuição das instâncias entre as classes, enquanto a seleção de variáveis prioriza atributos relevantes. A nível de algoritmo, ajustes nos pesos das classes, modificações e combinações de algoritmos (ex., *Ensembles*) são frequentemente reportadas. Já a abordagem híbrida consiste na combinação de técnicas de reamostragem com modificações nos algoritmos. No entanto, a literatura carece de informações sobre o momento ideal para aplicar essas técnicas e de um *workflow* completo de análise para a classificação de dados desbalanceados.

Benhar et al. (2020) argumentam que investigar a sequência de aplicação das técnicas pode influenciar os resultados da previsão. O desenvolvimento de um *workflow* para lidar com o desbalanceamento dos dados impacta também aspectos sobre a interpretabilidade dos resultados, uma vez que os algoritmos podem desconsiderar informações cruciais das classes minoritárias, resultando em uma interpretação incompleta ou distorcida dos dados.

Tabela 2. Revisões de literatura sobre o tema

Autores	Título	Descrição	Dominios de aplicação	Apresenta discussões sobre generalização	Apresenta discussões sobre a interpretabilidade
Haixiang *et al.* (2017)	*Learning from class-imbalanced data: Review of methods and applications*	A revisão examina estratégias para lidar com conjuntos de dados desbalanceados, expl*orando* técnicas de reamostragem, seleção e extração de variáveis, aprendizado sensível ao custo e algoritmos de classificação, incluindo *ensembles.*	diferentes áreas	não	não
Basha (2022)	*A Review on Imbalanced Data Classification Techniques*	Discute principais vantagens e desvantagens de técnicas de reamostragem aplicadas na literatura. Revisa alguns classificadores como *ensemble* e *bagging*.	diferentes áreas	não	não
Benhar *et al.* (2020)	*Data preprocessing for heart disease classification: A systematic literature review*	Discute técnicas de pré-processamento, destacando a importância da remoção de ruídos e normalização. O estudo sugere explorar a sequência de aplicação dessas técnicas para aprimorar a classificação. Também sugere tornar classificadores caixa-preta, como *Artificial Neural Networks* (ANN) e *Support Vector Machines* (SVM), mais interpretáveis.	doenças cardíacas	não	não
Ganganwar (2012)	*An overview of classification algorithms for imbalanced datasets*	Revisa abordagens a nível de dados e de algoritmo, e abordagens híbridas, destacando suas vantagens e desvantagens.	diferentes áreas	não	não
Kaur *et al.* (2019)	*A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions*	Discute algoritmos de ML, cobrindo estratégias a nível de dados e de algoritmo, e abordagens híbridas.	diferentes áreas	não	não
Vimalraj & Porkodi (2018)	*A Review on Handling Imbalanced Data*	Discute abordagens a nível de dados e de algoritmo, e abordagens híbridas.	diferentes áreas	não	não
Rezvani & Wang (2023)	*A broad review on class imbalance learning techniques*	Discute abordagens a nível de dados e de algoritmo, e abordagens híbridas. Revisa detalhadamente o uso do SVM para classificação desbalanceada.	diferentes áreas	não	não

Autores	Título	Descrição	Dominios de aplicação	Apresenta discussões sobre generalização	Apresenta discussões sobre a interpretabilidade
Maalouf & Trafalis (2011)	*Rare events and imbalanced datasets: an overview*	Discute abordagens a nível de dados e de algoritmo, e técnicas Baseadas em Kernel.	diferentes áreas	não	não
Maheshwari *et al.*, (2018)	*An Insight into Rare Class Problem: Analysis and Potential Solutions*	Discute os algoritmos de ML, cobrindo estratégias a nível de dados e de algoritmo, e abordagens híbridas.	diferentes áreas	não	não
Krawczyk (2016)	*Learning from imbalanced data: Progress in open challenges and future directions*	Discute abordagens a nível de dados e de algoritmo, e abordagens híbridas. Discute problemas de classificação, *clustering*, dados gerados em tempo real e *big data*.	diferentes áreas	não	não
Susan & Kumar (2021)	*The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art*	Discute abordagens a nível de dados.	diferentes áreas	não	não
He & Garcia (2009)	*Learning from Imbalanced Data*	Discute oportunidades e desafios sobre abordagens a nível de dados e técnicas baseados em Kernel.	diferentes áreas	não	não
Sun *et al.* (2009)	*Classification of imbalanced data: a review*	Revisão sobre os domínios de aplicação, natureza do problema, dificuldades dos algoritmos de aprendizagem padrão, medidas de avaliação, soluções de pesquisa relatadas, e o problema de desequilíbrio de classes na presença de várias classes.	diferentes áreas	Não, mas cita a importância de os dados de treinamento serem generalizáveis para os dados de teste.	não
Rekha *et al.* (2019)	*A Wide Scale Classification of Class Imbalance Problem and its Solutions: A Systematic Literature Review*	Identifica, classifica e compara métodos de pesquisa para lidar com esse problema. Analisa as forças e fraquezas das técnicas em cada nível (dados, algoritmos e abordagens híbridas).	diferentes áreas	não	não
Ghavidel & Pazos (2023)	*Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: a systematic review*	Discute sobre o pré-processamento, modelos preditivos robustos para classificação e as métricas de avaliação de resultados.	câncer de mama	não	não
Stafford *et al.* (2022a)	*A Systematic Review of Artifcial Intelligence and Machine Learning Applications to Infammatory Bowel Disease, with Practical Guidelines for Interpretation*	Avalia tipos de dados, aplicações e algoritmos comuns no campo de ML para doenças inflamatórias intestinais e avalia as mudanças na área nos últimos anos. Apresenta uma breve discussão a nível de dados e algoritmos.	Doença Inflamatória Intestinal	Sugestão de pesquisas futuras: construção de uma ferramenta para avaliar o potencial viés do *workflow* de ML.	Mencionam algoritmos que trazem um grau de interpretabilidade, e.g., arvores de decisão.

Além do *workflow*, Benhar et al. (2020) propõem a exploração de técnicas de interpretabilidade que tornem os algoritmos "caixa-preta", como *Artificial Neural Networks* (ANN), mais acessíveis e compreensíveis para os usuários finais, fornecendo uma visão mais clara do processo de tomada de decisão à partir do modelo.

Outra lacuna na literatura está associada à escassez de discussões sobre como avaliar a generalização dos resultados de classificação, a qual também pode ser comprometida em conjuntos de dados desbalanceados, uma vez que os modelos podem ter dificuldade em generalizar padrões das classes minoritárias devido à sua sub-representação nos dados de treinamento. Alguns autores (e.g., Stafford et al., 2022), sugerem que criar ferramentas para avaliação da generalização pode ser importante para minimizar viéses nos *workflows* de ML, especialmente em contextos clínicos, visando criar modelos mais robustos.

2.3 MÉTODO

Para propor um *workflow* de classificação para lidar com desbalanceamento dos dados, generalização e interpretabilidade de resultados, foi realizado um mapeamento de técnicas e *workflows* já utilizados na literatura. Seguiu-se as cinco etapas propostas por Arksey & O'Malley (2005): (i) elaborar questões de pesquisa, (ii) identificar estudos a elas relacionados, (iii) selecionar estudos relevantes, (iv) mapear dados, e (v) relatar os resultados.

O delineamento do *workflow* foi orientado pelas seguintes questões de pesquisa:

Q1. Como problemas de predição e classificação que lidam com dados desbalanceados são abordados na área da saúde? Como os autores que abordaram os problemas delineiam seus *workflows* analíticos para conduzir essas classificações considerando estratégias a nível de dados e híbridas?

Q2. Quais discussões, métodos, justificativas ou análises são propostas pelos autores para a interpretabilidade dos resultados da classificação em contextos de dados desbalanceados?

Q3. Quais discussões, métodos, justificativas ou análises são apresentados pelos autores em relação à generalização dos resultados ao lidar com classes de dados desbalanceadas?

Embora o foco da pesquisa esteja no desenvolvimento de um *workflow* genérico de previsão, como forma de garantir robustez no mapeamento de técnicas e *workflows* disponíveis pela literatura, seguiu-se o protocolo PRISMA (Figura 1) (Moher et al., 2009) para selecionar artigos, aplicando critérios como exclusão de duplicatas, foco em publicações revisadas por pares em inglês, e alinhamento do título, resumo e conteúdo com o tema.

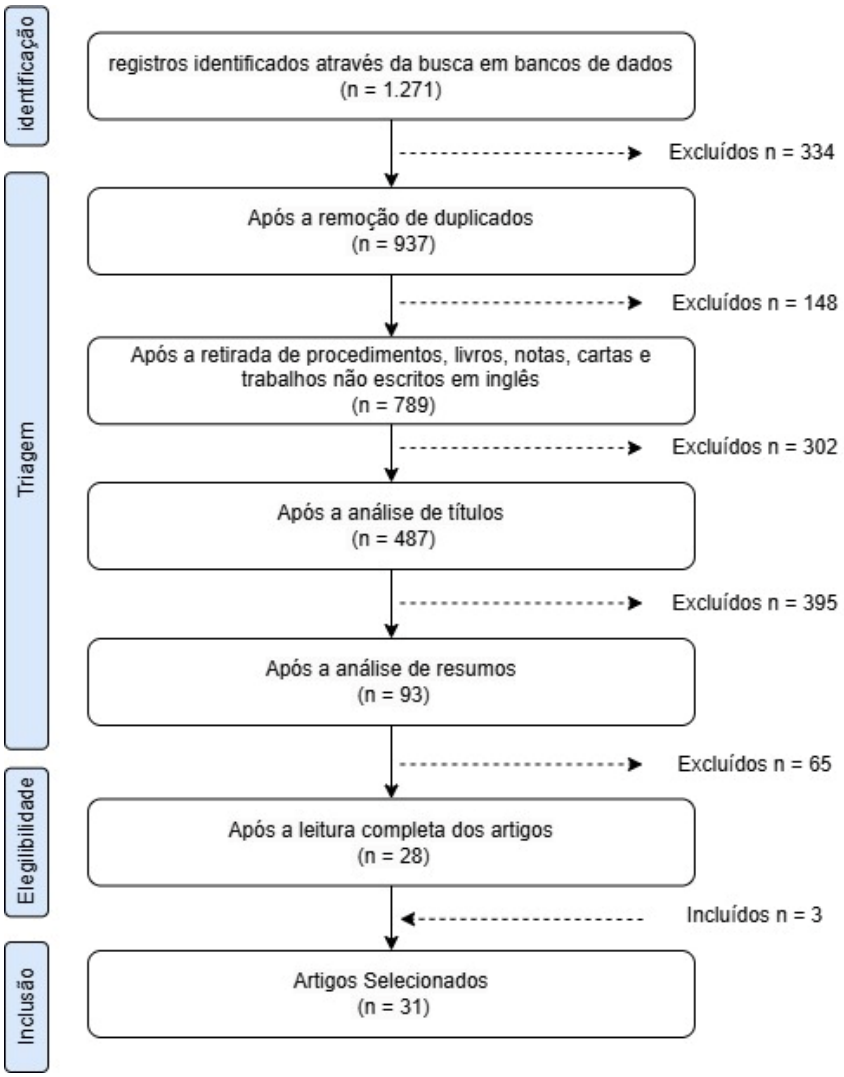

Figura 1. Fluxograma PRISMA

Foram realizadas buscas nas bases de dados Scopus, Web of Science e Pubmed, considerando as seguintes dimensões de pesquisa apresentadas na Tabela 3: 1. *prediction*, 2. *rare events*, 3. *health*, e 4. *interpretability or generalizability of the results*. Para combinar as dimensões de pesquisa, utilizou-se o operador booleano *AND*, e o sufixo de algumas palavras foi seguido de * a fim de capturar suas variações, buscou-se estudos publicado até maio de 2024. A busca resultou em 1.271 publicações.

Tabela 3. Dimensões de pesquisa e palavras-chaves

Dimensão 1: *prediction*		**Dimensão 2: *rare events***		**Dimensão 3: *health***		**Dimensão 4: *interpretability or generalizability***
predict *classifi** *learn**	***AND***	*"rare events"* *"imbalanced data"* *"unbalanced data"* *"skewed data"*	***AND***	*medical* *disease* *diagnos** *hospital* health*	***AND***	*general** *explain** *interpret**

Na fase de elegibilidade, restaram 937 artigos após a remoção de duplicados, 487 após a leitura de títulos, 93 após a análise de resumos e ao final foram mantidos 28 artigos que abordavam questões de generalização e interpretabilidade de resultados de previsão, excluindo aqueles que não tratavam dessas questões ou utilizavam apenas dados não estruturados. Mais 3 artigos foram incluídos após análise de referências, totalizando 31 estudos. Esses artigos foram mapeados e organizados em uma planilha detalhada, oferecendo uma visão abrangente sobre as técnicas e *workflows* utilizados nas previsões.

Foram detalhadamente analisadas as etapas de pré-processamento e modelagem dos dados e avaliação dos resultados de classificações, englobando o tipo de classificação realizada (binária ou multiclasse), proporção da classe minoritária nos dados, aplicação de técnicas de normalização, métodos usados para tratar dados ausentes, remoção de ruídos/*outliers*, correlações entre variáveis, natureza dos dados (e.g., categóricos, contínuos), codificação de dados categóricos, seleção de variáveis, comparação de técnicas de seleção de variáveis, algoritmos de ML aplicados, métodos para parametrização dos algoritmos de ML, atribuição de pesos para lidar com classes desbalanceadas, técnicas de reamostragem, estratificação por classe, divisão dos dados (treino, validação, teste), validação cruzada, interpretabilidade no contexto de ML e estratégia de generalização dos resultados.

2.4 RESULTADOS

Endereçando a questão de pesquisa Q1, explorou-se o processo para lidar com dados desbalanceados na área médica, visando avaliar o viés das aplicações de ML. Visa-se avaliar a consistência das etapas do processo, minimizando o viés e alcançando resultados confiáveis, generalizáveis e interpretáveis. No Apêndice A1 são apresentadas as etapas e algoritmos utilizados pelas pesquisas da literatura.

A Figura 2 traz a proposta de um *workflow* genérico que reduza possíveis vieses de previsão e melhore a interpretabilidade dos resultados.

A fase inicial corresponde ao pré-processamento de dados, contemplando as etapas de limpeza, transformação, balanceamento da base de dados e seleção de variáveis. Essas etapas garantem a integridade dos dados, facilitam a extração de padrões e reduzem o viés, gerando resultados mais precisos.

A etapa de limpeza de dados tem como objetivo corrigir erros, valores ausentes ou inconsistentes. A remoção de valores ausentes ou inconsistentes pode ser viável quando estes representam uma proporção pequena do conjunto de dados total, mas pode levar à perda de informações relevantes em caso contrário (Jia et al., 2022; Moslehi et al., 2022). Em contrapartida, preencher os valores ausentes com estimativas estatísticas pode introduzir ruído, especialmente quando a proporção de valores ausentes é alta (Jia et al., 2022). Nesses casos, técnicas de imputação mais avançadas, baseadas em algoritmos de ML são preferíveis, pois preservam melhor a estrutura dos dados (Jia et al., 2022).

Quando a base de dados inclui variáveis quantitativas e qualitativas, é importante realizar transformações nos dados para assegurar a interpretabilidade e eficácia dos algoritmos (Benhar et al., 2020; Ghavidel & Pazos, 2023). A normalização reescreve variáveis quantitativas em uma mesma escala, evitando que aquelas com magnitudes absolutas elevadas dominem o processo de modelagem (Cui et al., 2018). Algoritmos sensíveis à escala, como modelos lineares, podem ser influenciados pela amplitude das variáveis, resultando em coeficientes distorcidos e comprometendo a interpretabilidade dos resultados (Saarela et al., 2019). Transformações como *z-score* e *min-max* são

empregadas para normalizar os dados em uma escala comum (Kokkotis et al., 2022; Su et al., 2022).

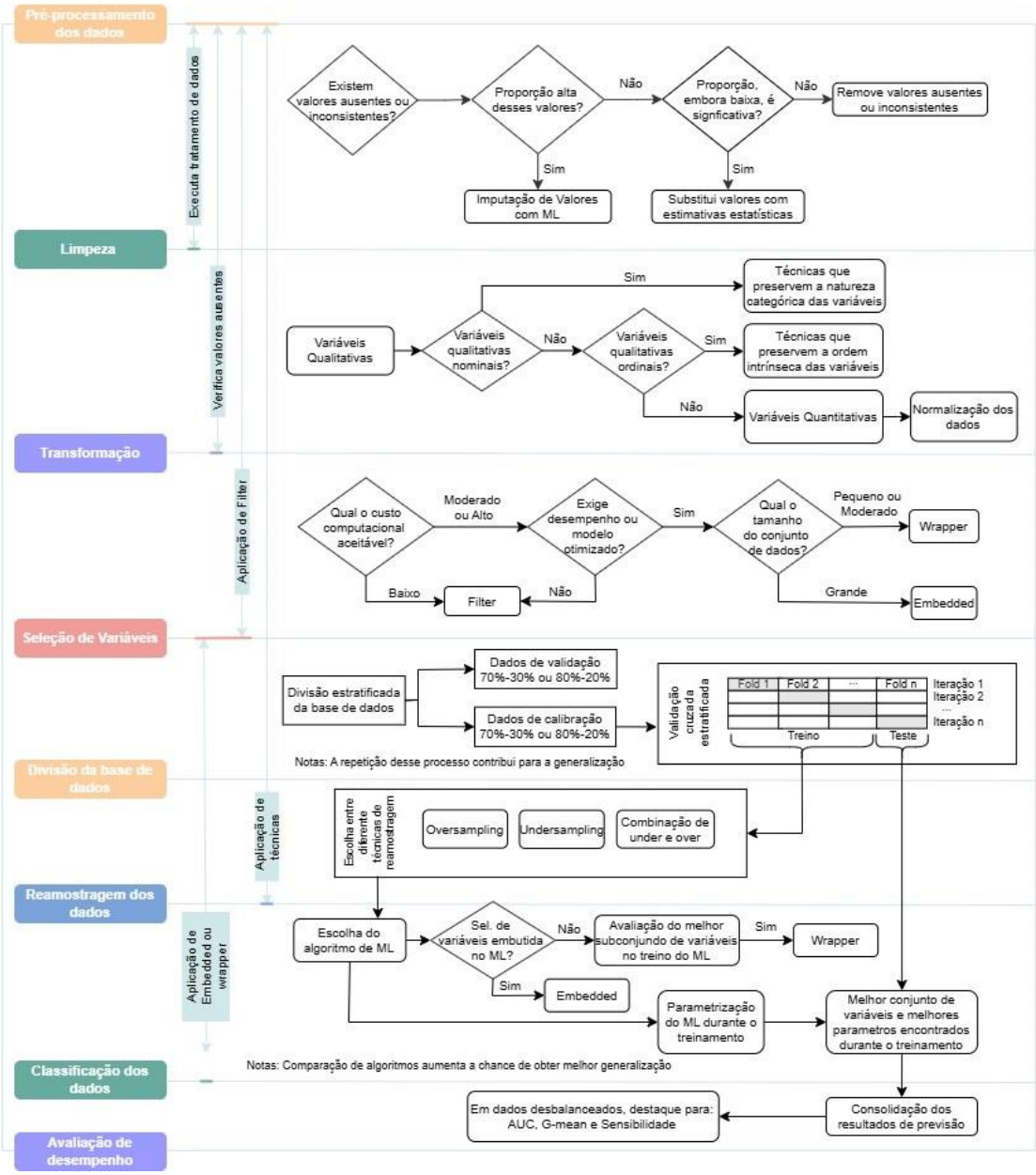

Figura 2. Workflow para classificação desbalanceada

Para variáveis qualitativas nominais, é importante utilizar técnicas de codificação apropriadas, pois os algoritmos de ML exigem entradas numéricas (Sedighi-Maman & Heath, 2022). O *One-Hot Encoding* é recomendado para manter a natureza categórica das variáveis, assegurando que a informação seja preservada sem impor ordens artificiais entre as categorias, o que mantém a interpretabilidade dos resultados (Sedighi-

Maman & Heath, 2022). No entanto, como apontado por Sedighi-Maman & Mondello (2021), a codificação *one-hot* aumenta o número de variáveis, resultando em maior complexidade do modelo e aumento no tempo de treinamento. Métodos de seleção de variáveis podem ser necessários para gerenciar essa complexidade.

A seleção de variáveis pode ser realizada ainda no pré-processamento quando se empregam métodos *filter*. Esses métodos avaliam, a partir de medidas estatísticas, a relevância de cada variável em relação à variável de saída (Johnson et al., 2022; Park et al., 2018). Embora existam também os métodos *embedded* e *wrapper* para a seleção de variáveis, estes não são considerados parte do pré-processamento dos dados. Sendo aplicados em fases posteriores do *workflow*, com o objetivo de otimizar o desempenho dos algoritmos de ML (Cui et al., 2018). Os métodos *embedded* integram a seleção de variáveis diretamente no processo de treinamento do modelo, ajustando pesos ou penalidades conforme necessários. Os métodos *wrapper* testam diferentes combinações de variáveis treinando e avaliando modelos repetidamente, selecionando o conjunto que proporciona o melhor desempenho. A escolha do método de seleção de variáveis depende das características do problema, das restrições computacionais e das metas de desempenho do modelo. Se a eficiência computacional for uma preocupação, métodos *filter* podem ser preferíveis, enquanto métodos *embedded* ou *wrapper* podem ser mais adequados se a precisão do modelo for priorizada e recursos computacionais adequados estiverem disponíveis (Ghavidel & Pazos, 2023; Cui et al., 2018).

Outra etapa importante do pré-processamento de dados é a reamostragem dos dados. Algoritmos treinados em conjuntos desbalanceados tendem a favorecer a classe majoritária, resultando em baixa sensibilidade e altas taxas de falsos negativos para a classe minoritária, a qual geralmente está relacionada ao desfecho de interesse, e.g., diagnósticos em que é mais importante identificar corretamente uma pessoa que está com a condição investigada (Cui et al., 2018; Jafari et al., 2022). Técnicas de reamostragem, como *undersampling*, *oversampling* e combinações de *undersampling* e *oversampling*, buscam mitigar esse viés, reduzindo erros de treinamento e melhorando a generalização do modelo. Jafari et al. (2022), Liu et al. (2022) e Chan et al. (2017) alertam que o uso de técnicas de reamostragem pode gerar confusão e desengajamento dos tomadores de decisão, pois esse processo cria observações artificiais ou oculta dados reais. Contudo,

Wang et al. (2021a) afirmam que essas preocupações podem ser atenuadas se a reamostragem for aplicada no momento adequado, ou seja, exclusivamente nos dados de treinamento. Dessa forma, evita-se o vazamento de informações, preservando a integridade dos dados de validação (Zhang et al., 2022). Ao contrário do que foi inicialmente sugerido por Chan et al. (2017), nessas condições a reamostragem auxilia no processo de aprendizado sem comprometer a interpretabilidade dos resultados.

Antes de realizar o pré-processamento, é necessário conhecer o funcionamento dos algoritmos de ML selecionados, pois nem sempre é necessário realizar todas as suas etapas de pré-processamento. Moslehi et al. (2022) mencionam que conduzir análises sem considerar valores ausentes em alguns algoritmos, *e.g.* regressão logística, pode gerar viés nos resultados e tornar algumas análises impossíveis. Em contraste, existem outros modelos, *e.g.*, Chan et al. (2017)'s *Prediction with Pattern Discovery* (ou Previsão com Descoberta de Padrões), que são tolerantes a dados ruidosos e faltantes.

Nas pesquisas selecionadas nessa revisão, não foi identificada uma sequência sugerida para a aplicação das técnicas de pré-processamento. Contudo, Benhar et al. (2020) sugerem a investigação da melhor sequência quando múltiplas tarefas ou técnicas de pré-processamento são empregadas, já que ela pode influenciar os resultados das previsões. É importante explorar como as técnicas de pré-processamento podem aprimorar a interpretabilidade dos classificadores, visto que a aplicação inadequada dessas técnicas pode resultar na exclusão de características importantes para a previsão, a exemplo das transformações ordinais em variáveis nominais, inadvertidamente atribuindo uma ordem de importância onde esta não existe (Sedighi-Maman & Heath, 2022).

A etapa subsequente ao pré-processamento dos dados corresponde à divisão da base de dados em porções de treinamento, teste e validação. Na literatura, essa etapa foi realizada de diferentes formas, mas comumente é sugerido primeiro dividir aleatoriamente o conjunto de dados em treino e teste (Brahma & Mukherjee, 2021), sendo que as proporções mais utilizadas são 70%/30% (Bai et al., 2023; Johnson et al., 2022; Luo et al., 2020; Moslehi et al., 2022; Saarela et al., 2019; Sedighi-Maman & Heath, 2022; Sedighi-Maman & Mondello, 2021; Wang et al., 2021b, 2023; Zhang et al., 2022;

Brahma & Mukherjee, 2021) e 80%/20% (Feng et al., 2022; Hakkoum, Idri, & Abnane, 2024; Hu et al., 2023; Liu et al., 2022; Yu et al., 2024). Posteriormente, a maioria das pesquisas aplicou a validação cruzada apenas nos dados de treino. Essa técnica divide os dados em k *folds* (i.e., partes iguais) e o modelo é treinado k vezes usando uma parte diferente como conjunto de teste e as $k - 1$ partes restantes como conjunto de treino; k = 5 e 10 são os valores mais utilizados (Bai et al., 2023; Feng et al., 2022; Hakkoum, Idri, & Abnane, 2024; Hu et al., 2023; Lu et al., 2022; Saarela, et al., 2019; Sedighi-Maman & Heath, 2022; Sedighi-Maman & Mondello, 2021; Wang et al., 2021a; Wang et al., 2021b; Yu et al., 2024; Johnson et al., 2022). Johnson et al. (2022) apontam que valores de k abaixo de 5 na validação cruzada aumentam a variância dos resultados.

Se os dados apresentarem classes altamente desequilibradas, a aplicação da estratificação durante a divisão, para garantir que todas as classes estejam representadas de forma proporcional em todos os conjuntos de dados, é fortemente recomendada. Omitir essa etapa pode resultar em instabilidade nas métricas de avaliação entre diferentes dobras e maior chance de obter modelos fracos ou sub-ótimos, especialmente na classificação de classes minoritárias (Wang et al., 2021a). A discussão sobre divisão dos dados e a forma como é avaliada a generalização dos modelos é aprofundada mais à frente neste artigo.

Na etapa de classificação, aplicam-se algoritmos de ML. Nesta revisão, não são discutidos os algoritmos mais comumente utilizados na literatura e de melhor desempenho, pois essas questões já foram abordadas em outros estudos (Tabela 2). O desempenho dos algoritmos pode variar de contexto para contexto. Quando não se tem uma pré-suposição sobre a distribuição dos dados, por exemplo, comparar diferentes algoritmos pode ser uma estratégia mais informativa do que simplesmente selecionar aqueles que já tiveram melhor desempenho em pesquisas anteriores (Bai et al., 2023; Wang et al., 2023). Outra etapa importante na classificação é a parametrização dos algoritmos, em sua maioria são não paramétricos, ou seja, não assumem dados seguindo alguma distribuição de probabilidade conhecida. Assim, é importante a escolha adequada dos parâmetros para garantir um bom desempenho dos modelos (Bai et al., 2023; Cui et al., 2018). Na literatura, validação cruzada e *grid-search* são utilizados para selecionar as melhores configurações de parâmetros (Cui et al., 2018; Kokkotis et al., 2022; Moslehi et al., 2022; Saarela et al., 2019).

O desempenho das previsões pode ser avaliado por diversas métricas (Brahma & Mukherjee, 2021). Em dados desbalanceados, a escolha da métrica é importante. Em problemas de classificação, a classe minoritária geralmente está relacionada ao desfecho de interesse e minimizar seus erros é essencial. Por exemplo, diagnosticar erroneamente um indivíduo doente como saudável (falso negativo) é mais grave que o contrário, e distinguir entre falsos positivos (erro do tipo I) e falsos negativos (erro do tipo II) é fundamental (Cui et al., 2018; Jafari et al., 2022; Kokkotis et al., 2022; Liu et al., 2022). Em cenários reais, a maioria dos algoritmos de ML não consegue minimizar, de forma simultânea e satisfatória, os erros do tipo I e II (Liu et al., 2022). Existe uma relação quase inversa entre as taxas de falsos positivos e de falsos negativos (Brahma & Mukherjee, 2021), i.e., reduzir a taxa de falsos negativos pode aumentar a taxa de falsos positivos e vice-versa, criando dificuldade em maximizar simultaneamente métricas como precisão e sensibilidade (Jafari et al., 2022). Métricas como o *F1-Score*, cujo cálculo utiliza a média harmônica entre precisão e sensibilidade, podem ser prejudicadas em situações de desbalanceamento extremo dos dados.

De acordo com Kokkotis et al. (2022), quando o conjunto de dados é extremamente desbalanceado é possível que algumas métricas informem mais sobre a distribuição de classes do que sobre o desempenho real dos modelos treinados. A acurácia tende a favorecer a classe majoritária, distorcendo o desempenho real do modelo; e.g., quando 90% das observações pertencem à classe majoritária, um modelo que classifica todas as pessoas como saudáveis vai apresentar uma acurácia de 90%, embora não tenha feito nenhuma distinção entre as classes (Jafari et al., 2022; Johnson et al., 2022; Santos et al., 2022). Kokkotis et al. (2022) recomendam considerar diversas métricas ao selecionar o algoritmo ideal de ML. Entre elas, destaca-se a *Area Under the Receiver Operating Characteristic Curve* (AUC) (Hu et al., 2023; Jafari et al., 2022; Jia et al., 2022; Lu et al., 2022; Park et al., 2018; Santos et al., 2022; Su et al., 2022; Zhang et al., 2022; Brahma & Mukherjee, 2021; Wang et al. 2021b) que oferece uma avaliação abrangente da capacidade discriminativa do modelo. Além disso, o *G-Mean*, que mede a média geométrica entre a sensibilidade e a especificidade (Cui et al., 2018; Kokkotis et al., 2022; Santos et al., 2022; Sedighi-Maman & Heath, 2022; Sedighi-Maman & Mondello, 2021b; Wang et al., 2023), surge como uma alternativa à AUC, uma vez que equilibra de forma eficaz o desempenho nas classes majoritária e minoritária. A

sensibilidade (Cui et al., 2018), que representa a proporção de casos positivos que são corretamente identificados como positivos pelo teste ou modelo em questão, também deve ser considerada. No diagnóstico, a sensibilidade mede a proporção de doentes corretamente identificados, assegurando que os pacientes recebam mais cedo o tratamento adequado (Cui et al., 2018).

Na sequência, explora-se o tema da interpretabilidade dos resultados de classificação, endereçando a questão de pesquisa Q2. A interpretabilidade em ML é a capacidade de entender como um modelo gera previsões, essencial para auxiliar a tomada de decisão. A partir de 2019 houve um aumento notável no interesse pela interpretabilidade na área médica, refletindo a necessidade de modelos mais transparentes neste domínio (Hakkoum, Idri, & Abnane, 2024).

No contexto da saúde, a capacidade dos modelos de ML em traçar fronteiras de classificação é relevante, mas a interpretabilidade clínica permite que os profissionais entendam os resultados de classificação (Chan et al., 2017). A interpretabilidade facilita a integração do conhecimento clínico nos padrões identificados pelos modelos, possibilitando decisões informadas e a personalização do cuidado com base nas características individuais dos pacientes (Chan et al., 2017; Mridha et al., 2023; Su et al., 2022).

Luo et al. (2020) discutem que a interpretabilidade atua como um sistema de apoio à decisão, auxiliando os profissionais de saúde a identificarem rapidamente fatores de risco e a planejar intervenções personalizadas, sem substituir o julgamento clínico. Essa abordagem otimiza o tempo dos profissionais e reduz o risco de negligenciar intervenções. Em certos cenários, a interpretabilidade pode ser priorizada em detrimento do poder preditivo, de modo que um modelo menos preciso possa ser preferível se oferecer uma compreensão mais clara da influência das variáveis na ocorrência de um evento (Santos et al., 2022). Essa preferência reflete a importância de compreender as relações subjacentes nos dados, ao invés de confiar apenas em previsões acuradas, mas pouco interpretáveis.

Algoritmos tradicionais como regressão logística (Zhang et al., 2022) *Least Absolute Shrinkage and Selection Operator* (LASSO) (Wang et al., 2021b), *Naïve Bayes*

(Chan et al., 2017), *Generalized Linear Model* (GLM) (Sedighi-Maman & Heath, 2022; Sedighi-Maman & Mondello, 2021) e *Decision Trees* (Zhang et al., 2022) têm sido amplamente adotados na prática clínica devido à sua interpretabilidade inerente. Esses algoritmos fornecem coeficientes ou rankings de importância das variáveis, o que permite aos profissionais entenderem como cada variável influencia as previsões (Brahma & Mukherjee, 2021; Chan et al., 2017). Essa transparência é essencial para validar os algoritmos e garantir que suas decisões sejam alinhadas com o conhecimento clínico existente.

Por outro lado, algoritmos considerados "caixas-pretas", como ANN, *Random Forest* e SVM, embora apresentem alto desempenho preditivo (Moslehi et al., 2022), enfrentam desafios de implementação devido à sua falta de transparência (Hakkoum, Idri, & Abnane, 2024; Jafari et al., 2022; Chan et al., 2017). A opacidade desses modelos dificulta a compreensão das decisões, o que pode gerar desconfiança e dificultar a adoção clínica.

Para mitigar o *trade-off* entre interpretabilidade e desempenho preditivo, observa-se um esforço crescente no desenvolvimento de técnicas que proporcionam interpretabilidade sem comprometer significativamente a precisão dos algoritmos de ML. Uma abordagem proeminente é a aplicação de técnicas interpretáveis em conjunto com qualquer algoritmo de ML (Hakkoum, Idri, & Abnane, 2024). Dentre essas técnicas, o *Shapley Additive Explanations* (SHAP), proposto por Lundberg e Lee (2017), tem ganhado destaque, sendo utilizado em 12 estudos nesta revisão de literatura.

O SHAP baseia-se na teoria dos valores de *Shapley*, um conceito da teoria dos jogos, e fornece uma análise de como cada variável de entrada de um algoritmo de ML contribui individualmente para a previsão. Para determinar a influência de uma variável específica, o SHAP calcula a variação na previsão de um algoritmo ao incluir ou excluir essa variável, levando em conta sua interação com outras variáveis. Assim, ele permite uma avaliação contextual da relevância de cada variável (Lundberg & Lee, 2017). O processo do SHAP gera uma matriz, denominada matriz de valores SHAP, em que as linhas representam as observações (ou instâncias) do conjunto de dados e as colunas representam as variáveis do modelo. Cada célula dessa matriz contém um valor SHAP,

que quantifica a contribuição de uma variável específica para a previsão individual daquela observação. Por exemplo, para uma previsão de diagnóstico, um valor SHAP elevado para uma variável como “idade” pode indicar que essa variável exerce um peso significativo no resultado para um paciente específico (Lundberg & Lee, 2017). Essa técnica é particularmente útil para algoritmos complexos, nos quais a interpretabilidade é geralmente limitada.

Na literatura, o SHAP tem sido empregado para trazer interpretabilidade a diversos algoritmos de ML, incluindo *Random Forest* (Bai et al., 2023; Su et al., 2022; Mridha et al., 2023; Hendawi et al., 2023), *Multi-Layer Perceptron* (MLP) (Kokkotis et al., 2022), *Deep Learning* (Jafari et al., 2022), *XGBoost* (Feng et al., 2022; Lombardi et al., 2023; Yu et al., 2024; Wang et al. 2021a) e *Ensemble* (Hu et al., 2023). Outras técnicas de interpretabilidade foram exploradas com diferentes graus de eficácia em Hakkoum, Idri, & Abnane (2024). A técnica *global surrogate* envolve a criação de um modelo simples e interpretável. Especificamente, um modelo simples—como *decision trees*—é treinado para prever as mesmas saídas que um modelo complexo a partir das mesmas entradas. Esse modelo substituto simplificado serve como uma representação aproximada do modelo original, permitindo compreender de forma geral como o modelo complexo toma suas decisões. No entanto, pode haver perda de precisão e fidelidade nas explicações, especialmente quando o modelo simples não captura todas as nuances e interações presentes no modelo original (Molnar, 2019).

A técnica de *Accumulated Local Effects* (ALE) mede o efeito de uma variável ao longo do domínio dos dados, permitindo *insights* sobre o impacto médio de variáveis, embora sua aplicação seja mais limitada a questões de interpretabilidade global (Apley & Zhu, 2020). Já o *Local Interpretable Model-agnostic Explanations* (LIME) cria explicações locais ao ajustar modelos lineares em torno de pontos de interesse, oferecendo boas explicações em casos específicos, mas sofrendo com variabilidade dos resultados (Ribeiro et al., 2016). Outras técnicas, como o *Model Agnostic Post Hoc Local Explanations* (MAPLE), combinam modelos de explicação local e global usando *decision forests*, permitindo uma flexibilidade maior, porém com menor clareza interpretativa para modelos muito complexos (Plumb et al., 2018). O *Local Rule-based Explanation* (LORE) utiliza regras locais para explicar as predições, sendo útil para entender decisões

complexas, mas menos eficaz em modelos não interpretáveis (Guidotti et al., 2018). Finalmente, o *Contextual Importance and Utility* (CIU) avalia a importância e a utilidade de cada variável no contexto específico de cada previsão, oferecendo um nível mais alto de explicação local, embora com limitações em termos de fidelidade e interpretabilidade global (Främling, 2022). Em uma comparação detalhada proposta por Hakkoum, Idri e Abnane (2024), dentre as técnicas citadas acima, o SHAP mostrou-se superior no desempenho da métrica de *faithfulness*, que avalia a precisão com que as explicações representam as verdadeiras contribuições das variáveis, e de *monotonicity*, ao garantir explicações mais consistentes em relação ao comportamento esperado do modelo.

Uma das vantagens do SHAP é a capacidade de fornecer explicações globais e locais sobre a importância das variáveis (Wang et al., 2021a). Explicações globais permitem identificar quais características do conjunto de dados exercem maior influência nas previsões gerais do modelo, oferecendo uma visão abrangente que pode orientar o desenvolvimento de políticas de saúde e o aprimoramento de modelos. Por sua vez, as explicações locais fornecem *insights* detalhados sobre como diferentes características de entrada afetam previsões individuais, o que é essencial para intervenções personalizadas e para a compreensão de casos específicos (Hendawi et al., 2023; Wang et al., 2021a). Essa dualidade torna o SHAP valioso na área médica, onde visão geral e detalhes individuais são importantes. Estudos recentes, como os de Mridha et al. (2023) e de Hendawi et al., (2023), também realizaram a integração de técnicas de interpretabilidade em ML. Essas pesquisas utilizaram o SHAP para fornecer explicações globais e, em seguida, aplicaram o LIME para obter explicações locais.

Em suma, apesar da escassez de estudos que tratem a interpretabilidade como um desafio técnico ou que apliquem técnicas de interpretabilidade exclusivamente a problemas de classificação com classes desbalanceadas, é evidente que a interpretabilidade permanece sendo um desafio significativo no domínio médico (Park et al., 2018). A transparência é essencial para que profissionais de saúde confiem e adotem ML em contextos clínicos. Nesse cenário, a aplicação de técnicas interpretáveis em conjunto com qualquer algoritmo de ML destaca-se como uma abordagem promissora. Essa estratégia combina a alta precisão de modelos complexos com a transparência, facilitando decisões médicas informadas.

Na sequência, explora-se o tema da generalização dos resultados ao lidar com classes de dados desbalanceadas, endereçando a questão de pesquisa Q3. A generalização é um aspecto crítico para modelos preditivos, que devem manter desempenho consistente em novas observações, sendo avaliada em dados não usados no treinamento (Jafari et al., 2022). Em cenários com classes desbalanceadas, a generalização é mais desafiadora, podendo levar a desempenho insatisfatório na detecção de classes minoritárias. Portanto, garantir a generalização é importante para a confiabilidade do modelo em situações reais. Nesta revisão, apenas 13 artigos discutiram ou mencionaram explicitamente a generalização de seus resultados.

Duas abordagens para melhorar a generalização são frequentemente citadas: *ensembles* e validação cruzada, ambas com características distintas e comumente empregadas em estudos para aprimorar a robustez dos modelos. Os *ensembles* combinam previsões de múltiplos modelos, captando padrões variados nos dados e, assim, aprimorando a precisão preditiva (Lu et al., 2022; Jia et al., 2022). Por outro lado, a validação cruzada permite uma avaliação mais consistente do modelo, ao dividir o conjunto de dados em múltiplos subconjuntos para treino e teste (Cui et al., 2018; Chan et al., 2017).

O uso de *Ensembles* pode aprimorar a generalização (Lu et al., 2022; Jia et al., 2022; Wang et al., 2021a; Wang et al., 2021b; Liu et al., 2020; West et al., 2005). Ao combinar resultados de diferentes modelos, erros individuais podem ser compensados, resultando em desempenho mais robusto e menos suscetível a variações nos dados (Wang et al., 2021b). A diversidade dos modelos que compõem o *ensemble* contribui para uma captura mais ampla dos padrões de dados, promovendo adaptação mais eficiente a novas observações (Jia et al., 2022). Contudo, a eficácia dos *ensembles* depende da qualidade dos modelos-base e da estratégia de combinação escolhida. Em certos contextos, modelos mais simples e isolados podem oferecer desempenho semelhante ou até superior, com um custo computacional reduzido. Portanto, é importante avaliar o custo-benefício do uso de *ensembles*, considerando a complexidade do problema e os recursos disponíveis.

Já a validação cruzada é empregada para avaliar a generalização, dividindo o conjunto de dados em múltiplas dobras ou *folds* (West et al., 2005; Cui et al., 2018; Chan

et al., 2017; Sedighi-Maman & Heath, 2022; Mridha et al., 2023). A técnica é aplicada ao conjunto completo de dados, comumente em divisões de 5 ou 10 *folds* (Mridha et al., 2023; Cui et al., 2018; Chan et al., 2017), dividindo aleatoriamente o conjunto completo em porções de calibração (entre 70% e 90%) e validação (entre 30% e 10%), seguidas de validação cruzada na parte de calibração (Sedighi-Maman & Heath, 2022; West et al., 2005). Em cenários com classes desbalanceadas, a sub-representação da classe minoritária em alguns *folds* pode resultar em um aprendizado enviesado para a classe majoritária, comprometendo a avaliação da capacidade de generalização do modelo. A validação cruzada estratificada, conforme Chan et al. (2017), mitiga esse problema, garantindo distribuição proporcional das classes em cada *fold.* O uso de poucos *folds*, especialmente em conjuntos pequenos, aumenta a variabilidade das métricas e favorece o *overfitting*, gerando estimativas distorcidas. Para contornar esses efeitos, alguns estudos recomendam a repetição aleatória do processo de divisão dos dados, aumentando a confiabilidade das estimativas. Por exemplo, Cui et al. (2018) realizaram 25 repetições, gerando 125 amostras de treino e validação. Chan et al. (2017) aplicaram 20 repetições, gerando 200 amostras de treino e validação. Já West et al. (2005) realizaram 20 repetições, gerando 100 amostras de treino e teste, além de 10 amostras de validação. Wang et al. (2021a) repetiram aleatoriamente 100 vezes o processo de dividir os dados em treino e teste, embora não tenham aplicado a validação cruzada. Essas estratégias visam aumentar a representatividade dos dados e melhorar a confiabilidade das estimativas de desempenho dos modelos.

Outra estratégia é a reamostragem nos *folds* de treinamento, como proposto por Cui et al. (2018), para garantir que cada *fold* seja representativo e proporcione uma avaliação mais estável. Wang et al. (2021b) destacam a importância da aplicação de técnicas para seleção de variáveis como estratégia eficaz para melhorar a generalização. Técnicas de seleção de variáveis e de reamostragem de dados devem ser aplicadas apenas ao conjunto de treinamento, a fim de evitar a introdução de viés ou *overfitting*. Executá-las no conjunto de teste comprometeria a generalização, inflacionando métricas de desempenho e gerando resultados não confiáveis (Wang et al., 2021b). Mridha et al. (2023) destacam a importância da parametrização dos algoritmos, incluindo técnicas como regularização e parada antecipada durante o treinamento do ML. A regularização adiciona uma penalidade aos parâmetros do modelo, controlando sua complexidade para

evitar que ele se ajuste demais aos dados de treinamento, o que reduz o risco de *overfitting*. Já a parada antecipada interrompe o treinamento assim que o desempenho em um conjunto de validação começa a piorar, mesmo que o modelo ainda melhore no conjunto de treinamento; essa interrupção previne que o modelo se ajuste excessivamente aos dados de treino, promovendo uma melhor generalização em dados novos.

Nesta revisão, não foram encontradas discussões aprofundadas sobre métodos para avaliar a generalização em conjuntos de dados desbalanceados. A maioria dos estudos apenas relatam o uso de métricas como a média dos valores de desempenho de, por exemplo, AUC, especificidade, sensibilidade e outras (Wang et al., 2023), acompanhadas de medidas de variabilidade como o desvio-padrão (Chan et al., 2017; Sedighi-Maman & Heath, 2022; Lu et al., 2022) e a amplitude dos valores observados (Wang et al., 2021a; West et al., 2005). Essas métricas fornecem uma avaliação básica da consistência do modelo em dados de validação; entretanto, são poucas as abordagens que tratam a generalização detalhadamente, especialmente em cenários de desbalanceamento.

2.5 CONCLUSÃO

A análise de dados desbalanceados no contexto médico requer atenção em cada etapa, do pré-processamento à modelagem e avaliação dos resultados. Este estudo aborda as etapas de um *workflow* genérico, com contribuições ao discutir e estruturar etapas necessárias para reduzir vieses preditivos, mantendo a confiabilidade e a interpretabilidade dos resultados — aspectos essenciais para a acurácia diagnóstica e a tomada de decisões clínicas.

A avaliação das técnicas de pré-processamento de dados, como limpeza, transformação e reamostragem, demonstrou impacto direto na qualidade e na generalização dos algoritmos de ML. Paralelamente, destacou-se a importância de validações robustas, como a validação cruzada estratificada e a seleção cuidadosa de variáveis, para evitar *overffiting* e promover estabilidade nas previsões. Em relação a interpretabilidade dos algoritmos, técnicas como o SHAP oferecem *insights* globais e locais sobre a influência das variáveis nos resultados dos modelos e promovem transparência necessária para uma maior confiança em algoritmos mais complexos.

Estudos futuros poderiam explorar a criação de *workflows* que adaptem técnicas de acordo com o nível de desbalanceamento de cada base de dados, buscando melhorias na consistência e na robustez dos modelos aplicados na área da saúde. Outra direção de pesquisa futura é a integração de *Natural Language Processing* (NLP) ou *Large Language Models* (LLMs) com métodos de interpretabilidade, como o SHAP, para simplificar as explicações fornecidas aos profissionais da saúde. LLMs, por exemplo, poderiam traduzir os resultados de SHAP em uma linguagem contextualizada, ressaltando o impacto das variáveis, sem exigir conhecimento avançado em ML ou estatística. Essa abordagem permitiria que profissionais da saúde recebessem explicações ajustadas ao nível de detalhamento necessário para cada caso.

Outra direção promissora reside no desenvolvimento de métodos robustos de avaliação da generalização, que vão além das métricas tradicionais e incluem visualizações que ilustrem a distribuição completa dos resultados. Essas visualizações poderiam destacar o desempenho nas classes minoritárias, fornecendo uma visão mais clara das variações e comportamentos dos modelos em diferentes regiões do espaço de dados. Tal abordagem aumentaria a confiabilidade dos modelos em cenários práticos e impulsionaria avanços na compreensão e no aprimoramento da generalização.

2.6 REFERENCES

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(4),* 1059–1086. https://doi.org/10.1111/rssb.12395

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology: Theory and Practice*, *8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Bai, B., Wu, Z., Weng, S., & Yang, Q. (2023). Application of interpretable machine learning algorithms to predict distant metastasis in osteosarcoma. *Cancer Medicine*, *12*(4), 5025–5034. https://doi.org/10.1002/cam4.5225

Basha, S. J., Madala, S. R., Vivek, K., Kumar, E. S., & Ammannamma, T. (2022). A Review on Imbalanced Data Classification Techniques. *2022 International Conference*

on Advanced Computing Technologies and Applications, ICACTA 2022. https://doi.org/10.1109/ICACTA54488.2022.9753392

Benhar, H., Idri, A., & L Fernández-Alemán, J. L. (2020). Data preprocessing for heart disease classification: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 195, 105635. https://doi.org/10.1016/j.cmpb.2020.105635

Ben-Israel, D., Jacobs, W. B., Casha, S., Lang, S., Ryu, W. H. A., de Lotbiniere-Bassett, M., & Cadotte, D. W. (2020). The impact of machine learning on patient care: A systematic review. *Artificial Intelligence in Medicine, 103,* 101785. https://doi.org/10.1016/j.artmed.2019.101785

Brahma, D., & Mukherjee, D. (2021). Early warning signs: targeting neonatal and infant mortality using machine learning. *Applied Economics*, *1*, 57–74.

Cabitza, F., & Campagner, A. (2021). The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *International Journal of Medical Informatics*, 153, 104510. https://doi.org/10.1016/j.ijmedinf.2021.104510

Chan, T. M., Li, Y., Chiau, C. C., Zhu, J., Jiang, J., & Huo, Y. (2017). Imbalanced target prediction with pattern discovery on clinical data repositories. *BMC Medical Informatics and Decision Making*, *17*(1). https://doi.org/10.1186/s12911-017-0443-3

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, *4*, 123–144. https://doi.org/10.1146/annurev-biodatasci-092820

Cui, S., Wang, D., Wang, Y., Yu, P. W., & Jin, Y. (2018). An improved support vector machine-based diabetic readmission prediction. *Computer Methods and Programs in Biomedicine*, *166*, 123–135. https://doi.org/10.1016/j.cmpb.2018.10.012

Feng, M., Zhang, J., Zhou, X., Mo, H., Jia, L., Zhang, C., Hu, Y., & Yuan, W. (2022). Application of an Interpretable Machine Learning Model to Predict Lymph Node Metastasis in Patients with Laryngeal Carcinoma. *Journal of Oncology*, *2022*. https://doi.org/10.1155/2022/6356399

Främling, K. (2022). *Contextual Importance and Utility: A Theoretical Foundation*. http://arxiv.org/abs/2202.07292

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4).

Ghavidel, A., & Pazos, P. (2023). Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: a systematic review. *Journal of Cancer Survivorship*. https://doi.org/10.1007/s11764-023-01465-3

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local Rule-Based Explanations of Black Box Decision Systems. *arXiv. https://doi.org/10.48550/arXiv.1805.10820*

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Hakkoum, H., Idri, A., & Abnane, I. (2024). Global and local interpretability techniques of supervised machine learning black box models for numerical medical data. *Engineering Applications of Artificial Intelligence*, *131*. https://doi.org/10.1016/j.engappai.2023.107829

Haynes, R. B., & Wilczynski, N. L. (2010). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: Methods of a decision-maker-researcher partnership systematic review. *Implementation Science, 5, 12. https://doi.org/10.1186/1748-5908-5-12*

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hendawi, R., Li, J., & Roy, S. (2023). A Mobile App That Addresses Interpretability Challenges in Machine Learning–Based Diabetes Predictions: Survey-Based User Study. *JMIR Formative Research*, *7*(1). https://doi.org/10.2196/50328

Hu, W., Jin, T., Pan, Z., Xu, H., Yu, L., Chen, T., Zhang, W., Jiang, H., Yang, W., Xu, J., Zhu, F., & Dai, H. (2023). An interpretable ensemble learning model facilitates early risk stratification of ischemic stroke in intensive care unit: Development and external validation of ICU-ISPM. *Computers in Biology and Medicine*, *166*. https://doi.org/10.1016/j.compbiomed.2023.107577

Jafari, H., Shohaimi, S., Salari, N., Kiaei, A. A., Najafi, F., Khazaei, S., Niaparast, M., Abdollahi, A., & Mohammadi, M. (2022). A full pipeline of diagnosis and prognosis of chronic disease risk using deep learning and Shapley values: The Ravansar county anthropometric cohort study. *PLoS ONE*, 17(1), e0262701. https://doi.org/10.1371/journal.pone.0262701

Jia, L., Wang, Z., Lv, S., & Xu, Z. (2022). PE_DIM: An Efficient Probabilistic Ensemble Classification Algorithm for Diabetes Handling Class Imbalance Missing Values. *IEEE Access*, *10*, 107459–107476. https://doi.org/10.1109/ACCESS.2022.3212067

Johnson, M., Albizri, A., & Simsek, S. (2022). Artificial intelligence in healthcare operations to enhance treatment outcomes: a framework to predict lung cancer prognosis. *Annals of Operations Research*, *308*(1–2), 275–305. https://doi.org/10.1007/s10479-020-03872-6

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, 52(4), 1–36. https://doi.org/10.1145/3343440

Kokkotis, C., Giarmatzis, G., Giannakou, E., Moustakidis, S., Tsatalas, T., Tsiptsios, D., Vadikolias, K., & Aggelousis, N. (2022). An Explainable Machine Learning Pipeline for Stroke Prediction on Imbalanced Data. *Diagnostics*, *12*(10). https://doi.org/10.3390/diagnostics12102392

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Liu, R., Wang, M., Zheng, T., Zhang, R., Li, N., Chen, Z., Yan, H., & Shi, Q. (2022). An artificial intelligence-based risk prediction model of myocardial infarction. *BMC Bioinformatics*, *23*(1). https://doi.org/10.1186/s12859-022-04761-4

Lombardi, A., Arezzo, F., Di Sciascio, E., Ardito, C., Mongelli, M., Di Lillo, N., Fascilla, F. D., Silvestris, E., Kardhashi, A., Putino, C., Cazzolla, A., Loizzi, V., Cazzato, G., Cormio, G., & Di Noia, T. (2023). A human-interpretable machine learning pipeline based on ultrasound to support leiomyosarcoma diagnosis. *Artificial Intelligence in Medicine*, *146*. https://doi.org/10.1016/j.artmed.2023.102697

Lu, H., Xue, G., Li, S., Mu, Y., Xu, Y., Hong, B., Huang, Q., Li, Q., Yang, P., Zhao, R., Fang, Y., Luo, Q., Zhou, Y., & Liu, J. (2022). An accurate prognostic prediction for aneurysmal subarachnoid hemorrhage dedicated to patients after endovascular treatment. *Therapeutic Advances in Neurological Disorders*, *15*. https://doi.org/10.1177/17562864221099473

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. http://arxiv.org/abs/1705.07874

Luo, G., Johnson, M. D., Nkoy, F. L., He, S., & Stone, B. L. (2020). Automatically explaining machine learning prediction results on asthma hospital visits in patients with asthma: Secondary analysis. *JMIR Medical Informatics*, *8*(12). https://doi.org/10.2196/21965

Maalouf, M., & Trafalis, T. B. (2011). Rare events and imbalanced datasets: An overview. *International Journal of Data Mining, Modelling and Management*, 3(4), 297–310.

Maheshwari, S., Jain, R. C., & Jadon, R. S. (2018). An insight into rare class problem: Analysis and potential solutions. *Journal of Computer Science*, 14(6), 777–792. https://doi.org/10.3844/jcssp.2018.777.792

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Molnar, C. (2019). *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/. http://leanpub.com/interpretable-machine-learning

Moslehi, S., Mahjub, H., Farhadian, M., Soltanian, A. R., & Mamani, M. (2022). Interpretable generalized neural additive models for mortality prediction of COVID-19 hospitalized patients in Hamadan, Iran. *BMC Medical Research Methodology*, *22*(1). https://doi.org/10.1186/s12874-022-01827-y

Mridha, K., Ghimire, S., Shin, J., Aran, A., Uddin, M. M., & Mridha, M. F. (2023). Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention. *IEEE Access*, *11*, 52288–52308. https://doi.org/10.1109/ACCESS.2023.3278273

Park, E., Chang, H. J., & Nam, H. S. (2018). A Bayesian network model for predicting post-stroke outcomes with available risk factors. *Frontiers in Neurology, 9, 699. https://doi.org/10.3389/fneur.2018.00699*

Plumb, G., Molitor, D., & Talwalkar, A. (2018). *Model Agnostic Supervised Local Explanations*. http://arxiv.org/abs/1807.02910

Rekha, G., Tyagi, A. K., & Reddy, V. K. (2019). A wide scale classification of class imbalance problem and its solutions: A systematic literature review. *Journal of Computer Science*, 15(7), 886–929. https://doi.org/10.3844/jcssp.2019.886.929

Rezvani, S., & Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing*, 143, 110415. https://doi.org/10.1016/j.asoc.2023.110415

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 97–101. https://doi.org/10.18653/v1/N16-3020

Saarela, M., Ryynänen, O. P., & Äyrämö, S. (2019). Predicting hospital associated disability from imbalanced data using supervised learning. *Artificial Intelligence in Medicine*, *95*, 88–95. https://doi.org/10.1016/j.artmed.2018.09.004

Santos, L. I., Camargos, M. O., D'Angelo, M. F. S. V., Mendes, J. B., Medeiros, E. E. C. de, Guimarães, A. L. S., & Palhares, R. M. (2022). Decision tree and artificial immune systems for stroke prediction in imbalanced data. *Expert Systems with Applications*, *191*. https://doi.org/10.1016/j.eswa.2021.116221

Sedighi-Maman, Z., & Heath, J. J. (2022). An Interpretable Two-Phase Modeling Approach for Lung Cancer Survivability Prediction. *Sensors*, *22*(18). https://doi.org/10.3390/s22186783

Sedighi-Maman, Z., & Mondello, A. (2021). A two-stage modeling approach for breast cancer survivability prediction. *International Journal of Medical Informatics*, *149*. https://doi.org/10.1016/j.ijmedinf.2021.104438

Spelmen V S, & Porkodi R. (2018). A Review on Handling Imbalanced Data. *International Conference on Current Trends towards Converging Technologies*, Coimbatore, India, 1-11, doi: 10.1109/ICCTCT.2018.8551020.

Stafford, I. S., Gosink, M. M., Mossotto, E., Ennis, S., & Hauben, M. (2022). A Systematic Review of Artificial Intelligence and Machine Learning Applications to Inflammatory Bowel Disease, with Practical Guidelines for Interpretation. *Inflammatory Bowel Diseases*, 28(10), 1573–1583. https://doi.org/10.1093/ibd/izac115

Su, P. Y., Wei, Y. C., Luo, H., Liu, C. H., Huang, W. Y., Chen, K. F., Lin, C. P., Wei, H. Y., & Lee, T. H. (2022). Machine Learning Models for Predicting Influential Factors of Early Outcomes in Acute Ischemic Stroke: Registry-Based Study. *JMIR Medical Informatics*, *10*(3). https://doi.org/10.2196/32508

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of Imbalanced Data: A Review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4), 687–719. https://doi.org/10.1142/S0218001409007326

Susan, S., & Kumar, A. (2021). The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art. *Engineering Reports*, 3(4), e12298. https://doi.org/10.1002/eng2.12298

Wang, K., Tian, J., Zheng, C., Yang, H., Ren, J., Li, C., Han, Q., & Zhang, Y. (2021). Improving risk identification of adverse outcomes in chronic heart failure using smote +enn and machine learning. *Risk Management and Healthcare Policy*, *14*, 2453–2463. https://doi.org/10.2147/RMHP.S310295

Wang, X., Ren, H., Ren, J., Song, W., Qiao, Y., Ren, Z., Zhao, Y., Linghu, L., Cui, Y., Zhao, Z., Chen, L., & Qiu, L. (2023). Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data. *Computer Methods and Programs in Biomedicine*, *230*. https://doi.org/10.1016/j.cmpb.2023.107340

Wang, X., Zhai, M., Ren, Z., Ren, H., Li, M., Quan, D., Chen, L., & Qiu, L. (2021). Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01471-4

West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research*, *162*(2), 532–551. https://doi.org/10.1016/j.ejor.2003.10.013

Yu, M., Yuan, Z., Li, R., Shi, B., Wan, D., & Dong, X. (2024). Interpretable machine learning model to predict surgical difficulty in laparoscopic resection for rectal cancer. *Frontiers in Oncology*, 14, 1337219. https://doi.org/10.3389/fonc.2024.1337219

Zhang, Y., Razbek, J., Li, D., Yang, L., Bao, L., Xia, W., Mao, H., Daken, M., Zhang, X., & Cao, M. (2022). Construction of Xinjiang metabolic syndrome risk prediction model based on interpretable models. *BMC Public Health*, *22*(1). https://doi.org/10.1186/s12889-022-12617-y

2.7 APÊNDICE A1

Autores (Ano)	Especialidade e problema de pesquisa	Taxa classe min.	Normalização	Imputação de dados faltantes	Remove ruído / outliers	Codificação de dados categóricos	Seleção de Variáveis.	Algoritmos de Machine Learning (ML) testados	Parametrização	Pesos para as classes	Técnicas de reamostragem e momento de aplicação	Estratifica classe	Repetição e parcionamento do dataset	Validação cruzada
Bai et al. (2023)	Ortopedia. Classificar pacientes com metástase	22,60%	-	MissForest	-	-	-	Comparação de: LR, SVM, GaussianNB, XGBoost, RF e KNN.	Grid-search aplicado no dataset de treino	-	ADASYN. Aplicado em todo dataset.	-	[70%, 0%, 30%]	[10-folds] no set de treino
Cui et al. (2018)	Endocrinologia. Previsão de readmissão hospitalar de pacientes com diabetes	-	Sim, não menciona qual	Não, foram removidas 1117 amostras com valores faltantes	-	Codificação binária	Combinação de filter e wrapper	Comparação de: SVM, LACE score, LR, Naïve Bayes, DT, ANN.	Algoritmo Genético (GA), utilizado apenas no SVM.	-	SMOTE. No set de treino	-	-	25 repetições x [5-folds] no dataset completo
Jia et al. (2022)	Endocrinologia. Utilizou 3 datasets para classificar pacientes com diabetes	34,98%, 20,68% e 35,13%	-	Comparação de: LMeGNB, média, Mediana, RF e NB	-	-	-	Ensemble probabilístico de: KNN, WKNN, SVM, LR, DT, RF, XGBoost e AdaBoost.	Parâmetros default de todos ML	-	K-means SMOTE. Aplicado no dataset completo	-	-	Comparou [2-folds], [5-folds] e [10-fold]. no dataset completo
Wang et al. (2023)	Pneumologia. Prever pacientes com doença pulmonar obstrutiva crônica.	9,08%	-	Não, foram removidas 841 amostras com valores faltantes	-	-	Embedded	Comparação de: LR, SVM, RF, XGBoost, LightGBM, NGBoost e Stacking.	Parâmetros default de todos ML	Sim	Cost-sensitive learning and SMOTE. No set de treino	Sim	100 repetições x [70%, 0%, 30%]	-
Kokkotis et al. (2022)	Neurologia. Prever a ocorrência de AVC.	2,00%	Standard Scaler function	Não, foram removidas amostras com valores faltantes	Sim, mas não menciona como.	-	-	Comparação de: LR, RF, XGBoost, KNN, SVM, MLP.	Grid-search aplicado no dataset de treino	-	RUS. No set de treino	Sim	-	[10-folds] no dataset completo
Moslehi et al. (2022)	Ambulatório. Identificação de óbitos em decorrência da COVID	40,00%	-	Comparação de: remoção de valores faltantes, média, KNN, MICE.	-	-	Embedded	Comparação de: GNAM, LR, RF, GAMs, GBDT, DNN.	Parametrização ajustada com validação cruzada	-	SMOTE. Aplicado no dataset completo	-	100 repetições x [70%, 0%, 30%]	-
Park et al. (2018)	Neurologia. Prever independência funcional aos 3 meses e a mortalidade após 1 ano do AVC.	26,40% e 12,03%	-	Não, foram removidas amostras com valores faltantes	-	-	Comparação de wrapper, filter, e nenhuma	TABNC	Hill-climbing searching	-	-	-	-	[10-folds] no dataset completo
Santos et al. (2022)	Previsão de AVC	1,89%	-	Não, foram removidas amostras com valores faltantes	-	Sim, mas não menciona como.	-	Comparação de: DT via GP e AutoHPO via DNN	Parametrização ajustada com validação cruzada	-	Artificial Immune Systems (AIS) combinado com One Sided Selection (OSS). No set de treino	-	-	[5-folds] no dataset completo
Lu et al. (2022)	Previsão prognóstica para hemorragia subaracnóidea aneurismática	17,39%	Standard Scaler function	-	-	Sim, One-Hot Encode	-	Comparação de: LR, SVM, RF, Ensemble MESA, e Bagging de LR, SVM E RF.	Grid-search	-	RUS. No set de treino	-	Dividiu, mas não informou a proporção	10 repetições x [5-folds] no set de treino
Liu et al. (2022)	Cardiologia. Predição de risco de infarto do miocárdio	6,15%	min-max	Não, foram removidas amostras com valores faltantes	-	-	Embedded	Comparação de: RF, GBDT, LR, e SVM.	-	-	Comparou ROS, RUS, easyensemble, and w-easy ensemble. No set de treino	Sim	[80%, 0%, 20%]	-
Brahma and Mukherjee (2022)	Mortalidade neonatal e infantil	~ 1,00%	-	-	-	-	-	Comparação de: LR, LR com LASSO, RF, and Boosted Classification Trees.	Testou diferentes parâmetros	-	Comparou RUSBoost com SMOTEBoost. No set de treino	-	Testou 2 divisões: [60%, 0%, 40%] e [70%, 0%, 30%].	-

Autores (Ano)	Especialidade e problema de pesquisa	Taxa classe min.	Normalização	Imputação de dados faltantes	Remove ruído / outliers	Codificação de dados categóricos	Seleção de Variáveis.	Algoritmos de Machine Learning (ML) testados	Parametrização	Pesos para as classes	Técnicas de reamostragem e momento de aplicação	Estratifica classe	Repetição e parcionamento do dataset	Validação cruzada
Wang et al. (2021)	Endocrinologia. Classificação de alto risco para diabetes mellitus	9,53%	-	Não, foram removidas amostras com valores faltantes	-	-	Embedded	Comparação de: Ensemble, LR e SVM.	Grid-search	-	SVM-SMOTE	-	100 repetições x [70%, 0%, 30%]	[9-folds] no set de treino
Feng et al. (2022)	Otorrinolaringologia. Previsão de Metástase Linfonodal em pacientes com Carcinoma de Laringe	30,24%	-	-	-	-	Embedded	Comparação de: LR, KNN, SVM, XGBoost, RF, LightGBM	-	-	-	-	[80%, 0%, 20%]	[10-folds] no set de treino
Yu et al. (2024)	Cirurgia colorretal. Dificuldade cirúrgica em ressecções laparoscópicas para câncer retal	17,57%	Standard Scaler function	Não, foram removidas amostras com valores faltantes	-	-	Embedded	Comparação de: LightGBM, CatBoost, XGBoost, LR, RF, MLP.	Grid-search	-	SMOTE. No set de treino	-	[80%, 0%, 20%]	[10-folds] no set de treino
Johnson, Albizri and Simsek (2020)	Oncologia. Prever o prognóstico do câncer de pulmão.	13,78%	-	-	-	-	Wrapper	Comparação de: LR, DT, RF, AdaBoost, ANN, Naïve Bayes.	Parametrização ajustada com validação cruzada	-	SMOTE. No set de treino	-	[70%, 0%, 30%]	[5-folds] no set de treino
Hakkoum, Idri and Abnane (2024)	5 datasets. 2 de diagnóstico de câncer de mama, retinopatia diabética, doença de Parkinson e coração.	-	-	Não, foram removidas amostras com valores faltantes	-	-	-	Comparação de: MLP, SVM, RF, XGBoost, Naïve Bayes	Particle Swarm Optimization (PSO)	-	SMOTE. No set de treino	-	[80%, 0%, 20%]	[10-folds] no set de treino
Mridha et. al (2023)	Previsão de AVC	4,87%	min-max	Média	Sim, mas não menciona como.	Sim, strings codificadas em números inteiros.	Filter	Comparação de: RF, XGBoost, KNN, LR, SVM, Naive Bayes	Parametrização ajustada com validação cruzada	-	SMOTE. No set de treino	-	Dividiu, mas não informou a proporção	[não informa os folds] no set de treino
Hendawi, Li & Roy (2023)	Endocrinologia. Previsões de diabetes	~ 38,00%	Standard Scaler function	-	Sim, utilizando Tomek Links	-	Embedded	Comparação de: DNN, RF, DT, Naive Bayes, Sequential Minimal Optimization, Java 48, KNN, LR	-	-	SMOTE.	-	Dividiu, mas não informou a proporção	-
Wang et al. (2021)	Cardiologia. Identificação do risco de desfechos adversos na insuficiência cardíaca crônica	9,95%	min-max	MissForest	-	Sim, One-Hot Encode	Wrapper	Comparação de: LR, KNN, SVM, RF e XGBoost.	Grid-search	-	SMOTE+ENN. No set de treino	Sim	100 repetições x [60%, 20%, 20%]	[5-folds]
West et al. (2021)	Oncologia. Diagnóstico do câncer de mama	34,47% e 23,73%	-	-	-	-	-	Ensemble com LDA, LR, MLP, MOE, RBF, CART, KNN e KD.	Testou diferentes parâmetros	-	-	-	Dividiu, mas não informou a proporção	[10-folds]
Lombardi et al. (2023)	Oncologia. Diagnóstico de leiomiossarcoma	11,76%	-	Não, foram removidas amostras com valores faltantes	-	Variáveis categóricas codificadas em números inteiros.	Embedded	Comparação de: RF, LR com Lasso, Logit Lasso, XGBoost	Parametrização ajustada com validação cruzada	-	Comparou SMOTE, ANSMOTE, ADASYN, DBSMOTE, MWMOTE, SLMOTE, RSLSMOTE. No set de treino	-	-	[3-folds]

Autores (Ano)	Especialidade e problema de pesquisa	Taxa classe min.	Normalização	Imputação de dados faltantes	Remove ruído / outliers	Codificação de dados categóricos	Seleção de Variáveis.	Algoritmos de Machine Learning (ML) testados	Parametrização	Pesos para as classes	Técnicas de reamostragem e momento de aplicação	Estratifica classe	Repetição e parcionamento do dataset	Validação cruzada
Zhang et al. (2022)	Epidemiologia. Predição de risco da síndrome metabólica	12,80%	-	Multiple imputation (MI)	-	-	-	Comparação de: LR e DT	-	-	Comparou ROS, SMOTE, RUS. No set de treino	-	[70%, 0%, 30%]	-
Su et al. (2022)	Neurologia. Prever a ocorrência de AVC. Escala de Rankin modificada (mRS) na alta hospitalar e deterioração hospitalar	6,00%	min-max	Não, foram removidas amostras com valores faltantes	-	Sim, One-Hot Encode	-	Comparação de: SVM, RF, LightGBM, DNN.	-	-	Comparou ROS, SMOTE, SMOTE-NC e RUS	-	-	[5-folds] no dataset completo
Saarela et al., 2019 (2019)	hospitalização de pacientes idosos	34,00%	Comparou: Standard Scaler function, min-max, e uso de nenhuma.	Criaram uma variável categórica para indicar valores faltantes de glicose, que indicavam que os níveis eram normais	-	Sim, One-Hot Encoding e Ordinal	Comparação de: Análise de componentes principais, Non-negative matrix factorization, e nenhuma.	Comparação de: LR, RF.	Grid-search aplicado no dataset de treino	Sim, comparou com técnicas de reamostragem	Comparou ROS, SMOTE, RUS, EasyEnsemble e nenhum. No set de treino	-	[70%, 0%, 30%]	[10-folds] no set de treino
Sedighi-Maman & Mondello (2021)	Oncologia. Predição da sobrevivência ao câncer de mama	-	-	Valores não numéricos foram imputados com a mediana, e níveis desconhecidos/NA foram unificados	Sim, utilizando distância de Cook's	Sim, One-Hot Encoding e Integer	LASSO e RF	Comparação de: GLM, XGBoost e MLP	-	-	Comparou SMOTE, RSLSMOTE, ADASYN. Aplicado no dataset de treino	-	[70%, 0%, 30%]	[5-folds] no set de treino
Hu et al. (2023)	Predizer mortalidade intra-hospitalar por AVC isquêmico na UTI	20,00% e 10,00%	Standard Scaler function	Variáveis com taxa faltante > 30% foram excluídas. Variáveis com taxa faltante < 5% foram substituidas pela média e moda	-	Variáveis categóricas foram codificadas em contagem e porcentagem.	Wrapper	Comparação de: LR, SVM, RF, XGBoost, LightGBM e MLP, e ensemble de todos os modelos	Grid-search	Sim	SMOTE, ROS, SMOTETomek, ADASYN. Aplicado no dataset de treino	-	[80%, 0%, 20%]	[10-folds] no set de treino
Chan et al. (2017)	Cardiologista e Cirurgião toráxico. Identificação de óbitos em conjunto de dados torácicos e cardíacos.	14,90% e 9,10%	-	-	ML proposta é tolerante a ruídos.	-		Comparação de: Prediction with pattern discovery), LR, Naïve Bayes, and DT.	Parâmetros default de todos ML	-	-	Sim	-	20 repetições x [10-folds] no dataset completo
Jafari et al. (2022)	Endocrinologia. Diagnóstico de diabetes mellitus tipo 2 (T2DM)	9,64%	Sim, não menciona qual	Regression Imputation	Substituídos por Regression Imputation	-	-	1-dimensional convolution DNN.	-	Sim	ROS. No set de treino	Sim	[80%, 10%, 10%]	-
Luo et al. (2020)	Pneumologia. Prever visitas hospitalares de pacientes com asma	6,00%	Standard Scaler function	-	Substituídos por valores nulos.	Sim, mas não menciona como.	-	XGBoost	-	-	-	-	[90%, 0%, 10%]	-
Sedighi-Maman & Heath (2022)	Oncologia. Predição da Sobrevivência ao Câncer de Pulmão	-	-	Não, foram removidas amostras com valores faltantes	Sim, utilizando distância de Cook's	Sim, One-Hot Encode	Embedded	Comparação de: General Linear Model (GLM), XGboost, e ANN	-	-	Comparou SMOTE, RSLSMOTE, BLSMOTE, ADASYN, MWMOTE. No set de treino	-	[70%, 0%, 30%]	[5-folds] no set de treino
Liu et al. (2020)	Nove conjuntos de dados médicos desbalanceados, predição de presença de doença	Entre 16,00% e 35,00%.	-	-	Sim, validação Cruzada Committees Filter (CVCF)	-	-	Ensemble SVM (ESVM) cada um com uma função de Kernel, combinado com SAGA para a otimização.	Grid-search	-	SMOTE. No set de treino	-	-	5 repetições x [10-folds] no set de treino

3 ARTIGO 2: MACHINE LEARNING FOR HIGH-RISK HOSPITALIZATION PREDICTION IN OUTPATIENT PATIENTS WITH DIABETES AT A TERTIARY HOSPITAL

Artigo aprovado no Archives of Endocrinology and Metabolism (2024)

4 ARTIGO 3: DECISION ANALYSIS FRAMEWORK FOR PREDICTING NO-SHOWS TO APPOINTMENTS USING MACHINE LEARNING ALGORITHMS

Artigo publicado no BMC Health Services Research (2023)

DOI: h10.1186/s12913-023-10418-6

5 CONSIDERAÇÕES FINAIS

Problemas relacionados à classificação de dados desbalanceados na área da saúde trazem desafios significativos para a implementação de algoritmos de ML, especialmente em cenários clínicos. Esses desafios incluem a dificuldade de generalizar os modelos para novas populações, a baixa representatividade de classes minoritárias e a necessidade de interpretabilidade para apoio na tomada de decisão médica. Esta pesquisa investigou essas questões sob diferentes perspectivas e propôs abordagens que contribuíssem para a generalização e a aplicabilidade prática de algoritmos de aprendizado supervisionado.

Inicialmente, esta tese realizou uma revisão sistemática da literatura sobre o tema, identificando as principais lacunas metodológicas e práticas no tratamento de dados desbalanceados na área médica. Foi evidenciado que, embora existam diversas técnicas para reamostragem, validação e seleção de variáveis, faltam *workflows* estruturados que garantam resultados robustos e interpretáveis. Dessa revisão, foi proposto um *framework* genérico que integra etapas essenciais para mitigar vieses e promover a eficácia dos modelos em contextos clínicos.

Com base no *framework* desenvolvido, foram realizados dois estudos de caso para validar as propostas teóricas. O primeiro estudo abordou a previsão de hospitalizações em pacientes com diabetes como um problema crítico para a gestão de recursos hospitalares. O uso de técnicas interpretativas, como SHAP, permitiu identificar variáveis-chave que influenciam o risco de hospitalização, oferecendo subsídios para decisões clínicas mais informadas. A aplicação do modelo possibilitou uma abordagem mais direcionada para o acompanhamento de pacientes em risco, com potencial de melhorar os resultados clínicos e otimizar a alocação de recursos.

O segundo estudo abordou a previsão de não comparecimentos (*no-show*) em consultas médicas, um problema que afeta a eficiência dos serviços de saúde. O modelo desenvolvido demonstrou que a combinação de técnicas de reamostragem e validação cruzada estratificada melhora a sensibilidade dos modelos, permitindo identificar

pacientes com maior probabilidade de ausência. Além disso, as análises indicaram que estratégias como *overbooking* poderiam ser mais bem informadas por esses modelos, otimizando a alocação de recursos hospitalares.

Os resultados apresentados nesta tese destacam três contribuições principais. Primeiramente, o *framework* proposto oferece uma abordagem estruturada para lidar com problemas de desbalanceamento em dados médicos, fornecendo diretrizes práticas para minimizar vieses e melhorar a generalização dos modelos. Em segundo lugar, os estudos de caso demonstraram que a aplicação do *framework* em problemas reais pode melhorar a eficiência dos serviços de saúde e apoiar a tomada de decisão clínica. Por fim, as técnicas de interpretabilidade discutidas e aplicadas, como SHAP, reforçam a importância de compreender os fatores que influenciam as previsões, contribuindo para a aceitação e adoção de modelos preditivos por profissionais da saúde.

Apesar das contribuições, algumas limitações devem ser reconhecidas. Esta pesquisa utilizou bases de dados específicas, o que restringe a generalização imediata dos resultados para outros contextos ou problemas de saúde. Além disso, a tese focou exclusivamente em dados estruturados e algoritmos de aprendizado supervisionado, não abrangendo outros tipos de dados ou técnicas mais avançadas, como aprendizado profundo. A ausência de uma análise econômica detalhada do impacto das propostas também limita a compreensão completa de seus benefícios.

Futuras pesquisas podem explorar a aplicação do *framework* proposto em outros contextos médicos, como a detecção de doenças raras ou o monitoramento de pacientes crônicos. O uso de dados multimodais, que combinem informações tabulares com imagens médicas ou textos de prontuários, poderia enriquecer as análises. Além disso, estudos futuros podem investigar abordagens híbridas que integrem aprendizado supervisionado e não supervisionado, ampliando as possibilidades de descoberta de padrões em dados desbalanceados.

Estudos futuros também poderiam explorar a criação de *framework* que adaptem técnicas de acordo com o nível de desbalanceamento de cada base de dados, buscando

melhorias na consistência e na robustez dos modelos aplicados na área da saúde. A etapa de estratificação, por exemplo, pode não ser suficiente em cenários altamente desbalanceados, a exemplo do artigo 2, onde a presença de comorbidades e diferentes faixas etárias não são heterogêneas. O desenvolvimento de validação cruzada estratificada por subgrupos clínicos poderia fornecer estimativas mais robustas, possibilitando identificar onde o modelo tende a errar com mais frequência.

Outra linha de pesquisa é examinar como o desbalanceamento afeta as técnicas de interpretabilidade, propondo métricas que avaliem se as explicações destacam variáveis importantes para a classe minoritária ou se a criação de "instâncias artificiais" é limitada pela escassez de exemplos. Além disso, integrar *Natural Language Processing* (NLP) ou *Large Language Models* (LLMs) com métodos interpretáveis como o SHAP pode simplificar a comunicação dos resultados aos profissionais da saúde, tornando as contribuições das variáveis acessíveis. Em paralelo, ferramentas de *software* que implementem o *workflow* proposto, com interfaces intuitivas, podem facilitar a adoção por médicos e gestores, traduzindo os resultados dos modelos em *insights* aplicáveis. Essas ferramentas também podem calibrar explicações conforme o risco clínico de cada classe e emitir "alertas interpretáveis" para sinalizar incertezas ou desvios de padrões de treinamento, contribuindo para maior confiança nos algoritmos.

Por fim, estudos futuros podem investigar outras formas de avaliar a generalização além de métricas pontuais, adotando visualizações que tornem evidentes as variações de desempenho nos diferentes estratos dos dados. A capacidade de identificar comportamentos atípicos ou zonas de maior erro amplia a confiabilidade dos algoritmos em cenários reais, apoiando tanto a segurança do paciente quanto a tomada de decisão informada.

Esta tese abordou questões teóricas e práticas relacionadas ao ML em dados desbalanceados na saúde, com foco na construção de modelos robustos e interpretáveis. Os resultados obtidos destacam a importância de *workflows* bem estruturados e adaptáveis, além do uso de técnicas de interpretabilidade, para promover a confiabilidade e a aceitação desses modelos em contextos clínicos.

Ao responder às questões de pesquisa propostas e ao alcançar os objetivos estabelecidos, esta pesquisa contribui para a literatura acadêmica e oferece ferramentas práticas para melhorar a eficiência e a eficácia na área da saúde. As abordagens propostas abrem caminho para novas investigações e aplicações que possam continuar a transformar os cuidados médicos e a gestão de recursos hospitalares.