Universidade Estadual Paulista "Júlio de Mesquita Filho"
Instituto de Biociências de Botucatu

Pós-Graduação em Ciências Biomoleculares e Farmacológicas – IBB

UNESP

FELIPE ALMEIDA FERNANDES

Modelo de Inteligência Artificial para a predição das Doenças Inflamatórias Intestinais, através de dados clínicos e epidemiológicos

Botucatu (SP)
2026

FELIPE ALMEIDA FERNANDES

Modelo de Inteligência Artificial para a predição das Doenças Inflamatórias Intestinais, através de dados clínicos e epidemiológicos

Dissertação apresentada à Universidade Estadual Paulista (UNESP), “Júlio de Mesquita Filho” Instituto de Biociências, campus de Botucatu para a obtenção do título de mestre em Ciências na Área de Farmacologia e Biotecnologia.

Área de concentração: Farmacologia e Biotecnologia

Orientador: Prof. Dr. José Celso Rocha

Coorientadora: Prof. Dra. Ligia Yukie Sassaki

Botucatu (SP)
2026

F363m

Fernandes, Felipe Almeida

Modelo de inteligência artificial para a predição das doenças inflamatórias intestinais, através de dados clínicos e epidemiológicos / Felipe Almeida Fernandes. -- , 2026

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu,

Orientador: José Celso Rocha

Coorientadora: Ligia Yukie Sassaki

1. Inteligência artificial Aplicações médicas. 2. Doenças inflamatórias intestinais. 3. Colite ulcerativa. 4. Crohn, Doença de. I.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor.

FELIPE ALMEIDA FERNANDES

Modelo de Inteligência Artificial para a predição das Doenças Inflamatórias Intestinais, através de dados clínicos e epidemiológicos

Dissertação apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu, para obtenção do título de Mestre em Ciências na área de Farmacologia e Biotecnologia.

Área de concentração: Farmacologia e Biotecnologia.

Data da defesa: 10/08/2026

BANCA EXAMINADORA:

_____

Orientador

Prof. Dr. José Celso Rocha

Departamento de Ciências Biológicas / UNESP / Campus de Assis – FCLAs

_____

Membro da banca (1)

Prof. Dr. Allan Felipe Fattori Alves

Departamento de Biofísica e Farmacologia / UNESP / Campus de Botucatu – IBB

_____

Membro da banca (2)

Prof. Dr. Júlio Pinheiro Baima

Departamento de Clínica Médica / Hospital das Clínicas da Faculdade de Medicina de Botucatu

AGRADECIMENTOS

Primeiramente, agradeço aos meus pais, Marcia Cristine Antunes Almeida Fernandes e Marcos Eduardo Fernandes, por todo o apoio, incentivo constante e por todos os esforços dedicados à minha formação acadêmica e pessoal.

Agradeço ao meu orientador, Professor Doutor José Celso Rocha, e à minha coorientadora, Professora Doutora Lígia Yukie Sassaki, por terem me aceitado como orientado, pela confiança em meu trabalho e pela dedicação em realizar este projeto.

Agradeço à Universidade Estadual Paulista (UNESP), ao programa de Pós-Graduação em Ciências Biomoleculares e Farmacológicas, ao laboratório FitoFarmaTec e ao laboratório LaMAp, pela infraestrutura disponibilizada, pelo suporte oferecido e pelos conhecimentos compartilhados durante o mestrado.

Por fim, agradeço aos meus demais familiares, em especial à minha esposa, Natalia Camargo Faraldo, pelo amor, compreensão, incentivo e apoio incondicional ao longo desta etapa importante da minha vida. Agradeço igualmente aos meus amigos, cuja presença e apoio emocional foram fundamentais.

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

As Doenças Inflamatórias Intestinais (DIIs), representadas pela Doença de Crohn (DC) e a Retocolite Ulcerativa (RCU), apresentam elevada heterogeneidade clínica, dificultando a previsão de sua evolução e a definição de estratégias terapêuticas. Nesse contexto, este estudo desenvolveu e avaliou modelos de redes neurais artificiais multicamadas (RNA-MLP) para classificar o tipo de DII e predizer desfechos cirúrgicos e medicamentosos. Foram utilizados dados do registro nacional do GEDIIB, composto por 2.751 pacientes e 932 variáveis clínicas. Após o pré-processamento, foram construídos modelos para diferenciação entre DC e RCU, ocorrência de cirurgia relacionada à DII e utilização de terapia biológica, considerando bases globais e específicas para cada doença, em versões originais e balanceadas. O desempenho foi avaliado por acurácia, precisão, recall, F1-score e ROC-AUC, enquanto a metodologia SHAP foi utilizada para identificar as variáveis clínicas de maior influência nas predições. Os resultados demonstraram o potencial dos modelos em reconhecer padrões clínicos complexos e auxiliar na classificação das DIIs e na predição de desfechos cirúrgicos e terapêuticos.

Palavras-chave: Doença Inflamatória Intestinal; Inteligência Artificial; Desfecho cirúrgico; Terapia Biológica.

ABSTRACT

Inflammatory Bowel Diseases (IBDs), represented by Crohn's Disease (CD) and Ulcerative Colitis (UC), are characterized by substantial clinical heterogeneity, making it challenging to predict disease progression and define therapeutic strategies. In this context, this study developed and evaluated multilayer artificial neural network models (ANN-MLP) to classify IBD type and predict surgical and medication-related outcomes. Data from the GEDIIB national registry, comprising 2,751 patients and 932 clinical variables, were used. After data preprocessing, models were developed to differentiate CD from UC, predict the occurrence of IBD-related surgery, and predict the use of biological therapy, considering global and disease-specific datasets in both original and balanced versions. Model performance was evaluated using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC), while SHAP was applied to identify the clinical variables with the greatest influence on model predictions. The results demonstrated the potential of these models to recognize complex clinical patterns and assist in IBD classification and the prediction of surgical and therapeutic outcomes.

Keywords: Inflammatory Bowel Diseases; Artificial Intelligence; surgical outcome; biological therapy

SUMÁRIO

1. INTRODUÇÃO GERAL 8

1.1 INTRODUÇÃO 8

2. MATERIAIS E MÉTODOS 13

2.1 DESENVOLVIMENTO DOS DATASETS 13

2.2 TRATAMENTO DO DATASET 16

2.2.1 Remoção de variáveis irrelevantes 16

2.2.2 Codificação de variáveis categóricas (One hot Encoding) 17

2.2.3 Balanceamento dos dados 17

2.3 REDES NEURAIS ARTIFICIAIS MULTICAMADAS 18

2.3.1 Algoritmo MLPClassifier 19

2.4 ANÁLISES ESTATÍSTICAS 20

2.4.1 Receiver Operating Charactristic e Area Under the Curve 21

2.4.2 Matriz de Confusão 22

2.4.3 Métricas de avaliação do modelo 23

2.4.3.1 Precision 23

2.4.3.2 Recall 24

2.4.3.3 F1-Score 24

2.4.3.4 Accuracy 24

2.4.4 Comparação entre os modelos pré e pós balanceamento 24

2.5 - GRIDSEARCHCV 25

2.6 - DESENVOLVIMENTO DO MODELO 26

3. RESULTADOS 28

4. DISCUSSÃO 53

5. CONCLUSÕES 60

REFERÊNCIAS 63

APÊNDICE I 68

APÊNDICE II 72

1. Introdução Geral

Esta dissertação, apresentada ao Programa de Mestrado em Ciências Biomoleculares e Farmacológicas, está estruturada em cinco seções. Na **Introdução** são apresentados o contexto das Doenças Inflamatórias Intestinais (DII), com ênfase na Doença de Crohn (DC) e na Retocolite Ulcerativa (RCU), abordando aspectos epidemiológicos, limitações dos métodos diagnósticos e terapêuticos atuais e o potencial das técnicas de inteligência artificial (IA) na área da saúde. A seção **Materiais e Métodos** descreve, de forma detalhada, o conjunto de dados utilizados, incluindo a origem dos registros clínicos e epidemiológicos, o processo de desenvolvimento e organização dos datasets e as etapas de pré-processamento. Além disso, são apresentados os fundamentos do desenvolvimento do modelo de IA, baseado em Redes Neurais Artificiais Multicamadas (RNA MLP), estratégias de otimização utilizadas e métricas de avaliação de desempenho. Em **Resultados**, são apresentados os desempenhos dos modelos desenvolvidos para a predição do tipo de DII (DC ou RCU) e para os desfechos cirúrgico e medicamentoso. Os resultados são organizados por meio de tabelas, matrizes de confusão e curvas ROC, dentre outras métrica, permitindo a comparação entre diferentes cenários de treinamento e blind teste, incluindo análises com e sem balanceamento das classes. A seção **Discussão dos resultados**, apresenta a análise dos resultados obtidos, considerando o desempenho dos modelos no contexto clínico e no cenário de desequilíbrio entre classes, além de estabelecer uma comparação com estudos prévios que aplicaram técnicas de IA às DIIs. Por fim, a seção **Conclusões** sintetiza os principais achados do estudo, destacando o potencial da aplicação da IA a dados clínicos e epidemiológicos de pacientes com DIIs, especialmente na distinção entre DC e RCU e na predição de desfechos cirúrgicos e medicamentosos, sendo este último representado, no presente estudo, pelo uso de terapia biológica. A seção também aborda a influência do desequilíbrio entre classes no desempenho dos modelos, a necessidade de validação em estudos prospectivos e em novos conjuntos de dados independentes, bem como os desafios metodológicos e clínicos relacionados à incorporação dessas ferramentas na prática clínica.

1.1 Introdução

As Doenças Inflamatórias Intestinais (DIIs), que incluem a Doença de Crohn (DC) e a Retocolite Ulcerativa (RCU), são condições crônicas caracterizadas por

inflamação do trato gastrointestinal com etiologia multifatorial envolvendo predisposição genética, disfunções imunológicas e fatores ambientais. A DC pode afetar qualquer parte do trato digestivo, enquanto a RCU é limitada ao cólon e ao reto. Ambas as doenças apresentam sintomas como dor abdominal, diarreia e perda de peso, impactando significativamente a qualidade de vida dos pacientes, Junior *et al*, 2023.

O aumento significativo na incidência e prevalência das DIIs, que incluem DC e a RCU, tem sido amplamente documentado em estudos recentes. De acordo com Quaresma *et al*, 2022, dados do Sistema Único de Saúde (SUS) indicam que, entre 2012 e 2020, a incidência de RCU aumentou de 5,7 para 6,9 por 100.000 habitantes, enquanto a de DC diminuiu de 3,7 para 2,7 por 100.000 habitantes. Em contrapartida, a prevalência de ambas as doenças apresentou aumento expressivo, com a RCU passando de 15,8 para 56,5 por 100.000 habitantes, e a DC de 12,6 para 33,7 por 100.000 habitantes no mesmo período.

Atualmente, não há um tratamento farmacológico definitivo para as DIIs. As DIIs possuem uma etiologia multifatorial, envolvendo interações entre fatores genéticos, ambientais, imunológicos e microbiológicos. Essa complexidade dificulta a identificação de alvos terapêuticos específicos e eficazes para todos os pacientes, Loddo *et al*, 2015. Embora as opções terapêuticas disponíveis como aminossalicilatos, imunossupressores, agentes biológicos e pequenas moléculas sejam utilizadas para induzir e manter a remissão, muitos pacientes não alcançam uma resposta sustentada, enfrentando recorrências e complicações ao longo do tempo. Walfish *et al*, 2024 mostra que, em ensaios clínicos, apenas 20% a 30% dos pacientes atingem remissão clínica, e entre esses, cerca de metade mantém a remissão ao longo do tempo. Estudos indicam que aproximadamente 30% a 40% dos pacientes não respondem adequadamente ao tratamento medicamentoso, necessitando de intervenções cirúrgicas devido à intratabilidade clínica ou complicações da doença, Zaltman *et al*, 2018. A ausência de biomarcadores confiáveis para prever a resposta ao tratamento e a progressão da doença dificulta a personalização das terapias e a identificação precoce de pacientes que poderiam se beneficiar de abordagens mais agressivas, Atreya *et al*, 2024. Esses dados evidenciam a necessidade de novas abordagens terapêuticas e estratégias de predição de risco para melhorar os desfechos clínicos e a qualidade de vida dos pacientes. Além disso, os métodos de

diagnósticos tradicionais para as DIIs, como colonoscopia e biópsias, são invasivos, dispendiosos e podem apresentar riscos, como perfuração intestinal. A biópsia, embora essencial para a confirmação diagnóstica, é limitada pela variabilidade inter-observador na interpretação dos resultados e pela dificuldade de obtenção de amostras representativas em áreas de difícil acesso do trato gastrointestinal, Loftus, 2007.

A inteligência artificial (IA) é um avanço tecnológico que permite que computadores e máquinas simulem capacidades humanas, como aprender, compreender, resolver problemas, tomar decisões, ser criativos e agir com certo grau de autonomia. Na prática, sistemas com IA conseguem, por exemplo, reconhecer objetos, entender e responder à linguagem humana, aprender com novas informações e experiências, fazer recomendações e até atuar de forma independente em algumas tarefas, Stryker, 2024. Sua finalidade é solucionar problemas complexos por meio da correta interpretação de dados externos, aprendendo com essas informações e aplicando esse conhecimento adquirido para executar tarefas e enfrentar desafios de maneira adaptativa e flexível, Kaplan *et al*, 2019.

A aplicação de técnicas de IA, como aprendizado de máquina (*Machine Learning* - ML*)* e aprendizado profundo (*Deep Learning* - DL), tem avançado significativamente na área médica. O ML é um subcampo da IA que se concentra no desenvolvimento de algoritmos capazes de aprender e melhorar automaticamente a sua acurácia, a partir de dados fornecidos, sem serem explicitamente programados para cada tarefa específica. Esses algoritmos analisam grandes volumes de dados para identificar padrões e realizar previsões ou decisões com base nesses padrões, Paixão *et al*, 2022. O DL, por sua vez, é uma subárea do ML que utiliza redes neurais artificiais (RNAs) profundas para modelar e resolver problemas complexos. Essas redes são modelos de ML que aprendem a partir de exemplos por meio do treinamento, no qual um algoritmo ajusta seus pesos para gerar saídas que representem a informação presente nos dados; elas são compostas por múltiplas camadas de neurônios artificiais (*Perceptrons*), que processam os dados de forma hierárquica e permitem extrair características de maior nível a partir de dados brutos. A Figura 1 apresenta um panorama visual de diferentes arquiteturas/tipos de RNAs comumente empregadas em ML. Por meio de diagramas simplificados, a imagem

ilustra como as RNAs podem variar na organização das camadas e nas conexões entre neurônios.

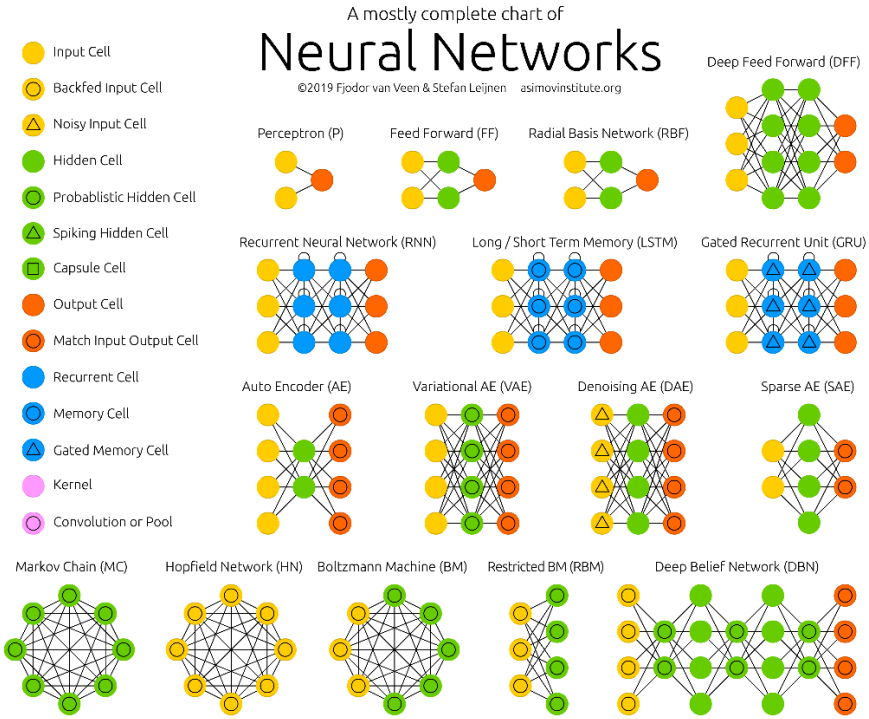

Figura 1: Arquiteturas de RNAs

Disponível em: https://www.asimovinstitute.org/neural-network-zoo/
Acesso em: 15/05/2025

O DL é particularmente eficaz em tarefas que envolvem grandes volumes de dados e alta complexidade, como reconhecimento de imagem e processamento de linguagem natural, Falcão *et al*, 2024. Essas tecnologias têm sido utilizadas para aprimorar diagnósticos, prever riscos e avaliar a gravidade de diversas doenças. Por exemplo, modelos de DL têm sido empregados para diagnosticar precocemente doenças gastrointestinais, aumentando a precisão diagnóstica e auxiliando na tomada de decisões clínicas, Sharma *et al*, 2023.

Outro exemplo, é a pesquisa desenvolvida por Zand *et al*, 2022 que traz o desenvolvimento e validação de modelos de ML para prever desfechos adversos em pacientes com DIIs, como hospitalizações, necessidade de cirurgias, uso prolongado de esteroides e início de terapias biológicas. Utilizando um conjunto de dados de 72.178 pacientes para treinamento e 69.165 para validação, os modelos alcançaram áreas sob a curva (AUC) entre 0,70 e 0,92, indicando alta acurácia na previsão desses

desfechos. Esses modelos podem ser aplicados para estratificação de risco e implementação de medidas preventivas em ambientes clínicos, contribuindo para uma abordagem mais personalizada no manejo das DIIs.

A perspectiva de que a IA pode revolucionar o campo da saúde, especialmente na gestão de doenças complexas como as DIIs, é cada vez mais reconhecida. Ao automatizar a análise de grandes volumes de dados e identificar padrões complexos, a IA tem o potencial de melhorar significativamente o diagnóstico, o prognóstico e o tratamento dessas condições, contribuindo para uma medicina mais personalizada e eficiente, Teim Junior *et al*, 2026.

No entanto, a análise das variáveis de entrada na IA, em nosso caso os dados dos pacientes, devem ser demasiadamente considerados. Atualmente, é possível utilizar métodos de Inteligência Artificial Explicável (*Explainable Artificial Intelligence – XAI)*, para identificar quais variáveis apresentam maior influência no processo de aprendizado e nas predições realizadas pelos modelos, Ali *et al,* 2023. Particularmente, a técnica SHAP (*SHapley Additive exPlanations)* atribui a cada variável um valor de contribuição para determinada predição, possibilitando identificar se sua presença favorece ou reduz a probabilidade de uma classe específica, Lundberg *et al,* 2017. Assim, o SHAP contribui para tornar os modelos de IA mais interpretáveis, permitindo avaliar quais características dos pacientes foram mais relevantes para os desfechos analisados e se essas associações são coerentes com o conhecimento clínico.

O objetivo da presente pesquisa, foi desenvolver modelos de ML, utilizando linguagem de programação Python® na aplicação de web de código aberto *Jupyter Notebook*, onde dados clínicos e epidemiológicos específicos dessas doenças foram aplicados para o treinamento de RNAs capazes de, ao final, prever o tipo de DII (DC ou RCU), desfecho cirúrgico e medicamentoso (definido nesta pesquisa como uso de terapia biológica). Tais modelos, poderão contribuir não apenas para o diagnóstico, prognóstico e tratamento mais eficaz dos pacientes, mas também na elaboração de políticas de saúde, tanto públicas quanto privadas, e no direcionamento de esforços e investimentos do setor de saúde para os pacientes com DII, especialmente no âmbito do SUS, que concentra a maior parte dos atendimentos no Brasil, mas ainda enfrenta importantes limitações devido à escassez de opções terapêuticas disponíveis, Quaresma *et al*, 2022.

Na busca de uma melhor interpretação dos resultados, também foi realizada uma análise de interpretabilidade dos resultados via técnica SHAP. Essa abordagem possibilitou identificar quais características clínicas e epidemiológicas dos pacientes exerceram maior influência na classificação do tipo de DII e dos desfechos cirúrgico e medicamentoso, contribuindo para uma compreensão mais transparente dos padrões aprendidos pelas RNAs e para a avaliação da coerência clínica dos resultados obtidos.

2. Materiais e Métodos

Nesta seção, serão detalhadas as etapas metodológicas desenvolvidas durante a pesquisa.

2.1 Desenvolvimento dos Datasets

Os *datasets* utilizados neste projeto foram fornecidos pela Organização Brasileira de Doença de Crohn e Colite (GEDIIB), organização responsável pela estruturação de um registro nacional voltado à caracterização clínica, epidemiológicas e assistencial de pacientes com DIIs no Brasil. A relevância desses *datasets* está relacionada não apenas ao seu volume amostral, mas também à sua origem nacional, multicêntrica e colaborativa, reunindo informações provenientes de diferentes instituições e serviços de saúde do país. Segundo Fróes *et al.* (2024), o Cadastro Nacional de Pacientes com DII foi desenvolvido com o objetivo de reunir informações de pacientes acompanhados em serviços públicos e privados de saúde, permitindo uma caracterização mais ampla do perfil epidemiológico das DIIs no Brasil. Os autores também destacam que, até então, havia escassez de dados clínicos nacionais detalhados sobre DII, especialmente considerando fatores associados à gravidade da doença e diferenças regionais no país.

Participaram do desenvolvimento da base de dados os seguintes membros e instituições: Ligia Yukie Sassaki (Faculdade de Medicina da Universidade Estadual Paulista - UNESP Botucatu), José Miguel Parente (UFPI - Hospital Universitário da Universidade Federal do Piauí), Renata de Sa Brito Froes (Gastromed), Roberto Kaiser Junior (Clínica Kaiser), Gilmara Pandolfo Zabot (Coloprocto Clínica do Aparelho Digestivo, Neogelia Pereira de Almeida Moraes (HGRS - Hospital Geral Roberto Santos), Adriana Ribas de Andrade (HSR - Hospital São Rafael), Francisco Guilherme Cancela e Penha (IPSEMG), Cyrla Zaltman (UFRJ - Universidade Federal do Rio de

Janeiro), Rogerio Serafim Parra (Hospital das Clínicas da Faculdade de Medicina - RPUSP), Genoile Oliveira Santana Silva (CLIAGEN - Clinica de Atenção em Gastroenterologia Especializada e Nutrição), Mardem Machado de Souza (Hospital Universitário Júlio Muller), Rafael Luis Luporni (UFSCar - Universidade Federal de São Carlos), Heitor Siffert Pereira de Souza (HUCFF - Hospital Universitario Clementino Fraga Filho), Sergio Lima Junior (Hospital Universitário João de Barros Barreto – UFPA). Os dados dos pacientes referem-se ao período de 2019 a 2024. O conjunto de dados está organizado em uma planilha contendo 1.456 registros relacionados a pacientes com DC e 1.295 registros relacionados a pacientes com RCU, somando um total de 2.751 pacientes. As variáveis integradas ao banco de dados compreendem aspectos clínicos e epidemiológicos fundamentais para a caracterização sociodemográfica, o acompanhamento evolutivo e a abordagem terapêutica das DIIs. O mapeamento reúne dados sobre a apresentação clínica e fenotípica da doença, comorbidades, intercorrências clínicas, tratamentos medicamentosos e intervenções cirúrgicas, conforme detalhados no Apêndice 1. Para fins de organização e caracterização do banco de dados, as variáveis foram agrupadas em diferentes categorias, de acordo com a natureza da informação clínica registrada. Essa estruturação permitiu melhor compreensão da composição do banco e facilitou as etapas posteriores de pré-processamento, seleção de variáveis e definição dos desfechos analisados. O primeiro grupo corresponde aos Dados do Atendimento, que contemplam informações gerais do paciente e do acompanhamento clínico, incluindo, por exemplo, o tipo de DII, presença de comorbidades, histórico de câncer e uso de tabaco. O segundo grupo reúne os registros referentes à Cirurgia, incluindo informações relacionadas à ocorrência ou não de procedimentos cirúrgicos, à via de realização da cirurgia, como laparotomia ou laparoscopia, e à sua classificação, como cirurgia abdominal ou perianal. As variáveis referentes às Complicações, que compõem o terceiro grupo, reúnem informações sobre eventos clínicos associados à evolução da doença. Esse conjunto inclui diferentes tipos de complicações, como infecção fúngica e herpes simples, além de outras intercorrências que podem refletir maior gravidade, complexidade clínica ou pior evolução durante o curso das DIIs. Por fim, o grupo Medicações inclui informações sobre os tratamentos utilizados, especialmente aquelas relacionadas ao uso de terapias específicas, como a terapia biológica. A Figura 2 apresenta os grupos mencionados anteriormente com alguns dos

registros como exemplo. Todos os 932 registros (variáveis), referentes aos pacientes utilizadas neste estudo, encontram-se descritas no Apêndice I.

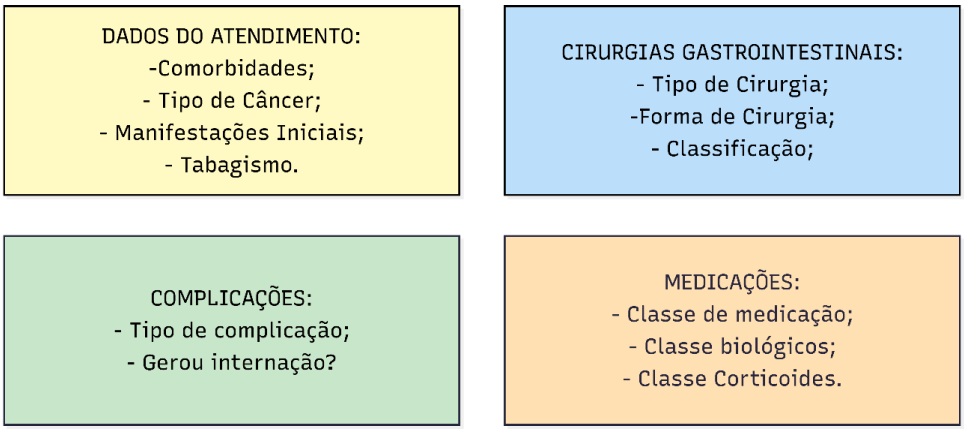

Figura 2: Estrutura das variáveis do banco de dados em diferentes categorias clínicas dos pacientes, incluindo dados do atendimento, cirurgias gastrointestinais, complicações e medicações, com exemplos de variáveis representativas de cada grupo.

Cada *dataset* foi importado no formato .xml (*Extensible Markup Language),* uma linguagem de marcação utilizada para estruturar dados de maneira hierárquica e legível, disponibilizada pelo REDCap no padrão CDISC ODM (*Operational Data Model*), amplamente empregado em estudos clínicos. Contudo, visando facilitar a análise e viabilizar a construção dos modelos de IA, foram desenvolvidos *notebooks* — documentos interativos que reúnem código executável, visualizações e texto explicativo — na aplicação *Jupyter Notebook,* utilizando linguagem de programação Python®, para realizar a conversão dos arquivos .xml para .csv (*Comma-Separated Values). A*lém do mapeamento de variáveis categóricas, da separação dos conjuntos de dados e da consolidação das planilhas, conforme apresentado no fluxograma da Figura 3. O primeiro documento, contém o código responsável por extrair informações específicas dos pacientes e dados epidemiológicos dos arquivos .xml, utilizando bibliotecas de manipulação e análise de dados, como o Pandas, e exportá-las para o formato .csv. Após a conversão dos datasets, o próximo documento tem como objetivo de mapear variáveis categóricas cujo valores variam de 0 a 10, por exemplo, aquelas que indicam o tipo de extensão anatômica da inflamação para RCU – e codifica-las por meio de *one hot encoding*. Assim, essas variáveis foram transformadas em colunas binárias, nas quais 1 indica presença e 0 indica ausência do desfecho. Posteriormente, foi desenvolvido outro documento que realiza a divisão dos dados em

80% para o aprendizado dos modelos e 20% para blind teste, destinado à avaliação de desempenho, resultando na base final utilizada para o desenvolvimento e validação do modelo.

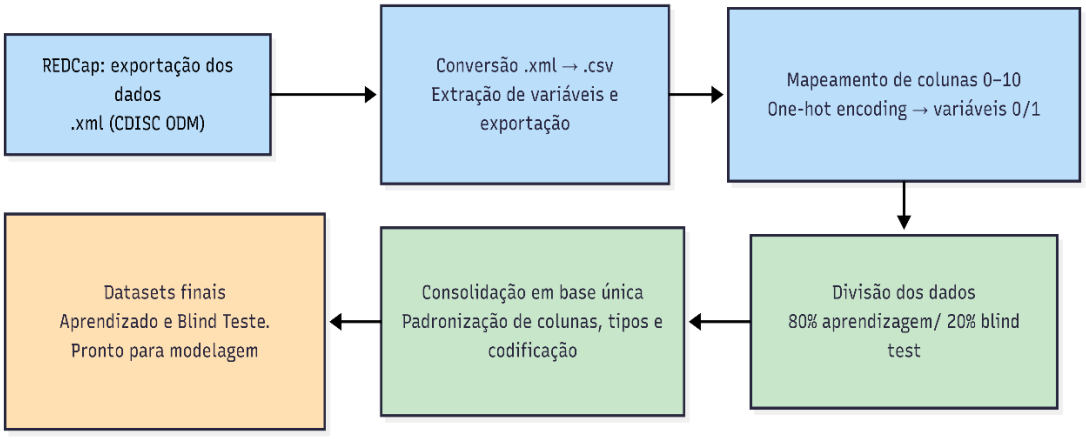

Figura 3: Fluxograma do desenvolvimento do dataset.

2.2 Tratamento do *dataset*

O tratamento de dados é uma etapa fundamental no processo de análise, sendo responsável por garantir qualidade, consistência e adequação dos dados para o treinamento do modelo, Landesman, 2024. Nesta pesquisa foram feitos os seguintes tratamentos:

2.2.1 Remoção de variáveis irrelevantes

A remoção de variáveis irrelevantes foi conduzida para reduzir ruído, aumentar a estabilidade do treinamento e minimizar o risco de vazamento de informação (*data leakage*). O *data leakage* ocorre quando informações que não estariam disponíveis no momento real da predição, ou que possuem relação direta com o desfecho a ser previsto, são indevidamente utilizadas no treinamento do modelo, Kaufman *et al*, 2012. No contexto de modelos preditivos clínicos, isso pode ocorrer, por exemplo, quando são incluídas variáveis registradas após a ocorrência do evento de interesse, ou altamente derivadas dele, levando o algoritmo a reconhecer indiretamente a resposta em vez de aprender padrões clínicos generalizáveis. Consequentemente, a presença dessas variáveis pode produzir métricas de desempenho artificialmente elevadas e menor capacidade de generalização em cenários reais de predição. Foram excluídas, portanto, variáveis consideradas sem valor analítico direto para o objetivo do estudo, incluindo datas aproximadas de término do uso de medicações e datas aproximadas de

internação. Também foram removidas variáveis relacionadas ao histórico familiar, devido à elevada proporção de valores ausentes, bem como variáveis indicadoras de “dado não disponível”, por representarem principalmente ausência de informação decorrente de falhas de preenchimento ou inconsistências no registro. Adicionalmente, atributos que poderiam introduzir viés por refletirem diretamente o resultado foram avaliados para exclusão, a fim de garantir que o modelo aprendesse padrões clínicos relevantes e generalizáveis.

2.2.2 Codificação de variáveis categóricas (*One hot Encoding)*

As variáveis categóricas foram codificadas para permitir sua utilização pelos algoritmos de ML, que requerem entradas numéricas estruturadas. Neste estudo, esse processo foi realizado por meio de um documento desenvolvido para identificar e mapear colunas categóricas com valores entre 0 e 10, convertendo-as por *one-hot encoding*. Com isso, cada categoria passou a ser representada por uma coluna binária, em que 1 indica a presença da característica e 0 sua ausência, evitando que os códigos originais fossem interpretados pelo modelo como uma escala numérica ordinal.

2.2.3 Balanceamento dos dados

Considerando que parte dos desfechos avaliados apresentava desequilíbrio entre classes, foi realizada uma etapa específica de balanceamento dos dados com o objetivo de reduzir o viés do modelo em favor da classe majoritária e melhorar a identificação da classe minoritária, geralmente a de maior interesse clínico. Essa etapa foi aplicada principalmente aos desfechos cirúrgico e medicamentoso, nos quais a distribuição assimétrica poderia inflar métricas globais e comprometer medidas por classe, como *Recall* e *F1-Score*.

O balanceamento foi conduzido de forma controlada e comparativa: os modelos foram treinados e avaliados tanto com a distribuição original quanto com a versão balanceada dos dados, permitindo quantificar o impacto dessa estratégia no desempenho.

Tabela 1: Distribuição dos dados para os desfechos

Desfecho	Dados	Classe	Treinamento	Blind Teste
DII	2751	DC RCU	1170 1043	286 252
Cirurgia (DC/RCU)	2751	Ausência Presença	1961 252	479 59
Cirurgia (DC)	1456	Ausência Presença	950 220	236 50
Cirurgia (RCU)	1295	Ausência Presença	1011 32	243 9
Cirurgia c/ bal (DC/RCU)	693	Ausência Presença	310 252	72 59
Cirurgia c/ bal (DC)	595	Ausência Presença	263 220	62 50
Cirurgia c/ bal (RCU)	95	Ausência Presença	43 32	11 9
Medicação (DC/RCU)	2751	Ausência Presença	1032 1181	251 287
Medicação (DC)	1456	Ausência Presença	307 863	61 225
Medicação (RCU)	1295	Ausência Presença	725 318	190 62
Medicação c/ bal (DC/RCU)	2626	Ausência Presença	1032 1081	251 262
Medicação c/ bal (DC)	779	Ausência Presença	307 339	61 72
Medicação c/ bal (RCU)	810	Ausência Presença	355 318	75 62

*bal=balanceamento

A Tabela 1 apresenta a distribuição das amostras por desfecho, classe e conjunto de dados, incluindo os cenários com e sem balanceamento, permitindo visualizar a estrutura das bases utilizadas e o grau de desequilíbrio entre as classes em cada situação analisada.

2.3 Redes Neurais Artificiais Multicamadas

As redes neurais artificiais multicamadas (RNAs MLP), também denominadas *Multilayer Perceptron (MLP)*, como o próprio nome indica, é composta por múltiplas camadas, Figura 4. Diferentemente da RNA de camada única, que resolve apenas problemas linearmente separáveis, as RNAs MLP conseguem lidar com problemas mais complexos e não linearmente separáveis. Isso é possível graças à adição de

uma ou mais camadas ocultas ao modelo. Essa estrutura é conhecida como rede neural *feed-forward,* caracterizada pelo fluxo unidirecional de informações entre as camadas. São usadas geralmente para reconhecimento de padrões, classificação de padrões de entrada, predição com base nas informações de entrada e aproximação, Sonawane *et al,* 2014.

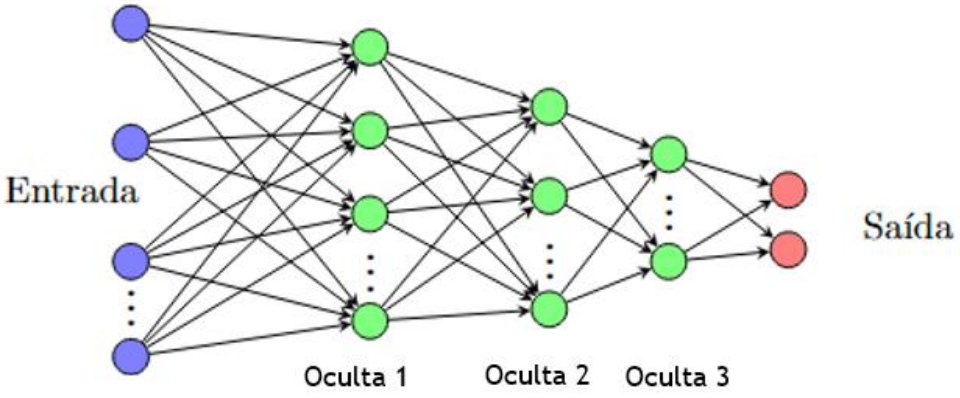

Figura 4: Exemplo de RNA MLP

O algoritmo de retropropagação (*backpropagation)* é amplamente utilizado para o treinamento de RNAs, especialmente em RNAs MLP. Esse método funciona comparando a saída gerada pela rede com a saída esperada (ou alvo), calculando o erro resultante dessa comparação. Em seguida, esse erro é realimentado na rede, ajustando os pesos das conexões com o objetivo de aproximar cada vez mais a saída produzida da saída desejada. Esse ciclo de ajuste é repetido diversas vezes, e a cada iteração o erro tende a diminuir, melhorando a precisão da rede. Esse procedimento iterativo é conhecido como o processo de 'aprendizado' da RNA, Markov, 2023.

2.3.1 Algoritmo *MLPClassifier*

O algoritmo *MLPClassifier* da biblioteca *Scikit-Learn*, foi selecionado para treinar os modelos e realizar a predição do tipo de DII (DC ou RCU), bem como dos desfechos cirúrgicos e medicamentoso, devido à sua capacidade de capturar relações não lineares complexas. Trata-se de um algoritmo de classificação baseado em RNA MLP, uma rede *feed-forward* composta por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída, conforme descrito no tópico 2.3.

Em uma RNA MLP, é importante distinguir parâmetros de hiperparâmetros. Os parâmetros são valores aprendidos diretamente a partir dos dados durante o treinamento da RNA – como os pesos e vieses (Bias) de cada camada – e são ajustados iterativamente para minimizar a função de perda, permitindo que o modelo represente o mapeamento entre as variáveis de entrada e saída. Já os hiperparâmetros são definidos previamente pelo desenvolvedor e controlam a capacidade e o comportamento do processo de aprendizado, incluindo a arquitetura da rede (número de camadas ocultas e neurônios), a função de ativação, o método de otimização, a taxa de aprendizado, critérios de convergência e mecanismos de regularização. Dessa forma, os parâmetros correspondem ao que o modelo de fato aprende a partir dos dados, enquanto os hiperparâmetros determina a configuração e a estratégia de treinamento que conduzem esse aprendizado, influenciando diretamente o equilíbrio entre desempenho no treinamento e capacidade de generalizar para dados novos, Deep Learning Book, 2025; Goodfellow *et al,* 2016.

Em resumo, os parâmetros são os “ajustes internos” que o modelo aprende automaticamente com os dados para fazer as previsões. Já os hiperparâmetros, são as configurações definidas antes do treino, que dizem como esse aprendizado vai acontecer. No *MLPClassifier*, essas configurações definem o “tamanho e o formato” da rede e a forma de treinamento, impactando diretamente o equilíbrio entre um modelo simples demais (não capta bem os padrões), complexo demais (pode “decorar” os dados) ou bem ajustado, com melhor chance de manter bom desempenho em novos pacientes/dados.

2.4 Análises estatísticas

Em dados clínicos e epidemiológicos como, por exemplo, nas DIIs é comum que o desfecho de interesse (por exemplo, cirurgia, óbito ou medicação) seja relativamente menos prevalente do que a ausência do desfecho, caracterizando um cenário de desequilíbrio de classe. Essa situação é consistente com o que se observa em problemas de predição em saúde, nos quais a prevalência de desfecho tende a ser menor do que a de “não-desfecho”. Nesses contextos, modelos preditivos podem ser “dominados” pela classe majoritária e acabar subestimando ou ignorando a classe minoritária (o desfecho), o que demanda atenção na escolha das métricas, na estratégia de modelagem e nas análises estatísticas, Blagus *et al*, 2018.

Em cenários como esses, onde há um desequilíbrio nas classes, métricas baseadas em Precisão *(Precision)* e Sensibilidade *(Recall)* tendem a ser mais informativa do que aquelas fundamentadas em sensibilidade e especificidade, como a curva de Característica de Operação do Receptor (ROC). Embora amplamente utilizada, a ROC deve ser interpretada com cautela em bases desbalanceadas, pois pode oferecer uma visão excessivamente otimista do desempenho do modelo, Saito *et al*, 2015.

Neste sentido para avaliar o desempenho dos modelos de forma mais robusta, utilizamos as métricas; área abaixo da curva ROC (AUC), precisão (*Precision),* sensibilidade (*Recall), F1-Score* e acurácia. Em especial, *Precision* e *Recall* foram priorizadas na interpretação dos resultados por refletirem com maior fidelidade a capacidade do modelo em identificar corretamente a classe minoritária (desfecho), reduzindo o risco de conclusões enviesadas pela predominância da classe majoritária. Além disso, a AUC e ROC foram reportadas como medidas globais de discriminação, porém sempre interpretada em conjunto com as métricas orientadas à classe positiva e com as matrizes de confusão, permitindo uma análise mais completa do comportamento do classificador em diferentes tipos de erros (falsos positivos e falsos negativos), mais relevantes no contexto clínico.

2.4.1 Receiver Operating Charactristic e Area Under the Curve

A curva de Característica de Operação do Receptor ou curva ROC (do inglês *Receiver Operating Characteristic*) é uma representação gráfica da performance de um classificador binário. Ela mostra a relação entre a taxa de verdadeiros positivos (TPR), no eixo Y, e a taxa de falsos positivos (FPR), no eixo X, para diferentes limiares de decisão, Fawcett, 2006.

A área abaixo da curva ROC ou AUC (do inglês *Area Under the Curve*) corresponde à probabilidade de o modelo atribuir um escore maior a um caso positivo do que a um caso negativo, quando ambos são escolhidos aleatoriamente. Um AUC próximo de 1,0 indica excelente separação entre classes, enquanto um valor de 0,5 indica performance aleatória, Li, 2024.

A Figura 5 a seguir, apresenta um modelo da interpretação das curvas ROC. Curvas mais próximas do canto superior esquerdo indicam melhor capacidade de discriminação, pois combinam alta sensibilidade com baixa taxa de falsos positivos,

enquanto curvas próximas à diagonal (classificador aleatório) sugerem desempenho semelhante ao acaso. A AUC sintetiza esse comportamento em um único valor entre 0 e 1: quanto maior a AUC, maior o poder discriminativo do modelo. Dessa forma, as anotações “Melhor” e “Pior” reforçam visualmente que, à medida que a curva se afasta da diagonal e se aproxima do topo à esquerda, o classificador tende a apresentar desempenho superior.

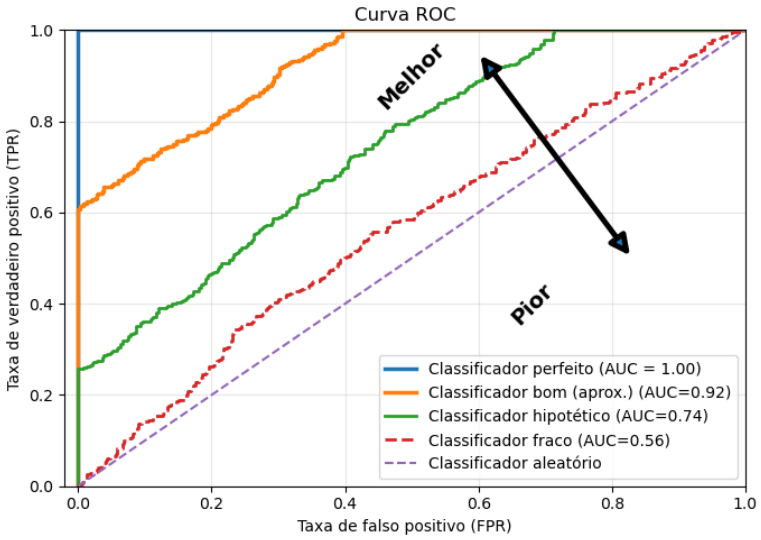

Figura 5: Modelos de curva ROC

2.4.2 Matriz de Confusão

Uma outra forma de avaliar os resultados gerados pelas RNAs MLP é por meio da matriz de confusão. Esse recurso permite comparar as predições do modelo com os valores reais (gabarito), evidenciando os acertos e erros de classificação. Ela mostra a quantidade de previsões corretas e incorretas, categorizadas por classe. Com isso, é possível calcular a acurácia global do sistema, bem como o desempenho por classe, Provost *et al*, 2013. A matriz de confusão organiza, em formato de tabela, a comparação entre os valores reais (linhas) e os valores previstos pelo modelo (colunas), como ilustrado na Figura 6. A célula Verdadeiro Positivo (TP) representa os casos em que o desfecho era positivo e foi corretamente previsto como positivo, enquanto o Verdadeiro Negativo (TN) indica os casos negativos corretamente classificados como negativos. Já os erros de classificação são representados por Falso Positivo (FP), quando o caso é negativo na realidade, mas o modelo prevê como positivo, e por Falso Negativo (FN), quando o caso é positivo na realidade, porém o modelo o prevê como negativo. Assim, a matriz permite visualizar de forma direta tanto

os acertos quanto os tipos de erro cometidos pelo classificador, servindo de base para métricas como acurácia, sensibilidade (*Recall*) e precisão.

	Previsto: Positivo	**Previsto: Negativo**	
Real: Positivo	Verdadeiro Positivo (TP)	Falso Negativo (FN)	Recall
Real: Negativo	Falso Positivo (FP)	Verdadeiro Negativo (TN)	Recall
	Precision	Precision	Acurácia

Figura 6: Matriz de Confusão

2.4.3 Métricas de avaliação do modelo

A avaliação do modelo foi realizada por meio de medidas derivadas da matriz de confusão, com destaque para *Precision, Recall, F1-Score e Accuracy*, por permitirem avaliar tanto o desempenho geral quanto a quantidade das predições das classes. Entretanto, não existem valores universais considerados ideais para essas métricas, uma vez que sua interpretação depende do contexto do problema e do custo associado aos erros de classificação. Em cenários com desbalanceamento entre classes, como frequentemente observado em dados clínicos, a acurácia pode apresentar resultados inflados, sendo recomendada a utilização de métricas como *Precision*, *Recall* e *F1-Score* para uma avaliação mais robusta do modelo, Saito *et al*, 2015. Em aplicações médicas, especialmente, costuma-se priorizar o *Recall* da classe positiva, a fim de minimizar falsos negativos, que podem representar falhas críticas na identificação de pacientes com necessidade de intervenção.

2.4.3.1 Precision

A *Precision* corresponde à proporção de verdadeiros positivos entre todas as predições positivas realizadas pelo modelo. Essa métrica é útil quando se busca avaliar a confiabilidade das previsões de “presença” do desfecho, priorizando a redução de falsos positivos, Manning *et al*, 2009. Assim, a precisão é calculada conforme equação 1, a seguir:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

2.4.3.2 Recall

O *Recall* (Sensibilidade) corresponde à proporção de verdadeiros positivos identificados pelo modelo em relação ao total de positivos reais. Essa métrica é relevante quando o objetivo é maximizar a detecção de casos com “presença” do desfecho, reduzindo a ocorrência de falsos negativos, Powers *et al*, 2020. Assim, o *Recall* é calculado conforme equação 2, a seguir:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

2.4.3.3 F1-Score

O *F1-Score* corresponde à média harmônica entre *Precision* e *Recall*, sintetizando em um único valor o equilíbrio entre a confiabilidade das predições positivas e a capacidade do modelo de recuperar os positivos reais. Essa métrica é útil quando se busca um desempenho mais balanceado entre falsos positivos e falsos negativos, Powers *et al*, 2020. O cálculo do *F1-Score* é calculado conforme a equação 3, a seguir:

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (3)$$

2.4.3.4 Accuracy

A *Accuracy* (acurácia) corresponde à proporção total de predições corretas do modelo, considerando tanto verdadeiros positivos quanto verdadeiros negativos, em relação ao total de observações. Essa métrica fornece uma visão geral do desempenho, mas deve ser interpretada com cautela quando há desequilíbrio de classes, pois pode ser influenciada pela classe majoritária Cabot *et al*, 2023. A acurácia é calculada conforme equação 4, a seguir:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

2.4.4 Comparação entre os modelos pré e pós balanceamento

A comparação das métricas de desempenho dos modelos antes e após o balanceamento dos dados foi realizada por meio do teste t de Student para amostras pareadas. As diferenças foram consideradas estatisticamente significativas quando o $p < 0,05$. Dessa forma, foi possível avaliar se o procedimento de balanceamento

promoveu alterações significativas no desempenho dos modelos analisados. Os resultados obtidos estão apresentados no Apêndice II.

2.5 - GridSearchCV

O *GridSearchCV* é um método de otimização de hiperparâmetros que realiza uma busca exaustiva sobre uma grade de valores previamente definida pelo desenvolvedor. Nesse processo, são testadas todas as combinações possíveis dos hiperparâmetros e cada configuração é avaliada por validação cruzada (*cross-validation*), selecionando-se aquela que apresenta o melhor desempenho médio. Essa abordagem é amplamente utilizada por tornar o ajuste do modelo padronizado e reprodutível, reduzindo a subjetividade e o risco de escolhas arbitrárias, Pedregosa *et al,* 2011. Conforme mostrado na Figura 7, o eixo X representa os valores possíveis do Hiperparâmetro 1 (a, b, c) e o eixo Y representa os valores do Hiperparâmetro 2 (x, y, z). Cada ponto no gráfico corresponde a uma combinação específica entre esses valores (por exemplo, b-y ou c-z) e, na prática, o *GridSearchCV* treina e avalia um modelo para cada ponto da grade, usando a validação cruzada para obter um desempenho médio mais confiável. Ao final, ele compara os resultados de todas as combinações testadas e seleciona aquela que apresenta a melhor métrica, retornando os melhores hiperparâmetros e o melhor modelo.

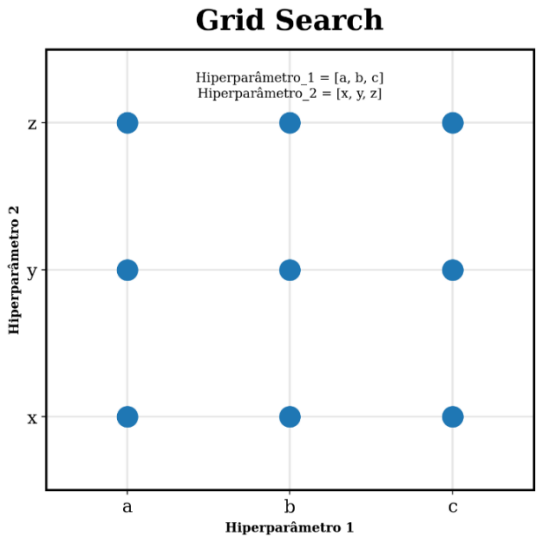

Figura 7: Ilustração do Grid Search para a seleção de hiperparâmetros.

2.6 - Desenvolvimento do modelo

Nesta pesquisa, foi proposto o uso de uma RNA do tipo MLP para realizar a classificação binária dos desfechos relacionados às DIIs, contemplando a distinção entre DC e RCU, bem como desfechos cirúrgicos e medicamentosos, lembrando que, em nosso estudo, este último é definido como uso de terapia biológica. Após o desenvolvimento e o pré-processamento dos datasets, empregou-se o *GridSearchCV* para a seleção dos hiperparâmetros mais adequados, com base no melhor desempenho médio em validação cruzada. Para essa etapa, foram avaliados os seguintes valores de hiperparâmetros para cada desfecho:

A) Número de camadas ocultas: define quantas etapas internas de processamento a rede terá entre a entrada e a saída. Mais camadas aumentam a capacidade de modelar padrões complexos, mas também elevam o risco de sobreajuste e o custo de treinamento. Para este estudo, foram avaliadas arquiteturas com até 3 camadas ocultas, a fim de comparar diferentes níveis de complexidade e identificar a configuração com melhor equilíbrio entre desempenho e generalização;

B) Número de neurônios por camada: define quantas “unidades de processamento” existem em cada camada oculta. Mais neurônios aumentam a flexibilidade do modelo para capturar relações nos dados, porém podem aumentar o sobreajuste e o tempo de treino. Neste estudo, foram avaliadas diferentes configurações de neurônios por camada (200,100,50), (100,50) e (200) com o objetivo de comparar arquiteturas de complexidade distinta e verificar quais delas oferecem melhor desempenho e maior estabilidade, considerando os desfechos analisados;

C) Número de épocas: define o limite de quantas vezes o algoritmo pode ajustar o modelo durante o treinamento. Valores maiores permitem mais ajustes, mas podem aumentar o tempo de treinamento e, em alguns casos, favorecer sobreajuste. Foram considerados valores (100, 300, 500 e 600) para garantir tempo suficiente de convergência em diferentes arquiteturas, sem restringir indevidamente o ajuste do modelo;

D) Função de ativação: define como a rede transforma os valores dentro de cada neurônio para introduzir não linearidade. Essa escolha influencia o tipo de padrão que a rede consegue aprender e a estabilidade do treinamento. Foram avaliadas as funções disponíveis no *MLPClassifier (tanh, relu, identity e logistic);*

E) Algoritmo de otimização: define o método usado para atualizar os parâmetros (pesos) durante o treinamento, buscando reduzir o erro do modelo. A escolha afeta velocidade de convergência, estabilidade e desempenho final. Foram testados os algoritmos disponíveis no *MLPClassifier (sgd, adam e lbfgs);*

F) Coeficiente de regularização: controla a intensidade da regularização (penalização) aplicada aos pesos para reduzir sobreajuste. Valores maiores impõem mais restrição ao modelo; valores menores permitem maior flexibilidade. Foram avaliados os valores (0,001 e 0,01) para comparar diferentes níveis de regularização e selecionar aquele que melhor preserva o desempenho em validação, especialmente em um contexto de múltiplas variáveis clínicas;

G) Taxa de aprendizado: define o tamanho do passo inicial das atualizações dos parâmetros durante o treinamento. Taxas muito altas podem dificultar a convergência; taxas muito baixas podem tornar o treinamento lento. Foram avaliados os valores (0,001 e 0,01), visando equilibrar estabilidade e velocidade de aprendizado, considerando que arquiteturas e os algoritmos de otimização diferentes podem responder de forma distinta a esse parâmetro;

H) Esquema de ajuste da taxa de aprendizado: define como a taxa de aprendizado se comporta ao longo do treinamento (por exemplo, manter fixa ou reduzir conforme o treino avança). Esse ajuste influencia a convergência e a estabilidade do aprendizado. Foram avaliadas as funções disponíveis no *MLPClassifier (constant, adaptive, invscaling);*

Inicialmente, foram utilizadas as combinações de hiperparâmetros identificadas como mais adequadas na busca em grade, descrita no tópico 2.5, contemplando

arquiteturas com 1, 2 e 3 camadas ocultas. Em seguida, cada modelo foi executado 300 vezes, a fim de garantir robustez na avaliação de desempenho. A partir desse conjunto de execuções, foram selecionados os três melhores resultados para cada desfecho e para cada arquitetura avaliada. Por fim, os modelos são comparados com base nas métricas definidas para este estudo (AUC-ROC, precisão, *Recall*, *F1-Score* e acurácia), juntamente com a análise das matrizes de confusão, permitindo interpretar o desempenho de forma integrada e alinhada ao contexto clínico e ao desequilíbrio entre classes.

A Figura 8 apresenta o fluxograma dessas etapas de desenvolvimento e avaliação dos modelos *MLPClassifier*, incluindo a seleção de hiperparâmetros via GridSearchCV, a execução dos modelos 300 vezes para cada desfecho (tipo de DII, cirúrgico e medicamentoso) e a seleção dos três melhores resultados.

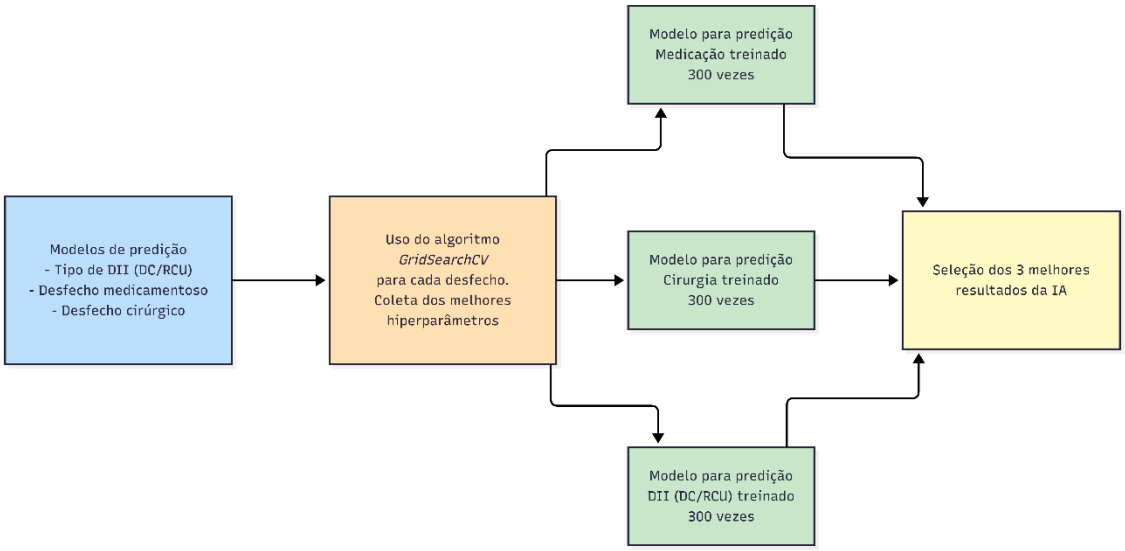

Figura 8: Fluxograma das etapas de desenvolvimento e avaliação dos modelos

3. Resultados

Neste tópico, são apresentados os resultados obtidos, com base nas métricas definidas no tópico 2.4, pelos três modelos selecionados e treinados com o *MLPClassifier* para a predição do tipo de DII (DC/RCU) e dos desfechos cirúrgicos e medicamentosos.

Nas Tabelas de 2 a 8, são apresentadas as métricas de desempenho dos modelos desenvolvidos para cada desfecho analisado, permitindo a comparação entre

as diferentes configurações avaliadas, tanto nos conjuntos de treinamento quanto de blind teste. Essas métricas são amplamente utilizadas em aplicações na área da saúde por permitirem avaliar não apenas o desempenho global do modelo, mas também sua capacidade de identificar corretamente eventos clínicos de interesse. Em particular, métricas como *Recall* são importantes em cenários clínicos, pois indicam a capacidade do modelo em detectar casos verdadeiros da condição avaliada, reduzindo a ocorrência de falsos negativos, que podem impactar diretamente a tomada de decisão clínica.

Tabela 2: Desempenho do melhor modelo de Inteligência Artificial por desfecho nas bases de treinamento e blind teste, estratificado por classe.

Desfecho	Base	Classe	Precision	Recall	F1-Score
DII	Treinamento	DC	99%	99%	99%
		RCU	99%	99%	99%
	Blind Teste	DC	99%	98%	99%
		RCU	98%	99%	99%
Cirurgia	Treinamento	Ausência	100%	100%	100%
		Presença	100%	100%	100%
	Blind Teste	Ausência	95%	94%	95%
		Presença	56%	58%	57%
Medicação	Treinamento	Ausência	75%	65%	70%
		Presença	72%	80%	76%
	Blind Teste	Ausência	83%	71%	77%
		Presença	78%	87%	82%

* >= 70% (Bom)/ >= 80% (Muito bom) / >= 90% (Excelente)

Para título de exemplo, considerando o desfecho cirurgia apresentado na Tabela 2, observa-se que na etapa de treinamento, o modelo da IA apresentou desempenho máximo, identificando os padrões de ausência e presença de cirurgia com 100% em *Precision*, *Recall* e *F1-Score*, indicando que, nos dados utilizados para treino, o modelo foi capaz de classificar corretamente todos os casos. Entretanto, ao avaliar o desempenho no conjunto de Blind Teste, nota-se redução nas métricas, especialmente para a classe presença cirurgia.

Para a classe ausência, a *Precision* foi de 95%, indicando que, entre os casos preditos pelo modelo como ausência de cirurgia, 95% estavam corretos. Já para a classe presença, a *Precision* foi de 56%, mostrando que, entre os casos classificados como presença de cirurgia, apenas 56% correspondiam de fato a pacientes com esse desfecho.

Em relação ao *Recall*, a classe ausência apresentou 94%, o que demonstra que o modelo conseguiu identificar corretamente 94% dos casos realmente negativos para cirurgia. Para a classe presença, o *Recall* foi de 58%, indicando menor capacidade de identificar corretamente os pacientes que de fato apresentavam o desfecho cirúrgico.

No *F1-Score*, que representa o equilíbrio entre *Precision* e *Recall*, a classe ausência alcançou 95%, evidenciando desempenho elevado e consistente. Em contrapartida, a classe presença apresentou *F1-Score* de 57%, reforçando que, apesar do excelente ajuste nos dados de treinamento, o modelo apresentou limitação na generalização para identificar adequadamente os casos positivos no conjunto de Blind Test.

Neste caso específico (desfecho cirurgia) o desbalanceamento entre as classes pode ter levado a essa limitação nos dados de blind test, como será explicado mais adiante.

Esse tipo de análise permite compreender de forma mais detalhada o comportamento do modelo em cada classe, indo além da avaliação global de desempenho.

Da mesma forma, essa interpretação pode ser realizada para as tabelas de 3 a 8, possibilitando uma análise mais completa e criteriosa dos resultados obtidos em cada desfecho.

Tabela 3: Comparação do desempenho do modelo de Inteligência Artificial para o desfecho "Cirurgia" com e sem balanceamento, nas bases de treinamento e blind teste, estratificado por classe.

Desfecho	Base	Classe	Precision	Recall	F1-Score
Cirurgia	Treinamento	Ausência	100%	100%	100%
		Presença	100%	100%	100%
	Blind Teste	Ausência	95%	94%	95%
		Presença	56%	58%	57%
Cirurgia Balanceado	Treinamento	Ausência	100%	100%	100%
		Presença	100%	100%	100%
	Blind Teste	Ausência	85%	83%	84%
		Presença	80%	81%	81%

Tabela 4: Comparação do desempenho do modelo de Inteligência Artificial para o desfecho "Cirurgia" no subgrupo DC, com e sem balanceamento, nas bases de treinamento e blind teste, estratificado por classe.

Desfecho	Base	Classe	Precision	Recall	F1-Score
Cirurgia DC	Treinamento	Ausência	97%	98%	97%
		Presença	91%	86%	89%
	Blind Teste	Ausência	92%	97%	94%
		Presença	83%	58%	68%
Cirurgia DC Balanceado	Treinamento	Ausência	96%	91%	94%
		Presença	90%	95%	93%
	Blind Teste	Ausência	87%	87%	87%
		Presença	84%	84%	84%

Tabela 5: Comparação do desempenho do modelo de Inteligência Artificial para o desfecho “Cirurgia” no subgrupo RCU, com e sem balanceamento, nas bases de treinamento e blind teste, estratificado por classe.

Desfecho	Base	Classe	Precision	Recall	F1-Score
Cirurgia RCU	Treinamento	Ausência	100%	100%	100%
		Presença	86%	89%	88%
	Blind Teste	Ausência	99%	97%	98%
		Presença	46%	67%	55%
Cirurgia RCU Balanceado	Treinamento	Ausência	97%	83%	89%
		Presença	82%	96%	89%
	Blind Teste	Ausência	100%	64%	78%
		Presença	69%	100%	82%

Tabela 6: Comparação do desempenho do modelo de Inteligência Artificial para o desfecho “Medicação” com e sem balanceamento, nas bases de treinamento e blind teste, estratificado por classe.

Desfecho	Base	Classe	Precision	Recall	F1-Score
Medicação	Treinamento	Ausência	75%	65%	70%
		Presença	72%	80%	76%
	Blind Teste	Ausência	83%	71%	77%
		Presença	78%	87%	82%
Medicação Balanceado	Treinamento	Ausência	78%	66%	71%
		Presença	71%	82%	76%
	Blind Teste	Ausência	83%	72%	77%
		Presença	76%	86%	81%

Tabela 7: Comparação de desempenho do modelo de Inteligência Artificial para o desfecho "Medicação" no subgrupo DC, com e sem balanceamento, nas bases de treinamento e blind teste, estratificado por classe.

Desfecho	Base	Classe	Precision	Recall	F1-Score
Medicação DC	Treinamento	Ausência	55%	7%	12%
		Presença	74%	98%	84%
	Blind Teste	Ausência	86%	10%	18%
		Presença	80%	100%	89%
Medicação DC Balanceado	Treinamento	Ausência	66%	40%	50%
		Presença	58%	80%	68%
	Blind Teste	Ausência	71%	39%	51%
		Presença	63%	86%	73%

Tabela 8: Comparação do desempenho do modelo de Inteligência Artificial para o desfecho "Medicação" no subgrupo RCU, com e sem balanceamento, nas bases de treinamento e blind teste, estratificado por classe.

Desfecho	Base	Classe	Precision	Recall	F1-Score
Medicação RCU	Treinamento	Ausência	77%	86%	81%
		Presença	57%	42%	48%
	Blind Teste	Ausência	84%	87%	86%
		Presença	56%	50%	53%
Medicação RCU Balanceado	Treinamento	Ausência	69%	72%	70%
		Presença	69%	65%	67%
	Blind Teste	Ausência	79%	76%	78%
		Presença	72%	76%	74%

As Figuras 9 a 21 reúnem as matrizes de confusão dos modelos avaliados para os diferentes desfechos do estudo. Por meio delas, é possível analisar de forma detalhada o padrão de classificação dos modelos, evidenciando os acertos e erros em cada classe, tanto no treinamento quanto no blind teste. No contexto da saúde, essa

análise é relevante, pois possibilita compreender os diferentes tipos de erro cometidos pelo modelo, o que é fundamental em aplicações clínicas, nas quais erros de classificação podem ter implicações distintas para o manejo do paciente, como atrasos no diagnóstico ou indicação inadequada de intervenções terapêuticas.

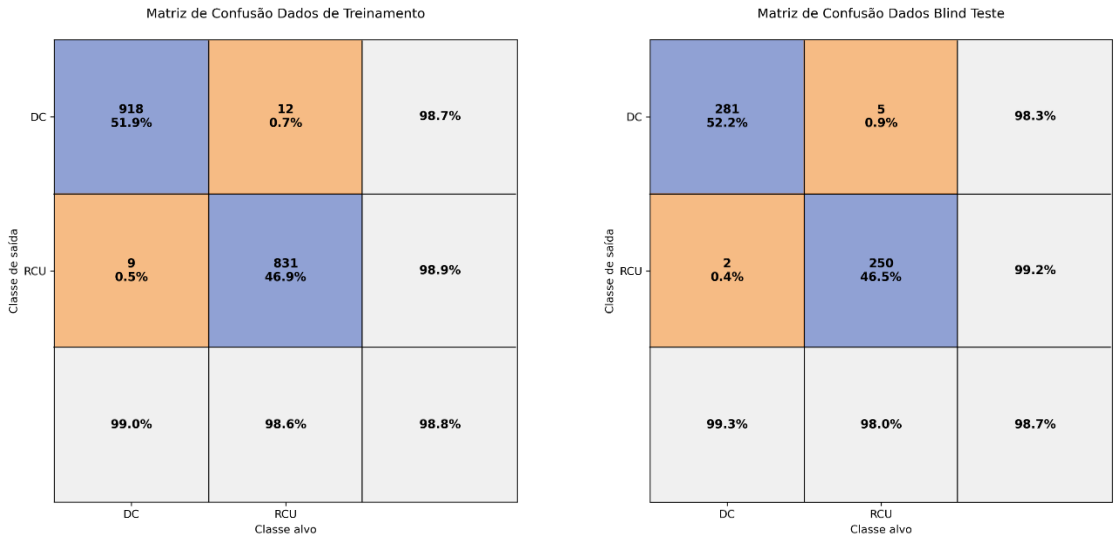

Figura 9: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Tipo de DII”.

Para um melhor entendimento da matriz de confusão, utilizaremos a figura 9 como exemplo modelo, para entendimento dos valores que aparecem na matriz. Nas linhas da matriz estão representados os valores reais de cada uma das classes e nas colunas estão representados os valores previstos para cada uma das classes. Os valores na diagonal azul representam os acertos e os valores na diagonal laranja os erros. As porcentagens dentro das células indicam a proporção em relação ao total de amostras. As margens à direita, em cinza, correspondem ao *Recall* por classe e as margens inferiores, também em cinza, correspondem à *Precision* por classe, enquanto o canto inferior direito, em cinza, apresenta a métrica global do modelo, cujo valor ideal é o mais próximo possível de 100%, embora sua interpretação deva considerar o contexto do problema e o balanceamento entre as classes.

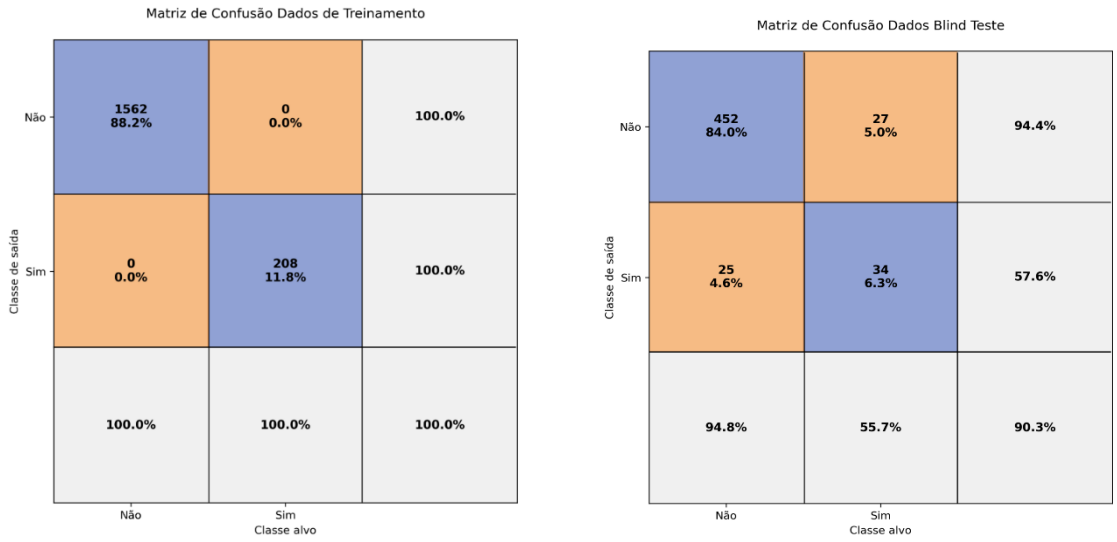

Figura 10: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Cirurgia”.

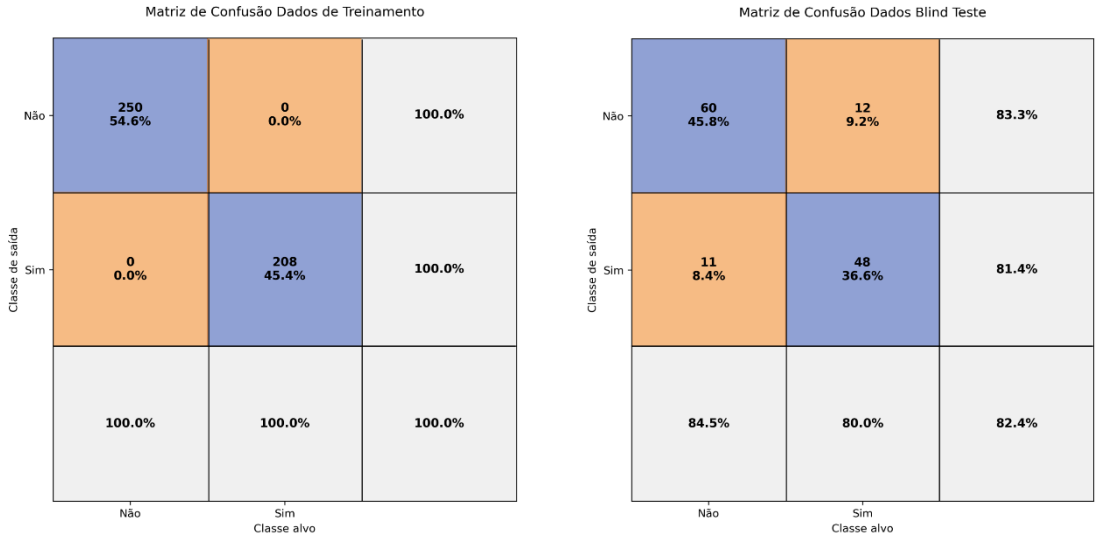

Figura 11: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Cirurgia”, utilizando dados balanceados.

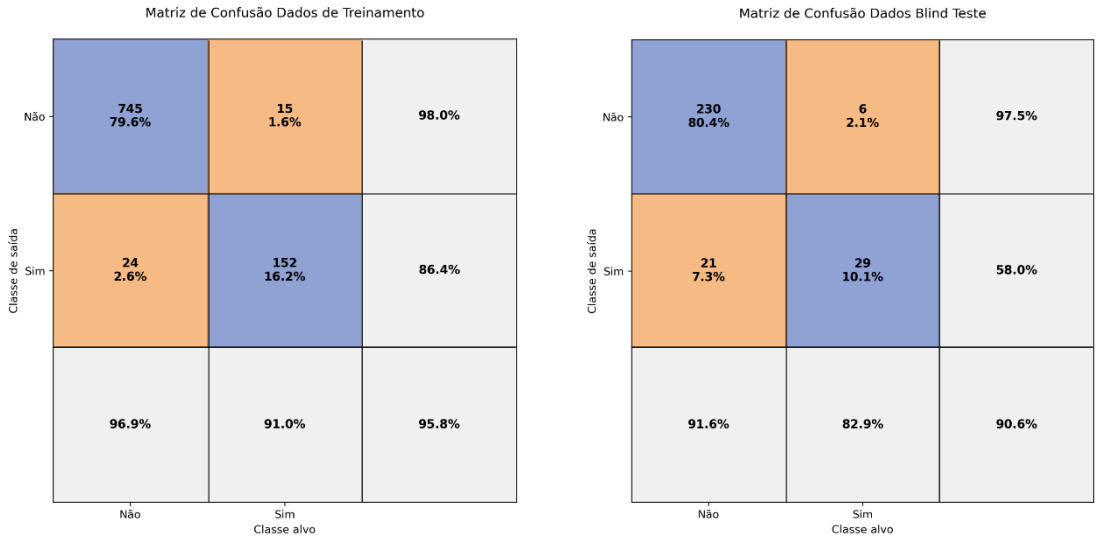

Figura 12: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Cirurgia” em DC.

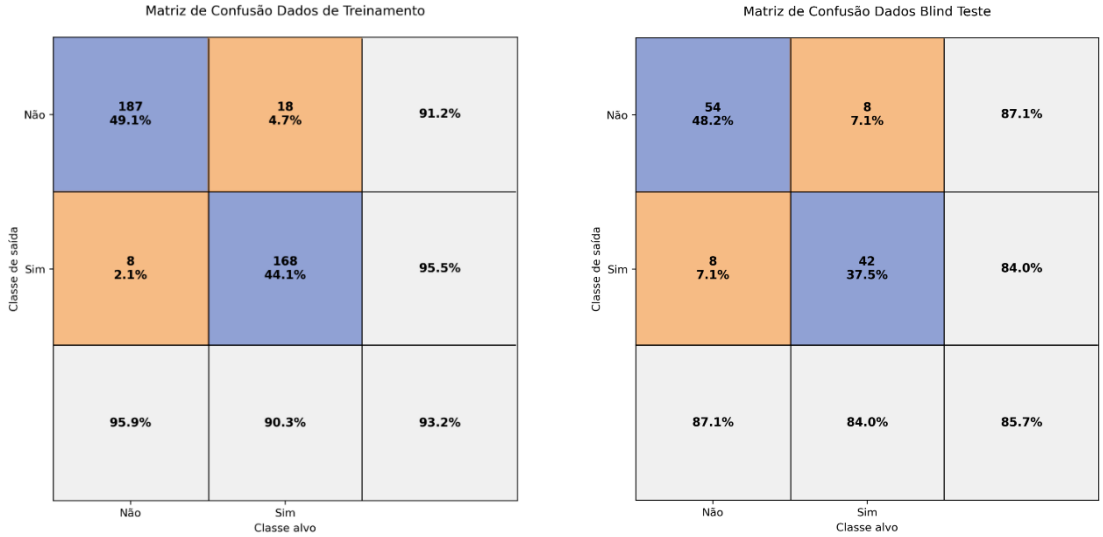

Figura 13: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Cirurgia” em DC, utilizando dados balanceados.

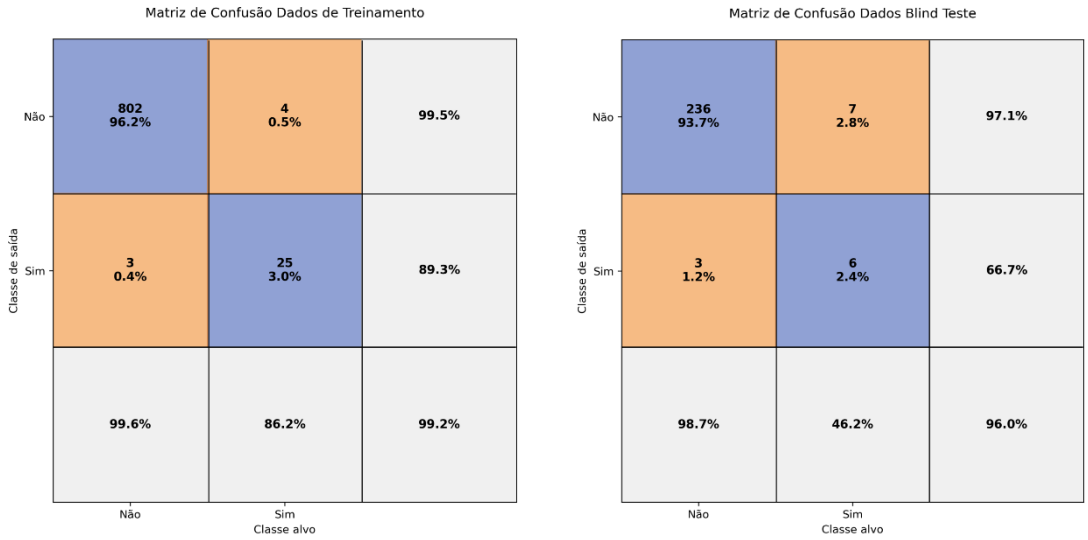

Figura 14: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Cirurgia” em RCU.

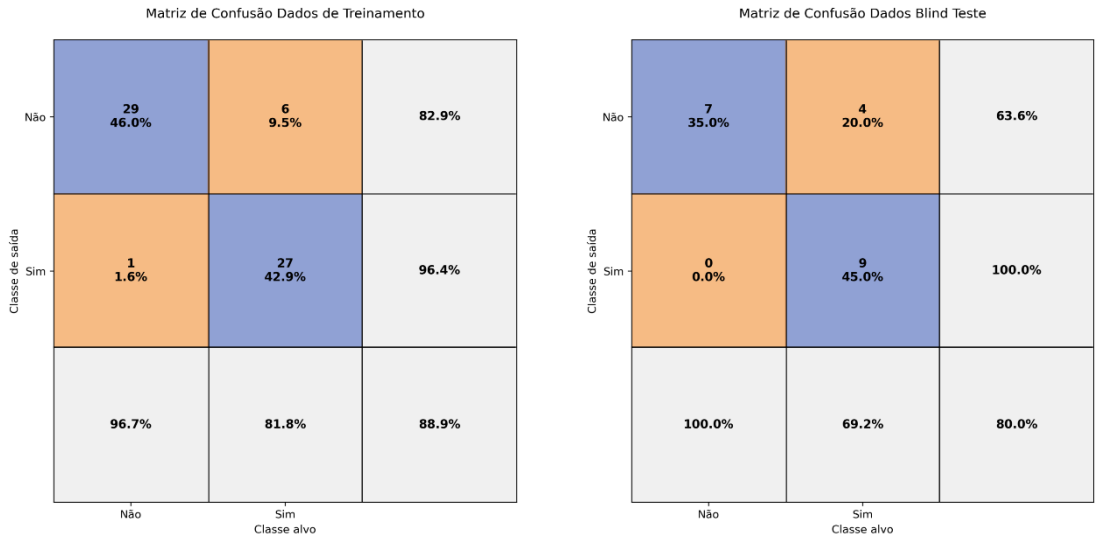

Figura 15: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Cirurgia” em RCU, utilizando dados balanceados.

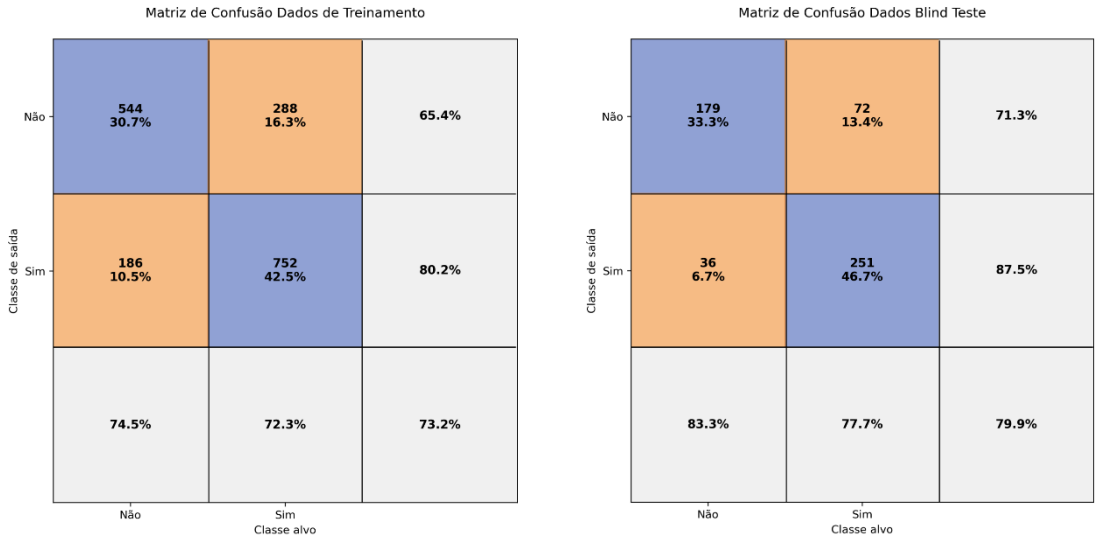

Figura 16: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Medicação”.

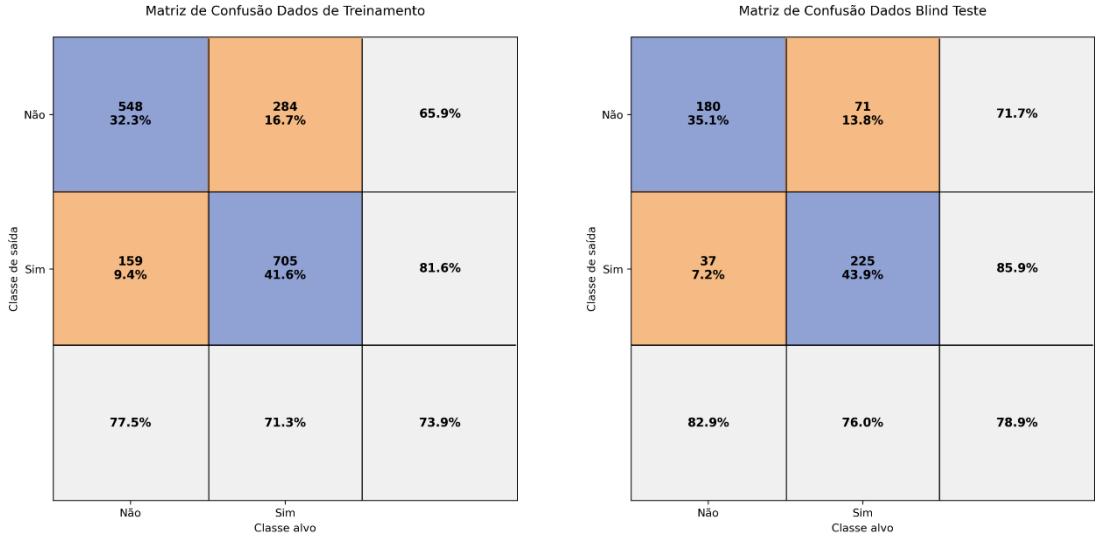

Figura 17: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Medicação”, utilizando dados balanceados.

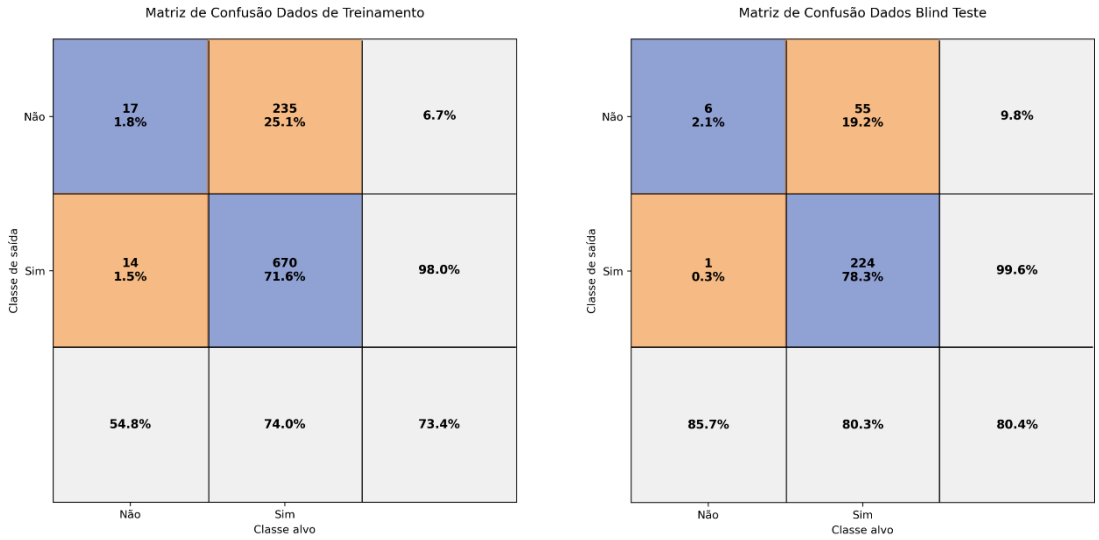

Figura 18: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Medicação” em DC.

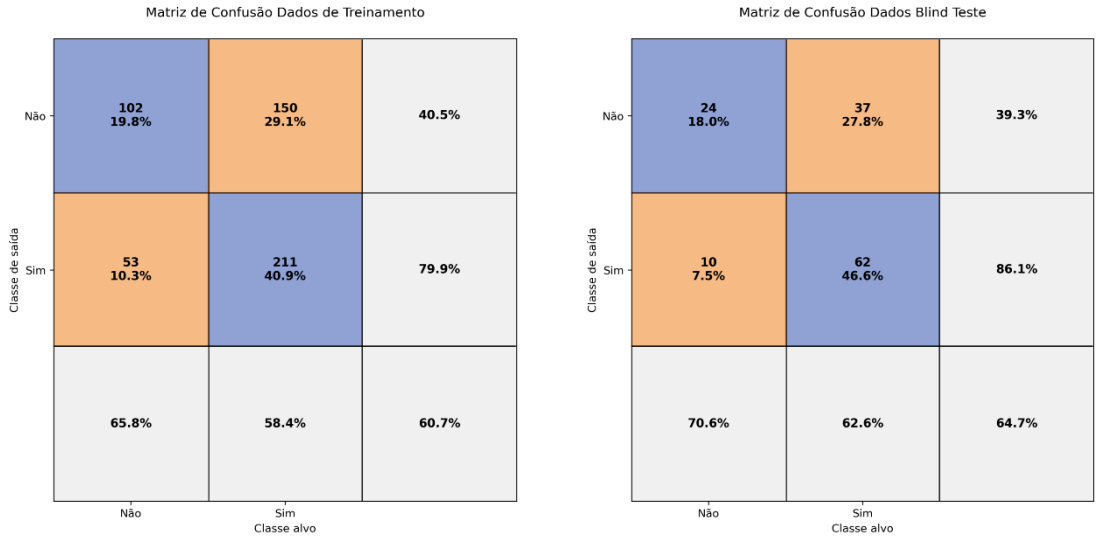

Figura 19: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Medicação” em DC, utilizando os dados balanceados.

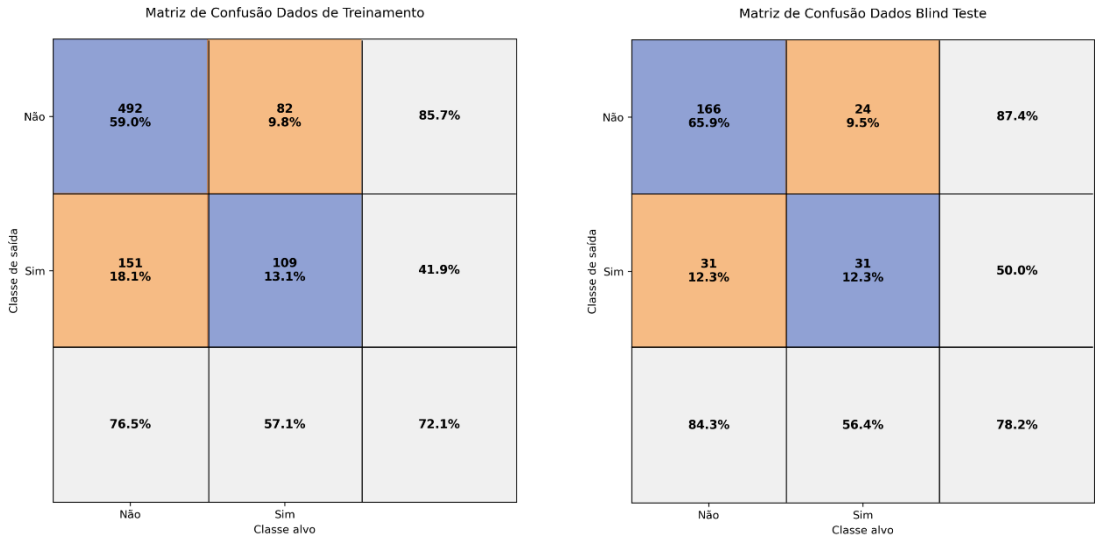

Figura 20: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Medicação” em RCU.

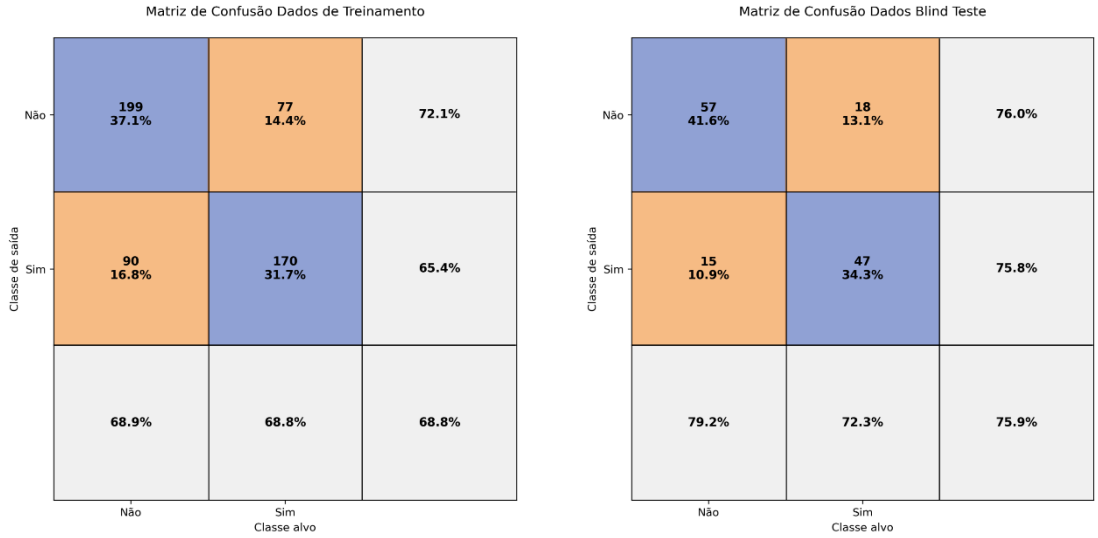

Figura 21: Matrizes de confusões dos dados de Treinamento e Blind Teste do melhor modelo de Inteligência Artificial para o desfecho “Medicação” em RCU, utilizando dados balanceados.

As Figuras 22 a 34 apresentam as curvas ROCs e os valores de AUC dos modelos avaliados para os diferentes desfechos do estudo. A curva ROC é amplamente utilizada em estudos clínicos e epidemiológicos que envolvem o desenvolvimento de modelos de IA, por permitir avaliar a capacidade discriminativa de modelos preditivos em diferentes limiares de decisão. A AUC fornece uma medida global da capacidade do modelo em distinguir entre classes, sendo útil em aplicações

na saúde para avaliar o desempenho de ferramentas de diagnóstico ou prognóstico. Dessa forma, a análise das curvas ROC permite estimar o potencial dos modelos em diferenciar corretamente pacientes com diferentes condições ou desfechos clínicos.

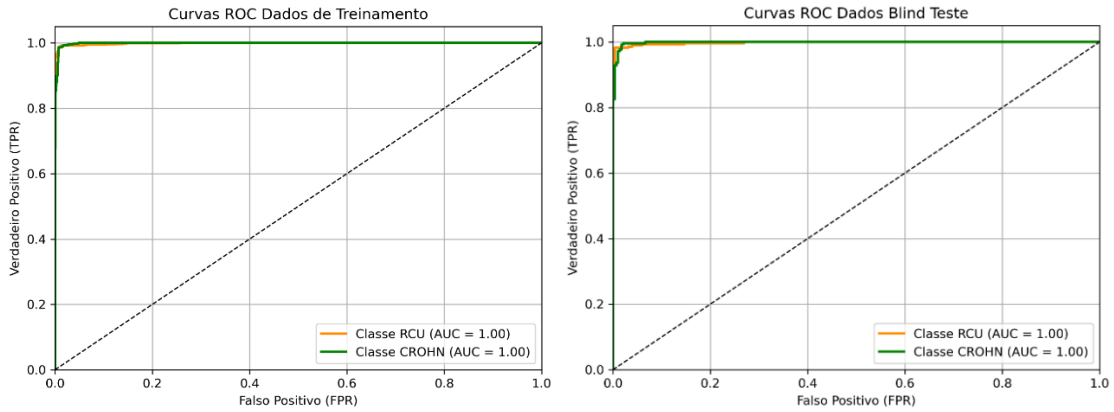

Figura 22: Curva ROC obtida para os dados de Treinamento e Blind teste para o melhor modelo de Inteligência Artificial para o desfecho “Tipo DII”.

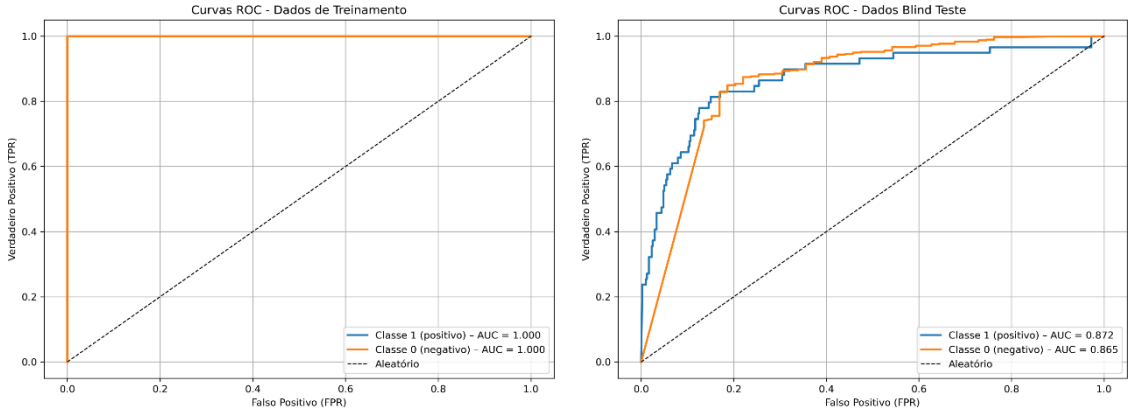

Figura 23: Curva ROC obtida para os dados de Treinamento e Blind teste para o melhor modelo de Inteligência Artificial para o desfecho “Cirurgia”.

Como exemplo da interpretação das curvas ROC apresentadas acima, observa-se que, na etapa de treinamento, foi obtida uma AUC de 1,0 para os dois desfechos, indicando capacidade discriminatória perfeita do modelo, ou seja, total habilidade em diferenciar corretamente as classes avaliadas nesse conjunto de dados. Nos dados de Blind Test para o desfecho cirurgia, a AUC foi de 0,872 para classe presença e de 0,865 para classe ausência, demonstrando que o modelo manteve a boa capacidade de discriminação mesmo em dados não vistos durante o treinamento. Esses valores indicam que a IA apresentou desempenho robusto na distinção entre

as classes, embora inferior ao observado no treinamento, o que é esperado em uma avaliação de generalização.

Nas demais figuras, a seguir, que representam a curva ROC e AUC, podem ser feitas as mesmas análises.

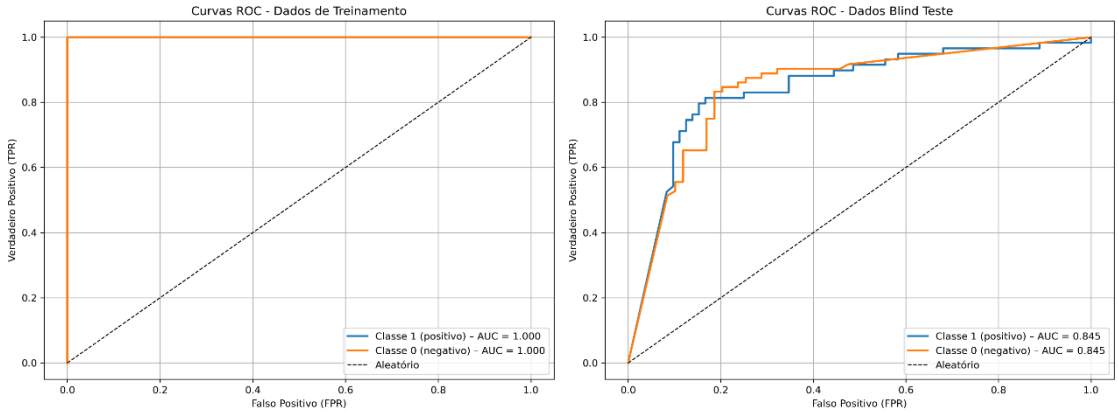

Figura 24: Curva ROC obtida para os dados de Treinamento e Blind teste para o melhor modelo de Inteligência Artificial para o desfecho “Cirurgia”, utilizando dados balanceados.

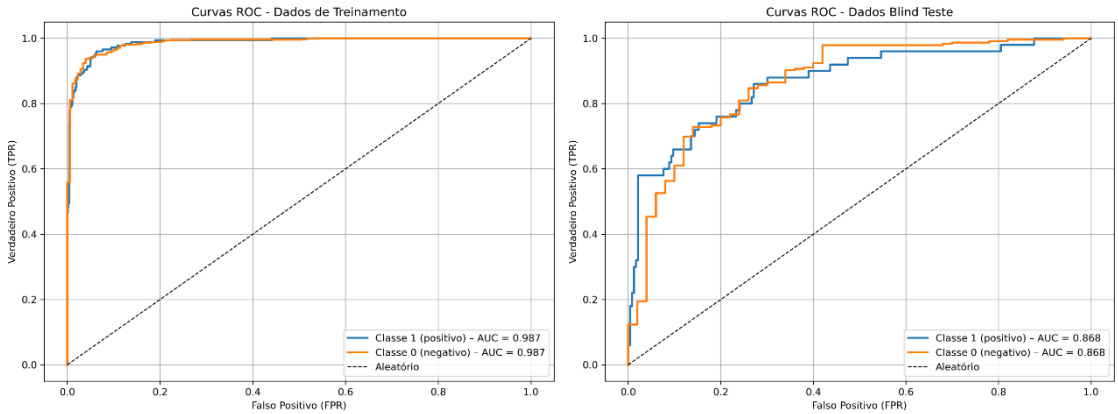

Figura 25: Curva ROC obtida para os dados de Treinamento e Blind teste para o melhor modelo de Inteligência Artificial para o desfecho “Cirurgia” em DC.

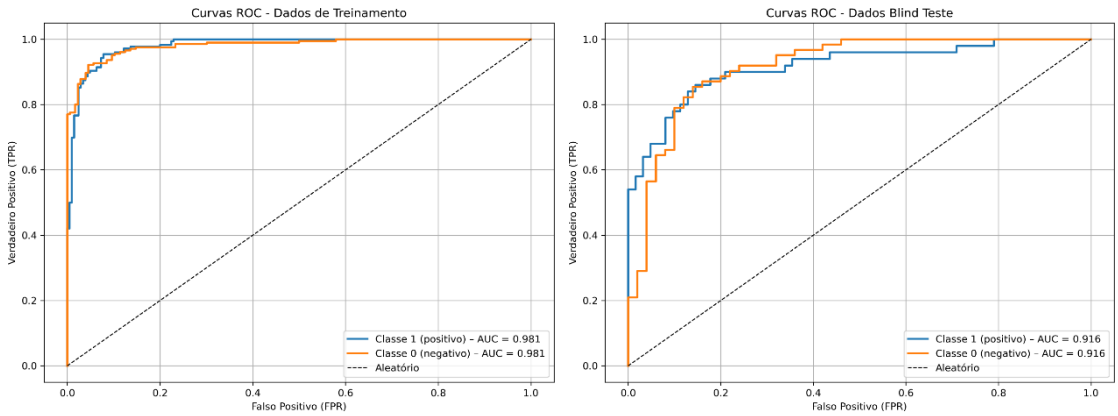

Figura 26: Curva ROC obtida para os dados de Treinamento e Blind teste para o melhor modelo de Inteligência Artificial para o desfecho “Cirurgia” em DC, utilizando dados balanceados.

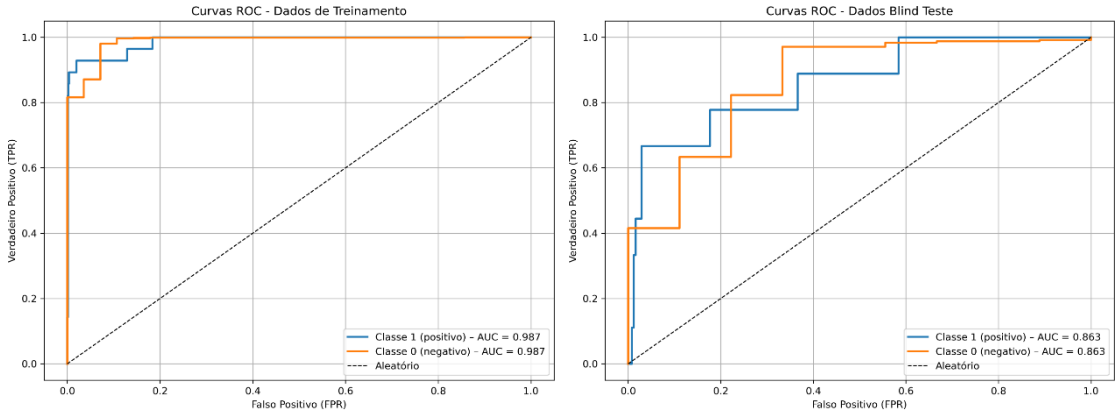

Figura 27: Curva ROC obtida para os dados de Treinamento e Blind teste para o melhor modelo de Inteligência Artificial para o desfecho “Cirurgia” em RCU.

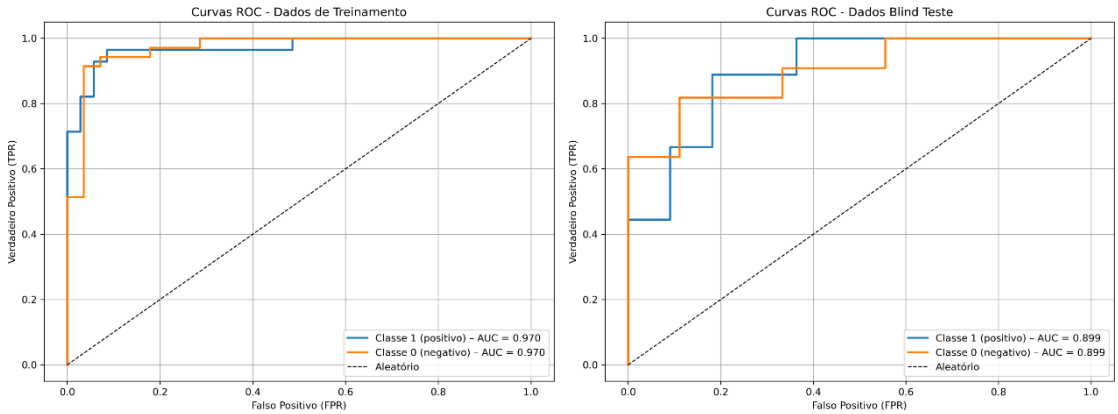

Figura 28: Curva ROC obtida para os dados de Treinamento e Blind teste para o melhor modelo de Inteligência Artificial para o desfecho “Cirurgia” em RCU, utilizando dados balanceados.

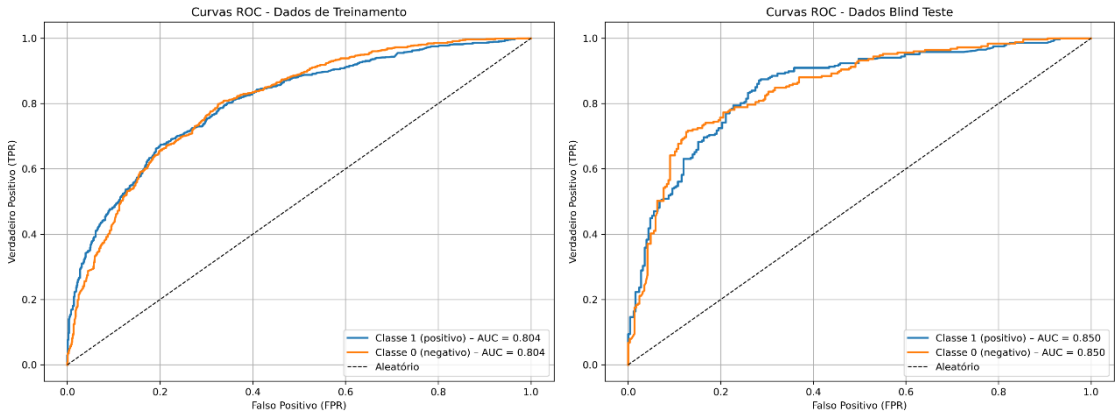

Figura 29: Curva ROC obtida para os dados de Treinamento e Blind teste para o melhor modelo de Inteligência Artificial para o desfecho “Medicação”.

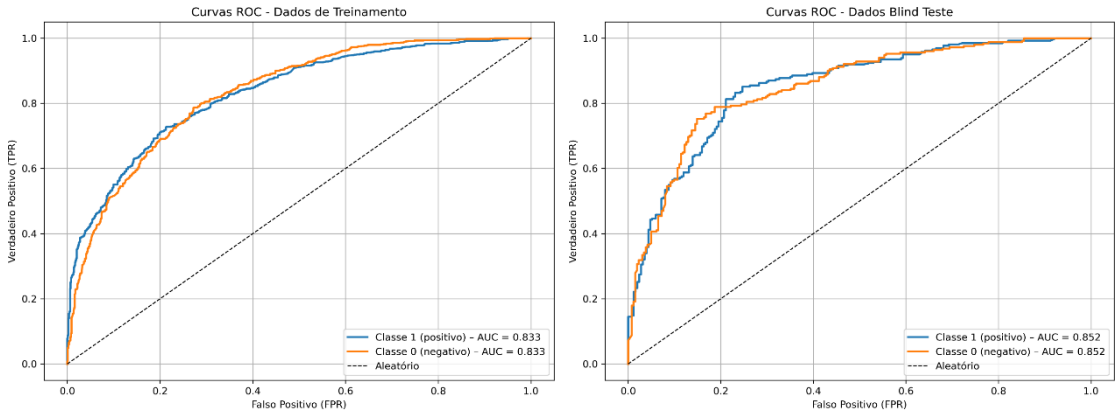

Figura 30: Curva ROC obtida para os dados de Treinamento do melhor modelo para o desfecho “Medicação”, utilizando dados balanceados.

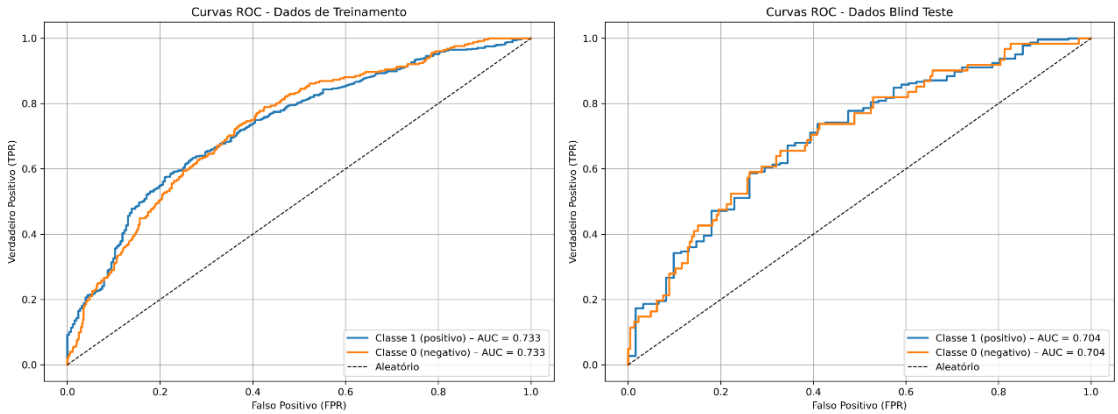

Figura 31: Curva ROC obtida para os dados de Treinamento do melhor modelo para o desfecho “Medicação” em DC.

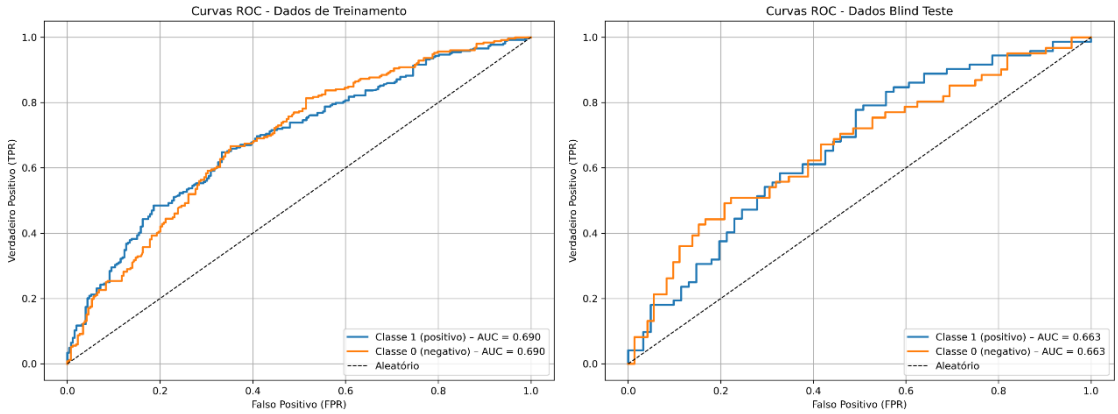

Figura 32: Curva ROC obtida para os dados de Treinamento do melhor modelo para o desfecho “Medicação” em DC, utilizando dados balanceados.

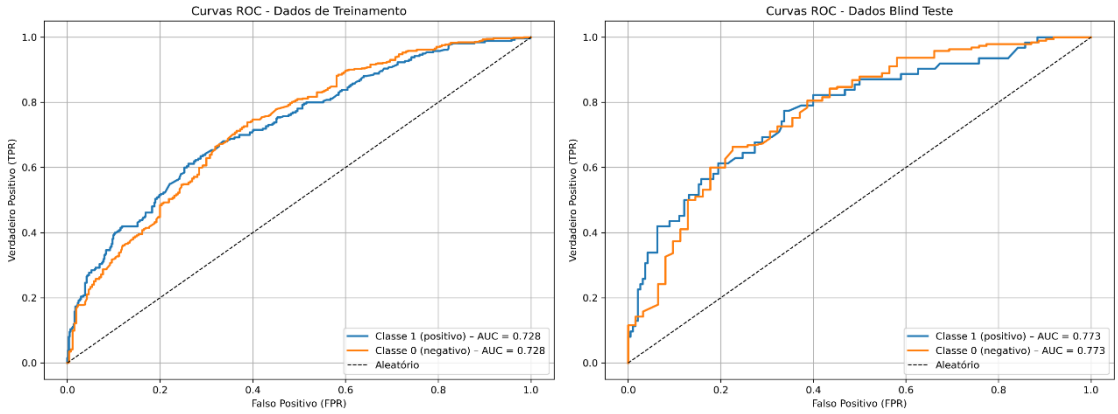

Figura 33: Curva ROC obtida para os dados de Treinamento do melhor modelo para o desfecho “Medicação” em RCU.

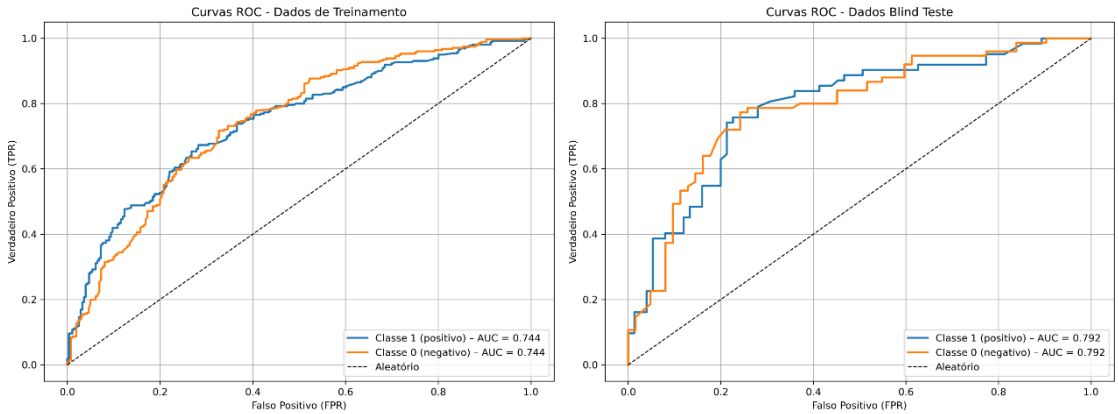

Figura 34: Curva ROC obtida para os dados de Treinamento do melhor modelo para o desfecho “Medicação” em RCU, utilizando dados balanceados.

Por fim, na Tabela 9, são apresentadas as métricas de desempenho de todos os modelos desenvolvidos para cada desfecho analisado. Essa tabela reúne os resultados obtidos nos diferentes cenários avaliados, incluindo a classificação entre DC e RCU, bem como os modelos voltados à predição dos desfechos cirúrgico e medicamentoso. A apresentação conjunta dessas métricas permite uma visão comparativa do desempenho dos modelos, considerando tanto a distribuição original dos dados quanto os cenários submetidos ao balanceamento, quando aplicável.

Tabela 9 : Desempenho do melhor modelo de Inteligência Artificial por desfecho em todos os cenários nas bases de treinamento e blind teste, estratificado por classe.

DESFECHO	BASE	CLASSE	PRECISION	RECALL	F1-SCORE
DII	Treinamento	DC	99%	99%	99%
		RCU	99%	99%	99%
	Blind test	DC	99%	98%	99%
		RCU	98%	99%	99%
Cirurgia	Treinamento	Ausência	100%	100%	100%
		Presença	100%	100%	100%
	Blind test	Ausência	95%	94%	95%
		Presença	56%	58%	57%
Cirurgia DC	Treinamento	Ausência	97%	98%	97%
		Presença	91%	86%	89%
	Blind test	Ausência	92%	97%	94%
		Presença	83%	58%	68%
Cirurgia RCU	Treinamento	Ausência	100%	100%	100%
		Presença	86%	89%	88%
	Blind test	Ausência	99%	97%	98%
		Presença	46%	67%	55%

Medicação	Treinamento	Ausência	75%	65%	70%
		Presença	72%	80%	76%
	Blind test	Ausência	83%	71%	77%
		Presença	78%	87%	82%
Medicação DC	Treinamento	Ausência	55%	7%	12%
		Presença	74%	98%	84%
	Blind test	Ausência	86%	10%	18%
		Presença	80%	100%	89%
Medicação RCU	Treinamento	Ausência	77%	86%	81%
		Presença	57%	42%	48%
	Blind test	Ausência	84%	87%	86%
		Presença	56%	50%	53%
Cirurgia bal	Treinamento	Ausência	100%	100%	100%
		Presença	100%	100%	100%
	Blind test	Ausência	85%	83%	84%
		Presença	80%	81%	81%
Cirurgia DC bal	Treinamento	Ausência	96%	91%	94%
		Presença	90%	95%	93%
	Blind test	Ausência	87%	87%	87%
		Presença	84%	84%	84%
Cirurgia RCU bal	Treinamento	Ausência	97%	83%	89%
		Presença	82%	96%	89%
	Blind test	Ausência	100%	64%	78%
		Presença	69%	100%	82%
Medicação bal	Treinamento	Ausência	78%	66%	71%
		Presença	71%	82%	76%
	Blind test	Ausência	83%	72%	77%
		Presença	76%	86%	81%
Medicação DC bal	Treinamento	Ausência	66%	40%	50%
		Presença	58%	80%	68%
	Blind test	Ausência	71%	39%	51%
		Presença	63%	86%	73%
Medicação RCU bal	Treinamento	Ausência	69%	72%	70%
		Presença	69%	65%	67%
	Blind test	Ausência	79%	76%	78%
		Presença	72%	76%	74%

As Figuras 35 a 42, reúnem as análises de interpretabilidade realizadas por meio da técnica SHAP para os modelos desenvolvidos nos diferentes desfechos do estudo. Por meio dessas figuras, é possível identificar quais variáveis apresentaram maior contribuição para as predições dos modelos, bem como compreender se sua presença aumentou ou reduziu a probabilidade de classificação em determinada classe. No contexto da saúde, essa análise é especialmente relevante, pois permite avaliar se os padrões aprendidos pelo modelo são coerentes com o conhecimento clínico das doenças analisadas. Dessa forma, a interpretação pelo SHAP contribui para aumentar a transparência dos modelos de IA auxiliando na compreensão dos fatores clínicos mais associados à discriminação entre DC e RCU e à predição dos desfechos cirúrgico e medicamentoso.

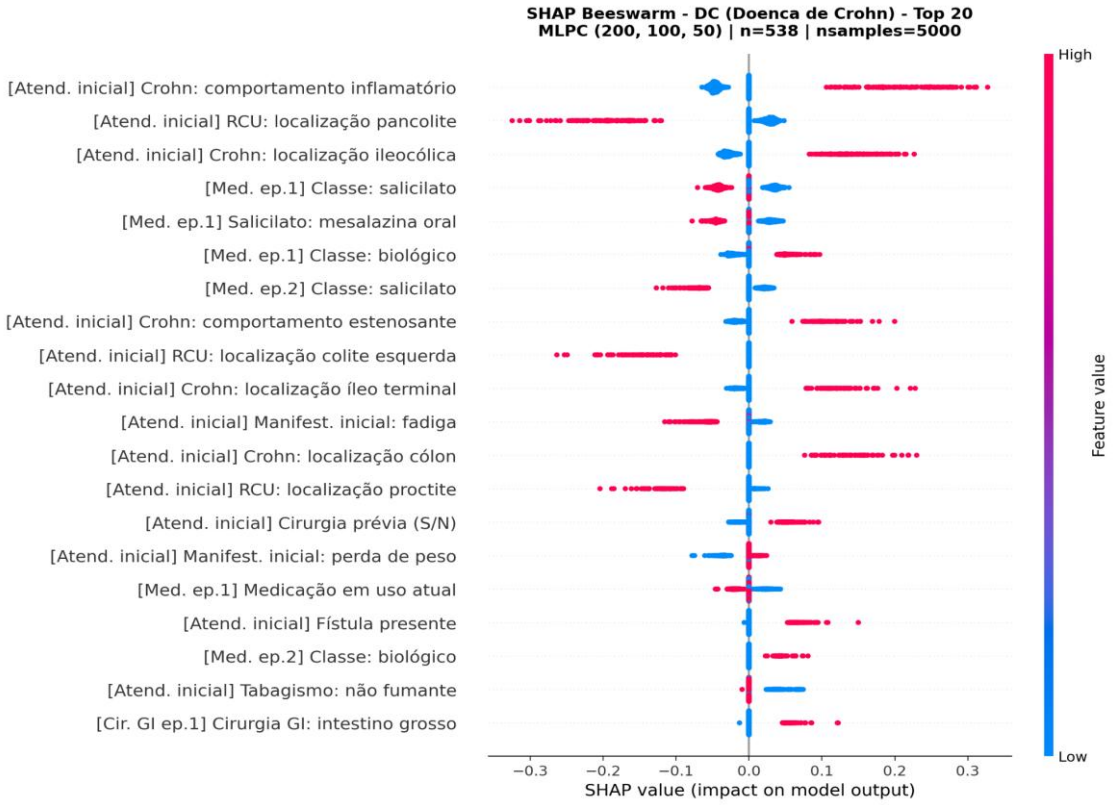

Figura 35: Interpretação baseada em SHAP das 20 principais variáveis no modelo de Inteligência Artificial para o desfecho “Tipo de DII” para DC.

Como exemplo da interpretação da análise SHAP apresentada anteriormente, observa-se que o gráfico permite avaliar não apenas quais variáveis foram mais importantes para o modelo, mas também a direção do efeito de cada uma sobre a predição. As variáveis posicionadas nas primeiras linhas exerceram maior influência

na saída do modelo, enquanto a distribuição dos pontos ao longo do eixo horizontal indica se sua presença aumentou ou reduziu a probabilidade de classificação como DC. Nesse sentido, pontos deslocados para a direita, com valores SHAP positivos, indicam contribuição favorável à classe DC, enquanto pontos à esquerda, com valores negativos, indicam menor probabilidade de classificação nessa classe. Além disso, a coloração dos pontos auxilia na interpretação da presença ou intensidade da variável, permitindo observar que determinadas características, quando presentes, direcionaram a predição para DC, enquanto outras deslocaram a classificação em sentido oposto. Dessa forma, a figura demonstra como o modelo ponderou diferentes informações clínicas para distinguir entre os tipos de DIIs, permitindo compreender o comportamento interno da predição e identificar se as variáveis utilizadas contribuíram para a classificação final.

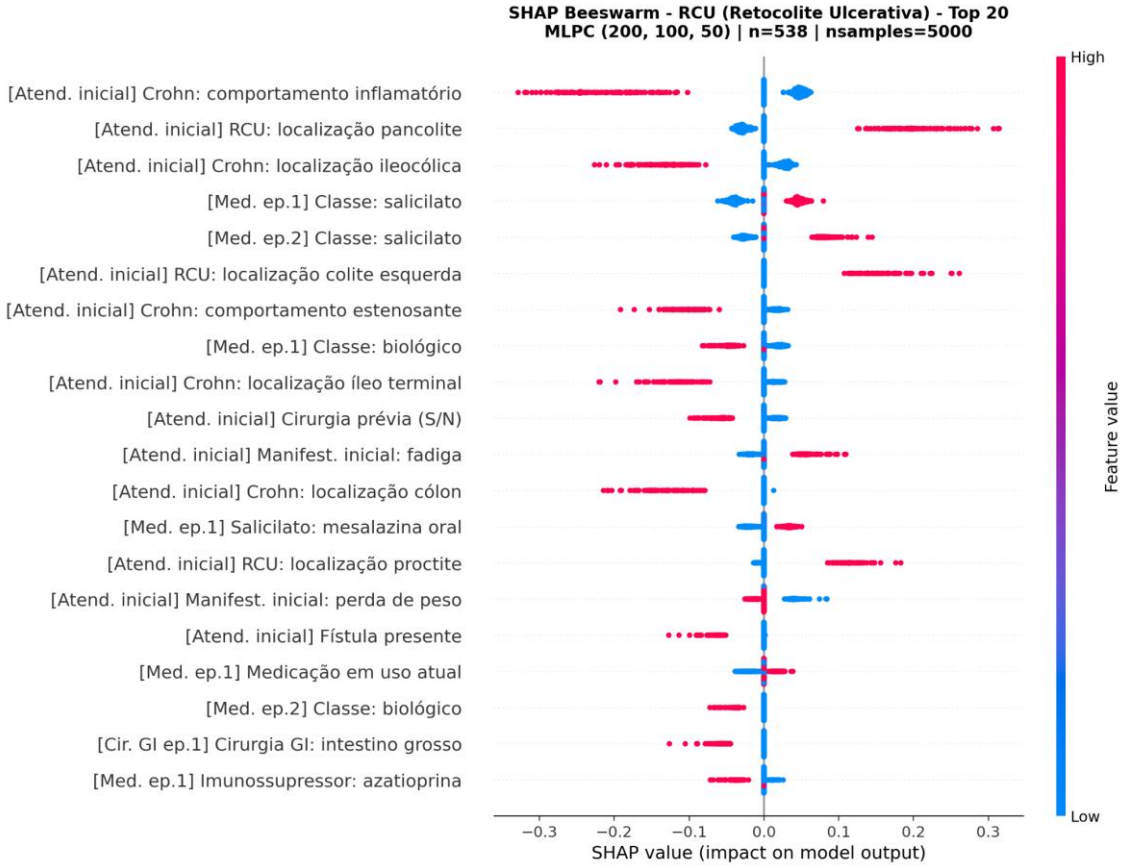

Figura 36: Interpretação baseada em SHAP das 20 principais variáveis no modelo de Inteligência Artificial para o desfecho “Tipo de DII” para RCU.

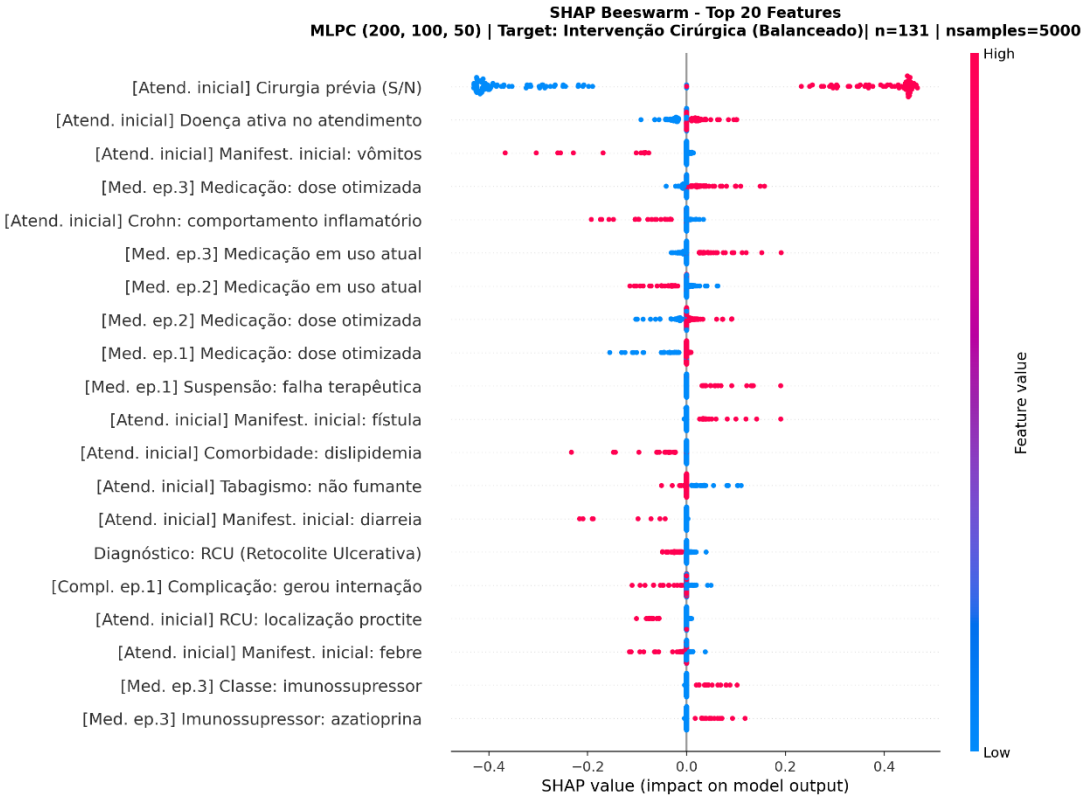

Figura 37: Interpretação baseada em SHAP das 20 principais variáveis no modelo de Inteligência Artificial para o desfecho “Cirurgia”, com dados balanceados.

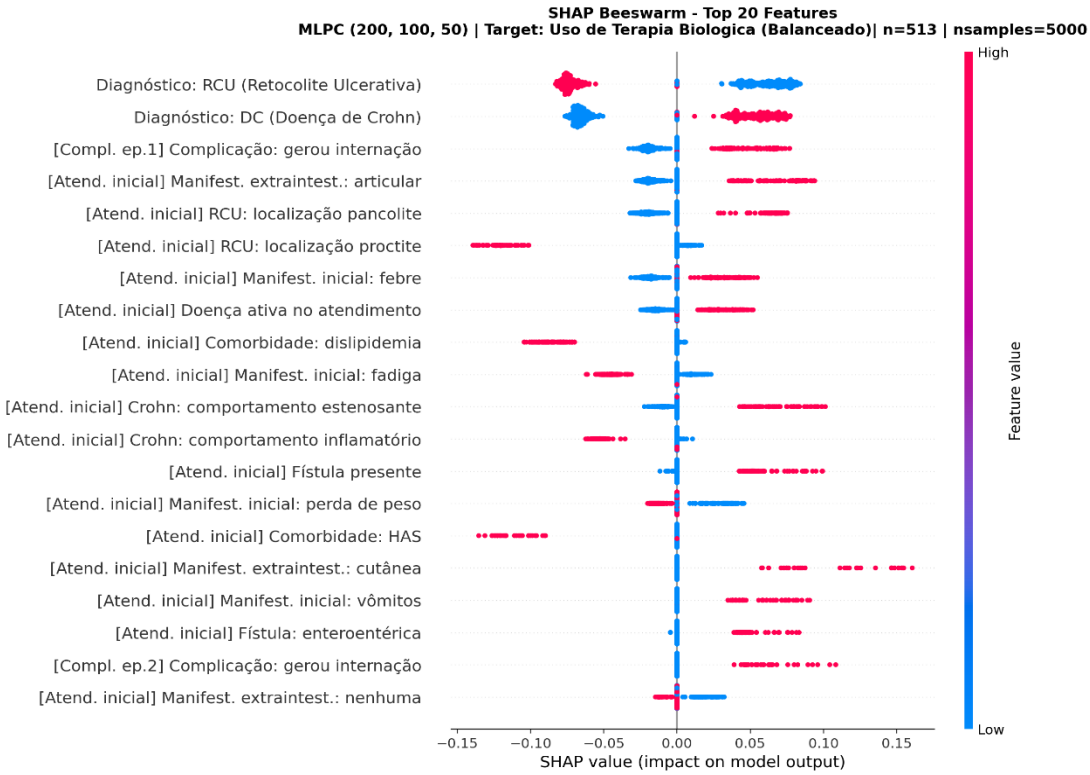

Figura 38: Interpretação baseada em SHAP das 20 principais variáveis no modelo de Inteligência Artificial para o desfecho “Medicação”, com dados balanceados.

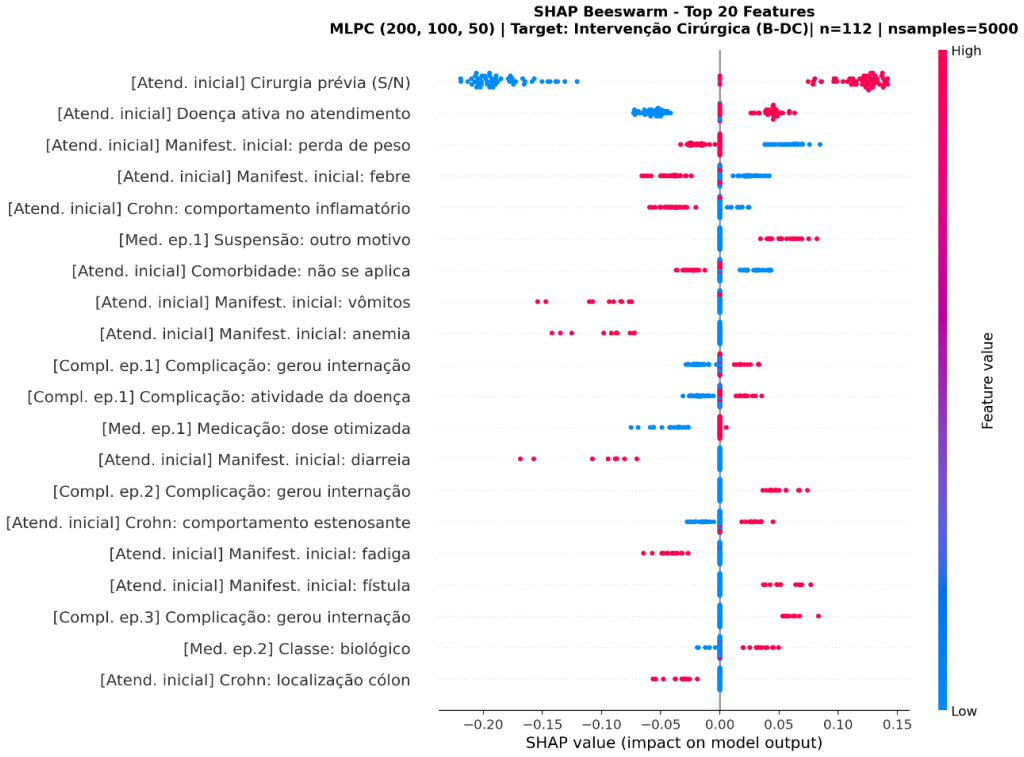

Figura 39: Interpretação baseada em SHAP das 20 principais variáveis no modelo de Inteligência Artificial para o desfecho “Cirurgia” no subgrupo DC, com dados balanceados

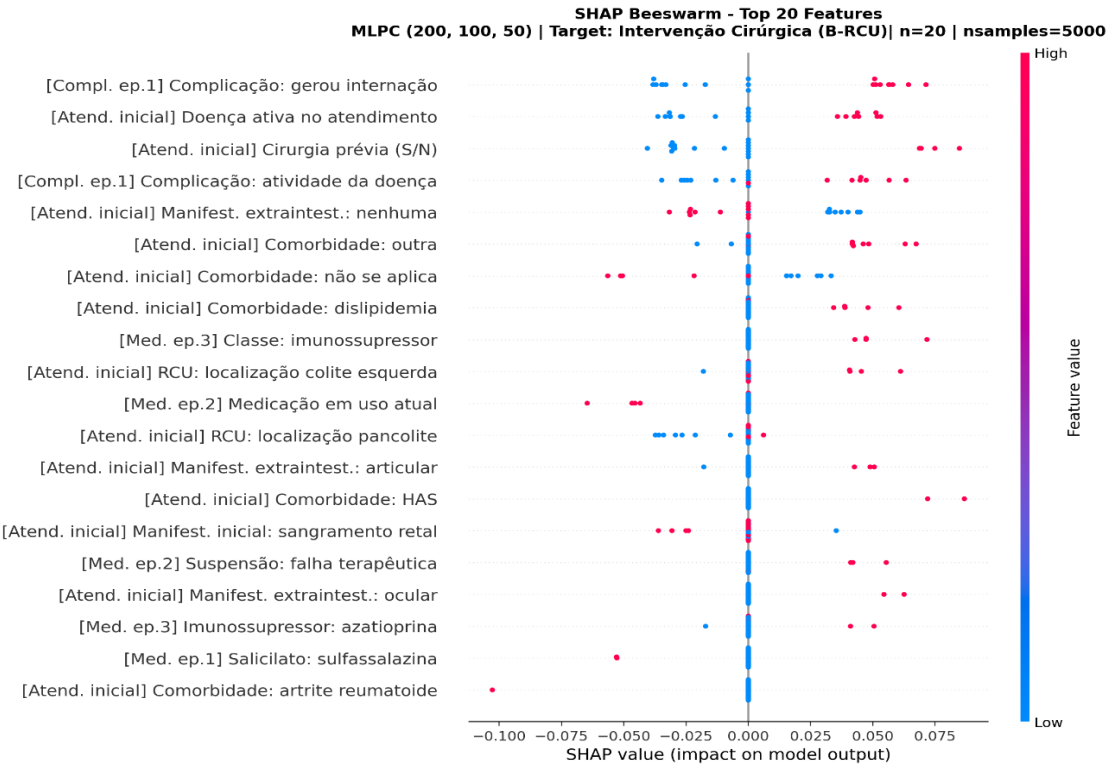

Figura 40: Interpretação baseada em SHAP das 20 principais variáveis no modelo de Inteligência Artificial para o desfecho “Cirurgia” no subgrupo RCU, com dados balanceados.

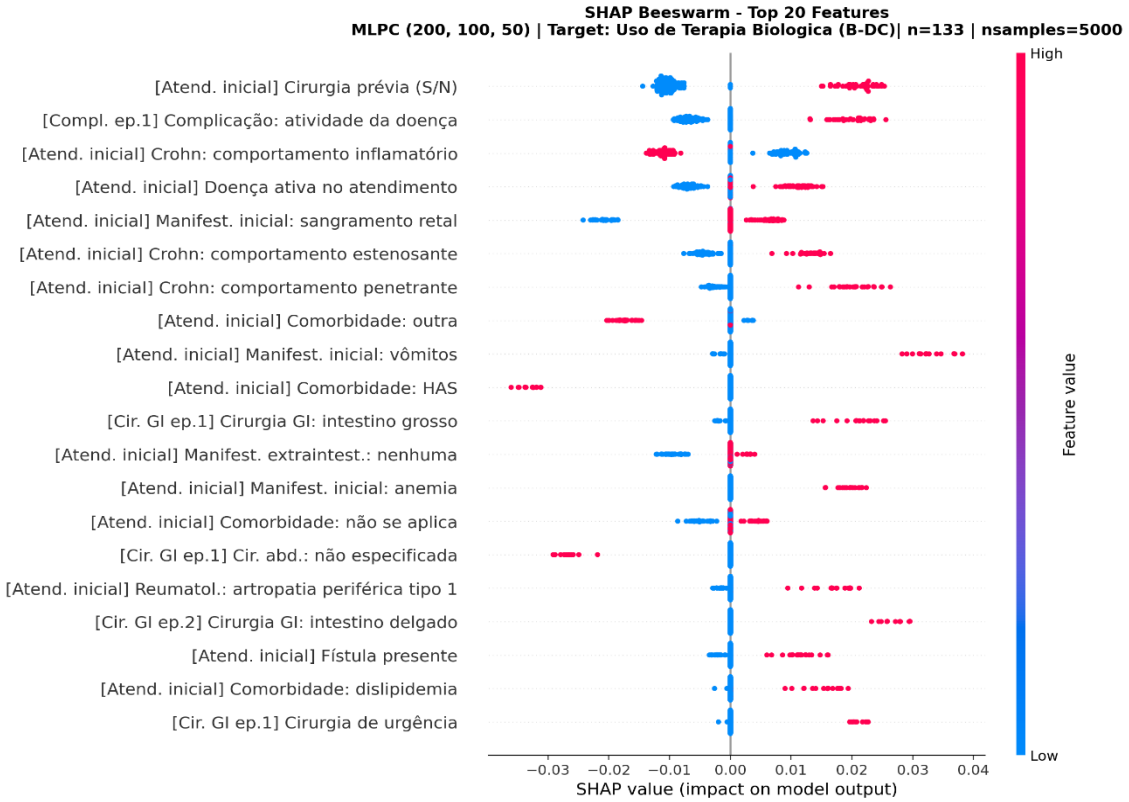

Figura 41: Interpretação baseada em SHAP das 20 principais variáveis no modelo de Inteligência Artificial para o desfecho “Medicação” no subgrupo DC, com dados balanceados.

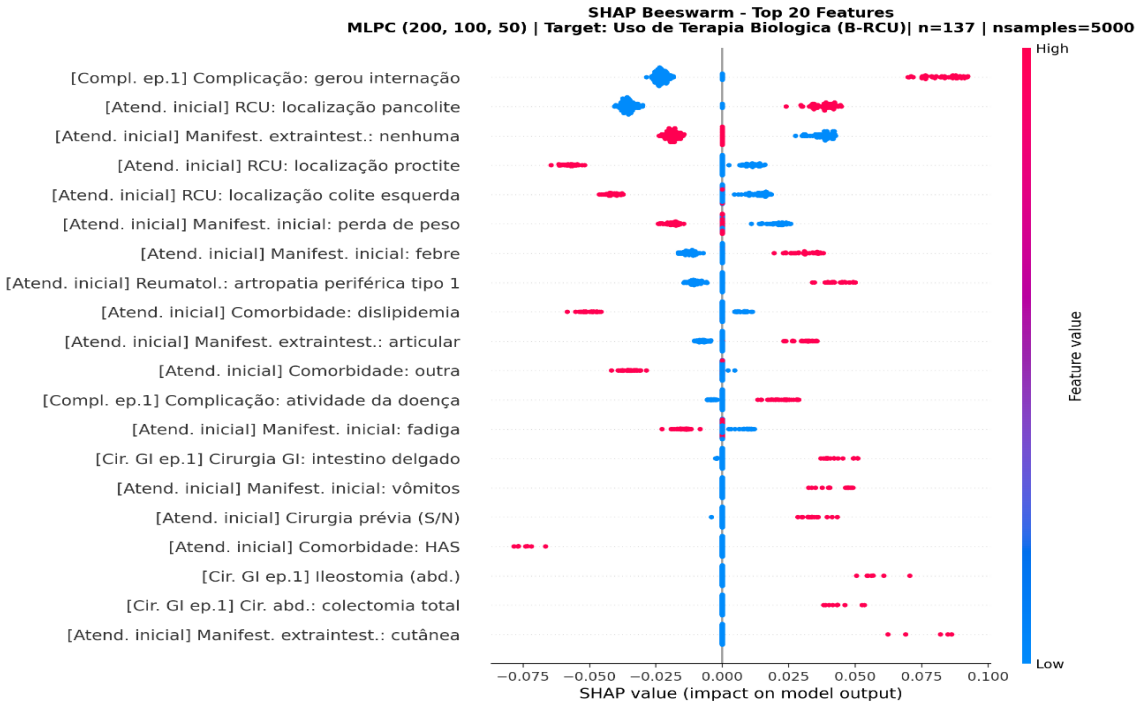

Figura 42: Interpretação baseada em SHAP das 20 principais variáveis no modelo de Inteligência Artificial para o desfecho “Medicação” no subgrupo RCU, com dados balanceados.

4. Discussão

De modo geral, os resultados evidenciaram que a predição do tipo de DII apresentou desempenho elevado, com métricas e AUC-ROC consistentemente altas nos conjuntos de treinamento e, sobretudo, de blind teste, sugerindo excelente capacidade discriminativa e bom potencial de generalização para esse desfecho. Para o tipo de DII, as AUCs obtidas, apresentadas na Figura 22 foram superiores a 0,97, com valores próximos de 1,0 no blind teste, indicando alta confiabilidade na distinção entre DC e RCU em novos casos clínicos. Esse desempenho também foi confirmado pelas matrizes de confusão (Figura 9), nas quais se observou predominância de verdadeiros positivos e verdadeiros negativos, com baixa ocorrência de falsos positivos e falsos negativos, além de métricas de avaliação próximas de 100% em ambos os conjuntos (Tabela 2). Esses achados são coerentes com a literatura, que aponta o avanço de métodos de IA no apoio ao diagnóstico e à estratificação em DIIs, seja com base em dados clínicos e laboratoriais, seja a partir de outras fontes, como endoscopia e biomarcadores, frequentemente com desempenho competitivo em relação a abordagens tradicionais, Gubatan *et al*, 2021.

Para os desfechos cirúrgico e medicamentoso, a análise demandou mais atenção devido ao desequilíbrio entre classes (ausência vs. presença do evento), condição comum em aplicações clínicas e que pode influenciar a leitura das métricas, especialmente quando a classe de interesse é menos frequente, Nguyen, 2021. Nesses casos, a predominância da classe ausência do desfecho pode mascarar o desempenho real do modelo quando se observam apenas medidas globais. Por essa razão, os modelos foram avaliados tanto com a distribuição original dos dados, que reflete a prevalência observada na prática clínica, quanto em cenários com balanceamento, visando melhorar a identificação da classe presença. Adicionalmente, foi aplicado o teste t de Student pareado para avaliar se as diferenças observadas nas métricas de desempenho entre os modelos treinados com dados originais e balanceados foram estatisticamente significativas, conforme descrito no Apêndice II. Para investigar essa questão de forma mais robusta, a avaliação foi conduzida em seis cenários distintos, contemplando o conjunto global de pacientes com DC e RCU, bem como os subgrupos separados por doença, todos analisados com e sem balanceamento. Em cada um desses cenários, os resultados de treinamento e blind teste foram comparados com base em AUC-ROC, precisão, *Recall*, *F1-Score* e

acurácia, além da análise das matrizes de confusão, permitindo quantificar o impacto do balanceamento sobre o desempenho global e principalmente, sobre a classificação da classe positiva.

No desfecho cirúrgico, em que a classe majoritária corresponde à ausência do evento, o balanceamento promoveu melhora importante na precisão e na sensibilidade para a classe presença, como mostrado nas Tabelas 3 a 5, indicando maior capacidade do modelo em detectar pacientes com ocorrência do desfecho. Além disso, os resultados do teste t de Student, apresentados na Tabela 1 do Apêndice II, demonstraram diferenças estatisticamente significativas entre as métricas obtidas com os dados originais e balanceados no blind test, em todos os cenários avaliados, reforçando que a melhora observada após o balanceamento não ocorreu apenas por variação aleatória entre as execuções. Esse comportamento é compatível com o que se espera em bases desbalanceadas: ao ajustar o modelo para reconhecer melhor a classe minoritária, tende-se a aumentar o *Recall* dessa classe, ainda que isso possa vir acompanhado de mudanças no equilíbrio entre os demais tipos de erro, Saito ***et al,*** 2015. Já no desfecho medicamentoso, o impacto do balanceamento variou conforme o cenário analisado. No conjunto global, a classe “presença” (Sim) é majoritária e as matrizes indicam predominância de acertos de verdadeiros positivos e verdadeiros negativos, com mudanças discretas após o balanceamento (Tabelas 6 a 8). Essa baixa variação também foi observada nos resultados do teste t de Student apresentados na Tabela 1 do Apêndice II, uma vez que, no cenário global, as diferenças entre as métricas obtidas com dados originais e balanceados no blind test não foram estatisticamente significativas. No subgrupo de pacientes com DC, o desequilíbrio favorece fortemente a classe “Sim” e, sem balanceamento, o modelo tende a prever “Sim” com frequência, resultando em muitos verdadeiros positivos, porém com baixa identificação da classe “Não” (verdadeiros negativos reduzido e falsos positivos elevados). Com o balanceamento, observa-se melhora na classificação da classe “Não” (aumento de verdadeiro negativo e redução de falso positivo), acompanhada de uma leve redução na detecção de “Sim” (Tabela 7). Entretanto, nesse cenário, a diferença observada para a precisão não foi estatisticamente significativa, indicando que a alteração nessa métrica deve ser interpretada com cautela (Tabela 1 Apêndice II). No subgrupo de pacientes com RCU, a classe majoritária é a “ausência” (Não). Nesse caso, sem balanceamento o modelo

acerta mais a classe “Não”, mas apresenta menor desempenho para “Sim”. Com o balanceamento, há melhora na identificação da presença (aumento de verdadeiros positivos e redução de falsos negativos), ainda que com algum aumento de falsos positivos (Tabela 8). Os resultados do teste t de Student no cenário RCU demonstraram diferenças estatisticamente significativas entre os modelos com dados originais e balanceados no blind test, reforçando que a melhora observada nesse subgrupo foi consistente entre as execuções (Tabela 1 Apêndice II). Esses achados reforçam que, em aplicações clínicas, a interpretação do desempenho do modelo não deve se restringir a métricas globais, mas considerar o impacto dos diferentes tipos de erro sobre a tomada de decisão.

Para a análise de interpretabilidade do modelo preditivo do desfecho tipo de DII para DC, a técnica SHAP evidenciou que as variáveis com maior contribuição para a classificação desse fenótipo de DII estiveram relacionadas principalmente ao comportamento clínico da doença e aos locais de acometimento intestinal (Figura 35). Entre os fatores mais associados à predição de DC, destacaram-se o comportamento inflamatório e o comportamento estenosante, compatíveis com padrões clínicos frequentemente observados nessa condição. Além disso, variáveis relacionadas à localização da doença, como acometimento ileocólico, íleo terminal e cólon, apresentaram importante influência sobre a saída do modelo. A presença de fístula também se destacou entre as variáveis associadas à predição de DC. De forma semelhante, a variável referente à cirurgia prévia apresentou contribuição positiva na classificação, o que pode ser explicado pelo fato de que parte dos pacientes com DC evolui com complicações que demandam intervenção cirúrgica, com possibilidade de múltiplos procedimentos ao longo do curso da doença. Entre as variáveis cirúrgicas, o envolvimento do intestino grosso também apareceu como fator relevante para o modelo. Em relação às variáveis terapêuticas, observou-se maior associação positiva com o uso de terapia biológica, o que é coerente com a prática clínica, uma vez que esses medicamentos são frequentemente empregados no manejo da DC, especialmente em casos moderados, graves ou complicados. Por outro lado, os derivados salicílicos, incluindo salicilato e mesalazina, apresentaram associação negativa com a predição de DC. Esse achado também apresenta plausibilidade clínica, considerando que tais medicamentos são menos utilizados no tratamento da DC quando comparados à RCU. Dessa forma, os resultados do SHAP demonstram

que o modelo atribuiu maior relevância a variáveis clinicamente compatíveis com o perfil da doença de Crohn, contribuindo para a interpretação dos padrões aprendidos pela RNA.

Para o modelo preditivo do desfecho tipo de DII para RCU, a técnica SHAP demonstrou que as variáveis com maior contribuição para a classificação desse fenótipo de DII estiveram relacionadas principalmente à extensão do acometimento colônico e ao perfil terapêutico dos pacientes (Figura 36). Entre as variáveis de localização, destacaram-se a pancolite, a colite esquerda e a proctite, achados compatíveis com a apresentação clínica da RCU, cuja inflamação se limita ao cólon e ao reto, com extensão variável ao longo do intestino grosso. Em relação às variáveis terapêuticas, a classe dos derivados salicílicos apresentou importante associação positiva com a predição de RCU. Entre as manifestações clínicas iniciais, a fadiga destacou-se como variável relevante para a predição de RCU. Dessa forma, os resultados obtidos pela análise SHAP indicam que o modelo atribuiu maior importância a variáveis compatíveis com o perfil clínico e terapêutico da RCU, especialmente aquelas relacionadas à extensão da doença no cólon e ao uso de derivados salicílicos.

Na análise de interpretabilidade do modelo preditivo para o desfecho cirúrgico, utilizando dados balanceados, a técnica SHAP evidenciou que as variáveis com maior contribuição para a classificação estiveram associadas a marcadores de maior gravidade e complexidade clínica da doença (Figura 37). Entre os fatores mais relevantes, destacaram-se a presença de cirurgia prévia, doença em atividade no início do acompanhamento, necessidade de otimização terapêutica, falha terapêutica, manifestação inicial com fístula e uso de imunossupressores. A variável referente à cirurgia prévia apresentou expressiva influência sobre a predição do desfecho cirúrgico, o que pode ser explicado pelo fato de que pacientes com histórico de intervenção cirúrgica frequentemente apresentam doença mais grave, recorrente ou complicada. De forma semelhante, a presença de doença em atividade no início do acompanhamento e a necessidade de otimização da medicação indicam maior dificuldade no controle clínico da DII, podendo refletir quadros com maior risco de evolução desfavorável. Além disso, a falha terapêutica, a presença de fístula e o uso de imunossupressores também se destacaram como variáveis relevantes para o modelo. Esses achados apresentam coerência clínica, uma vez que tais fatores geralmente estão associados a formas mais graves da doença, maior carga

inflamatória, complicações estruturais e necessidade de estratégias terapêuticas mais intensivas.

A análise de interpretabilidade do modelo preditivo para o desfecho medicamentoso, definido como uso de terapia biológica, utilizando dados balanceados, a técnica SHAP evidenciou que as variáveis com maior contribuição para a classificação estiveram relacionadas principalmente a marcadores de maior gravidade, atividade inflamatória e presença de complicações das DIIs (Figura38). Entre os fatores mais relevantes, destacaram-se o diagnóstico de DC, a ocorrência de complicações com necessidade de internação, manifestações extraintestinais articulares e cutâneas, pancolite na RCU, febre como manifestação inicial, doença em atividade no atendimento inicial, comportamento estenosante na DC, presença de fístula e necessidade de internação. A associação positiva entre DC e uso de terapia biológica apresenta coerência clínica, uma vez que essa condição frequentemente pode cursar com fenótipos mais complexos, incluindo comportamento estenosante, fístulas e complicações que demandam tratamento mais intensivo. De forma semelhante, a presença de complicações com internação e a doença em atividade no atendimento inicial indicam maior gravidade clínica e possível dificuldade de controle com terapias convencionais, favorecendo a indicação de terapias avançadas. Na RCU, a pancolite destacou-se como variável relevante para a predição do uso de terapia biológica, o que pode ser explicado pela maior extensão do acometimento colônico e pelo potencial aumento da carga inflamatória nesses pacientes. Além disso, manifestações extraintestinais, como acometimento articular e cutâneo, bem como febre entre os sintomas iniciais, também contribuíram para a classificação, sugerindo que o modelo atribuiu maior importância a variáveis associadas a doença sistemicamente mais ativa ou clinicamente mais complexa. Dessa forma, os resultados obtidos pelo SHAP indicam que o modelo identificou padrões compatíveis com maior necessidade de uso de terapia biológica, atribuindo maior relevância a variáveis relacionadas à gravidade, extensão da doença, atividade inflamatória, complicações e manifestações extraintestinais.

Na análise SHAP do modelo preditivo para o desfecho cirúrgico em pacientes com DC, utilizando dados balanceados, evidenciou que as variáveis com maior contribuição para a classificação estiveram associadas a marcadores de maior gravidade, atividade inflamatória e complexidade clínica da doença. Entre os fatores mais relevantes, destacaram-se a presença de cirurgia prévia, doença ativa no

atendimento inicial, suspensão de terapia medicamentosa, presença de atividade da doença, necessidade de internação por atividade inflamatória ou complicações, comportamento estenosante, presença de fístula e uso de terapia biológica (Figura 39). Os resultados do SHAP indicam que o modelo atribuiu maior relevância a variáveis compatíveis com formas mais graves e complicadas da DC, especialmente aquelas relacionadas à recorrência cirúrgica, atividade inflamatória persistente, internações, complicações estruturais e necessidade de tratamento avançado.

Na análise de interpretabilidade do modelo preditivo para o desfecho cirúrgico em pacientes com RCU, utilizando dados balanceados, a técnica SHAP evidenciou que as variáveis com maior contribuição para a classificação estiveram associadas a marcadores de maior gravidade, atividade inflamatória e dificuldade de controle clínico da doença (Figura 40). Entre os fatores mais relevantes, destacaram-se a presença de complicação com necessidade de internação, doença ativa no atendimento inicial, cirurgia prévia, atividade da doença como complicação, manifestação extraintestinal articular, uso de imunossupressores, localização em colite esquerda e falha terapêutica. Os resultados indicam que o modelo atribuiu maior relevância a variáveis compatíveis com formas mais graves ou de difícil controle da RCU, especialmente aquelas relacionadas à atividade inflamatória persistente, necessidade de internação, complicações clínicas, manifestações extraintestinais, falha terapêutica e uso de tratamento imunossupressor.

Para o modelo preditivo para o desfecho medicamentoso em pacientes com DC, definido como uso de terapia biológica, utilizando dados balanceados, a técnica SHAP evidenciou que as variáveis com maior contribuição para a classificação estiveram associadas a marcadores de maior gravidade, atividade inflamatória e comportamento complicado da doença (Figura 41). Entre os fatores mais relevantes, destacaram-se a necessidade de cirurgia prévia, presença de atividade da doença, doença ativa no atendimento inicial, sangramento retal, comportamento estenosante, comportamento penetrante e presença de fístula. Os resultados indicam que o modelo atribuiu maior relevância a variáveis compatíveis com formas mais graves e complexas da DC, especialmente aquelas relacionadas à atividade inflamatória persistente, complicações estruturais, recorrência cirúrgica e necessidade de terapias avançadas, como os imunobiológicos.

Por fim, na análise de interpretabilidade do modelo preditivo para o desfecho medicamentoso em pacientes com RCU, definido como uso de terapia biológica, utilizando dados balanceados, a técnica SHAP evidenciou que as variáveis com maior contribuição para a classificação estiveram associadas a marcadores de maior extensão da doença, atividade inflamatória, presença de complicações e manifestações sistêmicas (Figura 42). Entre os fatores mais relevantes, destacaram-se a presença de complicação com necessidade de internação, pancolite, febre como manifestação inicial, manifestação extraintestinal articular e presença de atividade da doença. Os resultados do SHAP sugerem que o modelo identificou padrões clinicamente coerentes com a necessidade de escalonamento terapêutico para imunobiológicos em pacientes com RCU.

Um estudo com proposta semelhante ao uso de dados rotineiros clínicos em DIIs é o de Kraszewski *et al*, 2021, que desenvolveu modelos de ML para apoiar a identificação de DC e RCU a partir de marcadores laboratoriais de rotina (sangue, urina e fezes) e características demográficas, utilizando 702 registros e comparando o desempenho do modelo com uma abordagem simples baseada apenas em proteína C-Reativa (PCR-CRP). Os autores relataram melhor desempenho para os modelos baseados em ML em relação ao marcador isolado, sugerindo que combinações de variáveis clínicas/laboratoriais rotineiras podem apoiar a tomada de decisão em DIIs. Em comparação, o presente estudo também parte de informações clínicas/epidemiológicas estruturadas, porém amplia o foco ao contemplar, além da distinção DC e RCU, a predição de desfechos medicamentosos e cirúrgicos, avaliados em múltiplos cenários (DC+RCU, apenas DC, apenas RCU; com e sem balanceamento) e com validação em blind teste, o que permite discutir não só a discriminação entre diagnósticos, mas também a utilidade do modelo para desfechos clínicos relevantes no acompanhamento.

Dessa forma, o presente estudo propõe o uso de técnicas de IA, aplicadas a dados clínicos e epidemiológicos de pacientes com DIIs (DC e RCU), com o objetivo de desenvolver modelos preditivos capazes de estimar o prognóstico e apoiar a estratificação de risco e auxiliar na tomada de decisão clínica. Que seja de nosso conhecimento, não foram identificados estudos que integrem, de forma conjunta, dados clínicos e epidemiológicos de pacientes com DIIs e as técnicas de Inteligência Artificial adotadas neste trabalho em um software voltado à predição de desfechos,

conforme a abordagem aqui proposta. Dessa forma, os modelos obtidos neste estudo apresentam potencial para aplicação na rotina clínica, com o objetivo de oferecer auxílio objetivo à tomada de decisão, auxiliando profissionais de saúde na classificação de casos e na estimativa de risco de desfechos em pacientes com DIIs.

5. Conclusões

A aplicação de IA a partir de dados clínicos e epidemiológicos de pacientes com doenças inflamatórias intestinais (DIIs) demonstrou potencial relevante para apoiar a prática clínica. Os modelos baseados em RNAs MLP apresentaram desempenho elevado para distinguir DC e RCU, sugerindo boa capacidade discriminativa e possibilidade de generalização. Para os desfechos cirúrgico e medicamentoso, a análise evidenciou que o desequilíbrio entre classes influencia diretamente a interpretação das métricas e a capacidade do modelo em identificar a presença do evento, reforçando a importância de avaliar os resultados por classe e de comparar cenários com distribuição original e com balanceamento.

Apesar dos resultados promissores, a implementação da IA na rotina clínica requer cautela e etapas adicionais. É fundamental que as abordagens sejam validadas em estudos prospectivos e em novos conjuntos de dados independentes, com definições consistentes e clinicamente relevantes dos desfechos, a fim de confirmar a robustez e a aplicabilidade dos modelos em diferentes cenários clínicos. Além disso, a incorporação de ferramentas de IA em DIIs depende da padronização do processamento e análise dos dados, do monitoramento de desempenho ao longo do tempo e de uma colaboração interdisciplinar entre profissionais de saúde e cientistas da computação, garantindo que os modelos sejam interpretáveis e úteis na tomada de decisão.

A utilização do SHAP contribuiu para aumentar a transparência dos modelos, permitindo não apenas avaliar seu desempenho preditivo, mas também compreender quais informações clínicas influenciaram suas classificações. Essa técnica é fundamental para a possível aplicação da IA na prática clínica, uma vez que modelos interpretáveis tendem a favorecer maior confiança por parte dos profissionais de saúde e possibilitam que os resultados sejam analisados em conjunto com o raciocínio clínico. Portanto, a interpretabilidade representa uma etapa importante para que ferramentas baseadas em IA possam futuramente atuar como apoio à tomada de

decisão em DIIs, auxiliando na estratificação de risco, na identificação de pacientes com maior probabilidade de evolução desfavorável e no direcionamento de estratégias terapêuticas individualizadas.

Dessa forma, este estudo contribui ao demonstrar a viabilidade da aplicação da inteligência artificial a dados clínicos e epidemiológicos estruturados para a construção de modelos preditivos em DIIs, bem como ao caracterizar, de forma sistemática, como diferentes cenários influenciam o desempenho desses modelos.

Como perspectivas futuras, destaca-se o desenvolvimento de uma interface gráfica voltada à utilização prática dos modelos, conforme ilustrado na Figura 43. A proposta de interface permitiria a inserção padronizada de dados clínicos, a seleção do modelo preditivo a ser aplicado e a visualização dos resultados de predição, incluindo estimativas de risco e opções de detalhamento individual dos pacientes. Dessa forma, a ferramenta poderia tornar o uso dos modelos mais acessível aos profissionais de saúde e favorecer sua futura integração ao contexto assistencial.

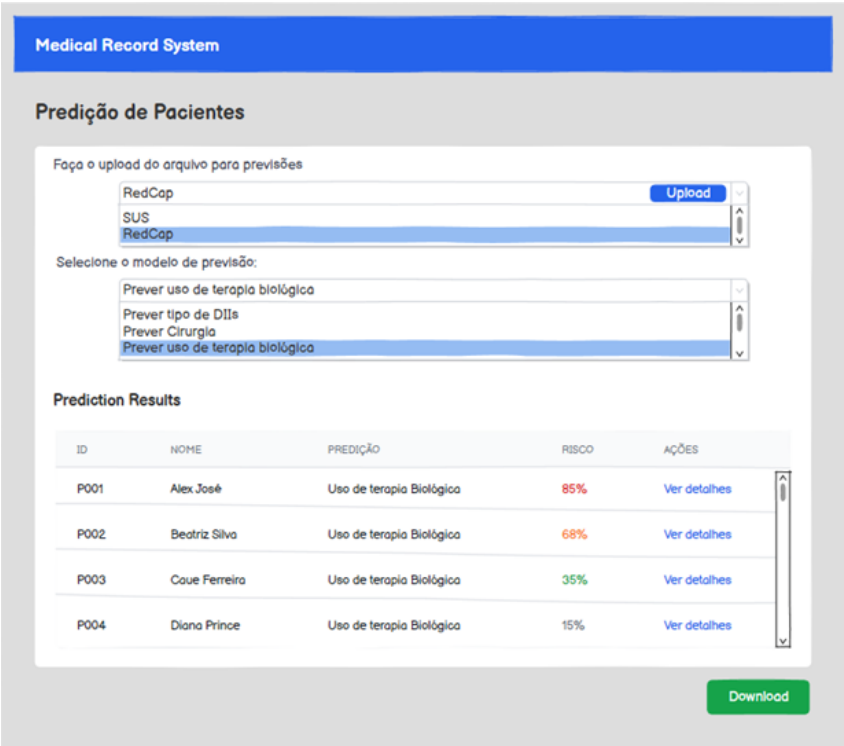

Figura 43: Protótipo de interface gráfica para aplicação dos modelos preditivos em pacientes com doenças inflamatórias intestinais.

Além disso, a avaliação de outros modelos de ML, como árvores de decisão, *Random Forest*, *Gradient Boosting*, *Support Vector Machine*, deve ser considerada. Essa ampliação metodológica permitirá comparar diferentes modelos quanto ao

desempenho preditivo, à interpretabilidade e à robustez em relação às RNAs MLP utilizadas neste estudo.

Estudos futuros devem priorizar a validação em novas bases de dados, a avaliação do impacto da aplicação da IA na prática clínica e o aprimoramento metodológico dessas ferramentas, a fim de aumentar sua confiabilidade e ampliar seu potencial de uso como suporte objetivo à estratificação de risco e ao planejamento terapêutico de pacientes com DIIs. O elevado número de variáveis presentes no banco de dados, totalizando 932 informações por paciente, constitui também num aspecto relevante a ser explorado em estudos futuros. A seleção das variáveis de maior importância, identificadas por meio da técnica SHAP e posteriormente validadas pelo corpo clínico, poderá orientar a construção de modelos de IA mais simplificados, interpretáveis e clinicamente aplicáveis, reduzindo a complexidade do algoritmo sem comprometer seu potencial preditivo.

REFERÊNCIAS

ALI, Sajid; ABUHMED, Tamer; EL-SAPPAGH, Shaker; MUHAMMAD, Khan; ALONSO-MORAL, Jose M.; CONFALONIERI, Roberto; GUIDOTTI, Riccardo; DEL SER, Javier; DÍAZ-RODRÍGUEZ, Natalia; HERRERA, Francisco. Explainable Artificial Intelligence (XAI): what we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, v. 99, artigo 101805, nov. 2023. DOI: 10.1016/j.inffus.2023.101805.

ATREYA, Raja; NEURATH, Markus F. Biomarkers for personalizing IBD therapy: the quest continues. Clinical Gastroenterology and Hepatology, v. 22, n. 7, p. 1353-1364, jul. 2024. DOI: 10.1016/j.cgh.2024.01.026.

BERGMANN, Victor. T de Studant: entendendo o Teste t de Student: do conceito à aplicação em testes A/B. *Medium*, 25 nov. 2025. Disponível em: https://medium.com/@vic.kunde/t-de-studant-15b000db59cc. Acesso em: 11 jun. 2026.

BLAGUS, Rok; GOEMAN, Jelle J. What (not) to expect when classifying rare events. *Briefings in Bioinformatics*, v. 19, n. 2, p. 341–349, 2018. DOI: 10.1093/bib/bbw107.

CABOT, John H.; ROSS, Elsie Gyang. Evaluating prediction model performance. *Surgery*, v. 174, n. 3, p. 723–726, 2023. DOI: 10.1016/j.surg.2023.05.023.

Data Science Academy. Deep Learning Book, 2025. Disponível em: <https://www.deeplearningbook.com.br/>. Acesso em: 10 Abril de 2025**.**

FALCÃO, Luiza C. J.; POZZEBON, Eliana; MICHELON, Ana L. B.; MOURA, Ana Maria M. O uso de deep learning para classificação de imagens com modelo pré-treinado: uma experiência na área da saúde. Em Questão, v. 30, e-135550, 2024. DOI: 10.1590/1808-5245.30.135550.

FAWCETT, Tom. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006. DOI: 10.1016/j.patrec.2005.10.010.

FRÓES, R. S. B., Andrade, A. R., Faria, M. A. G., de Souza, H. S. P., Parra, R. S., Zaltman, C., Dos Santos, C. H. M., Bafutto, M., Quaresma, A. B., Santana, G. O., Luporini, R. L., de Lima Junior, S. F., Miszputen, S. J., de Souza, M. M., Herrerias, G. S. P., Junior, R. L. K., do Nascimento, C. R., Féres, O., de Barros, J. R., Sassaki, L. Y., … Saad-Hossne, R. (2024). Clinical factors associated with severity in patients with

inflammatory bowel disease in Brazil based on 2-year national registry data from GEDIIB. *Scientific reports*, *14*(1), 4314. https://doi.org/10.1038/s41598-024-54332-1

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep Learning*. Cambridge: MIT Press, 2016. Disponível em: https://www.deeplearningbook.org. Acesso em: 16 jun. 2026.

GUBATAN, John; LEVITTE, Steven; PATEL, Akshar; BALABANIS, Tatiana; WEI, Mike T.; SINHA, Sidhartha R. Artificial intelligence applications in inflammatory bowel disease: emerging technologies and future directions. *World Journal of Gastroenterology*, [S. l.], v. 27, n. 17, p. 1920–1935, 7 maio 2021. DOI: 10.3748/wjg.v27.i17.1920.

JUNIOR T. S. A; PEREIRA M. B.; TEIXEIRA M. L. S.; SILVA V. V. S. Avanços recentes no tratamento da Retocolite Ulcerativa: Uma revisão abrangente *Revista Contemporânea*, v. 13, n. 32, p. 1–10, 2023.

KAPLAN, Andreas; HAENLEIN, Michael. Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, v. 62, n. 1, p. 15-25, 2019. DOI: 10.1016/j.bushor.2018.08.004.

KAUFMAN, Shachar; ROSSET, Saharon; PERLICH, Claudia; STITELMAN, Ori. Leakage in data mining: formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, New York, v. 6, n. 4, artigo 15, p. 1-21, dez. 2012. DOI: 10.1145/2382577.2382579.

KRASZEWSKI, Sebastian; SZCZUREK, Witold; SZYMCZAK, Julia; REGUŁA, Monika; NEUBAUER, Katarzyna. Machine learning prediction model for inflammatory bowel disease based on laboratory markers: working model in a discovery cohort study. *Journal of Clinical Medicine*, [S. l.], v. 10, n. 20, art. 4745, 16 out. 2021. DOI: 10.3390/jcm10204745.

LANDESMAN, Ronen. The importance of data cleaning in the analysis process. *E-Learn*, 12 ago. 2024. Disponível em: <https://e-learn.guide/the-importance-of-data-cleaning-in-the-analysis-process/>. Acesso em: 11 maio 2025.

LI, Jing. *Area under the ROC Curve has the most consistent evaluation for binary classification*. PLOS ONE, v. 19, n. 12, e0316019, 2024.

LODDO, Italia; ROMANO, Claudio. Inflammatory bowel disease: genetics, epigenetics, and pathogenesis. Frontiers in Immunology, v. 6, art. 551, 2 nov. 2015. DOI: 10.3389/fimmu.2015.00551.

LOFTUS, Edward V.. Emerging diagnostic methods in inflammatory bowel disease. Gastroenterology & Hepatology (N Y), v. 3, n. 4, p. 284-286, abr. 2007.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. In: GUYON, Isabelle et al. (ed.). *Advances in Neural Information Processing Systems 30*. Red Hook: Curran Associates Inc., 2017. p. 4768-4777. DOI: 10.5555/3295222.3295230.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. *An introduction to information retrieval*. Cambridge: Cambridge University Press, 2009. Disponível em: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf. Acesso em: 27 maio 2025.

MARKOV, K. *Multilayer Perceptron with Backpropagation*. Problems of Engineering Cybernetics and Robotics, 2023. Disponível em: https://www.iict.bas.bg/pecr/2023/80/2-PECR-13-22.pdf. Acesso em: 16 maio 2025.

NGUYEN, Nghia H.; PICETTI, Dominic; DULAI, Parambir S.; JAIRATH, Vipul; SANDBORN, William J.; OHNO-MACHADO, Lucila; CHEN, Peter L.; SINGH, Siddharth. Machine learning-based prediction models for diagnosis and prognosis in inflammatory bowel diseases: a systematic review. *Journal of Crohn's & Colitis*, [S. l.], v. 16, n. 3, p. 398–413, 7 set. 2021. DOI: 10.1093/ecco-jcc/jjab155.

PAIXÃO, Guilherme M. M.; ANDRADE, Gustavo F. S.; GOMES, Cristiano P.; *et al*. Machine learning na medicina: revisão e aplicabilidade. Arquivos Brasileiros de Cardiologia, v. 118, n. 1, p. 95-102, 2022. DOI: 10.36660/abc.20200596.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

POWERS, David M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*, 2020. arXiv:2010.16061v1. Disponível em: https://arxiv.org/abs/2010.16061v1. Acesso em: 27 maio 2025.

PROVOST, Foster; FAWCETT, Tom. *Data science for business: what you need to know about data mining and data-analytic thinking*. 1. ed. Sebastopol, CA: O'Reilly Media, 2013. 414 p. Disponível em: https://dl.acm.org/doi/10.5555/2564781. Acesso em: 27 maio 2025.

QUARESMA, Abel B.; DAMIAO, Aderson O. M. C.; COY, Claudio S. R.; *et al*. Temporal trends in the epidemiology of inflammatory bowel diseases in the public healthcare system in Brazil: a large population-based study. *The Lancet Regional Health – Americas*, [S. l.], v. 13, p. 100298, 2022. DOI: 10.1016/j.lana.2022.100298.

SAITO, Takaya; REHMSMEIER, Marc. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, v. 10, n. 3, e0118432, 2015. DOI: 10.1371/journal.pone.0118432.

SHARMA, Anju; KUMAR, Rajnish; GARG, Prabha. Deep learning-based prediction model for diagnosing gastrointestinal diseases using endoscopy images. International Journal of Medical Informatics, v. 177, art. 105142, set. 2023. DOI: 10.1016/j.ijmedinf.2023.105142.

SONAWANE, Jayshril S.; PATIL, D. R. Prediction of heart disease using multilayer perceptron neural network. In: ICICES 2014 (S.A. Engineering College), 2014, Chennai, India. *Anais* [...]. Chennai: IEEE, 2014. p. 1–6. ISBN 978-1-4799-3834-6

STRYKER, Cole; KAVLAKOGLU, Eda. O que é inteligência artificial (IA)?. *IBM Think*, [S. l.], [s.d.]. Disponível em: https://www.ibm.com/br-pt/think/topics/artificial-intelligence. Acesso em: 09 maio 2025.

TEIM JUNIOR, Adalton Cicero; SILVA, Fagner Augusto Ramos da; BRITO, Luiz Felipe Ferreira. Utilização da inteligência artificial na medicina. Revista FT (ISSN 1678-0817 Qualis/DOI), v. 28, ed. 134, maio 2024. DOI: 10.5281/zenodo.11235389. Disponível em: https://revistaft.com.br/utilizacao-da-inteligencia-artificial-na-medicina/. Acesso em: 09 maio 2025.

WALFISH, Aaron E.; COMPANIONI, Rafael Antonio Ching. Medicamentos para doença inflamatória intestinal. In: Manuais MSD (versão para profissionais de saúde). Revisado/corrigido: nov. 2023; modificado: jun. 2024. Disponível em: https://www.msdmanuals.com/pt/profissional/dist%C3%BArbios-

gastrointestinais/doen%C3%A7a-inflamat%C3%B3ria-intestinal-dii/medicamentos-para-doen%C3%A7a-inflamat%C3%B3ria-intestinal. Acesso em: 24 abr. 2025.

XU, Manfei; FRALICK, Drew; ZHENG, Julia Z.; WANG, Bokai; TU, Xiaoming M.; FENG, Changyong. The differences and similarities between two-sample t-test and paired t-test. *Shanghai Archives of Psychiatry*, Shanghai, v. 29, n. 3, p. 184-188, jun. 2017. DOI: 10.11919/j.issn.1002-0829.217070.

ZALTMAN, Cyrla; CHEBLI, Julio Maria Fonseca; TEIXEIRA, Magaly Gêmio; *et al*. (Orgs.). As doenças inflamatórias intestinais na atualidade brasileira: curso de atualização do GEDIIB na SBAD 2018. São Paulo: Office Editora, 2018. 98 p. ISBN 978-85-87456-47-5. Disponível em: https://gediib.org.br/wp-content/uploads/2019/07/Livro_As-Doencas-Inflamatorias-Intestinais-na-Atualidade-Brasileria-GEDIIB-2018.pdf. Acesso em: 24 abr. 2025.

ZAND, Aria; STOKES, Zack; SHARMA, Arjun; *et al*. Artificial intelligence for inflammatory bowel diseases (IBD); accurately predicting adverse outcomes using machine learning. Digestive Diseases and Sciences, v. 67, p. 4874-4885, 2022. DOI: 10.1007/s10620-022-07506-8.

APÊNDICE I

Varáveis utilizadas como entrada para a RNA MLP

ID	VARIÁVEIS	Desfecho DII	Desfecho Cirúgico	Desfecho Medicamentoso
1	Paciente evoluiu com intestino curto? 1:Sim 0:Não	X		X
2	Paciente evoluiu com incontinência fecal? 1:Sim 0:Não	X		X
3	Cirugia abd com necess de: COLOSTOMIA = ALÇA	X		X
4	Cirugia abd com necess de: COLOSTOMIA = TERMINAL	X		X
5	Cirugia abd com necess de: COLOSTOMIA = ILEOSTOMIA	X		X
6	Tipo de cirurgia abdominal: ENTERECTOMIA	X		X
7	Tipo de cirurgia abdominal: ILEOTIFLECTOMIA	X		X
8	Tipo de cirurgia abdominal: RESERVATÓRIO ILEAL	X		X
9	Tipo de cirurgia abdominal: COLOCTOMIA	X		X
10	Tipo de cirurgia abdominal: LAPAROTIMIA DIAGNÓSTICA	X		X
11	Tipo de cirurgia abdominal: APENDICECTOMIA	X		X
12	Tipo de cirurgia abdominal: DRENAGEM DE ABSCESSO	X		X
13	Tipo de cirurgia abdominal: OUTRA	X		X
14	Tipo de cirurgia abdominal: PASSAGEM DE SETON	X		X
15	Cirurgia Classificação: ABDOMINAL	X		X
16	Cirurgia Classificação: PERIANAL	X		X
17	Forma de cirurgia: LAPAROTOMIA	X		X
18	Forma de cirurgia: LAPAROSCOPIA	X		X
19	Tipo de cirurgia: ELETIVA	X	X	X
20	Tipo de cirurgia: URGENCIA	X	X	X
21	Tipo de cirurgia perianal: DILATAÇÃO ANAL	X		X
22	Tipo de cirurgia perianal: FISTULOTOMIA	X		X
23	Tipo de cirurgia perianal: COLOCAÇÃO DE SEDENHO	X		X
24	Tipo de cirurgia perianal: RETIRADA DE SEDENHO	X		X
25	Tipo de cirurgia perianal: DRENAGEM DE ABSCESSO	X		X
26	Tipo de cirurgia perianal: CURETAGEM DE TRAJETOS	X		X
27	Tipo de cirurgia perianal: OUTRO	X		X
28	Gerou internação? 1:Sim, 0:Não	X	X	X
29	Tipo de complicação ou atividade da doença: ATIVIDADE DA DOENÇA	X	X	X
30	Tipo de complicação ou atividade da doença: HERPES SIMPLES	X	X	X
31	Tipo de complicação ou atividade da doença: HERPES ZOSTER	X	X	X
32	Tipo de complicação ou atividade da doença: INFECÇÃO FUNGICA	X	X	X
33	Tipo de complicação ou atividade da doença: INFECÇÃO RESPIRATÓRIA	X	X	X
34	Tipo de complicação ou atividade da doença: INFECÇÃO URINÁRIA	X	X	X
35	Tipo de complicação ou atividade da doença: OUTRO	X	X	X
36	Tipo de complicação ou atividade da doença: PANCREATITE	X	X	X
37	Tipo de complicação ou atividade da doença: PNEUMONIA	X	X	X
38	Tipo de complicação ou atividade da doença: SINUSITE	X	X	X
39	Tipo de complicação ou atividade da doença: TUBERCULOSE	X	X	X
40	Paciente foi a óbito? 1:Sim, 0:Não	X	X	X
41	O óbito teve relação com a DII? 1:Sim, 0:Não	X	X	X
42	Paciente já fez alguma cirurgia? 1:SIM 0:NÃO	X	X	X
43	Comorbidades : HAS	X	X	X
44	Comorbidades : ICC	X	X	X
45	Comorbidades : DOENÇA ONCOLÓGICA	X	X	X
46	Comorbidades : DOENÇA CELIACA	X	X	X
47	Comorbidades : HEPATOPATIA CRÔNICA	X	X	X
48	Comorbidades : OBESIDADE	X	X	X
49	Comorbidades : DM	X	X	X
50	Comorbidades : SEM COMORBIDADES	X	X	X

51	Comorbidades : DPOC	X	X	X
52	Comorbidades : OUTRAS	X	X	X
53	Comorbidade oncológicas: CANCER COLORRETAL	X	X	X
54	Comorbidade oncológicas: CANCER DE PELE	X	X	X
55	Comorbidade oncológicas: CANCER DE PULMÃO	X	X	X
56	Comorbidade oncológicas: CANCER DE MAMA	X	X	X
57	Comorbidade oncológicas: CANCER DE PANCREAS	X	X	X
58	Comorbidade oncológicas: CANCER DE FIGADO	X	X	X
59	Comorbidade oncológicas: OUTROS	X	X	X
60	Tipo de câncer: COLORRETAL	X	X	X
61	Tipo de câncer: FIGADO	X	X	X
62	Tipo de câncer: MAMA	X	X	X
63	Tipo de câncer: PELE MELANOMA	X	X	X
64	Tipo de câncer: PELE N MELANOMA	X	X	X
65	Tipo de câncer: PROSTATA	X	X	X
66	Tipo de câncer: PULMÃO	X	X	X
67	Tipo de câncer: OUTROS	X	X	X
68	Tipo de câncer: PANCREAS	X	X	X
69	Tipo de câncer: HEMATOLÓGICO(LEUCEMIA/LINFOMA)	X	X	X
70	Tipo de câncer: VIAS BILIARES	X	X	X
71	Tipo de câncer: TIREÓIDE	X	X	X
72	Tipo de câncer: ESÔFAGO	X	X	X
73	Tipo de câncer: TRATO URINÁRIO	X	X	X
74	Tipo de câncer: APENDICE	X	X	X
75	Tipo de câncer: GINECOLÓGICO	X	X	X
76	Tipo de câncer: INTESTINO DELGADO	X	X	X
77	Paciente teve alfuma complicação ou atividade da doença que gerou internação? 1:Sim, 0:Não	X	X	X
78	Comportamento: NÃO ESTEOSANTE, NÃO PEDETRANTE	X	X	X
79	Comportamento: ESTENOSANTE	X	X	X
80	Comportamento: PENETRANTE	X	X	X
81	Comportamento: MODIFICADOR DE DOENÇA	X	X	X
82	Localização: ILEAL	X	X	X
83	Localização: COLÔNICA	X	X	X
84	Localização: ILEOCÓLICA	X	X	X
85	Localização: DOENÇA TGI SUPERIOR ISOLADA	X	X	X
86	Tipo de doença: CROHN	X	X	X
87	Tipo de doença: RCU	X	X	X
88	Quantidade de fistula: UMA	X	X	X
89	Quantidade de fistula: MAIOR QUE UM	X	X	X
90	Resultado do tratamento: FECHAMENTO COMPLETO	X	X	X
91	Resultado do tratamento: REFRATARIEDADE DO TRATAMENTO	X	X	X
92	Presença ou ausencia de fistula? 1:Presença, 0:Ausência	X	X	X
93	Tipo de fistula: PERIANAL	X	X	X
94	Tipo de fistula: ENTERO ENTÉRICA	X	X	X
95	Tipo de fistula: ENTERO COLÔNICA	X	X	X
96	Tipo de fistula: RETO VAGINAL	X	X	X
97	Tipo de fistula: OUTRO	X	X	X
98	Tipo de fistula Perianal: COMPLEXA	X	X	X
99	Tipo de fistula Perianal: SIMPLES	X	X	X
100	Tipo de tratamento: BIOLÓGICO	X	X	X

101	Tipo de tratamento: TRATAMENTO CIRÚRGICO	X	X	X
102	Tipo de tratamento: BIOLÓGICO + CIRURGIA	X	X	X
103	Tipo de tratamento: OUTROS	X	X	X
104	Manifestações Iniciais: ARTRALGIA	X	X	X
105	Manifestações Iniciais: CONSTIPAÇÃO	X	X	X
106	Manifestações Iniciais: DIARRÉIA	X	X	X
107	Manifestações Iniciais: DOR ABDOMINAL	X	X	X
108	Manifestações Iniciais: EMAGRECIMENTO	X	X	X
109	Manifestações Iniciais: ENTERORRAGIA	X	X	X
110	Manifestações Iniciais: FADIGA	X	X	X
111	Manifestações Iniciais: OUTRAS	X	X	X
112	Manifestações Iniciais: FEBRE	X	X	X
113	Manifestações Iniciais: HEMATOQUEZIA	X	X	X
114	Manifestações Iniciais: VÔMITO/NAUSEA	X	X	X
115	Manifestações Iniciais: AFTAS ORAIS	X	X	X
116	Manifestações Iniciais: DOENCA PERIANAL (FISTULA/ABCESSO OU FISSURA)	X	X	X
117	Manifestações Iniciais: TENESMO/PUXO/PROCTALGIA OU URGENCIA	X	X	X
118	Manifestações Iniciais: MUCO NAS FEZES/MUCORREIA	X	X	X
119	Manifestações Iniciais: ABDOME AGUDO INFLAMATORIO/APENDICITE	X	X	X
120	Manifestações Iniciais: ABDOME AGUDO PERFURATIVO/PERFURACAO INTESTINAL	X	X	X
121	Manifestações Iniciais: ABDOME AGUDO OBSTRUTIVO/OBSTRUCAO INTESTINAL	X	X	X
122	Manifestações Extras intestinais: REUMATOLÓGICAS	X	X	X
123	Manifestações Extras intestinais: DERMATOLÓGICAS	X	X	X
124	Manifestações Extras intestinais: OFTALMOLÓGICAS	X	X	X
125	Manifestações Extras intestinais: HEPATOBILIARES	X	X	X
126	Manifestações Extras intestinais: NEFROLÓGICAS	X	X	X
127	Manifestações Extras intestinais: TROMBOSE	X	X	X
128	Manifestações Extras intestinais: OUTRAS	X	X	X
129	Manifestações Extras intestinais: SEM MANIFESTAÇÃO	X	X	X
130	Manifestações Dermatológicas: AFTAS ORAIS	X	X	X
131	Manifestações Dermatológicas: ERITEMA NODOSO	X	X	X
132	Manifestações Dermatológicas: PIODERMA GANGRENOSO	X	X	X
133	Manifestações Dermatológicas: HIDRANITE SUPURATIVA	X	X	X
134	Manifestações Dermatológicas: PSORÍASE	X	X	X
135	Manifestações Hepatobiliares: CEP	X	X	X
136	Manifestações Hepatobiliares: COLELITÍASE	X	X	X
137	Manifestações Nefrológicas: NEFROLISTÍASE	X	X	X
138	Manifestações Hepatobiliares: AMILOIDOSE	X	X	X
139	Manifestações Oftálmicas: EPISCLERITE	X	X	X
140	Manifestações Oftálmicas: UVEÍTE	X	X	X
141	Manifestações Reumatológicas: ARTRALGIA/ARTRITE	X	X	X
142	Manifestações Reumatológicas: SACROILEÍTE	X	X	X
143	Manifestações Reumatológicas: ESPONDILITE ANQUILOSANTE	X	X	X
144	Manifestações Reumatológicas: ARTRITE IDIOPÁTICA JUVENIL POLIARTICULAR	X	X	X
145	Manifestações Reumatológicas: ARTRITE PSORIÁSICA	X	X	X
146	Manifestações Reumatológicas: ARTRITE RELACIONADA À ENTESITE	X	X	X
147	Manifestações Reumatológicas: ARTRITE REUMATÓIDE	X	X	X
148	Manifestações Reumatológicas: ESPONDILOARTRITE AXIAL NÃO RADIOGRÁFICA	X	X	X
149	Outras manifestações: ATRASO PUBERAL	X	X	X
150	Outras manifestações: ATRASO DE CRESCIMENTO	X	X	X

151	Outras manifestações: OSTEOPOROSE	X	X	X
152	Tipo de Trombose: ARTERIAL	X	X	X
153	Tipo de Trombose: VENOSA	X	X	X
154	Tipo de Trombose: TROMBOEMBOLISMO	X	X	X
155	Paciente já usou ou está em uso de alguma medicação para o tratamento da doença? 1:Sim, 0:Não	X	X	X
156	Extensão anatômica da inflamação: RCU DISTAL (PROCTITE OU PROCTOSIGMOIDITE DISTAL)	X	X	X
157	Extensão anatômica da inflamação: RCU ESQUERDA (INFLAMAÇÃO DA MUCOSA ATÉ A FLEXURA ESPLÊNICA, EVENTUALMENTE, ATÉ CÓLON TRANSVERSO DISTAL)	X	X	X
158	Extensão anatômica da inflamação: PANCOLITE (OU COLITE EXTENSA)	X	X	X
159	Idade ao início dos sintomas: VALOR	X	X	X
160	Tabagismo: EX-TABAGISTA	X	X	X
161	Tabagismo: Não	X	X	X
162	Tabagismo: Sim	X	X	X
163	Complicações da DII no câncer colorretal: NÃO	X	X	X
164	Complicações da DII no câncer colorretal: SIM	X	X	X
165	Houve necessidade de mais de dois ciclos no decorrer do tratamento?: NÃO	X	X	
166	Houve necessidade de mais de dois ciclos no decorrer do tratamento?: SIM	X	X	
167	Usou apenas no inicio do tratamento?: NÃO	X	X	
168	Usou apenas no inicio do tratamento?: SIM	X	X	
169	Classe med Biológicos: ADALIMUMABE (HUMIRA)	X	X	
170	Classe med Biológicos: CERTOLIZUMABE (CIMZIA)	X	X	
171	Classe med Biológicos: GOLIMUMAB (SIMPONI)	X	X	
172	Classe med Biológicos: INFLIXIMABE (REMICADE)	X	X	
173	Classe med Biológicos: INFLIXIMABE (REMSIMA)	X	X	
174	Classe med Biológicos: OUTRO	X	X	
175	Classe med Biológicos: USTEQUINUMABE (STELARA)	X	X	
176	Classe med Biológicos: VEDOLIZUMABE (ENTYVIO)	X	X	
177	Classe de medicação: BIOLÓGICOS	X	X	X
178	Classe med Corticóide: BUDESONIDA	X	X	
179	Classe med Corticóide: ENEMA	X	X	
180	Classe med Corticóide: HIDROCORTISONA	X	X	
181	Classe med Corticóide: OUTRO	X	X	
182	Classe med Corticóide: PREDNISONA	X	X	
183	Classe de medicação: CORTICÓDIDE	X	X	
184	Classe med Imunosupressor: 6-MERCAPTOPURINA	X	X	
185	Classe med Imunosupressor: AZATIOPRINA	X	X	
186	Classe med Imunosupressor: CICLOSPORINA	X	X	
187	Classe med Imunosupressor: METOTREXATO	X	X	
188	Classe med Imunosupressor: OUTRO	X	X	
189	Classe de medicação: IMUNOSSUPRESSOR	X	X	
190	Classe de medicação: OUTROS	X	X	
191	Classe med Salicilatos: MESALAZINA ENEMA	X	X	
192	Classe med Salicilatos: MESALAZINA ORAL	X	X	
193	Classe med Salicilatos: MESALAZINA SUPOSITORIO	X	X	
194	Classe med Salicilatos: OUTRO	X	X	
195	Classe med Salicilatos: SULFASSALAZINA	X	X	
196	Classe de medicação: SALICILATO	X	X	
197	Data aprox finalizou o uso da medicação: DATA	X	X	
198	Data aprox inicio do uso da medicação: DATA	X	X	
199	Dose padrão ou dose otimizada? 1:PADRÃO 2: OTIMIZADA	X	X	
200	Em uso atual? 1:Sim, 0:Não	X	X	
201	Qual o motivo da suspensão do medicamento?: EFEITO COLATERAL	X	X	
202	Qual o motivo da suspensão do medicamento?: FALHA TERAPEUTICA	X	X	
203	Qual o motivo da suspensão do medicamento?: OUTROS	X	X	
204	Qual o motivo da suspensão do medicamento?: PERDA DE ACESSO	X	X	

APÊNDICE II

Na Tabela 1, são apresentados os resultados do teste t de Student aplicado às métricas de desempenho dos modelos desenvolvidos para os desfechos cirúrgico e medicamentoso, considerando os resultados obtidos no blind test. Esse teste foi utilizado com o objetivo de comparar, de forma estatística, as métricas obtidas pelos modelos treinados com dados originais e com dados balanceados, verificando se as diferenças observadas entre esses dois cenários foram significativas ou se poderiam ser atribuídas à variação aleatória entre as execuções.

Tabela 1: Os valores são apresentados como média ± desvio padrão. As comparações entre as métricas obtidas nas 10 execuções com dados originais e balanceados no blind test foram realizadas por meio do teste t de Student pareado. A diferença em p.p. corresponde à diferença absoluta entre as médias, expressa em pontos percentuais. Valores de $p < 0.05$ foram considerados estatisticamente significativos.

Desfecho	Cenário	Métrica	Original média ± DP	Balanceado média ± DP	Diferença em p.p	Aumento relativo	T	P
Cirurgia	Geral	Precision	57.40% ±1.43	80.60% ±0.97	23.20	40.42%	71.0352	P<0.001
		Recall	57.10% ±1.20	80.30% ±0.95	23.20	40.63%	55.7246	P<0.001
		F1-Score	57.20% ±0.79	80.50% ±0.53	23.20	40.73%	77.6667	P<0.001
Cirurgia	DC	Precision	75.30% ±4.72	85.00% ±2.54	9.70	12.88%	5.5786	P<0.001
		Recall	57.00% ±2.71	81.20% ±2.15	24.20	42.46%	27.9231	P<0.001
		F1-Score	65.10% ±1.97	82.90% ±0.74	17.80	27.34%	33.3750	P<0.001
Cirurgia	RCU	Precision	41.90% ±4.75	68.50% ±5.10	26.60	63.48%	12.2784	P<0.001
		Recall	58.20% ±4.64	97.80% ±4.64	39.60	68.04%	22.0454	P<0.001
		F1-Score	48.60% ±4.01	80.40% ±2.27	31.80	65.43%	22.5110	P<0.001
Medicação	Geral	Precision	82.10% ± 0.74	82.20% ± 0.92	0.10	0.12%	0.3612	P>0.05
		Recall	71.60% ± 2.17	73.40% ± 2.27	1.80	2.51%	1.9140	P>0.05
		F1-Score	76.60% ± 1.07	77.60% ± 0.97	1.00	1.31%	2.1213	P>0.05
Medicação	DC	Precision	66.50% ±8.05	62.80% ±5.45	-3.70	-5.56%	-1.9558	P>0.05
		Recall	11.10% ±3.70	43.60% ±17.88	32.50	292.79%	4.8650	P<0.001
		F1-Score	18.70% ±5.54	49.70% ±13.48	31.00	165.78%	5.2940	P<0.001
		Precision	55.70%	64.30%	8.60	15.44%	6.9857	P<0.001

			±3.09	±3.80				
Medicação	RCU	Recall	47.40% ±2.07	80.50% ±2.72	33.10	69.83%	32.5793	P<0.001
		F1-Score	51.00% ±1.15	71.50% ±1.58	20.50	40.20%	76.2814	P<0.001