Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Desenvolvimento de um modelo preditivo para identificação de uso dos serviços de saúde com base nas características de brasileiros

Gabriel Felipe Souza Santos

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Helen de Cássia Sousa da Costa Lima

COORIENTAÇÃO:
Érica de Matos Reis Ferreira

Abril, 2025
João Monlevade–MG

Gabriel Felipe Souza Santos

Desenvolvimento de um modelo preditivo para identificação de uso dos serviços de saúde com base nas características de brasileiros

Orientador: Helen de Cássia Sousa da Costa Lima

Coorientador: Érica de Matos Reis Ferreira

Monografia apresentada ao curso de Sistemas de Informação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

Universidade Federal de Ouro Preto

João Monlevade

Abril de 2025

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

S237d Santos, Gabriel Felipe Souza.
Desenvolvimento de um modelo preditivo para identificação de uso dos serviços de saúde com base nas características de brasileiros. [manuscrito] / Gabriel Felipe Souza Santos. - 2025.
86 f.: il.: color., gráf., tab..

Orientadora: Profa. Dra. Helen de Cássia Sousa da Costa Lima.
Coorientadora: Ma. Érica de Matos Reis Ferreira.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Sistemas de Informação .

1. Alocação de recursos. 2. Aprendizado do computador. 3. Controle preditivo. 4. Mineração de dados (Computação). 5. Saúde pública. 6. Sistema Único de Saúde (Brasil). I. Lima, Helen de Cássia Sousa da Costa. II. Ferreira, Érica de Matos Reis. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 519.2:004.85

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Gabriel Felipe Souza Santos

Desenvolvimento de um modelo preditivo para identificação de uso dos serviços de saúde com base nas características de brasileiros

Monografia apresentada ao Curso de Sistemas de Informação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em 10 de abril de 2025.

Membros da banca

Doutor - Helen de Cássia Sousa da Costa Lima - Orientadora - Universidade Federal de Ouro Preto
Mestre - Érica de Matos Reis Ferreira - Coorientadora - Universidade Federal de Minas Gerais
Doutor - Alexandre Magno de Sousa - Universidade Federal de Ouro Preto
Doutor - Carlos Henrique Gomes Ferreira - Universidade Federal de Ouro Preto

Professora Helen de Cássia Sousa da Costa Lima, orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 09/05/2025.

Documento assinado eletronicamente por **Helen de Cassia Sousa da Costa Lima**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 09/05/2025, às 15:55, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0907943** e o código CRC **BA15FDC7**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

Dedico este trabalho aos meus pais, Fabricia Cristina de Souza e Glaucio Duarte dos Santos, que sempre me incentivaram a ter força, disciplina e a nunca desistir, ensinando-me os valores da dedicação e perseverança. A eles, meu mais sincero agradecimento por todo o apoio e incentivo ao longo dessa jornada.

Dedico também aos meus avós, Maria Gonçalves Souza e José Joaquim de Souza, que me receberam em sua casa e me ofereceram todo o suporte possível para que eu pudesse concluir minha graduação. Seu carinho, compreensão e apoio foram fundamentais para que eu chegasse até aqui.

Em especial, dedico este trabalho ao meu avô José Joaquim, que infelizmente não está mais presente para me ver concluir esta etapa tão importante da minha vida. Sua generosidade, força e exemplo de vida sempre serão inspiração para mim. Sem ele, nada disso teria sido possível, e levo comigo sua memória e todo o aprendizado que me deixou.

Este trabalho é para vocês, com todo meu amor e gratidão.

Agradecimentos

Agradeço à Universidade Federal de Ouro Preto por todos os ensinamentos e oportunidades oferecidos ao longo desta jornada acadêmica. Mesmo diante de grandes desafios, como a pandemia e dificuldades governamentais, a instituição manteve-se comprometida com a formação de seus alunos, proporcionando um ambiente de aprendizado e crescimento.

Meu profundo agradecimento à minha orientadora, Helen de Cássia Sousa da Costa Lima, por toda a dedicação, paciência e conhecimento compartilhado. Sua orientação foi essencial para o desenvolvimento deste trabalho, sempre me direcionando e estando disponível para aprimorar a pesquisa com seu olhar atento e criterioso.

Agradeço também à minha coorientadora, Érica de Matos Reis Ferreira, por sua valiosa contribuição na área da saúde. Sua presença e direcionamento foram fundamentais para enriquecer este estudo, trazendo novas perspectivas e sugestões que ajudaram a definir caminhos importantes para a pesquisa.

A todas essas pessoas e instituições que contribuíram para a realização deste trabalho, deixo aqui minha mais sincera gratidão.

"Amamos Vocês"

— Jose Joaquim de Souza (1956 – 2025),

Resumo

O sistema público de saúde brasileiro enfrenta desafios significativos, como orçamentos limitados, alta prevalência de doenças crônicas e envelhecimento populacional, fatores que intensificam a demanda por serviços e dificultam a alocação eficiente de recursos. Neste contexto, o presente estudo propõe desenvolver um modelo preditivo, utilizando dados da Pesquisa Nacional de Saúde (PNS) de 2019, com aplicação de técnicas avançadas de aprendizado de máquina, como *Random Forest*, *XGBoost* e redes neurais artificiais, visando antecipar padrões de utilização dos serviços de saúde e fornecer suporte estratégico à gestão do Sistema Único de Saúde (SUS). A análise identificou variáveis determinantes como idade, nível socioeconômico, presença de doenças crônicas e hábitos comportamentais, com destaque para a presença de planos de saúde, autoavaliação do estado de saúde e condições como hipertensão e diabetes. O algoritmo *XGBoost* apresentou o melhor desempenho, com acurácia de 78,5%, precisão e *recall* de 78,5%, e F1-Score de 78,4%, evidenciando sua robustez na identificação de padrões de uso. A aplicação do modelo permite o monitoramento contínuo da demanda e facilita intervenções preventivas mais assertivas por parte dos gestores públicos. A segmentação da população permitiu ainda a identificação de perfis distintos de usuários, favorecendo a formulação de políticas públicas mais direcionadas e eficientes.

Palavras-chave: modelo preditivo, aprendizado de máquina, saúde pública, SUS, gestão de recursos.

Abstract

The Brazilian public health system faces significant challenges, such as limited budgets, a high prevalence of chronic diseases, and an aging population, which increase the demand for healthcare services and complicate the efficient allocation of resources. In this context, this study proposes the development of a predictive model using data from the 2019 National Health Survey (PNS), applying advanced machine learning techniques—such as *Random Forest*, *XGBoost*, and artificial neural networks—with the goal of anticipating patterns of healthcare service utilization and supporting strategic decision-making within the Unified Health System (SUS). The analysis identified key predictors including age, socioeconomic level, chronic conditions, and behavioral habits, with special emphasis on health insurance coverage, self-assessed health status, and conditions such as hypertension and diabetes. Among the tested algorithms, *XGBoost* achieved the best performance, with 78.5% accuracy, 78.5% precision and recall, and an F1-Score of 78.4%, demonstrating its robustness in identifying healthcare usage patterns. The model enables continuous monitoring of demand and supports more targeted preventive interventions by public health managers. Furthermore, the segmentation of the population allowed the identification of distinct user profiles, supporting the development of more efficient and focused public health policies.

Keywords: predictive model, machine learning, public health, SUS, resource allocation.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AVC Acidente Vascular Cerebral

DATASUS Departamento de Informação e Informática do Sistema Único de Saúde

DORT Distúrbios Osteomusculares Relacionados ao Trabalho

IBGE Instituto Brasileiro de Geografia e Estatística

IMC Índice de Massa Corporal

MLP Multi-Layer Perceptron (*Perceptron de Múltiplas Camadas*)

PCA Principal Component Analysis (*Análise de Componentes Principais*)

PIB Produto Interno Bruto

PNS Pesquisa Nacional de Saúde

SUS Sistema Único de Saúde

SVC Support Vector Classifier (*Classificador de Vetores de Suporte*)

SVM Support Vector Machine (*Máquina de Vetores de Suporte*)

WCSS Within-Cluster Sum of Squares (*Soma dos Quadrados Dentro do Cluster*)

XGB XGBoost (*Extreme Gradient Boosting*)

Sumário

1 Introdução

O crescente desafio de gerir o sistema de saúde brasileiro tem sido amplamente reconhecido em um cenário de restrições orçamentárias e aumento contínuo da demanda por serviços de saúde. Dados do Instituto Brasileiro de Geografia e Estatística (IBGE) indicam que, em 2019, os gastos com saúde no Brasil representaram 9,6% do PIB, totalizando R$ 711,4 bilhões (IBGE, 2022). Esse valor representa um aumento em relação a anos anteriores, visto que, em 2015, os gastos com saúde correspondiam a 8,3% do PIB (MUNDIAL, 2021). Esse crescimento reflete a ampliação da cobertura de serviços e o aumento da complexidade dos atendimentos prestados no Sistema Único de Saúde (SUS).

Além disso, observa-se um crescimento significativo na utilização dos serviços de saúde. Entre 2010 e 2019, a demanda por serviços como consultas médicas, exames laboratoriais e terapias aumentou em aproximadamente 54% (Anahp, 2019). Esse aumento foi particularmente expressivo em atendimentos ambulatoriais e exames complementares, resultando em maior pressão sobre o sistema público de saúde. A crescente prevalência de doenças crônicas, como diabetes e hipertensão, combinada ao envelhecimento populacional, intensifica ainda mais essa demanda, tornando essencial o desenvolvimento de estratégias para a alocação eficiente de recursos (SIMõES; MEIRA; SANTOS, 2021).

1.1 Problema de pesquisa

A gestão eficiente dos recursos de saúde no Brasil é dificultada por vários fatores críticos, incluindo a ausência de ferramentas preditivas robustas que permitam identificar previamente os indivíduos ou grupos com maior probabilidade de utilizar os serviços de saúde. Essa lacuna resulta em desperdícios de recursos, ineficiências operacionais e impactos negativos na qualidade do atendimento ao paciente. Assim, a pergunta que norteia este trabalho é: *"Como desenvolver um modelo preditivo eficiente para identificar padrões de uso dos serviços de saúde com base nas características demográficas, socioeconômicas e de saúde da população brasileira?"*.

1.2 Objetivos

O objetivo principal deste estudo é desenvolver e validar um modelo preditivo capaz de identificar, com antecedência, indivíduos ou grupos propensos a demandar serviços de saúde. Para isso, os objetivos específicos incluem:

- Analisar padrões e tendências nos dados públicos de saúde, identificando os principais

fatores que influenciam o uso dos serviços de saúde;

- Construir um modelo preditivo utilizando algoritmos de aprendizado de máquina e técnicas de mineração de dados;
- Avaliar a eficácia do modelo desenvolvido, considerando métricas de predição e impacto na alocação eficiente de recursos.

1.3 Contribuições do estudo

Este estudo apresenta duas principais contribuições. A primeira refere-se à segmentação da população brasileira com base em características clínicas, comportamentais e demográficas, por meio de técnicas de agrupamento (PCA + K-means). Como resultado, foram identificados seis perfis distintos de indivíduos, com diferentes níveis de comorbidades e padrões de percepção de saúde, permitindo uma análise detalhada de como esses grupos acessam os serviços de saúde. Dentre os clusters, destacam-se perfis com múltiplas comorbidades e alta procura por serviços, bem como grupos aparentemente saudáveis que subutilizam o sistema de saúde mesmo diante de fatores de risco como obesidade.

A segunda contribuição é desenvolver um modelo preditivo para estimar a probabilidade de utilização dos serviços de saúde, utilizando algoritmos como *Random Forest*, *Support Vector Machine* (SVM), *Multi-Layer Perceptron* (MLP) e *Extreme Gradient Boosting* (XGB). O modelo XGBoost se destacou com acurácia de 78,5%, precisão e revocação de 78,5% e F1-score de 78,4%. A análise de importância das variáveis indicou que fatores como IMC, idade, presença de depressão e hipertensão são determinantes críticos. Esses resultados evidenciam o potencial do modelo para apoiar gestores públicos na alocação eficiente de recursos e na formulação de políticas mais direcionadas e preventivas.

1.4 Justificativa

A utilização de modelos preditivos para a gestão de recursos de saúde tem demonstrado resultados promissores em contextos internacionais (ZHAO; ATUN; ANINDYA, 2021; VOS et al., 2017). No Brasil, ferramentas analíticas robustas podem contribuir significativamente para a mitigação de desigualdades no acesso aos serviços de saúde e para o fortalecimento do SUS, sobretudo diante de desafios como a crescente demanda por serviços especializados e o envelhecimento populacional. Este estudo busca preencher essa lacuna, fornecendo um modelo preditivo adaptado às especificidades do contexto brasileiro.

1.5 Estrutura do trabalho

Este trabalho está estruturado em cinco capítulos. Após a introdução, o **Capítulo 2** apresenta a revisão da bibliográfica, destacando modelos preditivos na área da saúde. O **Capítulo 3** detalha a metodologia empregada. O Desenvolvimento e resultados e análises são apresentados no **Capítulo 5**, seguidos pelas conclusões e recomendações no **Capítulo 6**.

2 Referencial teórico

Este capítulo apresenta os fundamentos teóricos que sustentam a construção do modelo preditivo proposto neste trabalho. Serão discutidos os principais conceitos relacionados ao pré-processamento, à mineração de dados, ao agrupamento e à classificação, com o intuito de fornecer o embasamento necessário à compreensão das abordagens metodológicas adotadas.

2.0.1 Pré-processamento de dados

O pré-processamento de dados é uma etapa fundamental em projetos de ciência de dados, sendo responsável pela preparação dos dados para análise estatística ou modelagem preditiva. Estima-se que entre 60% e 80% do tempo total de desenvolvimento de projetos analíticos seja dedicado a essa fase, o que evidencia sua importância para a qualidade e confiabilidade dos modelos gerados (GARCIA; LUENGO; HERRERA, 2016).

No contexto da saúde pública, a complexidade e a heterogeneidade dos dados tornam o pré-processamento ainda mais crucial. A Pesquisa Nacional de Saúde (PNS) 2019, utilizada neste trabalho, reúne dados demográficos, socioeconômicos, clínicos e comportamentais da população brasileira, frequentemente obtidos por meio de autorrelato, o que pode introduzir omissões, ruídos e inconsistências (Instituto Brasileiro de Geografia e Estatística (IBGE), 2020).

As principais estratégias adotadas no pré-processamento incluem:

- **Filtragem e seleção de variáveis**: Essa etapa visa reduzir a dimensionalidade do conjunto de dados, eliminando atributos irrelevantes, redundantes ou com baixa variabilidade. A seleção de variáveis é fundamental para melhorar a interpretabilidade dos modelos e reduzir o risco de sobreajuste, além de contribuir para a eficiência computacional dos algoritmos (GUYON; ELISSEEFF, 2003).
- **Remoção de dados inválidos**: A exclusão de registros com informações inconsistentes ou ausentes em variáveis-chave é essencial para garantir a qualidade analítica da base. A presença de dados inválidos pode introduzir viés, distorcer padrões e comprometer a validade dos resultados obtidos (RAHM; DO, 2000).
- **Transformação de variáveis em binários**: A conversão de variáveis categóricas com respostas como “Sim” e “Não” para valores binários (1 e 0) é uma prática comum em aprendizado de máquina. Essa transformação permite o uso dessas variáveis em algoritmos que exigem entradas numéricas e facilita o cálculo de métricas estatísticas (HAN; KAMBER; PEI, 2011).

- **Junção de colunas**: A fusão de variáveis com informações semelhantes ou complementares visa simplificar a estrutura dos dados e evitar redundâncias. No contexto da saúde, essa estratégia permite consolidar o histórico clínico do paciente e gerar atributos mais representativos, como no caso da unificação de diagnósticos gestacionais com variáveis principais (SHICKEL et al., 2018).

- **Geração de variáveis derivadas**: Criar novos atributos a partir de variáveis existentes. Ela é fundamental para enriquecer o conjunto de dados e possibilitar que os modelos identifiquem padrões mais relevantes (World Health Organization, 2000).

- **Cálculo do peso amostral**: Em bases de dados com amostragem complexa, como a PNS, a utilização de pesos amostrais é imprescindível para garantir a representatividade dos resultados em relação à população-alvo. O uso correto dos pesos ajusta as estimativas para que reflitam adequadamente a estrutura demográfica nacional (BARROS; VICTORA, 2019).

- **Tratamento de valores ausentes**: O manejo de dados faltantes é essencial para manter a integridade da base e evitar perdas amostrais. Estratégias como substituição por valores padrão (-1 ou 0) são simples, porém eficazes, especialmente em modelos que não toleram valores nulos (SAAR-TSECHANSKY; PROVOST, 2007).

- **Codificação supervisionada (*Target Encoding*)**: Essa técnica consiste em substituir categorias por valores numéricos calculados com base na média da variável alvo para cada grupo. O *Target Encoding* preserva a correlação estatística entre a variável categórica e o desfecho, sendo útil em modelos supervisionados com variáveis de alta cardinalidade (MICCI-BARRECA, 2001). Além disso, evita a explosão dimensional típica do *one-hot encoding*.

- **Balanceamento da base com *NearMiss-1***: Quando há desbalanceamento entre as classes da variável alvo, algoritmos preditivos podem favorecer a classe majoritária. O *NearMiss-1* é uma técnica de *undersampling* que seleciona amostras da classe majoritária com base em sua menor distância para instâncias da classe minoritária. Isso melhora a capacidade do modelo de identificar corretamente ambas as classes, reduzindo o viés de classificação (MANI; ZHANG, 2003).

2.1 Agrupamento

O agrupamento (*clustering*) é uma técnica de aprendizado não supervisionado amplamente utilizada no campo da mineração de dados, com o propósito de identificar estruturas naturais dentro de um conjunto de dados sem a necessidade de rótulos previamente definidos (TAN; STEINBACH; KUMAR, 2013). Diferentemente dos métodos supervisionados, que se baseiam em categorias conhecidas para treinar os modelos, o

agrupamento permite explorar e descobrir padrões ocultos, promovendo segmentações que refletem a similaridade entre observações.

Essa abordagem é particularmente útil em contextos multidimensionais, como bases de dados de saúde pública, que reúnem informações clínicas, demográficas e socioeconômicas de grande diversidade. A partir da análise da proximidade entre os dados, é possível formar grupos que compartilham características semelhantes, permitindo análises mais direcionadas e detalhadas (HAN; KAMBER; PEI, 2011).

No campo da saúde, o agrupamento tem sido cada vez mais utilizado para caracterizar perfis populacionais e identificar grupos com necessidades específicas de atenção. Essa técnica contribui para a formulação de políticas públicas mais eficazes, apoiando a alocação racional de recursos, a definição de prioridades e o desenho de intervenções personalizadas (BERNAL; RESTREPO, 2018; LOPES; OLIVEIRA; MATOS, 2019). Por exemplo, ao segmentar a população com base em condições como doenças crônicas, faixa etária, ou acesso a serviços médicos, é possível desenvolver programas preventivos mais direcionados e eficientes.

2.1.1 Análise de componentes principais (PCA)

A Análise de Componentes Principais (PCA) é uma técnica estatística de redução de dimensionalidade que transforma um conjunto de variáveis correlacionadas em um novo conjunto de variáveis não correlacionadas, denominadas componentes principais (JOLLIFFE; CADIMA, 2016). Essa transformação permite representar os dados em um espaço de menor dimensão, preservando ao máximo a variância original, o que facilita a visualização e a interpretação dos padrões existentes.

No presente estudo, a PCA foi aplicada como etapa preliminar ao agrupamento, com o objetivo de reduzir a complexidade do conjunto de dados e eliminar redundâncias. Os principais elementos considerados foram:

- **Componentes principais**: cada componente representa uma combinação linear das variáveis originais e capta uma direção de máxima variância nos dados.
- **Variação explicada**: indica a proporção da variabilidade dos dados representada por cada componente.
- **Variação acumulada**: utilizada para definir o número de componentes a serem retidos, considerando um limite de explicação satisfatória (geralmente acima de 85% da variância total).

2.1.2 Determinação do número ideal de *clusters*: Método do Cotovelo

Para a definição do número ideal de *clusters*, foi empregado o *método do cotovelo* (*Elbow Method*), técnica amplamente utilizada em combinações com o algoritmo *K-means*. Esse método consiste na execução iterativa do algoritmo para diferentes valores de k (quantidade de grupos), calculando-se, para cada valor, a soma dos quadrados intra-cluster (WCSS — *Within-Cluster Sum of Squares*) (KODINARIYA; MAKWANA, 2013).

A partir da análise gráfica da relação entre k e o WCSS, o ponto de inflexão — visualmente semelhante a um "cotovelo" — indica o valor ótimo de *clusters*, pois representa o ponto a partir do qual o ganho em homogeneidade dos grupos se torna marginal. Essa escolha visa o equilíbrio entre complexidade do modelo e qualidade da segmentação, evitando tanto o subajuste (poucos clusters) quanto o sobreajuste (muitos clusters).

2.1.3 Validação dos grupos: Índice de Silhueta

Complementando a definição do número de *clusters*, foi utilizado o *Índice de Silhueta*, métrica que avalia o grau de coesão interna e separação entre grupos. O valor da silhueta varia entre -1 e 1, sendo que valores próximos de 1 indicam que os elementos estão bem agrupados em relação ao seu próprio grupo e bem separados dos demais (ROUSSEEUW, 1987).

O uso combinado do método do cotovelo com o índice de silhueta fornece uma base sólida para a escolha do número ideal de *clusters*, assegurando segmentações consistentes e interpretáveis.

2.2 Classificação

A tarefa de classificação consiste em atribuir uma etiqueta (ou classe) a uma instância com base em seus atributos observáveis, sendo uma das abordagens mais amplamente utilizadas em problemas de aprendizado de máquina supervisionado. Nesse paradigma, o modelo é treinado com um conjunto de dados rotulado, aprendendo a identificar padrões que distinguem entre categorias distintas (HAN; KAMBER; PEI, 2011).

No contexto da saúde pública, a classificação preditiva assume papel estratégico ao possibilitar a antecipação de demandas por serviços, a identificação de grupos populacionais de risco e a tomada de decisões mais eficientes quanto à alocação de recursos (HOSSEINI; CHEN; ATUN, 2018). A integração entre variáveis demográficas, clínicas e socioeconômicas amplia a capacidade de predição e fornece uma base robusta para o planejamento de políticas públicas direcionadas (TOPOL, 2019).

Neste trabalho, foram analisadas quatro abordagens comumente utilizadas em tarefas de classificação: *Random Forest*, *Support Vector Machine* (SVM), *Multi-Layer*

Perceptron (MLP) e *Extreme Gradient Boosting* (XGB).

2.2.1 Random Forest

O algoritmo *Random Forest* é uma técnica baseada em *ensemble learning*, que combina múltiplas árvores de decisão para obter um modelo mais robusto e preciso. Proposto por Breiman (2001) (BREIMAN, 2001), o método consiste na construção de diversas árvores com subconjuntos aleatórios de dados e variáveis, utilizando a estratégia de *bootstrap* e divisão aleatória de atributos.

Cada árvore fornece uma predição individual, e o resultado final da floresta é determinado por meio de votação majoritária (no caso de classificação). Essa abordagem reduz a variância dos modelos e melhora a capacidade de generalização, sendo particularmente eficaz em situações com grande número de variáveis e interações não lineares (LIAW; WIENER, 2002).

No campo da saúde, o *Random Forest* tem sido utilizado com sucesso para prever desfechos clínicos, identificar fatores de risco e apoiar a tomada de decisão médica, destacando-se também por sua capacidade de avaliar a importância relativa das variáveis preditoras (TOPOL, 2019).

2.2.2 Support Vector Machine (SVM)

A *Support Vector Machine* (SVM) é uma técnica de aprendizado supervisionado proposta por Vapnik e Cortes nos anos 1990 (CORTES; VAPNIK, 1995). Trata-se de um classificador robusto que busca identificar o hiperplano ótimo que maximiza a margem de separação entre classes distintas.

Em cenários onde os dados não são linearmente separáveis, a SVM recorre ao uso de funções *kernel*, que permitem mapear os dados para espaços de maior dimensionalidade, facilitando a separação (HEARST et al., 1998). Essa abordagem é especialmente vantajosa em problemas com poucos exemplos rotulados e grande número de atributos, características comuns em bases de saúde.

Neste trabalho, a técnica de SVM foi implementada por meio da classe `SVC` (*Support Vector Classifier*) da biblioteca `scikit-learn`, que fornece uma interface prática e parametrizável para aplicação do modelo em dados tabulares.

Apesar de sua eficiência teórica, modelos baseados em SVM podem apresentar limitações em situações com forte desbalanceamento entre classes, exigindo cuidados adicionais no processo de validação e ajustes dos hiperparâmetros (GENG; LIU; ZHANG, 2015).

2.2.3 Multi-Layer Perceptron (MLP)

As redes neurais artificiais constituem uma das vertentes mais poderosas do aprendizado de máquina, sendo especialmente indicadas para problemas com relações complexas e não lineares entre variáveis. Dentre suas diversas arquiteturas, destaca-se o *Multi-Layer Perceptron* (MLP), uma rede do tipo *feedforward* composta por múltiplas camadas ocultas e funções de ativação não lineares (LECUN; BENGIO; HINTON, 2015).

O MLP aprende por meio do algoritmo de retropropagação, ajustando os pesos das conexões para minimizar o erro de predição. Essa arquitetura é amplamente utilizada em aplicações biomédicas e epidemiológicas, como na previsão de readmissões hospitalares e na classificação de perfis de risco (SHICKEL et al., 2018).

Apesar de seu alto desempenho preditivo, redes neurais como o MLP apresentam desvantagens em relação à interpretabilidade, o que pode limitar sua adoção em contextos onde a transparência dos modelos é essencial, como na gestão de políticas públicas em saúde (BAXT, 1995).

2.2.4 XGBoost

O *eXtreme Gradient Boosting* (XGBoost) é um algoritmo baseado em árvores de decisão que adota a técnica de *gradient boosting* para a construção sequencial de modelos. Proposto por Chen e Guestrin (2016) (CHEN; GUESTRIN, 2016), o XGBoost tem se destacado por sua alta eficiência, capacidade de paralelização e controle eficaz de *overfitting*.

Diferente do *Random Forest*, o XGBoost constrói os modelos de forma aditiva, onde cada nova árvore busca corrigir os erros cometidos pelas anteriores. Além disso, incorpora regularizações L1 e L2 em sua função de custo, o que contribui para a robustez e generalização do modelo. A regularização L1, também conhecida como *Lasso*, adiciona à função de custo o valor absoluto dos coeficientes dos parâmetros, incentivando a esparsidade do modelo ao reduzir alguns coeficientes a zero, o que pode eliminar variáveis irrelevantes. Já a regularização L2, ou *Ridge*, utiliza o quadrado dos coeficientes, penalizando valores muito grandes e promovendo modelos mais suaves e menos propensos ao ajuste excessivo aos dados de treino. A combinação dessas penalizações permite ao XGBoost controlar a complexidade do modelo e evitar sobreajuste, especialmente em bases com muitos atributos.

No domínio da saúde, o XGBoost tem sido amplamente utilizado em tarefas como predição de risco cardiovascular, triagem de pacientes e alocação inteligente de recursos, com desempenho superior a muitos algoritmos tradicionais (NIELSEN, 2016).

2.2.5 Métricas de avaliação

No contexto da saúde pública, onde as decisões baseadas em modelos podem impactar diretamente a alocação de recursos e o atendimento à população, a escolha criteriosa das métricas de avaliação é ainda mais crítica (SIDEY-GIBBONS; SIDEY-GIBBONS, 2020).

As métricas fornece uma perspectiva diferente sobre o desempenho do modelo, sendo indicadas para diferentes contextos de desequilíbrio entre classes ou custos associados a falsos positivos e falsos negativos. Dentre as métricas mais utilizadas em problemas de classificação binária, destacam-se:

- **Acurácia** (*accuracy*): representa a proporção total de predições corretas sobre o total de observações. Embora amplamente utilizada, a acurácia pode ser enganosa em cenários com classes desbalanceadas, pois tende a favorecer a classe majoritária (FAWCETT, 2006).
- **Precisão** (*precision*): é a razão entre os verdadeiros positivos e o total de positivos preditos pelo modelo. Essa métrica é particularmente relevante quando o custo de um falso positivo é alto, como em diagnósticos médicos, onde uma indicação incorreta de doença pode gerar ansiedade e gastos desnecessários com exames (SAITO; REHMSMEIER, 2015).
- **Revocação** (*recall*): também chamada de sensibilidade, mede a proporção de verdadeiros positivos identificados corretamente pelo modelo. No contexto de saúde pública, essa métrica é essencial quando se busca garantir que todos os indivíduos de risco sejam corretamente identificados, minimizando o número de casos não detectados (POWERS, 2011).
- **F1-Score**: é a média harmônica entre precisão e revocação, fornecendo um equilíbrio entre as duas métricas. É especialmente útil em problemas com classes desbalanceadas, pois penaliza modelos que favorecem um indicador em detrimento do outro (GOUTTE; GAUSSIER, 2005).

Além das métricas mencionadas, a **matriz de confusão** também foi utilizada neste trabalho como instrumento visual para avaliar o desempenho do modelo. Essa matriz organiza as predições em quatro categorias: verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, permitindo uma análise mais granular dos acertos e erros cometidos pelo classificador.

A **importância das variáveis** também foi considerada como métrica complementar de avaliação, permitindo identificar quais atributos mais influenciam nas decisões do modelo. Técnicas como *Permutation Importance* foram utilizadas para mensurar o impacto de cada

variável sobre o desempenho preditivo, favorecendo a interpretabilidade e a transparência do modelo.

Como reforçado por (FERRI; HERNáNDEZ-ORALLO; MODROIU, 2009), não existe uma métrica universalmente superior, sendo necessário considerar os objetivos do estudo e os riscos associados a erros de classificação para definir as métricas mais apropriadas.

2.3 Tecnologias

A escolha das tecnologias utilizadas neste trabalho foi orientada pela necessidade de manipular grandes volumes de dados. Para isso, adotou-se uma abordagem híbrida, utilizando as linguagens R e Python, com execução primária no ambiente do **Google Colab**, uma plataforma baseada na nuvem que permite o desenvolvimento e execução de código Python com acesso facilitado a recursos computacionais e bibliotecas especializadas (BISONG, 2019).

As tecnologias foram organizadas segundo as etapas do estudo: pré-processamento dos dados, agrupamento e classificação. A seguir, detalham-se as bibliotecas empregadas em cada uma dessas fases.

2.3.1 Tecnologias no pré-processamento

O tratamento dos microdados da Pesquisa Nacional de Saúde (PNS) 2019 foi conduzido na linguagem R, escolhida por sua robustez em análises estatísticas e manipulação de dados amostrais complexos. As bibliotecas utilizadas foram:

- **PNSIBGE** (LIMA, 2023): biblioteca especializada na estrutura da PNS, permitindo importação e organização dos microdados com aplicação direta dos pesos amostrais.
- **survey** (LUMLEY, 2023): fornece suporte para análise estatística considerando o desenho amostral complexo da PNS, incluindo estratificação e conglomerados.
- **openxlsx** (SCHAUBERGER; WALKER, 2023): possibilita leitura e escrita de arquivos Excel de forma eficiente, sem dependências externas.
- **dplyr** (WICKHAM; FRANçOIS, 2023): facilita a manipulação de dados tabulares com sintaxe intuitiva para filtragem, agrupamento e transformação.
- **ggplot2** (WICKHAM, 2023a): base para construção de gráficos sofisticados, utilizada na visualização exploratória e apresentação dos dados.
- **corrplot** (WEI; SIMKO, 2023): permite gerar mapas de calor de correlação entre variáveis, úteis para avaliar colinearidade.

- **reshape2** (WICKHAM, 2023b): utilizada para reorganização da estrutura dos dados, convertendo formatos entre *wide* e *long*.
- **igraph** (CSARDI; NEPUSZ, 2023) e **ggraph** (PEDERSEN, 2023): empregadas na construção e visualização de grafos de correlação, destacando as relações estruturais entre variáveis.
- **ggrepel** (SLOWIKOWSKI, 2023): aprimora a legibilidade dos gráficos ao posicionar rótulos de forma inteligente.

2.3.2 Tecnologias no agrupamento

A etapa de agrupamento foi realizada em Python, linguagem amplamente utilizada em projetos de ciência de dados, em conjunto com bibliotecas voltadas à modelagem, visualização e análise exploratória:

- **NumPy** (HARRIS, 2023): base para operações vetorizadas e manipulação de matrizes numéricas de alta performance.
- **Pandas** (TEAM, 2023): biblioteca central para manipulação de *DataFrames*, utilizada no tratamento e organização dos dados.
- **Matplotlib** (HUNTER, 2023) e **Seaborn** (WASKOM, 2023): ferramentas para construção de gráficos estáticos e análises exploratórias visuais.
- **Plotly** (INC., 2023): utilizada para visualizações interativas, facilitando a interpretação de clusters e variáveis em múltiplas dimensões.
- **NetworkX** (HAGBERG; SCHULT; SWART, 2023): permite a análise de estruturas de grafos, representando relações entre variáveis.
- **Itertools** (FOUNDATION, 2023): módulo nativo do Python utilizado para gerar combinações e permutações entre variáveis.
- **Scikit-learn** (PEDREGOSA, 2023): principal biblioteca de aprendizado de máquina, da qual foram utilizadas:
  - **KMeans**: algoritmo de agrupamento baseado em similaridade.
  - **PCA (Principal Component Analysis)**: técnica de redução de dimensionalidade para visualização dos dados.
  - **silhouette_score** e **silhouette_samples**: métricas para avaliação da qualidade dos clusters.
  - **MinMaxScaler**: responsável pela normalização dos dados para manter as variáveis na mesma escala.

2.3.3 Tecnologias na classificação

Na etapa de classificação, os modelos preditivos foram desenvolvidos em Python com o suporte de bibliotecas voltadas para modelagem estatística, aprendizado de máquina e análise de desempenho.

- **NumPy** (HARRIS, 2023) e **Pandas** (TEAM, 2023): empregadas para estruturação, manipulação e tratamento dos dados de entrada dos modelos.
- **Matplotlib** (HUNTER, 2023), **Seaborn** (WASKOM, 2023) e **Plotly** (INC., 2023): utilizadas para criação de visualizações das métricas de avaliação e interpretação dos resultados.
- **tqdm** (AL., 2023): biblioteca para monitoramento de execuções iterativas por meio de barras de progresso.
- **scikit-learn** (PEDREGOSA, 2023): utilizada para a implementação dos algoritmos *Random Forest*, `SVC` (implementação da técnica *Support Vector Machine*) e *MLP*, bem como para:
  - divisão de dados (`train_test_split`),
  - validação cruzada (`StratifiedKFold`),
  - otimização de hiperparâmetros (`GridSearchCV`),
  - cálculo de métricas (`accuracy`, `precision`, `recall`, `f1-score`, `confusion_matrix`),
  - avaliação da importância das variáveis (`permutation_importance`).
- **xgboost** (CHEN; GUESTRIN, 2023): biblioteca especializada para o modelo *XGBoost*, amplamente adotada por seu desempenho superior em classificações com dados tabulares.
- **joblib** e **pickle**: responsáveis pela serialização dos modelos, garantindo reprodutibilidade e agilidade nos experimentos.
- **IPython.display**: utilizada para exibição clara de resultados em ambientes interativos como o Jupyter Notebook ou Google Colab.

2.4 Trabalhos relacionados

O uso de modelos preditivos na área da saúde tem crescido consideravelmente nas últimas décadas, impulsionado pelo avanço de tecnologias computacionais e pela maior disponibilidade de grandes volumes de dados. Tais modelos têm se mostrado promissores para prever demandas por serviços, otimizar a alocação de recursos e subsidiar decisões estratégicas em políticas públicas (BATES et al., 2014; HOSSEINI; CHEN; ATUN, 2018).

2.4.1 Fatores determinantes no uso de serviços de saúde

A literatura aponta que o uso de serviços de saúde é influenciado por uma combinação de fatores clínicos, demográficos e socioeconômicos. Doenças crônicas, como hipertensão, diabetes e obesidade, estão entre os principais determinantes da frequência ao atendimento médico (VOS et al., 2017). Além disso, aspectos como idade avançada, baixa escolaridade e menor renda são associados a maior utilização dos serviços, especialmente em contextos de desigualdade social (ZHAO; ATUN; ANINDYA, 2021).

No Brasil, estudos como o de Romero, Costa e Moraes (2018) destacam a elevada prevalência de condições como dor lombar crônica, que têm impacto significativo sobre a procura por atendimento na rede pública. Outros fatores relevantes incluem a percepção negativa do próprio estado de saúde e a ausência de plano de saúde, conforme evidenciado por Mendonça et al. (2021).

2.4.2 Uso da pesquisa nacional de saúde (PNS)

A Pesquisa Nacional de Saúde (PNS), conduzida pelo Instituto Brasileiro de Geografia e Estatística (IBGE), constitui uma das mais importantes fontes de dados sobre a saúde da população brasileira. Seus microdados abrangem informações detalhadas sobre condições clínicas, fatores de risco, acesso aos serviços e perfil socioeconômico da população (Instituto Brasileiro de Geografia e Estatística (IBGE), 2020).

Apesar de sua abrangência, a base apresenta limitações. Segundo Barros e Victora (2019), os dados autorrelatados podem estar sujeitos a viés de memória e subnotificação. Além disso, a sub-representação de determinados grupos pode comprometer a robustez estatística das análises preditivas, exigindo o uso de técnicas específicas de ponderação e correção amostral.

2.4.3 Modelos preditivos na saúde pública

Em escala global, a aplicação de modelos preditivos tem se mostrado eficaz para identificar populações vulneráveis, prever surtos epidemiológicos e direcionar recursos de forma mais estratégica (TOPOL, 2019; BATES et al., 2014). No Brasil, essas técnicas têm sido apontadas como ferramentas valiosas para melhorar a eficiência do Sistema Único de Saúde (SUS), especialmente diante de recursos limitados e alta demanda.

Pesquisas como a de Simões, Meira e Santos (2021) demonstraram a viabilidade de utilizar dados da PNS para prever padrões de atendimento e propor ações baseadas em evidências. Da mesma forma, Zhao, Atun e Anindya (2021) evidenciaram a capacidade de modelos baseados em aprendizado de máquina para estimar com precisão os custos de pacientes com doenças crônicas, possibilitando intervenções mais eficazes.

2.4.4 Avanços tecnológicos e desafios

Os avanços em ciência de dados e inteligência artificial possibilitaram melhorias significativas no armazenamento, processamento e análise de dados em saúde. No entanto, ainda existem desafios importantes a serem superados. Entre eles, destacam-se a necessidade de padronização dos dados, a explicabilidade dos modelos e a garantia da privacidade dos indivíduos (SHICKEL et al., 2018; LECUN; BENGIO; HINTON, 2015).

A integração da PNS com outras bases administrativas, como os sistemas do DATASUS, é uma estratégia recomendada para enriquecer os modelos preditivos e ampliar sua aplicabilidade prática (BARROS; VICTORA, 2019).

2.4.5 Perspectivas futuras

As perspectivas para o uso de modelos preditivos em saúde pública são amplamente promissoras. Tendências como o uso de dados em tempo real, o aprimoramento de técnicas de *deep learning* e o desenvolvimento de sistemas de apoio à decisão baseados em evidências devem fortalecer ainda mais essa área nos próximos anos (TOPOL, 2019).

Com a ampliação da digitalização da saúde e o fortalecimento da governança de dados, espera-se que esses modelos se tornem instrumentos centrais na formulação de políticas públicas, contribuindo para maior equidade, eficiência e qualidade nos serviços oferecidos à população.

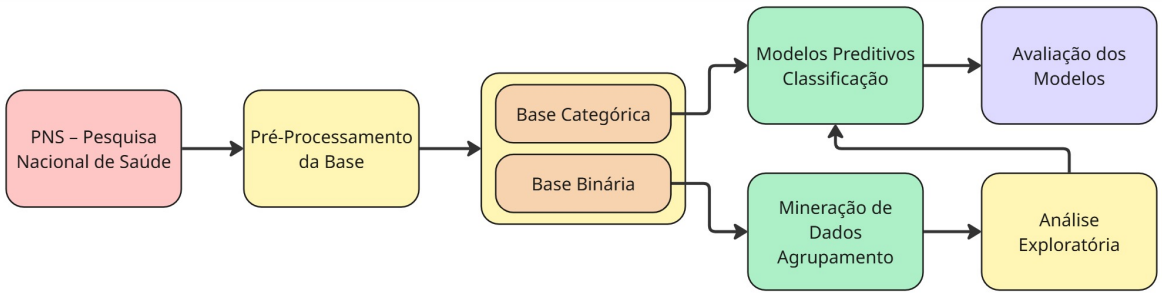
2.5 Considerações finais

A literatura revisada destaca o potencial dos modelos preditivos para otimizar a gestão dos sistemas de saúde. A combinação de dados públicos, algoritmos avançados e estratégias de implementação adaptadas à realidade brasileira pode contribuir significativamente para a melhoria da eficiência e equidade no acesso aos serviços de saúde. Além disso, a escolha dos algoritmos utilizados, fundamentada em estudos anteriores, reforça a robustez do modelo desenvolvido neste trabalho.

3 Metodologia

Este capítulo descreve, de forma sistemática, as etapas adotadas para a construção do modelo preditivo proposto, seguindo a estrutura metodológica apresentada na Figura 1. O fluxo contempla desde a obtenção da base de dados até a avaliação dos modelos preditivos.

Figura 1 – Fluxograma da metodologia aplicada no estudo.



Fonte: Elaborado pelo autor.

3.1 Obtenção da base de dados

O ponto de partida da metodologia foi a utilização dos dados da Pesquisa Nacional de Saúde (PNS) 2019, conduzida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em parceria com o Ministério da Saúde (Instituto Brasileiro de Geografia e Estatística (IBGE), 2020).

3.2 Pré-processamento dos dados

Na etapa de pré-processamento as principais atividades realizadas foram:

- **Filtragem e seleção de variáveis**: seleção criteriosa de atributos foi realizada com base em evidências da literatura, análise estatística e recomendações de especialistas da área da saúde.
- **Remoção de dados inválidos**: foram excluídos registros de indivíduos com menos de 18 anos, bem como observações com valores ausentes em variáveis essenciais como peso, altura e uso de serviços de saúde.

- **Transformação de variáveis em binários**: variáveis com respostas "Sim" e "Não" foram codificadas em valores numéricos (1 e 0, respectivamente), permitindo sua utilização por algoritmos que exigem entradas numéricas.
- **Junção de colunas**: diagnósticos de doenças como diabetes e hipertensão durante a gestação foram incorporados às variáveis principais, ampliando a representatividade dos históricos clínicos (SHICKEL et al., 2018). Como também as variáveis de uso de serviços de saúde e internações foram unificadas em uma única variável.
- **Geração de variáveis derivadas**: foram criadas novas variáveis, como faixas etárias, cálculo de IMC e grupos de percepção da saúde (World Health Organization, 2000).
- **Cálculo do peso amostral**: para assegurar a representatividade populacional, foram utilizados os pesos amostrais fornecidos pela PNS, conforme metodologia recomendada para análises com dados de amostragem complexa (BARROS; VICTORA, 2019).
- **Tratamento de valores ausentes**: valores ausentes foram tratados por substituição padrão — utilizando -1 para criação da base categórica numérica e 0 para criação da base binária — evitando a exclusão de registros e mantendo a integridade da amostra (SAAR-TSECHANSKY; PROVOST, 2007).
- **Codificação supervisionada (*Target Encoding*)**: Para variáveis categóricas, aplicou-se o *Target Encoding*, técnica que substitui as categorias por valores numéricos baseados na média da variável alvo para cada grupo.
- **Balanceamento da base com *NearMiss-1***: Como estratégia para lidar com o desbalanceamento entre classes da variável alvo, foi adotada a técnica de *undersampling NearMiss-1*, que seleciona exemplos da classe majoritária com base em sua proximidade com instâncias da classe minoritária (MANI; ZHANG, 2003).

3.3 Divisão em bases específicas

Para atender aos diferentes objetivos analíticos, o conjunto de dados foi particionado em duas versões:

- **Base binária**: contendo todos os atributos selecionados e tratados em variáveis binárias, utilizada na etapa de agrupamento.
- **Base categórica**: contendo todos os atributos selecionados e tratados em formato categórico numérico, destinada à modelagem preditiva.

3.4 Mineração de dados: Agrupamento

Com base na **base binária**, foi aplicado o algoritmo *K-means*, técnica amplamente utilizada para segmentação de dados pela similaridade entre observações (TAN; STEINBACH; KUMAR, 2013). O número ideal de clusters foi determinado por meio do *método do cotovelo* e do *índice de silhueta*, garantindo a formação de grupos coesos e bem separados (ROUSSEEUW, 1987).

3.5 Análise exploratória

Após o agrupamento, foi realizada uma análise exploratória dos clusters formados e da distribuição das variáveis, com ênfase na caracterização epidemiológica de cada grupo. Essa etapa permitiu contextualizar os perfis identificados e fundamentar as etapas posteriores de classificação e interpretação dos modelos.

3.6 Modelagem preditiva: Classificação

Com base na **base categórica**, foi desenvolvida uma modelagem preditiva para estimar a probabilidade de uso dos serviços de saúde. Foram avaliados quatro algoritmos de aprendizado supervisionado:

- **Random Forest** (BREIMAN, 2001): modelo baseado em múltiplas árvores de decisão, robusto a ruídos e interpretável.
- **Support Vector Machine** (SVM) (CORTES; VAPNIK, 1995): técnica eficaz em dados de alta dimensionalidade e com margens de separação não triviais. Neste trabalho, a implementação foi realizada por meio da classe `SVC` da biblioteca `scikit-learn`.
- **Multi-Layer Perceptron** (MLP) (LECUN; BENGIO; HINTON, 2015): rede neural com múltiplas camadas ocultas, capaz de capturar relações não lineares entre variáveis.
- **Extreme Gradient Boosting** (XGB) (CHEN; GUESTRIN, 2016): algoritmo de *gradient boosting* com alta capacidade preditiva e controle de *overfitting*.

Os modelos foram treinados com validação cruzada estratificada (*Stratified K-Fold Cross Validation*), promovendo a estabilidade das estimativas de desempenho.

3.7 Avaliação dos modelos

A avaliação dos modelos preditivos foi realizada com base nas métricas clássicas de desempenho para classificação binária (POWERS, 2020):

- **Acurácia**: mede a proporção total de previsões corretas.
- **Precisão**: indica a taxa de verdadeiros positivos entre os classificados como positivos.
- **Revocação (Recall)**: identifica a capacidade do modelo em detectar corretamente os casos positivos.
- **F1-Score**: média harmônica entre precisão e revocação, útil em cenários com desbalanceamento de classes.

Adicionalmente, utilizou-se a técnica de *Permutation Importance* para interpretar a contribuição individual de cada variável no desempenho dos modelos, fortalecendo a explicabilidade dos resultados e facilitando a aplicação prática no contexto da saúde pública.

Também foi aplicada a **matriz de confusão**, permitindo uma análise detalhada da distribuição entre verdadeiros e falsos positivos e negativos, o que contribui para a compreensão da qualidade das predições em cada classe.

Essas abordagens complementares foram fundamentais para garantir uma avaliação abrangente dos modelos, considerando tanto o desempenho quantitativo quanto a interpretabilidade dos resultados.

4 Caracterização da base

Este capítulo descreve o processo de construção e preparação da base de dados utilizada neste estudo, originada a partir da Pesquisa Nacional de Saúde (PNS) de 2019. São abordadas as etapas de coleta, filtragem e transformação das variáveis, bem como os procedimentos de pré-processamento, balanceamento das classes e definição das bases resultantes para as análises de agrupamento e classificação. O objetivo foi garantir a qualidade e a estruturação adequada dos dados, viabilizando a aplicação das técnicas de mineração de dados com maior precisão e coerência.

4.1 Coleção dados

No desenvolvimento deste trabalho, foi realizado o pré-processamento dos dados da Pesquisa Nacional de Saúde PNS de 2019, uma etapa essencial para a construção do modelo preditivo. Esta fase incluiu a aplicação de diversas técnicas de limpeza e transformação de dados, eliminando inconsistências, padronizando formatos e criando novas variáveis consideradas relevantes para a análise.

A preparação minuciosa dos dados teve como objetivo garantir a qualidade e a consistência da base, além de possibilitar a identificação de padrões e tendências nos dados. Esses elementos foram utilizados na etapa subsequente de clusterização e treinamento e avaliação o modelo preditivo. Ao final deste processo de pré-processamento, foi obtido um conjunto de dados estruturado e devidamente preparado, proporcionando maior precisão e confiabilidade aos resultados obtidos no decorrer da pesquisa.

4.1.1 Filtragem e seleção das variáveis

Após a importação e leitura dos dados da Pesquisa Nacional de Saúde PNS 2019, foram realizadas etapas de seleção e organização dos dados para criar um conjunto mais conciso e focado nas variáveis de interesse para a pesquisa. Estas etapas incluíram a filtragem das colunas relevantes e a renomeação de suas variáveis, conforme descrito abaixo.

As variáveis selecionadas na Tabela 1 foram renomeadas com títulos mais descritivos, visando melhorar a interpretação e a clareza das análises realizadas. Abaixo, apresentamos as variáveis selecionadas e respectivas descrições:

Tabela 1 – Descrição das variáveis selecionadas.

Nome	**Descrição**
Peso	Peso do morador selecionado
UF	Unidade Federativa UF de residência
Extrato	Estrato amostral da pesquisa
Zona	Zona de residência (1: Rural; 2: Urbana)
OrdemDomicilio	Número da ordem do domicílio na PNS
GrauEnsino	Grau de escolaridade do entrevistado
Renda	Faixa de renda mensal
CadastroUBS	Cadastro do domicílio na Unidade de Saúde da Família
AgenteComunitario	Frequência de visitas de agentes comunitários
NumeroOrdem	Número de ordem do morador
Sexo	Sexo do entrevistado (1: Homem; 2: Mulher)
Idade	Idade do entrevistado, em anos
Raca	Cor ou raça (1: Branca; 2: Preta; 3: Amarela; 4: Parda; 5: Indígena)
EstadoCivil	Estado civil do entrevistado
LerEscrever	Capacidade de leitura e escrita (1: Sim; 2: Não)
Escola	Frequência à escola (inclui pré-escola ao doutorado)
Trabalha	Situação de trabalho (1: Sim; 2: Não)
PlanoSaude	Possui plano de saúde particular
EstadoSaude	Autoavaliação do estado de saúde (1: Saúde Muito Bom, 2: Saúde Bom, 3: Saúde Regular, 4: Saúde Ruim, 5: Saúde Muito Ruim)
ServicoSaude	Procurou atendimento de saúde nas últimas duas semanas
Internacao	Internação nos últimos 12 meses
Alcool	Frequência de consumo de bebidas alcoólicas
ExercicioFisico	Prática de exercícios físicos (1: Sim; 2: Não)
Tabaco	Uso de tabaco (1: Sim, diariamente; 2: Sim, esporadicamente; 3: Não)
Peso	Peso corporal em quilos
Altura	Altura em metros
ColesterolAlto	Diagnóstico de colesterol alto (1: Sim; 2: Não)
AVC	Diagnóstico de AVC ou derrame
AVC_Limita	Grau de limitação causado pelo AVC (1: Não limita; 2: Um pouco; 3: Moderadamente; 4: Intensamente; 5: Muito intensamente)
Asma	Diagnóstico de asma (1: Sim; 2: Não; 3: Ignorado)

Nome	**Descrição**
Asma_Limita	Grau de limitação causado pela asma (1: Não limita; 2: Um pouco; 3: Moderadamente; 4: Intensamente; 5: Muito intensamente)
Artrite	Diagnóstico de artrite ou reumatismo
Artrite_Limita	Grau de limitação causado pela artrite ou reumatismo
DorColuna	Presença de problema crônico de coluna (1: Sim; 2: Não; 3: Ignorado)
DorColuna_Limita	Grau de limitação causado por problemas de coluna
DORT	Diagnóstico de DORT (distúrbio osteomuscular relacionado ao trabalho)
DORT_Limita	Grau de limitação causado pelo DORT
Depressao	Diagnóstico de depressão (1: Sim; 2: Não; 3: Ignorado)
Depressao_Limita	Grau de limitação causado pela depressão
Hipertensao	Diagnóstico de hipertensão arterial
Hipertensao_Limita	Grau de limitação causado pela hipertensão
Diabetes	Diagnóstico de diabetes (1: Sim; 2: Não; 3: Ignorado)
Diabetes_Limita	Grau de limitação causado pela diabetes
Cardiopatia	Diagnóstico de cardiopatia, como infarto ou insuficiência cardíaca
Cardiopatia_Limita	Grau de limitação causado pela cardiopatia
Cancer	Diagnóstico de câncer
Cancer_Limita	Grau de limitação causado pelo câncer
InsufRenal	Diagnóstico de insuficiência renal crônica
InsufRenal_Limita	Grau de limitação causado pela insuficiência renal
DoencaMental	Diagnóstico de doença mental
DoencaMental_Limita	Grau de limitação causado pela doença mental
DoencaPulmao	Diagnóstico de doença pulmonar crônica, como enfisema
DoencaPulmao_Limita	Grau de limitação causado por doença pulmonar crônica

Fonte: Produzido pelos autores.

4.2 Pré-processamento

4.2.1 Seleção de dados válidos

A primeira etapa do pré-processamento consistiu na seleção de dados válidos. Inicialmente, foi aplicado um filtro etário para manter no conjunto apenas indivíduos com 18 anos ou mais, de modo a garantir que a análise estivesse focada na população

adulta. Essa escolha justifica-se tanto do ponto de vista metodológico quanto da relevância epidemiológica, considerando que adultos apresentam maior autonomia no uso dos serviços de saúde e maior probabilidade de desenvolver condições crônicas.

Em seguida, foi realizada a remoção de registros com informações ausentes em variáveis consideradas essenciais para o estudo, especificamente as variáveis *Peso*, *Altura* e *Serviço de Saúde*. A ausência desses dados comprometeria tanto a modelagem quanto a integridade das análises descritivas e preditivas. A exclusão desses registros visou preservar a consistência da base e evitar viéses decorrentes de imputações inadequadas em variáveis críticas.

4.2.2 Transformação de variáveis categóricas binárias

Para viabilizar a aplicação de algoritmos de aprendizado de máquina, que exigem entradas numéricas, foram realizadas transformações em variáveis categóricas binárias. Respostas textuais como "Sim" e "Não" foram convertidas em valores numéricos, sendo representadas por 1 e 0, respectivamente. Essa transformação foi fundamental para que essas variáveis pudessem ser interpretadas pelos modelos computacionais sem perda de significado semântico.

4.2.3 Junção de colunas e reforço de variáveis clínicas

Primeiramente, as colunas *Internação* e *Serviço de Saúde* foram combinadas em uma única variável denominada *Serviço de Saúde*. Essa junção teve como finalidade simplificar a estrutura do conjunto de dados e unificar diferentes formas de utilização dos serviços médicos — tanto atendimentos ambulatoriais quanto internações — em uma única dimensão analítica.

Além disso, visando tornar mais representativo o histórico clínico dos indivíduos, foram incorporadas informações adicionais às variáveis Diabetes e Hipertensão. Casos em que o diagnóstico dessas condições ocorreu durante a gestação foram incluídos, mesmo que a doença não tenha persistido após esse período. Essa decisão foi adotada com base na literatura que recomenda considerar todo o histórico de exposição a condições de saúde relevantes em análises preditivas, sobretudo em contextos populacionais.

4.2.4 Geração de novas variáveis

Nesta etapa, novas variáveis foram geradas a partir da transformação de variáveis originais, com o objetivo de adaptar os dados para diferentes estratégias analíticas. A construção dessas variáveis teve como finalidade organizar o conjunto de dados em duas bases distintas: uma contendo exclusivamente variáveis binárias, voltada para os algoritmos

de agrupamento; e outra com variáveis numéricas categóricas, direcionada aos algoritmos de classificação.

As principais transformações realizadas foram:

A variável *Idade* foi categorizada em faixas etárias, como "18 a 29", "30 a 39", "40 a 59"e "60 ou mais", e convertida em variáveis binárias para a base de agrupamento. O mesmo processo foi aplicado às variáveis *Peso* e *Altura*, convertidas em categorias como "Peso 57 a 82"e "Altura 157 a 170", facilitando a modelagem com algoritmos sensíveis a escalas distintas.

A variável *Renda domiciliar per capita* também foi segmentada em faixas representativas de distribuição socioeconômica, como "Per capita 1", "Per capita 2", entre outras, refletindo a estratificação econômica dos indivíduos.

A variável *Estado Saúde* foi categorizada em faixas binárias com base nas opções de resposta disponíveis na Pesquisa Nacional de Saúde, respeitando a autoavaliação feita pelos próprios participantes. As categorias originais — “Saúde Muito Boa”, “Saúde Boa”, “Saúde Regular”, “Saúde Ruim” e “Saúde Muito Ruim” — foram convertidas em variáveis indicadoras, possibilitando a análise da percepção subjetiva de saúde da população.

No que diz respeito à região geográfica, a variável *UF* foi utilizada para criar indicadores binários para as cinco grandes regiões brasileiras: Norte, Nordeste, Centro-Oeste, Sudeste e Sul. Essa transformação permitiu uma análise mais agregada do fator territorial, respeitando as disparidades regionais de acesso à saúde no Brasil.

A variável *Raça* foi expandida em variáveis binárias para cada categoria (como branca, preta, parda, amarela e indígena), assegurando sua inclusão na base binária sem perda de informação.

Além disso, foi calculado o Índice de Massa Corporal (IMC), a partir das variáveis *Peso* e *Altura*, e posteriormente categorizado em faixas clínicas binárias, como “Peso normal”, “Sobrepeso” e “Obesidade”.

4.2.5 Cálculo do peso amostral

O cálculo dos pesos foi realizado com base nas variáveis *Unidade da Federação (UF)*, *Estrato*, *Peso amostral original (V0091)*, *Ordem do Domicílio* e *Número de Ordem*. Essas variáveis refletem a estrutura da amostragem da PNS e permitem ponderar cada observação de acordo com sua probabilidade de seleção, garantindo que as estimativas obtidas reproduzam fielmente a distribuição populacional do país.

4.2.6 Tratamento de valores ausentes

Para a base categórica destinada à modelagem preditiva por algoritmos de classificação, os valores ausentes foram substituídos pelo valor -1. Já na base binária utilizada nos algoritmos de agrupamento, os valores ausentes foram substituídos por 0. Essa diferenciação considerou as exigências específicas de cada tipo de algoritmo e evitou a exclusão de observações, como recomendado por Saar-Tsechansky e Provost (2007).

4.2.7 Codificação da variável alvo (*Target Encoding*)

Para aprimorar a representação de variáveis categóricas e garantir melhor desempenho dos algoritmos de classificação, foi aplicada a técnica de *Target Encoding* sobre variáveis qualitativas presentes na base de dados. Foram selecionadas as variáveis *Zona*, *Sexo*, *Raça*, e *Estado Civil*. Para cada uma dessas variáveis, calculou-se a média do valor da variável alvo *Serviço de Saúde* para cada categoria. Em seguida, os valores categóricos originais foram substituídos por essas médias numéricas.

4.2.8 Bases resultantes

Após a aplicação das etapas de pré-processamento descritas anteriormente, foram geradas duas bases de dados distintas, cada uma com características específicas para atender às diferentes técnicas utilizadas no estudo.

A **base binária** foi construída para ser utilizada nas etapas de agrupamento. Nela, todas as variáveis foram convertidas para representações binárias (0 e 1), incluindo aquelas originalmente contínuas ou categóricas, que foram transformadas em faixas indicadoras. Essa transformação aumentou consideravelmente o número de variáveis, permitindo uma análise mais granular de perfis populacionais com características semelhantes.

Já a **base categórica** (ou base real), utilizada para os modelos de classificação supervisionada, manteve as variáveis em formato numérico contínuo ou categórico codificado. Com um número menor de atributos, essa base permitiu a aplicação de algoritmos de classificação com maior eficiência computacional, sem comprometer a expressividade dos dados.

A Tabela 2 apresenta a comparação entre as duas bases quanto ao número de observações e variáveis.

Tabela 2 – Características das bases resultantes após o pré-processamento.

Base	**Linhas**	**Variáveis**	**Tipo de variáveis**
Binária	87.678	161	Binárias (0 ou 1)
Categórica (Reais)	87.678	61	Numéricas e categóricas

Fonte: Elaboração própria.

Nota: A base binária possui mais variáveis devido à conversão de variáveis contínuas e categóricas em faixas indicadoras.

Ambas as bases apresentaram o mesmo número de registros (87.678), porém com desequilíbrio entre as classes da variável-alvo — *uso de serviços de saúde*. A Tabela 3 mostra essa distribuição, evidenciando a predominância da classe que não utilizou os serviços.

Tabela 3 – Distribuição da variável-alvo: uso de serviços de saúde.

Classe	**Quantidade**	**Proporção (%)**
Usaram o serviço de saúde (1)	18.267	20,8%
Não usaram o serviço de saúde (0)	69.411	79,2%

Fonte: Elaboração própria.

Esse desbalanceamento pode prejudicar a capacidade dos algoritmos de classificação em aprender corretamente a identificar a classe minoritária. Por esse motivo, em etapas posteriores, técnicas como o *NearMiss-1* foram aplicadas para reequilibrar a base e otimizar os resultados dos modelos preditivos.

4.2.9 Balanceamento da base categórica

Para mitigar o problema do desbalanceamento da variável alvo deste estudo, *Serviço de Saúde*, foi aplicado o método de *undersampling* conhecido como **NearMiss**, especificamente a versão **NearMiss-1**. Dessa forma, mantém-se a estrutura dos dados enquanto se reduz o viés em relação à classe predominante.

Após a aplicação do *NearMiss-1*, o conjunto de dados resultante passou a ter distribuição equitativa entre as classes, com 50% dos indivíduos classificados como usuários dos serviços de saúde e 50% como não usuários. Esse balanceamento é essencial para promover um aprendizado mais justo e eficaz pelos modelos preditivos, aumentando a acurácia na identificação dos padrões de uso dos serviços.

A Tabela 4 apresenta a nova distribuição das classes após o balanceamento.

Tabela 4 – Distribuição balanceada da variável-alvo: uso de serviços de saúde.

Classe	**Quantidade**	**Proporção (%)**
Usaram o serviço de saúde (1)	22.254	50,0%
Não usaram o serviço de saúde (0)	22.254	50,0%

Fonte: Elaboração própria.

4.3 Agrupamento (*clustering*)

Nesta etapa, foram aplicadas técnicas de agrupamento aos dados da Pesquisa Nacional de Saúde (PNS) 2019, previamente pré-processados. O objetivo foi identificar perfis de indivíduos com características semelhantes em relação à presença de doenças crônicas e ao uso de serviços de saúde. Essa análise exploratória permitiu reconhecer padrões não supervisionados nos dados, contribuindo para uma compreensão mais aprofundada dos grupos populacionais com diferentes condições de saúde e seus comportamentos em relação ao atendimento médico.

4.3.1 Filtragem e seleção das variáveis

Para a realização do agrupamento, a base foi ajustada para conter somente as variáveis diretamente relacionadas a condições crônicas de saúde. A seleção dessas variáveis foi fundamentada em literatura especializada e na orientação de profissionais da área da saúde, garantindo a relevância clínica e epidemiológica dos atributos utilizados.

A filtragem teve como objetivo reduzir a complexidade do conjunto de dados, facilitando a interpretação dos agrupamentos gerados. As variáveis selecionadas refletem doenças e condições de saúde com potencial impacto sobre a demanda por serviços médicos, conforme listado a seguir:

As variáveis consideradas nesta etapa incluíram: hipertensão, diabetes, colesterol alto, cardiopatia, acidente vascular cerebral (AVC) ou derrame, asma, artrite ou reumatismo, distúrbios osteomusculares relacionados ao trabalho (DORT), depressão, doença mental, doença crônica no pulmão, câncer, insuficiência renal crônica, dor na coluna e obesidade. Todas essas condições apresentam reconhecida associação com maior demanda por cuidados médicos, justificando sua inclusão no processo de agrupamento.

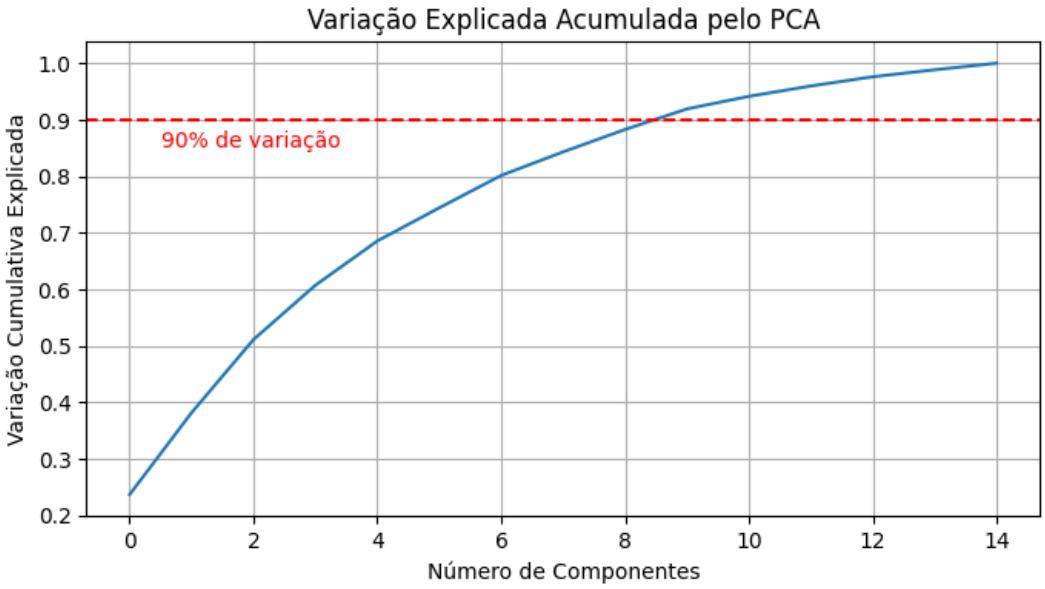
4.3.2 Análise de componentes principais (PCA)

Com o objetivo de reduzir a dimensionalidade dos dados e otimizar o desempenho dos algoritmos de agrupamento, aplicou-se a técnica de Análise de Componentes Principais (*Principal Component Analysis* – PCA). Esse procedimento transforma as variáveis originais

em um novo conjunto de componentes não correlacionados, que retêm a maior parte da variância dos dados.

A Figura 2 mostra que os 8 primeiros componentes principais são suficientes para explicar aproximadamente 90% da variância total dos dados. Após esse ponto, a contribuição dos componentes adicionais torna-se marginal, o que justifica a escolha de 10 componentes como critério de corte para a redução dimensional.

Figura 2 – Variação explicada acumulada pelos componentes principais.



Fonte: Produzido pelos autores.

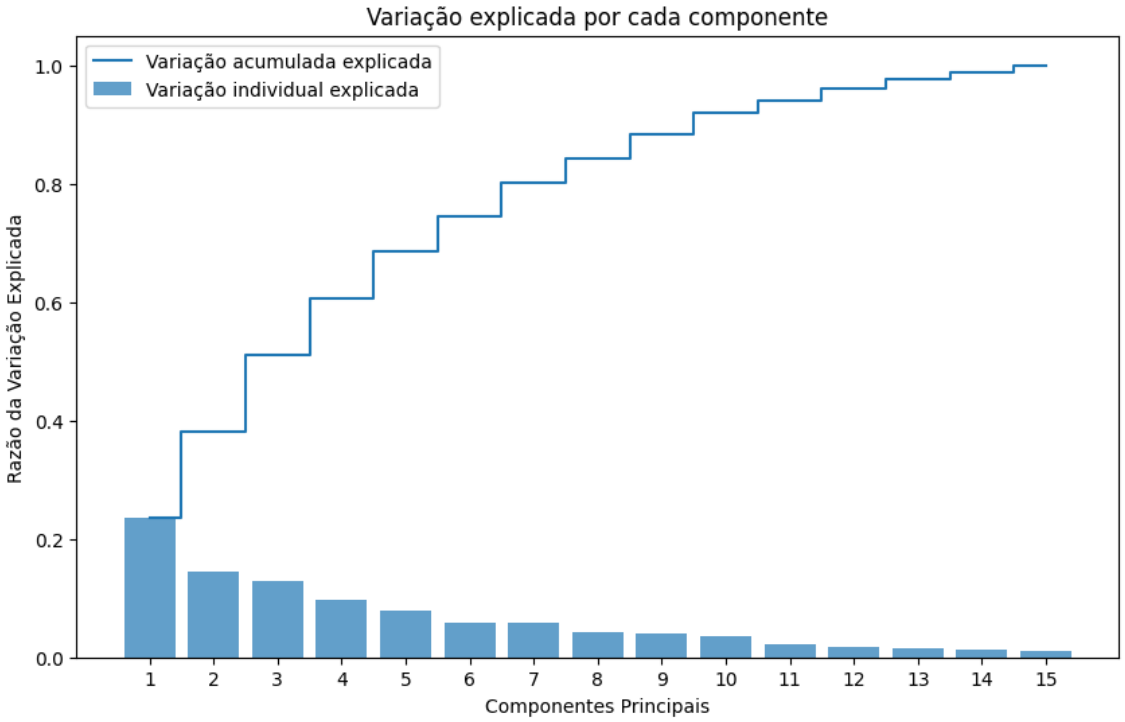
A Figura 3 reforça essa decisão ao indicar que o primeiro componente explica cerca de 20% da variância total, com os componentes seguintes apresentando declínio progressivo em suas contribuições. A partir do décimo componente, a variação individual explicada se estabiliza, indicando baixa relevância dos demais para a estrutura dos dados.

Com base nessa análise, a retenção de 8 componentes principais mostrou-se eficaz para representar a estrutura dos dados de forma compacta, sem perdas significativas de informação.

4.3.3 Determinação do número ideal de *clusters*: Método do Cotovelo

Para definir o número ideal de *clusters*, utilizou-se o método do cotovelo, que avalia a soma dos erros quadráticos intra-grupos (WCSS) em diferentes particionamentos. A Figura 4 mostra que o WCSS diminui rapidamente conforme o número de *clusters* aumenta, indicando que os dados estão sendo agrupados de forma mais eficiente. No entanto, a partir

Figura 3 – Variação explicada por componente principal.



Fonte: Produzido pelos autores.

de determinado ponto, essa redução se torna menos acentuada, formando um "cotovelo"na curva.

Neste estudo, o cotovelo foi identificado entre 4 e 6 *clusters*, sugerindo que essa faixa representa o ponto de equilíbrio entre a complexidade do modelo e a variância explicada. A escolha dentro desse intervalo visa otimizar a segmentação dos dados sem sobrecarregar o modelo com agrupamentos redundantes.

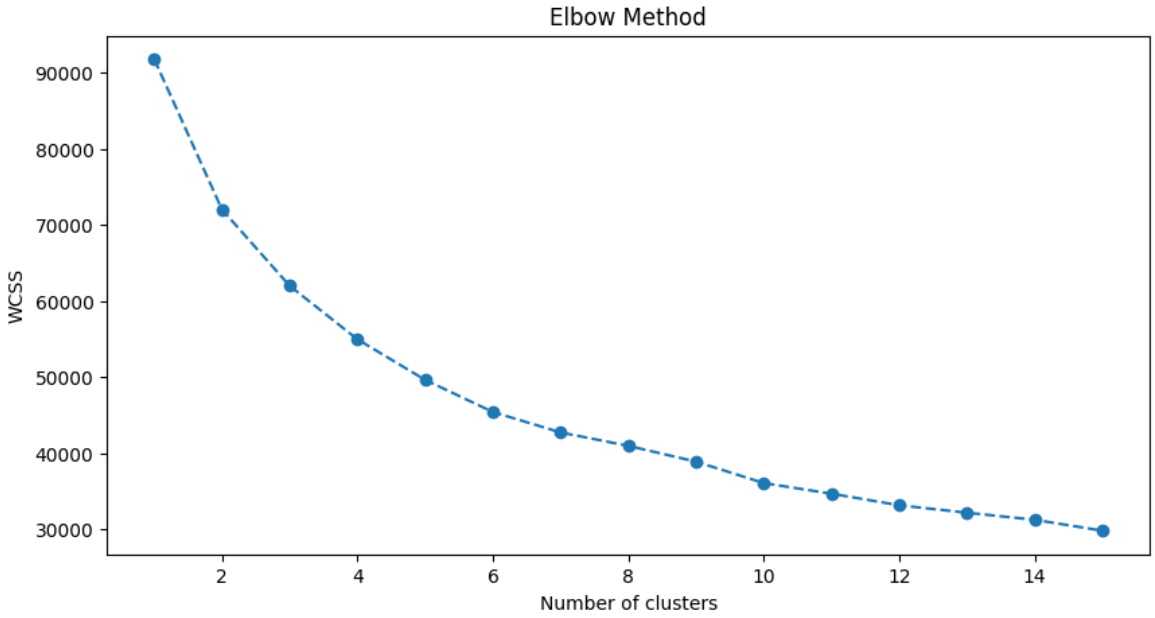
4.3.4 Aplicação do algoritmo K-Means

Nesta etapa, foi aplicado o algoritmo de agrupamento *K-Means* sobre os dados transformados por meio da Análise de Componentes Principais (PCA), com o objetivo de segmentar os indivíduos em grupos com perfis semelhantes de comorbidades. A escolha do número de *clusters* foi orientada pela análise do método do cotovelo e pelo *Silhouette Score*, sendo definido o uso de 6 *clusters* como configuração final, por apresentar o melhor equilíbrio entre coesão intra-grupo e separação entre grupos.

Os parâmetros utilizados no *K-means* foram:

- **Número de clusters (*n_clusters*)**: 6, com base na análise mencionada.

Figura 4 – Método do cotovelo para definição do número de clusters.



Fonte: Produzido pelos autores.

- **Inicialização (*init*)**: *k-means++*, garantindo uma inicialização eficiente e minimizando o risco de convergência para mínimos locais.
- **Número de inicializações (*n_init*)**: 50, assegurando maior robustez e estabilidade nos resultados.
- **Número máximo de iterações (*max_iter*)**: 1000, permitindo ao algoritmo convergir adequadamente mesmo em cenários complexos.
- **Semente aleatória (*random_state*)**: 42, para garantir a reprodutibilidade dos resultados.

Essa configuração permitiu uma segmentação eficiente dos dados, possibilitando a identificação de agrupamentos relevantes de indivíduos com padrões comuns de condições crônicas, contribuindo para uma compreensão mais aprofundada do perfil de uso dos serviços de saúde.

4.3.5 Distribuição dos *clusters*

Após a aplicação do algoritmo *K-Means*, os dados foram agrupados em seis *clusters*, conforme definido anteriormente. A Tabela 5 apresenta a distribuição das observações entre os grupos formados. O *cluster* 1 concentrou a maior parte dos indivíduos, totalizando

40.569 registros, seguido pelos *clusters* 2, 5, 3, 4 e 0, que apresentaram proporções menores e mais equilibradas.

Tabela 5 – Distribuição de observações por *cluster*.

Cluster	Número de Observações
0	7.341
1	40.569
2	11.585
3	8.523
4	8.415
5	11.245

Fonte: Elaboração própria.

A predominância do *cluster* 1 sugere a existência de um grupo majoritário com características comuns entre os indivíduos, enquanto os demais *clusters* representam subgrupos com perfis específicos, o que pode ser explorado em análises futuras para identificação de padrões relevantes associados ao uso dos serviços de saúde.

4.3.6 Validação dos grupos: Índice de Silhueta

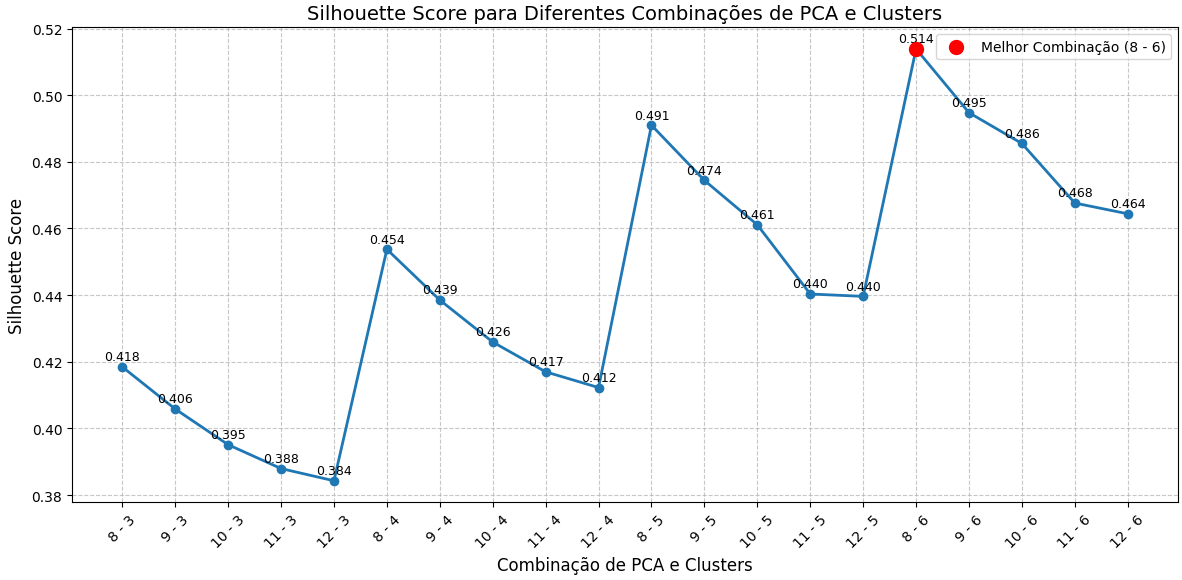
A análise de silhueta foi utilizada para avaliar a coesão e separação dos grupos gerados pelo algoritmo de agrupamento. Um coeficiente de silhueta mais próximo de 1 indica que os elementos estão bem ajustados ao seu *cluster*, enquanto valores próximos de 0 ou negativos apontam para classificações ambíguas ou incorretas.

Foram realizados 20 testes com diferentes combinações de componentes principais (PCA) e número de *clusters*, com o objetivo de identificar a configuração que maximizasse o *Silhouette Score*, sem comprometer a variância explicada. A Figura 5 resume os resultados dessas combinações. Observa-se que a melhor configuração foi alcançada com **8 componentes principais** e **6 *clusters***, apresentando o maior valor de *Silhouette Score* (51,39%).

A Figura 6 apresenta o gráfico de silhueta para a configuração selecionada. Os valores, em sua maioria, encontram-se acima de 0, indicando uma separação satisfatória entre os grupos formados. A visualização dos dados no espaço bidimensional das duas primeiras componentes principais revela uma boa distinção entre os *clusters*.

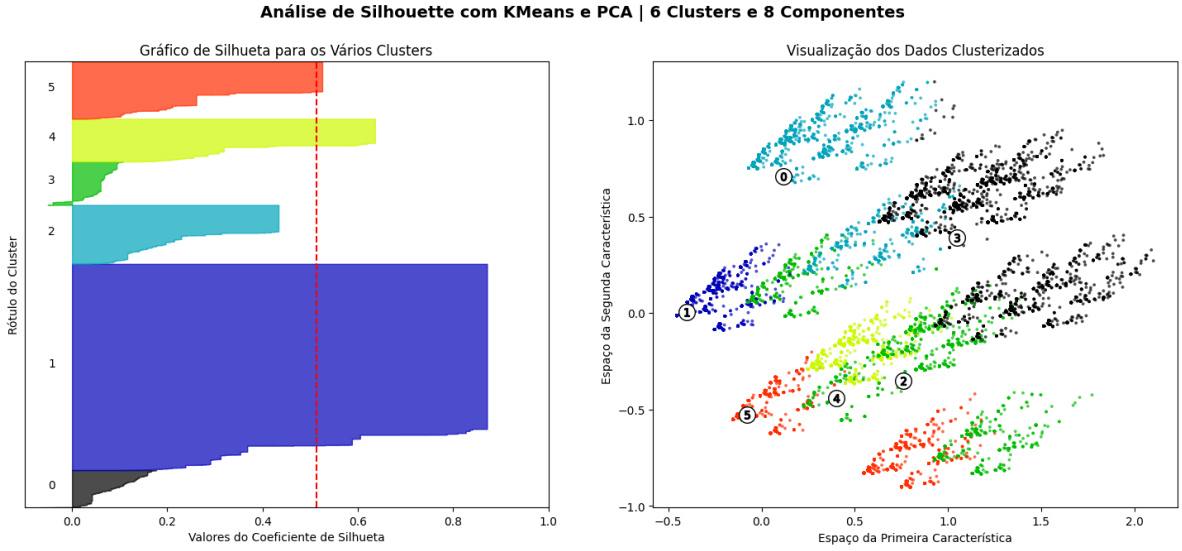
Apesar dos resultados positivos, foi identificado um desbalanceamento entre os grupos, com o *cluster* 1 concentrando a maior parte das observações. Os *clusters* 0, 2 e 3, por outro lado, apresentaram menor representatividade. Esse comportamento reflete a

Figura 5 – Testes de combinações de PCA e número de *clusters*.



Fonte: Produzido pelos autores.

Figura 6 – Análise de silhueta e visualização dos dados com PCA = 8 e 6 *clusters*.



Fonte: Produzido pelos autores.

heterogeneidade dos perfis de comorbidades presentes na base de dados e será considerado nas análises posteriores envolvendo o uso de serviços de saúde.

4.3.7 Erro do agrupamento

No contexto do algoritmo *K-means*, a *Within-Cluster Sum of Squares* (WCSS) representa a soma das distâncias quadráticas entre os pontos e os centroides dos *clusters* aos quais pertencem. Essa métrica avalia o grau de compactação dos grupos formados, sendo que valores menores indicam uma maior coesão interna.

Além disso, foi calculado o erro médio por ponto, obtido pela divisão do WCSS pelo número total de observações. Esse valor fornece uma estimativa da dispersão média dos pontos em relação ao centro de seus respectivos grupos.

Para a configuração final adotada, com **8 componentes principais** e **6 *clusters***, obteve-se um valor de WCSS de aproximadamente **37.821,75** e um **erro médio por ponto** de **0,4314**. Esses resultados indicam uma boa segmentação dos dados, com grupos bem definidos e pouca dispersão, o que reforça a adequação da configuração escolhida para representar os perfis de comorbidades presentes na base analisada.

4.4 Análise exploratória dos *clusters*

Com os *clusters* definidos por meio do algoritmo *K-means*, foi realizada uma análise exploratória para compreender os perfis formados com base em características relacionadas ao uso de serviços de saúde. O foco esteve em identificar padrões relevantes a partir das variáveis presentes na base da Pesquisa Nacional de Saúde de 2019.

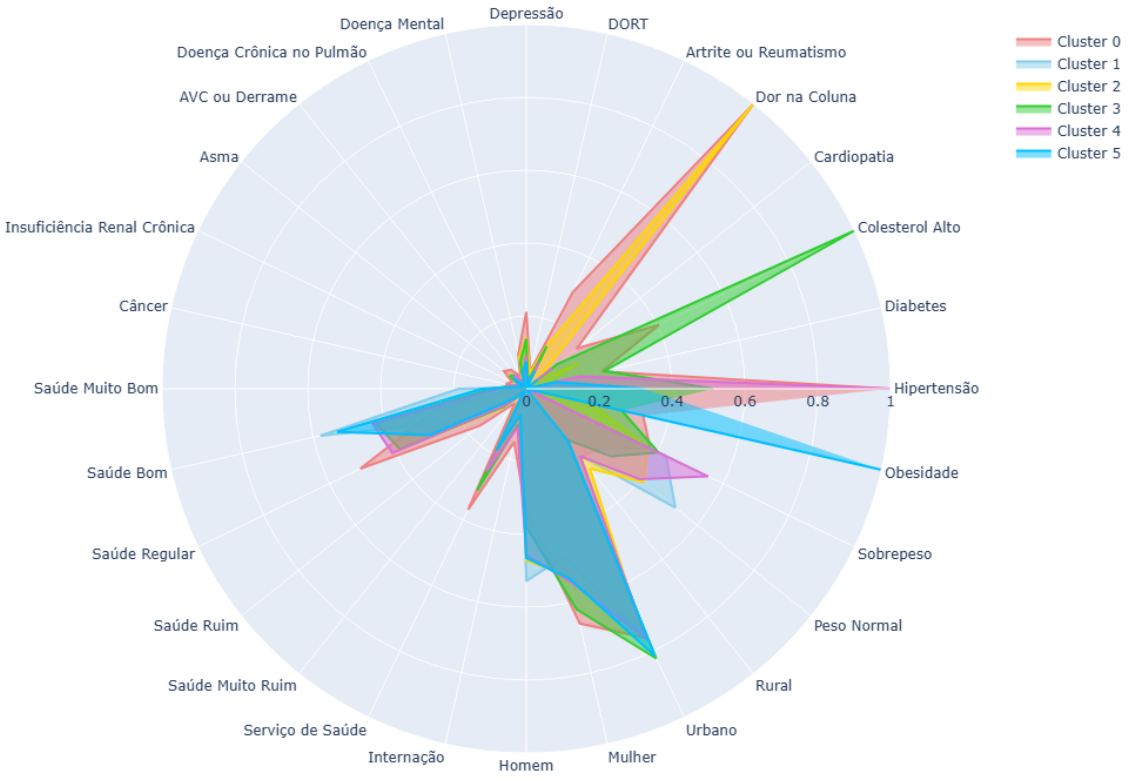
A análise considerou atributos representativos de condições clínicas, percepção de saúde, aspectos demográficos e variáveis diretamente associadas à utilização de serviços de saúde. Foram incluídas comorbidades como hipertensão, diabetes, colesterol alto, cardiopatia, dor na coluna, artrite ou reumatismo, DORT, depressão, doença mental, doença crônica no pulmão, AVC ou derrame, asma, insuficiência renal crônica e câncer. Também foram analisadas variáveis relacionadas ao estado de saúde (muito bom, bom, regular, ruim e muito ruim), ao uso de serviços (incluindo internações), além de informações demográficas (sexo e zona de residência) e condição física (peso normal, sobrepeso e obesidade).

As Figuras 7, 8, 9, 10, 11, 12, 13 apresentam os diagramas de radar com a média normalizada das variáveis para cada *cluster*. Esses gráficos permitem visualizar os perfis predominantes em cada grupo, facilitando a interpretação das características que mais influenciam a segmentação dos indivíduos.

4.4.1 Análise do *cluster* 0

O *cluster* 0 apresenta um perfil marcadamente caracterizado por altas prevalências de comorbidades. Destacam-se, principalmente, a hipertensão e a dor na coluna, ambas

Figura 7 – Diagrama de radar dos *clusters*.



Fonte: Produzido pelos autores.

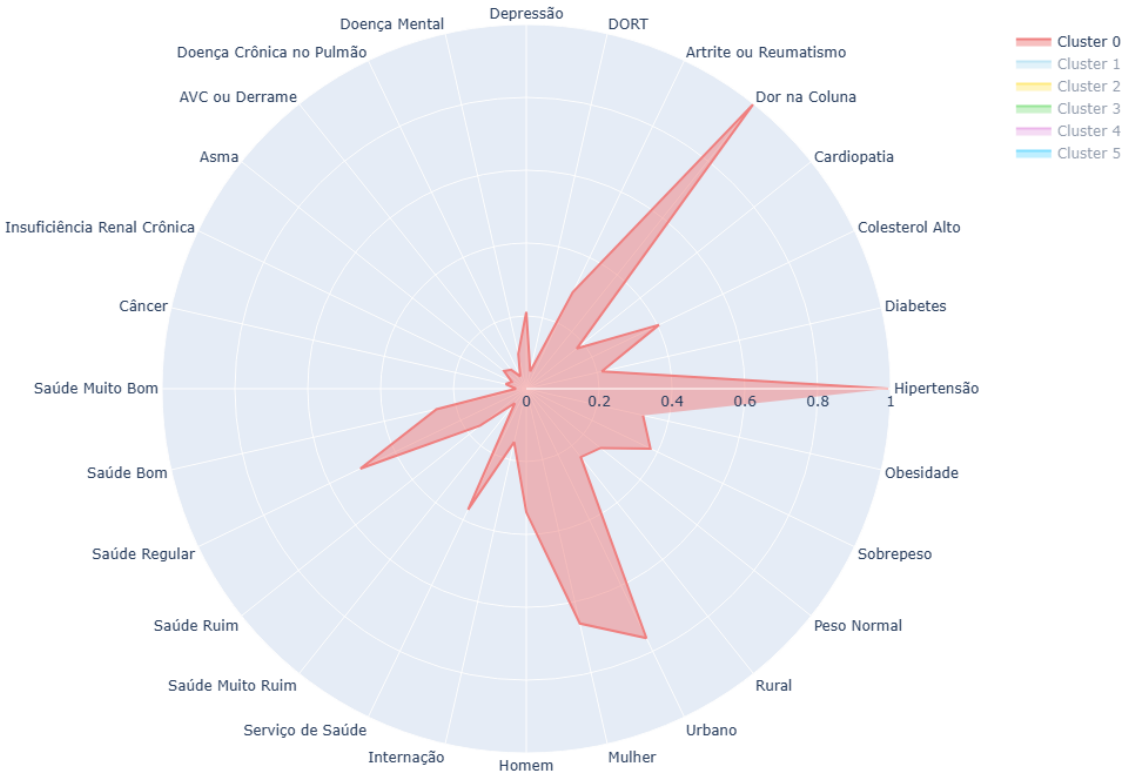
com valores próximos ao máximo da escala normalizada, indicando forte presença desses agravos entre os indivíduos desse grupo. Além disso, há incidência relevante de condições como artrite ou reumatismo, colesterol alto e obesidade.

Esse *cluster* também exibe níveis mais elevados de percepção negativa do estado de saúde. As categorias “saúde regular”, “ruim” e “muito ruim” são mais expressivas, sugerindo uma associação direta entre múltiplas comorbidades e uma autoavaliação de saúde debilitada.

Quanto ao uso de serviços de saúde, observa-se um padrão de maior utilização. Os indicadores de atendimento em serviços médicos e internações são mais elevados do que nos demais grupos, refletindo a maior demanda por cuidado e acompanhamento médico decorrente das condições de saúde enfrentadas.

Em relação ao perfil demográfico, nota-se um equilíbrio entre gêneros, com leve predominância feminina. Também há concentração urbana significativa e proporções

Figura 8 – Diagrama de radar do *cluster* 0.



Fonte: Produzido pelos autores.

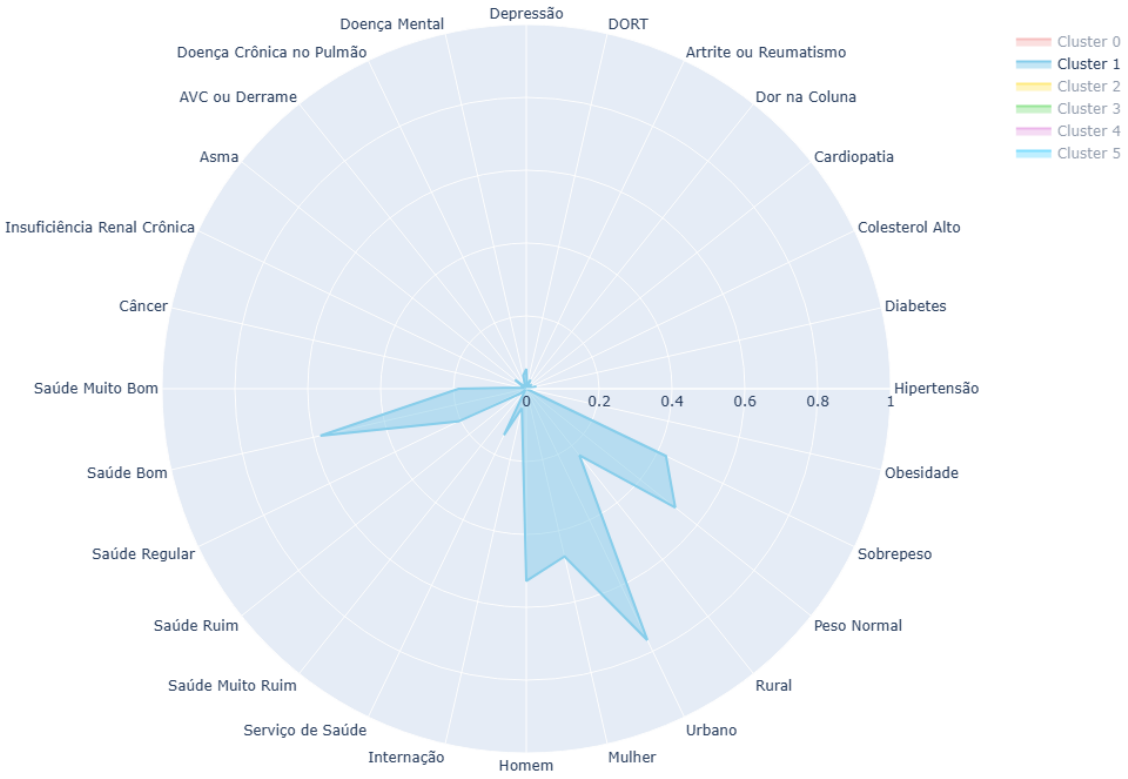
expressivas de sobrepeso e obesidade, o que pode estar relacionado ao surgimento ou agravamento das condições crônicas observadas.

4.4.2 Análise do *cluster* 1

O *cluster* 1 apresenta um perfil predominantemente saudável, com baixíssima incidência de comorbidades. As condições crônicas, como hipertensão, diabetes, colesterol alto, doenças respiratórias e mentais, apresentam valores muito próximos de zero, indicando que os indivíduos desse grupo possuem boa saúde geral.

Do ponto de vista da percepção de saúde, destaca-se uma elevada proporção de indivíduos que se classificam como possuidores de "saúde muito boa" ou "boa", especialmente a primeira, com valor expressivo entre os clusters. Esse padrão positivo se reflete também na baixa utilização dos serviços de saúde, com valores reduzidos tanto para atendimento em serviços médicos quanto para internações.

Figura 9 – Diagrama de radar do *cluster* 1.



Fonte: Produzido pelos autores.

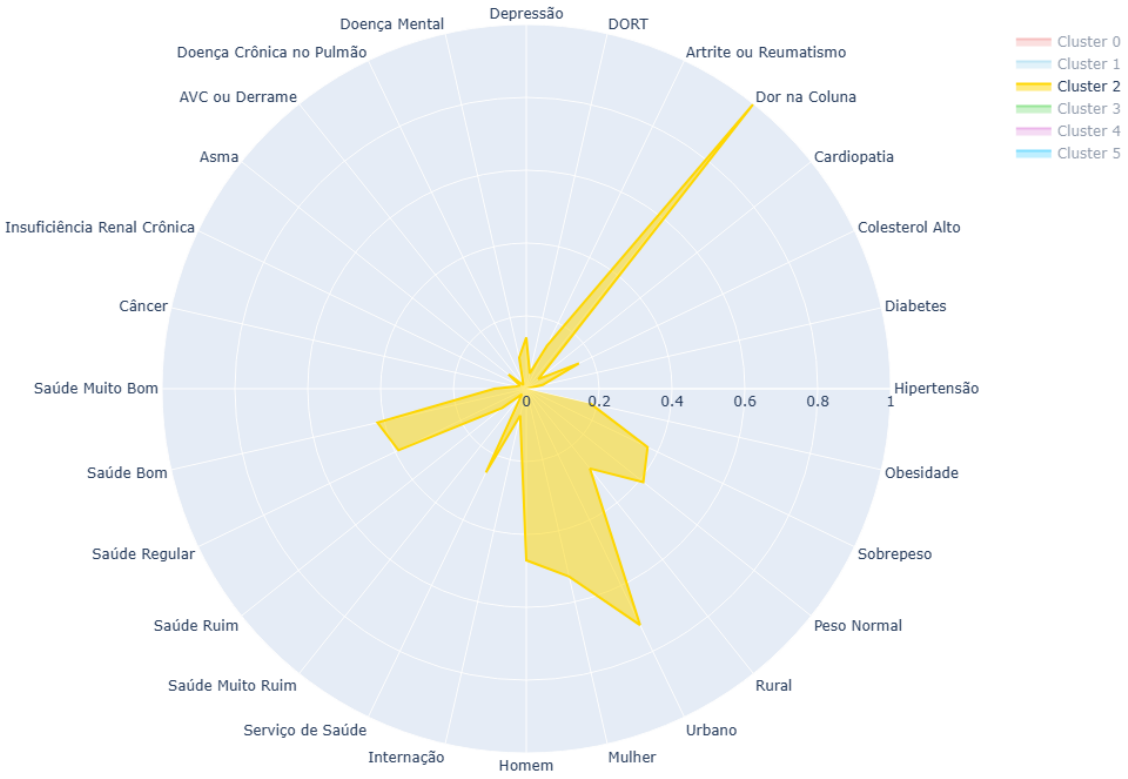
Em termos demográficos, o *cluster* 1 é composto majoritariamente por homens e apresenta predomínio de residentes em áreas urbanas. Além disso, observa-se um índice elevado de indivíduos com peso normal, e uma proporção moderada de pessoas com sobrepeso, com pouca incidência de obesidade.

4.4.3 Análise do *cluster* 2

O *cluster* 2 apresenta um perfil de saúde marcado pela alta prevalência de dor na coluna, presente em 100% dos indivíduos. Além disso, observa-se incidência relevante de obesidade e sobrepeso, o que pode estar associado às queixas musculoesqueléticas observadas. Hipertensão, diabetes e outras condições crônicas aparecem com baixa frequência, indicando que a principal condição limitante deste grupo está relacionada a dores físicas específicas, e não a doenças crônicas múltiplas.

A percepção de saúde nesse grupo é predominantemente positiva: grande parte dos indivíduos avalia sua saúde como “boa” ou “muito boa”. Ainda assim, uma fração da

Figura 10 – Diagrama de radar do *cluster* 2.



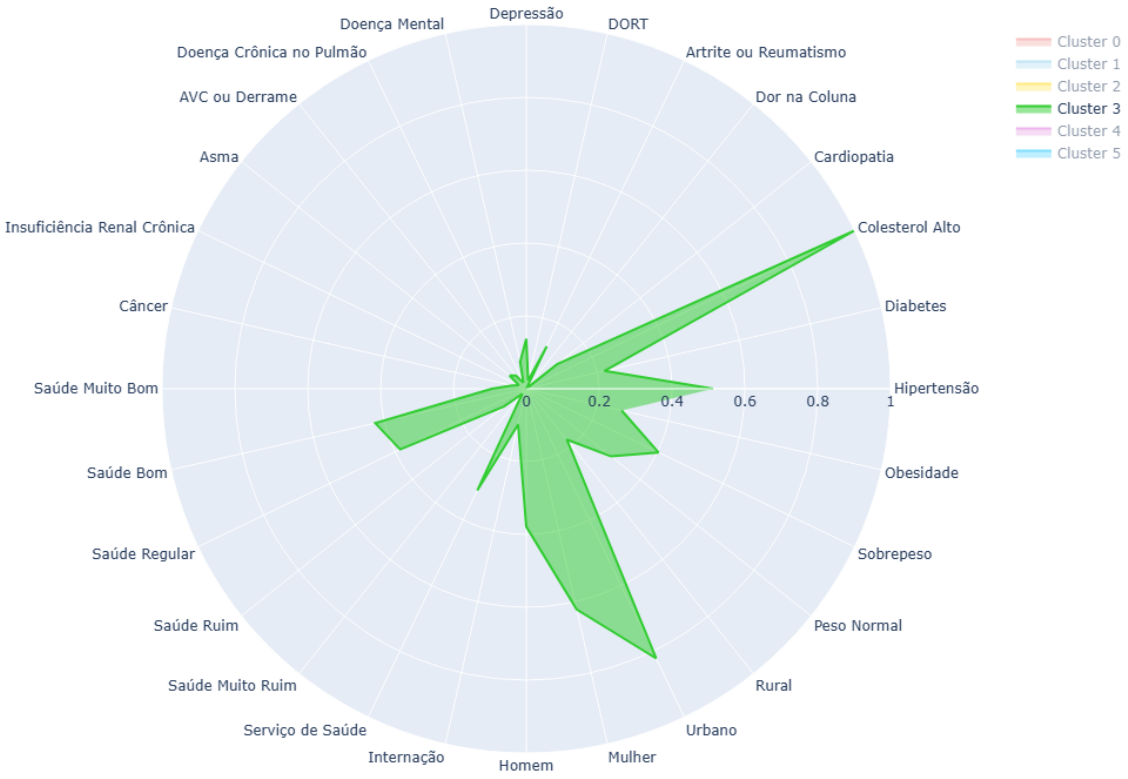
Fonte: Produzido pelos autores.

população apresenta avaliação regular, o que pode refletir os impactos funcionais das dores relatadas, mesmo na ausência de múltiplas comorbidades.

Em relação ao uso de serviços de saúde, os valores são discretos: há uma baixa taxa de internações e uma proporção moderada de indivíduos que utilizaram serviços ambulatoriais. Isso sugere que, embora o desconforto físico esteja presente, ele nem sempre é suficiente para motivar internações frequentes, sendo provavelmente tratado por meio de atendimentos pontuais ou automedicação.

Do ponto de vista demográfico, o grupo é composto majoritariamente por mulheres e residentes em áreas urbanas. Também se destaca a proporção considerável de indivíduos em idade produtiva, o que acentua o impacto potencial das dores crônicas na capacidade laboral e na qualidade de vida.

Figura 11 – Diagrama de radar do *cluster* 3.



Fonte: Produzido pelos autores.

4.4.4 Análise do *cluster* 3

O *cluster* 3 é caracterizado por uma prevalência elevada de colesterol alto, presente em 100% dos indivíduos, o que o diferencia fortemente dos demais grupos. Também se destacam as proporções consideráveis de hipertensão, obesidade e sobrepeso, indicando um perfil com múltiplos fatores de risco cardiovasculares. Outras condições, como diabetes e depressão, aparecem em menor grau, mas contribuem para a complexidade do estado de saúde do grupo.

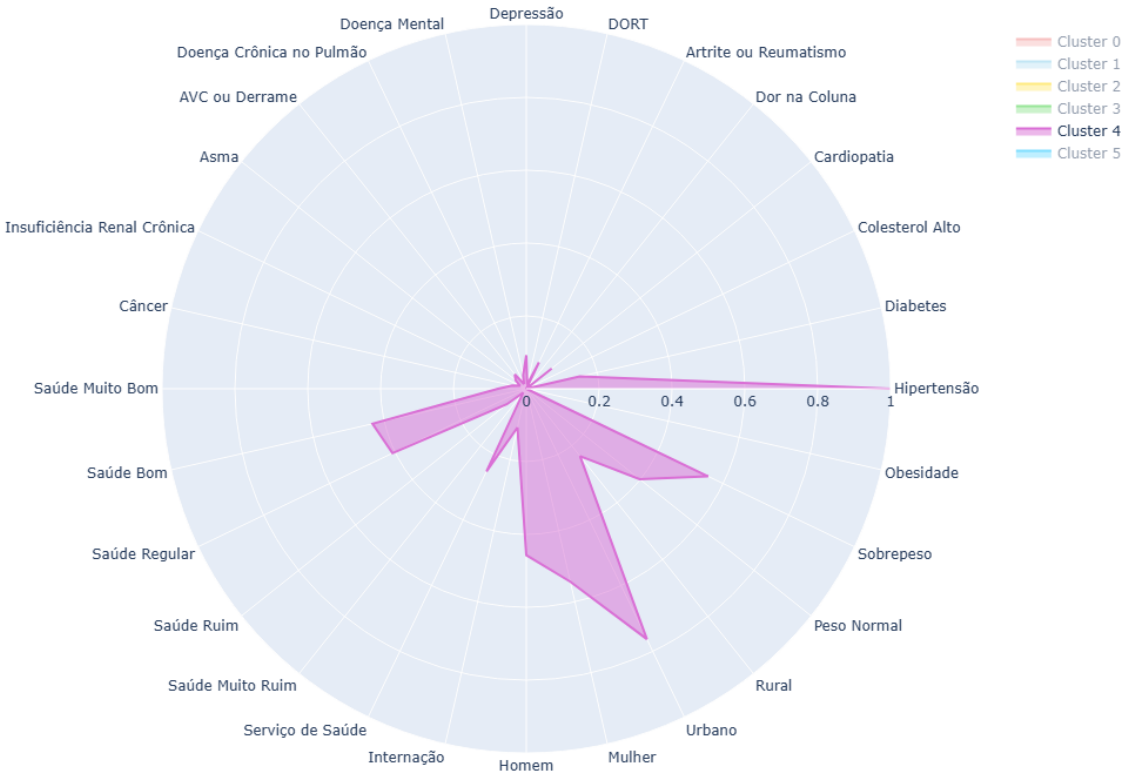
Esse grupo é composto majoritariamente por mulheres e indivíduos residentes em áreas urbanas, além de uma representação relevante da população em idade produtiva. Em relação à autopercepção de saúde, observa-se uma proporção expressiva de respostas positivas, com destaque para os que se consideram com “boa” ou “muito boa” saúde, o que pode contrastar com o quadro clínico mais crítico indicado pelas comorbidades.

A taxa de utilização de serviços de saúde é moderada, com uma parcela dos indiví-

duos relatando atendimento ambulatorial recente e uma baixa incidência de internações. Isso pode sugerir subutilização de serviços diante da real necessidade, possivelmente relacionada à percepção otimista da saúde ou a barreiras de acesso.

4.4.5 Análise do *cluster* 4

Figura 12 – Diagrama de radar do *cluster* 4.



Fonte: Produzido pelos autores.

O *cluster* 4 se destaca principalmente pela presença universal de hipertensão entre seus integrantes, além de uma proporção elevada de indivíduos com obesidade e sobrepeso. Esse perfil sugere um grupo com risco cardiovascular acentuado, ainda que outras comorbidades apareçam com menor intensidade. Condições como depressão, DORT e colesterol alto também estão presentes, mas em níveis inferiores aos observados em outros *clusters*.

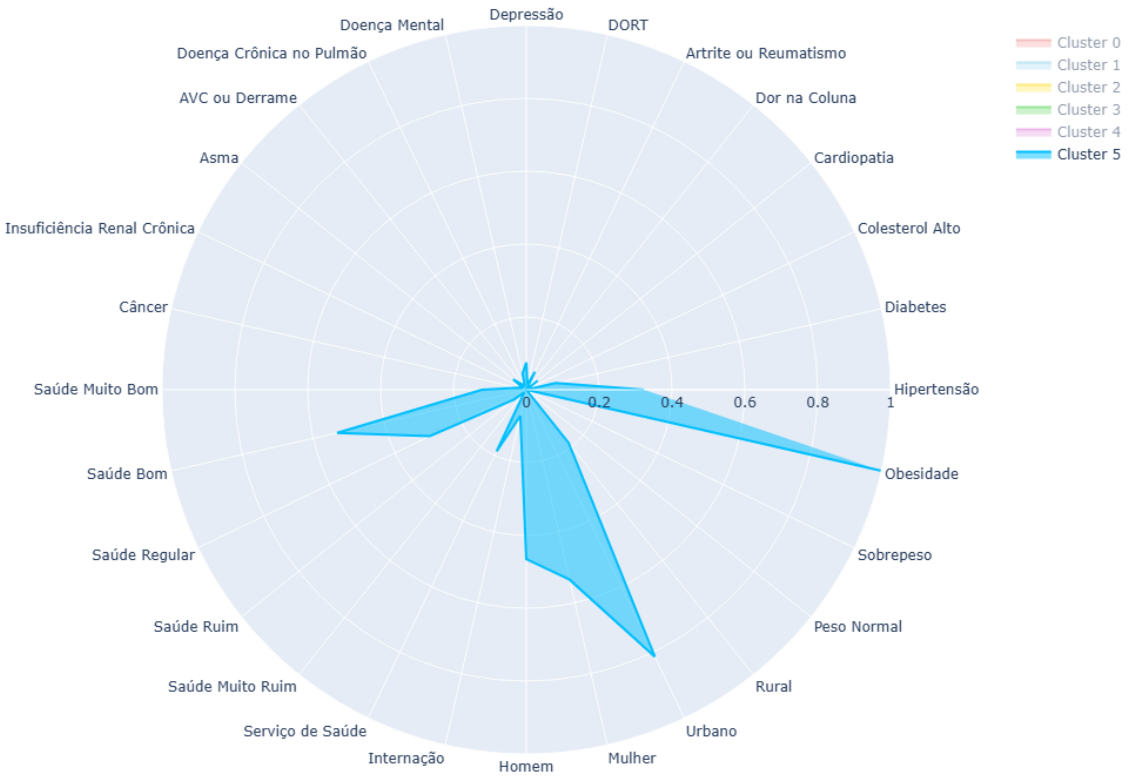
Em termos demográficos, há predominância de mulheres e indivíduos residentes em áreas urbanas. O grupo também apresenta uma boa percepção de saúde, com destaque para as respostas “boa” e “muito boa”. A autopercepção positiva, entretanto, pode contrastar

com o quadro clínico mais preocupante observado em razão das condições associadas ao excesso de peso e à hipertensão.

No que se refere ao uso de serviços de saúde, o grupo apresenta taxa moderada de procura por atendimento e baixa incidência de internações. Isso pode indicar uma menor utilização dos serviços diante de condições que exigiriam acompanhamento regular, ou ainda refletir uma possível subnotificação, ou barreiras de acesso ao cuidado contínuo.

4.4.6 Análise do *cluster* 5

Figura 13 – Diagrama de radar do *cluster* 5.



Fonte: Produzido pelos autores.

O *cluster* 5 caracteriza-se por uma prevalência significativa de obesidade, presente em 100% dos indivíduos, sendo este o principal fator de risco do grupo. Essa condição aparece acompanhada de proporções moderadas de hipertensão, sobrepeso e baixa prática de atividade física, indicando um perfil que exige atenção especial quanto à prevenção de doenças crônicas.

A maioria dos indivíduos apresenta autopercepção de saúde positiva, com destaque para os que classificam sua saúde como “boa” ou “muito boa”. Esse fator pode representar uma percepção otimista, ainda que coexistam condições fisiológicas preocupantes como o excesso de peso.

No aspecto demográfico, observa-se maior representatividade feminina, predomínio de residentes em áreas urbanas e um número considerável de indivíduos com perfil jovem ou adulto. Quanto ao uso dos serviços de saúde, as taxas são relativamente baixas tanto para atendimentos quanto para internações, o que pode estar associado à ausência de manifestações clínicas severas até o momento.

4.4.7 Comparação dos *clusters*

O mapa de calor apresentado na Figura 14 consolida os padrões identificados nos *clusters* gerados, facilitando a comparação entre os atributos e permitindo visualizar as principais características de cada grupo. A intensidade das cores reflete a média normalizada das variáveis dentro dos *clusters*, evidenciando diferenças marcantes entre os perfis formados.

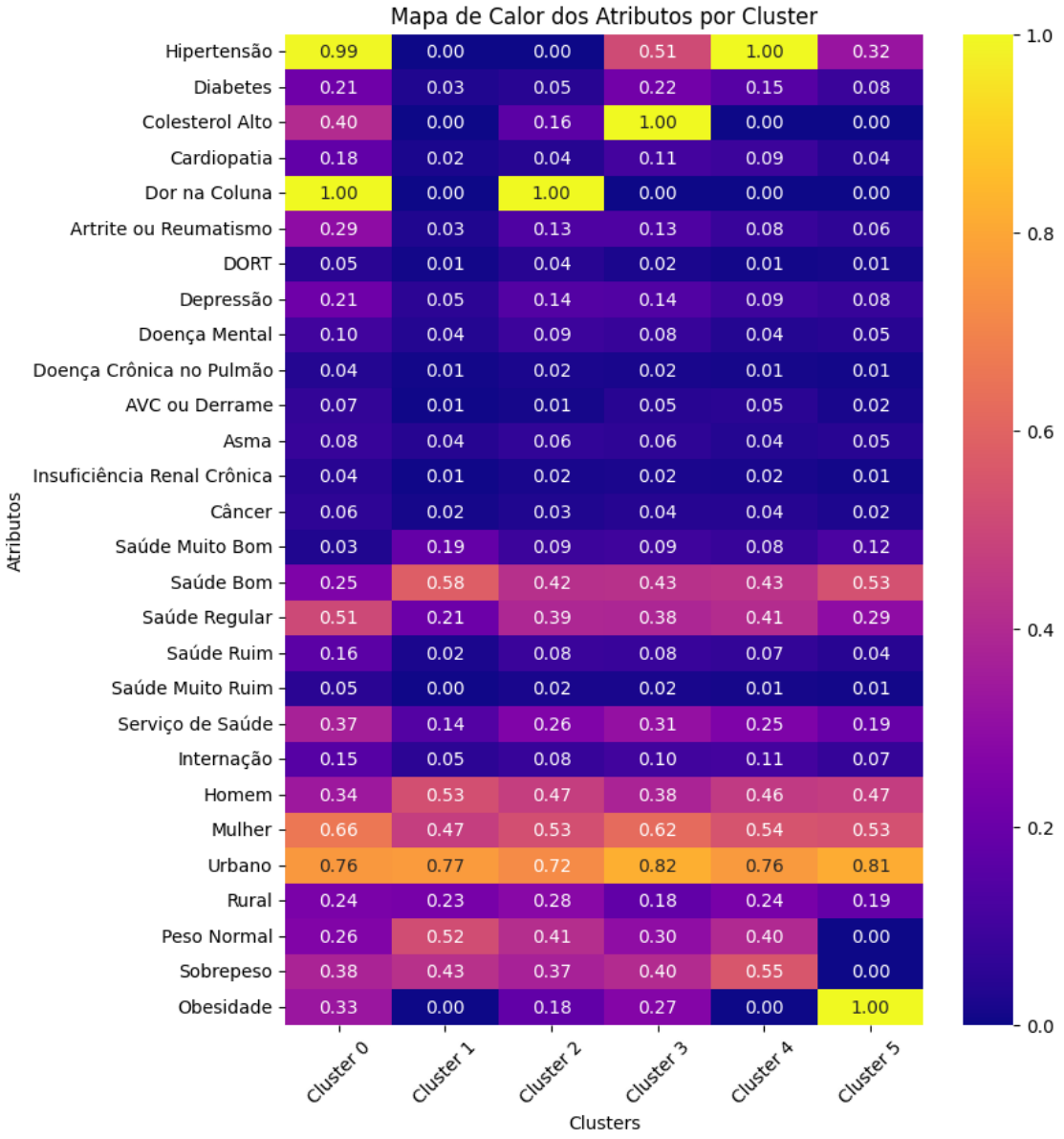
De forma geral, observam-se três grandes grupos com padrões distintos. O primeiro é composto por *clusters* com baixas prevalências de comorbidades e bons indicadores de saúde percebida, como é o caso dos *clusters* 0 e 5. O *cluster* 0 se destaca pela alta presença de indivíduos com dor na coluna e hipertensão, enquanto o *cluster* 5 apresenta a obesidade como principal marcador, com valores próximos a zero para comorbidades crônicas mais graves. Ambos demonstram baixa utilização dos serviços de saúde e percepções mais positivas sobre o estado geral de saúde.

O segundo grupo agrupa perfis mais vulneráveis clinicamente, como os *clusters* 1 e 4. Esses grupos concentram altas taxas de hipertensão (100%) e níveis elevados de diabetes, colesterol alto e outras doenças crônicas. Esses perfis estão associados a maior uso dos serviços de saúde, incluindo internações, além de uma percepção mais negativa do estado de saúde. Apesar disso, o *cluster* 4 possui índices um pouco mais favoráveis que o *cluster* 1, sugerindo um risco moderado.

O terceiro perfil é formado pelos *clusters* 2 e 3, que exibem características intermediárias. O *cluster* 2 se destaca pela dor crônica na coluna (100%), sobrepeso e maior incidência de problemas como depressão e artrite. Já o *cluster* 3 apresenta altos índices de colesterol alto, obesidade e sobrepeso, com menor presença de doenças crônicas mais graves. Ambos possuem níveis moderados de uso de serviços de saúde e percepção regular do estado de saúde.

Adicionalmente, os dados revelam predominância do ambiente urbano na maioria dos *clusters*, com leve variação entre os grupos. A distribuição por gênero é relativamente

Figura 14 – Mapa de calor dos *clusters*.



Fonte: Produzido pelos autores.

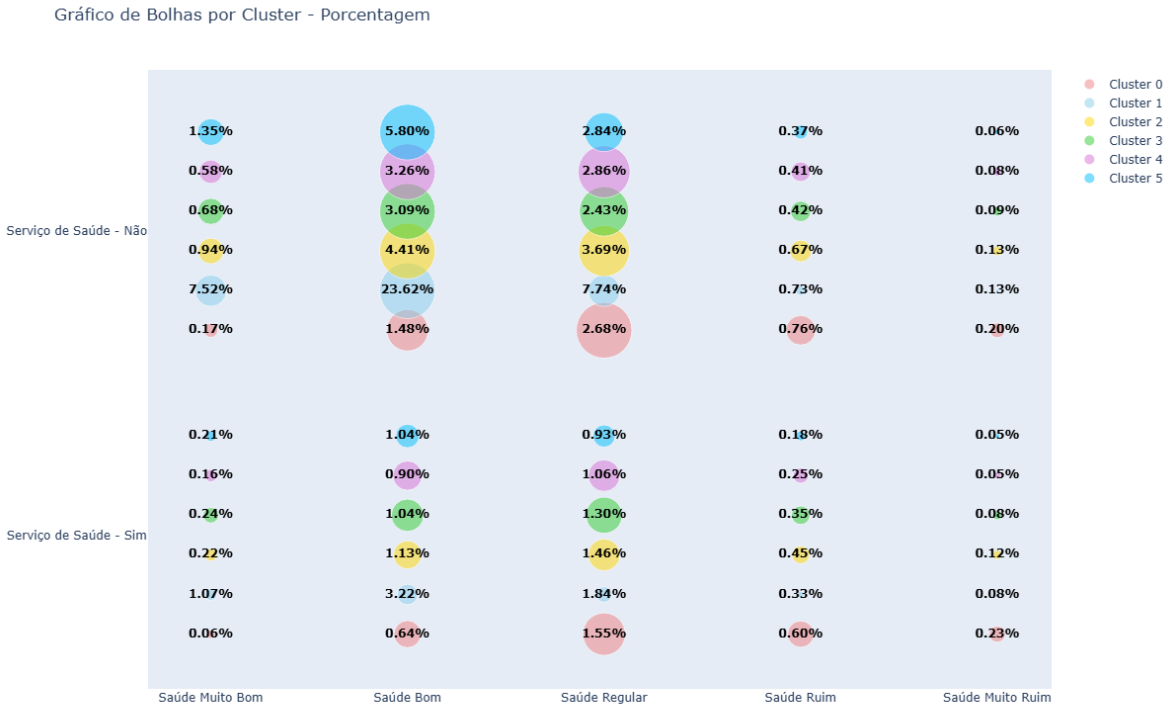
equilibrada, mas os *clusters* 3 e 5 possuem maior proporção de mulheres.

Essas diferenças reforçam a heterogeneidade dos perfis encontrados, com implicações diretas para o planejamento de políticas públicas e intervenções em saúde. A segmentação permite identificar grupos que demandam estratégias preventivas específicas, como controle de peso e incentivo à atividade física, e outros que necessitam de acompanhamento contínuo e gestão de múltiplas comorbidades.

4.4.8 Análise do uso do serviço de saúde em relação à percepção de saúde e aos *clusters*

O gráfico de bolhas apresentado na Figura 15 ilustra a relação entre a percepção do estado de saúde e o uso de serviços de saúde pelos indivíduos de cada *cluster*, permitindo uma visualização integrada do comportamento de diferentes grupos populacionais frente ao sistema de saúde.

Figura 15 – Gráfico de bolhas do uso do serviço de saúde em relação à percepção de saúde e aos *clusters*



Fonte: Produzido pelos autores.

De maneira geral, nota-se que a maioria dos indivíduos que relatam uma saúde "muito boa" ou "boa" não utilizou serviços de saúde no último ano, especialmente nos *clusters* 0 e 5, associados a perfis mais saudáveis ou com comorbidades ainda em estágio inicial. O *cluster* 5, por exemplo, reúne uma das maiores proporções de pessoas que se consideram saudáveis e que não buscaram atendimento, mesmo apresentando alta prevalência de obesidade, o que pode indicar um subuso de serviços ou uma percepção de risco ainda baixa.

O *cluster* 0, apesar de apresentar alguns fatores de risco como dor na coluna e hipertensão, também exibe um padrão de baixa utilização dos serviços de saúde entre

aqueles que avaliam positivamente sua saúde, sugerindo um perfil com menor gravidade clínica e maior autonomia.

Em contrapartida, o *cluster* 1 concentra as maiores proporções de indivíduos que acessaram os serviços de saúde, independentemente da percepção subjetiva de saúde, com destaque para os que avaliam sua saúde como regular ou ruim. Esse padrão está alinhado com o perfil de maior fragilidade clínica e prevalência de comorbidades como hipertensão e diabetes. Nesse grupo, o uso dos serviços parece estar mais relacionado à necessidade do que à percepção positiva ou preventiva.

O *cluster* 2 apresenta uma distribuição equilibrada, com destaque para a percepção de saúde "regular" entre os que usaram e não usaram serviços. Essa configuração sugere um padrão de uso moderado, possivelmente motivado por sintomas persistentes, como dor crônica e condições psicológicas, ainda que não tão graves quanto no *cluster* 1.

O *cluster* 3 mostra proporções relevantes tanto entre usuários quanto não usuários dos serviços de saúde, concentrando-se entre os que se percebem com saúde "boa" ou "regular". Esse comportamento pode estar associado a uma abordagem mais preventiva ou de monitoramento, principalmente devido à presença de obesidade e colesterol alto, ainda sem grandes impactos na percepção de saúde.

Por fim, o *cluster* 4, embora com menor destaque no gráfico, acompanha o padrão de maior utilização entre aqueles com percepção de saúde regular ou ruim, coerente com seu perfil clínico de risco moderado, notadamente pela hipertensão e obesidade.

5 Desenvolvimento

Este capítulo descreve o processo de desenvolvimento do modelo preditivo voltado à identificação do uso dos serviços de saúde pela população brasileira. As etapas compreendem desde o tratamento e a preparação dos dados até a modelagem e a avaliação dos algoritmos utilizados.

O objetivo central do modelo é contribuir para uma gestão mais eficiente dos recursos de saúde, fornecendo subsídios para a formulação de estratégias e programas de prevenção voltados a grupos populacionais com características específicas. Dessa forma, pretende-se oferecer uma ferramenta prática que auxilie gestores e profissionais da saúde na tomada de decisões baseadas em evidências, promovendo o aprimoramento das políticas públicas de cuidado e acesso aos serviços de saúde.

5.1 Filtragem e seleção das variáveis

A primeira etapa do desenvolvimento consistiu na filtragem e seleção criteriosa das variáveis do conjunto de dados, com o intuito de otimizar a performance dos algoritmos de aprendizado de máquina e evitar a introdução de ruído nos modelos.

Inicialmente, foi realizada uma análise de colinearidade e redundância entre os atributos, visando identificar variáveis altamente correlacionadas ou pouco informativas. Em paralelo, avaliou-se a proporção de valores ausentes nas colunas, representados por `-1`, que indicam respostas ignoradas, em branco ou não aplicáveis.

A Tabela 6 apresenta as variáveis com maior percentual de ausência de dados. Como critério de corte, foram excluídas todas as colunas com mais de 70% de valores ausentes, pois sua baixa completude comprometeria a consistência das análises e o desempenho dos modelos preditivos.

A remoção dessas variáveis redundantes ou com baixa completude resultou em um conjunto de dados mais enxuto e representativo, favorecendo o desempenho dos algoritmos e a interpretabilidade dos resultados na etapa de classificação. As variáveis selecionadas foram: Zona (urbana ou rural), Grau de Ensino, Renda domiciliar per capita, Cadastro em Unidade Básica de Saúde (UBS), Presença de Agente Comunitário de Saúde, Sexo, Idade, Raça, Estado Civil, Capacidade de ler e escrever, Frequência escolar, Situação de trabalho, Possuir plano de saúde, Estado de saúde autodeclarado, Utilização de serviços de saúde, Internações recentes, Consumo de álcool, Prática de exercícios físicos, Uso de tabaco, Diagnóstico de colesterol alto, Acidente Vascular Cerebral (AVC) ou derrame, Asma, Artrite ou reumatismo, Dor na coluna, Distúrbio Osteomuscular Relacionado ao

Tabela 6 – Colunas com alto percentual de valores ausentes ou ignorados.

Coluna	**% dos valores ausentes**
Doença Crônica no Pulmão Limita Atividades	98%
Insuficiência Renal Limita Atividades	98%
DORT Limita Atividades	98%
AVC ou Derrame Limita Atividades	97%
Câncer Limita Atividades	97%
Asma Limita Atividades	95%
Cardiopatia Limita Atividades	94%
Doença Mental Limita Atividades	94%
Diabetes Limita Atividades	91%
Artrite ou Reumatismo Limita Atividades	91%
Depressão Limita Atividades	90%
Dor na Coluna Limita Atividades	78%
Hipertensão Limita Atividades	74%

Fonte: Elaboração própria.

Trabalho (DORT), Depressão, Hipertensão, Diabetes, Cardiopatia, Câncer, Insuficiência renal crônica, Doença mental, Doença crônica no pulmão, Peso amostral e Índice de Massa Corporal (IMC) gerado na etapa de agrupamento.

5.2 Divisão de dados e treinamento dos modelos

A divisão dos dados é uma etapa essencial no desenvolvimento de modelos preditivos, pois assegura a separação entre os dados usados para treinamento e aqueles utilizados para avaliação. Esse procedimento visa garantir a imparcialidade na medição do desempenho dos modelos, prevenindo o problema de *overfitting* — quando o modelo aprende demais sobre os dados de treino, mas generaliza mal em dados novos.

Inicialmente, os dados foram organizados em variáveis independentes (features) e variável dependente (target), representada pelo uso de serviços de saúde. Além disso, foram incorporados os pesos amostrais da Pesquisa Nacional de Saúde (PNS) 2019, garantindo que cada observação contribua de forma proporcional à sua representatividade na população brasileira.

A etapa de treinamento foi conduzida com validação cruzada estratificada, que divide os dados em múltiplos subconjuntos (*folds*) mantendo a proporção original das classes da variável alvo em cada divisão. Essa técnica oferece maior robustez na avaliação do modelo, permitindo obter estimativas mais confiáveis de desempenho. A Tabela 7

apresenta dois exemplos de divisões adotadas ao longo do processo.

Tabela 7 – Distribuição de dados entre treinamento, validação e teste.

n_splits	Validação	Teste	Treinamento
5	4/15 (26,67%)	1/5 (20%)	8/15 (53,33%)
10	3/10 (30%)	1/10 (10%)	3/5 (60%)

Fonte: Produzido pelos autores.

Nota: Os dados representam a divisão dos conjuntos para validação cruzada. Os valores entre parênteses indicam as porcentagens correspondentes.

Durante o processo de validação cruzada, cada *fold* é avaliado independentemente, possibilitando a geração de métricas detalhadas como *accuracy*, *precision*, *recall* e *F1-Score*, além da matriz de confusão, proporcionando uma análise ampla do desempenho de cada modelo.

A otimização dos hiperparâmetros foi realizada por meio do método *GridSearchCV*, que testa sistematicamente diferentes combinações de parâmetros com o objetivo de encontrar a configuração que maximize o desempenho do modelo, com foco na métrica *F1-Score*, dada sua adequação para bases de dados desbalanceadas.

Essa abordagem foi implementada de forma genérica, possibilitando sua aplicação a diferentes algoritmos de classificação, incluindo *Random Forest*, *Support Vector Machine* (SVM) — implementada por meio da classe `SVC` —, *Multi-Layer Perceptron* (MLP) e *Extreme Gradient Boosting* (XGB). Isso permitiu a comparação sistemática entre os modelos, utilizando os mesmos critérios de validação, contribuindo para uma análise justa e rigorosa de seus desempenhos preditivos.

5.3 Classificação com Random Forest

Nesta etapa, foi aplicado o modelo de classificação *Random Forest*, com o objetivo de avaliar seu desempenho na predição do uso de serviços de saúde. A técnica de validação cruzada com 5 *folds* foi empregada, assegurando uma avaliação robusta e equilibrada do modelo. Abaixo estão listados os principais hiperparâmetros configurados no *GridSearchCV*:

- **n_estimators**: número de árvores na floresta (`50, 100`).
- **max_depth**: profundidade máxima de cada árvore (`None, 5, 10`).
- **min_samples_split**: número mínimo de amostras necessário para dividir um nó interno (`2, 5`).

- **min_samples_leaf**: número mínimo de amostras exigido em um nó folha (`1, 2`).
- **max_features**: número de *features* consideradas em cada divisão (`sqrt, log2`).
- **criterion**: função utilizada para medir a qualidade das divisões (`gini, entropy`).
- **max_leaf_nodes**: número máximo de nós folhas permitidos (`None, 10, 20`).
- **min_impurity_decrease**: redução mínima da impureza para permitir a divisão de um nó (`0.0, 0.01`).
- **bootstrap**: indica se a amostragem foi realizada com reposição (`True`).
- **class_weight**: pesos das classes para lidar com dados desbalanceados (`None, balanced`).
- **max_samples**: porcentagem de amostras usada para treinar cada árvore (`None, 0.5, 0.75`).

5.3.1 Análise do tempo de execução na validação cruzada

A Tabela 8 apresenta os tempos de execução registrados em cada *fold* durante o processo de validação cruzada aplicado ao modelo *Random Forest*. Observa-se uma leve variação entre os *folds*, possivelmente atribuída a variações no desempenho computacional do ambiente de execução.

Mesmo com essas diferenças, o tempo médio por *fold* foi de aproximadamente 27,5 segundos, o que demonstra boa estabilidade e eficiência computacional do modelo, considerando a complexidade da base e o volume de dados processados. Esse desempenho consistente é um indicativo positivo para a escalabilidade da abordagem proposta.

Tabela 8 – Tempo de execução na validação cruzada Random Forest.

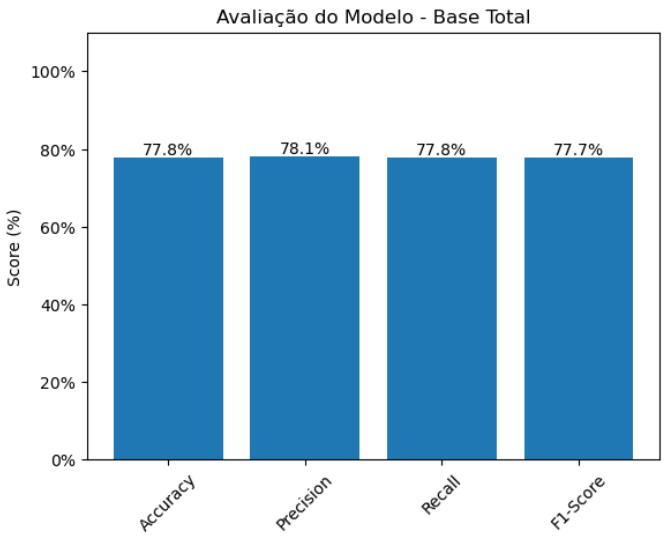
Fold	Tempo de Execução (s)	Tempo Acumulado (s)
1	30.79	30.79
2	27.06	57.85
3	26.79	84.64
4	26.51	111.15
5	26.36	137.51

Fonte: Produzido pelos autores.

5.3.2 Análise das métricas

Conforme apresentado na Figura 16, o modelo obteve uma acurácia de 77,8%, indicando que aproximadamente quatro em cada cinco previsões foram corretas. A métrica de precisão foi de 78,1%, o que demonstra que a maioria das classificações positivas feitas pelo modelo estava correta. A revocação alcançou 77,8%, evidenciando uma boa capacidade do modelo de identificar corretamente os indivíduos que utilizaram serviços de saúde. Por fim, o F1-Score, que representa o equilíbrio entre precisão e revocação, foi de 77,7%.

Figura 16 – Avaliação do modelo Random Forest.



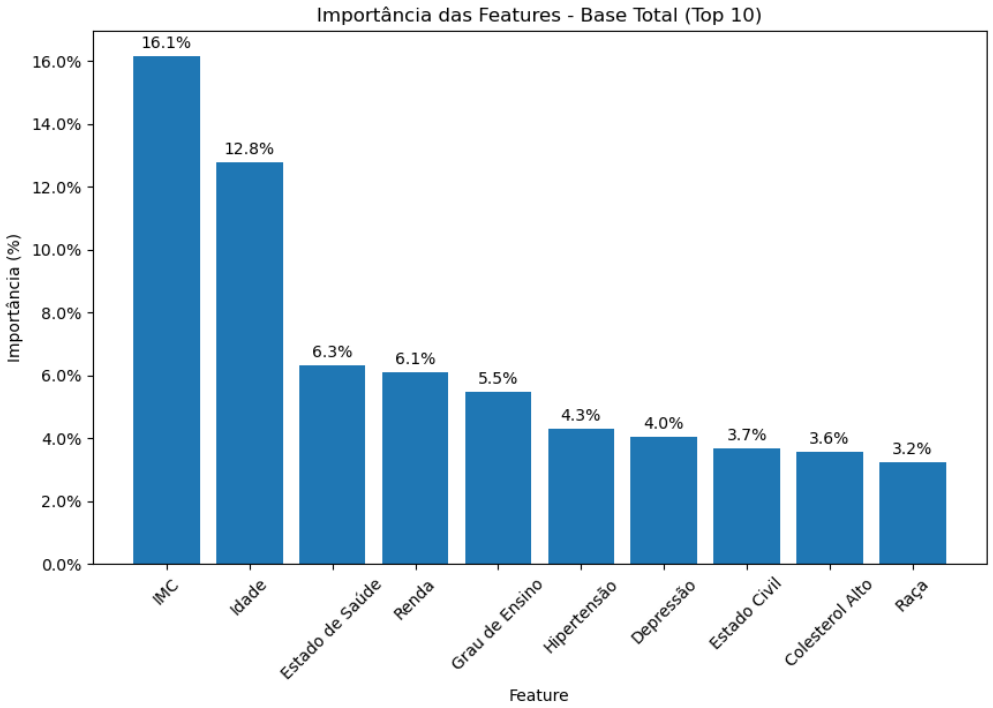
Fonte: Produzido pelos autores.

Esses resultados demonstram que o modelo apresentou desempenho sólido e consistente na tarefa de classificação. O equilíbrio entre as métricas indica que o classificador conseguiu minimizar tanto os falsos positivos quanto os falsos negativos, contribuindo para a confiabilidade das previsões. Assim, o modelo se mostra promissor como ferramenta de apoio à análise e à tomada de decisões em contextos de saúde pública.

5.3.3 Análise da importância das variáveis

Conforme ilustrado na Figura 17, o Índice de Massa Corporal (IMC) foi a variável com maior impacto no modelo, respondendo por 16,1% da importância total. Em seguida, destacaram-se a idade (12,8%), a renda domiciliar per capita (6,3%) e o grau de ensino (6,1%). Esses resultados evidenciam que aspectos fisiológicos e socioeconômicos exercem papel relevante na decisão de buscar atendimento em serviços de saúde.

Figura 17 – Top 10 variáveis mais importantes no Random Forest.



Fonte: Produzido pelos autores.

5.3.4 Análise da matriz de confusão

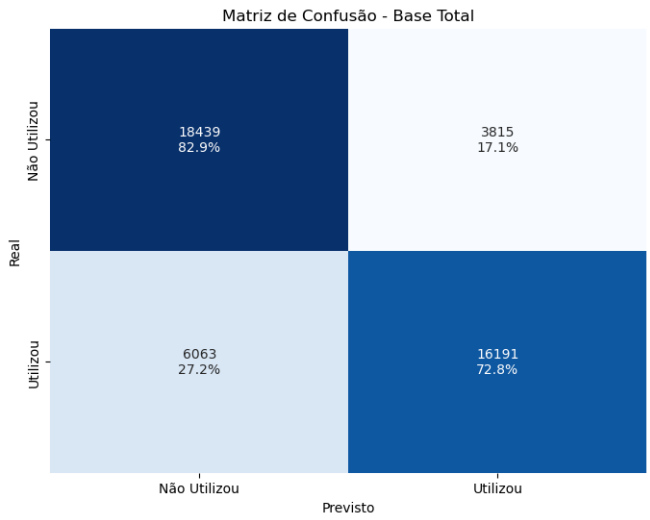
Conforme ilustrado na Figura 18, o modelo apresentou alta taxa de acerto para a classe dos indivíduos que não utilizaram os serviços de saúde, atingindo 82,9%. Para os que efetivamente utilizaram os serviços, a taxa de acerto foi de 72,8%, o que, embora inferior, ainda representa uma capacidade satisfatória de identificar corretamente os casos positivos.

Os resultados indicam que o modelo possui boa habilidade em classificar corretamente a maioria das observações, especialmente na classe majoritária. Embora a identificação da classe minoritária apresente desafios típicos de problemas desbalanceados, os índices alcançados reforçam a viabilidade prática do modelo para apoiar decisões baseadas em dados no contexto da saúde pública.

5.4 Classificação com Support Vector Machine (SVM)

Nesta etapa, foi aplicado o modelo *Support Vector Machine* (SVM), implementado por meio da classe `SVC` da biblioteca `scikit-learn`, com o objetivo de avaliar seu desempenho na predição do uso de serviços de saúde. A técnica de validação cruzada com 5 *folds* foi empregada, assegurando uma avaliação robusta e equilibrada do modelo. A seguir,

Figura 18 – Matriz de confusão Random Forest



Fonte: Produzido pelos autores.

estão listados os principais hiperparâmetros utilizados no `GridSearchCV`:

- **C**: parâmetro de regularização que controla o trade-off entre margem ampla e baixa taxa de erro (`0.1, 1, 10`).
- **kernel**: função que define o tipo de transformação aplicada aos dados (`poly, rbf`).
- **gamma**: coeficiente do kernel, aplicável aos tipos `rbf` e `poly` (`scale, auto`).
- **degree**: grau do polinômio quando o kernel `poly` é utilizado (`2, 3, 4`).

5.4.1 Análise do tempo de execução na validação cruzada

A Tabela 9 apresenta os tempos de execução registrados em cada *fold* durante a validação cruzada com o modelo *Support Vector Machine* (SVM), implementado por meio da classe `SVC`. Diferente dos demais classificadores, o modelo baseado em SVM apresentou um tempo de processamento consideravelmente maior, com média aproximada de 234 segundos por *fold*.

Esse comportamento é esperado, uma vez que a SVM depende da construção de uma matriz de similaridade entre os dados, cuja complexidade cresce quadraticamente com o número de amostras. Ainda assim, o modelo foi executado com sucesso em todas as divisões de validação, demonstrando viabilidade computacional mesmo com alto custo de tempo.

Tabela 9 – Tempo de execução na validação cruzada com SVM.

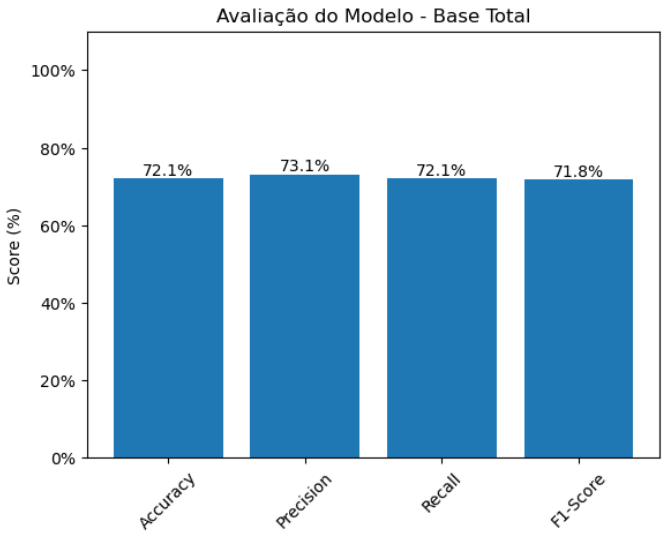
Fold	Tempo de Execução (s)	Tempo Acumulado (s)
1	232.81	232.81
2	234.60	467.41
3	231.44	698.85
4	235.01	933.86
5	235.79	1169.65

Fonte: Produzido pelos autores.

5.4.2 Análise das métricas

Conforme apresentado na Figura 19, o modelo baseado em *Support Vector Machine* (SVM), implementado com a classe `SVC`, obteve uma acurácia de 72,1%, indicando que o classificador acertou aproximadamente sete em cada dez previsões. A precisão foi de 73,1%, demonstrando que a maioria das classificações positivas realizadas pelo modelo estava correta. A revocação alcançou 72,1%, evidenciando a capacidade do modelo em identificar os indivíduos que utilizaram serviços de saúde. O F1-Score foi de 71,8%, representando um bom equilíbrio entre precisão e revocação.

Figura 19 – Avaliação do modelo SVM.



Fonte: Produzido pelos autores.

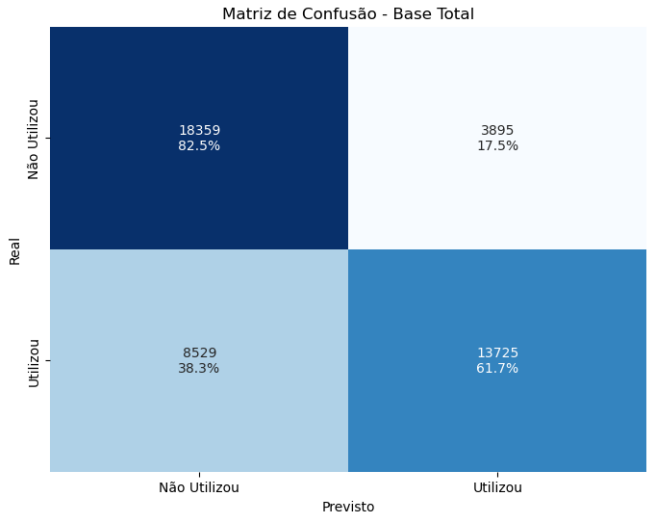
Os resultados obtidos demonstram que o modelo baseado em SVM apresentou desempenho consistente na tarefa de classificação binária, com métricas equilibradas e satisfatórias. Apesar do custo computacional mais elevado em comparação a outros modelos, o desempenho alcançado reforça sua aplicabilidade como uma ferramenta confiável para

previsão do uso de serviços de saúde, contribuindo para apoiar decisões baseadas em dados em contextos de saúde pública.

5.4.3 Análise da matriz de confusão

Conforme ilustrado na Figura 20, o modelo baseado em *Support Vector Machine* (SVM), implementado com a classe `SVC`, apresentou uma taxa de acerto de 82,5% na classificação dos indivíduos que não utilizaram os serviços de saúde. Para os que efetivamente utilizaram os serviços, a taxa de acerto foi de 61,7%, indicando um desempenho inferior na identificação da classe minoritária.

Figura 20 – Matriz de confusão SVC



Fonte: Produzido pelos autores.

Os resultados demonstram que o modelo baseado em SVM obteve desempenho satisfatório na identificação da classe majoritária, mas enfrentou limitações na detecção dos casos positivos. Essa diferença pode ser atribuída ao desbalanceamento das classes, característica comum em bases populacionais. Ainda assim, a técnica demonstrou capacidade de contribuir para a análise preditiva do uso de serviços de saúde, sendo possível aprimorar seu desempenho com ajustes adicionais e técnicas de balanceamento mais sofisticadas.

5.5 Classificação com Multi-Layer Perceptron (MLP)

Nesta etapa, foi aplicado o modelo de classificação *Multi-Layer Perceptron* (MLP), com o objetivo de avaliar seu desempenho na predição do uso de serviços de saúde. A técnica de validação cruzada com 5 *folds* foi empregada, assegurando uma avaliação robusta

e equilibrada do modelo. Abaixo estão listados os principais hiperparâmetros configurados para o modelo:

- **hidden_layer_sizes**: número de neurônios nas camadas ocultas (`(50,), (100,), (50, 50)`).
- **activation**: função de ativação nas camadas ocultas (`relu, tanh`).
- **solver**: algoritmo de otimização utilizado no treinamento (`adam, sgd`).
- **alpha**: termo de regularização L2 para evitar *overfitting* (`0.0001, 0.001`).
- **learning_rate**: estratégia de ajuste da taxa de aprendizado (`constant, adaptive`).
- **learning_rate_init**: taxa de aprendizado inicial (`0.001, 0.01`).

5.5.1 Análise do tempo de execução na validação cruzada

A Tabela 10 apresenta os tempos de execução registrados em cada *fold* durante o processo de validação cruzada aplicado ao modelo *Multi-Layer Perceptron (MLP)*. Observa-se uma variação significativa entre os *folds*, sendo o segundo *fold* o mais demorado, com 270,56 segundos.

O tempo médio por *fold* foi de aproximadamente 244 segundos, totalizando cerca de 20 minutos de execução para a validação completa. Essa duração reflete a complexidade computacional do MLP, que exige maior esforço de processamento para ajustar os pesos da rede neural ao longo de várias iterações. Apesar disso, os tempos observados foram compatíveis com o porte da base de dados utilizada, indicando viabilidade prática para aplicações similares.

Tabela 10 – Tempo de execução na validação cruzada MLP.

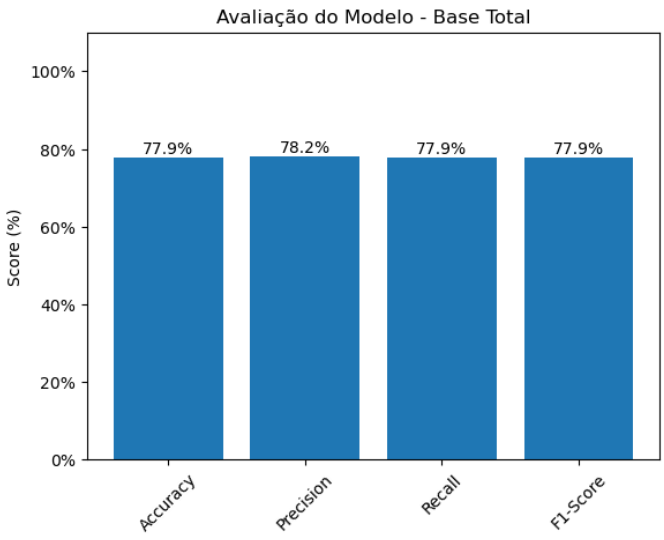
Fold	**Tempo de Execução (s)**	**Tempo Acumulado (s)**
1	260.14	260.14
2	270.56	530.70
3	261.21	791.91
4	215.47	1007.38
5	214.36	1221.74

Fonte: Produzido pelos autores.

5.5.2 Análise das métricas

Conforme apresentado na Figura 21, o modelo *Multi-Layer Perceptron (MLP)* obteve uma acurácia de 77,9%, indicando que acertou a maioria das classificações. A métrica de precisão foi de 78,2%, demonstrando que a maior parte das predições positivas feitas pelo modelo foi correta. O *Recall* atingiu 77,9%, evidenciando boa capacidade de identificação dos casos positivos reais. O *F1-Score* foi de 77,9%, reforçando o equilíbrio entre precisão e revocação.

Figura 21 – Avaliação do modelo MLP.



Fonte: Produzido pelos autores.

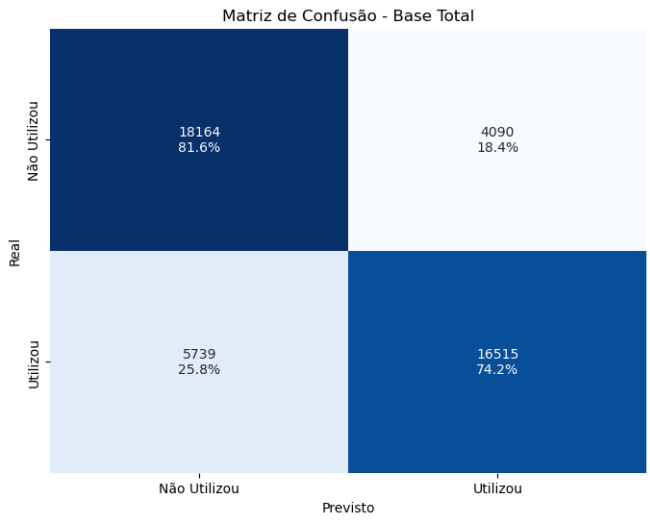
Os resultados demonstram que o classificador MLP apresentou desempenho estável e consistente, com métricas próximas às obtidas pelos demais modelos. A uniformidade entre as medidas reforça a capacidade do modelo de realizar previsões confiáveis, o que o torna uma alternativa viável para aplicações em saúde pública que envolvam a predição do uso de serviços de saúde.

5.5.3 Análise da matriz de confusão

Conforme ilustrado na Figura 22, o modelo apresentou uma taxa de acerto de 81,6% para a classe dos indivíduos que não utilizaram os serviços de saúde. Para os que utilizaram, o acerto foi de 74,2%, evidenciando boa capacidade do classificador em distinguir ambas as classes, mesmo diante do desafio de desbalanceamento entre elas.

Os resultados demonstram que o modelo MLP teve desempenho equilibrado, com boa identificação tanto da classe majoritária quanto da minoritária. Esse comportamento

Figura 22 – Matriz de confusão MLP



Fonte: Produzido pelos autores.

reforça sua aplicabilidade como ferramenta de apoio à previsão do uso de serviços de saúde, contribuindo para decisões mais assertivas no planejamento de políticas públicas.

5.6 Classificação com XGBoost

Nesta etapa, foi aplicado o modelo de classificação XGBoost (eXtreme Gradient Boosting), com o objetivo de avaliar seu desempenho na predição do uso de serviços de saúde. A técnica de validação cruzada com 5 *folds* foi empregada, assegurando uma avaliação robusta e equilibrada do modelo. A seguir, são apresentados os principais hiperparâmetros configurados para o modelo:

- **n_estimators**: número de árvores no modelo (`50, 100`).
- **max_depth**: profundidade máxima de cada árvore (`3, 5`).
- **learning_rate**: taxa de aprendizado que controla a contribuição de cada árvore (`0.01, 0.1`).
- **subsample**: fração de amostras utilizadas para o treinamento de cada árvore (`0.8, 1.0`).
- **colsample_bytree**: fração de variáveis consideradas em cada árvore (`0.8, 1.0`).
- **gamma**: redução mínima na função de perda exigida para uma divisão de nó (`0, 0.1, 0.5`).

- **reg_alpha**: regularização L1 para penalizar pesos elevados (`0, 0.1, 1`).
- **reg_lambda**: regularização L2 para evitar *overfitting* (`0, 0.1, 1`).

5.6.1 Análise do tempo de execução na validação cruzada

A Tabela 11 apresenta os tempos de execução registrados em cada *fold* durante a validação cruzada do classificador XGBoost. Observa-se que o tempo médio por *fold* foi de aproximadamente 3,6 segundos, totalizando 18,17 segundos ao final dos cinco ciclos de validação.

Esse desempenho reflete a alta eficiência computacional do XGBoost, mesmo considerando o volume expressivo de dados processados. A variação entre os *folds* foi mínima, indicando estabilidade no tempo de execução e confirmando a viabilidade do uso do algoritmo em contextos que exigem múltiplas execuções ou ajustes finos de parâmetros.

Tabela 11 – Tempo de execução na validação cruzada XGBoost.

Fold	**Tempo de Execução (s)**	**Tempo Acumulado (s)**
1	3.16	3.16
2	3.23	6.39
3	4.33	10.72
4	3.82	14.54
5	3.63	18.17

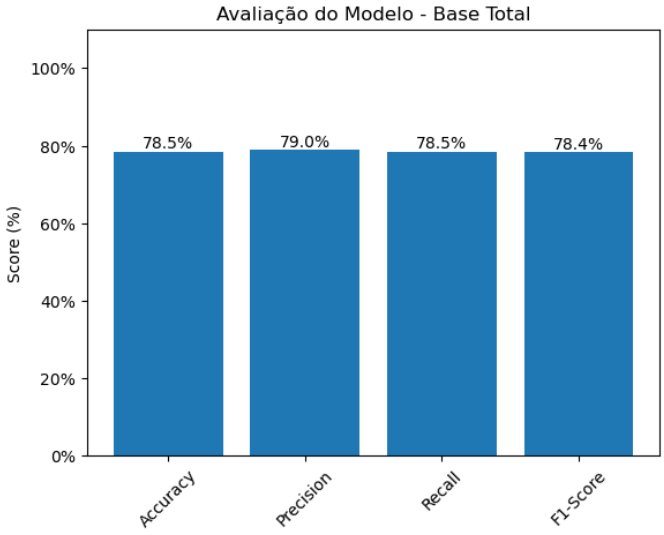
Fonte: Produzido pelos autores.

5.6.2 Análise das métricas

Conforme apresentado na Figura 23, o modelo XGBoost obteve uma acurácia de 78,5%, indicando que aproximadamente quatro em cada cinco previsões foram corretas. A métrica de precisão atingiu 79,0%, demonstrando que a maioria das classificações positivas realizadas pelo modelo estavam corretas. O *Recall* foi de 78,5%, evidenciando a boa capacidade do modelo em identificar corretamente os indivíduos que utilizaram serviços de saúde. Por fim, o *F1-Score*, que representa o equilíbrio entre precisão e revocação, alcançou 78,4%.

Os resultados obtidos demonstram que o modelo apresentou desempenho sólido e confiável na tarefa de classificação. O equilíbrio entre as métricas confirma a robustez do XGBoost para lidar com o problema proposto, mostrando-se eficaz na predição do uso dos serviços de saúde. O desempenho consistente reforça sua aplicabilidade como ferramenta de apoio à análise e à tomada de decisões em contextos de saúde pública.

Figura 23 – Avaliação do modelo XGBoost.



Fonte: Produzido pelos autores.

5.6.3 Análise da importância das variáveis

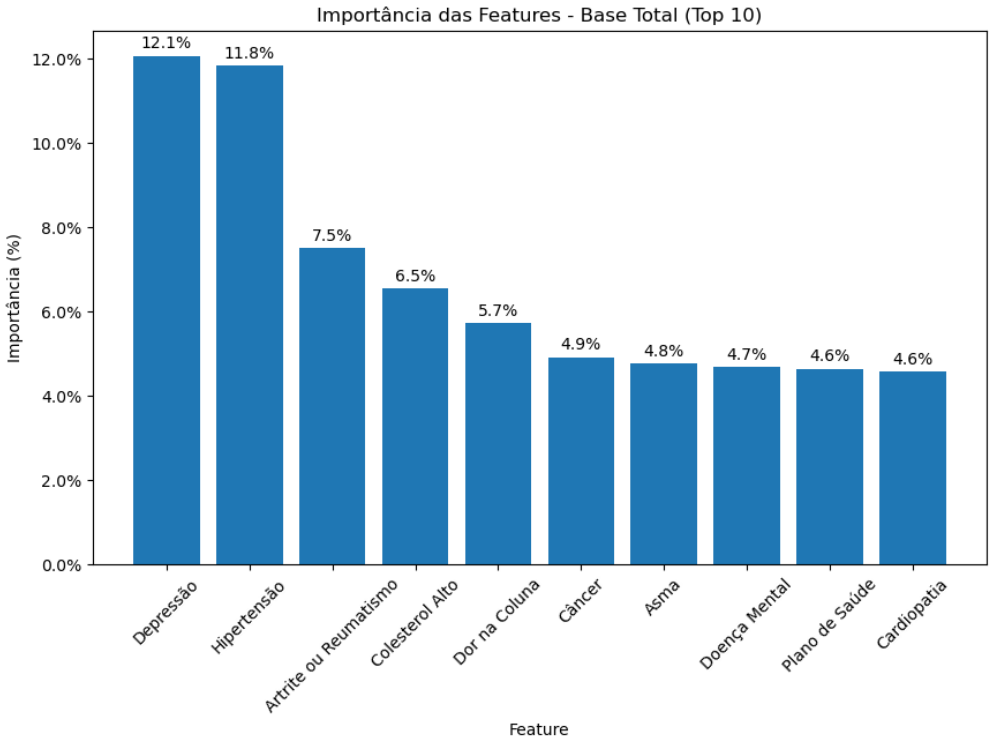
Conforme apresentado na Figura 24, as variáveis mais relevantes para o modelo XGBoost foram *Depressão* (12,1%), *Hipertensão* (11,8%) e *Artrite ou Reumatismo* (7,5%). Esses resultados indicam que condições crônicas e aspectos relacionados à saúde mental estão fortemente associados ao uso de serviços de saúde, desempenhando papel central na predição realizada pelo modelo.

5.6.4 Análise da matriz de confusão

Conforme ilustrado na Figura 25, o modelo XGBoost apresentou uma taxa de acerto de 85,0% para a classe dos indivíduos que não utilizaram os serviços de saúde. Para aqueles que efetivamente utilizaram os serviços, a taxa de acerto foi de 71,9%, o que representa uma performance sólida, ainda que inferior em comparação à classe majoritária.

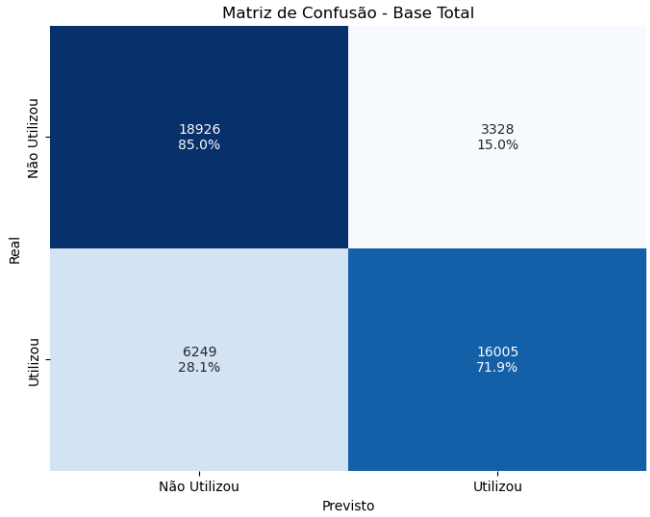
Os resultados evidenciam a efetividade do modelo em classificar corretamente a maioria das observações, com destaque para a classe majoritária. A taxa de falsos negativos, embora presente, manteve-se em patamar aceitável, indicando que o modelo possui boa capacidade discriminativa. Tais características reforçam o potencial do XGBoost como ferramenta auxiliar em estratégias preditivas na área da saúde pública, especialmente quando calibrado com atenção ao equilíbrio entre as classes.

Figura 24 – Top 10 variáveis mais importantes no XGBoost.



Fonte: Produzido pelos autores.

Figura 25 – Matriz de confusão XGBoost



Fonte: Produzido pelos autores.

6 Resultados e Discussão

6.1 Segmentação da população

A análise de agrupamento realizada com a técnica de *K-means*, combinada à redução de dimensionalidade via PCA, permitiu a identificação de seis grupos distintos de indivíduos com base em suas características clínicas, comportamentais e demográficas. Essa segmentação revelou padrões relevantes associados ao uso dos serviços de saúde e à percepção de saúde, conforme detalhado nas análises anteriores.

A Tabela 12 apresenta um resumo dos perfis identificados para cada *cluster*, destacando suas principais características e o percentual médio de indivíduos que utilizaram algum serviço de saúde.

Tabela 12 – Segmentação da população com base no uso dos serviços de saúde.

Cluster	**Perfil predominante**	**Uso do serviço de saúde**
0	Indivíduos com dor na coluna e hipertensão leve	15,1%
1	Idosos com múltiplas comorbidades crônicas	31,4%
2	Adultos com dor crônica e transtornos associados	25,7%
3	Indivíduos com sobrepeso e colesterol elevado	19,8%
4	Hipertensos obesos com percepção de saúde positiva	17,6%
5	Obesos com baixa prevalência de comorbidades	13,9%

Fonte: Elaborado pelos autores.

Os resultados revelam que o uso dos serviços de saúde está fortemente associado à presença de condições crônicas, como hipertensão, diabetes, colesterol alto e dor crônica, corroborando estudos prévios na literatura nacional e internacional (VOS et al., 2017; ORGANIZATION, 2019).

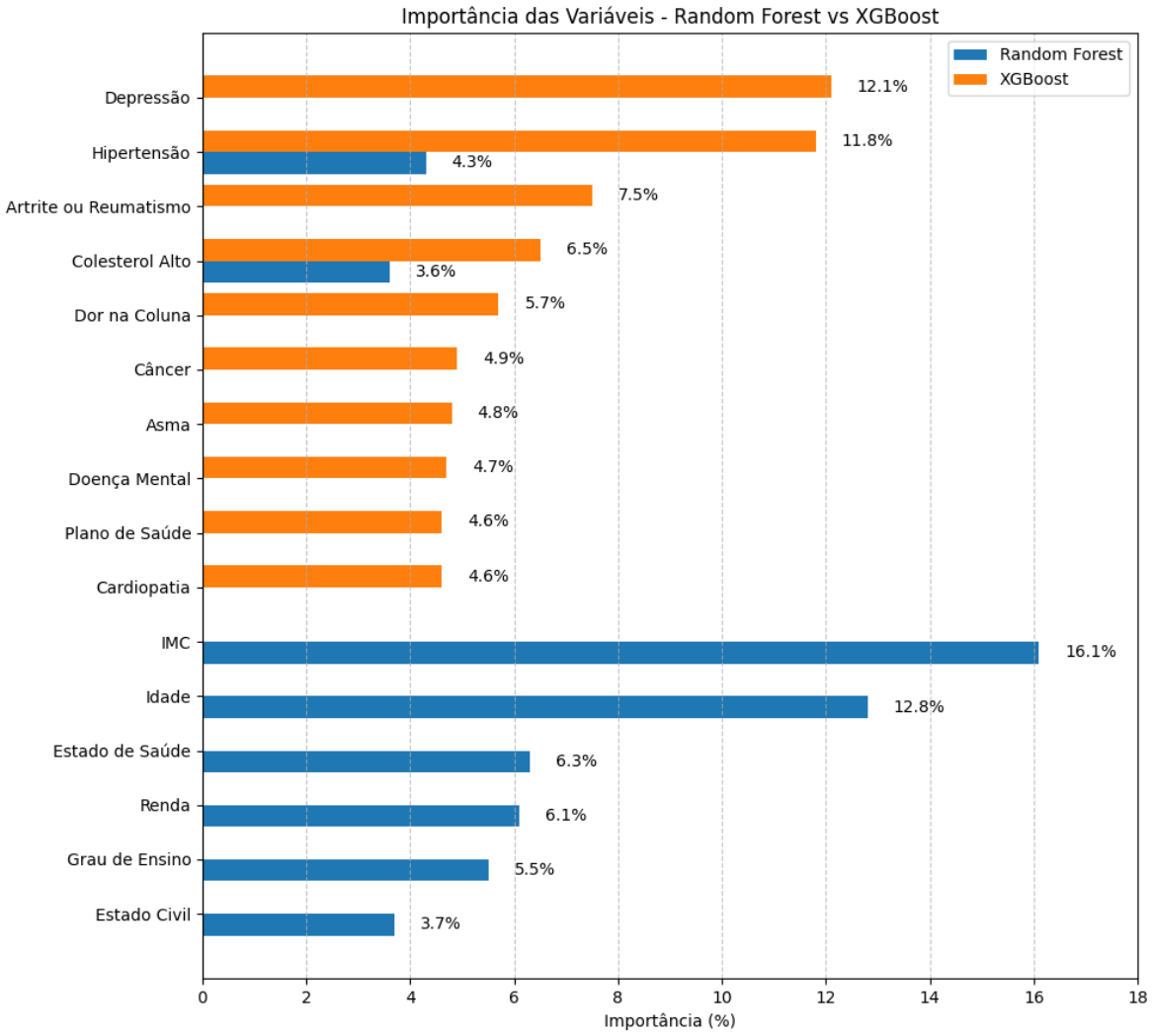
De modo geral, a segmentação obtida oferece subsídios valiosos para a formulação de estratégias de saúde pública mais específicas. Ao identificar perfis de risco, vulnerabilidade e comportamento, torna-se possível direcionar campanhas de prevenção, ações de monitoramento contínuo e melhorias no acesso aos serviços, com foco na integralidade do cuidado.

6.2 Análise de importância das variáveis

A análise de importância das variáveis foi realizada utilizando os modelos *Random Forest* e *XGBoost*, que permitem a extração dessas métricas. A Figura 26 apresenta as dez

variáveis mais relevantes.

Figura 26 – Importância das variáveis nos modelos Random Forest e XGBoost.



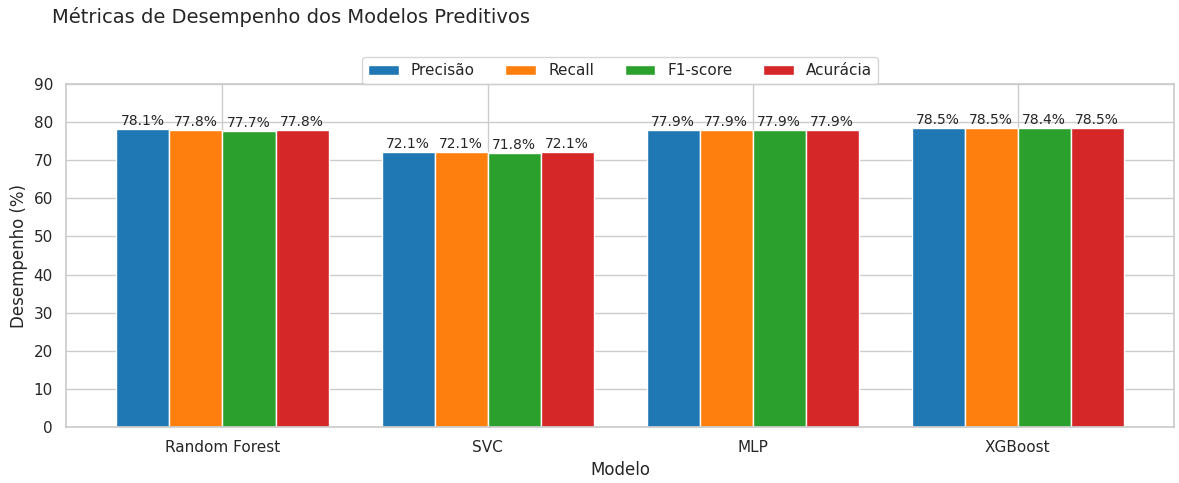
Fonte: Produzido pelos autores.

As variáveis mais relevantes diferem entre os modelos. O *XGBoost* destacou condições de saúde como **depressão e hipertensão**, fatores frequentemente associados ao maior uso de serviços médicos (MENDES; SANTOS, 2018; ROMERO; COSTA; MORAES, 2018). O *Random Forest*, por outro lado, priorizou características como **IMC e idade**, que são determinantes importantes em estudos epidemiológicos ((WHO), 2020).

6.3 Desempenho do modelo preditivo

Os modelos preditivos foram avaliados com base em métricas de desempenho. A Figura 27 apresenta os resultados.

Figura 27 – Métricas de desempenho dos modelos preditivos.



Fonte: Produzido pelos autores.

O *XGBoost* demonstrou o melhor desempenho geral, seguido de perto pelo *MLP*.

Observa-se que os modelos *Random Forest*, *MLP* e *XGBoost* apresentaram desempenhos bastante próximos, com métricas superiores a 77% em todas as avaliações. Esses resultados são consistentes com estudos recentes que indicam a superioridade desses algoritmos em tarefas de classificação (CHEN; GUESTRIN, 2016; LECUN; BENGIO; HINTON, 2015).

Um diferencial importante dos modelos *XGBoost* e *Random Forest* é a capacidade de fornecer medidas de importância das variáveis. Esse recurso é extremamente útil, pois permite interpretar quais fatores o modelo utilizou para realizar a predição. Por outro lado, embora a *MLP* não forneça diretamente essa interpretabilidade, ela apresentou o maior acerto para a classe de maior interesse do estudo — indivíduos que utilizaram os serviços de saúde — com acerto de 74,2% nessa categoria.

O modelo baseado em *Support Vector Machine* (SVM), implementado por meio da classe `SVC` do `scikit-learn`, apresentou desempenho inferior nas quatro métricas analisadas. Esse resultado pode estar associado à sensibilidade da técnica ao desbalanceamento da base, especialmente em algoritmos que se apoiam em margens de separação, como no caso da SVM. Apesar do uso de funções *kernel* para permitir classificações não lineares, o modelo demonstrou limitações frente aos demais algoritmos avaliados neste estudo.

6.4 Discussão geral

Os resultados obtidos reforçam a importância de fatores crônicos e comportamentais na predição do uso dos serviços de saúde. Como demonstrado, **idade, presença de**

doenças crônicas e estado de saúde autoavaliado foram determinantes significativos na previsão de demanda por serviços médicos. Esses achados corroboram estudos epidemiológicos anteriores (VOS et al., 2017; ORGANIZATION, 2019), que apontam que condições crônicas como hipertensão e obesidade são os principais preditores de uso frequente dos serviços de saúde.

A análise comparativa entre os modelos mostrou que *Random Forest*, MLP e XGB atingiram níveis similares de desempenho, cada qual com vantagens específicas. Em contraste, o modelo baseado em *Support Vector Machine* (SVM), implementado por meio da classe `SVC`, mostrou-se menos competitivo nas métricas avaliadas. Apesar disso, sua inclusão foi relevante para fins comparativos, evidenciando a importância da escolha adequada do algoritmo frente às particularidades da base de dados.

O uso de modelos baseados em aprendizado de máquina mostrou-se promissor na segmentação e previsão de padrões de utilização dos serviços, oferecendo suporte potencial para ações preventivas e alocação mais eficiente de recursos no **Sistema Único de Saúde (SUS)**.

7 Conclusão

A pesquisa realizada abordou o desafio da gestão eficiente dos recursos de saúde no Brasil, desenvolvendo um modelo preditivo para a identificação do uso dos serviços de saúde com base em características demográficas, socioeconômicas e clínicas da população brasileira. A partir da análise da Pesquisa Nacional de Saúde (PNS) de 2019, foram aplicadas técnicas de aprendizado de máquina e mineração de dados para construir um modelo robusto e interpretar padrões de utilização dos serviços médicos.

Os resultados obtidos demonstram que o modelo baseado no algoritmo *XGBoost* apresentou o melhor desempenho, superando outros métodos testados, como *Random Forest* e redes neurais. A validação cruzada indicou que o modelo é robusto, conseguindo generalizar suas previsões para diferentes subconjuntos de dados. Além disso, a análise de importância das variáveis evidenciou que fatores como idade, IMC, estado de saúde autodeclarado, presença de doenças crônicas (como hipertensão e diabetes) e acesso a planos de saúde são determinantes críticos no padrão de uso dos serviços médicos.

7.1 Impacto e aplicações

A capacidade de prever a demanda por serviços de saúde representa um avanço significativo na formulação de políticas públicas e na otimização da alocação de recursos. Com a implementação desse modelo preditivo, gestores de saúde poderão antecipar necessidades de atendimento, planejar melhor a distribuição de profissionais e infraestrutura, além de direcionar intervenções preventivas para populações vulneráveis.

Os resultados sugerem que grupos específicos, como idosos com doenças crônicas, indivíduos com obesidade e pessoas com menor acesso a planos de saúde, apresentam uma maior frequência de uso dos serviços médicos. Esse achado reforça a importância de políticas voltadas para esses segmentos, incluindo campanhas de prevenção, ampliação do acesso à atenção primária e monitoramento contínuo dessas populações.

Outro ponto relevante identificado foi a associação entre percepção de saúde e busca por atendimento. Indivíduos que autodeclararam sua saúde como "ruim"ou "muito ruim"demonstraram maior probabilidade de utilização dos serviços médicos, mesmo quando ajustado para outras variáveis clínicas e socioeconômicas. Isso ressalta a importância da subjetividade na experiência de saúde e sugere que intervenções psicológicas e educacionais podem desempenhar um papel fundamental na redução da carga sobre o sistema de saúde.

7.2 Limitações e trabalhos futuros

Apesar dos avanços obtidos, este estudo apresenta algumas limitações que devem ser abordadas em pesquisas futuras:

- **Representatividade da amostra**: A PNS 2019 oferece um panorama abrangente da saúde no Brasil, mas a sub-representação de determinados grupos populacionais pode introduzir viés nos resultados. Futuras pesquisas podem utilizar dados complementares para ampliar a diversidade da amostra.
- **Expansão da base de dados**: A inclusão de séries temporais permitiria avaliar tendências ao longo dos anos, melhorando a capacidade do modelo de prever mudanças na demanda por serviços de saúde.
- **Aprimoramento do modelo**: A implementação de técnicas de *deep learning* e modelos híbridos pode aumentar a precisão das previsões e identificar padrões mais complexos nos dados.
- **Desenvolvimento de uma interface para gestores**: A criação de um sistema interativo baseado no modelo preditivo permitiria que gestores e profissionais de saúde utilizassem as previsões em tempo real para tomada de decisões estratégicas.
- **Considerações éticas e interpretabilidade**: A transparência dos modelos preditivos deve ser priorizada, garantindo que as previsões sejam compreensíveis e justificáveis para evitar vieses indesejados na alocação de recursos.

Além disso, é essencial explorar técnicas de explicabilidade de modelos de aprendizado de máquina (*Explainable AI - XAI*) para permitir que profissionais da saúde e gestores compreendam as decisões do modelo, promovendo maior confiança em sua aplicação (MILLER, 2019).

7.3 Conclusão final

Os achados desta pesquisa evidenciam que modelos preditivos podem ser ferramentas valiosas para a gestão do sistema de saúde, possibilitando um uso mais eficiente dos recursos e promovendo um atendimento mais equitativo e acessível. A análise detalhada das variáveis e dos padrões de utilização dos serviços de saúde forneceu insights fundamentais para o aprimoramento das políticas públicas e para o desenvolvimento de estratégias que priorizem a prevenção e a intervenção precoce.

Recomenda-se que os resultados obtidos sejam considerados no desenvolvimento de políticas de saúde, especialmente no planejamento de estratégias voltadas para grupos de

risco identificados pelo modelo. Além disso, futuras colaborações entre pesquisadores, gestores e desenvolvedores de tecnologia poderão transformar esse modelo em uma ferramenta prática para aplicação no Sistema Único de Saúde (SUS).

Dessa forma, espera-se que este estudo contribua para a evolução do planejamento estratégico no setor da saúde, auxiliando na construção de um sistema mais sustentável e preparado para atender às crescentes demandas da população brasileira.

Referências

AL., C. da Costa-Luis et. *tqdm: A Fast, Extensible Progress Bar for Python and CLI*. 2023. <https://tqdm.github.io>. Acesso em 2024. Citado na página 27.

Anahp. *Gastos da saúde suplementar subiram R$ 83,6 bilhões em cinco anos*. 2019. Acesso em: 20 jan. 2025. Disponível em: <https://www.anahp.com.br/noticias/gastos-da-saude-suplementar-subiram-r-836-bilhoes-em-cinco-anos/>. Citado na página 15.

BARROS, A. J. D.; VICTORA, C. G. Measuring coverage in mnch: Monitoring equity in coverage of maternal, newborn, and child health interventions: A framework and indicators to improve equity measurement. *PLOS Medicine*, v. 10, n. 5, p. e1001390, 2019. Citado 4 vezes nas páginas 19, 28, 29 e 31.

BATES, D. W. et al. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, v. 33, n. 7, p. 1123–1131, 2014. Citado 2 vezes nas páginas 27 e 28.

BAXT, W. G. Application of artificial neural networks to clinical medicine. *The Lancet*, Elsevier, v. 346, n. 8983, p. 1135–1138, 1995. Citado na página 23.

BERNAL, O.; RESTREPO, J. Cluster analysis in health research: An overview and guidelines. *Revista Colombiana de Estadística*, v. 41, n. 2, p. 305–327, 2018. Citado na página 20.

BISONG, E. *Google Colaboratory*. Apress, Berkeley, CA, 2019. 59–64 p. Disponível em: <https://doi.org/10.1007/978-1-4842-4470-8_7>. Citado na página 25.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 22 e 32.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, p. 785–794, 2016. Citado 3 vezes nas páginas 23, 32 e 76.

CHEN, T.; GUESTRIN, C. *XGBoost: Scalable and Flexible Gradient Boosting*. 2023. <https://xgboost.readthedocs.io>. Acesso em 2024. Citado na página 27.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado 2 vezes nas páginas 22 e 32.

CSARDI, G.; NEPUSZ, T. *igraph: Network Analysis and Visualization*. 2023. <https://igraph.org/r/>. Acesso em 2024. Citado na página 26.

FAWCETT, T. An introduction to roc analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006. Citado na página 24.

FERRI, C.; HERNáNDEZ-ORALLO, J.; MODROIU, R. A. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, v. 30, n. 1, p. 27–38, 2009. Citado na página 25.

FOUNDATION, P. S. *itertools — Functions creating iterators for efficient looping*. 2023. <https://docs.python.org/3/library/itertools.html>. Acesso em 2024. Citado na página 26.

GARCIA, S.; LUENGO, J.; HERRERA, F. Data preprocessing in data mining. *Springer*, 2016. Citado na página 18.

GENG, Z.; LIU, H.; ZHANG, Y. A novel hybrid algorithm based on svm and k-means for health risk prediction. *Computers in Biology and Medicine*, Elsevier, v. 63, p. 160–172, 2015. Citado na página 22.

GOUTTE, C.; GAUSSIER, E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. *Proceedings of the 27th European Conference on Information Retrieval Research*, Springer, p. 345–359, 2005. Citado na página 24.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research*, v. 3, n. Mar, p. 1157–1182, 2003. Citado na página 18.

HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. *NetworkX: Python Software for the Creation, Manipulation, and Study of the Structure, Dynamics, and Functions of Complex Networks*. 2023. <https://networkx.org>. Acesso em 2024. Citado na página 26.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. [S.l.]: Elsevier, 2011. Citado 3 vezes nas páginas 18, 20 e 21.

HARRIS, C. R. e. a. *NumPy*. 2023. <https://numpy.org>. Acesso em 2024. Citado 2 vezes nas páginas 26 e 27.

HEARST, M. A. et al. Support vector machines. *IEEE Intelligent Systems and their Applications*, IEEE, v. 13, n. 4, p. 18–28, 1998. Citado na página 22.

HOSSEINI, M.; CHEN, D.; ATUN, R. Predicting healthcare utilization using machine learning: A systematic review. *BMJ Open*, v. 8, p. e020502, 2018. Citado 2 vezes nas páginas 21 e 27.

HUNTER, J. D. *Matplotlib: Python plotting*. 2023. <https://matplotlib.org>. Acesso em 2024. Citado 2 vezes nas páginas 26 e 27.

IBGE. Despesas com saúde em 2019 representam 9,6% do pib. *Agência IBGE Notícias*, 2022. Disponível em: <https://agenciadenoticias.ibge.gov.br/>. Citado na página 15.

INC., P. T. *Plotly for Python*. 2023. <https://plotly.com/python>. Acesso em 2024. Citado 2 vezes nas páginas 26 e 27.

Instituto Brasileiro de Geografia e Estatística (IBGE). *Pesquisa Nacional de Saúde 2019: Conceitos e Métodos*. Rio de Janeiro, Brasil, 2020. Disponível em: <https://www.ibge.gov.br>. Citado 3 vezes nas páginas 18, 28 e 30.

JOLLIFFE, I.; CADIMA, J. *Principal Component Analysis*. 2nd. ed. [S.l.]: Springer, 2016. Citado na página 20.

KODINARIYA, T. M.; MAKWANA, P. R. Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, v. 1, n. 6, p. 90–95, 2013. Citado na página 21.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, p. 436–444, 2015. Citado 4 vezes nas páginas 23, 29, 32 e 76.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R news*, Vienna, Austria: R Foundation for Statistical Computing, v. 2, n. 3, p. 18–22, 2002. Citado na página 22.

LIMA, I. *PNSIBGE: Importa e organiza os microdados da Pesquisa Nacional de Saúde (PNS)*. 2023. <https://github.com/iurilimajr/PNSIBGE>. Acesso em 2024. Citado na página 25.

LOPES, D. d. O.; OLIVEIRA, A. L. S.; MATOS, T. D. d. Agrupamento de usuários do sus com base em características clínicas e socioeconômicas. *Cadernos de Saúde Pública*, v. 35, n. 8, p. e00100418, 2019. Citado na página 20.

LUMLEY, T. *survey: analysis of complex survey samples*. 2023. <https://cran.r-project.org/package=survey>. Acesso em 2024. Citado na página 25.

MANI, I.; ZHANG, I. knn approach to unbalanced data distributions: a case study involving information extraction. In: *Proceedings of Workshop on Learning from Imbalanced Datasets, ICML*. [S.l.: s.n.], 2003. Citado 2 vezes nas páginas 19 e 31.

MENDES, E.; SANTOS, C. Fatores determinantes no uso de serviços de saúde no brasil. *Revista Brasileira de Epidemiologia*, v. 21, p. E180023, 2018. Citado na página 75.

MENDONçA, A. et al. Custos diretos da dor lombar em hospitais financiados pelo sistema Único de saúde. *Revista Pesquisa Fisioterapia*, v. 11, p. 181–189, 2021. Citado na página 28.

MICCI-BARRECA, D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explorations*, ACM, v. 3, n. 1, p. 27–32, 2001. Citado na página 19.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, Elsevier, v. 267, p. 1–38, 2019. Citado na página 79.

MUNDIAL, B. *World Development Indicators: Health Expenditure as % of GDP*. 2021. Banco Mundial - World Development Indicators. Acessado em: 05 fev. 2024. Disponível em: <https://databank.worldbank.org/source/world-development-indicators>. Citado na página 15.

NIELSEN, D. Gradient boosting machines: A tutorial. *arXiv preprint arXiv:1603.02754*, 2016. Disponível em: <https://arxiv.org/abs/1603.02754>. Citado na página 23.

ORGANIZATION, W. H. *World Health Organization: Health and aging statistics*. 2019. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>. Citado 2 vezes nas páginas 74 e 77.

PEDERSEN, T. L. *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. 2023. <https://ggplot2.tidyverse.org/reference/ggraph.html>. Acesso em 2024. Citado na página 26.

PEDREGOSA, F. e. a. *Scikit-learn: Machine Learning in Python*. 2023. <https://scikit-learn.org>. Acesso em 2024. Citado 2 vezes nas páginas 26 e 27.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2011. Citado na página 24.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2020. Citado na página 33.

RAHM, E.; DO, H. H. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, v. 23, n. 4, p. 3–13, 2000. Citado na página 18.

ROMERO, D.; COSTA, J.; MORAES, L. Prevalência, fatores associados e limitações relacionados ao problema crônico de coluna entre adultos e idosos no brasil. *Cadernos de Saúde Pública*, v. 34, p. e00012817, 2018. Citado 2 vezes nas páginas 28 e 75.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987. Citado 2 vezes nas páginas 21 e 32.

SAAR-TSECHANSKY, M.; PROVOST, F. Handling missing values when applying classification models. *Journal of Machine Learning Research*, v. 8, n. Jul, p. 1625–1657, 2007. Citado 3 vezes nas páginas 19, 31 e 39.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, v. 10, n. 3, p. e0118432, 2015. Citado na página 24.

SCHAUBERGER, P.; WALKER, A. *openxlsx: Read, Write and Edit XLSX Files*. 2023. <https://cran.r-project.org/package=openxlsx>. Acesso em 2024. Citado na página 25.

SHICKEL, B. et al. Deep learning in electronic health records: A systematic review. *Journal of the American Medical Informatics Association*, v. 25, p. 1419–1428, 2018. Citado 4 vezes nas páginas 19, 23, 29 e 31.

SIDEY-GIBBONS, J. A. M.; SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, v. 19, n. 1, p. 64, 2020. Citado na página 24.

SIMõES, T.; MEIRA, K.; SANTOS, J. dos. Prevalências de doenças crônicas e acesso aos serviços de saúde no brasil: evidências de três inquéritos domiciliares. *Ciência & Saúde Coletiva*, v. 26, n. 9, p. 3991, 2021. Citado 2 vezes nas páginas 15 e 28.

SLOWIKOWSKI, K. *ggrepel: Automatically Position Non-Overlapping Text Labels with ggplot2*. 2023. <https://cran.r-project.org/package=ggrepel>. Acesso em 2024. Citado na página 26.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. [S.l.]: Pearson Education, 2013. Citado 2 vezes nas páginas 19 e 32.

TEAM, T. P. D. *pandas: Python Data Analysis Library*. 2023. <https://pandas.pydata.org>. Acesso em 2024. Citado 2 vezes nas páginas 26 e 27.

TOPOL, E. *Deep medicine: How artificial intelligence can make healthcare human again*. [S.l.]: Basic Books, 2019. Citado 4 vezes nas páginas 21, 22, 28 e 29.

VOS, T. et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the global burden of disease study 2016. *Lancet*, v. 390, p. 1211–1259, 2017. Citado 4 vezes nas páginas 16, 28, 74 e 77.

WASKOM, M. *Seaborn: Statistical Data Visualization*. 2023. <https://seaborn.pydata.org>. Acesso em 2024. Citado 2 vezes nas páginas 26 e 27.

WEI, T.; SIMKO, V. *corrplot: Visualization of a Correlation Matrix*. 2023. <https://cran.r-project.org/package=corrplot>. Acesso em 2024. Citado na página 25.

(WHO), W. H. O. Obesity and overweight: Fact sheet. *World Health Organization*, 2020. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>. Citado na página 75.

WICKHAM, H. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. 2023. <https://ggplot2.tidyverse.org>. Acesso em 2024. Citado na página 25.

WICKHAM, H. *reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package*. 2023. <https://cran.r-project.org/package=reshape2>. Acesso em 2024. Citado na página 26.

WICKHAM, H.; FRANçOIS, R. *dplyr: A Grammar of Data Manipulation*. 2023. <https://cran.r-project.org/package=dplyr>. Acesso em 2024. Citado na página 25.

World Health Organization. *Obesity: preventing and managing the global epidemic*. 2000. Acesso em: 26 mar. 2025. Disponível em: <https://apps.who.int/iris/handle/10665/42330>. Citado 2 vezes nas páginas 19 e 31.

ZHAO, Y.; ATUN, R.; ANINDYA, K. Medical costs and out-of-pocket expenditures associated with multimorbidity in china: quantile regression analysis. *BMJ Global Health*, v. 6, p. e004042, 2021. Citado 2 vezes nas páginas 16 e 28.

Apêndices

Materiais elaborados pelo autor

Apêndice A — Notebook de Coleta e Pré-Processamento de Dados

Este notebook contém as rotinas desenvolvidas em linguagem Python e R para o tratamento e preparação dos dados da PNS 2019. As etapas contemplam:

- Importação e organização dos microdados da PNS com uso de pacotes estatísticos;
- Seleção e transformação de variáveis categóricas e contínuas;
- Cálculo do índice de massa corporal (IMC), categorização de idade e renda;
- Tratamento de valores ausentes e aplicação da técnica de *Target Encoding*;
- Balanceamento da base com o algoritmo *NearMiss-1*;
- Geração das bases binária e categórica para uso em agrupamento e classificação.

Apêndice B — Notebook de Agrupamento

Este notebook documenta a execução das técnicas de agrupamento dos dados de saúde pública com base nas variáveis clínicas. As rotinas incluem:

- Aplicação de *Principal Component Analysis* (PCA) para redução de dimensionalidade;
- Utilização do método do cotovelo e do índice de silhueta para escolha do número ideal de clusters;
- Aplicação do algoritmo *K-Means* para segmentação dos dados;
- Geração de gráficos de radar e mapa de calor para visualização dos perfis identificados;
- Análise qualitativa dos clusters com base em doenças crônicas e percepção de saúde.

Apêndice C — Notebook de Classificação

Este notebook apresenta as rotinas implementadas para o treinamento e avaliação dos modelos preditivos. As atividades documentadas envolvem:

- Separação das bases em treino, validação e teste com *Stratified K-Fold*;

- Treinamento dos algoritmos Random Forest, SVC, MLP e XGBoost;
- Otimização de hiperparâmetros com *GridSearchCV*;
- Avaliação de desempenho dos modelos com métricas como acurácia, precisão, recall e F1-score;
- Análise da importância das variáveis com *Permutation Importance*;
- Geração de matrizes de confusão e gráficos explicativos dos resultados.

Anexos

Outros materiais

Anexo A — Questionário da Pesquisa Nacional de Saúde (PNS) 2019

Este anexo apresenta o questionário original utilizado na Pesquisa Nacional de Saúde (PNS) 2019, disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). O questionário serviu como base primária para a estruturação dos dados utilizados neste trabalho, guiando a seleção de variáveis relacionadas às condições de saúde, hábitos de vida, acesso e uso de serviços de saúde pela população brasileira.

Fonte: IBGE, *Questionário da Pesquisa Nacional de Saúde 2019*.

Anexo B — Dicionário de Indicadores da PNS

Este material contém o dicionário de variáveis da base de microdados da Pesquisa Nacional de Saúde 2019, disponibilizado em formato Excel. O arquivo foi utilizado para interpretar os códigos das variáveis presentes na base de dados, possibilitando a correta identificação e transformação das informações originais para fins de análise estatística e modelagem preditiva.

Fonte: IBGE, *Dicionário de Variáveis da PNS 2019*.