UNIVERSIDADE FEDERAL DO MARANHÃO - UFMA
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIAS - CCET
COORDENAÇÃO DO CURSO DE ENGENHARIA DA COMPUTAÇÃO

BRENNO PACHECO CARNEIRO

PREVISÃO DE DOENÇAS CRÔNICAS EM IDOSOS COM APRENDIZADO DE MÁQUINA:

Uma abordagem de classificação multirrótulo utilizando dados do Estudo ELSI-Brasil

SÃO LUÍS - MA
2026

BRENNO PACHECO CARNEIRO

PREVISÃO DE DOENÇAS CRÔNICAS EM IDOSOS COM APRENDIZADO DE MÁQUINA:

Uma abordagem de classificação multirrótulo utilizando dados do Estudo ELSI-Brasil

Trabalho de Conclusão de Curso apresentado para obtenção do título de Bacharel em Engenharia da Computação, pela Universidade Federal do Maranhão - UFMA, Cidade Universitária: Campus Dom Delgado, São Luís, MA.

Orientador: Prof. Dr. Bruno Feres de Souza

SÃO LUÍS - MA

2026

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Pacheco Carneiro, Brenno.
PREVISÃO DE DOENÇAS CRÔNICAS EM IDOSOS COM APRENDIZADO DE MÁQUINA: uma abordagem de classificação multirrótulo utilizando dados do Estudo ELSI-Brasil / Brenno Pacheco Carneiro. - 2026.
53 p.

Orientador(a): Bruno Feres de Souza.
Curso de Engenharia da Computação, Universidade Federal do Maranhão, São Luís, 2026.

1. Doenças Crônicas Não Transmissíveis. 2. Multimorbidade. 3. Classificação Multirrótulo. 4. Aprendizado de Máquina. 5. ELSI-Brasil. I. Feres de Souza, Bruno. II. Título.

BRENNO PACHECO CARNEIRO

PREVISÃO DE DOENÇAS CRÔNICAS EM IDOSOS COM APRENDIZADO DE MÁQUINA:
Uma abordagem de classificação multirrótulo utilizando dados do Estudo ELSI-Brasil

Aprovada em: 28 / 01/ 2026.

BANCA EXAMINADORA

Prof. Dr. Bruno Feres de Souza (Orientador)
UFMA

Prof. Dr. Alex Oliveira Barradas Filho
UFMA

Prof. Dr. Paulo Rogério de Almeida Ribeiro
UFMA

SÃO LUÍS - MA
2026

AGRADECIMENTOS

Agradeço imensamente à minha família, minha mãe Sildemira Pacheco, meu pai Sérgio Victor e meu irmão André Luís Pacheco, por todo apoio incondicional ao longo dessa jornada. Vocês sempre estiveram ao meu lado, mesmo nos momentos mais difíceis, estendendo a mão quando eu já não acreditava em mim. Obrigado por nunca desistirem de mim.

Agradeço também à minha família estendida: tias, tios, primos, primas... Em especial, à minha vó Silvanira Pacheco e ao meu vô Ildemar Pacheco, que foram figuras paternas fundamentais durante a minha infância e formação.

À minha namorada, Emanuelle Victoria, minha companheira de vida, meu amor e minha força. Obrigado por estar comigo em cada passo, por me compreender, por me apoiar e por fazer parte dessa conquista.

Ao meu professor, coordenador, orientador e amigo Bruno Feres de Souza: meu sonho é um dia ser 1% do que o senhor representa como profissional e ser humano. O senhor é, para mim, exemplo, referência e inspiração. Obrigado por acreditar em mim.

RESUMO

O aumento da multimorbidade em idosos representa um desafio crescente, exigindo ferramentas preditivas que auxiliem e planejem ações na saúde pública. Este trabalho propõe uma abordagem baseada em aprendizado de máquina para predizer multimorbidade (presença simultânea de doenças crônicas) em idosos por meio de classificação multirrótulo. Utilizou-se a base da 2ª onda ELSI-Brasil, totalizando 9.617 indivíduos após o pré-processamento. Foram modeladas sete condições crônicas com prevalência ≥ 5%: hipertensão, diabetes, colesterol alto, artrite, depressão, osteoporose e problema crônico de coluna. O *pipeline* metodológico incluiu limpeza e recodificação de variáveis, definição de atributos (com ênfase em indicadores antropométricos), análise exploratória da coocorrência de doenças e modelagem com estratégias multirrótulo, comparando *Binary Relevance* e variações baseadas em cadeias de classificadores, utilizando *Random Forest* e *Support Vector Machine* como algoritmos base. A avaliação foi realizada por validação cruzada repetida (10-fold, 5 repetições), totalizando 50 avaliações independentes. Os resultados indicaram desempenho consistente entre os modelos, com *F1-Measure* entre 0,33–0,35, *Hamming Loss* aproximado de 0,237–0,243. Adicionalmente, evidenciou-se a relevância de variáveis clínicas e antropométricas (como IMC, Idade e RCQ) na predição das comorbidades. Conclui-se que a abordagem multirrótulo é viável na base ELSI-Brasil, oferecendo um caminho promissor para apoiar a estratificação de risco e o planejamento em saúde pública.

Palavras-chave: Doenças Crônicas Não Transmissíveis. Multimorbidade. Classificação Multirrótulo. Aprendizado de Máquina. ELSI-Brasil.

ABSTRACT

The rise of multimorbidity in elderly poses a growing challenge for public health, demanding predictive tools to assist and plan preventive actions. This work proposes a machine learning-based approach to predict multimorbidity (simultaneous presence of chronic diseases) in older adults using multi-label classification. Data from the second wave of the Brazilian Longitudinal Study of Aging (ELSI-Brasil) was used, comprising 9,617 individuals after pre-processing. Seven chronic conditions with prevalence ≥ 5% were modeled: hypertension, diabetes, high cholesterol, arthritis, depression, osteoporosis, and chronic back problems. The methodological pipeline included data cleaning, variable recoding, feature engineering (emphasizing anthropometric indicators), exploratory analysis of disease co-occurrence, and modeling using multi-label strategies. Binary Relevance and classifier chain-based variations were compared using Random Forest and SVM as base algorithms. Evaluation was performed using repeated cross-validation (10-fold, 5 repetitions), totaling 50 independent evaluations. Results indicated consistent performance across models, with *F1-Measure* ranging from 0.33–0.35, Hamming Loss approximately 0.237–0.243. Additionally, the relevance of clinical and anthropometric variables (such as BMI, Age, and WHR) in predicting comorbidities was evidenced. It is concluded that the multi-label approach is feasible on the ELSI-Brasil dataset, offering a promising path to support risk stratification and public health planning.

Keywords: Non-communicable Chronic Diseases. Multimorbidity. Multi-label Classification. Machine Learning. ELSI-Brasil.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
BR	Binary Relevance
CC	Classifier Chains
CMR	Classificação Multirrótulo
DBR	Dependent Binary Relevance
DCNTs	Doenças Crônicas Não Transmissíveis
ELSI-Brasil	Estudo Longitudinal da Saúde dos Idosos Brasileiros
FIOCRUZ	Fundação Oswaldo Cruz
MBR	Meta Binary Relevance
OMS	Organização Mundial da Saúde
RF	Random Forest
SVM	Support Vector Machine
UFMG	Universidade Federal de Minas Gerais
UTIML	Utilities for Multi-Label Learning

SUMÁRIO

1. INTRODUÇÃO

Com o aumento da expectativa de vida, surgem novos desafios para os sistemas de saúde. O envelhecimento da população está diretamente relacionado ao crescimento das Doenças Crônicas Não Transmissíveis (DCNTs), como diabetes, hipertensão e artrite. Essas doenças são comuns entre idosos e, muitas vezes, ocorrem simultaneamente, caracterizando quadros de comorbidade. Essa condição exige maior atenção médica e um uso mais intenso dos serviços de saúde, como internações frequentes e tratamentos prolongados (SIMIELI; PADILHA; TAVARES, 2019).

Diversos fatores agravam esse quadro nas populações mais vulneráveis, entre eles os maus hábitos alimentares, o sedentarismo, o consumo de substâncias prejudiciais à saúde e as desigualdades socioeconômicas. Diante desse cenário, o Ministério da Saúde criou o "Plano de Ações Estratégicas para o Enfrentamento das DCNTs no Brasil (2011-2022)", com foco na prevenção, vigilância e promoção da saúde (SIMIELI; PADILHA; TAVARES, 2019).

O avanço da tecnologia e da computação tem promovido transformações significativas na produção e aplicação do conhecimento científico, ampliando a integração entre diferentes áreas e possibilitando o desenvolvimento de soluções mais complexas e eficazes. Em especial, o setor da saúde tem sido fortemente impactado por essas inovações, beneficiando-se da convergência entre áreas como biologia, matemática, física e engenharia. Essa integração tem viabilizado avanços relevantes na medicina e na bioinformática, sobretudo no processamento e na análise de grandes volumes de dados clínicos e biomédicos.

Nesse contexto, a área da saúde tem se destacado como uma das maiores beneficiadas pelo uso do Aprendizado de Máquina (AM), capaz de analisar grandes volumes de dados e auxiliar na tomada de decisões. Segundo Faceli et al. (2011), o AM já é amplamente utilizado para o controle de epidemias, apoio a exames, monitoramento de pacientes e auxílio em diagnósticos, com destaque para a análise de dados obtidos por exames como eletrocardiogramas, tomografias e mamografias.

Um aspecto essencial nas aplicações médicas é que, além de precisão, os sistemas precisam ser interpretáveis. Isso significa que profissionais da saúde precisam compreender como o modelo chegou à determinada conclusão, já que erros em diagnósticos podem ter consequências graves, como atrasos no início do tratamento e complicações mais severas.

Ainda de acordo com Faceli et al. (2011), doenças como o diabetes são objeto de muitos estudos com uso de AM, pois um diagnóstico precoce pode evitar complicações futuras.

Com isso, o uso de tecnologias inteligentes na área médica tem se mostrado uma ferramenta poderosa na promoção da medicina preventiva.

Este trabalho apresenta uma abordagem para prever, de forma simultânea, quais doenças crônicas um idoso pode vir a apresentar, a partir de informações de estilo de vida, histórico médico e condições socioeconômicas disponíveis na base do Estudo Longitudinal da Saúde dos Idosos Brasileiros (ELSI-Brasil).

O problema é formulado como uma tarefa de Classificação Multirrótulo (CMR), na qual o desfecho é representado por um vetor binário que indica a presença ou ausência de cada doença crônica selecionada, permitindo que um mesmo indivíduo apresente múltiplas doenças concomitantes. Diferentemente das abordagens de classificação monorrótulo, a CMR possibilita a associação de vários rótulos a uma mesma instância, característica especialmente adequada à realidade da população idosa, frequentemente marcada pela presença de multimorbidade. Adicionalmente, a contagem de morbidades autorreferidas por participante é utilizada como medida derivada para análises descritivas de multimorbidade. A previsão desse perfil de doenças pode contribuir para o apoio a estratégias de prevenção e para o subsídio de intervenções clínicas mais oportunas e eficientes.

1.1 JUSTIFICATIVA

O envelhecimento da população representa um desafio crescente para os sistemas de saúde. À medida que a expectativa de vida aumenta, é comum que os idosos apresentem mais de uma doença crônica simultaneamente, como hipertensão, diabetes ou problemas cardiovasculares. Essas comorbidades afetam diretamente a qualidade de vida dessa parcela da população e geram uma sobrecarga nos serviços de saúde (SIMIELI; PADILHA; TAVARES, 2019). A capacidade de prever essas doenças de forma antecipada pode auxiliar os profissionais de saúde a planejarem tratamentos mais adequados, além de facilitar ações preventivas.

A base de dados do ELSI-Brasil oferece uma excelente oportunidade para esse tipo de análise, por conter informações detalhadas sobre saúde, hábitos e características sociais da população idosa.

Este trabalho se justifica por três principais motivos: contribuir com a melhoria da qualidade de vida dos idosos por meio da tecnologia; ajudar na tomada de decisão na área da saúde com base em dados confiáveis; e promover o uso de métodos inovadores, como a CMR, para lidar com problemas complexos e atuais.

Acredita-se que o desenvolvimento deste estudo trará benefícios tanto para a área acadêmica quanto para área da saúde, ao oferecer uma forma de auxílio à prevenção e ao diagnóstico de doenças crônicas em idosos.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Apresentar uma solução baseada em AM com foco na CMR para prever múltiplas doenças crônicas em pessoas idosas de forma simultânea, utilizando os dados do ELSI-Brasil.

1.2.2 Objetivos Específicos

- Sistematizar as informações mais importantes da base ELSI-Brasil, relacionadas à saúde, estilo de vida e aspectos sociais;
- Aplicar técnicas de CMR para prever quais doenças um mesmo idoso pode apresentar;
- Avaliar o desempenho dos modelos em prever corretamente os casos reais;

2. REFERENCIAL TEÓRICO

2.1 DOENÇAS CRÔNICAS NÃO TRANSMISSÍVEIS E ENVELHECIMENTO

As DCNTs representam um dos principais desafios de saúde pública na atualidade. Com características como longa duração, progressão lenta e causas multifatoriais, essas doenças se tornam ainda mais frequentes em um cenário de envelhecimento populacional, como o que ocorre no Brasil e em muitos outros países (MELO et al., 2022).

À medida que a população passa a viver mais, aumenta também a demanda por cuidados e por métodos que promovam o envelhecimento saudável. Enquanto no passado os seres humanos enfrentavam riscos físicos constantes, como lutas por sobrevivência, o estilo de vida atual é marcado pelo sedentarismo, dietas inadequadas, uso de substâncias prejudiciais e desigualdades sociais. Esses fatores estão fortemente relacionados ao crescimento das DCNTs, como aponta a Organização Mundial da Saúde (OMS) (2011).

Em 2022, as DCNTs foram responsáveis por 51,04% ($n = 755.747$) do total de óbitos registrados no Brasil, destacando-se as doenças do aparelho circulatório, o câncer, as doenças respiratórias crônicas e a diabetes mellitus tipo II (SOUZA et al., 2023).

Esses dados reforçam a importância de ações que deem prioridade a prevenção e o acompanhamento das DCNTs, principalmente entre os idosos. A compreensão do comportamento dessas doenças e a busca por soluções tecnológicas para enfrentá-las são essenciais para garantir mais qualidade de vida à população.

2.2. O ESTUDO LONGITUDINAL DA SAÚDE DOS IDOSOS BRASILEIROS

O ELSI-Brasil é uma pesquisa científica de base domiciliar e nacional criada para acompanhar o envelhecimento da população brasileira e compreender seus determinantes sociais, biológicos e psicológicos. O estudo é coordenado pela Universidade Federal de Minas Gerais (UFMG) em parceria com a Fundação Oswaldo Cruz (FIOCRUZ-MG), com financiamento do Ministério da Saúde (LIMA-COSTA et al., 2018).

De forma geral, o ELSI-Brasil reúne informações abrangentes sobre saúde física e mental, uso de serviços de saúde, condições socioeconômicas, bem-estar e medidas antropométricas. Essa amplitude torna o banco particularmente relevante para aplicações em AM, pois oferece variáveis capazes de capturar padrões complexos associados à multimorbidade (LIMA-COSTA et al., 2018).

Do ponto de vista metodológico, o ELSI-Brasil adota um desenho amostral complexo, estratificado e em múltiplos estágios (municípios, setores censitários e domicílios), assegurando representatividade de áreas urbanas e rurais, de municípios de diferentes portes, nas cinco grandes regiões do país. A coleta de dados é realizada em ondas, sendo que a segunda ocorreu entre 2019 e 2021 e abrangeu 9.949 indivíduos (LIMA-COSTA et al., 2018).

O ELSI-Brasil constitui uma base de dados de elevada robustez estatística, com destaque para a atualidade de sua segunda onda. Ao utilizar uma amostra nacionalmente representativa de idosos não institucionalizados, residentes na comunidade, os dados possibilitam o desenvolvimento de modelos preditivos com maior validade externa, permitindo a identificação de padrões de doenças crônicas que refletem, de maneira consistente, os desafios enfrentados pelo sistema de saúde brasileiro no contexto do envelhecimento populacional.

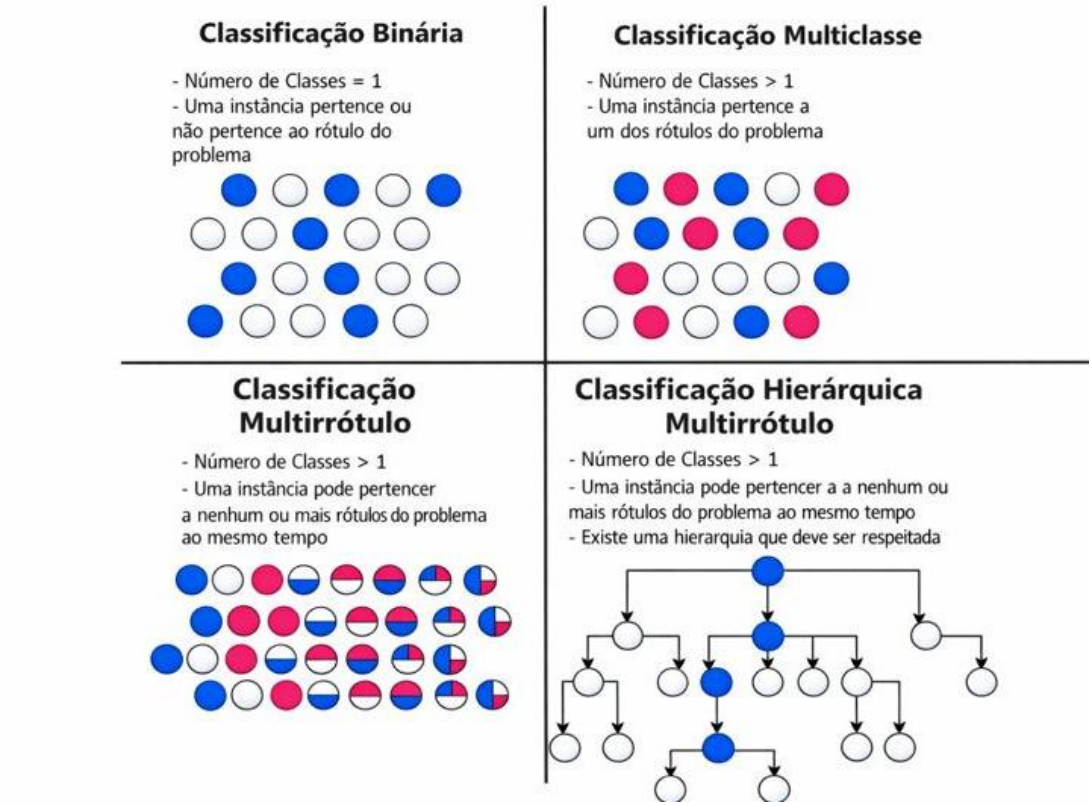
2.3 APRENDIZADO DE MÁQUINA E CLASSIFICAÇÃO MULTIRRÓTULO

O Aprendizado de Máquina (AM) é uma subárea da ciência da computação dedicada ao desenvolvimento de sistemas capazes de aprender a partir de dados. Segundo Han, Kamber e Pei (2011), essa abordagem possibilita a identificação de padrões e a realização de previsões por meio da análise de grandes volumes de informações. Ludermir (2021) ressalta que a qualidade dos dados exerce influência direta sobre a confiabilidade dos resultados, uma vez que conjuntos de dados bem estruturados favorecem a capacidade de generalização dos modelos. Nesse contexto, bases de dados abrangentes e confiáveis são fundamentais para aplicações do AM na área da saúde, como a previsão de doenças crônicas em idosos.

Entre os tipos de AM, destacam-se três categorias principais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. O aprendizado supervisionado é amplamente utilizado em problemas de classificação e regressão, pois opera sobre conjuntos de dados previamente rotulados, permitindo que os algoritmos aprendam padrões e generalizem esse conhecimento para novos casos (LUDERMIR, 2021).

No contexto dos problemas de classificação, diferentes configurações podem ser adotadas, como a classificação binária, a classificação multiclasse e a Classificação Multirrótulo (CMR). Enquanto nas abordagens binária e multiclasse cada instância é associada a um único rótulo, a CMR permite a atribuição simultânea de múltiplas categorias a uma mesma instância. Essa característica torna a CMR especialmente adequada para a modelagem de cenários nos quais coexistem múltiplas condições, como ocorre frequentemente na saúde da população idosa, marcada pela presença de multimorbidade.

Figura 1 - Tipos de classificação supervisionada



Fonte: Adaptado de GATTO (2023).

Na Figura 1, é apresentada uma comparação entre os principais tipos de classificação supervisionada, destacando suas diferenças conceituais e estruturais.

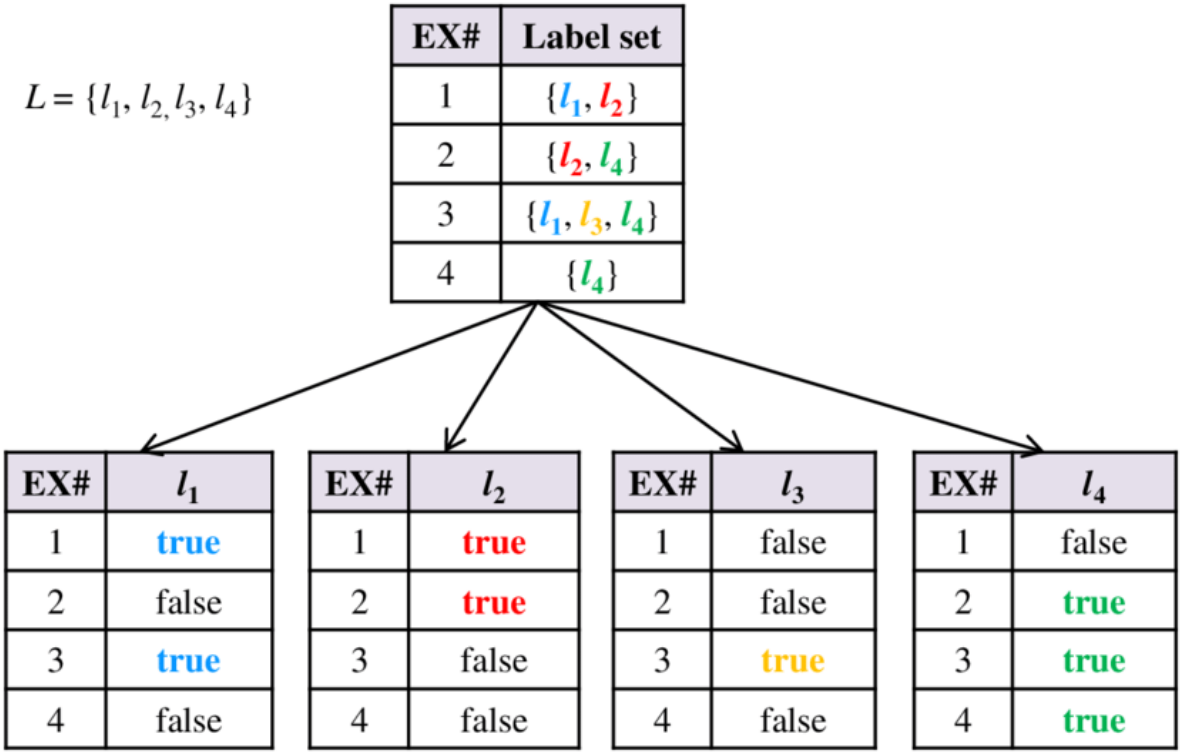
2.3.1 Estratégias de Transformação

A CMR diferencia-se da classificação monorrótulo por permitir que um único objeto seja associado simultaneamente a múltiplos rótulos de classe dentro de um conjunto (GONÇALVES, 2018). Para trabalhar com esse tipo de problema, uma das abordagens mais comuns é a Transformação do Problema (TP). Essa técnica consiste em converter o cenário multirrótulo em um ou mais problemas de classificação monorrótulo, viabilizando o uso de algoritmos clássicos como indutores base (GONÇALVES, 2018).

Dentre as estratégias de transformação, a mais difundida na literatura é o método de *Binary Relevance* (BR) (GONÇALVES, 2018). Sua premissa é simples e direta: decompor o problema original em L bases de dados binárias independentes, onde L representa o número total de rótulos distintos. Para cada rótulo, treina-se um classificador separado responsável por prever exclusivamente a relevância daquela classe específica (FACELI, 2011). A Figura 2

ilustra de forma conceitual esse processo de decomposição, evidenciando a criação de classificadores binários independentes para cada rótulo. A grande vantagem do BR reside em sua simplicidade teórica e complexidade linear, facilitando a implementação computacional. Entretanto, sua limitação crítica é ignorar as correlações entre os rótulos, já que as decisões são tomadas de forma isolada (FACELI, 2011). Em aplicações médicas, isso pode ser negativo, pois ignora as fortes correlações biológicas existentes entre certas comorbidades (GONÇALVES, 2018).

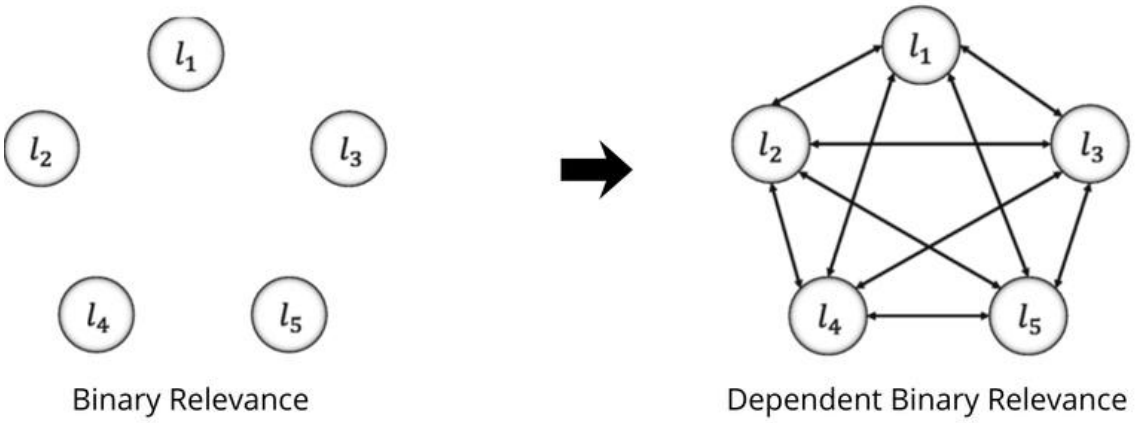
Figura 2 - Representação conceitual do método de Relevância Binária (Binary Relevance).



Fonte: VIDULIN (2012).

Para mitigar essa perda de informação sobre as relações entre as doenças, surgem estratégias mais robustas, como a *Dependent Binary Relevance* (DBR). Embora sua estrutura inicial se assemelhe ao BR, o DBR diferencia-se fundamentalmente por considerar a relação mútua entre os rótulos durante o processo preditivo (PAULA et al., 2022), conforme ilustrado na Figura 3.

Figura 3 - Comparação entre Binary Relevance (BR) e Dependent Binary Relevance (DBR)

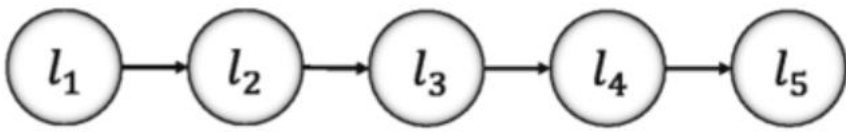


Fonte**:** Adaptado de Chen et al. (2020).

A técnica expande o espaço de atributos (*features*) de cada classificador individual: para prever um rótulo específico, o modelo utiliza como variáveis adicionais a informação real de todos os outros rótulos presentes no conjunto, exceto o alvo da predição corrente. Isso permite que algoritmos como *Random Forest* (RF) ou *Support Vector Machine* (SVM) aprendam como a presença de uma condição clínica influencia a probabilidade de outra (PAULA et al., 2022). Contudo, em cenários com muitos atributos, o desempenho do DBR pode ser superado por métodos que estruturam melhor essa dependência, especialmente na métrica de *subset accuracy*.

Uma evolução natural nesse sentido é denominada *Classifier Chains* (CC). O método busca superar a negligência do BR em relação à dependência entre rótulos treinando os classificadores de forma sequencial, criando uma "cadeia" (GONÇALVES, 2018), conforme ilustrado na Figura 4.

Figura 4 – Estrutura de Classifier Chains para classificação multirrótulo.



Fonte**:** Adaptado de Chen et al. (2020).

Durante o treinamento de cada elo, o classificador recebe os atributos originais somados às predições dos rótulos anteriores (GONÇALVES, 2013). Essa abordagem modela dependências de alto nível mantendo a eficiência computacional próxima ao BR. No domínio da saúde, isso significa que a predição de uma DCNT, como hipertensão, pode auxiliar estatisticamente na predição de uma condição correlata, como a insuficiência cardíaca (MORETTIN; SINGER, 2022). Vale ressaltar que a ordem dos rótulos na cadeia pode influenciar o desempenho, motivando o uso de otimizações via algoritmos genéticos (GONÇALVES, 2013).

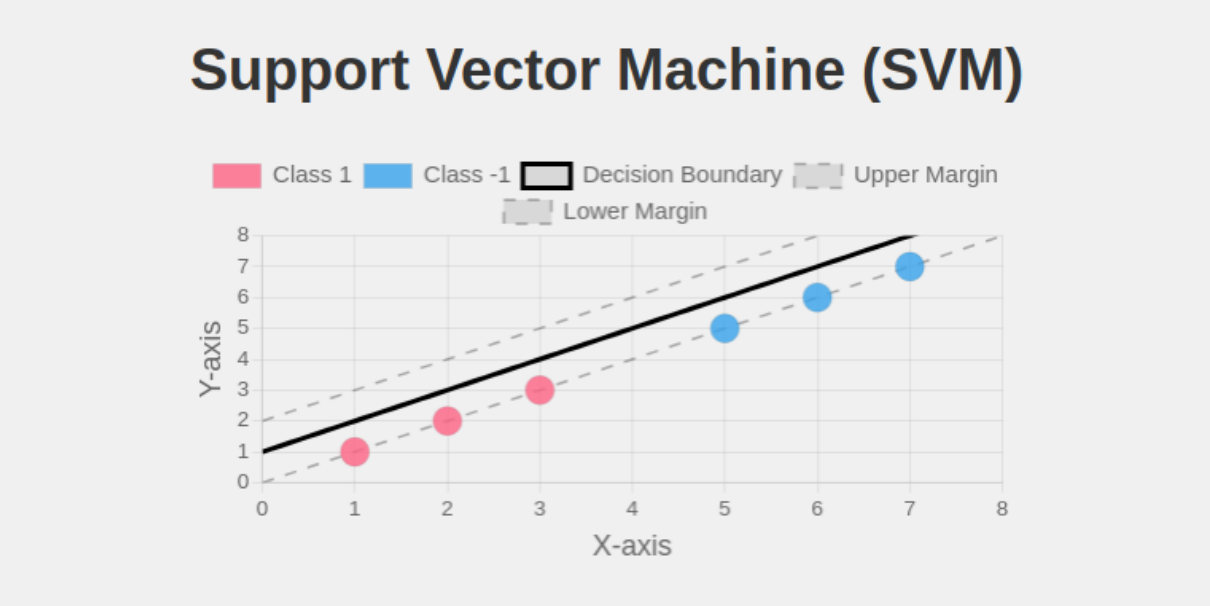
Por fim, no estado da arte dessas transformações, situa-se o *Meta Binary Relevance* (MBR). Baseado na técnica de *stacking* (generalização empilhada), o MBR visa aprimorar o desempenho combinando múltiplos modelos em duas camadas de aprendizado (GÉRON, 2019; RIVOLLI; CARVALHO, 2018). Na primeira fase, aplica-se o BR tradicional para gerar predições iniciais; na segunda, um meta-aprendiz (ou *blender*) utiliza essas predições como entrada para realizar a classificação final. Diferente do BR básico, a meta-camada do MBR consegue identificar e capturar padrões de coocorrência entre os rótulos, refinando o resultado (GÉRON, 2019). Essa abordagem de *ensemble* hierárquico é vantajosa na predição de multimorbidades pois aprende associações complexas a partir dos erros e acertos da primeira camada, sem a rigidez de uma ordem sequencial predefinida, como ocorre nas cadeias de classificadores (PAULA et al., 2022).

2.3.2 Classificadores Base

Na arquitetura da CMR, as estratégias de transformação atuam como uma camada de adaptação, mas o aprendizado efetivo dos padrões depende, em última instância, de algoritmos indutores binários ou multiclasse. Dentre as opções mais robustas e consolidadas na literatura para atuar como classificadores base, destacam-se a SVM e o RF.

O SVM destaca-se como um modelo versátil, capaz de realizar classificações lineares e não lineares, apresentando desempenho particularmente eficaz em conjuntos de dados complexos de pequeno e médio porte (GÉRON, 2019). Seu princípio fundamental é de natureza geométrica, uma vez que o algoritmo busca determinar um hiperplano separador que maximize a margem entre classes distintas. Conforme ilustrado na Figura 5, essa fronteira de decisão é definida de modo a manter a maior distância possível entre as instâncias pertencentes a diferentes classes, sendo denominados vetores de suporte os pontos localizados sobre os limites da margem (MORETTIN; SINGER, 2022).

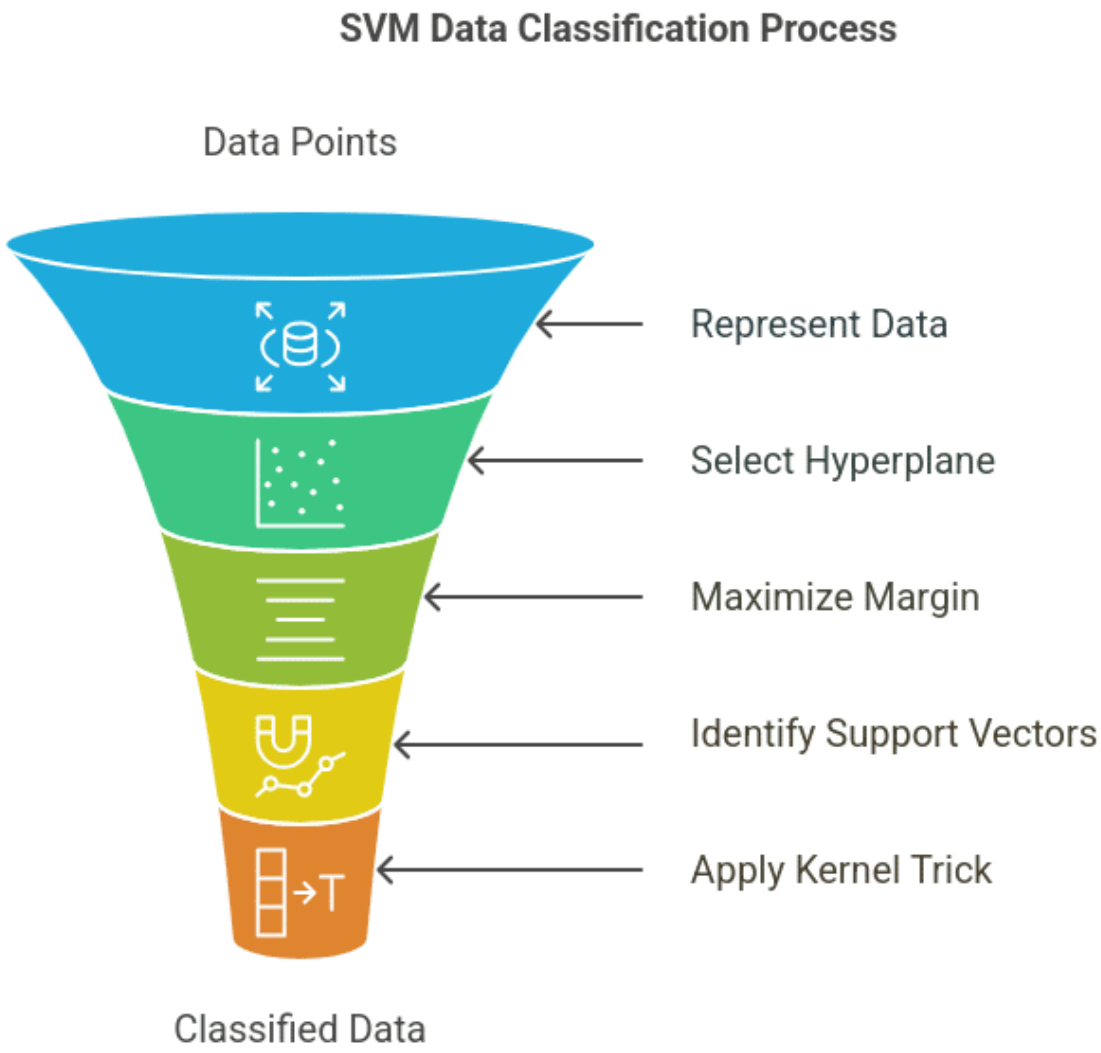
Figura 5 - Imagem com hiperplano, margens e vetores de suporte



Fonte: MyGreatLearning (2023).

Em situações nas quais os dados não são linearmente separáveis; cenário frequente em aplicações na área da saúde, a SVM recorre ao denominado kernel trick. Essa técnica permite o mapeamento dos dados do espaço original para um espaço de maior dimensão, no qual a separação linear se torna matematicamente viável (GÉRON, 2019). De forma complementar, a Figura 6 apresenta uma representação ilustrativa do processo de classificação do SVM, destacando conceitualmente a aplicação do kernel, a identificação dos vetores de suporte e a maximização da margem. Dentre os kernels mais utilizados na literatura, destacam-se o polinomial e o radial (Radial Basis Function – RBF) (DEISENROTH et al., 2020).

Figura 6 – Representação ilustrativa do processo de classificação do SVM

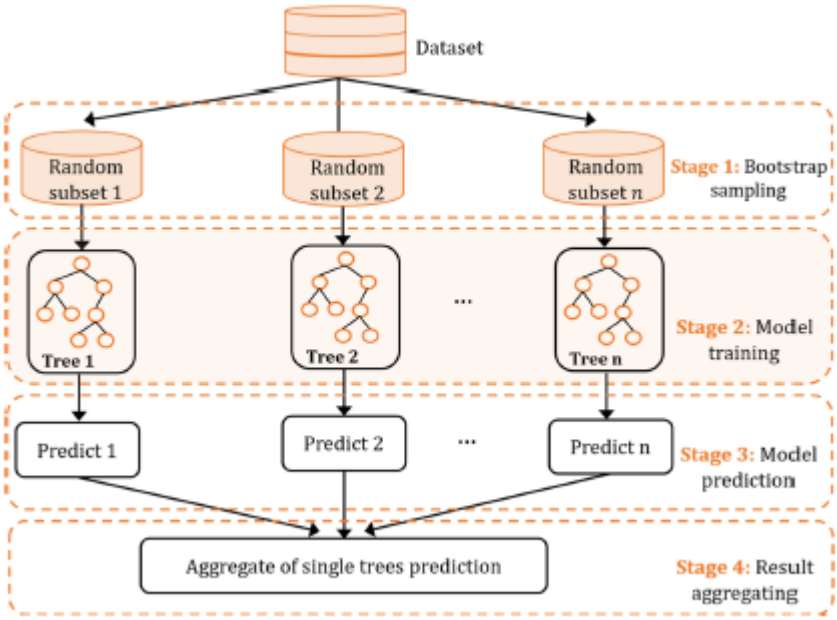


Fonte: MyGreatLearning (2023).

Cabe ressaltar que a SVM é sensível à escala das variáveis, tornando a normalização dos dados uma etapa indispensável no pré-processamento. Tal procedimento evita que atributos com magnitudes elevadas exerçam influência desproporcional sobre o modelo (GÉRON, 2019), sendo especialmente relevante em estudos envolvendo populações idosas, nos quais variáveis como idade e exames laboratoriais apresentam escalas numéricas significativamente distintas (MORETTIN; SINGER, 2022).

Já o RF trata-se de uma técnica de aprendizado por conjunto (ensemble learning) composta por múltiplas árvores treinadas de forma independente (GÉRON, 2019). Conforme ilustrado na Figura 7, o funcionamento do algoritmo baseia-se no método de bagging (Bootstrap Aggregating), no qual subconjuntos aleatórios do conjunto de dados são gerados para o treinamento de cada árvore. Adicionalmente, a RF introduz aleatoriedade na seleção das variáveis preditoras em cada nó da árvore (IZBICKI; SANTOS, 2020). Essa diversidade estrutural contribui para a redução da variância do modelo final, minimizando o problema de overfitting frequentemente observado em árvores de decisão individuais (IZBICKI; SANTOS, 2020).

Figura 7 – Funcionamento do algoritmo Random Forest



Fonte: Gobi, Haghnejad e Gharibzadeh (2022).

O resultado do RF é obtido por meio do voto majoritário das árvores (GÉRON, 2019). Além da alta precisão, suas principais vantagens incluem a robustez a ruídos e valores atípicos, bem como a capacidade de lidar automaticamente com interações complexas entre variáveis sem exigir modelagem manual (MORETTIN; SINGER, 2022).

No estudo da saúde do idoso, essa característica é fundamental para lidar com a multimorbidade, pois o algoritmo permite identificar a importância relativa de diferentes preditores, como IMC, pressão arterial, sexo e idade, na ocorrência simultânea de doenças crônicas. Essa capacidade de ranqueamento de variáveis auxilia na compreensão de padrões complexos que métodos tradicionais podem negligenciar (MORETTIN; SINGER, 2022).

2.3.3 Métricas de Avaliação

A avaliação de desempenho na CMR apresenta uma complexidade superior à da classificação tradicional. Enquanto no cenário monorrótulo um erro é absoluto, na CMR as predições podem estar parcialmente corretas: o modelo pode identificar acertadamente algumas doenças de um paciente e falhar em outras. Para capturar essa nuance, é necessário estabelecer formalmente a notação do problema. Considera-se um conjunto de dados D composto por N instâncias. Para cada indivíduo i, define-se Y_i como o conjunto de rótulos reais (doenças observadas) e Z_i como o conjunto de rótulos preditos pelo classificador. Alternativamente, essas classes podem ser representadas por vetores binários, onde o valor 1 indica a presença da condição clínica e 0 a sua ausência (PAULA et al., 2022; GONÇALVES, 2018).

A métrica mais intuitiva para lidar com esse cenário é a Acurácia Multirrótulo (*Accuracy*). Diferente da versão tradicional, ela é baseada em exemplos e calculada pela razão entre a interseção (acertos) e a união (total de rótulos ativos) dos conjuntos real e predito para cada instância, conforme a equação:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}$$

Trata-se de uma métrica de "sentido positivo" (quanto maior, melhor). Sua principal vantagem é penalizar menos severamente os acertos parciais, embora possa ser otimista em bases desbalanceadas se o modelo tender a prever apenas os rótulos mais frequentes (PAULA et al., 2022).

Em contrapartida, existe a métrica de *Subset Accuracy*, que representa o critério mais rigoroso da literatura:

$$SubsetAccuracy = \frac{1}{N}\sum_{i=1}^{N}[[Y_i = Z_i]]$$

Nesta equação, o símbolo [[···]] (colchete de Iverson) retorna 1 apenas se o conjunto predito Z_i for exatamente igual ao real Y_i. Por ignorar completamente os acertos parciais, a *Subset Accuracy* tende a apresentar valores baixos em cenários clínicos complexos, servindo mais como um indicador de perfeição do que de utilidade prática (GONÇALVES, 2018).

Para avaliar a taxa de erro, utiliza-se o *Hamming Loss*. Está métrica calcula a proporção média de rótulos classificados incorretamente, considerando tanto as omissões (falsos negativos) quanto as inclusões indevidas (falsos positivos):

$$HammingLoss = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \Delta Z_i|}{q}$$

Onde Δ representa a diferença simétrica entre os conjuntos e q é o número total de rótulos. Por ser uma métrica de perda, valores próximos de 0 indicam melhor desempenho. Apesar de ser a medida mais comum na literatura, ela possui a limitação de não diferenciar casos com muitos acertos daqueles com poucos, caso o número total de erros binários seja idêntico (GONÇALVES, 2018).

Quando se busca um equilíbrio entre a sensibilidade e a precisão do modelo, aplica-se a *F1-Measure* baseada em exemplos. Está métrica é a média harmônica entre *Precision* (quantos rótulos preditos são reais) ($\frac{|Y_i \cap Z_i|}{|Z_i|}$) e *Recall* (quantos rótulos reais foram capturados) ($\frac{|Y_i \cap Z_i|}{|Y_i|}$):

$$F1 = \frac{1}{N}\sum_{i=1}^{N} 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

A *F1-Measure* é especialmente robusta para lidar com a coocorrência de múltiplos rótulos por instância, sendo vital em problemas de saúde onde a omissão de um diagnóstico é crítica (PAULA et al., 2022).

Contudo, todas as métricas citadas até agora focam no desempenho por paciente (exemplo). No contexto específico de idosos e DCNTs, a escolha das métricas deve ser estratégica. Estudos indicam que classificadores como RF podem apresentar vantagens, dependendo do conjunto de dados e da configuração adotada, em comparação ao SVM em métricas baseadas em exemplos, como a *Accuracy* e a *F1-Measure*. Recomenda-se monitorar o *Hamming Loss* como medida de erro global por rótulo (quanto menor, melhor), pois ela quantifica a proporção média de classificações incorretas (PAULA et al., 2022; GONÇALVES, 2018).

3. METODOLOGIA

Este capítulo detalha o caminho metodológico adotado para a predição de multimorbidade, compreendendo desde o delineamento do estudo e a seleção da base de dados até o tratamento de variáveis e o protocolo de avaliação dos modelos. O fluxo de trabalho foi estruturado com o objetivo central de assegurar a rastreabilidade das decisões de pré-processamento, garantir a compatibilidade das variáveis com a literatura de referência e permitir a reprodutibilidade integral dos experimentos computacionais.

3.1. BASE DE DADOS E DELINEAMENTO DO ESTUDO

O estudo utiliza os dados da segunda onda do ELSI-Brasil, conduzida entre 2019 e 2021. Trata-se de uma pesquisa de base domiciliar com amostra representativa da população brasileira com 50 anos ou mais, que disponibiliza um amplo espectro de informações sociodemográficas, comportamentais e medidas clínicas. Embora o ELSI-Brasil seja longitudinal, este trabalho adota um delineamento transversal (recorte seccional da segunda onda), focado no desenvolvimento de modelos preditivos estáticos.

O problema foi formulado como uma tarefa de CMR, onde o objetivo é estimar a ocorrência simultânea de múltiplas DCNTs para um mesmo indivíduo. Cada doença foi tratada como um rótulo (*label*) binário, compondo um vetor que representa o perfil de multimorbidade do participante.

A amostra bruta inicial da segunda onda contava com 9.949 participantes. Para a constituição da base analítica, aplicou-se um critério de qualidade dos dados: foram mantidos apenas os registros com informação completa para todos os sete rótulos de doenças selecionados (detalhados na seção 3.2). Esse processo de limpeza resultou na exclusão de 332 registros (aproximadamente 3,34% da amostra) por ausência de resposta em pelo menos uma das condições alvo. Dessa forma, a base final para modelagem totalizou 9.617 instâncias válidas, garantindo a consistência dos vetores de saída para o treinamento dos algoritmos supervisionados.

3.2 SELEÇÃO DE VARIÁVEIS E DEFINIÇÃO DOS RÓTULOS

Para orientar a construção do espaço de atributos, este trabalho adotou como referência metodológica o estudo de Paula et al. (2022), aplicado ao contexto do ELSA-Brasil. Seguindo essa diretriz, a seleção de variáveis foi estruturada em dois eixos: a definição dos desfechos (as

doenças a serem preditas) e a escolha dos preditores (características sociodemográficas, comportamentais e clínicas), buscando reduzir arbitrariedades e garantir comparabilidade com a literatura de multimorbidade.

No que se refere às variáveis de desfecho (*labels*), realizou-se uma varredura sistemática no bloco das DCNTs do questionário individual. Para evitar a instabilidade estatística, estabeleceu-se um critério de corte baseado na frequência: apenas condições com prevalência igual ou superior a 5% na amostra foram mantidas para a etapa de modelagem.

A aplicação desse critério resultou na seleção de sete doenças crônica: Hipertensão, Problemas Crônicos de Coluna, Colesterol Alto, Artrite/Reumatismo, Diabetes, Osteoporose e Depressão. Condições com prevalência inferior ao limiar estabelecido, como AVC (4,44%), Insuficiência Cardíaca (4,31%) e Câncer (4,25%), foram excluídas do vetor alvo final, conforme detalhado na Tabela 1.

Tabela 1 - Prevalência das condições crônicas avaliadas para definição dos rótulos do estudo.

Identificador	Doença	Prevalência (%)
n28	Hipertensão arterial (pressão alta)	51,85%
n58	Problema crônico de coluna (dor nas costas, pescoço, lombalgia, ciática, vértebras ou disco)	31,91%
n44	Colesterol alto	22,64%
n56	Artrite ou reumatismo	19,89%
n35	Diabetes (açúcar no sangue)	18,09%
n57	Osteoporose	14,08%
n59	Depressão	12,80%
n52	Acidente vascular cerebral (derrame)	4,44%
n50	Insuficiência cardíaca	4,31%
n46	Infarto do coração	4,28%
n60	Câncer	4,25%
n55	Enfisema, bronquite crônica ou DPOC	3,46%
n59_2	Problema psiquiátrico (exclui depressão)	3,10%
n54	Asma	3,08%
n66	Ponte de safena / stent / angioplastia	2,86%
n63_2	Problema sério de memória ou demência (exclui Alzheimer)	2,81%
n61	Insuficiência renal crônica	2,34%
n48	Angina	1,91%
n63	Alzheimer	1,48%
n62	Parkinson	0,94%

Fonte: Autoria própria (2026).

Paralelamente, a seleção das variáveis preditoras (*features*) buscou espelhar os domínios conceituais de risco, integrando dados dos questionários individual e domiciliar. Foram mapeados itens correspondentes a três grandes grupos:

I. Sociodemográficos (sexo, idade, escolaridade, raça/cor, estado civil e renda familiar);
II. Estilo de Vida (tabagismo, consumo de álcool, atividade física, hábitos alimentares e comportamento sedentário);

III. Antropometria e Clínica (medidas de peso, altura, cintura, quadril e pressão arterial);

A extração desses dados envolveu a padronização dos identificadores originais do ELSI-Brasil para nomenclaturas legíveis, assegurando a rastreabilidade das informações que alimentariam a etapa subsequente do tratamento das variáveis.

3.3 TRATAMENTO E PREPARAÇÃO DOS DADOS

A preparação dos dados foi conduzida integralmente na linguagem R, utilizando os pacotes *tidyverse* e *mldr*, com o objetivo de padronizar codificações, tratar inconsistências e derivar medidas clinicamente relevantes a partir das variáveis brutas. O fluxo de pré-processamento resultou em uma base analítica otimizada para algoritmos de AM.

3.3.1 Tratamento de valores

A base do ELSI-Brasil utiliza codificações específicas para indicar ausência de informação ou situações especiais ('não sabe', 'recusou' ou 'não se aplica'), representadas por valores como 9, 99, 888 e -1. Nesta etapa, realizou-se a padronização global desses registros, convertendo todas as ocorrências de não resposta em valores ausentes (*missing values* - NA).

3.3.2 Variáveis sociodemográficos

As variáveis sociodemográficas foram tratadas de acordo com o tipo de dado e sua adequação aos métodos de modelagem. A variável sexo foi mantida como binária, seguindo o padrão original da base (Sexo: 1 = homem | 0 = mulher). A idade foi preservada como variável contínua, em anos, sem discretização, com o objetivo de reter o máximo de informação possível. A renda familiar foi mantida conforme sua codificação original, respeitando sua natureza quantitativa.

As variáveis escolaridade e escolaridade materna, que apresentam múltiplas categorias no questionário, foram recodificadas por meio de uma codificação ordinal em quatro níveis: 0 para sem instrução ou ensino fundamental incompleto; 1 para ensino fundamental completo; 2 para ensino médio completo; e 3 para ensino superior ou mais. Valores inválidos ou correspondentes a não resposta foram tratados como NA. Essa recodificação reduz a esparsidade dos dados e preserva um ordenamento educacional coerente, evitando a fragmentação excessiva de categorias.

As variáveis raça/cor e estado civil, de natureza categórica nominal, foram transformadas por meio de codificação dummy, gerando variáveis binárias para cada categoria. Esse procedimento é necessário, uma vez que a maioria dos algoritmos de AM não processa diretamente categorias textuais e, adicionalmente, evita a imposição de uma ordem artificial entre classes nominais. Para raça/cor, foram criadas variáveis como `Raca_Branca`, `Raca_Preta`, `Raca_Parda`, `Raca_Amarela` e `Raca_Indigena`. Para estado civil, foram geradas variáveis como `EC_Solteiro`, `EC_Casado`, `EC_Divorciado` e `EC_Viuvo`. Após a criação das variáveis dummy, as colunas originais foram removidas para evitar redundância.

Por fim, a variável número de filhos foi binarizada para indicar apenas a presença ou ausência de filhos, sendo codificada como 0 para indivíduos sem filhos e 1 para aqueles com pelo menos um filho. Essa estratégia reduz o espaço de estados e captura um aspecto familiar básico do indivíduo, alinhado ao foco preditivo do estudo.

3.3.3 Variáveis de estilo de vida e comportamento

As variáveis comportamentais foram binarizadas ou discretizadas em faixas interpretáveis, de acordo com sua natureza e relevância analítica. Para tabagismo e consumo de álcool, foram criadas ou ajustadas variáveis binárias que indicam tabagismo atual, tabagismo passado e consumo de álcool. Códigos correspondentes a não resposta foram tratados como NA.

A variável de atividade física vigorosa foi binarizada, sendo recodificada como 1 quando o indivíduo relatou a prática desse tipo de atividade e como 0 nos demais casos, mantendo-se NA para situações de não resposta.

Variáveis como caminhada, sono e consumo alimentar, originalmente numéricas ou com múltiplas categorias, foram agrupadas por meio de discretização em faixas, com o objetivo de reduzir ruído e facilitar a interpretação. O número de dias de caminhada foi categorizado em três níveis: 0 para indivíduos que não caminham; 1 para aqueles que caminham de 1 a 3 dias por semana; e 2 para os que caminham de 4 a 7 dias por semana. O consumo de frutas e vegetais foi organizado de forma análoga, com as categorias 0 para não consumo, 1 para consumo de 1 a 3 dias por semana e 2 para consumo de 4 a 7 dias por semana.

A qualidade do sono foi recodificada em uma escala invertida, de modo a garantir consistência interpretativa, com valores mais elevados indicando melhor condição quando aplicável. A variável dificuldade para dormir foi transformada em uma escala ordinal baseada na frequência do evento.

A lógica geral desses agrupamentos foi padronizar a interpretação das variáveis, de forma que, sempre que possível, valores mais altos representassem condições mais favoráveis ou maior intensidade, reduzindo ambiguidades durante o treinamento dos modelos.

3.3.4 Variáveis antropométricas e clínicas

O Índice de Massa Corporal (IMC) é um indicador amplamente utilizado para avaliar a relação entre peso corporal e estatura, sendo definido matematicamente pela seguinte expressão:

$$IMC = \frac{peso\ (kg)}{altura\ (m)^2}$$

Antes do cálculo, as variáveis peso e altura passaram por um processo de limpeza que incluiu o tratamento de códigos especiais, como 666, 777, 888 e 999, a correção de escalas numéricas inconsistentes e a exclusão de valores fora de faixas plausíveis. Além do IMC como variável contínua, foi criada a variável binária `IMC_Classificacao`, que identifica indivíduos com IMC dentro da faixa considerada ideal, entre 18,5 e 24,9, segundo recomendação da OMS, em contraste com aqueles fora desse intervalo.

As pressões arteriais sistólica e diastólica foram mantidas como variáveis contínuas após a limpeza dos dados, com a conversão de códigos de não resposta, como 777 e 999, para NA. Adicionalmente, foi derivada a variável ordinal `PA_Classificacao`, com três níveis clínicos: 0 para pressão normal, definida por pressão arterial sistólica inferior a 120 mmHg e pressão arterial diastólica inferior a 80 mmHg; 1 para condição limítrofe, caracterizada por pressão sistólica entre 120 e 139 mmHg ou pressão diastólica entre 80 e 89 mmHg; e 2 para hipertensão, definida por pressão sistólica igual ou superior a 140 mmHg ou pressão diastólica igual ou superior a 90 mmHg. Essa categorização complementa a informação dos valores contínuos, incorporando critérios clínicos amplamente utilizados.

A razão cintura–quadril foi calculada de acordo com a fórmula:

$$RCQ = \frac{cintura}{quadril}$$

Antes do cálculo, as medidas passaram por limpeza prévia, incluindo o tratamento de códigos como 888 e 999 e a exclusão de valores implausíveis. A partir da RCQ, foi criada a variável binária `Risco_RCQ`, considerando pontos de corte específicos por sexo, com RCQ >

0,90 para homens e RCQ > 0,80 para mulheres, em consonância com recomendações consolidadas para a avaliação de obesidade abdominal e do risco cardiometabólico.

3.3.5 Binarização das DCNTs e consistência dos rótulos

Cada DCNTs selecionada foi convertida para rótulo binário (0/1), com tratamento de não resposta (NA). No caso de hipertensão e diabetes, respostas "sim, apenas durante a gravidez" foram tratadas como NA para evitar mistura de critérios diagnósticos não equivalentes ao objetivo do estudo.

Após a padronização, condições crônicas que não fariam parte do conjunto final de rótulos (por exemplo: infarto, angina, insuficiência cardíaca, AVC, asma, câncer, doença respiratória, entre outras) foram removidas da base final de modelagem.

3.4 FORMULAÇÃO DO PROBLEMA E ESTRATÉGIAS DE CLASSIFICAÇÃO MULTIRRÓTULO

Nesta subseção, a causa da multimorbidade é transformada para o paradigma da CMR. Diferente da abordagem monorrótulo tradicional, que limita a análise a um único diagnóstico por vez, a CMR permite que o estado de saúde de um idoso seja representado como um perfil clínico integrado de diagnósticos simultâneos.

Para formular essa análise, cada indivíduo *i* extraído da base de dados ELSI-Brasil é descrito matematicamente por dois componentes fundamentais:

- Vetor de Características $\boldsymbol{x}_i \in \mathbb{R}^p$: Representa o perfil de entrada do participante. Onde o símbolo $\mathbb{R}$ indica que os dados são processados como valores numéricos, enquanto p representa a quantidade total de atributos preditores utilizados, como idade, IMC, escolaridade e histórico familiar.
- Vetor de Rótulos $\boldsymbol{y}_i \in \{0,1\}^L$: Representa o desfecho clínico para $L = 7$ DCNTs selecionadas. O conjunto $\{0,1\}$ define o caráter binário da resposta, onde 1 indica a presença da patologia e 0 a sua ausência.

Dessa forma, o objetivo central do aprendizado de máquina é aprender uma função preditiva $f: \mathbb{R}^p \longrightarrow \{0,1\}^L$ capaz de predizer simultaneamente quais doenças o indivíduo apresenta. Esta formulação é considerada mais informativa do que a classificação binária convencional "presença de multimorbidade (sim/não)", pois preserva explicitamente as combinações e as correlações existentes entre as patologias. A partir desse detalhamento, a

condição de multimorbidade pode ser derivada por meio de um critério de contagem, identificando-se o indivíduo com duas ou mais doenças através da condição:

$$\sum_{j=1}^{L} y_{ij} \geq 2$$

Foram avaliadas estratégias clássicas de transformação para problemas multirrótulo, conforme implementadas no pacote *utiml*. A abordagem BR treina um classificador independente para cada rótulo. O método DBR incorpora dependências entre rótulos por meio de mecanismos próprios da estratégia. O CC modela explicitamente a dependência entre rótulos ao encadear as predições de rótulos anteriores como entradas para os subsequentes. Já o MBR consiste em uma abordagem meta destinada a aprimorar o desempenho da predição multirrótulo.

A escolha do SVM e do RF como classificadores base fundamenta-se nas evidências apresentadas por Paula et al. (2022) no estudo ELSA-Brasil. Os autores justificam a seleção desses algoritmos pela sua eficácia em lidar com problemas de predição complexos e grandes volumes de dados.

O RF foi escolhido em função de sua robustez a não linearidades, interações entre variáveis e multicolinearidade, além de apresentar bom desempenho reportado na literatura em estudos sobre multimorbidade (Paula et al., 2022). O SVM foi empregado como modelo comparativo, dada sua capacidade de separação em espaços de alta dimensionalidade.

Nos experimentos, o RF foi configurado com 100 árvores (ntree $= 100$), explorando sua robustez a relações não lineares e à multicolinearidade, conforme descrito por Paula et al. (2022). O SVM foi utilizado como alternativa comparativa, mantendo seu papel de referência para cenários de alta dimensionalidade.

3.5 DELINEAMENTO EXPERIMENTAL PARA MÉTRICAS DE AVALIAÇÃO

Para a execução do experimento, a base de dados final foi inicialmente separada em duas estruturas distintas: a matriz de rótulos (Y), contendo as 7 DCNTs selecionadas como colunas binárias (0/1), como mostrado na Figura 8 a seguir, e a matriz de preditores (X), contendo as variáveis sociodemográficas, comportamentais e antropométricas/clínicas processadas. Essa separação é essencial para definir o caráter multirrótulo do problema e garantir que não haja vazamento de informação (*data leakage*) entre as entradas e as saídas durante o treinamento.

Figura 8 - Matriz de rótulos (Y) no RStudio

	Hipertensao	Diabetes	Colesterol Alto	Artrite	Depressao	Osteoporose	Problema cronico coluna
1	1	1	0	0	1	0	0
2	1	0	0	1	0	1	0
3	0	0	1	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	1	0	1	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	1
9	0	1	0	0	0	0	0
10	0	1	0	0	0	0	0
11	1	0	0	0	0	1	0
12	0	0	0	0	0	0	1
13	0	0	0	1	0	0	0
14	1	0	1	0	0	0	1

Showing 1 to 14 of 9,617 entries, 7 total columns

Fonte: Autoria própria (2026).

Com o objetivo de realizar uma comparação robusta entre os modelos, adotou-se o método de validação cruzada k-fold com $K = 10$, repetido 5 vezes, totalizando 50 avaliações por modelo. Em cada repetição, novas partições dos dados foram geradas e, em cada fold, o modelo foi ajustado no conjunto de treino e avaliado no conjunto de teste. Para garantir segurança computacional e reprodutibilidade, os resultados foram salvos de forma incremental em arquivo, permitindo a retomada do processo em caso de interrupções e evitando a recomputação de etapas já concluídas.

A avaliação de desempenho seguiu métricas consolidadas na literatura CMR, selecionadas para capturar diferentes dimensões do problema. Foram utilizadas a Hamming Loss, que quantifica a fração média de rótulos incorretamente preditos, sendo valores menores indicativos de melhor desempenho; a Subset Accuracy, métrica mais restritiva que exige o acerto exato de todo o vetor multirrótulo; a Accuracy multirrótulo, como medida agregada de desempenho; o *F1-Measure*, que combina precisão e revocação considerando o total de decisões realizadas. A escolha desse conjunto de métricas busca equilibrar medidas mais estritas, métricas tolerantes a acertos parciais e métricas sensíveis ao desequilíbrio entre classes.

De forma complementar à etapa de modelagem e com o intuito de subsidiar a interpretação dos resultados, foi conduzida uma Análise Exploratória de Dados (EDA) com três objetivos principais: caracterizar as prevalências individuais das DCNTs, descrever a distribuição do número de doenças por indivíduo e explorar padrões de coocorrência entre as condições. Para isso, foram elaborados gráficos de prevalência, boxplots da distribuição do

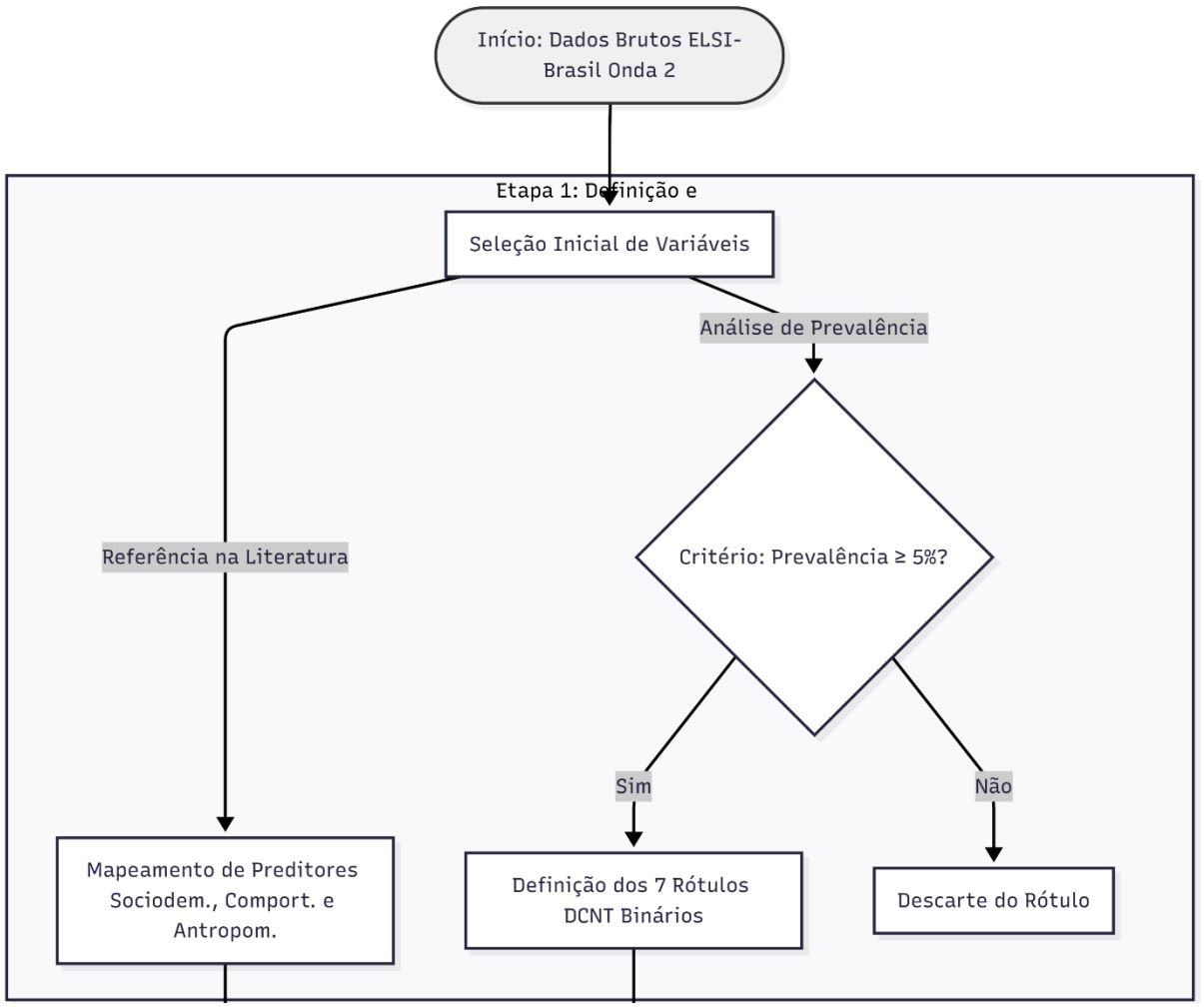
número de doenças, listagens das combinações de rótulos mais frequentes, matrizes de correlação de Pearson entre os rótulos e heatmaps condicionais, que representam a prevalência de comorbidades dado que o indivíduo apresenta uma condição de base, permitindo capturar a assimetria característica dessas relações. Adicionalmente, foi construída uma tabela de caracterização demográfica e de saúde, na qual os indivíduos foram estratificados de acordo com o número de doenças crônicas (0, 1, 2, 3 e 4 ou mais), contextualizando o fenômeno da multimorbidade na amostra analisada.

3.6 FLUXOGRAMA METODOLÓGICO

Para permitir a visualização detalhada de todas as etapas do processo, o fluxograma metodológico foi segmentado em três partes complementares. Essa divisão visa garantir a legibilidade das decisões tomadas em cada fase do *pipeline* do tratamento dos dados.

A Figura 9 ilustra a etapa inicial de seleção de variáveis e definição dos critérios de inclusão das doenças baseadas na prevalência.

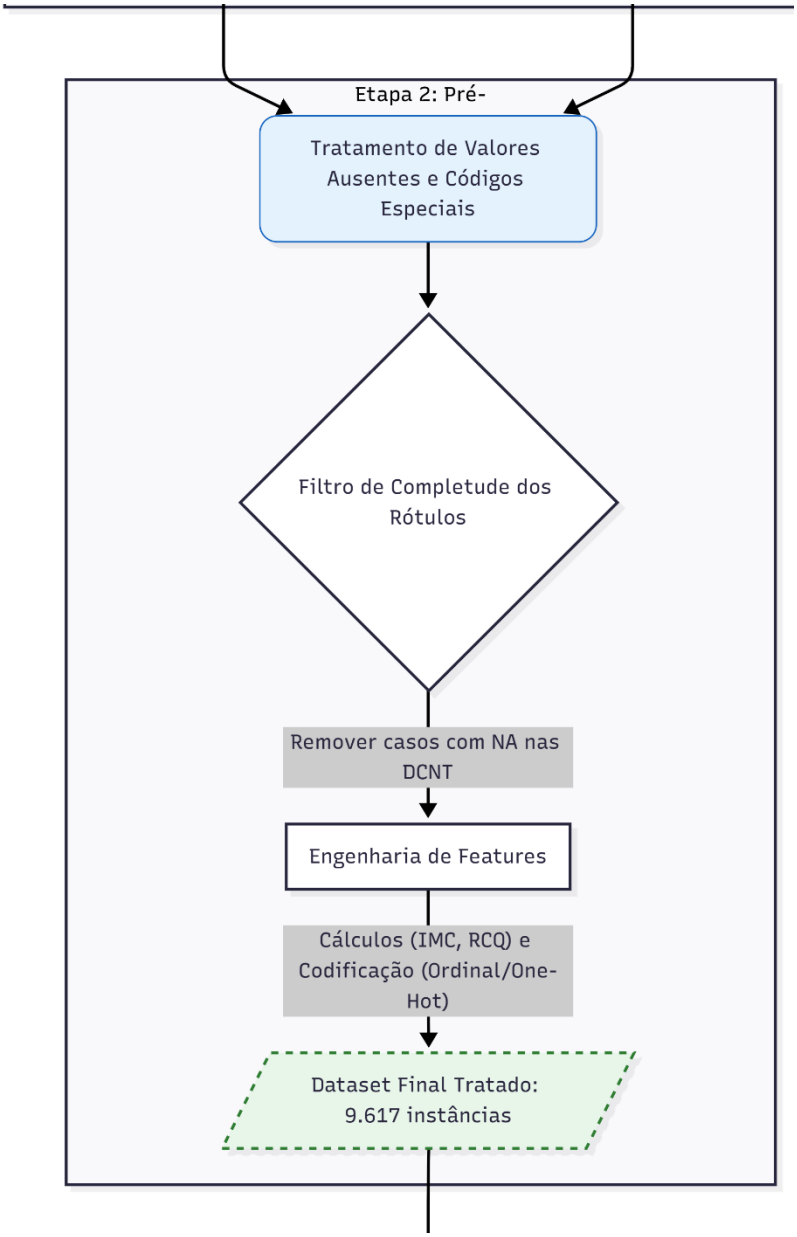
Figura 9 - Fluxo de seleção de variáveis e definição dos rótulos



Fonte: Autoria própria (2026).

A Figura 10 detalha o fluxo de pré-processamento, incluindo o tratamento de códigos especiais, o filtro de completude e a engenharia de atributos (*features*).

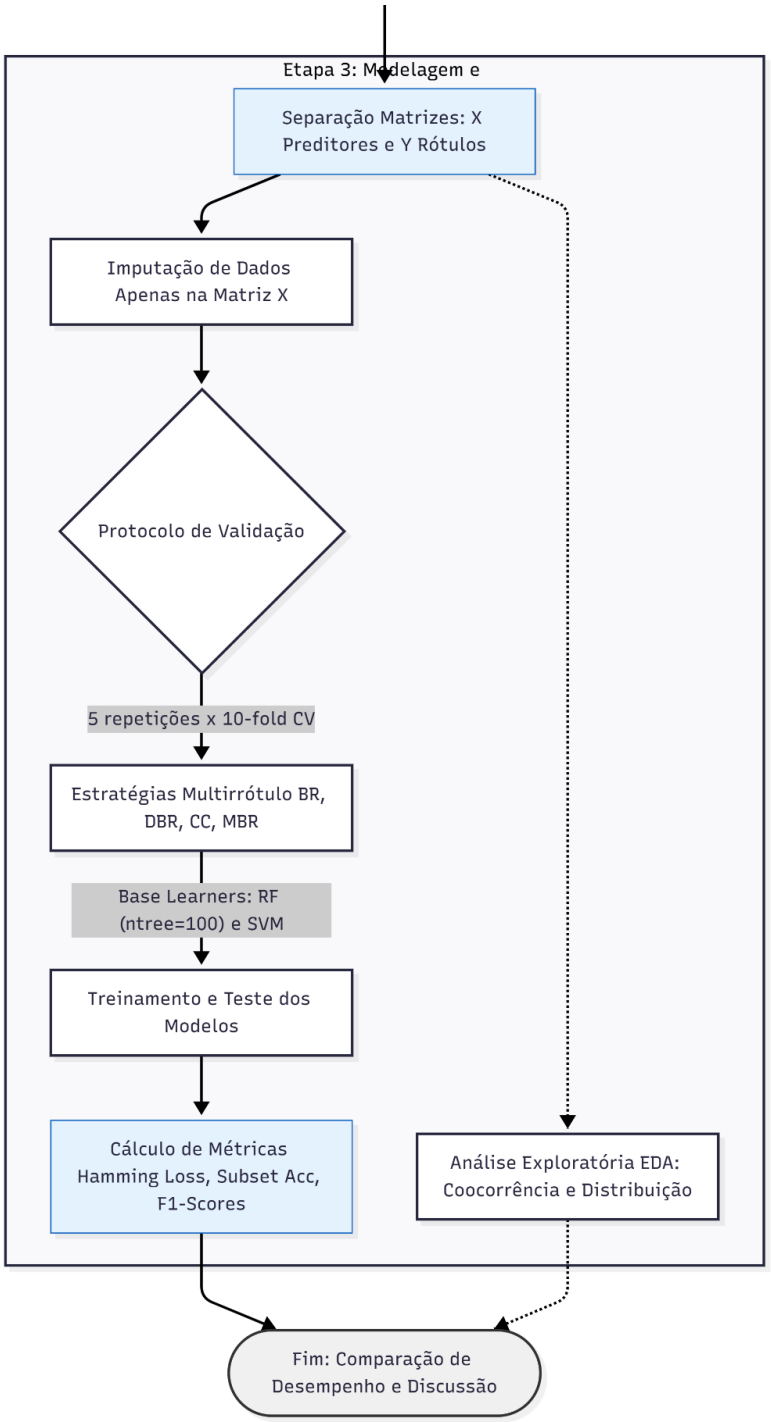
Figura 10 - Fluxo de tratamento de dados e engenharia de variáveis.



Fonte: Autoria própria (2026).

Por fim, a Figura 11 esquematiza o protocolo experimental de modelagem, desde a separação das matrizes e imputação de dados até o treinamento dos algoritmos multirrótulo e a avaliação por métricas de desempenho.

Figura 11 - Fluxo de modelagem, validação e avaliação.



Fonte: Autoria própria (2026).

4. RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados fundamentais deste trabalho, organizados para detalhar desde o perfil dos participantes até o desempenho técnico das combinações de estratégias de transformação (BR, CC, DBR e MBR) com os classificadores base (SVM e RF) que discutimos anteriormente. Inicialmente, descreve-se a caracterização da amostra e os padrões de prevalência e coocorrência das doenças no cenário estudado.

A Tabela 2 mostra as características sociodemográficas, comportamentais e clínico-antropométricas pelo número de doenças crônicas autorreferidas (0 a 4+). Essa visão permite identificar como as variáveis basais se comportam diante do aumento da carga de doença.

Tabela 2 - Características sociodemográficas e de saúde segundo número de doenças crônicas.

Características sociodemográficas e de saúde segundo número de condições crônicas.

	Total (N=9617)	0 Condições (N=2261)	1 Condição (N=2701)	2 Condições (N=2079)	3 Condições (N=1377)	4+ Condições (N=1199)
Idade (anos) (média (SD))						
Mean (SD)	66.3 (10.0)	64.1 (9.79)	66.0 (10.1)	66.9 (9.99)	67.6 (9.82)	68.4 (9.69)
Median [Min, Max]	65.0 [50.0, 109]	62.0 [50.0, 101]	64.0 [50.0, 109]	66.0 [50.0, 99.0]	67.0 [50.0, 109]	67.0 [50.0, 98.0]
Sexo						
Feminino	5692 (59.2%)	1037 (45.9%)	1450 (53.7%)	1285 (61.8%)	938 (68.1%)	982 (81.9%)
Masculino	3925 (40.8%)	1224 (54.1%)	1251 (46.3%)	794 (38.2%)	439 (31.9%)	217 (18.1%)
Escolaridade						
Fund. Incompleto/Sem inst.	3563 (37.0%)	742 (32.8%)	1001 (37.1%)	822 (39.5%)	521 (37.8%)	477 (39.8%)
Fund. Completo	3597 (37.4%)	819 (36.2%)	991 (36.7%)	790 (38.0%)	525 (38.1%)	472 (39.4%)
Médio Completo	1688 (17.6%)	473 (20.9%)	475 (17.6%)	323 (15.5%)	221 (16.0%)	196 (16.3%)
Superior ou mais	655 (6.8%)	190 (8.4%)	196 (7.3%)	133 (6.4%)	93 (6.8%)	43 (3.6%)
Missing	114 (1.2%)	37 (1.6%)	38 (1.4%)	11 (0.5%)	17 (1.2%)	11 (0.9%)
Escolaridade da mãe						
Fund. Incompleto/Sem inst.	7250 (75.4%)	1651 (73.0%)	2027 (75.0%)	1599 (76.9%)	1029 (74.7%)	944 (78.7%)
Fund. Completo	1288 (13.4%)	320 (14.2%)	354 (13.1%)	268 (12.9%)	211 (15.3%)	135 (11.3%)
Médio Completo	140 (1.5%)	38 (1.7%)	46 (1.7%)	34 (1.6%)	12 (0.9%)	10 (0.8%)
Superior ou mais	40 (0.4%)	13 (0.6%)	11 (0.4%)	8 (0.4%)	6 (0.4%)	2 (0.2%)
Missing	899 (9.3%)	239 (10.6%)	263 (9.7%)	170 (8.2%)	119 (8.6%)	108 (9.0%)
Estado Civil						
Solteiro(a)	1156 (12.0%)	343 (15.2%)	335 (12.4%)	210 (10.1%)	144 (10.5%)	124 (10.3%)
Casado(a)	5137 (53.4%)	1191 (52.7%)	1504 (55.7%)	1146 (55.1%)	722 (52.4%)	574 (47.9%)
Divorciado(a)	1203 (12.5%)	349 (15.4%)	319 (11.8%)	240 (11.5%)	148 (10.7%)	147 (12.3%)
Viúvo(a)	2121 (22.1%)	378 (16.7%)	543 (20.1%)	483 (23.2%)	363 (26.4%)	354 (29.5%)
Raça/Cor						
Branca	4448 (46.3%)	1116 (49.4%)	1237 (45.8%)	966 (46.5%)	619 (45.0%)	510 (42.5%)
Preta	1024 (10.6%)	220 (9.7%)	283 (10.5%)	227 (10.9%)	130 (9.4%)	164 (13.7%)
Parda	4051 (42.1%)	903 (39.9%)	1148 (42.5%)	869 (41.8%)	616 (44.7%)	515 (43.0%)
Amarela	26 (0.3%)	9 (0.4%)	5 (0.2%)	6 (0.3%)	4 (0.3%)	2 (0.2%)
Indígena	36 (0.4%)	7 (0.3%)	11 (0.4%)	8 (0.4%)	6 (0.4%)	4 (0.3%)
Missing	32 (0.3%)	6 (0.3%)	17 (0.6%)	3 (0.1%)	2 (0.1%)	4 (0.3%)
Tem Filhos						
Não	771 (8.0%)	227 (10.0%)	226 (8.4%)	156 (7.5%)	94 (6.8%)	68 (5.7%)
Sim	8801 (91.5%)	2014 (89.1%)	2459 (91.0%)	1920 (92.4%)	1280 (93.0%)	1128 (94.1%)
Missing	45 (0.5%)	20 (0.9%)	16 (0.6%)	3 (0.1%)	3 (0.2%)	3 (0.3%)
Classificação IMC						
Fora do Ideal	5797 (60.3%)	1168 (51.7%)	1546 (57.2%)	1312 (63.1%)	942 (68.4%)	829 (69.1%)
Ideal	2303 (23.9%)	683 (30.2%)	685 (25.4%)	458 (22.0%)	270 (19.6%)	207 (17.3%)
Missing	1517 (15.8%)	410 (18.1%)	470 (17.4%)	309 (14.9%)	165 (12.0%)	163 (13.6%)
Risco Cintura/Quadril						
Sem Risco	787 (8.2%)	274 (12.1%)	239 (8.8%)	131 (6.3%)	84 (6.1%)	59 (4.9%)
Com Risco	6884 (71.6%)	1480 (65.5%)	1872 (69.3%)	1553 (74.7%)	1057 (76.8%)	922 (76.9%)
Missing	1946 (20.2%)	507 (22.4%)	590 (21.8%)	395 (19.0%)	236 (17.1%)	218 (18.2%)
Fumante Atual						
Não	8402 (87.4%)	1906 (84.3%)	2343 (86.7%)	1848 (88.9%)	1228 (89.2%)	1077 (89.8%)
Sim	1215 (12.6%)	355 (15.7%)	358 (13.3%)	231 (11.1%)	149 (10.8%)	122 (10.2%)
Fumou no passado						
Não	7645 (79.5%)	1908 (84.4%)	2140 (79.2%)	1637 (78.7%)	1041 (75.6%)	919 (76.6%)
Sim	1954 (20.3%)	348 (15.4%)	557 (20.6%)	437 (21.0%)	334 (24.3%)	278 (23.2%)
Missing	18 (0.2%)	5 (0.2%)	4 (0.1%)	5 (0.2%)	2 (0.1%)	2 (0.2%)
Consumo de Álcool						
Não	7492 (77.9%)	1633 (72.2%)	2073 (76.7%)	1641 (78.9%)	1102 (80.0%)	1043 (87.0%)
Sim	2101 (21.8%)	617 (27.3%)	620 (23.0%)	436 (21.0%)	274 (19.9%)	154 (12.8%)
Missing	24 (0.2%)	11 (0.5%)	8 (0.3%)	2 (0.1%)	1 (0.1%)	2 (0.2%)
Atividade Física Vigorosa						
Não	7956 (82.7%)	1756 (77.7%)	2213 (81.9%)	1742 (83.8%)	1178 (85.5%)	1067 (89.0%)
Sim	1478 (15.4%)	446 (19.7%)	433 (16.0%)	301 (14.5%)	183 (13.3%)	115 (9.6%)
Missing	183 (1.9%)	59 (2.6%)	55 (2.0%)	36 (1.7%)	16 (1.2%)	17 (1.4%)
Frequência de Caminhada						
Não caminha	3981 (41.4%)	884 (39.1%)	1050 (38.9%)	834 (40.1%)	611 (44.4%)	602 (50.2%)
Pouco (1–3d)	2439 (25.4%)	562 (24.9%)	712 (26.4%)	541 (26.0%)	344 (25.0%)	280 (23.4%)
Frequente (4–7d)	3049 (31.7%)	765 (33.8%)	902 (33.4%)	672 (32.3%)	409 (29.7%)	301 (25.1%)
Missing	148 (1.5%)	50 (2.2%)	37 (1.4%)	32 (1.5%)	13 (0.9%)	16 (1.3%)
Qualidade do Sono						
Muito Ruim	302 (3.1%)	23 (1.0%)	50 (1.9%)	61 (2.9%)	76 (5.5%)	92 (7.7%)
Ruim	1184 (12.3%)	131 (5.8%)	262 (9.7%)	273 (13.1%)	222 (16.1%)	296 (24.7%)
Regular	2260 (23.5%)	423 (18.7%)	615 (22.8%)	512 (24.6%)	382 (27.7%)	328 (27.4%)
Boa	4628 (48.1%)	1278 (56.5%)	1404 (52.0%)	999 (48.1%)	556 (40.4%)	391 (32.6%)
Muito Boa	1206 (12.5%)	385 (17.0%)	363 (13.4%)	229 (11.0%)	140 (10.2%)	89 (7.4%)
Missing	37 (0.4%)	21 (0.9%)	7 (0.3%)	5 (0.2%)	1 (0.1%)	3 (0.3%)
Consumo de Frutas						
Não consome	1236 (12.9%)	340 (15.0%)	371 (13.7%)	241 (11.6%)	153 (11.1%)	131 (10.9%)
Baixo (1–3d)	2954 (30.7%)	682 (30.2%)	871 (32.2%)	646 (31.1%)	414 (30.1%)	341 (28.4%)
Frequente (4–7d)	5370 (55.8%)	1213 (53.6%)	1445 (53.5%)	1184 (57.0%)	806 (58.5%)	722 (60.2%)
Missing	57 (0.6%)	26 (1.2%)	14 (0.5%)	8 (0.4%)	4 (0.3%)	5 (0.4%)
Renda domiciliar (faixas)						
Faixas 1–7	8567 (89.1%)	1952 (86.3%)	2402 (88.9%)	1850 (89.0%)	1252 (90.9%)	1111 (92.7%)
Faixas 8–12	362 (3.8%)	97 (4.3%)	91 (3.4%)	85 (4.1%)	57 (4.1%)	32 (2.7%)
Faixas 13–20	139 (1.4%)	38 (1.7%)	39 (1.4%)	33 (1.6%)	20 (1.5%)	9 (0.8%)
Missing	549 (5.7%)	174 (7.7%)	169 (6.3%)	111 (5.3%)	48 (3.5%)	47 (3.9%)

Fonte: Autoria própria (2026).

Nota-se um claro envelhecimento: a idade média sobe de 64,1 anos (indivíduos sem doenças) para 68,4 anos (indivíduos com 4+ doenças), padrão compatível com a natureza cumulativa das DCNTs ao longo da vida. A distribuição por sexo se inverte drasticamente: enquanto os homens são maioria no grupo sem doenças (54,1%), as mulheres passam a predominar a partir do grupo com 2 doenças (61,8%), chegando a representar 81,9% daqueles com 4 ou mais doenças. Podemos sugerir que, nesta amostra do ELSI-Brasil, a multimorbidade é predominante no sexo feminino, refletindo fatores como maior longevidade e/ou diferenças na busca por serviços de saúde.

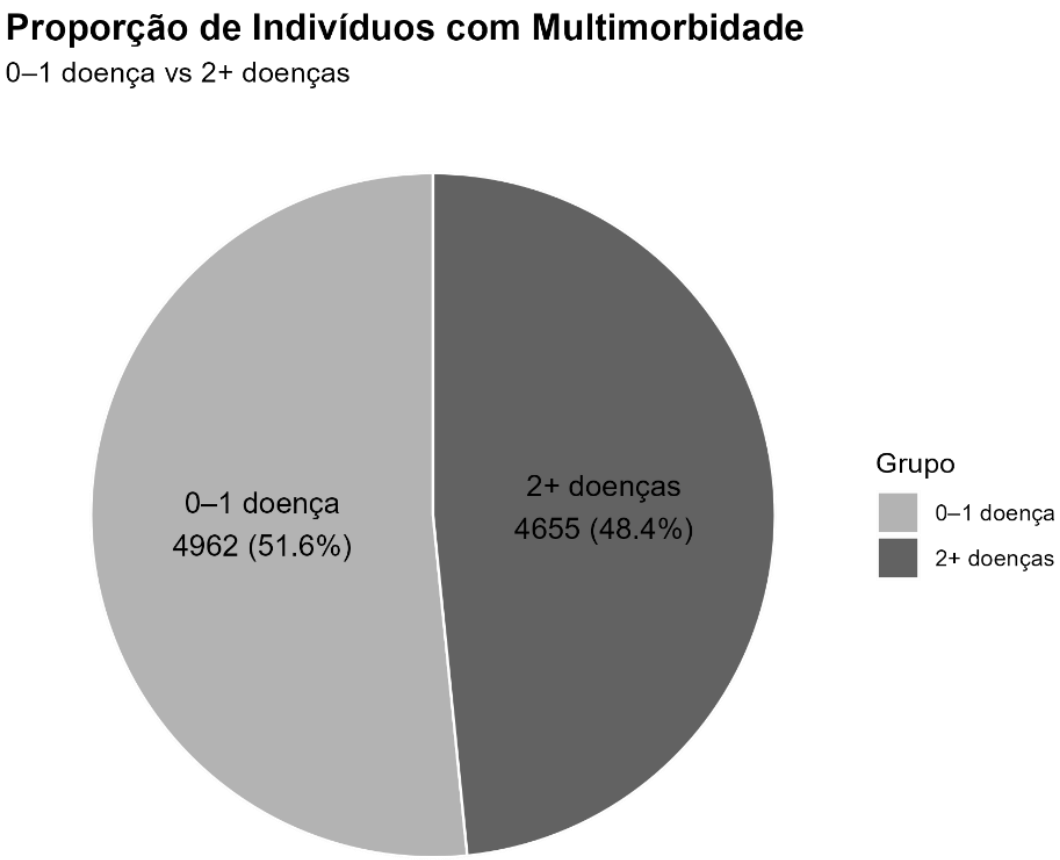
No aspecto socioeconômico, a baixa escolaridade, representada pelo ensino fundamental incompleto ou ausência de instrução, é predominante em todos os grupos, apresentando um leve aumento proporcional à medida que a quantidade de doenças aumenta. Semelhante ao que é verificado na renda domiciliar, onde as faixas mais baixas (de 0 até 7 salários-mínimos da época de coleta) concentram a maior parte da amostra em todas as categorias de multimorbidade. Esses dados reforçam a influência das categóricas sociais de saúde, sugerindo que a vulnerabilidade econômica pode atuar como um fator que favorece o acúmulo de condições crônicas na população idosa.

Os marcadores clínico-antropométricos reforçam a associação entre composição corporal e estado de saúde. Observa-se que a proporção de indivíduos com IMC fora da faixa considerada ideal aumenta progressivamente, passando de 51,7% entre aqueles sem doenças crônicas para 69,1% entre indivíduos com quatro ou mais condições. De forma semelhante, a prevalência de risco elevado segundo a razão cintura-quadril cresce de 65,5% para 76,9% ao longo desse mesmo gradiente. Esses resultados confirmam que o excesso de gordura, tanto aquela distribuída pelo corpo todo (medida pelo IMC) quanto a que se concentra na região da barriga (medida pela cintura), está fortemente ligado ao aparecimento de várias doenças ao mesmo tempo nos idosos. Na prática, isso mostra que o peso e a circunferência abdominal são indicadores fundamentais para identificar quem tem maior risco de desenvolver múltiplas doenças crônicas.

Quanto ao estilo de vida, observou-se que o grupo com maior número de doenças relata consumir menos álcool e praticar menos atividades físicas vigorosas. Esse comportamento sugere uma adaptação pós-diagnóstico: é provável que o aparecimento das doenças tenha forçado os idosos a mudarem seus hábitos (como parar de beber por recomendação médica), um fenômeno conhecido como causalidade reversa.

Expandindo a análise para o perfil geral da amostra (Figura **12** 12) e adotando a definição operacional de multimorbidade como a presença de duas ou mais condições crônicas (≥ 2), constatou-se que 4.655 participantes (48,4%) se enquadram nessa categoria, enquanto 4.962 (51,6%) possuem 1 ou nenhuma doença. Esse resultado indica que, mesmo considerando um conjunto restrito de sete DCNTs, a multimorbidade é altamente prevalente na população 50+ avaliada.

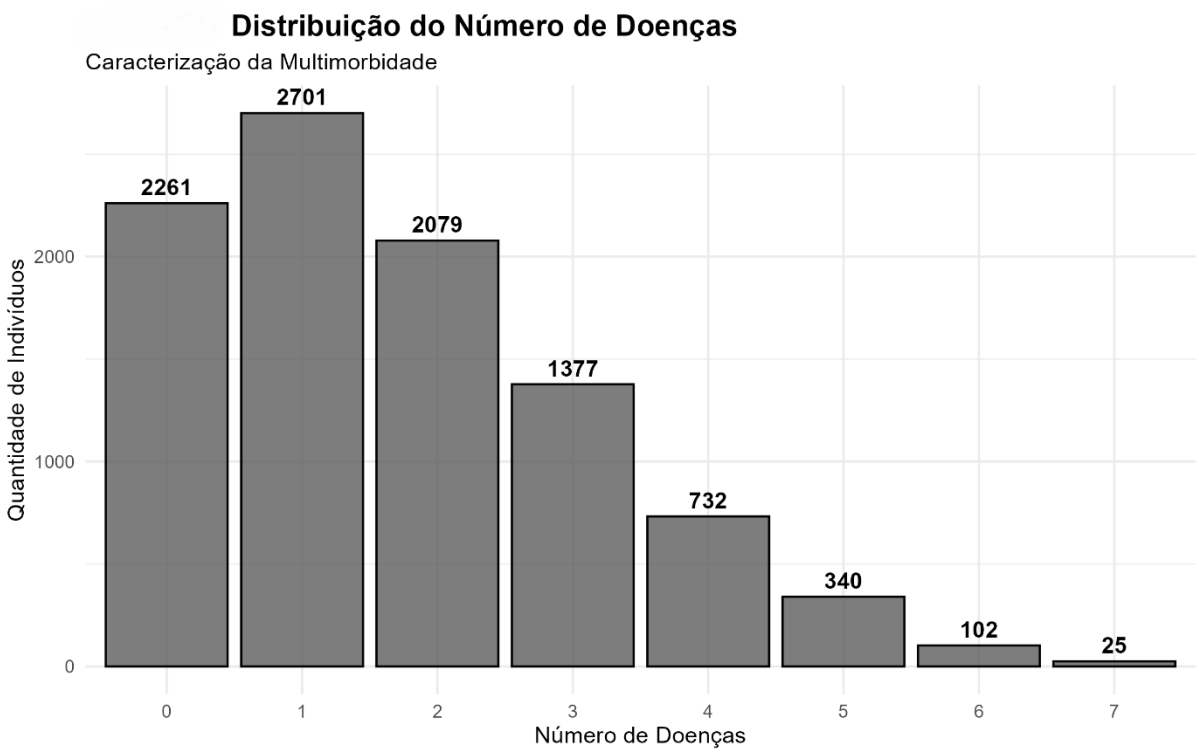
Figura 12 - Proporção de Indivíduos com Multimorbidade



Fonte: Autoria própria (2026).

Ao detalhar essa composição na Figura 13, que apresenta a distribuição do número de doenças (0 a 7) por indivíduo, nota-se uma concentração expressiva nas categorias iniciais: 2.261 participantes com nenhuma doença, 2.701 com apenas uma e 2.079 com duas. A partir desse ponto, observa-se uma cauda à direita com frequências decrescentes para 3 (1.377), 4 (732), 5 (340), 6 (102) e 7 doenças (25). Esse formato de distribuição é típico de fenômenos de carga de doença, nos quais uma parcela menor da amostra acumula a maior quantidade de condições simultâneas.

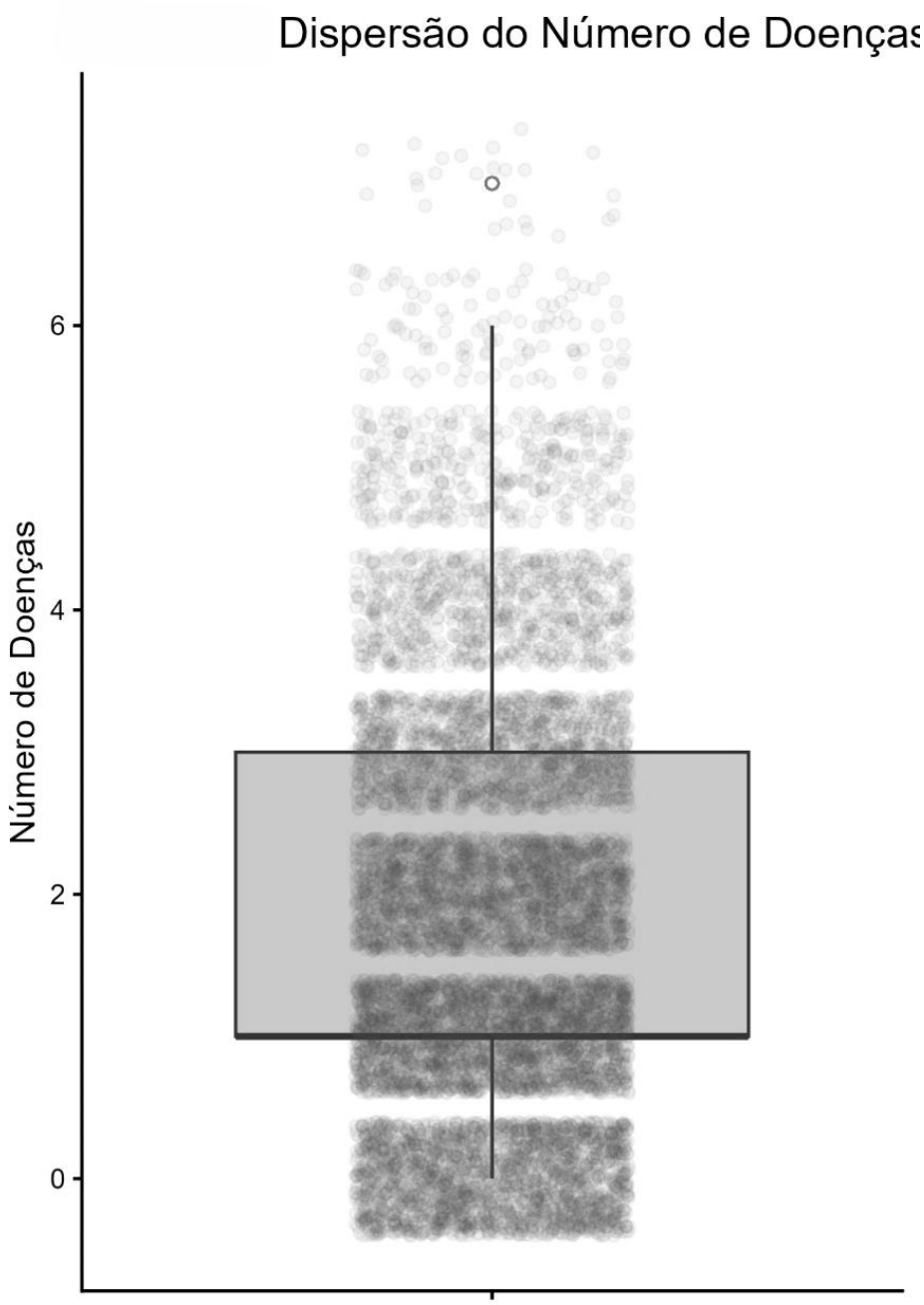
Figura 13 - Distribuição do Número de Doenças



Fonte: Autoria própria (2026).

Complementarmente, a análise de dispersão (Figura 14) apresenta mediana igual a 1 e intervalo interquartil de 1 a 3 doenças, confirmando a concentração da amostra em carga baixa a moderada. No entanto, há valores extremos que atingem até 7 doenças, o que demarca um subgrupo clinicamente relevante com alta multimorbidade e, por consequência, maior complexidade de cuidado.

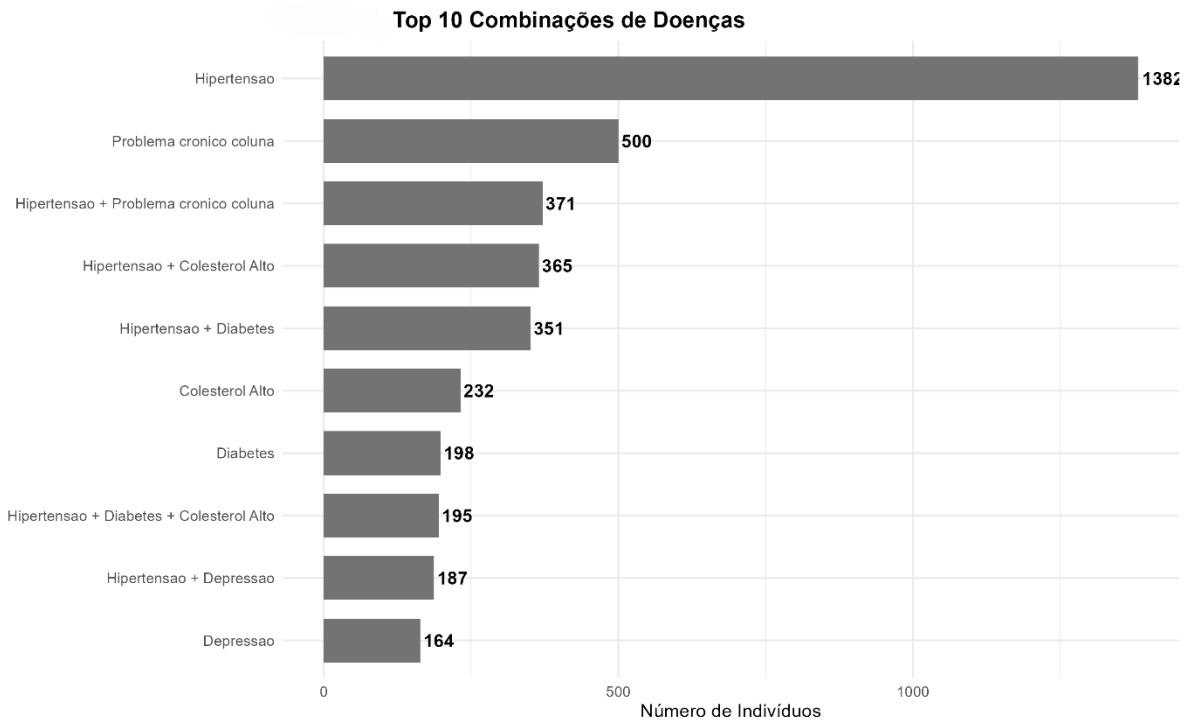
Figura 14 - Dispersão do Número de Doenças



Fonte: Autoria própria (2026).

Dada a natureza multirrótulo do problema, a análise estende-se para as combinações de doenças (*labelsets*). A Figura 15 destaca os dez padrões mais recorrentes, revelando que, embora condições isoladas (como hipertensão ou apenas problemas de coluna) ocupem as primeiras posições, as combinações binárias envolvendo a hipertensão são frequentes. Nota-se a forte associação desta com problemas crônicos de coluna, colesterol alto e diabetes. Essa onipresença da hipertensão nos diversos agrupamentos reforça seu papel como uma espécie de "doença eixo" na rede de multimorbidade da amostra.

Figura 15 - Top 10 Combinações de Doenças



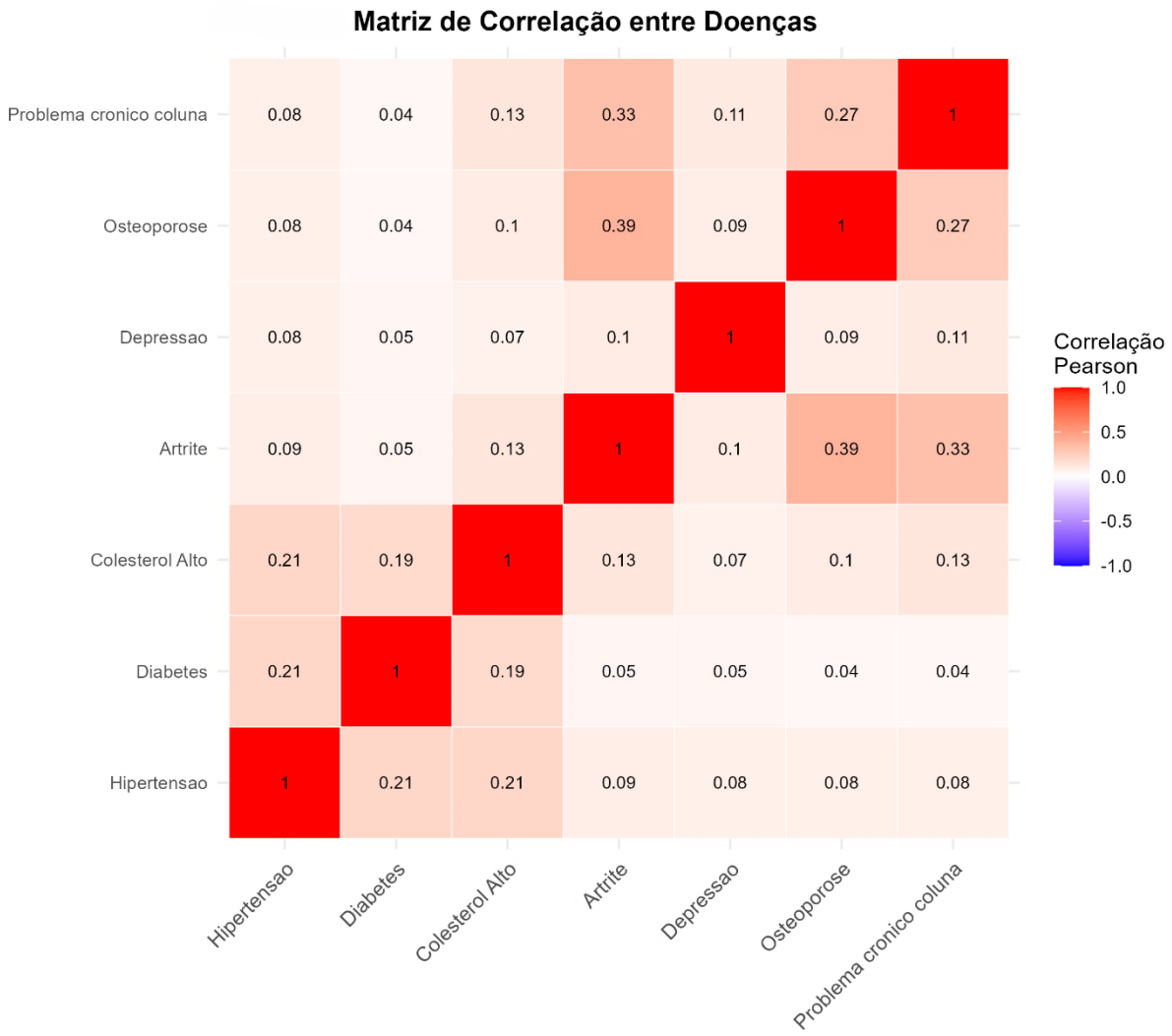
Fonte: Autoria própria (2026).

Para compreender como as doenças interagem entre si, a matriz de correlação (Figura 15) mediu a força das associações entre os pares de condições. De modo geral, as conexões variam de fracas a moderadas, o que é esperado ao se comparar doenças de naturezas tão distintas.

Apesar disso, os dados revelam dois agrupamentos claros. O primeiro agrupamento forma um núcleo musculoesquelético: as maiores correlações ocorrem entre artrite, osteoporose e problemas de coluna (chegando a 0,39), indicando que essas condições tendem a se manifestar em conjunto. O segundo grupo desenha um eixo cardiometabólico, conectando hipertensão, diabetes e colesterol. Embora as correlações aqui sejam mais discretas (em torno de 0,21), elas

confirmam estatisticamente a tendência clínica dessas doenças 'caminharem juntas' no paciente idoso.

Figura 15 - Matriz de Correlação entre Doenças



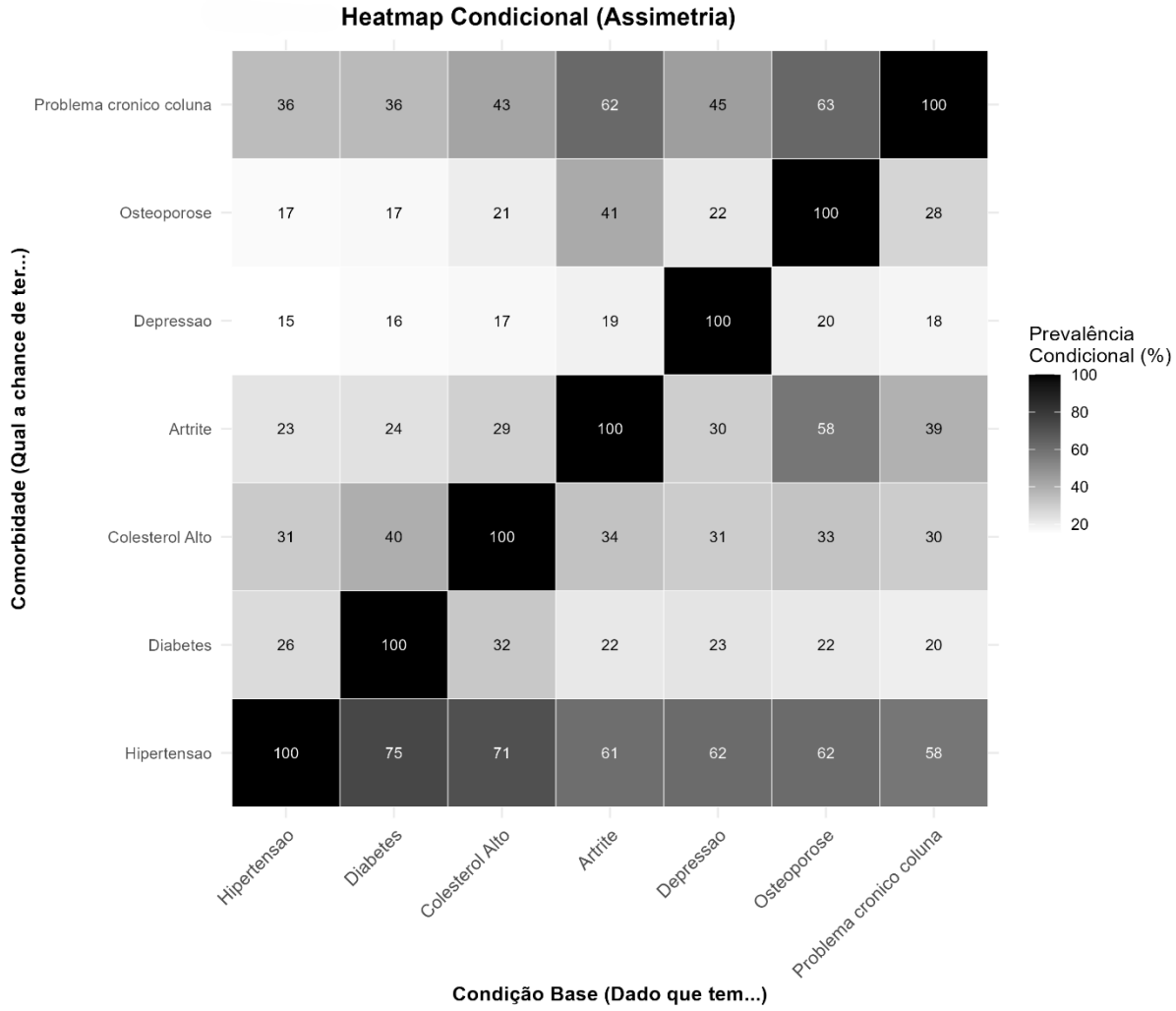
Fonte: Autoria própria (2026).

Enquanto a matriz de correlação expõe a força das associações, a análise condicional (Figura 16) revela a direção e a assimetria dessas relações, respondendo à pergunta: 'dado que o paciente tem a doença X, qual a probabilidade de ter a doença Y?'.

Esse *heatmap* evidencia desequilíbrios clínicos marcantes. Tome-se como exemplo o eixo metabólico: para um paciente diagnosticado com diabetes, a probabilidade de também ser hipertenso é altíssima, chegando a 75%. Já o caminho inverso é bem menos frequente: apenas 26% dos hipertensos apresentam diabetes. Essa discrepância confirma que a hipertensão atua como uma condição onipresente ou 'de base', enquanto o diabetes afeta um subgrupo mais específico. Padrões similares de assimetria observam-se no colesterol alto e na artrite. Na

prática, esse mapeamento identifica 'condições sentinela', cujo diagnóstico deve servir de alerta imediato para a investigação de outras comorbidades latentes.

Figura 16 - Heatmap Condicional



Fonte: Autoria própria (2026).

4.1 DESEMPENHO DOS MODELOS MULTIRRÓTULO

A consolidação do desempenho preditivo (Tabela 3), obtida após 50 rodadas de validação cruzada, indica estabilidade entre as abordagens avaliadas e um desempenho global moderado. As estratégias SVM-BR e MBR-RF apresentaram os melhores valores médios de *Accuracy* (0,289) e de *F-Measure* (0,351), sugerindo que ambas conseguem, em média, recuperar parte relevante do conjunto de rótulos por indivíduo. O modelo SVM-BR também obteve a menor *Hamming Loss* (0,237), indicando a menor taxa média de erros por rótulo (isto é, menor proporção de omissões e inclusões indevidas considerando todas as doenças). Por sua vez, a *Subset Accuracy* permaneceu próxima de 0,14 em todos os cenários, um resultado

esperado dada a severidade dessa métrica, que só contabiliza acertos quando o modelo prediz corretamente o conjunto completo das sete doenças simultaneamente para o mesmo indivíduo.

Tabela 3 - Desempenho médio dos modelos (50 folds) na predição multirrótulo de DCNTs (ELSI-Brasil, onda 2).

Método	Accuracy	Subset Accuracy	Hamming Loss	F-Measure
SVM-BR	0.289	0.143	0.237	0.351
MBR-RF	0.289	0.141	0.239	0.351
RF-BR	0.288	0.140	0.239	0.350
RF-CC	0.285	0.144	0.239	0.345
SVM-CC	0.284	0.144	0.239	0.344
SVM-DBR	0.280	0.141	0.240	0.339
RF-DBR	0.274	0.136	0.243	0.333

Fonte: Autoria própria (2026).

Ao comparar as estratégias de decomposição, observa-se que abordagens desenhadas para explorar dependências entre rótulos, como CC e variações DBR, não produziram ganhos consistentes em relação ao BR. Esse achado sugere que, para este conjunto de dados e configuração experimental, a informação adicional introduzida por modelos dependentes não foi suficiente para superar, de forma clara, uma abordagem mais simples com classificadores bem ajustados. Na prática, isso reforça o uso do BR como baseline competitivo, especialmente quando se considera o custo de implementação e interpretação de estratégias mais complexas.

Em termos de implicações, os resultados apontam que avanços futuros provavelmente dependerão menos da substituição do classificador base e mais de refinamentos no processo de modelagem, como calibração de limiares de decisão (*thresholds*), estratégias de tratamento de desbalanceamento e critérios de otimização alinhados às prioridades do contexto clínico (por exemplo, reduzir omissões de condições menos frequentes).

4.2 CONSOLIDAÇÃO DOS RESULTADOS

Integrando as diferentes camadas de análise, os dados da segunda onda do ELSI-Brasil revelam um cenário de alta complexidade epidemiológica: quase metade da amostra (48,4%) convive com a multimorbidade. Mais do que apenas uma contagem elevada, observou-se uma estrutura latente nos dados que organiza as doenças em dois grandes polos: um núcleo musculoesquelético (artrite, coluna, osteoporose) e um eixo cardiometabólico (hipertensão, diabetes, colesterol). Nesse contexto, a hipertensão assume um protagonismo inegável, atuando não apenas como a condição mais frequente, mas como uma espécie de "doença-pivô" presente na vasta maioria das combinações clínicas.

Sob a ótica do AM, essa complexidade se traduziu em um desafio de modelagem considerável. A análise de importância de variáveis confirmou que os algoritmos conseguiram capturar sinais biológicos relevantes: medidas antropométricas e hemodinâmicas, somadas à idade e determinantes sociais, formaram o conjunto de preditores com maior poder discriminativo. Isso válida a qualidade dos dados de entrada e a capacidade dos modelos de replicar, via dados, o conhecimento clínico estabelecido.

No entanto, o desempenho preditivo moderado e a pequena diferença entre estratégias simples (BR) e complexas (CC/DBR/MBR) sugerem que a multimorbidade é um fenômeno difuso. As correlações entre as doenças existem, mas não são determinísticas o suficiente para que modelos de encadeamento superem facilmente uma abordagem independente.

5. CONCLUSÃO

Este trabalho avaliou a viabilidade de prever multimorbidade em idosos por classificação multirrótulo (CMR) utilizando a segunda onda do ELSI-Brasil. Em vez de tratar multimorbidade como “contagem de doenças”, o estudo modelou a ocorrência simultânea de condições crônicas, com um pipeline reprodutível do preparo dos dados à validação dos modelos.

Foram selecionadas variáveis preditoras de três grupos (sociodemográficas, estilo de vida e antropometria) e definidas sete condições-alvo: Hipertensão, Diabetes, Colesterol alto, Artrite, Depressão, Osteoporose e Problema crônico de coluna. Após limpeza e imputação, foram incorporados indicadores antropométricos como IMC e razão cintura–quadril (RCQ), assegurando consistência para a etapa de modelagem.

A análise exploratória indicou alta carga de multimorbidade (48,4%) e padrões de coocorrência clinicamente plausíveis. Em particular, a hipertensão apareceu como condição frequente nas combinações, sugerindo associação com outras DCNTs cardiometabólicas. Esse comportamento reforça a adequação da abordagem multirrótulo para representar o problema.

Na modelagem, compararam-se estratégias BR, CC e variações DBR/MBR com classificadores base RF e SVM. Os resultados foram moderados e estáveis, com destaque para SVM-BR e MBR-RF, que atingiram os melhores valores médios de Accuracy (≈ 0,289) e F-Measure (≈ 0,351). A ausência de ganhos expressivos com métodos encadeados sugere que, neste conjunto de dados e configuração experimental, as dependências entre rótulos não foram suficientemente fortes para superar de forma clara a abordagem BR, que se mantém como um baseline competitivo.

Do ponto de vista computacional, o protocolo de validação cruzada repetida demandou mais de 12 horas em ambiente de hardware limitado, confirmando a viabilidade do experimento e indicando a necessidade de melhorias para escalabilidade (por exemplo, paralelização e otimização de memória).

Como contribuições, o trabalho entrega: um conjunto de dados tratado e documentado; uma caracterização objetiva da multimorbidade por coocorrência; e um benchmark de métodos multirrótulo no contexto do ELSI-Brasil. Como continuidade, recomenda-se investigar tratamento de desbalanceamento, calibração/ajuste de limiares e modelos capazes de capturar dependências de forma mais robusta.

Em síntese, a predição multirrótulo mostrou-se factível no ELSI-Brasil usando variáveis acessíveis, com potencial de apoio à estratificação de risco e vigilância em saúde,

desde que se considere o custo computacional e as limitações inerentes aos dados observacionais.

REFERÊNCIAS

CHEN, Yannan; WENG, Ming-Wei; WU, Shun-Xiang; CHEN, Bai-Hua; FAN, Yu-Ling; LIU, Jing-Hua. *An efficient stacking model with label selection for multi-label classification*. *Applied Intelligence*, v. 51, n. 9, p. 308–325, 2021. DOI: 10.1007/s10489-020-01807-z.

DEISENROTH, M. P.; FAISAL, A. A.; ONG, C. S. **Mathematics for Machine Learning**. Cambridge: Cambridge University Press, 2020.

DUNCAN, Bruce Bartholow et al. Doenças Crônicas Não Transmissíveis no Brasil: prioridade para enfrentamento e investigação. **Revista de Saúde Pública**, v. 46, supl., p. 126–134, 2012.

FACELI, K. et al. **Inteligência Artificial**: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

GATTO, Elaine Cecília. **Webinar:** Resolvendo problemas do mundo real com classificação multirrótulo: uma introdução. EmbarcadosTV, 2023. Disponível em: https://www.youtube.com/watch?v=d2yegaGI3iU. Acesso em: 15 abr. 2025.

GÉRON, A. **Mãos à Obra:** Aprendizado de Máquina com Scikit-Learn & TensorFlow. Rio de Janeiro: Alta Books, 2019.

GOBI, S.; HAGHNEJAD, A.; GHARIBZADEH, A. Development of new computational machine learning models for longitudinal dispersion coefficient determination: case study of natural streams, United States. [s.l.]: ResearchGate, 2022.

GONÇALVES, Eduardo Corrêa. **Introdução à Classificação Multirrótulo**. V Escola Regional de Sistemas de Informação do Rio de Janeiro, SBC, 1ª ed., 2018. ISBN 978-85-7669-456-4.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining:** concepts and techniques. 3. ed. San Francisco: Morgan Kaufmann, 2011.

IZBICKI, R.; SANTOS, T. M. **Aprendizado de Máquina**: uma abordagem estatística. São Carlos: [s.n.], 2020.

LIMA-COSTA, M. F. et al. Cohort Profile: The Brazilian Longitudinal Study of Ageing (ELSI-Brazil). **International Journal of Epidemiology**, v. 52, n. 1, p. e57-e65, fev. 2023. DOI: 10.1093/ije/dyac132.

LIMA-COSTA, M. F. et al. The Brazilian Longitudinal Study of Aging (ELSI-Brazil): Objectives and Design. **American Journal of Epidemiology**, v. 187, n. 7, p. 1345-1353, jul. 2018. DOI: 10.1093/aje/kwx387.

LUDERMIR, Teresa Bernarda**. Inteligência Artificial e Aprendizado de Máquina:** estado atual e tendências. Estudos Avançados, v. 35, n. 101, p. 85–102, 2021. DOI: 10.1590/s0103-4014.2021.35101.007.

MELO, Mônica Thalia Brito de et al. **Prevalência de Doenças Crônicas Não Transmissíveis em idosos do Nordeste:** uma revisão integrativa. Publicado em: 10 jan. 2023.

MORETTIN, P. A.; SINGER, J. M. **Estatística e Ciência de Dados**. Rio de Janeiro: LTC, 2022.

MYGREATLEARNING. **Introduction to Support Vector Machine**. Disponível em: https://www.mygreatlearning.com/blog/introduction-to-support-vector-machine/. Acesso em: 11 jan. 2026.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). **Global status report on noncommunicable diseases 2010.** Genebra: Organização Mundial da Saúde, 2011. Disponível em: https://apps.who.int/iris/handle/10665/44579. Acesso em: 04 jan. 2026.

PAULA, D. P. et al. Comparing machine learning algorithms for multimorbidity prediction: An example from the Elsa-Brasil study. **PLoS ONE**, v. 17, n. 10, p. e0275619, out. 2022.

RIVOLLI, A.; CARVALHO, A. C. P. L. F. utiml: Utilities for Multi-Label Learning. **The R Journal**, v. 10, n. 2, p. 24–38, 2018.

SANTOS, A. d. M. Investigando a combinação de técnicas de aprendizado semissupervisionado e classificação hierárquica multirrótulo. Dissertação (Mestrado) — Universidade Federal do Rio Grande doNorte, 2012.

SIMIELI, Isabela; PADILHA, Letícia Aparecida Resende; TAVARES, Cristiane Fernandes de Freitas. Realidade do envelhecimento populacional frente às doenças crônicas não transmissíveis. **Revista Eletrônica Acervo Saúde**, v. 11, n. 15, 2019. Disponível em: https://doi.org/10.25248/reas.e1511.2019. Acesso em: 20 mai. 2025.

SOUZA, L. K. C. de; FERREIRA, K. C.; RIBEIRO, G. C. de S.; VIANA NETA, I. S.; BARBOSA, A. J. M.; VALADÃO, P. A. da S.; FARIAS, T. M. Mortalidade prematura por doenças crônicas não transmissíveis no Nordeste do Brasil. In: **VI CONGRESSO BRASILEIRO DE MEDICINA DO ESTILO DE VIDA**, 2023. Anais [...]. Disponível em: https://publicacoes.cbmev.org.br/cbmev/article/view/39. Acesso em: 12 jan. 2026.

VIDULIN, Vedrana. **Searching for credible relations in machine learning**. 2012. 135 f. Tese (Doutorado em Ciência da Computação) – Jožef Stefan International Postgraduate School, Ljubljana, 2012.